Spring 2010

Local effects of limited recombination in Drosophila

Anna Ouzounian Williford
University of Iowa

LOCAL EFFECTS OF LIMITED RECOMBINATION IN DROSOPHILA

by

Anna Ouzounian Williford

An Abstract

Of a thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Biology
in the Graduate College of
The University of Iowa

May 2010

Thesis Supervisor: Associate Professor Josep M. Comeron

ABSTRACT

Recent years have witnessed the integration of theoretical advances in population genetics with large-scale analyses of complete genomes. As a result, a growing number of studies suggest the frequent occurrence of deleterious as well as adaptive mutations. Given the evidence for the widespread occurrence of selection, the finite sizes of natural populations, and the limited recombination in every genome, mutations under selection are expected to alter the fate of genetically linked mutations. The consequences of this non-independent behavior of mutations can be described by the Hill-Robertson effect in terms of the reduction in the effective population size (N_e). Reduction in the effective population size has two effects: 1) a reduction in levels of genetic variation and 2) a reduction in the effectiveness of selection that is manifested in an increased probability of fixation of deleterious mutations and a reduced probability of fixation of advantageous mutations. Changes in N_e that have previously been frequently associated with changes in recombination rate can also occur locally, in association with changes in the number of sites under selection even when the recombination rate remains uniform. The main objective of the work presented in this thesis is to investigate these local effects of the non-independent behavior of mutations on patterns of polymorphism and divergence in *Drosophila* using computer simulation and experimental approaches.

A computer simulation approach is developed to investigate the local consequences of linked selection on estimates of selection and the proportion of adaptive substitutions using the McDonald-Kreitman framework. The results suggest that even a high level of recombination is unlikely to remove all the effects of linked selection.

Ignoring these local linkage effects leads to misleading estimates of the intensity of selection and the proportion of adaptive substitutions.

Two predictions of the Hill-Robertson effect were tested empirically by examining patterns of polymorphism and divergence combined with codon bias estimates in genes with and without introns: 1) the effectiveness of selection and polymorphism levels are expected to be reduced in the center of the long coding sequence of genes without introns (the intragenic Hill-Robertson effect), and 2) introns are expected to function as modifiers of recombination thereby increasing the effectiveness of selection in the central region of the coding sequence of genes containing centrally located introns. The evidence from divergence and codon bias patterns in genes with a long coding sequence supports the presence of the intragenic Hill-Robertson effect. However, polymorphism levels do not show the expected decrease in the center of the coding sequence. With regard to the second prediction, results indicate that intron presence does not increase the effectiveness of selection at synonymous sites in the set of investigated genes. Rather, intron presence is associated with increased levels of adaptation at nonsynonymous sites. Further investigations are necessary to clarify the role of introns in mediating the increase in adaptation.

Abstract Approved: _____
Thesis Supervisor

_____
Title and Department

_____
Date

LOCAL EFFECTS OF LIMITED RECOMBINATION IN DROSOPHILA

by

Anna Ouzounian Williford

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Biology
in the Graduate College of
The University of Iowa

May 2010

Thesis Supervisor: Associate Professor Josep M. Comeron

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Anna Ouzounian Williford

has been approved by the Examining Committee
for the thesis requirement for the Doctor of Philosophy
degree in Biology at the May 2010 graduation.

Thesis Committee:
Josep M. Comeron, Thesis Supervisor

Ana Llopart

Bryant McAllister

John Manak

Terry Braun

To my families

ACKNOWLEDGMENTS

I am grateful to my first mentor, Dr. Barbara Stay, for her continuous support and care during my many years of study at The University of Iowa. I am in debt to Dr. Joseph Frankel who taught my first biology courses at The University of Iowa with an enthusiasm and clarity so much needed for a beginning student.

I am thankful to my advisor, Dr.Josep Comeron, from whom I learned so much about one of the most interesting areas in biology. If it were not for him, I would not have been able to appreciate, enjoy, and struggle with the rigor and the complexity of population genetics.

I thank Dr. Ana Llopart for enormous help with my research. She obtained most of the data analyzed in Chapter IV. The efficiency with which she works and manages a lab is something I have yet to learn how to do.

Special thanks go to my fellow graduate students, Ramesh Ratnappan and Derek Peters. I am convinced that very few people have had a similar experience with their lab mates. They are fantastic guys.

I would like to thank people outside the lab whose help I appreciate very much. Gerry Hehman was always willing to provide all the help I needed with sequencing and beyond. Matt Brockman and his team helped me multiple times with all sorts of computer issues. Phil Ecklund took care of all the paperwork and deadlines: without his constant reminders none of us would have made it to this stage.

My work would not be as stimulating if I did not have people to talk to about it. For their willingness to listen, question me, and give advice, I thank Gregory Landini and David Rudrauf.

Finally, I thank those who matter most to me: Kenny, Lydia, and Peter.

ABSTRACT

Recent years have witnessed the integration of theoretical advances in population genetics with large-scale analyses of complete genomes. As a result, a growing number of studies suggest the frequent occurrence of deleterious as well as adaptive mutations. Given the evidence for the widespread occurrence of selection, the finite sizes of natural populations, and the limited recombination in every genome, mutations under selection are expected to alter the fate of genetically linked mutations. The consequences of this non-independent behavior of mutations can be described by the Hill-Robertson effect in terms of the reduction in the effective population size (N_e). Reduction in the effective population size has two effects: 1) a reduction in levels of genetic variation and 2) a reduction in the effectiveness of selection that is manifested in an increased probability of fixation of deleterious mutations and a reduced probability of fixation of advantageous mutations. Changes in N_e that have previously been frequently associated with changes in recombination rate can also occur locally, in association with changes in the number of sites under selection even when the recombination rate remains uniform. The main objective of the work presented in this thesis is to investigate these local effects of the non-independent behavior of mutations on patterns of polymorphism and divergence in *Drosophila* using computer simulation and experimental approaches.

A computer simulation approach is developed to investigate the local consequences of linked selection on estimates of selection and the proportion of adaptive substitutions using the McDonald-Kreitman framework. The results suggest that even a high level of recombination is unlikely to remove all the effects of linked selection. Ignoring these local linkage effects leads to misleading estimates of the intensity of selection and the proportion of adaptive substitutions.

Two predictions of the Hill-Robertson effect were tested empirically by examining patterns of polymorphism and divergence combined with codon bias estimates

in genes with and without introns: 1) the effectiveness of selection and polymorphism levels are expected to be reduced in the center of the long coding sequence of genes without introns (the intragenic Hill-Robertson effect), and 2) introns are expected to function as modifiers of recombination thereby increasing the effectiveness of selection in the central region of the coding sequence of genes containing centrally located introns. The evidence from divergence and codon bias patterns in genes with a long coding sequence supports the presence of the intragenic Hill-Robertson effect. However, polymorphism levels do not show the expected decrease in the center of the coding sequence. With regard to the second prediction, results indicate that intron presence does not increase the effectiveness of selection at synonymous sites in the set of investigated genes. Rather, intron presence is associated with increased levels of adaptation at nonsynonymous sites. Further investigations are necessary to clarify the role of introns in mediating the increase in adaptation.

.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The main objective of the work presented here is to investigate local effects of non-independent behavior of mutations caused by the presence of selection and limited recombination (or linkage) in finite populations. The consequences of non-independent behavior of mutations can be described by the Hill-Robertson effect in terms of the reduction in the effective population size below the effective population size that characterizes a species as a whole. A reduction in the effective population size predicts a reduction in levels of genetic variation and a reduction in the effectiveness of selection. These two aspects of the mutational dynamics (polymorphism levels and fixation rates) are examined from different angles, using both computer simulation and experimental approaches. The Hill-Robertson effect is thus the unifying theme of the thesis.

Chapter I provides a brief overview of the developments in population genetics during 20^{th} century. I attempt to trace an evolution of ideas that can be broadly subdivided into three phases: 1) a series of fundamental theoretical investigations emphasizing the role of Darwinian selection, 2) the advance of the neutral theory that provided clear expectations for the behavior of mutations in the absence of selection thereby allowing the development of rigorous tests of neutralist and selectionist views of evolution, and 3) a shift towards the emphasis on the importance of natural (and frequently Darwinian) selection based on the empirical evidence obtained from a large amount of sequence data. More significantly, the influence of mutations under selection on the behavior of mutations at linked sites has been fully recognized in recent years; one

of the major tasks of the current research is to understand the extent to which interaction between linkage and selection governs evolutionary change.

Chapter II presents a computer simulation approach that is used as a tool for understanding the consequences of linkage in the presence of selection. I provide the description of the program and the model used to simulate sequence evolution. The computer simulation approach developed in this chapter is applied to investigate how the length of the coding sequence, recombination rates, and selection intensity influence the strength of the Hill-Robertson effect generated by weak selection.

Chapter III utilizes the computer simulation approach developed in Chapter II to investigate the local consequences of linked selection on estimates of selection and the proportion of adaptive substitutions using the McDonald-Kreitman framework. I show that even in areas of high recombination, the effects of linked selection on neutral sites are almost unavoidable. Consequently, estimates of selection and the proportion of adaptive substitutions might be compromised by employing methods that do not take into account local linkage effects.

Chapter IV presents the results of the empirical study designed to test the main predictions of the local Hill-Robertson effect. When the number of sites under selection is large, the effectiveness of selection is expected to vary locally, being reduced in the center of the coding region. Moreover, intron presence is expected to restore the effectiveness of selection in the central region by increasing physical distance between selected mutations, thereby effectively increasing recombination rates. These predictions are tested by examining patterns of polymorphism and divergence and estimates of codon bias in genes with and without introns. The results indicate that the effectiveness of

selection does vary across the coding sequence, supporting the intragenic Hill-Robertson effect. However, intron presence does not increase the effectiveness of selection at synonymous sites in the set of genes investigated. Rather, intron presence is associated with increased levels of adaptation at non-synonymous sites. This unexpected but interesting result should be further investigated with a larger data set.

To conclude the above summary, the common topic of the chapters presented in this thesis is the evaluation of *local* changes in the effectiveness of selection and polymorphism levels associated not with changes in recombination rates as in many previous studies but with changes in the number of linked sites under selection.

CHAPTER I

AN OVERVIEW OF THE PROGRESS IN POPULATION GENETICS

1.1 Population genetics: an historical note

The establishment of population genetics as a discipline was the result of long-lasting controversies concerning the nature of variation, principles of heredity, and the importance of natural selection that culminated in the mathematical account of evolution through natural selection in terms of Mendelian genetics (Provine 1971/2001). Beginning in the early 20th century, population genetics sought to quantify the process of evolution based on the presence of variation (provided by the input of mutations) and natural selection (provided by the differential fitness of genotypes). The main preoccupation of those times that has persisted up to the present is accounting for two major aspects of evolution: 1) the maintenance of variation within species upon which selection can operate and 2) the transformation of variation within species into variation between species in order to account for the continuous change in genetic composition, or evolutionary change itself (Ewens 2004). The clear formulation of these goals was preceded by years of bitter debates initiated immediately after the publication of *The Origin of Species*. What follows is a very brief summary of these debates based on the much more detailed account of the controversies given by Provine (1971/2001).

Neither the principles of inheritance nor the nature of variation was understood at that time. Nevertheless, the existence of variation was undeniable, but it appeared in two seemingly fundamentally different forms: there were small differences between individuals in continuously varying characters, and there was much more drastic, but less

frequently observed, discontinuous variation. Darwin envisioned evolution to be a gradual process, driven by natural selection acting on small continuous variations. He emphasized large amounts of variation that appeared each generation: if variation occurred at such a high rate, species would be unable to maintain their unique characteristics over long periods of time. To insure the preservation of a species as an identifiable unit over generations, Darwin adopted the blending theory of inheritance as a mechanism that accounted for species uniformity. At the same time, blending inheritance imposed a serious difficulty on the theory of natural selection: blending of variation would result in the removal of individual differences upon which selection was supposed to act. In opposition to Darwin with respect to the nature of variation stood his most avid supporters, Thomas Huxley and Francis Galton (and later William Bateson) who insisted that natural selection operates on discontinuous variation, or "sports". According to this view, evolution proceeds rapidly by mutational leaps as opposed to the slow and gradual accumulation of small differences, as advocated by Darwin and his followers, W. F.R. Weldon and Karl Pearson. After the rediscovery of Mendel's work, the controversy only intensified. Bateson argued that small differences have selective effects too small for natural selection to be effective and that the "swamping effect of intercrossing" would remove all the variation upon which selection could act. Discontinuous variation, on the other hand, had a large selective value and was not subject to blending inheritance, as Mendelian principles of heredity demonstrated. That is, Mendelian laws were thought to apply exclusively to discontinuous variation, and thus their rediscovery only alienated the gradualists even further. It was not until 1918, when Ronald Fisher demonstrated that continuously varying characters can be determined by a large number of Mendelian

factors, that the fundamental distinction between continuous and discontinuous variation disappeared, and Darwinian selection could be analyzed in terms of Mendelian principles of inheritance (Provine 1971/2001).

Fisher, Wright and Haldane are mainly credited for the development of population genetics. During the 1920's and 1930's their work focused on the evolutionary consequences of Mendelian genetics including the analysis of inbreeding, of the evolution of dominance, and of the effects of selection, as well as migration and random genetic drift, on changes in allele frequencies. The three worked independently and developed somewhat different views of evolution. Fisher believed evolution proceeded by the successive substitution of single mutations, and he emphasized the primary role of selection in large populations because selection in such populations can be very effective on mutations with even a very small selective advantage. Haldane generally agreed with Fisher, although he stressed the importance of mutations with large effects. Wright's view of evolution was based on interactions between population size, random genetic drift, migration, and selection acting not on individual genes but on interacting systems of genes or gene complexes (Provine 1971/2001; Ewens 2004).

Despite these major theoretical advances during the first half of 20th century that focused on the importance of Darwinian selection as the driving force of evolutionary change, the first molecular data obtained in the 1960s (Zuckerkandl and Pauling 1965; Harris 1966; Lewontin and Hubby 1966) stimulated the proposal of non-selective factors as the major evolutionary force governing the levels of variation within and between species.

1.2 The neutral and nearly neutral theories of molecular evolution

1.2.1 The neutral theory of molecular evolution

The neutral theory or the 'mutation-random genetic drift theory' of molecular evolution in its original form asserts that most evolutionary changes result from the fixation of neutral or nearly neutral mutations whose dynamics are governed by random genetic drift (Kimura 1983). It does not deny the role of Darwinian selection in adaptive evolution, but it proposes that adaptive substitutions are so rare that they contribute very little to the total number of substitutions (or the rate of molecular evolution). Likewise, most variation within species is provided by segregating neutral mutations. Beneficial mutations are assumed to be rare and thus do not contribute to variation within species. Deleterious mutations are compatible with the neutral theory because they are removed by purifying selection and do not contribute either to polymorphism within species or to divergence between species. Moreover, occurrence of either deleterious or beneficial mutations is not expected to affect levels of neutral variation. This means that a thorough understanding of the behavior of neutral mutations alone should provide an adequate description of the process of molecular evolution.

The proposal of the neutral theory was provoked mainly by two observations. First, comparative studies of protein sequences in mammals by Zuckerkandl and Pauling (1965) demonstrated high and constant rates of amino acid substitutions. Second, analysis of allozyme data in *Drosophila* (Lewontin and Hubby 1966) and human (Harris 1966) populations indicated the presence of high levels of genetic variation within these species. The initial argument presented by Kimura (1968) is based on the first

observation and the theoretical paper concerning substitutional load (Haldane 1957). Using estimates of the rate of amino acid substitution obtained by Zuckerkandl and Pauling, Kimura estimates the rate of nucleotide substitution to be about 1 substitution every 2 years or every 0.5 generations (assuming a generation time of 4 years in mammals), an estimate that far exceeds the number of beneficial substitutions that can be tolerated by a population according to the calculations of Haldane (1 substitution every 300 generations). The high observed substitution rate, Kimura argues, indicates that most substitutions must be neutral to avoid the creation of a substitutional load (Kimura 1968).

I will now discuss the quantitative features of molecular evolution, focusing on the main factors controlling the behavior of neutral mutations and mutations under selection in diploid populations. It is assumed that the mutation rate is low enough that each mutation occurs in a new site. This is known as the infinite site model (Kimura 1971).

In general, the rate of substitution per generation (k) is equal to the number of new mutations produced each generation ($2N\mu$) multiplied by the probability of their fixation (P_f). Kimura showed that if the majority of substitutions result from the fixation of neutral mutations, the rate of molecular evolution is limited only by the mutation rate and is independent of the population size (Kimura 1983). The probability of fixation of a neutral mutation is equal to its frequency in a population and for a new mutation, $P_f = 1/(2N)$ (where N is the number of diploid individuals in a population). The rate of substitution then is given by:

$$k = 2N\mu \cdot \frac{1}{2N} = \mu \qquad \text{(Equation 1)}$$

In this equation, μ is the mutation rate expressed as the number of mutations per generation per haploid genome; 2*N* is the number of genomes in a population, and k is the substitution rate in terms of the number of substitutions per generation. That is, $2N\mu$ new mutations appear each generation of which a fraction equal to 1/(2*N*) will reach fixation (Kimura 1971; Kimura 1983).

In contrast, the probability of the fixation of a mutation with a selection coefficient *s* (either beneficial or deleterious), assuming semidominance, is given by:

$$P_f = \frac{1-e^{-Sp}}{1-e^{-S}} \qquad \text{(Equation 2)}$$

where $S=4N_e s$ and p is the initial frequency of a mutation. For mutations with definite selective advantage, such that $4N_e s >> 1$, the probability of fixation is approximated by $2sN_e/N$ and the rate of substitution of beneficial mutations is expressed as:

$$k = 2Nv \cdot 2s\frac{N_e}{N} = 4N_e sv \qquad \text{(Equation 3)}$$

where v is the rate of beneficial mutations (Kimura 1971). This means that if the rate of molecular evolution is largely determined by the fixation of beneficial mutations, the observed constancy of the evolutionary rate implies that the product of svN_e remains constant for various lineages independently of the environment. Kimura argues that this situation is highly unlikely: the clock-like nature of amino acid substitutions would be a surprising property of advantageous mutations since they are expected to be fixed more frequently during episodes of environmental change. And if phenotypic and molecular evolution are driven by the same force (selection), it is not clear why the rate of molecular evolution is constant while the rate of phenotypic evolution is not (Kimura 1983).

Further support for the neutral theory came from the survey of the distribution of variable amino acids within proteins. King and Jukes (1969) showed that functionally important regions of proteins (determined on the basis of their biochemical and enzymatic properties) evolve slower than functionally less constrained regions. Furthermore, early molecular studies revealed that synonymous nucleotide changes occur at higher rates than non-synonymous nucleotide changes, that introns evolve at rates similar to or higher than synonymous sites, and that the rates of evolution at synonymous and non-synonymous sites in pseudogenes are similar to each other (Kimura 1977; Li *et al*., 1981; Kimura 1986). These observations provided the basis for the general rule that "…molecular changes that are less likely to be subject to natural selection occur more rapidly in evolution" (Kimura 1986).

Observations of high levels of heterozygosity in *Drosophila* (Lewontin and Hubby 1966) and humans (Harris 1966) provided yet another level of support for the neutral theory. Such high levels of variation could not be explained by the action of balancing selection in the form of heterozygote advantage because of the very large segregation load produced each generation (Kimura 1983). If segregating mutations are mostly neutral, variation within populations is not limited by segregational load. Neutral theory views polymorphism as a transient phase in molecular evolution. In 1971, Kimura and Ohta published "Protein polymorphism as a phase of molecular evolution" where the two phenomena of evolution, differences within and between species, were unified by the same mechanism of mutation and drift. Neutral mutations appear in a population and are either fixed or lost due to genetic drift. While segregating, these mutations constitute the variation within species; their fixation by random genetic drift contributes to differences

between species, and the rate of their fixation determines the rate of molecular evolution (Kimura and Ohta 1971a).

The convenient measure of the amount of genetic variation present in a population is heterozygosity (*H*), the probability that two randomly chosen alleles from a population are different. For neutral mutations, heterozygosity per site is given by:

$$H \approx 4N_e\,\mu \qquad \text{(Equation 4)}$$

where μ is the neutral mutation rate (Kimura 1971). This equation indicates that levels of variation under neutrality are determined by the effective population size and the neutral mutation rate. A population becomes more polymorphic as the effective population size increases due to the input of new mutations and the reduced effects of drift. Generally, the observed level of genetic variation is lower in smaller populations, although in large populations variation is usually lower than predicted on the basis of their estimated population sizes (see below).

Following the proposal of the neutral theory, both the empirical results and the theoretical arguments that begat the theory were subjected to sharp criticism. The points of debate concerned the constant rates of protein evolution, the lack of correspondence between levels of polymorphism and population size, and the plausibility of the genetic load arguments.

Gillespie has been a primary opponent of the claim that protein evolution proceeds at a constant rate, or that there is a molecular clock (Gillespie 1991). The constancy of the substitution rates can be evaluated statistically using the index of dispersion, R (the ratio of the variance in the number of substitutions to the average number), a method first used by Ohta and Kimura (1971). Under neutrality, R is expected

to be equal to 1; however multiple studies estimated the index of dispersion to be greater than 1, indicating that there is more variation in the rate of protein evolution than expected under the assumption of neutral evolution (Ohta and Kimura 1971; Langley and Fitch 1974; Gillespie 1984; Gillespie 1989). These results led Gillespie to conclude that the molecular "clock does not exist" and most replacement substitutions represent the fixation of advantageous mutations (Gillespie 1991). If it is indeed the case that the clock is episodic, this is difficult for the neutral theory to handle. There is just no way to accommodate apparent bursts of substitutions under the neutral theory according to which the substitution rate is independent of the population size.

The lack of the correspondence between levels of polymorphism and the population size, known as the paradox of variation, was pointed out by Lewontin (1974). The neutral theory predicts an increase in heterozygosity with an increase in population size (see Equation 4). However, since the observed heterozygosity from allozyme studies in various species available at that time falls within a narrow range, Lewontin concludes that "…we are required to believe that higher organisms including man, mouse, *Drosophila* and the horseshoe crab all have population sizes within a factor of 4 of each other. Moreover, other organisms, less well studied, including eels, wild grasses, chickens, and Pogonophora, give values of heterozygosity in the same range, so this extraordinary invariance in population size would appear to apply to all multicellular organisms. The patent absurdity of such a proposition is strong evidence against a neutralist explanation of observed heterozygosity." (Lewontin 1974, pp 208-210). Kimura's response to this problem is to emphasize the dependence of heterozygosity on the *effective* population size: levels of heterozygosity remain within the narrow range

because of the historical fluctuations in actual (census) population sizes that result in a much smaller effective population size (Kimura 1983).

Finally, observations of high rates of molecular evolution as well as high levels of genetic variation initially stimulated the proposal of the neutral theory mainly because they were incompatible with genetic load arguments. Substitutional and segregational loads are the result of variation in fitness: if many beneficial alleles are substituting at the same time, individuals that carry beneficial alleles at all loci will have a very high fitness that will necessitate the production of a very high number of offspring to maintain a stable population size. For example, in the case of substitutional load, the number of offspring that must be produced by the individual with maximum possible fitness assuming multiplicative fitness over all loci is about 10^{78} (Kimura and Ohta 1971b; Ewens 2004). Even if the assumption of multiplicative fitness is accepted (although it has been questioned by several authors), Ewens points out that the probability of the existence of an individual with the optimal genotype at each locus is "…so extremely small that a theory basing its numerical computations on the offspring requirement of such an individual must demand reconsideration." (Ewens 2004 p. 83).

1.2.2 The nearly neutral theory

A modification of the neutral theory came in the form of the nearly neutral theory. Ohta (1972; 1973; 1974; 1992) points out that the division of mutations into two categories, those that behave as neutral (subject to drift only) and those that behave as deleterious (subject to purifying selection only) is unlikely to reflect the biological distribution of the selection coefficients of mutations. She proposes that there is a

continuum in the strength of the deleterious effects of mutations, with a substantial fraction of slightly deleterious mutations. The key point of the theory is that the behavior of such mutations is controlled by the interplay between selection and drift (the product of the selection coefficient and the effective population size, $N_e s$): when selection is weak and the effect of drift is large (small $N_e s$), slightly deleterious mutations have a non-zero probability of fixation whereas mutations with the same selection coefficient in large populations (large $N_e s$) are efficiently removed by selection. The rate of evolution then is *not* independent of the population size, and the theory predicts a higher rate of evolution in small populations than in large ones, since slightly deleterious mutations are allowed to reach fixation in small populations but not in large ones. The neutral theory proposed by Kimura is centered around mutations whose behavior is governed only by random genetic drift. The rate of evolution of such mutations is independent of population size. This is also true for very weakly selected mutations, as long as $|N_e s|<0.05$ (Ohta and Kimura 1971), or using a more familiar expression, for mutations with $|s|<<1/(2N_e)$ where "much less" (<<) indicates at least one order of magnitude less. The substitution rate of such mutations is largely insensitive to changes in population size and is close to that of neutral mutations. For example, as can be inferred from Figure 1.1, for mutations with $s=10^{-6}$, the change in population size from 10^3 to 10^4 (change in $N_e s$ from 0.001 to 0.01) affects the rate of fixation only slightly (the rate of fixation is reduced by 0.2 and 2% respectively, relative to that of a neutral mutation). The nearly neutral theory emphasizes the importance of deleterious mutations whose rate of fixation is sensitive to changes in population size. This is the case as long as $N_e s \approx 1/2$ ($2N_e s \approx 1$). For example, for the same selection coefficient as above ($s=10^{-6}$), change in population size from 10^5 ($N_e s$

=0.1) to 10^6 ($N_e s$ =1) leads to reduction in fixation by 20 and 93% respectively, relative to the fixation rate of neutral mutation (see Figure 1.1). For such slightly deleterious mutations, the rate of substitution is negatively correlated with the effective population size. For $N_e s > 1$, change in population size has again little effect because selection is effective enough to prevent fixation of deleterious mutations (Figure 1.1).

The nearly neutral theory provides possible answers to the criticism raised against the neutral theory. This criticism was based on the empirical evidence against predicted constant rates of protein evolution as well as against the predicted correspondence between levels of polymorphism and population size.

With regard to protein evolution, the observed index of dispersion for non-synonymous substitutions is higher than 1 in mammals, indicating the episodic nature of such substitutions. Gillespie (1991) interprets this as evidence for adaptive evolution with bursts in non-synonymous substitutions in response to environmental change. Ohta (1995) however, argues that such bursts of substitutions can be explained by the nearly neutral theory: if a population experiences fluctuations in size in the course of its evolution, during the period when population size is small, slightly deleterious mutations will be fixed. These slightly deleterious mutations may follow by compensatory beneficial mutations resulting in episodes of increased rates of non-synonymous substitutions. This last point is an interesting feature of the nearly neutral theory that can be used to contrast it with the selectionist view. Nearly neutral theory incorporates beneficial mutations, however, they occur not in response to environmental change but in response to the genetic deterioration caused by the fixation of slightly deleterious mutations. That is, they play the role of maintaining the existing level of adaptation rather

than increasing it. Nearly neutral theory thus predicts a faster rate of evolution in small populations, while evolution driven by adaptive substitutions in response to environmental changes should be faster in large populations.

The nearly neutral theory also addresses Lewontin's paradox of variation. The contribution of slightly deleterious mutations to levels of variation within species is also expected to depend on population size. In small populations, the behavior of slightly deleterious mutations is controlled by the balance between mutation, selection, and drift, while in large populations these mutations are kept at low frequencies by efficient purifying selection. As population size increases, a larger fraction of mutations escapes genetic drift: the increase in selection effectiveness ($N_e s$) does not allow mutations to spread to high frequencies, imposing an upper limit on levels of heterozygosity (Ohta 1974; Ohta and Kimura 1975).

A summary of the main properties of selected mutations with respect to their contribution to the rate of evolution and population heterozygosity discussed above is given in Figure 1.2. For neutral mutations, the rate of substitution is equal to the mutation rate, and heterozygosity is equal to $4N_e\mu$. Relative to neutral mutations, advantageous mutations have higher rates of substitution and also contribute more to population heterozygosity than do neutral mutations, since they are less likely to be lost to genetic drift. Deleterious mutations have a narrow window of opportunity to contribute to the rate of evolution. The rate of substitution of these mutations is lower than that of neutral mutations and increases with decrease in population size. However, deleterious mutations can contribute substantially to genetic variation, even when they have no chance of being fixed. For example, mutations with $4N_e s = -8$ contribute only 0.003 of the substitutions

provided by neutral mutations but generate about 25% of neutral levels of heterozygosity (Figure 1.2).

It should be emphasized here that the neutral, nearly neutral, and selectionist views of evolution presented above have a common theme: all of them attempt to explain the observed patterns of polymorphism and divergence in terms of selection or drift affecting individual mutations; that is, they do not address the possibility that the dynamics of mutations may be influenced by other mutations segregating within a population. The shift toward an emphasis on the influence of mutations on the behavior of other mutations can be attributed to the description of the hitchhiking effect by Maynard Smith and Haigh (1974), at least in so far as their work stimulated multiple further studies in this area and revived the investigations of previous years. Inspired by Lewontin's paradox, these authors suggested that the fixation of advantageous mutations leads to a reduction in neutral variation by means of the hitchhiking of neutral mutations along with beneficial mutations. When a population reaches a certain size, selection becomes more important in controlling neutral variation than genetic drift, and levels of polymorphism become independent of population size. The paper was met with criticism by Ohta and Kimura who deemphasized the importance of linkage in reducing levels of variation (Ohta and Kimura 1975; 1976).

The situation changed when Begun and Aquadro (1992) reported a positive correlation between levels of nucleotide diversity and recombination rates in the genome of *D. melanogaster*. Together with previous studies documenting reduced levels of variation in areas of low recombination (Aguade *et al.* 1989; Stephan and Langley 1989; Begun and Aquadro 1991; Berry *et al.* 1991) the observed correlation identified

recombination as a major factor affecting levels of naturally occurring variation. Given that no correlation between recombination and divergence was observed (implying similar mutations rates in areas of high and low recombination), a simple neutral explanation can be ruled out. The explanation then must involve the interaction between linkage and selection (non-independent behavior of mutations), since in the absence of selection lack of recombination does not affect the average levels of variation. Moreover, the correlation was observed on the genome-wide scale, suggesting more than just isolated instances of the action of natural selection.

1.3 Interaction between linkage and selection

1.3.1 The Hill-Robertson effect

Long before the importance of selection and recombination for naturally occurring variation was empirically demonstrated, the impact of linkage on selection was investigated using a simulation approach. Hill and Robertson (1966) presented a two-locus computer simulation study indicating that the fixation probability of a beneficial mutation is reduced in the presence of another beneficial mutation at a *linked* locus. Their result can be understood in terms of a reduction in the effectiveness of selection (as a result of the interference between the two mutations), and it is often described as equivalent to a reduction in the effective population size (N_e). This so-called Hill-Robertson effect is expected to occur as long as two or more linked loci under selection are segregating (present as polymorphic sites), no matter whether mutations are beneficial or deleterious. A reduction in the effectiveness of selection associated with linkage between mutations under selection leads to a decrease in the fixation probability of

beneficial mutations and an increase in the fixation probability of deleterious mutations (Hill and Robertson 1966; Felsenstein 1974). Birky and Walsh (1988) presented a computer simulation study investigating the fixation rates of beneficial and deleterious mutations in the presence of linkage, confirming the Hill-Robertson effect: selection in the genetic background results in a reduction in the effectiveness of selection. In addition, they examined the case of linkage between selected and neutral mutations (both, analytically and with computer simulations) demonstrating that the rate of fixation of neutral mutations remains unaffected by linked selection (Birky and Walsh 1988).

It should be made clear that the Hill-Robertson effect is formulated in terms of the reduction in the effectiveness of selection because it was detected by the reduction in the probability of the fixation of beneficial mutations. The cause of this reduction in the effectiveness of selection is however a reduction in N_e (Felsenstein 1974; Felsenstein 1988), which may be considered the primary consequence of linkage in the presence of selection. Therefore, the Hill-Robertson effect is often defined as the reduction in N_e (relative to the species N_e) in the presence of selection and linkage; the tighter the linkage, the larger is the reduction in N_e and thus, the Hill-Robertson effect predicts a positive correlation between recombination rate and N_e. This formulation in terms of reduction in N_e is helpful because it provides a common explanation for different kinds of linked selection (Birky and Walsh 1988), be it positive or negative, weak or strong (for example, hitchhiking, background selection, interference selection; see below). This reduction in N_e has different consequences depending on whether mutations linked to a locus under selection are neutral or under selection. If linked sites are neutral, a reduction in N_e leads to reduction in polymorphism; if linked sites are under selection, a reduction

in N_e leads to both a reduction in polymorphism and a reduction in the effectiveness of selection (change in the fixation rate of mutations under selection). Thus, whenever a positive correlation between recombination and polymorphism (but not between recombination and divergence) is detected, the explanation involves the effect of selection on linked neutral sites. These are the cases of hitchhiking and background selection (Maynard Smith and Haigh 1974; Charlesworth *et al.* 1993), two kinds of linked selection that were originally investigated as mechanisms of the reduction of linked neutral polymorphism, and are equivalent to the Hill-Robertson effect when mutations linked to sites under selection are neutral. Whenever correlations between recombination and the effectiveness of selection (fixation rates) are detected, the explanation requires a case of the Hill-Robertson effect where mutations linked to sites under selection are also under selection (Hill and Robertson 1966). This is because when mutations behave independently from each other, neither polymorphism nor divergence is expected to correlate with recombination rates (given that the mutation rate does not change between areas of high and low recombination).

The distinctions between different kinds of linked selection are useful because they emphasize different selection regimes and different consequences of the reduction of N_e. In the following discussion, I will adhere to the following definitions of various kinds of linked selection that reflect the context in which these models were originally introduced. Hitchhiking (Maynard Smith and Haigh 1974): a mechanism of reduction in linked neutral variation as a result of a spread and fixation of strongly advantageous mutations. Background selection (Charlesworth *et al.* 1993): a mechanism of reduction in linked neutral polymorphism as a result of the elimination of strongly deleterious

mutations. Interference selection (Hill and Robertson 1966): the mechanism of reduction of effectiveness of selection as a result of the segregation in a population of two or more linked mutations under selection.

1.3.1.1 *Hitchhiking*

The significance of hitchhiking as a mechanism of reducing linked neutral polymorphism was first demonstrated by Maynard Smith and Haigh (1974), although there had been an earlier attempt to do so (Kojima and Schaeffer 1967). Stimulated by Lewontin's paradox of variation (Lewontin 1974), Maynard Smith used a deterministic, two-locus model to study the effect of the substitution process of a favorable mutation on linked neutral polymorphism. He showed that linked neutral variation is reduced as a result of linkage to a beneficial mutation. If advantageous mutations are fixed approximately at the same rate in areas of low and high recombination, neutral variation will be reduced to a greater extent in areas of low recombination, leading to a positive correlation between levels of neutral variation and recombination rates. Kaplan *et al.* (1989) presented a steady state analysis of hitchhiking with recurrent adaptive substitutions incorporating neutral mutation, random genetic drift, and recombination in their model. Their result confirms the power of hitchhiking in reducing linked neutral polymorphism. The magnitude of reduction in linked neutral polymorphism depends on the strength of selection, the recombination rate, and the substitution rate of advantageous mutations. The size of an area affected by a hitchhiking event depends on the ratio of the recombination rate to the selection coefficient, increasing when selection is high and/or

recombination is low. Wiehe and Stephan (1993) modeled the steady state hitchhiking process to obtain the expression for the reduction in heterozygosity below neutral levels:

$$\pi = \pi_0 \frac{r}{r+\kappa\gamma\lambda} \qquad \text{(Equation 5)}$$

where π is expected nucleotide diversity in the presence of hitchhiking, π_0 is the neutral levels in the absence of selection, r is the recombination rate per site per generation, γ is the selection intensity of beneficial mutations ($\gamma=2N_e s$), λ is the substitution rate of beneficial mutations and κ is a constant ($\approx$0.075) (Wiehe and Stephan 1993; Andolfatto 2007). This equation indicates that the steady state hitchhiking process is characterized by the product of selection intensity and the substitution rate of beneficial mutations. That is, the reduction in levels of neutral diversity can be explained by linkage to either rare, strongly beneficial mutations or frequent but weakly beneficial mutations. Using polymorphism data from *D. melanogaster*, Wiehe and Stephan (1993) estimated the product $\gamma*\lambda$ to be $>1.3*10^{-8}$. One of the main focal points of current studies is to obtain estimates of γ and λ separately, since these estimates will provide information about the rate and strength of adaptive substitutions from which the mutation rate of beneficial mutations can be inferred (see below; Andolfatto 2007; Macpherson *et al.* 2007; Jensen *et al.* 2008).

1.3.1.2 *Background selection*

Charlesworth *et al.* (1993) introduced the model known as "background selection" that describes a reduction in neutral variation associated with the continuous input of deleterious mutations. When a neutral mutation arises on chromosomes with deleterious mutations, it will be eliminated by natural selection. The only neutral

mutations that are allowed to drift to fixation are those that are maintained on mutation-free chromosomes. Therefore, the number of mutation-free chromosomes describes the N_e that determines levels of neutral variation under mutation-drift equilibrium. The effect of background selection depends on the recombination rate, the rate of deleterious mutations, and the selection coefficient of deleterious mutations (Hudson and Kaplan 1995; Nordborg *et al.* 1996). For the special case in which the neutral locus is located in the center of the region subject to deleterious mutations, the nucleotide diversity is given by:

$$\pi \approx \pi_0 exp\left(-\frac{U}{2hs+R}\right) \qquad \text{(Equation 6)}$$

where U is the diploid deleterious mutation rate in the region and R is the recombination rate between the ends of the region (Hudson and Kaplan 1995). Thus, as in the case of hitchhiking, background selection predicts a decrease in levels of variation with a decrease in recombination rates. The effect of the selection strength on neutral polymorphism depends on the degree of linkage or genetic distance between sites. For instance, weaker mutations remain as polymorphism longer and thus have a strong effect on the reduction in diversity when the linkage to neutral mutations is tight (Nordborg *et al.* 1996). For any given recombination rate, maximum reduction in polymorphism is expected with intermediate values of selection intensity. Mutations that are strongly deleterious are removed very quickly and do not reach sufficient frequencies to associate with a large number of neutral mutations, whereas weakly deleterious mutations remain in a population for a long time and are unable to considerably reduce linked neutral polymorphism (Charlesworth *et al.* 1995). These early studies focused on chromosome-

wide effects of background selection, investigating whether this mechanism could explain the genome-wide pattern of positive correlation between recombination rates and levels of variation.

