The Role of Balancing Selection in Maintenance of Natural Genetic Variation

Kerry Leigh Bubb

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Washington

2006

Program Authorized to Offer Degree:
Department of Genome Sciences

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Kerry Leigh Bubb

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of the Supervisory Committee:

____
Maynard V. Olson

Reading Committee:

____
Maynard V. Olson

____
Philip P. Green

____
Joseph Felsenstein

Date:______________________

In presenting this dissertation in partial fulfillment of the requirements for the doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of the dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to ProQuest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, 1-800-521-0600, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature____________________________

Date________________________________

University of Washington

Abstract

The Role of Balancing Selection in Maintenance of Natural Genetic Variation

Kerry Leigh Bubb

Chair of the Supervisory Committee:
Professor Maynard V. Olson
Department of Genome Sciences

There has been much speculation as to what role balancing selection has played in shaping the genetic diversity within species. We approached this problem in three different, but complementary ways. First, we engaged in a brute-force search in the human genome for regions with a high density of nucleotide polymorphism—a footprint of ancient balancing selection—by first computationally analyzing publicly available data and collecting additional data in candidate regions. We compare our findings with those expected in a neutrally evolving genome and find no indication of the action of long-term balancing selection outside of the well-studied MHC locus. We then examine the efficacy of two standard statistics used to test for selection. To do this, we implemented a program to simulate evolution of an entire population of haplotypes, going forward in time. We found that the statistics examined were limited in their ability to distinguish selection from neutrality, and also limited in the range of times under which a signal was detectable. Finally, we survey current knowledge of systems under balancing selection and synthesize a model which may be useful in identification of additional systems under this type of selection. Our model defines two distinct forms of balancing selection, both of which are heavily influenced by the action of convergent evolution. We conclude that

while balancing selection is rarely stable outside of its highly recurring role in detecting self vs non-self, it is common as a transient solution to acute environmental challenges.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

Tables Number Page

ACKNOWLEDGEMENTS

I wish to express sincere thanks to my thesis advisor, Maynard Olson, for providing inspiration, enthusiasm, and essential perspective for this project—but especially for being so supportive of my working mother status. I'd also like to thank Philip Green for his constant and generous supply of intellectual rigor, and Joseph Felsenstein for patiently and repeatedly explaining the vagaries of Population Genetics to me. Special thanks also to Mary Kuhner, Arnold Kas, Matthew Stephens, Charla Lambert and Eric Smith for many helpful conversations and suggestions, to the staff at the University of Washington Genome Center for generating the majority of the raw data on which this thesis is based, to my friends and colleagues in the Olson and Green labs for making graduate school a pleasant way to spend the past seven years, to Giselle Carcamo, Fatima Dilek and Seala Seidler for making me feel that leaving my children in their wonderful care might not be such a bad thing, and, of course, to my husband Matthai Philipose, not only for being my number one overall ally, but for learning much more about balancing selection than I'm sure he ever intended. This work was supported by the Genetics Training Grant, the Genome Training Grant, the Howard Hughes Medical Institute (PG) and Center of Excellence in Genome Science grant P50 HG02351 from the NIH National Human Genome Research Institute (MVO).

DEDICATION

To my Mother and Mother-in-Law,

whose childcare during my graduate school years
provided me with conclusive evidence for
“The Grandmother Hypothesis”

INTRODUCTION

Imagine the diversity of people you may encounter at a bus stop. Regardless of whether you are in a cosmopolitan city or a rural outpost, some characteristics seem to be consistently diverse among even small groups of people. There may be some men and some women. There may be some right-handed people and some left-handed people. Some people may prefer to wake up early, others to stay up late. Some may be risk-takers and others more conservative. Some may run faster, become more easily addicted to nicotine, or be more susceptible to certain diseases than others. It is natural to wonder whether or not this ubiquitous phenotypic variation is due to variation at a genetic level that is, for some reason, being preserved in the population across evolutionary time.

Natural selection is often thought of in terms of "survival of the fittest." As such, it is a homogenizing force, driving all the individuals of a population, generation after generation, towards some ideal fitness type. This type of selection—directional selection—is indeed probably the main reason that different species have evolved to fill the myriad niches occupied by contemporary organisms. However, within all natural populations, there remains some degree of genetic variation.

Biologists were confronted with this issue when the newly developed protein electrophoresis technology revealed a then unexpectedly high amount of variability in natural populations of humans and *Drosophila pseudoobscura* (HARRIS 1966; HUBBY and LEWONTIN 1966; LEWONTIN and HUBBY 1966). An initial hypothesis was that some sort of balancing selection was maintaining molecular diversity within the natural population. Soon thereafter, the neutral theory of evolution was developed (KIMURA 1968), which offered an explanation for most genetic diversity within natural populations that did not depend on any form of selection. The neutral theory quickly lay to rest suspicions of rampant balancing selection. But the question of exactly how much balancing selection exists, beyond a handful of prominent and largely undisputed cases, remained open. In this thesis, we attempt to address the issue of how common balancing selection is in metazoans.

In Part I, we describe a brute-force search in the human genome for regions containing a classic footprint of balancing selection. When multiple allelic variants of a gene are maintained for extended periods of time, the number of purely neutral nucleotide differences between the variants is expected to increase in proportion to the time span over which the variants have been maintained. The classic example of this effect involves the Major Histocompatibility Complex (MHC). Neutral sequence that is genetically linked to sites under balancing selection at the MHC locus have accumulated roughly 100 times more nucleotide differences are observed than elsewhere in the genome. The human genome is an ideal place to look because of the current abundance of publicly available sequence from multiple individuals. Despite a thorough, though not exhaustive search, we found no evidence that balancing selection has generated additional highly polymorphic loci in the human genome.

In Part II, we explore the limitations of the test statistic we used in Part I, as well as another commonly used test statistic, Tajima’s D, by using these tests on simulated data with known amounts of selection. The lack of programs that simulate data under selection has largely prevented these statistics from being tested. We implemented a “forward-evolving” algorithm that tracks the evolutionary events of a population of haplotypes moving forward in time. As opposed to other reverse-time methods, implementing selection in a forward-evolving program is straightforward. We identify the time points when these statistics are most useful in detecting balancing selection, but also reveal limits in their overall ability to distinguish neutrality from selection.

In Part III, we survey known genetic systems in an attempt to form an overarching picture of balancing selection. We conclude that there seem to be two discrete types of balancing selection—ancient and recent. Ancient forms almost invariably are associated with self-non-self discrimination; recent forms almost invariably involve maintenance of both functional and non-functional haplotypes in response to an acute environmental stress. In both cases, the balancing-selection system has usually evolved multiple times in different species, in the ancient cases, or within species, in the recent cases. We also point out why knowledge of these properties of balancing selection may be useful in clinical and perhaps commercial levels.

We conclude that, in all likelihood, balancing selection is evolutionarily unstable outside of systems related to self-non-self discrimination; however, we acknowledge the possible existence of long-term-balancing selection of single-site mutations that are impossible to distinguish from cases of convergent evolution.

PART I:

NO NEW LOCI UNDER ANCIENT BALANCING SELECTION IN THE HUMAN GENOME

(Originally published in Bubb KL 2006)

Introduction

Genomic regions with high densities of single-nucleotide polymorphisms (SNPs) may arise for one of three reasons: (1) a high mutation rate; (2) diversifying selection; or (3) chance. The latter category simply reflects neutral variation across the genome in the time elapsed since the most recent common ancestor of a genomic segment (i.e. deepest coalescence time). The level of diversity at any locus is often described in terms of the average pairwise divergence, or the probability that two randomly chosen haplotypes differ at a particular base in that locus (θ_π). Population-genetic theory predicts that, on average for neutrally evolving regions, the most dissimilar haplotypes, capturing the full depth of the evolutionary tree, will be about twice as divergent as a random pair of haplotypes (roughly $2\theta_\pi$) (THE CHIMPANZEE SEQUENCING AND ANALYSIS CONSORTIUM, 2000). The variance of each measurement (average pairwise divergence and maximal pairwise divergence) decreases as the size of the locus examined increases. Although there have been no systematic studies of regions of unusually high SNP density in the human genome, a few such regions have been discovered during the analysis of variation in important genes.

The most dramatic examples are in the HLA locus. In some regions of HLA, such as the DQB1-DQA1-DRB1 gene cluster, the sequence divergence between the most dissimilar haplotypes across tens of thousands of base pairs is nearly two orders of magnitude higher than θ_π (RAYMOND *et al.* 2005; STEWART *et al.* 2004). The elevated nucleotide diversity at HLA has been attributed to diversifying selection acting on functionally critical sites, coupled to extensive "hitchhiking" of neutral SNPs (HUGHES and YEAGER 1998; SLATKIN 2000). Similar explanations have been offered for other known regions with high SNP densities, although none is nearly as dramatic as HLA. For example, across the roughly 25 kbp of ABO sequence, the most dissimilar haplotypes are roughly three times as divergent as expected under neutrality ($6.25\theta_\pi$ compared to the neutral expectation of $2\theta_\pi$) (*SeattleSNPs. NHLBI Program for Genomic Applications, SeattleSNPs, Seattle, WA (URL: http://pga.gs.washington.edu*) [October 2005]). At an "ancient" 900 kbp inversion on chromosome 17, spanning many genes, the maximum pairwise difference is only about twice the genome average ($4.2\theta_\pi$) (STEFANSSON *et al.* 2005). This is an unusual case, because as recombination is absent between inverted haplotypes, the depth of the coalescent is the same across the entire region. Across the genome, the average pairwise diversity over tens of kilobase pairs rarely exceeds four times the mean ($4\theta_\pi$)

(NIEHS Environmental Genome Project, University of Washington, Seattle, WA (URL: http://egp.gs.washington.edu) [October 2005]). As a larger fraction of the genome is analyzed in multiple haplotypes, the number of regions that, by chance alone, have ancient coalescence times may overshadow the number of regions that are ‘deep’ due to selection. This preponderance of false positives may be true for many tests of selection based on characteristics of the coalescent tree.

In this study, we undertook a systematic search for segments of the human genome that have high SNP densities. In contrast to previous studies, our search was unbiased by functional considerations. We simply mined the same public databases of shotgun-sequencing reads that led to discovery of most of the SNPs presently in dbSNP. By paying careful attention—through computational filters and additional experimental sampling—to the many sources of false positives, we identified 16 loci that, next to HLA and ABO, now represent some of the highest SNP densities yet described in the human genome across regions of many thousands of base pairs. These regions do not seem to reflect the action of ancient balancing selection, but rather illustrate the significant “noise” level of the variation in neutral coalescent depth. Our results suggest a paucity of regions outside of HLA that are under long-term balancing selection, as detectable by polymorphism level. This result is in agreement with theoretical predictions that single-site balanced polymorphisms would not be detectable due to recombinational decay of the linked neutral region. Our analysis also highlights the various challenges associated with genome-wide scans for genomic segments under balancing selection.

Materials and Methods

Computational Filtering: The SNP Consortium (SACHIDANANDAM *et al.* 2001) reads (release 10) were downloaded from snp.cshl.org and ftp.ncbi.nih.gov. For each read, we required one fasta file, containing the base calls, and one quality file, containing the base-by-base, log-transformed probability that each call is incorrect (EWING and GREEN 1998). In some cases, paired fasta and quality files were downloaded, and in some cases, the chromatogram was downloaded and fasta and quality files were generated in-house using PHRED (EWING and GREEN 1998; EWING *et al.* 1998). The reads were then trimmed for quality such that each read consisted only of a region with >95% Q20 bp. Reads with less than 100 bp fitting this criterion were removed from the pipeline, leaving a total of 4,295,373 reads (see Figure 1.1 for filtering summary).

These trimmed reads were aligned to the human reference sequence (build30) using BLAST (ALTSCHUL *et al.* 1990). We retained those reads that matched the reference for at least 300 bp (3,334,762 reads) and in addition disagreed with the reference at 1% or more of the aligned basepairs (402,580 reads). Reads that contained any bases in repeat elements, as annotated on the University of California at Santa Cruz website (http://hgdownload.cse.ucsc.edu/goldenPath/hg16/bigZips/chromFaMasked.zip), were removed from further analysis, leaving 97,909 reads.

We then used cross_match (which is based on a word-nucleated banded Smith-Waterman algorithm; available at www.phrap.org) to align each read to its local genomic region. In the approximately 60,000 alignments that had a "minscore" of at least 100, we then looked for putative SNPs, defined as discrepancies between the read and the reference sequence in which both bases had a PHRED quality of ≥30 (corresponding to an error probability of < .001). (Note that much of the reference sequence has the artificial quality 99, which means that a human "finisher" manually annotated it as being correct.) To discount the effects of CpG mutational 'hotspots' (BIRD 1980), discrepancies in which one sequence had a CpG and the other a TpG or CpA were counted as 1/10th of a putative SNP, reflecting the roughly 10-fold higher mutation rate at CpGs (HWANG and GREEN 2004). We identified 6,395 alignments that had ≥1% putative SNPs. Primer3 (ROZEN and SKALETSKY 2000) was used on those 6,395 regions to pick primers for PCR amplification of the region; in 4,255 of the cases, primers were successfully picked. To avoid sequencing the same region multiple times, in cases where multiple reads implicated the same region as highly polymorphic, we PCR-amplified only one arbitrarily chosen amplicon per 10 kbp, for a total of 991 regions (see Figure 1.2a; see Supplementary Table 1 for list of primer sequences and expected amplicon lengths).

Experimental Filtering (PCR and Fosmid sequencing): PCR amplification was performed on a set of 10 DNA samples from self-identified African-Americans, obtained from the Coriell repository (NA17031, NA17032, NA17033, NA17034, NA17035, NA17036, NA17037, NA17038, NA17039, NA17040). PCR products amplified from these samples at the 991 test loci were sequenced (see Figure 1.2b). Polyphred (NICKERSON *et al.* 1997) identified 208 regions that had >3 SNPs of rank 3 or less, and the supporting trace data at these loci were manually examined. Loci in the HLA region, as defined by Stewart et al. (STEWART *et al.* 2004), were dropped from analysis at this point. We determined that 80 of the loci actually contained ≥1% SNPs with genotypes in approximate Hardy-Weinberg proportions (thereby indicating that they were not artifacts of co-amplified paralogous sequences), with a minimum of two occurrences of the minor allele for each SNP.

To test whether the high SNP density extended over several kilobase pairs, for each locus with ≥1% SNPs in this sample population (again counting possible CpG mutations as 1/10th SNP), we PCR amplified two flanking regions—one three kbp upstream and one three kbp downstream from the initial read—from 4 individuals from our diversity panel (NA03715, NA11373, NA11589, NA14660), 3 from the original African-American sample (NA17031, NA17038, NA17039) and one from a chimpanzee (NA03646), as shown in Figure 1.2c. Loci at which at least one of the flanking PCR sites had ≥0.7% SNPs, with a minimum of two occurrences of the minor allele per SNP, were considered further (80 loci). Based on visual inspection, we selected one SNP from each originally amplified locus as a "tag SNP" for purposes of defining a pair of diverged haplotypes; the tag SNP was required to be in apparent linkage disequilibrium with other SNPs in both the originally amplified locus and the flanking regions, and in general it was the SNP with the highest minor allele frequency (see asterisks in Figure 1.2). We sequenced additional individuals (NA01814, NA10471, NA10540, NA10543, NA10975, NA10976, NA11321, NA11521, NA13838, NA14663) in order to identify three samples from each of the two haplotypes defined as described above. For each candidate locus, we isolated the six haplotypes from fosmid libraries constructed from the identified individuals, using PCR assays targeted 10 kbp upstream and downstream from the site of the original read to identify fosmids sharing at least 20 kbp of overlapping sequence (RAYMOND *et al.* 2005). Isolated fosmids were then sequenced by standard shotgun-sequencing and finishing methods to an estimated error rate of <10^{-5} (the sequences have been deposited in Genbank, with accession numbers DQ384420-DQ384510).

Allele frequency information for each tag SNP by human subpopulations was obtained from the following samples, obtained from the Coriell depository: African-Americans (NA17101-NA17132), Caucasians (NA17201-NA17232), and Chinese-American (NA17733-NA17747, NA17749, NA17752-NA17759, NA17761-NA17762, NA17764- NA17769).

Calculating within-human divergence and human-chimpanzee divergence: For each region, we plotted a profile of the percent nucleotide difference between each pair of haplotypes in 5-kbp windows with a 100 bp slide (e.g., Figure 1.3). Human-chimpanzee divergence levels were calculated by comparing the human fosmid sequences to the orthologous chimpanzee draft sequence (UCSC build 1 version 1). Because the chimpanzee sequence is not ‘finished’, we used a POLYBAYES-like algorithm to estimate the probability that each discrepancy between a pair of sequences was a true discrepancy rather than a sequencing error (MARTH *et al.* 1999), and calculated the expected number of true discrepancies for each region.

Computer simulation and analysis: We used two neutral evolution models sets to simulate the evolutionary relationships among a sample of chromosomes. The first was a constant-population size, uniform-recombination-rate model implemented in the ms program (HUDSON 2002), referred to below as the “simple model”, which we used to simulate 30 haplotypes, arbitrarily defining the first as the reference haplotype, the second as the read haplotype and taking the remaining 28 haplotypes as the experimental sample. We estimated the genome-average θ (=$4N\mu$, where N is the effective population size and μ is the per-generation mutation rate) to use in these simulations from the number of segregating sites in all African-origin individuals in the 19 genes with the most sequence available [abcb1, app, blm, ctnna1, cyp19a1, ece1, gab1, map2k4, mapk9, mmp16, nos2a, pdlim1, poln, pten, rad18, rev1l, tjp1, tp53bp1, xrcc4] from the NIEHS SNPs project (NIEHS Environmental Genome Project, University of Washington, Seattle, WA (URL: http://egp.gs.washington.edu) [October 2005]). Our value (θ_S = 0.081%) is probably a slight underestimate, because the sequence includes exonic regions. Our value for the genome-average ρ (=4Nr, where r is the recombination rate) of 5×10^{-4} is consistent with current genome-average estimates (PRITCHARD and PRZEWORSKI 2001).

The second simulated data set was generated using a calibrated parameter-rich model. We simulated 38 haplotypes (22 African-American, six European, five Asian and five African haplotypes) using the cosi_package, maintaining all parameter values as described in the "bestfit" model (SCHAFFNER *et al.* 2005). We defined the European haplotype 1 as the reference and chose the read haplotype at random from European haplotype 4-6, African haplotype 3-5, or Asian haplotype 3-5. The remaining 28 simulated haplotypes correspond to our experimental sample.

In analyzing the simulated data, we attempted to mimic our real-data analysis as closely as possible (described as method (c) in RESULTS). Because some issues (particularly the presence of base-calling errors in the reads and the need to reduce computational burden) were relevant to the real-data analysis but not the simulations, some steps were done in a different order in the simulations. In particular, reads with at least 1% SNP differences relative to the reference could be identified in a single (late) step in the simulations, but required several steps in the real-data analysis, including an early step to identify reads with ≥1% discrepancies (including basecalling errors in addition to SNP differences) relative to the reference, as well as later steps involving experimental confirmation of the SNPs. As a result of this reordering of intermediate steps, some intermediate numbers from the simulations in Figure 1.1 are not directly comparable to the real-data results, but the bottom-line values (number of regions expected to be identified by our approach after applying all computational and experimental filters) should be.

In the simulations, the filtering process can be divided into two phases, consisting of the steps that do or do not enrich for highly polymorphic regions. First, we determined the number of reads that would have been expected to survive our data quality filters. We estimated this from the real data results: 24.3% of the reads had no repetitive elements; 62% of the reads had cross_match scores ≥100; we were able to design useful primers for 66.5% of the putatively highly polymorphic regions; 80% of our PCR amplifications produced useable sequence data; 50% of products were ≥300 bp (we subsequently required ≥3 SNPs / 300 bp). Multiplying all of these factors by the initial 3,334,762 trimmed reads with at least 300 bp aligned to the reference sequence, leaves 133,642 reads expected to pass the above filters (see Figure 1.1).

Second, we estimated how many highly polymorphic regions are expected among the surviving reads, which were regarded as unbiased with respect to polymorphism content, on the basis of our coalescent simulations (see schematic in Figure 1.2). Because of run-time constraints for the parameter-rich coalescence simulator, we divided the genome into one megabase "chromosomes" and randomly placed 1/3000th of the 133,642 reads on each chromosome. For each simulated read, we counted the number of nucleotide differences between the designated read and reference haplotypes in a read-size window of variable length, drawn from our empirical distribution of trimmed read-lengths (as in Figure 1.2a). We allowed only one highly polymorphic read (arbitrarily chosen) per 10-kbp window, discarding all others in that window. For those simulated reads with ≥1% nucleotide differences, we counted the number of SNPs within 20 African-American haplotypes, in the case of the parameter-rich simulation, or an arbitrary subset of 20 different haplotypes in the simple simulation (as in Figure 1.2b). We required that the minor allele for each of these SNPs be represented at least twice among the 20 haplotypes. For those regions that still had ≥1% SNP density by these criteria, we examined the number of SNPs in windows the same size as that used at the site of the original read, three kbp upstream and downstream from that site in six of the 20 test haplotypes and in the eight remaining unexamined haplotypes, as shown in Figure 1.2c (in the case of the parameter-rich simulation, these were European haplotypes 2-3, African-

American haplotypes 27-28, Asian haplotypes 1-2, and African haplotypes 1-2). If either the upstream or downstream window had ≥0.7% SNPs, again requiring that the minor alleles for each SNP appear at least twice, we scanned a 20-kbp window, centered on the original read position, for the 5-kbp window with the highest number of nucleotide differences between the most dissimilar haplotypes within this last set of haplotypes ("MAXDIV"). The numbers reported in Figure 1.1 are for the parameter-rich simulation, but the numbers for the simple simulation are similar (see Figure 1.7).

Results

Our initial data set consisted of pairwise comparison of approximately 4 million whole-genome-shotgun reads from the SNP Consortium (SACHIDANANDAM *et al.* 2001) to the human reference sequence. The main difficulty when looking for highly divergent regions with data of this type is the high frequency of false positives – indeed the vast majority of regions that initially appear to be highly divergent are false positives. Therefore, we developed an extensive filtering scheme (Figure 1.1) to increase the ratio of true to false positives. The types of false positives we encountered can be broken down into two classes: (i) misalignment of paralogous sequences, (ii) miscalled bases with atypically high quality values (i.e sequencing errors that are difficult to recognize).

We computationally filtered the reads in ways that eliminated most of these and then tested 991 of the most promising cases experimentally on a panel of 10 African-Americans to validate the SNPs in this region (see MATERIALS AND METHODS for details). Because the variance in SNP density due to the Poisson distribution alone is high within read-length-sized windows, which average 400 bp, we required that the highly variable segments extend over a minimum of 3 kbp in an attempt to avoid regions where a fortuitously high number of SNPs is not reflective of divergence time. We implemented this requirement by designing PCR assays 3 kbp upstream and downstream from the original read and resequencing these amplicons from a new panel of humans. This new panel comprised three of the original African-Americans, and four additional individuals from a diversity panel, in order to bias our positives toward those with high-minor-allele frequencies in multiple populations. Figure 1.2 illustrates the three steps of the filtering strategy in which most of the enrichment for highly polymorphic regions occurred. Finally, we sequenced at least 20 kbp of the most dissimilar haplotypes at the 16 non-HLA loci that passed all our filters. Typical results are shown in Figure 1.3, and characteristics of the 16 loci are given in Table 1.

Although maximum pairwise divergence is often well above the genome-average-pairwise divergence (indicated by the lowest dotted line in Figure 1.3;

0.081% for humans of recent African ancestry), and in places approaches that of the human-chimpanzee divergence, corroborative signs of balancing selection were absent from these regions. In no case was a human haplotype more closely related to a sampled chimpanzee haplotype than another of the sampled human haplotypes (i.e. there was no indication of trans-species polymorphism). Moreover, the interspecies divergences do not indicate unusual levels of mutation at any of these loci. Although traditional tests for balancing selection such as Tajima's D (TAJIMA 1989) and HKA (HUDSON *et al.* 1987), were often strongly positive in these regions, these statistics are highly correlated with polymorphism levels; hence, they are not independent tests of neutrality. Annotated genes were generally sparse, with most regions being 10's or 100's of kilobase pairs away from the nearest exon. For nine of the most divergent loci, we tested allele frequencies in three subpopulations and found that the FSTs (WEIR and HILL 2002) for these loci were not unusual as compared to the observed distribution for random sites (AKEY *et al.* 2002). The number of putative highly conserved bases in these regions (SIEPEL *et al.* 2005) was also not significantly different from randomly chosen regions (Table 1.1).

Given the lack of annotated gene features that might comprise targets for balancing selection, we sought to determine whether the levels of divergence observed were within the expectations of a neutral model. We chose as our key statistic the number of nucleotide differences in a 5-kbp window between the most dissimilar haplotypes ("MAXDIV"), in order to capture the amplitude of the polymorphism level over a significant genomic extent in the observed regions. Due to the complexity of the models we wish to consider, the analysis we present is based on coalescent simulations rather than explicit theory. We simulated the evolutionary relationship among a sample of haplotypes, using both a constant-population-size, uniform-recombination-rate model (referred to below as the "simple model"; (HUDSON 2002)) and a calibrated parameter-rich model (SCHAFFNER *et al.* 2005), and then looked for the frequency of occurrence of loci with characteristics similar to those we observed (see MATERIALS AND METHODS).

We analyzed ten genome equivalents of simulated data using each of three methods, illustrated schematically in Figure 1.4. In the first method (a), we made a histogram of MAXDIVs for each non-overlapping 5-kbp window in the genome. This distribution has a mean of $2\theta_\pi \times 5000$, which for our estimated θ_π of 0.081% is roughly 8. In the second method (b), we scanned each non-overlapping 20-kbp region for the 5-kbp window with the largest MAXDIV, as was done at our observed loci. Our last method (c) takes into account the acquisition bias in the choice of regions to examine (see MATERIALS AND METHODS and Figure 1.4 legend for details). We obtain results using both the simple and the parameter-rich simulators. The expected MAXDIV distributions for each of these analysis methods are illustrated in Figure 1.5, with the MAXDIV values for our loci indicated with red asterisks.

Overall, our data appear to be consistent with expectations based on the calibrated, parameter-rich coalescent model (see Figure 1.5). However, it is reassuring to note that the data also agree reasonably well with expectations based on a simple constant-population, uniform-recombination coalescent model; hence, the agreement between neutral theory and experiment is not highly sensitive to the details of the neutral model.

To examine the effect that varying recombination rates would have on the expectations under the simple coalescent model, we repeated the simple model simulations using recombination rates 1/2, 1/4, 1/8, 1/16, and 1/32 that in the initial simulation, and analyzed the simulated data according to method (c). The two main effects of decreasing the recombination rate were (i) lengthening the upper tail of the MAXDIV distribution (Figure 1.6), and (ii) increasing the expected number of regions of high polymorphism (Figure 1.7). The fact that the shapes of the distributions converge by around ρ/8, suggests that the results of our simulation—the expected numbers of highly polymorphic 5-kbp windows in a genome—are robust even to major uncertainties in the recombination parameter. [Note also that in the parameter-rich simulation, 12% of recombinations happened outside of hotspots, corresponding to a background rate of ρ/8.]

Two factors that we did not attempt to accommodate in our simulations, though there is ample evidence that both have had significant effects on the human-genome evolution, are varying mutation rates across the genome and the possibly larger ancestral population size than that estimated from the genome-wide average θ_π. Either of these factors would increase the variance of the distributions shown in Figure 1.5, thus making our findings even less unexpected under a null hypothesis of neutrality. For example, while the average divergence between the human and chimpanzee genome is 1.23%, the standard deviation for 100-kbp windows is in excess of 0.25% and rises rapidly with decreasing window size; this effect is thought mainly to be due to variation in mutation rates across the genome (THE CHIMPANZEE SEQUENCING AND ANALYSIS CONSORTIUM, 2005). We attempted to control for mutation rate by examining local human-chimpanzee divergences, which did tend to be slightly higher than the genome average although only one exceeded 2% (e.g., Table 1).

Ancestral population sizes on a multi-million-year time scale that exceed the usual estimates of about 10,000 would also increase the variance of the distributions shown in Figure 1.5, primarily by extending the upper tail of the distribution. It has been inferred that the size of the population ancestral to both humans and chimps was perhaps five times greater than estimates for the recent effective population size for humans (CHEN and LI 2001; TAKAHATA and SATTA 1997; TAKAHATA *et al.* 1995; WALL 2003). While it is not known when the effective population size of the human lineage shrank, it must have been long enough ago that it does not affect the coalescence time of the vast majority of sites in the genome (MARTH *et al.* 2004; SCHAFFNER *et al.* 2005; VOIGHT *et al.*

2005). If a locus had remained polymorphic since the time of the larger ancestral population size, the total coalescence time for that locus would reflect that larger population size and might be reflected in the few, very deep coalescences occurring fortuitously in the genome.

