Consequential Repetition:
Microsatellites as Targets of Natural Selection

By

Ryan James Haasl

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

Doctor of Philosophy
(Genetics)

at the
UNIVERSITY OF WISCONSIN–MADISON
2013

Date of final oral examination: 1/28/13

The dissertation is approved by the following members of the Final Oral Committee:
Bret A. Payseur, Associate Professor, Genetics
Colin N. Dewey, Associate Professor, Statistics
John F. Doebley, Professor, Genetics
Bret R. Larget, Associate Professor, Statistics
Donald M. Waller, Professor, Botany

For my two bright stars, Shannon and Charlotte.

CONTENTS

Contents ii

List of Figures vi

List of Tables ix

Acknowledgements x

Chapter 1: Introduction 1

Microsatellites, information content, and junk DNA 1

Populations, mutation, and neutral polymorphism 2

Complicated mutation 3

Neutral microsatellites 7

The non-neutral microsatellite 9

Functional microsatellite variation 10

Dangerous microsatellites 13

Genomic evidence in support of selection targeting microsatellites 14

Speculation surrounding the topic of microsatellite selection and the need for a careful treatment of its population and genomic consequences 16

Organization of the thesis 20

Chapter 2: Microsatellites as targets of natural selection 22

Summary 22

Introduction 22

Models and Simulation 26

Modeling the fitness surface of a microsatellite 26

Modeling the microsatellite mutation matrix 28

Rapid forward simulation of natural selection, mutation, and drift a microsatellite using a recursion equation 31

Results 32

Picturing mutation-selection-drift equilibrium at a microsatellite 32

The evolution of Friedreich's ataxia and its causative microsatellite 34

Population-level characteristics of microsatellite selection 39

Discussion 41

The role of mutational complexity in genomic scans for selection 41

Detecting microsatellite selection 43

The cost and duration of microsatellite selection are dependent on several factors 44

Inferring the origin and selective regime of Friedreich's ataxia 46

Extending models of the fitness surface to other mutliallelic variants 48

Methods 49

Characterizing the effects of microsatellite selection at the population level 51

Comparing the selective footprints of selection targeting SNP variants versus microsatellites 52

Chapter 3: The effects of microsatellite selection on linked variation 54

Summary 54

Introduction 55

Results 59

The spatial footprint of selection on microsatellites 59

The temporal footprint of selection on microsatellites 62

Patterns of genetic variation linked to a microsatellite candidate for selection 69

Discussion 73

The CA repeat in intron 1 of *HSD11B2* 79

Prospects for genome scans on microsatellites and other non-SNV targets . 80

Methods 82

Models of selection and mutation 82

Simulation 83

Summary statistics 86

Power analyses 87

Sequence analysis of the *HSD11B2* microsatellite 89

Chapter 4: Evidence of natural selection at human microsatellite loci 90

Introduction 90

Genotyping 1000 Genomes data 91

Reference data set: putatively neutral, intergenic dinucleotides 92

Rarity of coding dinucleotdies in the human genome and remarkable conservation of allele size across primates 96

Model selection and parameter inference 102

Bimodal models of the genotypic fitness surface 102

Inference protocol 104

Using cross-validation to assess the reliability of model choice 107

Using posterior predictive checks to assess the reliability of a specific model choice 109

Testing candidate microsatellites for selection 110

Allele frequency distributions and genomic context of the candidate microsatellites 110

Extended analysis of the intronic *GRIN2B* microsatellite 117

Analysis of the other five candidate genes 123

Discussion 132

Chapter 5: Conclusions 139

An inferential framework for detecting selection targeting microsatellites 139

Implications 141

Improving inference of microsatellite selection 144

References 147

Appendix 186

LIST OF FIGURES

Figure 1-1: Plausible relationships between allele size and gene expression . . . 18

Figure 1-2: Microsatellites and QTLs 20

Figure 2-1: Modeling mutation and selection at a microsatellite 28

Figure 2-2: Mutation-selection-drift equilibrium at a microsatellite under selection 34

Figure 2-3: Alternative demographic models for the origin of the Friedreich's ataxia mutation 36

Figure 2-4: Estimate of the genotypic fitness surface for the `GAA` microsatellite that causes Friedreich's ataxia 37

Figure 2-5: Cost and duration of selection for different models of microsatellite selection 40

Figure 2-6: The similarity of selective footprints left by soft sweeps and selection on microsatellites 44

Figure 3-1: The selective footprint of selection on single nucleotide variants and microsatellites 60

Figure 3-2: Statistical power to detect selection for summary statistics of the site frequency spectrum 64

Figure 3-3: Statistical power to detect selection for summary statistics of haplotype configuration 66

Figure 3-4: Variation in the haplotype configuration during and following a selective event 68

Figure 3-5: Changes to the statistics A – which quantifies uniqueness of the most common haplotype – during and following a selective event 69

Figure 3-6: Significant values of Zeng et al's E across a 2Mb region flanking a microsatellite in *HSD11B2* 73

Figure 4-1: Number of alleles versus allele size for 53 intergenic `CA` microsatellites across the world 93

Figure 4-2: Heterozyogsity versus allele size for 53 intergenic `CA` microsatellites across the world 94

Figure 4-3: Mean ratio of variance in allele size between seven 1000 Genomes populations, across 41 intergenic `CA` microsatellites 96

Figure 4-4: Amino acid conservation in the C-terminus of *FGFRL1* 102

Figure 4-5: Two models of a bimodal genotypic fitness surface 103

Figure 4-6: Observed allele frequency distribution of *GRIN2B* 111

Figure 4-7: Observed allele frequency distribution of *PAX3* 112

Figure 4-8: Observed allele frequency distribution of *SHBG* 113

Figure 4-9: Observed allele frequency distribution of *EGFR* 114

Figure 4-10: Observed allele frequency distribution of *AKR1B1* 115

Figure 4-11: Observed allele frequency distribution of *MMP9* 116

Figure 4-12 Posterior distributions of the mutation parameters ψ and ϕ and the selection parameters δ and g_u at the *GRIN2B* 121

Figure 4-13: The number of segregating sites less the number of unique haplotypes and Tajima's D along the 200kb sequence flanking the focal microsatellite in *GRIN2B* 122

Figure 4-14: Posterior predictive distributions for the *GRIN2B* microsatellite locus123

Figure 4-15: Posterior distributions for the *PAX3* microsatellite before and after adjustment by local linear regression 125

Figure 4-16: Fitness surfaces for the six candidate microsatellites 135

Figure A-2-1: Comparing linear and logistic models of microsatellite mutation 186

Figure A-2-2: Comparing individual-based and recursion-based simulations . . 187

Figure A-2-3: Posterior distributions for inference on Friedreich's ataxia 188

Figure A-2-4: Genetic load associated with Friedreich's ataxia 189

Figure A-3-1: Spatial footprint of selection measured by Fay and Wu's H_{FW} and Zeng et al.'s E 190

Figure A-3-2:Values of D for sequence flanking the *HSD11B2* microsatellite . . 191

Figure A-3-3: Values of H_{FW} for sequence flanking the *HSD11B2* microsatellite 192

Figure A-3-4: An extended view of K flanking the *HSD11B2* microsatellite . . . 193

Figure A-3-5: Values of H for sequence flanking the *HSD11B2* microsatellite . . 194

Figure A-3-6: Values of M for sequence flanking the *HSD11B2* microsatellite . . 195

Figure A-3-7: Observed allele frequency distribution for the *HSD11B2* microsatellite 196

Figure A-3-8: Observed allele frequency distribution for a `CA` microsatelilte 18kb downstream of gene *SLC24A5* 197

Figure A-3-9: Observed allele frequency distribution for a `CA` microsatelilte 1.4kb downstream of gene *CYP3A5* 198

Figure A-3-10: Mutation rate curves used in Chapter 3 199

Figure A-3-11: Sex-averaged recombination rate in the region flanking the *HSD11B2* microsatellite 200

Figure A-4-1: Distribution of the *PAX3* microsatellite genotypes in eight 1000 Genomes populations. 201

LIST OF TABLES

Table 2-1:Prior and posterior distributions for inference regarding Friedreich's ataxia . . . 38

Table 2-2 Simulated selective regimes and coefficients of determination for regression of C and d on Δ_{msat} . . . 40

Table 3-1: Allele frequency distribution of the `CA` repeat in intron 1 of gene *HSD11B2* . . . 71

Table 4-1: 1000 Genomes populations and their sample sizes . . . 92

Table 4-2: Polymorphism and divergence of human coding dinucleotides . . . 101

Table 4-3: Confusion matrix from cross-validation analysis of the seven models tested . . . 108

Table 4-4: Posterior probabilities for model selection at seven candidate microsatellites . . . 133

Table 4-5: Posterior probabilities for model parameters at six candidate microsatellites134

Table 4-6: Results from scans of the 400kb sequence flanking the six candidate microsatellites . . . 136

ACKNOWLEDGMENTS

> *It would all be done with keys on alphanumeric keyboards that stood for weightless, invisible chains of electronic presence or absence.*
>
> — Thomas Pynchon, Vineland (1990)

The work presented here required the constant, loving support of my family. In particular, my wife Shannon has been incredibly patient with me. She has allowed me to pursue my academic interests without complaint, for which I am eternally grateful. Her sense of humor, undying love for our new daughter, and bleeding heart liberalism are three items in a giant set comprising the things I love about her.

Our daughter Charlotte, the prairie sparrow, has been a joy to behold as I have completed my last year of the Genetics Ph.D. program. Her gentle laugh, knowing smile, and enthusiastic waves inspire me to be a better and harder working person. I cannot wait to watch you grow up! Love you, Charlie.

My mother Mary Lynn has of course been there from the start. Surely, one can trace the academic path I have taken back to an apartment above a furniture shop in a tiny Wisconsin town, where she exposed me to books and ideas that few children have the pleasure to experience at such an early age. My grandparents Leonard and Dawne helped raise me too, and their good wisdom and gentle souls are forever part of me.

I also want to thank my in-laws David and Sandra Schroederus, who, like their daughter, have put up with a man who apparently takes his time getting where he is going. Also, Sandy's help watching Charlotte over the past year is a direct reason that I was able to complete my thesis work this year. Thank you.

The mentorship of my advisor, Bret A. Payseur, has been transformative. Over the last five years, my scientific and professional skill set has grown tremendously thanks to his thoughtful guidance. I tell people all the time that Bret is a wonderful advisor, and it's

true! I am always amazed by the practical but inspirational advice he offers when I come to him with a problem. I wish I had a mind that worked like that. Moreover, I appreciate his willingness to let me pursue a variety of crazy ideas within the constraints of our research program. Some of them worked and are in this thesis. Others didn't and are not. I was never bored in the lab or by the work I did and I think that's supposed to be the way it goes, right?

Speaking of the lab, I am especially indebted to the help of Peicheng Jing and thankful for his constant smile. People always say that about people ... "they're always smiling" ... but it's actually true in this case. Thanks as well to all the lab members over the years who have been incredibly supportive and respectful of me and each other.

Finally, my advisory committee was incredible. I feel honored to have shared my work with such an accomplished group of scientists: Colin Dewey, John Doebley, Bret Larget, and Don Waller. In particular, Don Waller has proved critical to my continued scientific and professional development.

Inspired by Nature and computation, my promise to all of you is that I will continue this life of inquiry with a happy heart.

CHAPTER 1: INTRODUCTION

Microsatellites, information content, and junk DNA

Consider the following repetitive nucleotide sequence.

```
CACACACACACACACACACACACA
```

This type of repetitive sequence is commonly called a short tandem repeat, or, as I will refer to it throughout this thesis, a *microsatellite*. A microsatellite sequence repeats a nucleotide motif of 1-6 base pairs – e.g., the dinucleotide `CA` above.

Devoid of population and genomic context, a microsatellite sequence provides little to no information.Iterating across the sequence in steps of two base pairs yields no surprise. `CA` occurs with probability one and we therefore gain no information by observing the next `CA` in the series. In the parlance of information theory, this sequence has zero entropy, or complete predictability (Baldi and Brunak, 1998).

Given their rather featureless profile, it may seem difficult to contemplate a role for these pedestrian sequences in the ontogeny or function of an individual. Primarily for this reason, microsatellites are commonly relegated to the category of junk DNA (Trifonov, 1989), which encompasses the large non-coding fraction of most eurkaryotic genomes (Ohno, 1970). ***In this thesis, I consider a less dismissive possibility and develop methods to evaluate the hypothesis that an appreciable but unknown number of microsatellite loci are functional and therefore subject to natural selection.*** In the following introduction, I justify this exploration by placing the isolated microsatellite sequence above in the context of populations and genomes.

Reappraising the importance of ostensibly functionless DNA variation is timely. Last year, results published in a series of high-profile papers collectively suggested that a majority of the human genome is expressed and therefore potentially functional, despite the fact that only $\sim$ 2% of the genome consists of coding sequence (ENCODE Project Consortium, 2012). The increasing pace at which genomic data sets are being released ensures that bold conclusions such as this will continue to appear. In response, molecular and evolutionary geneticists must develop the means to assess the validity, and, then, the biological and evolutionary implications of such claims. The work presented here was performed in this inquisitive spirit. I model microsatellite selection (Chapter 2), assess its population (Chapter 2) and genomic level (Chapter 3) consequences, and develop and use methods that enable inference of selection targeting microsatellites (Chapters 2-4).

Populations, mutation, and neutral polymorphism

Next consider a series of microsatellites.

```
CACACACACACACA
CACACACACACACACACACACACACACACACA
CACACA
CACACACACA
```

Imagine these sequences represent a small microsatellite data set sampled from a population of lizards. Although each sequence considered on its own still provides no information, in the context of a population sample these orthologous sequences are no longer completely predictable. There is length variation at the locus and observing this sample provides information. It tells us something about the population, the microsatellite itself, and what we might expect if we were to sample another four chromosomes. For example, that four of four sequences are different in size suggests the locus is highly polymorphic. This in turn suggests the locus is highly mutable and/or that population size is large.

Complicated mutation

The ultimate origin of observed polymorphism is mutation. Microsatellite mutation is complicated on a number of levels. Indeed, the complexity of microsatellite mutation motivates much of the work presented in this thesis. I now detail the nature of this complexity using frequent comparisons to point mutation.

Mutational mechanism

The mechanism of microsatellite mutation is fundamentally different than point mutation. While spontaneous point mutations are caused by deamination of cytosine residues and copying errors during replication, the primary cause of microsatellite mutation is slipped-strand mispairing (Levinson and Gutman, 1987; Schlötterer and Tautz, 1992; Ellegren, 2004). During DNA replication, repetitive sequences are particularly vulnerable to ″slipping" on the template strand, which results in misalignment of the synthesized and template strands. If this primary mutation is not properly corrected by the DNA mismatch repair (MMR) system, a lasting mutation is sustained. As a result, defects in MMR genes lead to high effective rates of microsatellite mutation (Chang et al., 2001); resulting microsatellite instability is a key marker of numerous human cancers (Liu et al., 1995; Oki et al., 1999; de la Chapelle, 2003).

Mutational outcome

Microsatellite mutation increases or decreases the number of times its component nucleotide motif is repeated. While point mutation changes the sequence state of a single base pair, microsatellite mutation changes DNA in two dimensions. First, the sequence itself changes. Like point mutants, microsatellite-based sequence change can alter protein structure and downstream phenotype (Fondon and Garner, 2004) as well as sequence-dependent functions such as transcription factor binding (Contente et al., 2002; Martin et al., 2005). Second, the physical length of the DNA molecule is changed. As discussed in more detail below, the change in length associated with microsatellite mutation introduces a host of functional consequences that are more difficult or impossible to achieve by point mutation. For example, the size of a microsatellite has been shown to modulate nucleosome positioning (Godde and Wolfe, 1996; Sandman and Reeve, 1999; Tomita et al., 2002; Vinces et al., 2009), the formation of Z-DNA (Naylor and Clark, 1990; Rothenburg et al., 2001), and, in the case of genic microsatellites, the structure and splicing of RNA products (Tian et al., 2000; Galvao et al., 2001; Sobczak et al., 2003; Hefferon et al., 2004; Rozanska et al., 2007).

High mutation rate

Microsatellite mutation rate in the human genome is thought to range from $10^{-6} - 10^{-2}$ (Weber and Wong, 1993; Eckert and Hile, 2009), which is many orders of magnitude greater than current estimates of point mutation rate that range from 1 to 2.5×10^{-8} per site per generation (Nachman and Crowell, 2000; Lynch, 2010; Roach et al., 2010). High mutation rate is the main reason why population genetic models of point mutation are generally not applicable to microsatellite mutation. For example, the infinite sites model (Kimura, 1969) assumes that each site may only be hit by one mutation, while the infinite alleles model (Kimura and Crow, 1964) assumes that each mutation generates a distinct allele. Due to recurrent mutation, neither of these assumptions holds in the case of microsatellites, although the infinite alleles model may be approximately correct when microsatellite mutation rate is very low (Estoup et al., 2002; Haasl and Payseur, 2011). The stepwise mutation model (SMM; Ohta and Kimura, 1973; Kimura and Ohta, 1975) is commonly used in analytical and simulation-based research related to microsatellites. The SMM allows recurrent mutation and assumes that each mutation increases or decreases allele size by one repeat unit.

Variable mutation rate

Although microsatellite mutation rate is several orders of magnitude greater than point mutation rate, remarkable rate variation exists between loci and between species. Motif sequence (e.g., `CA` vs. `CG`) as well as motif length (e.g., di- vs. tri-nucleotides) are correlated with differences in mutation rate (Kelkar et al., 2008; Marriage et al., 2009; Payseur et al., 2011; Sun et al., 2012). Perhaps the most well-supported determinant of microsatellite mutation rate is allele size. In particular, a positive correlation exists between allele size and mutability (Goldstein and Clark, 1995; Wierdl et al., 1997; Brinkmann et al., 1998; Vigouroux et al., 2002; Leopoldino and Pena, 2003; Henke and Henke, 2006; Seyfert et al., 2008; Aandahl et al., 2012). The reason for this correlation seems clear; longer repeats have a greater chance to slip along the template strand during DNA replication. However, the exact form of the relationship between mutation rate and allele size is unclear. A recent, direct characterization of mutation at di- and tetranucleotide microsatellites supported a linear increase with size for both classes (Sun et al., 2012). Yet alleles sizes of less than 10 were not considered in this study. Polymorphism data that include the full spectrum of allele sizes suggest that the most dramatic increase in mutation rate occurs at an allele size < 10 (Brandstrom and Ellegren, 2008; Payseur et al., 2011). Interspecific differences in mutation rate are also common. Rubinsztein et al. (1995) showed that allele size distributions for orthologous microsatellites were significantly different among a number of closely related primates, implying recent evolution of differences in mutation rate. Schug et al. (1997) found an *average* microsatellite mutation rate in *Drosophila melanogaster* of 6.3×10^{-6}, which is dramatically lower than the estimated *range* of microsatellite mutation rates in most vertebrates. Similarly, estimated mutation rates of di- and trinucleotide repeats in the social amoeba *Dictyostelium discoideum* are only 1 to 6×10^{-6} (McConnell et al., 2007).

Mutlistep mutation and contraction bias

In addition to mutation rate variation, empirical data suggest considerable variation in the product of mutation. First, it appears that an appreciable fraction of mutations at microsatellites change allele size by more than one repeat unit (DiRienzo et al., 1994; Rubinsztein et al., 1995; Ellegren, 2000; Xu et al., 2000). This violates the key assumption of the SMM, which holds that all changes are one repeat unit in size. Various two-phase mutational models, including the generalized stepwise mutation model (GSM), address this by modeling mutational step size with a geometrical distribution, where parameter p is the probability of a single-step mutation. Second, in addition to higher mutation rates, longer microsatellite alleles appear to contract more frequently than they expand (Amos et al., 1996; Xu et al., 2000).

Interruptions by point mutation

A so-called perfect repeat is a microsatellite that does not deviate from the perfect repetition of its nucleotide motif. As microsatellites grow in length, however, they become increasingly large mutational targets. Thus, long microsatellites frequently contain single nucleotide variants that interrupt the otherwise perfect repeating sequence. The mutational properties of imperfect repeats are difficult to predict. For example, imagine a $\texttt{CA}_{15}$ repeat whose ninth `CA` becomes `TA` by point mutation. Does this interruption reduce mutation rate? By how much? Answers to these questions have not been forthcoming.

In Chapter 2, we present a model of microsatellite mutation that incorporates rate heterogeneity, multistep mutation, and contraction bias. In simulation, we assume that microsatellites are perfect repeats and in my analyses of empirical data we restrict myself to microsatellites that appear to be perfect repeats.

Neutral microsatellites

Neutral polymorphism

Despite the complexity of microsatellite mutation, a number of important theoretical results have been obtained for microsatellites. These results are all based on analytical derivations that assume the SMM and neutral evolution. Moran (1975) showed that the ordered distribution of allele frequencies at a locus, $\{p_i\} = p_2, p_3, p_4, ...$, where i is allele size, does not have a limiting distribution. In other words, $\{p_i\}$ wanders the i axis over time and mean allele size changes accordingly. However, the variance of $\{p_i\}$ does reach an equilibrium assuming that mutation rate and population size are constant (Moran, 1975). If we assume this equilibrium, set mean allele size to zero, and adjust other allele sizes accordingly, the resulting distribution is symmetric exponential in form (Beder, 1988; Valdes et al., 1993).

Numerous studies have explored the relationship between divergence time and genetic distance at a microsatellite, leading to the important result that this relationship is linear for certain genetic distances (Goldstein et al., 1995a,b; Zhivotovsky and Feldman, 1995; Sun et al., 2009). Pritchard and Feldman (1996) investigated the divergence measure s, which is the difference in allele size between to individuals. They derived formulas for the the expected value s^2 and its variance under a number of important demographic conditions, including structured and bottlenecked populations.

Despite its relative simplicity, the SMM leads to patterns of variation that are more difficult to interpret than those of sequence data. Most importantly, recurrent mutation at a microsatellite generates homoplastic alleles that are identical by state but not descent. Following the introduction of the infinite alleles and infinite sites models for sequence evolution, closed-form solutions for the sampling distribution and allele frequency spectrum of single nucleotide polymorphisms (SNPs) were introduced (Ewens, 1972; Watterson, 1975). Attempts to derive comparable distributions for microsatellites are hampered by recurrent mutation and the resultant homoplasy. Kimura and Ohta (1975, 1978) provided a frequency spectrum for alleles evolving under the SMM. However, their derivation assumed very low mutation rate relative to that of most microsatellites. More recently, Rosenberg and Jakobsson (2008) used analytical derivations and analysis of empirical data to show the strong correspondence between expected homozygosity and the frequency of the most frequent allele at a microsatellite locus. We used the results from simulations of microsatellite evolution following the SMM to show that the microsatellite allele frequency spectrum is well approximated by a gamma distribution across a broad range of mutation rates and sample sizes (Haasl and Payseur, 2010). Moreover, we showed that sample sizes of $n > 100$ chromosomes are required for the recovery of all but the rarest alleles, and developed three new estimators of $\theta = 4N_e\mu$ (where N_e is effective population size and μ is mutation rate at the locus). These estimators often possessed lower mean squared errors than other commonly used estimators of microsatellite θ, including the sophisticated and computationally intensive method of Beerli and Felsenstein (2001).

Microsatellites as genetic markers

Microsatellite variation was first described in the early 1980s (Miesfeld et al., 1981; Spritz, 1981) and the promise of microsatellites as a source of highly polymorphic genetic markers was soon realized (Tautz et al., 1986). A series of papers in 1989 showed it was possible to access these polymorphic data using the newly developed method of PCR (Litt and Luty, 1989; Tautz, 1989; Weber and May, 1989), which was much easier than the laborious procedure of blotting followed by probe hybridization. By the early 1990s, PCR-amplified microsatellites were being genotyped in a variety of species, including cows (Fries et al., 1990), museum bird specimens (Ellegren, 1991), cetaceans (Schlötterer et al., 1991), humans (Edwards et al., 1992), and a variety of other mammals (Stallings et al., 1991). In 1994, John

Avise predicted that,

> these highly-allelic Mendelian polymorphisms should find widespread application in various areas of population biology such as gene flow estimation and parentage assessment (Avise, 1994).

Indeed, early linkage maps were constructed using microsatellites (Dib et al., 1996; Broman et al., 1998), by 1997 the United States Federal Bureau of Investigation had selected 13 microsatellites as its primary means of DNA fingerprinting (Butler, 2006), and even today microsatellites enjoy widespread use in the characterization and identification of population structure and demographic change (Oliveira et al., 2006). Importantly, all of these applications rely on the truth of the assumption that each microsatellite marker evolves neutrally – i.e., that an allele neither increases nor decreases the survivorship and/or fecundity of an individual.

Historically, it was cost-prohibitive in non-model organisms to sequence sufficient DNA to uncover SNP variation that matched the variation found in a small number of microsatellites. The development of next-generation sequencing technology has changed this and microsatellites are now sometimes viewed as obsolete (Aitken et al., 2004; Morin et al., 2009). However, we recently found that substantially fewer microsatellite loci than SNPs are needed to identify the number of populations in a sample and the presence of population structure when divergence is recent (Haasl and Payseur, 2011). Regardless, the explosion in use of microsatellites as genetic markers beginning in the late 1990s and a long-standing assumption of neutrality reinforced by their simple, zero entropy sequences may have led most geneticists to overlook the possibility of a non-neutral microsatellite.

The non-neutral microsatellite

Now consider the previous series of microsatellites embedded in flanking sequence.

ACGCATGACGGCACACACACACACACAGCACAAGCTAG

ACGCATGACGGCACACACACACACACACACACACACACACACAGCACAAGCTAG

ACGCATGACGGCACACAGCACAAGCTAG

ACGCATGACGGCACACACACAGCACAAGCTAG

From a practical, bioinformatic perspective, flanking sequence aids the identification of orthologous microsatellites within genomes that are littered with microsatellites of the same motif. For this reason, two recent and promising algorithms for calling microsatellite genotypes from next generation sequencing data begin by discarding reads that have insufficient flanking sequence to reliably map them to reference sequence (Gymrek et al., 2012; Highnam et al., 2013). More fundamentally, the four sequences shown above clearly demonstrate that microsatellites are structural variants. The distance between the leading `A` and terminal `G` of each sequence changes with microsatellite allele size. As alluded to above, variation in the length dimension may have important functional consequences, which I now briefly review.

Functional microsatellite variation

In bacteria, phase variation refers to random and reversible switching of phenotypes on and off. The mechanism underlying phase variation is frequently microsatellite mutation (Gemayel et al., 2010). For example, mutation of non-triplet microsatellites in bacterial open reading frames may lead to a non-functional protein product by introducing a premature stop codon, thereby turning off any phenotypes associated with the protein. Subsequent mutations may lead to frame shifts that turn the phenotype back on. Phase variation is particularly prevalent in pathogenic bacteria (Moxon et al., 1994); the population heterogeneity generated by phase variation is thought to provide an array of potentially adaptive solutions in an unpredictable host environment. In the bacterium *Haemophilus influenzae*, mutation at a CAAT microsatellite in the gene *lic1* generates three distinct translation phases by shifting upstream translation initiation codons in and out of frame (Figure 1-1A; Weiser et al., 1989, 1990). The protein product of *lic1* is responsible for the addition of multiple polysaccharide groups to the outer membrane structure of the bacterium. Phase switching thereby alters the presentation of epitopes on its surface, which aids survival of the bacterium in its human host (Weiser et al., 1989). Also in *H. influenzae*, mutation at a TA microsatellite found in the common promoter to the divergently orientated genes *hifA* and *hifB* is responsible for phase variation in the expression of fimbriae on the membrane surface (van Ham et al., 1993). In this case phase variation of the fimbrial phenotype is regulated at the transcriptional level.

The first studies to demonstrate a correlation between microsatellite variation in gene expression were performed in the early 1980s (Russell et al., 1983; Hamada et al., 1984). We now know that microsatellite length variation may be a common means of regulating gene expression (Rockman and Wray, 2002) and that it does so in a variety of important ways. First, some microsatellites provide actual binding sites for transcription factors. Expansion and contraction of these microsatellites directly affects the number of binding sites available to the regulatory protein. A TAAA repeat in the *nadA* promoter of the bacterium *Neisseria meningitidis* modulates binding of transcription factor IHF thereby changing gene expression (Martin et al., 2005). In humans, variation at a TCC microsatellite in the promoter of the epidermal growth factor (EGF) gene alters expression through its effect on the binding efficiency of transcription factor Sp1 (Johnson et al., 1988). Second, when a microsatellite that lies between regulatory and functional elements of a promoter mutates, it changes the physical distance between these elements. In many cases, particularly in pathogenic bacteria, spacing between elements affects levels of gene expression (Willems et al., 1990; Yogev et al., 1991; van Ham et al., 1993). Third, microsatellite variation affects chromatin structure by controlling nucleosome positioning. Vinces et al. (2009) found that nucleosome-free sequences in human and yeast promoters are enriched for microsatellites and other repetitive elements, suggesting that microsatellites inhibit nucleosome formation in critical regions of gene promoters. Furthermore, experimental manipulation of the length of promoter microsatellites in yeast demonstrated a relationship between microsatellite length and gene expression (Vinces et al., 2009). Fourth, microsatellite length can affect the physical properties of DNA sequence. Gebhardt et al. (1999) found that length variation of a microsatellite in the first intron of the epidermal growth factor receptor (EGFR) gene is correlated with transcription levels. Further modeling revealed that longer stretches of the `CA` microsatellite incurred high bendability to the polymorphic region, suggesting that length variation can help bring the promoter into close proximity with a putative regulatory protein that binds near the `CA` repeat (Gebhardt et al., 2000). Finally, variation in the length of *transcribed* microsatellites can also have functional consequences. For example, non-classic presentations of cystic fibrosis are linked to the expansion of a `TG` repeat in the gene *CFTR*, which is thought to perturb wild-type RNA splicing (Hefferon et al., 2004).

Microsatellite variation has also been linked to an interesting variety of disparate phenotypes. In the fungus *Neurospora crassa*, a polyglutamine repeat in the N-terminus of the protein WC-1 is necessary for circadian clock function (Froehlich et al., 2002). The circadian period of *N. crassa* varies across natural accessions; in turn, this variation is correlated with the length of the polyglutamine repeat, which is coded for by a trinucleotide microsatellite (Michael et al., 2007). Moreover, circadian period and microsatellite length were both correlated with the latitude of accession sampling locations. Fondon and Garner (2004) documented a striking correlation between the snout morphology of domestic dog breeds and the length of a compound microsatellite in the gene *Runx-2*. Interestingly, examination of canine skulls over the last century showed that evolution of facial morphology in domestic dog breeds was accomplished rapidly. This suggests that phenotypes linked to microsatellite variation may be uniquely primed to respond rapidly to selective pressures such as the artificial selection imposed by dog breeders. In the vole genus *Microtus*, the length of a microsatellite in the 5′UTR of the vasopressin receptor gene *avpr1a* regulates expression of the gene (Hammock and Young, 2005). Furthermore, populations of montane voles carry substantially longer copies of this microsatellite than congeneric prairie voles. Unlike montane voles, prairie voles form strong pair bonds. The authors attribute this socio-behavioral difference to the difference in microsatellite allele size and subsequent expression difference between the two species and suggest that intraspecific variation at the locus may cause subtler variation in other behavioral traits.

Dangerous microsatellites

Although the preceding examples suggest microsatellite loci are often advantageous to the organisms that carry them, it must be remembered that these are inherently unstable variants. Consider that an exonic dinucleotide repeat will almost always cause a frame shift whenever it mutates. Though largely dependent on genomic context, mutation of these and other microsatellites can produce substantial deleterious effects. Most positive examples of microsatellite length variation have been reported in prokaryotes and non-human eukaryotes, presumably because they are easier to manipulate experimentally. To the contrary, most deleterious examples of microsatellite variation come from the human genetics literature.

A large number of human diseases are collectively referred to as trinucleotide expansion disorders (Orr and Zoghbi, 2007). As the name implies, these diseases are caused by expansions of trinucleotide repeats. In some cases, trinucleotide expansion limits protein production itself. For example, expansion of a `CGG` microsatellite in the 5′ UTR of the *FMR1* gene causes aberrant methylation that inhibits transcription of the gene and results in Fragile X Syndrome (Fu et al., 1991). Friedreich's ataxia (Cossee et al., 1997) is caused by an expansion of a GAA repeat in the first intron of the gene *FXN*, which interferes with transcriptional elongation. Intriguingly, in a number of trinucleotide expansion disorders, the expansion mutations are a gain-of-function mutation. For example, muscular dystrophy (DM1) is caused by the expansion of a CTG repeat in the 3'UTR of the gene (*DMPK*; Brook et al., 1992; Mahadevan et al., 1992). Expanded tracts of CUG in the transcribed mRNA cause binding of two RNA-binding proteins that affect alternative splicing. Interestingly, *DMPK* itself is spliced in a normal fashion, but the CUG-bound RNA-binding proteins induce alternative splicing in a number of other genes that ultimately lead to the phenotypes associated with DM1, including insulin resistance, muscle wasting, and cognitive defects. Huntington's disease is caused by the expansion of a CAG repeat in the gene *HTT*, which codes for a polyglutamine (poly-Q) amino acid sequence (Bhide et al., 1996). Beyond a threshold length, fragments of poly-Q repeat are left behind following post-translational of the protein. Over time these fragments aggregate, forming inclusions in the axons and dendrites of neurons, thereby physically disrupting the flow of vesicles containing neurotransmitters. Cognitive difficulties ensue and death occurs 10-15 years following the initial onset of symptoms Orr and Zoghbi (2007).

Genomic evidence in support of selection targeting microsatellites

Studies that uncover specific examples of functional and deleterious microsatellites provide candidate microsatellites for positive or negative selection. Many research groups have used genomic data to investigate distributions of microsatellites across different genomic compartments, within and among species, in search of statistical evidence for microsatellite selection.

A comparison between the genome of Craig Venter (Levy et al., 2007) and the human genome reference sequence at 2.8 million microsatellite loci revealed that 2.84% of 1.67

million intergenic loci differed between the sequences, whereas only 0.2% of 31,764 exonic loci differed (Payseur et al., 2011). This significant (Fisher's exact test, $P < 10^{-15}$), 14-fold difference between intergenic and coding microsatellites suggests a selective constraint on the coding microsatellites. In a comparison of 32,448 trinucleotide repeats throughout the human genome, Kozlowski et al. (2010) found that a small number of motifs were highly overrepresented in exons, while others were highly underrepresented. These results suggest that motifs themselves may be targets of both positive and purifying selection. In a survey of all perfect trinucleotide repeats in the human transcriptome in three human genomes, Molla et al. (2009) found that the fraction of polymorphic triplet repeats was substantially greater than the polymorphic fraction among exonic repeats for allele sizes of less than 14. Intriguingly, however, the authors found that among longer alleles, the polymorphic fraction was equal regardless of position. This suggests that at the population level, higher mutational pressure associated with longer allele sizes can overpower selection against new mutants alleles – i.e., short alleles are kept in check because they mutate infrequently, while long alleles escape selective pressure through frequent mutation.

Comparing microsatellites across 17 vertebrates Buschiazzo and Gemmell (2010) found that the decline in conservation from human to more divergent species was slower for microsatellites in exons than in UTRs, and slower for microsatellites in UTRs than intergenic microsatellites. This suggests microsatellite size in coding regions is more likely to be maintained by natural selection, ostensibly due to functional consequences. In a related study that analyzed the conservation of microsatellites across the mammalian clade, Sawaya et al. (2012) found that microsatellites near transcription start sites of genes were often highly conserved. Moreover, the distance from a microsatellite to the nearest transcription start site was a good predictor of their measure of conservation. In a more focused study of microsatellite conservation, (Riley and Krieger, 2009) identified dinucleotide microsatellites in the UTRs of humans that were flanked by sequence that was identical between human and the distantly related marsupial. Of the 22 genes with dinucleotide microsatellites meeting these criteria, 18 were critical to mammalian nervous system development, suggesting an intriguing but unclear relationship between dinucleotide variation and nervous system function in mammals.

In *Saccharomyces cerevisiae*, Vinces et al. (2009) showed that $\sim 25\%$ of gene promoters contained microsatellites or other tandem repeats. Comparing expression levels across the roughly 5,700 promoters in the yeast genome, they found that promoters containing repetitive elements had expression levels that were significantly more divergent among eight yeast strains than promoters lacking repetitive elements. The abundance of repetitive elements in these key regulatory regions suggest they are effective regulators of transcription, whose presence may be selected for over time. Also in *S. cerevisiae*, ORFs are significantly enriched for trinucleotides while dinucleotides are very rare; the pattern is reversed in non-ORF sequence (Young et al., 2000). This finding supports a selective constraint against the emergence and expansion of non-triplet repeats in coding regions – presumably selection against frameshift mutations – that has been documented in the genomes of numerous species (Field and Wills, 1998; Bachtrog et al., 1999; Dokholyan et al., 2000; Metzgar et al., 2000; Scotti et al., 2000)

Speculation surrounding the topic of microsatellite selection and the need for a careful treatment of its population and genomic consequences

Expanding upon earlier conjectures that repetitive sequences in general might impact gene regulation (Britten and Davidson, 1969; Georgiev, 1969; Britten and Davidson, 1971; Zuckerkandl, 1974), a dedicated group of scientists has promulgated a “tuning knob" metaphor/hypothesis for microsatellite-mediated control of gene expression and other quantitative traits (Trifonov, 1989; King et al., 1997; Trifonov, 2004; King, 2012). In this metaphor, microsatellite allele size is the tuning knob; the quantitative trait – usually gene expression – changes incrementally in response to incremental changes to allele size (Figure 1-1A). Thus, in response to a changing environment, natural selection may act upon microsatellite length variation to smoothly bring about the optimal level of gene expression with minimal selective cost (Kashi et al., 1997; King, 2012). Furthermore, the hypothesis supposes that the genetic load associated with frequent mutation at microsatellites is minimal because the effect on the quantitative trait is minimal (Kashi et al., 1997; King et al., 1997). If this assumption holds, it means that functional quantitative variation can accrue via mutation during times of ecological stasis at minimal selective cost.

One potential problem with this argument is the assumption of a smooth and gradual relationship between microsatellite allele size and the value of the correlated quantitative trait. In the case of gene expression, numerous empirical studies do support this notion (e.g., Peters et al., 1999; Warpeha et al., 1999; Yamada et al., 2000; Vinces et al., 2009). Yet other studies have identified threshold allele sizes on either side of which a single, distinct phenotype is revealed. For example, an in vitro assay of gene expression using a luciferase reporter demonstrated that the length of a compound microsatellite in the promoter of human gene *PAX6* had a binary effect on gene expression (Okladnova et al., 1998). Allele sizes > 29 showed greater than two-fold increases in expression relative to smaller alleles. In the case the proper metaphor might be a switch (Figure 1-1B), while the phase variation described earlier might be called a repeating switch (Figure 1-1C).

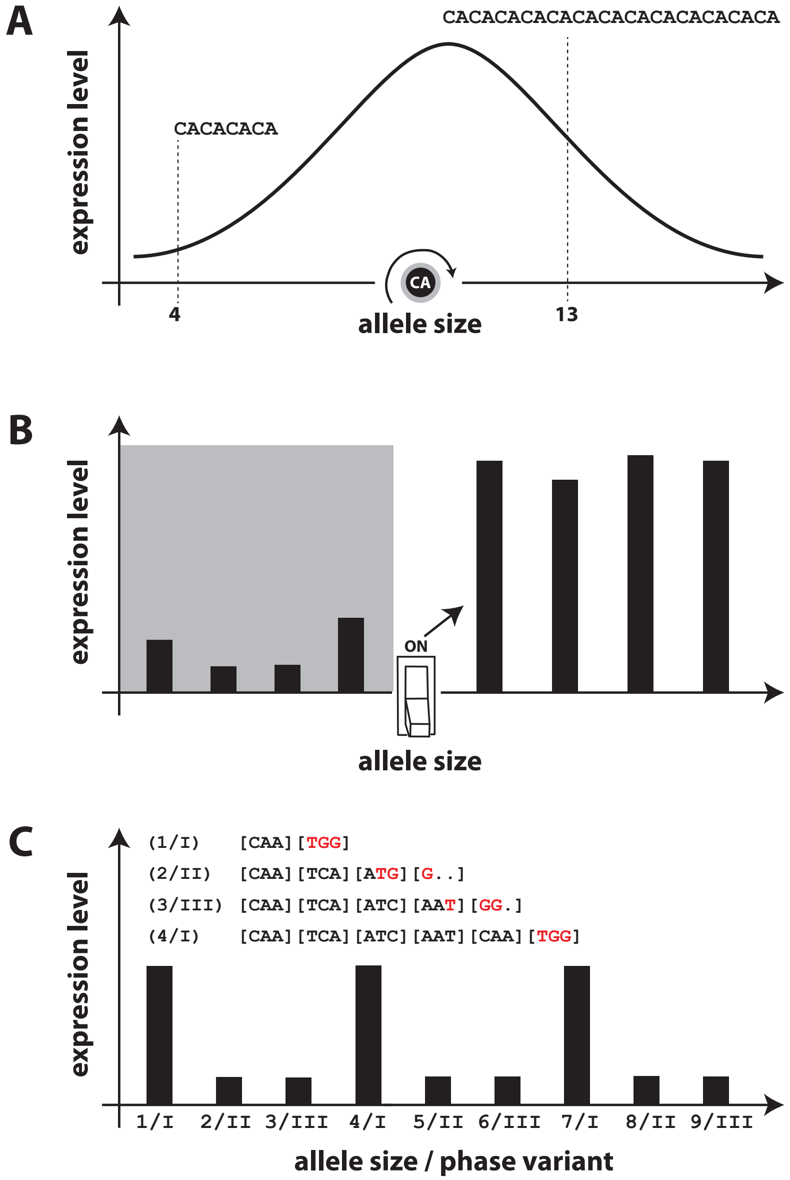
A
expression level
CACACACACACACACACACACACACACA
CACACACA
CA
4
allele size
13
B
expression level
ON
allele size
C
expression level
(1/I) [CAA][TGG]
(2/II) [CAA][TCA][ATG][G..]
(3/III) [CAA][TCA][ATC][AAT][GG.]
(4/I) [CAA][TCA][ATC][AAT][CAA][TGG]
1/I
2/II
3/III
4/I
5/II
6/III
7/I
8/II
9/III
allele size / phase variant

Figure 1-1 Plausible relationships between allele size and gene expression

A more serious deficiency of the tuning knob hypothesis is that no explicit mention is ever made of fitness values, other than the supposition that genetic load is minimal. The examples of functional microsatellite variation mentioned in this chapter suggest that microsatellites may indeed be targets of natural selection. However, to advance the study of non-neutral microsatellite evolution beyond the level of verbal reasoning, it is necessary to model the fitness of microsatellites. As an illustrative example of the importance of this point, consider the following statement from a recent review of the topic:

> As any given TR [tandem repeat, e.g., microsatellite] locus has a wide range of allelic variations, they allow digital rather than binary fine-tuning of specific phenotypes such as binding activities or transcription levels ... gradual, quantitative changes are often believed to depend on multiple genes, or QTLs, and complex genetic interactions. However, the examples summarized in this review show that TRs may in fact provide a simple, monogenic mechanism underlying quantitative changes (Gemayel et al., 2010)

Thus, microsatellites are described as monogenic quantitative trait loci. But how does selection act on such a locus? Consider a species in which height is determined additively based on the sum of short and tall alleles at three diallelic loci (Figure 1-2A). In this case truncation selection for tallness (only individuals taller than the dashed line mate) will move the population towards taller average height because tall parents possess discrete alleles associated with greater height. The situation is trickier in the case of selection for a “monogenic quantitative trait", by virtue of the digital tuning mentioned by Gemayel et al. (2010). Let gene expression increase linearly with allele size, and for convenience let the measure of gene expression equal allele size (Figure 1-2B). Furthermore, let the microsatellite be an autosomal locus in a diploid organism and total gene expression be additive. Thus, 7/8 genotype yields gene expression of 15. Finally, let the optimal level of gene expression be 13. Under these conditions, a possible genotypic fitness surface is shown in Figure 1-2C, where white is highest relative fitness and darker shades represent lower genotypic fitness. In this case, selection for optimal gene expression is impossible or at least highly inefficient All possible offspring of the parental types circled in red in Figure 1-2C possess suboptimal genotypes (Figure 1-2D).