1.3.1.3 *Interference selection*

Interference selection causes the reduction in the effectiveness of selection as a result of the segregation of two or more linked mutations under selection (Hill and Robertson 1966). Although the original investigation focused on the interference between beneficial mutations, most of the further analyses of interference selection focused on weak selection, stimulated by the observation of the non-random usage of synonymous codons or codon bias (Ikemura 1985; Sharp and Li 1987; Akashi 1995). Synonymous mutations are most appropriate for the study of interference since they comprise a sizable fraction of all segregating mutations within coding regions and are physically clustered, allowing for little recombination between them. These features provide the best conditions for interference between mutations. Moreover, the fact that fixation probabilities of weakly selected mutations are sensitive to small changes in N_e allows for the detection of predicted changes in the effectiveness of selection by measuring the changes in codon bias or rates of synonymous evolution.

Within the last decade, evolutionary and genomic consequences of interference selection generated by weakly selected mutations have been examined in much detail (Comeron *et al.* 1999; McVean and Charlesworth 2000; Comeron and Kreitman 2002). Given a sufficient number of linked sites under selection, interference between mutations leads to a reduction in the effectiveness of selection (decrease in codon bias or increase in

the rate of synonymous evolution, *Ks*) and a reduction in polymorphism levels at sites under selection relative to the expectations in the absence of interference. These effects depend on the interplay between three factors: the recombination rate, the number of sites under selection, and the selection intensity of weak mutations. The strength of interference selection increases with a decrease in recombination rates and an increase in the number of sites under selection (McVean and Charlesworth 2000; Comeron and Kreitman 2002). In addition, in the presence of many weakly selected sites, linked neutral polymorphism can be substantially reduced in the absence of recombination. Increasing the recombination rate decreases the effect of interference, although even very high levels of recombination do not restore neutral levels completely (Comeron and Kreitman 2002). These results indicate that weak selection *alone* can generate a positive correlation between the recombination rate and the effectiveness of selection as well as between the recombination rate and levels of neutral polymorphism. This expected correlation between recombination rate and effectiveness of selection (as measured by codon bias) has been supported by empirical studies in *Drosophila* (Kliman and Hey 1993; Comeron *et al.* 1999; Hey and Kliman 2002; Larracuente *et al.* 2008). A more detailed description of the Hill-Robertson effect generated by weakly selected mutations will be given in Chapter II.

Interference selection is not restricted to weakly selected mutations. If deleterious and advantageous mutations are common, they will interfere with each other and with weakly selected mutations nearby. The interference between strongly deleterious mutations has been suggested to explain the smaller reduction in neutral polymorphism than that expected under the standard background selection model in regions that lack or

have very low levels of recombination, such as the dot chromosomes of *Drosophila* species (Betancourt *et al.* 2009; Kaiser and Charlesworth 2009). Interference between strongly deleterious or strongly advantageous mutations and weakly selected mutations leads to a reduction in the effectiveness of selection at linked weakly selected sites. These results were obtained by an analytical approach (Stephan *et al.* 1999; Kim 2004) and are supported by empirical observation of a negative correlation between the rate of protein evolution and codon bias (Betancourt and Presgraves 2002; Andolfatto 2007; Larracuente *et al.* 2008). Strongly deleterious mutations are also expected to interfere with the fixation of beneficial mutations (Peck 1994; Barton 1995). If adaptation at non-synonymous sites is reduced as a result of interference between deleterious and beneficial mutations or between beneficial mutations alone, a positive correlation between protein evolution and the recombination rate is expected. Such a correlation has been documented in *Drosophila* (Betancourt and Presgraves 2002; Presgraves 2005). Moreover, the extent to which neutral polymorphism is reduced as a result of linkage to mutations under selection differs depending on the presence or absence of interference between selected mutations (Kim and Stephan 2000; Kim and Stephan 2003; Kaiser and Charlesworth 2009). It is clear that if interference is in fact pervasive, the dynamics of mutations in a population can no longer be understood as a collection of independent trajectories of mutations characterized by their specific selection coefficients.

The influence of the Hill-Robertson effect, encompassing all three types of linked selection discussed above, on patterns of variation and rates of evolution in general, or specifically on rates of adaptive evolution, depends on the mutation rates and the strength of selected mutations in combination with their recombinational environment. Although

very accurate estimates of recombination rates can be obtained experimentally with enough hard work, it is much harder to obtain accurate estimates of the rates at which mutations under selection arise and of the fitness effects associated with these mutations. The best information we have at our disposal that can be utilized to answer these questions is DNA sequence data. But the observed polymorphism and divergence data most likely reflect the combination of various selective regimes, and distinguishing between them has proved to be difficult. Still, the analysis of sequence data provides at least a rough understanding of the action of natural selection across genomes.

1.3.2 Factors affecting the strength of linked selection

I will now address the issue of the distribution of deleterious, neutral, and advantageous mutations. The best estimates are available for the class of strongly deleterious mutations. For protein-coding sequences, the fraction of deleterious mutations can be obtained by examining the numbers of synonymous (*Ks*) and non-synonymous (*Ka*) substitutions. The assumption here is that substitutions at synonymous sites are the result of the fixation of neutral mutations. Substitutions at non-synonymous sites represent fixations of neutral, advantageous, and slightly deleterious mutations. Thus, the ratio of *Ka/Ks* gives the estimate of the fraction of substitutions that are not strongly deleterious. In humans and *Drosophila*, the *Ka/Ks* ratio is about 0.3 and 0.16, respectively. The fraction of deleterious mutations that do not reach fixation is then 1-*Ka/Ks* implying that at least 70% and 84% of non-synonymous mutations are strongly deleterious in humans and *Drosophila*, respectively (Eyre-Walker and Keightley 2007). Recent studies in *Drosophila* report surprisingly high levels of constraint in non-coding

DNA as well; more than 50% of mutations have been estimated to be deleterious (Andolfatto 2005; Halligan and Keightley 2006). Since there is about 4 times as much non-coding than coding DNA in *Drosophila*, the total fraction of deleterious mutations is estimated to be ~0.58. This fraction was used to arrive at the estimate of the deleterious mutation rate per diploid genome, U≈1.2 (Haag-Liautard *et al.* 2007). In contrast to the fly genome, only about 5% of the mammalian genome is under selective constraint (Waterson *et al.* 2002; Cooper *et al.* 2004), although there are still twice as many deleterious mutations in intergenic sequences than has been identified in genic regions (Gaffney and Keightley 2006). The estimates of the distribution of fitness effects are complicated by the uncertainty in the choice of the specific distribution of selection coefficients (Loewe and Charlesworth 2006), although there is a general agreement on the average weak selection acting on segregating variants on the order of $s \approx 10^{-5}$ (Sawyer *et al.* 2003; Loewe *et al.* 2006) and the fraction of neutral non-synonymous mutations, which is generally less than 20% (Fay *et al.* 2001; Loewe and Charlesworth 2006; Loewe *et al.* 2006). Although the estimates of selection for deleterious mutations vary depending on the methods used and should not be considered as final estimates, together with high levels of constraints in proteins and non-coding DNA, they suggest that deleterious mutations occur at high rates and can be strongly and weakly deleterious, implying that both background selection caused by strongly deleterious mutations and interference selection between weakly selected mutations might have a considerable effect on levels of variation and divergence.

Estimating the proportion of advantageous mutations is even more complicated. While there are distinct classes of mutations that can be used to investigate purifying

selection (non-synonymous sites) or weak selection (synonymous sites), no such clear-cut set of sites is available for investigating the action of positive selection. Estimates of the rate and the strength of beneficial mutations have been mainly achieved by employing methods based on combined polymorphism and divergence data (McDonald and Kreitman 1991; Sawyer and Hartl 1992; Smith and Eyre-Walker 2002) or methods that rely on the effects of positive selection on linked neutral variation (Wiehe and Stephan 1993).

The prediction of the neutral theory that levels of polymorphism are proportional to levels of divergence forms the theoretical basis of tests for selection that utilize polymorphism and divergence data. This approach was used by McDonald and Kreitman (1991) to detect adaptive evolution in protein sequences. The proposed test (MK test) compares the polymorphism to divergence ratio (r_{pd}) obtained for synonymous and non-synonymous sites assuming that synonymous mutations are neutral and non-synonymous mutations are strongly deleterious, neutral, or advantageous. In the absence of advantageous mutations, the ratios of polymorphism to divergence for synonymous and non-synonymous sites should be equal since strongly deleterious mutations contribute neither to polymorphism nor to divergence. If any of the non-synonymous mutations are driven to fixation by positive selection, the ratio of polymorphism to divergence at non-synonymous sites will be lower than that of synonymous sites, indicating the excess of non-synonymous substitutions. This test can be further used to estimate the fraction of adaptive substitutions (Smith and Eyre-Walker 2002) as well as the average selection intensity acting on non-synonymous mutations (Sawyer and Hartl 1992; Bustamante *et al.* 2002). In addition, the test can be applied to detect selection in non-coding regions by

subdividing sequence data into neutral and putatively selected classes of sites (Kohn *et al.* 2004; Andolfatto 2005). The application of this test and its modified versions revealed the genome-wide action of positive selection with estimates of the proportion of adaptive substitution in *Drosophila* species ranging between 25 and 95% for protein sequences (Smith and Eyre-Walker 2002; Sawyer *et al.* 2003; Bierne and Eyre-Walker 2004; Sawyer *et al.* 2007; Shapiro *et al.* 2007) and between 20 and 60% for non-coding regions (Kohn *et al.* 2004; Andolfatto 2005; Bachtrog 2008; Haddrill *et al.* 2008; Sella *et al.* 2009). Estimates of selection inferred from the methods based on combined polymorphism and divergence are generally small, with the average selection intensity $N_e s < 10$ for beneficial mutations within protein sequences (Sawyer *et al.* 2003; Andolfatto 2005; Sawyer *et al.* 2007) and $N_e s <1$ for mutations within non-coding regions (Andolfatto 2005).

Another way to approach the issue of the rate and the strength of adaptive substitutions is to consider the effect of selection on linked neutral sites. Wiehe (1993) derived an expression that describes the reduction in neutral polymorphism in the presence of recurrent adaptive substitutions. He showed that the magnitude of reduction primarily depends on the product of the rate (λ) and strength (γ; $\gamma=2N_e s$) of adaptive substitutions (see Equation 5). Assuming that reduction in neutral polymorphism is caused entirely by the fixation of advantageous mutations, it is possible to estimate the product $\lambda\gamma$ that generates the observed levels of polymorphism (Wiehe and Stephan 1993). To estimate the rate of adaptive substitutions (λ) and their selection intensity (γ) separately, Andofatto (2007) expresses the rate of adaptive substitutions as a product of the rate of protein evolution ($K_a/2T$) and the proportion of adaptive substitutions (α);

$\lambda=\alpha K_a/2T$, where T is the species divergence time in generations. Using estimates of α = 50% obtained from the McDonald-Kreitman approach, he estimates the rate of adaptive substitutions in *D.melanogaster* to be approximately $7.5*10^{-10}$ per generation per site (1 substitution every 200 generations) in protein-coding regions with an average selection coefficient of $s=10^{-5}$ ($\gamma=2N_es \approx 40$). Macpherson *et al.* (2007) utilized yet another signature of selective sweeps, that of spatial heterogeneity in levels of linked neutral polymorphism. The basic idea is that a genomic region with more hitchhiking events should produce lower and more heterogeneous levels of polymorphism as well as higher levels of non-synonymous divergence than genomic regions with less adaptive evolution. Utilizing this expected relationship Macpherson *et al.* (2007) analyzed large-scale polymorphism and divergence data from *D.simulans* and estimated much lower rates of adaptive substitutions ($3.6*10^{-12}$ per generation per site or 1 substitution every 3000 generations) but much stronger selection intensities of adaptive mutation, s=0.01 ($\gamma=2N_es \approx 40000$). The difference in the estimates is likely to reflect the difference in the scale of the analysis and/or differences in the methodologies employed. The point of the above examples is to show that there is very little confidence about the estimates of hitchhiking parameters (the rate of adaptive substitutions and the strength of the selection). Moreover, tests based on the relationship between neutral polymorphism and recombination rates (tests that rely on Wiehe's formula) explicitly do not take into account background selection as a possible mechanism that can also reduce linked neutral variation. Despite the uncertainty about the strength and the rate of adaptive substitutions, there is no doubt that positive selection occurs to some extent as revealed by the results of the MK test.

There are two more widely used methods to detect the presence of selection. The first is based on the departure of the frequency spectrum of neutral mutations from that expected under neutrality, although it can only detect relatively recent selective events. The second is based on the comparison of non-synonymous and synonymous substitution rates and can detect the occurrence of adaptive substitutions in the relatively distant past, but it is limited in application to protein-coding sequences. Thus, although it is difficult to infer the rate of adaptive substitutions using these methods, they do reveal the presence of selection on gene to gene bases.

Methods that utilize the frequency spectrum to detect selection rely on the fact that selection alters the distribution of the frequency classes of linked neutral mutations. Fixation of advantageous mutations in the absence of recombination leads to a complete removal of linked variation; subsequent accumulation of new mutations produces a skew in the frequency spectrum towards low-frequency variants. In the presence of recombination, linked neutral variation is not eliminated completely, and the frequency spectrum of new mutations (derived variants as determined from an outgroup) is skewed either towards high or low frequencies depending on the initial association with the advantageous mutation. That is, positive selection in the presence of recombination produces a bipartite frequency spectrum (Fay and Wu 2005). The excess of low-frequency variants can be detected by Tajima's *D* statistic that compares low-frequency and intermediate-frequency variants (Tajima 1989). However, weak background selection (Charlesworth *et al.* 1995; Gordo *et al.* 2002) and interference selection (Tachida 2000; Comeron and Kreitman 2002) may also lead to an excess of mutations at low frequency. The excess of high-frequency variants on the other hand is a unique signature of positive

selection; this is the basis of the *H* statistic that compares high-frequency and intermediate-frequency variants (Fay and Wu 2000). This signature of positive selection is very brief because high-frequency variation drifts towards fixation soon after the hitchhiking event. This means that only very recent adaptive substitutions (< 0.2 N generations ago) can be detected by this test (Fay and Wu 2005).

The action of positive selection can also be inferred if the rate of evolution at the region of interest is higher than the rate of evolution at neutral sites, because the fixation probability of beneficial mutations is higher than that of neutral mutations. This idea is implemented in tests that compare the rate of substitution at non-synonymous sites (*Ka*) with the rate of substitution at synonymous sites (*Ks*) under the assumption that synonymous sites are neutral. If the ratio of *Ka/Ks* is significantly greater than one, adaptive evolution is inferred. This test is conservative since much of a protein sequence is under purifying selection, leading to much lower *Ka* than *Ks* values. Even if some amino acids are driven to fixation by positive selection, the *Ka/Ks* ratio can be less than 1. This means that there must be multiple recurrent adaptive substitutions to detect positive selection by this method. To overcome this problem, maximum likelihood tests have been developed that allow variation in *Ka/Ks* among codons. For example, one can compare the likelihood of the data under a model that allows for some codons with *Ka/Ks* >1 with the likelihood of the data under a model where no positive selection is allowed, making it possible to infer the action of positive selection for proteins with overall *Ka/Ks* <1 (reviewed in Yang and Bielawski 2000). Initial studies suggested that proteins evolving under positive selection belong to specific functional groups: for example, genes involved in immunity and reproduction comprise the largest set of genes that show

adaptive evolution (Yang and Bielawski 2000). However, recent analysis of protein-coding genes from 12 sequenced *Drosophila* genomes indicates that although essential genes have a higher proportion of codons subject to strong purifying selection, there is no difference in the proportion of codons with *Ka/Ks* >1 between essential and various non-essential classes of genes (Larracuente *et al.* 2008), indicating that genes belonging to different functional classes are equally likely to evolve adaptively. Overall, these analyses indicate that genomes might contain a substantial enough number of mutations under selection for the Hill-Robertson effect to be a relevant phenomenon in biological genomes.

As discussed above, in addition to the rate and the strength of selected mutations, recombination rates greatly affect the extent to which selection influences polymorphism levels and the effectiveness of selection at linked sites. Variation in recombination rates is observed at multiple levels: recombination rates vary between species, between individuals, between sexes and across the genome (Singh *et al.* 2009). In *Drosophila melanogaster*, recombination rates in females is on average 3 cM/Mb and zero in males, with a sex-average recombination rate of 1.5 cM/Mb. Recombination is nearly absent on the fourth chromosome and generally suppressed near the centromeres ranging between near 0 and more than 5 cM/Mb (Nachman 2002). Levels of recombination differ between *Drosophila* species as well: the recombination rates of closely related species *D.simulans* and *D.mauritiana* are approximately 30 and 80% higher, respectively, than in *D.melanogaster* (True *et al.* 1996). The average recombination rate in the more distantly-related *D. pseudoobscura* is about 4 times larger than in *D.melanogaster* (Ortiz-Barrientos *et al.* 2006). A few studies examining fine-scale recombination rates in

Drosophila have revealed considerable heterogeneity in levels of recombination. For example, in *D. melanogaster*, a 3.5-fold variation in crossover frequency is observed within a 1.2 Mb region (Singh *et al.* 2009), and a much more pronounced variation of about 40-fold (1.4 to 52 cM/Mb) within a 2 Mb region is reported in *D. pseudoobscura* (Cirulli *et al.* 2007). The large-scale pattern of recombination rates in mammals shows variation across chromosomes with reduction in recombination around the centromeres and its increase near the telomeres. At the smaller scale, the recombination profile in mammals is characterized by the presence of recombinational hotspots where large sequence blocks with low levels of recombination are interrupted by small regions (as small as 1kb) of very high recombination rates (Nachman 2002). Comparison of recombination rates between humans, mouse and rat reveals higher (~1.2 cM/Mb vs 0.6 cM/Mb) as well as more heterogeneous recombination levels in humans (Jensen-Seaman *et al.* 2004). Thus, there is a sufficient variation in recombination rates across and between genomes to detect the consequences of the Hill-Robertson effect in the presence of selection.

In summary, the above discussion is aimed at pointing out two fundamental features of biological genomes. First, all genomes examined so far lack free recombination; this is simply the consequence of the organization of genetic material—a sequence of linked base pairs that makes up the chromosomes. That is, even biologically high rates of recombination will not be able to fully uncouple the behavior of mutations at neighboring sites. Second, examination of sequence data within last two decades indicates that multiple sites within a genome are under selection. These two features ensure the non-independent behavior of linked mutations. Selection at one site will

influence the dynamics of the mutations nearby. Because selection in effect reduces the effective population size of linked sites, the immediate consequence of linkage is the genome-wide correlation between recombination rates and N_e. However, selection will also affect levels of polymorphism and the effectiveness of selection locally, even in areas of uniform recombination rates. It is the investigation of these local effects of limited recombination that is the focus of this thesis. This work is mainly motivated by the previous studies of the local variation in N_e: on the one hand, local variation in N_e can account for local patterns of effectiveness of selection and the exon-intron structure among genes. On the other hand, local variation in N_e might alter the estimates of selection and of the rate of adaptive evolution that are mostly based on methods that assume independence between sites.

1.3.3 Consequences of limited recombination: local effects of linked selection

The general Hill-Robertson effect can produce detectable consequences both at genome-wide and local scales. Most of the above discussion of the general features of linked selection describes changes in N_e that follow changes in recombination rates across chromosomes. Such changes in N_e are detected by genome-wide correlations between 1) recombination rates and polymorphism levels observed in various *Drosophila* species (Stephan and Langley 1989; Begun and Aquadro 1992; Begun *et al.* 2007), mice (Nachman 1997), tomatoes (Stephan and Langley 1998), goatgrasses (Dvorak *et al.* 1998), maize (Tenaillon *et al.* 2001), and humans (Nachman 2001); and 2) recombination rates and the effectiveness of selection (Kliman and Hey 1993; Comeron *et al.* 1999;

Betancourt and Presgraves 2002; Presgraves 2005; Larracuente *et al.* 2008; Betancourt *et al.* 2009). On the other hand, the local Hill-Robertson effect refers to changes in N_e that occur across very short distances, often differentiating single genes, single exons or even codon positions across the same exon. In this situation, selection can generate variation in N_e across genomic regions with the same recombination rate. These local consequences of selection have been invoked in the explanation of codon bias patterns and exon-intron structure (Comeron *et al.* 1999; Comeron and Kreitman 2002; Loewe and Charlesworth 2007). In addition, this short-range variation in N_e may affect estimates of selection and the proportion of adaptive substitutions.

Most work on local variation in N_e focused on the investigation of weakly selected mutations. Kliman and Hey (1993) were the first to demonstrate that the effectiveness of selection is reduced in areas of low recombination by analyzing codon bias in 385 genes from the *Drosophila melanogaster* genome. This study however did not find a correlation between codon bias and recombination across the full range of recombination rates, as predicted by polymorphism analyses that did find such a general relationship (Begun and Aquadro 1992). The predicted correlation was observed by Comeron *et al*. (1999) who detected a strong positive relationship between codon bias and recombination, but only for short proteins; genes encoding long proteins were shown to have lower codon bias and higher rates of synonymous divergence. The length of the coding sequence has been since acknowledged to be an important factor contributing to the observed patterns of codon bias and synonymous evolution, at least in *Drosophila* (Larracuente, 2008). On the basis of their results and earlier observations, Comeron *et al.* (1999) suggested that interference between weakly selected sites can account for many

of the observed patterns of codon bias in *Drosophila* as a consequence of the interplay between recombination, selection, and gene-specific features such as gene length and intron-exon structures. In support of this view, Comeron *et al.* (1999) also presented a computer simulation study demonstrating local effects of interference between many weakly selected sites.

Further simulations have investigated the full range of evolutionary and genomic consequences of interference between many weakly selected sites (McVean and Charlesworth 2000; Tachida 2000; Comeron and Kreitman 2002; Comeron *et al.* 2008). Although an increase in recombination rates decreases interference, its effect on polymorphism levels is still detectable even with a very high recombination rate (for *Drosophila* species) when the number of sites under selection is large. Importantly, interference between weakly selected sites influences the behavior of linked neutral mutations (Tachida 2000; Comeron and Kreitman 2002): neutral polymorphism levels are reduced and the frequency of neutral mutations is skewed towards rare variants.

Hill-Robertson (HR) interference has also been proposed to play a role in shaping gene and genome architecture (Comeron 2001). One of the predictions of HR is a reduction in the effectiveness of selection in the center of long coding sequences with non-zero levels of recombination (Comeron and Kreitman 2002; Comeron and Guthrie 2005). If this reduction is in fact caused by interference, the effect of which depends on tight linkage between sites under selection, it follows that if such tight linkage could be broken, the effectiveness of selection in the center of the coding region would increase. Comeron and Kreitman (2000; 2002) suggested that introns might function as modifiers of recombination, reducing the interference between weakly selected mutations from the

adjacent exons. Simulation results and the comparison of codon bias between intronless genes and genes with centrally located introns show that intron presence results in increased effectiveness of selection in the middle of the coding sequence. Thus, local interference can be important in shaping population as well as genomic features. The effect of intron presence on the effectiveness of selection is further addressed in Chapter IV using an experimental approach.

It should be emphasized that interference between weakly selected mutations is not the only selective regime that predicts the patterns described above. Recently, Loewe and Charlesworth (2007) argued that deleterious mutations are frequent enough and selection intensities (s) of deleterious mutations are low enough (but with $N_e s > 1$) that the possibility of a very local influence of background selection on N_e cannot be dismissed. Using analytical results for the reduction of N_e by background selection, they computed the magnitude of reduction in N_e across a single gene using realistic values of mutation, recombination, and gene conversion rates for *Drosophila melanogaster*. The results strongly suggest that background selection can result in variation of N_e across genes and influence adjacent neutral regions. This means that linkage effects caused only by deleterious mutations may also explain, alone or in combination with weaker selection under mutation-selection-drift equilibrium, a number of the observed patterns of codon bias in *Drosophila*.

Local changes in N_e are also predicted by hitchhiking. Fixation of advantageous mutations leaves a distinctive spatial signature on linked neutral variation. While in the absence of recombination all neutral variation is removed, in the presence of recombination neutral variation is expected to be reduced the most at sites near the

selection target. As the distance from the selected sites increases, levels of variation return to their equilibrium values. The size of an area affected by a hitchhiking event depends on the ratio of recombination rate to selection coefficient (~ 0.1 x s/r), increasing when selection is high and/or recombination is low. But even the fixation of a strongly advantageous mutation is expected to produce local differences in levels of neutral variation in areas of non-reduced recombination rates as the position under study moves away from the site under selection. For example, a mutation under strong selection (s~0.001, $N_e s \sim 1 \times 10^3$ for *Drosophila*) will leave about 10% of variation 500 bp away from the selected site in areas with a normal rate of recombination in *Drosophila melanogaster* (Stephan *et al.* 1992; Fay and Wu 2000). Mutations under weaker selection with higher recombination rates will produce less reduction in levels of neutral polymorphism near the selection target, but will recover to equilibrium values within shorter distances, generating more pronounced local changes in polymorphism levels. This means that in areas of high recombination N_e can vary across short distances in the domain of single genes and adjacent regions with moderate/strong selection. Local reduction in N_e due to hitchhiking effects is also suggested by a negative correlation between levels of synonymous polymorphism and the rate of amino acid evolution observed in genes located in areas of high recombination (Andolfatto 2007). Average levels of synonymous polymorphism are reduced by about 20% in these regions, indicating that high recombination does not remove the influence of linked selection on neutral variation completely.

One of the important results of the studies reviewed above is the existence of local effects of linked selection on neutral variation in areas with normal and high

recombination. No matter what kind of selection operates, neutral variation adjacent to sites under selection will be reduced, recovering to original levels within short physical distances. If we imagine a neutral region adjacent to a coding sequence, linkage effects are expected to alter N_e across the entire area with the largest reduction at the neutral sites that are genetically centrally positioned relative to the sites under selection. The magnitude of this reduction depends on the number of sites under selection, the kind of selection, and the distance from sites under selection. Since levels of neutral divergence are largely unaffected by linked selection (Birky and Walsh 1988), ratios of polymorphism to divergence (r_{pd}) at neutral sites will also vary across the entire region (Comeron and Kreitman 2002). This is significant because, as discussed above, the comparison of polymorphism to divergence ratios between neutral and putatively selected sites forms the basis of the widely-used McDonald-Kreitman test that can be used to estimate the fraction of adaptive substitutions (α) as well as the selection intensity (γ; $\gamma = 4N_e s$) of mutations (McDonald and Kreitman 1991; Sawyer and Hartl 1992; Bustamante *et al.* 2002; Smith and Eyre-Walker 2002). However, these estimates of α and γ rely on the independence of mutations, an assumption that does not hold in the presence of selection, even in areas of high recombination (see above).

The application of the McDonald-Kreitman test requires two classes of sites. Polymorphism to divergence ratio at neutral sites is used as a reference to assess the action of selection in a region which is putatively under selection. The presence of selection is detected by the deviation of the polymorphism to divergence ratio at sites under selection from the polymorphism to divergence ratio at neutral sites ($r_{pd\text{-neutral}} \neq r_{pd\text{-under study}}$). Within putatively selected regions, there can be neutral sites as well as sites

under selection; and in the absence of independence between mutations, these neutral sites will be affected by linked selection and will appear to evolve under positive selection in comparison to neutral sites that are not affected by selection ($r_{pd\text{-linked\_neutral}} < r_{pd\_\text{neutral}}$). This means that the polymorphism to divergence ratio for a complete region is expected to be lower than expected under independence of sites, leading to underestimates of negative selection and overestimates of positive selection when neutral sites unaffected by selection are used as a neutral reference. Notice that this particular problem is unlikely to be a matter of concern for the original McDonald-Kreitman framework when the two classes of sites are intermingled, since linked selection will affect neutral synonymous and neutral non-synonymous polymorphism equally. However, this problem may have detectable consequences when the McDonald-Kreitman framework is used to estimate the proportion of adaptive substitutions (α) in genomic regions where neutral sites and sites under selection are spatially separated. In this case, linked selection will affect the neutral polymorphism of the two regions differently, generating different ratios of polymorphism to divergence for a neutral region that is used as a neutral standard and for neutral sites within the region of interest. That is, in the absence of site-independence, the fundamental assumption of the McDonald-Kreitman test that neutral sites from the two regions contribute equally to polymorphism is violated (Williford and Comeron, 2010). This problem will be addressed in detail in chapter III.

In summary, given limited recombination and frequent selective events across genomes, N_e is expected to vary 1) with changes in recombination rate and 2) with changes in the number of sites under selection. Variation in N_e associated with changes in recombination rate has been studied extensively using analytical, simulation-based, and

experimental approaches. Local variation in N_e associated with changes in the number of sites under selection has received considerably less attention. The main goal of the following chapters is to investigate two consequences of *local* changes in N_e: 1) local change in neutral polymorphism levels and its consequences for estimates of selection at non-coding DNA, and 2) local change in the effectiveness of selection associated with intron presence/absence. The consequences of the local Hill-Robertson effect for estimates of selection are studied in chapter III using a computer simulation approach developed in the following chapter (Chapter II). Chapter IV presents the results of the experimental study aimed at evaluating the effect of intron presence/absence on the effectiveness of selection.

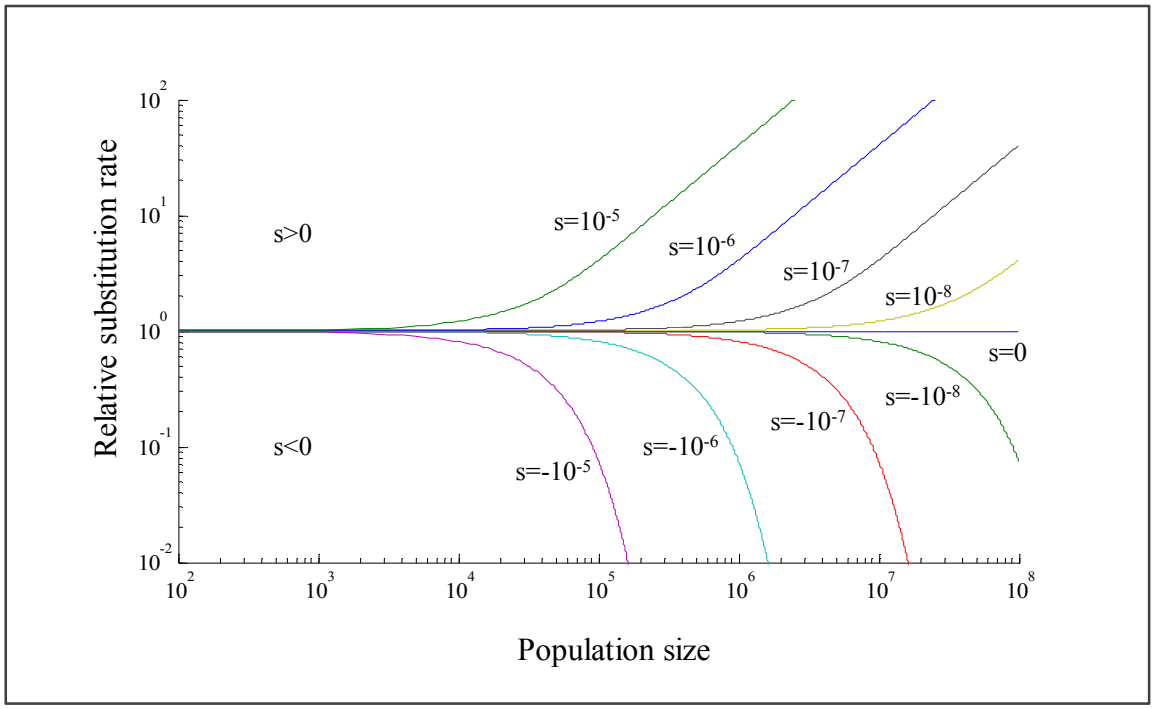

Figure 1.1. Dependence of the fixation rate of weakly selected mutations on population size. (Modified from Hurst 2009).

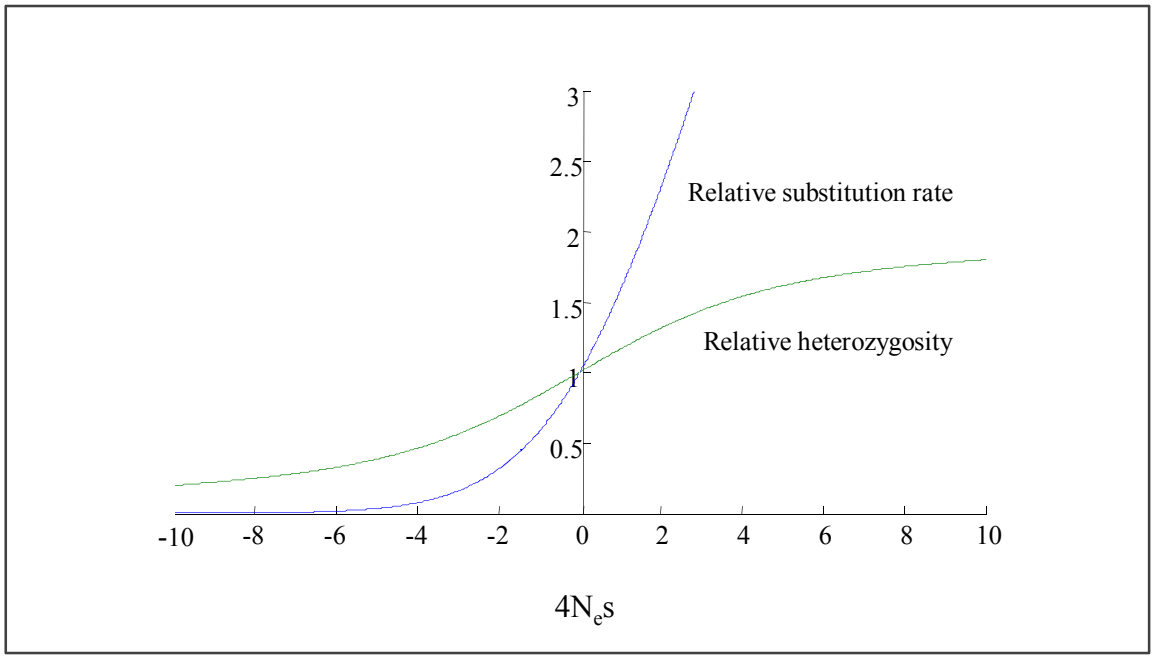

Figure 1.2. Contribution of selected mutations to substitution rate and heterozygosity relative to that of neutral mutations. Modified from Kimura (1983).

CHAPTER II

INVESTIGATION OF THE HILL-ROBERTSON EFFECT USING COMPUTER SIMULATION APPROACH

2.1 Introduction

The quantitative features of evolutionary processes can be examined using theoretical and experimental approaches. Ideally, the two approaches should complement each other. Theoretical models should be built on biologically relevant assumptions that are based on experimental data, while the most informative or comprehensive interpretations of the experimental results are only possible if they are evaluated within an appropriate theoretical framework.

Theoretical approaches can be further subdivided into analytical and simulation-based treatments of evolutionary processes. The analytical approach refers to the mathematical description of the behavior of mutations (etc.) in which solutions to various equations give a precise (or approximate) expression describing the particular properties of mutations (etc.), for example, probability of fixation, rate of evolution, polymorphism levels, and so on. This approach was used extensively in the first half of the 20th century by Fisher, Wright, and Haldane during the development of the basic framework for the treatment of changes in allele frequencies over time under different conditions. Analytical treatments of evolutionary processes also formed the basis of Kimura's and Ohta's work on neutral and nearly neutral theories of molecular evolution, although computer simulations were frequently used to complement the results obtained by analytical methods. The definite advantage of computer simulations is evident in situations in which

analytical solutions are not available. This is frequently the case when the behavior of mutations at multiple loci under different selection pressures and with various degrees of recombination is of interest.

This chapter introduces a multi-locus simulation approach that incorporates mutations, random genetic drift, various selection pressures, as well as different recombination rates to assess the extent to which non-independence between sites affects the behavior of mutations relative to the case of free recombination, or independence of sites. Since the Hill-Robertson (HR) effect refers to non-independent behavior of mutations in the presence of selection and random genetic drift, the computer simulation scheme presented here can be used to gain insights into the strength of the HR effect expected under a wide range of conditions.

2.2 Computer simulations

2.2.1 The model and its assumptions

The evolution of a diploid population is simulated according to the Fisher-Wright model. The model assumes a population of constant population size (N) with non-overlapping generations and random mating. Mutation and recombination events are introduced into the population each generation. The next generation is obtained by sampling N individuals with replacement from the previous generation, with the probability proportional to their relative fitness. The mutational process allows for two allelic states at each site and mimics the infinitely many sites model according to which new mutations always occur at a new site. Semidominant selection is assumed where each mutation has a selective advantage s in a heterozygous state and $2s$ in a homozygous

state. Fitness over sites is multiplicative, meaning that the fitness of each diploid individual is determined by the product of the fitness at individual sites.

Simulations are aimed at investigating the behavior of mutations under weak and moderate/strong selection. The behavior of these two types of mutations was mimicked according to the infinitely many sites model with irreversible mutations (for moderate/strong selection) and with reversible mutations (for weak selection).

Under the assumption of the infinitely many sites model and of no interference between sites, the probability of fixation of a mutation under selection can be approximated by:

$$PrFix = \frac{\gamma}{2N(1-e^{-\gamma})} \quad \text{(Equation 2.1)}$$

where $\gamma = 4N_e s$ and N is the number of diploid individuals in a population (Kimura 1983; Sawyer and Hartl 1992). At equilibrium, polymorphism levels (as measured by the number of segregating sites per site) for mutations with selection intensity γ are given by:

$$Pol = 4N_e \mu \int \frac{(1-x^n-(1-x)^n)(1-e^{\gamma(x-1)})}{x(1-x)(1-e^{-\gamma})}\, dx \quad \text{(Equation 2.2)}$$

where μ is the mutation rate per site per generation, x is the frequency of a mutation and n is the sample size (Bustamante *et al.* 2002). Figure 2.1 illustrates the change in the probability of fixation and polymorphism levels with changes in selection intensity (γ) for mutations under negative and positive selection. The probability of fixation increases with the increase in γ, while polymorphism levels reach a maximum at $2*4N_e\mu$.