The HLA locus serves as an important positive control for our genome-wide screen. Regions identified by the computational filter were enriched for the HLA region, as expected. Since the HLA locus, as commonly defined, comprises less than 0.1% of the genome and only a fraction of the locus is highly polymorphic, we expect less than 1 of the 991 computationally identified regions to land within this locus by chance. In contrast, we observed sixteen. Without prior knowledge of HLA annotation, the sixteen HLA 'hits' implicated 11 genes (RFP, HLA-F, DRB3, DRB1, DQA1, DQB1, DQA2, DQB2, DOA, DPA1 and DPB1) as candidates for balancing selection due to elevated polymorphism (Figure 1.8). Most of these genes are in extremely polymorphic HLA sub-regions (STEWART *et al.* 2004).

Discussion

Although the regions we identified are among the most polymorphic regions yet reported in the human genome outside of the HLA and ABO loci, we did not find any evidence that balancing selection had maintained this variation. Our results are consistent with theoretical predictions (WIUF *et al.* 2004) and recent experimental results (ASTHANA *et al.* 2005) both of which suggest that only under very special circumstances can balancing selection persist over millions of years, while also leaving a signature of high local polymorphism on the human genome. Our choice of test statistic (maximum divergence over a 5 kbp interval) limits our ability to detect certain types of balanced haplotypes, such as those that in which the balanced haplotypes have been eroded by recombination to a few nucleotides. If we were looking instead for evidence of ancient admixture with Neanderthal—longer regions of moderately high polymorphism levels—our test statistic would need to reflect these expected properties. In all scans of this type, there is a trade-off between discarding true positives by filtering too aggressively and being overwhelmed with false positives when the filtering is not aggressive enough.

The larger question posed by our data is whether or not the use of DNA polymorphism data to detect targets of long-term balancing selection in the human genome is futile regardless of the details of the methodology employed. Despite the abundance of theoretical and experimental literature on balancing selection, there are only three well documented instances of trans-species balancing selection among eukaryotes: (i) the MHC (HUGHES and YEAGER 1998), (ii) the gene responsible for middle-wavelength and long-wavelength color vision in New World monkeys (SURRIDGE and MUNDY 2002), and (iii) the self-incompatibility (SI) loci in plants, which have arisen multiple times independently (CASTRIC and VEKEMANS 2004). Beyond these examples, one could argue that the sex chromosomes, which have arisen multiple times (FRASER and HEITMAN

2005), are the oldest and most prevalent example of balancing selection. However, the first three examples are more relevant for our study.

There is strong evidence that adaptive immunity, a process in which the MHC is definitively involved, evolved multiple times (LITMAN *et al.* 2005). Hence, like self-incompatibility in plants, the MHC may be another case of convergent evolution of balancing selection. What do these genetic loci have in common? We argue that for balancing selection to be evolutionarily stable it must be frequency-dependent (as opposed to heterotic), and, for it to be detectable, there must be at least two physically linked loci that are each under balancing selection, thereby enabling neutral mutations to accumulate over a substantial region due to selection against most recombination events in the interval (SLATKIN 2000). These conditions seem to have been met rarely during human evolution—in our screen, the HLA region was unique in the density of highly polymorphic "hits." While any type of selection that favors maintenance of more than one allele is, by definition, balancing selection, there are multiple mechanisms through which a balance of alleles can be maintained. The most widely recognized mechanism is heterozygote advantage, as in the textbook example of sickle-cell anemia. Although the sickle-cell allele raises the overall fitness of the population, a significant fraction of individuals have decreased survival and reproductive rates as a consequence of this one allele—a phenomenon that has been described as genetic or segregational load. There are two indications that such systems may not be stable. First, a new allele under balancing selection may rise in frequency more quickly than a new allele under positive selection—even one which, in equilibrium state, confers a greater fitness benefit on the population. This is because when a new allele is at a low frequency, the fitness advantage of the heterozygote is most important, while the lower fitness of homozygotes is not yet very relevant. For example, despite the fact that multiple hemoglobinopathy-related alleles (including the one responsible for sickle-cell anemia) have arisen independently in response to selective pressure by malaria, an allele exists (HbC) that is protective against malaria in the homozygous state and more weakly in the heterozygous state, as well. Neither state is associated with hemoglobinopathy. Given enough time under continued selective pressure, it is expected that this allele would sweep through the at-risk region and increase the total population fitness (MODIANO *et al.* 2001). Second, in general, one can imagine some combination of gene duplication and regulatory modifications that would allow all individuals to have the benefits of both alleles of a gene under balancing selection (SPOFFORD 1969), as is illustrated by the evolution of separate middle-wavelength and long-wavelength-color-vision genes in Old World monkeys and Great Apes. In contrast, frequency-dependent selection does not require a steady-state-fitness differential, and, therefore, confers less load on a population (KOJIMA 1971). Consequently, this type of balancing selection is probably more stable than instances that depend on heterozygote advantage.

If a balancing-selection system is sustainable, but depends on a single SNP, the genomic region in which two haplotypes are preserved will erode, due to recombination, ultimately making it impossible to recreate an accurate coalescent tree with SNP data (WIUF *et al.* 2004). However, if there are two physically linked sites that, in certain combinations, produce balanced haplotypes, the neutral sites between them will reflect the divergence time of the balanced haplotypes. The sites must be nearby, however, to avoid recombinational erosion (BARTON and NAVARRO 2002; KELLY and WADE 2000). Although there is evidence for some degree of genomic clustering of co-expressed genes in mammals at the megabasepair scale, which may largely reflect local chromatin characteristics (HURST *et al.* 2004), we suspect that sites at which recombination is strongly suppressed by selection against recombinant haplotypes are rare.

In each of our three best examples, there is epistasis between physically linked sites. In the form of SI in which compatibility is determined by parental genotype ("sporophytic SI"), the male and female determinants are tightly linked, both physically and phylogenetically (HISCOCK and MCINNIS 2003; SATO *et al.* 2002). In the form of SI in which compatibility is determined by gametic genotype alone ("gametophytic SI"), only the female determinant has been identified. For that locus, there are two hypervariable regions within a single gene that are thought to interact functionally (FRANKLIN-TONG and FRANKLIN 2003). The color vision alleles in New World monkeys also appear to display balancing selection at multiple intragenic sites, as there are multiple mutational differences in different exons that distinguish functional alleles (SHYUE *et al.* 1998). At the MHC locus, there are many associations between "ancestral haplotypes" covering multiple genes and disease susceptibilities (PRICE *et al.* 1999). It appears that, as in the evolution of the sex chromosomes (FRASER and HEITMAN 2005), the MHC locus has acquired functionally related sets of genes whose gene products interact (KELLEY *et al.* 2005).

While there are frequent claims for balancing selection at other loci in the literature, the plausibility of most of these cases depends on scenarios for heterozygote advantage. Thus far, the best case for balancing selection in the human genome based solely on greater-than-expected coalescence time is at the locus controlling ABO blood-type, specifically between the A and B alleles. ABO is an interesting example because, although it has been known to be polymorphic for over 100 years due to its relevance in blood transfusion, its primary evolutionary function remains elusive. The lack of a strongly deleterious genotype satisfies our first proposed criterion that there should be little genetic load. The initial suggestion of long-term balancing selection came from the fact that the AB antigen-antibody phenotype is present in many primates, including some New World monkeys (BLANCHER *et al.* 2000). Furthermore, it has been shown biochemically that only two nucleotides, separated by 6 bp, differentiate the A allele from the B allele (YAMAMOTO and HAKOMORI 1990), and these two nucleotides demonstrate apparent trans-species polymorphism within humans, chimpanzees and gorillas (MARTINKO *et al.* 1993). In contrast, the O allele appears to have arisen multiple times in humans but is rare in non-human primates. When intronic sequence of humans, gorillas and chimpanzees is compared, there is no evidence for trans-species polymorphism of linked neutral sites, so it has been argued that the two functional polymorphisms reflect convergent evolution (O'HUIGIN *et al.* 1997). However, if the balanced haplotype is just 8-bp long, it would behave as a single site and have only modest effects on flanking polymorphism levels (WIUF *et al.* 2004); the 6 exonic nucleotides between the functional polymorphisms certainly cannot hold enough neutral mutation to provide an accurate estimate of divergence time. Indeed, while polymorphism levels are high in the ABO region—with a MAXDIV of 49, which approaches human-chimpanzee divergence levels—there is no evidence for trans-species polymorphism outside the 8-bp haplotype (*SeattleSNPs. NHLBI Program for Genomic Applications, SeattleSNPs, Seattle, WA (URL: http://pga.gs.washington.edu) [October 2005]*). Thus, while we cannot conclude that ABO is another example of trans-species balancing selection, the possibility exists that it is an “invisible” example that cannot be detected by polymorphism studies.

We hypothesize that balancing selection most frequently arises in transient situations when the environment changes rapidly. Balancing-selection systems may largely be evolutionary “band-aids” that survive only until a more stable strategy arises, based on gene duplication and divergence, or the rise of a more evolutionarily successful allele. This view is reminiscent of arguments supporting the less-is-more hypothesis (OLSON 1999); indeed, many suspected examples of recent balancing selection involve maintenance of non-functional or sub-functional alleles in the population (e.g., Δccr5, ΔF508, HbS).

This brief analysis suggests that long-term balancing selection may simply be rare in humans and other organisms with similar biology and evolutionary histories. Certainly, this conclusion is compatible with the results of our search for targets of long-term-balancing selection in the human genome. Nonetheless, the question still arises as to whether or not we failed to identify such targets simply because we had too little data to analyze. Would we have fared better, for example, if the entire genome were sequenced across 20 human haplotypes? While we cannot exclude that possibility, we suspect that identification of genes under long-term-balancing selection will remain a gene-by-gene process, based largely on functional evidence, and not greatly accelerated by genomic analysis because (i) the phenomenon itself is rare, and (ii) compatible balancing selection between physically linked loci—a requirement for generating a detectable genomic fingerprint—is also rare. Nonetheless, the fact that balancing selection systems have arisen multiple independent times and involve core functions of multicellular, sexually reproducing organisms (e.g., combating pathogens and avoiding selfing) suggests that, while rare, balancing selection has had major effects on the evolution of metazoan organisms.

Table 1.1. Characteristics of the sixteen highly divergent loci identified in our screen, along with ABO and HLA for comparison.

				minor allele freq (#haps sampled)[b]						
				Chinese-	Caucasian-	African-		human-chimp	#conserved bp[d]	
locus	chr,band	build35.1 position	MAXDIV[a]	American	American	American	FST	divergence	(p-value)[e]	annotation
bs001	1q42.13	chr1:225,180,594-225,185,593	17	N/A	N/A	N/A	N/A	0.92%	12 (0.68)	Adjacent to 774-bp terminal coding exon of RHOU, which contains no known polymorphisms
bs002	1q43	chr1:234,960,748-234,965,747	41	0.820 (50)	0.688 (64)	0.469 (64)	0.20	1.50%	23 (0.62)	500 kbp from nearest annotated gene
bs003	3p25.3	chr3:11,615,674-11,620,673	39	N/A	N/A	N/A	N/A	2.00%	214 (0.22)	Contains a 172-bp internal coding exon of VGLL4, which has no high-minor-allele-frequency polymorphisms
bs004	5q33.2	chr5:152,911,118-152,916,117	38	0.078 (64)	0.516 (64)	0.145 (62)	0.47	1.18%	39 (0.51)	In the middle of 150-kbp intron of GRIA1
bs005	3q26.31	chr3:176,595,380-176,600,379	19	N/A	N/A	N/A	N/A	1.09%	158 (0.29)	In the middle of 130-kbp intron of NAALADL2
bs006	4q31.22	chr4:147,247,730-147,252,729	22	N/A	N/A	N/A	N/A	0.93%	27 (0.51)	20 kbp downstream of LOC152485
bs007	4q35.2	chr4:189,958,640-189,963,639	49	0.417 (60)	0.766 (64)	0.476 (62)	0.15	1.19%	8 (0.63)	In exon of GENSCAN-predicted gene with no mRNA support and no homology between predicted amino acid sequence and known proteins
bs008	7p21.3	chr7:8,407,241-8,412,240	31	0.233 (60)	0.048 (62)	0.328 (64)	0.60	1.50%	205 (0.21)	200 kbp away from nearest annotated gene
bs009[c]	7q21.13	chr7:88,955,055-88,960,054	63	0.621 (58)	0.741 (58)	0.414 (58)	0.23	1.63%	0 (1)	In the middle of 300-kbp intron of NXPH1
bs010	7q34	chr7:141,399,700-141,404,699	37	0.617 (60)	0.583 (60)	0.435 (62)	0.24	1.30%	129 (0.31)	5 kbp downstream of TRY1, which has no coding known polymorphisms
bs011	7q35	chr7:146,451,078-146,456,077	33	N/A	N/A	N/A	N/A	1.51%	19 (0.60)	In the middle of 100-kbp intron of CNTNAP2
bs012	8p23.2	chr8:3,887,626-3,892,625	45	N/A	N/A	N/A	N/A	4.70%	0 (1)	10 kbp downstream of CSMD1 exon, which has no known polymorphisms
bs013	8q11.21	chr8:50,278,353-50,283,352	47	0.391 (64)	0.468 (62)	0.177 (62)	0.42	1.22%	0 (1)	120 kbp downstream of nearest predicted gene with mRNA support.
bs014	8q12.1	chr8:57,926,636-57,931,635	47	0.567 (60	0.406 (64)	0.133 (60)	0.43	1.68%	73 (0.39)	100 kbp upstream of nearest predicted gene with mRNA support.
bs015	8q22.1	chr8:96,581,055-96,586,054	25	N/A	N/A	N/A	N/A	1.12%	440 (0.06)	300 kbp away from nearest annotated gene
bs016	8q24.23	chr8:138,682,749-138,687,748	48	0.170 (62)	0.161 (62)	0.452 (62)	0.47	1.42%	425 (0.06)	100 kbp upstream of nearest predicted exon with mRNA support
ABO	9q34.2	chr9:133,160,949-133,165,948	49	N/A	N/A	N/A	N/A	1.06%	138 (0.33)	Contains last four exons of ABO
HLA	6p21.32	chr6:32,686,220-32,691,219	409	N/A	N/A	N/A	N/A	0.96%	0 (1)	Midway between DRB1 and DQA1, which are separated by 50 kbp

[a] number of nucleotide differences in this 5 kbp window between the most dissimilar haplotypes tested

[b] allele defined by a single "tag" SNP, chosen as described in the text; the "minor allele" is the allele that is less frequent in the African-American population

[c] resequenced original site, but probe for identifying appropriate fosmids was in 3 kbp upstream region

[d] posterior probability of conservation >0.9 - according to Siepel et al. 2005

[e] probability that a randomly chosen 5 kbp window from the same chromosome would contain at least that number of conserved bp in that window

FIGURE 1.1.—Analysis pipelines used to identify regions of the genome with high polymorphism in real and simulated data. Numbers shown for the simulated pipeline are those generated using a parameter-rich coalescent model (SCHAFFNER *et al.* 2005). Black arrows indicate steps that enrich for highly polymorphic regions, indicated schematically by black boxes following these steps. Gray arrows are accompanied by the percentage of reads that passes through that particular filter. That value is used in the simulated pipeline. As is described more fully in MATERIALS AND METHODS, the order of the filters differed between the real-data and simulated pipelines.

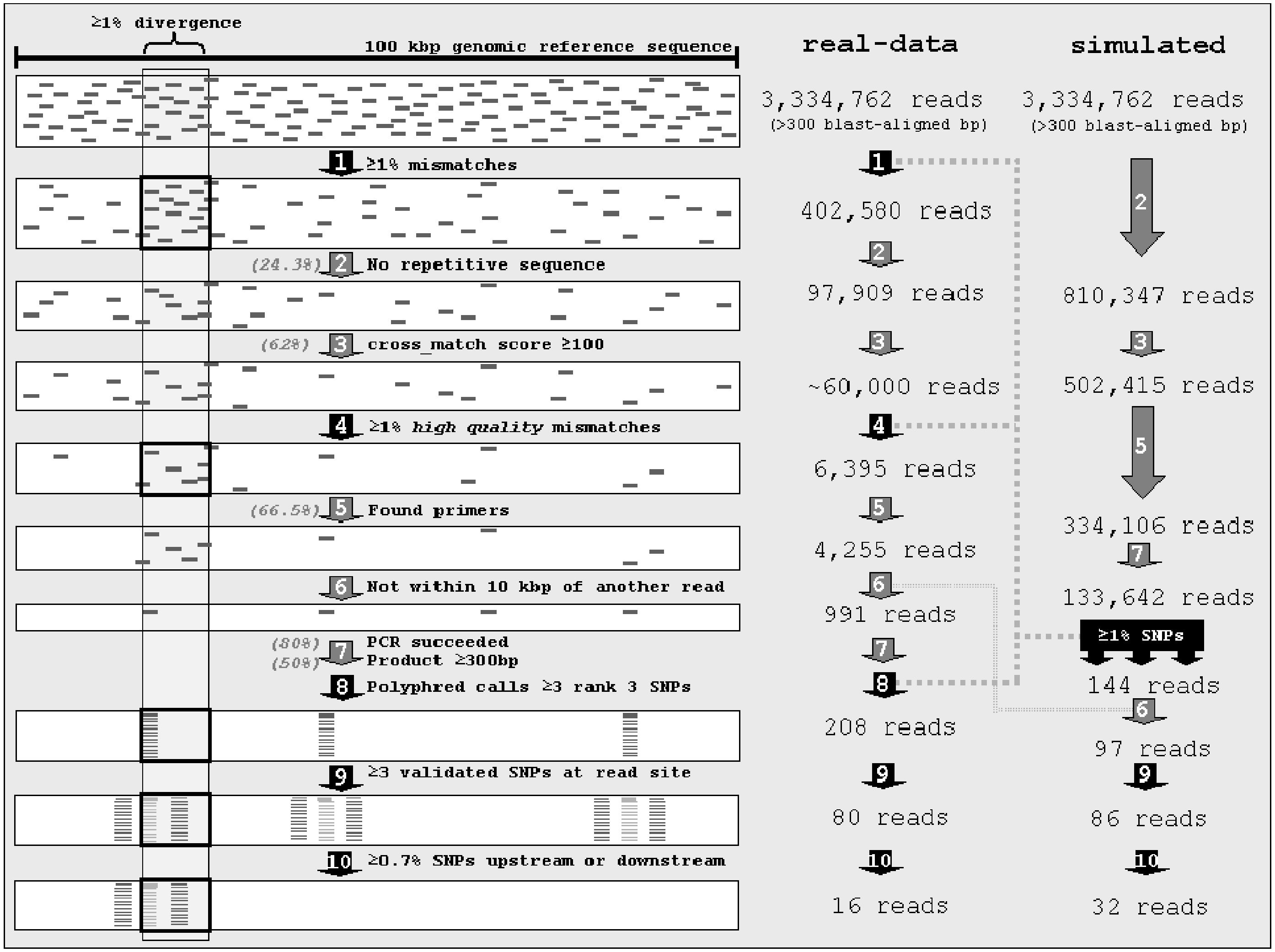
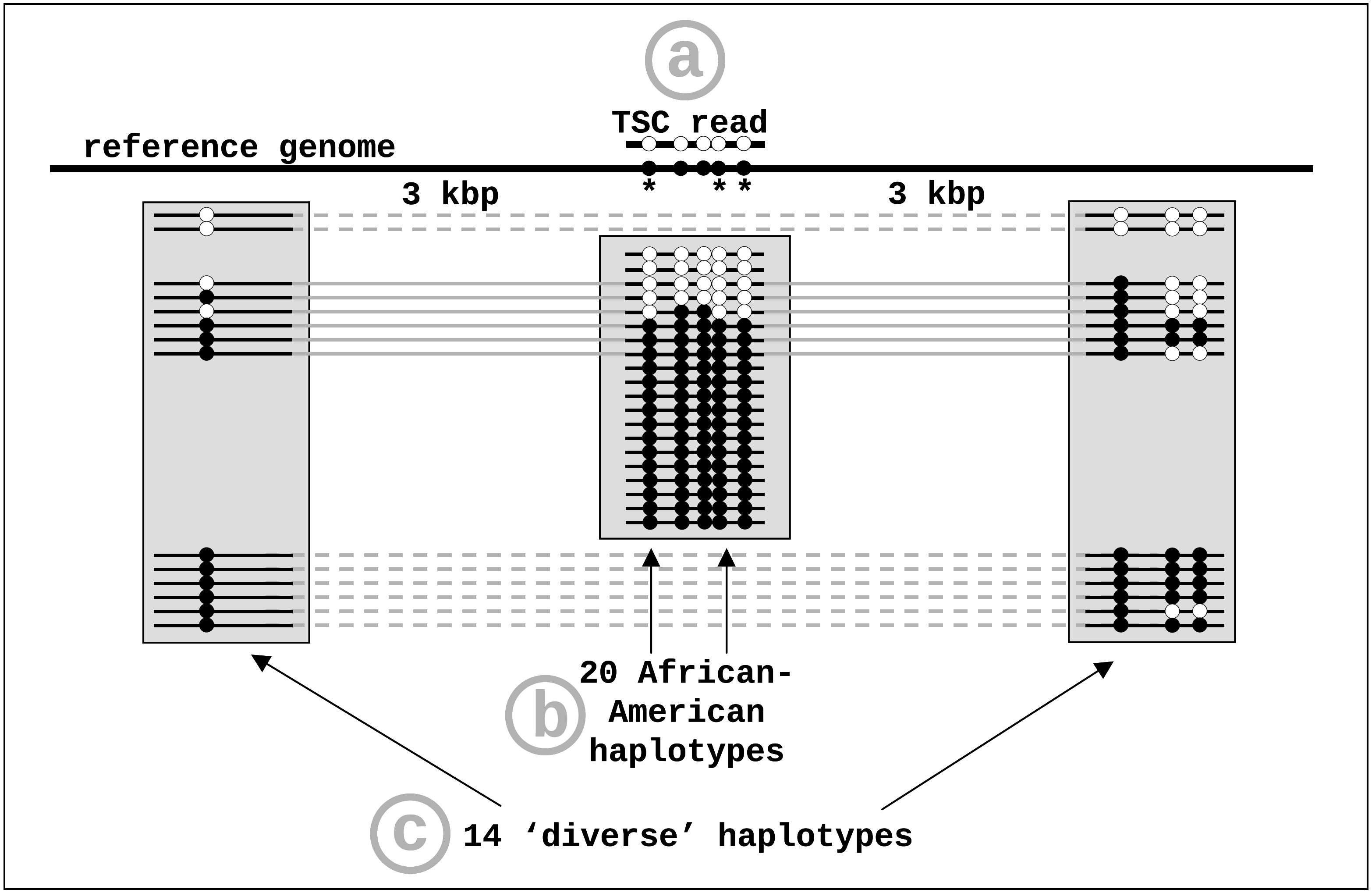

≥1% divergence
100 kbp genomic reference sequence
real-data
simulated
3,334,762 reads
(>300 blast-aligned bp)
3,334,762 reads
(>300 blast-aligned bp)
1 ≥1% mismatches
402,580 reads
2
(24.3%) 2 No repetitive sequence
97,909 reads
810,347 reads
(62%) 3 cross_match score ≥100
~60,000 reads
502,415 reads
4 ≥1% high quality mismatches
6,395 reads
5
(66.5%) 5 Found primers
4,255 reads
334,106 reads
6 Not within 10 kbp of another read
991 reads
7
133,642 reads
≥1% SNPs
(80%) 7 PCR succeeded
(50%) Product ≥300bp
8 Polyphred calls ≥3 rank 3 SNPs
144 reads
6
208 reads
97 reads
9 ≥3 validated SNPs at read site
80 reads
86 reads
10 ≥0.7% SNPs upstream or downstream
16 reads
32 reads

FIGURE 1.2.—Illustration of key filtering techniques used to find extended regions of high polymorphism: (a) original alignment of the SNP Consortium read ("TSC read") to the human reference genome; (b) the SNP-confirmation step in which the region was amplified from the genomic DNA of 10 self-identified African-Americans and resequenced (20 haplotypes); (c) the SNP discovery step 3 kbp upstream and downstream from the original read, based on a panel including three of the previous African-Americans (haplotypes indicated with solid gray lines) and four additional individuals from a diversity panel (haplotypes indicated with dashed gray lines). Note that SNPs were not typed for the four additional individuals at the site of the original read alignment. For each SNP, the major and minor alleles are indicated as black and white balls, respectively. Asterisks indicate potential "tag" SNPs used in the subsequent-fosmid-isolation step.


a
TSC read
reference genome
3 kbp
* * *
3 kbp
20 African-American haplotypes
b
c
14 'diverse' haplotypes

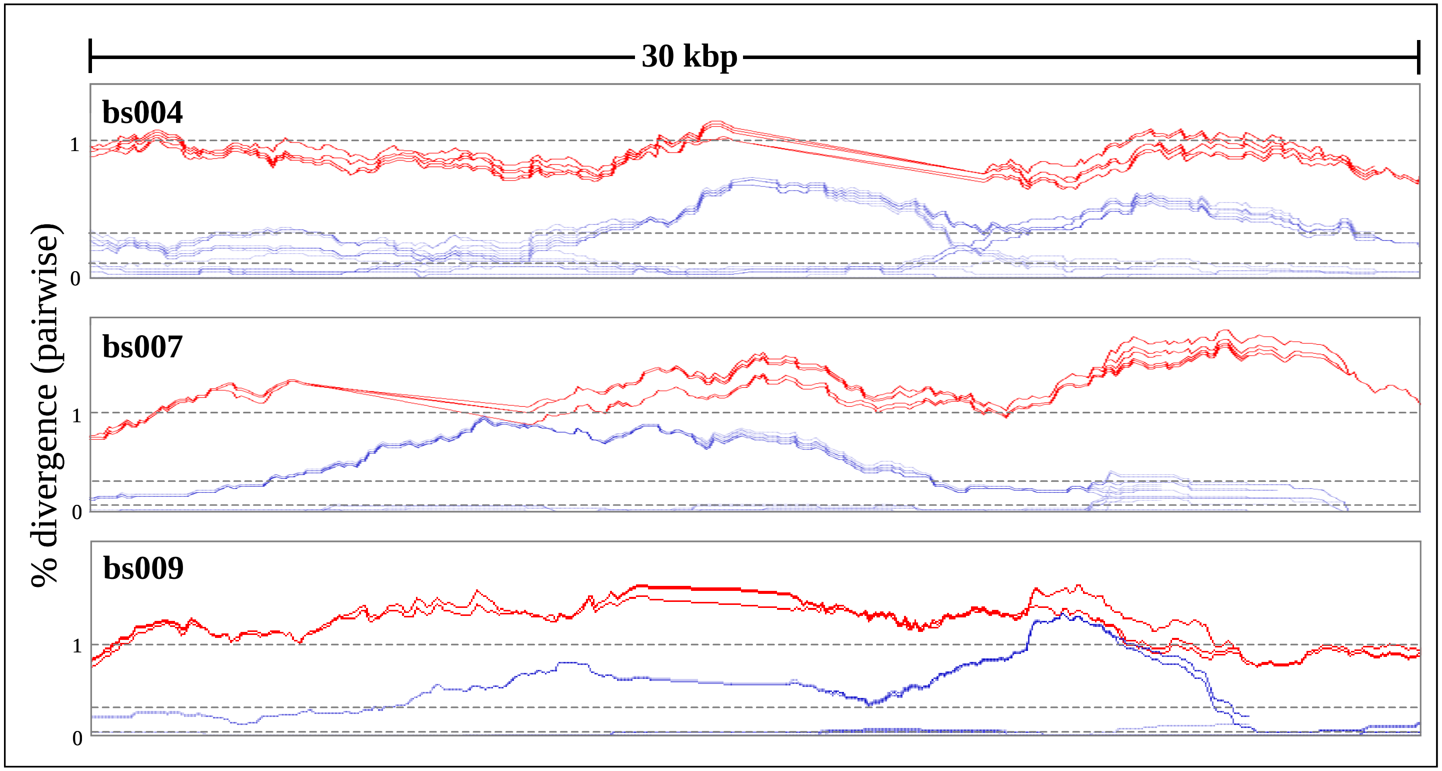


FIGURE 1.3.—Pairwise divergences in 5-kbp sliding windows (offset = 100 bp) over a 30-kbp genomic span for three loci. Blue lines indicate human-human comparisons; red lines indicate human-chimpanzee comparisons. In each panel, upper, middle and lower dotted lines represent pairwise divergences of 1%, 0.3% and 0.081%. The latter value is the genome-wide average divergence between two randomly sampled sequences. Straight edges indicate interpolation of the human-chimpanzee comparisons across regions in which chimpanzee sequence is lacking.