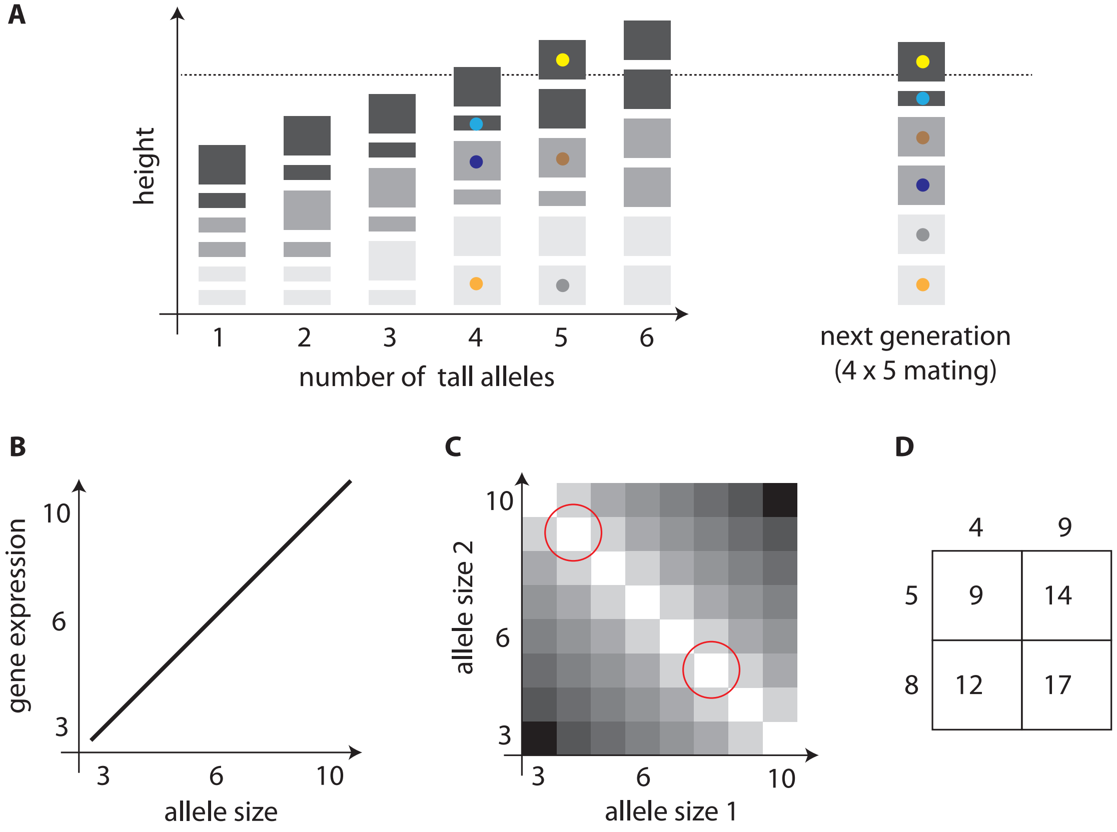

Figure 1-2

This example illustrates the shortcomings of the verbal arguments in favor of microsatellite selection offered to date. The overarching goal of this thesis is to place the concept of microsatellite selection on a more solid footing and provide solutions for identifying selection targeting microsatellites.

Organization of the thesis

In Chapter 2, I introduce four models of natural selection on microsatellites. In addition, I present two models of microsatellite mutation that incorporate many of the complicated aspects of microsatellite mutation discussed above. I then detail a simulation algorithm for rapid generation of microsatellite samples using these models of selection and mutation.

These tools enable inference of microsatellite selection and exploration of the population-level consequences of microsatellite selection. I demonstrate these capabilities by inferring the evolutionary history of the microsatellite that causes Friedreich's ataxia and comparing the cost and duration of selection under different scenarios of microsatellite selection. This chapter was recently published in *Molecular Biology and Evolution* (Haasl and Payseur, 2013), and, with the exception of some formatting details, has not been altered from the published version.

Chapter 3 examines the effects of microsatellite selection on linked variation. In particular, I compare the relative power of statistics based on the site frequency spectrum and haplotype configuration to detect microsatellite selection and SNP-based selective sweeps. Results from this study suggest that patterns of sequence variation can be used to scan for microsatellite selection, thereby offering an inference method that is complementary to the one presented in Chapter 2.

Chapter 4 presents results from the first analysis of microsatellites genotyped in 200 participants of the 1000 Genomes Project from eight populations around the world. We examine the characteristics of 53 intergenic, ostensibly neutral, microsatellites across the world. Then, I investigate patterns of polymorphism at 20 long, exonic dinucleotides and use use the inferential methods described in Chapter 2 and 3 to assess the evidence of natural selection at several strong candidates for selection.

Finally, in Chapter 5 I briefly summarize the implications of the work presented here as well as possible future improvements to the models and methods presented in this thesis.

CHAPTER 2: MICROSATELLITES AS TARGETS OF NATURAL SELECTION

Summary

The ability to survey polymorphism on a genomic scale has enabled genome-wide scans for the targets of natural selection. Theory that connects patterns of genetic variation to evidence of natural selection most often assumes a diallelic locus and no recurrent mutation. Although these assumptions are suitable to selection that targets single nucleotide variants, fundamentally different types of mutation generate abundant polymorphism in genomes. Moreover, recent empirical results suggest mutationally complex, multiallelic loci including microsatellites and copy number variants are sometimes targeted by natural selection. Given their abundance, the lack of inference methods tailored to the mutational peculiarities of these types of loci represents a notable gap in our ability to interrogate genomes for signatures of natural selection. Previous theoretical investigations of mutation-selection balance at multiallelic loci include assumptions that limit their application to inference from empirical data. Focusing on microsatellites, we assess the dynamics and population-level consequences of selection targeting mutationally complex variants. We develop general models of a multiallelic fitness surface, a realistic model of microsatellite mutation, and an efficient simulation algorithm. Using these tools we explore mutation-selection-drift equilibrium at microsatellites and investigate the mutational history and selective regime of the microsatellite that causes Friedreich's ataxia. We characterize microsatellite selective events by their duration and cost, note similarities to sweeps from standing point variation, and conclude it is premature to label microsatellites as ubiquitous agents of efficient adaptive change. Together, our models and simulation algorithm provide a powerful framework for statistical inference, which can be used to test the neutrality of microsatellites and other multiallelic variants.

Introduction

Genomic scans for natural selection are now ubiquitous and target a variety of subject species (Oleksyk et al., 2010; Strasburg et al., 2012). Despite their promise, however, positive results from separate scans of the same species can show limited overlap (Biswas and Akey, 2006; Akey, 2009) and a relatively small number of unambiguously positive results have been gathered (e.g., *LCT* and *G6PD*, Tishkoff et al., 2001; Bersaglieri et al., 2004). Indeed, the prevalence of genomic scans has revealed a number of biological and demographic factors that complicate the intuitive simplicity of the selective sweep model (Maynard Smith and Haigh, 1974) and are likely to confound statistical tests for selection that assume a homogeneous genome. For example, statistics like Tajima's D (Tajima, 1989) may fail to identify selection targeting standing variation (Innan and Kim, 2005; Przeworski et al., 2005), yet produce false positives in response to demographic change (Nielsen et al., 2005; Li, 2011).

A complication that has received little attention is the role diverse mutational mechanisms play in the dynamics and signatures of selection. This oversight is noteworthy since a large fraction of genetic variation is of a fundamentally different mutational nature than a single nucleotide polymorphism (SNP), which is assumed to arise from a single, unique mutation under the infinite sites model (ISM; Kimura, 1969). Though SNPs are the most common type of polymorphism, several mutationally complex structural variants – including micro and minisatellites, copy number variants (CNVs), and transposable elements – are abundant in genomes (Ellegren, 2004; Korbel et al., 2007; Huang et al., 2010). Reliable detection of natural selection across the full complement of mutationally heterogeneous loci will require models (mutational and selective) appropriate to each non-SNP variant.

Here, we focus on microsatellites. Found throughout the genomes of prokaryotes and eukaryotes, microsatellites are defined as sequential repeats of a 1-6 nucleotide motif. The mutation rate at microsatellites generally exceeds that of point mutation by several orders of magnitude (Bhargava and Fuentes, 2010), which leads to recurrent mutation that violates the ISM on which much of the theoretical work regarding SNP-based selection is based (Maynard Smith and Haigh, 1974; Hermisson and Pennings, 2005). Thanks to their early adoption in forensic analysis (Hampikian et al., 2011), genetic map construction (e.g., Broman et al., 1998; Kong et al., 2002), and population genetic inference (e.g., Navascués et al., 2009; Goldberg and Waits, 2010), more is known about microsatellite mutation than other non-SNP variants. For these reasons, microsatellites provide a model system for studying the effects of non-ISM mutation on the inference of natural selection.

Microsatellites have long been used as markers in population genetics and forensic analysis because they are often highly variable (Oliveira et al., 2006). An implicit assumption underlying the use of microsatellites as diagnostic markers is that they evolve neutrally. However, recent studies have identified functional microsatellites that affect the fitness of an individual (Kashi and King, 2006; Gemayel et al., 2010). Putatively (dys)functional microsatellites are primarily located in or near genic regions, where a change in the number of times the motif is repeated (hereafter referred to as *allele size*) is hypothesized to modify gene expression or change protein sequence (Wren et al., 2000; Li et al., 2004; Gemayel et al., 2010). Synthesizing the results of more than 500 individual experiments, Rockman and Wray (2002) concluded that as much as 20% of cis-regulation in humans is mediated by variation in repetitive elements including microsatellites. More recently, Vinces et al. (2009) provided strong experimental evidence for eukaryotic gene regulation via microsatellites. In *Saccharomyces cerevisiae*, the authors demonstrated rapid and effective selection for change in gene expression that was mediated by concomitant change in the allele size of a promoter microsatellite. In exons, changes in protein sequence caused by microsatellite mutation can drive rapid morphological evolution. For example, profound evolution of the snout morphology of domestic dog breeds was accomplished in less than a century through artificial selection acting on the length of a compound microsatellite in the gene *Runx2* (Fondon and Garner, 2004). The presence of microsatellites in coding regions can also present substantial hazard for organisms. For example, most mutations of non-triplet microsatellites in protein coding regions cause frame shifts, which can eliminate protein function. Furthermore, hyperexpansion of trinucleotide repeats in genic regions cause numerous human diseases such as Fragile X syndrome (Kremer et al., 1991), Friedreich's ataxia (Durr et al., 1996), and Huntington's disease (Huntington's Disease Collaborative Research Group, 1993).

Though these empirical examples show that repetitive elements can be functional, a few authors have suggested that repetitive variants including microsatellites may be ubiquitous agents of efficient adaptive evolution (Trifonov, 1989; King, 1994; Kashi et al., 1997; King et al., 1997; King, 1999; Fondon and Garner, 2004; Trifonov, 2004; Kashi and King, 2006; King and Kashi, 2009). In general, they argue that if small changes in allele size at a microsatellite correspond to incremental changes in the value of a quantitative trait such as gene expression, then high mutation at a microsatellite should generate a reservoir of quantitative trait variation to be drawn on in times of ecological stress. Although theoretical and empirical studies have focused on the use of microsatellites markers to detect selective sweeps targeting linked variation (Wiehe, 1998; Schlötterer, 2002; Nair et al., 2003; Rockman et al., 2005), a paucity of research addresses the topic of direct microsatellite selection. An objective, inferential framework to test the neutrality of microsatellites is absent.

Natural selection at a microsatellite is perhaps best considered in the context of mutation-selection balance. While the action of natural selection tends to increase mean fitness of the population, mutation acts in constant opposition to this increase by producing less fit alleles. Previous theoretical treatments of mutation-selection dynamics at loci with multiple alleles make assumptions that limit their application to inference from microsatellite data. Both Crow and Kimura (1970) and Clark (1998) assume the infinite alleles model of mutation (Kimura and Crow, 1964), which is inappropriate to microsatellite mutation unless the selective event of interest is recent enough or mutation rate is low enough to limit recurrent mutation and resultant homoplasy. Several studies have investigated mutation-selection balance at a locus mutating according to the stepwise mutation model (SMM) (Moran, 1976; Kingman, 1977; Moran, 1977; Bürger, 1988, 1998); the SMM is a simple but appropriate model for microsatellite mutation (Ohta and Kimura, 1973). However, these studies make several assumptions that limit their practical use: haploidy, deterministic evolution, and, often, that a single allele is most fit.

The models of selection and mutation presented here empower exploration of diverse selective and mutational dynamics at microsatellites in diploids. We also describe a rapid simulation algorithm, which makes it simple to generate thousands of sample datasets. Together, models and simulation provide a reasonable framework to: (1) test the neutrality of individual microsatellite loci, which is simply assumed in most studies that use microsatellite markers; (2) evaluate claims regarding the importance and prevalence of selection targeting microsatellites, and; (3) investigate the population-level consequences of selection targeting microsatellites. Although we focus on microsatellites as a molecular model system, our models and simulation algorithm should be portable to other classes of multiallelic loci such as CNVs assuming a variant-specific mutational matrix can be constructed.

Models and Simulation

Modeling the fitness surface of a microsatellite

We present four models for the fitness surface of a microsatellite locus: additive, multiplicative, dominant, and recessive. Using four parameters – key allele size (x), threshold effect (δ), and lower and upper gradient effects (g_l and g_u) – the fitness surface is constructed in two steps. Regardless of model, the first step is to calculate a vector of allelic fitnesses. Let a_i represent an allele of size i and let $w(a_i)$ be its fitness. Initially, set $w(a_i) = 1,\ i = 2, 3, 4...$ Then, a detrimental effect of allele a_i on fitness is indicated by $w(a_i) < 1$. The sign of threshold effect δ determines which set of alleles are subject to its effect. When negative, it reduces the fitness of all alleles $< x$ equally; when positive, it reduces the fitness of all alleles $> x$ equally. More specifically, when δ is negative add δ to $w(a_i)$ for all a_i where $i < x$. When δ is positive subtract δ from $w(a_i)$ for all a_i where $i > x$. Gradient effects g_l

and g_u affect the fitness of alleles of size $i < x$ and $i > x$, respectively. When negative, these parameters decrease fitness as distance from x increases and vice-versa. To realize these effects, add $g_l|x - i|$ to $w(a_i)$ for all a_i where $i < x$ and $g_u|x - i|$ to $w(a_i)$ for all a_i where $i > x$. Finally, lethal alleles are represented by a relative fitness of zero. For all i considered, set $w(a_i) = 0$ if $w(a_i) < 0$ after the previous calculations are performed. The second step is to construct the diploid fitness surface in a model-specific manner. Let $w(a_i a_j)$ be the fitness of the diploid genotype containing alleles of size i and j. Under additive and multiplicative models, $w(a_i a_j)$ equals the sum or product of the fitnesses $w(a_i)$ and $w(a_j)$, respectively. Under the dominant model, deleterious effects are dominant. Thus, genotypic fitness is calculated as the minimum fitness of the two component alleles: $w(a_i a_j) = \min\left(w(a_i), w(a_j)\right)$. Under the recessive model, deleterious effects are recessive. Thus, genotypic fitness is equal to the maximum fitness of the component alleles: $w(a_i a_j) = \max\left(w(a_i), w(a_j)\right)$. For all four models, the fitness surface is normalized by dividing each $w(a_i a_j)$ by $\max(w(a_i a_j))$. Figure 2-1A shows a schematic of fitness surface construction.

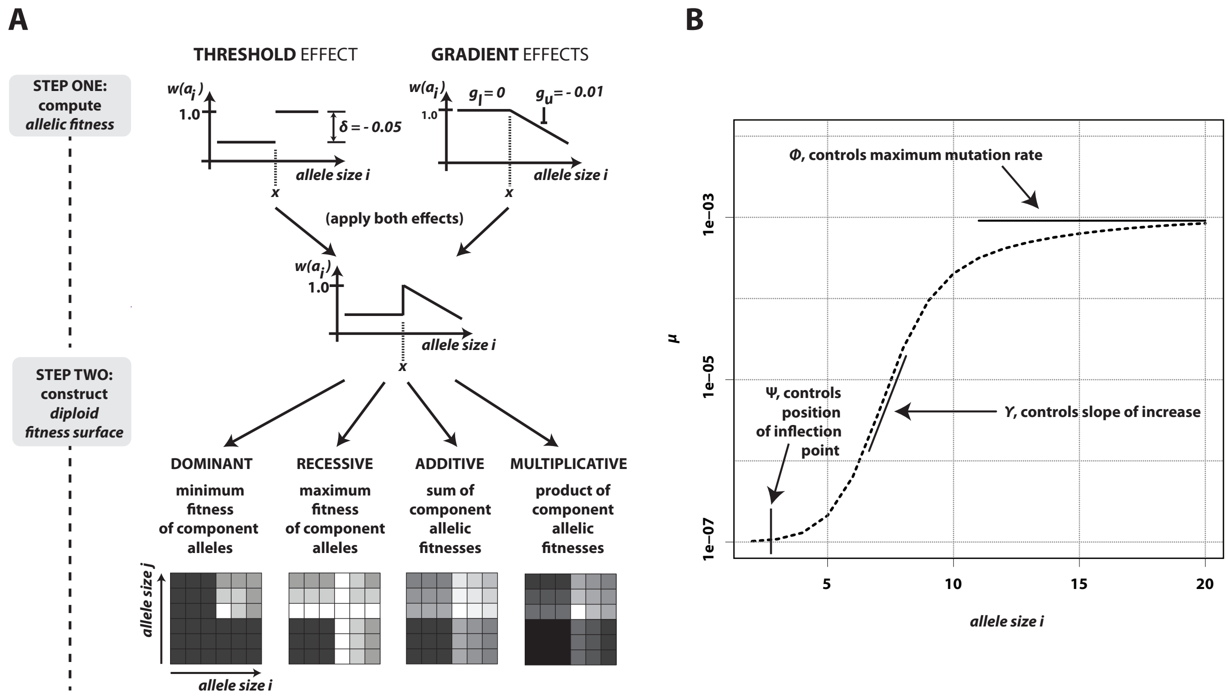
A
THRESHOLD EFFECT
GRADIENT EFFECTS
STEP ONE: compute allelic fitness
w(a_i)
1.0
δ = - 0.05
g_l = 0
g_u = - 0.01
allele size i
x
(apply both effects)
STEP TWO: construct diploid fitness surface
DOMINANT
minimum fitness of component alleles
RECESSIVE
maximum fitness of component alleles
ADDITIVE
sum of component allelic fitnesses
MULTIPLICATIVE
product of component allelic fitnesses
allele size j
B
Φ, controls maximum mutation rate
Ψ, controls position of inflection point
Y, controls slope of increase
μ
1e−03
1e−05
1e−07
5
10
15
20
allele size i

Figure 2-1: Modeling mutation and selection at a microsatellite. (A) The diploid fitness surface is constructed in two steps. First, allelic fitnesses are calculated by combining the threshold and gradient effects associated with the values of parameters δ, g_l, and g_u. Second, the vector of allelic fitnesses is used to compute the fitness surface (genotypic fitnesses) in a model-specific manner. (B) Allele-specific mutation rate is defined as a basic logistic function modified by three parameters whose values control the allele size where mutation rate begins to increase (ψ), the slope of increase (γ), and the maximum mutation rate (ϕ).

Modeling the microsatellite mutation matrix

A positive correlation between allele size and mutation rate is supported by mutational studies (Goldstein and Clark, 1995; Wierdl et al., 1997; Brinkmann et al., 1998; Schlötterer et al., 1998; Vigouroux et al., 2002; Leopoldino and Pena, 2003; Henke and Henke, 2006; McConnell et al., 2007; Seyfert et al., 2008; Marriage et al., 2009; Sun et al., 2012), analyses of polymorphism data (Ellegren, 2000; Legendre et al., 2007; Brandstrom and Ellegren, 2008; Kelkar et al., 2008; Payseur et al., 2011), and model-based inference (Aandahl et al., 2012).

Several studies have modeled this size-dependent aspect of microsatellite mutation rate using a linear or polynomial function of allele size (Kruglyak et al., 1998; Calabrese et al., 2001; Sibly et al., 2001). However, genome-wide analyses of polymorphism data further suggest that mutation rate increases rapidly over a short range of allele sizes after which mutation rate appears to asymptote (Brandstrom and Ellegren, 2008; Payseur et al., 2011). This characteristic suggests that a logistic function might be a reasonable alternative model for allele-specific mutation rate. We use three parameters to modify the logistic function and control allele-specific mutation rate: ψ controls the position of the upward inflection point of mutation rate on the allele-size axis, ϕ controls maximum mutation rate, and γ controls the slope of increase in mutation rate (Figure 2-1B). Following the general formula for the logistic function, allele specific mutation rate μ is:

$$\mu(g, \psi, \phi, \gamma) = 10 \exp\left[\frac{\phi(1 - e^{-g\gamma})}{1 + 10^{\psi} e^{-g}} - 7\right], \quad g \geq 2, \tag{1}$$

where g is current allele size. Recent studies suggest a linear increase in mutation rate with allele size (Aandahl et al., 2012; Sun et al., 2012). A linear model of mutation rate requires only two parameters, slope b and intercept a:

$$\mu(g, a, b) = \begin{cases} a + bg & \text{if } a + bg > 0 \\ 0 & \text{otherwise}, \qquad g \geq 2, \end{cases}$$

Note that negative values of a can lead to $\mu = 0$ for small allele sizes. Indeed, based on human mutation data and assuming a linear model of allele-specific mutation rate, Sun et al. (2012) infer negative intercepts for di- and tetranucleotide microsatellites and therefore $\mu = 0$ for small alleles. Although μ is likely minimal for small allele sizes at most microsatellite loci, it is almost certainly non-zero. Therefore, we use the logistic model in the remainder of this study because it allows realistic, non-zero mutation rates for the smallest allele sizes and can recapitulate mutation curves derived from the linear model for larger allele sizes (Appendix, Figure A-2-1). We note, however, that any previous mutational model translated into a stochastic matrix may be used in the algorithm detailed below.

Under the SMM, transition probabilities for mutation from size g to size $h = l \pm 1$ are:

$$P_{gh} = \begin{cases} \mu/2 & \text{if } g \neq h \\ 1-\mu & \text{if } g = h, \quad g \geq 2, h \geq 2, \end{cases}$$

where μ is determined using equation (1). To model departures from the SMM, we specified two additional parameters. First, we used parameter c to control contraction bias – the empirically observed tendency for longer alleles to contract more frequently than expand (Amos et al., 1996; Xu et al., 2000). Let $Z(c, g) = P(\text{contraction}) = 1 - 1/(2cg^2 + 2)$, $0.5 \leq Z < 1.0$, where g is current allele size and $0 \leq c < \infty$ (though for most loci, reasonable values of c will not exceed 0.01). Z has a horizontal asymptote at 1. When $Z = 0.5$ ($c = 0$), there is no contraction bias; when Z is near one, most mutations reduce allele size. Second, we used parameter m to model multi-step mutation. Specifically, step size $k \sim \text{Geometric}(m)$, where m is the probability of single step mutation. When $c = 0$ and $m = 1$, mutation reduces to the standard SMM.

Finally, a stochastic matrix M comprising transition probabilities $\{P_{gh}\}$ from size g to h is computed:

$$P_{gh} = \begin{cases} \mu Z \times P(k = |g-h|) = \mu Z \times m(1-m)^{|g-h|-1} & \text{if } g < h \\ \mu(1-Z) \times P(k = |g-h|) = \mu(1-Z) \times m(1-m)^{|g-h|-1} & \text{if } g > h \\ 1-\mu & \text{if } g = h \end{cases}, \tag{2}$$

where μ is computed using equation (1) and $\sum_{h=2}^{\infty} P_{gh} = 1$, $g \geq 2$.

Rapid forward simulation of natural selection, mutation, and drift at a microsatellite using a recursion equation

Edwards (2000) corrected Wright's equation for the change in allele frequencies at a multiallelic locus in response to natural selection (Wright, 1937). This difference equation specifies the change in allele frequencies after one generation of natural selection:

$$\Delta\vec{p} = \Delta \begin{bmatrix} p_1 \\ p_2 \\ \dots \\ p_n \end{bmatrix} = \frac{1}{2\bar{w}} \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & \dots & -p_1p_n \\ -p_2p_1 & p_2(1-p_2) & \dots & -p_2p_n \\ \dots & \dots & \ddots & \dots \\ -p_np_1 & -p_np_2 & \dots & -p_n(1-p_n) \end{bmatrix} \begin{bmatrix} \frac{\partial\bar{w}}{\partial p_1} \\ \frac{\partial\bar{w}}{\partial p_2} \\ \dots \\ \frac{\partial\bar{w}}{\partial p_n} \end{bmatrix}, \tag{3}$$

where p_i is the frequency of allele a_i , $\bar{w}$ is mean fitness, and the partial derivative $\frac{\partial\bar{w}}{\partial p_i}$ is equal to twice the marginal fitness of allele a_i, $w^*(a_i)$.

We express the vector of allele frequencies after one generation of selection and mutation as a recursion equation:

$$\vec{p}_{t+1} = \mathbf{M}^T(\vec{p}_t + \frac{\mathbf{C}}{2\bar{w}}\nabla\bar{w}), \tag{4}$$

where $\mathbf{M}^T$ is the transpose of the mutation matrix (eqn. 2), $\vec{p}_t$ is the vector of current allele frequencies, $\mathbf{C}$ is the covariance matrix on the RHS of equation (3), and $\nabla\bar{w}$ is the gradient vector of partial derivatives on the RHS of equation (3). In the following algorithm, we use repeated application of equation (4) with multinomial sampling to simulate evolution of microsatellite allele frequencies subject to mutation, selection, and drift:

A0 Set $t = 0$ and $\vec{p}_0$ to the starting vector of allele frequencies.

A1 For each allele a_i, calculate marginal fitness $w^*(a_i)$ and $\frac{\partial\bar{w}}{\partial p_i} = 2 \times w^*(a_i)$.

A2 Calculate $\bar{w}$ and **C**.

A3 (Selection and mutation) Use equation (4) to find $\vec{p}_{t+1}$.

A4 (Reproduction and drift) Use multinomial sampling to draw a sample of size $2N_e$ based on probabilities $\vec{p}_{t+1}$, where N_e is effective diploid population size.

A5 Use the sample from [A4] to recalculate $\vec{p}_{t+1}$.

A6 Repeat steps [A1]-[A5] for the number of generations desired.

If steps [A4] and [A5] are skipped, thereby disregarding drift, steps [A1]-[A3] may be repeated until $|\vec{p}_{(t+1)} - \vec{p}_{(t)}| < \epsilon$, where ϵ is an appropriately small threshold (we used $\epsilon = 1/2N_e$). Then, current $\vec{p}_{(t)}$ provides an approximation of the allele frequencies at mutation-selection balance.

To assess accuracy, we compared the outcome of simulations using algorithm **A** to the outcome of forward, individual-based simulations. In forward simulations, all $2N_e$ copies of the allele were followed; each generation consisted of selection on diploid individuals, mutation of the surviving alleles, and reproduction by random sampling of surviving alleles until $2N_e$ copies were obtained. For the comparison of recursion and forward simulations, we used a representative set of parameter values: dominant model, $\delta = 0.05$, $g_l = -0.001$, $g_u = 0$, $\phi = 3.5$, $\psi = 1.5$, $\gamma = 0.15$, $m = 1$, $c = 0$. We performed the comparison for two distinct populations sizes: $N_e = 500$ or 10000.

Results

Picturing mutation-selection-drift equilibrium at a microsatellite

Forward simulations following algorithm **A** generated samples highly similar to those produced using much slower individual-based simulations (Appendix, Figure A-2-2). The contour plots in Figure 2-2,A-C each summarize the frequency distribution of a single allele over time and across 1000 replicate simulations using algorithm **A**. Equilibrium between mutation, selection, and drift eventually becomes apparent across replicates. The frequency of the key allele at mutation-selection balance (obtained by a single simulation in the absence of drift) was 0.9864. For a diploid population size of $N_e = 10000$ the key allele slowly approaches mutation-selection equilibrium in all 1000 replicates (Figure 2-2A). The effect of drift is minimal, but does cause key allele frequency to oscillate about its equilibrium frequency at mutation-selection balance. When $N_e = 500$ (Figure 2-2B), however, the effect of drift dominates. In a large fraction of simulations (31%), frequency of the key allele at 4500 generations is < 0.2. Figure 2-2C shows the frequency distribution of the next-most-fit allele (size 7) across the same 1000 replicates shown in Figure 2-2A. Comparing panels A and C of Figure 2-2, we can intuit the chronology of selective effects resulting from the topology of the multiallelic fitness surface (Figure 2-2D). Initially, the frequencies of both alleles increase because the large fitness penalty imposed on alleles of size > 8 by threshold effect $\delta = 0.05$ rapidly eliminates these alleles from the population. After ~ 50 generations, however, only alleles of size $<= 8$ remain and the gradient parameter $g_l = -0.001$ begins to slowly eliminate alleles of size ≤ 7.

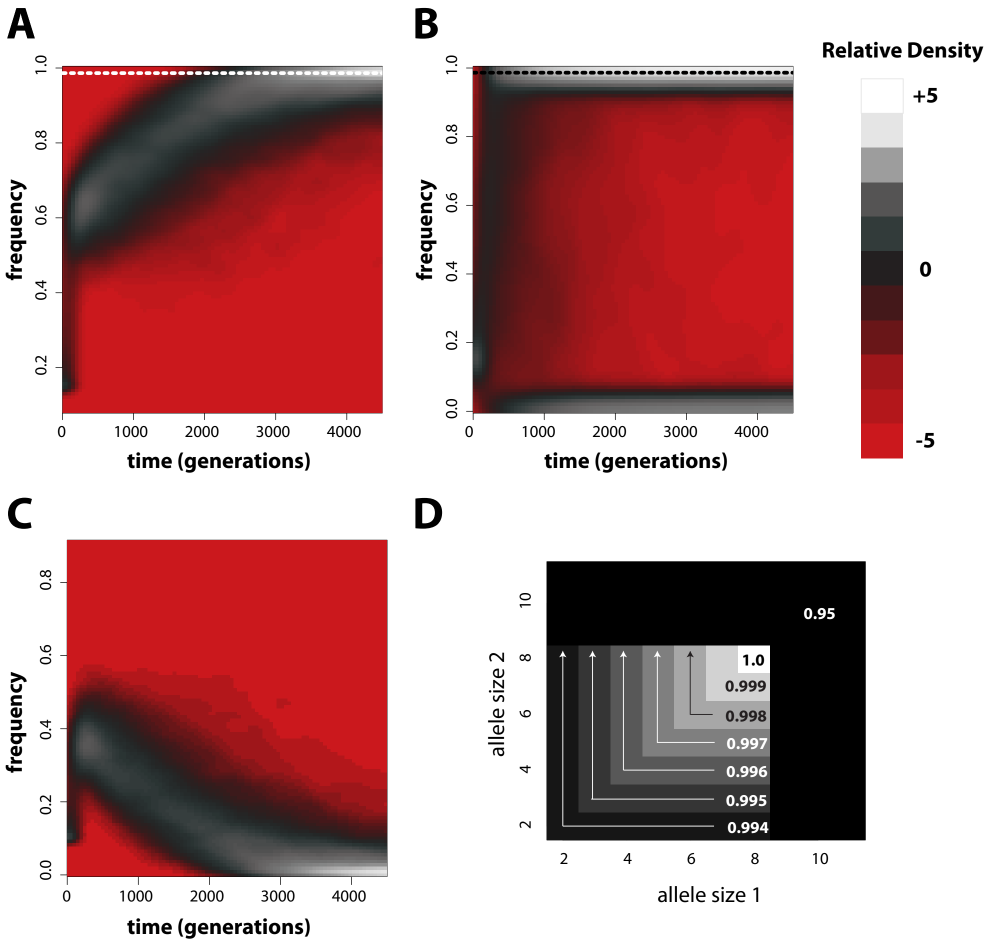

Figure 2-2: Mutation-selection-drift equilibrium for a microsatellite under selection. (A) The joint distribution of key allele (size = 8) frequency versus time for 1000 replicates at a selected microsatellite locus. In this case, the key allele is also the most fit and its frequency at mutation-selection equilibrium is 0.9684 (dashed line). The simulated selective regime was: dominant model, $x = 8$, $\delta = 0.05$, $g_l = -0.001$, $g_u = 0$. Simulated mutational parameters were: $\phi = 3.5$, $\psi = 1.5$, $\gamma = 0.15$, $m = 1.$, and $c = 0$. Diploid population size $N_e = 10000$. (B) The same as (A) for 1000 simulations where $N_e = 500$. (C) Derived from the same simulations as (A), the joint distribution of the frequency of allele size 7 versus time is shown. This allele is the next most-fit allele according to the modeled selective regime. (D) The fitness surface used in the simulations underlying panels A-C.

The evolution of Friedreich's ataxia and its causative microsatellite

To demonstrate the utility of the fitness models described here, we applied the recessive model to inference of parameters concerning the origin and selective regime of the human disease Friedreich's ataxia (FRDA). FRDA is caused by the hyperexpansion of a GAA repeat in the first intron of the autosomal gene frataxin (FXN; Campuzano et al., 1996) and is the most common inherited ataxia among individuals of Western European ancestry (Pandolfo, 2008). Four size-based classes of GAA allele are generally identified: short normal (SN) with allele size < 12, long normal (LN) with allele size between 12 and 33, premutation (P) with allele size between 34 and 60, and expanded (E) with allele size > 60. Affected individuals are homozygous for an E allele; age of onset and severity of the disease increase with the size of the smaller allele in affected genotypes (Durr et al., 1996). Patterns of linkage disequilibrium with nearby SNPs support the hypothesis that a single 18-repeat allele (and the LN class with it) originated from a rare doubling mutation of a 9-repeat allele (Cossee et al., 1997; Monticelli et al., 2004). Subsequently, LN alleles likely proliferated via ordinary mutation (Montermini et al., 1997), eventually generating larger P alleles that are vulnerable to hyperexpansion (size ≥ 34). E-class alleles mutate roughly 85% of the time and while the expansion/contraction ratio is even in females, nearly all mutations of E alleles in males are contractions (Pianese et al., 1997). The geographic distribution of non-SN alleles and analyses of linkage disequilibrium suggest that a unique SN-to-LN mutation took place in Africa (Colombo and Carobene, 2000). Based on measures of LD in modern Europeans, one study dated the origin of the first LN allele at 682 +/- 203 generations ago (Colombo and Carobene, 2000). However, the authors acknowledge this may be an underestimate. Their method assumed equilibrium population dynamics, but migration from Africa to Europe incurred a population bottleneck that would have slowed decay of LD, thereby skewing the estimate of allele age towards more recent times. In our simulation-based inference, we allowed both African and European origins of the LN class to be simulated (Figure 2-3).

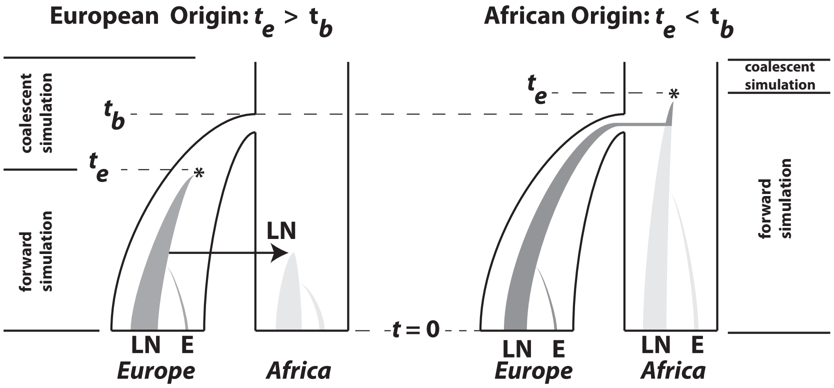

Figure 2-3: The demographic model for FRDA inference. Outer trees indicate population size. Inner shaded trees represent the frequencies of LN and E class alleles. Parameters t_b (bottleneck time) and t_e (time of LN class origin) were drawn from uniform prior distributions before the start of each simulation. The relation between these parameter values distinguished between two historical possibilities. When $t_e > t_b$ (left), the bottleneck occurred before the emergence of the first LN allele. In this case, the LN and E alleles observed in Northern Africa on the same haplotypic background as European LN and E alleles can only be explained by back-migration to Africa (pointed arrow). When $t_e < t_b$ (right), LN emergence takes place in Africa and is subsequently carried to Europe by members of a founding population. Note that only simulations where LN alleles survived to modern day ($t = 0$) were retained and that the post-divergence African population was not simulated. Coalescent simulation was used to simulate starting distributions of genetic variation; forward simulations as detailed here were used to progress from time t_e to $t = 0$.

Posterior point estimates and 95% credible intervals for all parameters of interest are found in Table 2-1, while graphical comparisons of prior and posterior distributions for each estimate are shown in Figure A-2-3 (Appendix). Our median estimate of the age of the anomalous SN-to-LN doubling event is 1494 generations ago with a credible interval of 840-2593 generations ago. Figure 2-4 shows the estimated fitness surface of the causative `GAA` repeat assuming median values of δ and g_u from posterior distributions. After normalizing the fitness surface by assigning a fitness of 1.0 to all genotypes with at least one allele less than 34 in size, the relative fitness of the most deleterious genotype, (1500/1500) is 0.105. All genotypes in which both alleles are of size ≥ 34 have relative fitness $<= 0.984$.

Despite very low fitness of affected genotypes, the low frequency of E alleles in the observed Western European population and the recessivity of the disease suggest the selective toll of FRDA is minimal. This expectation was confirmed by additional simulation; across 1000 simulations using median parameter estimates, maximum realized genetic load was only $\sim$ 1.2e-04 (Appendix, Figure A-2-4).

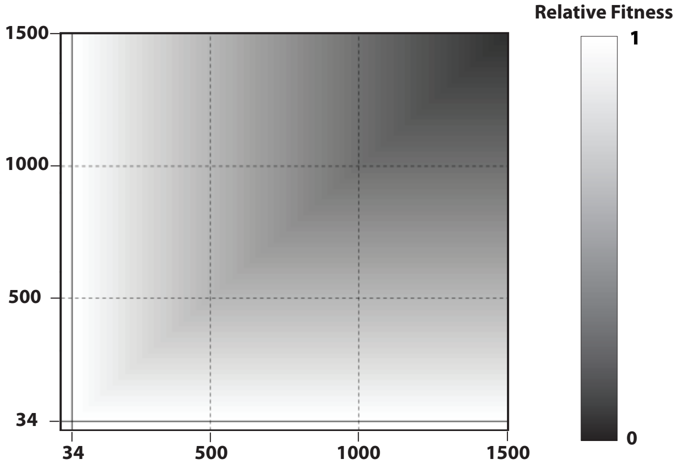

Figure 2-4: Estimate of the fitness surface for the GAA repeat that causes Friedreich's ataxia. This estimate is based on median selective parameter values from their posterior distributions. The solid black lines are drawn at allele size 34. We assumed that all genotypes with at least one allele of size < 34 had a relative fitness of 1. The least fit genotype on the graph – $1500/1500$ – has an estimated fitness of only 0.104.

Approximate posterior densities on the mutational parameters ϕ, ψ, and γ were relatively narrow (Table 2-1; Figure A-2-3). Using the median estimates of these parameters to calculate allele-specific rates of μ_{STR}, we estimate that alleles < size 12 mutate at rates $< 1e - 03$. However, alleles of size > 12 were inferred to be extremely mutable, peaking at $\mu_{\text{STR}} \simeq 0.1$ for alleles of size > 24. These results suggest that modeling allelic-specific mutation rate is an important part of characterizing selection targeting microsatellites.

<table>
<tr><th></th><th></th><th></th><th colspan="2">| |selection| |</th><th colspan="3">| | |mutation| | |</th><th>population growth</th></tr>
<tr><th></th><th>t_e</th><th>t_b</th><th>g_u</th><th>δ</th><th>ϕ</th><th>ψ</th><th>γ</th><th>α</th></tr>
<tr><td><b><i>prior</i></b></td><td>(-4000, -475)</td><td>(-4000, -1000)</td><td>(-0.0015, 0.)</td><td>(0., 0.04)</td><td>(4.5, 7.)</td><td>(1.5, 4.)</td><td>(0.05, 0.4)</td><td>(-0.003, 0.)</td></tr>
<tr><td><b><i>posterior</i></b></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>median</td><td>-1494</td><td>-2645</td><td>-0.0006</td><td>0.0157</td><td>6.27</td><td>2.74</td><td>0.17</td><td>-0.0016</td></tr>
<tr><td>2.5 percentile</td><td>-2593</td><td>-3705</td><td>-0.0012</td><td>0.</td><td>5.66</td><td>1.98</td><td>0.1</td><td>-0.003</td></tr>
<tr><td>97.5 percentile</td><td>-840</td><td>-1656</td><td>0.</td><td>0.028</td><td>6.93</td><td>3.7</td><td>0.31</td><td>-0.0001</td></tr>
</table>

Table 2-1: Prior distributions and posterior estimates for parameters relevant to the microsatellite causative of Friedreich's ataxia. All prior distributions were uniform on the specified interval. In addition, the listed priors are narrower than those used for the first 10,000 simulations.

