The Effects of Natural Selection and Random Genetic Drift

in Structured Populations

by

Takahiro Maruki

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2011 by the
Graduate Supervisory Committee:

Yuseob Kim, Co-Chair
Jesse E. Taylor, Co-Chair
Priscilla E. Greenwood
Philip W. Hedrick
Michael S. Rosenberg

ARIZONA STATE UNIVERSITY

December 2011

ABSTRACT

Building mathematical models and examining the compatibility of their theoretical predictions with empirical data are important for our understanding of evolution. The rapidly increasing amounts of genomic data on polymorphisms greatly motivate evolutionary biologists to find targets of positive selection. Although intensive mathematical and statistical studies for characterizing signatures of positive selection have been conducted to identify targets of positive selection, relatively little is known about the effects of other evolutionary forces on signatures of positive selection. In this dissertation, I investigate the effects of various evolutionary factors, including purifying selection and population demography, on signatures of positive selection. Specifically, the effects on two highly used methods for detecting positive selection, one by Wright's F_{ST} and its analogues and the other by footprints of genetic hitchhiking, are investigated. In Chapters 2 and 3, the effect of purifying selection on F_{ST} is studied. The results show that purifying selection intensity greatly affects F_{ST} by modulating allele frequencies across populations. The footprints of genetic hitchhiking in a geographically structured population are studied in Chapter 4. The results demonstrate that footprints of genetic hitchhiking are significantly influenced by geographic structure, which may help scientists to infer the origin and spread of the beneficial allele. In Chapter 5, the stochastic dynamics of a hitchhiking allele are studied using the diffusion process of genetic hitchhiking conditioned on the fixation of the beneficial allele. Explicit formulae for the conditioned two-locus diffusion process of genetic hitchhiking are derived and stochastic aspects of

genetic hitchhiking are investigated. The results in this dissertation show that it is essential to model the interaction of neutral and selective forces for correct identification of the targets of positive selection.

ACKNOWLEDGMENTS

I would like to thank Drs. Yuseob Kim and Jesse Taylor for mentoring my dissertation studies. Their understanding of my research interests made it possible for me to pursue the subjects in this dissertation. I also would like to thank other members of the dissertation committee, Drs. Priscilla Greenwood, Philip Hedrick, and Michael Rosenberg for their help. In addition to my committee members, I received help from Dr. Sudhir Kumar in the dissertation studies and would like to appreciate his support. Finally, I would like to thank my colleagues in the Center for Evolutionary Medicine and Informatics and School of Life Sciences at ASU and my family.

TABLE OF CONTENTS

Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER 1: INTRODUCTION

Population genetics plays an important role in our understanding of the mechanisms of evolution. Major evolutionary processes such as speciation usually take millions of years and are difficult to directly observe. Therefore, building mathematical models of evolutionary processes and comparing their theoretical predictions with empirical data of polymorphism in extant populations are necessary for understanding the mechanisms of evolution. In this dissertation, I study the interacting effects of random genetic drift and natural selection on patterns of polymorphism in populations. All of the models in this dissertation are stochastic and include random genetic drift, which is inherent in every finite population. Although intensive mathematical and statistical studies have been conducted to study the effects of neutral and selective forces on the pattern of polymorphisms (Kimura 1983, Gillespie 1991), relatively little is known about the effects of other evolutionary forces on signatures of positive selection. With the rapidly increasing genomic data of polymorphisms in various organisms, active researches to identify targets of positive selection are conducted (Akey et al. 2002, Harr et al. 2002, Akey et al. 2004, Carlson et al. 2005, Nielsen et al. 2005, Kelley et al. 2006, Voight et al. 2006, Sabeti et al. 2007, Williamson et al. 2007, Nielsen et al. 2009, Pickrell et al. 2009, Grossman et al. 2010). Although many regions have recently been reported as targets of positive selection, the overlap of them among different genomic studies is low (Nielsen et al. 2007, Akey 2009). Many of the targets found in one study are not found in another study. This inconsistency indicates false positive and negative detections of targets of positive

selection by current methods. To correctly understand the mechanisms of evolution and apply the knowledge in practical fields such as medicine and conservation, accurate description of signatures of positive selection is critical. Recent genomic studies identified evidence for the presence of both positive and negative (purifying) selection in the genome of various organisms (e.g., Bustamante et al. 2005, Clark et al. 2007, Drosophila 12 genomes Consortium 2007, Barreiro et al. 2008, Boyko et al. 2008, Halligan et al. 2011). Moreover, many of these studies found purifying selection appears to be much more frequent than positive selection. Therefore, purifying selection may severely affect signatures of positive selection.

In this dissertation, I examine how the mathematical and statistical methods for detecting positive selection are affected by the presence of other evolutionary forces including purifying selection and population demography. In Chapters 2 and 3, I investigate the effect of purifying selection on Wright's F_{ST} and its analogues. The increasing amounts of genome-wide data of polymorphisms in various populations are being intensively used by evolutionary biologists to identify targets of population-specific positive selection (local adaptation). The distribution of F_{ST} at genomic positions is described and outliers in the distribution are considered to be potential targets of local adaptation (reviewed in Akey 2009). This approach is based on the assumption that the vast majority of positions are under the same evolutionary forces. This assumption is violated when different positions are under different degree of functional constraints. Therefore, it is important to investigate how the distribution of F_{ST} at

genomic positions is affected by the strength of purifying selection. In Chapter 2, the significance of the effect of purifying selection on F_{ST} is empirically demonstrated with genomic data of human SNPs. F_{ST} is defined and calculated at individual SNP sites in the human genome and the effect of purifying selection intensity on the distribution of F_{ST} is investigated. Recent studies on the genome-wide pattern of polymorphism in humans support widespread existence of purifying selection in the human genome (e.g., Williamson et al. 2005, Yampolsky et al. 2005, Boyko et al. 2008, Lohmueller et al. 2011). Therefore, it is important to investigate how purifying selection affects F_{ST} at human SNP sites. Assuming that the difference in evolutionary rates among sites reflects the difference in functional constraints, the strength of purifying selection at an individual SNP site is estimated by the evolutionary rate at the site (Kumar et al. 2009). The genome sequences of 36 mammalian species are used for estimating the evolutionary rate at an individual SNP site. The results suggest that stronger purifying selection diminishes F_{ST}, which is reflected as a positive correlation between evolutionary rates and F_{ST}, at human SNP sites. This relationship between purifying selection and F_{ST} is found to result from the dependence of F_{ST} on minor allele frequencies across populations. Statistical and mathematical analyses are conducted to investigate this dependence of F_{ST} on minor allele frequencies across populations. Statistical and mathematical analyses are conducted to investigate the dependence of F_{ST} on minor allele frequencies. In addition to that in identification of targets of local adaptation, the dependence of

F_{ST} on minor allele frequencies is shown to severely affect the use of F_{ST} in inference of historical amounts of gene flow.

Motivated by the empirical observation, computer simulations of population differentiation under purifying selection are conducted to predict the effect in other organisms in Chapter 3. The effects of various evolutionary forces including mutation and migration rates on F_{ST} between populations at bi-allelic loci are investigated. To understand and generalize the results observed in Chapter 2, it is important to simulate the process of genetic differentiation between populations and examine the effects of various evolutionary forces on F_{ST} at bi-allelic loci. In particular, the effect of purifying selection on F_{ST} is intensively investigated in two different models of population demography under different parameter values. The effects on its application in identification of targets of local adaptation as well as estimation of population demographic parameters are investigated. The results show that stronger purifying selection diminishes F_{ST} with wide range of parameter values. This effect of purifying selection on F_{ST} is found to be severe when the migration rate between populations is low, which predicts the effect observed in humans may be more significant in other organisms with stronger population structure.

In Chapters 4 and 5, I investigate footprints of positive selection left by genetic hitchhiking. When a beneficial allele rapidly increases in frequency, alleles at different loci closely linked to the beneficial allele also rapidly increase by genetic hitchhiking. Genetic hitchhiking helps scientists to infer whether a

fixation of an allele was caused by recent positive directional selection because it leaves characteristic footprints in genetic regions close to the target of selection. Because genetic hitchhiking tends to reduce genetic variation near the target of positive selection, evolutionary biologists identify potential targets of positive selection by finding regions with low genetic variation (Carlson et al. 2005, Nielsen et al. 2005, Williamson et al. 2007, Nielsen et al. 2009). In Chapter 4, I investigate how the beneficial allele and hitchhiking effect spread in a subdivided population. Although intensive mathematical and statistical studies have been conducted to characterize the footprints of genetic hitchhiking, most of them were conducted in a single panmictic population. However, a natural population is subdivided into demes because of geographic structure and limited amounts of migration occur among them. Therefore, in order for the beneficial allele to spread through demes, it needs to spread by migration (Morjan and Rieseberg 2004). Unlike previous studies (Slatkin and Wiehe 1998, Santiago and Caballero 2005), the effects of the geographic structure of a population on the spread of the beneficial allele and genetic hitchhiking are investigated in cases where $2Nm > 1$ but $m < s$, where $2N$, m, and s are the effective size of each deme, migration rate, and selection coefficient, respectively. These cases are important, because geographic structure of a population may not be detected at isolated neutral loci when $2Nm > 1$ but may have a significant effect on genetic hitchhiking when $m < s$. The results show the significance of 'hidden' geographic structure on genetic hitchhiking.

In Chapter 5, I investigate the stochastic dynamics of a hitchhiking allele during the process of a selective sweep. The dynamics is studied forward in time in a population using a diffusion process of genetic hitchhiking. With the increasing availability of time-series data of polymorphisms in various organisms, it is important to study the dynamics of the hitchhiking process as a function of time in a population. The diffusion process is conditioned on the fixation of the beneficial allele to investigate its effect on the hitchhiking process. When an allele is introduced by a mutation in a population, it is most likely to be lost, just by chance, by random genetic drift, even when it is beneficial. Therefore, conditioning the process of genetic hitchhiking on the fixation of the beneficial allele has significant effects on the dynamics of alleles. Previous mathematical studies formulated the effect of the conditioning on the dynamics of the beneficial allele (Griffiths 2003, Etheridge et al. 2006, Pfaffelhuber et al. 2006, Eriksson et al. 2008). I formulate the effect of the conditioning on the dynamics of alleles at the neutral locus in a diffusion approximation of the hitchhiking model. Stochastic aspects of genetic hitchhiking such as 'reverse hitchhiking' are quantitatively investigated.

In summary, the results in this dissertation show the importance of interacting effects of various evolutionary forces on patterns of polymorphisms in natural populations. Signatures of positive selection described by currently available methods are shown to be significantly affected by confounding factors such as purifying selection and population demography. Modeling the interaction of various evolutionary forces and examining the compatibility of the theoretical

predictions with actual empirical data are essential for our understanding of the mechanisms of evolution.

CHAPTER 2: EMPIRICAL DETERMINATION OF THE DEPENDENCE OF POPULATION DIFFERENTIATION MEASURES ON MINOR ALLELE FREQUENCIES AT HUMAN SNP SITES

Abstract

Wright's F_{ST} and other similar statistics are often used by evolutionary biologists to quantify the degree of population differentiation at genetic loci. Their applications include identification of potential targets of local adaptation and estimation of the amount of gene flow between populations. Therefore, understanding their biological and mathematical properties is important for correct interpretation of empirical data. In this research, I determine the dependence of population differentiation measures on minor allele frequencies at SNP sites, using genomic data of human SNPs in protein coding regions. In particular, it is shown that the maximum of F_{ST} at a site is a monotonically increasing function of the minor allele frequency. Because of this property, F_{ST} at sites with low minor allele frequencies are inevitably limited to low values. As a result, purifying selection at sites decreases F_{ST} values because it decreases minor allele frequencies. This is shown empirically as a positive correlation between F_{ST} and site-specific long-term evolutionary rates measured from multi-species alignments. Furthermore, it is shown that F_{ST} can be highly overestimated using data with ascertainment biases, where polymorphic sites are identified in a sample of a few sequences and then allele frequencies at the identified sites are examined in a larger sample, due to this property. The finding in this research shows that

we need to correct the difference in minor allele frequencies, when we compare F_{ST} at different sites to evaluate the degree of genetic differentiation between populations.

Introduction

Wright's F_{ST} **(**Wright 1951) and other similar statistics are widely used by evolutionary biologists in order to quantify population structures and estimate migration rates between populations. F_{ST} measures the degree of genetic differentiation between populations by showing the proportion of between-population components of genetic variation (heterozygosity) in the total population. In recent years, the artificial dependence of F_{ST} on within-population heterozygosity was criticized and led to the formation of new measures of population differentiation, which are supposedly independent of within-population heterozygosity (Hedrick 2005; Jost 2007, 2008). Those studies were motivated by empirical observations of unreasonably low values of F_{ST} at loci with high mutation rates such as microsatellites. The new statistics removed the unreasonably low upper limit of F_{ST} at high-diversity loci (Jost 2008, Meirmans and Hedrick 2011). The issue of the dependence of F_{ST} on allele frequencies at high-diversity loci has been actively discussed but considered to cause few problems at low-diversity loci, including single nucleotide polymorphism (SNP) sites (Meirmans and Hedrick 2011, Whitlock 2011). However, in this research, I determine that F_{ST} is also highly dependent on allele frequencies at low-diversity loci but the nature of the dependence is different from that at high-diversity loci.

This problem was found when I examined the relationship between F_{ST} and evolutionary rates at SNP sites in protein-coding regions in the human genome. The nature and biological significance of the problem are discussed in this research.

Materials and Methods

Analysis of SNPs in protein-coding regions

The relationship between F_{ST} and evolutionary rates is investigated at SNP sites in protein-coding regions. I analyzed 15,432 nonsynonymous SNPs (nSNPs) from 6,494 genes and 18,001 synonymous SNPs (sSNPs) from 7,549 genes in the data set published by Lohmueller et al. (2008). This data set contains resequencing data of allele frequencies in African American (AA) and European American (EA) populations. The average number of chromosomes resequenced in AA and EA populations at the sites is 28 and 37, respectively. Each SNP site was classified into a CpG or non-CpG site based on the context in the dinucleotides in the reference human genome sequence (hg 19). If a site is C followed by G or G preceded by C, it is classified into a CpG site. Otherwise, the site is classified into a non-CpG site. At each SNP site, F_{ST} is calculated from sample allele frequencies in AA and EA populations. The absolute rate of evolution (r) is also estimated at each SNP site by mapping sequence differences among 36 mammalian species onto their evolutionary tree (Figure 1A) and dividing the inferred number of nucleotide substitutions by the total time summed over all tree branches (see Kumar et al. 2009). The alignment of nucleotide sequences of 36

mammalian species was made for this purpose, following the procedures outlined in Kumar et al. (2009).

Calculation of F_{ST} at each SNP site

I use the formula by Nei for defining F_{ST} at each SNP site. F_{ST} between two demes at a bi-allelic locus is given by

$$F_{ST} = \frac{H_T - H_S}{H_T} = 1 - \frac{H_S}{H_T}, \quad (2.1)$$

where $H_T = 2 \cdot \frac{p_1+p_2}{2} \cdot \left(1 - \frac{p_1+p_2}{2}\right)$, $H_S = \frac{2 \cdot p_1 \cdot (1-p_1) + 2 \cdot p_2 \cdot (1-p_2)}{2}$, and p_1 and p_2 are frequencies of an allele in demes 1 and 2, respectively (Nei, 1977). H_T and H_S are the heterozygosity in the total population and the average heterozygosity across subpopulations (demes), respectively. F_{ST} is estimated from the sample by the following equation:

$$\hat{F}_{ST} = \frac{\hat{H}_T - \hat{H}_S}{\hat{H}_T}, \quad (2.2)$$

where $\hat{H}_T = H_T + \frac{\hat{H}_S}{4\tilde{n}}$ and $\hat{H}_S = \frac{2\tilde{n}}{2\tilde{n}-1} H_S$. $\tilde{n}$ is the harmonic mean of the sample size across populations (Nei and Chesser, 1983).

Sliding window analysis of SNPs

To show the relationship among estimates of population differentiation, evolutionary rates (*r*) and MAF, SNP sites are binned according to *r*/MAF such that each bin contains 1,000 SNP sites except for *r* = 0 and the bin with the

highest r/MAF. Then, data on sites with negative values of population differentiation estimates are removed before the subsequent analyses.

Examination of the effect of ascertainment biases

I examined the effect of ascertainment biases on F_{ST}, using the HapMap phase I data (Altshuler et al. 2005). The effect of ascertainment biases is examined by comparing minor allele frequencies (MAF) and F_{ST} at SNP sites in ENCODE regions with those at SNP sites outside ENCODE regions. F_{ST} between CEU and YRI populations at each site on chromosome 7 is calculated from their allele frequencies. There are three 500kb ENCODE regions, ENm010 (26,699,793-27,199,792, NCBI build 34 coordinates), ENm013 (89,395,718-89,895,717), and ENm014 (126,135,436-126,632,577), on chromosome 7 in the data. Those sites, where allele frequencies are available in both of the two populations and polymorphism is observed, are used for the subsequent analysis.

Results

Attributes of the SNPs in the Lohmueller et al. data set

Figure 2.1B shows the distribution of SNPs among genes. A vast majority of genes contain single or a few SNPs but some genes contain several SNPs. The distribution of evolutionary rates at SNP sites is shown in Figure 2.1C. Overall, evolutionary rates at nSNP sites are lower than those at sSNP sites ($P < 10^{-15}$ in t-test). The distribution at nSNP sites is very different from that at sSNP sites: The distribution is highly skewed to the right at nSNP sites, whereas it is much less

skewed at sSNP sites. Because of the difference, nSNP and sSNP sites are analyzed separately in the subsequent results.

The effect of purifying selection on F_{ST} at SNP sites

Assuming that the long-term evolutionary rate (r) is mainly determined by functional constraint, the evolutionary rate at a site represents the strength of purifying selection specific to the site. Therefore, I investigate the relationship between the strength of purifying selection and F_{ST} at SNP sites, by using evolutionary rates as proxies for the strength of purifying selection at sites. Figure 2.2 shows the relationship between r and F_{ST} at nSNP sites. There is a strong positive correlation between them ($P < 10^{-25}$ in Pearson correlation analysis). The strong correlation remains when nSNPs at CpG and non-CpG sites are analyzed separately (Table 2.1). Therefore, the results indicate that stronger purifying selection leads to less population differentiation between populations at SNP sites (see also, Barreiro et al. 2008). There is also a positive, but weaker correlation between r and F_{ST} at sSNP sites (Figure 2.3, $P < 10^{-2}$). I found that these positive correlations between r and F_{ST} at SNP sites result from dependence of F_{ST} on minor allele frequencies (MAF). MAF in this study is defined to be the frequency of the allele rarer in the total population that consists of all of the populations under examination. A strong positive correlation between MAF and F_{ST} is observed at nSNP sites (Figure 2.4, correlation coefficient = 0.49, $P < 10^{-25}$). Similar patterns are seen when other widely used measures of population differentiation such as Weir and Cockerham's θ (Weir and Cockerham 1984) and

Jost's *D* (Jost 2008), are used (Figure 2.5). Furthermore, the positive correlation between *r* and F_{ST} at nSNP sites disappears when the effect of minor allele frequencies is controlled in partial correlation analysis ($P = 0.74$). These results indicate that stronger purifying selection reduces F_{ST} at SNP sites, which results in the observation of the positive correlation between *r* and F_{ST}, just because it reduces frequencies of mutant alleles (Subramanian and Kumar 2006).

The dependence of F_{ST} on MAF exists regardless of the presence of purifying selection because it is observed even in simulated data of polymorphic sites under neutral evolution (results shown in Chapter 3). In addition, the nonsignificant result in the correlation analysis between *r* and F_{ST} at CpG sSNP sites is explained by the fact that the positive correlation between *r* and MAF does not exist in this class of SNP sites (Figure 2.6).

The effect of ascertainment biases on F_{ST} at SNP sites

I also examined the effect of ascertainment biases on F_{ST}, as widely used human SNP data sets suffer from ascertainment biases (Clark et al. 2005). When polymorphic sites are discovered in samples of only a few sequences and then size of samples is expanded, the average MAF in such samples becomes much higher compared to that in bias-free samples, because sites with higher MAF are more likely to be identified in small samples. Therefore, due to the dependence of F_{ST} on MAF, different values of F_{ST} can be seen from different data sets of SNPs with different degree of ascertainment biases. Using allele frequencies in CEU and YRI populations in HapMap phase I data (Altshuler et al. 2005), I compared

estimates of F_{ST} at SNP sites in ENCODE regions with those of F_{ST} at SNP sites outside ENCODE regions on human chromosome 7 (see Materials and Methods). Table 2.2 shows that the average MAF at SNP sites in ENCODE regions, which suffer from few ascertainment biases, is much lower than that at SNP sites outside ENCODE regions. As a result, the average F_{ST} in ENCODE regions is much higher than that outside ENCODE regions (0.071 compared to 0.042). If researchers infer historical amounts of gene flow from the average F_{ST} values, for example, by using the commonly used equation $4Nm = (1 - F_{ST})/F_{ST}$, the ascertainment bias observed here leads to an underestimation of the migration rate by approximately 43 %.

The nature of the dependence of population differentiation measures on MAF

I further investigated the nature of the dependence of F_{ST} and other similar statistics on MAF. Consider F_{ST} between two demes at a bi-allelic locus, which is defined by equation 2.1 above. Let M be the MAF, $M = \frac{p_1+p_2}{2}$, where p_1 and p_2 are frequencies in demes 1 and 2, respectively, of the allele, which is minor in the total population. Then, it can be shown that the maximum F_{ST} with given M (, which is reached when $p_1 = 0$ and $p_2 = 2M$ or $p_1 = 2M$ and $p_2 = 0$,) is

$$F_{ST(\max)} = \frac{M}{1-M}, \tag{2.3}$$

which is a monotonically increasing function of M. Therefore, only a small value of F_{ST} is possible at a SNP site with low MAF, while scientists generally expect that F_{ST} takes a value between zero and one. When MAF decreases, H_T and H_S in

equation 2.1 above both decrease. However, H_T decreases more than H_S with a given decrease in MAF, because the former is a second order function of MAF, while the latter is approximately a first order function of MAF when the mutation rate is low. This explains why F_{ST} decreases when MAF decreases at SNP sites.