The dynamics of weakly selected mutations (Figure 2.1 C,D) is simulated according to the mutation-selection-drift (MSD) model (Li 1987; Bulmer 1991) that was developed to investigate synonymous codon usage. In this model, the two alleles are

referred to as preferred (beneficial) and unpreferred (deleterious). In simulations, this situation is modeled by allowing reversible mutations. That is, at equilibrium, there will be a fraction of sites with the preferred variant and a fraction of sites with the unpreferred variant. A new mutation introduces the unpreferred variant at a site with the preferred variant and the preferred variant at a site with the unpreferred variant. The expected frequency of sites with the preferred variant (P) is given by:

$$P = \frac{e^{\gamma}V}{e^{\gamma}V+U} \qquad \text{(Equation 2.3)}$$

where $\gamma = 4N_e s$, V is the mutation rate from the unpreferred to the preferred state and U is the mutation rate in the reverse direction (Bulmer 1991). Polymorphism levels in this case are determined by the contribution of segregating unpreferred mutations at sites with the preferred variant and segregating preferred mutations at sites with the unpreferred variant:

$$Pol = P * Pol(-\gamma) + (1 - P) * Pol(+\gamma) \quad \text{(Equation 2.4)}$$

where P is given by equation 2.3, $Pol(-\gamma)$ is given by equation 3.2 with the negative sign of γ and $Pol(+\gamma)$ is given by equation 2.2 with the positive sign of γ. The frequency of sites with the preferred variant increases with the increase in γ until a plateau is reached when all sites are in the preferred state, while polymorphism levels decline since most of the segregating mutations are represented by the unpreferred variant (Figure 2.2).

According to the diffusion theory, the selection process on a time scale proportional to N can be completely described by the initial frequency of a mutant allele and the parameters $N_e\mu$, $N_e s$, $N_e r$ (Hill and Robertson 1966; Ewens 2004). That is, the dynamics of mutations and their effects in large (natural) populations can be studied by

simulating the evolution of relatively small population sizes as long as the product of $N_e\mu$, N_er and N_es remains constant. The range of these scaled parameters are chosen to approximate those found in natural populations of *Drosophila*: $4N_e\mu$: 0.03 and 0.04 (Moriyama and Powell 1996; Andolfatto 2007; Begun *et al.* 2007); $4N_es$: -100 to 100 (Eyre-Walker 2006; Loewe and Charlesworth 2006; Andolfatto 2007); N_er : 0 to 0.1 - this range is equivalent to 0 - 10 cM/Mb assuming $N_e=10^6$ (Comeron and Kreitman 2002; Nachman 2002; Ortiz-Barrientos *et al.* 2006; Singh *et al.* 2009).

2.2.2 Description of the program

The code is written in the C++ programming language. All simulations were performed in two steps. First, the initial population was generated from random sequences and was brought to mutation-random genetic drift balance by repeated rounds of mutations and random sampling to create a new generation. Second, the population at mutation-drift equilibrium from step one was used as an input population for further evolution in the presence or absence of selection and various recombination rates. Each simulation proceeded until the population reached equilibrium (5000 *N* generations). From this point on, the population was sampled every *N* or 5 *N* generations for at least 5000 *N* generations, and various properties of the population/sample were calculated.

A diploid population of size *N* is represented by a two-dimensional vector defined by the number of sequences/chromosomes in a population (2*N*) and the length of each sequence/chromosome (*L*). The first step is to generate a population in mutation-drift equilibrium. Initially, the population is composed of 2*N* identical chromosomes represented by a random sequence of zeros and ones. This means that no more than two

alleles can segregate at each site, reflecting the fact that in natural populations, the mutation rate is low enough that the same site is unlikely to be mutated more than once while it is segregating. That is, sites with more than two alleles are only rarely observed. The mutational process is modeled by a Poisson distribution, with the average number of mutations, $\lambda = 2N\mu L$, where μ is the mutation rate per site per generation. Each generation, an average λ of mutations is added to a population by randomly choosing a chromosome and a position within a chromosome provided that the position is fixed (a non-polymorphic site). This procedure mimics the infinitely many sites model without having to simulate a very large number of sites. To create the next generation, $2N$ chromosomes are selected randomly with replacement from the previous generation. This procedure simulates random genetic drift, or random fluctuations in allele frequencies due to sampling effects. After 10 000 N generations, when the population reaches mutation-drift equilibrium where the continuous input of mutations is balanced by the loss of mutations due to random genetic drift, the population sequence data is recorded and used as an input file for further evolution under various conditions.

The flowchart of the program for the simulated evolution of populations in the presence of selection is shown in Figure 2.3. Three user-defined input files are provided at the beginning of each simulation. The first input file contains the population vector at mutation-drift equilibrium as described above. The second input file specifies additional parameters, such as recombination rate, population sample size, divergence time between populations to be compared, and the window size for the analysis of population and evolutionary parameters. The third input file provides the distribution of fitness effects. Each sequence of length L can be subdivided into at most 4 classes of sites that can have

different selection coefficients. For example, depending on the conditions investigated, a user-defined number of sites can be assigned to receive neutral mutations as well as mutations evolving under positive, negative, or weak selection. Various properties of a population/sample (polymorphism, divergence, fixation probability) can be analyzed separately for each class. In addition, analysis can be done separately for a specified window size. For example, 1000 base pairs of neutral sequence adjacent to sites under selection can be subdivided into 4 regions of 250 bp, and sample statistics can be calculated for each region separately to assess spatial changes in levels of polymorphism.

Each simulation proceeds for at least 10 000 N generations. Each generation, mutations are added in the same way as described above for the initial simulation step. Chromosomes are randomly paired to form diploid individuals. Recombination then takes place where the number of recombination events per population is drawn from a Poisson distribution with an average, $\lambda=NrL$, where r is the recombination rate per site per generation. Each generation, the fitness of diploid individuals is calculated assuming multiplicative fitness over sites. The next generation is formed by choosing N individuals from the previous generation with a probability proportional to their relative fitness. A number of population and sample properties (from a randomly chosen sample of 10 chromosomes) is obtained every N or 5 N generations. The description of the program output is given below:

θ: measure of genetic variation (polymorphism levels/heterozygosity) based on the number of segregating sites, i.e., the probability of two randomly chosen sequences being different. It is estimated from simulations by counting the number of segregating sites (S) in a sample of 10 chromosomes (n) and the following equation:

$$\theta = \frac{S}{a_n L}$$

where $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$ and L is the number of sites.

Polymorphism: number of polymorphic sites, obtained from simulations by counting the number of segregating sites in a sample of 10 chromosomes.

Divergence: number of fixed sites, obtained from simulations by counting the number of fixed differences between a sample of 10 chromosomes and a sequence separated from the sample by 10*N* generations.

P: proportion of sites with the preferred variant, obtained from simulations for mutations under weak selection by counting the number of sites with the preferred variant in the ancestral state.

PrFix: probability of fixation, obtained from simulations by calculating the ratio of the number of mutations that are fixed in a population to the number of mutations that are added to a population.

2.2.3 Agreement between theoretical predictions and simulation results

In order to evaluate the accuracy of the simulation procedure, computer simulations were performed with the set of parameters chosen to simulate the *independent* behavior of mutations. In this case, simulation results are expected to match closely the theoretical predictions derived under the assumptions of the infinitely many sites model when sites are independent. The agreement between the theory and the simulations is shown below for the case of weak and moderate/strong selection. Results presented for a case of weak selection serve as a check for the simulations that are used to

investigate the Hill-Robertson effect generated by weakly selected mutations (section 2.3 below). The agreement between the simulation results and the theory in the case of moderate/strong selection confirms the accuracy of the simulation procedure that is used in Chapter III to evaluate local effects of linked selection.

2.2.3.1 *Weak selection*

Table 2.1 lists simulation parameters used to assess the agreement between simulation results and theoretical predictions in the case of weak selection. A number of simulation parameters were chosen to match those used in the previous studies investigating the Hill-Robertson effect generated by weakly selected mutations (McVean and Charlesworth 2000; Comeron and Kreitman 2002). Simulations of short sequences with high recombination rates were performed to minimize the effect of possible interference between selected sites.

Table 2.1 Simulation parameters used to assess the accuracy of simulations

	N	L	$4N_e\mu$	$N_e r$	γ ($4N_e s$) range
Weak Selection	100	100	0.04	0.4	0 to 10
Moderate/Strong Selection	200	1	0.03	0.1	-100 to +100

Two measures were used to assess the accuracy of simulations: the frequency of sites with a preferred variant (P, Figure 2.4 A) and polymorphism levels based on the number of segregating sites (θ, Figure 2.4 B). There is a reasonably good agreement between simulation results and the theoretical predictions for both neutral and selected mutations, indicating that the simulation procedure mimics the expected behavior of

mutations in the presence of genetic drift and selection fairly well (Figure 2.4). The slight deviations from the expectations can be attributed to the residual effect of interference since the simulated sequence contained 100 sites under selection.

2.2.3.2 *Moderate/strong selection*

Table 2.1 lists simulation parameters used to assess the agreement between simulation results and theoretical predictions for the case of moderate/strong selection. The probability of fixation of selected mutations obtained from simulations was compared to the theoretical predictions of the infinitely many sites model. To avoid the possibility of interference between mutations under selection, only one site under selection was introduced within L=500 bp. To minimize the amount of noise associated with the analysis of a single site, 6 independent runs were performed. As the range of investigated selection intensities (γ) increases, simulation of a larger population size is required (see below). To minimize the effects associated with small population sizes, the population size was increased to N=200.

In addition to the probability of fixation given by the Equation 2.1 (which is an approximation used when s is small), simulation results were checked against the expectations in a more general case given by Kimura's equation (Kimura 1983):

$$PrFix = \frac{1-e^{-\gamma p}}{1-e^{-\gamma}} \qquad \text{(Equation 2.5)}$$

where $\gamma = 4N_e s$ and p is the initial frequency of a mutation which is equal to $1/(2N)$ for a new mutation.

However, the true probability of fixation lies between the probability of fixation obtained from the Equation 2.5 and the probability of fixation obtained from the Equation 2.6 shown below (Moran 1962; Felsenstein 1988):

$$PrFix = \frac{1-e^{-4N_e s/(1+s)p}}{1-e^{-4N_e s/(1+s)}} \qquad \text{(Equation 2.6)}$$

where s is the selection coefficient of a mutation and p is its initial frequency.

Equation 2.1 is a special case of Equation 2.5 and is widely used in the literature (Bulmer 1991; Sawyer and Hartl 1992; Bustamante *et al.* 2002). The use of the approximate expression is justified when s is small or N is large, which is a likely situation in natural populations. However, verification of the simulation procedure with small N should be done against the probability of fixation that lies between that given by the Equation 2.5 and the Equation 2.6 (I will refer to it as Moran's probability of fixation). Figure 2.5 shows the agreement between simulation results and theoretical expectations according to equations 2.1, 2.5 and Moran's probability of fixation. There is a very good match between simulations and Moran's probability of fixation for a full range of parameters investigated demonstrating that the simulation procedure successfully mimics the expected behavior of mutations. The discrepancy between simulation results and values obtained from equations 2.1 and 2.5 is observed for large values of γ ($4N_e s > 50$) because a small population size (N=200) and a large s ($s >$ 0.0625) were used in simulations. In natural populations, such small population sizes and large selection coefficients are unlikely, and the approximation (Equation 2.1) gives accurate values of the probability of fixation. At the practical level, the investigation of large selection intensities ($4N_e s > 50$) requires simulations of larger population sizes. For

example, with N=2000, the differences between values obtained from the three equations for γ =100 mostly disappear (Figure 2.6) so that results of the simulations based on N=2000 would reflect the dynamics of mutations under selection in natural populations more accurately than those based on N=200.

2.3 Applications of the computer simulation approach: investigation of the Hill-Robertson effect generated by weakly selected mutations

2.3.1 Introduction

In its most basic formulation, the Hill-Robertson effect refers to the non-independent behavior of mutations in finite populations in the presence of selection. The original investigation (Hill and Robertson, 1966) focused on the interference between advantageous mutations segregating at two loci. The non-independent behavior of mutations is observed whenever the two mutations are present in a population at the same time and recombination between the two loci is not sufficiently high to eliminate the effect of selection at one locus on the behavior of a mutation under selection at another locus. Such interference between mutations under selection is detected by the reduction in the probability of fixation of advantageous mutations. Since the ultimate fate of a mutation (or the effectiveness of selection) depends on the product of the effective population size (N_e) (which determines the magnitude of genetic drift) and selection coefficient (s), the observed reduction in fixation probability can be interpreted as a reduction in the effectiveness of selection as a result of the decrease in the effective population size (N_e). This means that the presence and the consequences of the Hill-Robertson effect in biological genomes can be investigated most effectively using a class

of mutations whose behavior is sensitive to changes in N_e. This class of mutations contains weakly selected mutations, and in *Drosophila* as well as in many other species, weak selection can be investigated by studying synonymous mutations and synonymous codon usage (Kliman and Hey 1993; Moriyama and Hartl 1993; Hartl *et al.* 1994; Akashi 1995; Akashi 1996; Moriyama and Powell 1998; Comeron *et al.* 1999; Comeron and Kreitman 2002). The investigation of the consequences of the Hill-Robertson effect generated by weakly selected mutations is motivated by the fact that the classical model of weak selection (MSD) cannot account for a number of observed features of synonymous mutations. First, there is a positive correlation between codon bias and recombination rates (Kliman and Hey 1993; Comeron *et al.* 1999). Second, the length of the coding sequence is positively correlated with the synonymous substitution rate and negatively correlated with codon bias (Comeron and Aguade 1996; Powell and Moriyama 1997; Moriyama and Powell 1998; Comeron *et al.* 1999). Third, there is a positive correlation between levels of synonymous polymorphism and codon bias (Moriyama and Powell 1996; Powell and Moriyama 1997). Whether or not these observed features of synonymous mutations can be at least partially accounted for by interference between synonymous mutations can be assessed by computer simulation studies in which the consequences of multiple weakly selected mutations can be examined under various conditions.

2.3.2 Computer Simulation Results

Below I show the results of the computer simulations performed as described in section 2.2 focusing on the consequences of the Hill-Robertson effect for levels of

polymorphism and the effectiveness of selection. These results are in qualitative agreement with previous studies (McVean and Charlesworth 2000; Comeron and Kreitman 2002).

2.3.2.1 *The Hill-Robertson effect results in the reduction of the frequency of sites with the beneficial/preferred variant*

In the absence of interference, according to the MSD model, the frequency of sites with the preferred variant increases with an increase in the strength of selection intensity (Figure 2.2 A). Introduction of multiple linked sites under selection generates the Hill-Robertson effect, that is, the frequency of sites with the preferred variant is reduced. This result is equivalent to the reduction in the effectiveness of selection and is the consequence of the reduction in the fixation probability of preferred mutations and of the increase in the fixation probability of unpreferred mutations. The strength of the HR effect increases as the recombination rate decreases and the number of sites under selection increases. Figure 2.7 A shows the effect of increasing the number of sites under selection (with complete linkage) on the frequency of sites with the preferred variant (P). For any given selection intensity, the longer the sequence under selection, the greater the reduction in P. But the largest reduction in P is observed with intermediate selection intensities (Figure 2.7 B). This is because the expected increase in P with selection strength is not linear but is rapid for small selection intensities ($\gamma < 2$) and very slow for larger selection intensities ($\gamma > 5$). Figure 2.7 C shows theoretical expectations when N_e is reduced by a factor f (f = 0.75 or f = 0.5), with an expected maximum reduction in P

when γ is between 1 and 4. That is, the reduction in *P* due to the Hill-Robertson effect can be approximated by the consequences of reduced N_e.

Increasing the recombination rate reduces the HR effect. When the number of sites under selection is small, high levels of recombination mostly restore the frequency of preferred sites expected in the absence of interference. But for long sequences under selection, even realistically high levels of recombination are not sufficient to completely remove the effects of interference (Figure 2.8 A).

The frequency of sites with preferred variants is analogous to the frequency of preferred codons: the higher the frequency of preferred variant/optimal codon, the stronger the codon bias. The results presented above predict the reduction in codon bias with decrease in recombination rates and with increase in the length of the coding sequence. The analysis of the *D.melanogaster* genome shows qualitatively similar results (Figure 2.8 B) suggesting that the observed positive correlation between codon bias and recombination rate and negative correlation between codon bias and gene length might be at least partly explained by the Hill-Robertson effect generated by weak selection acting on synonymous mutations.

2.3.2.2 *The Hill-Robertson effect results in the reduction of the polymorphism levels of weakly selected mutations*

In the absence of interference, according to the theoretical predictions of the MSD model, polymorphism levels of sites under selection decrease with an increase in selection intensity (Figure 2.2 B). Interference between selected mutations is expected to result in the reduction of polymorphism levels if the Hill-Robertson effect can be

rationalized in terms of the reduction in N_e. In general (and as predicted), levels of polymorphism decrease with increase in the number of sites under selection and with decrease in the rates of recombination. The maximum effect of interference is again observed for intermediate selection intensity with up to a 40% reduction in θ when 10000 completely linked sites are under selection with $\gamma = 2$. High levels of recombination nearly completely restore polymorphism levels expected in the absence of interference (Figure 2.9). This consequence of the Hill-Robertson effect is consistent with the observed positive correlation between codon bias and synonymous polymorphism: low codon bias associated with the reduction in the effectiveness of selection is the result of the reduction in N_e that also leads to reduced polymorphism levels.

2.3.2.3 *The Hill-Robertson effect generates local variation in levels of polymorphism and the effectiveness of selection*

The dependence of the magnitude of the Hill-Robertson effect on the number of sites under selection predicts the so called 'center effect' where the central region of the sequence under selection is expected to show a larger reduction in the effectiveness of selection and levels of polymorphism (Comeron and Kreitman 2002). This effect is expected because in areas with non-zero recombination rates, the central region is affected by more sites under selection than lateral regions, and the interference between mutations will be reduced more effectively toward the ends of the sequence than in the middle. With zero and high levels of recombination, no such heterogeneity across the region under selection is expected to occur. Figure 2.10 illustrates "center effect" for intermediate/low levels of recombination (Nr=0.004) detected in the long sequence

(L=2500 bp) under selection (γ=2). In this case, the polymorphism levels are reduced in the center of the region by ~ 4%. The effectiveness of selection also varies across the region under selection with intermediate levels of recombination (Figure 2.11). There is very slight but detectable reduction in the frequency of sites with preferred variant (P) in the center of the region, ~2% (Figure 2.11 A). This reduction is the consequence of a decrease in the probability of fixation of a preferred variant and an increase in the probability of fixation of an unpreferred variant (Figure 2.11 B,C).

Importantly, polymorphism levels of adjacent neutral regions are also notably affected by weakly selected mutations (Figure 2.10). Under the conditions investigated here, neutral polymorphism is reduced the most in the region immediately adjacent to the region under selection (by ~ 17%) and approaches neutral expectations about 2 kb away from the selection target. In general, the strongest reduction in neutral polymorphism occurs under the same conditions that maximize the Hill-Robertson effect: large number of sites under selection and low recombination rate. Increasing recombination leads to a gradual increase in levels of neutral polymorphism with faster recovery to neutral levels at sites that are farthest away from the region under selection. This result indicates that reduction in neutral polymorphism levels is not a unique consequence of strong selection (hitchhiking or background selection) but is a general consequence of directional selection of any strength. That is, unless neutral regions are located in genomic regions devoid of sites under selection, their polymorphism levels are unlikely to reflect the true neutral levels of genetic variation.

2.4 Summary

Computer simulations based on the code described in section 2.2.2 have been applied to investigate the effect of many weakly selected sites on patterns of polymorphism and the effectiveness of selection. There are three variables that jointly determine the strength of the Hill-Robertson effect generated by such weakly selected mutations: 1) the number of sites under selection (L), 2) the recombination rate ($N_e r$), and 3) the selection intensity (γ). With a large number of sites under selection and sufficiently low recombination rates, interference between mutations generates patterns of polymorphism and the effectiveness of selection consistent with those observed for synonymous mutations. For any given number of sites under selection, codon bias and polymorphism levels decrease with decrease in recombination rates. This effect can be referred to as the long-range Hill-Robertson effect. On the other hand, the short-range Hill-Robertson effect describes local changes in levels of polymorphism and the effectiveness of selection associated with changes in the number of sites under selection in areas of uniform recombination.

The observation of the "center effect" suggests a possible advantage of introns as modifiers of recombination in areas of low/intermediate recombination rates. Computer simulations as well as empirical studies provide strong support for this hypothesis (Comeron and Kreitman 2002; Comeron and Guthrie 2005), although the experimental approach is complicated by the presence of strongly selected mutations. The possible role of introns in reducing the intragenic Hill-Robertson effect will be examined in detail in Chapter IV.

The detection of local effects of selection on neutral polymorphism (Figure 2.10) indicates that polymorphism levels may vary within short distances generating heterogeneous neutral polymorphism to divergence ratios. Such non-uniformity of polymorphism to divergence ratios violates the fundamental assumption of the McDonald-Kreitman test widely used to detect and estimate the rate of adaptive evolution. The simulation approach successfully applied here in the case of weak selection will be extended in the next chapter to include stronger selection intensities in order to examine the extent to which estimates of selection and the proportion of adaptive substitutions are affected by linked selection.

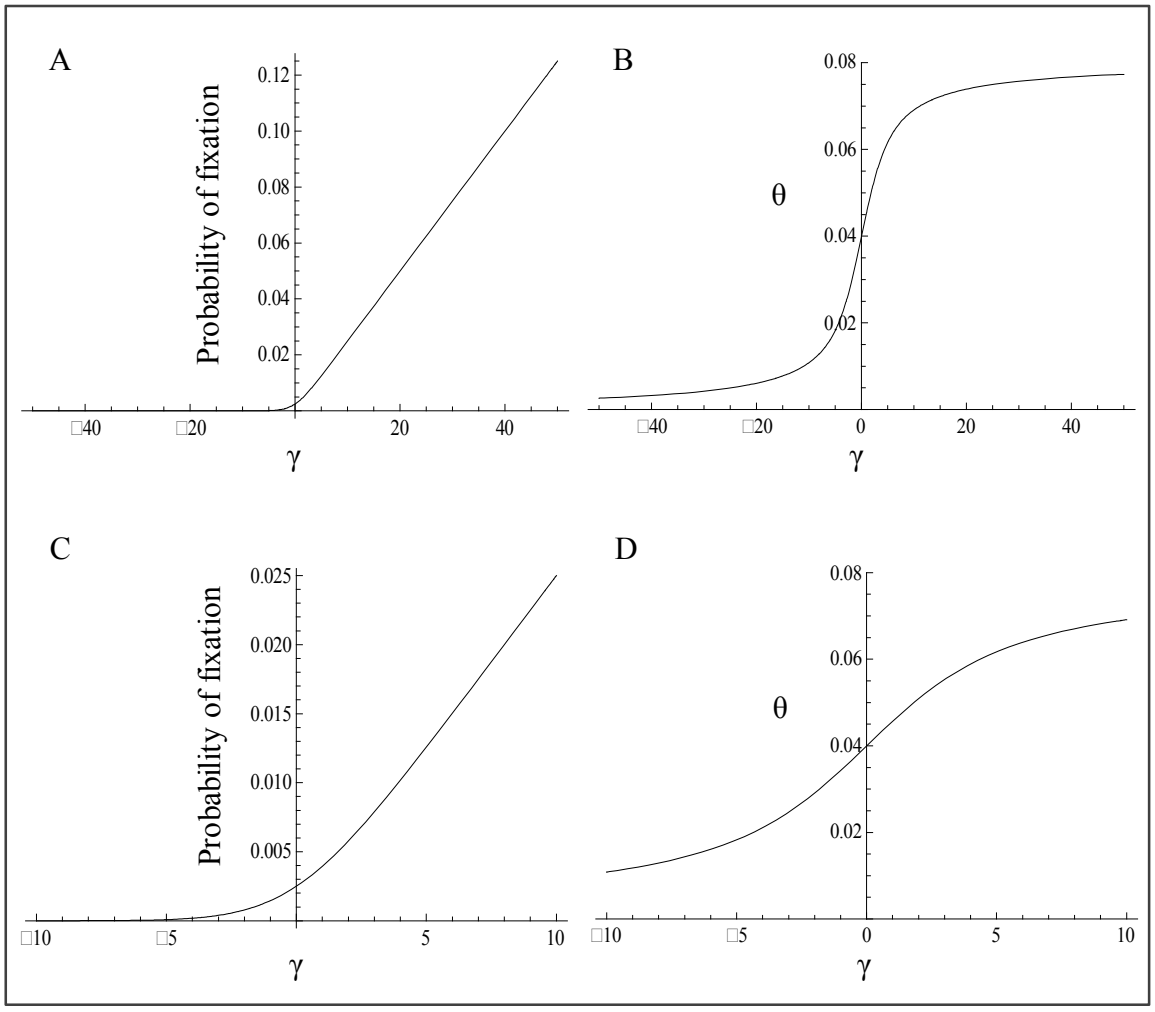

Figure 2.1. Expectations under infinitely many sites model in the absence of interference. A, C. Probability of fixation with neutral expectation corresponding to $1/(2N)$ with N=200. B, D. Polymorphism levels (θ) with neutral expectation of 0.04. The range of γ values in C and D is reduced to show the expectations for mutations within the range of weak selection.

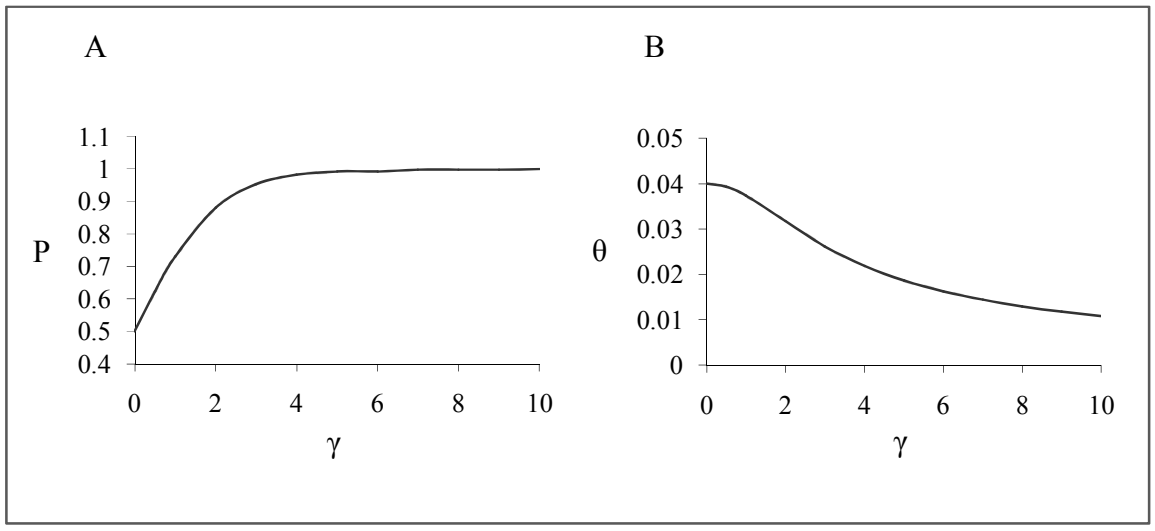

Figure 2.2. Expectations under mutation-selection-drift (MSD) equilibrium in the absence of interference. A. Frequency of sites with the preferred variant (*P*). B. Polymorphism levels (θ) with neutral expectations of θ=0.04.

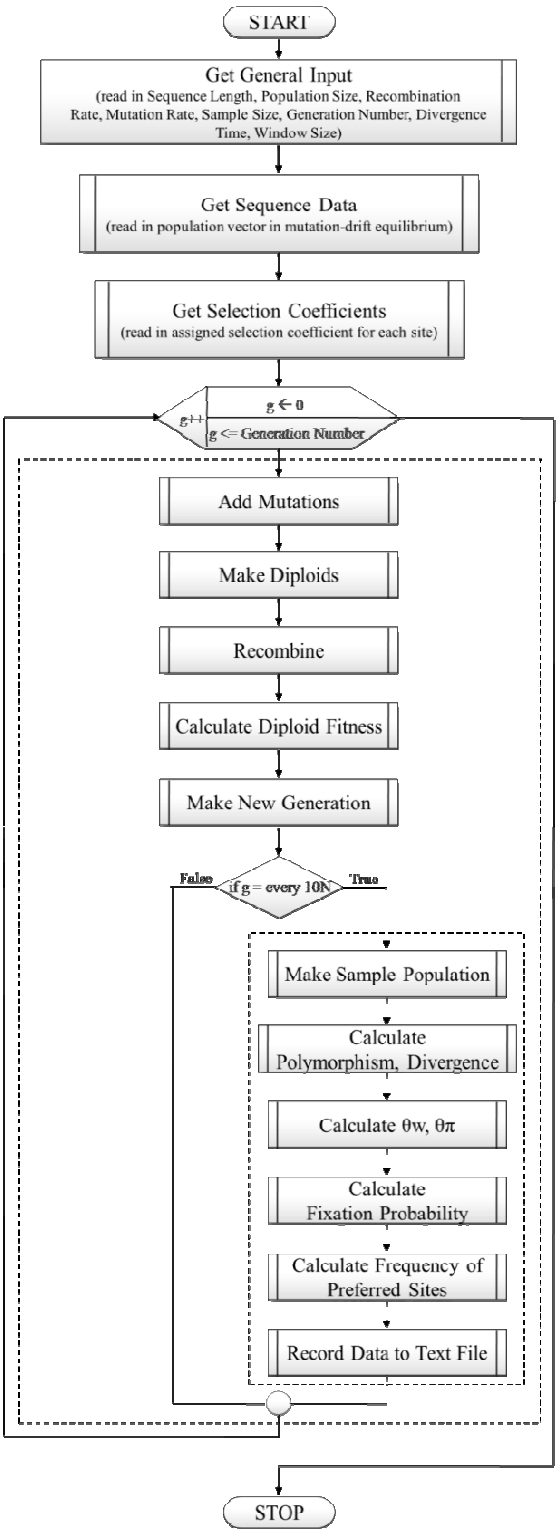

Figure 2.3. Flowchart of the program showing main biologically relevant functions.

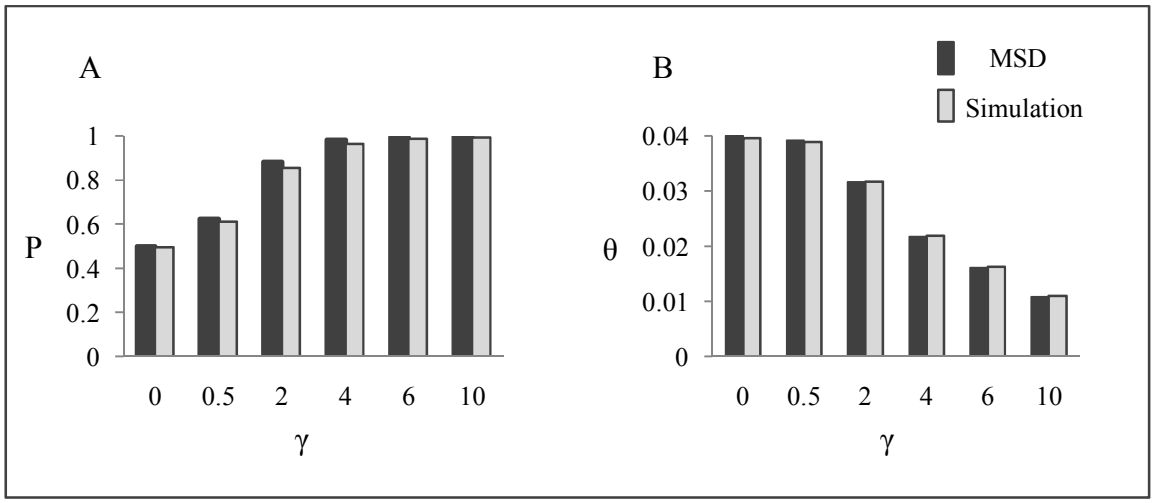

Figure 2.4. Agreement between simulation results and expectations under mutation-selection-drift (MSD) equilibrium in the absence of interference. A. Frequency of sites with the preferred variant (*P*). B. Polymorphism level (θ) based on the number of segregating sites in a sample of 10 chromosomes (with neutral expectations of θ=0.04).

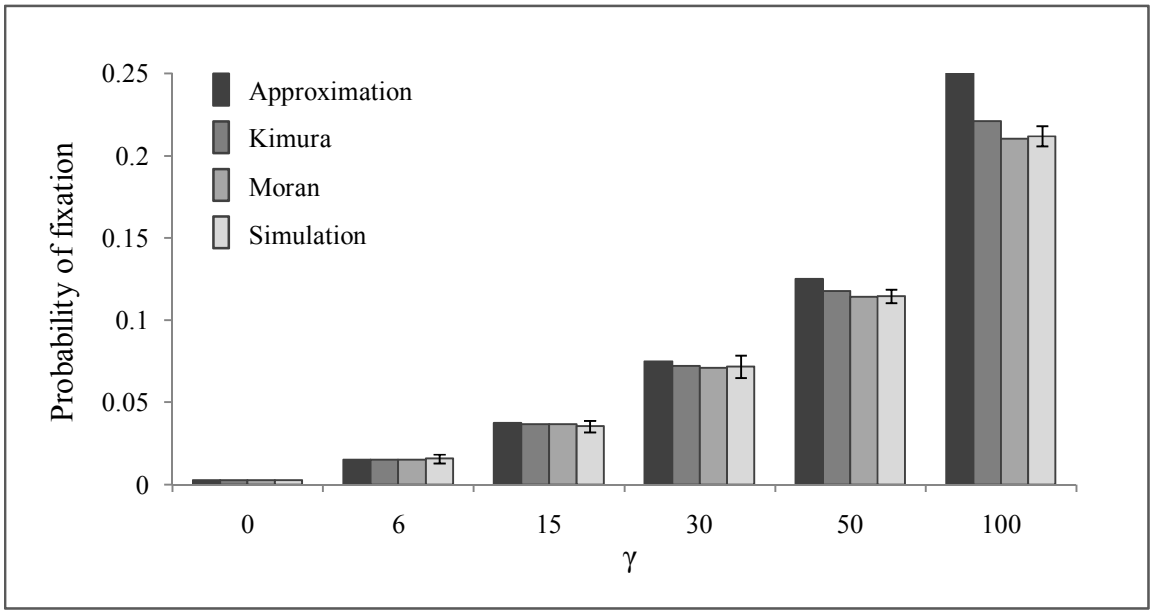

Figure 2.5. Agreement between simulation results (*N*=200) and expectations under infinitely many sites model. Moran's probability of fixation shown here lies exactly between the values obtained from the Equations 2.5 and 2.6 (see text). Error bars associated with simulation results show +/- 2 standard deviations determined on the basis of the results from 6 independent runs.

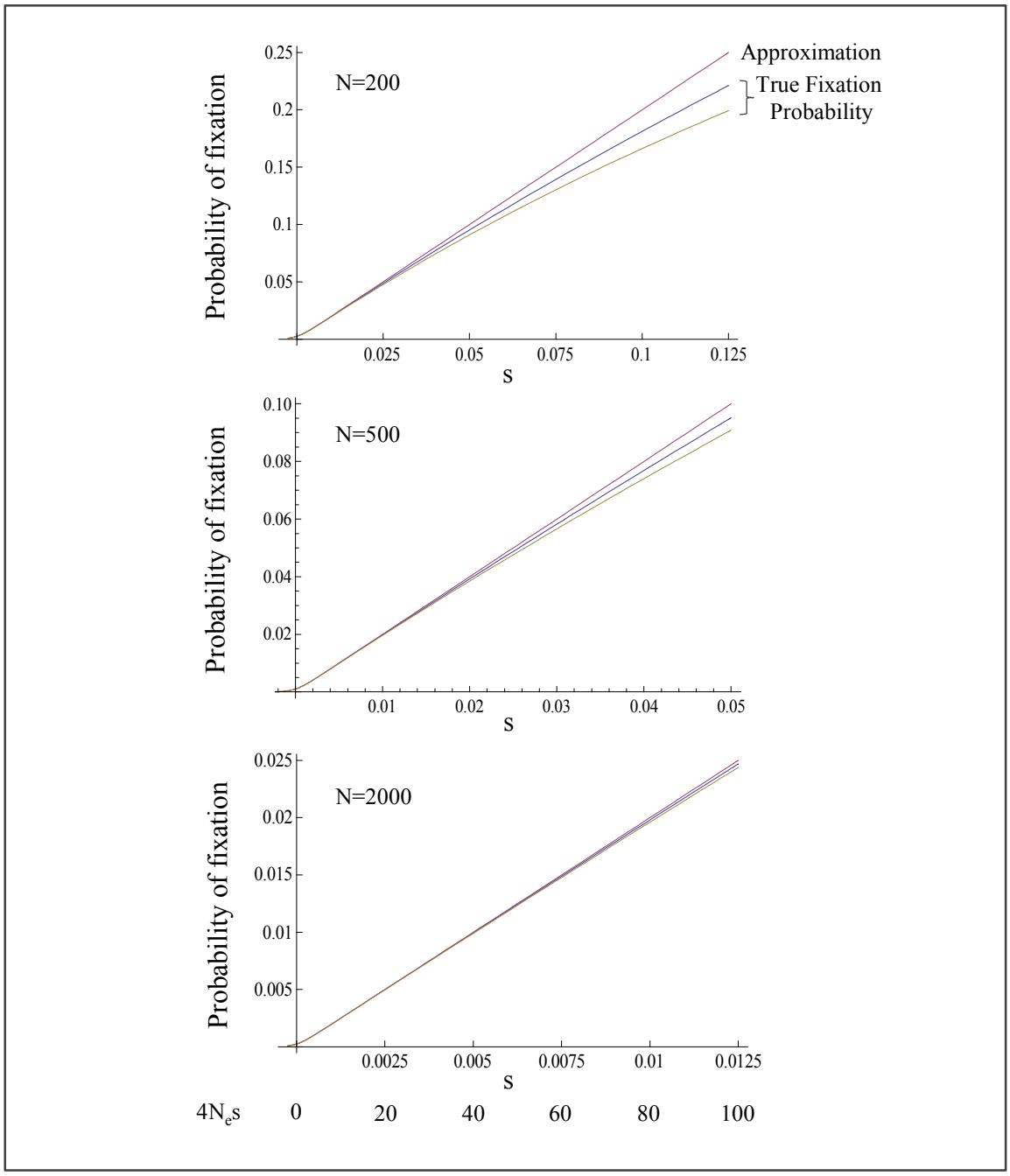

Figure 2.6. Discrepancies between theoretical expectations of fixation probability. The three curves on each plot from top to bottom correspond to Equation 2.1, Equation 2.5 and Equation 2.6 (see text).