Population-level characteristics of microsatellite selection

We quantified distance between the starting distribution of allele frequencies and those at mutation-selection balance as Δ_{msat} (see Material and Methods). For all selective regimes tested (Table 2-2), regression of d on Δ_{msat} and cost of selection C on Δ_{msat} were significant ($P < 1e - 05$). Values of r^2 associated with regression analyses (Table 2-2) suggest that Δ_{msat} is an important determinant of both the cost and duration of selection in a population. Interestingly, the influence of Δ_{msat} on the cost of selection is largely independent of selective strength. Comparing additive regimes A1 and A2, the rate at which C increases in response to increases in Δ_{msat} is identical for both scenarios, despite five-fold greater values of g_l and g_u in regime A2 ($P = 0.915$; ANCOVA: H_0: slopes identical; Figure 2-5A). Although the intercepts of the best-fit lines for regimes A1 and A2 are significantly different (P=0.021), it is visibly evident that the average increase in C associated with regime A1 is very minimal (Figure 5A). These results agree with those for diallelic loci, where, except for very strong selection, increases in selective strength do not affect C (Haldane, 1957). Different models of microsatellite selection can lead to selective events with very different characteristics (Figure 2-5B). For example, dominant and recessive selective regimes produced selective events of greater duration than those of additive and multiplicative selection regimes. In addition, populations evolving under the multiplicative regime M2 obtained mutation-selection equilibrium in roughly half the time of populations evolving according to selective regime A2, despite identical parameter values. Finally, for all selective regimes simulated, greater than 70% of replicates fell to the left of the hard sweep line in Figure 5B. This region of the graph corresponds to soft selective sweeps on SNPs, where the starting frequency of the beneficial variant is $> 1/2N_e$.

						r^2	
regime	model	x	g_l	g_u	δ	C on Δ_{msat}	d on Δ_{msat}
A1	additive	11	-0.01	-0.01	0.	0.74	0.56
A2	additive	11	-0.05	-0.05	0.	0.59	0.29
M1	multiplicative	11	-0.05	-0.05	0.	0.64	0.38
D1	dominant	11	-0.01	-0.01	0.	0.36	0.43
R1	recessive	11	0.	-0.01	-0.025	0.74	0.77

Table 2-2 Simulated selective regimes and coefficients of determination for regression of C and d on Δ_{msat}.

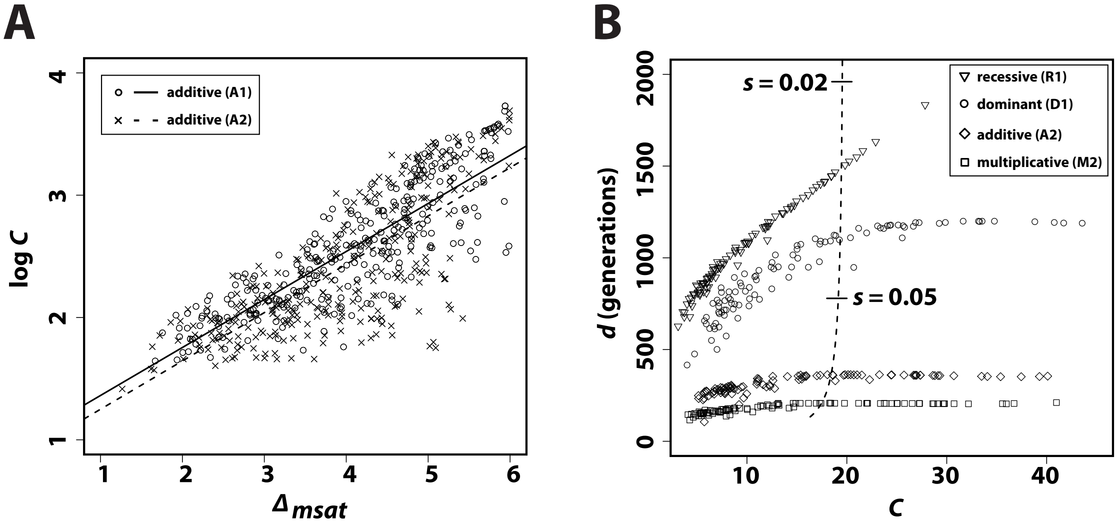

Figure 2-5: Cost and duration of microsatellite selection. (A) Regression of log C on Δ_{msat} for additive regimes A1 and A2 (Table 2-2). The results of 250 deterministic simulations are shown. The only difference between replicates of the same regime was the starting distribution of allele frequencies, which was generated using neutral coalescent simulation. Δ_{msat} quantifies the difference between starting allele frequencies and those at mutation-selection balance. Best fit lines for both regimes are drawn. (B) Duration of selection versus cost of selection for regimes R1, D1, A2, and M2; 250 deterministic replicates each. The dashed line is drawn from deterministic simulations of a hard, SNP-based selective sweep (dominance coefficient $h = 0.5$). The line is interpolated but based on thousands of simulations, each with a different value of s. Two values of s are indicated on the dashed line.

Discussion

The role of mutational complexity in genomic scans for selection

Standard genomic scans for selection assume that natural selection is the only locus-specific force active in the genome. The effects and/or rates of mutation, recombination, and demography are assumed to be homogeneous across the genome. This paradigm is attractive because it implies that anomalous patterns of genetic diversity must be attributable to the action of natural selection. However, while genomic scans have identified a handful of loci clearly subject to natural selection in humans, meta-analyses of genomic scans in humans do not yield ready consensus (Akey, 2009). One reason for this is likely the characterization of genomes as a monolithic sequence. Methods that ignore interlocus heterogeneity caused by factors other than natural selection bear reduced statistical power to detect selection and could suffer elevated false positive rates. In particular, studies that differ in terms of sample, sample size, markers, etc ... will often yield distinct or conflicting results.

Here, we focused on a common source of heterogeneity that is seldom considered: frequent, recurrent mutation. We used microsatellites as a model form of variation for this purpose. Implementing our models of direct selection on microsatellites revealed the danger in assuming that high-density SNP data are capable of detecting selection at non-SNP variants. In Figure 2-5, the majority of simulated selective events targeting microsatellites fall to the left of the line denoting a starting frequency of $1/2N_e$ for SNP selection. In other words, selective events targeting microsatellites will frequently resemble soft sweeps on SNPs, which are nearly impossible to detect using statistics based on the site frequency spectrum (Pennings and Hermisson, 2006b). Indeed, simulations of linked diversity in the case of direct selection on a microsatellite corroborate the analogy to soft sweeps; soft sweeps and direct microsatellite selection generate minimal selective footprints in their wake, at least as measured by summaries of the site frequency spectrum (Figure 2-6).

In general, this is most likely due to the fact that recurrent mutation causes an advantageous microsatellite allele to become associated with a variety of haplotypic backgrounds.

Pennings and Hermisson (2006a,b) developed a model of positive selection that did include recurrent mutation (following an infinite alleles model) and found that resultant soft sweeps were detected with high statistical power using measures of linkage disequilibrium (LD). The authors attributed this power to the fact that each individual mutation to the beneficial allele was likely to bring with it a distinct ancestral haplotype whose genetic associations (LD) were unlikely to decay during the selective period. Though it warrants further investigation, there are several reasons to suspect that the encouraging results of Pennings and Hermisson (2006b) may not hold for the detection of microsatellite selection: (1) favored microsatellite alleles will frequently be drawn from standing variation, suggesting that selected alleles will already lie on genetic backgrounds of partially decayed LD; (2) the population mutation rate of microsatellites, $\theta = 4N_e\mu$, will generally be much higher than the values considered by Pennings and Hermisson (2006a,b), leading to very frequent recurrent mutation; (3) back mutation, ignored by Pennings and Hermisson (2006a,b), will be common at microsatellites, and; (4) considerable variation in allelic fitness may often exist at non-neutral microsatellite loci, which can undermine the effectiveness of tests for selection based on LD (Pennings and Hermisson, 2006a).

In general, lessons learned from studies based on the infinite sites or infinite alleles models of mutation will not hold for microsatellites and other genetic variants created by complex mutation. Therefore, it seems prudent to develop models of selection and mutation tailored to the peculiarities of these variants. Otherwise, even strong instances of selection on many forms of genetic variation that are less commonly considered will be difficult or impossible to detect.

Detecting microsatellite selection

The complex nature of multiallelic selection makes detecting evidence of natural selection at microsatellites a challenging task. As discussed above, the standard genomic scan for selection will generally be a poor approach for detecting microsatellite targets of selection. Furthermore, the absence of genome-wide microsatellite data currently precludes full genomic scans for microsatellite selection (though see Gymrek et al. (2012)). Yet we believe testing candidate microsatellites for evidence of selection provides one way forward. In this sense, testing for microsatellite selection may actually prove an easier task, since microsatellite loci are well defined while genomic scans for positive selection proceed under the assumption that all nucleotides could be of adaptive consequence. Also, a locus-specific test of microsatellite neutrality should be helpful to empiricists, where the presumed neutrality of microsatellite markers is rarely tested.

One approach to testing candidate microsatellites for selection is to embrace their potential complexity and use simulation based inference procedures. We have demonstrated that a simple implementation of ABC inference using our models and simulation algorithm was sufficient to provide novel insights regarding evolution of the microsatellite underlying Friedreich's ataxia (see below). However, direct selection on a microsatellite and selection on a tightly linked SNP both cause reductions in microsatellite variation (Slatkin, 1995a). Thus, full implementation of our models in the inference of microsatellite selection requires a means to distinguish between direct and linked selection. One possibility is to examine levels of linked diversity in sequence flanking the subject microsatellite. Since most instances of microsatellite selection appear most similar to selection on standing SNP variation (Figure 2-5), direct microsatellite selection should most often reduce variance at the microsatellite while leaving a minimal selective footprint in linked sequence diversity (Figure 2-6).

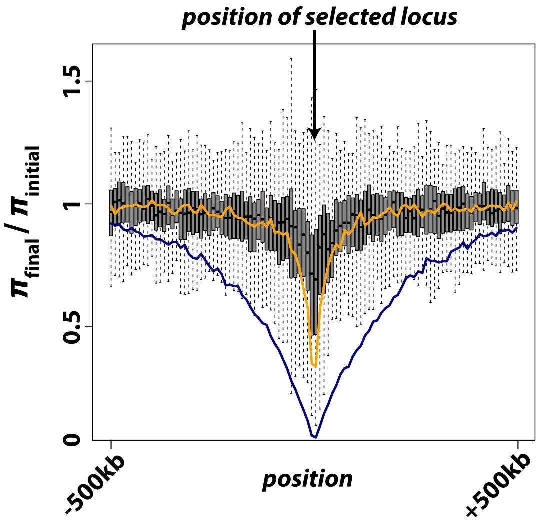

Figure 2-6: Results from 250 independent simulations each of additive selection on a microsatellite, a soft sweep (p_0 on the interval [0.1, 0.2]), or a hard sweep ($p_0 = 1/2N_e$), where p_0 is the starting frequency of the beneficial SNP variant. The y-axis plots $\pi_{final}/\pi_{initial}$, where final nucleotide diversity (π_{final}) was calculated from a sample of $n = 100$ chromosomes either at the time of fixation of the beneficial variant (SNP selection) or when mutation-selection-drift equilibrium was achieved (microsatellite selection). In all selection scenarios, the target of selection was located at the center of a 1Mb sequence. Boxplots summarize the results from simulations of microsatellite selection in non-overlapping 10kb windows (rectangles are interquartile distances). Colored lines plot the mean value of $\pi_{final}/\pi_{initial}$ across simulations for soft sweep (orange) and hard sweep (blue) simulations.

The cost and duration of microsatellite selection are dependent on several factors

A recent study of experimental evolution unequivocally demonstrated that rapid adaptive responses are possible when the selected target is a repetitive element with high mutation rate (Vinces et al., 2009). This result supports hypotheses that microsatellites provide reservoirs of potentially adaptive alleles and that frequent recurrent mutation provides the opportunity for rapid adaptive response to environmental change (Kashi et al., 1997;

King et al., 1997; Trifonov, 2004; Kashi and King, 2006; King and Kashi, 2009; Gemayel et al., 2010). Yet it is premature to claim that repetitive elements such as microsatellites are truly ubiquitous agents of efficient adaptive change; their capacity as drivers of adaptive change appears contingent on several factors. First, the efficiency of adaptive response is dependent upon the selective regime imposed by ecological change. We found that 99% of the replicates of microsatellite selection under regime A1 take longer to reach equilibrium than those of regime A2. Yet there is not a significant difference between initial variance in allele size in the A1 and A2 replicates ($P = 0.643$). In other words, the difference in efficiency of adaptive responses demonstrated by A1 and A2 replicates is not due to insufficient accumulation of standing variation at the selected locus but the relatively flatter fitness surface under scenario A1. As another example, consider the substantial difference in the efficiency of selection between regimes R1 and M1. While the duration of selection is < 200 generations for all replicates of M1 selective events, it can take > 1500 generations to obtain mutation-selection balance under the R1 regime (Figure 2-5B). Second, efficiency of the selective response of a microsatellite is dependent on the starting distribution of allele frequencies. For both A1 and A2 scenarios, replicates with the highest selective costs (Figure 2-5A) and longest durations of selection were also among the set of replicates with the highest values of Δ_{msat} (Figure 2-5A). In many of these cases, the most fit allele was not present in the population at the start of selection. Thus, the accumulation of standing variation at a microsatellite prior to environmental change will only lead to a more efficient selective response if the new selective regime selects for alleles in the vicinity of the current allele distribution. Some hypotheses that advocate the efficacy of selection on repetitive elements do make this very assumption, such as the “tuning knob" model of Trifonov (2004). Finally, as shown in Figure 2-2B, small population size can lead to an appreciable probability that a population will segregate the most beneficial allele at near-zero frequency despite high rates of mutation. This suggests that potential efficiencies of adaptation via microsatellite will be difficult to obtain in small, imperiled populations.

Inferring the origin and selective regime of Friedreich's ataxia

Our estimated date for the anomalous SN-to-LN mutation is more than double that of a previous estimate (Colombo and Carobene, 2000), which was calculated as a simple function of LD and recombination fraction at several linked loci (Risch et al., 1995). As the authors discussed, however, their estimate may be biased toward more recent estimates. Indeed, we believe a substantially more ancient estimate of LN emergence is supported by a variety of evidence. First, near-perfect linkage disequilibrium with nearby variants (Cossee et al., 1997; Monticelli et al., 2004) and a noticeable gap between observed frequency distributions of SN and LN alleles (Monticelli et al., 2004) support the hypotheses that: (1) the current pool of LN, P, and E class alleles is derived from a single, anomalous mutation of an SN allele and (2) broadening of the SN allele range by standard mutation has not contributed to the current pool of LN alleles. We incorporated these hypotheses in our inference procedure by rejecting any simulation in which all descendants of the single, initial LN allele were lost. In this case, we found it nearly impossible to generate LN and E frequencies comparable to empirical frequencies in less than 1000 generations, even when mutation rate of LN alleles was very high. Only two of the 500 best simulated samples used to compute posterior distributions had $t_e > -1000$. Second, E class alleles are limited to Northern Africa, the Middle East, and Western Europe. Coupled with the hypothesis of a single LN origin, this fact recommends the parsimonious hypothesis that LN emergence took place somewhere in Northern Africa and subsequently spread with immigrants to the Near East and Europe. If a Northern African origin of the LN class is true, it necessitates that the mutation occurred > 2000 generations ago, as the Eurasian expansion likely took place on the order of 40kya (Liu et al., 2006a). On the other hand, we interestingly found that 93% of the best fitting simulations had $t_e > t_b$ – i.e., LN emergence took place in the bottlenecked European population (Fig S3). If true, this historical hypothesis necessitates the back-migration of LN class alleles to Africa (Figure 2-3).

To our knowledge, the fitness surface presented in Figure 2-4 is the first estimate of its kind for a microsatellite that causes a human trinucleotide disorder. The topography of this surface agrees with clinical observations. First, decreasing fitness with increasing size of the smallest E allele in a genotype (i.e., negative g_u) agrees with the observation that decreased age of onset and increased severity of symptoms are correlated with the size of the smaller allele in affected individuals (Durr et al., 1996). Second, a positive value of δ agrees with the fact that all individuals with two E alleles experience some impairment. Relative fitnesses of genotypes in which both alleles are > 1100 repeats are very low (< 0.35). However, the occurrence of these genotypes in nature must be very rare. Using standard formulas for expected homozygosity and conditional probability, the probability of a 1100+/1100+ genotype is only

$$\begin{aligned} E\{\text{freq. 1100+/1100+ genotype}\} &= P(\text{size} > 1100)^2 = \{P(\text{size} > 1100 \mid \text{E}) \;\; P(\text{E})\}^2 \\ &= (0.095 \times 0.01)^2 = 9e-07, \end{aligned}$$

where $P(\text{E})$ is the marginal probability of an E class allele. Thus, we expect only one in 1.1 million people of European ancestry to carry these highly deleterious genotypes. Although natural selection acts upon variation at the GAA repeat in FXN, it has had very minor impact on the evolution of the microsatellite relative to mutation and drift (Appendix, Figure A-2-4).

We inferred remarkable heterogeneity in mutation rates for the FXN microsatellite.

While SN alleles are predicted to mutate within the range of mutation rates generally cited for microsatellites ($10^{-06} - 10^{-03}$), the median estimate of μ for larger LN alleles was on the order of 10^{-1}. The absence of empirical examples of LN alleles on more than one haplotypic background as well as the discontinuity in the observed frequency distribution between SN and LN class ranges support the idea that SN alleles mutate quite slowly. If SN alleles mutated at very high rates, they would likely invade LN allele space thereby linking LN alleles to a diversity of haplotypic backgrounds. Also, our simulations indicate that a very high mutation rate of LN alleles is required for the rapid increase in frequency of LN alleles from $1/2N_e$ to 0.1675 (even in 1000+ generations).

Although the qualitative patterns implied by our parameter estimates seem reasonable, the absolute quantitative estimates presented here should be treated with caution. For example, these estimates possess little value if the seemingly well-supported assumption that there was a single LN origin does not hold. Furthermore, our model of the European bottleneck (Figure 2-3) overlooks the fact that the colonization of Europe and other regions likely included serial serial bottlenecks (Liu et al., 2006a; DeGiorgio et al., 2009). Our main motivation for including this example was to point out the potential value of our models and simulation algorithm to population genetic inference. Indeed, we believe analysis of the FXN locus that used African and Eurasian samples as well as more detailed summary statistics could provide a high-resolution portrait of the evolution of Friedreich's ataxia and its causative locus.

Extending models of the fitness surface to other multiallelic variants

Our models could be applied to other multiallelic variants. Copy number variants (CNVs) are polymorphisms in the number of repeats of 1kb-1Mb DNA segments. Recently CNVs have been implicated in disease and other phenotypic variation (Cooper et al., 2007; Nair et al., 2008), most likely due to differences in dosage of genes contained within the repeated segments (Stankiewicz and Lupski, 2010). The mutational mechanism leading to the generation and variation of CNVs is far from settled (Hastings et al., 2009). Nevertheless, CNVs resemble microsatellites in several ways. They are repetitive elements that mutate in a complicated manner and whose allele size may affect fitness. CNV analogues to the models reported here could similarly be used in inference regarding selection on these variants, which are of increasing interest to the human genetics community. While selective models could be ported directly, construction of a realistic mutational model would likely be difficult. However, a variety of mutational models could be combined with the selective models reported here to enable simulation-based investigation of the population-level consequences of different mutational mechanisms.

Methods

Modeling Friedreich's ataxia and inferring parameters of interest

In modeling FRDA evolution, we assumed the following: (1) recessive model of natural selection; (2) key allele $x = 34$; (3) effective population size of the affected, modern day Western European population is $N_e = 10000$; (4) an historical demographic model in which an African population of $N_e = 10,000$ gives rise to a bottlenecked founding population that undergoes exponential population growth at rate α (Figure 2-3; parameter t_b specifies the time of the bottleneck); (5) no selection against allele sizes < 34; (6) $g_u \leq 0$ – i.e., the fitness of alleles of size greater than 34 (key allele size) could only decline with increasing allele size; (7) single origin of an LN allele at size 18; (8) mutation of SN and LN alleles follows the mutation model outlined above; (9) gender-specific differences in hyperexpansion mutations follow a 50/50 mixture model of male and female mutational distributions (Pianese et al., 1997); (10) P and E alleles hyperexpand with probability 0.85; and (11) with probability 0.15, E alleles undergo no change and P alleles are subject to normal mutation probabilities.

We used approximate Bayesian computation (ABC; Beaumont et al., 2002) to estimate parameter values of interest. Frequencies of SN and LN alleles were estimated from 400 chromosomes sampled from Europeans in two studies (Montermini et al., 1997; Monticelli et al., 2004), while E frequencies were estimated from 332 chromosomes sampled from Europeans in two separate studies (Durr et al., 1996; Pianese et al., 1997). Following the ABC paradigm, we estimated parameter values by comparing empirical frequencies to those generated by simulation.

For each simulation, we drew random values of parameter t_e – the emergence time of the first LN allele – as well as seven other parameters: t_b, α, g_u, δ, ϕ, ψ, and γ. Constant values of $c = 0$ and $m = 0.95$ were used. All prior distributions were uniform (Table 2-1). Note that the prior for t_e includes more recent time points that that of t_b. This allowed the emergence of the first LN allele to occur in the founding European population rather than the ancestral African population. Although haplotype data indicate this is a less parsimonious hypothesis, we allow simulation of this hypothesis because it is possible that the first LN allele emerged in the European population and back-migrated to Northern Africa (Figure 2-3). To increase the efficacy of simulation effort, we refined initial prior distributions based on the results of 10,000 pilot simulations. These narrower priors are the ones listed in Table 2-1. We ran 100,000 total simulations with these priors. Each simulation began with a coalescent phase (Figure 2-3). At time t_e, a single SN allele was converted to a size 18 LN allele. Then, forward simulation following algorithm **A** proceeded until $t = 0$ (modern day); note, however, that N_e changed through time and that the post-divergence African population pictured in Figure 2-3 was not directly simulated. At $t = 0$, a sample of $n = 400$ chromosomes was taken from the population. 100,000 total simulations were run. We restarted a replicate whenever all descendants of the single size 18 allele were lost. Thus all results are conditioned on survival of this lineage as supported by linkage analysis (Cossee et al., 1997; Monticelli et al., 2004). Empirical and simulated samples were summarized using six summary statistics: total frequencies of LN and E alleles and the proportion of E-class alleles found on the size intervals (60,500], (500,700], (700,900], and ≥ 1100. Observed values of these summary statistics were 0.1675, 0.01, 0.146, 0.17, 0.293, and 0.095, respectively. We retained all simulated samples and used weighted local linear regression (Beaumont et al., 2002) with a tolerance of 0.005 (0.5% of simulations) as implemented in the R package *abc* (Csillery et al., 2012) to estimate approximate posterior distributions for the parameters of interest. Parameters were log-transformed for regression and back-transformed post-regression.

Characterizing the effects of microsatellite selection at the population level

To compare population-level consequences of microsatellite selection, we simulated representative selective regimes for each of the four models described above (Table 2-2; 250 replicates each). Each replicate of a given selective regime began with a random starting distribution of allele frequencies, generated using neutral coalescent simulation in MARKSIM (Haasl and Payseur, 2011). Simulations were deterministic and mutation parameters were constant across all simulated regimes: $\phi = 5$, $\psi = 2$, $\gamma = 0.3$, $c = 0$, $m = 0.95$. For each replicate, we calculated: (1) the duration of selection, d, which was was the time in generations from the onset of selection until mutation-selection equilibrium was achieved (defined as the first generation when the sum of allele frequencies at the selected locus was less than $1/2N_e$ [2]); (2) the cost of selection, $C = \sum_{t=1}^{d} 1 - \bar{w}$ (Haldane, 1957), and; (3) the distance between starting allele frequencies and those at mutation-selection equilibrium, Δ_{msat}. The last metric was calculated as:

$$\Delta_{msat} = \sum_{x \in \mathcal{S}} \sum_{y \in \mathcal{E}} |x - y| p_x p_y, \tag{5}$$

where $\mathcal{S}$ is the set of starting allele sizes, $\mathcal{E}$ is the set of equilibrium allele sizes, and $p_\bullet$ is allele frequency. Thus, Δ_{msat} weights the distance between each starting and equilibrium allele by the product of their frequencies, $p_x p_y$, which can be thought of as the probability that a starting allele of size x will be replaced by an allele of size y by the time of equilibrium. Finally, for comparison, we calculated d and C for hard selective sweeps, where the beneficial single nucleotide variant started at a frequency of 5e-05[3]. In all simulations of SNP selection, the dominance coefficient $h = 0.5$. We simulated values of the selection coefficient s ranging from 0.001 to 0.1 in increments of 0.001.

Comparing the selective footprints of selection targeting SNP variants versus microsatellites.

We ran 250 independent simulations each for three different selective scenarios: (1) additive selection on a microsatellite ($g_l = g_u = -0.05$, $\delta = 0.$; (2) a soft sweep (p_0 on the interval [0.1, 0.2]), and; (3) a hard sweep ($p_0 = 1/2N_e$), where p_0 is the starting frequency of the beneficial SNP variant. For each type of selection, the 250 simulations started with an array of independently generated SNP variation along a 1MB sequence using MS (Hudson, 2002) embedded in MARKSIM (Haasl and Payseur, 2011). We then added the beneficial SNP variant or microsatellite to the exact center of the 1MB sequence. Next,

[2] although these were deterministic simulations, this definition of equilibrium implicitly assumes $N_e = 10000$

[3] although these were deterministic simulations, this starting frequency implicitly assumes $N_e = 10000$

we used forward simulations, in which the order of events was selection, reproduction and recombination, and mutation. Simulations finished when fixation of the beneficial variant occurred (SNP-based selection) or the selected microsatellite reached mutation-selection-drift equilibrium.To simulate reproduction and recombination, two chromosomes from those remaining after selection were chosen at random to represent the "father" and two to represent the "mother". For each of these two pairs we then tested the pair for recombination (rate 1.25cM/Mb). If recombination was indicated, we then tested to see if a recombinant chromatid was inherited. If so, we chose the position of the breakpoint at random. From each parent, then, an offspring inherited a random recombinant or non-recombinant chromosome. Reproduction continued until the constant population size of $N_e = 10000$ was reached. During the mutation phase, new SNPs arose at random positions at a Poisson-distributed rate of 0.0125 (10^6 bases $\times \mu = 1.25e - 08$). Microsatellites mutated according to the logistic model described in this paper with $\phi = 5$, $\psi = 2$, $\gamma = 0.3$, $c = 0$, and $m = 0.95$. For both soft and hard sweeps, selection parameter $s = 0.05$ and dominance parameter $h = 0.5$.

CHAPTER 3: THE EFFECTS OF MICROSATELLITE SELECTION ON LINKED VARIATION

Summary

Theory that underlies standard genome-wide scans for selection assumes that mutation follows the infinite sites model (ISM). This assumption may limit the ability of genome-wide scans to identify selected targets that do not follow the ISM, such as microsatellites and other copy number variants. We focus on microsatellites, which bear high mutation rates and are therefore subject to frequent recurrent mutation. We compare the power of statistics that summarize the site frequency spectrum or distribution of haplotypes to detect selection on microsatellites with their power to detect selection on single nucleotide variants (SNVs) that follow the ISM. We find that microsatellites with high mutation rates are generally difficult to detect using statistics based on the site frequency spectrum. However, we also identify a notable exception to these pessimistic results. Selection on microsatellites generates a joint signal of selection comprising a low number of haplotypes and an intermediate number of segregating sites. In fact, when conditioned on the number of segregating sites, the number of haplotypes provides greater power to detect selection on highly mutable microsatellites than any other tested scenario of selection. Due to recurrent and back mutation, the favored microsatellite allele does not fix under most circumstances. We hypothesize that continued generation of deleterious alleles by mutation followed by their elimination may generate long-lasting reductions in linked sequence diversity. The magnitude of this reduction will be positively correlated with mutation rate at the selected microsatellite. Finally, we apply insights gained from simulation results to the analysis of an intronic microsatellite in the human gene *HSD11B2*, which multiple lines of evidence suggest is a good candidate for direct selection. We find that anomalous patterns of variation at the subject microsatellite are likely due to linked selection rather than direct selection on the microsatellite. In general, knowledge of the relative powers of different statistics to detect selection on different types of genetic variants – including our identification of a pattern in linked variation that is particularly sensitive to selection on rapidly mutating microsatellites – should aid the detection and characterization of selection on non-ISM variants in two important ways. First, this knowledge should enable the performance of informed, full genome scans for microsatellite selection that do not require initial sampling of microsatellite variation itself. Second, as the case of *HSD11B2* illustrates, patterns of linked variation can be used to corroborate or weaken the case for natural selection supported by other lines of evidence. Our results also suggest that, if computationally tractable, a joint analysis of microsatellite and linked variation may provide a powerful method to infer selection at candidate microsatellite loci.

Introduction

The genome-wide scan for selection is a powerful and promising method in the toolkit of the evolutionary biologist. Results from scans for selection can provide remarkable knowledge: the regions of the genome that have been among the most critical to the evolution of a population or species. For this reason and because whole genome sequencing is becoming increasingly inexpensive, the genome-wide scan for selection first envisioned forty years ago (Lewontin and Krakauer, 1973) has now become commonplace (Biswas and Akey, 2006; Akey, 2009; Oleksyk et al., 2010; Strasburg et al., 2012). Moreover, scans for selection have lived up to their promise by identifying interesting examples of selection in a variety of species, including parallel evolution in divergent freshwater populations of threespine stickleback (Hohenlohe et al., 2010), local positive selection for a derived allele in the pigmentation gene *SLC24A5* in Europeans (Lamason et al., 2005), and selection targeting ion transport and metal detoxification genes in populations of *Arabodopsis lyrata* growing in inhospitable serpentine soils (Turner et al., 2008, 2010). As access to genomic data for an increasingly broad swath of phylogenetic diversity accrues, it behooves the evolutionary genetics community to understand the patterns of genome-wide polymorphism in as complete a way as possible. In particular, are there targets of selection that are potentially overlooked by the scan for selection as currently practiced?

One particularly appealing feature of the genome-wide scan for selection is its ostensibly unbiased nature. Abstaining from a priori specification of candidate targets of selection, the genome-wide scan interrogates the majority of genomic regions without reference to their potential biological function. Absent alternative explanations such as demographic change, anomalous patterns of polymorphism may be attributed to natural selection. However, it is now clear that the models and statistics underlying genome-wide scans for selection may lead to biased result sets. For example, selection from standing variation often fails to significantly distort patterns of genetic variation as measured by the site frequency spectrum (SFS; Innan and Kim, 2005; Przeworski et al., 2005). Thus, standard genome-wide scans are biased towards identifying selective targets derived from new mutation. Similarly, selection on a polygenic trait may fail to significantly distort patterns of genetic variation linked to any one component gene (Pritchard et al., 2010). Therefore genome-wide scans may also be plagued by a bias towards the identification of genetic variants responsible for variation in Mendelian traits. Finally, Teshima et al. (2006) found that selective sweeps are more difficult to identify when the selected allele is recessive and concluded this will lead genome-wide scans to produce an unrepresentative set of potential selective targets.

These and other biases associated with scans for selection have received substantial attention (e.g., Hermisson and Pennings, 2005; Hancock et al., 2010b,a). Here we investigate a bias that is seldom considered. Namely, the methods of statistical genetics used to detect selection assume that positively selected variants emerge according to the infinite sites model (ISM; Kimura, 1969). In other words, on the time scale of a selective event, the beneficial single nucleotide variant (SNV) arises only once. Violations of the ISM in the context of sweeps targeting SNVs have been investigated – e.g., recurrent mutation without back mutation (Pennings and Hermisson, 2006b,a). However, the genome is mutationally complex and functional variants in the genome are not limited to SNVs. Micro- and minisatellites, copy number variants, and transposable elements are all abundant in genomes (Ellegren, 2004; Korbel et al., 2007; Huang et al., 2010) and possess mutational rates and processes that are notably different from point mutation. It is not clear how selection targeting these variants will affect linked variation. Selection targeting non-SNVs may affect linked variation in a fundamentally different manner than posited by the canonical model of selective sweeps (Maynard Smith, 1976) or fail to affect it altogether. In either case, standard genomic scans would categorically fail to detect a variety of important selective targets.

To address the effect of non-ISM mutation on inference of selection, we focus on microsatellites, the best-studied class of common genetic variation that mutates in a non-ISM manner (Ohta and Kimura, 1973; Levinson and Gutman, 1987; Weber and Wong, 1993). Microsatellites are sequential repeats of a 1-6 nucleotide motif. Microsatellite mutation increases or decreases the number of repeats and occurs at a rate exceeding that of point mutation by several orders of magnitude (Bhargava and Fuentes, 2010). Importantly, high microsatellite mutation rate causes recurrent mutation, back mutation, and multiallelism at microsatellite loci (Ellegren, 2004).

Long considered to be non-functional genetic variants, a growing body of evidence suggests that a subset of microsatellites are functional. Numerous studies have identified a correlation between microsatellite variation at genic microsatellites and levels of gene expression (Rockman and Wray, 2002; Vinces et al., 2009; Gemayel et al., 2010). In pathogenic bacteria, mutation of microsatellites in open reading frames or their promoters cause phase variation by which phenotypes are turned on and off (Weiser et al., 1989; Moxon et al., 1994). Other microsatellites have been implicated in circadian clock regulation (Michael et al., 2007), drought tolerance in barley (Nevo et al., 2005), and skeletal morphology in domestic dog breeds (Fondon and Garner, 2004). Microsatellite variation is often deleterious as well. Expansions of genic microsatellites cause a number of human neurological diseases (Orr and Zoghbi, 2007) as well as canine epilepsy (Lohi et al., 2005). These examples suggest that some microsatellites may be targets of positive and negative natural selection.

The selective regime of a multiallelic microsatellite is necessarily more complex than that of a diallelic single nucleotide polymorphism (SNP). In conjunction with its complicated mutational properties, a microsatellite therefore represents a substantially different selective target than a SNV. Recently, we developed biologically realistic models of the diploid fitness surface at a non-neutral microsatellite (Haasl and Payseur, 2013). These models were inspired by empirically observed corrleations between microsatellite length and gene expression (see Elmore et al. (2012) for an experimental investigation of the functions that relate allele size and gene expression in *Aspergillus flavus*). In most studied examples, the plot of gene expression versus allele size is a concave (e.g., Peters et al., 1999) or convex (e.g., Vinces et al., 2009) bell shaped curve or a step-like graph in which expression increases or decreases suddenly at a threshold allele size (e.g., Okladnova et al., 1998; Yamada et al., 2000). In other words, the relationship between allele size and gene expression may be divided into smooth and discontinuous examples. It therefore seems reasonable to model the genotypic fitness surface of a non-neutral microsatellite as either (1) a hill-like function in which one genotype is optimal and relative fitness declines in all directions from around this optimal genotype (Additive and Multiplicative fitness models; Figure 2-1A), or (2) a surface that contains sharp divisions between high and low fitness genotypes (Dominant and Recessive models; Figure 2-1A).

In this study, we use the Additive model (described in detail in Methods) to investigate the selective footprint of microsatellite selection on linked variation. We vary mutation rate and selective strength, conduct comparisons to multiple scenarios of selection on SNVs, and examine selective footprints through time. We compare the statistical power of various summary statistics to identify instances of SNV and microsatllite selection. We find that summaries of the SFS provide comparatively low power to detect selection at microsatellites, particularly when mutation rate is high. However, summaries of the haplotype distribution can provide moderate-to-high power to detect selection on microsatellites, even for an appreciable number of generations after mutation-selection equilibrium is achieved at the locus. In particular, when conditioned on the number of segregating sites, the number of haplotypes can provide considerable power to detect selection targeting highly mutable microsatellites.

Results

The spatial footprint of selection on microsatellites

SFS-based statistics

On average, SFS-based statistics were more sensitive to a hard sweep than selection on microsatellites. The spatial footprint of selection as measured by Tajima's D (Tajima, 1989) is shown in Figure 3-1A,B; these measures were taken immediately after fixation of the favored SNV or achievement of mutation-selection equilibrium at a selected microsatellite. In the case of microsatellite selection, the mean value of D was flat around zero (black line) except for a minor deflection at the position of the targeted microsatellite. This result contrasts sharply with the deep trough in mean D seen in simulations of a hard sweep (purple line).

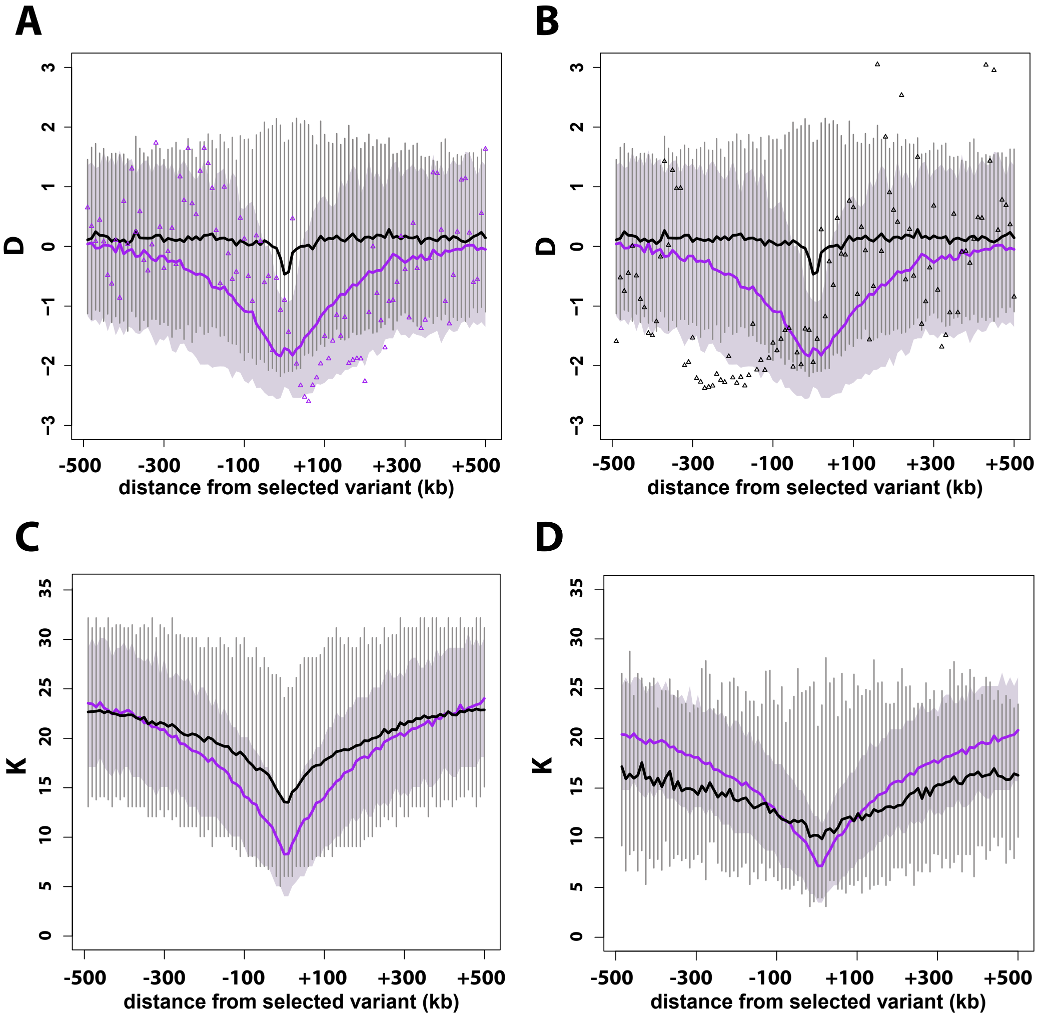

Figure 3-1: The spatial footprint of a hard sweep compared with that of selection on a microsatellite. (A-B) Tajima's D summarized across 500 simulations of a hard sweep ($s = 0.05, h = 0.5$) or selection on a microsatellite (additive model, $\phi = 5$, $g = -0.05$). D was measured in the generation following fixation of the beneficial SNV (hard sweep) or achievement of mutation-selection equilibrium (microsatellite selection). Purple and black lines mark the mean value of D across simulations of a hard sweep and microsatellite selection, respectively. The 5-95% interquantile range of D is marked by a light purple cloud (hard sweep) or vertical gray bars (microsatellite selection). Triangles mark values of D at each window for one simulation (purple: hard sweep; black: microsatellite selection). (C) The number of haplotypes K. Colors are the same as in (A-B). (D) Same as (C), except only microsatellites with values of Δ_{msat} in the top 10% of all simulations are included.

However, the mean value of D across 500 simulations is somewhat misleading. In Figure 3-1A the values of D for one simulation of a hard sweep are superimposed upon the distribution of D across all replicates. In keeping with previous results (Kim and Stephan, 2002), we found that the downward deflection in D was often asymmetrical relative to the selected SNV. That is, more severe declines in D were observed to the right of the SNV than the left (Figure 3-1A) or vice-versa. Individual simulations of microsatellite selection often demonstrated more dramatic departures from the mean value of D. Superimposed on the same summary of 500 simulations as in Figure 3-1A, Figure 3-1B shows a simulation in which Tajima's D was primarily deflected downwards to the left of the selected microsatellite. Although this is qualitatively similar to the hard sweep case, the width of the trough in D values is much wider. In addition, this replicate of microsatellite selection affected linked variation at a much longer range than in the hard sweep case, with values of $D < -2$ in excess of 300kb from the selected microsatellite. Also of note, in this same simulation replicate we observed highly positive values of D to the right of the selected microsatellite, which illustrates the comparatively higher variance in D and other summary statistics associated with microsatellite selection. Many simulations of microsatellite selection that used parameter values identical to those illustrated in Figure 3-1B (except for starting allele frequency distribution and the identity of the favored allele size, which were drawn randomly) only generated moderately positive and/or negative values of D across the entire simulated 1Mb sequence. Thus, microsatellite selection produced a highly variable and often very weak effect on the values of SFS-based statistics such as D. However, when D was driven negative by microsatellite selection, the decreases were often substantial, expansive, and long-ranged.

Spatial patterns of Fay and Wu's H_{FW} (Fay and Wu, 2000) and Zeng et al.'s E (Zeng et al., 2006) were qualitatively similar to those observed in D (Appendix, Figure A-3-1). Namely, mean values of these statistics also deviated little from neutral expectations in the case of microsatellite selection. Like D, however, both statistics exhibited greater intra- and inter-replicate variation for microsatellite selection. Also, when significant values of these statistics did emerge in simulations of microsatellite selection, they were often expansive and very strong (Figure A-3-1, right hand column). Finally, high values of Δ_{msat} were associated with a greater number of significant windows for both H_{FW} and E.