It is possible to make the maximum value of F_{ST} one irrespective of MAF by developing a new measure of population differentiation, $F'_{ST} = F_{ST}/F_{ST(max)}$. However, when it is applied to nSNP sites in Lohmueller et al. data set, a negative correlation is observed between MAF and F'_{ST} (Figure 2.7). A similar negative correlation is seen when $|p_1 - p_2|/(p_1 + p_2)$ is used as a measure of population differentiation. These may be explained by considering the relationship between the age of polymorphism and actual degree of population differentiation at the locus. When a derived allele is young and has a low frequency, it is likely to be confined in the deme where it originated. Therefore, the dependence of population differentiation measures on MAF seems inevitable even when they are normalized such that their maxima are independent from MAF.

Discussion

Wright's F_{ST} is widely used by scientists for quantifying population structures with increasing data of genomic sequences. Its applications include estimation of the amounts of historical gene flow and identification of potential targets of local adaptation. Therefore, it is important to understand its properties and how the degree of genetic differentiation between populations is quantified by F_{ST}. This research determined the dependence of population differentiation measures on

MAF and demonstrated why it is problematic when they are applied to empirical data. While recent debates on properties on F_{ST} focused on decreasing F_{ST} with increasing heterozygosity at loci with high mutation rates (Hedrick 2005; Jost 2007, 2008; Meirmans and Hedrick 2011; Whitlock 2011), this research finds increasing F_{ST} with increasing heterozygosity (MAF) at loci with low mutation rates. This artificial dependence of F_{ST} on MAF identified here explains the negative correlation between purifying selection intensity and F_{ST} at human SNP sites. Stronger purifying selection, which is reflected as lower evolutionary rates, leads to lower MAF and therefore lower F_{ST} at human SNP sites. This dependence of F_{ST} on MAF should now be considered when interpreting and comparing results from population genomic studies. For example, it is important when inferring the difference in female and male migration rates from SNP data. F_{ST} at sites on the Y chromosome and that at sites in the mitochondrial genome are compared to detect the difference (Seielstad et al. 1998). It is necessary to compare F_{ST} at sites with similar minor allele frequencies to correctly detect the difference. It is also important when inferring the difference in genetic differentiation among populations between different sets of populations. For example, African populations are known to have higher MAF across populations than non-African populations (Tishkoff and Kidd 2004). This may lead to incorrect inference of greater degree of genetic differentiation among African populations compared to that among non-African populations. Recently, a number of studies examined the genomic distribution of F_{ST} at SNP sites in order to identify positions under local adaptation (Akey et al. 2002, Izagirre et al. 2006,

Lohmueller et al. 2006, Norton et al. 2007, Myles et al. 2008, Pickrell et al. 2009). These studies are based on the idea that positions under local adaptation show greater degree of population differentiation compared to that under migration-genetic drift equilibrium (Lewontin and Krakauer 1973). Therefore, an outlier of F_{ST} is considered as a candidate of a position under local adaptation. However, the results shown here strongly suggest that false positive or negative detection of positions under local adaptation is likely if the dependence of F_{ST} on MAF is not taken into account. This is particularly true when the approach is taken in organisms with low migration rates, because the artifact becomes more severe when F_{ST} values are limited by low MAF despite the actual high differentiation between populations. The correct identification of positions under local adaptation is approached by pooling SNPs with similar MAFs and making groups of distributions.

A. Raw data

Data class	No. of SNPs	Coefficient	*p*-value
nSNPs at non-CpG sites	11,064	0.10	$P < 10^{-23}$
nSNPs at CpG sites	4,368	0.07	$P < 10^{-5}$
sSNPs at non-CpG sites	11,625	0.08	$P < 10^{-17}$
sSNPs at CpG sites	6,376	-0.02	$P = 0.1$

B. Sliding window

Data class	Coefficient	*p*-value
nSNPs at non-CpG sites	0.62	$P < 10^{-2}$
nSNPs at CpG sites	0.54	$P = 0.058$
sSNPs at non-CpG sites	0.93	$P < 10^{-8}$
sSNPs at CpG sites	-0.26	$P = 0.28$

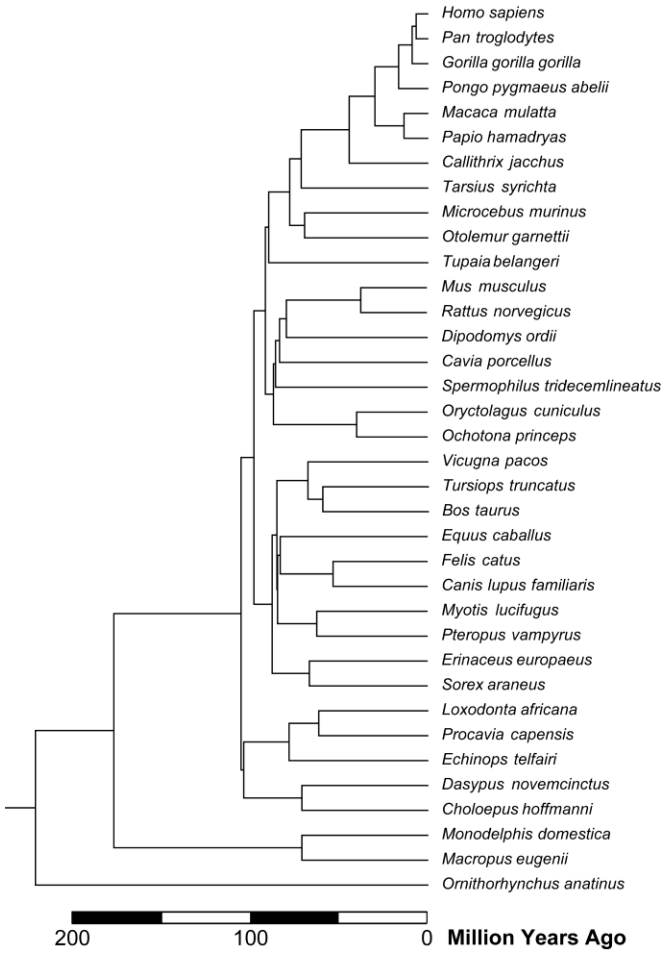
Table 2.1. Pearson correlation analysis of raw data (A) and sliding window analysis (B) between evolutionary rates (r) and F_{ST} at SNP sites in the Lohmueller et al. (2008) data set. Data on sites with negative values of F_{ST} are removed in the sliding window analysis.

Region	No. of SNPs	MAF	F_{ST}
ENCODE	3,555	0.131 ± 0.0044	0.042 ± 0.0020
Outside ENCODE	150,539	0.200 ± 0.0008	0.071 ± 0.0004

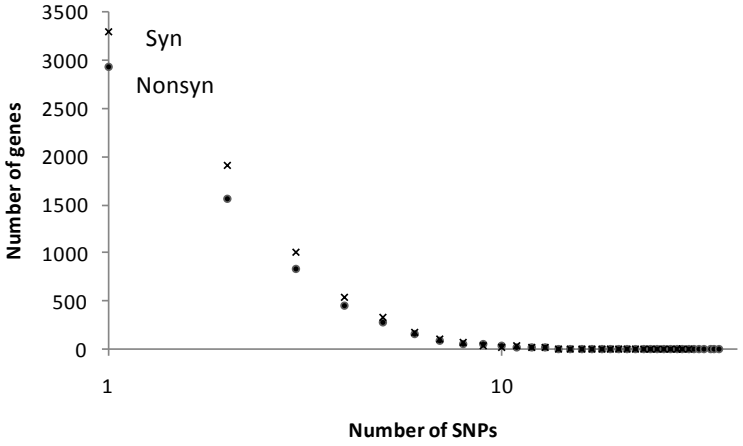
Note. Allele frequencies in CEU and YRI populations at SNP sites on chromosome 7 in HapMap phase I data are used. Average $\pm$ two standard error are shown for MAF and F_{ST}.

Table 2.2. Comparison of minor allele frequency (MAF) across populations and F_{ST} at SNP sites in ENCODE regions with those at SNP sites outside ENCODE regions.

A. Evolutionary time-tree of 36 mammalian species used for estimating evolutionary rates (Kumar et al. 2009).



B. The frequency distribution of the number of nonsynonymous (Nonsyn) and synonymous (Syn) SNPs in human protein-coding genes.



C. The frequency distribution of nucleotide evolutionary rates at nonsynonymous (Nonsyn) and synonymous (Syn) SNP sites.

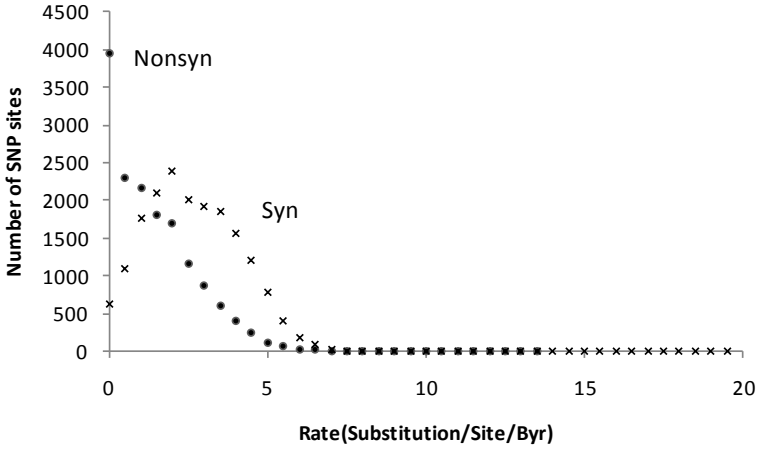
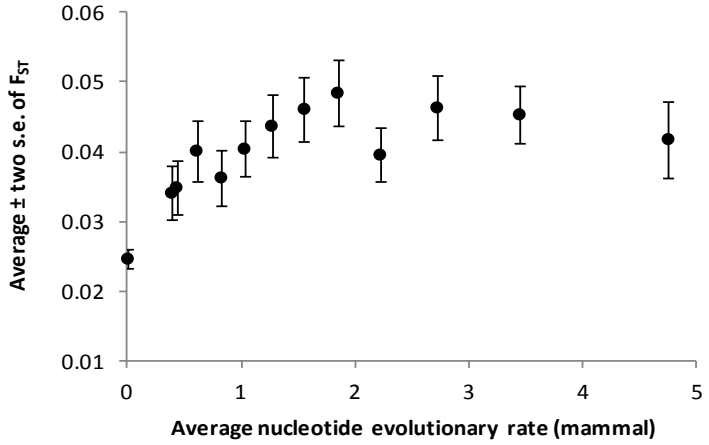
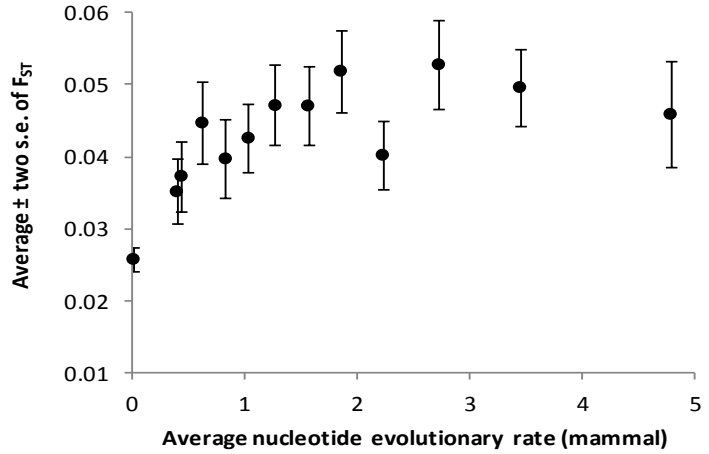


Figure 2.1. Properties of the data analyzed. The scale on the x-axis in B is a logarithm of base ten. The average rates at nonsynonymous and synonymous SNP sites in C are 1.20 and 2.39 substitution/site/Byr, respectively.

A. Overall (nSNP)



B. Non-CpG (nSNP)



C. CpG (nSNP)

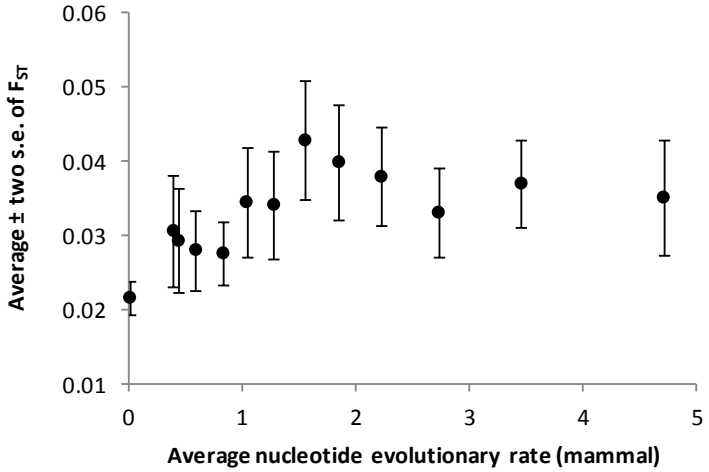
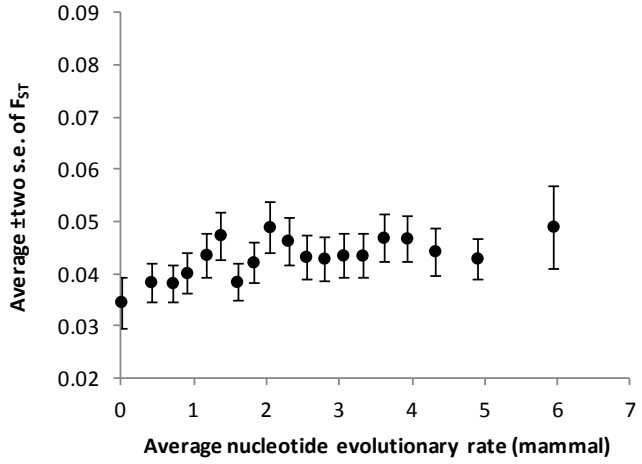
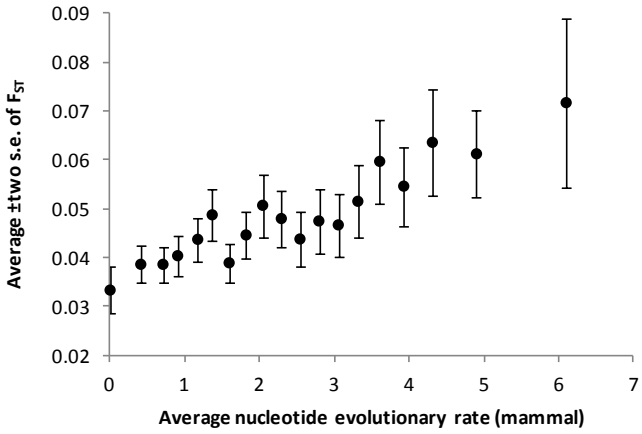


Figure 2.2. The relationship between average evolutionary rates (r) and F_{ST} between African American (AA) and European American (EA) populations at nonsynonymous SNP (nSNP) sites in the Lohmueller et al. (2008) data set. SNP sites were binned according to r. Each bin contains 1,000 SNPs (except for $r = 0$ and the bin with the highest r) and its average F_{ST} (and ± 2 standard error) is plotted. Data on sites with negative values of F_{ST} are removed.

A. Overall (sSNP)



B. Non-CpG (sSNP)



C. CpG (sSNP)

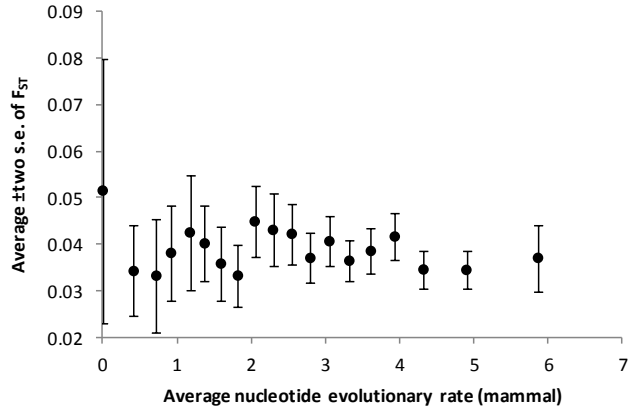
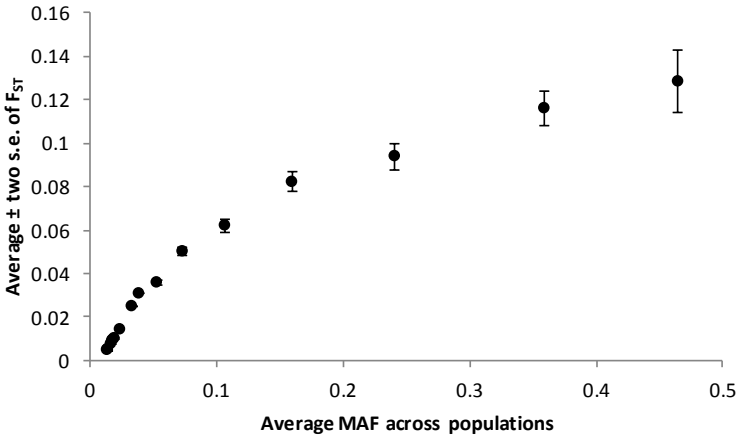
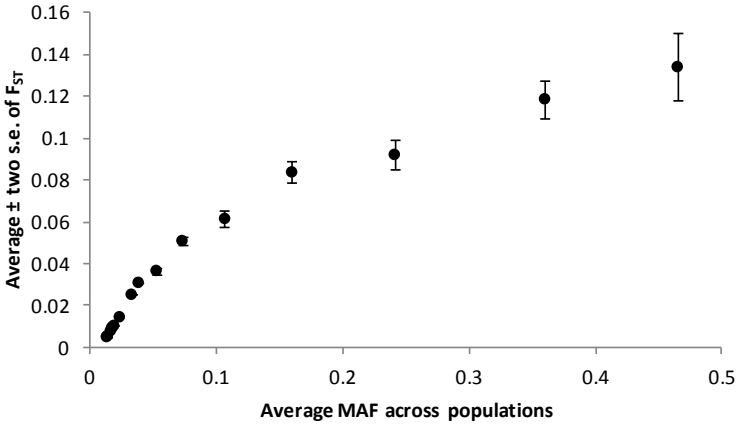


Figure 2.3. The relationship between average evolutionary rates (r) and F_{ST} between African American (AA) and European American (EA) populations at synonymous SNP (sSNP) sites in the Lohmueller et al. (2008) data set. SNP sites were binned according to r. Each bin contains 1,000 SNPs (except for $r = 0$ and the bin with the highest r) and its average F_{ST} (and ± 2 standard error) is plotted. Data on sites with negative values of F_{ST} are removed.

A. Overall (nSNP)



B. Non-CpG (nSNP)



C. CpG (nSNP)

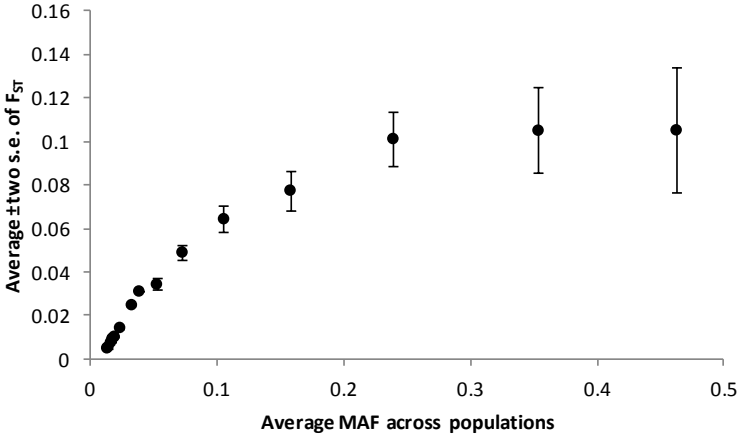
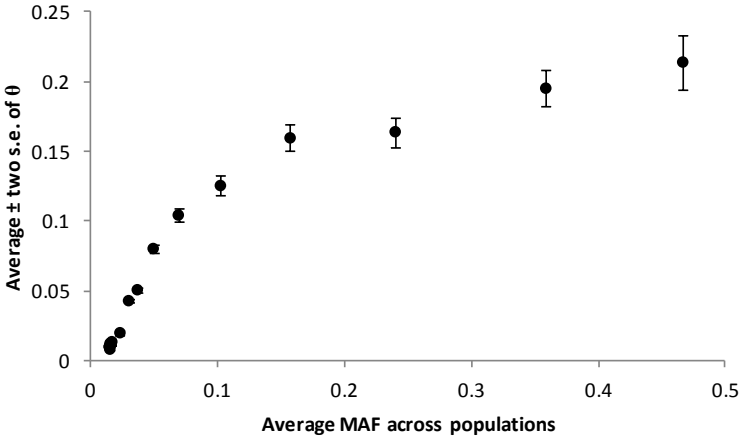


Figure 2.4. The relationship between average minor allele frequency (MAF) across populations and F_{ST} between African American (AA) and European American (EA) populations at nonsynonymous SNP (nSNP) sites in the Lohmueller et al. (2008) data set. SNP sites were binned according to MAF. Each bin contains 1,000 SNPs (except for the bin with the highest MAF) and its average F_{ST} (and ± 2 standard error) is plotted. Data on sites with negative values of F_{ST} are removed.