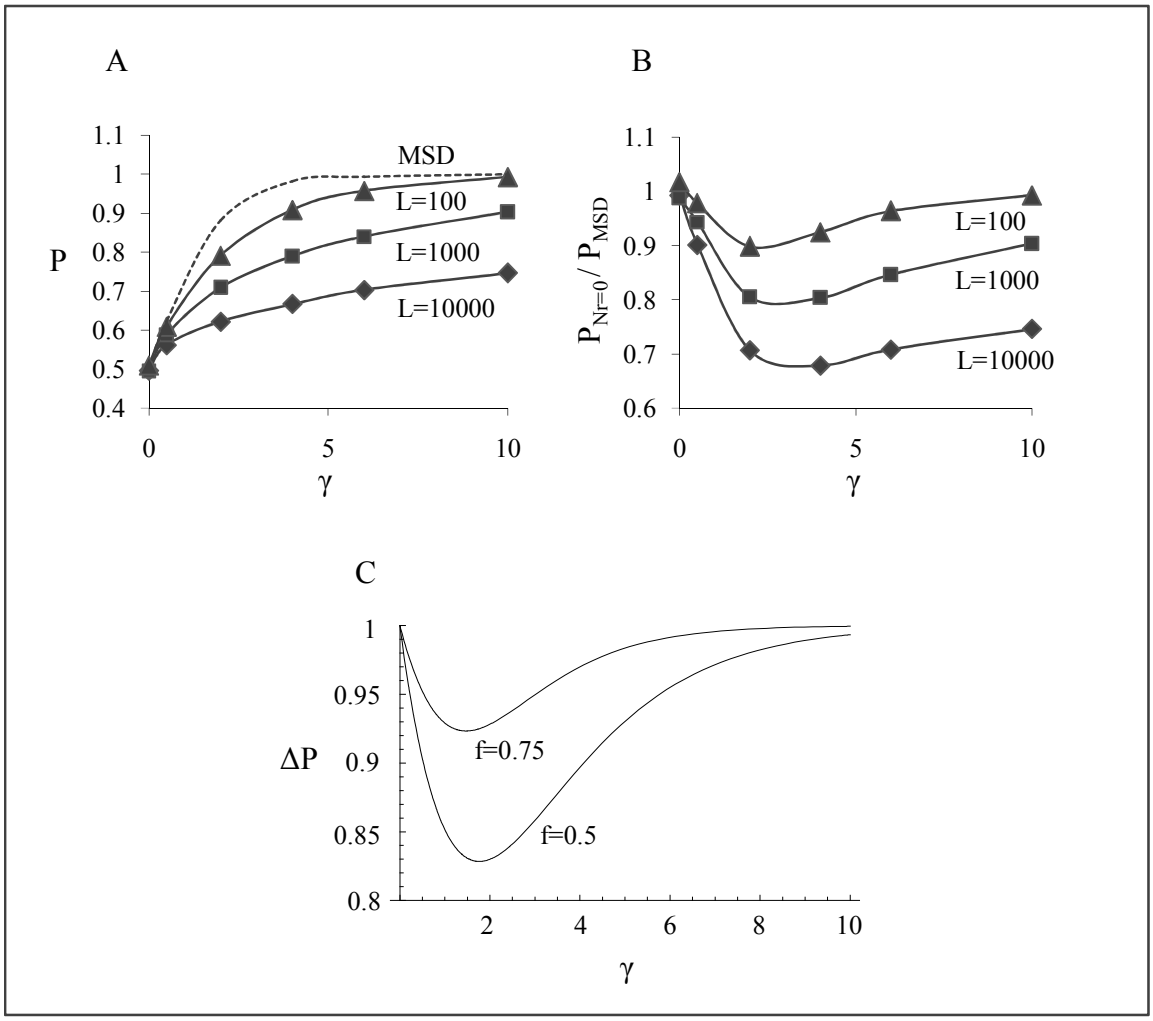

Figure 2.7. Dependence of the frequency of sites with preferred variant (P) on the number of sites under selection and selection intensity. A. Simulation results showing the influence of the number of sites under selection (L) and selection intensity (γ) with complete linkage (Nr=0) on the frequency of sites with preferred variant (P). B. Simulation results showing P relative to its expected value under mutation-selection-drift (MSD) model. C. MSD predictions of the expected reduction in P (ΔP) when N_e is reduced by a factor f (f=0.75 or f=0.5).

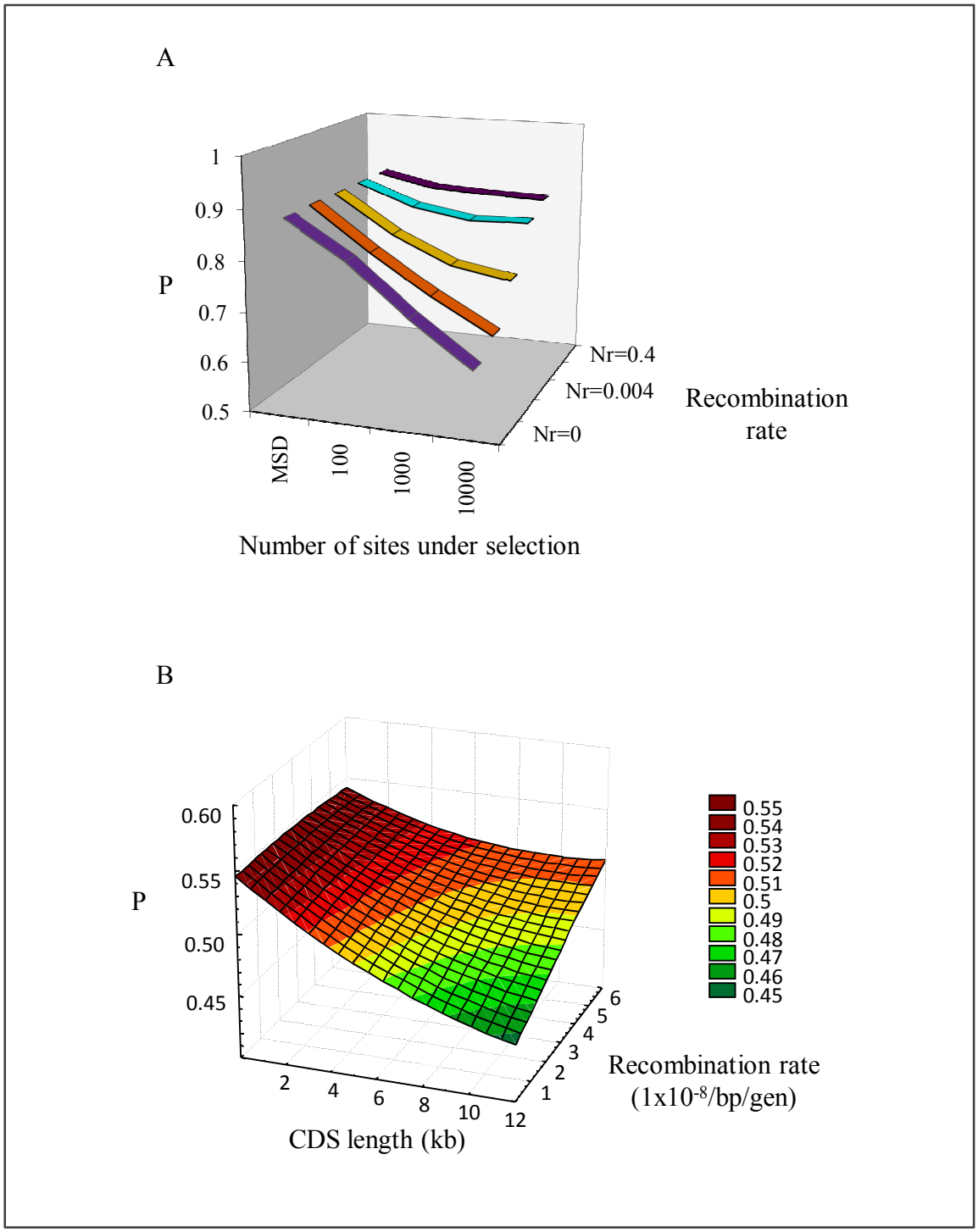

Figure 2.8. Dependence of the frequency of sites with preferred variant (P) on the number of sites under selection and recombination rate. A. Simulation results with $\gamma=2$ ($\gamma=4N_e s$). B. Observed relationship between the length of coding sequences (CDS), recombination rate and the frequency of preferred codons (P) in *Drosophila melanogaster* (Comeron *et al.* 2008).

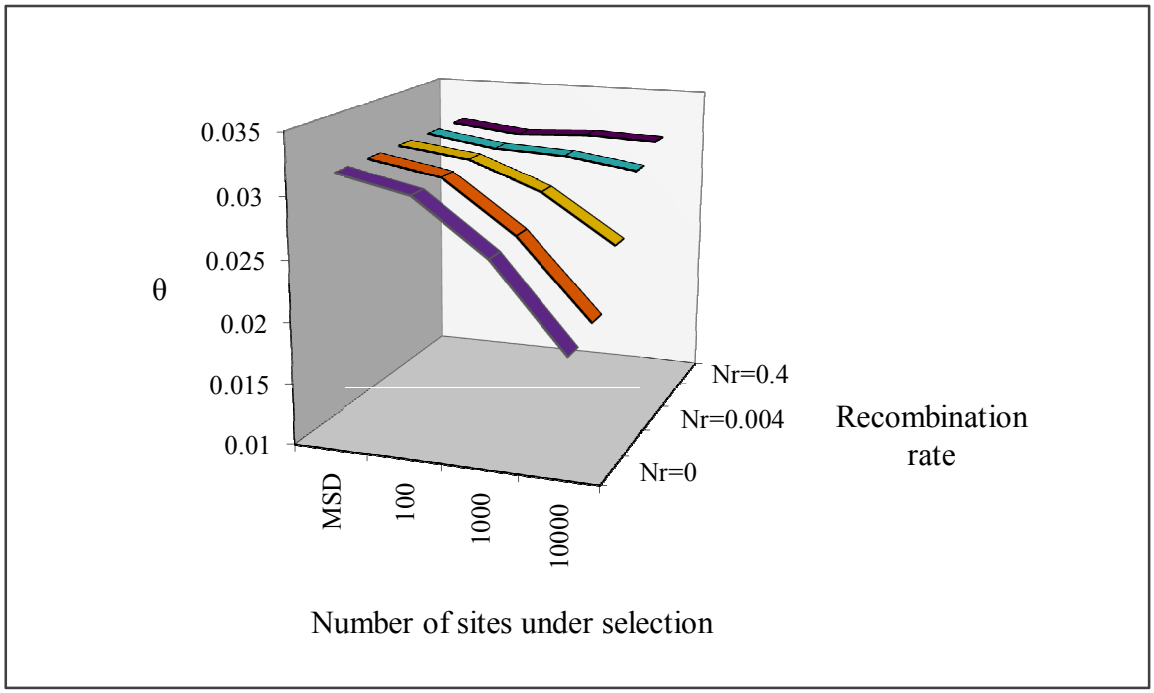

Figure 2.9. Simulation results showing the effect of the number of sites under selection (γ=2, γ=$4N_e s$) on polymorphism levels (θ).

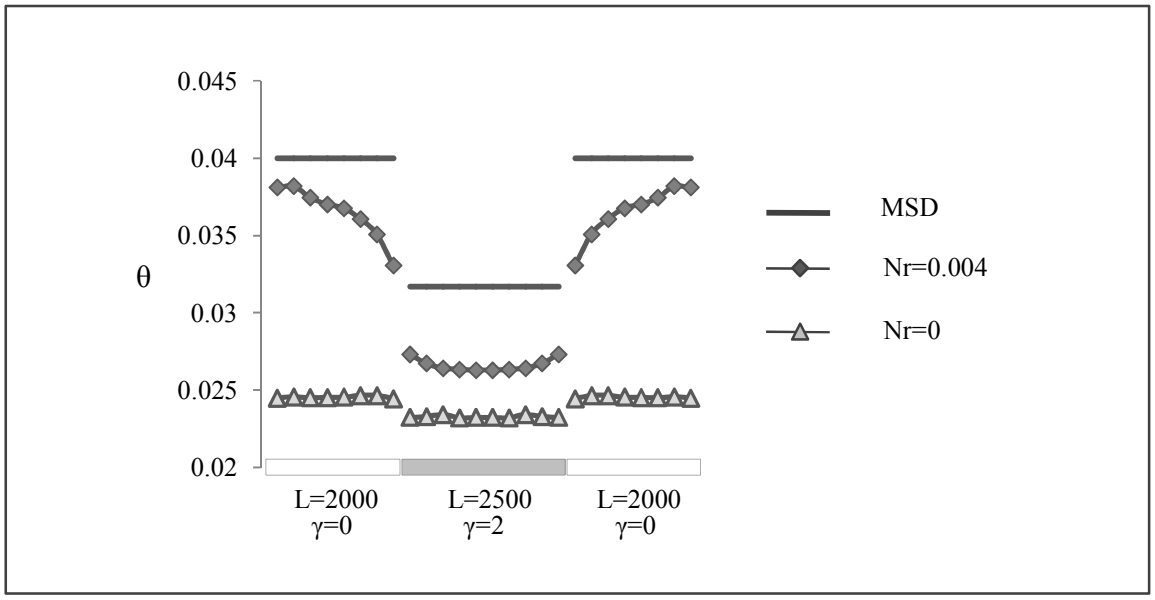

Figure 2.10. Simulation results showing heterogeneous distribution of polymorphism levels (θ) across a region of 2500 sites under uniform selection (γ=2) and at adjacent neutral sites.

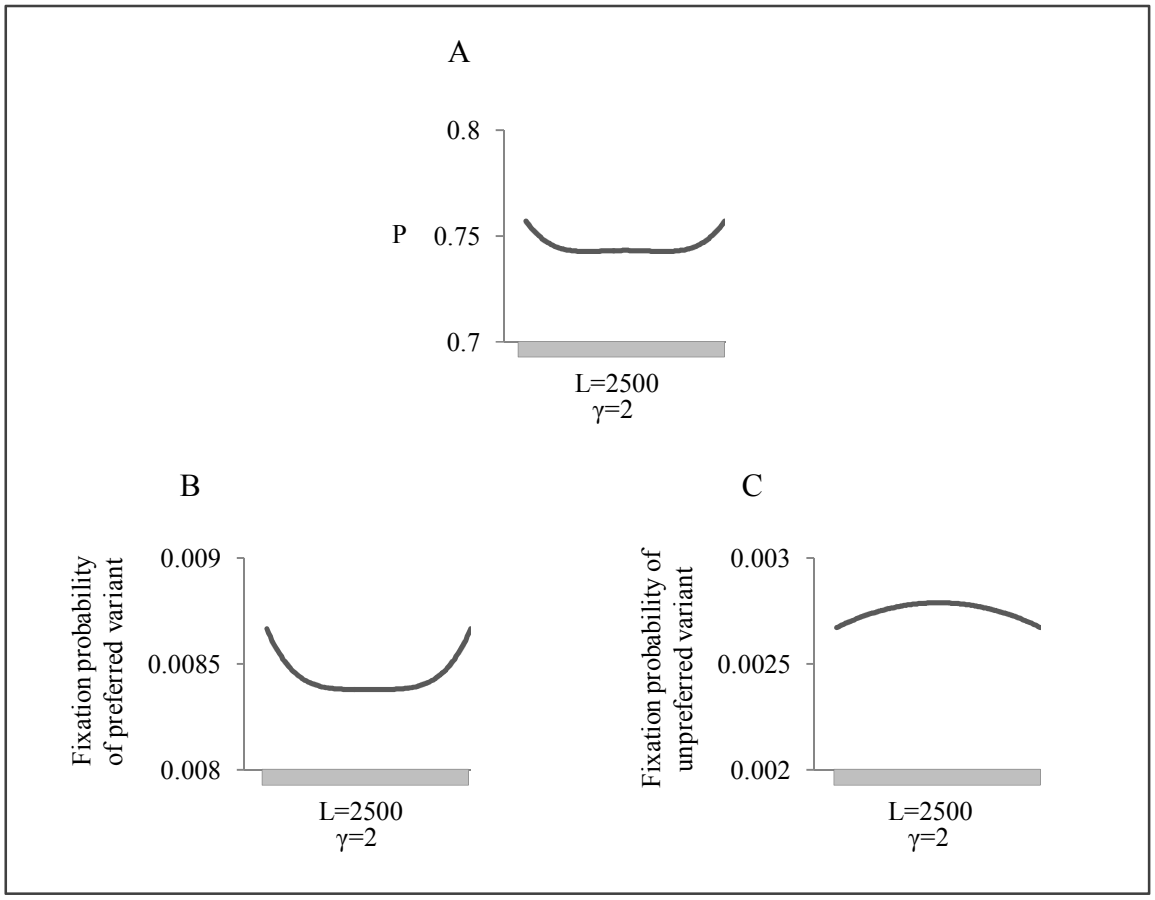

Figure 2.11. Simulation results showing heterogeneous distribution of effectiveness of selection across a region of 2500 sites under uniform selection ($\gamma = 2$) and partial linkage (Nr=0.004). A. Frequency of sites with preferred variant (P). B. Probability of fixation of preferred variant. C. Probability of fixation of unpreferred variant.

CHAPTER III

THE EFFECT OF PARTIAL LINKAGE ON ESTIMATES OF SELECTION USING MCDONALD-KREITMAN FRAMEWORK

3.1 Introduction

The wealth of DNA sequence data that became available in recent years is providing an increasing understanding of the tempo and mode of evolution in many species, mostly due to the analysis of population variation at the nucleotide level. Population genetic studies are quantifying the rate of deleterious and beneficial mutations as well as the magnitude of selection, depicting detailed landscapes of selective events across genomes and differentiating events at coding and non-coding regulatory sequences.

A common and particularly successful methodology to disentangle selective from neutral tendencies is the combination of measures of standing genetic variation within populations (polymorphism) and of fixed differences between species (divergence) at two classes of sites, one evolving neutrally, the other putatively under selection (Hudson *et al.* 1987; McDonald and Kreitman 1991; Sawyer and Hartl 1992). Such tests of neutral evolution involve the comparison of the ratio of polymorphism to divergence (r_{pd}) data between these two classes of sites assuming that non-selective parameters influencing r_{pd}, such as the effective population size (N_e) and divergence time, are equivalent for both classes of sites and therefore any possible difference can be assigned to selection. An application of this approach under McDonald-Kreitman (MK) framework can be used to estimate the fraction of mutations driven to fixation by positive selection (α) and/or the

average strength of selection (γ; $\gamma = 4N_e s$) expressed by the product of N_e and the selection coefficient s (Sawyer and Hartl 1992; Bustamante *et al.* 2002; Fay *et al.* 2002; Smith and Eyre-Walker 2002; Bierne and Eyre-Walker 2004).

Estimates of α based on the MK approach in *Drosophila* species suggest that adaptive protein evolution may be common enough to explain a substantial percentage of all amino acid differences between species, with estimates of α between 25 and 94% depending on the dataset, *Drosophila* species and methodology applied (Fay *et al.* 2002; Smith and Eyre-Walker 2002; Sawyer *et al.* 2003; Bierne and Eyre-Walker 2004; Andolfatto 2005; Welch 2006; Andolfatto 2007; Sawyer *et al.* 2007; Shapiro *et al.* 2007; Bachtrog 2008; Sella *et al.* 2009). Studies of *Drosophila* non-coding sequences not only reveal a large fraction of sites to be functionally relevant (Andolfatto 2005), but a high fraction of differences between species at introns (20%) and UTRs (60%) might also be driven by adaptive evolution (Kohn *et al.* 2004; Andolfatto 2005; Bachtrog 2008; Haddrill *et al.* 2008; Sella *et al.* 2009).

These estimates of γ and α under the MK framework assume evolutionary independence of mutations under the infinitely many sites model (Kimura 1969; Crow and Kimura 1970; Sawyer and Hartl 1992; Bustamante *et al.* 2002). However, in the presence of partial linkage, selection at one site affects levels of polymorphism and overall efficacy of selection at adjacent sites. This dual effect can be rationalized in terms of a reduction in N_e (the Hill-Robertson effect) (Robertson 1961; Hill and Robertson 1966; Felsenstein 1974; Birky and Walsh 1988). Classically, models of selection and linkage have assumed that the genomic units associated with linkage apply to large genomic regions involving many genes, with long-range variation in N_e across genomes

associated with differences in recombination rates. Population and evolutionary analyses in *Drosophila* have confirmed these long-range linkage effects, reporting reduced polymorphism (and r_{pd}) and estimates of γ in genomic regions with severely reduced or absent recombination in many species (Begun and Aquadro 1992; Kliman and Hey 1993; Langley *et al.* 1993; Munte *et al.* 1997; Takano 1998; Andolfatto and Przeworski 2001; Betancourt and Presgraves 2002; Bachtrog 2003; Presgraves 2005; Zhang and Parsch 2005; Haddrill *et al.* 2007; Larracuente *et al.* 2008; Betancourt *et al.* 2009).

Linkage effects can also act at a very local level in regions of intermediate or high recombination. Simulation studies have shown that local HR could explain variation in N_e between genes as well as across genes and influence adjacent sequences (Comeron *et al.* 1999; McVean and Charlesworth 2000; Comeron and Kreitman 2002; Loewe and Charlesworth 2007). Genomic and evolutionary analyses support detectable gene-specific and intragenic HR effects in *Drosophila* species in regions with high recombination, with exon length and intron-exon structures having a significant influence on protein evolution and codon usage bias (Comeron *et al.* 1999; Comeron and Kreitman 2000; Comeron and Kreitman 2002; Qin *et al.* 2004; Comeron and Guthrie 2005; Comeron *et al.* 2008; Larracuente *et al.* 2008).

Notably, studies that have estimated a large percentage of sites under positive selection clustered at exons and adjacent regulatory regions describe the very conditions where local linkage effects are predicted to occur even in regions with high recombination in species such *Drosophila* (Comeron and Kreitman 2002; Loewe and Charlesworth 2007). Thus, linkage effects acting at a very local scale generate scenarios

in which the two types of mutations compared under the MK framework might be under different local N_e challenging the application of a MK framework to non-intermingled types of mutations.

The simulation approach described in the previous chapter is applied here to investigate how spatial patterns of local reduction in linked neutral polymorphism affect r_{pd}, and consequently, estimates of γ and α. Simulations are expanded to explore complex scenarios that incorporate different combinations of neutral, beneficial and deleterious mutations, and incorporate both weak and moderate selection. The parameters of the simulations are chosen to mimic exons and adjacent non-coding sequences with high recombination rates that incorporate gene conversion. The results show that if selection (either positive or negative) is indeed pervasive, local linkage effects are almost unavoidable which in turn affect estimates of selection and the rates of adaptive evolution.

3.2 Materials and methods

A diploid population of N=200 individuals was simulated as described in Chapter II. Every chromosome consisted of L number of sites, with a variable fraction of sites evolving under neutral, positive or negative selection. Each generation, the total number of mutations and recombination events were drawn from a Poisson distribution with mean $2N\mu L$ and NrL respectively where μ is the mutation rate per site per generation and r is the recombination rate between adjacent sites per generation. Whenever neutral sites or weakly selected third codon positions were simulated, the mutation process allowed two allelic states with reversible mutations. The rest of the sites were assigned to be

under directional selection and reversible mutations were not allowed. Genic (semidominant) selection and multiplicative fitness over sites were assumed. Population parameters were estimated based on 10 randomly chosen chromosomes, taken every 5*N* generations.

Because the goal here is to quantify possible linkage effects in genomic regions where these effects have been assumed to be minimal or absent, most of the analyses focused on conditions equivalent to high rates of recombination for *D. melanogaster*. Assuming an N_e for *D. melanogaster* of 1.5-2x10^6 (McVean and Vieira 2001; Andolfatto 2007) and the average rates of crossover in region with high recombination ~2x10^{-8}/bp/generation (Charlesworth 1996; Comeron *et al.* 1999; Hey and Kliman 2002; Singh *et al.* 2005) in *D. melanogaster*, a scaled recombination rate (the product of population size and recombination rate *r*) of $Nr = 0.04$ per bp/generation represents an average region with high recombination for this species taking into account crossover only. To investigate the possible consequences of linkage in regions with very high recombination including both gene conversion and maximum crossover rates in a species like *D. melanogaster* ($r_{c\text{-}max} = 4\text{x}10^{-8}$) the simulations were performed with $Nr = 0.1$ between adjacent sites.

The intensity of selection ($\gamma = 4N_e s$) was estimated using an extension of the McDonald-Kreitman test implemented in the web-based program MKPRF (http://cbsuapps.tc.cornell.edu/mkprf.aspx) (Bustamante *et al.* 2002; Barrier *et al.* 2003; Bustamante *et al.* 2003). Application of the MK test to two regions requires polymorphism and divergence data from a putatively neutral region (used as a neutral standard) and from the region of interest that is under selection. The following

modification of the traditional terminology used in MK test is adopted here; for the neutral region, P_{neut} is defined as the number of polymorphic sites within a species and D_{neut} as the number of fixed sites between species. The corresponding numbers of polymorphism and divergence for the region of interest are given by $P_{interest}$ and $D_{interest}$. In order to evaluate the effects of linkage on selection estimates, polymorphism and divergence values obtained from simulations were used as input to MKPRF program. For each of the different classes of sites under study, polymorphism values were obtained by analyzing a sample of 10 random chromosomes while divergence estimates were obtained by comparing the population sample and a single randomly chosen sequence from a population that has evolved independently for 10 *N* generations. Under this scenario, simulations under complete neutrality for 500 sites (hereafter referred to as a *true* neutral standard) generate levels of polymorphism per site (θ) of 0.03 and a ratio of polymorphism to divergence (r_{pd}) of 0.78. The divergence between two randomly chosen sequences from each population is ~ 13%. These polymorphism and divergence values are similar to those estimated from empirical data for *D.melanogaster/D.simulans* comparison (Andolfatto 2005; Andolfatto 2007; Begun *et al.* 2007).

The proportion of sites fixed by positive selection (α) was estimated following Smith and Eyre-Walker (2002) using polymorphism and divergence values from simulations, with $\alpha = 1 - \frac{D_{\text{neut}} P_{\text{interest}}}{D_{\text{interest}} P_{\text{neut}}}$. This estimate is referred to as α_{EW}. The values of α were also obtained directly from the simulations, by counting the actual fraction of mutations driven to fixation that were under positive selection; this value describes the *true* α in the simulations and is referred to as α_{sim}.

In order to compare estimates of γ and α obtained using polymorphism and divergence values from simulations with those in the absence of linkage effects, we calculated the expected number of polymorphisms and fixed substitutions using sampling formulas of Sawyer and Hartl (1992). Assuming independence between sites, the same mutation rate for neutral and non-neutral mutations, semidominant selection, and random genetic drift, the expected values for neutral sites (*neut*) and sites under selection (*sel*) are given by

$$E(D_{neut}) = 4N_e\mu\, l_{neut}\left(t + \frac{1}{m} + \frac{1}{n}\right)$$

$$E(P_{neut}) = 4N_e\mu\, l_{neut}[L(m) + L(n)]$$

$$E(D_{sel}) = 4N_e\mu\, l_{sel}\left(\frac{\gamma}{1 - e^{-\gamma}}\right)[t + G(m) + G(n)]$$

$$E(P_{sel}) = 4N_e\mu\, l_{sel}\left(\frac{\gamma}{1 - e^{-\gamma}}\right)[F(m) + F(n)]$$

Where

$$L(m) = \sum_{i=1}^{m-1} \frac{1}{i}$$

$$G(m) = \int_0^1 (1 - x)^{m-1}\, \frac{1 - e^{-\gamma x}}{\gamma x}\, dx$$

$$F(m) = \int_0^1 \frac{1 - x^m - (1 - x)^m}{1 - x}\, \frac{1 - e^{-\gamma x}}{\gamma x}\, dx$$

$$\gamma = 4N_e s$$

In these equations, *m* and *n* stand for the number of sequences sampled from each species and *t* is the divergence time since the last common ancestor between species, in multiples of the haploid effective population size. We calculate *t* using r_{pd} for neutral sites

that were simulated to approximate divergence between *D.melanogaster* and *D.simulans* (~13% in our simulations). Neutral r_{pd} value of 0.78 gives *t* value of 2.5. This *t* value and assigned values of γ were used to calculate the expected numbers of polymorphism and divergence for sites under selection. Because simulated regions of interest contain both neutral sites and sites under selection, the expected values for neutral and non-neutral sites were combined to obtain expected polymorphism and divergence values for a complete region of interest.

3.3 Results and discussion

The main interest here is to investigate the consequences of local reduction in neutral polymorphism caused by linkage to sites under selection for estimates of selection ($\gamma=4N_e s$) and proportion of adaptive substitutions (α) within the McDonald-Kreitman framework. The results presented here are relevant to any empirical investigation that seeks to evaluate the strength of selection and the rate of adaptive evolution at non-coding DNA using this approach. More often than not, there is no *a priori* knowledge of the proportion of sites that evolve neutrally within non-coding DNA, and in these cases the complete region of interest is analyzed. In these analyses the influence of sites under selection on linked neutral sites cannot be determined, although such non-independence between sites can, in principle, affect overall estimates of γ and α. Here I report the general tendencies in estimates of γ and α in the presence of linkage effects. Situations in which non-coding DNA contains various fractions of neutral and non-neutral sites are examined and the analysis of neutral sites and complete regions is performed separately in order to identify the contribution of neutral sites to overall estimates. I begin with the

extreme scenarios in which γ and α at the non-coding region of interest are estimated using a *true* neutral standard, mimicking the choice of distal intergenic regions or pseudogene areas as a neutral standard. I then focus on estimates of γ and α at neutral regions that are adjacent to a coding sequence with the emphasis on local changes in neutral r_{pd} across the region. Finally, the last section describes the effect of linkage and the choice of neutral standard on estimates of γ and α at non-coding regions under selection.

3.3.1 Effect of linkage on estimates of γ and α at non-coding DNA when a true neutral region is used as a neutral standard

In this section I am focusing on estimates of selection (γ) and the proportion of adaptive substitution (α) at non-coding 500 bp region of interest using independently simulated *true* neutral region as a neutral standard. Three different scenarios are investigated where a region of interest contains neutral sites together with 1) 1% of sites under positive selection; 2) 10 to 66% of sites under negative selection and 3) 1% of sites under positive selection with 10% to 66% of sites under negative selection (see Table 3.1 for simulation parameters)

3.3.1.1 *Region of interest with neutral and advantageous mutations*

In spite of clear evidence for adaptive evolution in *Drosophila*, it has been difficult to obtain consistent estimates of the rate and the strength of adaptive substitutions (reviewed in Eyre-Walker 2006; Sella *et al.* 2009). Simulations were performed assuming that beneficial mutations are frequent (1% of sites is under positive

selection) and selection intensity ranges from weak to moderate ($4N_es$ between 6 and 100). These parameters most closely approximate estimates of Andolfatto (2007) who reports adaptive substitution rates of $7.5 * 10^{-10}$ per site per generation and selection intensity of $4N_es$ =80.

It is well known that fixation of advantageous mutations leads to a reduction in linked neutral polymorphism (Smith and Haigh 1974; Kaplan *et al.* 1989; Fay and Wu 2000). The width of the region affected by hitchhiking and the magnitude of reduction in neutral polymorphism depends on the recombination rate as well as the strength of advantageous mutations. Under conditions investigated here, even with high levels of recombination (Nr=0.1), neutral polymorphism is reduced by ~ 4% and ~ 44% in the presence of advantageous mutations with selection intensity $4N_es$ =15 and $4N_es$=100 respectively (Figure 3.1 A). This reduction in neutral polymorphism leads to reduction in r_{pd} at linked neutral sites since linked selection affects levels of neutral divergence much less than levels of polymorphism. As a consequence, neutral sites within the region of interest appear to evolve under positive selection (Figure 3.2 A, Table 3.2). The magnitude of positive selection estimates at neutral sites increases with the increase in the strength of advantageous mutations. For example, increasing selection strength of beneficial mutations from $4N_es$=15 to $4N_es$=100, increases positive selection estimates at neutral sites from 0.1 to 1.6 when Nr=0.1. As expected, reducing recombination rate increases the positive estimates of selection at neutral sites as a result of a reduction in linked neutral polymorphism (Figure 3.1 A, Figure 3.2 A).

Estimates of selection for sites under positive selection remain largely unaffected by linkage (Figure 3.2 B, Table 3.3). However, with stronger selection ($\gamma > 50$), selection

estimates are slightly lower than those expected if sites were fully independent providing evidence for weak interference between beneficial mutations.

Estimates of selection for the complete region are based on the contribution of both, neutral sites and sites under positive selection. But contribution of neutral sites to total polymorphism is substantially reduced by linkage leading to the reduction of polymorphism to divergence ratio for a complete region relative to that expected under independence of sites. This reduction leads to overestimates of positive selection (Figure 3.2 C, Table 3.4). For example, selection intensities for the region of interest containing beneficial mutations with $4N_e s$=100 assuming independence of sites are expected to be ~2, while in the presence of partial linkage this value is ~4 in case of Nr=0.1 and ~6 when recombination is reduced to Nr=0.04 (Figure 3.2 C). Notice that these estimates of selection are obtained assuming that all mutations within the region of interest have the same weak average intensity of selection (Sawyer and Hartl 1992; Bustamante *et al.* 2002) and do not reflect true selection intensity of advantageous mutations.

3.3.1.2 *Region of interest with neutral and deleterious mutations*

Given that most mutations are deleterious, it is worth examining if partial linkage within regions containing deleterious mutations can have detectable effects on estimates of selection. Loewe and Charlesworth (2007) showed that background selection alone can lead to local reduction in neutral polymorphism within a single gene. Similar reduction in neutral polymorphism is expected to occur within non-coding regions under purifying selection, potentially affecting estimates of selection. Following estimates of selection on

non-synonymous mutations reported by Loewe *et al*. (2006), simulations are performed with different fractions of deleterious mutations with selection intensity ranging from $4N_es$ =-6 to $4N_es$=-100 (Table 3.1).

First, similarly to cases with advantageous mutations described above, neutral sites within the regions containing deleterious mutations appear to evolve under positive selection (Table 3.5). Selection estimates for neutral sites increase with the increase in number of sites under negative selection. These positive estimates are the consequence of reduction in linked neutral polymorphism as a result of background selection (Figure 3.1 B). Under the conditions investigated, estimates of $4N_es$ range between 0.1 and 0.2 for neutral sites in the presence of 66% of sites under negative selection (Table 3.5). Clearly, these estimates are much smaller than estimates for neutral sites in the presence of advantageous mutations (see Figure 3.2 A) since deleterious mutations are quickly eliminated by selection.

Second, the effect of deleterious mutations on linked neutral polymorphism is most pronounced with moderate selection, with maximum reduction in neutral polymorphism by ~6% when 66% of sites are under $4N_es$ between -15 and -50 (Figure 3.1 B). This observation is consistent with the expected dynamics of deleterious mutations. Their effect on neutral polymorphism is expected to be strongest at intermediate selection intensities (Nordborg *et al.* 1996).

Estimates of selection for the complete region of interest remain in rather narrow range ($4N_es$ = -1 to + 0.1) and are largely independent of the number of sites under selection or the strength of deleterious mutations (Figure 3.3, Table 3.5). This is expected because the method for estimates of selection assumes that all mutations have the same

weak average intensity of selection. In the case of strongly deleterious mutations that contribute very little to polymorphism and hardly anything to divergence, selection estimates are expected to be close to zero (Figure 3.3 C). On the other hand, when 66% of sites are under weaker selection ($4N_e s$=-6), selection for the complete sequence is estimated to be close to -1. In this case, weakly selected sites contribute more to total polymorphism than they do to the reduction of linked neutral polymorphism, resulting in higher r_{pd} than that of a *true* neutral standard, generating negative estimates of selection (Figure 3.3 A).

Reduction of linked neutral polymorphism below true neutral levels also leads to estimates of selection that are always less negative than expected if sites were independent (Figure 3.3, Table 3.5). However, because background selection has much smaller effect on linked neutral polymorphism than does hitchhiking (Figure 3.1 A,B), the deviation of overall selection estimates from the expectations under independence of sites is less for regions with deleterious mutations than for regions with advantageous mutations.

3.3.1.3 *Region of interest with neutral, deleterious and advantageous mutations*

Recent studies argue for pervasive natural selection in *Drosophila* including evidence for substantial adaptive evolution (reviewed in Sella *et al.* 2009). In light of these findings, perhaps the most realistic case to consider is the one that incorporates neutral, deleterious as well as advantageous mutations. Introduction of 1% of sites under positive selection in the presence of deleterious mutations yields selection estimates at

neutral sites that are similar to selection estimates in the presence of beneficial mutations only (Table 3.2). This is not surprising given that background selection has much smaller effect on neutral polymorphism than does hitchhiking. Increasing number of deleterious sites increases estimates of positive selection at neutral sites when such mutations cause an additional decrease in neutral polymorphism compared to that caused by advantageous mutations only (Table 3.2, Figure 3.1 C).

Estimates of selection at positive sites are only slightly affected by the presence of deleterious mutations (Table 3.3). The difference between selection estimates in the presence and absence of deleterious mutations suggests that there is weak interference between deleterious and advantageous mutations that leads to the reduction in estimates of selection at positive sites in the presence of deleterious mutations (Table 3.3).

Selection estimates for the complete region of interest reflect the contribution of neutral, advantageous and deleterious mutations to total polymorphism and divergence. When the region of interest contains a large fraction of sites under weak negative selection, deleterious mutations contribute more to total polymorphism than advantageous mutations to total divergence leading to larger ratio of polymorphism to divergence for a complete region relative to neutral standard. This increase in r_{pd} generates negative estimates of selection for a complete region of interest (Figure 3.4 A). On the other hand, deleterious mutations under stronger selection will contribute less to total polymorphism than advantageous mutations to total divergence generating lower r_{pd} value than that of neutral standard resulting in positive estimates of selection for a complete region (Figure 3.4 B and C, Table 3.4). However, because linkage leads to reduction in neutral polymorphism, r_{pd} for the complete region will be always smaller

than that expected under independence of sites leading to overestimates of positive selection and underestimates of negative selection (Figure 3.4, Table 3.4).