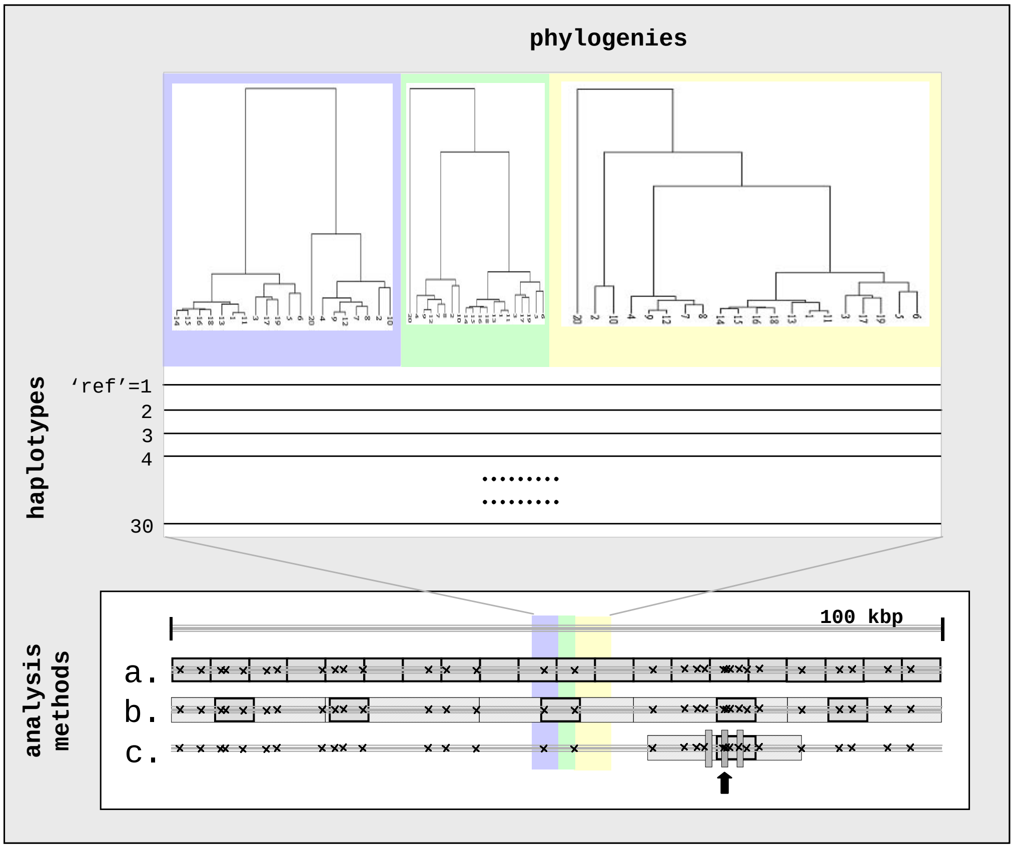


FIGURE 1.4.—Description of simulation and three alternate methods of analysis. The top portion illustrates the evolutionary relationship among 30 haplotypes of a population for a segment of genomic sequence. For those 30 haplotypes, there are two changes in their evolutionary relationship in this segment, due to ancestral recombination events. The sites of these ancestral recombinations are represented by edges between adjacent color blocks, which contain slightly differing phylogenies. The lower portion illustrates three alternate methods of analyzing the simulated genomes. (a) The number of nucleotide differences between the most dissimilar haplotypes ("MAXDIV") within each non-overlapping 5-kbp window is reported. (b) For each non-overlapping 20-kbp window, the MAXDIV of the most divergent 5-kbp window is reported. (c) For each 20-kbp window that satisfied the computational filtering requirements summarized in Figure 1.1, the MAXDIV of the most divergent 5-kbp window is reported (see MATERIALS AND METHODS for details). In order to simulate the filtering steps, three test regions (see small vertical boxes) were established in the center of each 20-kbp window, corresponding to the positions of the original read and sites 3 kbp upstream and downstream from this position.

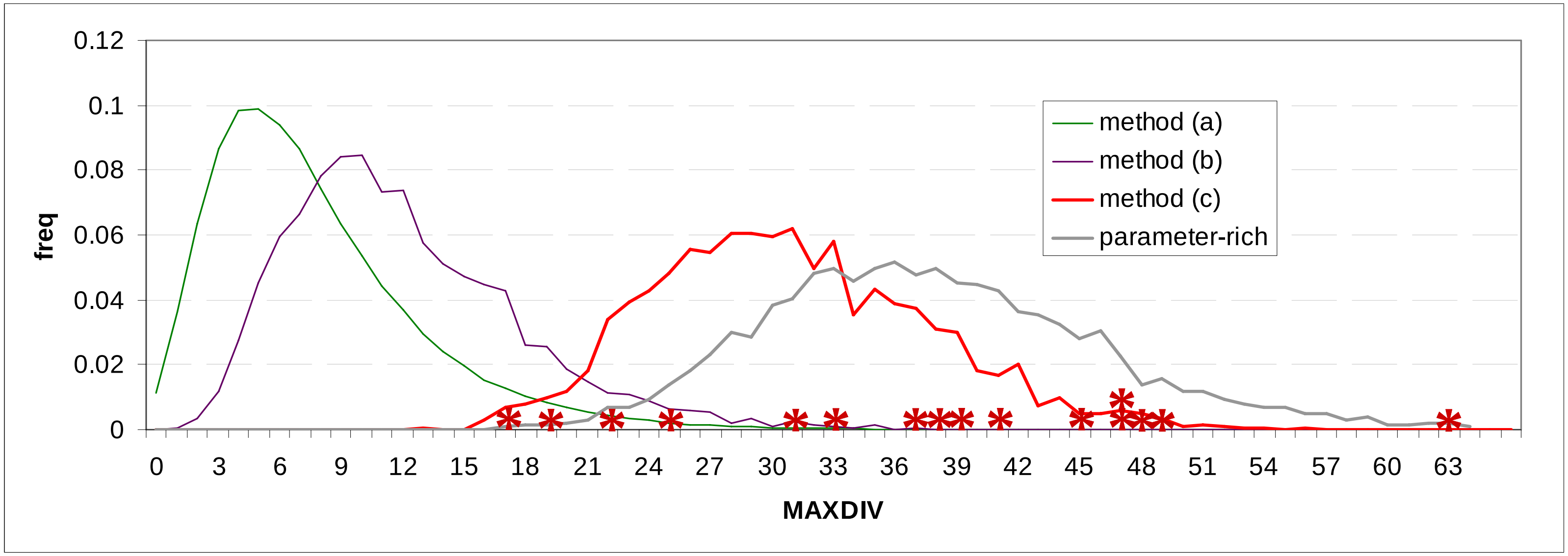


FIGURE 1.5.—Comparison of observed loci with simulated MAXDIV distributions. Curves labeled method (a-c) were generated by analyzing data simulated under the simple coalescent model, using the analysis methods illustrated in Figure 1.4. The curve labeled "parameter-rich" was generated by analyzing data simulated under the parameter-rich coalescent model using method (c). See MATERIALS AND METHODS for details of both simulation models and analysis methods. For each analysis method, a histogram was produced and then normalized such that the bar areas sum to one. Red asterisks indicate the MAXDIV of the 16 loci for which we obtained extended sequence. The smoothness of the curve for method (a) reflects a higher number of windows analyzed in the ten genomes with this method.

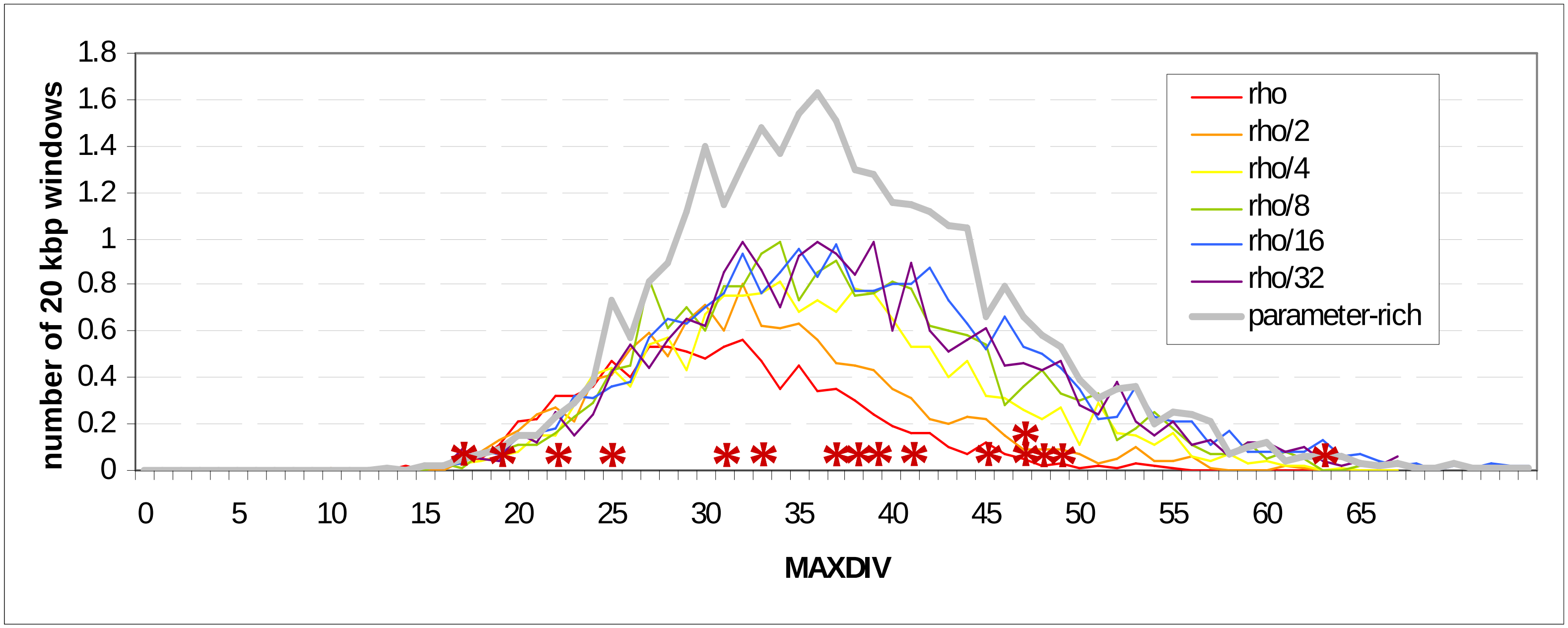


FIGURE 1.6.—The effect of varying recombination rates (rho) on simulated MAXDIV distributions. All distributions were generated using method (c). The curves for different multiples of rho (where rho=5×10^{-4}) used data simulated under the simple model. The curve labeled “parameter-rich” was generated by analyzing data simulated under the parameter-rich coalescent model. Red asterisks indicate the MAXDIV of the 16 loci for which we obtained extended sequence.

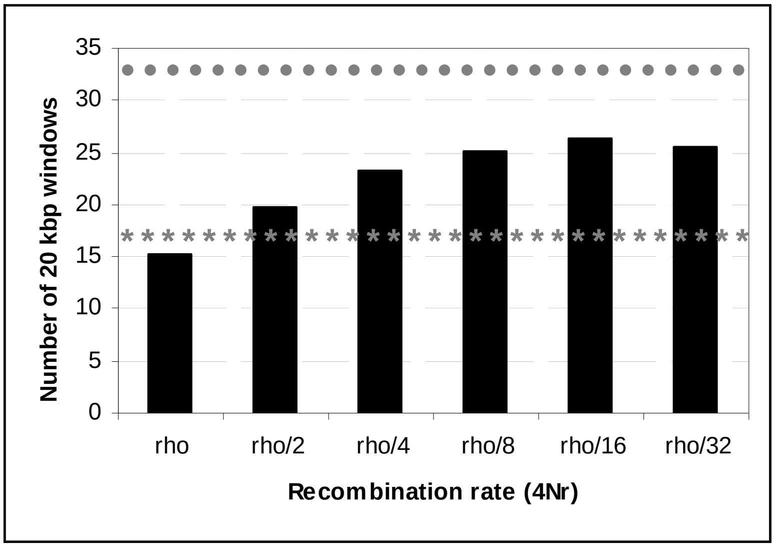


FIGURE 1.7.—Numbers of 20-kbp windows found per simulated genome for simple coalescent models with varying recombination rates and the parameter-rich coalescent model. Gray asterisk line indicates the number of 20-kbp windows identified in our real-data screen. Dotted gray line indicates the number of 20-kbp windows found per simulated genome for the parameter-rich model. Analyses used method (c), illustrated in Figure 1.4.

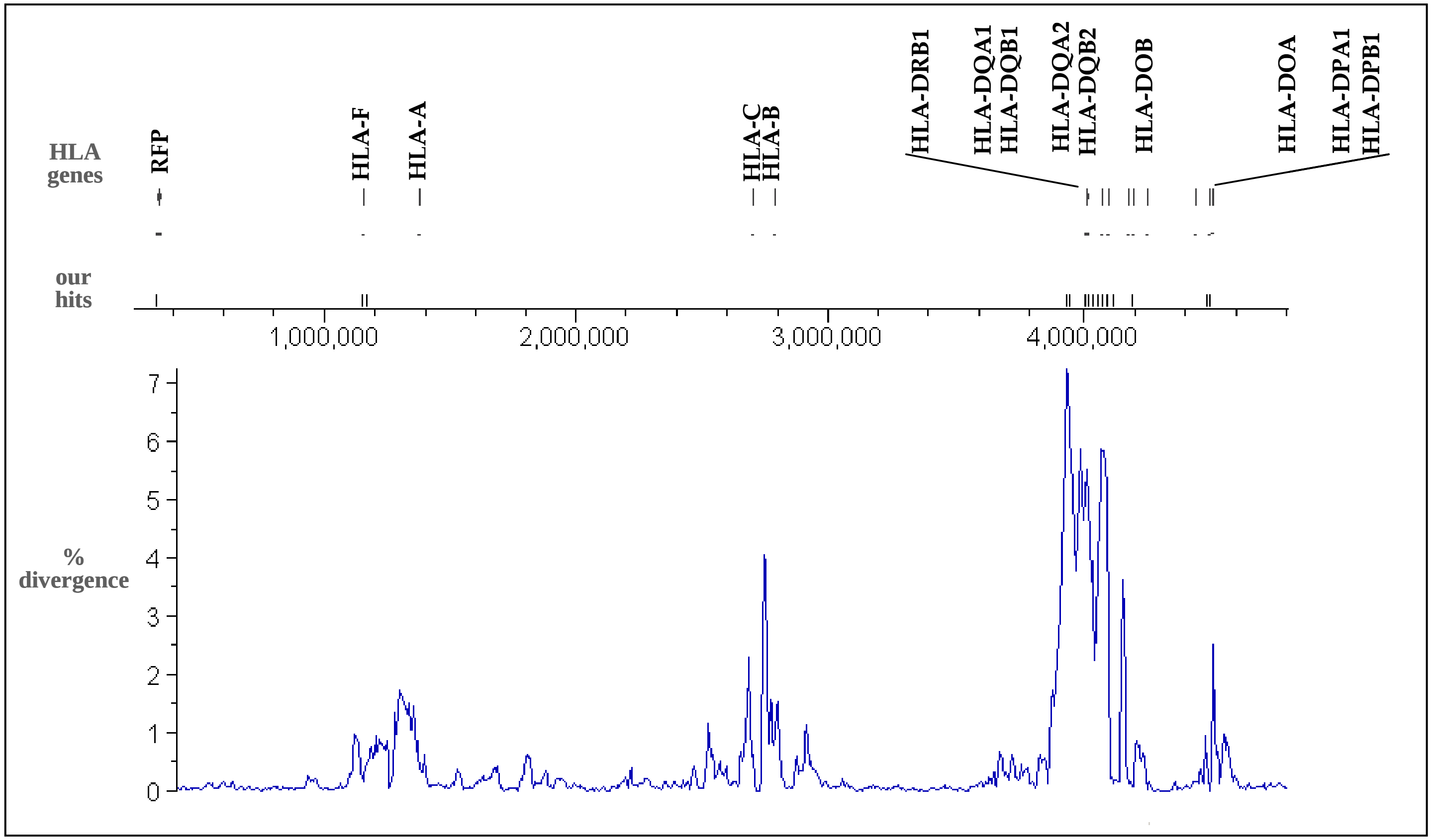


FIGURE 1.8.—Sites within the HLA locus that our computational filter indicated as putative highly polymorphic. The profile of the pairwise divergence for 10-kbp sliding windows with 5 kbp offsets of two HLA haplotypes, 6-COX and PGF (STEWART *et al.* 2004), is plotted, with select HLA genes and our hits indicated with black vertical lines above the scale bar.

PART II:

EXAMINATION OF THE BEHAVIOR OF SELECTION TEST STATISTICS ON SIMULATED DATA

Introduction

Detecting signatures of selection is currently a major goal for genome analysts, mainly because of the recent availability of large amounts of sequence data. It is therefore critical that attention be paid to the methods with which these signatures are being identified.

Selection can be thought of as a parameter—like recombination rate or migration rate—that has shaped the evolution of a population. As such, it is common to measure the strength of selection using so-called *selection coefficients*, often denoted as s, and a measure of dominance, often denoted as h. For example, under positive selection, an individual that is homozygous for the derived allele may be a fraction s more fit than an individual that is homozygous for the ancestral allele. We would then write the fitness of the ancestral homozygote as 1 and the fitness of the derived homozygote as $1+s$. The fitness of the heterozygote, in the case of positive selection, is anywhere between these two values—depending on the degree of dominance of the derived allele (eg: $1+sh$, for $0 \leq h \leq 1$). Balancing selection is often modeled as heterozygote advantage, in which the fitness of the heterozygote is greater than that of either homozygote.

For example, the fitnesses of the ancestral homozygote, the heterozygote and the derived homozygote could be represented as 1, 1+*s*, and 1+*sh*, respectively, for $s > 0$ and $-(1/s) \leq h \leq 1$. A special case of balancing selection is symmetric balancing selection, in which the fitness of the homozygotes is the same (h=0). These parameters—the relative fitnesses of the genotypes—are reflected in the evolutionary history of the population.

Theoreticians have developed many test statistics for inferring selection. Some are based directly on the number of coding changes (NEI and GOJOBORI 1986; YANG *et al.* 2000). Others involve comparisons of observed evolutionary trees with trees expected under neutrality, using various summary statistics, including FST (LEWONTIN and KRAKAUER 1973) and linkage disequilibrium (e.g., TISHKOFF *et al.* 2001). In this study, we concentrate on two test statistics that are of particular relevance when searching for loci that might be under balancing selection: Tajima's D (TAJIMA 1989) and maximum pairwise divergence ("MAXDIV").

A locus under positive selection is expected to have both a decreased Tajima's D and MAXDIV score, compared to a locus under neutrality. This is because when the descendants of one haplotype rapidly replace all other haplotypes in the population, two things happen: (1) the overall diversity decreases, and (2) the variance in the number of nucleotide differences between any pair of haplotypes in the population decreases. The more quickly the derived, positively selected allele goes to fixation, the more dramatic these effects will be. In contrast, if it takes many generations for the derived, positively selected allele to get to fixation—close to the expected number of generations back to the tMRCA under neutrality—then the effects will be very slight. Similarly, as time progresses, a locus that at one point underwent a “selective sweep” (i.e., a positively selected allele went to fixation), will again look neutral no matter how strong the historical selective sweep was.

A locus under balancing selection, on the other hand, is expected to have an increased Tajima’s D and MAXDIV score, compared to a locus under neutrality. This is because when two functionally distinct of haplotypes are maintained in a population over time, the resulting evolutionary tree is likely to have two divergent clades. The effects are the opposite of those produced by positive selection: (1) an overall increase in diversity, and (2) an increase in the variance in the number of nucleotide differences between any pair of haplotypes in the population. Specifically, there will be two types of haplotype pairs, those between members of the same clade and those between members of different clades. Over time, the divergence between the two clades increases, but recombination between the two clades homogenizes the region surrounding the selected site, making test statistics look as they would on a neutral region.

Often, tools for inferring population parameters are developed in tandem with programs that simulate populations affected by those parameters; hence, the parameters under which simulated data are generated can be compared to inferences made. While many evolutionary features have been studied in this way—including exponential growth (KUHNER *et al.* 1998), recombination (KUHNER *et al.* 2000) and migration (BEERLI and FELSENSTEIN 2001)—development of a program simulating selection has proven more challenging, and thus tests for selection are largely untested.

Recently two groups have proposed algorithms for simulating selection (KRONE and NEUHAUSER 1997; NEUHAUSER and KRONE 1997; SPENCER and COOP 2004). Both methods track a sample of extant haplotypes and their ancestors, moving backward in time. These methods are extensions of coalescent theory, which in its basic formulation is much better suited to the analysis of drift, population dynamics, and migration than any form of selection.

Although the "backward" nature of coalescent simulators is elegant and time-efficient, advances in computational speed make it currently feasible to run reasonably large forward-time simulations (for early descriptions of such algorithms, see CROSBY 1973; for early descriptions of such algorithms, see FRASER and BURNELL 1970). In contrast with "backward" simulations of selection, simulations of selection that move forward in time are algorithmically straightforward. With this in mind, we implemented a simulator that simply tracks the evolutionary events (mutation, recombination and drift) of all the haplotypes in a population, including a function that preferentially selects haplotypes for the next generation.

Our aim was to quantify the behavior of Tajima's D and MAXDIV using our "forward-evolving" simulator. As should be expected, we observe an overall positive correlation between Tajima's D and MAXDIV test statistics (Figure 2.1). Additionally, we explore the distribution of test scores at various points in the life of a derived allele under selection.

Methods

IMPLEMENTATION OF EVOLUTION UNDER NEUTRALITY

Before describing the specifics of our treatment of selection implementation, we first describe the steps in a forward neutral simulation (Figure 2.2). At each generation, a number of mutations, *k*, is chosen for the entire population stochastically from a binomial distribution,

$$P(k) = \binom{n}{k} p^k q^{n-k}$$

where *P*(*k*) = probability of *k* mutations, *n* = (number of haplotypes in the entire population) × (number of nucleotides per haplotype), and *p* = mutation rate (per nucleotide, per generation), and *q* = 1 - *p*. The mutations are then distributed uniformly throughout the haplotypes. Next, a number of recombination events is chosen for the entire population stochastically from a binomial distribution (*n* = same as above, *p* = recombination rate), and the recombination points are distributed uniformly through the population. Recombination always involves swapping the outer sequence (or list of events) of two haplotypes, as indicated in Figure 2.2. For example, when simulating a 5-kbp region, all recombinations occurring from nucleotides 1 - 2,500 result in haplotypes exchanging the first part of their sequence. Implicit in this model is the assumption that a selected site is an invisible nucleotide (or non-recombining nucleotides) in the center of the simulated haplotype, which never mutates and is never recombinationally disrupted. Finally, drift is simulated by sampling the next generation, with replacement, from the current generation. Note that this implementation is a slight departure from a strict Wright-Fisher model; once a haplotype has undergone recombination, the original, non-recombined haplotype is no longer available for sampling in subsequent generations.

To decrease run time in simulating a human-population, we decreased the total number of haplotypes in our simulated effective population by 10^2 and increased per-generation mutation rate (μ) and recombination rate (r) by 10^2, so that θ ($4N_e\mu$) and ρ ($4N_er$) remain the same. Because of this device, everything (mutation, recombination and drift) happens 100x faster than it would if N, μ, and r were un-scaled (ROBERTSON 1970); hence, each simulation-round actually simulates 100 generations.

IMPLEMENTATION OF EVOLUTION UNDER SELECTION

Selection is implemented in three phases: (i) generation of a starting neutral population, (ii) generation of an appropriate derived-allele frequency trajectory, and (iii) actual simulation of haplotypes during the selection phase.

To generate a starting neutral population, we could have either used a backward simulator such as ms (HUDSON 2002) or simply relied on a neutral "burn-in" phase. Our program currently does the latter. We start with a population of identical haplotypes, then run the forward simulator until the haplotypes have the characteristics of a neutrally evolving population. We determined that a burn-in time of 10^3 simulation-rounds generates an equilibrium neutral population similar to one generated by the gold-standard backward coalescence simulator, as measured by Tajima's D and MAXDIV (Figures 2.3 and 2.4).

We then generate an appropriate derived-allele frequency trajectory. This is an essential step, since the most likely fate of a new allele—even one that is selectively favored—is rapid extinction. We wanted to ensure that we are simulating only "successful" cases of selection, in which the new allele rises to its equilibrium frequency, before engaging in the time-consuming mutation and recombination steps. In the case of positive selection, the equilibrium frequency of the derived allele is simply 1.0. In the case of balancing selection as described above (the fitnesses of the ancestral homozygote, the heterozygote and the derived homozygote are 1, 1+*s*, and 1+*sh*, respectively, for $s > 0$ and $–(1/s) \leq h \leq 1$), the equilibrium frequency of the derived allele is 1/(2-*h*). In the special case of symmetric balancing selection (*h*=0), which we examine here, the equilibrium frequency of the derived allele is 0.5. For cases of balancing selection, we also require that both derived and ancestral alleles be maintained in the population until the simulation is terminated at a user-defined point. To achieve an appropriate derived-allele frequency trajectory, we simply simulate allele frequency trajectories over and over, rejecting all that don't satisfy our criteria. Once we have a successful trajectory (for example, Figures 2.5 and 2.6), we simply record it for use in the next step, during which the actual haplotype simulation takes place.

The steps to simulate the allele trajectory are as follows. Assume that "A" is the positively selected derived allele, and "a" is the ancestral allele. The number of type "A" alleles in generation *i*+1 is determined using a binomial distribution, where *n* = the number of haplotypes in the entire population, and *p* = the number of type "A" alleles expected in the next generation based on the frequency of "A" in generation *i* and the selection coefficients of the genotypes. Note that unlike the rates of mutation and recombination events, the expected rate of increase per simulation-round for a new, positively selected allele is not dependent on N_e but only on the selection coefficients, *s* and *h*. Therefore, when we use a selection coefficient *s* for a certain number of simulation-rounds, we are really simulating with coefficient 100*s* in terms of actual generations.

Finally, we perform the actual simulation of haplotypes during the selection phase. The only difference between simulation with and without selection is in how haplotypes are sampled from the preceding generation. In the case of neutrality—as described above—generation *i*+1 is simply sampled randomly from generation *i*, with replacement. In the case of selection, we assume a bi-allelic polymorphism, which creates two allelic classes, each of which is independently sampled with replacement. For example, if in one generation there are 3 “a” alleles (green, in Figure 2.7) and 2 “A” alleles (purple, in Figure 2.7), and in the next generation, the pre-calculated allele-frequency trajectory calls for 2 “a” (green) alleles and 3 “A” (purple) alleles, we simply sample 2 “a” alleles, with replacement, from the current 3 “a” alleles, and 3 “A” alleles from the current 2 (illustrated in Figure 2.7).

Results

We first tested our simulation results on data undergoing positive selection under a model in which the ancestral and derived alleles are co-dominant. In our first set of simulations, we stopped each of the 1,000 simulations when a certain derived allele frequency was reached, and then calculated each of the test statistics. In our next set of simulations, we stopped each of the 1,000 simulations a certain number of simulation-rounds after the equilibrium frequency—in this case, 1.0—was reached, and then calculated each of the test statistics. Results are shown in Figure 2.8.

When the positive selection is strong (*s*=0.5), both Tajima's D and MAXDIV begin decreasing when the fraction of derived alleles in the population gets above 0.5. Tajima's D reaches its minimum just before fixation (0.9), while MAXDIV reaches its minimum at fixation (1.0). When the positive selection is weak (*s*=0.01), Tajima's D and MAXDIV begin decreasing later—when the derived allele frequency reaches around 0.7—because the evolutionary relatedness of the haplotypes begins to recover a neutral-like state even during the sweep phase; the minima for Tajima's D and MAXDIV are the same as for the case of strong selection. Interestingly, Tajima's D and MAXDIV both return to neutral-state levels sometime between 100 and 1,000 simulation-rounds after fixation of the derived allele. This is consistent with our observation that by 10^3 simulation-rounds (equivalent to 10^5 real-time generations), a population that initially consisted of identical haplotypes resembles one that is neutrally evolving (Figures 2.3 and 2.4).

We then tested our simulation results on data undergoing "strong" symmetrical balancing selection. In this case, "strong" balancing selection means that neither the ancestral nor the derived allele has gone to fixation at the point when the simulation is stopped. This requirement appears have a much bigger effect than the relative advantage of the heterozygote, because the Tajima's D and

MAXDIV curves we observe looked similar for systems in which the fitness advantage of the heterozygote ranged from 0.01 to 0.5 greater than either homozygote. The main difference is in the probability that a system of weak balancing selection will survive many generations (as opposed to either going to fixation or being lost, due to drift). As in the case of positive selection, in our first set of simulations we stopped each of the 1,000 simulations when a certain derived allele frequency was reached, and then calculated each of the test statistics. In our next set of simulations, we stopped each of the 1,000 simulations a certain number of simulation-rounds after the equilibrium frequency—in this case, 0.5—was reached, and then calculated each of the test statistics. Results are shown in Figure 2.9.

Under “strong” symmetrical balancing selection, both Tajima’s D and MAXDIV begin to decrease slightly when equilibrium is reached, then increase to a maximum about 1,000 simulation-rounds after equilibrium is reached. The reason for the initial dip is the same as for the subsequently more dramatic dip in these statistics for a positively selected allele; in the early stages of the rise in frequency of a derived allele, balancing selection and positive selection scenarios are identical. As the allele stays at equilibrium, the statistics increase until recombination erodes away the signal. Tajima’s D and MAXDIV both return to neutral-state levels sometime between 1,000 and 10,000 simulation-rounds past equilibrium.