A. Weir and Cockerham's θ



B. Jost's *D*

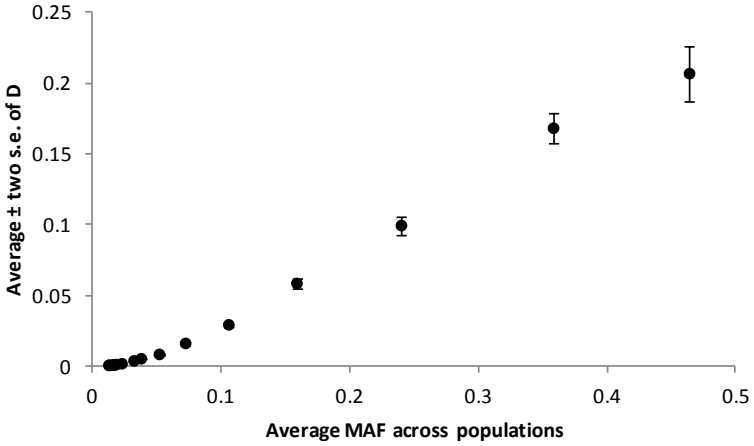


Figure 2.5. The relationship between average minor allele frequency (MAF) across populations and Weir and Cockerham's θ (A) or Jost's *D* (B) between African American (AA) and European American (EA) populations at nonsynonymous SNP (nSNP) sites in the Lohmueller et al. (2008) data set. Each bin contains 1,000 SNPs (except for the bin with the highest MAF) and its average MAF (and ±2 standard error) is plotted. Data on sites with negative values of theta or *D* are removed.

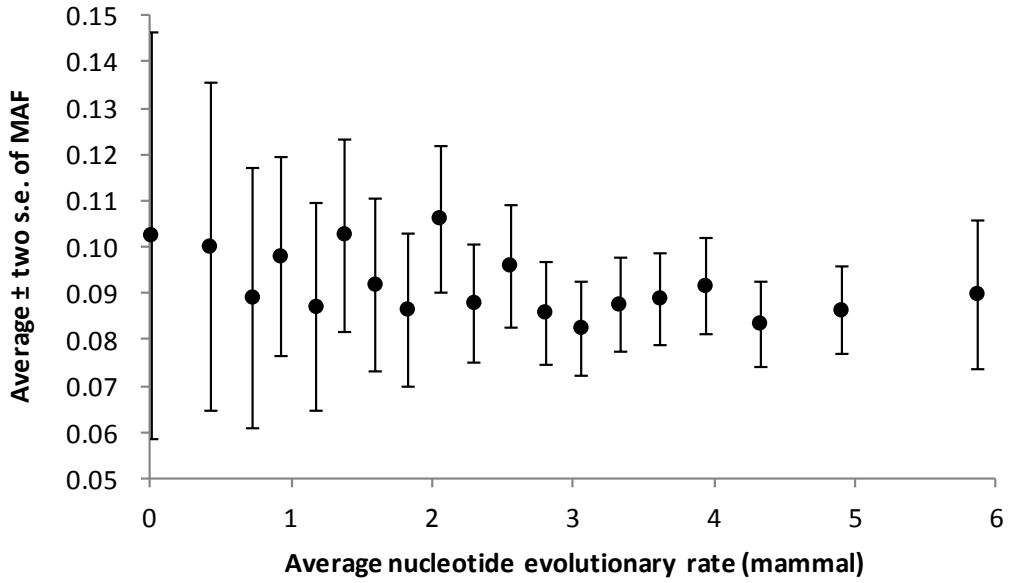


Figure 2.6. The relationship between average evolutionary rates (*r*) and minor allele frequency (MAF) across populations at CpG synonymous SNP (sSNP) sites in the Lohmueller et al. data set (2008). SNP sites were binned according to *r*. Each bin contains 1,000 SNPs (except for $r = 0$ and the bin with the highest *r*) and its average MAF (and ± 2 standard error) is plotted. Data on sites with negative values of F_{ST} are removed.

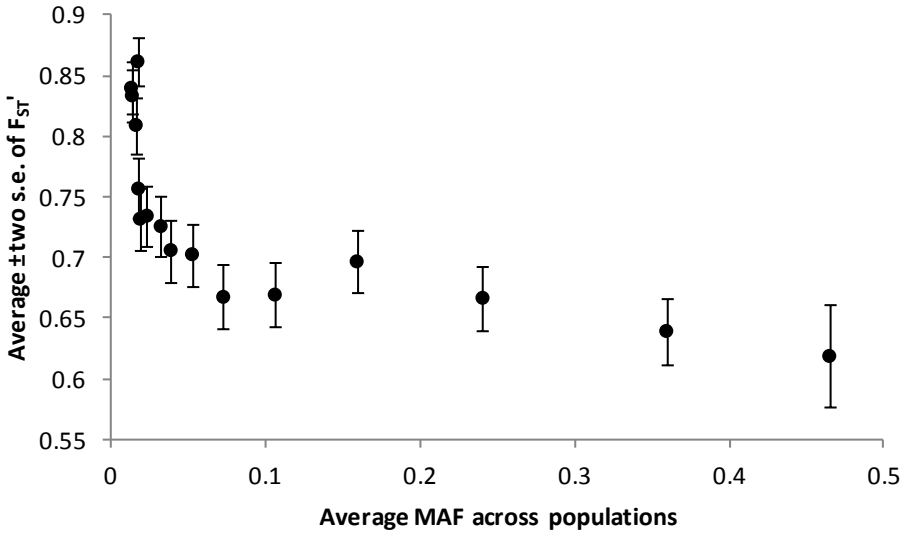


Figure 2.7. The relationship between average minor allele frequency (MAF) across populations and F_{ST}' between African American (AA) and European American (EA) populations at nonsynonymous SNP (nSNP) sites in the Lohmueller et al. (2008) data set. SNP sites were binned according to MAF. Each bin contains 1,000 SNPs (except for the bin with the highest MAF) and its average F_{ST}' (and ± 2 standard error) is plotted.

CHAPTER 3: SIMULATION OF GENETIC DIFFERENTIATION BETWEEN POPULATIONS UNDER PURIFYING SELECTION

Abstract

Quantifying the degree of genetic differentiation between populations is important. Wright's F_{ST} and similar statistics are often used by evolutionary biologists for this purpose. However, previous studies showed that F_{ST} is dependent on allele frequencies and may not measure the actual differentiation between populations. In this study, I explore the effects of various evolutionary forces on allele frequencies and F_{ST} at bi-allelic loci by computer simulations. The effects of neutral and selective forces on F_{ST} are examined in subdivided population and population split models. The results show that purifying selection greatly affects F_{ST} by modulating minor allele frequencies across populations in both models. Overall, stronger purifying selection lowers minor allele frequencies and $F_{ST.}$ This results in severe underestimation of migration rates and time since the population split under the presence of widespread purifying selection, when they are estimated by F_{ST}, assuming neutral equilibrium. The difference in minor allele frequencies must be considered when evaluating the difference in genetic differentiation between pairs of populations by F_{ST}.

Introduction

Quantifying the amount of genetic differentiation between populations is important for several reasons. First, it is important for understanding the

mechanisms of evolution. Scientists measure genetic differentiation among populations to understand how much genetic variation in a species is due to variation among populations. Genetic differentiation between geographically distant populations plays an important role in major evolutionary processes, including speciation (Beaumont 2005). Second, it is important for practical purposes. For example, measuring genetic difference between populations is important for understanding the genetic cause of the difference in susceptibility to a disease between human populations (e.g., Lohmueller et al. 2006, Myles et al. 2008, Amato et al. 2009). In conservation biology, understanding the genetic differences between populations is important when making strategies for maintaining genetic variation of an endangered species (e.g., Palumbi 2003, Pearse and Crandall 2004, Wang 2004, Charruau et al. 2011). Wright's F_{ST} (Wright 1951) and its analogues are widely used by evolutionary biologists for measuring genetic differentiation between populations (reviewed in Holsinger and Weir 2009). The amount of genetic variation found in the total population, which is made by combining all of the populations under examination, is partitioned into between-populations and within-population components in these statistics. Then, F_{ST} represents the proportion of the between-populations component of the genetic variation in the total population. F_{ST} is a convenient measure because it quantifies the degree of genetic differentiation at various kinds of loci in various organisms. However, recent studies have demonstrated that values of F_{ST} are dependent on allele frequencies and suggest that F_{ST} may not be a good measure of the actual degree of genetic differentiation between populations. Specifically,

F_{ST} takes unreasonably low values at loci with high mutation rates such as microsatellites, when the actual amount of genetic differentiation between populations is high (Charlesworth 1998, Hedrick 1999, Long and Kittles 2003, Hedrick 2005, Jost 2008). This problem of F_{ST} at loci with high mutation rates occurs because the maximum value of F_{ST} inevitably becomes low when within-populations genetic variation is high. However, the nature of the dependence of F_{ST} on allele frequencies is different at loci with low mutation rates including SNP sites. In the previous chapter, I empirically demonstrated that F_{ST} at human SNP sites is dependent on minor allele frequencies (MAF) across populations. The maximum value of F_{ST} between populations is a monotonically increasing function of MAF at bi-allelic loci and F_{ST} at human SNP sites takes higher values when minor allele frequencies are higher.

In this study, I further investigate the nature of the dependence of F_{ST} on minor allele frequencies at bi-allelic loci. In-depth simulation of genetic differentiation between populations is conducted to examine how the change in evolutionary forces influences values of F_{ST}. This is important for understanding how the dependence on minor allele frequencies at bi-allelic loci influences F_{ST} at different loci in different organisms. Because evolutionary biologists often infer historical amounts of gene flow and time since population separation using F_{ST} (e.g., Weir and Hill 2002, Ramachandran et al. 2005, Cox et al. 2008), two different models of demography, the population subdivision and population split models, are studied. In particular, I focused on the effect of purifying selection on

F_{ST} under these demographic models. The results show that purifying selection significantly affects F_{ST} in both models by modulating minor allele frequencies.

Materials and Methods

Purifying selection in a subdivided population model

The population in the model is a subdivided population of a diploid organism that consists of two demes of effective size N_1 and N_2. There are two alleles, A_1 and A_2, at the locus under purifying selection. The derived allele is assumed to be negatively selected with selection coefficient t and dominance coefficient h in both of the demes. Mutation occurs at rate u per generation from allele A_1 to allele A_2, and vice versa, in both demes. Migration occurs at rate m per generation between the demes. Reproduction occurs according to the Wright-Fisher model in each of the demes.

Purifying selection in a population split model

An ancestral population of diploid effective size N_a is split into two demes of diploid effective size N_1 and N_2 in the model. There are two alleles, A_1 and A_2, at the locus under purifying selection. After the split, the frequencies of an allele in the two daughter demes are determined by binomial sampling with replacement with probability of sampling equal to the allele frequency in the ancestral population. The deleterious allele is assumed to be negatively selected with selection coefficient t and dominance coefficient h in every population. Mutation occurs at rate u per generation from allele A_1 to allele A_2, and vice versa, in every

population. No migration occurs between the demes following the population split. Reproduction occurs according to the Wright-Fisher model in every population.

The simulation of purifying selection in a subdivided population model

The above model of purifying selection in a subdivided population is simulated by a forward-in-time frequency-based simulation. The initial frequency of the deleterious allele is sampled from a beta distribution with parameters $a = b = 4N_i u$ in deme i. This is the stationary distribution of allele frequencies in the diffusion limit of the neutral Wright-Fisher model in a panmictic population of constant size (Otto and Day 2007). When the deleterious alleles are weakly selected against, the stationary distribution of allele frequencies is expected to be similar to the neutral equilibrium. In contrast, when selection is strong, the correct stationary distribution will be quickly reached even if the initial frequencies are sampled from the neutral equilibrium. In either case, the equilibrium process can be studied by taking a sufficiently long burn-in period at the onset of each simulation. Ten pairs of beta-distributed allele frequencies are used and ten runs of simulation are conducted for each set of parameter values unless stated otherwise. Each generation consists of deterministic changes in allele frequencies by mutation, migration, and selection, followed by the random sampling of surviving individuals using a binomial number generator.

The simulation of purifying selection in a population split model

Purifying selection in the population split model explained above is simulated by a forward-in-time frequency-based simulation. The initial frequency of the deleterious allele in the ancestral population is specified by beta distribution with parameters a = b = $4N_a u$. After 50,000 generations, the ancestral population is split into two demes of diploid effective size N_1 and N_2 by random sampling that uses a random binomial number generator. Time since the population split is denoted as T. Each generation consists of deterministic changes in allele frequencies by mutation and selection, followed by the random change by the step of random sampling that uses a random binomial number generator.

Calculation of F_{ST}

F_{ST} at a bi-allelic locus between populations is defined as described in Chapter 2 (page 11). F_{ST} is estimated from a sample of size 30 per deme according to equation 2.2. A sample of alleles is obtained by random sampling that uses a binomial random number generator. In the subdivided population model, F_{ST} is estimated from a sample every 100 generations, if there is polymorphism in the sample, after the burn-in period of $8N_{\text{larger}}$ generations, where N_{larger} is the diploid effective size of the larger deme. In the population split model, F_{ST} is estimated from a sample after T generations from the split, if there is polymorphism in the sample. Unless stated otherwise, F_{ST} is estimated from a sample in the following results.

Calculation of effective sample size of time-series data in the simulation of purifying selection in a subdivided population

Because the time-series data of F_{ST} and MAF are highly correlated, depending on parameter values, in the model of purifying selection in a subdivided population, standard errors of F_{ST} and MAF are calculated by taking the square root of the variance divided by the effective sample size. Effective sample size of time-series data of F_{ST} and MAF is calculated as the sum of that in each of the chains by the function 'effectiveSize' in the R package 'coda'.

Results

The relationship between MAF and F_{ST} at a locus under neutral evolution

Figure 3.1 shows the relationship between MAF and F_{ST} between demes at a locus under neutral evolution in the subdivided population model. There is a strong correlation between them (correlation coefficient = 0.29, $P < 10^{-15}$ when $2Nm = 2$ in Pearson correlation analysis). This correlation becomes stronger when the migration rate is lower (correlation coefficient = 0.60, $P < 10^{-15}$ when $2Nm = 0.4$). Figure 3.2 shows the effect of the mutation rate on F_{ST} between demes at a locus under neutral evolution in the subdivided population model. F_{ST} first increases and then decreases with the increase in the mutation rate. MAF monotonically increases with the increase in the mutation rate. Comparing the figures for F_{ST} and MAF as functions of the mutation rate, there is a positive correlation between them when MAF is less than 0.25. On the other hand, there is a negative

correlation between them when MAF is greater than 0.25. Figure 3.2 also shows the effect of ascertainment biases on MAF and F_{ST}. Ascertainment biases exist in the sample because F_{ST} is calculated if there is polymorphism at the locus (see Materials and Methods). The minimum MAF in the sample is 1/60, whereas it is 1/(4N) in the total population. MAF and F_{ST} estimated from sample allele frequencies are higher than those calculated from population allele frequencies. That is, ascertainment biases increase MAF and F_{ST}.

The effect of the migration rate on F_{ST} at a locus under purifying selection in a subdivided population model

Figure 3.3 shows the effect of the migration rate on F_{ST} between demes at a locus under purifying selection in the subdivided population model. As empirically shown in Chapter 2, there is a negative correlation between the strength of purifying selection and F_{ST} (correlation coefficient = -0.18, $P < 10^{-15}$ when $2Nm$ = 2). This effect of purifying selection on F_{ST} becomes larger when the migration rate is smaller.

The effect of the dominance coefficient on F_{ST} at a locus under purifying selection in a subdivided population model

Figure 3.4 shows the effect of the dominance coefficient on F_{ST} between demes at a locus under purifying selection in the subdivided population model. When the dominance of the deleterious allele increases, MAF decreases. As a consequence,

there is a negative correlation between the dominance coefficient and F_{ST} (correlation coefficient = -0.09, $P < 10^{-15}$ when $2Nt = 4$ and $2Nm = 2$).

The effect of the difference in deme size on F_{ST} at a locus under purifying selection in a subdivided population model

Figure 3.5 shows the effect of purifying selection on F_{ST} between demes when the deme sizes are unequal in the subdivided population model. When one of the demes is one hundred times smaller than the other deme, purifying selection increases F_{ST} compared to that under neutral evolution. There is a strong correlation between MAF and F_{ST} in these cases too. When MAF increases, F_{ST} also increases.

The effect of purifying selection on F_{ST} in a population split model

Figure 3.6 shows the effect of purifying selection on F_{ST} between demes in the population split model. There is a negative correlation between the strength of purifying selection and F_{ST} in this model too (correlation coefficient = -0.25, $P < 10^{-15}$). Again, there is a strong correlation between MAF and F_{ST}. Notice that no migration between demes after the population split is assumed in this model. As a result, the effect of the purifying selection intensity on F_{ST} is high.

Discussion

Given the dependence of F_{ST} on allele frequencies, it is important to study how various evolutionary forces affect F_{ST} by modulating allele frequencies. F_{ST} was originally defined at bi-allelic loci (Wright 1951) and is considered to be suitable

when it is measured at SNP sites (Meirmans and Hedrick 2011, Whitlock 2011). The amount of genome-wide data of SNPs is rapidly increasing in various organisms and the use of F_{ST} at SNP sites is expected to become more important (Helyar et al. 2011). In the previous study, I showed that the maximum of F_{ST} between populations at bi-allelic loci is an increasing function of MAF and F_{ST} is limited to low values when MAF is low. In this study, effects of neutral and selective forces on MAF and F_{ST} at bi-allelic loci are investigated.

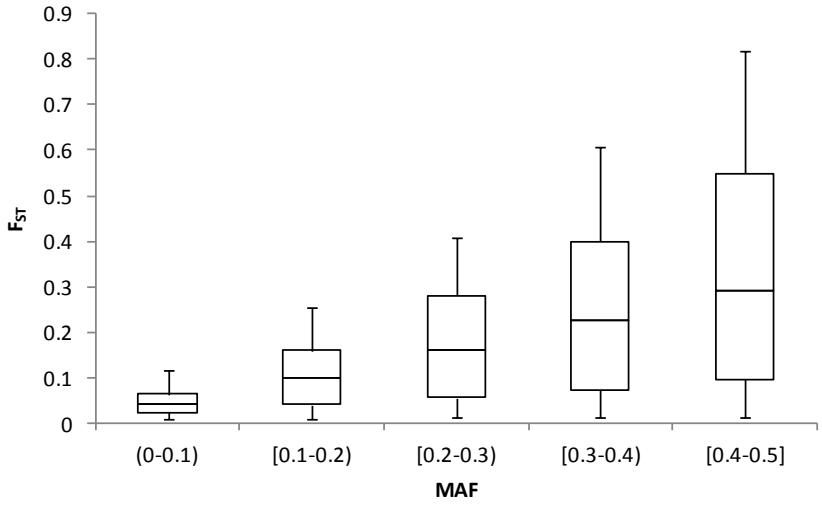
The dependence of F_{ST} on MAF becomes severe when the migration rate is low. This is because F_{ST} becomes inevitably low when MAF is low despite the actual high differentiation between populations. The 95 % confidence interval of F_{ST} under neutral equilibrium is highly dependent on MAF when the migration rate is low (Figure 3.1B). Therefore, the difference in MAF among sites must be considered when potential targets of local adaptation are identified as F_{ST} outliers. The effect of mutation is small when the migration rate is relatively high ($2Nm > 1$). However, when the migration rate is low ($2Nm < 1$), the mutation rate has a significant effect on F_{ST}. F_{ST} first increases and then decreases, whereas MAF monotonically increases, when the mutation rate increases. The pattern of F_{ST} as a function of the mutation rate observed in this research is somewhat surprising, because previous theoretical studies predicted that higher mutation rates decrease F_{ST} (e.g., Takahata and Nei 1984, Wilkinson-Herbots 1998). Figure 3.7 shows a comparison of the expected F_{ST} estimated using the population allele frequencies simulated in this research with the expected value of F_{ST} calculated, using the analytical expression, $1/(1+16Nm+8Nu)$. F_{ST} in this research is lower than the

theoretical F_{ST}, especially when the mutation rate is low. This discrepancy appears to be due to the different approaches used to calculate the expected value of this statistic in these simulations and in previous analytical studies. F_{ST} is defined at each site in this research by equation 2.1. On the other hand, previous studies calculated the expected value of F_{ST} by setting $F_{ST} = (E[H_T] - E[H_S])/E[H_T]$. The discrepancy is due to the fact that $E[(H_T - H_S)/H_T] \neq (E[H_T] - E[H_S])/E[H_T]$, because the expected value of the ratio is generally not equal to the expected value of the numerator to that of the denominator. Another possible cause of the discrepancy is that F_{ST} is defined at bi-allelic loci in this research, whereas the infinite alleles model is assumed in the previous studies. In fact, a recent study on genetic differentiation between two populations at bi-allelic loci reports a pattern of a measure of population differentiation as a function of the mutation rate similar to that found here (Dewar et al. 2011). Although the ascertainment scheme simulated in this study is different from that in HapMap data, the maximum MAF of 1/60 imposed in the sample raises sample F_{ST} compared with population F_{ST}, which is qualitatively consistent with the effect of ascertainment biases on F_{ST} observed in Chapter 2.

Strong purifying selection lowers frequencies of deleterious mutations and thus MAF and F_{ST}. This effect becomes higher when the migration rate is lower. Again, this is because the actual high differentiation with low migration rates is limited by the maximum possible value of F_{ST} when MAF is low. The dominance coefficient of the deleterious allele affects F_{ST} also by modulating MAF. Interestingly, when there is a large difference in size between populations, MAF

and F_{ST} can be increased by purifying selection from neutral values. This is because purifying selection dominates in the larger population, whereas random genetic drift dominates in the smaller population, which may lead to larger difference in frequencies of the deleterious allele between them. Stronger purifying selection also diminishes MAF and F_{ST} in the population split model. This effect becomes severe when the populations diverged a long time ago. Overall, the results in this study show that migration rates and the time since a population split could be substantially underestimated if the effects of purifying selection on F_{ST} statistics are neglected. Comparison of the degree of genetic differentiation between pairs of populations by F_{ST} needs to be made using sites with similar minor allele frequencies.