Haplotype-based statistics

On average, the decline in the absolute number of haplotypes (K) in response to positive selection on a SNV was greater than the decline associated with microsatellite selection (Figure 3-1C). Unlike D, however, the average decline in K was comparable between the two types of selected target. Furthermore, limiting consideration of microsatellite selection to the 10% of simulations with the highest values of Δ_{msat} – which quantifies the difference between allele frequencies at the start of selection and mutation-selection equilibrium – we observed a much broader selective footprint in the case of microsatellite selection (Figure 3-1D). Indeed, the *average* value of K in this case was lower than the 1% quantile of all neutral simulations as far as 200kb removed from the selected microsatellite. As with SFS-based statistics, microsatellite selection resulted in greater inter- and intra-replicate variability in haplotype-based statistics. This fact is evident in the much broader inter-quantile (5%-95%) ranges of K for simulated microsatellite selection (Figure 3-1C,D).

The temporal footprint of selection on microsatellites

SFS-based statistics

The power of SFS-based statistics to detect selection varied considerably over time and by selective target (Figure 3-2). For both hard and soft sweeps, D increased to high statistical power by the time of fixation of the favored SNV. The power afforded by D showed little sign of declining at the last time point sampled (2000 generations = 0.05 $4N_e$ generations post-fixation). On the other hand, the power of H_{FW} declined precipitously following fixation of the favored SNV, particularly in the case of a hard sweep (Figure 3-2C). Finally, E provided high power to detect selection, but only following fixation of the favored SNV (Figure 3-2E).

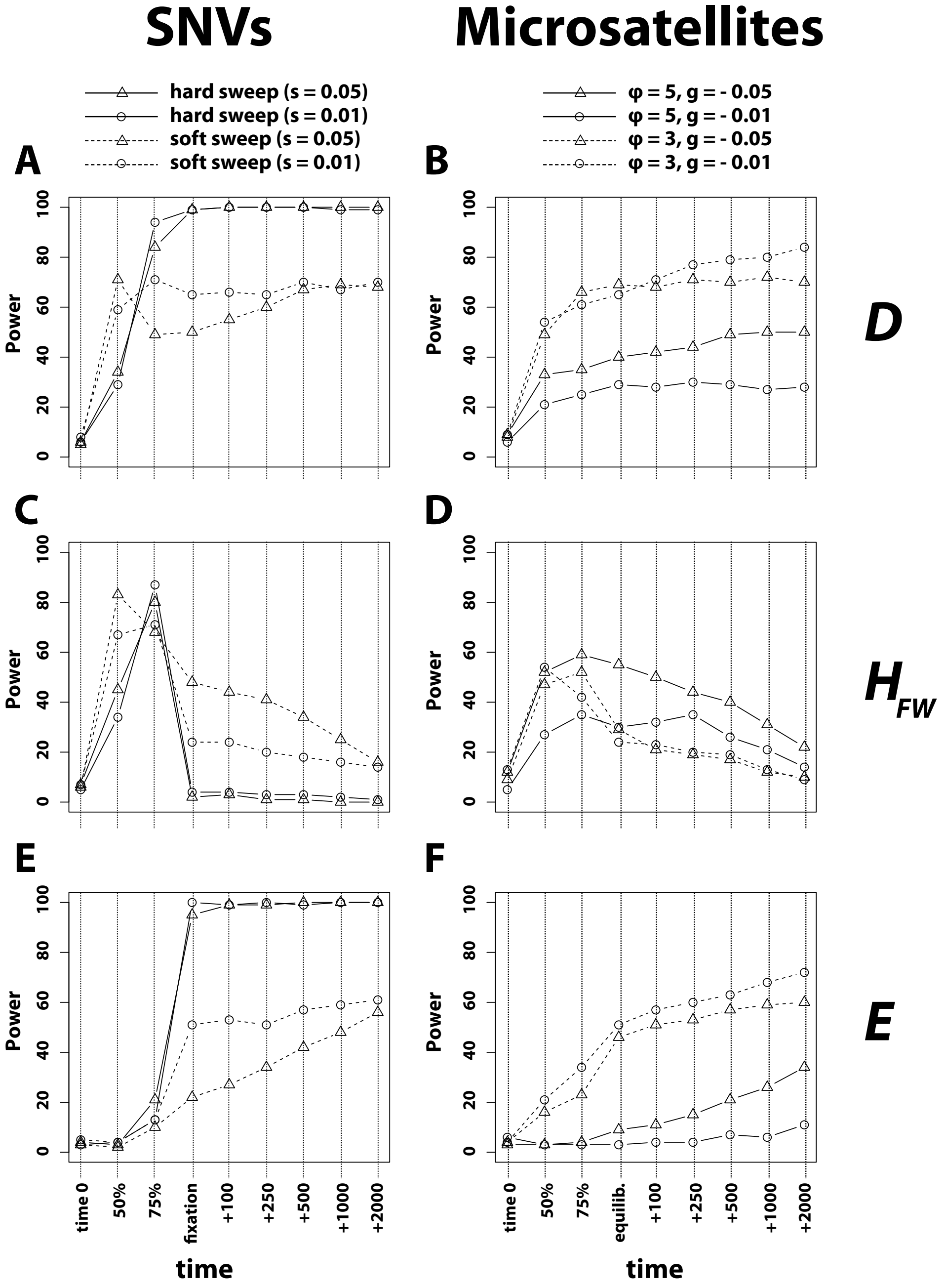

Figure 3-2: Statistical power of statistics that summarize the site frequency spectrum. Power to detect sweeps targeting SNVs is shown in the left column, while power to detect scenarios to microsatellite selection is shown in the right column. (A-B) The power of Tajima's D. (C-D) The power of Fay and Wu's H_{FW}. (E-F) The power of Zeng et al.'s E. Time points sampled are: time 0 = the generation before selection begins; 50% = half the time to fixation/equilibrium; 75% = three-quarters the time to fixation/equilibrium; fixation/equilibrium = one generation after fixation or mutation-selection equilibrium; +X = X generations after fixation or mutation-selection equilibrium.

The power of these same statistics to detect microsatellite selection was comparatively muted. D and E showed increasingly high power to detect selection after mutation-selection equilibrium was achieved, particularly when the mutation rate of the selected microsatellite was low (dashed lines, Figure 3-2B,F). However, the power of these two statistics was always less than their power to detect hard sweeps. Moreover, high rates of mutation at a selected microsatellite substantially depressed the power of D and E to detect selection. For example, E only began to register selection 1000 generations after mutation-selection equilibrium was achieved when $\phi = 5$ (solid lines, Figure 3-2F). Similar to soft sweeps, H_{FW} maintained power to detect microsatellite selection after mutation-selection equilibrium, though power was low to moderate (cf. Figure 3-2C and 3-2D).

Haplotype-based statistics

Both haplotype diversity, H, and frequency of the most frequent haplotype, M, maintained intermediate-to-high power to detect selection long after fixation in the case of positive selection targeting a SNV (Figure 3-3C,E). Conversely, the power of K declined rapidly following fixation of the beneficial SNV. In the case of a hard sweep, the statistical power of K declined to near-zero following fixation. Note, however, that the absolute number of haplotypes (K) did decline sharply in response to a hard sweep (Figure 3-1C,D) and remained low for hundreds to thousands of generations following fixation. On the other hand, K provided intermediate-to-high power to detect microsatellite selection before and after mutation-selection equilibrium was achieved (Figure 3-3B). Unlike other statistics, the power of K to detect microsatellite selection was markedly higher when mutation rate of the targeted microsatellite was high. Both H and M demonstrated intermediate-to-high power to detect microsatellite selection, although lower power than for hard sweeps (Figure 3-3,C-F).

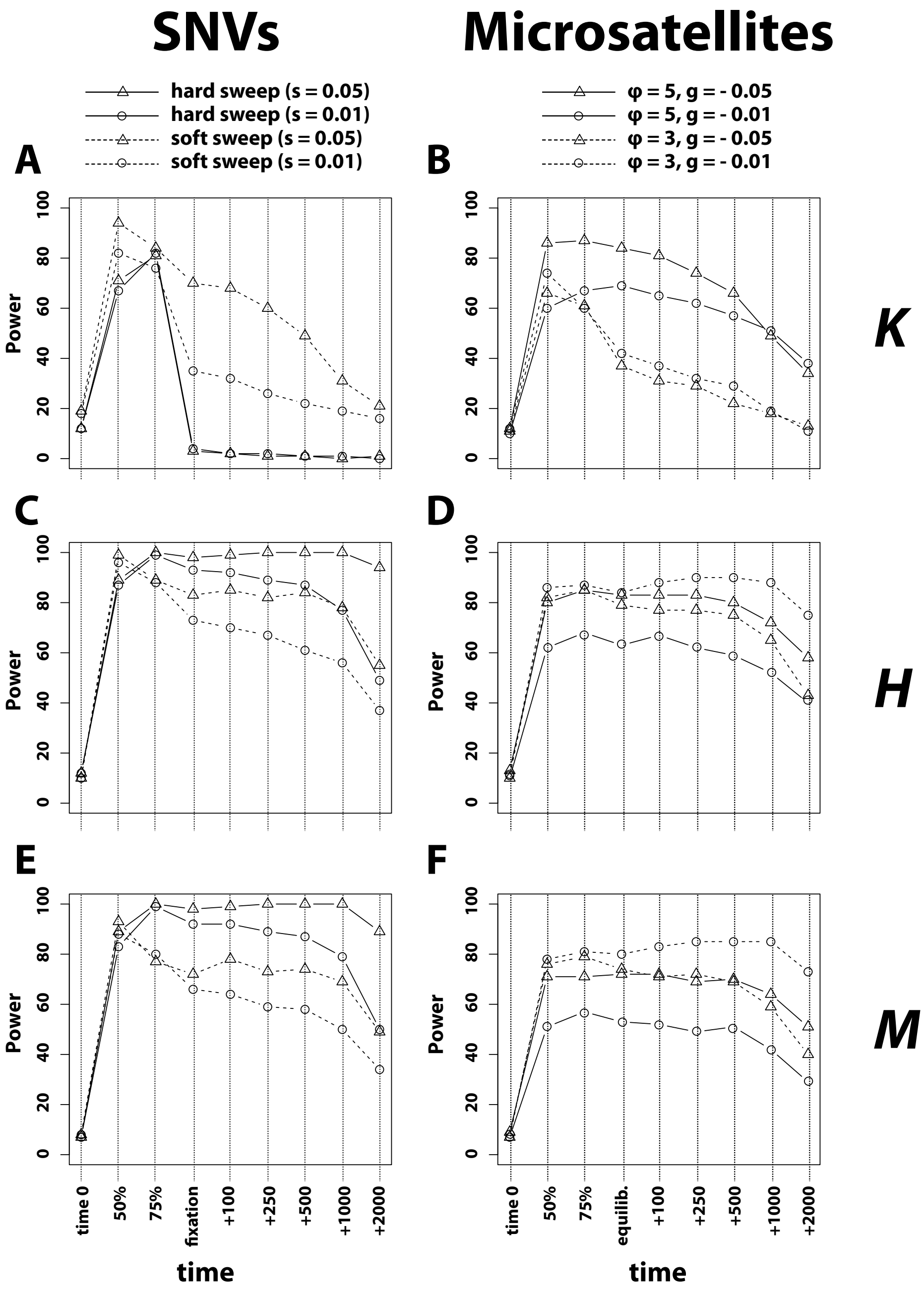

Figure 3-3: Statistical power of statistics that summarize the distribution of haplotypes. Power to detect sweeps targeting SNVs is shown in the left column, while power to detect scenarios to microsatellite selection is shown in the right column. (A-B) The power of K. (C-D) The power of H. (E-F) The power of M. Time points sampled are the same as in Figure 3-2.

Haplotype configuration and the uniqueness of the most common haplotype relative to other haplotypes

Haplotype configuration differed markedly among selective scenarios and selective targets (Figure 3-4). As expected, a hard sweep and strong selection ($s = 0.05$) drove a single haplotype to near fixation, implying a drastic loss of diversity that facilitated comparatively easy detection of hard sweeps using SFS-based statistics (Figure 3-2A,E). For a hard sweep with $s = 0.01$ the most common haplotype only obtained an average frequency of $\sim 60\%$ at fixation. Soft sweeps and selection on microsatellites with high mutation rate ($\phi = 5$) produced haplotype configurations in which the three most common haplotypes all had frequencies greater than 10% on average. In other words, multiple haplotypes became common. Of the scenarios tested, weak selection on a microsatellite with high mutation rate ($\phi = 5$, $g = -0.01$) induced the least change to the haplotype configuration at mutation-selection equilibrium. SFS-based statistics were particularly poor at detecting this selective scenario (Figure 3-2B,D,F) while haplotype-based statistics were surprisingly effective. K, H, and M all detected microsatellite selection under the $\phi = 5$, $g = -0.01$ scenario with $> 50\%$ power at most time points before and after mutation-selection equilibrium (Figure 3-3B,D,F). Selection on microsatellites with low mutation generated haplotype configurations similar to those of soft sweeps. Namely, frequencies of the most common haplotypes rose rapidly followed by a relatively static configuration for hundreds to thousands of generations.

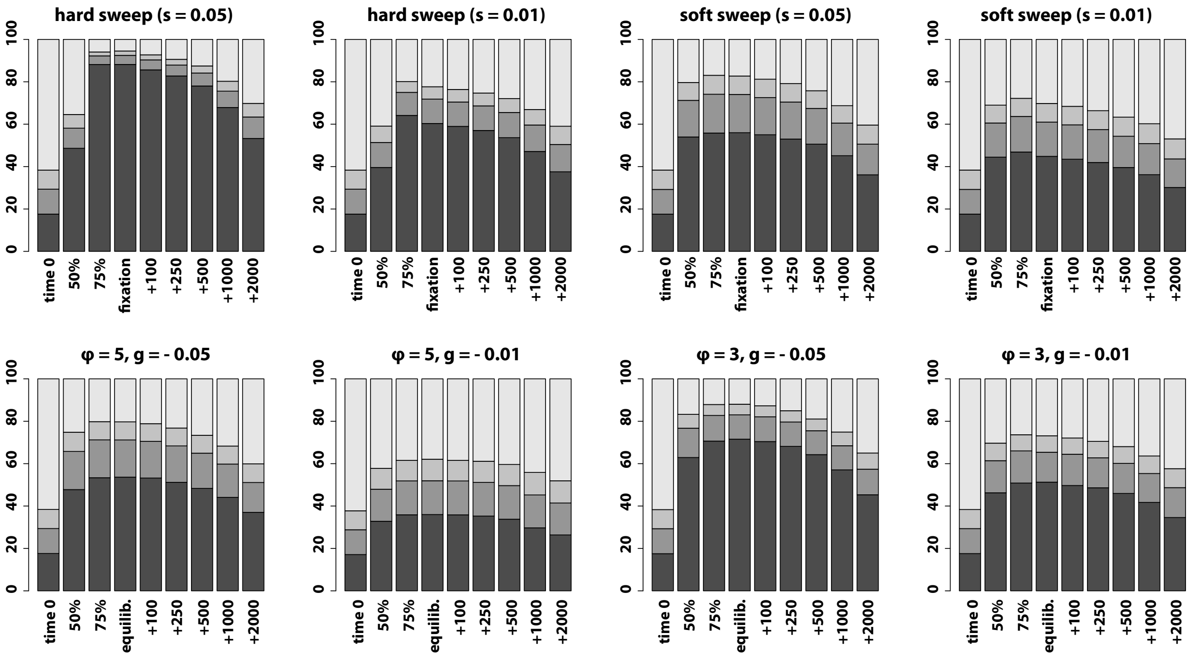

Figure 3-4: Changes in haplotype configuration through time. Each panel is labelled with the corresponding selective scenario and proportions illustrated are average proportions across 500 simulations each. The proportions of the sample of the first, second, and third most common haplotypes are shaded in decreasingly dark shades of gray. The proportion of the remaining haplotypes is shaded white. Time points sampled are the same as in Figure 3-2.

In addition to the configuration of haplotypes, we examined the uniqueness of the most common haplotype using the statistic A (Figure 3-5), which is the average number of pairwise differences between the most common haplotype and all other haplotypes (see Methods). Hard sweeps ($s = 0.05$ or $s = 0.01$) consistently drove A towards its lower bound of 1, which indicates that most secondary haplotypes only differed from the most common haplotype at one site. On average, other forms of selection reduced A, but not to the degree as hard sweeps. These results suggest that the statistic A might be used to differentially diagnose hard sweeps. More importantly, however, low values of A indicate that haplotypes share a recent common ancestor, while the much higher values of A observed in cases of selection on microsatellites with high mutation rates indicate that the remaining haplotypes are deeply divergent from one another.

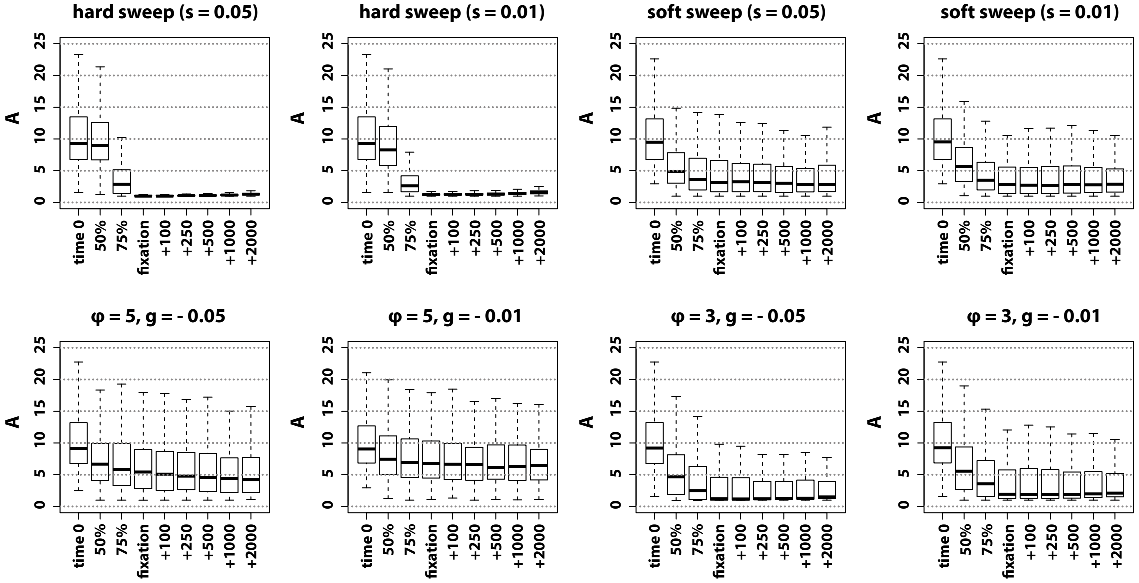

Figure 3-5: Changes in A through time. A measures the average number of pairwise differences between the most common haplotype and all other haplotypes in a sample. Each boxplot summarizes 500 simulations of the scenario listed above each panel for a specific time point. Time points are the same as in Figure 3-2. Box plot whiskers extend to the 5% quantile and 95% quantile, while the box marks the interquartile range and the horizontal bar is the median.

Patterns of genetic variation linked to a microsatellite candidate for selection

The differences between different forms of microsatellite and SNV-based selection reported here are helpful in an empirical context. Both direct selection on a microsatellite and selection on a linked variant are likely to reduce allele size variance at a microsatellite (Wiehe, 1998). The inference methodology introduced in Chapter 2 tested between four models of selection and one of neutrality. None of the models incorporated a linked selective sweep. Therefore it should be possible to obtain a highly positive result for selection using this method when the actual cause of the anomalous allele frequency distribution at the microsatellite is linked selection. Although we do not present a formal test, knowledge of the strengths and weaknesses of different statistics for different selective targets should help corroborate or dismiss putative selective events on microsatellites.

Based on numerous in vitro studies, Rockman and Wray (2002) documented several human microsatellites where allele size was correlated with the expression of a gene in cis. We genotyped several of these microsatellite loci in eight 1000 Genome populations. Characteristics of the population-specific allele frequency distributions at one of these loci – *HSD11B2* – suggested it may be a target of local additive selection of the type simulated in this study (Table 3-1). First, median allele size was $\texttt{CA}_{15}$ in African populations (LWK and YRI), but shifted to $\texttt{CA}_{18}$ in all six non-African populations sampled. Second, frequency of the median $\texttt{CA}_{18}$ allele was > 0.4 for all non-African populations, and considerably higher in others (e.g., 0.72 in CEU and CHB), suggesting this allele may provide a selective advantage in non-African populations. Allele frequencies greater and less than $\texttt{CA}_{18}$ generally decreased smoothly in both directions, suggesting an additive selective regime with $\texttt{CA}_{18}$ as the favored allele. Finally, values of R_{ST} (Slatkin, 1995b) suggest that differences in the allele frequency distributions between African and non-African populations are not solely due to population structure. R_{ST} between pooled African and non-African samples (0.327) was substantially greater than that of any of 50 intergenic microsatellite loci genotyped in the same eight populations. At the same time, only eight of the 50 intergenic loci had lower values of R_{ST} between the pooled European populations (TSI, CEU, FIN) and the pooled Asian populations (GIH, CHB). In other words, allele frequencies are unusually distinct between African and non-African samples yet unusually similar among non-African populations at the *HSD11B2* locus. These data suggest it is unlikely that the observed allele frequency differences are solely due to historical human migration and resulting population structure. We were interested in determining whether patterns of linked genetic variation supported our hypothesis of microsatellite selection or pointed to alternatives such as selection on a SNV.

population	11	12	13	14	15	16	17	18	19	20	21	mean	median	variance
LWK	0	2	0	21	6	2	10	5	4	0	0	15.5	15	3.60
YRI	1	2	4	6	11	2	12	4	2	0	0	15.5	15	3.79
TSI	0	0	0	0	3	3	8	27	4	1	0	17.6	18	1.08
CEU	0	0	0	0	1	2	6	35	3	1	0	17.8	18	0.61
FIN	0	0	0	0	1	4	18	28	7	2	0	17.7	18	0.89
GIH	0	1	0	2	5	3	4	30	4	1	0	17.3	18	2.35
CHB	0	0	0	1	1	1	1	39	7	4	0	18.1	18	0.99
MXL	0	2	0	0	1	1	7	14	5	3	1	17.7	18	3.49

Table 3-1: Allele size frequency distribution at the CA repeat in intron 1 of human gene *HSD11B2* in eight population samples from the 1000 Genomes Project.

Using empirical null distributions that accounted for the low rate of recombination at the *HSD112B* locus and population bottlenecks associated with extra-African migration (see Methods), we interrogated a 2Mb sequence centered on the *HSD112B* microsatellite in African (YRI, Yoruba), European (GBR, Great Britain), Asian (CHB, China), and American (MXL, Mexican-Americans in the United States) populations ($n = 100$ chromosomes each) using the same SFS- and haplotype-based statistics examined in our simulations. Patterns were similar in other continental populations of the 1000 Genomes Project and for ease of presentation we present results from these four populations only. E was significantly negative at numerous windows flanking the focal microsatellite in all three non-African populations, while no signal was found in the YRI sample (Figure 3-6). D was significantly negative at numerous windows surrounding the focal microsatellite in the MXL population, and for a small number of windows in the CHB and GBR populations (Appendix, Figure A-3-2). Again, none of the 500 windows were associated with significant values of D in the YRI sample. H_{FW} was not significant at any window within 750kb of the focal microsatellite in any of the populations (Appendix, Figure A-3-3). Although there is an obvious extended depression in the absolute number of haplotypes roughly centered on the focal microsatellite in all three non-African populations (Appendix, Figure A-3-4), only the CHB sample returned multiple significant values of the K statistic. However, windows obtaining significance were scattered across the sampled 2Mb region and most dense near the end of the examined sequence, nearly 1Mb away from the focal microsatellite. Numerous windows across the 1Mb flanking the focal microsatellite were also found to be significant in the CHB sample for haplotype diversity, H; in this case, significant windows flank the focal microsatellite rather evenly (Appendix, Figure A-3-5). The GBR and MXL populations only returned one significant window each for H, although all three non-African populations clearly demonstrated dramatic absolute reductions in H relative to YRI. Finally, M was significant at > 100 windows in all three non-African populations and at only one window in the YRI population (Appendix, Figure A-3-6). Notably, values of M regularly exceeded 80 (out of 100 sampled haplotypes) in non-African populations while rarely exceeding 50 in the YRI sample.

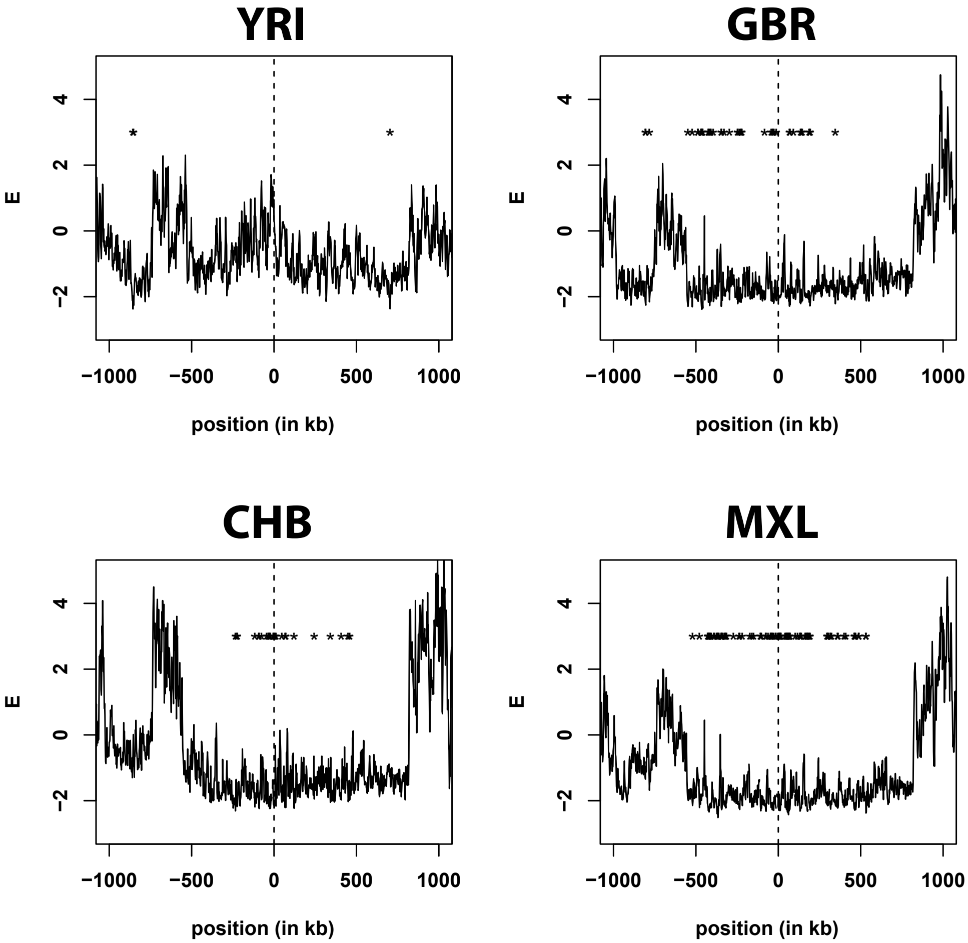

Figure 3-6: Values of E for the 2Mb region flanking the studied *HSD11B2* microsatellite, which is located at position 0 of each plot. Plotted values are for 10kb sliding windows (4kb jumps). Windows with significant values of E are marked by asterisks towards the top of each plot. From the top left and moving clockwise, the populations shown are are YRI, GBR, MXL, and CHB.

Discussion

Microsatellites targeted by selection may depart from the standard selective sweep model (Maynard Smith, 1976) in two important ways, both of which should decrease power to detect selection on microsatellites. First, high mutation rate at a selected microsatellite increases the frequency of recurrent mutation and back mutation. By assuming the ISM, the standard selective sweep model requires that all copies of the favored variant at fixation are identical by descent. When this assumption holds, selection is comparatively easy to detect because the selected variant is tagged by its original haplotypic background, which rises in frequency with the selected variant and generates a concomitant crash in sequence diversity. To the contrary, if a selected locus experiences common recurrent and back mutation in violation of the ISM, all copies of the favored variant need not be identical by descent. For example, many copies of the most fit allele at a microsatellite locus targeted by selection may be recent products of mutation from less fit alleles rather than direct descendants of the first chromosome to carry the favored allele size. Thus, a favored microsatellite allele may exist on several different haplotypic backgrounds, making it more difficult to detect the presence of selection using statistics that rely on substantial deformations of the site frequency spectrum. The negative correlation between the prevalence of recurrent mutation and power to detect selection is demonstrated by our results for D and E, which provide very low power when microsatellite mutation rate is high (solid lines, Figure 3-2B,F).

On average, selection on a microsatellite with high mutation rate fails to drive a single haplotype to frequencies greater than 0.5 (Figure 3-4). More importantly, haplotypes existing in a population at mutation-selection equilibrium are highly divergent from each other ($A > 5$ on average). In contrast, hard sweeps greatly reduce haplotypic diversity as the favored variant rises to fixation (Figure 3-4). Moreover, haplotypes that remain after fixation are all highly similar; most only differ from the most frequent haplotype at a single site as evidenced by values of $A \sim 1$ (Figure 3-5). Thus, while haplotypes following fixation due to a hard sweep primarily differ from each other due to recent point mutation, we infer that most differences between haplotypes at mutation-selection equilibrium in the microsatellite case reflect the deeply divergent ancestries of the haplotypes. These differences help explain why haplotype-based statistics provide more power than SFS-based statistics to detect microsatellite selection. While haplotype diversity is substantially reduced by selection on a microsatellite (i.e., H and K go down), linked sequence diversity across the SFS may remain rather high due to the divergent ancestries/sequences of the surviving haplotypes. Pennings and Hermisson (2006b) obtained similar results in their investigation of soft sweeps with recurrent mutation.

The second potential departure from the standard selective sweep model is that selection on microsatellites may often include selection on standing variation. Indeed, several authors have posited that microsatellites represent important targets of selection because they accumulate functional variation that can be drawn upon immediately when environmental conditions change (Kashi et al., 1997; King et al., 1997; Trifonov, 2004). While recurrent mutation leads to association of the selected variant with multiple divergent haplotypes during the course of a selective event, selection on standing variation describes a situation in which the same association exists at the onset of selection.

We used Δ_{msat} to quantify the distance between the allele frequency distribution of a microsatellite when selection begins and at mutation-selection equilibrium. We previously showed that this distance is positively correlated with the duration and cost of microsatellite selection (Haasl and Payseur, 2013). Here, we find that Δ_{msat} also influences the selective footprint left by microsatellites under selection. Low values of Δ_{msat} weaken the selective footprint (Figure A-3-1) and vice-versa (Figures 3-1,A-3-1). High values of Δ_{msat} indicate there is little to no overlap between starting and equilibrium allele frequency distributions. In many such cases, the favored allele does not yet exist in the population when selection begins. Once the favored allele is discovered via mutation, however, it quickly rises in frequency, making it less likely that it will become linked to a variety of haplotypic backgrounds. Conversely, low values of Δ_{msat} often indicate that the favored allele has existed for some time under neutrality and thus obtained appreciable frequency. In short, low values of Δ_{msat} indicate a situation analogous to selection on standing variation, while high values of Δ_{msat} indicate a situation analogous to selection on new variation. Thus, Δ_{msat} impacts the effect size of selection on linked genetic variation in addition to the cost and duration of selection. Given its importance to selective dynamics and because the starting allele frequency distribution is unavailable in most empirical situations, the starting distribution of allele sizes (or its proxy, Δ_{msat}) presents a troubling nuisance parameter for inference of microsatellite selection.

The outlook for detecting microsatellite selection using patterns of linked variation may seem bleak. With appreciable probability, instances of selection on microsatellite variation will include both recurrent mutation and selection on standing variation. This concern is realized in the case of SFS-based statistics, for which statistical power to detect selection never exceeds 50% when mutation rate is high (Figure 3-2). On the other hand, haplotype-based statistics yield moderate-to-high power to detect microsatellite selection regardless of mutation rate. Of particular note is the long-lived power of K to detect selection on microsatellites with high mutation rates. This result runs counter to the other five statistics, for which microsatellites with low mutation rate are either easier to detect or yield comparable power to microsatellites with high mutation rate. Surprisingly, K provides greater statistical power to detect selection on microsatellites with high mutation rates than it does to detect soft or hard sweeps post-fixation. To explain this difference in statistical power, recall that we determined the significance of observed K by comparing it to null distributions of K conditioned on the observed value of S. Although hard sweeps dramatically reduce K (Figure 3-1), they also reduce S. Under neutrality, when a sample bears low S it is also expected to harbor fewer distinct haplotypes. Thus, low S, low K conditions are characteristic of a hard sweep after fixation, but are hardly unexpected under the null hypothesis of neutrality. Conversely, microsatellite selection can link the favored allele to multiple divergent haplotypes, particularly when mutation rate is high.

Thus, selection on microsatellites with high mutation rates produces a combination that is unexpected under neutrality: intermediate S and low K. Importantly, simple demographic change may also fail to produce this unexpected pattern; population bottlenecks should decrease both S and K while expansions should increase S and K.

Fixation of a beneficial SNV terminates the transient selective phase and its associated effect on linked diversity. However, unless selection is strong and mutation rate is low, fixation of the favored variant at a selected microsatellite does not occur (Haasl and Payseur, 2013). Instead, new mutation continuously introduces less fit alleles to the population – i.e., mutation-selection equilibrium is achieved rather than fixation. The constant production of less fit microsatellite alleles ensures that selection continues to act at the locus, thereby eliminating less fit alleles along with their linked variants. These conditions are analogous to background selection (Charlesworth et al., 1993). In other words, continuous selection on microsatellites with high mutation rates may cause long-term reductions in linked sequence diversity. For example, non-triplet repeats in exons and trinucleotide microsatellites that underlie human triplet expansion disorders may cause local depressions in linked sequence diversity if mutation rate is great enough to generate substantial numbers of deleterious alleles. Moreover, higher mutation rates at a selected microsatellite will cause more frequent production of deleterious alleles and concomitant elimination of linked diversity. This predicts that mutation rate among genic microsatellites will be negatively correlated with flanking sequence diversity.

Finally, the spatial pattern of selective footprints left by microsatellite selection was unusual in that it was often much wider and long-ranged than in the case of selection on a SNV. One obvious reason for the greater width of selective footprints left by microsatellite selection is the rapidity with which microsatellite selection may act. The additive selective regime simulated here appears to be particularly efficient (Haasl and Payseur, 2013). Fast selective events leave little time for recombination to degrade linked haplotypes, thereby leaving more extensive selective footprints than slow selective events. However, selection scenarios where g was weak (-0.01) produced similar spatial patterns. When g is this low, the time to mutation-selection equilibrium increases substantially. In addition, intra-replicate variability of the monitored statistics was rather large for microsatellite selection (Figure 3-1, A-3-1). Therefore, the possibility remains that a more complicated dynamic specific to microsatellite selection is responsible for the observed patterns.

We also note that although some empirical results suggest that additive or multiplicative models are the most biologically plausible forms of microsatellite selection (Vinces et al., 2009; Gemayel et al., 2010), the dynamics of microsatellite selection are not known. Different selective regimes may produce selective footprints far different from those suggested by the results of our simulations. In addition, it is difficult to equalize selective strength between the scenarios of microsatellite and SNV-based selection as the parameters used to impose selection – s for SNVs and g for microsatellite selection – have different interpretations. Thus, there is some concern that differences between the power of the statistics observed in our simulations of SNV and microsatellite selection may reflect differences in simulated selective strength rather than the divergent mutational and selective regimes. However, we note that mutation had a greater influence on the power of different statistics to detect microsatellite selection than the choice of selection parameter g – e.g., in Figure 3-2A,B,D, dashed lines (low mutation, high and low values of g) are more similar to one another than solid lines (high mutation, high and low values of g). The same is true of haplotype configuration (Figure 3-4). This suggests that mutational dynamics have a greater influence on the selective footprint left by microsatellite selection than the value of the selection parameter, minimizing the effect of possible disparities between selective strength.

The CA repeat in intron 1 of *HSD11B2*

In vitro assays suggest that expression of *HSD11B2* is modulated by the allele size of a CA microsatellite in intron 1 of the gene (Agarwal et al., 2000; White et al., 2000). Furthermore, individuals with type 1 diabetes mellitus are enriched for allele sizes less than CAx18 (Lavery et al., 2002) and Williams et al. (2005) found that Sardinians who bore two alleles of size CAx18 or greater had significantly lower values of a biomeasure associated with hypertension. We found that the allele frequency distributions of non-African populations were characterized by similar shifts in the median (from CAx15 to CAx18) and reductions in variance compared with sampled African populations (Table 3-1). Together, these facts suggest the *HSD11B2* microsatellite might be a target of direct selection.

However, in light of our simulation results, numerous factors suggest the anomalous patterns of variation in non-African populations are better explained as a consequence of selection targeting a linked SNV than direct selection on the *HSD11B2* microsatellite. First, our simulations indicate that K is the statistic with the greatest power to differentially detect selection on a microsatellite. However, despite low values of K surrounding the *HSD11B2* locus (Figure A-3-4), few windows in any of the four populations examined were associated with significant values of K when conditioned on S. Instead empirical data suggest a hard sweep or population bottleneck, both of which decrease S in addition to K. Second, all three non-African populations possessed significant values of E at numerous windows in the vicinity of the *HSD11B2* microsatellite (Figure 3-6), while the African YRI population did not. While E possesses high power to detect hard selective sweeps immediately following fixation of the favored variant, it provided the lowest power of any statistic to detect microsatellite selection in our simulations (Figure 3-2F). Third, values of M are very high across the region flanking the *HSD11B2* locus. Although our simulation results suggest M possesses intermediate power to detect both soft sweeps and microsatellite selection, these types of selection are generally characterized by lower though significant values of M (Figure 3-4). Very high values of M, such as those observed in the three non-African populations (Figure A-3-7) are suggestive of a hard selective sweep or recent bottleneck. Finally, the observed reduction in variation at the *HSD11B2* microsatellite in non-African populations (Table 3-1) is consistent with patterns observed at other microsatellites linked to likely targets of selection in humans such as *SLC24A5* and *CYP3A5*, where a reasonably variable microsatellite in Africa is reduced to a nearly invariant microsatellite in the rest of the world (Appendix, Figures A-3-7, A-3-8, A-3-9). Thus, data are consistent with selection on a variant linked to the examined microsatellite. Interestingly, the *HSD11B2* is positioned near the center of a 1Mb region identified as a site of positive selection in six independent genome-wide scans (Akey, 2009), which provides increased support for the linked selection hypothesis. Regardless, the microsatellite itself seems an unlikely target of selection.

Prospects for genome scans on microsatellites and other non-SNV targets

The genome of a species comprises numerous types of genetic variants with a variety of mutational mechanisms and rates. Due to their simplicity and abundance, single nucleotide polymorphisms receive the most empirical and theoretical attention. As a result, methods used to detect selection were specifically developed to detect anomalies in sequence data that are expected when selection targets a SNV. Whether or not selection on variants with different mutational properties will produce similar effects on sequence variation is unclear.

In general, our findings are encouraging. Statistics developed to detect selection on SNVs possess some power to detect microsatellite selection. Unfortunately, while different statistics have different power profiles depending on the selected target, a significant value of one statistic does not definitively indicate that a SNV was targeted by selection rather than a microsatellite, or vice versa. It would be advantageous if a genome scan could not only identify significant regions but provide us with some indication of the type of selected target. Importantly, we find that the statistic K may help in this regard. When conditioned on the number of segregating sites, K provides higher statistical power to detect selection on rapidly mutating microsatellites than any of the other types of selection simulated. In other words, K might function in genomic scans to discriminate between selection on SNVs and microsatellites.

The discriminatory power of K could be increased in two different ways. First, a joint statistic that includes K may be helpful. Zeng et al. (2006) used a joint D and H statistic as the basis for a test for selection that was largely insensitive to the confounding effects of demography. Similarly, a significant value of K and non-significant value of A might be powerful at specifically detecting microsatellite selection. Second, in the case of candidate microsatellites, a more sophisticated inference procedure might be employed. We previously used approximate Bayesian computation (ABC) in conjunction with a recessive model of selection to infer the evolutionary history of the microsatellite that causes Friedreich's ataxia (Haasl and Payseur, 2013). In this case, we only relied on patterns of variation at the microsatellite itself. Simultaneous simulation of sequence and microsatellite variation would allow the addition of sequence summary statistics to comparisons of the empirical and simulated data sets within the ABC framework. An advantage to this second approach is that it does not require the computation of null distributions on joint statistics such as (D, H) or (A, K). A disadvantage is that it cannot be scaled up to the genome-wide scale.

Finally, the power analyses presented here specifically apply to a single model of microsatellite selection. Whether or not our results are broadly applicable to other types of non-SNV targets is uncertain. As an example, copy number variants (CNVs) share many attributes with microsatellites. Although their repeated segments range from 1kb to several

Mb, CNVs are tandem repeats subject to recurrent mutation that can increase or decrease the number of repeats. When mutation occurs via meiotic nonallelic homologous recombination, the mutation rate of CNVs is thought to be comparable to that of microsatellites ($\sim 10^{-4}$; Conrad and Hurles, 2007). Moreover, CNVs are sometimes functional – e.g., a positive correlation exists between the number of copies of the amylase gene *AMY1* and starch content of diet in humans (Perry et al., 2007). Due to these similarities, it is tempting to suggest that our results are applicable to CNVs as well. However, uncertainty regarding the mutation rates of CNVs as well as their much larger physical scale should lend caution to this assessment.

Methods

Models of selection and mutation

SNPs

We considered a diallelic SNP where the relative fitness of allele B was greater than that of allele b. To model positive selection at the locus, we used an additive selective regime in which relative genotypic fitnesses were $w(\text{BB}) = 1$, $w(\text{Bb}) = 1 - hs$, and $w(\text{bb}) = 1 - s$. We set dominance coefficient $h = 0.5$, and selection coefficient s to either 0.05 or 0.01. We assumed a constant per-site point mutation rate of $2.5e - 08$ and mutation followed the infinite sites model (Kimura, 1969).

Microsatellites

Let a_i represent a microsatellite allele with i repeats of a nucleotide motif (we refer to this as allele size i). We focus on a simple instance of the additive model of microsatellite selection presented in Haasl and Payseur (2013) in which a single allele size, x, is most fit with relative fitness $w(a_x) = 1$. The relative fitness of each allele is then defined as $w(a_i) = 1 - g|x - i|$, where g is the gradient parameter. For example, if $g = -0.01$ and $x = 10$, then alleles of sizes 9 and 11 each have a relative fitness of 0.99. The relative fitness of genotype $a_i a_j$ was then calculated as $w(a_i a_j) = [w(a_i) + w(a_j)]/2$.