Discussion

This study could be easily expanded to examine (i) cases of asymmetric balancing selection, (ii) other selection test statistics, including, for example, those using linkage disequilibrium, (iii) more extreme selection coefficients, (iv) additional time points, (v) larger samples taken from the population.

What does the current analysis reveal about what sort of selective events we are most likely to detect in humans using these statistics? Firstly, the most recent successful cases of positive selection are most likely to have conferred the greatest selective advantage, simply because the time from original mutation to fixation of the derived allele is smallest for cases of strong selection. For example, a derived allele that increases the fitness of the heterozygote by 2%, and the derived homozygote by 4% compared to the ancestral homozygote, is expected to take around 10,000 years (500 generations) to go to fixation. [Note that while the probability that an allele will go to fixation is dependent on N_e, the expected number of generations it takes a selected allele to go to fixation, assuming that it does, is not dependent on N_e.] More selectively advantageous alleles would be expected to take even less time. Secondly, the most ancient detectable cases of successful positive selection are less than 2 million years (1000 simulation-rounds) old. These figures suggest that we should not mistake our ability to detect strong recent positive selection for an overabundance of strong recent positive selection (for example, correlating with agriculture 10,000 years ago). Similarly, we should not mistake a lack of signal more than 2 million years ago for a paucity of ancient positive selection events (for example, before the rise of ‘modern’ homo sapiens).

Balancing selection, on the other hand, is most readily identified, by these statistics, when it originated roughly 2 million years ago (1,000 simulation-rounds ago). It takes roughly that amount of time to accumulate the divergence between balanced haplotype clades necessary to produce a signal. If the mutation arose much before then, recombination between balanced haplotypes would have homogenized regions flanking the actual selected site, decaying the signal.

Even more importantly, there seems to be a great deal of overlap between the distributions of both Tajima’s D and MAXDIV under neutrality and at the peak of their signal under selection. For example, most of the Tajima’s D scores for our most positively selected allele when the signal is strongest (Figure 2.8c, derived allele frequency = 0.9) are well within the 95 percent confidence interval of the Tajima’s D distribution for a locus under neutrality. The MAXDIV score appears to do a little better at its nadir, with most of the MAXDIV scores outside of the 95 percent confidence interval for a locus under neutrality (Figure 2.8d, derived allele frequency = 1.0). Unfortunately, neither statistic is in this most informative state for very long—by 100 simulation-rounds after equilibrium, even the mean MAXDIV is well within the 95% confidence interval. Of course, for weaker positive selection the signals are even less distinct from neutrality. Likewise, under strong balancing selection, while the median score for both statistics is above the 95% confidence interval under neutrality (Figure 2.9a, b, 1e3 generations after equilibrium), there is a great deal of overlap between the 95% confidence intervals.

From this brief analysis, it is clear that because we don't know the actual number of selective events—and know even less about their selection coefficients—it is critical to keep in mind not only the probability of getting a equal or greater test scores under neutrality (p-value), but also the number of neutral loci in the entire genome expected to have an equal or greater test score, often called the False Discovery Rate when analyzing loci suspected of having been under selection.

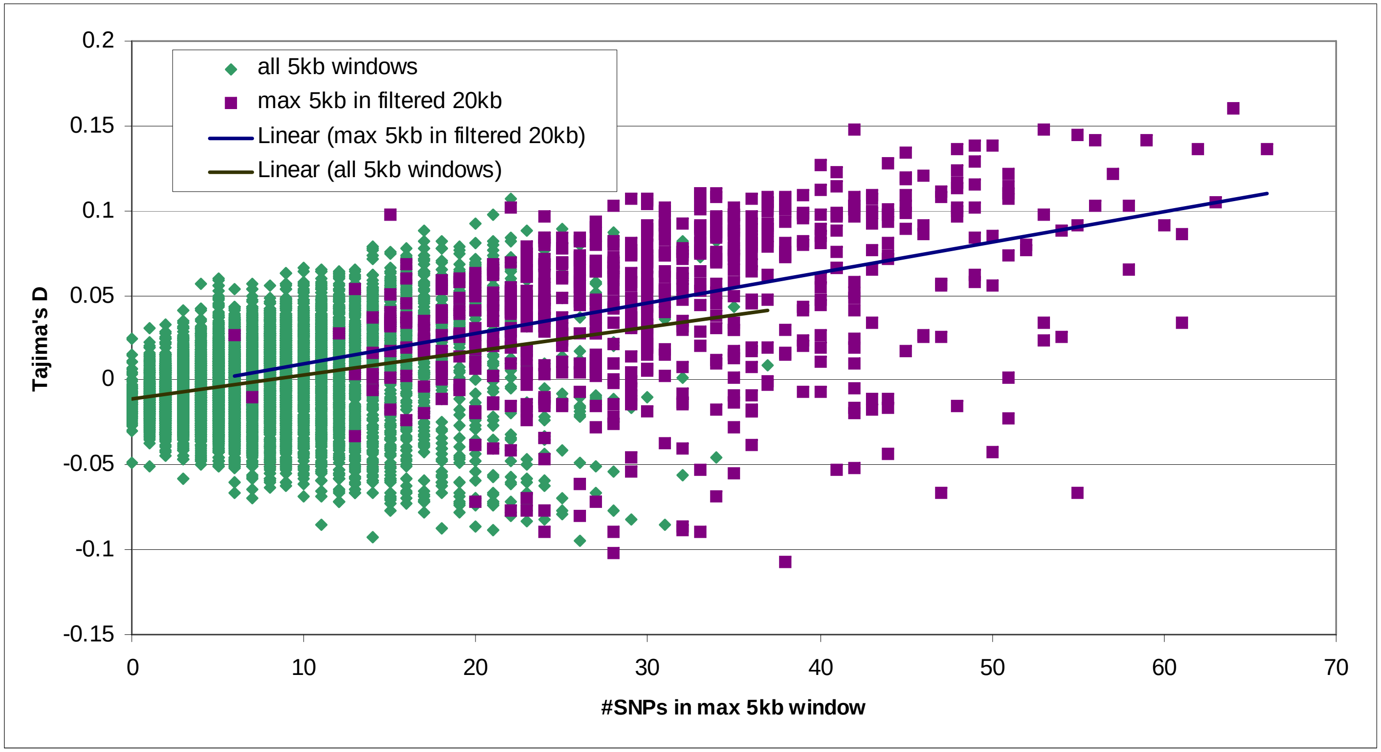


FIGURE 2.1.— Correlation between Tajima's D and MAXDIV. Each dot represents one 5-kbp genomic window, simulated under a simple neutral model of coalescence (ref Hudson). For the green dots, all non-overlapping 5-kbp windows were examined in simulated 1-Mbp genomes (as in method (a), Figure 1.4). For the purple dots, we enriched for highly polymorphic regions by selecting only those 5-kbp windows that were identified using simulated, highly-polymorphic reads (as in method (c), Figure 1.4).

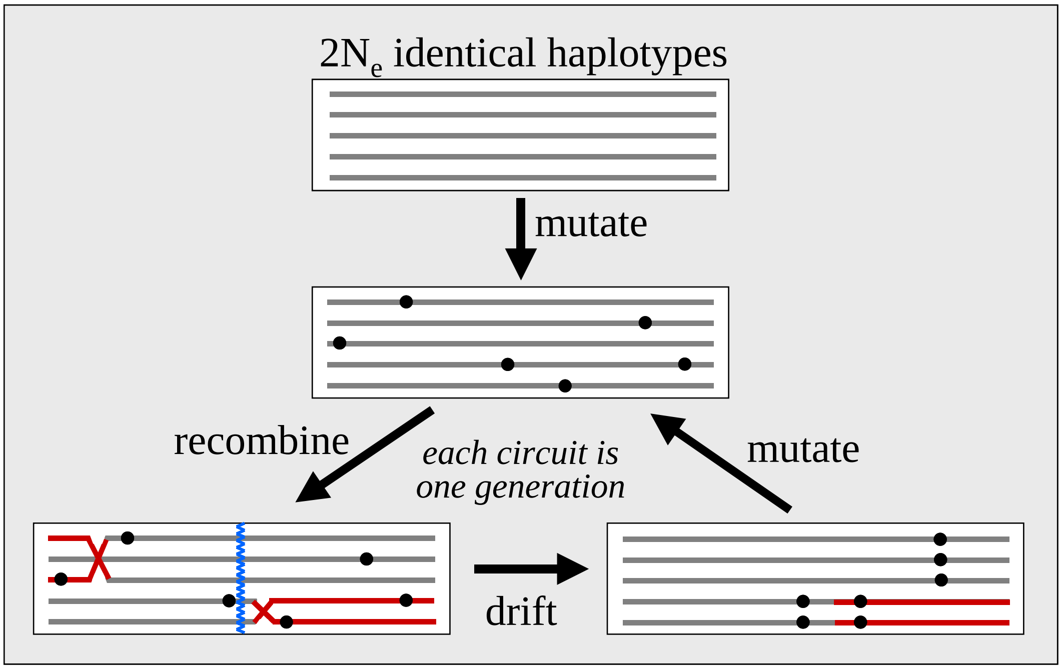


FIGURE 2.2.—Schematic of the forward-time evolution simulator under neutrality. Initially, $2N_e$ identical haplotypes are created. Then the population of haplotypes begins a cycle of mutation, recombination, and drift, each cycle representing one generation, or simulation-round, of simulation. In the box representing newly-recombined haplotypes, a blue zigzag line shows the "continental divide" of the sequence; as illustrated by the swapped red lines, recombination always results in exchange of peripheral sequence, never including the blue zigzag center.

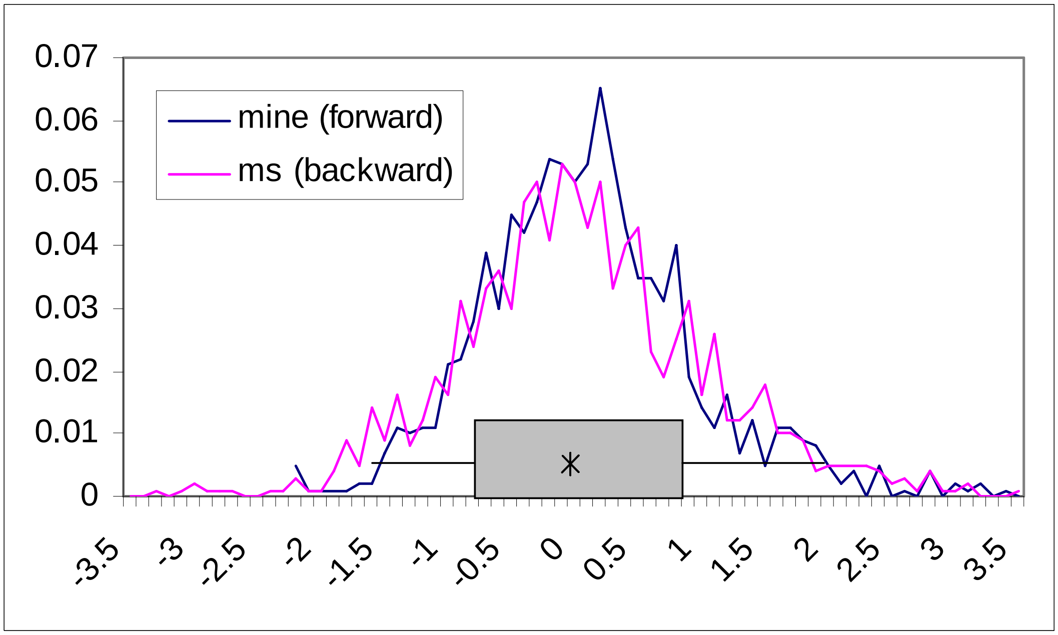


FIGURE 2.3.—Comparison of Tajima’s D scores for 1,000 5-kbp windows simulated under the forward coalescence simulator and a standard “backward” coalescence simulator (HUDSON 2002). The grey box indicates the middle 65% of the data; the ‘whiskers’ indicate the middle 95% of the data; the star in the grey box indicates the median value.

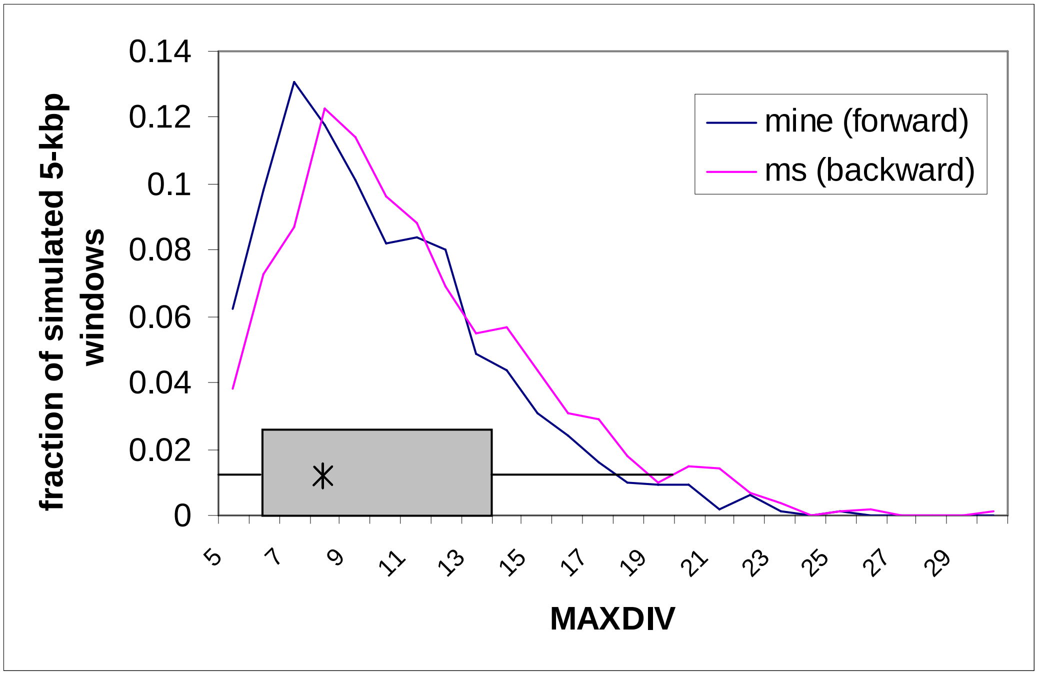


FIGURE 2.4.—Comparison of MAXDIV scores for 1,000 5-kbp windows simulated under the forward coalescence simulator and a standard “backward” coalescence simulator (HUDSON 2002). The grey box indicates the middle 65% of the data; the ‘whiskers’ indicate the middle 95% of the data; the star in the grey box indicates the median value.

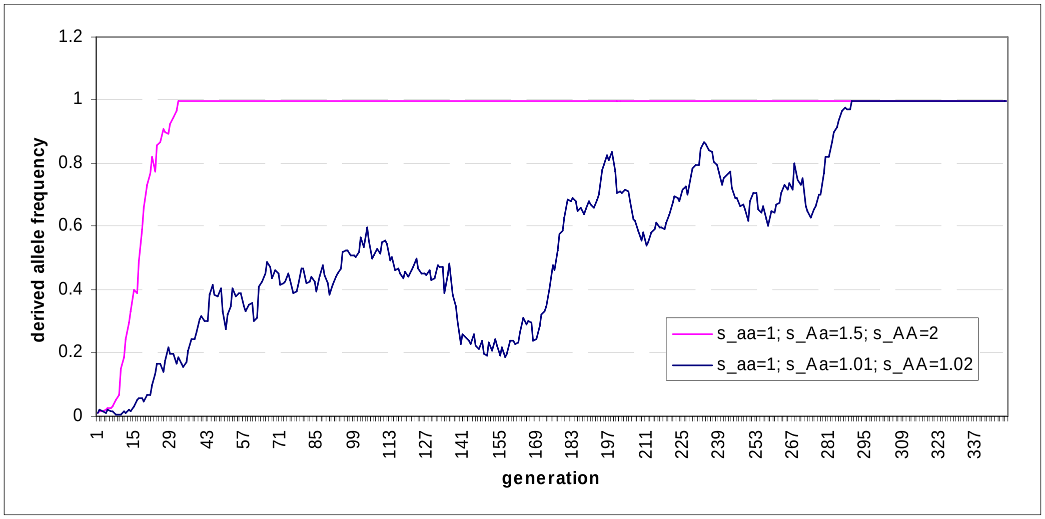


FIGURE 2.5.— Two “successful” examples of the trajectories by which positively selected derived alleles went to fixation during simulation.

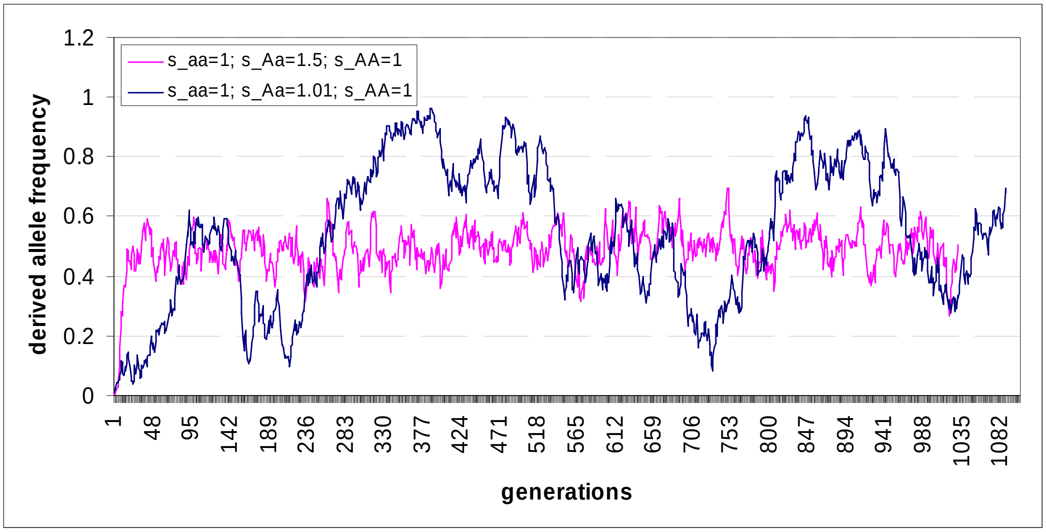
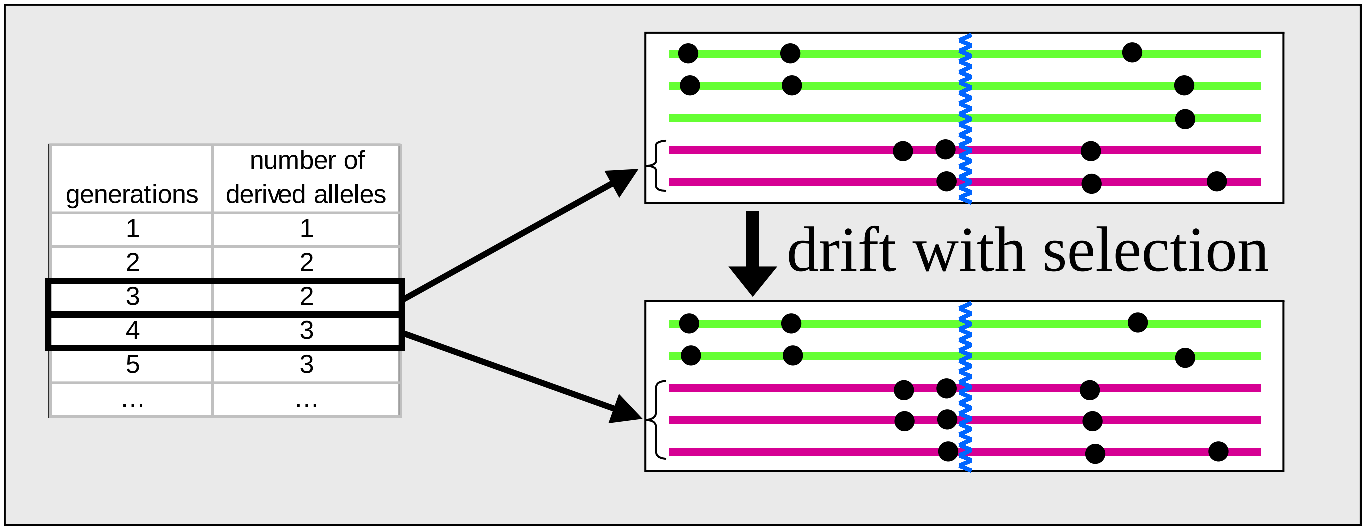


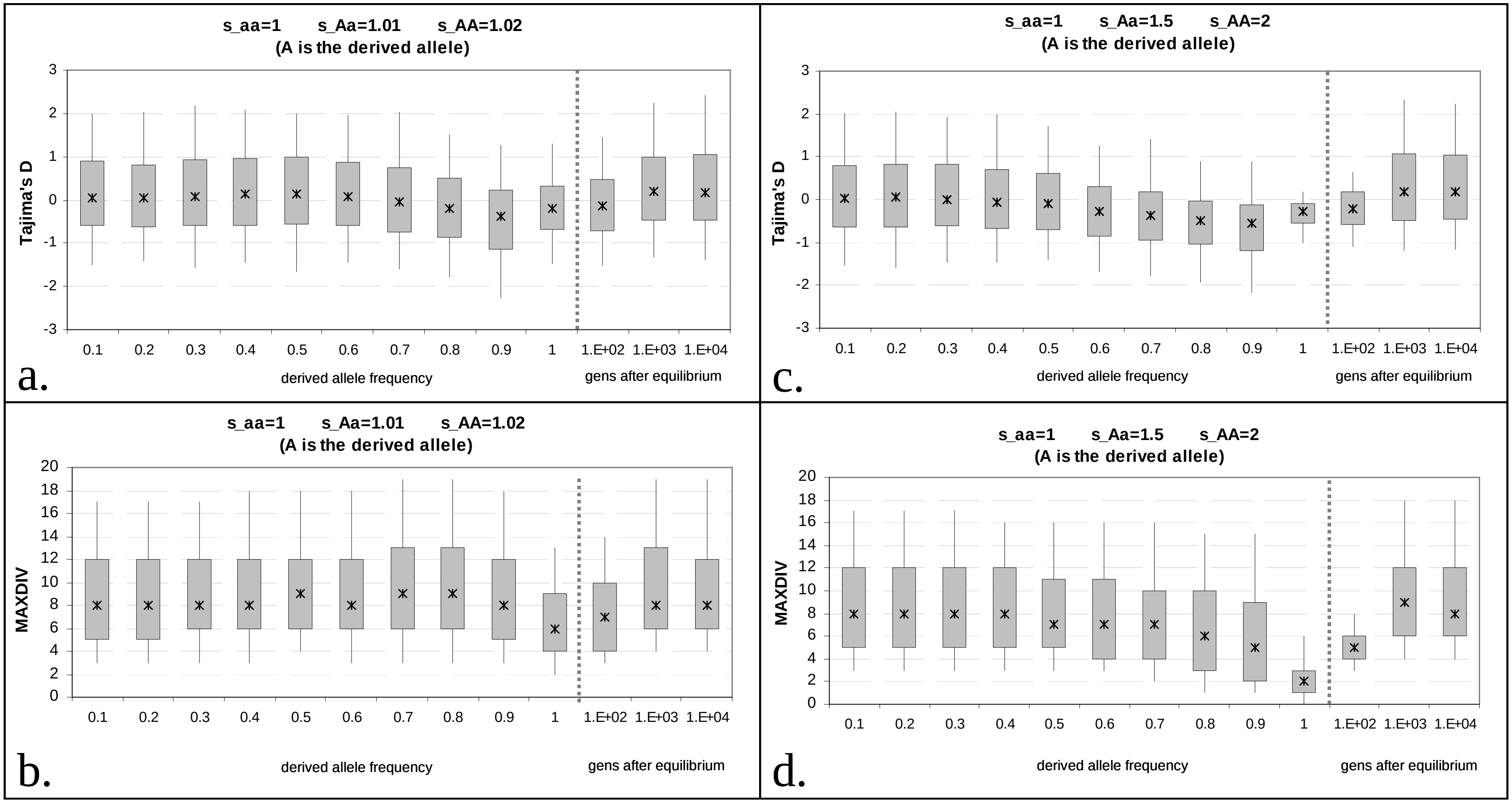
FIGURE 2.6.— Two “successful” examples of the trajectories by which derived alleles under strong symmetric balancing selection went to equilibrium frequency (i.e., neither allele in the bi-allelic system has gone to fixation at the ending timepoint).



generations	number of derived alleles
1	1
2	2
3	2
4	3
5	3
…	…

FIGURE 2.7.— Schematic of the drift-with-selection phase of the forward-time evolution simulator. The table on the left represents the pre-calculated allele-frequency trajectory, translated into actual numbers of derived alleles for the population per generation. As directed by the table, there are two derived (purple) alleles in generation 3 (the top boxed population of haplotypes) and three derived (purple) alleles in generation 4 (the bottom boxed population of haplotypes). The blue zigzag line indicates the location of an invisible selected nucleotide (or non-recombining block of nucleotides) that is both immutable and not transferable to other alleles via recombination. This invisible selected site is essentially the tag that identifies the allelic classes—in this case as either green or purple.

FIGURE 2.8.—Effect of positive selection on test statistics (a) Tajima's D, and (b) MAXDIV over time. The grey box indicates the middle 65% of the data; the 'whiskers' indicate the middle 95% of the data; the star in the grey box indicates the median value. Boxplots in the "derived allele frequency" section are generated from simulations that were stopped the first time the derived allele frequency reached the specified minor allele frequency. Boxplots in the "generations after equilibrium" section are generated from simulations stopped *n* generations after equilibrium—in this case, a derived allele frequency of 1.0—was reached.


s_aa=1 s_Aa=1.01 s_AA=1.02
(A is the derived allele)
Tajima's D
3
2
1
0
-1
-2
-3
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.E+02
1.E+03
1.E+04
derived allele frequency
gens after equilibrium
a.
s_aa=1 s_Aa=1.5 s_AA=2
(A is the derived allele)
Tajima's D
c.
s_aa=1 s_Aa=1.01 s_AA=1.02
(A is the derived allele)
MAXDIV
20
18
16
14
12
10
8
6
4
2
0
b.
s_aa=1 s_Aa=1.5 s_AA=2
(A is the derived allele)
MAXDIV
d.

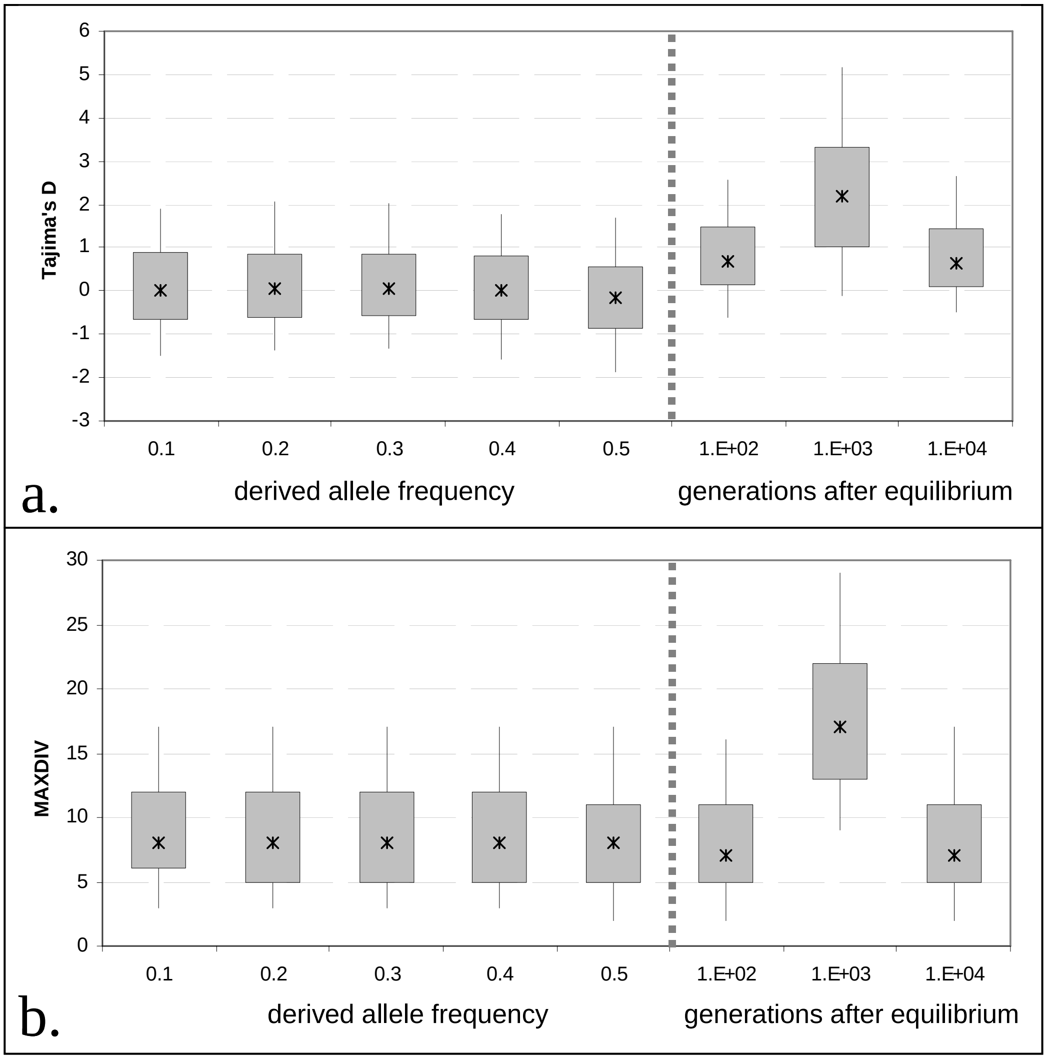


FIGURE 2.9.--Effect of strong balancing selection on test statistics (a) Tajima's D, and (b) MAXDIV over time. Each boxplot represents 1,000 5-kbp simulations. The grey box indicates the middle 65% of the data; the 'whiskers' indicate the middle 95% of the data; the star in the grey box indicates the median value. Boxplots in the "derived allele frequency" section are generated from simulations that were stopped the first time the derived allele frequency reached the given minor allele frequency. Boxplots in the "generations after equilibrium" section are generated from simulations stopped *n* generations after equilibrium—in this case, a derived allele frequency of 0.5—was reached.