A. $2Nm = 0.4$



B. $2Nm = 2$

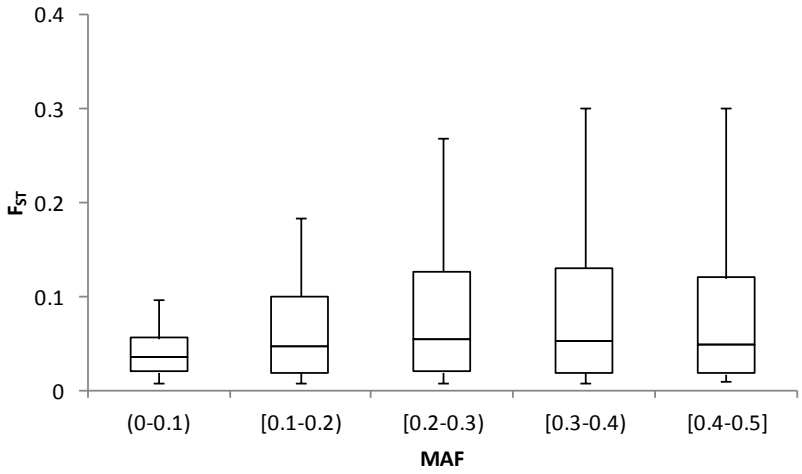
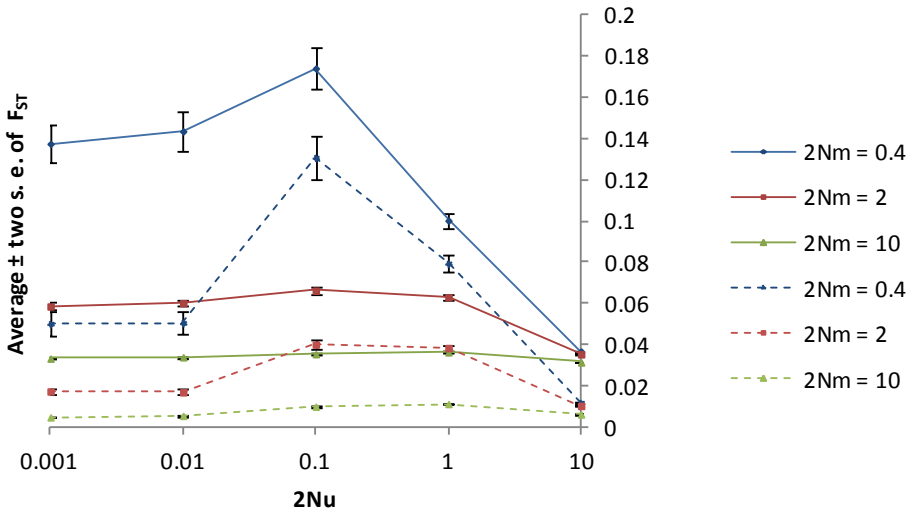


Figure 3.1. Box plots of F_{ST} between two demes with $2Nm = 0.4$ (A) and $2Nm = 2$ as functions of the minor allele frequency (MAF) across populations in the subdivided population model, where m is the migration rate. The parameter values used: $N_1 = N_2 = N = 10^4$, $t = 0$ and $u = 5 \cdot 10^{-6}$, where N_i, t, and u are the

diploid effective size of deme i, selection coefficient against the deleterious allele, and mutation rate, respectively. The median is shown as the horizontal line dividing the box. The whiskers stop at 5th and 95th percentiles.

A. F_{ST}



B. MAF

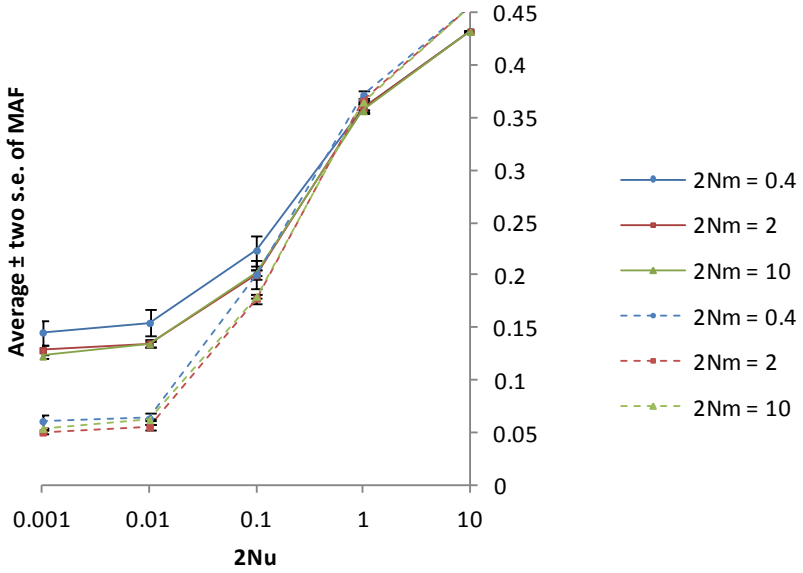
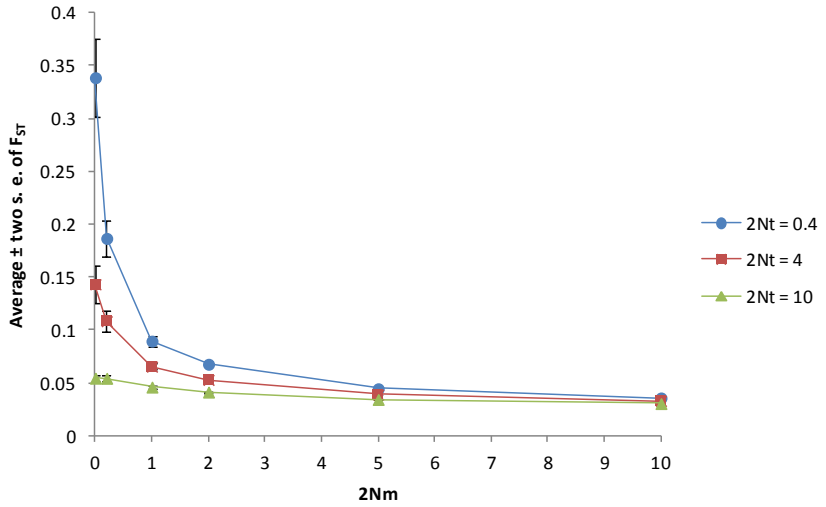


Figure 3.2. F_{ST} between two demes (A) and MAF across demes (B) at a locus under neutral evolution as a function of the mutation rate, $2Nu$, where u is the mutation rate, in the subdivided population model. The continuous and dashed lines are used for sample and population F_{ST}/MAF, respectively. The average and

two standard error are shown as points and bars, respectively. The parameter values used: $N_1 = N_2 = N = 10^4$ and $t = 0$, where N_i and t are the diploid effective size of deme i and selection coefficient against the deleterious allele, respectively. 1,000 simulation replicates, each with 100 pairs of initial allele frequencies are run. A burn-in period of 500,000 generations is taken before data are recorded. The number of recorded data for each set of parameter values: 10^5.

A. F_{ST}



B. MAF

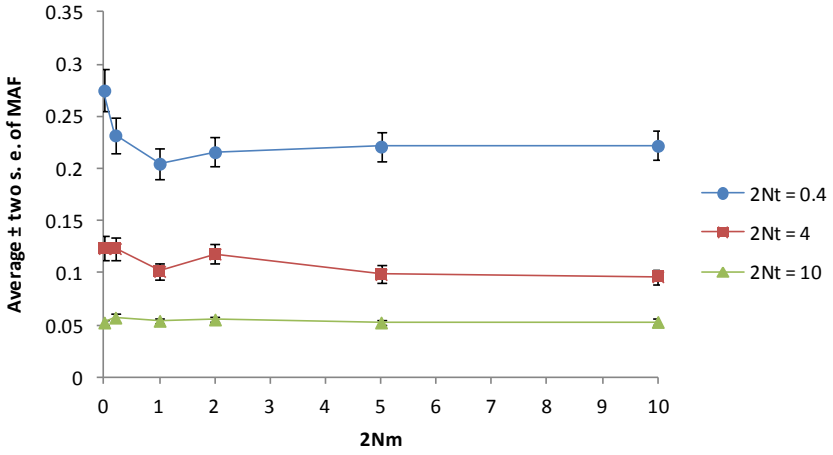
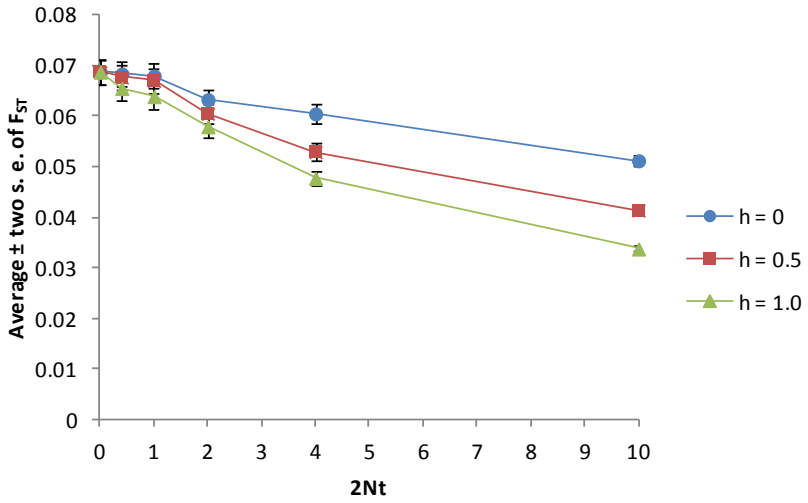


Figure 3.3. The effect of the migration rate on F_{ST} between two demes (A) and MAF (B) at a locus under purifying selection in the subdivided population model. The parameter values used: $N_1 = N_2 = N = 10^4$, $u = 5 \cdot 10^{-6}$ and $m = 10^{-4}$, where N_i, u, and m are diploid effective size of deme i, mutation rate, and migration rate, respectively. The average and two standard error are shown as points and bars, respectively. The number of recorded data for each set of parameter values: 10^5.

A. F_{ST}



B. MAF

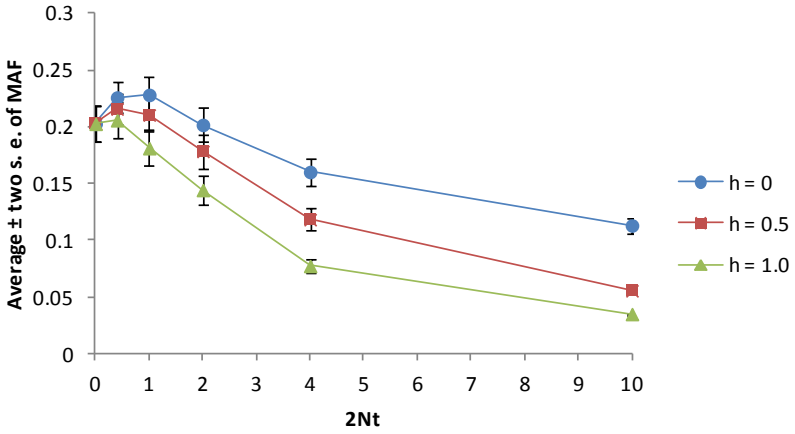
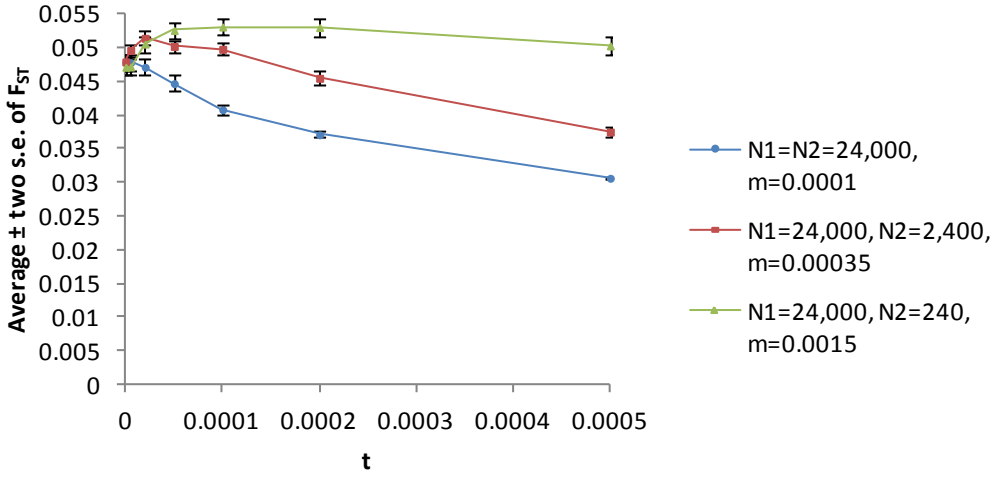


Figure 3.4. The effect of the dominance coefficient on F_{ST} between two demes (A) and MAF (B) at a locus under purifying selection in the subdivided population model. The parameter values used: $N_1 = N_2 = N = 10^4$, $u = 5 \cdot 10^{-6}$ and $m = 10^{-4}$, where N_i, u, and m are diploid effective size of deme i, mutation rate, and migration rate, respectively. The average and two standard error of F_{ST}/MAF

are shown as points and bars, respectively. The number of recorded data for each set of parameter values: 10^5.

A. F_{ST}



B. MAF

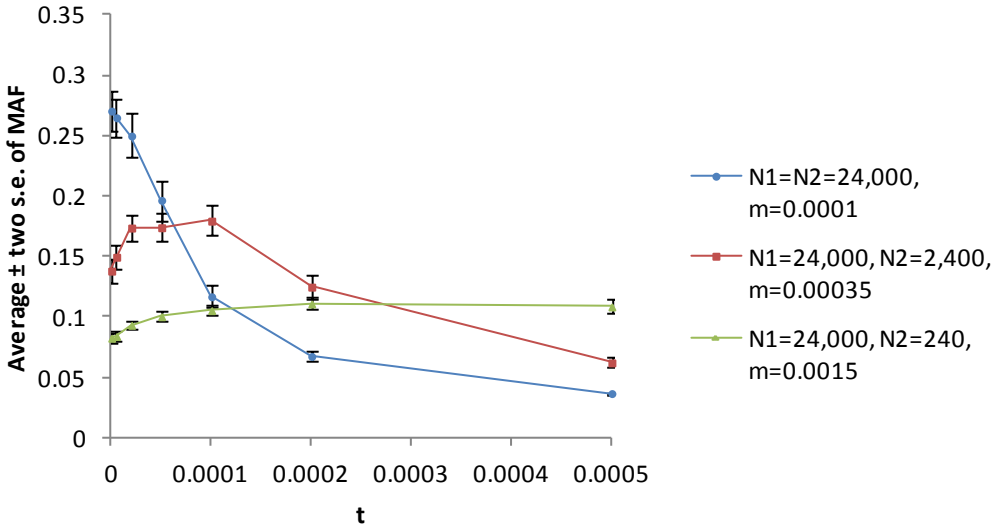
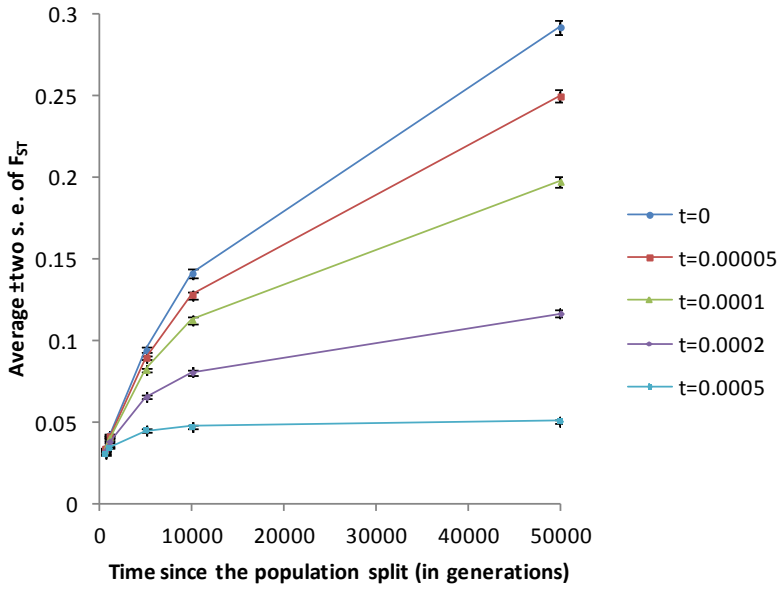


Figure 3.5. The effect of unequal deme sizes on F_{ST} between two demes (A) and MAF (B) at a locus under purifying selection in the subdivided population model. $u = 5 \cdot 10^{-6}$, where u is the mutation rate, is used. The average and two standard

error of F_{ST}/MAF are shown as points and bars, respectively. The number of recorded data for each set of parameter values: 10^5.

A. F_{ST}



B. MAF

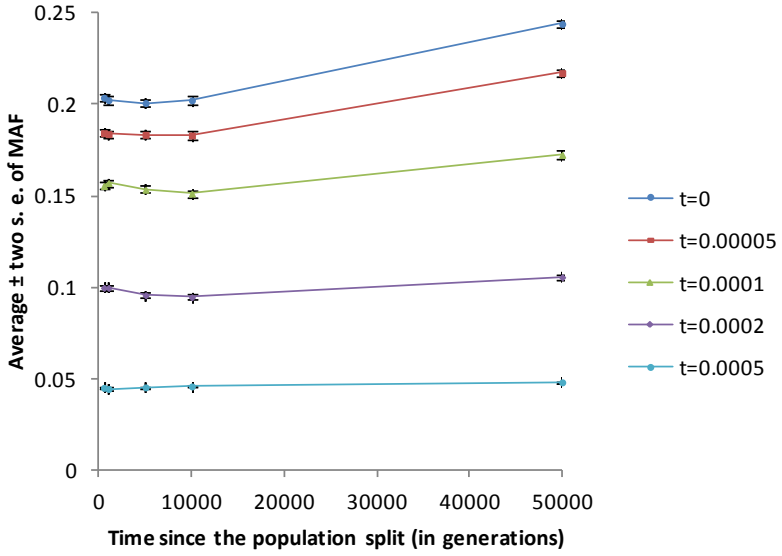
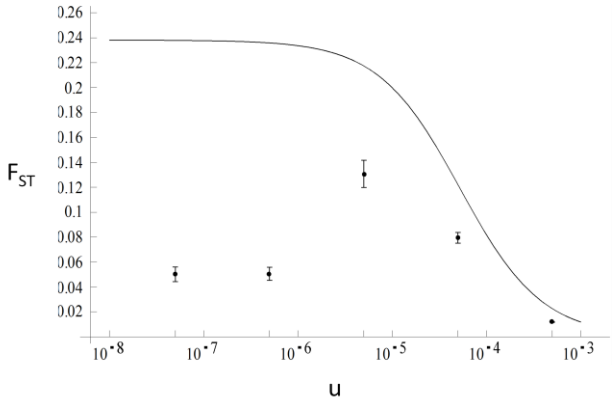


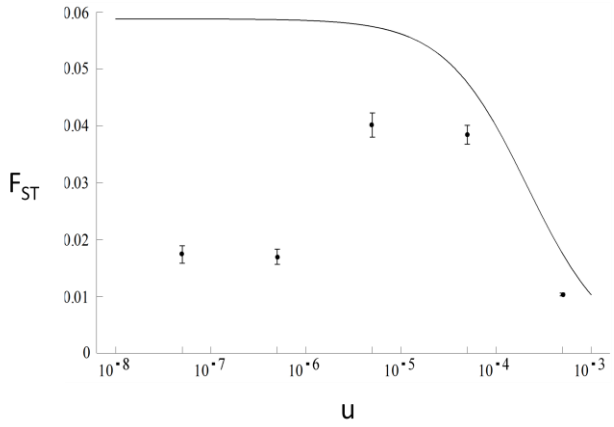
Figure 3.6. F_{ST} between demes (A) and MAF (B) at a locus under purifying selection in the population split model. The parameter values used: N_a = 24,000,

N_1 = 24,000, N_2 =7,700, $u = 5 \cdot 10^{-6}$, h = 0.5, where N_a, N_i, u, and h are the diploid effective size of the ancestral population, diploid effective size of deme i, mutation rate, and dominance coefficient, respectively. The average and two standard error of F_{ST}/MAF are shown as points and bars, respectively. The number of recorded data for each set of parameter values: $2 \cdot 10^4$.

A. $2Nm = 0.4$



B. $2Nm = 2$



C. $2Nm = 10$

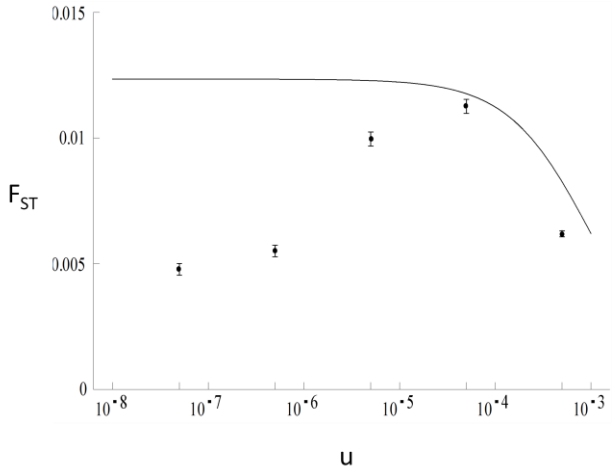


Figure 3.7. Comparison of population F_{ST} obtained by the simulations in this research and theoretical F_{ST} by previous researchers as a function of the mutation rate u with different migration rates $2Nm = 0.4$ (A), 2 (B), and 10 (C), where N and m are diploid effective size of each deme and migration rate, respectively. The simulation results are identical to those shown in Figure 3.2. The average and two standard error are shown as points and bars, respectively for the simulation results. The theoretical F_{ST} is shown by a curve.