We used a logistic model of microsatellite mutation rate (Haasl and Payseur, 2013) in which mutation rate is a monotonically increasing though nonlinear function of allele size. The model requires specification of three parameters: ψ controls the allele size at which mutation rate begins to increase, ϕ controls the maximum mutation rate, and γ controls the slope of increase. In all simulations, we used $\psi = 2$ and $\gamma = 0.15$ while ϕ was either set to 3 or 5 (Appendix, Figure A-3-10). Mutation was symmetric, equally likely to increase or decrease allele size. Mutational step size followed a geometric distribution with $p = 0.95$ – i.e., 95% of mutations were single step.

Simulation

We performed exact, forward-in-time simulations and assumed a constant population size of $N_e = 10,000$ (20,000 chromosomes) programmed in C++. 500 replicates of each distinct evolutionary scenario were run. In the case of SNV selection, we noted the generation at which the beneficial SNV became fixed in the population. In simulations of microsatellite selection, we noted the equilibrium generation, which we defined as the first generation for which the difference between the frequency of the most fit allele a_x and its frequency at mutation-selection balance (determined in the absence of genetic drift; Haasl and Payseur, 2013) was less than $1/2N = 5e - 05$. Simulated sequences were either 1Mb or 30kb in length.

Neutral, pre-selection phase

For each simulation replicate, we used neutral coalescent simulations implemented in MS (Hudson, 2002) to obtain a starting population of 20,000 chromosomes ($N_e = 10000$ diploids). We then extracted the genealogy corresponding to the exact center of the simulated 1Mb or 30kb sequence. In the case of microsatellite selection, we input this genealogy to our program MARKSIM (Supplementary text; Haasl and Payseur, 2011), which output a starting microsatellite allele for each chromosome. In all cases, we specified allele size of 8 as the MRCA of the genealogy. The only significance of this allele size was that it was sufficiently large to provide modest mutability at the locus, which more often than not resulted in a microsatellite locus that entered the selective phase as polymorphic. The microsatellite locus was placed at the exact center of the simulated sequence and the size of the most fit allele was determined randomly on the interval [8,20]. Thus, for many replicates the most-fit allele did not exist in the population when selection began. For simulations of SNP-based sweeps from standing variation, we also used the genealogy corresponding to the center of the simulated sequence. We searched this tree for a bipartition that allowed us to generate a new SNP at the center of the sequence with a minor allele frequency on the interval [0.1, 0.15]. The minor allele was treated as the beneficial SNV. In simulations of a hard selective sweep, we simply placed a single copy of a beneficial SNV at the center of one random chromosome. All other chromosomes carried the less fit ancestral allele.

Selection phase

The selective phase proceeded as follows:

1. Set generation counter to 1.

2. SELECTION. Determine which of the 10,000 individuals survive based on the genotypic fitness of the selected SNP or microsatellite genotype.

3. REPRODUCTION and RECOMBINATION. Use the pool of survivors from step 1, and repeat the following steps until 10,000 offspring are generated:

   - randomly choose two parent individuals
   - determine if recombination occurs; if so, perform crossover, yielding 2 recombinant and 2 non-recombinant chromosomes
   - choose one chromosome from each parent for inheritance by the offspring

4. MUTATION. For each chromosome of the next generation, randomly determine how many (if any) new SNPs arise (Poisson-distributed) and at what position(s). Check for mutation at the microsatellite.

5. (SNP selection only) If the beneficial SNV is lost, set generation counter to 1 and start selective phase over from the original set of starting chromosomes.

6. Determine if fixation (SNVs) or equilibrium (microsatellites) has been achieved. Increment generation counter and return to step 2.

We stopped simulations of 1Mb sequence at the point of fixation/equilibrium. For simulations of 30kb sequence, we simulated 2000 additional generations beyond the point of fixation/equilibrium following. In the case of SNV selection, post-fixation generations did not require performance of step 2.

Sampling

At each sampling timepoint, we pulled a random sample of 50 individuals (100 chromosomes) from the population. For 1Mb simulations, we only sampled the population upon fixation/equilibrium. For simulations of 30kb sequence, we sampled every generation prior to fixation/equilibrium, and then at the following timepoints: fixation/equilibrium and 100, 250, 500, 1000, and 2000 generations afterwards.

Measuring the distance between starting and equilibrium allele frequencies at a microsatellite targeted by selection

For a microsatellite under selection, we previously showed that the duration and cost of selection are positively correlated with the distance between the starting allele frequencies and those at mutation-selection equilibrium (Haasl and Payseur, 2013). Because the most-fit allele size and the starting distribution of allele sizes were randomly determined for each replicate, this distance varied between replicates. We quantified this consequential distance as

$$\Delta_{msat} = \sum_{x \in \mathcal{S}} \sum_{y \in \mathcal{E}} |x - y| p_x p_y,$$

where $\mathcal{S}$ is the set of starting allele sizes, $\mathcal{E}$ is the set of equilibrium allele sizes, and $p_\bullet$ is allele frequency. Equilibrium alleles and their frequencies were determined using a single deterministic simulation for the appropriate selective and mutational parameter values.

Summary statistics

We calculated the following statistics for all simulations: (1) Tajima's D (Tajima, 1989); (2) Fay and Wu's H_{FW} (Fay and Wu, 2000); (3) Zeng *et al*'s E (Zeng et al., 2006); (4) number of distinct haplotypes, K; (5) haplotype diversity, H, and; (6) count of the most frequent allele, M. The first three statistics contrast separate estimators of the scaled mutation rate $\theta = 4N_e\mu$. Although these estimators possess identical expectations at mutation-drift equilibrium, they diverge from each other in characteristic ways under non-equilibrium conditions due to dependencies on different partitions of the frequency spectrum (Zeng et al., 2006). The final three statistics summarize the distribution of sampled haplotypes. Each statistic was separately calculated for each non-overlapping 10kb window in the simulated sample of 1Mb or 30kb sequences.

We also defined a seventh statistic that was useful for distinguishing a hard sweep from other scenarios of selection, such as soft sweeps and microsatellite selection:

$$A = \sum_{i=2}^{K} n_i d(h_1, h_i) \ / \ \sum_{i=2}^{K} n_i$$

where h_i is the ith haplotype in the ordered (by descending frequency) set of observed haplotypes, $d(h_i, h_j)$ is the number of pairwise differences between haplotypes i and j, and n_i is the count of haplotype i. In words, A is the average number of pairwise differences between the most common haplotype and all secondary haplotypes. For samples with more than one haplotype, the lower bound on A is unity, which indicates all secondary haplotypes differ from the most common haplotype at a single site. When only one haplotype is present in the sample, A is undefined.

Power analyses

Scaled mutation and recombination parameters, θ and ρ, respectively, can vary widely across the genome. Unfortunately, equilibrium values of the statistics we measured here depend on the values of these two parameters. To incorporate empirical uncertainty regarding θ and ρ, we computed empirical null distributions for each statistic based on 10^6

neutral coalescent simulations of 10kb sequences ($n = 100$) in MS (Hudson, 2002), which each began with independent draws from uniform prior probability densities for θ and ρ. We considered reasonable ranges of these parameters for human: recombination rates between 0.75 and 2.0 cM/Mb, per-site point mutation rates μ between 5e-09 and 2.5e-08, and effective population sizes N_e between 10000 and 25000. For a 10kb sequence, these imply priors of $\theta \sim$ [2, 25] and $\rho \sim$ [3, 20]. The empirical distribution for each statistic was conditioned on the number of segregating sites, S, and was simply the distribution of the statistic across the subset of simulated 10kb windows in which $S = s$.

We calculated power using the results from 30kb simulations, in which the selective target was positioned at the midpoint of the 30kb sequence. For each statistic, we tested each of the three non-overlapping 10kb windows for significance and counted selection as detected if one or more of the three windows produced a significant result. The positive selection modelled here is expected to shift each statistic in one, specific direction. Therefore, all tests were one-tailed. Values of statistics such as Tajima's D, which is expected to decrease in response to positive selection, were deemed significant if they ranked below the $\alpha = 0.05/3 = 0.0167$ quantile of the appropriate empirical distribution. M, on the other hand, is expected to increase in response to positive selection and was deemed significant if its rank was greater than or equal to the $1 - 0.05/3 = 0.9833$ quantile of the appropriate empirical distribution. We calculated the power of a statistic as the fraction of 500 replicates in which selection was detected by the statistic.

Genotyping a CA microsatellite in intron 1 of human gene *HSD11B2*

DNA samples from 200 participants in the 1000 Genomes Project were obtained from Coriell Institute for Medical Research (Camden, NJ). Genotyping of the CA microsatellite in the first intron of *HSD11B2* (chr16:67467591-67467631; GRCh37/hg19 assembly) was performed by Prevention Genetics (Marshfield, WI) using capillary electrophoresis. Genotypes were recovered from 193 of the 200 individuals for 386 total alleles.

Sequence analysis of the *HSD11B2* microsatellite

For the 2Mb sequence flanking the intronic *HSD11B2* microsatellite, we downloaded all variant calls for all participants of the 1000 Genomes Project from the YRI (Yoruban), GBR (British), CHB (Han Chinese), and MXL (Mexican-American) sample populations. We randomly selected 100 chromosomes from each population. We moved a sliding 10kb window across the 2Mb sequence in steps of 4kb. For each window and for each population separately, we calculated K, H, M, D, H_{FW}, and E. To compute empirical null distributions for these statistics we ran coalescent simulations of 2Mb sequences in MS ($n = 100$ chromosomes), splitting each simulated sequence into 200 individual 10kb sequences. The value of θ for the 2Mb sequence in each simulation was drawn from the uniform prior [400, 5000]. Recent estimates of the sex-averaged recombination rate in human suggest the 2Mb locus centered on *HSD11B2* is characterized by very little recombination in African, European, and Asian populations (Appendix, Figure A-3-11; Kong et al., 2010). In light of these estimates, we considered 0.01-0.5 cM/Mb to be a reasonable range for the recombination rate at the locus. The locus-wide value (2Mb) of ρ was therefore drawn from the uniform prior [8, 800]. To partially control for demographic change, we also included a population bottleneck in each simulation. Bottlenecks began and ended instantaneously. Bottleneck strength β was drawn from prior $\beta \sim$ [0.025, 0.05]. Bottleneck time τ was drawn from prior $\tau \sim$ [0.125, 1], where time is in units of $4N_e$ generations. All bottlenecks ended after 10 generations. All tests of significance were one-tailed, with critical value $\alpha = 0.05/500 = 0.0001$, or 0.9999 for the upper-tail test on M (Bonferroni correction for multiple tests, 498 10 kb windows).

CHAPTER 4: EVIDENCE OF NATURAL SELECTION AT HUMAN MICROSATELLITE LOCI

Introduction

Numerous authors have speculated on the possibility that microsatellites are targeted by selection (King et al., 1997; Fondon and Garner, 2004; Trifonov, 2004; Kashi and King, 2006; Gemayel et al., 2010). However, our understanding of the mechanics and consequences of selection on microsatellites is lacking. The previous chapters have demonstrated the complexity of modeling natural selection and mutation at microsatellite loci. However, the ideas and results in these chapters also offer hope that microsatellite selection can be detected. In this chapter, we present results from studies aimed at doing just that.

We recently genotyped microsatellites in participants of the 1000 Genomes Project (1000 Genomes Project Consortium, 2010), which has now sequenced the genomes of more than 2000 individuals around the world using low-coverage (2-4x), next-generation sequencing. These data are not suitable to the recovery of most microsatellite genotypes. However, by genotyping microsatellites of interest in 1000 Genomes participants using standard methods, we now have sequence and reliable microsatellite data for 200 individuals at these microsatellites.

In addition to genotyping more than 50 intergenic microsatellites as a reference data set, we genotyped microsatellites that were of interest for a number of different reasons. In this chapter, we analyze patterns of polymorphism in the putatively neutral intergenic microsatellites and then focus on two sets of microsatellites that are candidates as targets of natural selection. First, we examine polymorphism at long, dinuclotide microsatellites that overlap with exons in the human genome. These are inherently dangerous microsatellites, because most mutations of these microsatellites will generate frameshifts. Second, we use and extend the models and methods of Chapter 2 in order to characterize the evidence for selection at six candidate microsatellites. Length variation at five of these six candidates has been shown to affect gene expression. The sixth candidate is of interest due to its low variance in a very large sample, despite the large allele size of its most common allele.

Together, these analyses highlight the complexity of microsatellite evolution and the challenges involved in detecting selection at microsatellites. However, they also demonstrate the unique character of microsatellite variation and mutation that makes these variants a intriguing topic of study.

Genotyping 1000 Genomes data

We genotyped nearly 200 microsatellites in participants of the 1000 Genomes Project. DNA samples were obtained from the Coriell Institute for Medical Research (Camden, NJ) and genotyping was performed using capillary electrophoresis at Prevention Genetics (Marshfield, WI). As a quality control measure, we included one individual twice among the samples sent to Prevention Genetics. Technicians at Prevention Genetics were blind to the identity of this individual. In all cases, genotypes of this individual were identical between replicates.

To enable conversion of raw data (reported in base pairs) into actual allele sizes, for each locus five random homozygotes (as determined by electrophoresis) were sequenced using Sanger sequencing at Prevention Genetics. These data provided the calibration necessary to convert fragment sizes to allele sizes. Furthermore, sequencing did not uncover any errors in the original genotyping data.

The 200 sampled individuals were drawn from eight worldwide populations. Throughout this chapter, we use the standard three-letter abbreviations for these populations (Table 4-1). With the exception of one father-mother-daughter trio each in the YRI and CEU

population	location	abbreviation	sample size
Yoruban	Ibadan, Nigeria	YRI	25
Luhya	Webuye, Kenya	LWK	25
CEPH	NW European-Americans in Utah	CEU	25
Toscani	Italy	TSI	25
Finnish	Finland	FIN	31
Gujarati	Indians in Houston, TX	GIH	25
Han Chinese	Beijing, China	CHB	27
Mexican	Los Angeles, CA	MXL	17

Table 4-1: 1000 Genomes populations and their sample sizes (in number of individuals)

populations, all individuals were unrelated. Of the nearly 200 microsatellites analyzed, in this chapter we report on the analysis of 53 intergenic microsatellites, 20 dinucleotide microsatellites that overlap with exons, and six microsatellites that are candidates for selection.

Reference data set: putatively neutral, intergenic dinucleotides

We genotyped 53 autosomal CA microsatellites whose length in the human genome reference sequence ranged from 6 to 29. 5 of the 53 microsatellites were monomorphic across the entire sample of 200 individuals. For each population sample, we plotted number of alleles versus mean allele size (Figure 4-1) and observed heterozygosity versus mean allele size (Figure 4-2). Using linear regression for the number of alleles plot and logistic regression for the heterozygosity plot, we calculated best-fit lines for each population. A clear dichotomy is evident between the trend lines for African and non-African populations on both plots, with the two African populations possessing higher values for number of alleles and heterozygosity on average. In some instances, this visible difference is statistically significant. For example, although the intercepts of all best-fit lines in Figure 4-1 are not statistically different from one another, the slopes are significantly different in some cases (ANCOVA: MXL vs. YRI, $P = 0.0012$; MXL vs. LWK, $P = 0.0034$; CEU vs. YRI, $P = 0.0097$; CEU vs. LWK, $P = 0.0186$).

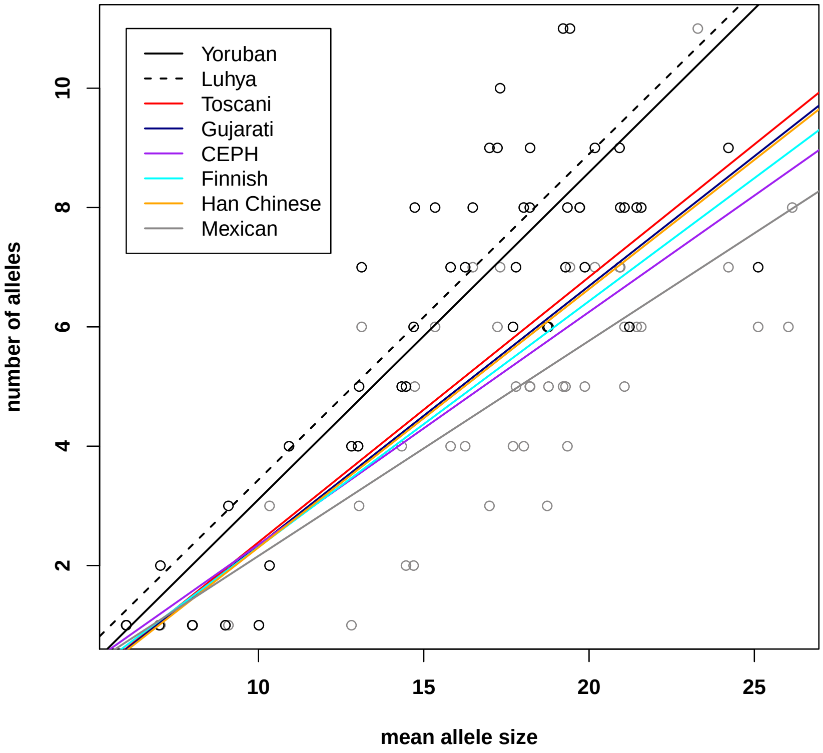

Figure 4-1: Heterozyogsity versus allele size for 53 intergenic CA microsatellites across the world.

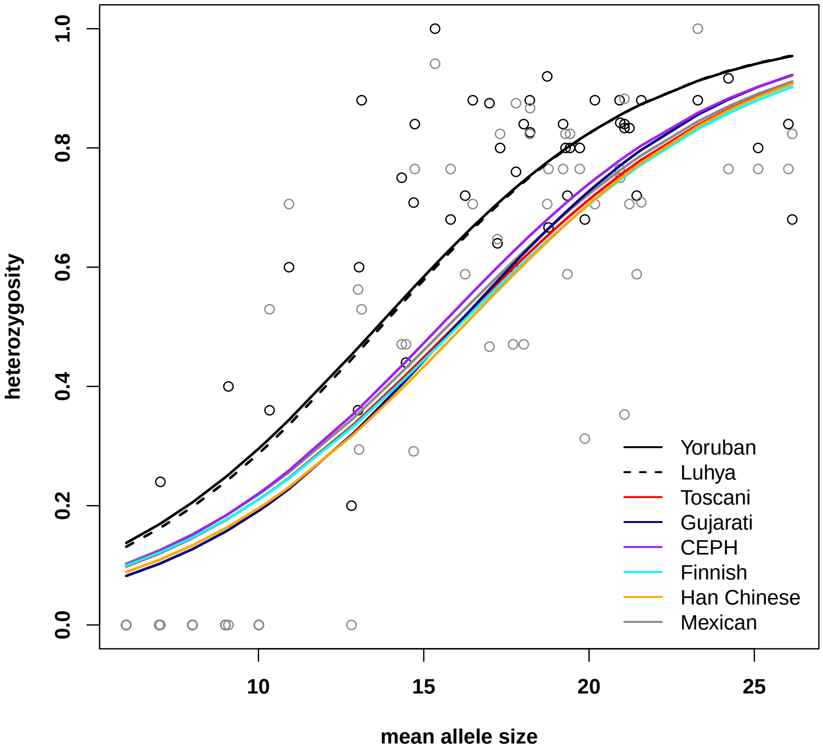

Figure 4-2: Heterozyogsity versus allele size for 53 intergenic CA microsatellites across the world.

Because we do not expect that the mutation rate of the analyzed loci has changed across the world, the differences between African and non-African populations in terms of heterozygosity and number of alleles is likely due to larger effective populations sizes (N_e) in African populations. To assess the validity of this hypothesis, for each intergenic microsatelite with heterozygosity > 0.2 in each of the eight populations sampled ($n = 41$), we calculated variance in allele size (V_{AS}). $2V_{AS}$ estimates $\theta = 4N_e\mu$, where θ is the population-scaled mutation rate and μ is microsatellite mutation rate per-locus/per-generation (Pritchard and Feldman, 1996). Assuming that μ is constant across populations, the ratio of V_{AS} between two populations yields an estimate of the ratio of their effective population sizes:

$$\frac{V_{AS_1}}{V_{AS_2}} = \frac{8N_{e_1}\mu}{8N_{e_2}\mu} = \frac{N_{e_1}}{N_{e_2}}$$

We calculated the mean and standard error of this ratio across all 41 loci for all pairwise comparisons between the sampled populations (with the exception of MXL, which is a population admixed from divergent source populations; Figure 4-3). As expected, we found that the two African populations are estimated to have effective population sizes much greater than those of non-African populations (roughly 3-to-1 ratios, Figure 4-3). This result agrees with the a recent study by McEvoy et al. (2011), who used empirical measures of linkage disequilbrium to estimate the N_e of 17 worldwide populations and found most African-to-non-African ratios of effective size were roughly between 2- and 3-fold. Moreover, we estimated the N_e ratio between the two African populations sampled – YRI and LWK – to be 1.007. McEvoy et al. (2011) also estimated near equality in size between these populations, with a ratio of 1.07. These results support the hypothesis that higher diversity at microsatellite loci in Africans is most likely due to long-term greater N_e.

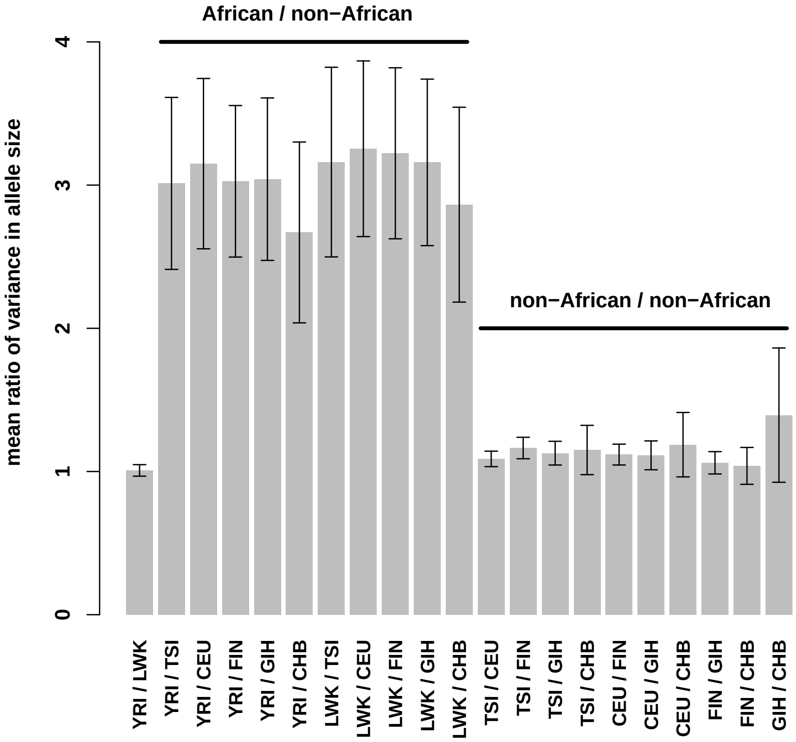

Figure 4-3: Mean ratio of variance in allele size between sampled 1000 Genomes populations, across 41 intergenic CA microsatellites. Bars are standard errors.

Rarity of coding dinucleotdies in the human genome and remarkable conservation of allele size across primates

We searched the human genome reference sequence for perfect-repeat dinucleotides that overlapped exons and were ≥ 7 in length. Only 20 dinucleotide microsatellites met these criteria, of which 18 were fully contained in exons of annotated genes (Table 4-2). Of the remaining two dinucleotide microsatellites one overlapped an intron/UTR border in the gene *RBM5* and the other was present in the putative protein-coding open reading frame

C3orf27. We used the Multiz Alignment and Conservation track of the UCSC Genome Browser (Meyer et al., 2013) to find the allele size of each orthologous microsatellite in *Pan troglodytes* (chimpanzee), *Gorilla gorilla*, *Pongo pygmaeus* (orangutan), *Nomascus leucogenys* (gibbon), *Macaca mulatta* (rhesus macaque), and *Callithrix jacchus* (marmoset) (Table 4-2).

Frameshifts resulting from the mutation of exonic, dinucleotide microsatellites should constrain the expansion of these microsatellites. Indeed, numerous details of the dinucleotide data set reveal the strength and nature of the selective constraint on dinucleotides in coding sequence. First, there are at least 105,584 perfect-repeat dinucleotide microsatellites of size $\geq$ 7 in the human genome (Payseur et al., 2011). Assuming that 2% of the human genome consists of coding sequence, the neutral expectation is that 2,112 dinucleotides should overlap with coding sequence. Thus our observation of 20 dinucleotides overlapping exons is highly unlikely ($\chi^2 = 2017$, df = 1, $P < 10^{-15}$).

Second, polymorphism is virtually non-existent within the data set. Only two of the 19 fully exonic dinucleotides were polymorphic in our sample of 1000 Genomes participants ($n > 380$ chromosomes in all cases). Of these, only one was in an annotated gene (*CCAR10)*, where two alleles out of 400 sampled were of size 6; these were found in two heterozygous individuals from the Yoruban (YRI) population. In a sample of just $n = 2$, Payseur et al. (2011) found that 12.5% of perfect-repeat dinucleotides of allele size 7 were polymorphic. Given that sample size approaches or equals 400 at all loci in the dinucleotide data set, this lack of polymorphism is striking. Apparent release of selective constraint is observed in the dinuclotide microsatellite that overlaps an exon/UTR border in the gene *RBM5*. Of the 20 dinucleotides identified, it bears the highest polymorphism and shows considerable divergence across primates.

Third, excluding the dinucleotide microsatellites in *RBM5* and *C3orf27*, 18 of 18 dinucleotides microsatellites are monomorphic across the great apes, while 16 of 18 dinucleotide repeats are conserved across primates back to the common ancestor of marmoset and the other primates included. These data suggest that all 18 dinucleotide microsatellites were present in their current state > 12 million years ago, which is the minimum estimated human-orangutan divergence time (Glazko and Nei, 2003). Moreover, 16 of the microsatellites appear to have existed in their current state > 32 million years ago, the minimum estimated human-marmoset divergence time (Glazko and Nei, 2003).

Finally, intergenic loci of similar length show very different patterns of polymorphism and divergence compared to the dinucleotides that overlap exons (Table 4-1). For example, variance in allele size across primates for the two intergenic dinuclotide of size 7 that we sampled are 2.57 and 9.48. In comparison, excluding the *RBM5* and *C3orf27* dinucleotides, the exonic dinucleotide microsatellites have an average variance in allele size of just 0.022 across the primates sampled. Although one of the sampled intergenic dinucleotides is monomorphic in our sample, divergent allele sizes were found in the other great apes (family *Hominidae*). Moreover, the other intergenic dinucleotide of reference length 7 had observed heterozygosity of 0.061 in our sample. As allele size increases beyond 7, polymorphism in our data set increases dramatically and the divergence between allele sizes in human and other primates becomes much greater. Indeed, the intergenic data suggest that an allele size of seven may represent the maximum allele size of an exonic dinucleotide that is easily tolerated by primate species. If allowed to exist at high frequencies in a population, allele sizes ≥ 8 are likely to mutate frequently, contributing greatly to genetic load.

The most remarkable coding dinucleotide was found in exon 10 of *FGFRL1*. The allele size of this microsatellite is 10, suggesting it should be subject to considerable mutational pressure. However, it was monomorphic in our sample of 400 alleles and is conserved to marmoset (though orangutan and marmoset each have one internal point mutation). *FGFRL1* is a member of the fibroblast growth factor receptor (FGFR) family. Unlike the canonical receptors of this family (*FGFR1-4*), FGFRL1 lacks the intracellular tyrosine-kinase domain that is instrumental to the signal transduction capacity of the other FGFRs (Trueb et al., 2003; Zhuang et al., 2009). However, in its place, FGFRL1 possesses a "peculiar histidine-rich motif that does not share much homology with any known protein" (Zhuang et al., 2009), which acts as a negative regulator of the Ras/Raf/Erk signaling pathway (Zhuang et al., 2011). Importantly, the *FGFRL1* microsatellite in question codes for the most highly conserved portion of this domain (Figure 4-4).

Mutations in members of the canonical *FGFR*s lead to craniosynostosis syndromes, chondrodysplasias, and cancer (Knowles, 2007; Wilkie, 2005). FGFRL1 itself is involved in kidney and bone morphogenesis in humans (Trueb, 2011). Interestingly, the first mutation discovered in *FGFRL1* was a frameshift mutation in the histidine-rich intracellular domain in a patient who exhibited a host of skeletal abnormalities, including craniosynostosis (Rieckmann et al., 2009). Also of note is the valine residue in mouse directly preceding the highlighted sequence in Figure 4-4, which is apparently also due to a frameshift mutation (Trueb et al., 2003).

The evolution of these long dinucleotide microsatellites is interesting. Their rarity seems to confirm their capacity for detrimental effect. One can imagine two categories or stages of selection related to the evolution of these rare repeats. The first category is *preventative*, in which non-triplet microsatellites are restricted from expanding to sizes associated with high mutability. Presumably the emergence of the few longer coding dinucleotides documented here resulted from stochastic escapes from this form of selection. The second category of selection might be characterized as *damage-control*, in which the monomorphism of an already-long, non-triplet microsatellite is maintained through selection against individuals carrying a mutated allele. Counter-productively, this form of selection would also act against down-mutations. Thus, once long allele size is obtained, these rare microsatellites may be trapped at their current size despite the genetic load they impart. Another possibility is that these few dinucleotides are inherently less mutable than suggested by the polymorphism and divergence data from intergenic microsatellites of similar length. In fact there is some evidence that flanking sequence does occasionally modify mutation rate (e.g., Chung et al., 2010), although the specific details of a relationship between flanking sequence and mutation rate are lacking.

We believe a more-objective, less-heuristic study of these interesting repetitive elements is warranted. The development of a microsatellite version of the standard HKA test (Hudson et al., 1987), which uses the comparison of divergence and polymorphism to infer selection, would be of interest in this regard. However, the HKA test assumes an objective measure of divergence is available. Although Goldstein et al. (1995a) and Goldstein et al. (1995b) provide a microsatellite-specific distance metric, it is unclear that genetic distance would continue to accrue along the very long branches connecting some of the primate species included in this data set. Regardless, the HKA test with its simultaneous use of polymorphism and divergence data should prove inspirational in the development of an objective microsatellite test that would allow us to approximate the probability of the data shown in Table 4-1.

gene	chromosome	motif	intragenic	heterozygosity	refseq	*Pt*	*Gg*	*Pp*	*Nl*	*Mm*	*Cj*
CCDC30	1	AG	exon	0	8	8	8	8	(8)	8	6
CACNA1E	1	AG	exon	0	7	7	7	7	7	7	7
WDR64	1	CA	exon	0	7	7	7	7	7	7	(7)
ATP6V1C2	2	AG	exon	0	7	7	7	7	7	7	7
RBM5	3	AG	exon/UTR	0.04	11	10	(12)	15	8	10	3
MAGI1	3	CT	exon	0	7	7	7	7	7	7	7
C3orf27	3	CT	exon	0.005	9	9	9	6	(9)	(5)	(14)
FIP1L1	4	AG	exon	0	7	7	7	7	7	7	7
FGFRL1	4	CA	exon	0	10	10	10	(10)	–	10	(10)
IK	5	AG	exon	0	7	7	7	7	7	7	7
PAPD7	5	CA	exon	0	7	7	7	7	(7)	(7)	7
LUC7L2	7	AG	exon	0	7	(7)	(7)	–	(7)	((7))	((7))
MKI67	10	CA	exon	0	7	7	7	7	7	(7)	(7)
CCAR1	10	AG	exon	0.01	7	7	7	7	7	7	(7)
ACIN1	14	CT	exon	0	7	7	7	7	7	7	7
SLTM	15	CT	exon	0	7	7	7	7	7	7	7
NAGPA	16	CG	exon	0	7	7	7	7	7	7	(7)
BRD7	16	CT	exon	0	7	7	7	7	6	6	6
HMGXB4	22	AG	exon	0	7	7	7	7	7	7	7
TXLNG	X	AG	exon	0	7	7	7	7	7	7	(7)
intergenic	*13*	*CA*	–	*0.061*	*7*	*8*	*4*	*12*	*(8)*	*(10)*	*13*
intergenic	*4*	*CA*	–	*0*	*7*	*8*	*11*	*9*	–	*(7)*	*(7)*
intergenic	*18*	*CA*	–	*0.01*	*8*	*7*	*6*	*(6)*	–	*6*	–
intergenic	*6*	*CA*	–	*0.23*	*9*	*(13)*	*(10)*	*7*	*7*	*(11)*	*7*
intergenic	*2*	*CA*	–	*0.43*	*10*	*12*	*6*	*(25)*	*22*	*((21))*	*5*

Table 4-2: Coding dinucleotide microsatellites of allele size ≥ 7 in the human genome reference sequence (refseq). For comparison, five intergenic dinucleotides are also shown. Reference sequence lengths are also reported for the following species: *Pt*: *Pan troglodytes*, chimpanzee; *Gg*: *Gorilla gorilla*; *Pp*: *Pongo pygmaeus*, orangutan; *Nl*: *Nomascus leucogenys*, gibbon; *Mm*: *Macaca mulatta*, rhesus macaque; *Cj*: *Callithrix jacchus*, marmoset. Lengths in parentheses are imperfect repeats, where the number of parenthetical sets equals the number of internal imperfections.

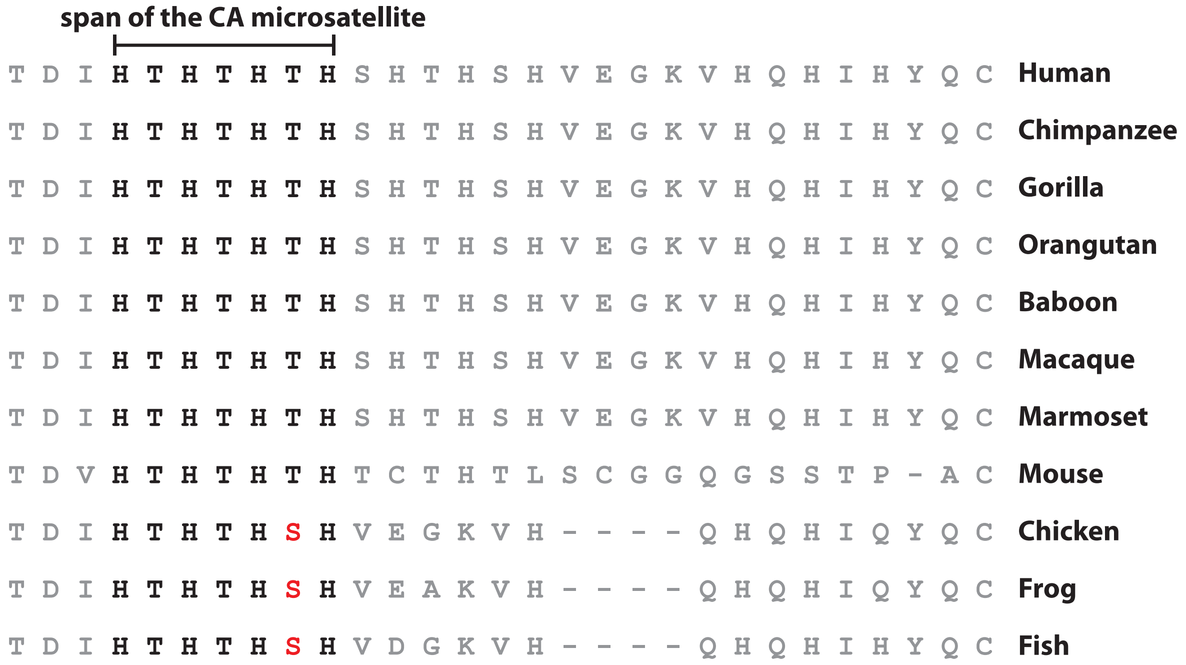
span of the CA microsatellite
T D I H T H T H T H S H T H S H V E G K V H Q H I H Y Q C Human
T D I H T H T H T H S H T H S H V E G K V H Q H I H Y Q C Chimpanzee
T D I H T H T H T H S H T H S H V E G K V H Q H I H Y Q C Gorilla
T D I H T H T H T H S H T H S H V E G K V H Q H I H Y Q C Orangutan
T D I H T H T H T H S H T H S H V E G K V H Q H I H Y Q C Baboon
T D I H T H T H T H S H T H S H V E G K V H Q H I H Y Q C Macaque
T D I H T H T H T H S H T H S H V E G K V H Q H I H Y Q C Marmoset
T D V H T H T H T H T C T H T L S C G G Q G S S T P - A C Mouse
T D I H T H T H S H V E G K V H - - - - Q H Q H I Q Y Q C Chicken
T D I H T H T H S H V E A K V H - - - - Q H Q H I Q Y Q C Frog
T D I H T H T H S H V D G K V H - - - - Q H Q H I H Y Q C Fish

Figure 4-4 Amino acid conservation in the C-terminus of FGFRL1. The $\texttt{CA}_{10}$ microsatellite conserved across primates codes for a highly conserved histidine-rich intracellular domain. The cysteine residue at the right of the alignment is the terminal amino acid of FGFRL1.

Model selection and parameter inference

Bimodal models of the genotypic fitness surface

A number of the candidate loci for selection genotyped in 1000 Genomes populations exhibited striking bimodal distributions. None of the four models of selection presented in Chapter 2 are capable of producing bimodal fitness surfaces. Thus, we thought it prudent to develop bimodal fitness models. This does not imply that bimodal allele frequency distributions require selection to occur. For example, a non-selective explanation for bimodal frequency distributions is population structure. Nevertheless, the frequency of bimodal distributions in empirical data suggest it is important to include models that are capable of capturing this aspect of the data.

The two bimodal models of microsatellite fitness are mixtures of two normal distributions. Each normal distribution is described by two parameters: mean (in terms of allele size) and variance. The mixture of the two normal distributions is then modified in two ways. First the parameter rel controls the relative height of each mode (i.e., each distribution) and the parameter sel is used to rescale the y-axis (fitness axis) so that its maximum values is 1 and the minimum fitness is equal to $1 - sel$. Figure 4-5 demonstrates construction of the final genotypic fitness surface for the *bimodal-additive* and *bimodal-recessive* models.

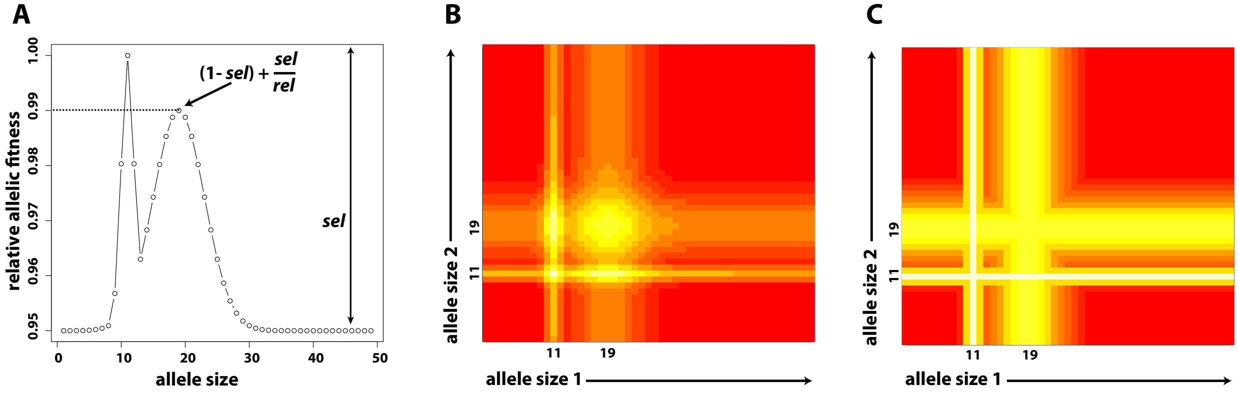

Figure 4-5 The bimodal-additive and bimodal-recessive models. (A) A mixture model of two Normal distributions, $N(11, 1)$ and $N(19, 16)$, is the starting point for computing allelic fitnesses. The density of this mixture distribution is converted into relative fitness using the parameters rel and sel. First, density is rescaled such that the lowest density on the graph corresponds to a fitness of $1 - sel$. In this example, $sel = 0.05$, so the lowest fitness on the graph equals 0.95. Next, rel is used to adjust the relative fitnesses of the two modes. Together with sel, rel specifies the fitness of the first mode relative to the second mode. In this example, $rel = 1.25$ and the relative height of the second mode equals $(1 - sel) + \frac{sel}{rel} = 0.99$. (B) The genotypic fitness surface corresponding to the allelic fitness surface in (A) under the bimodal-additive model. This model specifies that relative genotypic fitnesses are half the sum of the fitnesses of their two component alleles. (C) The genotypic fitness surface corresponding to the allelic fitness surface in (A) under the bimodal-recessive model. This model specifies that relative genotypic fitnesses are equal to the greater fitness of the two component alleles. In (B) and (C), white corresponds to a fitness of one and red to a fitness of $1 - sel$.

Inference protocol

Due to the complexity of the selection models used here, we adopted an approximate Bayesian computation (ABC) scheme for model selection and parameter inference (Beaumont et al., 2002; Beaumont, 2008). The intuitive principle underlying this form of simulation-based inference is that models and parameter values that generate simulated data sets similar to the observed data set are preferable to those that do not. We now specify our method of of inference in some detail.

1. PREPARATION: We calculated observed summary statistics from the observed allele frequency distribution: (1) number of alleles, na; (2) heterozygosity, hz; (3) mean allele size, $\bar{x}$; (4) minimum allele size, min; (5) maximum allele size, max; (6) variance in allele size v_{as}, and; (7) skewness of allele size s_{as}. We also calculated an eighth summary statistic, $|AFD|$, which is the absolute difference between the observed frequency distribution and a simulated frequency distribution. $|AFD|$ takes a value of zero when simulated and observed frequencies match perfectly and a maximum value of two, when there is no overlap between the alleles present in the observed and simulated data sets. $|AFD| = 0$ was the observed value of this statistic. Next, we determined uniform prior distributions on the mutational parameters ϕ, ψ, and γ (Chapter 2). We assumed that $c = 0.001$ and $g = 0.95$ – i.e., minimal contraction bias and a 95% chance of single-step mutation. We then determined uniform priors for the selection parameters specific to each of the models. For the four selection models detailed in Chapter 2, the parameters included x, δ, g_l, and g_u. For the bimodal models introduced in the previous section, the parameters were m_1 and m_2 (the position of the two modes in terms of allele size), sd_1 and sd_2 (the standard deviation of each Normal distribution in the mixture distribution), rel, and sel. Unless otherwise noted, we assumed effective population size $N_e = 10000$ (Schaffner et al., 2005). The

duration of the selective event, t, was treated as a nuisance parameter and pulled from a uniform prior of [200,3000] generations.