PART III:

THE ROLE OF BALANCING SELECTION – A PROPOSED MODEL

Introduction

While it seems likely that the number of systems under balancing selection is limited for any given species (e.g., ASTHANA *et al.* 2005; BUBB KL 2006), current methods for detecting balancing selection based on coalescence-based analysis of locally-affected genomic segments, as discussed in the previous section and illustrated by other investigators (THORNTON 2005; WIUF *et al.* 2004), are inadequate. Nevertheless, it seems worthwhile to pursue systems of balancing selection because of the importance—from agricultural to medical—of the examples currently known (e.g., BOYES *et al.* 1997; FIX 2003; HUGHES and NEI 1992; LIBERT *et al.* 1998; LIU *et al.* 2004). In this section, I offer an updated model of balancing selection, emphasizing a description of characteristics which might be used to discover additional systems under this type of selection.

I begin by briefly reviewing the history of the idea of balancing selection, then move on to a description of our current model, drawing on both theoretical predictions and empirical evidence from a wide variety of metazoans. The model includes two major facets, both relevant to new detection techniques. First, systems under balancing selection can be naturally dissected into two temporally distinct classes. Second, the role of convergent evolution in most systems of balancing selection is elucidated. We finish with a discussion of some potentially useful applications of this model.

Background

Genetic variability within a species is central to Darwin's original theory, as it provides the raw material on which selection can act, by rapidly increasing the fraction of individuals in the population with optimal fitness. As such, Darwin's theory of selection came in conflict with observations among plant and animal breeders of inbreeding depression, which is essentially an extreme version of survival—or in this case reproduction—of the fittest (EAST 1936; JONES 1917), a phenomenon that was replicated and more rigorously quantified in *Drosophila* (reviewed in SIMMONS and CROW 1977). It was this apparent conflict between theory and empirical observation that initiated the debate as to the reason for genetic variability within natural populations—essentially a debate between rampant balancing selection, in which heterozygotes are generally more fit than homozygotes, and a "classical hypothesis", in which inbreeding fortuitously increases the number of deleterious-recessive homozygotes (DOBZHANSKY 1955). A major ideological problem with the "balanced hypothesis" was the high degree of accompanying genetic, or segregational "load" inherent in maintaining large numbers of individuals of sub-optimal fitness (CROW 1958; MULLER 1950). However, the large amount of diversity in natural populations—direct observations of which were made possible by advances in the ability to measured differences in the electrophoretic properties of proteins (e.g., LEWONTIN and HUBBY 1966)—bolstered the notion that selection could be actively maintaining natural genetic variability.

Much of the controversy died down with the development of the ‘neutral theory’ (KIMURA 1968), which offers a robust mathematical explanation for the existence of genetic variation in the absence of any selective force. But a few fantastic examples, namely at the hemoglobin and MHC loci, kept interest in balancing selection alive—some have argued to the detriment of evolutionary biology as a whole (WILLIAMS 2003).

In recent years, coinciding with advances in the ability to acquire abundant nucleotide sequence information, there has been a revival of claims for selection, including those for balancing selection. There are many forms of balancing selection, and many tests that work with some forms and not others (HEDRICK 1998). The very definition of balancing selection—any type of selection that favors maintenance of genetic variability within an interbreeding population—is semantically precarious. For example, human subpopulations, which have varying degrees of outbreeding, may have unusual allele frequencies because of the effects of local positive selection (e.g., LIVINGSTONE 1984; YOSHIURA *et al.* 2006). Furthermore, in all but the most extreme cases, it remains possible that current allele distributions arose through genetic drift during and following population bottlenecks. In such cases, although multiple genetic variants are maintained in the larger, more-or-less interbreeding population, it is not considered balancing selection. Despite recent advances in our ability to collect and analyze genetic data, it remains unclear how pervasive balancing selection actually is in the natural world.

Discrete Classes of Systems Under Balancing Selection

Cases of balancing selection in metazoans can be divided surprisingly neatly into two categories, depending on the age of origin of the derived balanced allele. In one category, we have examples in which the time to most recent common ancestor (tMRCA) of all the alleles greatly exceeds that for typical neutral loci, and in the other category, tMRCA is less than that for most neutral loci. This dichotomy is predicted by theory (see Figures 3.1 and 3.2) and also supported by the handful of observed cases of balancing selection among metazoans. In this article, we term those cases with tMRCA greater than twice the mean for neutral alleles as “ancient,” and those with less than twice the mean tMRCA as “recent,” examples of balancing selection.

Probably the most familiar place for any discussion of ancient balancing selection to begin is with the MHC locus. There is an abundant literature on MHC, which I will not attempt to review here, but two aspects of the MHC system are particularly relevant to the discussion of what is required for a balanced locus to be ancient. First, although it now plays a central role in helping the body distinguish foreign proteins from pathogens, the MHC system may have been co-opted from a system for recognizing foreign tissue from individuals of the same species. This model, which is commonly called Burnet's hypothesis (Burnet 1971), was initially based on observations that the selective advantage of viviparity—parasitism of offspring on the parent, which occurs in many metazoans—requires relaxation and refinement of a simple self/non-self detection mechanism. Evolution of such a system may have been the first step toward development of adaptive immunity. Burnet's hypothesis continues to have support among contemporary immunologists (Cohn 1994; Janeway 1992; Klein 1982). Second, the MHC loci involves clustering of many genes essential to self/non-self discrimination. In and of itself, proximity of the genes results in decreased recombination among them, but there is also evidence for significant additional suppression of recombination in this region (Raymond et al. 2005). These properties of the MHC are relevant to our discussion since all confirmed cases of loci under "ancient" balancing selection in metazoans are (i) involved in self/non-self recognition and (ii) physically linked. The plant immune system, for example, although much less well understood, is another rich source of loci under ancient balancing selection having these two properties. (Boyes *et al.* 1997; Holub 2001; Meyers *et al.* 2005; Michelmore and Meyers 1998; Van der Hoorn *et al.* 2002).

A more dramatic example of a large haplotype under ancient balancing selection is the sex chromosome. Sex chromosomes can be thought of as one polymorphic locus with two macro alleles, one of which is fully dominant over the other. A heterozygote has one phenotype, one type of homozygote has another, and the remaining homozygote is usually lethal (in mammals, the phenotypes are male and female, respectively). The function of this 'locus' is not immediately obvious. Of course it is involved in maintenance of sexual reproduction, which is thought to be essential for preventing accumulation of deleterious mutations via Muller's ratchet (HAIGH 1978; MULLER 1964). However, the irony of the sex chromosome is that a byproduct of maintaining different genders is often generation of non-recombining genetic units like the Y-chromosome. There must be additional advantages to having separate genders.

Three principal theories attempt to explain the abundance of independently evolved systems of dioecy (existence of two genders) in metazoans. The first (Figure 3.3a) is the observation, predicted by theoreticians, that having gametes of different sizes ("anisogamy") tends to result in higher rates of fertilization among free-spawning organisms (BULMER and PARKER 2002; COX and SETHIAN 1984; DUSENBERY 2000; PARKER 1978; PARKER et al. 1972). Female and male individuals can be thought of, respectively, simply as specialized large and small gamete producers—a trait that is strongly under balancing selection. A second theory is that producing different gametes in different organisms is advantageous because it prevents the most severe form of inbreeding—selfing (Figure 3.3b). This theory is most abundantly explored in the plant literature, where there exist alternate and more explicit solutions to the selfing problem, the self-incompatibility (SI) systems, which also appear to be maintained by balancing selection (ANDERSON 1984; CHARLESWORTH 1985; CHARLESWORTH 2002; FREEMAN 1997). A third theory is that dioecy increases the fitness of the population by allowing it to exploit the environment more efficiently via sexual dimorphism (Figure 3.3c). While an environment with an expanding "niche width" can often lead to speciation (BOLNICK 2001; ROUGHGARDEN 1972; TURNER et al. 2001), specialization of different genders to different niches can allow one species to appropriate more resources without incurring the genetic load of generating a balancing selection system de novo (BOLNICK and DOEBELI 2003). All three of these rationales for the widespread occurrence of dioecy in metazoans appear to contribute to the breeding systems typically found in vertebrates.

Three other notable cases with credible evidence that observably polymorphic traits are being maintained by ancient balancing selection are the ADH locus in fruit flies (ASHBURNER 1998; GRELL 1965; HUDSON *et al.* 1987; KREITMAN 1983; STEPHENS and NEI 1985), the X-linked color vision locus in New World monkeys (see Figure 3.4; BOISSINOT *et al.* 1998; HUNT *et al.* 1998; JACOBS 1996; MORGAN *et al.* 1992), and the locus controlling ABO blood type in humans (KODA et al. 2000; SAITOU and YAMAMOTO 1997). However, in each of these cases, while the implicated locus has a significantly higher-than-average divergence level, it is difficult to get a reliable estimate of the tMRCA of the balanced alleles because the haplotypes have been decayed by recombination and/or obfuscated by gene conversion.

The majority of examples of more recent balancing selection systems are in human. The main reason for this is not that more SNP data is collected in humans than any other organism, but because the selected sites are associated with a disease (i.e. self-reporting) phenotype. The most obvious case of this is the hemoglobinopathy sickle-cell anemia (refs), but there are many other claims for balancing selection in the literature. The majority of them originate with an observation of high incidence of congenital disease in a certain population. Disease-causing mutations in the cystic fibrosis gene, for example, have an allele frequency of close to 4% in Caucasian populations. These deleterious alleles are thought to be at such high frequency because heterozygotes are resistant to typhoid (PIER et al. 1998; VAN DE VOSSE et al. 2005), a strong selective force in Europe at least since the time of the ancient Egyptians. Even more common, although less eminent because the resulting disease is manageable by phlebotomy alone, are alleles causing hereditary hemochromatosis. Among Caucasian Americans, approximately 4-7% of alleles are disease-causing (MOALEM et al. 2002). It is thought that heterozygotes, because they are more efficient at absorbing iron from their gut, are more resistant to anemia and enteric bacterial infections. The worldwide contender for most common, serious inherited disorder is congenital adrenal hyperplasia (21-hydroxylase mutation), with a disease-causing allele frequency of roughly 5% in all world populations studied to date (WITCHEL et al. 1997). While mutant-allele homozygotes have serious fitness problems, including ambiguous genitalia and decreased fertility due to altered hormone levels, clinical tests show that carriers have more rapid cortisol response than controls. Still another example that was discovered because of its involvement in pathogen resistance is the CCR5 gene; homozygotes for a null mutation in CCR5 (CCR5-Δ32) are highly resistant to infection by HIV-1. However, allele frequencies across different subpopulations indicate that balancing selection, rather than positive selection, is likely to be the major force acting on this locus. (BAMSHAD et al. 2002; LIBERT et al. 1998; WOODING et al. 2005). Challenges to selective interpretations of CCR5-Δ32 (SABETI *et al.* 2005) illustrate the inherent difficulty of reaching definitive conclusions.

There are other self-reporting cases of phenotypic variability in humans, not associated with genetic disease, but which may have genetic bases and are therefore suspected of being maintained by balancing selection. Many of these loci are not yet firmly associated with defined genes; examples include handedness (BILLIARD et al. 2005; FAURIE and RAYMOND 2005), mental illness (HARPENDING and COCHRAN 2002; WANG et al. 2004), athletic propensity (YANG et al. 2003), and homosexuality (CAMPERIO-CIANI et al. 2004)). In some cases, suspicions of balancing selection have been disproved (eg: prion proteins (BROOKFIELD 2003; KREITMAN and DI RIENZO 2004; MEAD et al. 2003; SEABURY et al. 2004; SOLDEVILA et al. 2005) and diurnal preference (NADKARNI et al. 2005)). There are undoubtedly additional examples of functional polymorphisms maintained by balancing selection among non-humans, but because we are less attuned to differences within members of other species, candidate traits are less readily identified than in humans.

The Role of Convergent Evolution in Balancing Selection

I would next like to argue that convergent evolution is a common accessory in balancing selection systems, and point out why recognition of this association could be useful.

Although the very notion of convergent evolution is imprecise (see Figure 3.5), there are two general types of convergent evolution. In one case, alterations of orthologous genes lead to convergent evolution of the same trait; in the other case, the process depends on changes in non-orthologous genes. In this section, I highlight how these two types of convergent evolution are frequently observed in recent and ancient cases of balancing selection, respectively.

The most apparent case of convergent evolution among instances of long-term balancing selection is with the sex chromosomes. As indicated above, sex chromosomes have evolved multiple times independently in both the plant and animal kingdoms (reviewed in FRASER and HEITMAN 2005). Self-incompatibility systems also arose multiple times and in multiple ways (FRANKLIN-TONG and FRANKLIN 2003; HISCOCK and MCINNIS 2003). Recently, the balanced alleles responsible for a trait long-suspected of being under ancient balancing selection based on trans-specific (human-chimpanzee) phenotypic polymorphism were shown to have arisen independently on different haplotypic backgrounds in each of the human and chimpanzee lineages (KIM and DRAYNA 2005; WOODING *et al.* 2006; WOODING *et al.* 2004). This example highlights the inherent risk of conflating ancient balancing selection with convergent evolution of more recently evolved balancing selection systems—particularly for small (in the limit, single nucleotide) haplotypes that are subject to significant recombinational decay over long periods of time (WIUF *et al.* 2004). As illustrated in Figure 3.5, both the locus controlling color vision in New World monkeys and the locus controlling ABO blood type may be such examples. Although each locus has evidence for long-term balancing selection based on polymorphism levels, as discussed above, the potential for extensive recombinational decay of the small balanced haplotypes (23 and 8 nucleotides, respectively) make it difficult to determine with certainty whether these balancing selection systems are trans-specific (BOISSINOT *et al.* 1998; MARTINKO *et al.* 1993; O'HUIGIN *et al.* 1997; SHYUE *et al.* 1998).

An interesting consequence of convergent evolution of systems under balancing selection, particularly in the case of sex chromosomes, is that we can see evolution in action—there are extant organisms in various phases of sex chromosome evolution (see Figure 3.6, taken from VYSKOT and HOBZA 2004). What lessons can be extracted from these examples? For one thing, most examples of evolving sex chromosomes are consistent with a theoretical prediction that functional units under compatible balancing selection (e.g., NAVARRO and BARTON 2002; SLATKIN 2000) will become linked, with suppression of recombination between balanced alleles, resulting in large blocks of high nucleotide divergence. More generally, and probably more importantly, the evolution of sex chromosomes illustrates the importance of looking for genetic characteristics of a locus—such as linked highly-polymorphic functional elements and suppressed recombination—rather than simply looking for direct orthologs, suggesting a predictive tool for finding new loci under balancing selection.

The best illustration of the power of this tool is the recent discovery of a histocompatibility locus in a protochordate. The major histocompatibility locus has well-established role in adaptive immunity of jawed vertebrates, but as its name suggests, it was first discovered via tissue transplant rejection (MARSH 2000). While no homolog for the MHC has been found outside of jawed vertebrates, there is good evidence for the existence of adaptive immunity outside the jawed-vertebrate clade (ALDER *et al.* 2005; PANCER 2000). Recently, a highly polymorphic histocompatibility locus—with a polymorphic receptor less than 200 kbp away—has been mapped in a protochordate (DE TOMASO et al. 2005). Although the protochordate histocompatibility protein is structurally similar to MHC proteins, the genes are not orthologous, suggesting independent (i.e. convergent) evolution.

It seems entirely possible that mechanisms for both prevention of selfing (sex chromosomes and SI systems) and combating pathogens may be of such importance that each has arisen multiple times in metazoans; however, there could be a lingering molecular interplay between these two mechanisms, as implied by Burnet's hypothesis (BURNET 1971). A role for the MHC in reproduction occurs at many levels, including mate choice (EGGERT *et al.* 1998; JORDAN and BRUFORD 1998), gamete-directed immune response (BOHRING and KRAUSE 2005; OHL and NAZ 1995), and actual sperm-egg binding (MORI et al. 2000). Fernandez et al. provided a good review of these topics (FERNANDEZ *et al.* 1999).

Among the examples of short-term balancing selection, the best known case of convergent evolution involves the multiple independent origins of alleles that are protective against malaria in heterozygous state, but cause hemoglobinopathies (e.g., sickle-cell and multiple types of thalassemia) when homozygous. Figure 3.7 (taken from VOGEL 1997) illustrates the geographic distribution of some of the alleles that appear to fit this model (for more comprehensive reviews, see FIX 2003; LIVINGSTONE 1984). Other examples are less clear cut, but the theme of convergent evolution persists. Among the examples cited above, hereditary hemochromatosis can be caused by compound heterozygosity for the major 'bad' allele and another allele, H63D, which appears to have arisen multiple times on separate haplotypes (ROCHETTE et al. 1999). In congenital adrenal hyperplasia, although a flanking pseudogene is a constant source of mutation via gene conversion, each subpopulation seems to have one common mutant allele. This pattern suggests that mutant alleles are not common simply because of recurrent mutation but because of separate episodes of positive selection. Even at the cystic fibrosis gene, although there is one common mutant allele, ΔF508, that accounts for close to 70% of all mutant alleles among Europeans, other mutant alleles have risen to frequencies as high as 36% in certain subpopulations, suggestive of independent selection events (ESTIVILL et al. 1997). In other less well studied systems that may have undergone balancing selection, it is unclear whether high occurrence of a 'balanced' phenotype in various subpopulations indicates convergent or divergent evolution.

One striking similarity between mutant alleles of more recently evolved systems of balancing selection is that most of these derived, balanced alleles are non- or sub-functional. This observation suggests that many examples of balancing selection conform to the 'less-is-more' hypothesis, which argues that rapid evolution is most easily, and commonly, accomplished through loss-of-function mutations (OLSON 1999). Balancing selection that maintains both one or more loss-of-function mutations and an ancestral functional allele can be a "stopgap" evolutionary measure until mutations arise that confer the selective advantages of balanced loss-of-function alleles without conferring the associated genetic load. Two prominent examples that appear to fit this model involve the Hb and CCR5 loci. While multiple sub-functional hemoglobin alleles have arisen in various subpopulations, there is one allele, HbC, that appears to be both fully functional and also to confer resistance to complications from malaria (MODIANO et al. 2001). At the CCR5 locus, many of the more common haplotypes result in decreased protein expression, associated with slower AIDS progression. In chimpanzees however, which become infected with HIV-1 but do not progress to AIDS, there is evidence for positive selection on one haplotype. (WOODING et al. 2005). The predominant chimpanzee haplotype, which corresponds to the "ancestral" haplotype in humans, is somewhat protective against progression to AIDS in African-Americans, but not in Caucasians (GONZALEZ et al. 1999). One plausible explanation for the distribution of human and chimpanzee genotypes at the CCR5 locus is that the ancestral haplotype acquired an optimal mutation on the chimpanzee lineage that swept through the entire population (WOODING et al. 2005).

The transient nature of balancing selection is predicted by theory. When a new allele is generated, the rate at which it is expected to rise in frequency initially depends almost entirely on the fitness of the heterozygote. Therefore, a new allele that confers high fitness in the heterozygous state could propagate—and reach equilibrium frequency—more quickly than a new allele that has lower fitness in heterozygotes but will, when fixed, confer a higher fitness to the population as a whole. This effect can be shown with simulations, as illustrated in Figure 3.8.

Although probably more common among cases of transient balancing selection, there are some examples of apparent long-term maintenance of one functional and one non- or sub-functional allele. In plants, for example, a main mechanism for generating males is by making hermaphrodites 'female-sterile' (FREEMAN 1997). Also, many of the presence/absence polymorphisms in plant resistance genes appear to be of ancient origin (SHEN et al. 2006).

How Will (and Why Should) We Find More Cases of Balancing Selection?

There are three major ways of looking for selection, (i) mapping of obviously polymorphic traits and genetic diseases of high frequency (e.g., De Tomaso et al. 2005; Kim et al. 2003), (ii) re-sequencing candidate genes and examining them for signs of selection (e.g., Akey et al. 2004; Livingston et al. 2004), (iii) scanning genomes without using prior annotation knowledge (e.g., Beaumont and Balding 2004; Bubb et al. 2006; Carlson et al. 2005; Cork and Purugganan 2005; Voight et al. 2006). Each of these methods has its own strengths and weaknesses.

Although instinctively appealing, mapping the genes responsible for traits that are known to be polymorphic in multiple subpopulations, or that are suspected to have a high minor-allele frequency because of a high incidence of an associated genetic disease in certain subpopulations, has multiple problems. First, it is often unclear whether the observed phenotypic diversity has genetic or environmental causes. Even if a genetic basis for the diversity can be established, it is difficult to map traits whose inheritance patterns are not Mendelian or “near Mendelian.” Finally, it is tautological to perform tests for selection, particularly those based on allele frequencies, on loci associated with traits chosen for their high degree of phenotypic polymorphism.

Testing candidate genes for signatures of selection by re-sequencing them in multiple haplotypes is an appealingly direct method of detecting selection, and indeed, has had some success (ref?). Although the functions of candidate genes, by definition, are typically known, there is usually not prior phenotypic association with variant alleles. Hence, convincing evidence for selection must come from accompanying epidemiological studies, which are often of unknown relevance to post-environmental conditions. More basically, this method cannot identify truly novel selective processes, including those acting on genes of unknown function, which at this point account for a large fraction of any genome.

Whole-genome scans for regions with a signature consistent with balancing selection have the obvious advantage of being able to identify novel loci. However, a major drawback to whole-genome scanning is the difficulty of applying appropriate corrections for multiple-hypothesis testing; most eukaryotic genomes are large, presenting a lot of potential for false positives. And, of course, it is uncertain wheter or not recognizable phenotypes can be associated with newly identified loci, even if the genotypic evidence for balancing selection is quite strong.

It may be advantageous to build on the lessons of the examples discussed above when looking for new instances of balancing selection. For example, long-term balancing selection appears often to (i) affect self/non-self recognition, and (ii) involve functionally related genes that are physically linked, often accompanied by high polymorphism and/or suppressed recombination. Hence, perhaps these characteristics of genetic loci should be targeted during new searches for the effects of balancing selection. An example of a practical application of such analysis may be to elucidate the basis of compatibility between 'scion' (branch) and 'rootstock' (host tree) in plant grafting, which is central to fruit-tree agriculture. Unlike in humans or protochordates, where histocompatibility is rapidly apparent, 'incompatibility' between scion and rootstock can take years to appear. Tests such as tensile strength of a graft are inconsistent measures of compatibility: one union may produce fruit bountifully yet snap in a strong wind, while another union may never produce fruit but may be structurally sound. The molecular mechanisms of compatibility, including whether or not they actually exist, are wholly unknown (MOORE 1984; PEDERSEN 2005). These factors, in addition to the long generation time of trees, suggest that a tailored genomic analysis may be more productive than traditional genetic studies.

Similarly, because more recently evolved systems of balancing selection often involve maintenance of a non- or sub-functional allele, it may be advantageous to look for these properties alone, particularly if such alleles are present at a significant population frequency (CONRAD *et al.* 2006; HINDS *et al.* 2006; MCCARROLL *et al.* 2006). There is a practical interest in such searches, since when hypomorphic or null alleles confer some health advantage, they may suggest novel drug targets. It is typically easier to mimic loss-of-function mutations pharmacologically than to create new functions or compensate for deleterious loss-of-function mutations. This approach is already being attempted with antagonists to CCR5 (e.g., RUFF et al. 2003), and to some extent in experimental malaria vaccines (KWIATKOWSKI 2005; SMITH *et al.* 2002; YAZDANI *et al.* 2004).

Of course, finding novel instances of balancing selection is also of purely theoretical interest. Recent examples might afford us a better understanding of the details of gene annotation (i.e. how many ways can you break a gene?). Ancient examples might illuminate multiple ways of evolving new function (i.e. how many genes can you use to perform a given function?). It seems likely to be worthwhile to skim the cream of novel balancing selection examples from different species before investing too heavily in human. Learning about balancing selection in other organisms may even be of ultimate use to humans because of convergent evolutionary systems regarding common pathogens, but also just regarding our dependence on the natural world—for example, through agriculture. Perhaps I should end this discussion with a precaution against ‘seeing’ rampant selection where it does not exist, a common error before the neutral theory gained force (KREITMAN and DI RIENZO 2004). Particularly in responding to the current deluge of nucleotide sequence data, there is reason to be cautious in perceiving evidence for any type of selection.

Coalescence theory provides an excellent way to approximate the expected time to most recent common ancestor (tMRCA) of any sample of neutral haplotypes in a fixed-size population. The theory relies on a population model in which individual haplotypes occur in discrete generations each having an equal probability of contributing to the next generation, as illustrated in the left panel; the same evolutionary tree is shown in the right panel in a more familiar format. The number of haplotypes in such a model that most closely approximates the demography of an actual diploid population is 2Ne, since each diploid individual has two haplotypes. As indicated in the left panel, most haplotypes in previous generations are not ancestors of extant haplotypes.

In the simplest case of exactly two haplotypes, the probability that both have the same ancestor of the $2N_e$ haplotypes in the previous generation is simply $1/2N_e$. The expected tMRCA for two alleles, is the reciprocal—$2N_e$. The expected tMRCA for k alleles sampled from a population with $2N_e$ haplotypes approaches $4N_e$ and can be derived as follows.

In any given generation, the probability of coalescence between any two haplotypes is:

$$1 – \mathrm{Pr(no\ coalescences)}$$

Pr(no coalescences) is the probability that the second haplotype did not come from the same ancestor as the first, and the third haplotype did not come from the same ancestor as either the first or second, etc…:

$$(1 - 1/2N_e) \times (1 - 2/2N_e) \times (1 - 3/2N_e) \times \ldots \times (1 - (k-1)/2N_e)$$

Ignoring higher-order terms, this product simplifies to:

$$1 – k(k-1)/4N_e$$

So the probability of coalescence between any two ancestral samples in a given generation is:

$$1 – (1 – k(k-1)/4N_e) \ = \ k(k-1)/4N_e$$

Therefore, the expected time until the next coalescent event, given k samples, is:

$$4N_e/k(k-1)$$

To get the tMRCA, the time between all coalescent events must be summed:

$$4N_e/k(k-1) + 4N_e/(k-1)(k-2) + 4N_e/(k-2)(k-3) + \ldots + 4N_e/2$$

This can be simplified to:

$$4N_e\ (1 - 1/k)$$

As k increases, this rapidly approaches $4N_e$.