CHAPTER 4: SPREAD OF A BENEFICIAL ALLELE AND THE HITCHHIKING EFFECT IN A SUBDIVIDED POPULATION

Abstract

Detecting and analyzing the fixation of an allele caused by positive directional selection are important for understanding the mechanisms of evolution. Genetic hitchhiking leaves footprints of such fixation events, which help scientists to infer past scenarios of positive directional selection. Although a great deal of advance has been recently made in the development of models of genetic hitchhiking, most of the existing models assume a single random-mating population. Many of natural populations show evidence of geographic structure. Spread of a beneficial allele can be delayed by limited amounts of gene flow due to the geographic structure in a natural population. Such delays may have a significant impact on footprints of genetic hitchhiking. Therefore, I investigate the effects of geographic structure of a population on the spread of a beneficial allele and resulting footprints of genetic hitchhiking. Island and stepping-stone models are used for the population structure and simulation results under the two different models are compared to study the effect of distance-dependent migration observed in many natural populations. The strength of the hitchhiking effect is measured as degree of the reduction in the average heterozygosity at neutral loci closely linked to the selected locus. Unlike previous studies on genetic hitchhiking in a subdivided population, cases with large amounts of migration are investigated in this study. Specifically, it is assumed that $2Nm > 1$ but $m < s$,

where 2*N*, *m*, and *s* are effective population size in each deme, migration rate, and selection coefficient, respectively. These conditions are important, because geographic structure of a population may not be detected by polymorphism at isolated neutral loci but a significant effect on genetic hitchhiking as a result of the delay in the spread of a beneficial allele may be observed. The results show that the hitchhiking effect in the total population is diminished by the geographic structure because of the increased opportunities for recombination between selected and neutral loci due to the increased time taken for the fixation of the beneficial allele. Also, the strength of the hitchhiking effect in each deme is shown to decrease with the increase in the distance between the deme and origin of the beneficial mutation. These results suggest that close examination of footprints of genetic hitchhiking may reveal 'hidden' geographic structure of a population, which is difficult to detect from polymorphism at isolated neutral loci.

Introduction

When a beneficial allele is introduced by a mutation and its frequency rapidly increases by positive directional selection, frequencies of alleles on the same chromosome as the beneficial allele, at loci closely linked to the selected locus also rapidly increase. This effect is called the hitchhiking effect (Maynard-Smith and Haigh 1974) or a selective sweep. Genetic hitchhiking leaves footprints of positive directional selection, which provide means for identifying and analyzing recent scenarios of positive selection (reviewed in Nielsen 2005, Sabeti et al. 2006, Thornton et al. 2007, Akey 2009, Stephan 2010). A number of mathematical

studies on genetic hitchhiking have recently been conducted and they provide detailed theoretical prediction on the characteristics signatures and tools for detecting past scenarios of positive directional selection (Maynard Smith and Haigh 1974, Kaplan et al. 1989, Stephan et al. 1992, Fay and Wu 2000, Kim and Stephan 2002, Hermission and Pennings 2005, Etheridge et al. 2006). However, the major results were obtained in models in a single panmictic population, where geographic structure is ignored.

Natural populations show geographic structure and are composed of several demes. When migration is geographically limited, mating occurs more often among individuals geographically close to each other. The effect of the geographic structure of a population can be examined in simple models, where a number of demes, each of which is panmictic, are connected by limited amounts of migration (Wright 1940). In these models, the migration rate, m, represents the proportion of migrants coming from other demes in a deme per generation. One of the major results in spatially structured populations is that, if m is sufficiently large such that $2Nm > 1$, where N is diploid effective size of a deme, polymorphism at loci under neutral evolution is well homogenized and the total population appears to be panmictic under mutation-migration-genetic drift equilibrium (Slatkin 1987). Therefore, even when populations are actually structured ($m << 0.5$), their geographic structures may not have significant effects on patterns of polymorphism at isolated neutral loci, when $2Nm < 1$.

However, if an evolutionary process rapidly occurs at a time scale shorter than that of neutral coalescence, limited amounts of migration may have significant effects on footprints of genetic hitchhiking even when $2Nm > 1$. Let s be the selection coefficient for the beneficial allele such that the relative fitness of the allele is $1 + s$. Then, the spread of the beneficial allele is expected to be affected by limited amounts of migration when $m < s$, regardless of the value of Nm. That is, a significant effect of the geographic structure on the frequency path of the beneficial allele is expected in a subdivided population when $m < s$. Then, the resulting footprints of genetic hitchhiking may be different from those in a panmictic population. For example, Barton (2000) predicted that the effect of genetic hitchhiking in a subdivided population would diminish due to the longer time taken for the fixation of the beneficial allele compared to that in a panmictic population.

A few mathematical models of genetic hitchhiking in a subdivided population have been previously developed (Slatkin and Whiehe 1998, Santiago and Caballero 2005). They focused on cases with small amounts of migration among demes ($2Nm << 1$), where the fixation of the beneficial allele occurs only in one deme at a given time. They showed that, if there is initially little genetic differentiation among demes at neutral loci close to the selected locus, the degree of genetic differentiation is increased at the loci linked to the selected locus by genetic hitchhiking. Therefore, Wright's F_{ST} increases from small to intermediate values at the linked loci affected by a selective sweep (Slatkin and Wiehe 1998, Bierne 2010). If, on the other hand, the initial degree of the genetic

differentiation among demes at linked loci is high, the genetic differentiation at the linked loci is decreased by a selective sweep, because common alleles are more likely to hitchhike along with the beneficial allele when recombination is limited. Therefore, F_{ST} decreases from large to small values at the linked loci by a selective sweep in this case (Santiago and Caballero 2005). These studies were useful in analyses of footprints of positive directional selection in organisms with low migration rates, including *Drosophila ananassae* (Stephan et al. 1998, Baines et al. 2004, Das et al. 2004) and *Mytilus edulis* (Faure et al. 2008).

In collaboration with Dr. Kim, I investigate cases with more frequent migration among demes ($2Nm > 1$), where long-term neutral polymorphism appears to be similar to that under panmictic population, in this study. In these cases, fixation processes of the beneficial allele in different demes can occur at the same time. This biological condition is important, because the geographic structure of a population may have a significant effect on footprints of positive directional selection, which is not seen at isolated neutral loci. The results here show the importance of this 'hidden' geographic structure on the pattern of polymorphism shaped by a rapid evolutionary process.

Materials and Methods

The model

A schematic figure of a selective sweep in a subdivided population in comparison with that in a panmictic population is shown as Figure 4.1. A positively selected

allele, *B*, rapidly increases along with a hitchhiking allele *A* at a linked locus in the first deme. Such an association is broken down by recombination events, which allow some amounts of polymorphism to remain at the linked locus after a selective sweep. The major differences between genetic hitchhiking in a subdivided population and that in a panmictic population are described in Figure 4.1. First, the time taken for the fixation of allele *B* is longer in a subdivided population, because it takes some time until *B* introduced by migration from the first deme survives stochastic loss due to genetic drift in the second deme. Second, opportunities for recombination to break down the association of the alleles are expected to increase in a subdivided population. This is because the opportunities monotonically decrease as the frequency of *B* increases in a panmictic population, while there are multiple times when the frequency of *B* is low in a deme in a subdivided population. This may result in a weaker effect of genetic hitchhiking due to the increased breakdowns of association of the alleles in a subdivided population. Importantly, different alleles, either *A* or *a*, can be the hitchhiking allele in the second deme, depending on which *B*-bearing chromosome, *A*-*B* or *a*-*B* migrates and establishes in the second deme.

In this research, the haploid population consists of K demes, each of which has effective size $2N$. Demes are structured according to the circular stepping-stone model if $K > 2$, unless stated otherwise. Demes are indexed by numbers from 1 to K, indicating their spatial order. Demes 1 and K are next to each other such that the demes form a circular structure. Generations are non-overlapping and each generation consists of four biological processes of selection,

recombination, migration, and random genetic drift. In the migration process, the migration rate m specifies the proportion of migrants in a deme coming from the neighboring demes ($m/2$ from deme i-1 and $m/2$ from deme i+1 in deme i). When K=2, migrants in a deme come from the other deme. Two bi-allelic loci on the same chromosome, one positively selected and the other neutral, are employed to model the process of genetic hitchhiking. Recombination occurs at rate r per generation between the two loci. The beneficial allele B with selective advantage s is introduced by a mutation from the ancestral allele b at the selected locus on a randomly chosen chromosome in deme 1. When this beneficial mutation occurs, there is polymorphism at the neutral locus with two alleles in frequencies p_0 and 1-p_0, respectively, in each deme. After a selective sweep, p_0 in deme j changes to p_j. Then, heterozygosity in the total population after a selective sweep, $H^{(T)}$, is given by $H^{(T)} = 2p(1-p)$, where $p = \sum_j p_j/K$. The hitchhiking effect in the total population is described in this study and is measured by the ratio $H^{(T)}/\tilde{H}$, where $\tilde{H} = 2\,p_0(1-p_0)$.

The simulation

The above discrete-time model is simulated by a forward-in-time frequency based simulation. Each generation consists of deterministic changes in haplotype frequencies by selection, recombination, and migration, followed by the random change by the step of random sampling that uses a random binomial number generator (Kim and Wiehe 2009). p_0 is given as a fixed value (0.2) for all demes. The initial distribution of the allele frequencies was also specified by the

distribution under mutation-migration-genetic drift equilibrium (obtained by a separate forward-in-time simulation) but there was little difference in the results when the hitchhiking effect was measured by $H^{(T)}/\widetilde{H}$ (results not shown). The beneficial mutation occurs on a randomly chosen chromosome such that the probability for a neutral allele to become the hitchhiking allele is equal to its frequency. If the beneficial allele is lost, the simulation replicate is repeated from the beginning until its fixation occurs in the total population. All simulation results in this research are based on 10,000 replicates for each set of parameter values.

Results

Spread of the beneficial allele in a subdivided population with two demes

The frequency of the beneficial allele rapidly increases in a deme by positive directional selection. In order for the allele to spread across the demes, it needs to be introduced to other demes by migration. However, even when the allele is introduced to another deme by migration, it is lost just by chance by random genetic drift with high probability. Therefore, there may be a 'delay' in the fixation of the allele in a subdivided population compared to the case in a panmictic population of equal size. In order to analyze the hitchhiking effect in a subdivided population, I first examined how much delay in the fixation of the beneficial allele is caused by the geographic structure of a population. The deme, where the beneficial allele is introduced by a mutation, is defined as deme 1. The beneficial allele is assumed to eventually be fixed in the total population.

Let X_i (T) be the frequency of the beneficial allele B in deme i at time T, which is measured forward in generation and defined to be zero at the time of the beneficial mutation. Define $\check{T}_i = max_T(X_i\,(T) > 0$ and $X_i(T-1) = 0)$ such that $\check{T}_i$ shows the time when allele B survives the genetic drift and is established in deme i. Then, the 'delay' in the spread of allele B is defined by

$$\delta = \left|\check{T}_2 - \check{T}_1\right|. \tag{4.1}$$

(When $m << s$, $\check{T}_2 > \check{T}_1$ in most cases. However, when the migration rate is higher, deme 1 may lose allele B and later receive the allele from deme 2. In this case, the initial roles of demes 1 and 2 are switched and the delay is defined to be $\check{T}_1 - \check{T}_2$.)

Figure 4.2 shows the delay as a function of the migration rate, with two different values of selection coefficients. As expected, the delay becomes larger when selection coefficients and migration rates are smaller.

Spread of the hitchhiking effect in a subdivided population with two demes

Next, the spread of the hitchhiking effect is examined in a subdivided population that consists of two demes. As explained above, the degree of the effect is measured as the ratio of the average heterozygosity at linked neutral loci after a selective sweep to that at linked loci before a selective sweep. Heterozygosity is the probability that two randomly chosen chromosomes are different at the locus under examination. In order to investigate the pattern of the hitchhiking effect distributed over the demes in detail, three kinds of heterozygosity with different

sampling schemes, $H^{(11)}$, $H^{(22)}$, and $H^{(12)}$, are examined. $H^{(11)}$, $H^{(22)}$, and $H^{(12)}$, are the average heterozygosity at linked loci after a selective sweep when two chromosomes are sampled from deme 1 only, deme 2 only, and both of the demes, respectively. Because of the assumption that $2Nm > 1$, all of these three types of the average heterozygosity at neutral loci before a selective sweep are given by $\widetilde{H}$. Then, assuming two chromosomes are randomly sampled from the demes, the heterozygosity ratio in the total population has the following relationship with the ratios of three different kinds of heterozygosity:

$$H^{(T)}/\widetilde{H} = 1/4 \cdot (H^{(11)}/\widetilde{H}) + 1/2 \cdot (H^{(12)}/\widetilde{H}) + 1/4 \cdot (H^{(22)}/\widetilde{H}) \qquad (4.2)$$

The effect of the geographic structure on genetic hitchhiking may be the greatest in deme 2, where the beneficial allele is introduced by migration. Therefore, the hitchhiking effect in deme 2, measured as $H^{(22)}/\widetilde{H}$, is first examined. Figure 4.3 shows the heterozygosity ratio as a function of a scaled recombination rate (r/s). Compared to that in a corresponding single panmictic population, the heterozygosity ratio in a subdivided population is higher, which means the hitchhiking effect is diminished in a subdivided population. Figure 4.4 shows how much the heterozygosity ratio is higher in a subdivided population compared to that in a corresponding panmictic population. The increase in the heterozygosity ratio (decrease in the hitchhiking effect) in a subdivided population is greater when the migration rate is decreased.

The effect of the geographic structure on genetic hitchhiking is similar when two chromosomes are sampled from both of the demes (Figure 4.5). That is,

geographic structure of a population diminishes the hitchhiking effect. However, when two chromosomes are sampled from deme 1 only, the hitchhiking effect becomes stronger by the geographic structure (Figure 4.6). A smaller migration rate results in decrease in the heterozygosity ratio in a subdivided population compared to that in a corresponding panmictic population in this case. This may be understood when the initial frequency of the beneficial allele is considered, under the assumption that migration of the beneficial allele from deme 2 to deme 1 is rare. Under the assumption, the initial frequency of the allele in a subdivided population is effectively half of that in the panmictic population, which is formed by combining demes 1 and 2. The effect of genetic hitchhiking becomes stronger when the initial frequency of the beneficial allele is higher, because the time taken for the fixation of the allele becomes shorter (Barton, 2000). Therefore, the hitchhiking effect in the origin of the beneficial allele is increased when $m << s$ but this increase of the hitchhiking effect disappears when m approaches s.

The hitchhiking effect in the total population as a function of the scaled recombination rate is shown in Figure 4.7. The geographic structure of a population increases the heterozygosity ratio and thus diminishes the overall effect of genetic hitchhiking. Figure 4.8 shows how much the heterozygosity ratio is increased by the geographic structure compared to that in a corresponding panmictic population. The effect of the geographic structure on the overall effect of genetic hitchhiking increases when the migration rate decreases. This effect of the geographic structure is significant only when m/s is small. When m/s

approaches 0.1, there is little effect of the geographic structure on genetic hitchhiking.

The case in a subdivided population with ten demes

The effect of geographic structure on polymorphism in a population becomes stronger when the number of demes is more than two. There is large difference in the migration patterns between island and stepping-stone models when the number of demes, K, is greater than two, whereas they are identical when K is equal to two. Therefore, it is important to investigate the effect of geographic structure on genetic hitchhiking when K is greater than two. As an example, I explore the case where K is equal to ten. As in the case with two demes, the origin of the beneficial allele is defined as deme 1 and the beneficial allele is assumed to be fixed in the total population. Figure 4.9 shows the time taken for the beneficial allele to be fixed in the total population as a function of the migration rate. The time greatly increases compared to that in a corresponding panmictic population with the decrease in the migration rate when $m << s$. On the other hand, the difference is small when $m \geq s$. The effect of geographic structure on genetic hitchhiking in the total population is similar to that in the case with two demes (Figure 4.10A). The effect of the geographic structure on the spread of the beneficial allele is stronger than that on the hitchhiking effect: The increase in the fixation time and heterozygosity ratio from those in a corresponding panmictic population, when $m/s = 10^{-2}$, are 68% and 34%, respectively (compare Figures 4.9 and 4.10A). The pattern in the strength of the

hitchhiking effect in each deme is also similar to that seen in the case with two demes: The hitchhiking effect is the strongest in the origin of the beneficial allele (deme1) and diminishes as the distance of a deme from deme 1 increases (Figure 4.10A).

To examine the effect of the distance-dependent migration mode in the stepping-stone model, corresponding results with the same parameter values in Wright's island model are shown as controls in Figure 4.10B. In the island model, migrants into a deme come from all of the other demes equally and therefore, there is no distance-dependence in the migration mode. Figure 4.10B shows that the overall effect of the geographic structure on genetic hitchhiking in the island model is similar to that in the stepping-stone model. The hitchhiking effect in the total population diminishes with smaller migration rates. The hitchhiking effect becomes stronger in the deme of origin of the beneficial allele but becomes weaker in the other demes (results in only three out of nine demes are shown in Figure 4.10B) when $m/s << 1$. However, these patterns are much clearer in the stepping-stone model. Therefore, the heterogeneous pattern of the hitchhiking effect across demes is expected to be seen in organisms with distance-dependent migration mode. More detailed results concerning the heterogeneous pattern of the hitchhiking effect across demes in the stepping-stone model is shown in Table 1. In the table, $H^{(ij)}$ is the average value of the heterozygosity at the linked locus when two chromosomes are sampled from demes *i* and *j*. The coefficient of variation, cv_{ij}, is given by $cv_{ij} = \sqrt{Var[H^{(ij)}]}/H^{(ij)}$. F_{ST} is the average value of

Wright's F_{ST} at the linked locus and is given by $F_{ST} = (H^{(T)} - H^{(S)})/H^{(T)}$, where $H^{(T)}$ is the heterozygosity in the total population and $H^{(S)} = (\sum_{i=1}^{K} H^{(i)})/K$. The results show that the decrease in migration rates result in greater increase in F_{ST} compared to that in the coefficient of variation. Note that the simulation assumes the same initial allele frequencies at the linked locus in all demes ($F_{ST} = 0$). As reported by Slatkin and Wiehe (1998) and Bierne (2010), the effect of population structure on genetic differentiation at the linked locus is highly dependent on the recombination rate: For a given *m*, intermediate values of *r*/*s* (0.03-0.1) give the largest values of F_{ST}.

Discussion

Geographic and demographic structures of natural populations have significant effects on evolutionary genetic processes and therefore patterns of polymorphism including footprints of genetic hitchhiking (Jensen et al. 2005, Nielsen et al. 2005, Li and Stephan 2006, Kim and Gulisija 2010, Stephan 2010). This study examined how the geographical structure of natural populations affects the spread of the beneficial allele and hitchhiking effect in simple models of population subdivision. Footprints of genetic hitchhiking were shown to be affected even by relatively weak population structure such that its impact may not influence the patterns of neutral polymorphism. The geographic structure of a population modulates footprints of genetic hitchhiking in several important ways.

First, the hitchhiking effect is diminished in a subdivided population if the migration rate is much smaller than the selection coefficient. As briefly argued by

Barton (2000), this is because the opportunities for recombination to break down the association between beneficial and neutral alleles increase when the time taken for the fixation of the beneficial allele increase in a geographically structured population. This result indicates the strength of selection estimated under models of genetic hitchhiking in a panmictic population (Kim and Stephan 2002, Thornton et al. 2007) may be underestimated. Furthermore, the effect of the geographic structure on the spread of the beneficial allele was shown to be even greater than that on the hitchhiking effect. These results indicate that the strength of selection estimated from the chromosomal span of reduced polymorphism may be greater than that estimated from the age of the sweeping haplotype inferred from rare mutations (see, for example, Sáez et al. 2003, Meikeljohn et al. 2004, and Xue et al. 2006). However, it is not clear whether such difference can be detected with reasonable statistical power.

Another important result is the negative relationship between the strength of the hitchhiking effect in a deme and its distance from the origin of the beneficial mutation. This is because of the increased time taken for the fixation of the beneficial allele and therefore opportunities for recombination in demes where the allele is introduced by migration. Slatkin and Wiehe (1998) and Bierne (2010) demonstrated that the degree of genetic differentiation at linked loci can be increased by genetic hitchhiking and suggested that this is because a selective sweep leaves heterogeneous polymorphism across populations, which is in accordance with the results here. The heterogeneous outcomes in a subdivided population result when positive directional selection occurs faster than migration.

This gradient of the footprints of genetic hitchhiking may help scientists to infer the origin of the beneficial mutation and patterns of migration difficult to detect at isolated neutral loci.

The results in this study were obtained in simple models of a structured population, where demes have equal size and the population structure remains the same during the process. However, natural populations experience complex demographic changes. For example, an ancestral population is split into several demes distributed over geographic space, which is known to have important impacts on polymorphism. Nonetheless, the results in this study should be applicable to natural populations. If a beneficial allele spreads very rapidly across demes by strong selection, the hitchhiking process would occur under effectively constant demographic structure, because major demographic changes occur at a time scale much longer than that of a selective sweep. The results here apply as long as demes are genetically homogeneous when a selective sweep begins. This condition would be met if the demes under study recently derived from an ancestral population and are genetically similar to each other.

This study suggests that other aspects of the footprints of genetic hitchhiking, such as those in site frequency spectrum and linkage disequilibrium, may be also significantly affected by the geographic structure of a population. Site frequency spectrum and linkage disequilibrium are highly dependent on the shape of the coalescent trees at the linked loci, which is strongly influenced by the shape of the genealogy at the selected locus (Fay and Wu 2000, Kim and Nielsen

2004, McVean 2007, Pfaffelhuber et al. 2008). For example, Pfaffelhuber et al. (2006) showed that approximating the genealogy at the selected locus by a Yule process corrects the error introduced by the assumption of the simplified star-like genealogy. Population subdivision is likely to result in great deviation from the star-like genealogy that cannot be handled by the Yule process. In particular, the deviation is expected to have significant impacts when the force of selection is stronger than that of migration. Further studies are necessary to characterize the footprints of genetic hitchhiking in a subdivided population.

r/s	m/s	$H^{(11)}$ (cv_{11})	$H^{(33)}$ (cv_{33})	$H^{(66)}$ (cv_{66})	$H^{(13)}$ (cv_{13})	$H^{(16)}$ (cv_{16})	$H^{(36)}$ (cv_{36})	F_{ST}
0.01	0.01	0.041 (1.81)	0.057 (1.72)	0.080 (1.51)	0.057 (1.78)	0.082 (1.76)	0.089 (1.65)	0.068
0.01	0.03	0.042 (1.66)	0.054 (1.60)	0.078 (1.43)	0.052 (1.59)	0.070 (1.55)	0.075 (1.49)	0.037
0.01	0.1	0.047 (1.49)	0.054 (1.45)	0.069 (1.33)	0.052 (1.43)	0.061 (1.36)	0.064 (1.35)	0.015
0.01	0.3	0.045 (1.40)	0.051 (1.39)	0.057 (1.33)	0.051 (1.38)	0.054 (1.33)	0.055 (1.34)	0.0047
0.01	1	0.049 (1.34)	0.049 (1.34)	0.049 (1.34)	0.049 (1.34)	0.049 (1.34)	0.049 (1.33)	0.00086
0.001	0.1	0.0049 (4.18)	0.0054 (4.27)	0.0077 (4.06)	0.0053 (4.03)	0.0066 (3.99)	0.0069 (4.03)	0.0060
0.003	0.1	0.014 (2.52)	0.016 (2.51)	0.022 (2.36)	0.016 (2.44)	0.019 (2.42)	0.020 (2.37)	0.0088
0.01	0.1	0.047 (1.49)	0.054 (1.45)	0.069 (1.33)	0.052 (1.43)	0.061 (1.36)	0.064 (1.35)	0.015
0.03	0.1	0.119 (0.95)	0.133 (0.92)	0.161 (0.84)	0.130 (0.92)	0.149 (0.87)	0.155 (0.86)	0.022
0.1	0.1	0.246 (0.55)	0.258 (0.52)	0.281 (0.43)	0.257 (0.52)	0.276 (0.48)	0.279 (0.46)	0.021
0.3	0.1	0.312 (0.23)	0.313 (0.21)	0.316 (0.17)	0.315 (0.20)	0.319 (0.16)	0.319 (0.16)	0.0098
1	0.1	0.317 (0.14)	0.317 (0.14)	0.318 (0.14)	0.319 (0.11)	0.321 (0.10)	0.320 (0.10)	0.0063

Table 4.1. The heterogeneous pattern of footprints of genetic hitchhiking across demes in the stepping stone model with parameter values $K = 10$, $s = 0.01$, $2N = 10^5$, where K, s, and $2N$ are number of demes, selection coefficient, and effective size of each deme, respectively (reproduced from Kim and Maruki 2011).