In summary, the estimates of selection in the presence of partial linkage deviate from those expected under the assumption of site independence primarily as a result of reduction in neutral polymorphism in the presence of selection. It is worth pointing out once again that selection estimates obtained by MK test implemented in MKPRF can only serve as rough indicators of the direction of selection, and hardly of its magnitude because of the underlying assumption that all mutations within the region have the same average intensity and direction of selection (Sawyer and Hartl 1992; Bustamante *et al.* 2002). Whenever the region contains a combination of advantageous and deleterious mutations, the overall estimates of selection reflect the presence of positive selection when deleterious mutations are strongly deleterious but will miss the action of positive selection in the presence of many slightly deleterious mutations (Figure 3.4, Table 3.4) as has been pointed out previously.

3.3.1.4 *Effect of partial linkage on estimates of the proportion of adaptive substitutions*

Another way to assess the magnitude of adaptive evolution is to estimate the proportion of substitutions fixed by positive selection. Estimates of the proportion of adaptive substitutions rely on the assumption that mutations are neutral, strongly deleterious or strongly advantageous (Smith and Eyre-Walker 2002). In this case, reduced polymorphism to divergence ratio relative to the neutral standard should indicate the excess of divergence due to fixation of advantageous mutations. Given the results

presented above, it is clear that reduction in linked neutral polymorphism will contribute to further reduction in r_{pd} leading to overestimates of α. Simulation results presented below suggest that these expected overestimates of α can indeed be considerable in regions containing advantageous mutations. In regions with both, advantageous and deleterious mutations, deviations of α estimates from real values depend greatly on the number and selection intensity of the deleterious mutations.

To evaluate how linkage effects influence commonly used estimate of α ($\alpha_{EW} = 1 - \frac{D_{neut}P_{interest}}{D_{interest}P_{neut}}$), α_{EW} is compared with the real value of α obtained directly from simulations (α_{sim}). Figure 3.5 A and B shows both estimates of α for a complete 500 bp region containing 1% of sites under positive selection using an independently simulated neutral region as a neutral standard. Given these conditions, α_{EW} tends to overestimate the true proportion of adaptive substitutions, α_{sim}. With Nr =0.1, α_{sim} is overestimated by ~ 45% when advantageous mutations with γ=100 are present. Reducing recombination rate to Nr=0.04 generates α_{EW} values that overestimate the true proportion of adaptive substitutions by ~ 60% (Figure 3.5 A,B; Table 3.6). These inflated values of α as measured by α_{EW} are the direct consequence of the reduction in neutral polymorphism within the region of interest as a result of linkage to sites under positive selection. As in the case of selection overestimates described in the previous section, neutral sites linked to sites under selection have lower r_{pd} than true neutral sites used as a neutral standard and if analyzed separately, produce positive estimates of α (Figure 3.5 C; Table 3.6). These neutral sites that appear to evolve under positive selection cause the overestimates of α. Estimates of α and selection intensity are expected to be inflated whenever fixation of beneficial mutations generates local reduction in linked neutral polymorphism. Even

with a single site under selection within 500 bp region, the same trend is evident (Figure 3.6). The above results, however, apply only to regions containing neutral and advantageous mutations. Such regions are unlikely to represent an average non-coding region given the evidence for reduced levels of divergence and presence of deleterious mutations in non-coding regions (Andolfatto 2005; Haddrill *et al.* 2008). Estimates of α for regions containing both, advantageous and deleterious mutations are shown in Figure 3.7 and Table 3.7. In the presence of deleterious mutations, the difference between α_{EW} and α_{sim} depends greatly on the number and selection intensity of deleterious mutations. The magnitude and direction of the deviation from α_{sim} is controlled by the contribution of deleterious sites to total polymorphism. Accurate estimates of α can be obtained whenever the reduction in neutral polymorphism due to linkage to sites under selection is compensated by polymorphisms contributed by deleterious mutations (Figure 3.7 B). When deleterious mutation are under weak selection ($4N_es$=-6), α_{sim} is consistently underestimated because deleterious mutations contribute substantially to total polymorphism (Figure 3.7 A). But when selection intensity of deleterious mutations is large, such mutations contribute little to total polymorphism and divergence and the total polymorphism is controlled mostly by neutral mutations. But neutral polymorphism is reduced relative to that expected if sites were independent as a result of linkage to advantageous mutations generating overestimates of the true proportion of adaptive substitutions (Figure 3.7 C).

In summary, the degree of α overestimate depends on the distribution of fitness effects within the region of interest. Including polymorphism contribution from neutral sites that is reduced by background selection or hitchhiking will inflate α estimates while

including polymorphism contribution from weakly selected sites will result in underestimates of α. The larger the fraction of neutral sites and the smaller the fraction of sites under weak negative selection the larger the overestimate of α will be. Since we do not know the distribution of fitness effects in the region of non-coding DNA of interest, it is problematic to know to what degree α_{EW} overestimates or underestimates α_{sim}. It has been suggested that in order to avoid underestimates of α, rare polymorphisms should be excluded from the analysis (Fay *et al.* 2001; Andolfatto 2005; Charlesworth and Eyre-Walker 2008). This solution, however, will not always lead to accurate estimates of α since in some cases inclusion of the polymorphism contribution from sites under negative selection counterbalances the reduction in neutral polymorphism and hence, the exclusion of rare polymorphism may lead to overestimates of α.

3.3.2 Effect of linkage on estimates of γ and α at neutral non-coding DNA adjacent to the coding sequence

In the presence of recombination, the magnitude of the reduction in linked neutral polymorphism is expected to decline as the distance from sites under selection increases. This dynamics at neutral sites predicts stronger linkage effects at neutral regions immediately adjacent to the coding region than at neutral regions further away. In this section I am interested in the estimates of selection (γ) and the proportion of adaptive substitution (α) at a non-coding neutral region that is adjacent to the coding sequence focusing on local changes in r_{pd} across the simulated region for adjacent neutral and synonymous sites. The results of the simulations reported here are more realistic than those described in the previous section since, in practice, we do not have access to *true*

neutral standard, but instead use synonymous sites as a proxy of neutral evolution (although estimates based on *true* neutral are included here for comparison). Here I explore how the choice of neutral standard within the coding sequence (neutral 3rd codon position or weakly selected 3rd codon position) affects estimates of γ and α at an adjacent neutral region. Two sets of simulations are performed in which 1kb of neutral sequence is followed by 1 kb of coding sequence which includes sites under negative (γ=-50), or sites under negative (γ=-50) and positive (1%, γ=50) selection with 3rd codon position under weak selection (γ=2), or neutral (γ=0). In order to have more realistic distribution of fitness effects, simulated coding regions include weakly selected mutations (12% of non-synonymous sites, γ=2) in addition to strongly deleterious or advantageous mutations (Figure 3.8).

3.3.2.1 *Effect of selection on linked neutral polymorphism*

In spite of high recombination rates (Nr=0.1), reduction in polymorphism at neutral sites adjacent to sites under selection is detectable across 1 kb of the simulated region. Neutral polymorphism levels are reduced by ~ 5% and ~13% at sites immediately adjacent to the coding region without and with sites under positive selection respectively. The reduction of neutral polymorphism is determined by the strength of selection at 1st and 2nd position, but is largely unaffected by whether 3rd position is neutral or weakly selected. This local and non-uniform reduction in polymorphism is reflected in local changes in r_{pd} that predicts different and non-zero estimates of γ and α for neutral regions located immediately next to or further away from the coding sequence (Figure 3.8. C-F).

When neutral sites are interspersed within the coding sequence (neutral 3rd codon position) the polymorphism is reduced by ~10 and ~25% on average for the coding region without and with sites under positive selection respectively. Thus, interspersed neutral sites have lower r_{pd} than adjacent neutral sites and both values are lower than r_{pd} of *true* neutral sites that are unaffected by selection (Figure 3.8, A, B). For sites under weak selection in the absence of linkage, both polymorphism and divergence are lower than neutral values that results in higher r_{pd} than for neutral sites. Linkage to sites under negative or positive selection leads to reduction in polymorphism and increase in divergence, lowering r_{pd} relative to that expected under the assumption of site independence. However, with high levels of recombination, even reduced r_{pd} at weakly selected sites can be higher than adjacent neutral r_{pd} in the presence of deleterious non-synonymous mutations (Figure 3.8 A). When the coding sequence contains sites under positive selection, r_{pd} at weakly selected sites is lower than adjacent neutral r_{pd} (Figure 3.8 B). Consequently, estimates of selection and proportion of adaptive substitutions at adjacent neutral sites using weakly selected third position will differ depending on whether the coding sequence contains sites under positive selection or not (Figure 3.8, C-F).

In either case, whether 3rd position is neutral or under weak selection, there is a "center effect" with more pronounced reduction in r_{pd} at the center of the coding sequence than at the edges. Approximately 3% and 5% reduction in r_{pd} at synonymous sites is detected for a central region relative to the lateral regions of the coding sequence without and with sites under positive selection respectively. These results are consistent with those reported previously (Nordborg *et al.* 1996; McVean and Charlesworth 2000;

Comeron and Kreitman 2002; Comeron and Guthrie 2005; Loewe and Charlesworth 2007) and indicate that selection leads to reduction in local N_e: the magnitude of the change in N_e varies locally across the region even in areas of constant high recombination rates depending on the type and the strength of selection.

The change in local N_e also varies with the number of sites under selection and can be caused by deleterious or advantageous mutations alone in the absence of non-synonymous weakly selected sites within the coding sequence. For example, Figure 3.9 shows changes in polymorphism levels for adjacent neutral and neutral synonymous sites within coding sequence of different length (0.5 kb and 2.5 kb) containing either sites under negative selection (66% of sites, γ=-50; Figure 3.9 A,B) or sites under negative and positive selection (1% positive, γ=50; Figure 3.9 C,D). Clearly, the length of the coding sequence is another important factor that contributes to changes in N_e and, consequently, neutral r_{pd} ratios across the region. Such variation in N_e is not taken into account by MK approach but is expected to influence the estimates of γ and α.

3.3.2.2 *Estimates of γ and α at adjacent neutral region*

Local changes in neutral $r_{pd,}$ caused by the action of selection nearby, alter the null expectations for estimates of γ *and* α at neutral regions. Depending on the choice of neutral standard, neutral regions adjacent to the coding sequence may appear to be the subject of weak positive (when r_{pd} (adjacent)< r_{pd} (neutral standard)) or weak negative selection (when r_{pd} (adjacent) > r_{pd} (neutral standard)). The precise magnitude of positive and negative estimates is determined by the type and strength of selection, by the number of sites under selection and by the choice of neutral standard. For the case presented in

Figure 3.8, two trends are evident: 1) when true neutral standard is used, estimates of γ and α at adjacent neutral non-coding region will always be positive, and 2) when neutral 3rd position is used as a neutral standard, estimates of γ and α will always be negative. When 3rd position is under selection (as frequently is the case), either negative or positive estimates can be obtained depending on the precise distribution of fitness effects within the coding region (Figure 3.8 C-F). These results suggest that it is almost unavoidable that truly neutral sites will appear as either weakly positive or weakly negative when sites under selection are present nearby.

3.3.3 Effect of linkage on estimates of γ and α at non-coding DNA under selection

The final case to be considered is the effect of linkage and the choice of neutral standard on estimates of selection and proportion of adaptive substitutions at non-coding regions under selection. This is the situation most frequently investigated in recent experimental studies (Andolfatto 2005; Begun *et al.* 2007; Haddrill *et al.* 2008) in which synonymous sites are used as a proxy for neutral sites and are used as neutral standard to assess the strength of selection and the extent of adaptive evolution at non-coding regions. Here I evaluate the extent of variation in estimates of γ and α for non-coding regions that are obtained using different neutral standards. I investigate two non-coding regions: 1) non-coding region (L=500 bp) with 33% of sites under negative selection (γ =-50), and 2) non-coding region (L=500 bp) with 33% of sites under negative selection (γ =-50) and 1% of sites under positive selection (γ =50). Estimates of γ and α at these regions are obtained using either independently simulated true neutral standard or

synonymous sites that are either neutral or under weak selection from the adjacent coding regions described in section 3.3.2: coding regions (L=1 kb) that either contain only deleterious and weakly selected non-synonymous sites or also include 1% of sites under positive selection (γ=50).

The action of positive selection is inferred whenever r_{pd} of the region of interest is lower than r_{pd} of the putatively neutral standard. This condition can be met in the absence of sites under positive selection in two ways. First, positive selection estimates (based on *true* neutral standard) can be obtained as a result of background selection effect on linked neutral polymorphism that reduces r_{pd} of non-coding DNA below that of the neutral standard (not shown here). Second, positive selection estimates in the absence of advantageous mutations can be obtained when the 3rd codon position used as a neutral standard is under weak selection and comes from the coding region that itself does not contain sites under positive selection (Figure 3.10). Under conditions investigated here, estimates of α can be as high as 19%. On the other hand, if the coding region is under positive selection, negative estimates of γ and α are obtained (Figure 3.10). At the practical level, inferring the magnitude of selection using synonymous sites should be done after careful consideration of possible selection at synonymous sites themselves as well as at non-synonymous sites since the latter may substantially influence r_{pd} of synonymous sites. Overall, estimates of the proportion of adaptive substitutions for non-coding region containing only deleterious and neutral sites ranges between -23% and +19% depending on the choice of neutral standard (Figure 3.10 A). Estimates of selection follow similar trend (Figure 3.10 B).

Similarly, for non-coding region containing sites under positive selection, estimates of γ and α also vary, ranging from 27% to 50% for α values when the real α obtained from simulations is 34% (Figure 3.10 A, B). Although the extent of underestimate or overestimate of selection is not large in this case, it illustrates the sensitivity of selection estimates to linkage effects. Given that relatively little is known about the distribution of fitness effects within the investigated regions (especially the ones involving positive selection), it is very difficult to predict the extent to which selection is underestimated/overestimated in the empirical studies.

3.4 Summary

The main purpose of this study is to evaluate commonly used method for estimates of γ and α at non-coding DNA based on the extension of the McDonald-Ktreitman test (McDonald and Kreitman 1991; Sawyer and Hartl 1992; Bustamante *et al.* 2002). Previous studies investigated potential difficulties with application of this test focusing on the effects of changes in population size (Eyre-Walker 2002; Eyre-Walker and Keightley 2009), presence of weakly deleterious mutations (Fay *et al.* 2002; Charlesworth and Eyre-Walker 2008; Eyre-Walker and Keightley 2009) and problems associated with combining data from multiple loci (Smith and Eyre-Walker 2002; Welch 2006; Shapiro *et al.* 2007).

In this study I focused on the fundamental assumption of the original McDonald-Kreitman test: the ratio of polymorphism to divergence at neutral synonymous sites is equal to the ratio of polymorphism to divergence at neutral nonsynonymous sites. This assumption is appropriate within the original McDonald-Kreitman framework when two

classes of sites are intermingled since linked selection will affect neutral synonymous and neutral non-synonymous polymorphism equally. However, this assumption is violated within extended version of McDonald-Kreitman test when neutral and selected classes of sites are spatially separated. In this case, linked selection will affect neutral polymorphism of the two regions differently, generating different ratios of polymorphism to divergence for a neutral region that is used as neutral standard and neutral sites within the region of interest. I now briefly summarize how this spatially heterogeneous effect of linked selection on neutral polymorphism affects estimates of selection and proportion of adaptive substitutions.

Unequal ratios of polymorphism to divergence at neutral sites within neutral standard and region of interest can be generated in two ways: 1) neutral r_{pd} within the region of interest is reduced as a result of linked selection; 2) neutral r_{pd} of the neutral standard is reduced as a result of linked selection when synonymous neutral sites are used as neutral standard or r_{pd} of the neutral standard is inflated when synonymous sites used as neutral standard are under weak selection.

Whenever neutral standard is represented by *true* neutral sites, r_{pd} for this region will not be affected by sites under selection. However, the region of interest is likely to contain fractions of sites under different selection intensity as well as a fraction of neutral sites. Polymorphism at neutral sites will be always reduced to various degrees depending on how many sites under selection are nearby and the strength of selection. If the region of interest contains neutral and deleterious sites, negative selection for the complete region will be always underestimated. If the region of interest contains neutral and advantageous sites, positive selection and the proportion of adaptive substitutions will be

overestimated. Whenever the region of interest contains neutral, deleterious and advantageous sites, the proportion of adaptive substitutions can be overestimated or underestimated depending on contribution of deleterious sites to total polymorphism. In some cases, polymorphisms contributed by deleterious sites will compensate for the reduction in polymorphism at neutral sites resulting in more or less accurate estimates of α. This implies that removal of rare polymorphisms will not always result in more accurate estimates of α because after the exclusion of polymorphism contribution from deleterious sites, analysis of the data set with reduced levels of neutral polymorphism will lead to inflated estimates of α.

When synonymous sites are used as neutral standard, the situation is more complicated because both regions can be affected by linked selection to a different degree. When synonymous sites are neutral, polymorphism is reduced as a result of linkage to sites under selection within the coding region. The degree of this reduction depends on the length of the coding sequence and the type and the strength of selection acting on the coding sequence. This means that r_{pd} of this "neutral" region will be reduced relatively to *true* neutral region leading to underestimate of positive selection and overestimate of negative selection. Recent studies report reduction in synonymous polymorphism by 10-20% as a result of either hitchhiking (Andolfatto 2007; Shapiro *et al.* 2007; Bachtrog 2008) or background selection (Loewe and Charlesworth 2007). These values are similar to those generated by simulations in this study: background selection reduces polymorphism at neutral 3rd codon position by up to 13% when the coding region is long (Figure 3.9 A, B). In the presence of adaptive mutations, synonymous polymorphism is reduced up to 35% (Figure 3.9 C, D). These results suggest

that the reduction of polymorphism at synonymous sites might be a general feature of *Drosophila* genome implying that r_{pd} of synonymous sites is unlikely to be that of *true* neutral sites. The two effects of linked selection acting on the region of interest and the neutral standard can in principle generate similar local N_e, but it is difficult to know when this is the case without knowing the number of sites under selection and the strength of selection within putatively neutral and putatively selected regions.

However, synonymous sites are frequently under weak selection (Kliman and Hey 1993; Akashi 1995; Comeron *et al.* 1999; Comeron 2006). In the absence of linkage, weakly selected synonymous sites have higher r_{pd} relative to neutral sites leading to overestimates of positive selection and underestimates of negative selection. Analyzed regions are usually sampled from areas of high recombination to minimize linkage effects, but these areas are also areas with most efficient selection and thus, stronger codon bias (Kliman and Hey 1993; Comeron *et al.* 1999). Adequate tests must be performed to make sure that synonymous sites are not subject to weak selection. For example, Andolfatto (2005) performs reanalysis of his data using a subset of coding regions with the lowest codon bias, minimizing the possibility that estimates of α are affected by weak selection at synonymous sites. It is possible to use only preferred to preferred or unpreferred to unpreferred synonymous changes as neutral standard (Haddrill *et al.* 2008). However, in the presence of non-synonymous sites under selection even these neutral changes are unlikely to have neutral r_{pd}. It appears that 3rd codon position rarely conforms to neutral expectations and is likely to bias selection estimates.

Simulation results presented here also imply that genomic location of non-coding DNA may affect estimates of selection. It is expected that regions immediately upstream

of the coding sequences will be affected by sites under selection within the adjacent exon. This effect is likely to be detectable only when linkage effects are strong. For example, in the case shown in Figure 3.10, the location of the non-coding sequence has little effect on estimates of selection since considerable linkage effects are only detected immediately upstream of the coding sequence (Figure 3.8). However, more pronounced reduction in neutral polymorphism adjacent to the coding sequence (as in Figure 3.9 D) is expected to alter estimates of γ and α at non-coding regions that are adjacent to the coding region relatively to the estimates for the same non-coding regions which are located further upstream of the coding sequence. One possible artifact of this situation is the inflation of γ and α estimates at UTRs that results solely from proximal location of UTRs to the coding sequence.

The results presented in this chapter demonstrate the consequences of selection operating in finite populations with loose linkage. Even in areas of high recombination for *Drosophila melanogaster*, local linkage effects are still detectable. Linked neutral polymorphism is always reduced locally resulting in heterogeneous distribution of neutral polymorphism to divergence ratios, introducing unanticipated problem for the application of McDonald-Kreitman test to two spatially separated regions.

Table 3.1. Simulation parameters for non-coding regions

L=500 $4N_e\mu$=0.03		
Advantageous (Nr=0.1 and 0.04)	% adv. sites	1
	$4N_es$	6, 15, 30,50,100
Deleterious (Nr=0.1)	% del. sites	10, 33, 66
	$4N_es$	-6,-15,-30,-50,-100
Deleterious and Advantageous (Nr=0.1)	% del. sites	10, 33, 66
	% adv. sites	1
	$4N_es$ (del./adv.)	-6/6, -15/15, -30/30, -50/50, -100/100

Table 3.2. Estimates of selection for neutral sites in the presence of 1% of sites under positive selection or in the presence of 1% of sites under positive selection with different fractions of sites under negative selection

	$4N_es$	% adv	$4N_es$	% del with 1% adv		
		1%		10%	33%	66%
Simulation	6	0.035	6/-6	0.031	0.117	0.165
IndSites		0.025		0.025	0.025	0.024
Simulation	15	0.103	15/-15	0.120	0.173	0.302
IndSites		0.024		0.025	0.025	0.025
Simulation	30	0.320	30/-30	0.311	0.371	0.463
IndSites		0.026		0.025	0.024	0.026
Simulation	50	0.680	50/-50	0.670	0.699	0.788
IndSites		0.025		0.025	0.024	0.026
Simulation	100	1.612	100/-100	1.658	1.721	1.678
IndSites		0.026		0.025	0.025	0.025

Table 3.3. Estimates of selection for sites under positive selection in the presence of 1% of sites under positive selection or in the presence of 1% of sites under positive selection with different fractions of sites under negative selection

	$4N_es$	% adv	$4N_es$	% del with 1% adv		
		1%		10%	33%	66%
Simulation	6	6.0	6/-6	6.2	5.9	5.8
IndSites		6.0		6.0	6.0	6.0
Simulation	15	15.4	15/-15	14.8	14.7	14.7
IndSites		15.0		15.0	15.0	15.0
Simulation	30	29.6	30/-30	28.4	29.1	29.0
IndSites		30.1		30.1	30.1	30.1
Simulation	50	47.4	50/-50	48.3	45.8	46.1
IndSites		50.0		50.0	50.1	50.1
Simulation	100	89.2	100/-100	86.0	87.9	87.0
IndSites		99.8		99.8	99.7	99.7

Table 3.4. Estimates of selection for a complete region (L=500 bp) in the presence of 1% of sites under positive selection or in the presence of 1% of sites under positive selection with different fractions of sites under negative selection

	$4N_es$	% adv	$4N_es$	% del with 1% adv		
		1%		10%	33%	66%
Simulation	6	0.152	6/-6	0.034	-0.240	-1.103
IndSites		0.141		0.026	-0.312	-1.166
Simulation	15	0.432	15/-15	0.409	0.320	0.001
IndSites		0.358		0.322	0.212	-0.105
Simulation	30	1.006	30/-30	1.019	1.139	1.424
IndSites		0.705		0.733	0.825	1.119
Simulation	50	1.861	50/-50	1.928	2.210	3.125
IndSites		1.141		1.225	1.502	2.413
Simulation	100	4.274	100/-100	4.545	5.399	7.846
IndSites		2.164		2.358	3.004	5.235

Table 3.5. Estimates of selection for neutral sites and a complete region (L=500 bp) in the presence of different fractions of sites under negative selection

	Neutral Sites				All Sites		
		% del			% del		
	$4N_eS$	10%	33%	66%	10%	33%	66%
Simulation	-6	0.027	0.098	0.171	-0.099	-0.423	-0.980
IndSites		0.025	0.026	0.026	-0.100	-0.484	-1.469
Simulation	-15	0.030	0.092	0.228	-0.040	-0.191	-0.804
IndSites		0.027	0.024	0.025	-0.042	-0.260	-0.937
Simulation	-30	0.046	0.093	0.200	0.007	-0.071	-0.431
IndSites		0.025	0.026	0.024	-0.013	-0.139	-0.565
Simulation	-50	0.019	0.131	0.205	-0.004	0.026	-0.198
IndSites		0.026	0.025	0.026	0.002	-0.078	-0.364
Simulation	-100	0.043	0.110	0.126	0.033	0.062	-0.067
IndSites		0.026	0.024	0.027	0.015	-0.027	-0.185

Table 3.6. Estimates of the proportion of adaptive substitutions (α) in the presence of 1% of sites under positive selection

	All Sites Nr=0.1				
$4N_eS$	6	15	30	50	100
α_{sim}	0.048	0.103	0.181	0.263	0.408
α_{EW}	0.035	0.106	0.230	0.369	0.593
	All Sites Nr=0.04				
$4N_eS$	6	15	30	50	100
α_{sim}	0.048	0.103	0.183	0.268	0.410
α_{EW}	0.048	0.136	0.286	0.442	0.671
	Neutral Sites				
$4N_eS$	6	15	30	50	100
α_{EW} Nr=0.1	0.003	0.022	0.079	0.163	0.333
α_{EW} Nr=0.04	0.016	0.055	0.146	0.258	0.464

Table 3.7. Estimates of the proportion of adaptive substitutions (α) in the presence of 1% of sites under positive selection and different fractions of sites under negative selection

$4N_eS$		6/-6	15/-15	30/-30	50/-50	100/-100
10% del	α_{sim}	0.052	0.114	0.197	0.283	0.432
	α_{EW}	0.002	0.101	0.232	0.378	0.609
33%del	α_{sim}	0.065	0.143	0.244	0.343	0.502
	α_{EW}	-0.080	0.080	0.255	0.414	0.652
66% del	α_{sim}	0.108	0.239	0.388	0.509	0.663
	α_{EW}	-0.410	-0.007	0.304	0.509	0.736

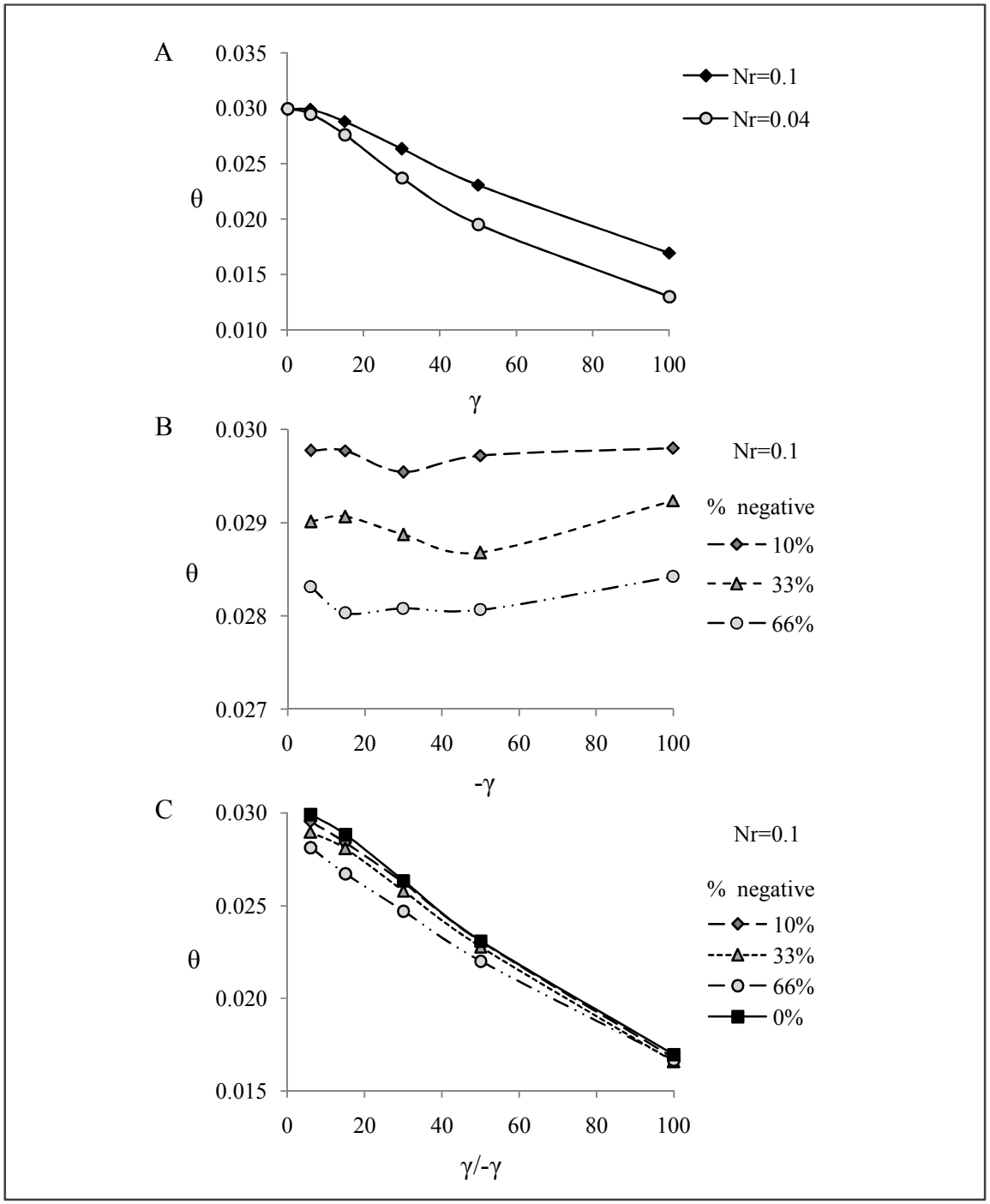

Figure 3.1. Effect of selection on linked neutral polymorphism within simulated 500 bp region. A. Effect of hitchhiking (1% of sites under positive selection). B. Effect of background selection. C. Levels of neutral polymorphism in the presence of 1% of sites under positive selection and various fractions of sites under negative selection (joint effect of hitchhiking and background selection). *X* –axis indicates the strength of selection of advantageous and/or deleterious mutations ($\gamma=4N_e s$).

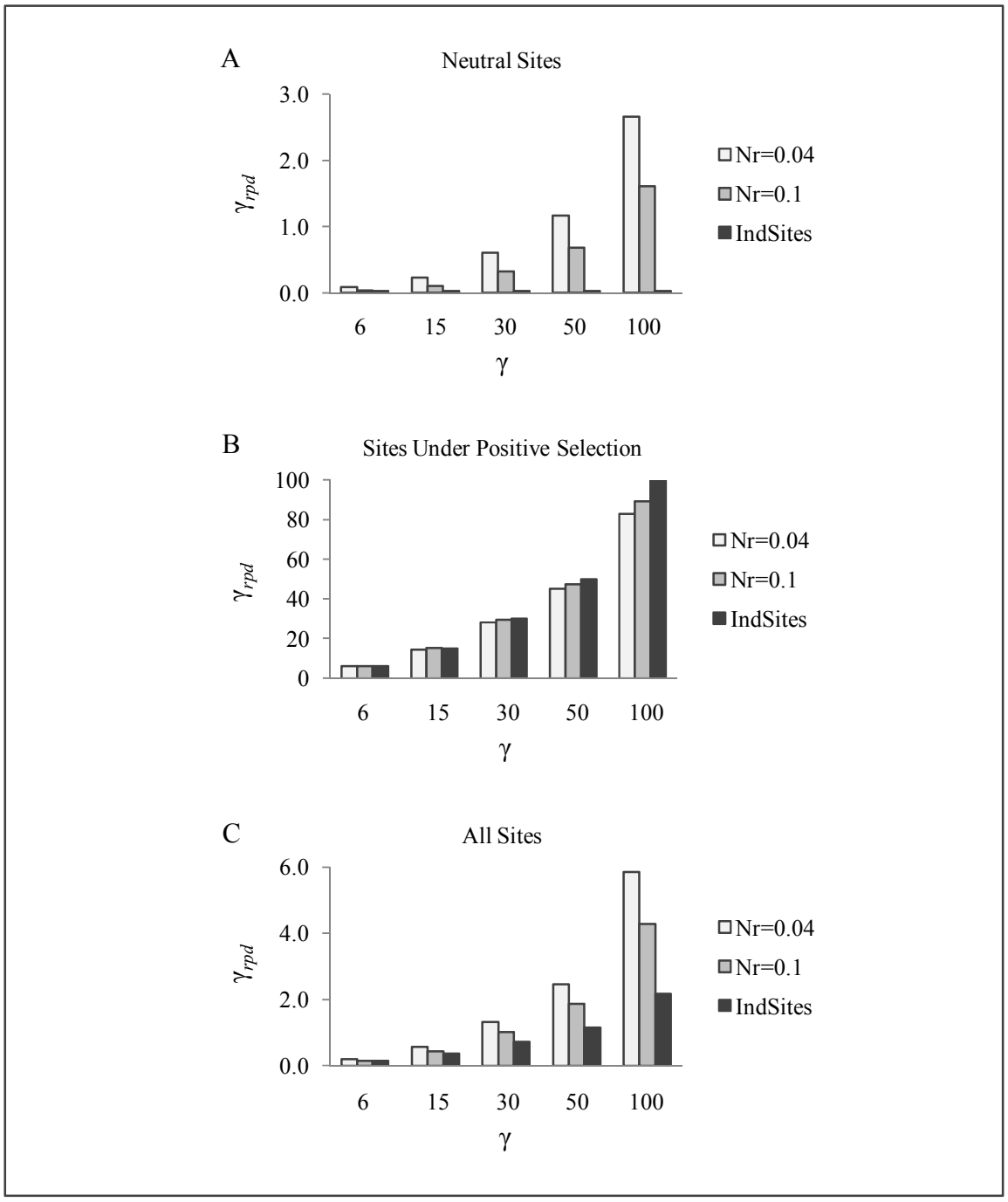

Figure 3.2. Effect of linkage on estimates of selection in the presence of advantageous mutations. A. Selection estimates for neutral sites linked to sites under positive selection. B. Selection estimates for sites under positive selection. C. Selection estimates for a complete region of interest containing 1% of sites under positive selection. *X* –axis indicates the strength of selection of advantageous mutations ($\gamma=4N_e s$).

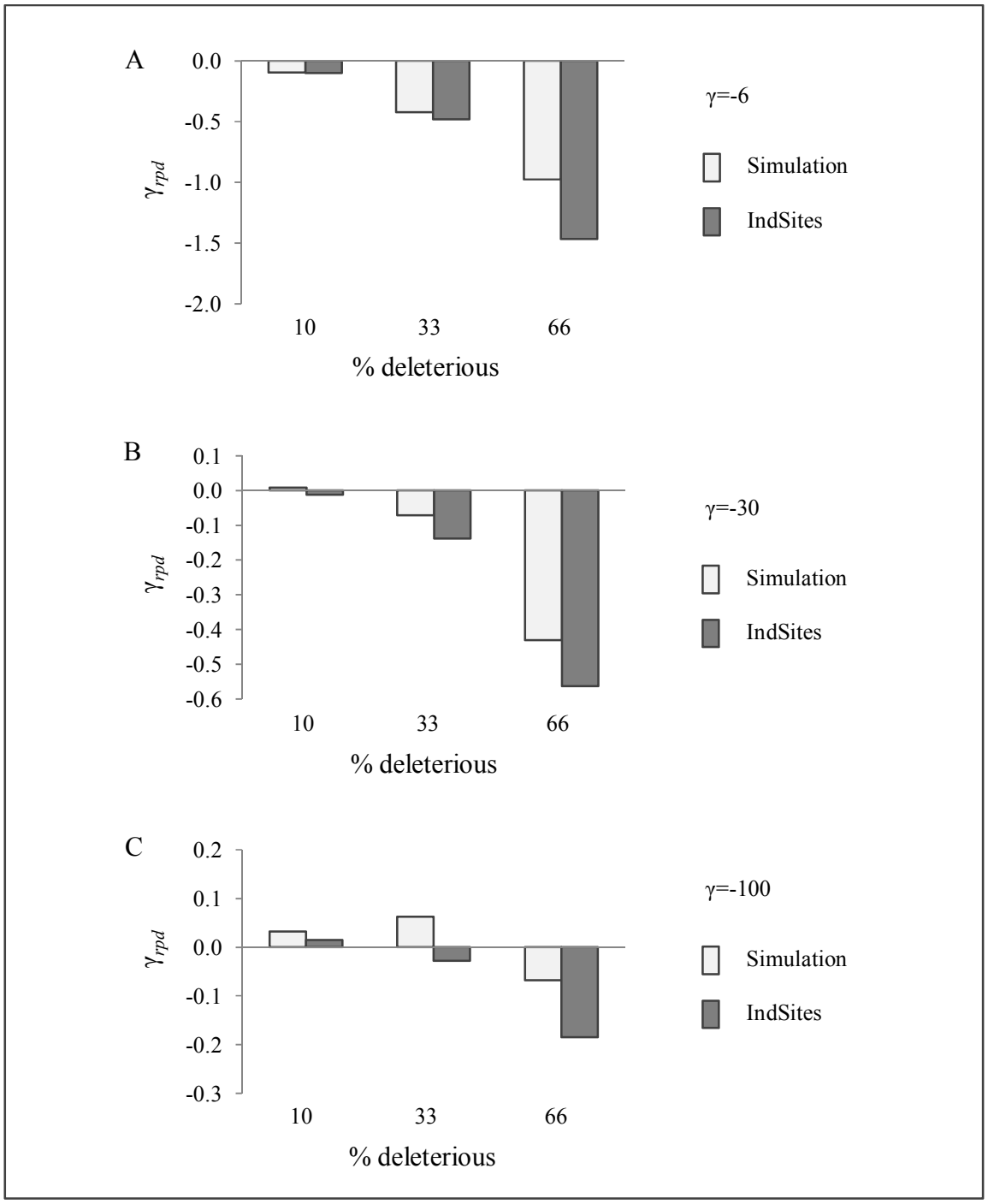

Figure 3.3. Effect of linkage on estimates of selection in the presence of different fractions of deleterious mutations. A. Deleterious mutations with $\gamma=4N_es=-6$. B. Deleterious mutations with $\gamma=4N_es=-30$. C. Deleterious mutations with $\gamma=4N_es=-100$.