2. CALIBRATION: We ran 250,000 simulations per model, where the models were: additive, multiplicative, dominant, recessive, neutral (Chapter 2); bimodal-additive and bimodal-recessive (introduced in the previous section). Simulations were performed using the algorithm described in Chapter 2. Parameter values for each simulation replicate were drawn from the model-specific prior distributions. For the starting allele frequency distribution of each simulation, we used a skew normal distribution whose mean, dispersion, and skew were randomly chosen from priors on each of these parameters. For each simulation, we recorded the eight summary statistics and parameter values simulated, which included model as a categorical parameter.

3. MODEL SELECTION: We used the R package *abc* (Csillery et al., 2012) and its function *postpr* to calculate posterior probabilities for each of the seven models simulated. This function treats the parameter index as the response variable in a logistic regression analysis (Fagundes et al., 2007; Beaumont, 2008). In this extension of standard approximate Bayesian computation (ABC), the distance between observed and simulated summary statistics is used to identify the best simulations. We used a tolerance of 0.0025, meaning that the best ~ 500 simulations (in terms of low observed-to-simulated distances) were used in the regression.

4. PARAMETER ESTIMATION: In order to estimate parameter values, we ran additional simulations, which assumed the model that obtained the most posterior support in the previous step.

    a) *GRIN2B*: For this locus only, we ran a more sophisticated ABC protocol, in which we used non-uniform priors and performed simulations in a Markov

Chain Monte Carlo framework (ABC-MCMC; Marjoram et al., 2003; Wegmann et al., 2009). We used information from simulations of the chosen model in the calibration phase to identify informative lognormal, normal, or uniform priors for each parameter. We then ran 10 non-communicating chains of 1 million iterations each, where the starting parameter values of each chain were chosen from prior distributions. For each parameter value, the proposed value for the next iteration of the chain was chosen from a normal distribution centered on the current value of the parameter with variance equal to 0.05 variance of the parameter's prior distribution. Proposed parameter values in the chain were accepted if both (1) the distance between the observed and simulated summary statistics was less than tolerance value $\epsilon = 0.2$, and (2) a randomly generated number on the interval (0,1) was less than the probability $h = (1, \frac{\pi(\theta')}{\pi(\theta)})$, where π is the prior distribution of parameter θ, and θ' is the proposed parameter value (algorithm F in Marjoram et al. (2003)). In the context of ABC inference, the MCMC step can simply be viewed as a more efficient means of collecting successful simulations than a blind rejection method (Wegmann et al., 2009). Thus, each time a proposed set of parameter values was accepted, we recorded these values along with the summary statistics calculated from the data set simulated from these parameter values. Results from all 10 chains were pooled into a result set of 408,105 entries (roughly 8.1% accepted parameter proposals). We used the function "abc" in the R package *abc* to estimate parameter values via weighted local linear regression (Beaumont et al., 2002). We assumed a tolerance of 0.001 so that results from roughly 500 simulations were used in parameter estimation.

b) To estimate parameter values for the other five candidate microsatellites, we ran an additional 2 million simulations assuming the model chosen in the calibration phase. All priors were uniform; however, we decreased the range of prior distributions based on results from the calibration phase. The results from all simulations were pooled and we used the "abc" function on this result set to estimate parameters using local linear regression.

Using cross-validation to assess the reliability of model choice

To assess the ability of our method to distinguish between the neutral model and six selection models, we used leave-one-out cross-validation. First, we simulated 100,000 data sets for each model. The parameters used in each simulation were drawn from broad, uniform distributions. For each round of cross-validation, a random data set was selected and treated as the observed data set. The remaining 699,999 data sets were then treated as the simulated data sets and model selection was performed logistic regression as specified above, with tolerance equal to 0.05. We treated the model with the greatest posterior probability as the chosen model. For each model, 200 separate data sets were tested as the observed data set.

We used the results of cross-validation analyses to construct a confusion matrix (Table 4-3 Hastie et al., 2009). In this matrix, rows correspond to the true model and columns to the model chosen. Thus, off-diagonal elements tally instances of misclassification.

Results of the cross-validation analysis indicate that neutral data sets are misclassified as targets of selection 22.5% of the time (bottom row, Table 4-3). Thus, choice of a non-neutral model must be treated with some caution. Nearly all misclassifications of neutral simulations were cases where the bimodal-additive or bimodal-recessive model was chosen. This can be explained by the fact that, in general, these two models cause increases in variance that can mimic neutral distributions. The four models detailed in Chapter 2 tend to decrease variance in allele size drastically, often producing allele frequency distributions

	add	**mult**	**dom**	**rec**	**bma**	**bmr**	**neu**
add	**25**	56	20	35	31	20	13
mult	26	**90**	22	19	27	11	5
dom	20	41	**45**	27	22	32	13
rec	15	27	17	**89**	12	26	14
bma	3	14	11	12	**80**	49	31
bmr	0	1	3	9	23	**138**	26
neu	0	0	1	2	23	19	**155**

Table 4-3: Confusion matrix from cross-validation analysis of the seven models tested. *add* = additive, *mult* = multiplicative, *dom* = dominant, *rec* = recessive, *bma* = bimodal-additive, *bmr* = bimodal-recessive, *neu* = neutral

that are very different then neutral models when mutation rate is high.

Table 4-3 also indicates that our method does a poor job of choosing between models of selection. There are two alternative explanations for this imprecision. First, distinct sets of parameter values can generate similar fitness surfaces under different selective models. In these cases, it can be difficult to choose the correct selective model. Second, the selective fitness models capture biological reality in different ways. Thus, split support for two or more selective models may suggest that each model captures different aspects of the data.

The answer to the simpler question, "Is it neutral?" is particularly important to biologists who use microsatellites as assumed neutral markers. We can assess the ability of our method to answer this question by collapsing the confusion matrix into a binary choice between selection and neutrality. In this context 102 of 1200 data sets generated under a model of selection were misclassified as neutral (a 8.5% false negative rate), while 45 of 200 neutral data sets were classified as selected (a 22.5% false positive rate). In the context of choosing neutral genetic markers, the false negative rate is of greater importance. While a false positive result will reject a marker for use, a false negative result will lead to the use of a non-neutral marker. The relatively low false negative rate of our method suggests it could be quite useful in the context of marker selection.

Improving model selection results will require the use of additional data and/or additional summary statistics. Cross-validation analysis can help assess the advantage to adding another summary statistic. For example, we ran an identical analysis in which the kurtosis of the allele frequency distribution was added as a ninth summary statistic. Inclusion of this statistic decreased the accuracy of model selection, indicating it should not be included in data analysis.

Using posterior predictive checks to assess the fit of a specific model choice

Posterior predictive distributions (PPDs) provide an important means of assessing the fit of a chosen model and its parameter estimates (Hoff, 2009). To generate a PPD for a particular model, we pull a random set of parameters from this joint posterior distribution of all parameters. A simulation is run using this parameter set and the resulting data are summarized and recorded. The PPD is simply the distribution of these summary statistics across many independent simulations. If the fit of the model and parameter estimates is good, observed values of the summary statistics should lie within a high density regions of the PPD. In short, a posterior predictive check asks whether the estimated model and parameters are likely to generate a data set similar to the original, empirical data.

We performed posterior predictive checks for the *GRIN2B* locus only. In order to do this, we estimated parameters using the ABC-MCMC approach detailed above for the neutral, additive, multiplicative, dominant, and recessive models. For each model, we approximated the form of the joint posterior distribution of all parameters with a multivariate skew normal distribution and drew 1000 parameter sets from this distribution using the R package *sn* (Azzalini, 2011). We used each of these parameter sets to generate 1000 data sets for each model, which we then summarized using the same summary statistics as above. We then compared observed values of summary statistics to their PPDs, and asked whether the best fitting PPDs were those associated with the model chosen during the model selection phase of our method.

Testing candidate microsatellites for selection

Allele frequency distributions and genomic context of the candidate microsatellites

For each of the six candidate microsatellites, we next present the observed allele frequency distributions for each sampled population as well as a graph that illustrates the genomic position of the candidate microsatellite (red bar) relative to its proximate gene (Figures 4-6 through 4-11). The scale of each gene map is noted and exons are represented by thicker bars.

Glutamate receptor, ionotropic, N-methyl D-aspartate 2B (*GRIN2B*)

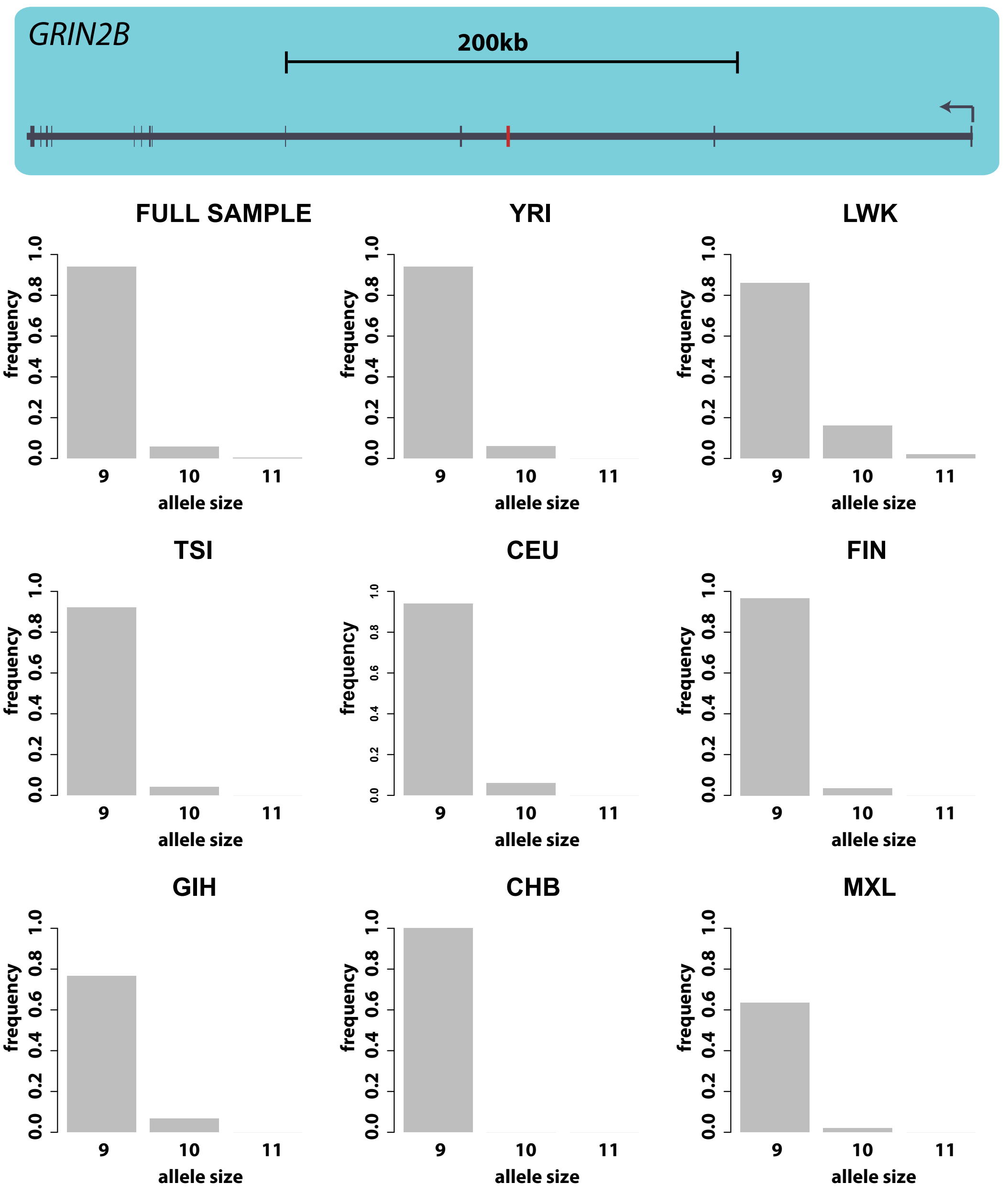

Figure 4-6

Paired box 3 *PAX3*

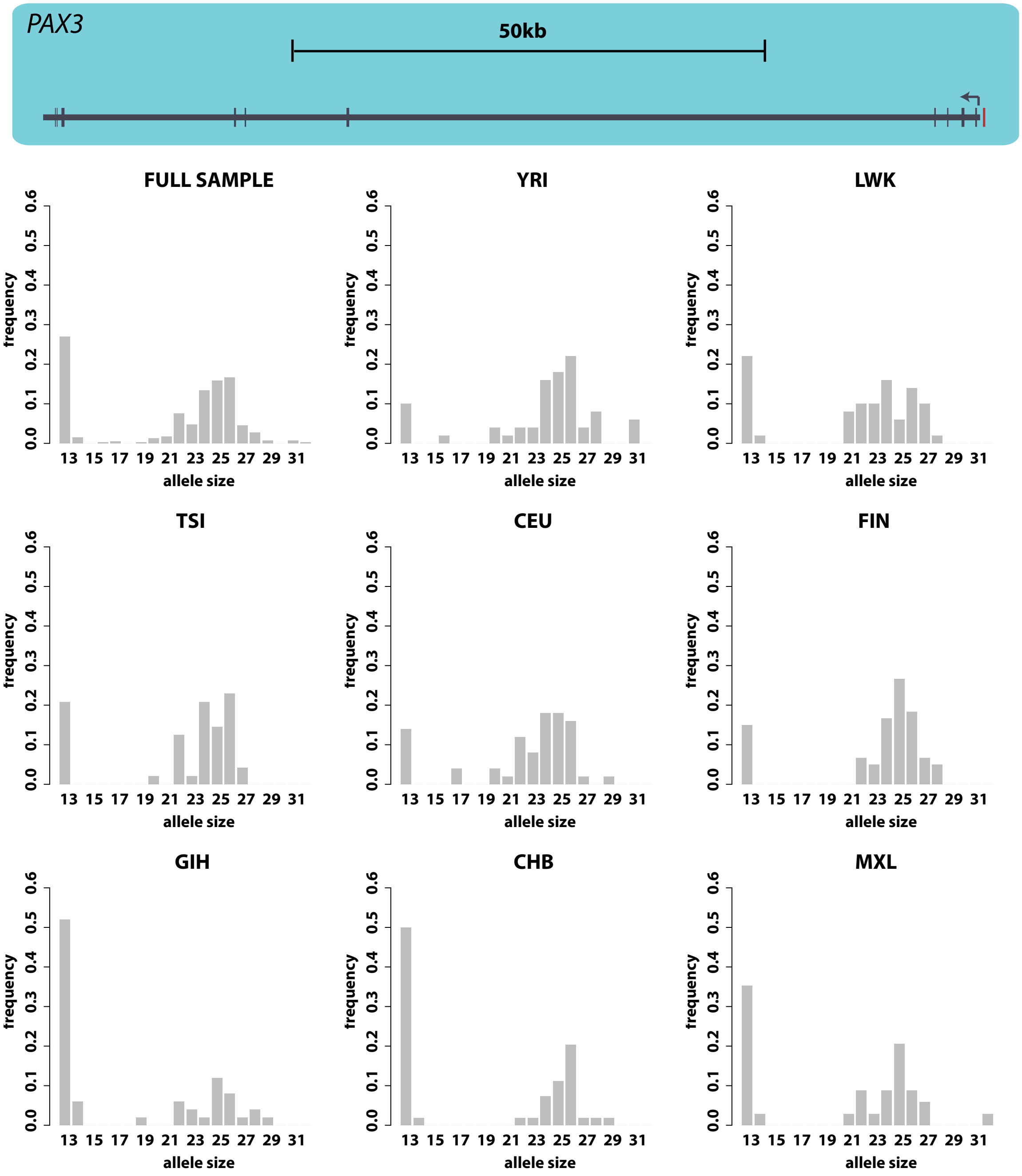

Figure 4-7

Sex hormone-binding globulin (*SHBG*)

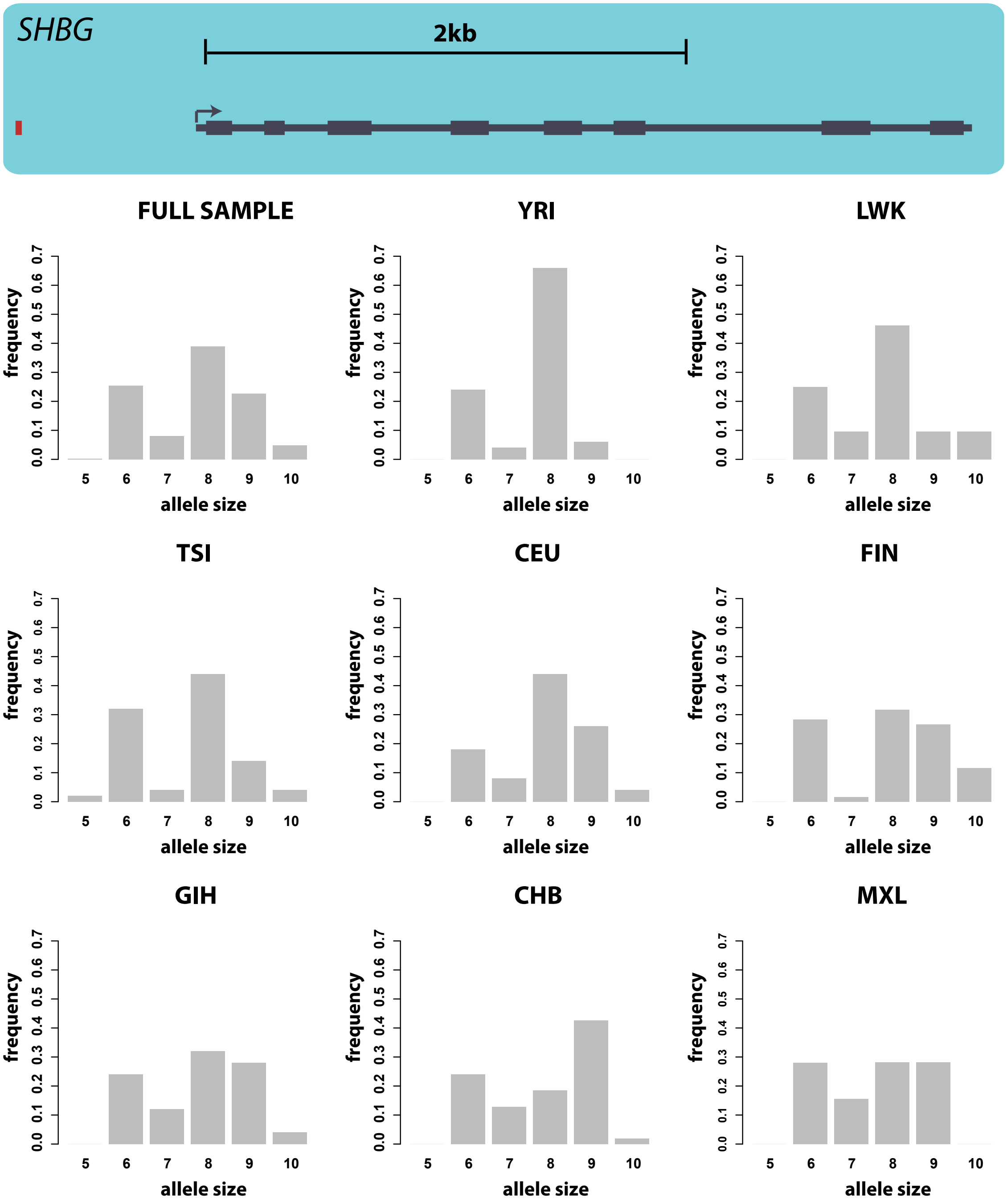

Figure 4-8

Epidermal Growth Factor Receptor (*EGFR*)

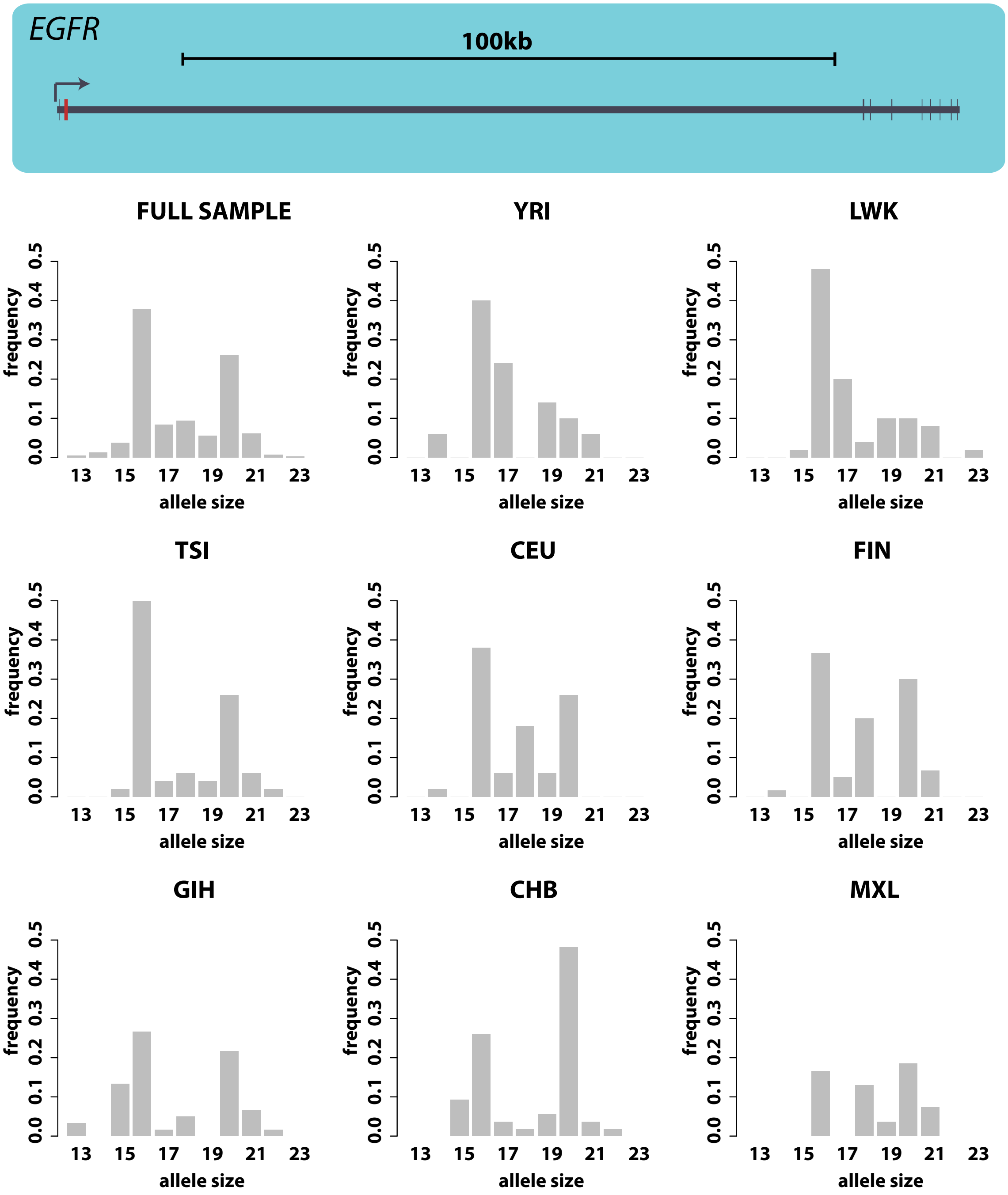

Figure 4-9

Aldose reductase *AKR1B1*

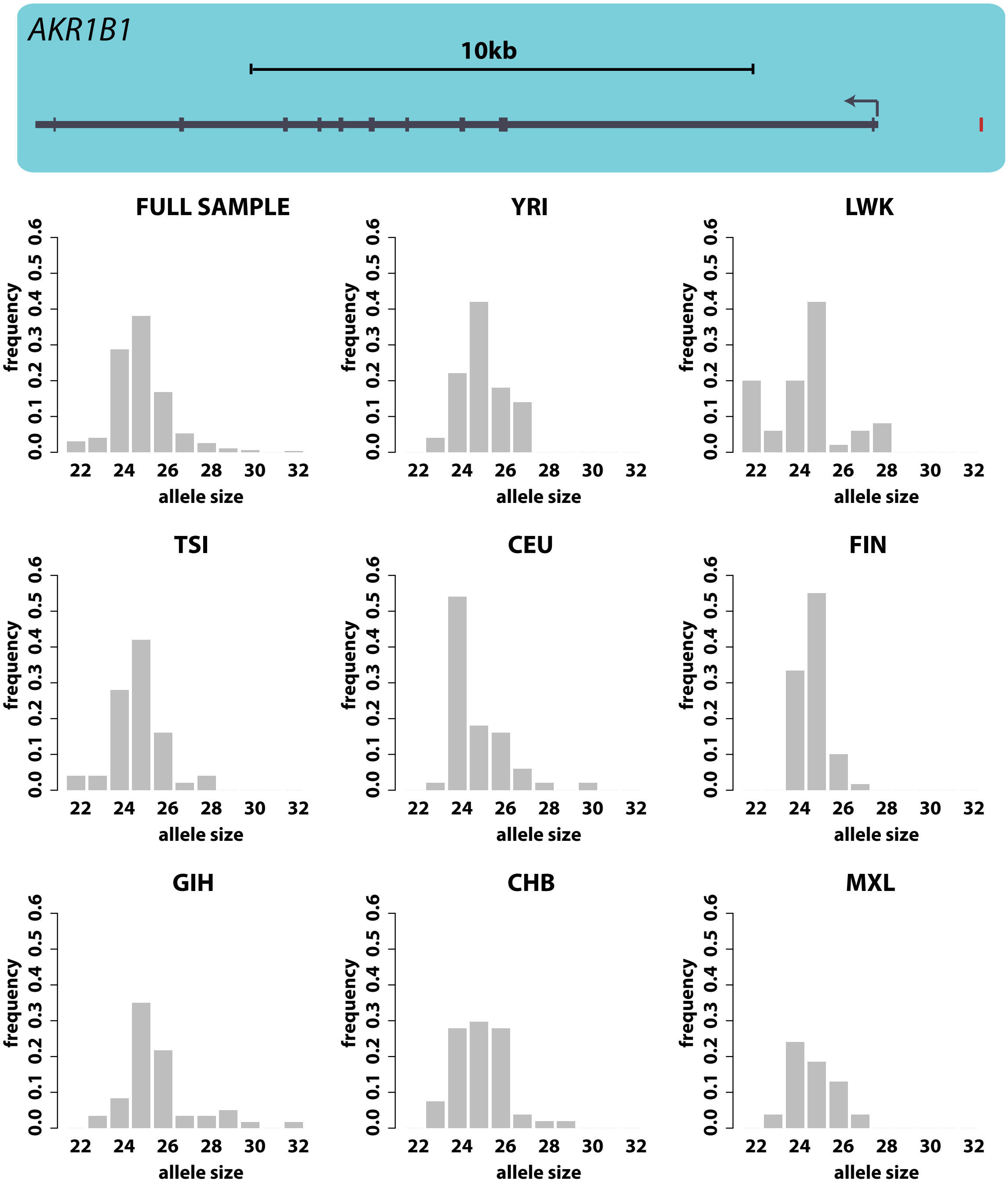

Figure 4-10

Matrix metallopeptidase 9 (*MMP9*)

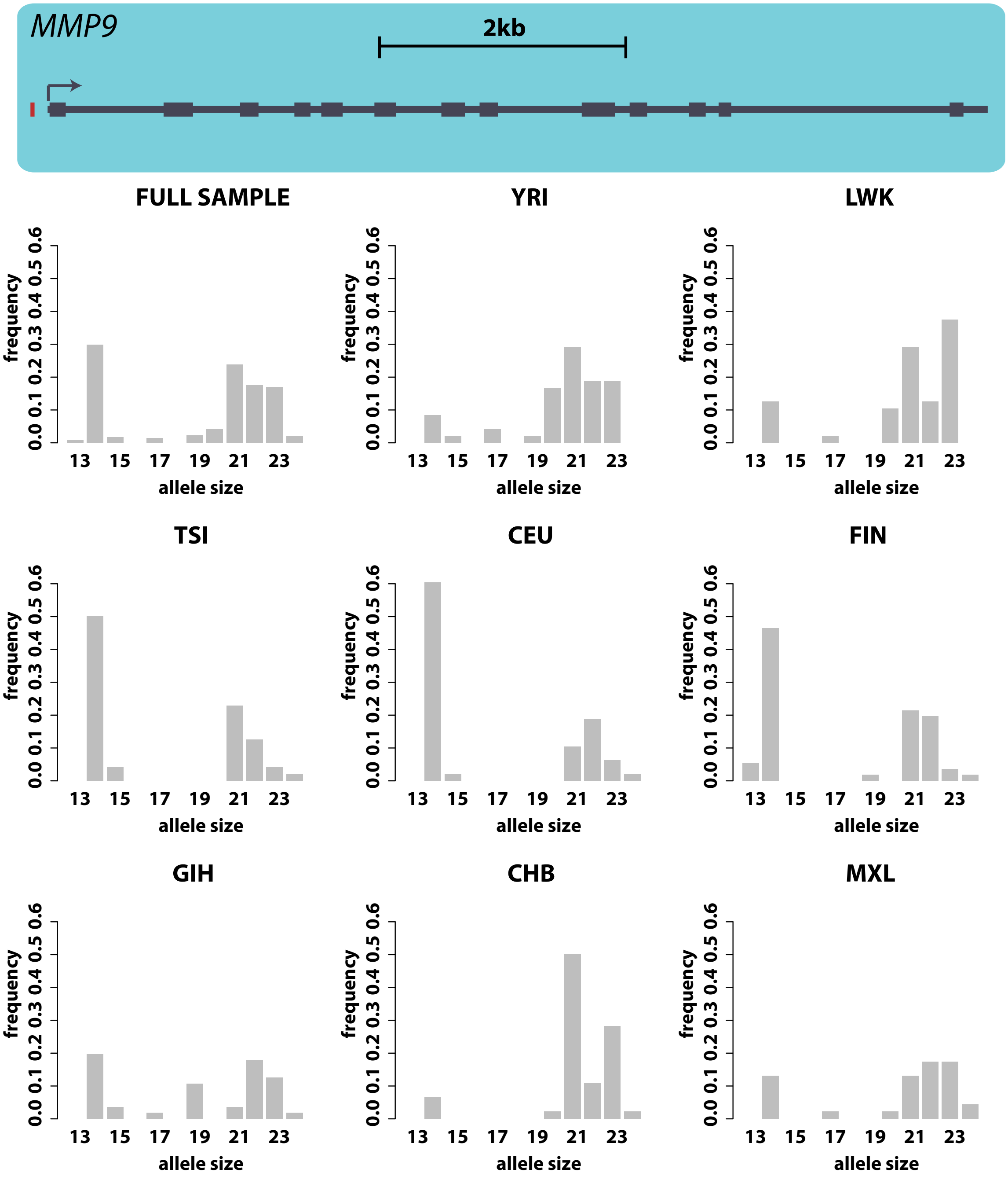

Figure 4-11

Extended analysis of the intronic *GRIN2B* microsatellite

Biological context

GRIN2B codes for a subunit of the glutamate-gated ion channels known as *N*-methyl-D-asparate receptors (NMDARs), which are found in the postsynaptic density of neurons in the central nervous system (Lau and Zukin, 2007). NMDARs are heterodimers generally comprising two NR1 and two NR2 subunits. While the identity of the NR1 subunits remains the same throughout human life, the identity of the NR2 subunits changes during early development. In the transition from prenatal to postnatal development, the NR2 subunits transition from primarily NR2B subunits (coded for by *GRIN2B*) to NR2A subunits (coded for by *GRIN2A*) in a process known as the NR2A-NR2B developmental switch (Sheng et al., 1994). Although NR2A receptors become predominant in the postnatal brain, NR2B subunits still exist and the ratio of NR2A to NR2B subunits is an important determinant of synaptogenesis as well as synaptic motility and plasticity (Barria and Malinow, 2005; Gambrill and Barria, 2011). In general, *GRIN2B* expression is important to learning and memory. In mice, overexpression of *GRIN2B* in the forebrain enhances memory and learning ability (Tang et al., 1999), while deficient expression of *GRIN2B* leads to spatial learning deficits (Sakimura et al., 1995). In humans, multiple genome-wide association studies link *GRIN2B* to deficits in memory formation as well as dyslexia (e.g., Ludwig et al., 2010), and in a recent study examining the effects of *GRIN2A* and *GRIN2B* polymorhpisms, the authors conclude,

> These data ... strongly suggest that any disturbance in the number and/or composition of NMDA receptors has profound effects on neuronal development and activity in humans (Endele et al., 2010).

Analysis, results, and implications

Unlike the other candidate microsatellites analyzed here, no studies have been performed to test for a correlation between allele size of the ATG repeat in the second intron of *GRIN2B* and its expression. Instead, *GRIN2B* became a candidate due to results from the lab of collaborator Jim Weber. In an initial sample of 368 long microsatellites in 143 individuals, Weber and colleagues found six trinucleotide microsatellites that were nearly monomorphic and of allele size > 8. One of these six loci is the *GRIN2B* locus studied here. Because trinucleotide repeats of this size were rarely monomorphic in their extensive microsatellite data sets, the Weber group typed the six low variation microsatellites in a much larger sample of nearly 1000 African-, Mexican-, and European-Americans. In this sample of nearly 2,000 chromosomes the frequencies of allele sizes 9, 10, and 11 at the *GRIN2B* locus were 0.955, 0.043, and 0.003, respectively. In our sample of 1000 Genomes individuals, we observed nearly identical allele frequencies (Figure 4-6) in all eight populations sampled.

Due to lack of inter-population variation at the locus, we pooled alleles from the Weber data set with ours and analyzed this pooled data set. We found overwhelming support for the dominant model (Table 4-4). Due to the large sample size and the power of the ABC-MCMC method, we were able to recover meaningful estimates of important mutation and selection parameters (Figure 4-12, Table 4-5). The results point to a locus whose low variability is maintained by a combination of selection and low mutation. Mode estimates of mutation parameters ϕ and ψ indicate that mutation rate for alleles of sizes 9 and 10 is $< 10^{-5}$. Importantly, despite these very low estimates of mutation rate for alleles of this size, the neutral model received no support. Instead, estimated parameter values, including $\delta = -0.063$, specify strong selection against any genotype that is hetero- or homozygous for an allele of size < 9. In addition, the estimated negative value of parameter g_u specifies that allelic fitness declines linearly with allele size > 9. In summary, our results suggest that although low mutation rate reduces the production of deleterious alleles, it is necessary to invoke selection against alleles longer and shorter than 9 in order to explain the observed allele frequency distribution.

To discount the possibility that reduced variance in allele size at the *GRIN2B* locus is due to selection on linked variation, we tested the significance of the summary statistics focused on in Chapter 3. First, we calculated empirical distributions of each statistic conditioned on the number of segregating sites as detailed in Chapter 3. Unlike the *HSD11B2* locus analyzed in Chapter 3, however, the 2Mb sequence flanking the *GRIN2B* microsatellite is characterized by a higher-than-average recombination rate (Kong et al., 2010) of roughly 1.4 cM/Mb. We therefore used a prior of 1-2cM/Mb for the simulations used to compute the empirical distribution. In Chapter 3, we found that the number of haplotypes, K (when conditioned on S), possessed reasonable power to mark selection on microsatellites. In particular, the combination of low K and intermediate S was largely specific to selection on microsatellites. To better visualize evidence of this signal, we plot $S - K$ for the English (GBR) population in the left panel of Figure 4-13 as an example. An obvious peak in $S - K$ is evident just downstream of the position of the *GRIN2B* microsatellite. In fact, the two windows comprising the apex of this peak have significant values of K ($P < 0.001$). Similar peaks at the same position are present in all eleven 1000 Genomes populations for which sequence data are currently available, although they are not always statistically significant. None of the other five summary statistics had significant values in any population for any of the 10kb windows flanking the *GRIN2B* microsatellite. As an example, the right panel of Figure 4-13 shows the consistently flat value of Tajima's D across the same region in the GBR population.

We used posterior predictive checks to assess the fit of the dominant model and the estimated parameter values. We visualize the results of these posterior predictive checks by showing joint PPDs for two pairs of summary statistics – skew and heterozygosity

(Figure 4-14A) and the frequencies of allele sizes 9 and 10 (Figure 4-14B). As shown in these graphs, the dominant model clearly outperforms the other four models tested. By far the worst PPD for these two pairs of statistics was generated by simulation from the posterior distributions estimated for the neutral model. The neutral model was fundamentally unable to recapitulate the highly unusual *GRIN2B* data set, regardless of mutational parameter values.

The uniformity of distribution at the *GRIN2B* microsatellite across the world suggests continuing selective pressure that is cosmopolitan in nature. Our results suggest that mutation by itself provides insufficient explanation of the data. In particular, the complete absence of size 8 alleles across the world is difficult to explain using anomalous mutation rate alone. Even if mutation is extremely low at the locus, we still expect it to be symmetrical, meaning that mutation of the abundant size 9 allele should produce size 8 alleles (unobserved) as often as size 10 alleles (observed). Their absence suggests selective pressure against alleles of this size, which our inference supports.

We hypothesize that the *GRIN2B* microsatellite modulates expression of the gene. As related above, results from studies in mice and humans both stress the importance of *GRIN2B* expression to learning and memory. Thus, studies that investigate the correlation between allele size of the *GRIN2B* microsatellite and gene expression are warranted.

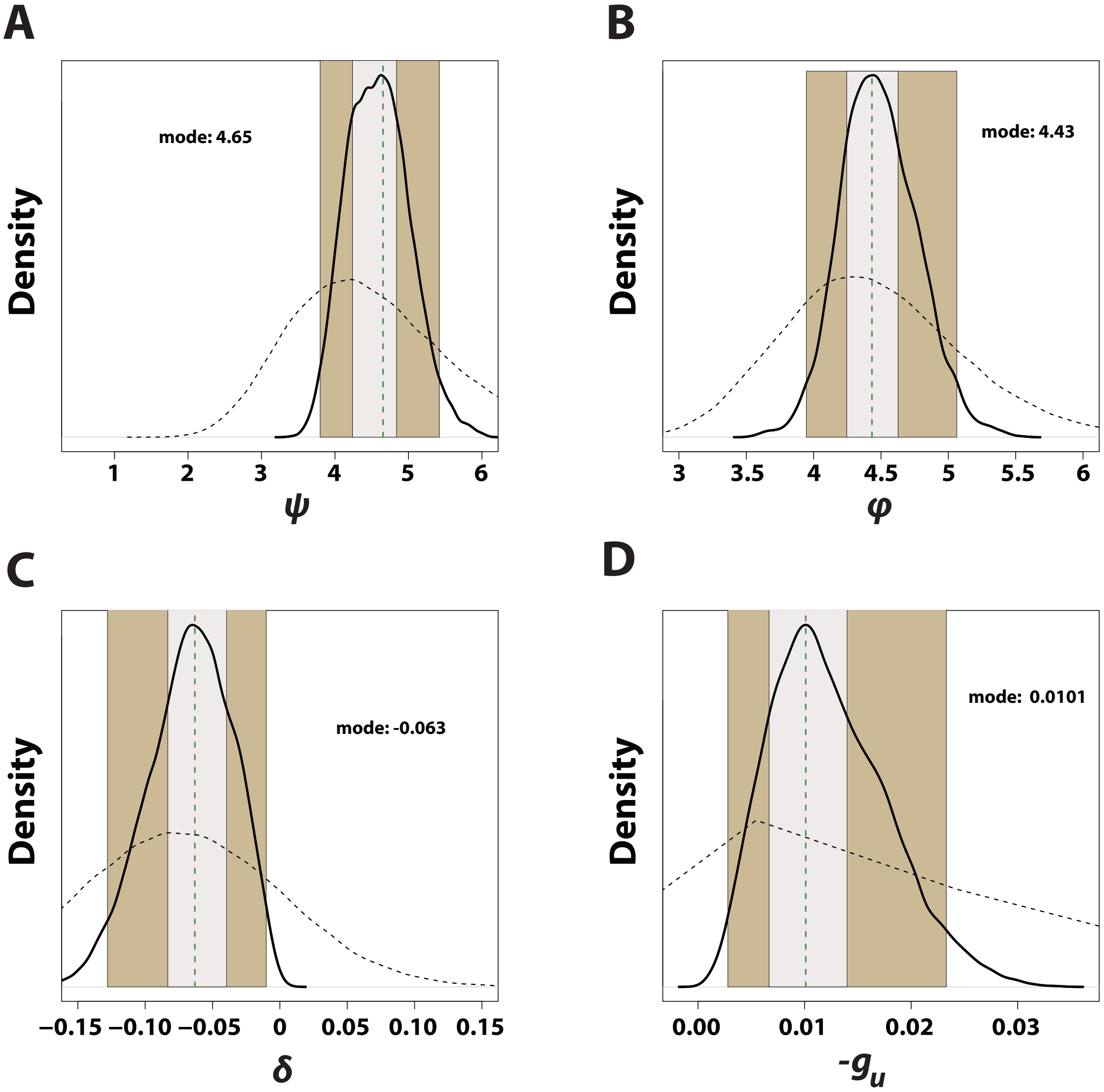

Figure 4-12 Posterior distributions from the analysis of the *GRIN2B* locus. (A) Mutation parameter ψ. (B) Mutation parameter ϕ. (C) Selection parameter δ. (D). Selection parameter g_u. Solid black lines = posterior distribution following local linear regression; dashed lines = prior distribution; light tan box = 25-75% credible interval; dark tan box = 2.5-97.5% credible interval.