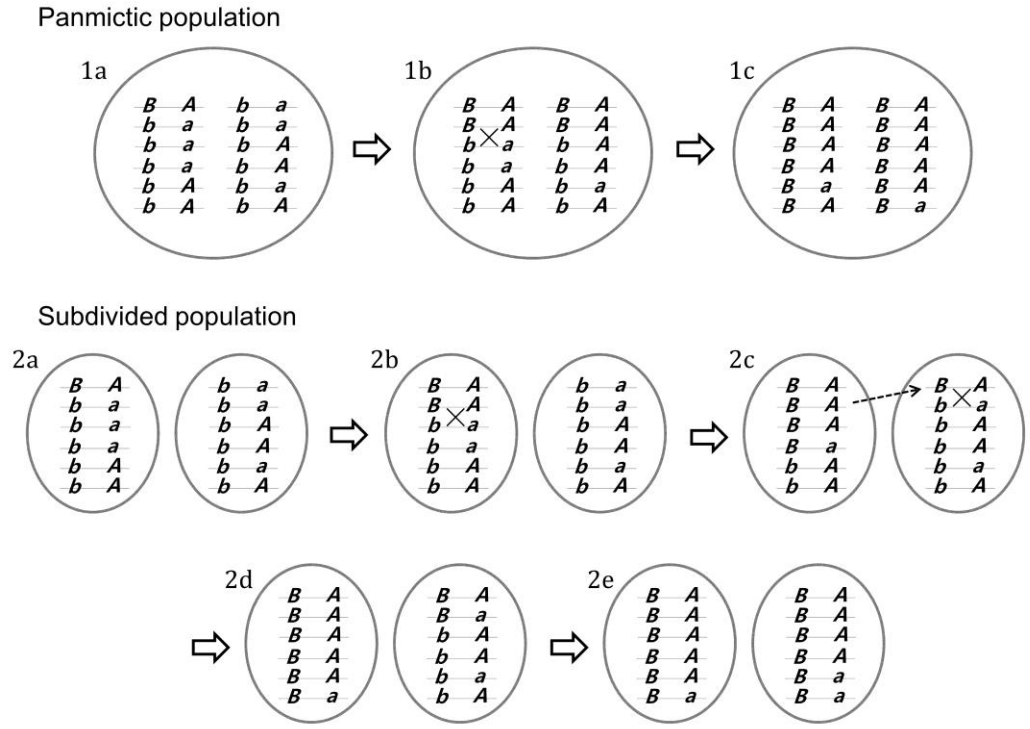


Figure 4.1. Comparison of the two-locus model of genetic hitchhiking in a panmictic population with that in a subdivided population (reproduced from Kim and Maruki 2011). Chromosomes shown in each population carry alleles at a selected locus (ancestral allele *b* or beneficial allele *B*) and a neutral locus (allele *A* or allele *a*). Allele *b* is initially fixed in the panmictic population. The beneficial allele is introduced by a mutation on a chromosome and the hitchhiking allele *A* becomes associated with allele *B* (stage 1a above). Then, the frequency of haplotype *BA* rapidly increases by positive directional selection when the amounts of recombination between the two loci are limited. At stage 1b, a *BA* chromosome recombines with a *ba* chromosome (indicated by "×"), which allows the increase of haplotype *Ba* in the population. Allele *a* thus survives the

wipeout but exists in low frequency when allele *B* is fixed (stage 1c). In a subdivided population, the rapid increase of allele *B*, along with allele *A* initially occurs only in the first deme (stages 2a and 2b). While allele *B* increases, its association with allele *A* is broken by recombination (stage 2b) and a chromosome carrying alleles *B* and *A* migrates to the second deme and starts increasing there (stage 2c). Association between alleles *B* and *A* is also broken by recombination in the second deme (stage 2c). Allele *B* is fixed in the first deme while it is still in intermediate frequency in the second deme (stage 2d). When allele *B* is fixed in the second deme (stage 2e), the frequency of allele *a* is low in both demes. Note that, a *Ba* instead of *BA* chromosome also can migrate at stage 2c , in which case allele *a* becomes dominant in the second deme and much less change in the overall allele frequencies at the neutral locus in the total population results after a selective sweep.

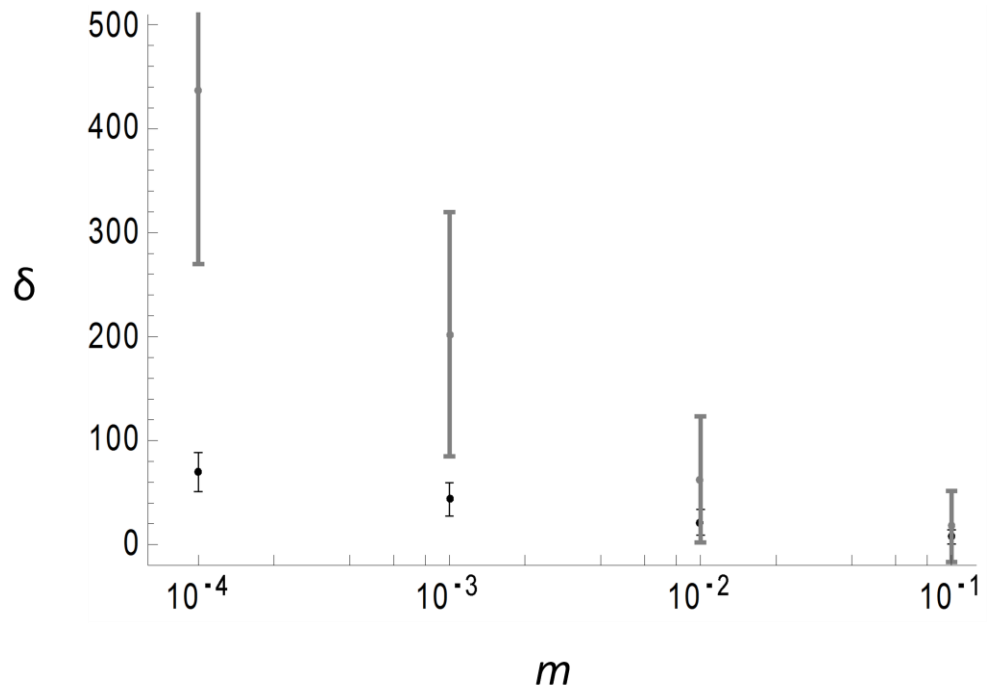


Figure 4.2. Delay (δ) in the spread of a beneficial allele as a function of the migration rate *m* with parameter values $K = 2$, $s = 0.01$ (gray) or 0.1 (dark) and $2N = 10^4$ (reproduced from Kim and Maruki 2011). Average ± standard error of δ are shown for each *m*.

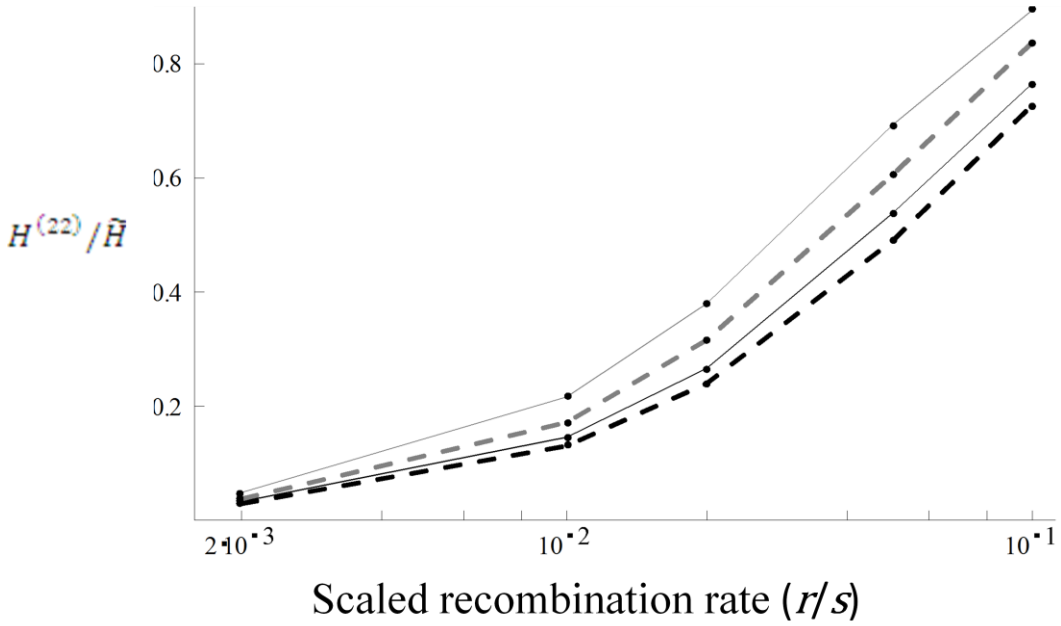


Figure 4.3. The effect of genetic hitchhiking on heterozygosity, when two chromosomes are randomly sampled from deme 2, measured by the average heterozygosity ratio $H^{(22)}/\tilde{H}$, as a function of the scaled recombination rate with parameter values $K = 2$, $s = 0.01$ (dark) or 0.1 (gray), $2N = 10^5$, and $m/s = 0.01$ (reproduced from Kim and Maruki 2011). Continuous and dashed curves show the hitchhiking effect in the subdivided population and that in a corresponding panmictic population of effective size $4N$, respectively.

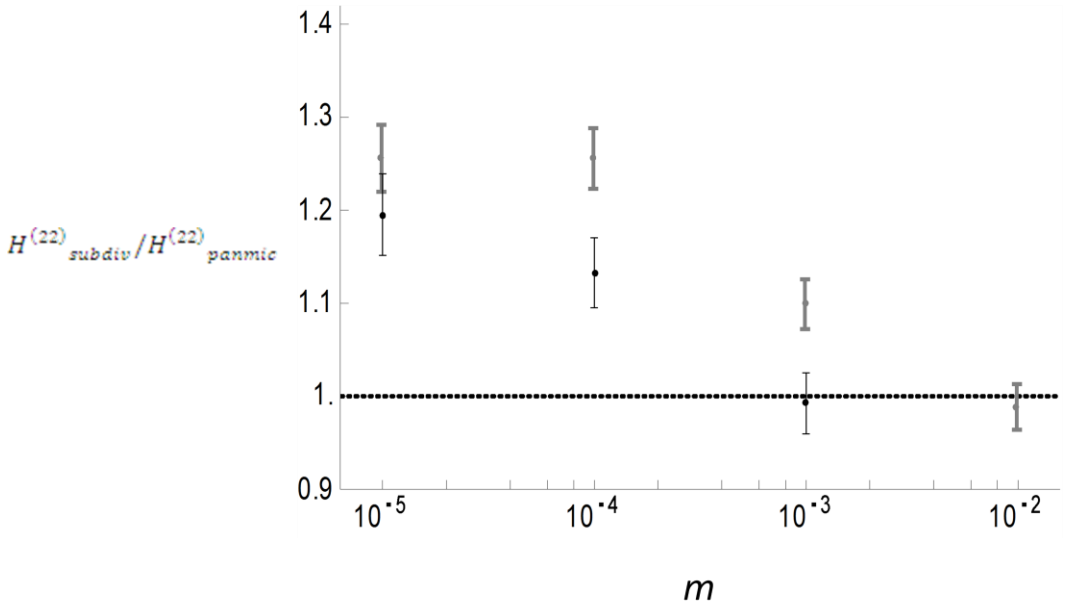


Figure 4.4. The effect of population subdivision on the hitchhiking effect, when two chromosomes are randomly sampled from deme 2, measured as the average heterozygosity ratio $H^{(22)}_{subdiv}/H^{(22)}_{panmic}$, as a function of the migration rate m with parameter values $K = 2$, $2N = 10^5$, $r/s = 0.01$, $s = 0.01$ (dark) or 0.1 (gray) (reproduced from Kim and Maruki 2011). Average ± 2 standard error of $H^{(22)}_{subdiv}/H^{(22)}_{panmic}$ are shown for each m.

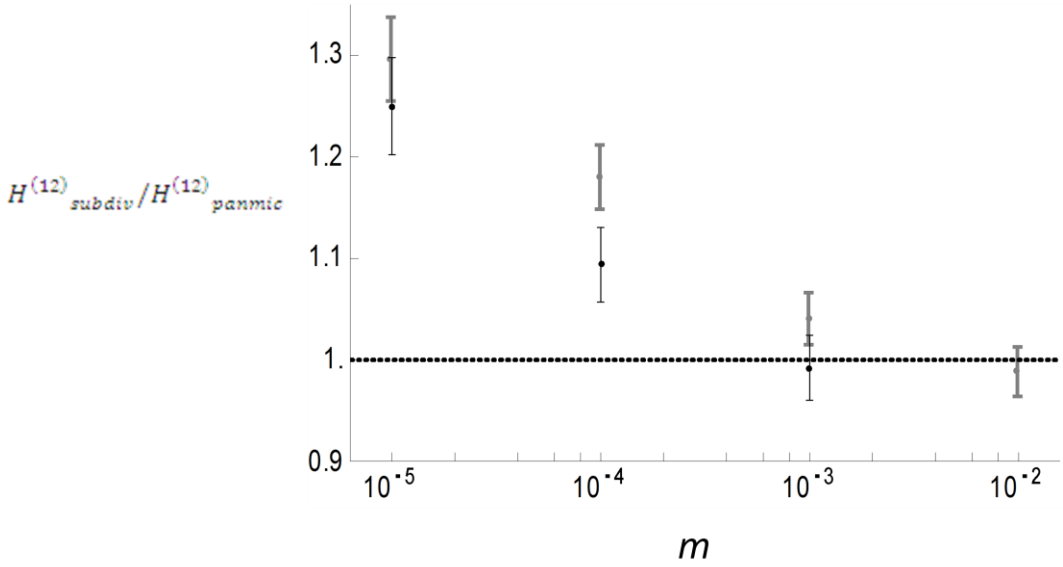


Figure 4.5. The effect of population subdivision on the hitchhiking effect, when two chromosomes are randomly sampled from demes 1 and 2, measured as the average heterozygosity ratio $H^{(12)}_{subdiv}/H^{(12)}_{panmic}$, as a function of the migration rate m with parameter values $K = 2$, $2N = 10^5$, $r/s = 0.01$, $s = 0.01$ (dark) or 0.1 (gray) (reproduced from Kim and Maruki 2011) . Average ± 2 standard error of $H^{(12)}_{subdiv}/H^{(12)}_{panmic}$ are shown for each m.

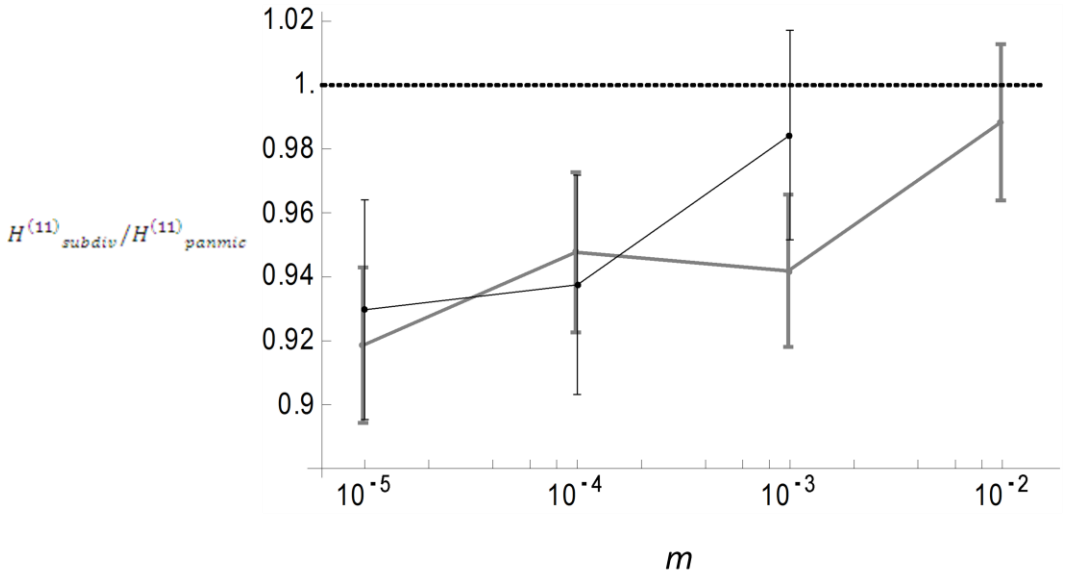


Figure 4.6. The effect of population subdivision on the hitchhiking effect, when two chromosomes are randomly sampled from deme 1, measured as the average heterozygosity ratio $H^{(11)}_{subdiv}/H^{(11)}_{panmic}$, as a function of the migration rate m with parameter values $K = 2$, $2N = 10^5$, $r/s = 0.01$, $s = 0.01$ (dark) or 0.1 (gray) (reproduced from Kim and Maruki 2011). Average ± 2 standard error of $H^{(11)}_{subdiv}/H^{(11)}_{panmic}$ are shown for each m.

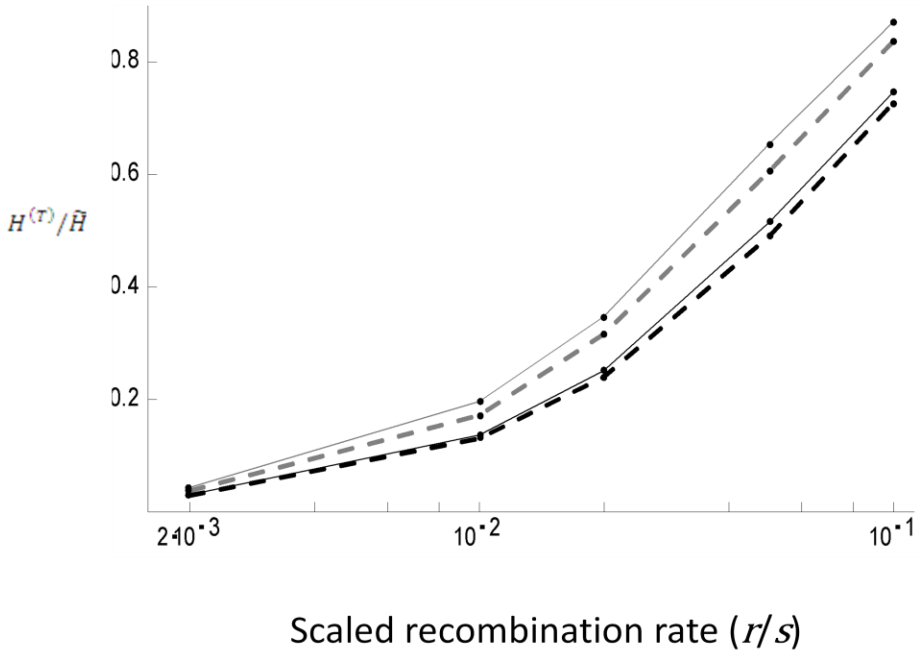


Figure 4.7. The effect of genetic hitchhiking on heterozygosity, when two chromosomes are randomly sampled from the total population, measured as the average heterozygosity ratio $H^{(T)}/\tilde{H}$, as a function of the scaled recombination rate r/s with parameter values $K = 2$, $2N = 10^5$, $m/s = 0.01$, $s = 0.01$ (dark) or 0.1 (gray) (reproduced from Kim and Maruki 2011). Continuous and dashed curves show the hitchhiking effect in the subdivided population and that in a corresponding panmictic population of effective size $4N$, respectively.