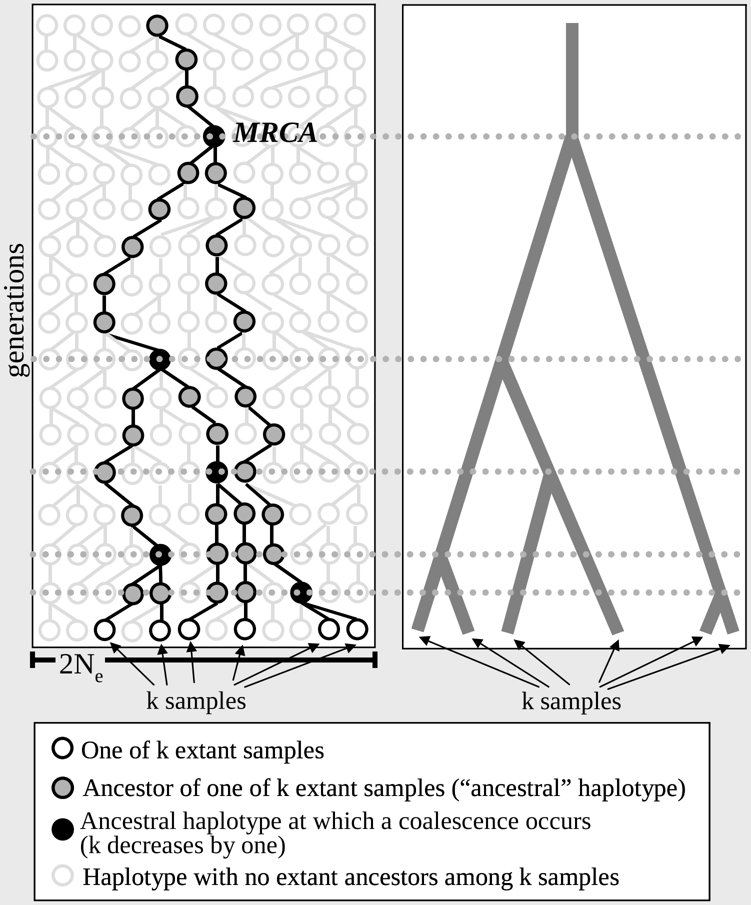
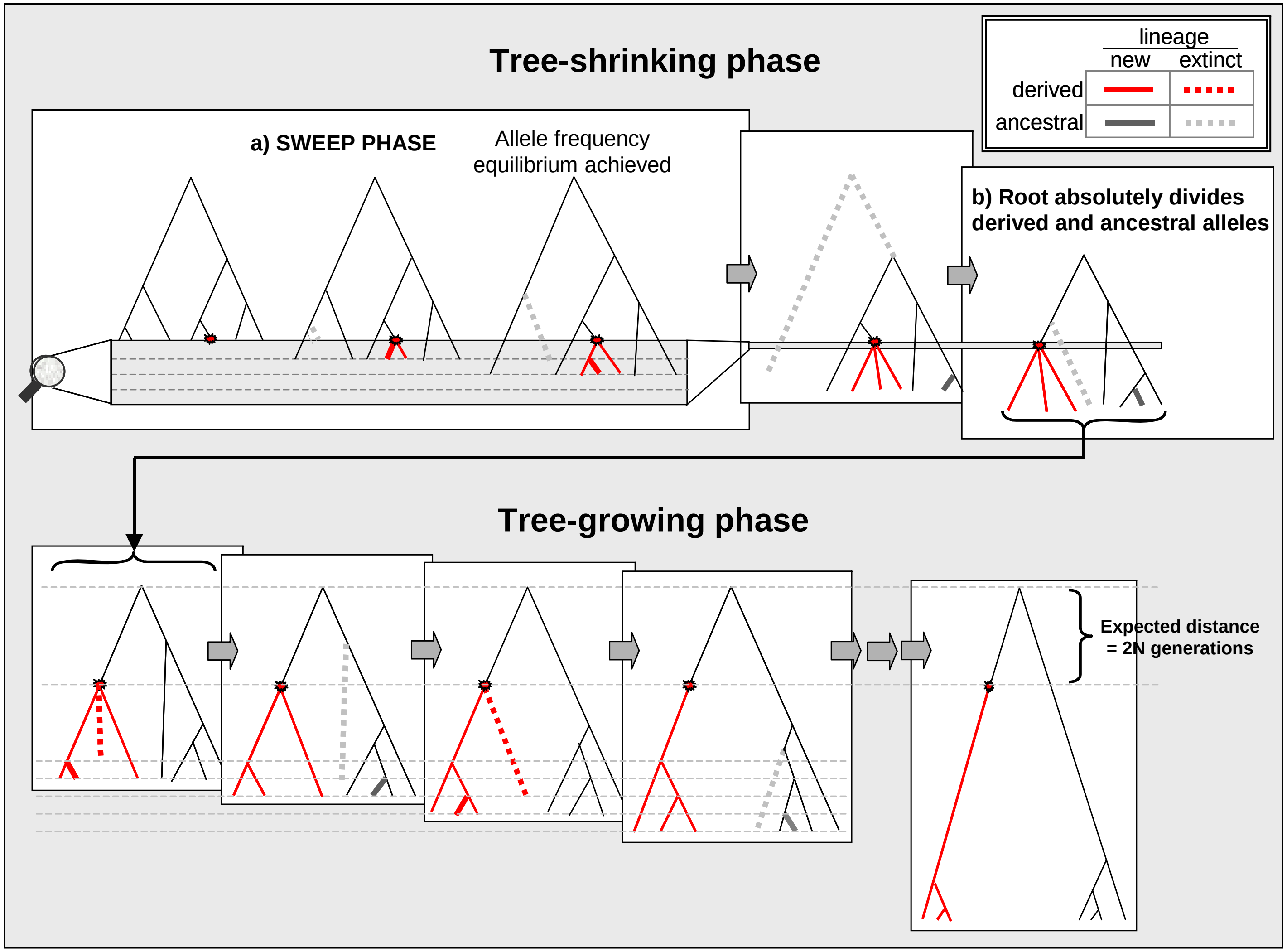

MRCA
generations
2Ne
k samples
k samples
○ One of k extant samples
◍ Ancestor of one of k extant samples (“ancestral” haplotype)
● Ancestral haplotype at which a coalescence occurs
(k decreases by one)
○ Haplotype with no extant ancestors among k samples

FIGURE 3.1. Micro-review of coalescent theory.

FIGURE 3.2. Effect of a new balanced mutation on tree shape and depth. This schematic illustrates how a new mutation that initiates a system of balancing selection would be expected to affect a previously neutral tree over time. In the initial ‘sweep’ phase (a), the new, ‘derived’ allele rapidly rises in frequency (depending mainly on the fitness advantage of the heterozygote), displacing some of the older, ‘ancestral’ alleles. The grey box shows a zoomed-in portion of the evolving trees. This process can eliminate some more distantly related ancestral clades, such that when the equilibrium frequency of the derived allele is reached, the new root of the tree is much more recent than it was prior to the ‘sweep’ phase. Soon after equilibrium allele frequencies are reached, one branch coming off the root leads exclusively to derived alleles, the other to ancestral alleles (b). Thereafter, the derived and ancestral branches diverge as long as the system of balancing selection is maintained, with neutral mutation and drift occurring independently within each clade. The branch connecting the newly arisen derived allele to the root can be thought of as a randomly drawn sample from the original neutral tree. Hence, the expected time from the new mutation to the root is approximately $2N_e$ generations, as it would be for any two randomly chosen alleles, where N_e is the effective population size.


Tree-shrinking phase
lineage
new
extinct
derived
ancestral
a) SWEEP PHASE
Allele frequency
equilibrium achieved
b) Root absolutely divides
derived and ancestral alleles
Tree-growing phase
Expected distance
= 2N generations

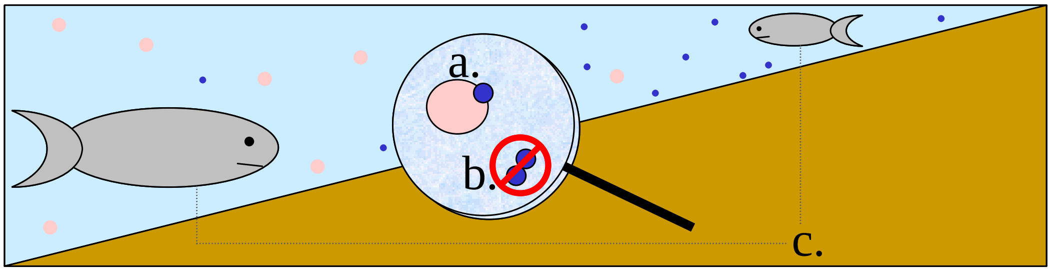


FIGURE 3.3.—Illustration of three selective advantages to dioecy. (a) Different sized gametes ("anisogamy") increases rate of fertilization (ref). (b) Same-gamete-type incompatibility prevents severest form of inbreeding—selfing. (c) Sexual dimorphism allows one species to optimize for two different niches.

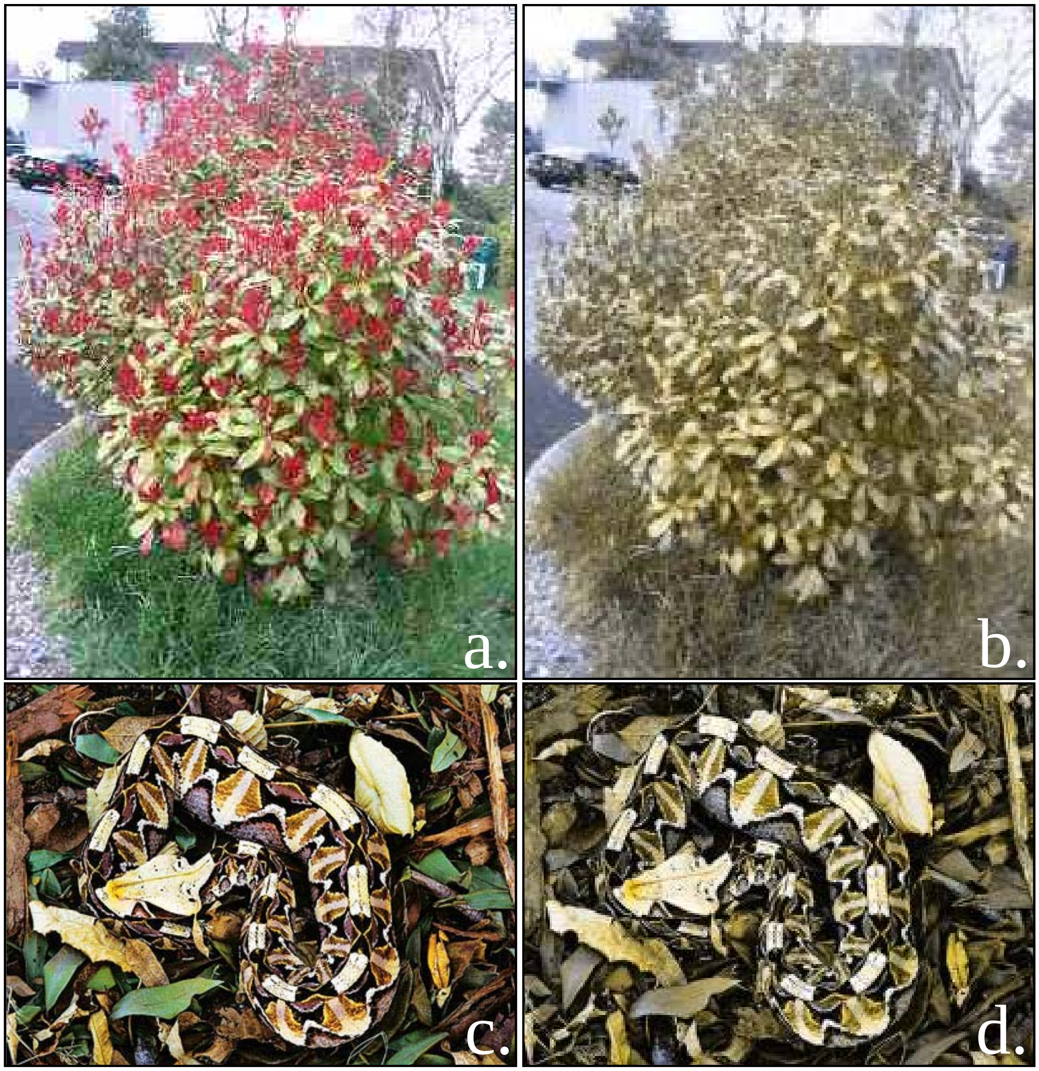
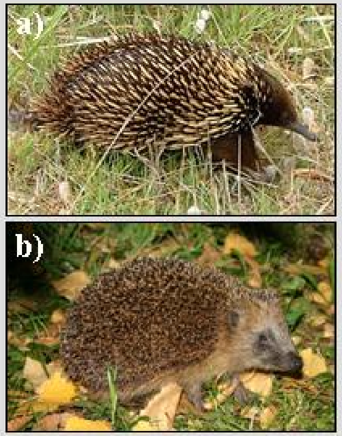
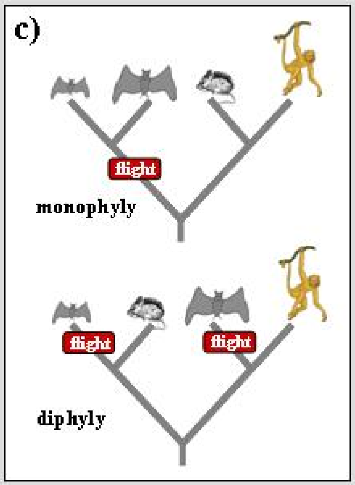


FIGURE 3.4.—Suspected disadvantage and advantage of color-blindness. (a) Full-color image of a bush with new red leaves, which are more tender and nutritious. (b) Image of bush as seen by an individual with color-blindness (deuteranopia). (c) Full-color image of a snake among leaves. (d) Image of snake as seen by an individual with color-blindness (deuteranopia). Color-blind individuals are thought to see camouflaged objects better than fully trichromatic individuals (ref), perhaps because they are less visually distracted by patterns similarities between foreground and background images. All color-blind images were generated at http://www.etre.com/tools/colourblindsimulator/.



Convergent evolution is a general term describing the phenomenon similar traits evolving independently in different evolutionary lineages. Originally, it was detected because distantly related organisms often have similar characteristics, ranging from general body form (for example, (a) the marsupial echidna and (b) the placental hedgehog) to very specific features, such as eye lenses or wings. Convergent evolution was the major confounding factor when building evolutionary trees in the pre-molecular era. For example, as recently as the 1990's, there was debate as to whether frugivorous, non-echolocating 'megabats' and insectivorous, echolocating 'microbats' were monophyletic or diphyletic (c). Recent molecular evidence indicates that bats are indeed monophyletic (i.e. bat morphology did not evolve twice). More surprisingly, megabats and microbats are not even discrete clades—echolocation has evolved multiple times via convergent evolution.



Molecular technology has resolved much of the early phylogenic confusion. At a molecular level, we can more precisely define two types of convergent evolution. In the first type, different genes evolve similar function. An example of this process is the recently discovered histocompatibility locus in protochordates that, while non-orthologous to the MHC genes in vertebrates, is likewise responsible for tissue rejection among incompatible individuals. In the second type, orthologous genes evolve the same function via different mutations. A within-species example of this process is the array of hemoglobinopathy-related alleles that have arisen in response to selective pressure from malaria. A recent cross-species example appears to be the bitter-taste receptor (TAS2R) in humans and chimpanzees.

There are still ambiguous cases in which, despite extensive molecular evidence, we cannot be certain whether the functional mutation arose multiple times (convergently) or once in a common ancestor. Prominent examples include the gene underlying the ABO blood groups in Great Apes (d) and the gene coding for red or green color vision in New World monkeys (e). These two examples both depend on functional polymorphisms in which phenotype is determined by specific tightly linked mutations (represented by white and black balls in the illustrations) that could have either arisen independently, along different lineages, or in a common ancestor. In cases such as these, flanking neutral mutations have not allowed reliable dating of the functional mutations since over time, recombination has eroded relevant genomic linkage disequilibrium in the regions.

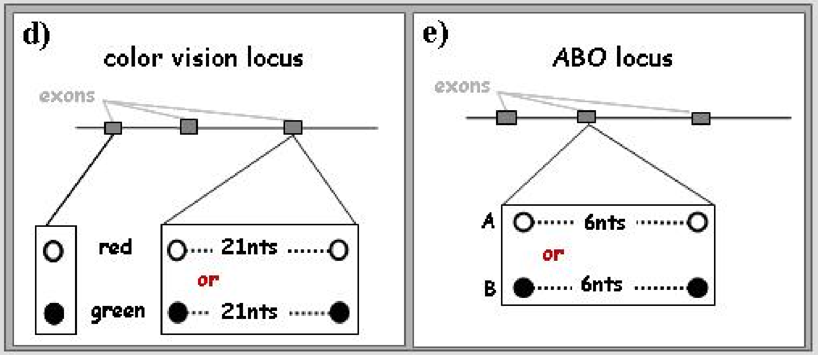


FIGURE 3.5. Micro-review of convergent evolution.

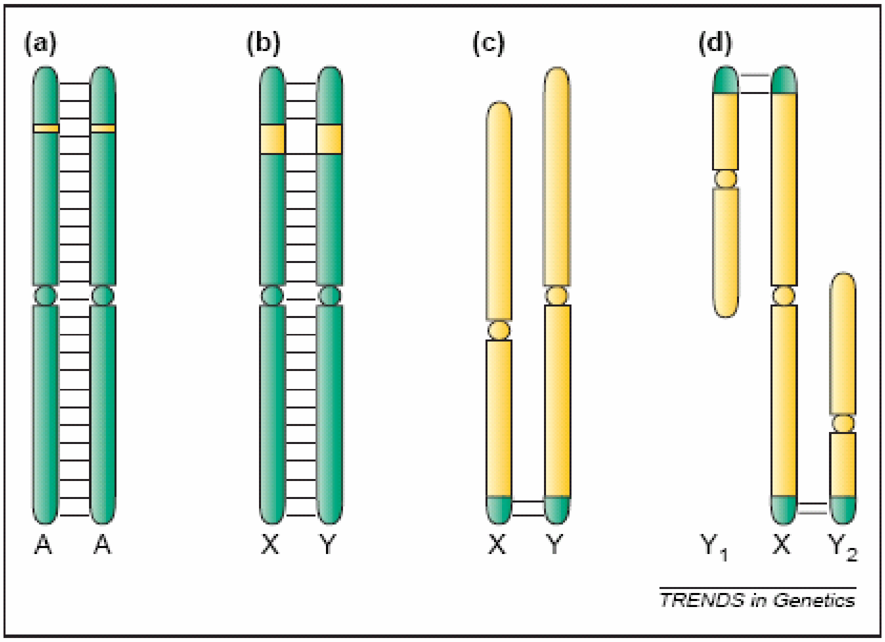


FIGURE 3.6.— Main steps in the evolution of sex chromosomes (taken directly from VYSKOT and HOBZA 2004): “A sketch of the four main steps in the evolution of plant sex chromosomes, from the most primitive (left) to evolutionarily advanced and degenerate (right). **(a)** Squirting cucumber (Ecballium elaterium) with single locus-based sex determination and no observed blockage of genetic recombination observed. ‘A’ stands for autosomes. **(b)** Papaya (Carica papaya) possesses homomorphic sex chromosomes X and Y, with a short non-recombining region on the Y chromosome, MSY. **(c)** White campion (Silene latifolia) has large sex chromosomes with a Y that is largely non-recombining but looks euchromatic. **(d)** Sorrel (Rumex acetosa) has polymorphic sex chromosomes with two different Y chromosomes, both are constitutively heterochromatic.”

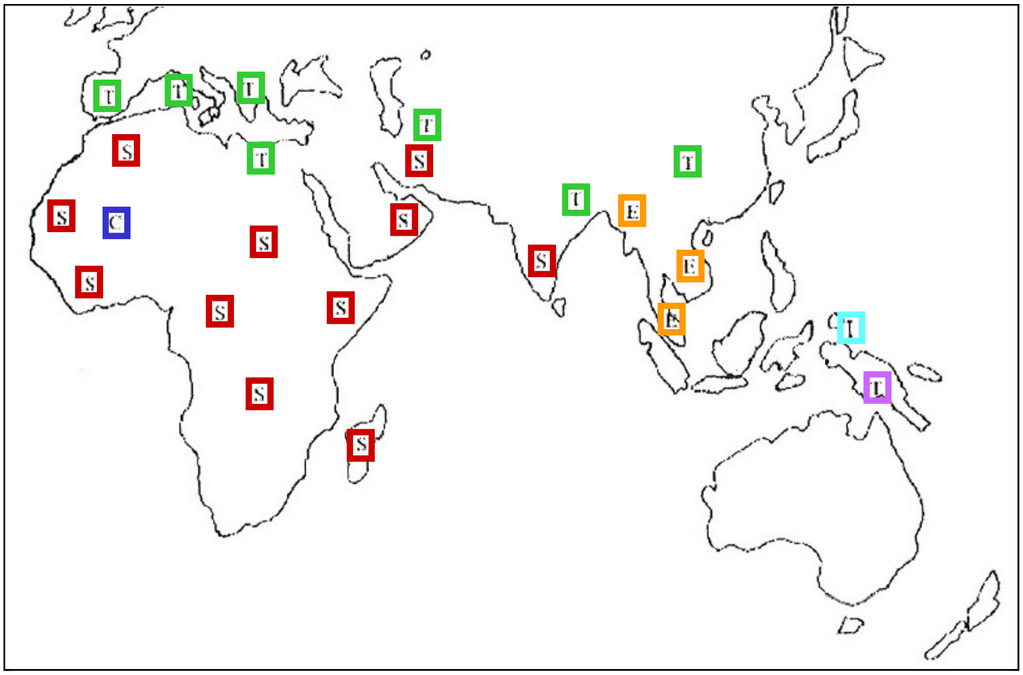


FIGURE 3.7.—World distribution of mutant beta-globin alleles (taken from VOGEL 1997). In homozygous state, each of these alleles results in some form of hemoglobinopathy, with the exception of the Hb-C allele, boxed in dark blue.

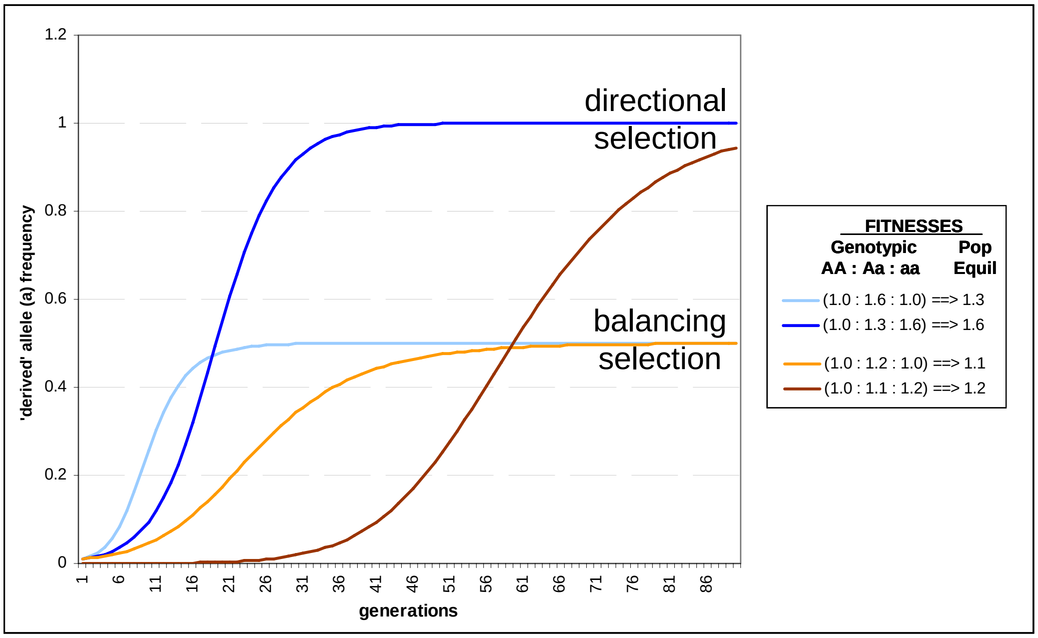


FIGURE 3.8.—Expected allele frequency change for newly generated alleles under additive directional selection and symmetric balancing selection. Under additive directional selection, genotypes (AA : Aa : aa) have the following fitness ratios (1 : 1+x : 1+2x); under symmetric balancing selection, the ratio is (1 : 1+x : 1). The blue and brown sets of curves, respectively, describe scenarios in which the maximal genotypic fitness is 60% or 20% greater than the minimal genotypic fitness. Although the directional cases result in a higher population-equilibrium fitness, the rate at which a new derived allele increases in frequency is entirely driven by the magnitude of heterozygous advantage.

BIBLIOGRAPHY

AGUILAR, A., G. ROEMER, S. DEBENHAM, M. BINNS, D. GARCELON *et al.*, 2004 High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. Proc Natl Acad Sci U S A **101:** 3490-3494.

AKEY, J. M., M. A. EBERLE, M. J. RIEDER, C. S. CARLSON, M. D. SHRIVER *et al.*, 2004 Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol **2:** e286.

AKEY, J. M., G. ZHANG, K. ZHANG, L. JIN and M. D. SHRIVER, 2002 Interrogating a high-density SNP map for signatures of natural selection. Genome Res **12:** 1805-1814.

ALDER, M. N., I. B. ROGOZIN, L. M. IYER, G. V. GLAZKO, M. D. COOPER *et al.*, 2005 Diversity and function of adaptive immune receptors in a jawless vertebrate. Science **310:** 1970-1973.

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS and D. J. LIPMAN, 1990 Basic local alignment search tool. J Mol Biol **215:** 403-410.

ARAKI, H., D. TIAN, E. M. GOSS, K. JAKOB, S. S. HALLDORSDOTTIR *et al.*, 2006 Presence/absence polymorphism for alternative pathogenicity islands in Pseudomonas viridiflava, a pathogen of Arabidopsis. Proc Natl Acad Sci U S A **103:** 5887-5892.

ARANZANA, M. J., S. KIM, K. ZHAO, E. BAKKER, M. HORTON *et al.*, 2005 Genome-Wide Association Mapping in Arabidopsis Identifies Previously Known Flowering Time and Pathogen Resistance Genes. PLoS Genet **1:** e60.

ASHBURNER, M., 1998 Speculations on the subject of alcohol dehydrogenase and its properties in Drosophila and other flies. Bioessays **20:** 949-954.

ASTHANA, S., S. SCHMIDT and S. SUNYAEV, 2005 A limited role for balancing selection. Trends Genet **21:** 30-32.

AUSUBEL, F. M., 2005 Are innate immune signaling pathways in plants and animals conserved? Nat Immunol **6:** 973-979.

BAMSHAD, M. J., S. MUMMIDI, E. GONZALEZ, S. S. AHUJA, D. M. DUNN *et al.*, 2002 A strong signature of balancing selection in the 5' cis-regulatory region of CCR5. Proc Natl Acad Sci U S A **99:** 10539-10544.

BARRETT, S. C., 2002 The evolution of plant sexual diversity. Nat Rev Genet **3:** 274-284.

BARTON, N. H., and A. NAVARRO, 2002 Extending the coalescent to multilocus systems: the case of balancing selection. Genet Res **79:** 129-139.

BAUM, J., R. H. WARD and D. J. CONWAY, 2002 Natural selection on the erythrocyte surface. Mol Biol Evol **19:** 223-229.

BEAUMONT, M. A., and D. J. BALDING, 2004 Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol **13:** 969-980.

BEAUMONT, M. A., W. ZHANG and D. J. BALDING, 2002 Approximate Bayesian computation in population genetics. Genetics **162:** 2025-2035.

BERNATCHEZ, L., and C. LANDRY, 2003 MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? J Evol Biol **16:** 363-377.

BERTONE, P., V. STOLC, T. E. ROYCE, J. S. ROZOWSKY, A. E. URBAN *et al.*, 2004 Global identification of human transcribed sequences with genome tiling arrays. Science **306:** 2242-2246.

BEYRER, C., A. W. ARTENSTEIN, S. RUGPAO, H. STEPHENS, T. C. VANCOTT *et al.*, 1999 Epidemiologic and biologic characterization of a cohort of human immunodeficiency virus type 1 highly exposed, persistently seronegative female sex workers in northern Thailand. Chiang Mai HEPS Working Group. J Infect Dis **179:** 59-67.

BILLIARD, S., C. FAURIE and M. RAYMOND, 2005 Maintenance of handedness polymorphism in humans: a frequency-dependent selection model. J Theor Biol **235:** 85-93.

BIRD, A. P., 1980 DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res **8:** 1499-1504.

BLANCHER, A., M. E. REID and W. W. SOCHA, 2000 Cross-reactivity of antibodies to human and primate red cell antigens. Transfus Med Rev **14:** 161-179.

BODMER, W. F., 1962 Linkage and recombination in evolution. Advances in Genetics.

BOHRING, C., and W. KRAUSE, 2005 The role of antisperm antibodies during fertilization and for immunological infertility. Chem Immunol Allergy **88:** 15-26.

BOISSINOT, S., Y. TAN, S. K. SHYUE, H. SCHNEIDER, I. SAMPAIO *et al.*, 1998 Origins and antiquity of X-linked triallelic color vision systems in New World monkeys. Proc Natl Acad Sci U S A **95:** 13749-13754.

BOLNICK, D. I., 2001 Intraspecific competition favours niche width expansion in Drosophila melanogaster. Nature **410:** 463-466.

BOLNICK, D. I., 2004 Can intraspecific competition drive disruptive selection? An experimental test in natural populations of sticklebacks. Evolution Int J Org Evolution **58:** 608-618.

BOLNICK, D. I., and M. DOEBELI, 2003 Sexual dimorphism and adaptive speciation: two sides of the same ecological coin. Evolution Int J Org Evolution **57:** 2433-2449.

BOYES, D. C., M. E. NASRALLAH, J. VREBALOV and J. B. NASRALLAH, 1997 The self-incompatibility (S) haplotypes of Brassica contain highly divergent and rearranged sequences of ancient origin. Plant Cell **9:** 237-247.