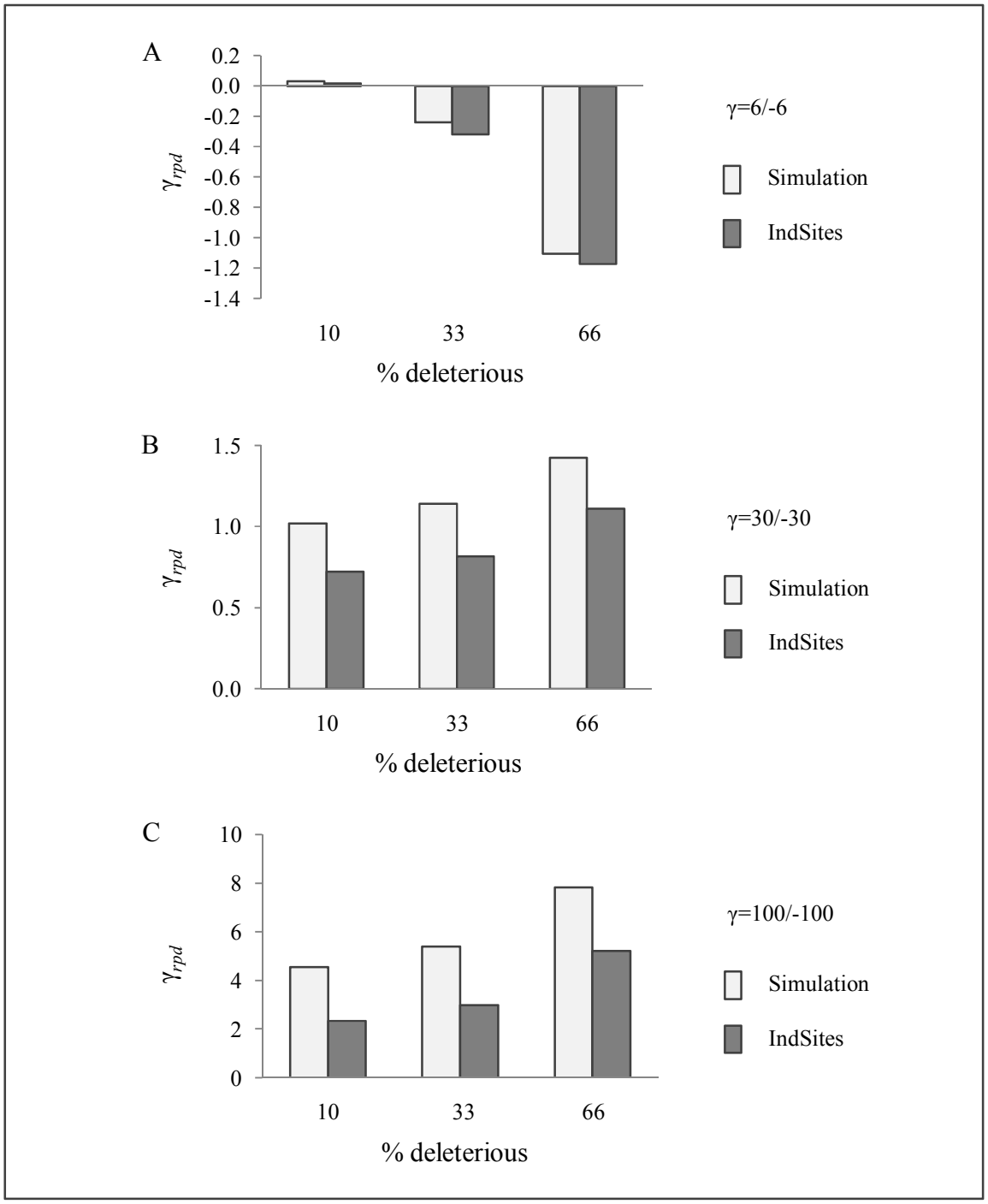

Figure 3.4. Effect of linkage on estimates of selection in the presence of 1% of sites under positive selection and different fractions of sites under negative selection. A. Advantageous/deleterious mutations with $\gamma=4N_e s=6/-6$. B. Advantageous/deleterious mutations with $\gamma=4N_e s=30/-30$. C. Advantageous/deleterious mutations with $\gamma=4N_e s=100/-100$.

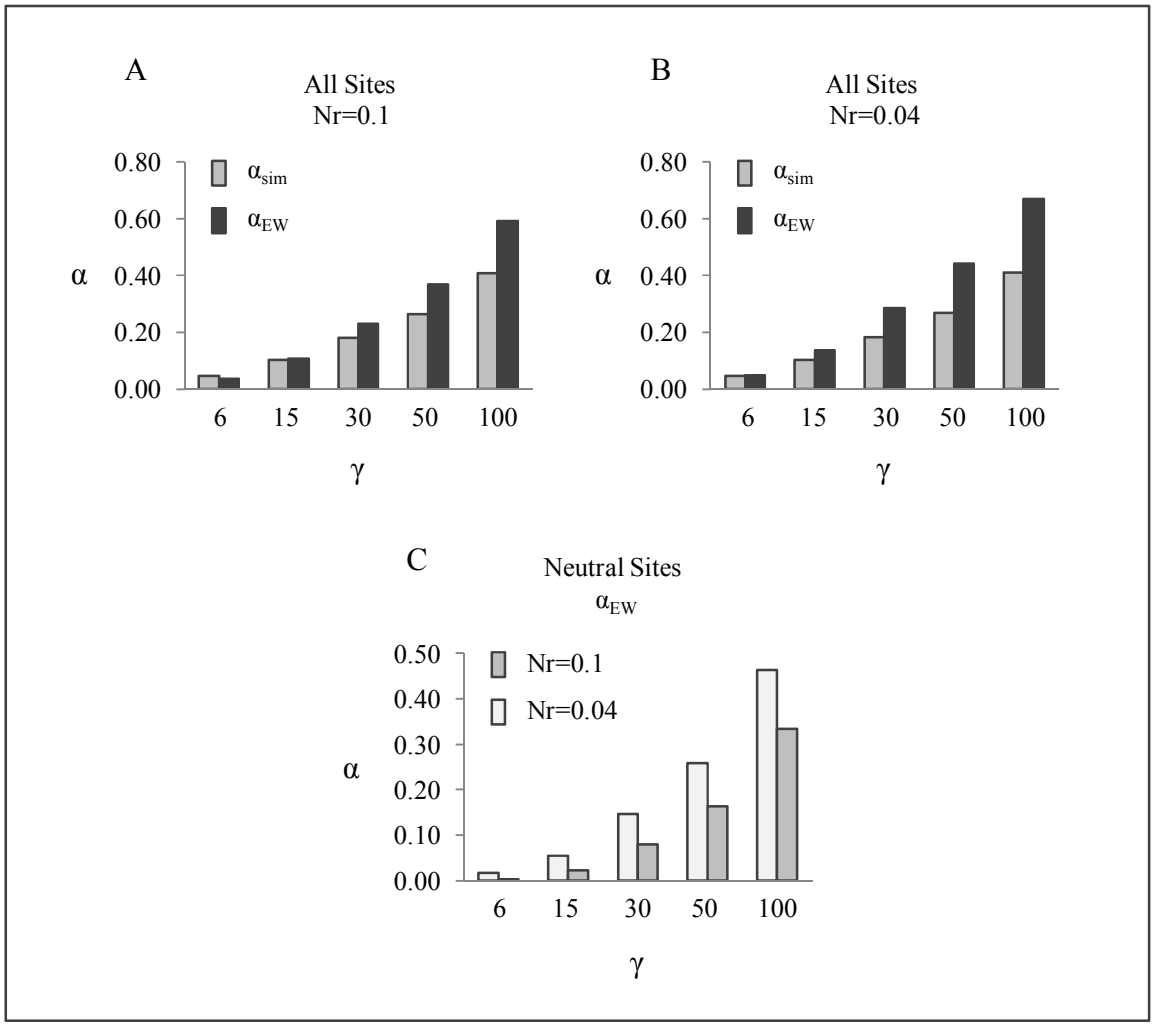

Figure 3.5. Effect of linkage on estimates of the proportion of adaptive substitutions (α) in the presence of 1% of sites under positive selection. A. Estimates of α_{sim} and α_{EW} for a complete region with Nr=0.1. B. Estimates of α_{sim} and α_{EW} for a complete region with Nr=0.04. C. Estimates of α_{EW} for neutral sites alone. *X* –axis indicates the strength of selection of advantageous mutations ($\gamma=4N_e s$).

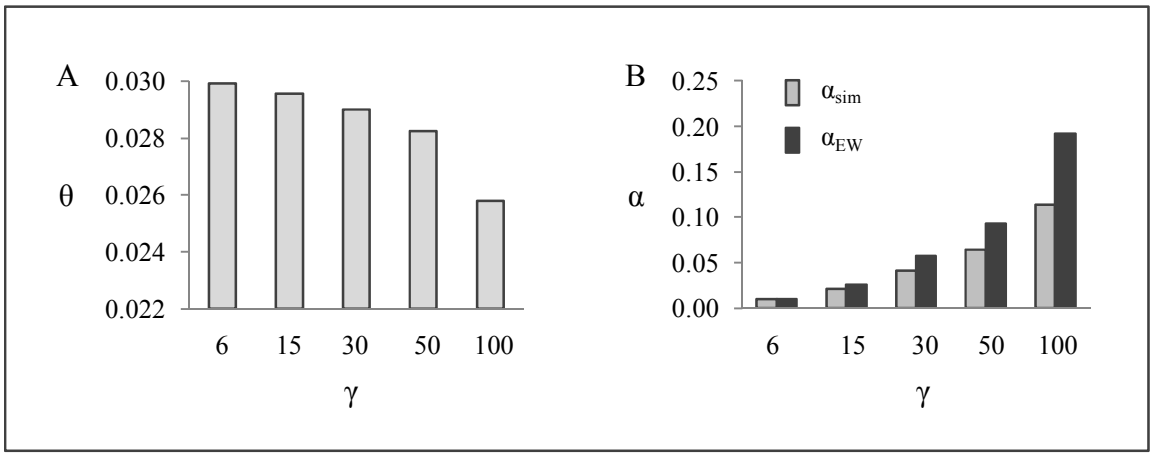

Figure 3.6. Effect of linkage on neutral polymorphism and estimates of the proportion of adaptive substitutions (α) for a complete region in the presence of a single site under positive selection within simulated 500 bp region. A. Neutral polymorphism levels with neutral expectations of θ=0.03. B. Estimates of α_{sim} and α_{EW} for a complete region. Nr=0.1. *X* –axis indicates the strength of selection of advantageous mutations ($\gamma=4N_e s$).

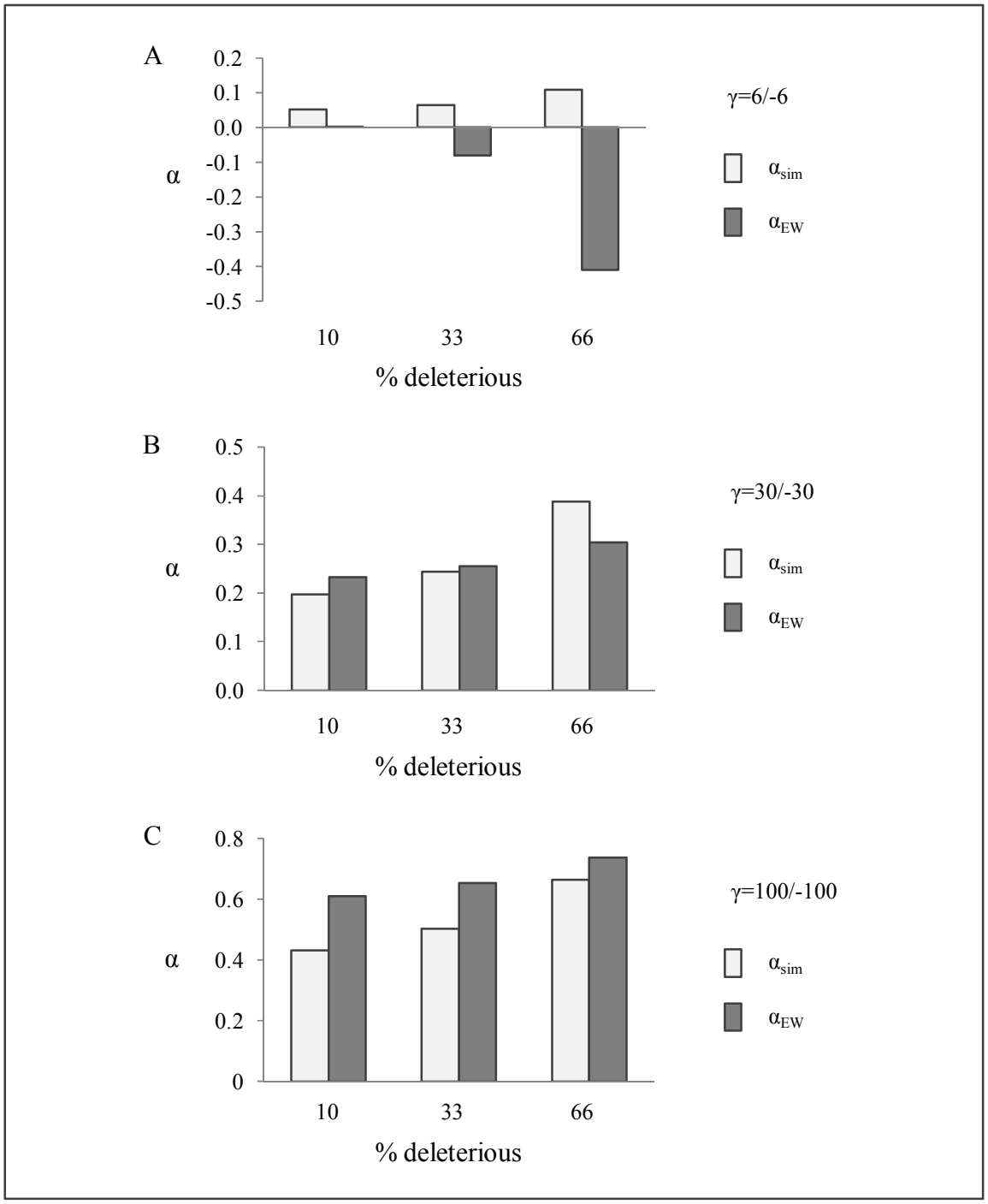

Figure 3.7. Effect of linkage on estimates of the proportion of adaptive substitutions (α) in the presence of 1% of sites under positive selection and different fractions of deleterious mutations. A. Advantageous/deleterious mutations with $\gamma=4N_e s=6/-6$. B. Advantageous/deleterious mutations with $\gamma=4N_e s=30/-30$. C. Advantageous/deleterious mutations with $\gamma=4N_e s=100/-100$.

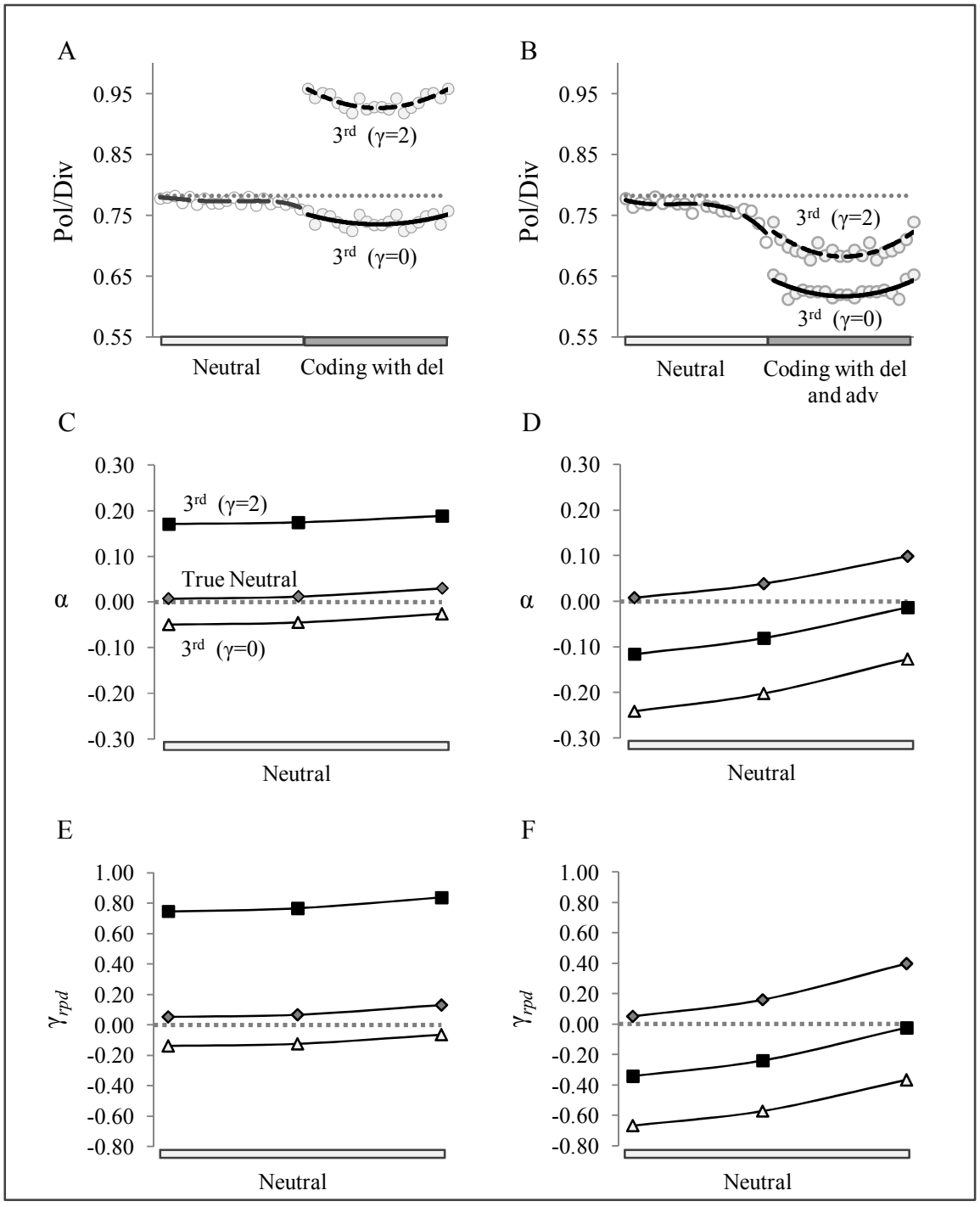

Figure 3.8. Effect of linked selection on estimates of γ and α at neutral regions adjacent to the coding sequence. Simulated region includes 1 kb of neutral sequence (γ=0) followed by 1 kb of the coding region. A,B. Trendlines of r_{pd} across simulated region for adjacent neutral and synonymous sites. C,D. Estimates of the proportion of adaptive substitutions (α_{EW}) and selection (E,F) at adjacent neutral sites using either true neutral, weakly selected or neutral 3rd positions as neutral standard.

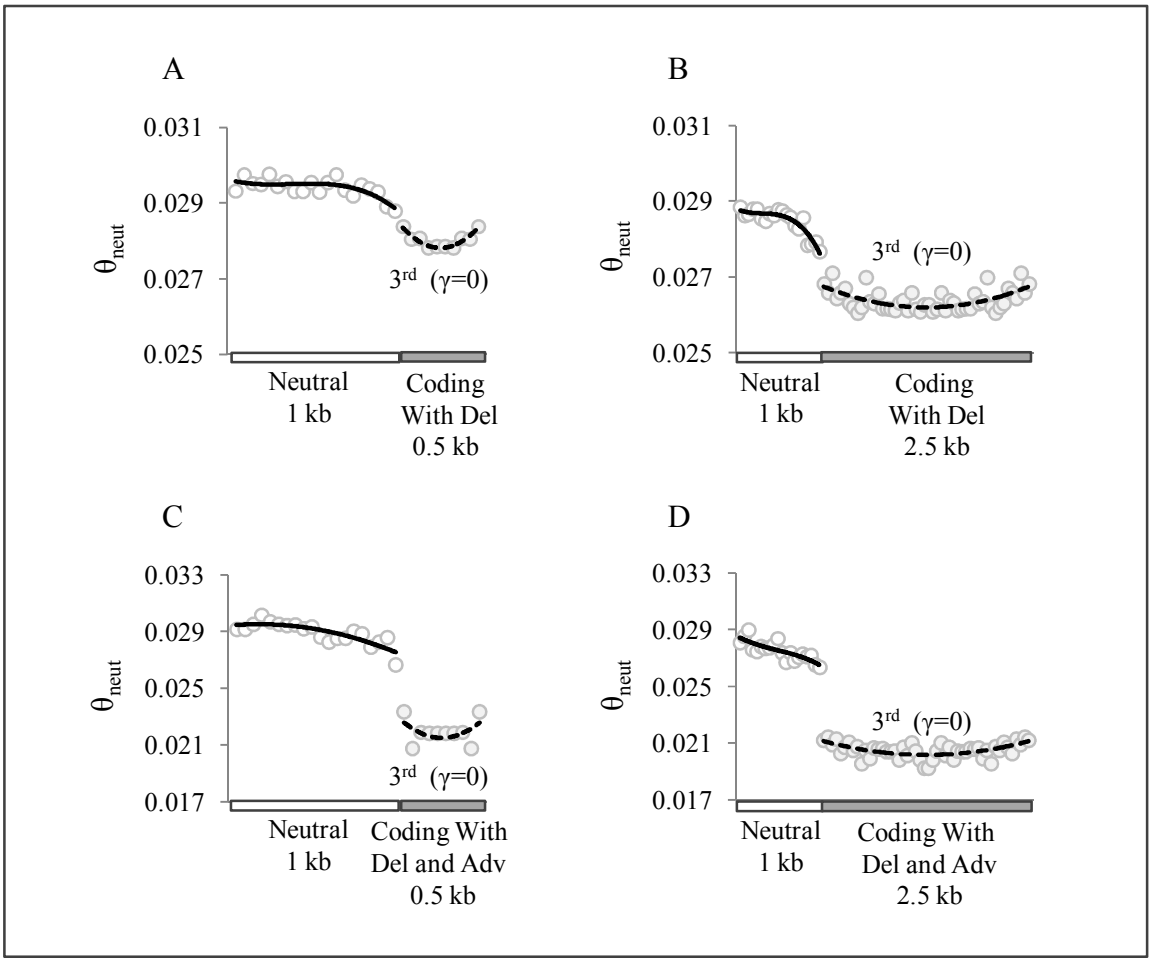

Figure 3.9. Effect of the number of sites under selection on variation in neutral polymorphism levels (N_e). Simulated region contains 1 kb of neutral sequence followed by either 0.5kb or 2.5 kb of the coding region with neutral 3rd positions. Neutral polymorphism levels for neutral sites across the simulated region when coding region contains sites under negative selection (66%, γ=-50; A, B) or sites under negative (66%, γ=-50) and positive selection (1%, γ=50; C, D).

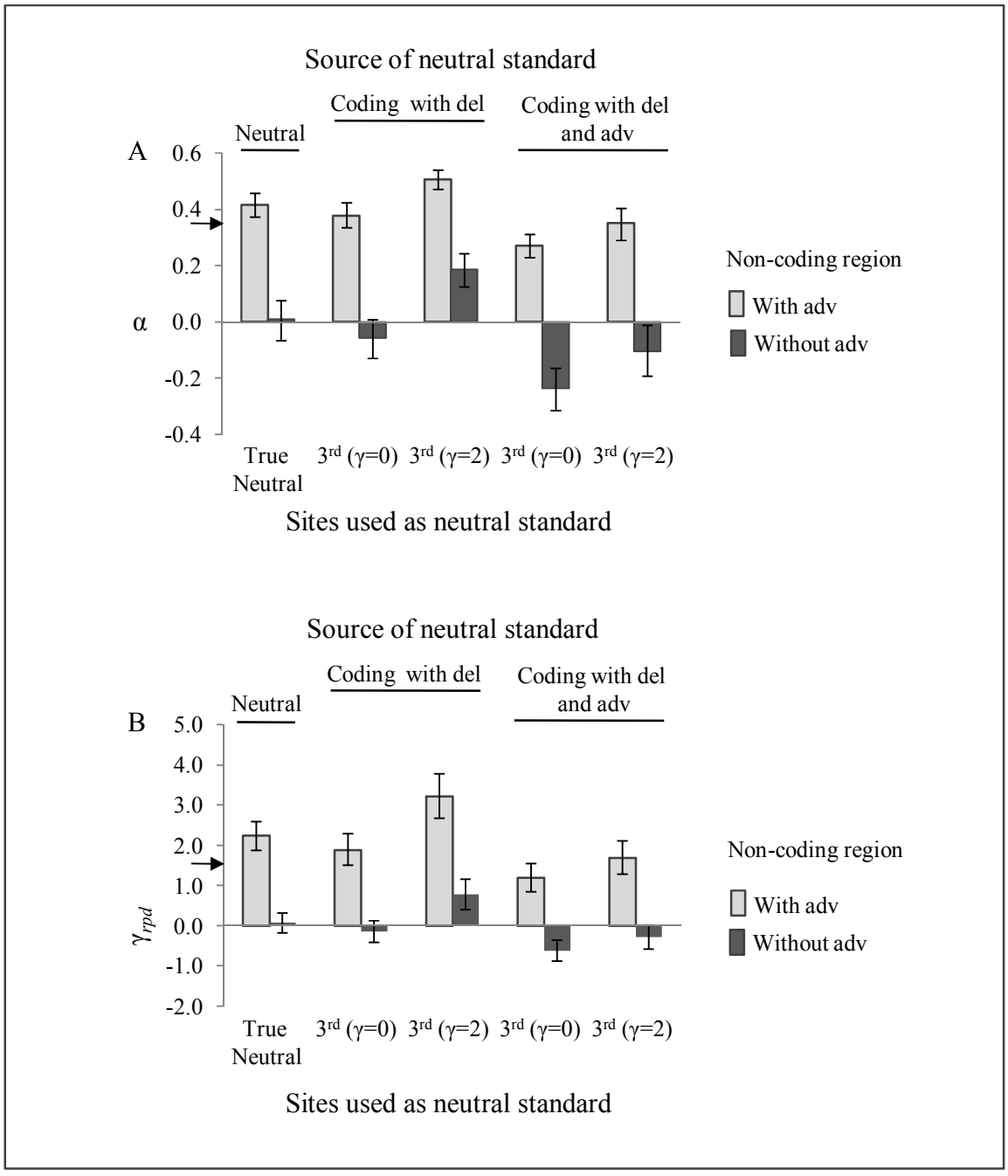

Figure 3.10. Effect of linkage on estimates of γ and α at non-coding DNA under selection. Estimates of α (A) and γ (B) at non-coding regions without any sites under positive selection (dark gray bars) and at non-coding region containing sites under positive selection (light gray bars) using different neutral standards. Neutral standard is represented by either true neutral sites or synonymous sites that are either neutral or under weak selection taken from the coding regions with or without sites under positive selection (see text for details).

CHAPTER IV

INTRON PRESENCE AND EFFECTIVENESS OF SELECTION

4.1 Introduction

Over the last 20 years there has been continuous accumulation of empirical evidence in favor of frequent occurrence of natural selection in *Drosophila* (Sella *et al.* 2009). Moreover, given the lack of free recombination in biological genomes, sites under selection are expected to alter the behavior of genetically linked neutral and selected mutations. This lack of independent behavior of mutations can be described by the Hill-Robertson effect (Hill and Robertson 1966; Felsenstein 1974).The main consequence of selection in the presence of limited recombination in finite populations is a reduction in N_e manifested in 1) a reduction in polymorphism levels and 2) a reduction in the effectiveness of selection.

In order to evaluate whether Hill-Robertson effect is relevant to the behavior of mutations in real genomes, most studies have focused on the relationship between recombination rates and either polymorphism levels or the effectiveness of selection (measured by *Ka*, *Ks* or codon bias). Assuming equal fitness effects of mutations (same selection coefficient) and no mutagenic effects of recombination, a positive correlation between recombination rate and either polymorphism or the effectiveness of selection indicates the action of the Hill-Robertson effect.

The behavior of weakly selected mutations is sensitive to changes in N_e and these mutations, represented by synonymous sites, have been used extensively in early investigations of the Hill-Robertson effect. The first empirical evidence indicating

increase in the effectiveness of selection as a result of an increase in N_e came from the analysis of codon bias in *Drosophila melanogaster* in areas of genomes with different recombination rates (Kliman and Hey 1993). But it was the discovery of the negative correlation between codon bias and the length of the coding sequence (Comeron 1997; Powell and Moriyama 1997; Comeron *et al.* 1999) that shifted the attention from changes in N_e associated with recombination rates to much more localized changes in N_e associated with the number of sites under selection. That is, the Hill-Robertson effect can have detectable local consequences, where the effectiveness of selection (measured by codon bias) is reduced more in longer CDS than in shorter CDS in areas with the same recombination rate. This observation, together with the detection of negative correlation between intron length and recombination (Carvalho and Clark 1999) stimulated the proposal of the hypothesis that introns might function as modifiers of recombination (Comeron and Kreitman 2000). That is, intron presence increases the physical distance between selected mutations thereby effectively increasing recombination rates. Thus, longer introns are beneficial in areas of lower recombination where the effects of interference are stronger. This hypothesis would gain maximal support if the two major predictions of the Hill-Robertson effect could be empirically verified: 1) central region of the coding sequence should have a reduced N_e, and 2) intron presence should remove/lessen heterogeneity in N_e across the coding region. A reduction in N_e in the center of genes with long CDS has been experimentally supported by detecting reduced levels of codon bias, reduced selection intensity on synonymous mutaitons and increased synonymous substitution rates, *Ks* (Comeron and Kreitman 2002; Comeron and Guthrie 2005; Larracuente *et al*., 2008). The second prediction has also gained support from the

analysis of codon bias in genes with and without introns showing reduced codon bias in the center of the coding sequence in genes without introns but not in genes with a centrally located intron (Comeron and Kreitman 2002). Furthermore, computer simulations studies detect differences in the effectiveness of selection with short intron sequences (<100 bp) and suggest that the required intron size to produce the equivalent increase in the effectiveness of selection is smaller in areas of non-reduced recombination (~ 0.4 cM/Mb for *Drosophila*) than in areas of lower recombination (~ 0.04 cM/Mb). Overall, these results indicate that the Hill-Robertson effect generated by weakly selected mutations may at least in part account for the distribution of introns across the genome and influence exon-intron structure of genes.

It should be noted that the above discussion is relevant to the situations when interference is generated by weakly selected mutations. A recent simulation study (Loewe and Charlesworth 2007) suggests that background selection alone can account for the heterogeneous distribution of codon bias across long CDS by reducing N_e in the center of the coding sequence. But independently of the cause of the reduction in N_e, both scenarios predict that intron presence should increase the effectiveness of selection at synonymous sites in the center of the coding region. A somewhat more complicated scenario is suggested by the presence of adaptive mutations. There have been a number of studies indicating increased levels of adaptation in areas of high recombination (Betancourt and Presgraves 2002; Presgraves 2005; Zhang and Parsch 2005; Shapiro *et al.* 2007). The changes in effectiveness of selection are detected across broad range of recombination rates (Betancourt and Presgraves 2002) suggesting that adaptive mutations may respond to small changes in N_e. The consequence of fixation of adaptive mutations

is the reduction in N_e at linked sites leading to a decrease in the effectiveness of selection at synonymous sites (Betancourt and Presgraves 2002; Andolfatto 2007). That is, codon bias is expected to be reduced in fast-evolving genes, indicating that selection cannot be simultaneously effective in fixation of beneficial mutations and maintenance of optimal codon usage. Under this scenario, intron presence may play a role in enhancing adaptive evolution rather than increasing codon bias.

To evaluate the role of introns in alleviating interference between mutations further, the previous study investigating intragenic Hill-Robertson effect (Comeron and Guthrie 2005) is expanded to include polymorphism and divergence data for genes with and without introns from three *Drosophila* species: *D. melanogaster*, *D. simulans* and *D. yakuba*. The genes chosen for analysis are grouped into four sets based on the length of the coding sequence and the presence/absence of introns: 1) short genes (~550 bp) without introns 2) short genes (~550 bp) with introns 3) long genes (>2500 bp) without introns and 4) long genes (>2500 bp) with introns. The large amount of data collected from these species should provide most of the information needed to test major predictions of the local Hill-Robertson effect and the proposed role of introns in reducing the interference between mutations from neighboring exons.

4.2 Materials and methods

4.2.1 *Drosophila* lines

DNA isolation and sequencing

A total of 71 genes were sequenced in 43 *Drosophila* lines: 11 *D.melanogaster*, 16 *Drosophila simulans* and 16 *Drosophila yakuba*. The genes are grouped into four sets based on the length of the coding sequence and the presence/absence of introns: 1) 19

genes with short CDS (~550 bp) without introns 2) 18 genes with short CDS (~550 bp) with introns 3) 21 genes with long CDS (>2500 bp) without introns and 4) 13 genes with long CDS (>2500 bp) with introns (Tables 4.1, 4.2, 4.3, 4.4). DNA was isolated from adult flies using protocols and reagents from Stratagene. For short genes, complete CDS was amplified by polymerase chain reaction (PCR). For long genes, three DNA fragments (~ 600 bp each) of the CDS were amplified: 5' region, central region and 3' region. For genes containing introns, cDNA was used to amplify central regions around the intron. The PCR products were cleaned up using Wizard MagneSil PCR System (Promega, Madison, WI.). Both strands were sequenced directly from PCR products using Applied Biosystems Big Dye Terminator chemistry on ABM prism 3100/3730 Genetic Analyzers (Applied Biosystems, Foster City, CA.). Gene definitions were obtained from the *Drosophila* genome (FlyBase, Version FB2010_02, R5.25). The conservation of gene structure in *D.simulans* and *D.yakuba* was confirmed using UCSC Genome Browser Gateway. Only genes with no evidence of multiply spliced forms that could affect the analysis were investigated.

4.2.2 Recombination rates and gene expression levels

All genes are chosen from genomic areas with recombination rate > 0.5 cM/Mb according to estimates of recombination rates obtained by Comeron (1999). Expression data (from whole adult fly) for each gene was obtained from FlyAtlas database. Whenever the sequenced gene had multiple CDSs, the expression data was obtained for the analyzed CDS only by selecting gene expression data generated with the probe that matches CDS of interest (probe location was obtained with GBrowse option in FlyBase.

4.2.3 Analysis of polymorphism, divergence and codon bias

The data set contains 11 *D.melanogaster* lines, 16 *D.simulans* lines and 16 *D.yakuba* lines. In order to eliminate the influence of polymorphic sites on estimates of divergence, ancestral sequences of *D.melanogaster*, *D.simulans* and *D.yakuba* were reconstructed using PAML 4.4 (Yang 1997; Yang 2007). The reconstructed sequences were then used to estimate the number of synonymous and non-synonymous substitutions per sites according to Comeron (1995), using publically available software K-Estimator V6.1. Polymorphism levels measured by heterozygosity (θ_w) and average pairwise diversity (θ_π), Tajima's *D* (Tajima 1989) as well as counts of the number of polymorphic and fixed sites were calculated using a program written by J.M. Comeron.

Codon usage bias was measured by the effective number of codons (ENC) using CodonW program. The ENC index measures deviation from equal usage of synonymous codons, ranging from 20 (highest codon bias - when only one codon is used for each amino acid) to 61 (lowest codon bias - when different codons are used equally).

4.3 Results and discussion

In order to evaluate the hypothesis of intron-mediated reduction of interference between selected mutations, I examined the evidence from polymorphism and divergence data combined with codon bias estimates.

Divergence and polymorphism data were obtained for four sets of genes: two sets of genes with short CDS (~ 550 bp) with and without introns, and two sets of genes with long CDS (> 2500 bp) with and without introns (Tables 4.1- 4.4). Overall, 37 genes with short CDS and 34 genes with long CDS were analyzed. The analysis of genes with long

CDS was restricted to three regions: 3' region, central region and 5' region, about 550 bp each. For these long genes the intragenic polymorphism and divergence patterns were also examined. A total of ~ 82 kb of CDS was sequenced in each of 43 *Drosophila* lines: 11 *D.melanogaster*, 16 *D. simulans* and 16 *D.yakuba* lines. The aligned data set included ~ 20 kb for genes with short CDS and ~62 kb for genes with long CDS.

Analyzed genes are located in genomic regions with non-reduced recombination rates (> 0.5 cM/Mb) and levels of recombination do not differ significantly between sets of genes investigated (Mann-Whitney U test: $Z = 1.79$, $p>0.05$ for genes with short CDS with and without introns; $Z=-1.24$, $p>0.05$ for genes with long CDS with and without introns ; $Z=-1.48$, $P>0.05$ for genes with short and long CDS).

Expression levels of genes with short CDS containing introns do not differ significantly from the expression levels of genes with short CDS without introns (Mann-Whitney U test, $Z=1.66$, $p>0.05$). Similarly, there is no significant difference between expression levels of genes with long CDS with and without introns (Mann-Whitney U test, $Z=0.69$, $p>0.05$). However, expression levels of genes with short CDS are significantly higher than those of genes with long CDS (Mann-Whitney U test, $Z=2.3$, $p<0.05$). Given a positive correlation between codon bias and gene expression levels (Shields *et al.* 1988), comparison of genes with short and long CDS might not be informative about the strength of the Hill-Robertson effect. For this reason, I primarily focus on the comparison of genes with and without introns within sets of genes with short and long CDS.

4.3.1 Comparison of genes with short CDS with and without introns

One of the major factors that determine the strength of the Hill-Robertson effect is the number of sites under selection, with more sites generating stronger effect. Simulation results suggest that interference is expected to occur across short coding sequences under some conditions (Figure 3.9, Chapter III), but the magnitude of such effects is very small (~ 2%). According to the working hypothesis, if there are in fact detectable intragenic effects within genes with short CDS, intron presence is expected to generate patterns distinct from those observed in genes without introns. However, results presented below indicate that there is no difference between short genes with and without introns in any of the investigated aspects of the data. The estimates of synonymous (*Ks*) and non-synonymous substitutions (*Ka*) per site for each gene are given in Tables 4.5, 4.6. Comparison of synonymous and non-synonymous divergence between genes with short CDS indicates no significant difference between genes with and without introns (Table 4.7). Likewise, the analysis of codon bias as measured by the effective number of codons, or ENC (Table 4.8 gives ENC values for each gene) shows no significant difference between genes with and without introns for any of the species compared (Table 4.9). These results suggest that intron presence has no detectable effect on the evolution of short genes. The results of polymorphism analysis are consistent with the results obtained from the divergence data. No significant difference in polymorphism levels are observed between genes with and without introns for either synonymous or non-synonymous sites (Table 4.10). All together, the set of analyzed genes with and without introns offers no

evidence that would suggest that introns have a detectable effect on the effectiveness of selection when the coding sequence is short (<600 bp).