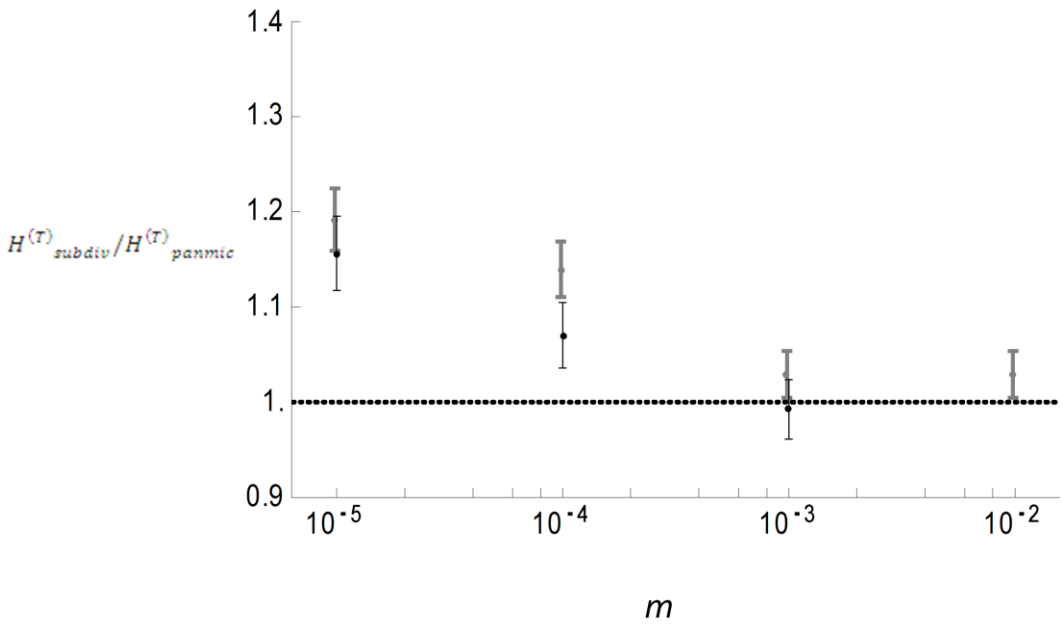


Figure 4.8. The effect of population subdivision on the hitchhiking effect, when two chromosomes are randomly sampled from the total population, measured as the average heterozygosity ratio $H^{(T)}{}_{subdiv}/H^{(T)}{}_{panmic}$, as a function of the migration rate m with parameter values $K = 2$, $2N = 10^5$, $r/s = 0.01$, $s = 0.01$ (dark) or 0.1 (gray) (reproduced from Kim and Maruki 2011). Average ± 2 standard error of $H^{(11)}{}_{subdiv}/H^{(11)}{}_{panmic}$ are shown for each m.

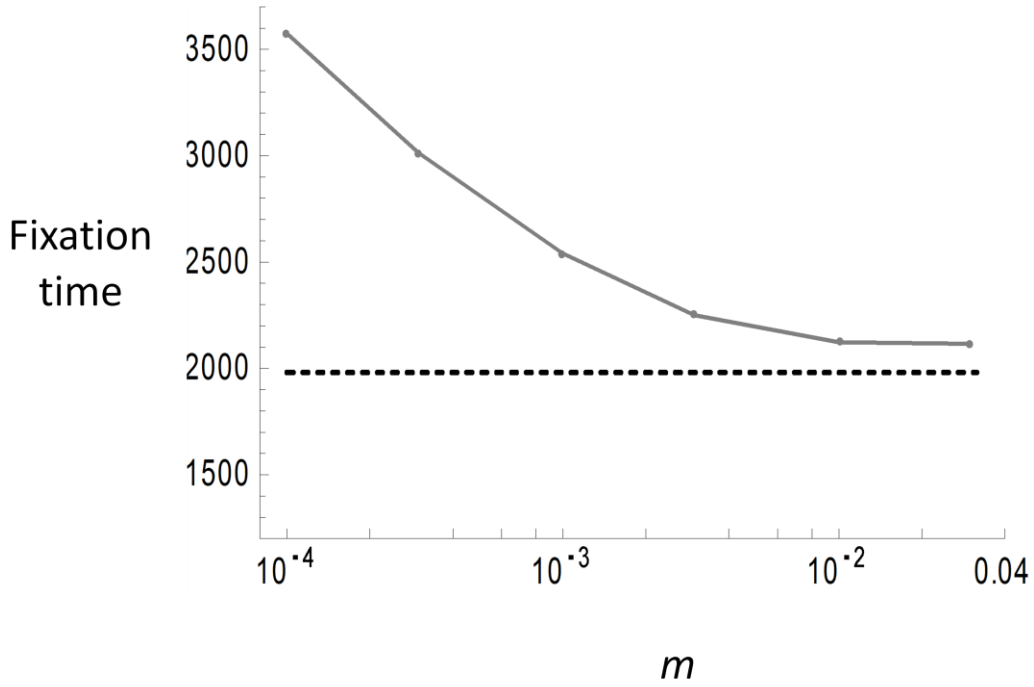


Figure 4.9. The average time taken for the beneficial allele to be fixed in the total population as a function of the migration rate with parameter values $K = 10$, $2N = 10^5$, and $s = 0.01$ (reproduced from Kim and Maruki 2011). The expected fixation time in a corresponding panmictic population, $2 \log(4NKs)/s$, is shown by the dashed line.

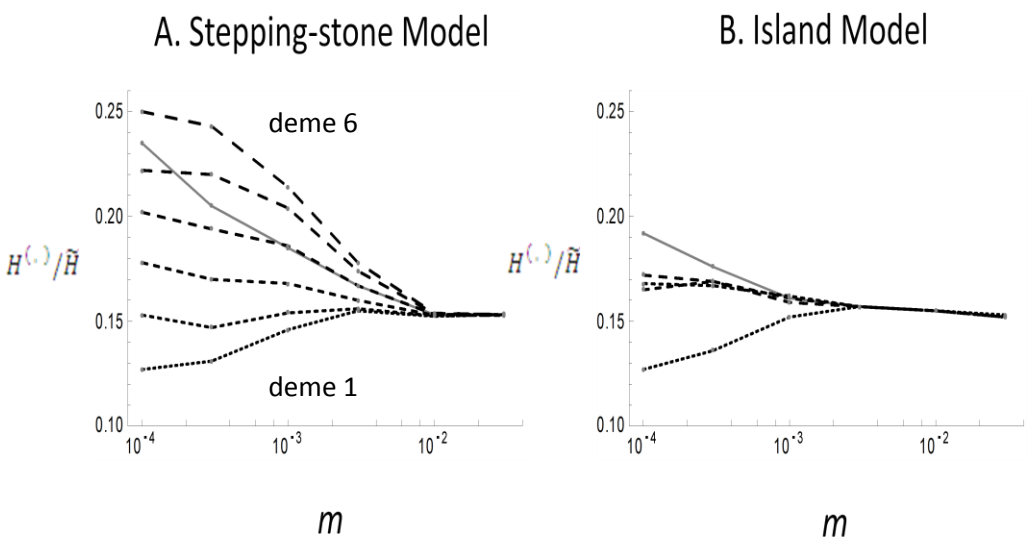


Figure 4.10. The hitchhiking effect in each deme (from deme 1 to deme 6 shown by curves with increasing dash sizes) and total population (gray curve), measured as the average heterozygosity ratio $H^{(\cdot)}/\tilde{H}$, as a function of the migration rate *m*, in the stepping-stone model (A) and island model (B) with parameter values K = 10, $2N = 10^5$, s = 0.01(identical to those in Figure 4.9), and r/s = 0.01 (reproduced from Kim and Maruki 2011). Note that deme 6 is the farthest from deme 1, where the beneficial mutation occurs, in the stepping-stone model.

CHAPTER 5: THE STOCHASTIC DYNAMICS OF A HITCHHIKING ALLELE

Abstract

Conditioning a Wright-Fisher model on the fixation of the beneficial allele has a large impact on the dynamics of the alleles. In this research, I investigate the effect of the conditioning on the dynamics of alleles at a neutral locus linked to the selected locus. The conditioned process is explicitly derived in the diffusion process of a selective sweep in the Wright-Fisher model. The stochastic dynamics of a hitchhiking allele are examined in the conditioned diffusion process and some interesting aspects of the process including 'reverse hitchhiking' are revealed.

Introduction

When a beneficial allele rapidly increases in frequency during the process of a selective sweep, neutral or deleterious alleles linked to the beneficial allele also increase in frequency by genetic hitchhiking (Maynard Smith and Haigh 1974). Unless recombination occurs between selected and linked loci, the fixation of the genetic background on which the beneficial mutation occurred results after a selective sweep. Development of mathematical models of genetic hitchhiking has been motivated by two main reasons. First, because polymorphism surrounding targets of positive directional selection tends to be low, scientists can identify potential targets of positive directional selection (reviewed in Thornton et

al. 2007) and estimate the strength and timing of the sweeps (e.g. Kim and Stephan 2002, Przeworski 2003). Second, depending on the relative frequencies of selective sweeps and recombination, genetic hitchhiking could be an important force shaping the genomic pattern of polymorphism within species (e.g. Gillespie 2000, Begun et al. 2007). For both purposes, in-depth quantitative descriptions of footprints of genetic hitchhiking are necessary to correctly interpret empirical data of nucleotide sequences and a number of mathematical models have been built with these goals in mind (Maynard Smith and Haigh 1974, Ohta and Kimura 1975, Kaplan et al. 1989, Stephan et al. 1992, Braverman et al. 1995, Fay and Wu 2000, Gillespie 2000, Kim and Stephan 2002, Przerorski 2002).

Although the first mathematical model of genetic hitchhiking was deterministic (Maynard Smith and Haigh 1974), subsequent studies have emphasized that stochastic models are needed to describe the footprints of genetic hitchhiking in natural populations (Barton 1998, Przeworski 2003, Durrett and Schweinsberg 2004, Jensen et al. 2005, Pfaffelhuber et al. 2006, Teshima et al. 2006). This is because the dynamics of polymorphism in a finite population subject to random genetic drift can easily deviate from the theoretical predictions of deterministic models. Much of the recent development of mathematical models of genetic hitchhiking has been motivated by the need for distinguishing footprints of selective sweeps from stochastic fluctuations of polymorphism across the genome (e.g. Jensen et al. 2005, Jensen et al. 2007, Pavlidis et al. 2010). The dynamics of a beneficial allele is greatly affected by random genetic drift. For example, several studies have demonstrated that the mean time taken for the

fixation of the beneficial allele is shorter when it is conditioned on fixation in a stochastic model compared to that in a corresponding deterministic model (Barton 1998, Durrett and Schweinsberg 2004, Eriksson et al. 2008). As a result, the effect of genetic hitchhiking is expected to be stronger in stochastic models, because, on average, the opportunities for recombination to break down the association between beneficial and linked alleles decrease.

Because empirical data of nucleotide sequences are usually available only in a sample of small size at present time, most recent studies of genetic hitchhiking have used coalescent processes to study the distribution of polymorphism after a selective sweep. The increasing availability of the empirical data in larger samples at various time points motivates this study to describe the dynamics of hitchhiking alleles in a diffusion process. For example, the genomic data of polymorphism in large samples are increasing in various organisms including humans (e.g., Altshuler et al. 2010). Serially-sampled data sets for rapidly evolving organisms such as HIV (e.g., Drummond et al. 2003) provide opportunities for analyzing the dynamics of a selective sweep as a function of time. In addition, improvements in ancient DNA techniques provide serial genetic data for some taxa with much lower mutation rates including humans and several domesticated species (e.g., Fehren-Schmitz et al. 2011). In this study, I investigate the stochastic dynamics of the hitchhiking allele by conditioning the diffusion process of genetic hitchhiking on the fixation of the beneficial allele. The mathematical technique used for this conditioning is known as Doob's h-transform and has been used by several authors to study the

dynamics of polymorphism at a single locus in diffusion processes (e.g., Kimura and Ohta 1969, Maruyama 1974, Watterson 1977, Griffiths 2003, Pfaffelhuber et al. 2006). The diffusion process of selected and neutral loci in a selective sweep model is conditioned and therefore its effect at the neutral locus is formulated in this study. This makes it possible to efficiently analyze the diffusion process either by simulating the process itself or by numerically solving the associated Kolmogorov forward equation. Furthermore, the dynamics of a hitchhiking deleterious allele can be examined in the framework of the model here.

Materials and Methods

The model

The population in the model is a diploid population of effective size N. There are two loci on the same chromosome, one of which is selected and the other neutral. Recombination occurs at rate c per generation between the two loci. No dominance in either of the alleles is assumed and the relative fitnesses of the three genotypes *BB*, *Bb*, and *bb* are specified as $1 + 2s$, $1 + s$, and 1, respectively. There are two alleles at the neutral locus and let these be denoted when the beneficial mutation occurs at time t = 0. I call the neutral allele initially associated with the beneficial allele the hitchhiking allele *A* although, as shown later, recombination events early in the process can cause the other allele *a* to increase to a higher frequency than *A*. Let q_0 denote the frequency of allele *A* at time zero. Then, the initial frequencies of the four haplotypes *AB*, *Ab*, *aB*, and *ab*

are $1/(2N)$, $q_0 - 1/(2N)$, 0, and $1 - q_0$, respectively. Reproduction occurs according to the Wright-Fisher model.

Conditioning the hitchhiking process on the fixation of the beneficial allele

The diffusion process in the above model is derived and expressed by its infenetesimal generator. Then, the diffusion process is conditioned on the fixation of the beneficial allele by Doob's h-transform.

Simulation of the conditioned diffusion process

The conditioned diffusion process is used for a discrete-time simulation that iterates deterministic change in haplotype frequencies identified in the conditioned diffusion process followed by random change by the step of random sampling that uses a binomial random number generator.

Results

Diffusion process in the model

Let D_i and D_{ij} denote $\frac{\partial}{\partial_{x_i}}$ and $\frac{\partial}{\partial_{x_i}\partial_{x_j}}$, respectively. Then, provided that the population size N is sufficiently large and the time is measured in units of $2N$ generations, the changes in the haplotype frequencies can be approximated by a three dimensional diffusion process with the following infinitesimal generator:

$$Lf = \sum_i D_i b_i(x) f + \frac{1}{2}\sum_{i,j} D_{ij} a_{ij}(x) f, \qquad (5.1)$$

where

$$b(x) = \begin{pmatrix} \alpha x_1\{1-(x_1+x_3)\} - \gamma\{x_1-(x_1+x_2)(x_1+x_3)\} \\ -\alpha x_2(x_1+x_3) + \gamma\{x_1-(x_1+x_2)(x_1+x_3)\} \\ \alpha x_3\{1-(x_1+x_3)\} + \gamma\{x_1-(x_1+x_2)(x_1+x_3)\} \end{pmatrix}, \quad (5.2)$$

$$a(x) = \begin{pmatrix} x_1(1-x_1) & -x_1x_2 & -x_1x_3 \\ -x_1x_2 & x_2(1-x_2) & -x_2x_3 \\ -x_1x_3 & -x_2x_3 & x_3(1-x_3) \end{pmatrix}. \quad (5.3)$$

$b(x)$ and $a(x)$ are the infinitesimal drift vector and the diffusion matrix, respectively, and $\alpha = 2Ns$ and $\gamma = 2Nc$ are the scaled selection coefficient and recombination rate, respectively.

Conditioning the diffusion process on the fixation of the beneficial allele

The diffusion process conditioned on the fixation of the beneficial allele is derived by applying Doob's h-transform to the process shown above. Let $h(x_1, x_3)$ be the fixation probability of the beneficial allele *B* starting from the frequency $x_1 + x_3$. Because the fixation probability of the beneficial allele does not depend on the frequencies of the alleles at the neutral locus, $h(x_1, x_3)$ can be calculated using the theory of one-dimensional diffusion processes (Durrett 2008) and is given by the following equation:

$$h(x_1, x_3) = \frac{1-e^{-2\alpha(x_1+x_3)}}{1-e^{-2\alpha}} \quad (5.4)$$

Then, the infinitesimal generator of the diffusion process conditioned on the fixation of *B* is obtained from the h-transform via the following equation:

$$L^h f = \frac{1}{h(x_1, x_3)} L(hf) \quad (5.5)$$

By substituting equations 5.1 and 5.4 into equation 5.5 and then simplifying, the drift vector $b^h(x)$ and diffusion matrix $a^h(x)$ of the conditioned process are identified as follows:

$$b^h(x) =$$

$$\begin{pmatrix} \alpha x_1\{1-(x_1+x_3)\}\coth\{\alpha(x_1+x_3)\} - \gamma\{x_1-(x_1+x_2)(x_1+x_3)\} \\ -\alpha x_2(x_1+x_3)\coth\{\alpha(x_1+x_3)\} + \gamma\{x_1-(x_1+x_2)(x_1+x_3)\} \\ \alpha x_3\{1-(x_1+x_3)\}\coth\{\alpha(x_1+x_3)\} + \gamma\{x_1-(x_1+x_2)(x_1+x_3)\} \end{pmatrix}, \quad (5.6)$$

$$a^h(x) = a(x)\ . \quad (5.7)$$

Thus the conditioned diffusion process is found from the original process by substituting the scaled selection coefficient α by the frequency-dependent term, $\coth\{\alpha(x_1+x_3)\}$, whenever it appears. By using Ito's change-of-variables formula, the generator of the conditioned process can be also expressed in terms of marginal frequencies of alleles *A* and *B*, x_A and x_B, and the linkage disequilibrium coefficient between the selected and neutral loci, $D = x_{AB} - x_A x_B$, as follows:

$$b^h(x,t) = \begin{pmatrix} \alpha D\coth(\alpha x_B) \\ \alpha x_B(1-x_B)\coth(\alpha x_B) \\ D\{\alpha(1-2x_B)\coth(\alpha x_B) - (\gamma+1)\} \end{pmatrix}, \quad (5.8)$$

$$a^h(x,t) = \begin{pmatrix} x_A(1-x_A) & D & D(1-2x_A) \\ D & x_B(1-x_B) & D(1-2x_B) \\ D(1-2x_A) & D(1-2x_B) & F \end{pmatrix}, \quad (5.9)$$

where $F = x_B x_A(1-x_A)(1-x_B) + D(1-2x_A)(1-2x_B) - D^2$(Innan 2003).

Simulation of the conditioned diffusion process

One of the advantages of the conditioned diffusion process defined by equations 5.6 and 5.7 is that it enables efficient simulations of genetic hitchhiking. If the unconditioned process defined by equations 5.2 and 5.3 is used to conduct the simulations, we need to discard those frequency paths where the beneficial allele is lost. In contrast, the beneficial allele is guaranteed to be fixed in the conditioned process. Table 1 shows the ratio of the average time taken to simulate 10,000 selective sweeps using the unconditioned process to that using the conditioned process. As expected, the efficiency of the conditioned process increases when the fixation probability of the beneficial allele decreases, when selection is weaker or the initial frequency of the beneficial allele is lower. For example, the conditioned process is more than four times more efficient than the original process when $\alpha = 2$, $\gamma = 5$, and $2N = 2 \cdot 10^5$.

Figure 5.1A shows five examples of frequency paths of the hitchhiking allele *A* with parameter values $2N = 2 \cdot 10^6$, $\alpha = 200$, $\gamma = 5$, and $q_0 = 0.1$. In most cases, the frequency of allele *A* rapidly increases through time as a result of genetic hitchhiking. However, occasionally, the frequency of allele *A* first increases and then decreases to a value smaller than q_0. This may occur if recombination generates a *Ba* gamete early in the process of a selective sweep. I call this event reverse hitchhiking. Figure 5.1B shows paths of the linkage disequilibrium coefficient *D* for the same replicates of the process as those in Figure 5.1A. There is a strong correlation between the change of the frequency of allele *A* and that of *D* when $2N$ is large. In particular, when reverse hitchhiking occurs, *D* becomes

negative throughout most of the process. Figure 5.2 shows estimates of the probability of reverse hitchhiking obtained by running 10^5 simulation replicates of the conditioned process. In these simulation replicates, the initial frequency of the hitchhiking allele was specified such that the probability that an allele becomes the hitchhiking allele is proportional to its frequency given by the stationary distribution under mutation-drift equilibrium. As expected, reverse hitchhiking does not occur frequently and its probability increases when the selective advantage of the beneficial allele decreases or the recombination rate between the selected and neutral loci increases.

Discussion

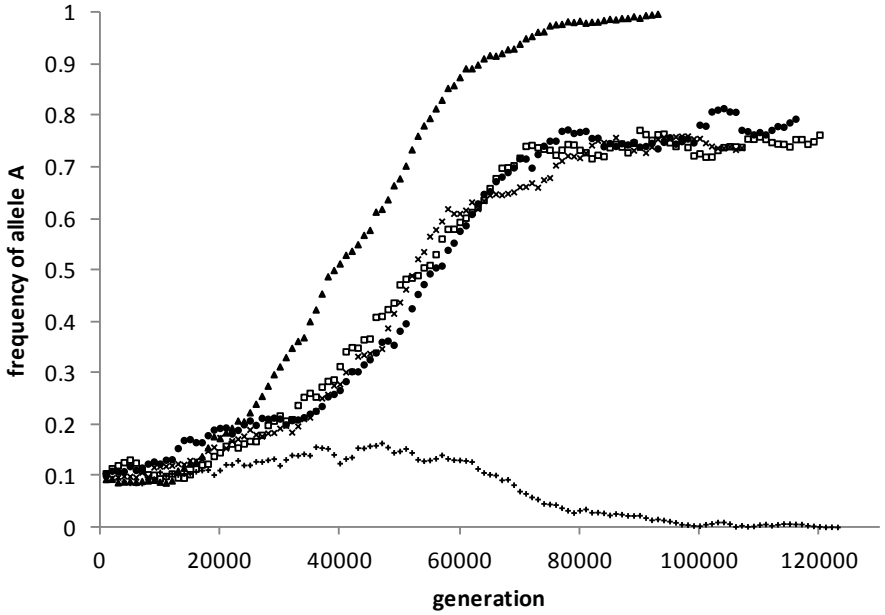
The use of conditioned diffusion processes to model the effect of genetic hitchhiking at linked loci could be extended in several directions. First, the diffusion process of a selective sweep with multiple loci linked to the positively selected locus can be conditioned by Doob's h-transform provided that the linked loci are neutral. The conditioned diffusion process of such models is heuristically expected to differ from the unconditioned process in the same way as that in the two-locus model shown in this study: every scaled selection coefficient α in the original diffusion process is replaced by a frequency dependent term $\alpha\coth(\alpha x_B)$, where x_B is the frequency of the beneficial allele *B*, in the conditioned diffusion process. This is because the multi-dimensional process is conditioned on the fixation of the beneficial allele at the selected locus only and the effect of the conditioning on the process should be the same as long as linked loci are neutral.