BROOKFIELD, J. F., 2003 Human evolution: a legacy of cannibalism in our genes? Curr Biol **13:** R592-593.

BUBB, K. L., D. BOVEE, D. BUCKLEY, E. HAUGEN, M. KIBUKAWA *et al.*, 2006 Scan of human genome reveals no new loci under ancient balancing selection. Genetics.

BUJAS, M., B. BERIC and A. KAPAMADZIJA, 1988 Incidence of infertility of immune origin in a group of marriages with unexplained infertility. Hum Reprod **3:** 301-302.

BULMER, M. G., and G. A. PARKER, 2002 The evolution of anisogamy: a game-theoretic approach. Proc Biol Sci **269:** 2381-2388.

BURGOYNE, P. S., 1998 The mammalian Y chromosome: a new perspective. Bioessays **20:** 363-366.

BURNET, F. M., 1971 "Self-recognition" in colonial marine forms and flowering plants in relation to the evolution of immunity. Nature **232:** 230-235.

BUSACCA, M., F. FUSI, C. BRIGANTE, N. DOLDI, M. SMID *et al.*, 1989 Evaluation of antisperm antibodies in infertile couples with immunobead test: prevalence and prognostic value. Acta Eur Fertil **20:** 77-82.

CAICEDO, A. L., B. A. SCHAAL and B. N. KUNKEL, 1999 Diversity and molecular evolution of the RPS2 resistance gene in Arabidopsis thaliana. Proc Natl Acad Sci U S A **96:** 302-306.

CAMPERIO-CIANI, A., F. CORNA and C. CAPILUPPI, 2004 Evidence for maternally inherited factors favouring male homosexuality and promoting female fecundity. Proc Biol Sci **271:** 2217-2221.

CARLSON, C. S., D. J. THOMAS, M. A. EBERLE, J. E. SWANSON, R. J. LIVINGSTON *et al.*, 2005 Genomic regions exhibiting positive selection identified from dense genotype data. Genome Res **15:** 1553-1565.

CASTRIC, V., and X. VEKEMANS, 2004 Plant self-incompatibility in natural populations: a critical assessment of recent theoretical and empirical advances. Mol Ecol **13:** 2873-2889.

CHARLESWORTH, B., and D. CHARLESWORTH, 1997 Rapid fixation of deleterious alleles can be caused by Muller's ratchet. Genet Res **70:** 63-73.

CHARLESWORTH, B., and D. CHARLESWORTH, 1999 The genetic basis of inbreeding depression. Genet Res **74:** 329-340.

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

CHARLESWORTH, D., 2002 Plant sex determination and sex chromosomes. Heredity **88:** 94-101.

CHARLESWORTH, D., 2004 Sex determination: balancing selection in the honey bee. Curr Biol **14:** R568-569.

CHARLESWORTH, D., 2006 Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet **2:** e64.

CHARLESWORTH, D., C. BARTOLOME, M. H. SCHIERUP and B. K. MABLE, 2003 Haplotype structure of the stigmatic self-incompatibility gene in natural populations of Arabidopsis lyrata. Mol Biol Evol **20:** 1741-1753.

CHARLESWORTH, D., B. CHARLESWORTH and G. MARAIS, 2005 Steps in the evolution of heteromorphic sex chromosomes. Heredity **95:** 118-128.

CHARLESWORTH, D., B. K. MABLE, M. H. SCHIERUP, C. BARTOLOME and P. AWADALLA, 2003 Diversity and linkage of genes in the self-incompatibility gene family in Arabidopsis lyrata. Genetics **164:** 1519-1535.

CHEN, F. C., and W. H. LI, 2001 Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet **68:** 444-456.

THE CHIMPANZEE SEQUENCING AND ANALYSIS CONSORTIUM, 2005 Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437:** 69-87.

CHRISTENSEN, G. L., I. P. IVANOV, S. P. WOODING, J. F. ATKINS, A. MIELNIK *et al.*, 2006 Identification of polymorphisms and balancing selection in the male infertility candidate gene, ornithine decarboxylase antizyme 3. BMC Med Genet **7:** 27.

CLARK, V. J., N. J. COX, M. HAMMOND, C. L. HANIS and A. D. RIENZO, 2005 Haplotype structure and phylogenetic shadowing of a hypervariable region in the CAPN10 gene. Hum Genet **117:** 258-266.

COHN, M., 1994 The wisdom of hindsight. Annu Rev Immunol **12:** 1-62.

CONRAD, D. F., T. D. ANDREWS, N. P. CARTER, M. E. HURLES and J. K. PRITCHARD, 2006 A high-resolution survey of deletion polymorphism in the human genome. Nat Genet **38:** 75-81.

COOP, G., and R. C. GRIFFITHS, 2004 Ancestral inference on gene trees under selection. Theor Popul Biol **66:** 219-232.

CORK, J. M., and M. D. PURUGGANAN, 2005 High-diversity genes in the Arabidopsis genome. Genetics **170:** 1897-1911.

COX, P. A., and J. A. SETHIAN, 1984 Search, encounter rates, and the evolution of anisogamy. Proc Natl Acad Sci U S A **81:** 6078-6079.

CROSBY, J. L., 1973 *Computer simulation in genetics*. Wiley, New York.

CROW, J. F., 1958 Some possibilities for measuring selection intensities in man. Hum Biol **30:** 1-13.

CULLEN, M., S. P. PERFETTO, W. KLITZ, G. NELSON and M. CARRINGTON, 2002 High-resolution patterns of meiotic recombination across the human major histocompatibility complex. Am J Hum Genet **71:** 759-776.

THE CYSTIC FIBROSIS GENETIC ANALYSIS CONSORTIUM, 1994 Population variation of common cystic fibrosis mutations. Hum Mutat **4:** 167-177.

DANCHIN, E., V. VITIELLO, A. VIENNE, O. RICHARD, P. GOURET *et al.*, 2004 The major histocompatibility complex origin. Immunol Rev **198:** 216-232.

DANCHIN, E. G., L. ABI-RACHED, A. GILLES and P. PONTAROTTI, 2003 Conservation of the MHC-like region throughout evolution. Immunogenetics **55:** 141-148.

DANCHIN, E. G., P. GOURET and P. PONTAROTTI, 2006 Eleven ancestral gene families lost in mammals and vertebrates while otherwise universally conserved in animals. BMC Evol Biol **6:** 5.

DANCHIN, E. G., and P. PONTAROTTI, 2004 Statistical evidence for a more than 800-million-year-old evolutionarily conserved genomic region in our genome. J Mol Evol **59:** 587-597.

DANCHIN, E. G., and P. PONTAROTTI, 2004 Towards the reconstruction of the bilaterian ancestral pre-MHC region. Trends Genet **20:** 587-591.

DE TOMASO, A. W., S. V. NYHOLM, K. J. PALMERI, K. J. ISHIZUKA, W. B. LUDINGTON *et al.*, 2005 Isolation and characterization of a protochordate histocompatibility locus. Nature **438:** 454-459.

DE TOMASO, A. W., and I. L. WEISSMAN, 2004 Evolution of a protochordate allorecognition locus. Science **303:** 977.

DOBZHANSKY, T., 1955 A review of some fundamental concepts and problems of population genetics. Cold Spring Harb Symp Quant Biol **20:** 1-15.

DOBZHANSKY, T. C., KRIMBAS, C. AND KRIMBAS, M.G., 1960 Genetics of natural populations. XXX. Is the genetic load in Drosophila pseudoobscura a mutational or balanced load? Genetics **45:** 741-753.

DOMINY, N. J., and P. W. LUCAS, 2001 Ecological importance of trichromatic vision to primates. Nature **410:** 363-366.

DULAI, K. S., J. K. BOWMAKER, J. D. MOLLON and D. M. HUNT, 1994 Sequence divergence, polymorphism and evolution of the middle-wave and long-wave visual pigment genes of great apes and Old World monkeys. Vision Res **34:** 2483-2491.

DULAI, K. S., M. VON DORNUM, J. D. MOLLON and D. M. HUNT, 1999 The evolution of trichromatic color vision by opsin gene duplication in New World and Old World primates. Genome Res **9:** 629-638.

DUSENBERY, D. B., 2000 Selection for high gamete encounter rates explains the success of male and female mating types. J Theor Biol **202:** 1-10.

EASON, D. D., R. T. LITMAN, C. A. LUER, W. KERR and G. W. LITMAN, 2004 Expression of individual immunoglobulin genes occurs in an unusual system consisting of multiple independent loci. Eur J Immunol **34:** 2551-2558.

EAST, E., 1936 Heterosis. Genetics **21:** 375-397.

EGGERT, F., W. MULLER-RUCHHOLTZ and R. FERSTL, 1998 Olfactory cues associated with the major histocompatibility complex. Genetica **104:** 191-197.

ENARD, W., M. PRZEWORSKI, S. E. FISHER, C. S. LAI, V. WIEBE *et al.*, 2002 Molecular evolution of FOXP2, a gene involved in speech and language. Nature **418:** 869-872.

ESTIVILL, X., C. BANCELLS and C. RAMOS, 1997 Geographic distribution and regional origin of 272 cystic fibrosis mutations in European populations. The Biomed CF Mutation Analysis Consortium. Hum Mutat **10:** 135-154.

FAURIE, C., and M. RAYMOND, 2005 Handedness, homicide and negative frequency-dependent selection. Proc Biol Sci **272:** 25-28.

FELSENSTEIN, J., 1972 The substitutional load in a finite population. Heredity **28:** 57-69.

FELSENSTEIN, J., 1992 Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet Res **59:** 139-147.

FERNANDEZ, N., J. COOPER, M. SPRINKS, M. ABDELRAHMAN, D. FISZER *et al.*, 1999 A critical review of the role of the major histocompatibility complex in fertilization, preimplantation development and feto-maternal interactions. Hum Reprod Update **5:** 234-248.

FIX, A. G., 2003 Simulating hemoglobin history. Hum Biol **75:** 607-618.

FOERSTER, K., K. DELHEY, A. JOHNSEN, J. T. LIFJELD and B. KEMPENAERS, 2003 Females increase offspring heterozygosity and fitness through extra-pair matings. Nature **425:** 714-717.

FRANKLIN-TONG, N. V., and F. C. FRANKLIN, 2003 Gametophytic self-incompatibility inhibits pollen tube growth using different mechanisms. Trends Plant Sci **8:** 598-605.

FRASER, A. S., and D. BURNELL, 1970 *Computer Models in Genetics*. McGraw-Hill, NY.

FRASER, J. A., and J. HEITMAN, 2005 Chromosomal sex-determining regions in animals, plants and fungi. Curr Opin Genet Dev.

FREEMAN, C. D., DOUST, J.L., EL-KEBLAWY, A, MIGLIA, K.J. AND MCARTHUR, E.D., 1997 Sexual specialization and inbreeding avoidance in the evolution of dioecy. The Botanical Review **63:** 65.

FREY, F. M., 2004 Opposing natural selection from herbivores and pathogens may maintain floral-color variation in Claytonia virginica (Portulacaceae). Evolution Int J Org Evolution **58:** 2426-2437.

FUKUDA, N., K. YOMOGIDA, M. OKABE and K. TOUHARA, 2004 Functional characterization of a mouse testicular olfactory receptor and its role in chemosensing and in regulation of sperm motility. J Cell Sci **117:** 5835-5845.

GABRIEL, S. B., S. F. SCHAFFNER, H. NGUYEN, J. M. MOORE, J. ROY *et al.*, 2002 The structure of haplotype blocks in the human genome. Science **296:** 2225-2229.

GARRIGAN, D., and P. W. HEDRICK, 2003 Perspective: detecting adaptive molecular polymorphism: lessons from the MHC. Evolution Int J Org Evolution **57:** 1707-1722.

GARTE, S., 2003 Locus-specific genetic diversity between human populations: an analysis of the literature. Am J Hum Biol **15:** 814-823.

GAUDIERI, S., R. L. DAWKINS, K. HABARA, J. K. KULSKI and T. GOJOBORI, 2000 SNP profile within the human major histocompatibility complex reveals an extreme and interrupted level of nucleotide diversity. Genome Res **10:** 1579-1586.

GEHRING, W. J., 2004 Historical perspective on the development and evolution of eyes and photoreceptors. Int J Dev Biol **48:** 707-717.

GLEMIN, S., T. GAUDE, M. L. GUILLEMIN, M. LOURMAS, I. OLIVIERI *et al.*, 2005 Balancing selection in the wild: testing population genetics theory of self-incompatibility in the rare species Brassica insularis. Genetics.

GOMEZ, G. A., and E. HASSON, 2003 Transpecific polymorphisms in an inversion linked esterase locus in Drosophila buzzatii. Mol Biol Evol **20:** 410-423.

GONZALEZ, E., M. BAMSHAD, N. SATO, S. MUMMIDI, R. DHANDA *et al.*, 1999 Race-specific HIV-1 disease-modifying effects associated with CCR5 haplotypes. Proc Natl Acad Sci U S A **96:** 12004-12009.

GONZALEZ, E., R. DHANDA, M. BAMSHAD, S. MUMMIDI, R. GEEVARGHESE *et al.*, 2001 Global survey of genetic variation in CCR5, RANTES, and MIP-1alpha: impact on the epidemiology of the HIV-1 pandemic. Proc Natl Acad Sci U S A **98:** 5199-5204.

GOULD, S. J., 1970 Dollo on Dollo's law: irreversibility and the status of evolutionary laws. J Hist Biol **3:** 189-212.

GRELL, E. H., JACOBSON, K.B., MURPHY, J.B., 1965 Alcohol dehydrogenase in Drosophila melanogaster: isozymes and genetic variants. Science **149:** 80-82.

GROSBERG, R. K., and M. W. HART, 2000 Mate selection and the evolution of highly polymorphic self/nonself recognition genes. Science **289:** 2111-2114.

HALDANE, J. B. S., 1957 The Cost of Natural Selection. Journal of Genetics **55:** 511-524.

HARPENDING, H., and G. COCHRAN, 2002 In our genes. Proc Natl Acad Sci U S A **99:** 10-12.

HARRIS, H., 1966 Enzyme polymorphisms in man. Proceedings of the Royal Society of London. Series B, Biological Sciences **164:** 298-310.

HARTL, D. L., 2000 *A Primer of Population Genetics*. Sinauer Associates, Inc., Sunderland, MA.

HEDRICK, P. W., 1998 Balancing selection and MHC. Genetica **104:** 207-214.

HINDS, D. A., A. P. KLOEK, M. JEN, X. CHEN and K. A. FRAZER, 2006 Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet **38:** 82-85.

HISCOCK, S. J., and S. M. MCINNIS, 2003 Pollen recognition and rejection during the sporophytic self-incompatibility response: Brassica and beyond. Trends Plant Sci **8:** 606-613.

HOEKSTRA, R. F., 2005 Evolutionary biology: why sex is good. Nature **434:** 571-573.

HOFF, C., I. THORNEYCROFT, F. WILSON and M. WILLIAMS-MURPHY, 2001 Protection afforded by sickle-cell trait (Hb AS): what happens when malarial selection pressures are alleviated? Hum Biol **73:** 583-586.

HOLUB, E. B., 2001 The arms race is ancient history in Arabidopsis, the wildflower. Nat Rev Genet **2:** 516-527.

HOULE, D., 1998 How should we explain variation in the genetic variance of traits? Genetica **102-103:** 241-253.

HUBBY, J. L., and R. C. LEWONTIN, 1966 A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in Drosophila pseudoobscura. Genetics **54:** 577-594.

HUDSON, R. R., 1991 Gene genealogies and the coalescent process. Oxford Surveys in Evolutionary Biology **7:** 1-44.

HUDSON, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18:** 337-338.

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

HUGHES, A. L., M. K. HUGHES, C. Y. HOWELL and M. NEI, 1994 Natural selection at the class II major histocompatibility complex loci of mammals. Philos Trans R Soc Lond B Biol Sci **346:** 359-366; discussion 366-357.

HUGHES, A. L., and M. NEI, 1988 Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335:** 167-170.

HUGHES, A. L., and M. NEI, 1989 Ancient interlocus exon exchange in the history of the HLA-A locus. Genetics **122:** 681-686.

HUGHES, A. L., and M. NEI, 1989 Evolution of the major histocompatibility complex: independent origin of nonclassical class I genes in different groups of mammals. Mol Biol Evol **6:** 559-579.

HUGHES, A. L., and M. NEI, 1989 Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc Natl Acad Sci U S A **86:** 958-962.

HUGHES, A. L., and M. NEI, 1990 Evolutionary relationships of class II major-histocompatibility-complex genes in mammals. Mol Biol Evol **7:** 491-514.

HUGHES, A. L., and M. NEI, 1992 Maintenance of MHC polymorphism. Nature **355:** 402-403.

HUGHES, A. L., and M. NEI, 1992 Models of host-parasite interaction and MHC polymorphism. Genetics **132:** 863-864.

HUGHES, A. L., T. OTA and M. NEI, 1990 Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Mol Biol Evol **7:** 515-524.

HUGHES, A. L., B. PACKER, R. WELCH, S. J. CHANOCK and M. YEAGER, 2005 High level of functional polymorphism indicates a unique role of natural selection at human immune system loci. Immunogenetics **57:** 821-827.

HUGHES, A. L., and M. YEAGER, 1998 Natural selection and the evolutionary history of major histocompatibility complex loci. Front Biosci **3:** d509-516.

HUGHES, A. L., and M. YEAGER, 1998 Natural selection at major histocompatibility complex loci of vertebrates. Annu Rev Genet **32:** 415-435.

HUNT, D. M., K. S. DULAI, J. A. COWING, C. JULLIOT, J. D. MOLLON *et al.*, 1998 Molecular evolution of trichromacy in primates. Vision Res **38:** 3299-3306.

HURST, L. D., C. PAL and M. J. LERCHER, 2004 The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet **5:** 299-310.

HWANG, D. G., and P. GREEN, 2004 Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci U S A **101:** 13994-14001.

JACOBS, G. H., 1996 Primate photopigments and primate color vision. Proc Natl Acad Sci U S A **93:** 577-581.

JANEWAY, C. A., JR., 1992 The immune system evolved to discriminate infectious nonself from noninfectious self. Immunol Today **13:** 11-16.

JOCKUSCH, E. L., and K. A. OBER, 2004 Hypothesis testing in evolutionary developmental biology: a case study from insect wings. J Hered **95:** 382-396.

JONES, D., 1917 Dominance of Linked Factors as a Means of Accounting for Heterosis. Genetics **2:** 466-479.

JONES, G. A. T., E.C., 2006 The evolution of echolocation in bats. Trends in Ecology and Evolution **21:** 149-156.

JORDAN, W. C., and M. W. BRUFORD, 1998 New perspectives on mate choice and the MHC. Heredity **81 (Pt 3):** 239-245.

KALISH, R. B., S. VARDHANA, N. J. NORMAND, M. GUPTA and S. S. WITKIN, 2006 Association of a maternal CD14 -159 gene polymorphism with preterm premature rupture of membranes and spontaneous preterm birth in multi-fetal pregnancies. J Reprod Immunol.

KASAHARA, M., J. NAKAYA, Y. SATTA and N. TAKAHATA, 1997 Chromosomal duplication and the emergence of the adaptive immune system. Trends Genet **13:** 90-92.

KAUSERUD, H., G. P. SAETRE, O. SCHMIDT, C. DECOCK and T. SCHUMACHER, 2006 Genetics of self/nonself recognition in Serpula lacrymans. Fungal Genet Biol.

KELLEY, J., L. WALTER and J. TROWSDALE, 2005 Comparative genomics of major histocompatibility complexes. Immunogenetics **56:** 683-695.

KELLY, J. K., and M. J. WADE, 2000 Molecular evolution near a two-locus balanced polymorphism. J Theor Biol **204:** 83-101.

KHALTURIN, K., Z. PANZER, M. D. COOPER and T. C. BOSCH, 2004 Recognition strategies in the innate immune system of ancestral chordates. Mol Immunol **41:** 1077-1087.

KIDD, K. K., A. J. PAKSTIS, W. C. SPEED and J. R. KIDD, 2004 Understanding human DNA sequence variation. J Hered **95:** 406-420.

KIM, U. K., and D. DRAYNA, 2005 Genetics of individual differences in bitter taste perception: lessons from the PTC gene. Clin Genet **67:** 275-280.

KIM, U. K., E. JORGENSON, H. COON, M. LEPPERT, N. RISCH *et al.*, 2003 Positional cloning of the human quantitative trait locus underlying taste sensitivity to phenylthiocarbamide. Science **299:** 1221-1225.

KIMURA, M., 1968 Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet Res **11:** 247-269.

KLEIN, J., 1982 *Immunology: the science of self-nonself discrimination.* Wiley, New York.

KODA, Y., H. TACHIDA, M. SOEJIMA, O. TAKENAKA and H. KIMURA, 2000 Ancient origin of the null allele se(428) of the human ABO-secretor locus (FUT2). J Mol Evol **50:** 243-248.

KOHN, M., H. KEHRER-SAWATZKI, W. VOGEL, J. A. GRAVES and H. HAMEISTER, 2004 Wide genome comparisons reveal the origins of the human X chromosome. Trends Genet **20:** 598-603.

KOJIMA, K., 1971 The distribution and comparison of "genetic loads" under heterotic selection and simple frequency-dependent selection in finite populations. Theor Popul Biol **2:** 159-173.

KONG, A., D. F. GUDBJARTSSON, J. SAINZ, G. M. JONSDOTTIR, S. A. GUDJONSSON *et al.*, 2002 A high-resolution recombination map of the human genome. Nat Genet **31:** 241-247.

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

KREITMAN, M., and A. DI RIENZO, 2004 Balancing claims for balancing selection. Trends Genet **20:** 300-304.

KRONE, S. M., and C. NEUHAUSER, 1997 Ancestral Processes with Selection. Theor Popul Biol **51:** 210-237.

KROYMANN, J., and T. MITCHELL-OLDS, 2005 Epistasis and balanced polymorphism influencing complex trait variation. Nature **435:** 95-98.

KULKARNI, P. S., S. T. BUTERA and A. C. DUERR, 2003 Resistance to HIV-1 infection: lessons learned from studies of highly exposed persistently seronegative (HEPS) individuals. AIDS Rev **5:** 87-103.

KWIATKOWSKI, D. P., 2005 How malaria has affected the human genome and what human genetics can teach us about malaria. Am J Hum Genet **77:** 171-192.

LAHN, B. T., and D. C. PAGE, 1997 Functional coherence of the human Y chromosome. Science **278:** 675-680.

LAHN, B. T., and D. C. PAGE, 1999 Four evolutionary strata on the human X chromosome. Science **286:** 964-967.

LAZARUS, R., D. VERCELLI, L. J. PALMER, W. J. KLIMECKI, E. K. SILVERMAN *et al.*, 2002 Single nucleotide polymorphisms in innate immunity genes: abundant variation and potential role in complex human disease. Immunol Rev **190:** 9-25.

LEHMAN, N., 2003 A case for the extreme antiquity of recombination. J Mol Evol **56:** 770-777.

LENORMAND, T., and J. DUTHEIL, 2005 Selection on Sex Cells Favors a Recombination Gender Gap. PLoS Biol **3:** 1.

LEVITAN, D. R., 1996 Effects of gamete traits on fertilization in the sea and the evolution of sexual dimorphism. Nature **382:** 153-155.

LEWONTIN, R. C., and J. L. HUBBY, 1966 A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. Genetics **54:** 595-609.

LIBERT, F., P. COCHAUX, G. BECKMAN, M. SAMSON, M. AKSENOVA *et al.*, 1998 The deltaccr5 mutation conferring protection against HIV-1 in Caucasian populations has a single and recent origin in Northeastern Europe. Hum Mol Genet **7:** 399-406.

LITMAN, G. W., 2005 Histocompatibility: colonial match and mismatch. Nature **438:** 437-439.

LITMAN, G. W., J. P. CANNON and L. J. DISHAW, 2005 Reconstructing immune phylogeny: new perspectives. Nat Rev Immunol **5:** 866-879.

LITMAN, G. W., J. P. CANNON and J. P. RAST, 2005 New insights into alternative mechanisms of immune receptor diversification. Adv Immunol **87:** 209-236.

LIU, Z., P. H. MOORE, H. MA, C. M. ACKERMAN, M. RAGIBA *et al.*, 2004 A primitive Y chromosome in papaya marks incipient sex chromosome evolution. Nature **427:** 348-352.

LIVINGSTON, R. J., A. VON NIEDERHAUSERN, A. G. JEGGA, D. C. CRAWFORD, C. S. CARLSON *et al.*, 2004 Pattern of sequence variation across 213 environmental response genes. Genome Res **14:** 1821-1831.

LIVINGSTONE, F. B., 1984 The Duffy blood groups, vivax malaria, and malaria selection in human populations: a review. Hum Biol **56:** 413-425.

LONG, A. D., R. F. LYMAN, A. H. MORGAN, C. H. LANGLEY and T. F. MACKAY, 2000 Both naturally occurring insertions of transposable elements and intermediate frequency polymorphisms at the achaete-scute complex are associated with variation in bristle number in Drosophila melanogaster. Genetics **154:** 1255-1269.

MACDONALD, S. J., and A. D. LONG, 2005 Prospects for identifying functional variation across the genome. Proc Natl Acad Sci U S A **102 Suppl 1:** 6614-6621.

MANGANO, A., E. GONZALEZ, R. DHANDA, G. CATANO, M. BAMSHAD *et al.*, 2001 Concordance between the CC chemokine receptor 5 genetic determinants that alter risks of transmission and disease progression in children exposed perinatally to human immunodeficiency virus. J Infect Dis **183:** 1574-1585.

MARSH, S. G. E., PARHAM, P., BARBER, L.D., 2000 *The HLA FactsBook.* Academic Press.

MARSHALL, C. R., E. C. RAFF and R. A. RAFF, 1994 Dollo's law and the death and resurrection of genes. Proc Natl Acad Sci U S A **91:** 12283-12287.

MARTH, G. T., E. CZABARKA, J. MURVAI and S. T. SHERRY, 2004 The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics **166:** 351-372.

MARTIN, L. D., and T. J. MEEHAN, 2005 Extinction may not be forever. Naturwissenschaften **92:** 1-19.

MARTINI, A., and G. R. BURGIO, 1999 Tolerance and auto-immunity: 50 years after Burnet. Eur J Pediatr **158:** 769-775.

MARTINKO, J. M., V. VINCEK, D. KLEIN and J. KLEIN, 1993 Primate ABO glycosyltransferases: evidence for trans-species evolution. Immunogenetics **37:** 274-278.

MCCARROLL, S. A., T. N. HADNOTT, G. H. PERRY, P. C. SABETI, M. C. ZODY *et al.*, 2006 Common deletion polymorphisms in the human genome. Nat Genet **38:** 86-92.

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

MCNICHOLL, J. M., and N. PROMADEJ, 2004 Insights into the role of host genetic and T-cell factors in resistance to HIV transmission from studies of highly HIV-exposed Thais. Immunol Res **29:** 161-174.

MEAD, S., M. P. STUMPF, J. WHITFIELD, J. A. BECK, M. POULTER *et al.*, 2003 Balancing selection at the prion protein gene consistent with prehistoric kurulike epidemics. Science **300:** 640-643.

MENGE, A. C., and O. BEITNER, 1989 Interrelationships among semen characteristics, antisperm antibodies, and cervical mucus penetration assays in infertile human couples. Fertil Steril **51:** 486-492.

MENGE, A. C., N. E. MEDLEY, C. M. MANGIONE and J. W. DIETRICH, 1982 The incidence and influence of antisperm antibodies in infertile human couples on sperm-cervical mucus interactions and subsequent fertility. Fertil Steril **38:** 439-446.

MEYERS, B. C., S. KAUSHIK and R. S. NANDETY, 2005 Evolving disease resistance genes. Curr Opin Plant Biol **8:** 129-134.

MICHELMORE, R. W., 2003 The impact zone: genomics and breeding for durable disease resistance. Curr Opin Plant Biol **6:** 397-404.

MICHELMORE, R. W., and B. C. MEYERS, 1998 Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res **8:** 1113-1130.

MICHON, P., I. WOOLLEY, E. M. WOOD, W. KASTENS, P. A. ZIMMERMAN *et al.*, 2001 Duffy-null promoter heterozygosity reduces DARC expression and abrogates adhesion of the P. vivax ligand required for blood-stage infection. FEBS Lett **495:** 111-114.

MILLER, R. D., M. S. PHILLIPS, I. JO, M. A. DONALDSON, J. F. STUDEBAKER *et al.*, 2005 High-density single-nucleotide polymorphism maps of the human genome. Genomics.

MOALEM, S., M. E. PERCY, T. P. KRUCK and R. R. GELBART, 2002 Epidemic pathogenic selection: an explanation for hereditary hemochromatosis? Med Hypotheses **59:** 325-329.

MODI, W. S., and D. CREWS, 2005 Sex chromosomes and sex determination in reptiles Commentary. Curr Opin Genet Dev.

MODIANO, D., G. LUONI, B. S. SIRIMA, J. SIMPORE, F. VERRA *et al.*, 2001 Haemoglobin C protects against clinical Plasmodium falciparum malaria. Nature **414:** 305-308.