4.3.2 Comparison of genes with long CDS with and without introns

Examination of patterns of polymorphism and divergence in genes with long CDS is expected to be more informative about the possible role of introns in alleviating interference. Long genes contain a large number of sites under selection and multiple evidence suggest that the length of the coding sequence examined in this data set (L>2500 bp) should be sufficient to detect interference between mutations. In addition to simulations studies (Comeron *et al.* 1999; Comeron and Kreitman 2002; Loewe and Charlesworth 2007) there are evidence from empirical studies (Comeron and Kreitman 2002; Qin *et al.* 2004; Comeron and Guthrie 2005; Larracuente *et al*. 2008) documenting heterogeneous distribution of divergence and codon bias across coding sequences, with elevated *Ks* and reduced codon bias in the central region of CDS. If such heterogeneity is the result of interference, intron presence (if relevant) should lead to a reduction in *Ks* and an increase in codon bias in the center of the CDS by increasing the effectiveness of selection as a result of an increase in local N_e.

The estimates of *Ks* and *Ka* for each gene across long CDS with and without introns are given in Table 4.11 and Table 4.12. The summary of the data is presented in Table 4.13. To provide an overall picture of the divergence patterns, Figure 4.1 shows the results based on the whole phylogeny. The first thing to notice is that genes without introns behave differently than genes with introns in two ways: 1) the distribution of *Ks*

values across CDS differs between genes with and without introns, and 2) genes with introns appear to have elevated levels of nonsynonymous divergence.

The analysis of the whole phylogeny indicates that in the set of genes without introns, the central region has a significantly higher *Ks* compared to lateral regions: *Ks* (center)=0.290 and *Ks* (lateral)=0.247 ($p < 0.05$, Wilcoxon Signed-Rank test). This significant increase in Ks values is observed for both, *mel-yak* and *sim-yak* comparison (Table 4.13). There is no significant difference in *Ks* values between lateral regions. In the set of genes with introns, no such heterogeneity in *Ks* values is detected across the coding sequence for either each species pair or for the whole phylogeny (Table 4.13, Figure 4.1). With respect to non-synonymous evolution, *Ka* values do not change significantly across the coding sequence in genes with and without introns (Table 4.13, Figure 4.1). Similar divergence patterns have been previously reported for genes with long CDS without introns based largely on the same data set (Comeron and Guthrie 2005). An increase in *Ks* in the center of the coding region in genes without introns suggests the action of the Hill-Robertson effect: decrease in N_e in the central region leads to a reduction in the effectiveness of selection at synonymous sites. Intron presence appears to remove local changes in N_e since no differences in *Ks* are observed across coding sequences of genes with introns.

However, the comparison of *Ks* values between genes with and without introns indicates that there is no difference in synonymous divergence in the central region (P=0.834, Mann-Whitney U test) and no overall difference in synonymous divergence when all regions are combined (P=0.063, Mann-Whitney U test), suggesting similar levels of the effectiveness of selection at synonymous sites in genes with and without

introns (Table 4.14). Evaluation of codon bias across and between genes with and without introns corroborates the results obtained from divergence data. Overall, codon bias patterns as measured by ENC (see Tables 4.15 and 4.16 for ENC values per gene) mirror those of synonymous divergence. There is a significant reduction in codon bias (increase in ENC) in the center of the coding region in genes without introns, but not in genes with introns (Table 4.17, Figure 4.2). This result is consistent with local reduction in N_e as predicted by the Hill-Robertson effect. However, just as seen with the divergence data, codon bias does not differ significantly between the central regions of long genes with and without introns (Table 4.18). This suggests that intron presence does not increase the effectiveness of selection at synonymous sites, at least in the set of investigated genes.

The lack of overall difference in the effectiveness of selection at synonymous sites between genes with and without introns does not exclude the possibility that increase in *Ks* and a reduction in codon bias observed in central regions of long genes without introns is the result of interference. In this case, the crucial evidence should come from polymorphism data since the Hill-Robertson effect predicts a reduction in N_e and consequently, a reduction in polymorphism in the center of the coding sequence. Alternatively, the observed patterns of synonymous divergence and codon bias can be explained by the relaxation of selection at synonymous sites of the central codons. If this is the case, polymorphism levels are expected to be higher in the central region than in the lateral regions of the coding sequence. Analysis of synonymous polymorphism across coding regions indicates no significant difference between central and lateral regions for each species separately (Table 4.19). However, when values are combined from all

species, polymorphism levels are significantly higher in the center of the coding sequence of genes without introns (Wilcoxon Signed- Rank test , Z=2.27, P = 0.0232), but not in the center of the CDS of genes with introns (Z=1.63, P=0.1031). In addition, if selection is relaxed in the center of the coding regions, the ratio of polymorphism to divergence should be reduced relative to that observed in the lateral regions. However, the ratios of polymorphism to divergence do not show heterogeneous distribution across the coding sequence and there is no significant difference between the three regions (Table 4.20). Moreover, if there is a relaxation of selection on internal codons, similar patterns of codon bias and *Ks* should be seen in genes with introns, but this is not the case. It is also possible that mutation rate is increased in the center of the long genes leading to elevated rates of synonymous polymorphism and divergence: in this case there should also be an increase in *Ka* in the center of the long genes without introns, but this is not observed. Thus, although divergence data indicates presence of interference, the result from the analysis of polymorphism data does not, but offers no alternative explanation either. Moreover, it is difficult to reconcile homogeneous selection intensity at synonymous sites (indicated by similar ratios of synonymous polymorphism to divergence) with heterogeneous codon bias across the coding sequence. Analysis of a larger data set is required to clarify the relationship between Pol/Div and codon bias in long genes without introns. In genes with introns, the ratio of polymorphism to divergence also does not vary significantly across the coding sequence (Table 4.20) and there is no difference between Pol/Div ratios between long genes with and without introns (Mann-Whitney U test, Z=0.03, P= 0.9761), consistent with no differences in *Ks* and codon bias between the two sets of genes. Overall, analysis of synonymous sites

provides conflicting evidence with regard to intragenic Hill-Robertson effect; divergence data argues in favor of interference while polymorphism data argues against it. However, every line of evidence suggests that intron presence does not mediate an increase in the effectiveness of selection at synonymous sites.

Although long genes with and without introns evolve similarly with respect to synonymous sites, patterns of nonsynonymous evolution differ between the two sets of genes. Genes with introns have significantly higher *Ka* values than genes without introns in every species pair comparison and in the analysis of the whole phylogeny (Table 4.14, Figure 4.1). The *Ka/Ks* ratio is also significantly elevated in genes with introns compared to that of genes without introns (Table 4.21). Elevated *Ka* levels signal either increase in fixation of non-synonymous slightly deleterious mutations or increase in fixation of non-synonymous beneficial mutations. The first interpretation implies that intron presence reduces the effectiveness of selection and in this case *Ks* in genes with introns should be higher than *Ks* in genes without introns, but this is not observed. If elevated *Ka* reflects increase in fixation of beneficial mutations, intron presence must then reduce interference between beneficial and deleterious mutations (or between beneficial mutations) increasing the effectiveness of selection on non-synonymous advantageous mutations. If this is the case, fixation of beneficial mutations would result in a reduction of N_e at linked sites leading to an increase in fixation of slightly deleterious mutations and a reduction in polymorphism. That is, the lack of significant difference in *Ks*, codon bias, and synonymous polymorphism between long genes with and without introns (Table 4.14, Table 4.18, Table 4.19) could be accounted for by more frequent fixation of beneficial mutations in genes with introns.

To understand the cause of the difference in *Ka/Ks* between genes with and without introns, I estimated the proportion of adaptive substitution (α) in genes with and without introns (McDonald and Kreitman 1991; Smith and Eyre-Walker 2002). If elevated *Ka* in genes with introns is the result of fixation of slightly deleterious mutations, there should be an excess of non-synonymous polymorphism and the estimates of α should be smaller than that for genes with long CDS without introns. To minimize the effects of selection at synonymous sites on estimates of α, synonymous codons were separated into preferred and unpreferred: mutations from unpreferred to unpreferred codon are assumed to be more neutral than those from preferred to unpreferred or from unpreferred to preferred and were used as a neutral standard. The proportion of adaptive substitutions was estimated for each gene separately and for combined data sets. The results suggest that the proportion of adaptive substitutions is higher in genes with introns than without introns (Table 4.22). The values for each gene based on the total phylogeny are shown in Table 4.23. Two genes out of 21 genes (~10%) without introns and 4 genes out of 12 (~ 30%) show a significant departure from neutral expectations. After sequential Bonferroni correction for multiple tests, one gene remains significant in the set of genes without introns and 3 in the set of genes with introns. Detection of higher proportion of adaptive substitutions in genes with introns is consistent with the intron-mediated increase in the effectiveness of selection on beneficial mutations.

If intron presence effectively increases recombination rate, comparison of genes with and without introns is analogous to the comparison of genes in lower and higher areas of recombination. A number of empirical studies that focused on such comparison

do report higher levels of adaptation in regions of higher recombination (Betancourt and Presgraves 2002; Presgraves 2005; Zhang and Parsch 2005; Shapiro *et al.* 2007). Whenever polymorphism data is available, the general approach in these studies is to demonstrate positive correlation between silent variation and recombination rates. Such correlation indicates increase in N_e with increase in recombination rates. If higher levels of adaptation are detected in areas of high recombination, it is attributed to an increase in the effectiveness of selection associated with higher N_e. In the data set of long genes examined there is no significant correlation between neutral variation (as measured by polymorphism to divergence ratio at unpreferred to unpreferred sites) and recombination rate, indicating no changes in N_e associated with changes in recombination rate. That is, higher levels of adaptation detected in long genes with introns cannot be attributed to an increase in N_e associated with recombination rates as such, but can be attributed to an increase in N_e generated by intron presence as a result of an increase in the *effective* rate of recombination. Although an attractive hypothesis, the limited analysis of the data set does not allow to draw definite conclusions with regard to intron-mediated adaptation hypothesis. I also want to emphasize that the data set investigated here contains genes located on X chromosome. Moreover, 2 out of the four genes with introns and 1 out of 2 genes without introns that individually show departure from neutral expectation are located on X chromosome. Given that X-linked genes might experience different selection pressures (Baines *et al.* 2008), higher levels of adaptation in those genes might have nothing to do with intron presence. Larger data set with separate analysis of autosomal and X-linked gene is needed to evaluate this hypothesis properly.

4.3.3 Comparison between genes with short and long CDS

Genes with short CDS have higher expression levels than genes with long CDS. This means that the comparison between the two sets of genes cannot be used to assess the role of CDS length on the strength of the Hill-Robertson effect. Given the strong correlation between gene expression and codon bias (Shields *et al.* 1988), it is expected that genes with short CDS will have higher ENC values (lower codon bias) than genes with long CDS in the absence of interference. Genes with long CDS in the analyzed data set do show significantly lower codon bias (Mann-Whitney U test, $Z= 5.77$, $p<0.0001$) than genes with short CDS as well as elevated *Ks* values (Mann-Whitney U test, $Z=2.16$, $p< 0.05$), as expected. However, neither levels of synonymous polymorphism (Mann-Whitney U test, $Z= 1.86$, $p=0.0629$) nor polymorphism to divergence ratios (Mann-Whitney U test, $Z=0.03$, $p=0.9761$) differ significantly between short and long genes. This lack of correspondence between codon bias and pol/div ratios is also observed in long genes without introns (see above). Although genes have been chosen from non-reduced areas of recombination it is possible that genes within each set show considerable variation in polymorphism levels making it difficult to detect expected differences in polymorphism to divergence ratios between short and long genes and across long genes without introns. With respect to levels of adaptation, in contrast to genes with long CDS, no departure from neutral expectation is detected for short genes using combined data set (Table 4.22), possibly due to higher levels of expression (Larracuente *et al.* 2008).

4.4 Summary

A number of simulation and empirical studies indicates the relevance of the Hill-Robertson effect to protein evolution, both at synonymous and non-synonymous sites. These studies have focused on evaluating changes in the effectiveness of selection associated with changes in N_e generated by differences in recombination rates. Relatively few studies have attempted to evaluate the extent to which Hill-Robertson effect might influence intragenic patterns of evolution at synonymous and non-synonymous sites.

Here I examined polymorphism and divergence data in combination with codon bias estimates in sets of genes with short and long CDS with and without introns in attempt to answer two questions. First, is there evidence in favor of intragenic Hill-Robertson effect? Second, does intron presence mediate the increases in effectiveness of selection?

Analysis of genes with long CDS without introns provides conflicting evidence with respect to the presence of intragenic Hill-Robertson effect. Increased *Ks* and reduced codon bias in the center of the coding sequence argues in its favor. However, lack of reduction in polymorphism in the center of the coding region argues against intragenic Hill-Robertson effect. Moreover, lack of differences in polymorphism to divergence ratios across the coding sequence indicates no changes in effectiveness of selection between the three regions – a result that is not compatible with the observed differences in codon bias across the coding sequence. Analysis of a larger data set is required to obtain conclusive results.

The effect of intron presence was evaluated by the comparison of genes with long CDS with and without introns. Although intron presence removes heterogeneity in *Ks*

and codon bias across the coding region, there is no difference in overall effectiveness of selection at synonymous sites between genes with and without introns. This is an unexpected result since if intron presence had no effect on protein evolution, we would see similar patterns of *Ks* and codon bias in genes with and without introns. The analysis of non-synonymous sites indicates that genes with introns show higher levels of adaptation. This finding can explain unexpected lack of differences in *Ks*, codon bias and levels of synonymous polymorphism between genes with and without introns. Thus, it appears that intron presence might increase the effectiveness of selection on adaptive mutations. This possibility warrants further investigation: information obtained from the data set investigated here is not sufficient for a conclusive evaluation of intron-mediated adaptation hypothesis.

Table 4.1. List of genes with long CDS containing introns

Gene Name	Locus	Genomic Position	CDS Length (bp)	Intron Number	Intron Size (bp)	Total Intron Size (bp)	Intron Density**
big bang*	FBgn0087007	3L:70E1-70D7	7911/4651	4	984/90418/68/75	984	0.21
APC-like	FBgn0015589	3R:98E5-98E6	7251	2	213/60	273	0.04
CG7065	FBgn0030091	X:8C13-8C14	3693	1	1060	1060	0.29
tamas	FBgn0004406	2L:34D1-34D1	3435	2	55/59	114	0.03
RpI135	FBgn0003278	2L:21C2-21C2	3387	2	62/59	121	0.04
Caf1-180	FBgn0030054	X:7F8-7F8	3549	2	63/62	125	0.04
CG15311	FBgn0030182	X:9B2-9B3	3483	2	1400/1583	2983	0.86
Usp7	FBgn0030366	X:11A1-11A2	3387	2	114/56	170	0.05
CG8915	FBgn0030833	X:15F4-15F4	2928	2	102/61	163	0.06
CG15431	FBgn0031602	2L:24F1-24F1	3078	2	69/75	144	0.05
Bre1	FBgn0086694	3L:64E8-64E8	3132	2	114/52	166	0.05
CG11070	FBgn0028467	2L:26F5-26F6	2979	2	60/66	126	0.04
CG6040	FBgn0038679	3R:91F2-91F2	5352	5	1832/99/67/57/61	2116	0.40

* First two exons (L=4651 bp) separated by intron (L=984 bp) were analyzed

** Intron density =Total Intron Length/CDS Length

Table 4.2. List of genes with short CDS containing introns

Gene Name	Locus	Genomic Position	CDS length (bp)	Intron Number	Intron Size (bp)	Total Intron Size (bp)	Intron Density**
Arf72A	FBgn0000115	3L:72C1-72C1	540	2	97/268	365	0.68
CG15098	FBgn0034398	2R:55F6-55F6	558	3	79/192/180	451	0.81
san	FBgn0024188	2R:47F5-47F5	552	2	126/232	358	0.65
CG13086	FBgn0032770	2L:37D2-37D2	546	1	406	406	0.74
corn*	FBgn0259173	3L:65E5-65E6	3321/394	6	66/62/58/54/455/53	53	0.13
CG31800	FBgn0051800	2L:37B10-37B10	546	3	57/54/58	169	0.31
CG10795	FBgn0034626	2R:E1-57E1	534	1	65	65	0.12
Scp2	FBgn0020907	3R:89B20-89D4	552	3	1616/278/1528	3422	6.20
CG16817	FBgn0037728	3R:85E3-85E3	552	3	475/71/66	612	1.11
CG2789	FBgn0031263	2L:21E2-21E2	555	1	59	59	0.11
CG30185	FBgn0050185	2R:59E3-59E3	537	2	68/71	79	0.15
CG16799	FBgn0034538	2R:57A9-57A9	537	3	1066/52/134	1252	2.33
Cpr62Bc	FBgn0035281	3L:62B7-62B7	540	2	97/573	670	1.24
Chd64	FBgn0035499	3L:64A6-64A7	564	2	6509/162	6671	11.83
CG12918	FBgn0033477	2R:46D9-46D9	567	2	100/194	294	0.52
Cpr49Ah	FBgn0033731	2R:49A3-49A3	570	1	219	219	0.38
CG2113	FBgn0035384	3L:63A2-63A2	549	1	58	58	0.11
RpL11	FBgn0013325	2R:56D7-56D7	552	3	128/60/214	402	0.73

* Last two exons of corn-PC (L=394 bp) separated by intron (L=53 bp) were analyzed.

** Intron density =Total Intron Length/CDS Length.

Table 4.3. List of genes with long CDS without introns

Gene Name	Locus	Genomic Position	CDS Length (bp)
Wnt5	FBgn0010194	X:17B6-17C1	3012
lin	FBgn0002552	2R:44F6_44F7	2574
CG10321	FBgn0034643	2R:57F5-57F5	2505
CG14411	FBgn0030582	X:12F5-12F5	2529
Atg9	FBgn0034110	2R:53B1-53B2	2535
Brd8	FBgn0039654	3R:99A1-99A1	2616
kek1*	FBgn0015399	2L:34A1-34A1	2433/2367
Snoo*	FBgn0085450	2L:28D3-28D3	3669/2670
iHog	FBgn0031872	2L:27C6-27C6	2658
kek2	FBgn0015400	2L:32F4-32F4	2682
CG13350	FBgn0033890	2R:50C23-50C23	2685
gprs*	FBgn0024232	2R:53C13-53C14	3906/2898
CG17075	FBgn0031239	2L:21B6-21B6	2904
CG5669	FBgn0039169	3R:95F2-95F3	2904
Ranbp21	FBgn0031051	X:18D12-18D13	3723
CG18265	FBgn0036725	3L:74C2-74C2	3966
Tollo	FBgn0029114	3L:71C1-71C1	4038
18 wheeler	FBgn0004364	2R:56F8-56F8	4155
mus101	FBgn0002878	X:12B4-12B4	4275
Toll-7	FBgn0034476	2R:56E4-56E4	4338
Atg2	FBgn0044452	3L:62F3-62F4	5718

* These genes contain introns. The first number indicates total CDS length and the second number indicates the length of the analyzed long exon.

Table 4.4. List of genes with short CDS without introns

Gene Name	Locus	Genomic Position	CDS Length (bp)
CG10839	FBgn0028858	2L:35E1-35E1	546
PGRP-SC1a	FBgn0043576	2R:44E2-44E2	555
CG7197	FBgn0035866	3L:66C5-66C5	537
Roughened	FBgn0004636	3L:62B7-62B7	552
Ubc-E2H	FBgn0029996	X:7D6-7D6	549
NC2β	FBgn0028926	2L:35B8-35B8	549
PGRP-SC2	FBgn0043575	2R:44E2-44E2	552
E(spl)	FBgn0000591	3R:96F10-96F10	537
CG15696	FBgn0038833	3R:93A2-93A2	537
CG13308	FBgn0035932	3L:66D15-66D15	546
CG4764	FBgn0031310	2L:21F1-21F1	546
mRpL12	FBgn0011787	3L:66E5-66E5	546
CG34388*	FBgn0085417	3R:88C1-88C1	1656/729
CG17304	FBgn0038267	3R:88D5-88D5	531
Bro	FBgn0013755	3L:62A9-62A9	639
Vhl	FBgn0041174	2R:47E5-47E6	534
ATPsyn-d	FBgn0016120	3R:91F1-91F1	534
CG34366*	FBgn0085395	2L:30B1-30B1	2823/632
HLHm7	FBgn0002633	3R:96F10-96F10	558

* These genes contain introns. The first number indicates total CDS length and the second number indicates the length of the analyzed short exon.

Table 4.5. Synonymous (*Ks*) and non-synonymous (*Ka*) substitutions per site in genes with short CDS without introns

Short No Introns	*mel-sim*		*mel-yak*		*sim-yak*	
Gene Name	*Ks*	*Ka*	*Ks*	*Ka*	*Ks*	*Ka*
CG10839	0.106	0.008	0.222	0.022	0.236	0.021
PGRP-SC1a	0.040	0.003	0.060	0.003	0.058	0.000
CG7197	0.033	0.000	0.164	0.000	0.175	0.000
Roughened	0.044	0.000	0.083	0.000	0.080	0.003
Ubc-E2H	0.107	0.000	0.119	0.000	0.049	0.000
NC2β	0.073	0.000	0.226	0.000	0.170	0.000
PGRP-SC2	0.116	0.009	0.285	0.003	0.234	0.006
E(spl)	0.035	0.009	0.079	0.012	0.055	0.009
CG15696	0.074	0.009	0.243	0.017	0.227	0.014
CG13308	0.116	0.018	0.351	0.069	0.308	0.073
CG4764	0.068	0.000	0.135	0.000	0.104	0.000
mRpL12	0.064	0.000	0.192	0.011	0.186	0.011
CG34388	0.050	0.004	0.050	0.027	0.081	0.031
CG17304	0.065	0.016	0.156	0.025	0.130	0.026
Bro	0.065	0.002	0.338	0.008	0.292	0.006
Vhl	0.080	0.006	0.228	0.009	0.259	0.003
ATPsyn-d	0.079	0.005	0.197	0.005	0.129	0.010
CG34366	0.126	0.008	0.190	0.024	0.139	0.020
HLHm7	0.035	0.005	0.181	0.007	0.168	0.002

Table 4.6. Synonymous (*Ks*) and non-synonymous (*Ka*) substitutions per site in genes with short CDS with introns

Short With Introns	*mel-sim*		*mel-yak*		*sim-yak*	
Gene Name	*Ks*	*Ka*	*Ks*	*Ka*	*Ks*	*Ka*
Arf72A	0.020	0.000	0.136	0.000	0.114	0.000
CG15098	0.080	0.057	0.174	0.063	0.161	0.045
san	0.109	0.005	0.182	0.013	0.138	0.018
CG13086	0.078	0.005	0.396	0.037	0.384	0.032
corn	0.078	0.004	0.199	0.015	0.140	0.012
CG31800	0.112	0.014	0.376	0.018	0.361	0.011
CG10795	0.042	0.006	0.281	0.009	0.225	0.003
Scp2	0.039	0.004	0.089	0.004	0.093	0.000
CG16817	0.135	0.008	0.216	0.015	0.221	0.012
CG2789	0.101	0.000	0.351	0.010	0.306	0.010
CG30185	0.106	0.008	0.146	0.008	0.121	0.005
CG16799	0.090	0.012	0.269	0.034	0.220	0.028
Cpr62Bc	0.035	0.000	0.081	0.000	0.056	0.000
Chd64	0.011	0.000	0.063	0.000	0.052	0.000
CG12918	0.094	0.000	0.280	0.005	0.263	0.005
Cpr49Ah	0.164	0.000	0.227	0.005	0.248	0.005
CG2113	0.185	0.011	0.338	0.033	0.288	0.026
RpL11	0.038	0.000	0.084	0.000	0.067	0.000

Table 4.7. Synonymous (*Ks*) and non-synonymous (*Ka*) substitutions per site in genes with short CDS with and without introns

Short CDS	No introns		With Introns		p-value *	
	Ks	*Ka*	*Ks*	*Ka*	*Ks*	*Ka*
mel-sim	0.072	0.005	0.084	0.007	0.478	0.881
mel-yak	0.184	0.013	0.216	0.015	0.352	0.617
sim-yak	0.162	0.012	0.192	0.012	0.497	0.865
Total Divergence	0.209	0.015	0.246	0.017	0.308	0.764

* Mann-Whitney U test.

Table 4.8. ENC values (codon bias measure) for genes with short CDS with and without introns

Short No Introns		ENC		Short With Introns		ENC	
Gene Name	*mel*	*sim*	*yak*	Gene Name	*mel*	*sim*	*yak*
CG10839	50.16	50.64	54.00	Arf72A	40.53	39.56	42.73
PGRP-SC1a	26.76	26.57	27.28	CG15098	42.82	41.85	39.63
CG7197	46.76	45.30	49.11	san	45.10	46.43	43.03
Roughened	37.35	38.74	36.95	CG13086	51.58	51.55	53.46
Ubc-E2H	44.46	43.23	41.08	corn	56.00	51.84	52.75
NC2β	38.97	33.69	34.71	CG31800	61.00	61.00	61.00
PGRP-SC2	33.42	31.26	36.17	CG10795	38.28	37.15	38.81
E(spl)	40.77	38.57	37.74	Scp2	36.09	35.48	36.30
CG15696	52.22	54.28	49.93	CG16817	42.47	42.88	39.43
CG13308	57.12	57.63	52.73	CG2789	46.21	42.49	40.95
CG4764	33.05	29.65	30.95	CG30185	38.44	38.14	38.81
mRpL12	40.25	40.53	40.66	CG16799	47.59	47.31	51.05
CG34388	48.01	52.70	49.11	Cpr62Bc	34.46	30.09	30.05
CG17304	40.55	42.80	37.76	Chd64	30.74	29.93	32.04
Bro	60.95	58.77	53.34	CG12918	48.29	46.40	50.16
Vhl	37.93	37.75	44.13	Cpr49Ah	49.58	47.08	48.52
ATPsyn-d	30.20	28.85	29.39	CG2113	50.73	56.76	51.94
CG34366	36.94	33.70	34.68	RpL11	35.58	34.00	36.07
HLHm7	40.69	40.90	44.76				

Table 4.9. Comparison of ENC values for genes with short CDS with and without introns

Short CDS	*mel*	*sim*	*yak*	All Species
ENC No Introns	41.92	41.35	41.29	41.52
ENC With Introns	44.19	43.33	43.71	43.74
P-value*	0.40	0.46	0.40	0.18

* Mann-Whitney U test.

Table 4.10. Summary of polymorphism data for genes with short CDS with and without introns

Short CDS		Aligned	Analyzed	Synonymous Sites					Nonsynonymous sites				
	n	(bp)	(bp)	Total	S	θ	π	Taj D	Total	S	θ	π	Taj D
No Introns													
mel	11	9966	9582	2759	91	0.0111	0.0123	0.450*	6823	23	0.0012	0.0012	0.293 (NS)
sim	16	9966	9693	2793	146	0.0158	0.0171	0.358 (NS)	6900	27	0.0012	0.0011	-0.188(NS)
yak	16	9966	9354	2699	218	0.0247	0.0199	-0.792*	6655	45	0.002	0.0016	-0.858*
With Introns													
mel	11	9711	9543	2701	80	0.0101	0.0106	0.237(NS)	6842	30	0.0015	0.0014	-0.287 (NS)
sim	16	9711	9387	2663	148	0.0168	0.0189	0.564*	6724	33	0.0015	0.0012	-0.777*
yak	16	9711	9516	2695	229	0.0256	0.0234	-0.375*	6821	48	0.0021	0.0016	-1.006*

*Indicates significant deviation from neutral expectations for Tajima's D. No significant difference is detected in either synonymous or nonsynonymous θ between genes with and without introns

Table 4.11. Synonymous (*Ks*) and non-synonymous (*Ka*) substitutions per site in lateral and central regions of genes with long CDS without introns

Long	*mel-sim*						*mel-yak*						*sim-yak*					
No Introns	5' region		central		3' region		5' region		central		3' region		5' region		central		3' region	
Gene Name	*Ks*	*Ka*	*Ks*	*Ka*	*Ks*	*Ka*	*Ks*	*Ka*	*Ks*	*Ka*	*Ks*	*Ka*	*Ks*	*Ka*	*Ks*	*Ka*	*Ks*	*Ka*
Wnt5	0.053	0.005	0.067	0.000	0.059	0.003	0.170	0.007	0.146	0.006	0.204	0.005	0.146	0.007	0.145	0.006	0.219	0.008
lin	0.084	0.013	0.117	0.005	0.056	0.000	0.157	0.032	0.284	0.013	0.262	0.000	0.141	0.024	0.201	0.009	0.223	0.000
CG10321	0.093	0.007	0.123	0.023	0.069	0.007	0.153	0.008	0.308	0.035	0.237	0.014	0.139	0.009	0.222	0.021	0.184	0.005
CG14411	0.133	0.014	0.123	0.004	0.158	0.023	0.279	0.048	0.222	0.006	0.301	0.048	0.243	0.050	0.232	0.006	0.229	0.048
Atg9	0.074	0.002	0.080	0.002	0.116	0.009	0.152	0.024	0.187	0.006	0.241	0.028	0.189	0.022	0.172	0.004	0.145	0.019
Brd8	0.060	0.016	0.153	0.012	0.117	0.010	0.255	0.025	0.487	0.018	0.189	0.019	0.245	0.025	0.478	0.021	0.226	0.016
kek1	0.047	0.007	0.115	0.002	0.099	0.005	0.164	0.008	0.300	0.006	0.168	0.005	0.164	0.011	0.338	0.004	0.141	0.009
Snoo	0.094	0.003	0.064	0.004	0.050	0.014	0.176	0.003	0.155	0.016	0.160	0.035	0.132	0.000	0.118	0.012	0.138	0.030
iHog	0.081	0.011	0.141	0.009	0.125	0.010	0.169	0.029	0.346	0.009	0.301	0.016	0.228	0.028	0.293	0.004	0.316	0.007
kek2	0.128	0.000	0.116	0.002	0.012	0.000	0.227	0.007	0.228	0.005	0.154	0.008	0.165	0.005	0.187	0.002	0.154	0.008
CG13350	0.164	0.002	0.130	0.012	0.121	0.007	0.322	0.011	0.228	0.017	0.260	0.013	0.245	0.013	0.259	0.020	0.285	0.015
gprs	0.204	0.000	0.085	0.011	0.061	0.002	0.426	0.003	0.222	0.019	0.134	0.007	0.333	0.003	0.221	0.013	0.136	0.005
CG17075	0.078	0.015	0.061	0.002	0.052	0.008	0.324	0.036	0.263	0.017	0.184	0.020	0.270	0.032	0.211	0.015	0.204	0.017
CG5669	0.067	0.013	0.078	0.004	0.129	0.012	0.196	0.024	0.210	0.014	0.170	0.017	0.171	0.022	0.237	0.009	0.143	0.014
Ranbp21	0.055	0.003	0.130	0.009	0.119	0.005	0.189	0.005	0.261	0.004	0.244	0.003	0.150	0.008	0.173	0.005	0.220	0.009
CG18265	0.056	0.010	0.085	0.010	0.061	0.003	0.138	0.030	0.232	0.022	0.219	0.034	0.089	0.020	0.169	0.019	0.201	0.031
Tollo	0.120	0.000	0.060	0.000	0.049	0.002	0.259	0.002	0.220	0.000	0.130	0.007	0.282	0.002	0.206	0.000	0.112	0.005
18 wheeler	0.047	0.004	0.065	0.000	0.073	0.000	0.160	0.009	0.202	0.002	0.150	0.000	0.141	0.007	0.203	0.002	0.110	0.000
mus101	0.098	0.012	0.093	0.011	0.142	0.011	0.248	0.019	0.346	0.035	0.275	0.036	0.242	0.023	0.241	0.033	0.199	0.042
Toll-7	0.017	0.003	0.088	0.000	0.103	0.000	0.167	0.010	0.159	0.000	0.147	0.010	0.148	0.008	0.198	0.000	0.137	0.010
Atg2	0.083	0.004	0.109	0.013	0.113	0.003	0.249	0.004	0.275	0.026	0.272	0.004	0.201	0.000	0.294	0.013	0.223	0.000

Table 4.12. Synonymous (*Ks*) and non-synonymous (*Ka*) substitutions per site in lateral and central regions of genes with long CDS with introns

Long With	*mel-sim*						*mel-yak*						*sim-yak*					
Introns	5' region		central		3' region		5' region		central		3' region		5' region		central		3' region	
GeneName	*Ks*	*Ka*	*Ks*	*Ka*	*Ks*	*Ka*	*Ks*	*Ka*	*Ks*	*Ka*	*Ks*	*Ka*	*Ks*	*Ka*	*Ks*	*Ka*	*Ks*	*Ka*
big bang	0.117	0.023	0.069	0.015	0.080	0.007	0.273	0.033	0.190	0.027	0.226	0.038	0.259	0.024	0.143	0.011	0.144	0.041
CG7065	0.125	0.024	0.162	0.043	0.114	0.023	0.244	0.068	0.235	0.076	0.340	0.077	0.169	0.076	0.136	0.045	0.272	0.073
APC-like	0.084	0.004	0.082	0.014	0.105	0.005	0.215	0.010	0.192	0.025	0.315	0.014	0.213	0.007	0.161	0.015	0.303	0.014
Caf1-180	0.095	0.045	0.100	0.012	0.066	0.030	0.262	0.086	0.381	0.033	0.213	0.059	0.203	0.083	0.367	0.047	0.179	0.056
CG15311	N/A	N/A	0.093	0.026	0.113	0.011	N/A	N/A	0.269	0.064	0.284	0.026	N/A	N/A	0.266	0.054	0.230	0.019
tamas	0.178	0.012	0.137	0.011	0.136	0.004	0.285	0.036	0.270	0.031	0.358	0.013	0.262	0.039	0.257	0.024	0.252	0.013
RpI135	0.138	0.006	0.116	0.006	0.094	0.000	0.210	0.023	0.318	0.029	0.243	0.002	0.196	0.021	0.309	0.028	0.255	0.002
Usp7	0.196	0.002	0.137	0.002	0.188	0.012	0.336	0.022	0.248	0.010	0.317	0.044	0.232	0.024	0.161	0.008	0.190	0.036
CG8915	N/A	N/A	0.119	0.023	0.087	0.055	N/A	N/A	0.233	0.054	0.267	0.098	N/A	N/A	0.282	0.046	0.304	0.095
CG15431	0.076	0.020	0.114	0.005	0.065	0.015	0.138	0.037	0.119	0.009	0.191	0.040	0.099	0.036	0.117	0.008	0.233	0.039
Bre1	0.055	0.000	0.084	0.002	0.082	0.000	0.151	0.002	0.121	0.004	0.213	0.003	0.127	0.002	0.103	0.002	0.154	0.003
CG11070	0.068	0.007	0.111	0.014	0.114	0.004	0.364	0.065	0.293	0.090	0.296	0.028	0.328	0.061	0.300	0.089	0.323	0.031
CG6040	0.126	0.011	0.078	0.012	0.074	0.000	0.217	0.026	0.188	0.028	0.170	0.005	0.138	0.023	0.123	0.028	0.136	0.005

Table 4.13. Synonymous (*Ks*) and non-synonymous (*Ka*) substitutions per site in lateral and central regions in genes with long CDS with and without introns

Long CDS	*Ks*			p value*		*Ka*			p value*	
	5' region	Central	3' region	5'-3'	Sides-Center	5' region	Central	3' region	5'-3'	Sides-Center
No Introns										
mel-sim	0.087	0.099	0.090	0.529	0.174	0.007	0.006	0.006	0.624	0.667
mel-yak	0.218	0.251	0.210	0.881	0.047	0.016	0.013	0.016	0.803	0.259
sim-yak	0.194	0.229	0.188	0.992	0.019	0.015	0.010	0.014	0.631	0.078
Total Divergence	0.250	0.290	0.244	0.535	0.019	0.019	0.015	0.018	0.976	0.174
With Introns										
mel-sim	0.114	0.108	0.102	0.223	0.984	0.014	0.014	0.013	0.050	0.660
mel-yak	0.245	0.235	0.264	0.298	0.156	0.037	0.037	0.034	0.384	0.984
sim-yak	0.202	0.210	0.229	0.549	0.226	0.036	0.031	0.033	0.223	0.472
Total Divergence	0.281	0.276	0.297	0.610	0.136	0.043	0.041	0.040	0.298	0.873

*Wilcoxon Signed-Rank test

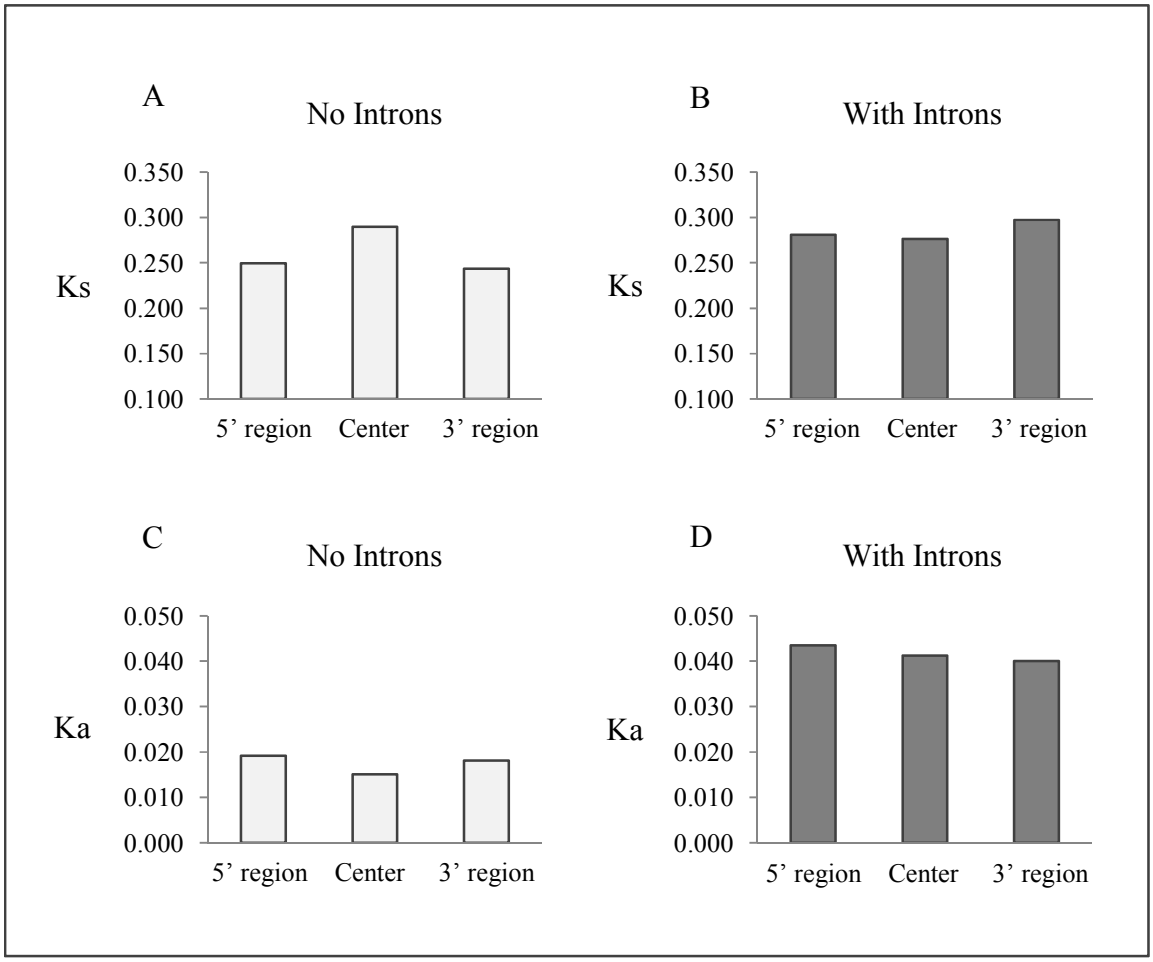

Figure 4.1. Synonymous (*Ks*) and non-synonymous (*Ka*) substitutions per site (total divergence) in genes with long CDS with and without introns.