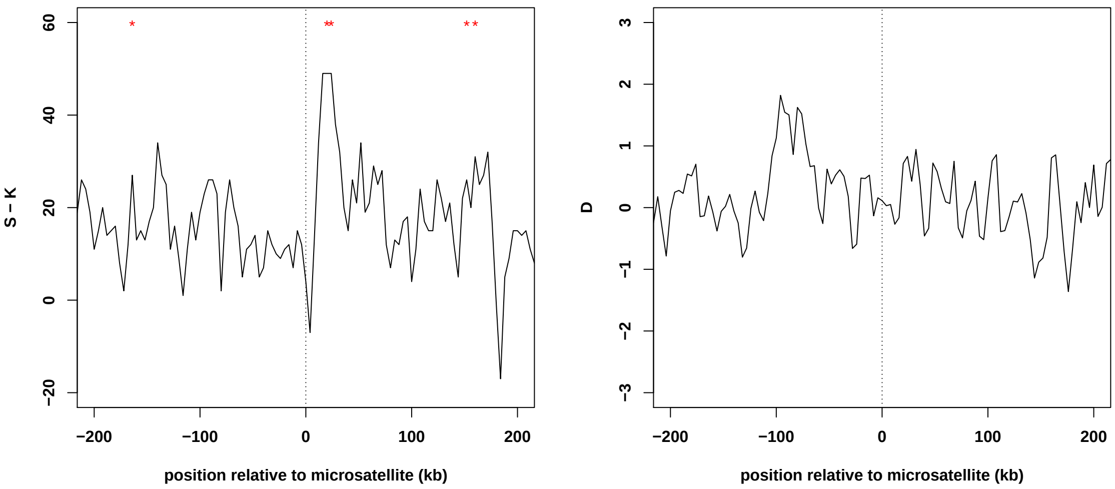

Figure 4-13 $S - K$ and D in the English (GBR) population sample for the 400kb sequence flanking the analyzed *GRIN2B* microsatellite. Red asterisks indicate significant windows for the K statistic. No windows were significant for D.

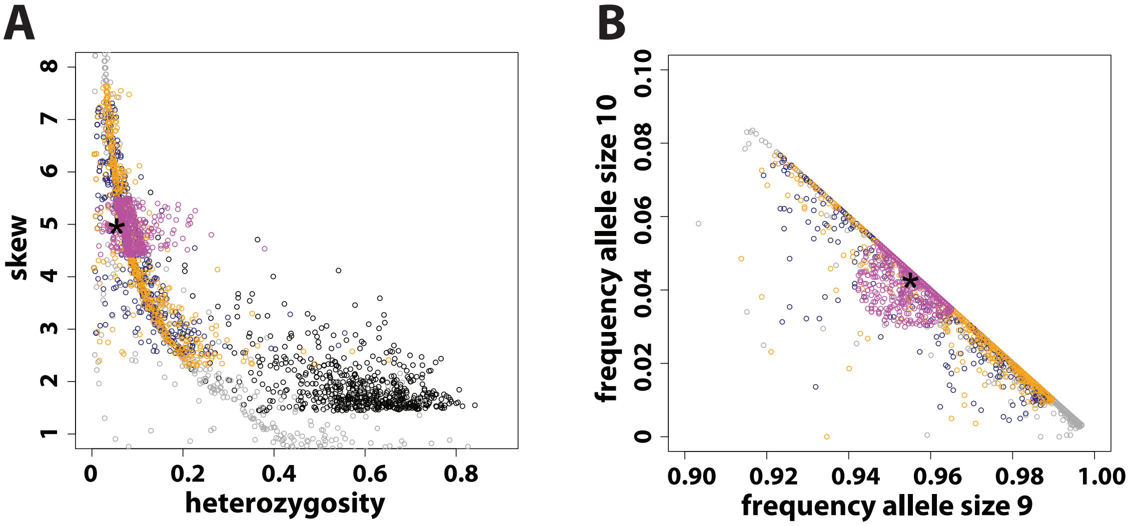

Figure 4-14 Random samples from joint posterior predictive distributions (PPDs) for the dominant (magenta), recessive (gray), additive (orange), multiplicative (navy), and neutral (black) models demonstrate the superior fit of the dominant model. (A) Joint PPD of skew and heterozygosity across posterior predictive simulations. (B) Joint PPD of the allele frequencies for sizes 9 and 10. Black asterisks mark the observed values of these joint statistics. For clarity, only the 10% of simulations with the best match to the observed values for each model are shown. Note that the results from the best neutral simulations are outside the ranges graphed in (B).

Analysis of the other five candidate genes

Paired box 3 (*PAX3*)

PAX3 is a transcription factor that plays an important role in fetal development and is associated with the diseases Waardenburg syndrome and alveolar rhabdomyosarcoma (Macina et al., 1995). Although alveolar rhabdomyosarcoma is caused by a translocation that includes a portion of *PAX3*, Waardenburg syndrome seems to be caused by differences in gene dosage, as a wide variety of mutations exhibit similar Waardenburg phenotypes (Tassabehji et al., 1994). In addition, Baldwin et al. (1995) found that although numerous mutations within *PAX3* lead to Waardenburg syndrome, deletion of the entire gene produced similar phenotypes as individual point mutations. In this context, it is interesting that the length of the CA promoter microsatellite has been shown to correlate with gene expression (Okladnova et al., 1999).

Regardless, the worldwide distribution of the *PAX3* promoter microsatellite is striking (Figure 4-7) and highly similar to the distribution recovered by Okladnova et al. (1999). In every sampled population, a bimodal allele frequency distribution was observed. Particularly anomalous is the high frequency of an isolated size 13 allele. We found high support for bimodal-additive model at the *PAX3* locus ($P(M \mid S) = 0.965$), with negligible support for the recessive and neutral models (Table 4-4). Estimates of mutation and selection parameters were generally tight (Figure 4-15, Table 4-5).

The mode estimates of the selection parameters produce a complex genotypic fitness surface on which a 13/13 genotype has the greatest relative fitness (Figure 4-16). Indeed any 13/- genotype possesses relatively high fitness. However, nearby genotypes that do not carry a size 13 allele (e.g., 14/15 or 12/12) have much lower fitness. Indeed, the observed distribution of haplotypes at the *PAX3* confirms the absence of genotypes such as 14/15 across the world (Appendix, Figure A-4-1). Although it appears that allele frequencies are currently at a worldwide equilibrium, the estimated fitness surface implies that the path taken to this equilibrium may have incurred significant selective cost to populations.

Given that the two modes are at allele sizes of 13 and roughly 26, it is tempting to hypothesize that the second mode results from a rare but neutral doubling of a size 13 allele. Then, due to higher mutation rate at this long allele, the allele frequency distribution began to diffuse away from the original size of 26. However, this fails to explain the isolation and high frequency of the size 13 allele, which itself should be quite mutable. Given the high frequency of size 13 alleles, the total absence of alleles shorter than size 13 is difficult to explain by invoking neutral explanations alone.

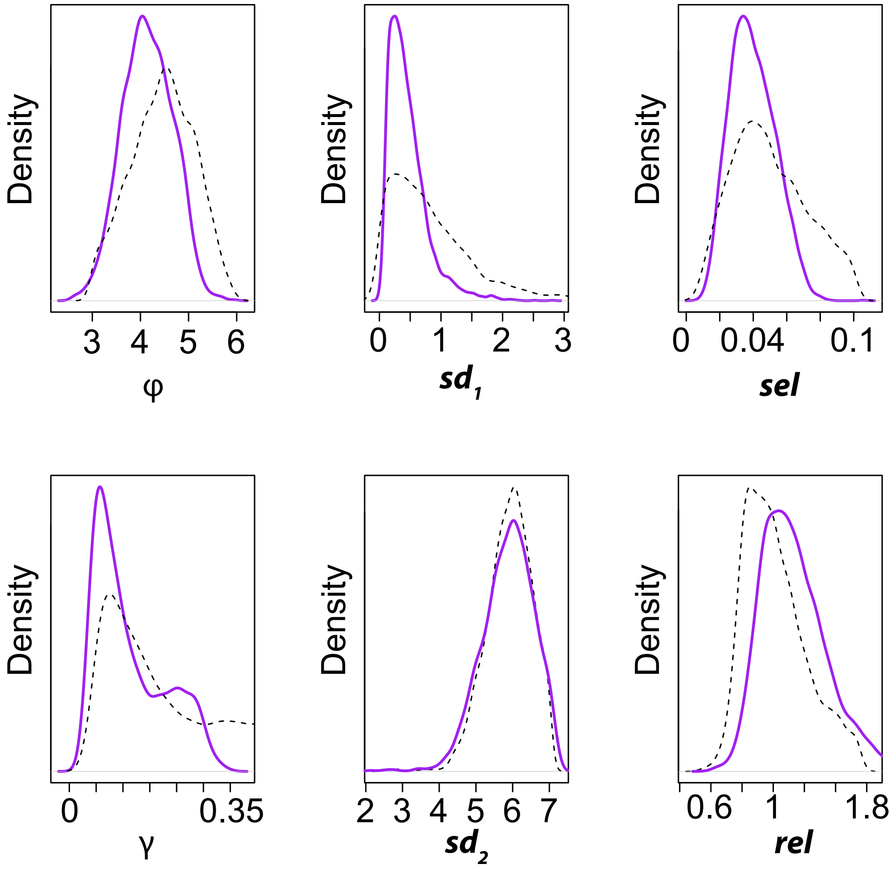

Figure 4-15 Posterior distributions for the *PAX3* microsatellite before and after adjustment by local linear regression.

Sex hormone-binding globulin (*SHBG*)

The allele frequency distribution of the pentanucleotide `TAAAA` in the promoter of *SHBG* (sex hormone-binding globulin) is also notable for its bimodality. Although allele sizes 6 and 8 are common in populations across the world, the intermediate allele size of 7 is noticeably lower in all populations (Figure 4-8). Moreover, low values of R_{ST} for all population comparisons suggest that the allele frequency distributions are objectively similar. Due to the lack of divergence between populations at the *SHBG* locus, we analyzed the allele frequency distribution of the full sample. Model selection based on the results of the rejection algorithm gave greatest support to the *bimodal-additive* model ($P(M \mid S) = 0.682$), and appreciable though much less support to the *neutral* model ($P(M \mid S) = 0.153$). As with the other candidate microsatellites that were analyzed using the less intensive rejection-only method, the posterior distribution on the mutational ψ parameter provided little resolution relative to the prior distribution. However, the mode estimate of ϕ of 4.68 (Table 4-5) suggests that mutation rate is intermediate at the locus with a maximum mutation rate of $\sim 5\text{x}10^{-3}$ for longer alleles. Given the observed allele frequency distributions, for the bimodal models we assumed that the two modes were located at allele sizes of 6 and 8. The mode estimate of $rel = 1.21$ indicates that the size 6 allele is roughly 20% more fit than the size 8 allele, while the mode estimate of $sel = 0.042$ suggests relatively strong selection against genotypes that contain alleles of less than size 5 or greater than size 11 (Figure 4-16). Interestingly, the estimated genotypic fitness surface is highest for alleles containing a size 6 allele, despite its relatively lower frequency in our sample (Figure 4-16).

In blood plasma, *SHBG* binds the sex steroids testosterone and estradiol and regulates their access to target tissues. Hogeveen et al. (2001) were the first to report a correlation between allele size at the promoter microsatellite of *SHBG* and transcriptional activity. Using luciferase reporter constructs, they found that allele sizes 7-10 all possessed 6-fold greater activity than allele size 6, which nearly silenced the gene.

Numerous studies suggest that the analyzed `TAAAA` microsatellite in the promoter of *SHBG* affects SHBG serum levels, and, as a result, human reproductive health. Polycystic ovary syndrome (PCOS) is a leading cause of female infertility, as it can result in frequent failure of the ovaries to release an oocyte during menstrual cycles or complete absence of menstrual cycles (Azziz et al., 2004). SHBG serum levels are frequently low in women with PCOS and a case-control study of PCOS found a statistically significant enrichment for `TAAAA` repeats greater than eight in size, which were in turn correlated with lower levels of serum SHBG (Xita et al., 2003). Similarly, in both control and PCOS women, another study found that genotypes with one or more alleles of size 6 had higher levels of SHBG in blood serum (Ferk et al., 2007). In another result suggesting the promoter microsatellite as a potential molecular target of fecundity selection, Xita et al. (2005) found that the size 6 allele was correlated with an earlier first menstrual cycle. In men, carriers of a size 6 allele are associated with higher levels of SHBG, and, as a consequence, testosterone (Vanbillemont et al., 2009). In a recent study of idiopathic infertility in men, infertile men were enriched for allele sizes greater than 8 and carriers of size 9 alleles had a 2.82 increased risk of infertility (Safarinejad et al., 2011). The authors also found statistically significant positive correlations between SHBG plasma concentrations and sperm count, motility, and morphology.

Thus biochemical and association studies suggest that SHBG serum concentration in men and women declines with allele size of the promoter `TAAAA` microsatellite (though note the in vitro assay of Hogeveen et al. (2001) suggests the exact opposite relationship). Moreover, low levels of serum SHBG (associated with allele sizes >8, and, in some cases simply >6) are associated with PCOS, late menarche, and infertility in males. Together, these results suggest the promoter microsatellite may be a target of fecundity selection. Although these biological facts comport with the higher fitness of size 6 alleles in our estimated genotypic fitness surface (Figure 4-16), they do not offer an explanation for the relatively lower frequency of size 7 alleles across the world. It seems unlikely that this feature can be explained by human demographic or migration history, as the same deficit of size 7 alleles is observed across the world (Figure 4-8) and very low divergence in the allele frequency distributions of the eight sampled populations as measured by R_{ST}. Although it appears that longer alleles are detrimental to the reproductive health of both sexes, the possibility remains that sexual antagonism might help explain the bimodal nature of the distribution at *SHBG*. A more sophisticated model of selection that incorporates potential sexual conflict may go further in explaining the peculiar low frequency of size 7 alleles.

Epidermal Growth Factor Receptor (*EGFR*)

Overexpression of *EGFR* has been associated with efficacy of treatment and tumorigenesis in a variety of epithelial cancers (Normanno et al., 2006), including pancreatic cancer (Lemoine et al., 1992), breast cancer (Buerger et al., 2000), and head and neck cancers (Amador et al., 2004). Moreover, the `CA` microsatellite in the first intron of *EGFR* has been shown to modulate transcriptional activity. Gebhardt et al. (1999) showed that expression declined with allele sizes > 16 in vitro. As measured by pre-mRNA expression in cell lines, however, gene expression peaked at allele sizes of 16 *and* 20. Based on molecular modeling, it has been hypothesized that longer repeats in general increase the flexibility of the DNA molecule, thereby bringing the promoter and a putative repressor binding site into close proximity (Gebhardt et al., 2000).

A wide variety of correlations between allele size and disease risk have been reported. Notable examples include: (1) higher incidence of asthma with allele sizes ≤ 16 (Wang et al., 2006); (2) a negative correlation between glioma risk and allele size (Costa et al., 2011); (3) a specific correlation between $\mathtt{CA}_{19}$ and susceptibility to somatic mutations in the *EGFR* tyrosine kinase domain, which themselves play a critical role in the development of non-small cell lung cancer (Liu et al., 2011), and; (4) increased pancreatic cancer survival (from 13 to 30 months) when the sum of allele sizes was ≥ 36 (Tzeng et al., 2007). More than anything, these results may demonstrate the danger of using the allele sizes at highly polymorphic microsatellites in association studies. It may prove too easy to find a partition of allele sizes that produces a significant correlation by chance. In fact, a number of studies have disputed the existence of an association between allele size at the EGFR intron 1 microsatellite and disease (McKay et al., 2002; Etienne-Grimaldi et al., 2005; Frolov et al., 2010).

We only analyzed the pooled sample of non-African populations, because these populations exhibited similar bimodality in their allele frequency distributions (Figure 4-9). The allele sizes of these two modes at size 16 and 20 are notable for their correspondence with the two peaks in gene expression found in the in vivo assay of Gebhardt et al. (1999). We found strong support for the bimodal-additive model with $P(M \mid S) = 0.869$ and no support for the neutral model (Table 4-4). Estimates for the sd_1 and sd_2 parameters were very similar, at 1.11 and 1.32, respectively (Table 4-5). As a result, the estimated genotypic fitness surface for the *EGFR* microsatellite is highly symmetrical with a tight cluster of high fitness values for genotypes containing alleles between the sizes of 16 and 20 (Figure 4-16).

Aldose reductase *AKR1B1*

AKR1B1 catalyzes the reduction of numerous aldehydes. Because AKR1B1 reduces the aldehyde form of glucose to sorbitol and because diabetic individuals frequently produce significant amounts of sorbitol, high expression of AKR1B1 can lead to very high concentrations of sorbitol that cause numerous complications in diabetic individuals including retinopathy and neuropathy. Reporter constructs of the CA microsatellite found in the promoter of *AKR1B1* revealed that $\texttt{CA}_{23}$ alleles are associated with significantly higher expression of *AKR1B1* than alleles $\texttt{CA}_{24-26}$ (Heesom et al., 1998; Ikegishi et al., 1999). Diabetics who developed neuropathy were less likely to possess a $\texttt{CA}_{25}$ allele (Heesom et al., 1998), which is the most frequent allele size in our sample (Figure 4-16). Also, the $\texttt{CA}_{23}$ allele is significantly enriched in patients with diabetic retinopathy (Ichikawa et al., 1999; Ikegishi et al., 1999; Demaine et al., 2000; Abhary et al., 2010). Interestingly, the $\texttt{CA}_{23}$ allele is at much lower frequency than alleles greater than $\texttt{CA}_{23}$ in all populations sampled (Figure 4-10).

Because R_{ST} between pooled African and non-African samples was only 0.0019, we analyzed all genotypes as a single data set. We found substantial support for the dominant model, ($P(M \mid S) = 0.869$) and no support for the neutral model (Table 4-4). After running simulations that assumed the dominant model and estimating parameters, allele size 24 was unanimously supported as the key allele. Results regarding the threshold parameter δ were uninformative (Table 4-5), suggesting the lack of a threshold effect at this locus. However, the mode value of the g_l parameter was -0.024, which specifies a rapid decline in allelic fitness for alleles < 24 in size. Finally a modest decrease in fitness was also predicted for allele sizes greater than 24, as specified by the mode estimate of $g_u = -0.009$. Together, these results are in keeping with the observed allele frequency distribution, which drops off sharply below allele size 24 and decreases at a slower rate above allele size 24 (Figure 4-10).

It seems unlikely that increased susceptibility to diabetes-related disorders would present sufficient selective pressure to drive worldwide distributions of the *AKR1B1* microsatellite in such similar directions. After all, diabetes is most prevalent in individuals > 65 years of age, and only recently has the incidence of diabetes started increasing around the world (Wild et al., 2004). If selection is indeed acting on this microsatellite, it would seem that the selective pressure remains a mystery.

Matrix metallopeptidase 9 (*MMP9*)

In all eight populations, we observed a bimodal distribution of the *MMP9* promoter microsatellite, with one mode at $\texttt{CA}_{14}$, a second between $\texttt{CA}_{21}$ and $\texttt{CA}_{23}$, and a near absence of alleles of sizes $\texttt{CA}_{16-18}$. Highly similar distributions have been observed in other studies, including samples of African-American (Ferrand et al., 2002), Caucasian-American (Jean et al., 1995), English (Zhang et al., 2001), Finnish (Yoon et al., 1999), Sardinian (Nelissen et al., 2000), and Swedish (Nelissen et al., 2000) populations. Interestingly, two studies of the Japanese population recovered unimodal allele frequency distributions that lacked the lower mode at allele size 14 (Shimajiri et al., 1999; Maeda et al., 2001). However, these distributions agree with the much lower frequency of the size 14 allele in the Han Chinese population (CHB) relative to all other populations we sampled (Figure 4-11). Model selection overwhelmingly supported the *bimodal-additive* model with a posterior probability of $P(M \mid S) = 0.9997$. Similar to the *PAX3* microsatellite, the estimated fitness surface for *MMP9* suggests it is a complex target of selection that has obtained mutation-selection-drift equilibrium across the world (Figure 4-16).

The matrix metalloproteinase family of enzymes aid breakdown of the extracellular matrix. The enzyme MMP9 is the largest member of the family and serves important roles in reproduction, growth and development (den Steen et al., 2002). In addition, due to its ability to degrade components of the extracellular matrix including collagens, MMP9 plays an important role in tumor cell invasion and metastasis (e.g., Hiratsuka et al., 2002; Huang et al., 2002) as well as vascular diseases (e.g., Chandrasekar et al., 2006).

The *MMP9* promoter microsatellite is thought to influence transcriptional activity due to its proximity to the transcription initiation site and the fact that it lies intermediate to two consensus binding sites for the transcription factor AP-1 (den Steen et al., 2002). Indeed, a number of studies have shown that length variation at this microsatellite affects transcriptional activity using reporter assays in a variety of cells (Himelstein et al., 1998; Peters et al., 1999; Shimajiri et al., 1999; Maeda et al., 2001; Ferrand et al., 2002).

Together, these facts suggest that expression levels of *MMP9* as controlled by the analyzed promoter microsatellite may be important to development as well as the risk of disease.

Testing the possibility of selection on linked variation

Because selection on linked variants affects microsatellite variation (Slatkin, 1995a; Wiehe, 1998), it is important to test for the possibility that the anomalous and putatively non-neutral patterns of polymorphism at the candidate microsatellites are not due to linked selection. As detailed in Chapter 3, direct selection on microsatellites should have different effects on linked variation. As with the *GRIN2-B* locus, we downloaded sequence data for the 400kb of flanking sequence surrounding each of the candidate microsatellites. Then we calculated the six statistics studied in Chapter 3 for significant deviation from neutrality by comparing them to empirical distributions that accounted for local recombination rate. Results are presented in Table 4-6. At all six candidate loci substantial numbers of 10kb windows in the flanking 400kb sequence show significant values of K. This was true of all popuations, except the African populations of LWK and YRI. In addition, M and H_{FW} were found to be significant at several windows.

The most important conclusion to be drawn from these results is that D and E are not significant for any of the loci in any population. This suggests that selection on a linked variant is not the cause of anomalous patterns of microsatellite variation at the candidate loci. It is tempting to suggest that significant values of K provide further evidence of microsatellite selection at these loci given the results of Chapter 3, which suggest K is particularly good at detecting microsatellite selection (Figure 3-3B). However, the lack of significant K in African populations suggests that K may in fact be detecting demographic change in non-African populations.

Discussion

The microsatellites studied here highlight the roles that genomic context and historical contingency play in shaping the varied states and distributions that microsatellites occupy

locus	additive	multiplicative	dominant	recessive	bimodal-additive	bimodal-recessive	neutral
GRIN2B	0.006	0.026	**0.954**	0.014	0	0	0
PAX3	0	0	0	0.016	**0.965**	0	0.019
SHBG	0.005	0	0.066	0.028	**0.682**	0.072	0.153
EGFR	0.031	0.0001	0.0147	0.086	**0.869**	0.0003	0.0001
AKR1B1	0.008	0.038	**0.785**	0.0003	0.023	0.146	0
MMP9	0	0	0.003	0.0005	**0.9997**	0	0

Table 4-4: Posterior probabilities, $P(M \mid S)$, for each model. Maximum posterior probability is in boldface. Note the bimodal models were not simulated for *TLR2*.

locus	ϕ	δ	$\mathbf{g_l}$	$\mathbf{g_u}$
	logN(1.48, 0.15)	N(0.075, 0.075)	[0,0]	logN(-4, 0.5)*
GRIN2B	4.43	-0.063	0	-0.0101
	(3.92,5.06)	(0.009,0.128)	(0,0)	(-0.021, -0.023)

locus	ϕ	δ	$\mathbf{g_l}$	$\mathbf{g_u}$	$\mathbf{sd_1}$	$\mathbf{sd_2}$	**rel**	**sel**
	[3, 6]	[-0.1, 0.1]	[-0.03, 0.03],	[-0.03, 0.03]	[0.5, 5]	[0.5, 5]	[0.25, 1.75],	[0., 0.1]
PAX3	4.17	–	–	–	0.38	5.95	1.24	0.039
	(3.2, 5.13)	–	–	–	(0.09, 1.32)	(4.48, 7.02)	(1.0, 1.82)	(0.017, 0.068)
SHBG	4.68	–	–	–	1.19	3.18	1.21	0.042
	(3.69, 5.93)	–	–	–	(0.12, 3.32)	(1.57, 4.69)	(0.77, 1.52)	(0.024, 0.067)
EGFR	4.52	–	–	–	1.11	1.32	0.93	0.053
	(3.88, 5.08)	–	–	–	(0.62, 2.22)	(0.49, 3.06)	(0.61, 1.41)	(0.008, 0.098)
AKR1B1	5.55	-0.005	-0.024	-0.009	–	–	–	–
	(5.16,5.78)	(-0.07,0.1)	(-0.029, -0.0097)	(-0.015, 0.003)	–	–	–	–
MMP9	3.93	–	–	–	0.83	3.65	0.98	0.044
	(3.14, 4.80)	–	–	–	(0.56, 1.79)	(2.06, 4.93)	(0.85, 1.19)	(0.008, 0.088)

Table 4-5: Posterior probabilities on model parameters. For *GRIN2B*, informative, non-uniform priors were used. These are designated in the first line of the table (logN = lognormal distribution; N = normal distribution). Priors for the other five models were identical uniform distributions. For each posterior, mode and (2.5%, 97.5%) credible interval are listed. *because lognormal distributions are only supported for positive values, this prior is actually for $-g_u$

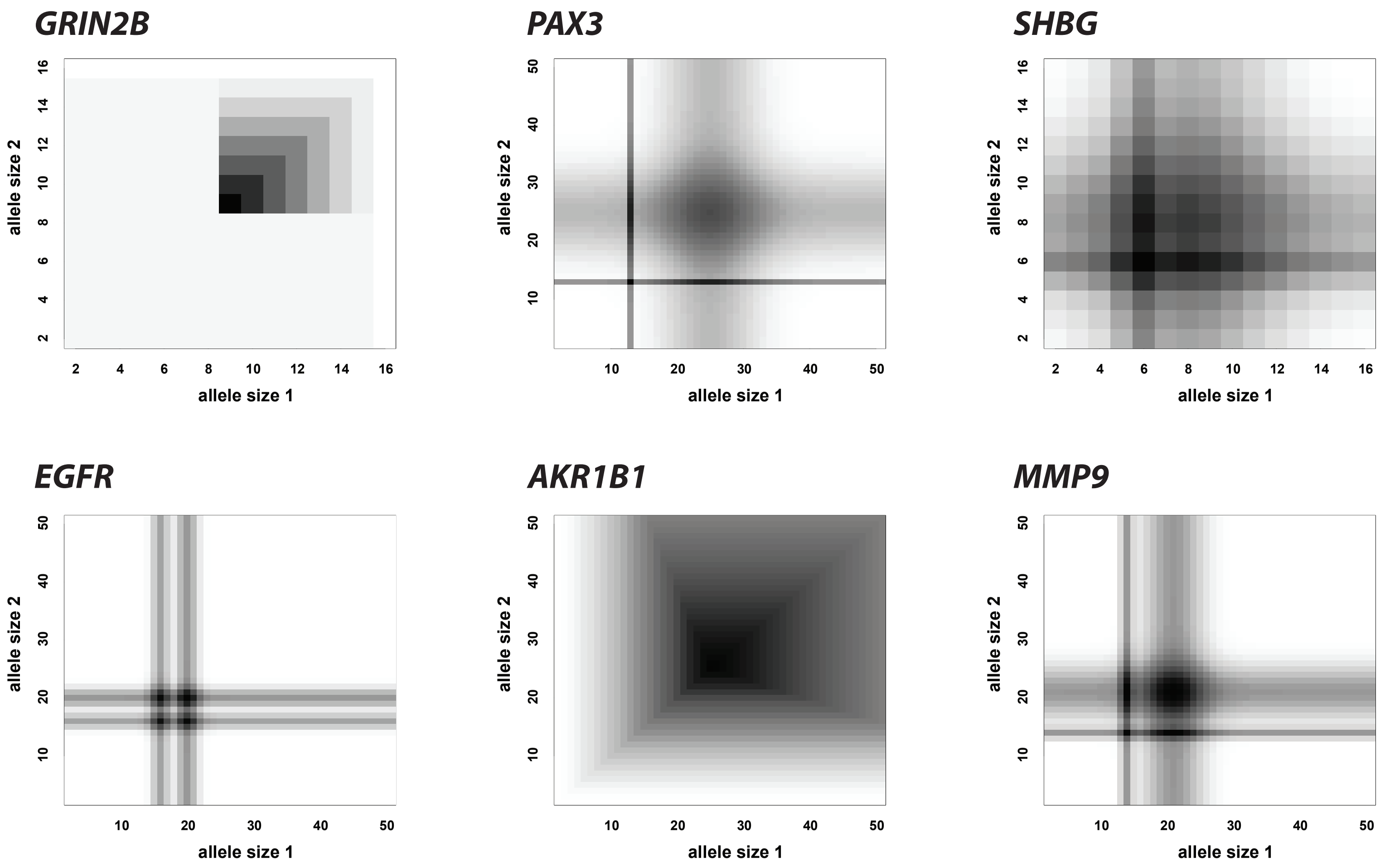

Figure 4-16 Fitness surfaces for the six candidate microsatellites drawn based on the posterior modes of selective parameters. Black is maximal relative fitness of 1, while lighter shades of gray represent decreasing fitness. Color scales are not equivalent between graphs.

locus	population	K	H	M	D	H_{FW}	E
GRIN2B	GBR	5	0	0	0	1	0
PAX3	CHB	22	0	8	0	2	0
EGFR	TSI	18	0	5	0	0	0
AKR1B1	JPT	31	0	18	0	7	0
MMP9	MXL	7	0	4	0	2	0
SHBG	FIN	14	0	14	0	14	0

Table 4-6: For select populations, the number of significant 10kb windows for six summary statistics in the 400kb sequence surrounding the six candidate microsatellites. K = number of haplotypes, H = haplotype diversity, M = count of the most frequent haplotype, D = Tajima's D,H_{FW} = Fay & Wu's H, E = Zeng et al.'s E

in the organisms inhabiting Earth today. For example, genomic context is the most obvious reason that dinucleotides are rare in exons. In the context of an exon, these microsatellites are deleterious because they cause frameshifts. Therefore, their expansion should be subject to negative selection. However, this negative selective pressure against long, exonic dinucleotides does not explain the few examples of exonic dinucleotides that do exist in the human genome, and, seemingly, in primates separated by tens of millions of years of divergence (Table 4-2). Are these in fact historical oddities that by chance escaped negative selective pressure in the genome of a common ancestor to primates or vertebrates? Or, is there an unseen commonality to these loci that might explain their presence?

In contrast, the genomic context of intergenic microsatellites should free them from most selective constraints. As a result, patterns of polymorphism at intergenic microsatellites enable us to infer that mutation rate increases with allele size (Figures 4-1, 4-2). In addition, the historical migration of these variants along with their human carriers allows us to infer that effective population size in Africa is nearly triple that of non-African populations (Figure 4-3). Their putatively neutral status also makes them invaluable as reference microsatellites. For example, the remarkable monomorphism of the $\texttt{CA}_{10}$ microsatellite in

FGFRL1 is made all the more remarkable by comparison to intergenic microsatellites of the same motif and size, which exhibit strikingly different patterns of polymorphism and divergence (Table 4-2).

Comparing among the candidate microsatellites, we find that four of the six candidates were identified as targets of bimodal-additive selection, while the other two were found to be targets of dominant selection. This supports the supposition that similar microsatellite function (regulation of gene transcription) can be instantiated and subsequently selected upon in a number of different ways. In this sense, the purported versatility of microsatellites as agents of great phenotypic variation (King et al., 1997; Fondon and Garner, 2004) is supported.

The fact that all six microsatellite candidates were identified as targets of selection by our method may simply reflect the fact that they were good candidates. Indeed, their allele frequency distributions are unusual. In particular, the distributions of the *MMP9* and *PAX3* candidate microsatellites might be considered errors if not for numerous studies that have uncovered similar bimodal distributions at these loci. However, cross-validation analysis indicates that the choice of non-neutral models should be treated with some caution, as 22.5% of neutral data sets were classified as target of natural selection. In addition, the most frequently identified model of selection among the candidate microsatellites (bimodal-additive) is the very model for which neutral data sets were most frequently mistaken (Table 4-3).

These results stress the importance of posterior predictive checks. Our application of posterior predictive checks to the *GRIN2B* microsatellite demonstrate that these analyses can substantially bolster confidence in the selected model and estimated parameter values. In the case of *GRIN2B*, posterior predictive checks showed that the neutral model was incapable of producing data sets similar to the observed allele frequencies (Figure 4-14).

Ultimately, the results presented in this chapter demonstrate that inference of microsatellite selection requires attention to detail. The observation of dramatically bimodal allele frequency distributions required that we develop models that could capture this type of variation. Biological details regarding expression of *SHBG* suggest that although we identify bimodal-additive selection as causative of the observed data, a model of sexual antagonism may provide a better fit. In the case of exonic dinucleotides, divergence data and comparison to intergenic loci turn a data set with no variation into a data set of great interest. In other words, the very complexities that make microsatellites interesting subjects of study require that we account for those complexities. Future inference of selection on microsatellites can be improved by even greater attention to detail, which is the subject of the final, brief chapter of this thesis.

CHAPTER 5: CONCLUSIONS

An inferential framework for detecting selection targeting microsatellites

Repetitive sequences were first shown to modulate gene expression in the early 1980s (Russell et al., 1983; Hamada et al., 1984). Since that time, experimental evidence has accumulated in support of the existence of non-neutral, functional microsatellites (Gemayel et al., 2010). During the same period, numerous authors have speculated that microsatellites and other repetitive sequences are important sources of adaptive genetic variation (King et al., 1997; Kashi and King, 2006). Still, a method for statistically testing claims of microsatellite selection has been lacking. In this thesis, models of microsatellite selection were introduced (Chapters 2 and 4) and used to infer selection from empirical data (Chapter 4). This work represents the first constructive step towards filling the gap between speculation and objective testing of the hypotheses engendered by this speculative thought.

The inferential framework found in Chapters 2-4 of this thesis is multifaceted. First, our work provides a detailed method for testing the neutrality of candidate microsatellites. This approach comprises: (1) realistic models of microsatellite mutation (Chapter 2); (2) biologically plausible models of microsatellite selection (Chapters 2 and 4); (3) rapid simulation of data sets according to these models (Chapter 2), and; (4) adoption of the approximate Bayesian computation (ABC) approach to simulation (Pritchard et al., 1999; Marjoram et al., 2003; Wegmann et al., 2009), model choice (Fagundes et al., 2007; Beaumont, 2008), and parameter estimation (Beaumont et al., 2002).

Over the last decade, ABC has become an increasingly popular means to perform statistical inference in situations where full likelihood methods are impossible due to a lack of relevant analytical results (Beaumont, 2010; Csillery et al., 2010). However, it is worth remembering that ABC is an approximate method; posterior distributions on models and parameters should be evaluated in this context. Moreover, model selection within the framework of ABC (Beaumont, 2008) has been the subject of some theoretical criticism (Robert et al., 2011). In general, concerns regarding the accuracy of ABC-based inference are best addressed using cross-validation procedures (Table 4-3) and posterior predictive checks (e.g., Figure 4-14). These methods assess whether different models can be distinguished using a given set of summary statistics as well as the ability of estimated parameter values to recapitulate observed data. These same approaches have been used in recent studies to address the concerns of Robert et al. (2011) (e.g., Barker et al., 2012; Dutech et al., 2012; Sousa et al., 2012).

Second, the results of the simulation study presented in Chapter 3 suggest it may be possible to perform scans for microsatellite targets of selection using sequence data alone. We discovered that the number of haplotypes, K, when conditioned on the number of segregating sites, S, provides greater power to detect selection on microsatellites with high mutation rate than to detect hard or soft sweeps (Figure 3-3A,B). A genomic scan for microsatellite selection would therefore begin with the identification of windows with anomalous values of $K \mid S$ in publicly available data such as those provided by the 1000 Genomes Project (1000 Genomes Project Consortium, 2010). Long microsatellites near these windows would then be treated as candidates for selection. Finally, genotype data from these microsatellites would be analyzed using our method of ABC-based inference in an attempt to corroborate the putative selective footprint indicated by the anomalous value of $K \mid S$. As we show in Chapters 3 (Figure 3-6) and 4 (Table 4-6), the reverse order of analysis can also be performed – i.e., we can attempt to corroborate support for selection on a candidate microsatellite by subsequently calculating values of K in sequence adjacent to this locus.

A primary challenge to using the $K \mid S$ statistic in genomic scans is the generation of an appropriate null distribution via simulation. In addition to selection, K is sensitive to neutral processes such as demographic change and recombination. As we speculated in Chapter 3, conditioning on S may negate the confounding effects of demographic change because K and S will respond in similar ways to changes in population size (e.g., both are expected to increase in response to population expansion). However, conditioning K on S does not account for differences in recombination rate across the genome; a recombination event can easily leave S unchanged while increasing K by 2. Although we incorporated fine-scale estimates of recombination rate (Kong et al., 2010) in the generation of null distributions, further simulations are required to properly assess the effect of failing to account for recombination rate appropriately.

Finally, using the example of exonic dinuclotide microsatellites, we showed that the combination of polymorphism and divergence data at microsatellite loci can help identify microsatellites that have apparently been subject to selective constraints for tens of millions of years (Table 4-2). Others have used divergence data to characterize the conservation of microsatellite variation across mammals (Buschiazzo and Gemmell, 2010; Mularoni et al., 2010; Sawaya et al., 2012). Intuitively, however, claims of conservation and selective constraint that are based on interspecific data only are bolstered by the observation that a microsatellite is also invariant in a sizeable intraspecific sample. We garnered additional support for the hypothesis that exonic nucleotides are under selective constraint by comparing: (1) polymorphism and divergence at these loci to intergenic microsatellites of similar size and motif, and; (2) the observed number of exonic dinucleotides to those expected based on their genome-wide distribution (Payseur et al., 2011). As we discuss in Chapter 4, development of a statistical test that makes use of microsatellite polymorphism and divergence data – i.e., a microsatellite-specific analog to the HKA test (Hudson et al., 1987) – would provide a more direct means to reject the null hypothesis of neutral evolution.

Implications

Models of microsatellite selection and methods used to detect it must contend with the challenge of high mutation rates at microsatellite loci. Recurrent mutation, multi-step mutation, mutation rate dependent on length and motif; these and other complications invalidate assumptions that are reasonable in the context of selection on single nucleotide variants (Chapter 1). For example, long microsatellites violate assumptions of the infinite sites model (Kimura, 1969) because they are expected to experience frequent recurrent mutation (Chapter 3). We therefore developed models and methods that explicitly incorporate the complications of microsatellite mutation.

Besides providing us with a means to test the neutrality of microsatellites, our models of selection and mutation coupled with the simulation algorithm we developed allowed us to explore the genomic and population-level consequences of selection on microsatellites. In Chapter 3, we examined genome-level consequences. We found that microsatellite selection often affects linked variation in a manner similar to that of soft sweeps targeting single nucleotide variants. On average, microsatellite selection affects the site frequency spectrum minimally, while its effect on the haplotype configuration of proximate windows is more pronounced (Figure 3-1). Yet we also found that microsatellite selection occasionally produces deep, long-ranged selective footprints in summaries of the site frequency spectrum (Figure 3-1B). Furthermore, the effect of microsatellite selection on linked variation is correlated with the difference between the allele frequency distribution present at the start of selection and that at mutation-selection equilibrium (Figure 3-1D). Together, these results all suggest that the genomic consequences of microsatellite selection are complicated and contingent upon nuisance parameters such as the starting vector of allele frequencies.

Our investigation of the population-level consequences of microsatellite selection show that our models of microsatellite selection in diploids do increase the fitness of a population.

As such, our models are a marked improvement over the non-quantitative models previously proposed in the literature, which do not necessarily lead to adaptive evolution (e.g., Figure 1-2,B-D). We also found that distinct models of selection can affect population fitness in very different ways. For example, the duration of selection – the time between the onset of selection and mutation-selection equilibrium – was markedly higher for dominant and recessive models than additive and multiplicative models (Figure 2-5B). Also, regardless of model, both the cost and duration of selection were significantly correlated with the difference between starting and equilibrium distributions of allele frequencies (Table 2-2). In the context of inference, we can treat the unknown starting distribution as a nuisance parameter by integrating results over numerous starting distributions. In the broader context, however, these results suggest that selection on microsatellites and other repetitive sequences may be less common than some authors have speculated. In short, the selective regimes of non-neutral microsatellites and their starting frequency distributions are both unknown; yet both have a dramatic and significant effect on the efficiency and cost of selection. In our opinion, these unknowns should temper speculation that microsatellites are ubiquitous agents of adaptive evolution.

It this thesis, we investigated a very small fraction of microsatellite variation in the human genome. Importantly, the models and methods documented here empower exploration of greater numbers of microsatellite loci, which could provide answers to broader questions of microsatellite evolution. A question of particular interest is: *What is the contribution of microsatellites to adaptive evolution?* Although all six candidate microsatellites analyzed in Chapter 4 enjoyed strong support for selection, it cannot be said that these loci were an unbiased sample of genomic microsatellite variation. If the same analyses were applied to 10,000 randomly selected microsatellites in the human genome, what fraction would be inferred as targets of natural selection? Though certainly less than 100%, the actual figure remains unknown. While our research corroborates the existence of microsatellite loci whose variation has been shaped by non-neutral evolution, the small number of loci examined precludes us from making a genome-wide assessment of microsatellite selection.

The primary obstacle to identifying the fraction of non-neutral microsatellites in genomes is a lack of population genomic microsatellite data. Encouragingly, increased availability of high-coverage sequence data coupled with the methods of Gymrek et al. (2012) and Highnam et al. (2013) should soon make it possible to produce large data sets of microsatellite polymorphism that are relatively free of ascertainment bias. Although it would be difficult to apply ABC-based inference at the genome-wide scale, a polymorphism-divergence test would be eminently applicable to a data set comprising human polymorphism data and microsatellite lengths observed in the reference genomes of comparative species. Furthermore, scans for microsatellite selection can be performed today using publicly available sequence data (with the caveats detailed in the previous section).

Microsatellites are still widely used as putatively neutral markers in non-model organisms (Jennings et al., 2011; Gardner et al., 2011). Yet, without a means to test the neutrality of microsatellite markers, it is possible that researchers frequently use non-neutral microsatellites to test hypotheses of importance to conservation and evolutionary biology. In this context, parameter estimation is of less importance than the binary choice between neutrality and selection. Although our ABC-based method of model choice does a relatively poor job at choosing between different models of selection (Table 4-3), the false negative rate (identifying a non-neutral microsatellite as neutral) was only 8.5%. This suggests that our method could be applied immediately to improve the choice of neutral microsatellite markers in population genetic studies.

Improving inference of microsatellite selection

A number of improvements to the methods detailed in this thesis are possible in the short term. First, the allele frequency distributions shown in Figures 4-6 through 4-11 make it clear that inter-population variability is sometimes large. In studies of selection targeting single nucleotide variants, it is important to control for demographic influences on genetic variation as these forces can lead to false positives (Zeng et al., 2006; **?**). Although we limited our analyses to pooled samples of populations that exhibited low pairwise R_{ST} in an attempt to avoid the confounding influence of demography, it would be advantageous to explicitly incorporate demography in our ABC-based inference procedure. Some suggest that ABC may be used to jointly infer selection and demography (Li et al., 2012; Crisci et al., 2012). Yet we believe that the joint estimation of selection, demography, *and* microsatellite mutation asks too much. Rather, we propose a two-step method in which the first step is to estimate demographic history from sequence data or co-opt a previously estimated model of demographic history (e.g., Ramachandran et al., 2005; Schaffner et al., 2005; Liu et al., 2006b; DeGiorgio et al., 2009). In the second step, we would assume an estimated history, incorporating its details into simulations of mutation and selection.