It may also be possible to use conditioned diffusions to study the effect of genetic hitchhiking on a linked weakly deleterious allele. Several empirical studies have found potential examples of genetic hitchhiking of a deleterious allele (e.g., Bachtrog 2004, Williamson et al. 2007, Chun and Fay 2011). The stochastic dynamics of the deleterious hitchhiking allele can be naturally studied in the framework of this model, at least by simulation, by introducing purifying selection at the linked locus. How the footprints of genetic hitchhiking are modulated by the existence of purifying selection is an important topic that needs to be addressed, given the accumulating molecular evidence of widespread existence of purifying selection in the genome of various organisms, including humans.

$2N$	α	γ	Ratio of the average time taken
200,000	200	5	0.75
200,000	20	5	0.57
200,000	2	5	0.23
20,000	200	5	0.84
20,000	20	5	0.65
20,000	2	5	0.28
2,000	200	5	0.91
2,000	20	5	0.76
2,000	2	5	0.39
200	200	5	1.00
200	20	5	0.88
200	2	5	0.53

Table 5.1. The ratio of the average simulation time taken with the conditioned process to that with the original process. A total of 10^4 simulation replicates are run to calculate the average time.

A. Frequency of the hitchhiking allele *A*



B. Linkage disequilibrium coefficient *D*

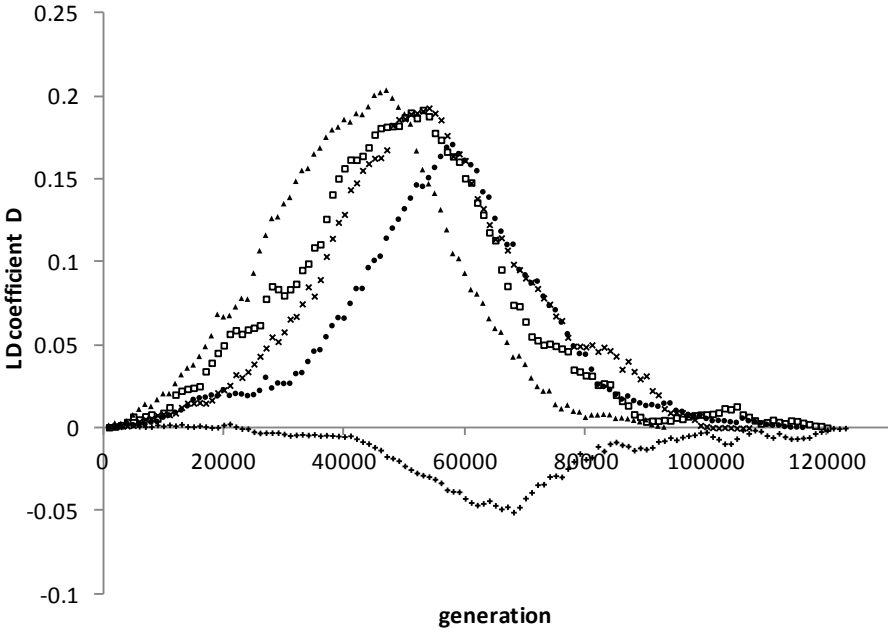


Figure 5.1. Sample paths of the frequency of the hitchhiking allele *A* (A) and linkage disequilibrium coefficient *D* (B) in five replicates of the simulation with parameter values $2N = 2 \cdot 10^6$, $\alpha = 200$, $\gamma = 5$, and $q_0 = 0.1$.

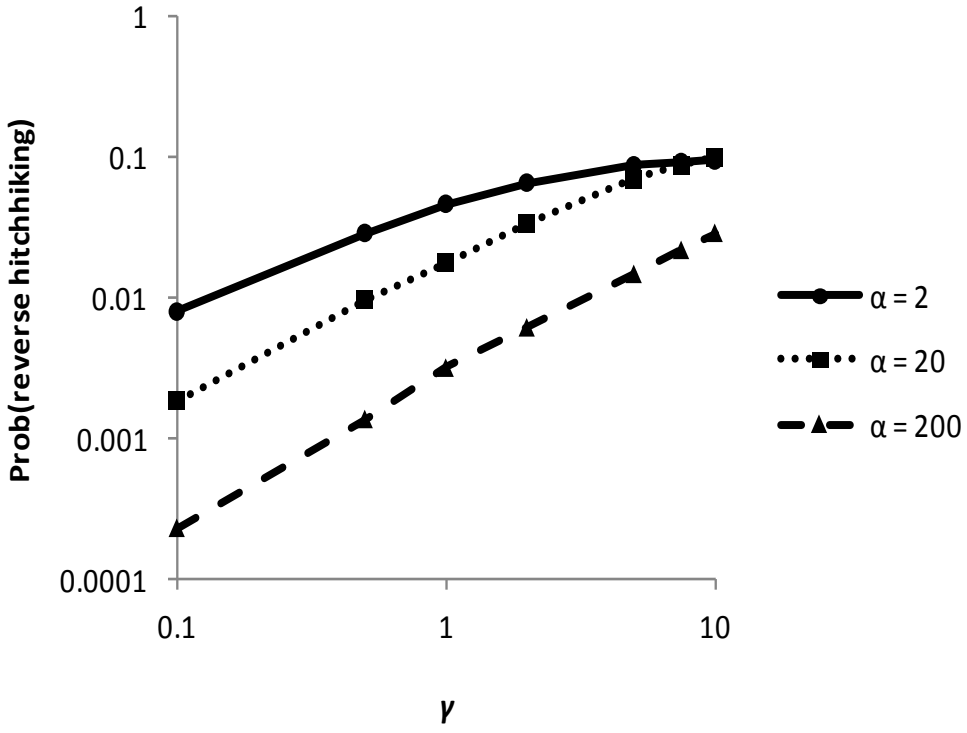


Figure 5.2. Probability of reverse hitchhiking as a function of α and γ, where α and γ are the scaled selection coefficient and recombination rate, respectively, estimated by the simulation. The logarithmic scale with base ten is used on the axes. Reverse hitchhiking is defined to be the case where the frequency of the hitchhiking allele after a selective sweep, q_T, is smaller than its initial value, q_0. The stationary distribution under mutation-drift equilibrium is applied to q_0. $2N = 2 \cdot 10^4$, $\alpha = 2, 20, 200$, $\gamma = 0.1, 0.5, 1, 2, 5, 7.5, 10$ are used. A total of 10^5 simulation replicates are run for each set of parameter values to estimate the probability.

LITERATURE CITED

Akey, J. M. 2009. Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Research.* **19:** 711-722.

Akey, J. M., G. Zhan, *et al*. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* **12**: 1805-1814.

Akey, J. M. *et al*. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biology*. **2**: 1591-1599.

Altshuler, D.L. D. *et al*. 2005. A haplotype map of the human genome. *Nature*. **437**: 1299-1320.

Altshuler, D. L., R. M. Durbin, *et al*. 2010. A map of human genome variation from population-scale sequencing. *Nature*. **467**: 1061-1073.

Amato, R. *et al*. 2009. Genome-wide scan for signatures of human population differentiation and their relationship with natural selection, functional pathways and diseases. *PLoS One* **4**: e7927.

Bachtrog, D. 2004. Evidence that positive selection drives Y-chromosome degeneration in *Drosophila miranda*. *Nature Genetics*. **36**: 518-522.

Baines, J. F., A. Das, S. Mousset, and W. Stephan. 2004. The role of natural selection in genetic differentiation of worldwide populations of *Drosophila ananassae*. *Genetics.* **168**: 1987-1998.

Barton, N. H. 1998. The effect of hitch-hiking on neutral genealogies. *Genetics Research*. **72:** 123-133.

Barton, N. H. 2000. Genetic hitchhiking. *Philosophical Transactions of the Royal Society B: Biological Sciences.* **355:** 1553.

Barreiro, L. B., G. Laval, *et al*. 2008. Natural selection has driven population differentiation in modern humans. *Nature Genetics*. **40**: 340-345.

Beaumont, M. A. 2005. Adaptation and speciation: what can F_{ST} tell us? *Trends Ecol. Evol*. **20**: 435-440.

Begun, D. J. *et al.* 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans. PLoS Biol.* **5:** e310.

Bierne, N. 2010. The distinctive footprints of local hitchhiking in a varied environment and global hitchhiking in a subdivided population. *Evolution.* **64:** 3254-3272.

Boyko, A. R. *et al*. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* **4**: e1000083.

Braverman, J. M. *et al.* 1995. The hitchhiking effect on the site frequency-spectrum of DNA polymorphisms. *Genetics.* **140**: 783-796.

Bustamante, C. D. *et al*. 2005. Natural selection on protein-coding genes in the human genome. *Nature.* **437**: 1153-1157.

Carlson, C. S. *et al*. 2005. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res*. **15**: 1553-1565.

Charlesworth, B. 1998. Measure of divergence between populations and the effect of forces that reduce variability. *Mol. Biol. Evol.* **15**: 538-543.

Charruau, P., C. *et al*. 2011. Phylogeography, genetic structure and population divergence time of cheetahs in Africa and Asia: evidence for long-term geographic isolates. *Molecular Ecology*. **20**: 706-724.

Chun, S. and Fay, J. C. 2011. Evidence for hitchhiking of deleterious mutations within the human genome. *PLoS Genet.* **7**: e1002240.

Clark, A. G., M. J. Hubisz, *et al*. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res*. **15**: 1496-502.

Clark, R. M., *et al*. 2007. Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science*. **317**: 338-342.

Cox, M. P., A. E. Woerner, J. D. Wall, and M. F. Hammer. 2008. Intergenic DNA sequences from the human X chromosome reveal high rates of global gene flow. *BMC Genetics.* **9**: 76.

Das, A., Mohanty, S., and W. Stephan. 2004. Inferring population structure and demography of *Drosophila ananassae* from multilocus data. *Genetics.* **168:** 1975-1985.

Dewar, R. C. *et al*. 2011. Predictions of single-polymorphism differentiation between two populations in terms of mutual information. *Molecular Ecology*. **20**: 3156-3166.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*. **450**: 203-218.

Drummond, A. J. *et al.* 2003. Measurably evolving populations. *Trends Ecol. Evol.* **18**: 481-488.

Durrett, R. 2008. Probability models for DNA sequence evolution. Springer-Verlag, New York.

Durrett, R., and J. Schweinsberg. 2004. Approximating selective sweeps. *Theoretical Population Biology*. **66:** 129-138.

Eriksson, A., P. *et al.* 2008. An accurate model for genetic hitchhiking. *Genetics* **178**: 439-451.

Etheridge, A., P. Pfaffelhuber and A. Wakolbinger. 2006. An approximate sampling formula under genetic hitchhiking. *The Annals of Applied Probability.* **16:** 685-729.

Faure, M. F., P. David, F. Bonhomme and N. Bierne. 2008. Genetic hitchhiking in a subdivided population of *Mytilus edulis*. *BMC Evolutionary Biology*. **8:** 164.

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

Fehren-Schmitz, L. *et al*. 2011. Diachronic investigations of mitochondrial and Y - chromosomal genetic markers in pre-Columbian Andean highlanders from south Peru. *Annals of Human Genetics*. **75**: 266-283.

Gillespie, J. H. 1991. The causes of molecular evolution. New York: Oxford University Press.

Gillespie, J. H. 2000. Genetic drift in an infinite population: The pseudohitchhiking model. *Genetics.* **155**: 909-919.

Griffiths, R. C. 2003. The frequency spectrum of a mutation, and its age, in a general diffusion model. *Theoretical Population Biology*. **64:** 241-251.

Grossman, S. R., *et al*. 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*. **327**: 883-886.

Halligan, D. L., *et al*. 2011. Positive and negative selection in murine ultraconserved noncoding elements. *Molecular Biology and Evolution*. **28**: 2651-2660.

Harr, B. *et al*. 2002. Hitchhiking mapping: A population-based fine-mapping strategy for adaptive mutations in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA.* **99**: 12949-12954.

Hedrick, P. W. 1999. Perspective: Highly variable loci and their interpretation in evolution and conservation. *Evolution*. **53**: 313-318.

Hedrick, P. W. 2005. A standardized genetic differentiation measure. *Evolution*. **59**: 1633-1638.

Helyar, S. J. *et al*. 2011. Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources*. **11**: 123-136.

Hermisson, J., and P. S. Pennings. 2005. Soft sweeps molecular population genetics of adaptation from standing genetic variation. *Genetics.* **169:** 2335-2352.

Holsinger, K. E., and B. S. Weir. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nat. Rev. Genet.* **10**: 639-650.

Innan, H. 2003. A two-locus gene conversion model with selection and its application to the human RHCE and RHD genes. *Proc. Natl. Acad. Sci. USA.* **100:** 8793-8798.

Izagirre, N., I. Garcia, *et al*. 2006. A scan for signatures of positive selection in candidate loci for skin pigmentation in humans. *Mol. Biol. Evol*. **23**: 1697-706.

Jensen, J. D., Y. Kim, V. B. Dumont, C. F. Aquadro and C. D. Bustamante. 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics.* **170:** 1401-1410.

Jensen, J. D., *et al.* 2007. On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. *Genetics.* **176:** 2371-2379.

Jost, L. 2007. Partitioning diversity into independent alpha and beta components. *Ecology*. **88:** 2427-2439.

Jost, L. 2008. G_{ST} and its relatives do not measure differentiation. *Molecular Ecology*. **17**: 4015-4026.

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

Kelley, J. L. *et al*. 2006. Genomic signatures of positive selection and the limits of outlier approaches. *Genome Res*. **16**: 980-989.

Kim, Y., and D. Gulisija. 2010. Signatures of recent directional selection under different models of population expansion during colonization of new selective environments. *Genetics.* **184:** 571-585.

Kim, Y., and Maruki, T. 2011. Hitchhiking effect of a beneficial mutation spreading in a subdivided population. *Genetics*. **189**: 213-226

Kim, Y., and R. Nielsen. 2004. Linkage disequilibrium as a signature of selective sweeps. *Genetics.* **167:** 1513-1524.

Kim, Y., and W. Stephan. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics.* **160:** 765-777.

Kim, Y., and T. Wiehe. 2009. Simulation of DNA sequence evolution under models of recent directional selection. *Briefings in Bioinformatics.* **10:** 84-96.

Kimura, M., and T. Ohta. 1969. Average number of generations until fixation of a mutant gene in a finite population. *Genetics*. **61:** 763-771.

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

Kumar, S., M. P. Suleski, *et al*. 2009. Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and damaging personal amino acid mutations. *Genome Res*. **19**: 1562-1569.

Lewontin, R. C. and J. Krakauer. 1973. Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. *Genetics*. **74**: 175-195.

Li, H., and W. Stepham. 2006. Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genet.* **2:** e166.

Lohmueller, K. E., M. M. Mauney, *et al*. 2006. Variants associated with common disease are not unusually differentiated in frequency across populations. *Am. J. Hum. Genet*. **78**: 130-136.

Lohmueller, K. E., A. R. Indap, *et al*. 2008. Proportionally more deleterious genetic variation in European than in African populations. *Nature*. **451**: 994-997.

Lohmueller, K. E. *et al*. 2011. Natural selection affects multiple asapects of genetic variation at putatively neutral sites across the human genome. *PLoS Genetics*. **7**: e1002326.

Long, J. C., and Kittles, R. A. 2003. Human genetic diversity and the nonexistence of biological races. *Hum. Biol.* **75**: 449-471.

Maruyama, T. 1974. Age of an allele in a finite population. *Genetical Research*. **23**: 137-143.

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

McVean, G. 2007. The structure of linkage disequilibrium around a selective sweep. *Genetics.* **175:** 1395-1406.

Meiklejohn, C. D., Y. Kim, D. L. Hartl and J. Parsh. 2004. Identification of a locus under complex positive selection in *Drosophila simulans* by haplotype mapping and composite-likelihood estimation. *Genetics.* **168:** 265-279.

Meirmans, P. G., and P. W. Hedrick. 2011. Assessing population structure: F_{ST} and related measures. *Molecular Ecology Resources*. **11**: 5-18.

Morjan, C. L., and L. H. Rieseberg. 2004. How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. *Molecular Ecology*. **13**: 1341-1356.

Myles, S., D. Davison, *et al*. 2008. Worldwide population differentiation at disease-associated SNPs. *BMC Medical Genomics*. **1**: 22.

Nei, M. 1977. *F*-statistics and analysis of gene diversity in subdivided populations. *Ann. Hum. Genet*. **41**: 225-233.

Nei, M. and R. K. Chesser. 1983. Estimation of fixation indexes and gene diversities. *Ann. Hum. Genet*. **47**: 253-259.

Nielsen, R. 2005. Molecular signatures of natural selection. *Annual Review of Genetics.* **39:** 197-218.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark *et al.* 2005. Genomic scans for selective sweeps using SNP data. *Genome Research.* **15:** 1566-1575.

Nielsen, R. *et al*. 2007. Recent and ongoing selection in the human genome. *Nat. Rev. Genet.* **39**: 197-218.

Nielsen, R. *et al*. 2009. Darwinian and demographic forces affecting human protein coding genes. *Genome Res*. **19**: 838-849.

Norton, H. L., R. A. Kittles, *et al*. 2007. Genetic evidence for the convergent evolution of light skin in Europeans and east Asians. *Mol. Biol. Evol*. **24**: 710-722.

Ohta, T. and M. Kimura. 1975. Effect of selected linked locus on heterozygosity of neutral alleles (Hitch-hiking effect). *Genetical Research*. **25**: 313-326.

Otto, S. P. and Day, T. 2007. A biologist's guide to mathematical modeling in ecology and evolution. Princeton University Press.

Palumbi, S. R. 2003. Population genetics, demographic connectivity, and the design of marine reserves. *Ecological Applications.* **13**: S146-S158.

Pavlidis, P. *et al.* 2010. Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. *Genetics.* **185:** 907-922.

Pearse, D. E., and K. A. Crandall. 2004. Beyond F_{ST}: Analysis of population genetic data for conservation. *Conservation Genetics*. **5**: 585-602.

Pfaffelhuber, P., B. Haubold and A. Wakolbinger. 2006. Approximate genealogies under genetic hitchhiking. *Genetics.* **174:** 1995-2008.

Pfaffelhuber, P., A. Lenhert and W. Stephan. 2008. Linkage disequilibrium under genetic hitchhiking in finite populations. *Genetics.* **179:** 527-537.

Pickrell, J. K., G. Coop, *et al*. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res*. **19**: 826-837.

Przeworski, M. 2002. The signature of positive selection at randomly chosen loci. *Genetics*. **160**: 1179-1189.

Przeworski, M. 2003. Estimating the time since the fixation of a beneficial allele. *Genetics*. **164**: 1667-1676.

Ramachandran, S., *et al*. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. USA.* **102**: 15942-15947.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly *et al.* 2006. Positive natural selection in the human lineage. *Science.* **312:** 1614-1620.

Sabeti, P. C. *et al*. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature*. **449**: 913-919.

Sáez, A. G., A. Tatarenkov, E. Barrio, N. H. Becerra, and F. J. Ayala. 2003. Patterns of DNA sequence polymorphism at Sod vicinities in *Drosophila melanogaster*: unraveling the footprint of a recent selective sweep. *Proc. Natl. Acad. Sci. USA.* **100:** 1793-1798.

Santiago, E., and A. Caballero. 2005. Variation after a selective sweep in a subdivided population. *Genetics.* **169:** 475-483.

Slatkin, M. 1987. Gene flow and the geographic structure of natural populations. *Science.* **236:** 787-792.

Slatkin, M., and T. Wiehe. 1998. Genetic hitch-hiking in a subdivided population. *Genetical Research.* **71:** 155-160.

Seielstad, M. T., E. Minch, *et al*. 1998. Genetic evidence for a higher female migration rate in humans. *Nature Genetics*. **20**: 278-280.

Stephan, W. 2010. Detecting strong positive selection in the genome. *Molecular Ecology Resources.* **10**: 863-872.

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. *Theoretical Population Biology.* **41:** 237-254.

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

Subramanian, S. and Kumar, S. 2006. Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. *BMC Genomics.* **7**: 306.

Takahata, N. and M. Nei. 1984. F_{ST} and G_{ST} statistics in the finite island model. *Genetics***. 107**: 501-504.

Teshima, K. M. *et al.* 2006. How reliable are empirical genomic scans for selective sweeps? *Genome Research*. **16**: 702-712.

Thornton, K. R., J. D. Jensen, C. Becquet and P. Andolfatto. 2007. Progress and prospects in mapping recent selection in the genome. *Heredity.* **98:** 340-348.

Tishkoff, S. A., and K. K. Kidd. 2004. Implications of biogeography of human populations for 'race' and medicine. *Nature Genetics*. **36**: S21-S27.

Wang, J. L. 2004. Application of the one-migrant-per-generation rule to conservation and management. *Conservation Biology*. **18**: 332-343.

Watterson, G. A. 1977. Reversibility and age of an allele II. Two-allele models, with selection and mutation. *Theoretical Population Biology*. **12:** 179-196.

Weir, B. S. and C. C. Cockerham. 1984. Estimating *F*-statistics for the analysis of population structure. *Evolution.* **38**: 1358-1370.

Weir, B. S., and W. G. Hill. 2002. Estimating *F*-statistics. *Annu. Rev. Genet*. **36**: 721-750.

Whitlock, M. C. 2011. G_{ST}' and *D* do not replace F_{ST}. *Molecular Ecology*. **20**: 1083-1091.

Wilkinson-Herbots, H. M. 1998. Genealogy and subpopulation differentiation under various models of population structure. *J. Math. Biol.* **37**: 535-585.

Williamson, S. H. *et al*. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl. Acad. Sci. USA* **102**: 7882-7887.

Williamson, S. H., M. J. Hubisz, A. G. Clark, B. A. Payseur, C. D. Bustamante *et al.* 2007. Localizing recent adaptive evolution in the human genome. *PLoS Genet*. **3:** e90.

Wright, S. 1940. Breeding structure of populations in relation to speciation. *American Naturalist*. **74:** 232-248.

Wright, S. 1951. The genetical structure of populations. *Annals of Eugenics*. **15**: 323-354.

Xue, Y., A. Daly, B. Yngvadottir, M. Liu, G. Coop *et al.* 2006. Spread of an inactive form of caspase-12 in humans is due to recent positive selection. *Am. J. Hum. Genet*. **78:** 659-670.

Yampolsky, L. Y. 2005. Distribution of the strength of selection against amino acid replacements in human proteins. *Hum. Mol. Genet.* **14**: 3191-3201.