Table 4.14. Probabilities that *Ks* and *Ka* values do not differ between genes with long CDS with and without introns

Long CDS	p values* for *Ks* comparison				p values* for *Ka* comparison			
	5' region	Central	3' region	All regions	5' region	Central	3' region	All regions
mel-sim	0.103	0.478	0.478	0.057	0.177	0.007	0.337	0.004
mel-yak	0.285	0.834	0.018	0.070	0.020	0.003	0.077	<0.0001
sim-yak	0.810	0.459	0.048	0.430	0.023	0.006	0.066	<0.0001
Total Divergence	0.259	0.834	0.037	0.063	0.028	0.003	0.089	<0.0001

*Mann-Whitney U test

Table 4.15. ENC values (codon bias measure) for genes with long CDS without introns

Long CDS No Introns	ENC								
	mel			*sim*			*yak*		
Gene Name	5' region	central	3' region	5' region	central	3' region	5' region	central	3' region
Wnt5	47.07	55.37	46.44	41.91	54.74	44.68	49.03	52.48	47.83
lin	40.52	45.77	32.40	41.07	42.25	31.58	48.11	48.35	36.68
CG10321	44.70	46.31	43.84	40.19	48.20	42.97	37.68	45.27	42.09
CG14411	47.67	42.91	43.46	41.80	43.15	38.36	42.84	43.03	41.29
Atg9	57.74	59.12	48.12	59.63	57.51	43.95	53.19	59.92	48.22
Brd8	33.14	49.77	49.79	36.25	51.40	52.88	37.13	58.49	46.23
kek1	39.51	46.61	55.60	39.39	49.01	48.05	37.76	48.90	44.54
Snoo	47.60	56.22	54.13	50.48	56.47	51.74	48.00	54.93	58.62
iHog	50.31	55.30	53.13	52.85	54.75	51.06	54.19	56.67	45.49
kek2	45.56	46.24	45.69	42.73	43.39	45.77	38.98	44.37	43.27
CG13350	48.57	48.11	49.39	44.23	55.57	49.67	44.52	47.04	50.86
gprs	54.01	56.90	55.90	49.59	52.87	53.31	52.80	51.64	50.73
CG17075	53.65	52.92	47.88	50.68	52.96	47.28	54.77	48.33	46.18
CG5669	56.84	57.40	41.41	52.92	50.86	40.39	49.77	53.71	44.07
Ranbp21	32.88	42.20	38.74	32.11	40.13	37.79	32.28	43.29	41.41
CG18265	48.90	53.34	57.46	47.14	52.40	54.01	47.49	47.33	53.40
Tollo	42.44	52.63	42.37	44.20	51.58	44.93	34.70	45.40	44.87
18 wheeler	46.84	39.10	42.31	42.74	39.99	44.91	42.39	36.91	40.00
mus101	46.60	52.87	47.02	44.09	52.89	43.88	44.08	51.85	45.49
Toll-7	46.95	51.56	50.16	46.83	50.76	49.29	52.99	47.85	49.81
Atg2	38.78	52.79	45.12	38.71	54.59	44.79	43.30	51.16	39.17

Table 4.16. ENC values (codon bias measure) for genes with long CDS with introns

Long CDS With Introns	ENC								
	mel			*sim*			*yak*		
Gene name	5' region	central	3' region	5' region	central	3' region	5' region	central	3' region
big bang	59.48	54.10	60.98	58.68	55.73	58.74	56.01	50.09	53.72
CG7065	47.33	49.77	46.00	45.48	42.42	61.00	44.83	36.19	56.09
APC-like	48.17	57.76	49.10	47.85	54.67	49.12	43.93	51.22	48.45
Caf1-180	51.45	43.21	47.13	49.63	45.64	40.29	44.63	46.42	38.70
CG15311	N/A	51.06	57.09	N/A	53.85	56.28	N/A	52.13	52.22
tamas	50.50	49.64	47.22	49.11	49.98	44.53	45.77	50.47	43.28
RpI135	53.26	61.00	50.72	53.64	60.53	55.41	46.46	54.45	49.86
Usp7	49.09	44.97	46.91	48.89	40.85	40.62	44.15	42.81	42.86
CG8915	N/A	53.32	54.57	N/A	55.92	51.42	N/A	47.48	50.24
CG15431	50.20	45.97	46.05	54.34	45.46	45.44	47.17	46.80	44.08
Bre1	47.71	37.96	36.54	47.57	40.90	35.76	50.43	40.50	40.23
CG11070	42.91	52.34	48.48	40.52	53.84	46.68	39.63	48.65	46.03
CG6040	45.04	51.03	39.76	44.27	47.99	39.80	41.76	50.20	42.27

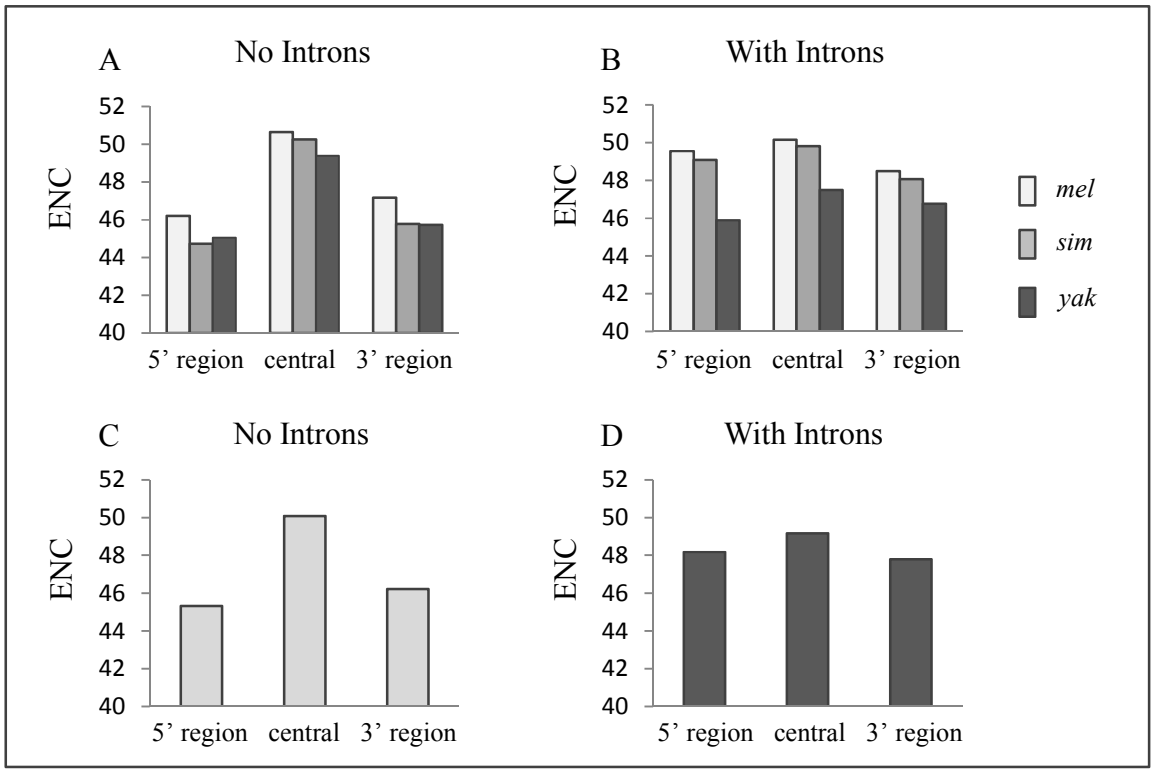

Figure 4.2. ENC values for genes with long CDS with and without introns. A,B. ENC values for each region and species separately. C,D. ENC values averaged over the three species.

Table 4.17. ENC values for genes with long CDS with and without introns

Long CDS	ENC			p value*	
	5' region	Central	3' region	5'-3'	Sides-Center
No Introns					
mel	46.20	50.64	47.16	0.529	0.002
sim	44.74	50.26	45.78	0.263	0.0002
yak	45.05	49.38	45.73	0.576	0.004
All species	45.33	50.09	46.22	0.242	< 0.0001
With Introns					
mel	49.56	50.16	48.50	0.095	0.660
sim	49.09	49.83	48.08	0.384	0.472
yak	45.89	47.49	46.77	0.944	0.390
All species	48.18	49.16	47.79	0.124	0.298

*Wilcoxon Signed-Rank Test

Table 4.18. Probabilities that ENC values do not differ between genes with and without introns

Long CDS	p value*			
	5' region	Central	3' region	All regions
mel	0.103	0.667	0.549	0.034
sim	0.043	0.944	0.503	0.358
yak	0.780	0.435	0.596	0.294
All species	0.034	0.358	0.294	0.180

*Mann-Whitney U test

Table 4.19. Summary of polymorphism data for genes with long CDS with and without introns

Long CDS			Aligned	Analyzed	Synonymous Sites					Nonsynonymous sites				
		n	(bp)	(bp)	Total	S	θ	π	Taj D	Total	S	θ	π	Taj D
No Introns														
mel	5' region	11	13131	12129	3587	122	0.0116	0.0125	0.350(NS)	8542	34	0.0014	0.0011	-1.017*
	center	11	13251	13167	3795	148	0.0133	0.0147	0.509*	9372	32	0.0012	0.0012	0.157 (NS)
	3' region	11	12903	12114	3550	112	0.0108	0.0107	-0.020(NS)	8564	18	0.0007	0.0006	-0.763 (NS)
sim	5' region	16	13131	12723	3762	220	0.0176	0.0199	0.568*	8961	38	0.0013	0.0014	0.426(NS)
	center	16	13251	12564	3623	263	0.0219	0.0245	0.527*	8941	27	0.0009	0.0009	0.108(NS)
	3' region	16	12903	12471	3645	248	0.0205	0.0216	0.225(NS)	8826	47	0.0016	0.0014	-0.501(NS)
yak	5' region	16	13131	12849	3805	316	0.0250	0.0219	-0.546*	9044	72	0.0024	0.0015	-1.669*
	center	16	13251	13041	3770	380	0.0304	0.0246	-0.834*	9271	74	0.0024	0.0021	-0.586*
	3' region	16	12903	12354	3615	363	0.0303	0.0257	-0.659*	8739	57	0.0020	0.0014	-1.143*
With Introns														
mel	5' region	11	6672	6549	1875	55	0.0100	0.0104	0.158(NS)	4674	26	0.0019	0.0021	0.464(NS)
	center	11	8244	7992	2285	82	0.0123	0.0132	0.360(NS)	5707	30	0.0018	0.0021	0.670(NS)
	3' region	11	8172	7977	2332	51	0.0075	0.0077	0.154(NS)	5645	33	0.0020	0.0017	-0.610(NS)
sim	5' region	16	6672	6483	1862	137	0.0222	0.0227	0.098(NS)	4621	35	0.0023	0.0023	-0.053(NS)
	center	16	8244	7881	2257	134	0.0179	0.0192	0.328(NS)	5624	42	0.0023	0.0018	-0.823*
	3' region	16	8172	7848	2296	111	0.0146	0.0156	0.294(NS)	5552	26	0.0014	0.0014	-0.024(NS)
yak	5' region	16	6672	6534	1878	157	0.0252	0.0197	-0.950*	4656	62	0.0040	0.0029	-1.192*
	center	16	8244	8124	2335	224	0.0289	0.0240	-0.743*	5789	60	0.0031	0.0019	-1.647*
	3' region	16	8172	7821	2295	232	0.0305	0.0262	-0.610*	5526	82	0.0045	0.0035	-0.962*

*Indicates significant deviation from neutral expectations for Tajima's D. No significant difference is detected in either synonymous or nonsynonymous θ between regions of genes with and without introns as well as between genes with and without introns

Table 4.20. Synonymous and nonsynonymous Pol/Div values across coding sequence of genes with long CDS

Long CDS	Synonymous Pol/Div values			p-value		
	5'	center	3'	5-center	center-3	5'-3'
No Introns						
mel-sim	2.100	2.056	2.186	1.000	0.719	0.308
mel-yak	0.907	0.947	0.986	0.529	0.667	0.308
sim-yak	1.282	1.376	1.472	1.000	0.529	0.285
Total tree	1.249	1.299	1.376	0.617	0.529	0.174
With Introns						
mel-sim	2.010	1.643	1.280	0.589	0.142	0.223
mel-yak	0.801	0.978	0.801	0.114	0.156	0.741
sim-yak	1.349	1.253	1.178	0.298	0.226	0.741
Total tree	1.226	1.211	1.026	0.298	0.136	0.259
Long CDS	Nonsynonymous Pol/Div values			p-value		
	5'	center	3'	5-center	center-3	5'-3'
No Introns						
mel-sim	1.400	1.167	1.570	0.352	0.711	0.960
mel-yak	0.723	1.003	0.610	0.889	0.080	0.352
sim-yak	0.784	1.073	0.970	0.857	0.873	0.719
Total tree	0.858	1.062	0.901	0.667	0.881	0.675
With Introns						
mel-sim	1.504	1.138	0.907	>0.05	>0.05	>0.05
mel-yak	0.643	0.454	0.642	0.897	0.280	>0.05
sim-yak	0.711	0.622	0.626	0.459	0.984	>0.05
Total tree	0.788	0.620	0.675	0.522	0.542	>0.05

*Wilcoxon Signed-Rank Test

Table 4.21. Comparison of *Ka/Ks* between genes with long CDS with and without introns

Long CDS	*mel-sim*	*mel-yak*	*sim-yak*	Total Tree
Ka/Ks No Introns	0.071	0.067	0.065	0.069
Ka/Ks With Introns	0.127	0.146	0.156	0.146
P-value*	0.047	<0.001	<0.001	<0.001

*Mann-Whitney U test

Table 4.22. Proportion of adaptive substitutions in genes with long and short CDS with and without introns

	mel-sim	*mel-yak*	*sim-yak*	total tree
Long CDS				
No Introns				
5' region	0.386	0.181	0.397	0.327
center	0.422	-0.095	0.195	0.158
3' region	0.363	0.359	0.414	0.383
All regions	0.389	0.159	0.348	0.296
p-value*	0.0013	0.121	<0.0001	0.0003
With Introns				
5' region	0.468	0.314	0.580	0.487
center	0.519	0.630	0.515	0.573
3' region	0.354	0.236	0.551	0.407
All regions	0.448	0.417	0.546	0.489
p-value*	0.0003	<0.0001	<0.0001	<0.0001
Short CDS				
No Introns	0.172	0.045	0.196	0.133
p-value*	0.597	0.852	0.361	0.508
With Introns	0.197	0.119	0.13	0.117
p-value*	0.516	0.594	0.569	0.558

*G test of independence for the ratios of polymorphism to divergence for putatively neutral and nonsynonymous sites based on the combined polymorphism and divergence counts for all regions

Table 4.23. Proportion of adaptive substitutions in each gene with long CDS

	No Introns		With Introns		
Gene Name	α	p-value*	Gene Name	α	p-value*
Wnt5	0.135	1.000	big bang	0.344	0.352
lin	-0.045	1.000	CG7065	0.868	1.15×10^{-6}*/*
CG10321	0.658	0.063	APC-like	0.040	1.000
CG14411	0.732	0.002*/*	Caf1-180	0.424	0.093
Atg9	0.706	0.003*/NS	CG15311	-0.174	0.701
Brd8	0.444	0.141	tamas	0.020	1.000
kek1	0.209	0.791	RpI135	0.692	0.015*/NS
Snoo	0.254	0.656	CG8915	0.783	0.004*/*
iHog	-0.150	0.849	CG15431	0.258	0.546
kek2	0.476	0.341	Bre1	0.762	0.173
CG13350	-0.108	1.000	CG11070	0.798	2×10^{-5}*/*
gprs	0.360	0.462	CG6040	-0.458	0.611
CG17075	-2.733	0.023			
CG5669	-0.583	0.312			
Ranbp21	-0.500	0.569			
CG18265	0.226	0.672			
Tollo	0.263	0.714			
18 wheeler	-0.086	1.000			
mus101	0.331	0.347			
Toll-7	-0.113	1.000			
Atg2	0.196	0.813			

/ Fisher's exact test before and after sequential Bonferroni correction

CONCLUSION

The work presented in this thesis examines the short-range consequences of the Hill-Robertson effect in *Drosophila*. Such interference has been proposed to shape local patterns of polymorphism and divergence as well as a number of features of genomic architecture.

Using computer simulations, I show that linkage effects are not removed completely even in areas of high recombination. Consequently, estimates of selection and the proportion of adaptive substitutions using methods that do not take into account these effects may be misleading when applied to situations in which putatively neutral and selected sites are spatially separated. Whether or not trends observed by the simulation approach are detectable in the *Drosophila* genome should be evaluated with empirical studies. Further simulation-based investigations can be carried out in order to examine the extent to which the removal of rare polymorphisms alters estimates of the proportion of adaptive substitutions.

The limited analysis of the results obtained from the experimental study of genes with and without introns does not permit one to draw definite conclusions with regard to the effect of intron presence on the effectiveness of selection. Given higher rates of adaptation associated with genes containing introns, it would be informative to perform similar analyses with a larger data set in which slow- and fast-evolving genes could be analyzed separately. The examination of the slow-evolving set would allow the assessment of the role of introns in enhancing the effectiveness of selection on synonymous sites, while the examination of the fast-evolving data set would allow a

more careful evaluation of the intron-mediated adaptation hypothesis. In this respect, the simulation-based approach can be employed to investigate the conditions under which intron presence can increase levels of adaptation.

REFERENCES

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

Akashi, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. Genetics **139:** 1067-1076.

Akashi, H., 1996 Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. Genetics **144:** 1297 - 1307.

Andolfatto, P. and M. Przeworski, 2000 A genome-wide departure from the standard neutral model in natural populations of *Drosophila*. Genetics **156:** 257-268.

Andolfatto, P. and M. Przeworski, 2001 Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*. Genetics **158:** 657-665.

Andolfatto, P., 2005 Adaptive evolution of non-coding DNA in *Drosophila*. Nature **437:** 1149-1152.

Andolfatto, P., 2007 Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melangoaster* genome. Genome Res. **17**:1755-1762.

Bachtrog, D., 2003 Adaptation shapes patterns of genome evolution on sexual and asexual chromosomes in *Drosophila*. Nature Genet. **34:** 215-219.

Bachtrog, D., 2008 Positive selection at the binding sites of the male-specific lethal complex involved in dosage compensation in *Drosophila*. Genetics **180:** 1123-1129.

Baines, J., S. Sawyer, D. Hartl and J. Parsch, 2008 Effects of X-linkage and sex-biased gene expression on the rate of adaptive protein evolution in *Drosophila*. Mol Biol Evol **25:** 1639-1650.

Barrier, M., C. Bustamante, J. Yu and M. Purugganan, 2003 Selection on rapidly evolving proteins in the *Arabidopsis* genome. Genetics **163:** 723-733.

Barton, N., 1995 Linkage and the limits to natural selection. Genetics **140:** 821-841.

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

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

Begun, D., A. Holloway, K. Stevens, L.Hillier, Y. Poh *et al.*, 2007 Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila* simulans. PLoS Biol **5:** e310.

Berry, A., J. Ajioka and M. Kreitman, 1991 Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. Genetics **129:** 1111-1117.

Betancourt, A. and D. Presgraves, 2002 Linkage limits the power of natural selection in *Drosophila*. Proc Natl Acad Sci USA **99:** 13616 - 13620.

Betancourt, A., J. Welch and B. Charlesworth, 2009 Reduced effectiveness of selection caused by a lack of recombination. Curr Biol **19:** 655-660.

Bierne, N. and A. Eyre-Walker, 2004 The genomic rate of adaptive amino acid substitution in *Drosophila*. Mol Biol Evol **21:** 1350-1360.

Birky, C. and J. Walsh, 1988 Effects of linkage on rates of molecular evolution. Proc Natl Acad Sci USA **85:** 6414-6418.

Bulmer, M., 1991 The selection-mutation-drift theory of synonymous codon usage. Genetics **129:** 897 - 907.

Bustamante, C., R. Nielsen, S. Sawyer, K. Olsen, M. Purugganan *et al.*, 2002 The cost of inbreeding in *Arabidopsis*. Nature **416:** 531 - 534.

Bustamante, C., R. Nielsen and D. Hartl, 2003 Maximum likelihood and Bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data. Theor Popul Biol **63:** 91-103.

Carvalho, A. and A. Clark, 1999 Genetic recombination: Intron size and natural selection. Nature **401:** 344-344.

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

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

Charlesworth, D., B. Charlesworth and M. T. Morgan, 1995 The Pattern of Neutral Molecular Variation Under the Background Selection Model. Genetics **141:** 1619-1632.

Charlesworth, J. and A. Eyre-Walker, 2008 The McDonald-Kreitman test and slightly deleterious mutations. Mol Biol Evol **25:** 1007-1015.

Cirulli, E., R. Kliman and M. Noor, 2007 Fine-scale crossover rate heterogeneity in *Drosophila pseudoobscura*. J Mol Evol **64:** 129-135.

Comeron, J. M., 1995 A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. J Mol Evol **41:** 1152-1159.

Comeron, J. M. and M. Aguade, 1996 Synonymous substitutions in the Xdh gene of *Drosophila*: heterogeneous distribution along the coding region. Genetics **144:** 1053-1062.

Comeron, J. M., 1997 *Estudi de la variabilitat nucleotidica a Drosophila: regio Xdh a D. subobscura. Ph.D. Thesis*. Universitat de Barcelona, Barcelona, Spain.

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

Comeron, J. M. and M. Kreitman, 2000 The correlation between intron length and recombination in *Drosophila*. Dynamic equilibrium between mutational and selective forces. Genetics **156:** 1175-1190.

Comeron, J. M., 2001 What controls the length of noncoding DNA? Curr Opin Genetics Dev **11:** 652-659.

Comeron, J. M. and M. Kreitman, 2002 Population, evolutionary and genomic consequences of interference selection. Genetics **161:** 389-410.

Comeron, J. M. and T. B. Guthrie, 2005 Intragenic Hill-Robertson interference influences selection intensity on synonymous mutations in *Drosophila*. Mol Biol Evol **22:** 2519-2530.

Comeron, J.M., 2006 Weak selection and recent mutational changes influence polymorphic synonymous mutations in humans. Proc Natl Acad Sci USA **103:** 6940-6945.

Comeron, J. M., A. Williford and R. M. Kliman, 2008 The Hill-Robertson effect: evolutionary consequences of weak selection and linkage in finite populations. Heredity **100:** 19-31.

Cooper, G. M., M. Brudno, E. A. Stone, I. Dubchak, S. Batzoglou *et al.*, 2004 Characterization of evolutionary rates and constraints in three mammalian genomes. Genome Res **14:** 539-548.

Crow, J. F. and K. Kimura, 1970 *An Introduction to Population Genetics Theory*. Harper and Row, New York.

Dvorak, J., M.-C. Luo and Z.-L. Yang, 1998 Restriction fragment length polymorphism and divergence in the genomic regions of high and low recombination in self-fertilizing and cross-fertilizing Aegilops species. Genetics **148:** 423-434.

Ewens, W. J., 2004 *Mathematical Population Genetics I. Theoretical Introduction*. Springer, New York.

Eyre-Walker, A., 2002 Changing effective population size and the McDonald-Kreitman test. Genetics **162:** 2017-2024.

Eyre-Walker, A., 2006 The genomic rate of adaptive evolution. Trends Ecol Evol **21:** 569-575.

.
Eyre-Walker, A. and P. D. Keightley, 2007 The distribution of fitness effects of new mutations. Nature Rev Genet **8:** 610-618.

Eyre-Walker, A. and P. D. Keightley, 2009 Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol **26**: 2097-2108.

Fay, J. C. and C. I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405-1413.

Fay, J. C., G. J. Wyckoff and C.-I. Wu, 2001 Positive and negative selection on the human genome. Genetics **158:** 1227-1234.

Fay, J. C., G. J. Wyckoff and C. I. Wu, 2002 Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. Nature **415:** 1024-1026.

Fay, J. C. and C.-I. Wu, 2005 Detecting hitchhiking from patterns of DNA polymorphism pp. 65-77 in *Selective Sweep*, edited by D. NURMINSKY. Landes Bioscience/Eurekah.com, Georgetown.

Felsenstein, J., 1974 The evolutionary advantage of recombination. Genetics **78:** 737-756.

Felsenstein, J., 1988 Sex and evolution of recombination, pp. 74-86 in *The evolution of sex. An examination of current ideas.*, edited by R. E. MICHOD and B. R. LEVIN. Sinauer associates inc., Sunderland.

Gaffney, D. J. and P. D. Keightley, 2006 Genomic selective constraints in murid noncoding DNA. PLoS Genet **2:** e204.

Gillespie, J. H., 1984 The molecular clock may be an episodic clock. Proc Natl Acad Sci USA **81:** 8009-8013.

Gillespie, J., 1989 Lineage effects and the index of dispersion of molecular evolution. Mol Biol Evol **6:** 636-647.

Gillespie, J. H., 1991 *The causes of molecular evolution*. Oxford University Press, Oxford.

Gordo, I., A. Navarro and B. Charlesworth, 2002 Muller's ratchet and the pattern of variation at a neutral locus. Genetics **161:** 835-848.

Haag-Liautard, C., M. Dorris, X. Maside, S. Macaskill, D. Halligan *et al.*, 2007 Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. Nature **445:** 82 - 85.

Haddrill, P. R., D. L. Halligan, D. Tomaras and B. Charlesworth, 2007 Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. Genome Biol **8:** R18.

Haddrill, P. R., D. Bachtrog and P. Andolfatto, 2008 Positive and negative selection on noncoding DNA in *Drosophila simulans*. Mol Biol Evol **25:** 1825-1834.

Haldane, J., 1957 The cost of natural selection. Journal of Genetics **55:** 511-524.

Halligan, D. L. and P. D. Keightley, 2006 Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. Genome Res **16:** 875-884.

Harris, H., 1966 Enzyme polymorphisms in man. Proc R Soc Lond Ser B **164:** 298-310.

Hartl, D. L., E. N. Moriyama and S. A. Sawyer, 1994 Selection intensity for codon bias. Genetics **138:** 227-234.

Hey, J. and R. M. Kliman, 2002 Interactions between natural selection, recombination and gene density in the genes of *Drosophila*. Genetics **160:** 595-608.

Hill, W. G., and A. Robertson, 1966 The effects of linkage on limits to artificial selection. Genet Res **8:** 269-294.

Hudson, R. R., M. Kreitman and M. Aguade, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153-159.

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

Hurst, L., 2009 Genetics and the understanding of selection. Nature Rev Genet **10:** 83-93.

Ikemura, T., 1985 Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol **2:** 13-34.

Jensen-Seaman, M. I., T. S. Furey, B. A. Payseur, Y. Lu, K. M. Roskin *et al.*, 2004 Comparative recombination rates in the rat, mouse, and human genomes. Genome Res **14:** 528-538.

Jensen, J. D., K. R. Thornton and P. Andolfatto, 2008 An approximate bayesian estimator suggests strong, recurrent selective sweeps in *Drosophila*. PLoS Genet **4:** e1000198.

Kaiser, V. B. and B. Charlesworth, 2009 The effects of deleterious mutations on evolution in non-recombining genomes. Trends Genet **25:** 9-12.

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The "Hitchhiking Effect" revisited. Genetics **123:** 887-899.

Kim, Y. and W. Stephan, 2000 Joint effects of genetic hitchhiking and background selection on neutral variation. Genetics **155:** 1415-1427.

Kim, Y. and W. Stephan, 2003 Selective sweeps in the presence of interference among partially linked loci. Genetics **164:** 389-398.

Kim, Y., 2004 Effect of strong directional selection on weakly selected mutations at linked sites: implication for synonymous codon usage. Mol Biol Evol **21:** 286 - 294.

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

Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61:** 893-903.

Kimura, M., 1971 Theoretical foundation of population genetics at the molecular level. Theor Popul Biol **2:** 174-208.

Kimura, M., 1977 Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature **267**: 275-276.

Kimura, M. and T. Ohta, 1971a Protein polymorphism as a phase of molecular evolution. Nature **229:** 467-469.

Kimura, M. and T. Ohta, 1971b *Theoretical aspects of population genetics*. Princeton University Press, Princeton.

Kimura, M., 1983 *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge.

Kimura, M., 1986 DNA and the neutral theory. Philos Trans R Soc Lond Ser B **312:** 343-354.

King, J. L. and T. H. Jukes, 1969 Non-Darwinian evolution. Science **164:** 788-798.

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

Kohn, M. H., S. Fang and C.-I. Wu, 2004 Inference of positive and negative selection on the 5' regulatory regions of *Drosophila* genes. Mol Biol Evol **21:** 374-383.

Kojima, K. I. and H. E. Schaeffer, 1967 Survival process of linked genes. Evolution **21:** 518-531.

Langley, C. H. and W. M. Fitch, 1974 An examination of the constancy of the rate of molecular evolution. J Mol Evol **3:** 161-177.

Langley, C. H., J. MacDonald, N. Miyashita and M. Aguade, 1993 Lack of correlation between interspecific divergence and intraspecific polymorphism at the suppressor of forked region in *Drosophila melanogaster* and *Drosophila simulans*. Proc Natl Acad Sci USA **90:** 1800-1803.

Larracuente, A. M., T. B. Sackton, A. J. Greenberg, A. Wong, N. D. Singh *et al.*, 2008 Evolution of protein-coding genes in *Drosophila*. Trends Genet **24:** 114-123.

Lewontin, R. C. and J. L. Hubby, 1966 A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of

heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics **54:** 595-609.

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

Li, W-H., T. Gojobori and M. Nei, 1981 Pseudogenes as a paradigm of neutral evolution. Nature **292**: 237-239.

Li, W-H., 1987 Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J Mol Evol **24:** 337-345.

Loewe, L. and B. Charlesworth, 2006 Inferring the distribution of mutational effects on fitness in Drosophila. Biology Letters **2:** 426-430.

Loewe, L., B. Charlesworth, C. Bartolome and V. Noel, 2006 Estimating selection on nonsynonymous mutations. Genetics **172:** 1079-1092.

Loewe, L. and B. Charlesworth, 2007 Background selection in single genes may explain patterns of codon bias. Genetics **175:** 1381-1393.

Macpherson, J. M., G. Sella, J. C. Davis and D. A. Petrov, 2007 Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in *Drosophila*. Genetics **177:** 2083-2099.

Maynard Smith, J. and J. Haigh, 1974 The hitch-hiking effect of a favorable gene. Genet Res **23:** 23-35.

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

McVean, G. A. and B. Charlesworth, 2000 The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics **155:** 929-944.

McVean, G. A. and J. Vieira, 2001 Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. Genetics **157:** 245-257.

Moran, P. A. P., 1962 *The Statistical Processes of Evolutionary Theory*. Clarendon Press, Oxford.

Moriyama, E. and J. Powell, 1996 Intraspecific nuclear DNA variation in *Drosophila*. Mol Biol Evol **13:** 261-277.

Moriyama, E. and J. Powell, 1998 Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. Nucleic Acids Res **26:** 3188-3193.

Moriyama, E. N. and D. L. Hartl, 1993 Codon usage bias and base composition of nuclear genes in *Drosophila*. Genetics **134:** 847-858.

Munte, A., M. Aguadé and C. Segarra, 1997 Divergence of the yellow gene between Drosophila melanogaster and D. subobscura: recombination rate, codon bias and synonymous substitutions. Genetics **147:** 165-175.

Nachman, M. W., 1997 Patterns of DNA variability at X-linked loci in *Mus domesticus*. Genetics **147:** 1303-1316.

Nachman, M. W., 2001 Single nucleotide polymorphisms and recombination rate in humans. Trends Genet **17:** 481-485.

Nachman, M. W., 2002 Variation in recombination rate across the genome:evidence and implications. Curr Opin Genetics Dev **12:** 657-663.

Nordborg, M., B. Charlesworth and D. Charlesworth, 1996 The effect of recombination on background selection. Genet Res **67:** 159-174.

Ohta, T., 1972 Population size and rate of evolution. J Mol Evol **1:** 305-314.

Ohta, T., 1973 Slightly deleterious mutant substitutions in evolution. Nature **246:** 96-98.

Ohta, T., 1974 Mutational pressure as the main cause of molecular evolution and polymorphism. Nature **252:** 351-354.

Ohta, T., 1992 The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst **23:** 263-286.

Ohta, T., 1995 Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J Mol Evol **40:** 56-63

Ohta, T. and M. Kimura, 1971 On the constancy of the evolutionary rate of cistrons. J Mol Evol **1:** 18-25.

Ohta, T. and M. Kimura, 1975 The effect of selected linked locus on heterozygosity of neutral alleles (the hitch-hiking effect). Genet Res **25:** 313-326.

Ohta, T. and M. Kimura, 1976 Hitch-hiking effect - counter reply. Genet Res **28:** 307-308.

Ortiz-Barrientos, D., A. Chang and M. Noor, 2006 A recombinational portrait of the *Drosophila pseudoobscura* genome. Genet Res **87:** 23 - 31.

Peck, J. R., 1994 A Ruby in the rubbish: beneficial mutations, deleterious mutations and the evolution of sex. Genetics **137:** 597-606.

Powell, J. R. and E. N. Moriyama, 1997 Evolution of codon usage bias in *Drosophila*. Proc Natl Acad Sci USA **94:** 7784-7790.

Presgraves, D. C., 2005 Recombination enhances protein adaptation in *Drosophila melanogaster*. Curr Biol **15:** 1651-1656.

Provine, W. B., 1971/2001 *The origins of theoretical population genetics*. University of Chicago Press, Chicago.

Qin, H., W. B. Wu, J. M. Comeron, M. Kreitman and W.-H. Li, 2004 Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. Genetics **168:** 2245-2260.

Robertson, A., 1961 A theory of limits in artifical selection. Proc R Soc Ser B **153:** 234-249.

Sawyer, S., R. Kulathinal, C. Bustamante and D. Hartl, 2003 Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. J Mol Evol **57:** S154 - 164.

Sawyer, S. A. and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics **132:** 1161-1176.

Sawyer, S. A., J. Parsch, Z. Zhang and D. L. Hartl, 2007 Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. Proc Natl Acad Sci USA **104:** 6504-6510.

Sella, G., D. A. Petrov, M. Przeworski and P. Andolfatto, 2009 Pervasive natural selection in the *Drosophila* genome? PLoS Genet **5:** e1000495.

Shapiro, J., W. Huang, C. Zhang, M. Hubisz, J. Lu *et al.*, 2007 Adaptive genic evolution in the *Drosophila* genomes. Proc Natl Acad Sci USA **104:** 2271 - 2276.

Sharp, P. and W. Li, 1987 The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol **4:** 222-230.

Shields, D., P. Sharp, D. Higgins and F. Wright, 1988 "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol **5:** 704-716.

Singh, N., C. Aquadro and A. Clark, 2009 Estimation of fine-scale recombination intensity variation in the white–echinus interval of *D. melanogaster*. J Mol Evol **69:** 42-53.

Singh, N. D., P. F. Arndt and D. A. Petrov, 2005 Genomic heterogeneity of background substitutional patterns in *Drosophila melanogaster*. Genetics **169:** 709-722.

Smith, J. M. and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet Res **23:** 23-35.

Smith, N. G. and A. Eyre-Walker, 2002 Adaptive protein evolution in *Drosophila*. Nature **415:** 1022-1024.

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

Stephan, W., T. H. E. Wiehe and M. W. Lenz, 1992 The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory. Theor Popul Biol **41:** 237-254.

Stephan, W. and C. H. Langley, 1998 DNA Polymorphism in Lycopersicon and crossing-over per physical length. Genetics **150:** 1585-1593.

Stephan, W., B. Charlesworth and G. McVean, 1999 The effect of background selection at a single locus on weakly selected, partially linked variants. Genet Res **73:** 133-146.

Tachida, H., 2000 Molecular evolution in a multisite nearly neutral mutation model. J Mol Evol **50:** 69-81.

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

Takano, T. S., 1998 Rate variation of DNA sequence evolution in the *Drosophila* lineages. Genetics **149:** 959-970.

Tenaillon, M. I., M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley *et al.*, 2001 Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. mays L.). Proc Natl Acad Sci USA **98:** 9161-9166.

True, J., J. Mercer and C. Laurie, 1996 Differences in crossover frequency and distribution among three sibling species of *Drosophila*. Genetics **142:** 507 - 523.

Waterson, R. H., K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril *et al.*, 2002 Initial sequencing and comparative analysis of the mouse genome. Nature **420:** 520-562.

Welch, J. J., 2006 Estimating the genomewide rate of adaptive protein evolution in *Drosophila*. Genetics **173:** 821-837.

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

Williford, A. and J.M. Comeron, 2010 Local effects of limited recombination: historical perspective and consequences for population estimates of adaptive evolution. J Heredity **101**: S127-S134.

Yang, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci **13:** 555-556.

Yang, Z. and J. P. Bielawski, 2000 Statistical methods for detecting molecular adaptation. Trends Ecol Evol **15:** 496-503.

Yang, Z., 2007 PAML 4: Phylogenetic analysis by Maximum Likelihood. Mol Biol Evol **24:** 1586-1591.

Zhang, Z. and J. Parsch, 2005 Positive correlation between evolutionary rate and recombination rate in *Drosophila* genes with male-biased expression. Mol Biol Evol **22:** 1945-1947.

Zuckerkandl, E. and L. Pauling, 1965 Evolutionary divergence and convergence in proteins, pp. 97-166 in *Evolving Genes and Proteins*, edited by V. BRYSON and H. J. VOGEL. Academic Press, New York.