MOHAN, H., S. YADAV, U. SINGH, A. KADIAN and P. MOHAN, 1990 Circulating iso and auto antibodies to human spermatozoa in infertility. Indian J Pathol Microbiol **33:** 161-165.

MORGAN, M. J., A. ADAM and J. D. MOLLON, 1992 Dichromats detect colour-camouflaged objects that are not detected by trichromats. Proc Biol Sci **248:** 291-295.

MORI, T., M. W. GUO, E. SATO, T. BABA, S. TAKASAKI *et al.*, 2000 Molecular and immunological approaches to mammalian fertilization. J Reprod Immunol **47:** 139-158.

MUIRHEAD, C. A., 2001 Consequences of population structure on genes under balancing selection. Evolution Int J Org Evolution **55:** 1532-1541.

MUIRHEAD, C. A., N. L. GLASS and M. SLATKIN, 2002 Multilocus self-recognition systems in fungi as a cause of trans-species polymorphism. Genetics **161:** 633-641.

MULLER, H. J., 1950 Our load of mutations. Am J Hum Genet **2:** 111-176.

MYERS, S., L. BOTTOLO, C. FREEMAN, G. MCVEAN and P. DONNELLY, 2005 A fine-scale map of recombination rates and hotspots across the human genome. Science **310:** 321-324.

NADKARNI, N. A., M. E. WEALE, M. VON SCHANTZ and M. G. THOMAS, 2005 Evolution of a Length Polymorphism in the Human PER3 Gene, a Component of the Circadian System. J Biol Rhythms **20:** 490-499.

NAGEL, R. L., 1990 Innate resistance to malaria: the intraerythrocytic cycle. Blood Cells **16:** 321-339; discussion 340-329.

NAKAJIMA, T., S. WOODING, Y. SATTA, N. JINNAI, S. GOTO *et al.*, 2005 Evidence for natural selection in the HAVCR1 gene: high degree of amino-acid variability in the mucin domain of human HAVCR1 protein. Genes Immun **6:** 398-406.

NASRALLAH, J. B., 2005 Recognition and rejection of self in plant self-incompatibility: comparisons to animal histocompatibility. Trends Immunol **26:** 412-418.

NASRALLAH, M. E., P. LIU, S. SHERMAN-BROYLES, N. A. BOGGS and J. B. NASRALLAH, 2004 Natural variation in expression of self-incompatibility in Arabidopsis thaliana: implications for the evolution of selfing. Proc Natl Acad Sci U S A **101:** 16070-16074.

NAVARRO, A., and N. H. BARTON, 2002 The effects of multilocus balancing selection on neutral variability. Genetics **161:** 849-863.

NEI, M., X. GU and T. SITNIKOVA, 1997 Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci U S A **94:** 7799-7806.

NEKRUTENKO, A., and W. H. LI, 2000 Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Res **10:** 1986-1995.

NEUHAUSER, C., and S. M. KRONE, 1997 The genealogy of samples in models with selection. Genetics **145:** 519-534.

NOONAN, J. P., J. LI, L. NGUYEN, C. CAOILE, M. DICKSON *et al.*, 2003 Extensive linkage disequilibrium, a common 16.7-kilobase deletion, and evidence of balancing selection in the human protocadherin alpha cluster. Am J Hum Genet **72:** 621-635.

NORDBORG, M., T. T. HU, Y. ISHINO, J. JHAVERI, C. TOOMAJIAN *et al.*, 2005 The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol **3:** e196.

NORDBORG, M., and H. INNAN, 2003 The genealogy of sequences containing multiple sites subject to strong selection in a subdivided population. Genetics **163:** 1201-1213.

NORMAN, P. J., M. A. COOK, B. S. CAREY, C. V. CARRINGTON, D. H. VERITY *et al.*, 2004 SNP haplotypes and allele frequencies show evidence for disruptive and balancing selection in the human leukocyte receptor complex. Immunogenetics **56:** 225-237.

O'HUIGIN, C., A. SATO and J. KLEIN, 1997 Evidence for convergent evolution of A and B blood group antigens in primates. Hum Genet **101:** 141-148.

OHL, D. A., and R. K. NAZ, 1995 Infertility due to antisperm antibodies. Urology **46:** 591-602.

OLSON, M. V., 1999 When less is more: gene loss as an engine of evolutionary change. Am J Hum Genet **64:** 18-23.

OLSON, M. V., A. KAS, K. BUBB, R. QUI, E. E. SMITH *et al.*, 2004 Hypervariability, suppressed recombination and the genetics of individuality. Philos Trans R Soc Lond B Biol Sci **359:** 129-140.

OTA, T., 1975 Statistical analyses of Drosophila and human protein polymorphisms. Proc Natl Acad Sci U S A **72:** 3194-3196.

OTA, T., T. SITNIKOVA and M. NEI, 2000 Evolution of vertebrate immunoglobulin variable gene segments. Curr Top Microbiol Immunol **248:** 221-245.

OTTO, S. P., and T. LENORMAND, 2002 Resolving the paradox of sex and recombination. Nat Rev Genet **3:** 252-261.

PANCER, Z., 2000 Dynamic expression of multiple scavenger receptor cysteine-rich genes in coelomocytes of the purple sea urchin. Proc Natl Acad Sci U S A **97:** 13156-13161.

PANCER, Z., C. T. AMEMIYA, G. R. EHRHARDT, J. CEITLIN, G. L. GARTLAND *et al.*, 2004 Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. Nature **430:** 174-180.

PANCER, Z., N. R. SAHA, J. KASAMATSU, T. SUZUKI, C. T. AMEMIYA *et al.*, 2005 Variable lymphocyte receptors in hagfish. Proc Natl Acad Sci U S A **102:** 9224-9229.

PANNELL, J. R. A. V., M., 2006 Evolution **60:** 660-673.

PARHAM, P., and T. OHTA, 1996 Population biology of antigen presentation by MHC class I molecules. Science **272:** 67-74.

PARKER, G. A., 1978 Selection on non-random fusion of gametes during the evolution of anisogamy. J Theor Biol **73:** 1-28.

PARKER, G. A., R. R. BAKER and V. G. SMITH, 1972 The origin and evolution of gamete dimorphism and the male-female phenomenon. J Theor Biol **36:** 529-553.

PETTIGREW, J. D., 1991 Wings or Brain? Convergent Evolution in the Origins of Bats. Systematic Zoology **40:** 199-216.

PHILLIPS, M. S., R. LAWRENCE, R. SACHIDANANDAM, A. P. MORRIS, D. J. BALDING *et al.*, 2003 Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. Nat Genet **33:** 382-387.

PIER, G. B., M. GROUT, T. ZAIDI, G. MELULENI, S. S. MUESCHENBORN *et al.*, 1998 Salmonella typhi uses CFTR to enter intestinal epithelial cells. Nature **393:** 79-82.

POLLEY, S. D., W. CHOKEJINDACHAI and D. J. CONWAY, 2003 Allele frequency-based analyses robustly map sequence sites under balancing selection in a malaria vaccine candidate antigen. Genetics **165:** 555-561.

POTASHNIK, G., D. KLEINMAN, V. INSLER, S. ALBOTIANO, M. GLEZERMAN *et al.*, 1988 Results of in vitro fertilization in women with antisperm antibodies in serum, cervical mucus, and follicular fluid. J In Vitro Fert Embryo Transf **5:** 199-201.

PRICE, P., C. WITT, R. ALLCOCK, D. SAYER, M. GARLEPP *et al.*, 1999 The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases. Immunol Rev **167:** 257-274.

PRITCHARD, J. K., and M. PRZEWORSKI, 2001 Linkage disequilibrium in humans: models and data. Am J Hum Genet **69:** 1-14.

PRUGNOLLE, F., A. MANICA, M. CHARPENTIER, J. F. GUEGAN, V. GUERNIER *et al.*, 2005 Pathogen-driven selection and worldwide HLA class I diversity. Curr Biol **15:** 1022-1027.

PTAK, S. E., D. A. HINDS, K. KOEHLER, B. NICKEL, N. PATIL *et al.*, 2005 Fine-scale recombination patterns differ between chimpanzees and humans. Nat Genet **37:** 429-434.

RAST, J. P., C. T. AMEMIYA, R. T. LITMAN, S. J. STRONG and G. W. LITMAN, 1998 Distinct patterns of IgH structure and organization in a divergent lineage of chrondrichthyan fishes. Immunogenetics **47:** 234-245.

RAYMOND, C. K., A. KAS, M. PADDOCK, R. QIU, Y. ZHOU *et al.*, 2005 Ancient haplotypes of the HLA Class II region. Genome Res **15:** 1250-1257.

RAYMOND, C. K., S. SUBRAMANIAN, M. PADDOCK, R. QIU, C. DEODATO *et al.*, 2005 Targeted, haplotype-resolved resequencing of long segments of the human genome. Genomics.

REED, F. A., J. M. AKEY and C. F. AQUADRO, 2005 Fitting background-selection predictions to levels of nucleotide variation and divergence along the human autosomes. Genome Res **15:** 1211-1221.

RICHMAN, A., 2000 Evolution of balanced genetic polymorphism. Mol Ecol **9:** 1953-1963.

ROBERTSON, A., 1970 A theory of limits in artificial selection with many linked loci, pp. 246-288 in *Mathematical Topics in Population Genetics*, edited by K. KOJIMA. Springer-Verlag, Berlin.

ROCHETTE, J., J. J. POINTON, C. A. FISHER, G. PERERA, M. ARAMBEPOLA *et al.*, 1999 Multicentric origin of hemochromatosis gene (HFE) mutations. Am J Hum Genet **64:** 1056-1062.

ROUGHGARDEN, J., 1972 Evolution of Niche Width. American Naturalist **106:** 683-718.

RUFF, M. R., M. POLIANOVA, Q. E. YANG, G. S. LEOUNG, F. W. RUSCETTI *et al.*, 2003 Update on D-ala-peptide T-amide (DAPTA): a viral entry inhibitor that blocks CCR5 chemokine receptors. Curr HIV Res **1:** 51-67.

SABATER-LLEAL, M., J. M. SORIA, J. BERTRANPETIT, L. ALMASY, J. BLANGERO *et al.*, 2005 Human F7 sequence is split into three deep clades that are related to FVII plasma levels. Hum Genet**:** 1-11.

SABETI, P. C., E. WALSH, S. F. SCHAFFNER, P. VARILLY, B. FRY *et al.*, 2005 The case for selection at CCR5-Delta32. PLoS Biol **3:** e378.

SACHIDANANDAM, R., D. WEISSMAN, S. C. SCHMIDT, J. M. KAKOL, L. D. STEIN *et al.*, 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature **409:** 928-933.

SAITOU, N., and F. YAMAMOTO, 1997 Evolution of primate ABO blood group genes and their homologous genes. Mol Biol Evol **14:** 399-411.

SATO, K., T. NISHIO, R. KIMURA, M. KUSABA, T. SUZUKI *et al.*, 2002 Coevolution of the S-locus genes SRK, SLG and SP11/SCR in Brassica oleracea and B. rapa. Genetics **162:** 931-940.

SAVAGE, A. E., and J. S. MILLER, 2006 Gametophytic self-incompatibility in Lycium parishii (Solanaceae): allelic diversity, genealogical structure, and patterns of molecular evolution at the S-RNase locus. Heredity.

SAVAN, R., A. AMAN, K. SATO, R. YAMAGUCHI and M. SAKAI, 2005 Discovery of a new class of immunoglobulin heavy chain from fugu. Eur J Immunol **35:** 3320-3331.

SAVOLAINEN, O., C. H. LANGLEY, B. P. LAZZARO and H. FR, 2000 Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing Arabidopsis lyrata and the selfing Arabidopsis thaliana. Mol Biol Evol **17:** 645-655.

SCHAFFNER, S. F., C. FOO, S. GABRIEL, D. REICH, M. J. DALY *et al.*, 2005 Calibrating a coalescent simulation of human genome sequence variation. Genome Res **15:** 1576-1583.

SCHIERUP, M. H., D. CHARLESWORTH and X. VEKEMANS, 2000 The effect of hitch-hiking on genes linked to a balanced polymorphism in a subdivided population. Genet Res **76:** 63-73.

SCHIERUP, M. H., A. M. MIKKELSEN and J. HEIN, 2001 Recombination, balancing selection and phylogenies in MHC and self-incompatibility genes. Genetics **159:** 1833-1844.

SCHIERUP, M. H., X. VEKEMANS and D. CHARLESWORTH, 2000 The effect of subdivision on variation at multi-allelic loci under balancing selection. Genet Res **76:** 51-62.

SCHILTHUIZEN, M., 2000 Dualism and conflicts in understanding speciation. Bioessays **22:** 1134-1141.

SCHLIEWEN, U. K., D. TAUTZ and S. PAABO, 1994 Sympatric speciation suggested by monophyly of crater lake cichlids. Nature **368:** 629-632.

SEABURY, C. M., R. L. HONEYCUTT, A. P. ROONEY, N. D. HALBERT and J. N. DERR, 2004 Prion protein gene (PRNP) variants and evidence for strong purifying selection in functionally important regions of bovine exon 3. Proc Natl Acad Sci U S A **101:** 15142-15147.

SHEN, J., H. ARAKI, L. CHEN, J. Q. CHEN and D. TIAN, 2006 Unique evolutionary mechanism in R-genes under the presence/absence polymorphism in Arabidopsis thaliana. Genetics.

SHRIVER, M. D., G. C. KENNEDY, E. J. PARRA, H. A. LAWSON, V. SONPAR *et al.*, 2004 The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum Genomics **1:** 274-286.

SHYUE, S. K., S. BOISSINOT, H. SCHNEIDER, I. SAMPAIO, M. P. SCHNEIDER *et al.*, 1998 Molecular genetics of spectral tuning in New World monkey color vision. J Mol Evol **46:** 697-702.

SIEPEL, A., G. BEJERANO, J. S. PEDERSEN, A. S. HINRICHS, M. HOU *et al.*, 2005 Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res **15:** 1034-1050.

SIMMONS, M. J., and J. F. CROW, 1977 Mutations affecting fitness in Drosophila populations. Annu Rev Genet **11:** 49-78.

SLATKIN, M., 2000 Balancing selection at closely linked, overdominant loci in a finite population. Genetics **154:** 1367-1378.

SMITH, T. G., K. AYI, L. SERGHIDES, C. D. MCALLISTER and K. C. KAIN, 2002 Innate immunity to malaria caused by Plasmodium falciparum. Clin Invest Med **25:** 262-272.

SOEJIMA, M., H. TACHIDA, M. TSUNEOKA, O. TAKENAKA, H. KIMURA *et al.*, 2005 Nucleotide sequence analyses of human complement 6 (C6) gene suggest balancing selection. Ann Hum Genet **69:** 239-252.

SOLDEVILA, M., A. M. ANDRES, A. RAMIREZ-SORIANO, T. MARQUES-BONET, F. CALAFELL *et al.*, 2005 The prion protein gene in humans revisited: Lessons from a worldwide resequencing study. Genome Res.

SPASSKY, B., SPASSKY, N., PAVLOVSKY, O., KRIMBAS, M.G., KRIMBAS, C. AND DOBZHANSKY, T., 1960 Genetics of natural populations. XXVIII. The magnitude of the genetic load in Drosophila pseudoobscura. Genetics **45:** 723-740.

SPENCER, C. C., and G. COOP, 2004 SelSim: a program to simulate population genetic data with natural selection and recombination. Bioinformatics **20:** 3673-3675.

SPOFFORD, J. B., 1969 Heterosis and the Evolution of Duplications. American Naturalist **103:** 407-432.

SPOFFORD, J. B., 1972 A heterotic model for the evolution of duplications. Brookhaven Symp Biol **23:** 121-143.

STASKAWICZ, B. J., M. B. MUDGETT, J. L. DANGL and J. E. GALAN, 2001 Common and contrasting themes of plant and animal diseases. Science **292:** 2285-2289.

STEFANSSON, H., A. HELGASON, G. THORLEIFSSON, V. STEINTHORSDOTTIR, G. MASSON *et al.*, 2005 A common inversion under selection in Europeans. Nat Genet **37:** 129-137.

STEPHENS, J. C., and M. NEI, 1985 Phylogenetic analysis of polymorphic DNA sequences at the Adh locus in Drosophila melanogaster and its sibling species. J Mol Evol **22:** 289-300.

STEWART, C. A., R. HORTON, R. J. ALLCOCK, J. L. ASHURST, A. M. ATRAZHEV *et al.*, 2004 Complete MHC haplotype sequencing for common disease gene mapping. Genome Res **14:** 1176-1187.

SURRIDGE, A. K., and N. I. MUNDY, 2002 Trans-specific evolution of opsin alleles and the maintenance of trichromatic colour vision in Callitrichine primates. Mol Ecol **11:** 2157-2169.

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

TAKAHATA, N., 1990 A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. Proc Natl Acad Sci U S A **87:** 2419-2423.

TAKAHATA, N., and Y. SATTA, 1997 Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. Proc Natl Acad Sci U S A **94:** 4811-4815.

TAKAHATA, N., and Y. SATTA, 1998 Selection, convergence, and intragenic recombination in HLA diversity. Genetica **102-103:** 157-169.

TAKAHATA, N., Y. SATTA and J. KLEIN, 1992 Polymorphism and balancing selection at major histocompatibility complex loci. Genetics **130:** 925-938.

TAKAHATA, N., Y. SATTA and J. KLEIN, 1995 Divergence time and population size in the lineage leading to modern humans. Theor Popul Biol **48:** 198-221.

TAKAYAMA, S., H. SHIBA, M. IWANO, H. SHIMOSATO, F. S. CHE *et al.*, 2000 The pollen determinant of self-incompatibility in Brassica campestris. Proc Natl Acad Sci U S A **97:** 1920-1925.

TANEICHI, A., H. SHIBAHARA, Y. HIRANO, T. SUZUKI, H. OBARA *et al.*, 2002 Sperm immobilizing antibodies in the sera of infertile women cause low fertilization rates and poor embryo quality in vitro. Am J Reprod Immunol **47:** 46-51.

TEELING, E. C., M. SCALLY, D. J. KAO, M. L. ROMAGNOLI, M. S. SPRINGER *et al.*, 2000 Molecular evidence regarding the origin of echolocation and flight in bats. Nature **403:** 188-192.

TEELING, E. C., M. S. SPRINGER, O. MADSEN, P. BATES, J. O'BRIEN S *et al.*, 2005 A molecular phylogeny for bats illuminates biogeography and the fossil record. Science **307:** 580-584.

TEPPER, B. J., 1998 6-n-Propylthiouracil: a genetic marker for taste, with implications for food preference and dietary habits. Am J Hum Genet **63:** 1271-1276.

THOMAS, J. D., AND BARRETT, C.H., 1981 Selection for Outcrossing, Sexual Selection, and the Evolution of Dioecy in Plants. The American Naturalist **118:** 443-449.

THORNTON, K., 2005 Recombination and the properties of Tajima's D in the context of approximate likelihood calculation. Genetics.

TIAN, D., H. ARAKI, E. STAHL, J. BERGELSON and M. KREITMAN, 2002 Signature of balancing selection in Arabidopsis. Proc Natl Acad Sci U S A **99:** 11525-11530.

TSUI, L. C., 1992 Mutations and sequence variations detected in the cystic fibrosis transmembrane conductance regulator (CFTR) gene: a report from the Cystic Fibrosis Genetic Analysis Consortium. Hum Mutat **1:** 197-203.

TURELLI, M., and N. H. BARTON, 2004 Polygenic variation maintained by balancing selection: pleiotropy, sex-dependent allelic effects and G x E interactions. Genetics **166:** 1053-1079.

TURNER, G. F., O. SEEHAUSEN, M. E. KNIGHT, C. J. ALLENDER and R. L. ROBINSON, 2001 How many species of cichlid fishes are there in African lakes? Mol Ecol **10:** 793-806.

VAN DE VOSSE, E., S. ALI, A. W. VISSER, C. SURJADI, S. WIDJAJA *et al.*, 2005 Susceptibility to typhoid fever is associated with a polymorphism in the cystic fibrosis transmembrane conductance regulator (CFTR). Hum Genet **118:** 138-140.

VAN DER BIEZEN, E. A., and J. D. JONES, 1998 Plant disease-resistance proteins and the gene-for-gene concept. Trends Biochem Sci **23:** 454-456.

VEKEMANS, X., and M. SLATKIN, 1994 Gene and allelic genealogies at a gametophytic self-incompatibility locus. Genetics **137:** 1157-1165.

VERRELLI, B. C., J. H. MCDONALD, G. ARGYROPOULOS, G. DESTRO-BISOL, A. FROMENT *et al.*, 2002 Evidence for balancing selection from nucleotide sequence analyses of human G6PD. Am J Hum Genet **71:** 1112-1128.

VIENNE, A., T. SHIINA, L. ABI-RACHED, E. DANCHIN, V. VITIELLO *et al.*, 2003 Evolution of the proto-MHC ancestral region: more evidence for the plesiomorphic organisation of human chromosome 9q34 region. Immunogenetics **55:** 429-436.

VOGEL, F. A. M., A.G., 1997 *Human Genetics: Problems and Approaches.* Springer-Verlag, Berlin and Heidelberg.

VOIGHT, B. F., A. M. ADAMS, L. A. FRISSE, Y. QIAN, R. R. HUDSON *et al.*, 2005 Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc Natl Acad Sci U S A **102:** 18508-18513.

VOIGHT, B. F., S. KUDARAVALLI, X. WEN and J. K. PRITCHARD, 2006 A map of recent positive selection in the human genome. PLoS Biol **4:** e72.

VOSSHALL, L. B., 2004 Olfaction: attracting both sperm and the nose. Curr Biol **14:** R918-920.

VUJISIC, S., S. Z. LEPEJ, L. JERKOVIC, I. EMEDI and B. SOKOLIC, 2005 Antisperm antibodies in semen, sera and follicular fluids of infertile patients: relation to reproductive outcome after in vitro fertilization. Am J Reprod Immunol **54:** 13-20.

VYSKOT, B., and R. HOBZA, 2004 Gender in plants: sex chromosomes are emerging from the fog. Trends Genet **20:** 432-438.

WALL, J. D., 1999 Recombination and the power of statistical tests of neutrality. Genetical Research **74:** 65-79.

WALL, J. D., 2003 Estimating ancestral population sizes and divergence times. Genetics **163:** 395-404.

WALL, J. D., and R. R. HUDSON, 2001 Coalescent simulations and statistical tests of neutrality. Mol Biol Evol **18:** 1134-1135; author reply 1136-1138.

WALL, J. D., and J. K. PRITCHARD, 2003 Assessing the performance of the haplotype block model of linkage disequilibrium. Am J Hum Genet **73:** 502-515.

WANG, E., Y. C. DING, P. FLODMAN, J. R. KIDD, K. K. KIDD *et al.*, 2004 The genetic architecture of selection at the human dopamine receptor D4 (DRD4) gene locus. Am J Hum Genet **74:** 931-944.

WANG, N., J. M. AKEY, K. ZHANG, R. CHAKRABORTY and L. JIN, 2002 Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. Am J Hum Genet **71:** 1227-1234.

WARR, G. W., 1995 The immunoglobulin genes of fish. Dev Comp Immunol **19:** 1-12.

WATERSTON, R. H., K. LINDBLAD-TOH, E. BIRNEY, J. ROGERS, J. F. ABRIL *et al.*, 2002 Initial sequencing and comparative analysis of the mouse genome. Nature **420:** 520-562.

WATSON, F. L., R. PUTTMANN-HOLGADO, F. THOMAS, D. L. LAMAR, M. HUGHES *et al.*, 2005 Extensive diversity of Ig-superfamily proteins in the immune system of insects. Science **309:** 1874-1878.

WEIGEL, D., and M. NORDBORG, 2005 Natural variation in Arabidopsis. How do we find the causal genes? Plant Physiol **138:** 567-568.

WEINIG, C., L. A. DORN, N. C. KANE, Z. M. GERMAN, S. S. HALLDORSDOTTIR *et al.*, 2003 Heterogeneous selection at specific loci in natural environments in Arabidopsis thaliana. Genetics **165:** 321-329.

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

WILLIAMS, A. F., and A. N. BARCLAY, 1988 The immunoglobulin superfamily--domains for cell surface recognition. Annu Rev Immunol **6:** 381-405.

WILLIAMS, R. C., 2003 The mind of primitive anthropologists: hemoglobin and HLA, patterns of molecular evolution. Hum Biol **75:** 577-584.

WINCKLER, W., S. R. MYERS, D. J. RICHTER, R. C. ONOFRIO, G. J. MCDONALD *et al.*, 2005 Comparison of fine-scale recombination rates in humans and chimpanzees. Science **308:** 107-111.

WITCHEL, S. F., P. A. LEE, M. SUDA-HARTMAN, M. TRUCCO and E. P. HOFFMAN, 1997 Evidence for a heterozygote advantage in congenital adrenal hyperplasia due to 21-hydroxylase deficiency. J Clin Endocrinol Metab **82:** 2097-2101.

WIUF, C., K. ZHAO, H. INNAN and M. NORDBORG, 2004 The probability and chromosomal extent of trans-specific polymorphism. Genetics **168:** 2363-2372.

WONG, G. K., D. A. PASSEY, Y. HUANG, Z. YANG and J. YU, 2000 Is "junk" DNA mostly intron DNA? Genome Res **10:** 1672-1678.

WOODING, S., 2004 Natural selection: sign, sign, everywhere a sign. Curr Biol **14:** R700-701.

WOODING, S., B. BUFE, C. GRASSI, M. T. HOWARD, A. C. STONE *et al.*, 2006 Independent evolution of bitter-taste sensitivity in humans and chimpanzees. Nature **440:** 930-934.

WOODING, S., U. K. KIM, M. J. BAMSHAD, J. LARSEN, L. B. JORDE *et al.*, 2004 Natural selection and molecular evolution in PTC, a bitter-taste receptor gene. Am J Hum Genet **74:** 637-646.

WOODING, S., A. C. STONE, D. M. DUNN, S. MUMMIDI, L. B. JORDE *et al.*, 2005 Contrasting effects of natural selection on human and chimpanzee CC chemokine receptor 5. Am J Hum Genet **76:** 291-301.

YAHIAOUI, N., S. BRUNNER and B. KELLER, 2006 Rapid generation of new powdery mildew resistance genes after wheat domestication. Plant J.

YAMAMOTO, F., H. CLAUSEN, T. WHITE, J. MARKEN and S. HAKOMORI, 1990 Molecular genetic basis of the histo-blood group ABO system. Nature **345:** 229-233.

YAMAMOTO, F., and S. HAKOMORI, 1990 Sugar-nucleotide donor specificity of histo-blood group A and B transferases is based on amino acid substitutions. J Biol Chem **265:** 19257-19262.

YANG, N., D. G. MACARTHUR, J. P. GULBIN, A. G. HAHN, A. H. BEGGS *et al.*, 2003 ACTN3 genotype is associated with human elite athletic performance. Am J Hum Genet **73:** 627-631.

YAZDANI, S. S., A. R. SHAKRI, P. MUKHERJEE, S. K. BANIWAL and C. E. CHITNIS, 2004 Evaluation of immune responses elicited in mice against a recombinant malaria vaccine based on Plasmodium vivax Duffy binding protein. Vaccine **22:** 3727-3737.

YAZER, M. H., 2005 What a difference 2 nucleotides make: a short review of ABO genetics. Transfus Med Rev **19:** 200-209.

YODER, J. A., R. T. LITMAN, M. G. MUELLER, S. DESAI, K. P. DOBRINSKI *et al.*, 2004 Resolution of the novel immune-type receptor gene cluster in zebrafish. Proc Natl Acad Sci U S A **101:** 15706-15711.

YOSHIURA, K., A. KINOSHITA, T. ISHIDA, A. NINOKATA, T. ISHIKAWA *et al.*, 2006 A SNP in the ABCC11 gene is the determinant of human earwax type. Nat Genet **38:** 324-330.

ZAN, Q., B. WEN, Y. HE, Y. WANG, S. XU *et al.*, 2006 Complete sequence data support lack of balancing selection on PRNP in a natural Chinese population. J Hum Genet.

ZECHNER, U., M. WILDA, H. KEHRER-SAWATZKI, W. VOGEL, R. FUNDELE *et al.*, 2001 A high density of X-linked genes for general cognitive ability: a run-away process shaping human evolution? Trends Genet **17:** 697-701.

ZEH, J. A., and D. W. ZEH, 2006 Outbred embryos rescue inbred half-siblings in mixed-paternity broods of live-bearing females. Nature **439:** 201-203.

ZHANG, K., J. M. AKEY, N. WANG, M. XIONG, R. CHAKRABORTY *et al.*, 2003 Randomly distributed crossovers may generate block-like patterns of linkage disequilibrium: an act of genetic drift. Hum Genet **113:** 51-59.

VITA

Kerry Leigh Bubb was born in Syracuse, New York and raised in upstate New York. She earned a Bachelor of Sciences from Cornell University. She now lives in Seattle with her husband and two children.