Simultaneous simulation of sequence and microsatellite data represents another tractable advance that could improve the current ABC-based inference procedure. Concurrent simulation of both types of variation would allow us to use the summary statistics investigated in Chapter 3. This is important because the allele frequency distribution at the microsatellite locus can only be summarized in a limited number of ways. Also, joint simulation of sequence and microsatellite variation may help resolve the difficulty our method currently has distinguishing between different forms of microsatellite selection (Table 4-3). Concurrent simulation of sequence and microsatellite variation would also sidestep the issue of generating a separate empirical distribution for the sequence data; in the context of ABC, we simply retain (possibly regression-adjusted) parameter values that produce a good match between observed and simulated summary statistics. Moreover, joint simulation of sequence and microsatellite variation would allow the introduction of a selective sweep model that would capture the effect of SNV-based selection on linked sequence and linked microsatellites. In this way, the hypothesis of indirect selection could be tested side-by-side with models of direct microsatellite selection.

Having studied the complexities of microsatellite mutation and selection in some detail, we are less certain than others that microsatellites provide a vast reservoir of adaptive genetic variation. However, as detailed in the Chapter 1, microsatellites are interesting variants that mutate in two dimensions, and, as a consequence, can produce genetic and phenotypic effects that are much more difficult to achieve with simple point mutation. Furthermore, our results do show that some microsatellites are targets of positive and purifying selection. To ignore the adaptive and maladaptive potential of specific microsatellite variation is therefore inadvisable.

Regardless of their frequency as drivers of adaptive evolution, microsatellites represent yet another form of variation in the diversity of the world. Thus, as we have found, study of their evolutionary dynamics adds to our understanding of biological evolution no matter what their primary functions may be.

REFERENCES

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.

Aandahl, R. Z., J. F. Reyes, S. A. Sisson, and M. M. Tanaka. 2012. A model-based Bayesian estimation of the rate of evolution of VNTR loci in *Mycobacterium tuberculosis*. *PLoS Computational Biology* 8:e1002573.

Abhary, Sotoodeh, Kathryn P Burdon, Kate J Laurie, Stacey Thorpe, John Landers, Lucy Goold, Stewart Lake, Nikolai Petrovsky, and Jamie E Craig. 2010. Aldose reductase gene polymorphisms and diabetic retinopathy susceptibility. *Diabetes Care* 33(8):1834–1836.

Agarwal, A. K., G. Giacchetti, G. Lavery, H. Nikkila, M. Palermo, M. Ricketts, C. McTernan, G. Bianchi, P. Manunta, P. Strazzullo, F. Mantero, P. C. White, and P. M. Stewart. 2000. CA-repeat polymorphism in intron 1 of *HSD11B2*: effects on gene expression and salt sensitivity. *Hypertension* 36:187–194.

Aitken, Nicola, Steven Smith, Carsten Schwarz, and Phillip A Morin. 2004. Single nucleotide polymorphism (SNP) discovery in mammals: a targeted-gene approach. *Mol Ecol* 13(6):1423–1431.

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

Amador, Maria L, Darin Oppenheimer, Sofia Perea, Anirban Maitra, George Cusatis, George Cusati, Christi Iacobuzio-Donahue, Sharyn D Baker, Raheela Ashfaq, Chris Takimoto, Arlene Forastiere, and Manuel Hidalgo. 2004. An epidermal growth factor receptor intron 1 polymorphism mediates response to epidermal growth factor receptor inhibitors. *Cancer Res* 64(24):9139–9143.

Amos, W., S. J. Sawcer, R. W. Feakes, and D. C. Rubinsztein. 1996. Microsatellites show mutational bias and heterozygote instability. *Nature Genetics* 13:390–391.

Avise, J.C. 1994. *Molecular markers, nautral history and evolution*. Kluwer Academic Publishers.

Azzalini, A. 2011. *R package `sn`: The skew-normal and skew-t distributions (version 0.4-17)*. Università di Padova, Italia.

Azziz, Ricardo, Keslie S Woods, Rosario Reyna, Timothy J Key, Eric S Knochenhauer, and Bulent O Yildiz. 2004. The prevalence and features of the polycystic ovary syndrome in an unselected population. *J Clin Endocrinol Metab* 89(6):2745–2749.

Bachtrog, D., S. Weiss, B. Zangerl, G. Brem, and C. Schlötterer. 1999. Distribution of dinucleotide microsatellites in the *Drosophila melanogaster* genome. *Mol Biol Evol* 16(5): 602–610.

Baldi, P., and S. Brunak. 1998. *Bioinformatics: The machine learning approach*. 1st ed. The MIT Press.

Baldwin, C. T., C. F. Hoth, R. A. Macina, and A. Milunsky. 1995. Mutations in PAX3 that cause Waardenburg syndrome type I: ten new mutations and review of the literature. *Am J Med Genet* 58(2):115–122.

Barker, Brittany S, Javier A Rodríguez-Robles, Vani S Aran, Ashley Montoya, Robert B Waide, and Joseph A Cook. 2012. Sea level, topography and island diversity: phylogeography of the Puerto Rican red-eyed coquì, *Eleutherodactylus antillensis*. *Mol Ecol* 21(24): 6033–6052.

Barria, Andres, and Roberto Malinow. 2005. NMDA receptor subunit composition controls synaptic plasticity by regulating binding to CaMKII. *Neuron* 48(2):289–301.

Beaumont, M. A. 2008. Joint determination of topology, divergence time, and immigration in population trees. In *Simulations, genetics, and human prehistory*, ed. S. Matsumura, P. Forster, and C. Renfrew, 135–154. McDonald Institute for Archaeological Research.

———. 2010. Approximate Bayesian computation in evolution and ecology. *Annu. Rev. Ecol. Evol. Syst.* 41:379–406.

Beaumont, M. A., W. Zhang, and D. J. Balding. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035.

Beder, B. 1988. Allele frequencies given the sample's common ancestral type. *Theoretical Population Biology* 33:126–137.

Beerli, P., and J. Felsenstein. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences USA* 98:4563–4568.

Bersaglieri, T., P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner, J. A. Drake, M. Rhodes, D. E. Reich, and J. N. Hirschhorn. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics* 74:1111–1120.

Bhargava, A., and F. F. Fuentes. 2010. Mutational dyamics of microsatellites. *Molecular Biotechnology* 44:250–266.

Bhide, P. G., M. Day, E. Sapp, C. Schwarz, A. Sheth, J. Kim, A. B. Young, J. Penney, J. Golden, N. Aronin, and M. DiFiglia. 1996. Expression of normal and mutant huntingtin in the developing brain. *J Neurosci* 16(17):5523–5535.

Biswas, S., and J. M. Akey. 2006. Genomic insights into positive selection. *Trends in Genetics* 22:437–446.

Brandstrom, M., and H. Ellegren. 2008. Genome-wide analysis of microsatellite polymorphism in chicken circumventing the ascertainment bias. *Genome Research* 18:881–887.

Brinkmann, B., M. Klintschar, F. Neuhuber, J. Huhne, and B. Rolf. 1998. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *American Journal of Human Genetics* 62:1408–1415.

Britten, R.J., and E.H. Davidson. 1969. Gene regulation for higher cells: a theory. *Science* 165:349–357.

———. 1971. Repetitive and non-repetitive DNA sequences and a speculation on origins of evolutionary novelty. *Quarterly Review of Biology* 46:111–138.

Broman, K. W., J. C. Murray, V. C. Sheffield, R. L. White, and J. L. Weber. 1998. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *American Journal of Human Genetics* 63:861–869.

Brook, J. D., M. E. McCurrach, H. G. Harley, A. J. Buckler, D. Church, H. Aburatani, K. Hunter, V. P. Stanton, J. P. Thirion, and T. Hudson. 1992. Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. *Cell* 68(4):799–808.

Buerger, H., F. Gebhardt, H. Schmidt, A. Beckmann, K. Hutmacher, R. Simon, R. Lelle, W. Boecker, and B. Brandt. 2000. Length and loss of heterozygosity of an intron 1 polymorphic sequence of *EGFR* is related to cytogenetic alterations and epithelial growth factor receptor expression. *Cancer Res* 60(4):854–857.

Bürger, R. 1988. Mutation-selection balance and continuum-of-allele models. *Mathematical Biosciences* 91:67–83.

———. 1998. Mathematical properties of mutation-selection models. *Genetica* 102/103: 279–298.

Buschiazzo, Emmanuel, and Neil J Gemmell. 2010. Conservation of human microsatellites across 450 million years of evolution. *Genome Biol Evol* 2:153–165.

Butler, John M. 2006. Genetics and genomics of core short tandem repeat loci used in human identity testing. *J Forensic Sci* 51(2):253–265.

Calabrese, P. P., R. T. Durrettt, and C. F. Aquadro. 2001. Dyanmics of microsatellite divergence under stepwise mutation and proportional slippage/point mutation models. *Genetics* 159:839–852.

Campuzano, V., L. Montermini, M. D. Molto, L. Pianese, M. Cossee, F. Cavalcanti, E. Monros, F. Rodius, A. Monticelli, F. Zara, J. Canizares, H. Koutnikova, S. I. Bidichandani, C. Gellera, A. Brice, P. Trouillas, G. DeMichelle, A. Filla, R. DeFrutos, F. Palau, P. I. Patel, S. Di Donato, J. L. Mandel, S. Cocozza, M. Koenig, and M. Pandolfo. 1996. Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* 271:1423–1427.

Chandrasekar, Bysani, Srinivas Mummidi, Lenin Mahimainathan, Devang N Patel, Steven R Bailey, Syed Z Imam, Warner C Greene, and Anthony J Valente. 2006. Interleukin-18-induced human coronary artery smooth muscle cell migration is dependent on NF-kappaB- and AP-1-mediated matrix metalloproteinase-9 expression and is inhibited by atorvastatin. *J Biol Chem* 281(22):15099–15109.

Chang, D. K., D. Metzgar, C. Wills, and C. R. Boland. 2001. Microsatellites in the eukaryotic dna mismatch repair genes as modulators of evolutionary mutation rate. *Genome Research* 11:1145–1146.

de la Chapelle, Albert. 2003. Microsatellite instability. *N Engl J Med* 349(3):209–210.

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

Chung, H., Claudia G Lopez, Joy Holmstrom, Dennis J Young, Jenny F Lai, Deena Ream-Robinson, and John M Carethers. 2010. Both microsatellite length and sequence context determine frameshift mutation rates in defective dna mismatch repair. *Hum Mol Genet* 19(13):2638–2647.

Clark, A. G. 1998. Mutation-selection balance with multiple alleles. *Genetica* 102/103: 41–47.

Colombo, R., and A. Carobene. 2000. Age of the intronic GAA triplet repeat expansion mutation in friedreich ataxia. *Human Genetics* 106:455–458.

Conrad, Donald F, and Matthew E Hurles. 2007. The population genetics of structural variation. *Nat Genet* 39(7 Suppl):S30–S36.

Contente, Ana, Alexandra Dittmer, Manuela C Koch, Judith Roth, and Matthias Dobbelstein. 2002. A polymorphic microsatellite that mediates induction of *PIG3* by *p53*. *Nat Genet* 30(3):315–320.

Cooper, G. M., D. A. Nickerson, and E. E. Eichler. 2007. Mutational and selective effects on copy-number variants in the human genome. *Nature Genetics* 39:S22–S29.

Cossee, M., M. Schmitt, V. Campuzano, L. Reutenauer, C. Moutout, J.-L. Mandel, and M. Koenig. 1997. Evolution of the Friedreich's ataxia trinucleotide repeat expansion: Founder effect and premutations. *Proceedings of the National Academy of Sciences, USA* 94: 7452–7457.

Costa, Bruno Marques, Marta Viana-Pereira, Ricardo Fernandes, Sandra Costa, Paulo Linhares, Rui Vaz, Céline Pinheiro, Jorge Lima, Paula Soares, Ana Silva, Fernando Pardal, Júlia Amorim, Rui Nabiço, Rui Almeida, Carlos Alegria, Manuel Melo Pires, Célia Pinheiro, Ernesto Carvalho, Pedro Oliveira, José M Lopes, and Rui M Reis. 2011. Impact of *EGFR* genetic variants on glioma risk and patient outcome. *Cancer Epidemiol Biomarkers Prev* 20(12):2610–2617.

Crisci, Jessica L, Yu-Ping Poh, Angela Bean, Alfred Simkin, and Jeffrey D Jensen. 2012. Recent progress in polymorphism-based population genetic inference. *J Hered* 103(2): 287–296.

Crow, J. F., and M. Kimura. 1970. *An introduction to population genetic theory*. Harper and Row.

Csillery, K., O. Francois, and M. G. B. Blum. 2012. abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution* 3:475–479.

Csillery, Katalin, Michael G B Blum, Oscar E Gaggiotti, and Olivier Francois. 2010. Approximate bayesian computation (ABC) in practice. *Trends Ecol Evol* 25(7):410–418.

DeGiorgio, M., M. Jakobsson, and N. A. Rosenberg. 2009. Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from africa. *Proceedings of the National Academy of Sciences USA* 106:16057–16062.

Demaine, A., D. Cross, and A. Millward. 2000. Polymorphisms of the aldose reductase gene and susceptibility to retinopathy in type 1 diabetes mellitus. *Invest Ophthalmol Vis Sci* 41(13):4064–4068.

Dib, C., S. Faure, C. Fizames, D. Samson, N. Drouot, A. Vignal, P. Millasseau, S. Marc, J. Hazan, E. Seboun, M. Lathrop, G. Gyapay, J. Morissette, and J. Weissenbach. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 380:152–154.

DiRienzo, A., A. C. Peterson, J. C. Garza, A. M. Valdes, M. Slatkin, and N. B. Freimer. 1994. Mutational processes of simple-sequence repeat loci in human populations. *Proceedings of the National Academy of Sciences USA* 91:3166–3170.

Dokholyan, N. V., S. V. Buldyrev, S. Havlin, and H. E. Stanley. 2000. Distributions of dimeric tandem repeats in non-coding and coding dna sequences. *J Theor Biol* 202(4): 273–282.

Durr, A., M. Cossee, Y. Agid, V. Campuzano, C. Mignard, C. Penet, J.-L. Mandel, A. Brice, and M. Koenig. 1996. Clinical and genetic abnormalities in patients with Friedreich's ataxia. *New England Journal of Medicine* 335:1169–1175.

Dutech, C., B. Barrès, J. Bridier, C. Robin, M. G. Milgroom, and V. Ravigné. 2012. The chestnut blight fungus world tour: successive introduction events from diverse origins in an invasive plant fungal pathogen. *Mol Ecol* 21(16):3931–3946.

Eckert, K.A., and S.E. Hile. 2009. Every microsatellite is different: intrinsic dna features dictate mutatgenesis of common microsatellites present in the human genome. *Molecular Carcenogenesis* 48:379–388.

Edwards, A., H. A. Hammond, L. Jin, C. T. Caskey, and R. Chakraborty. 1992. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics* 12(2):241–253.

Edwards, A. W. F. 2000. Sewall Wright's equation $\Delta q = (q(1-q)\partial w/\partial q)/2w$. *Theoretical Population Biology* 57:67–70.

Ellegren, H. 1991. DNA typing of museum birds. *Nature* 354(6349):113.

———. 2000. Heterogeneous mutation processes in human microsatellite DNA sequences. *Nature Genetics* 24:400–402.

———. 2004. Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics* 5:435–445.

Elmore, Martha H, John G Gibbons, and Antonis Rokas. 2012. Assessing the genome-wide effect of promoter region tandem repeat natural variation on gene expression. *G3 (Bethesda)* 2(12):1643–1649.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.

Endele, Sabine, Georg Rosenberger, Kirsten Geider, Bernt Popp, Ceyhun Tamer, Irina Stefanova, Mathieu Milh, Fanny Kortüm, Angela Fritsch, Friederike K Pientka, Yorck Hellenbroich, Vera M Kalscheuer, Jürgen Kohlhase, Ute Moog, Gudrun Rappold, Anita Rauch, Hans-Hilger Ropers, Sarah von Spiczak, Holger Tönnies, Nathalie Villeneuve, Laurent Villard, Bernhard Zabel, Martin Zenker, Bodo Laube, André Reis, Dagmar Wieczorek, Lionel Van Maldergem, and Kerstin Kutsche. 2010. Mutations in *GRIN2A* and *GRIN2B* encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes. *Nat Genet* 42(11):1021–1026.

Estoup, Arnaud, Philippe Jarne, and Jean-Marie Cornuet. 2002. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Mol Ecol* 11(9):1591–1604.

Etienne-Grimaldi, M-C., S. Pereira, N. Magné, J-L. Formento, M. Francoual, X. Fontana, F. Demard, O. Dassonville, G. Poissonnet, J. Santini, R-J. Bensadoun, P. Szepetowski, and G. Milano. 2005. Analysis of the dinucleotide repeat polymorphism in the epidermal growth factor receptor (*EGFR*) gene in head and neck cancer patients. *Ann Oncol* 16(6): 934–941.

Ewens, W. J. 1972. The sampling theory of selectively neutral alleles. *Theoretical Population Biology* 3:87–112.

Fagundes, Nelson J R, Nicolas Ray, Mark Beaumont, Samuel Neuenschwander, Francisco M Salzano, Sandro L Bonatto, and Laurent Excoffier. 2007. Statistical evaluation of alternative models of human evolution. *Proc Natl Acad Sci U S A* 104(45):17614–17619.

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

Ferk, Polonca, Natasa Teran, and Ksenija Gersak. 2007. The (TAAAA)n microsatellite polymorphism in the *SHBG* gene influences serum SHBG levels in women with polycystic ovary syndrome. *Hum Reprod* 22(4):1031–1036.

Ferrand, Pedro E, Samuel Parry, Mary Sammel, George A Macones, Helena Kuivaniemi, Roberto Romero, and Jerome F Strauss. 2002. A polymorphism in the matrix metalloproteinase-9 promoter is associated with increased risk of preterm premature rupture of membranes in African Americans. *Mol Hum Reprod* 8(5):494–501.

Field, D., and C. Wills. 1998. Abundant microsatellite polymorphism in saccharomyces cerevisiae, and the different distributions of microsatellites in eight prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces. *Proc Natl Acad Sci U S A* 95(4):1647–1652.

Fondon, J. W., and H. R. Garner. 2004. Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences USA* 101:18058–18063.

Fries, R., A. Eggen, and G. Stranzinger. 1990. The bovine genome contains polymorphic microsatellites. *Genomics* 8(2):403–406.

Froehlich, Allan C, Yi Liu, Jennifer J Loros, and Jay C Dunlap. 2002. White collar-1, a circadian blue light photoreceptor, binding to the frequency promoter. *Science* 297(5582): 815–819.

Frolov, Andrey, J. Spencer Liles, Andrew V Kossenkov, Ching-Wei D Tzeng, Nirag Jhala, Peter Kulesza, Shyam Varadarajulu, Mohamad Eloubeidi, Martin J Heslin, and J. Pablo Arnoletti. 2010. Epidermal growth factor receptor (*EGFR*) intron 1 polymorphism and clinical outcome in pancreatic adenocarcinoma. *Am J Surg* 200(3):398–405.

Fu, Y. H., D. P. Kuhl, A. Pizzuti, M. Pieretti, J. S. Sutcliffe, S. Richards, A. J. Verkerk, J. J. Holden, R. G. Fenwick, and S. T. Warren. 1991. Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell* 67(6):1047–1058.

Galvao, R., L. Mendes-Soares, J. Camara, I. Jaco, and M. Carmo-Fonseca. 2001. Triplet repeats, RNA secondary structure and toxic gain-of-function models for pathogenesis. *Brain Research Bulletin* 56:191–201.

Gambrill, Abigail C, and Andres Barria. 2011. NMDA receptor subunit composition controls synaptogenesis and synapse stabilization. *Proc Natl Acad Sci U S A* 108(14): 5855–5860.

Gardner, Michael G, Alison J Fitch, Terry Bertozzi, and Andrew J Lowe. 2011. Rise of the machines–recommendations for ecologists when using next generation sequencing for microsatellite development. *Mol Ecol Resour* 11(6):1093–1101.

Gebhardt, F., H. Bürger, and B. Brandt. 2000. Modulation of *EGFR* gene transcription by a polymorphic repetitive sequence–a link between genetics and epigenetics. *Int J Biol Markers* 15(1):105–110.

Gebhardt, F., K. S. Zänker, and B. Brandt. 1999. Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *J Biol Chem* 274(19):13176–13180.

Gemayel, R., M. D. Vinces, M. Legendre, and K. J. Verstrepen. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual Review. Genetics* 44:445–477.

Georgiev, G.P. 1969. On the structural organization of the operon and the regulation of RNA synthesis in animal cells. *Journal of Theoretical Biology* 25:473–490.

Glazko, Galina V, and Masatoshi Nei. 2003. Estimation of divergence times for major lineages of primate species. *Mol Biol Evol* 20(3):424–434.

Godde, J. S., and A. P. Wolfe. 1996. Nucleosome assembly on CTG triplet repeats. *Journal of Biological Chemistry* 271:15222–15229.

Goldberg, C. S., and L. P. Waits. 2010. Comparative landscape genetics of two pond-breeding amphibian species in a highly modified agricultural landscape. *Molecular Ecology* 19:3650–3663.

Goldstein, D. B., and A. G. Clark. 1995. Microsatellite variation in north american populations of drosophila melanogaster. *Nucelic Acids Research* 23:3882–3886.

Goldstein, D. B., A. R. Linares, L. L. Cavalli-Sforza, and M. W. Feldman. 1995a. An evaluation of genetic distances for use with microsatellite loci. *Genetics* 139:463–471.

———. 1995b. Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceedings of the National Academy of Sciences USA* 92:6723–6727.

Gymrek, M., D. Golan, S. Rosset, and Y. Erlich. 2012. lobSTR: a short tandem repeat profiler for personal genomes. *Genome Research* 22:1154–1162.

Haasl, R. J., and B. A. Payseur. 2010. The number of alleles at a microsatellite defines the allele frequency spectrum and facilitates fast, accurate estimation of θ. *Molecular Biology and Evolution* 27:2702–2715.

———. 2011. Multi-locus inference of population structure: a comparsison between single nucleotide polymorphisms and microsatellites. *Heredity* 106:158–171.

———. 2013. Microsatellites as targets of natural selection. *Molecular Biology and Evolution* 30:285–298.

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

van Ham, S. M., L. van Alphen, F. R. Mooi, and J. P. van Putten. 1993. Phase variation of *H. influenzae* fimbriae: transcriptional control of two divergent genes through a variable combined promoter region. *Cell* 73(6):1187–1196.

Hamada, H., M. Seidman, B. H. Howard, and C. M. Gorman. 1984. Enhanced gene expression by the poly(dT-dG).poly(dC-dA) sequence. *Mol Cell Biol* 4(12):2622–2630.

Hammock, E. A., and L. J. Young. 2005. Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science* 308:1630–1634.

Hampikian, G., E. West, and O. Akselrod. 2011. The genetics of innocence: analysis of 194 U.S. DNA exonerations. *Annual Review. Genomics and Human Genetics* 12:97–120.

Hancock, A. M., G. Alkorta-Aranburu, D. B. Witonsky, and A. DiRienzo. 2010a. Adaptations to new environments in humans: the role of subtle allele frequency shifts. *Phiolosophical Transactions of the Royal Society of London B: Biological Sciences* 365:2459–2468.

Hancock, A. M., D. B. Witonsky, E. Ehler, G. Alkorta-Aranburu, C. Beall, A. Gebremedhin, R. Sukernik, G. Utermann, J. Pritchard, G. Coop, and A. DiRienzo. 2010b. Colloquium paper: human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proceedings of the National Academy of Sciences, USA* 107:S2:8924–8930.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. Springer.

Hastings, P. J., J. R. Lupski, S. M. Rosenberg, and G. Ira. 2009. Mechanisms of change in gene copy number. *Nature Reviews Genetics* 10:551–564.

Heesom, A. E., A. Millward, and A. G. Demaine. 1998. Susceptibility to diabetic neuropathy in patients with insulin dependent diabetes mellitus is associated with a polymorphism at the 5′ end of the aldose reductase gene. *J Neurol Neurosurg Psychiatry* 64(2):213–216.

Hefferon, T. W., J. D. Groman, C. E. Yurk, and G. R. Cutting. 2004. A variable dinucleotide repeat in the *CFTR* gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. *Proceedings of the National Academy of Sciences, USA* 101: 3504–3509.

Henke, L., and J. Henke. 2006. Supplemented data on mutation rates in 33 autosomal short tandem repeat polymorphisms. *Journal of Forensic Science* 51:446–447.

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

Highnam, Gareth, Christopher Franck, Andy Martin, Calvin Stephens, Ashwin Puthige, and David Mittelman. 2013. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res* 41(1): e32.

Himelstein, B. P., E. J. Lee, H. Sato, M. Seiki, and R. J. Muschel. 1998. Tumor cell contact mediated transcriptional activation of the fibroblast matrix metalloproteinase-9 gene: involvement of multiple transcription factors including Ets and an alternating purine-pyrimidine repeat. *Clin Exp Metastasis* 16(2):169–177.

Hiratsuka, Sachie, Kazuhiro Nakamura, Shinobu Iwai, Masato Murakami, Takeshi Itoh, Hiroshi Kijima, J. Michael Shipley, Robert M Senior, and Masabumi Shibuya. 2002. *MMP9* induction by vascular endothelial growth factor receptor-1 is involved in lung-specific metastasis. *Cancer Cell* 2(4):289–300.

Hoff, P.D. 2009. *A first course in bayesian statistical methods*. Springer.

Hogeveen, K. N., M. Talikka, and G. L. Hammond. 2001. Human sex hormone-binding globulin promoter activity is influenced by a (TAAAA)n repeat element within an Alu sequence. *J Biol Chem* 276(39):36383–36390.

Hohenlohe, Paul A, Susan Bassham, Paul D Etter, Nicholas Stiffler, Eric A Johnson, and William A Cresko. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced rad tags. *PLoS Genet* 6(2):e1000862.

Huang, C. R. L., A. M. Schneider, Y. Lu, T. Niranjan, P. Shen, M. A. Robinson, J. P. Steranka, D. Valle, C. I. Civin, T. Wang, S. J. Wheelan, H. Ji, J. D. Boeke, and K. H. Burns. 2010. Mobile interspersed repeats are major structural variants in the human genome. *Cell* 141: 1171–1182.

Huang, Suyun, Melissa Van Arsdall, Sean Tedjarati, Marya McCarty, Wenjuan Wu, Robert Langley, and Isaiah J Fidler. 2002. Contributions of stromal metalloproteinase-9 to angiogenesis and growth of human ovarian carcinoma in mice. *J Natl Cancer Inst* 94(15): 1134–1142.

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. Aguadé. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116(1):153–159.

Huntington's Disease Collaborative Research Group. 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* 72:971–983.

Ichikawa, F., K. Yamada, S. Ishiyama-Shigemoto, X. Yuan, and K. Nonaka. 1999. Association of an (A-C)n dinucleotide repeat polymorphic marker at the 5′-region of the aldose reductase gene with retinopathy but not with nephropathy or neuropathy in japanese patients with type 2 diabetes mellitus. *Diabet Med* 16(9):744–748.

Ikegishi, Y., M. Tawata, K. Aida, and T. Onaya. 1999. Z-4 allele upstream of the aldose reductase gene is associated with proliferative retinopathy in Japanese patients with NIDDM, and elevated luciferase gene transcription in vitro. *Life Sci* 65(20):2061–2070.

Innan, H., and Y. Kim. 2005. Pattern of polymophism after strong artificial selection in a domestication event. *Proceedings of the National Academy of Sciences* 101:10667–10672.

Jean, P. L. St, X. C. Zhang, B. K. Hart, H. Lamlum, M. W. Webster, D. L. Steed, A. M. Henney, and R. E. Ferrell. 1995. Characterization of a dinucleotide repeat in the 92 kDa type IV collagenase gene (*CLG4B*), localization of CLG4B to chromosome 20 and the role of CLG4B in aortic aneurysmal disease. *Ann Hum Genet* 59(Pt 1):17–24.

Jennings, T. N., B. J. Knaus, T. D. Mullins, S. M. Haig, and R. C. Cronn. 2011. Multiplexed microsatellite recovery using massively parallel sequencing. *Mol Ecol Resour* 11(6):1060–1067.

Johnson, A. C., Y. Jinno, and G. T. Merlino. 1988. Modulation of epidermal growth factor receptor proto-oncogene transcription by a promoter site sensitive to S1 nuclease. *Mol Cell Biol* 8(10):4174–4184.

Kashi, Y., D. King, and M. Soller. 1997. Simple sequence repeats as a source of quantitative genetic variation. *Trends in Genetics* 13:74–78.

Kashi, Y., and D.G. King. 2006. Simple sequence repeats as advantageous mutators in evolution. *Trends in Genetics* 22:253–259.

Kelkar, Y.D., S. Tyekucheva, F. Chiaromonte, and K.D. Makova. 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Research* 18: 30–38.

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

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

Kimura, M., and J. F. Crow. 1964. The number of alleles that can be maintained in a finite population. *Genetics* 49:725–738.

Kimura, M., and T. Ohta. 1975. Distribution of allelic frequencies in a finite population under stepwise production of neutral alleles. *Proceedings of the National Academy of Sciences USA* 72:2761–2764.

———. 1978. Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proceedings of the National Academy of Sciences USA* 75:2868–2872.

King, D. G. 1999. Modeling selection for adjustable genes based on simple sequence repeats. *Annals of the New York Academy of Sciences* 870:396–399.

King, D. G., and Y. Kashi. 2009. Heretical DNA sequences? *Science* 326:229–230.

King, David G. 2012. Indirect selection of implicit mutation protocols. *Ann N Y Acad Sci* 1267:45–52.

King, D.G. 1994. Triplet repeat DNA as a highly mutable regulatory mechanism. *Science* 263:595–596.

King, D.G., M. Soller, and Y. Kashi. 1997. Evolutionary tuning knobs. *Endeavour* 21:36–40.

Kingman, J. F. C. 1977. On the properties of bilinear models for the balance between genetic mutation and selection. *Mathematical Proceedings of the Cambridge Philosophical Society* 81:443–453.

Knowles, Margaret A. 2007. Role of FGFR3 in urothelial cell carcinoma: biomarker and potential therapeutic target. *World J Urol* 25(6):581–593.

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, and S. A. Gudjonsson. 2002. A high-resolution recombination map of the human genome. *Nature Genetics* 31:241–247.

Kong, A., G. Thorleifsson, D. F. Gudbjartsson, G. Masson, A. Sigurdsson, A. Jonasdottir, G. B. Walters, A. Jonasdottir, A. Gylfason, and K. T. Kristinsson et al. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* 467: 1099–1103.

Korbel, J. O., A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert, J. F. Simons, P. M. Kim, D. Palejev, N. J. Carriero, L. Du, B. E. Taillon, Z. Chen, A. Tanzer, A. C. Saunders, J. Chi, F. Yang, N. P. Carter, M. E. Hurles, S. M. Weissman, T. T. Harkins, M. B. Gerstein, M. Egholm, and M. Snyder. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318:420–426.

Kozlowski, Piotr, Mateusz de Mezer, and Wlodzimierz J Krzyzosiak. 2010. Trinucleotide repeats in human genome and exome. *Nucleic Acids Res* 38(12):4027–4039.

Kremer, E. J., M. Pritchard, M. Lynch, S. Yu, K. Holman, E. Baker, S.T. Warren, D. Schlessinger, G. R. Sutherland, and R. I. Richards. 1991. Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. *Science* 252:1711–1714.

Kruglyak, S., R. T. Durrett, M. D. Schug, and C. F. Aquadro. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proceedings of the National Academy of Sciences USA* 95:10774–10778.

Lamason, Rebecca L, Manzoor-Ali P K Mohideen, Jason R Mest, Andrew C Wong, Heather L Norton, Michele C Aros, Michael J Jurynec, Xianyun Mao, Vanessa R Humphreville, Jasper E Humbert, Soniya Sinha, Jessica L Moore, Pudur Jagadeeswaran, Wei Zhao, Gang Ning, Izabela Makalowska, Paul M McKeigue, David O'donnell, Rick Kittles, Esteban J Parra, Nancy J Mangini, David J Grunwald, Mark D Shriver, Victor A Canfield, and Keith C Cheng. 2005. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* 310(5755):1782–1786.

Lau, C. Geoffrey, and R. Suzanne Zukin. 2007. NMDA receptor trafficking in synaptic plasticity and neuropsychiatric disorders. *Nat Rev Neurosci* 8(6):413–426.

Lavery, G. G., C. L. McTernan, S. C. Bain, T. A. Chowdhury, M. Hewison, and P. M. Stewart. 2002. Association studies between the HSD11B2 gene (encoding human 11beta-hydroxysteroid dehydrogenase type 2), type 1 diabetes mellitus and diabetic nephropathy. *European Journal of Endocrinology* 146:553–558.

Legendre, M., N. Pochet, T. Pak, and K.J. Verstrepen. 2007. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Research* 17:1787–1796.

Lemoine, N. R., C. M. Hughes, C. M. Barton, R. Poulsom, R. E. Jeffery, G. Klöppel, P. A. Hall, and W. J. Gullick. 1992. The epidermal growth factor receptor in human pancreatic cancer. *J Pathol* 166(1):7–12.

Leopoldino, A. M., and S. D. Pena. 2003. The mutational spectrum of human autosomal tetranucleotide microsatellites. *Human Mutation* 21:71–79.

Levinson, G., and G. A. Gutman. 1987. High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. *Nucleic Acids Res* 15(13):5323–5338.

Levy, S., G. Sutton, P.C. Ng, A.L. Halpern, B.P. Walenz, N. Axelrod, J. Huang, E.F. Kirkness, G. Denisov, Y. Lin, J.R. MacDonald, A.W. Pang, M. Shago, T.B. Stockwell, A. Tsiamouri, V. Bafna, V. Bansal, S.A. Kravitz, D.A. Busam, K.Y. Beeson, T.C. McIntosh, K.A. Remington, J.F. Abril, J. Gill, J. Borman, Y.H. rogers, M.E. Frazier, S.W. Scherer, R.L. Strausber, and J.C. Venter. 2007. The diploid genome sequence of an individual human. *PLoS Biology* 5:e254.

Lewontin, R. C., and J. Krakauer. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74(1):175–195.

Li, H. 2011. A new test for detecting recent positive selection that is free from the confounding impacts of demography. *Molecular Biology and Evolution* 28:365–375.

Li, Junrui, Haipeng Li, Mattias Jakobsson, Sen Li, Per Sjödin, and Martin Lascoux. 2012. Joint analysis of demography and selection in population genetics: where do we stand and where could we go? *Mol Ecol* 21(1):28–44.

Li, Y. C., A. B. Korol, T. Fahima, and E. Nevo. 2004. Microsatellites within genes: structure, function, and evolution. *Molecular Biology and Evolution* 21:991–1007.

Litt, M., and J. A. Luty. 1989. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am J Hum Genet* 44(3): 397–401.

Liu, B., N. C. Nicolaides, S. Markowitz, J. K. V. Willson, R. E. Parsons, J. Jen, N. Papadopolous, P. Peltomaki, A. Delachapelle, S. R. Hamilton, K. W. Kinzler, and B. Vogelstein. 1995. Mismatch repair gene defects in sporadic colorectal cancers with microsatellite instability. *Nature Genetics* 9:48–55.

Liu, H., F. Prugnolle, A. Manica, and F. Balloux. 2006a. A geographically explicit genetic model of worldwide human-settlement history. *American Journal of Human Genetics* 79: 230–237.

Liu, Hua, Franck Prugnolle, Andrea Manica, and François Balloux. 2006b. A geographically explicit genetic model of worldwide human-settlement history. *Am J Hum Genet* 79(2):230–237.

Liu, Wanqing, Lijun He, Jacqueline Ramírez, Soundararajan Krishnaswamy, Rajani Kanteti, Yi-Ching Wang, Ravi Salgia, and Mark J Ratain. 2011. Functional *EGFR* germline polymorphisms may confer risk for EGFR somatic mutations in non-small cell lung cancer, with a predominant effect on exon 19 microdeletions. *Cancer Res* 71(7):2423–2427.

Lohi, Hannes, Edwin J Young, Susan N Fitzmaurice, Clare Rusbridge, Elayne M Chan, Mike Vervoort, Julie Turnbull, Xiao-Chu Zhao, Leonarda Ianzano, Andrew D Paterson, Nathan B Sutter, Elaine A Ostrander, Catherine André, G. Diane Shelton, Cameron A Ackerley, Stephen W Scherer, and Berge A Minassian. 2005. Expanded repeat in canine epilepsy. *Science* 307(5706):81.

Ludwig, Kerstin U, Darina Roeske, Stefan Herms, Johannes Schumacher, Andreas Warnke, Ellen Plume, Nina Neuhoff, Jennifer Bruder, Helmut Remschmidt, Gerd Schulte-Körne, Bertram Müller-Myhsok, Markus M Nöthen, and Per Hoffmann. 2010. Variation in *GRIN2B* contributes to weak performance in verbal short-term memory in children with dyslexia. *Am J Med Genet B Neuropsychiatr Genet* 153B(2):503–511.

Lynch, Michael. 2010. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A* 107(3):961–968.

Macina, R. A., F. G. Barr, N. Galili, and H. C. Riethman. 1995. Genomic organization of the human *PAX3* gene: DNA sequence analysis of the region disrupted in alveolar rhabdomyosarcoma. *Genomics* 26(1):1–8.

Maeda, S., M. Haneda, B. Guo, D. Koya, K. Hayashi, T. Sugimoto, K. Isshiki, H. Yasuda, A. Kashiwagi, and R. Kikkawa. 2001. Dinucleotide repeat polymorphism of matrix metalloproteinase-9 gene is associated with diabetic nephropathy. *Kidney Int* 60(4):1428–1434.

Mahadevan, M., C. Tsilfidis, L. Sabourin, G. Shutler, C. Amemiya, G. Jansen, C. Neville, M. Narang, J. Barceló, and K. O'Hoy. 1992. Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene. *Science* 255(5049):1253–1255.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré. 2003. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences, USA* 100:15324–15328.

Marriage, T. N., S. Hudman, M. E. Mort, M. E. Orive, R. G. Shaw, and J. K. Kelly. 2009. Direct estimation of the mutation rate at dinucleotide microsatellite loci in *Arabadopsis thaliana* (*Brassicaceae*). *Heredity* 103:310–317.

Martin, Patricia, Katherine Makepeace, Stuart A Hill, Derek W Hood, and E. Richard Moxon. 2005. Microsatellite instability regulates transcription factor binding and gene expression. *Proc Natl Acad Sci U S A* 102(10):3800–3804.

Maynard Smith, J., and J. Haigh. 1974. The hitch-hiking effect of a favourable gene. *Genetical Research* 23:23–35.

Maynard Smith, John. 1976. What determines the rate of evolution? *American Naturalist* 110:331–338.

McConnell, R., S. Middlemist, C. Scala, J. E. Strassmann, and D. C. Quell. 2007. An unusually low microsatellite mutation rate in *Dictyostelium discoideum*, an organism with unusually abundant microsatellites. *Genetics* 177:1499–1507.

McEvoy, Brian P, Joseph E Powell, Michael E Goddard, and Peter M Visscher. 2011. Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Res* 21(6):821–829.

McKay, J. A., L. J. Murray, S. Curran, V. G. Ross, C. Clark, G. I. Murray, J. Cassidy, and H. L. McLeod. 2002. Evaluation of the epidermal growth factor receptor (*EGFR*) in colorectal tumours and lymph node metastases. *Eur J Cancer* 38(17):2258–2264.

Metzgar, D., J. Bytof, and C. Wills. 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res* 10(1):72–80.

Meyer, Laurence R, Ann S Zweig, Angie S Hinrichs, Donna Karolchik, Robert M Kuhn, Matthew Wong, Cricket A Sloan, Kate R Rosenbloom, Greg Roe, Brooke Rhead, Brian J

Raney, Andy Pohl, Venkat S Malladi, Chin H Li, Brian T Lee, Katrina Learned, Vanessa Kirkup, Fan Hsu, Steve Heitner, Rachel A Harte, Maximilian Haeussler, Luvina Guruvadoo, Mary Goldman, Belinda M Giardine, Pauline A Fujita, Timothy R Dreszer, Mark Diekhans, Melissa S Cline, Hiram Clawson, Galt P Barber, David Haussler, and W. James Kent. 2013. The UCSC genome browser database: extensions and updates 2013. *Nucleic Acids Res* 41(D1):D64–D69.

Michael, Todd P, Sohyun Park, Tae-Sung Kim, Jim Booth, Amanda Byer, Qi Sun, Joanne Chory, and Kwangwon Lee. 2007. Simple sequence repeats provide a substrate for phenotypic variation in the *Neurospora crassa* circadian clock. *PLoS One* 2(8):e795.

Miesfeld, R., M. Krystal, and N. Arnheim. 1981. A member of a new repeated sequence family which is conserved throughout eucaryotic evolution is found between the human delta and beta globin genes. *Nucleic Acids Res* 9(22):5931–5947.

Molla, M., A. Delcher, S. Sunyaev, C. Cantor, and S. Kasif. 2009. Triplet repeat length bias and variation in the human transcriptome. *Proceedings of the National Academy of Sciences, USA* 106:17095–17100.

Montermini, L., E. Andermann, M. Labuda, A. Richter, M. Pandolfo, F. Cavalcanti, L. Pianese, L. Iodice, G. Farina, A. Monticelli, M. Turano, A. Filla, Giuseppe De Michelle, and S. Cocozza. 1997. The Friedreich ataxia GAA triplet repeat: premutation and normal alleles. *Human Molecular Genetics* 6:1261–1266.

Monticelli, A., M. Giacchetti, I. De Biase, L. Pianese, M. Turano, M. Pandolfo, and S. Cocozza. 2004. New clues on the origin of the Friedreich ataxia expanded alleles from the analysis of new polymorphisms closely linked to the mutation. *Human Genetics* 114: 458–463.

Moran, P. A. P. 1975. Wandering distributions and the electrophoretic profile. *Theoretical Population Biology* 8:318–330.

———. 1976. Global stability of genetic systems governed by mutation and selection. *Mathematical Proceedings of the Cambridge Philosophical Society* 80:331–336.

———. 1977. Global stability of genetic systems governed by mutation and selection. II. *Mathematical Proceedings of the Cambridge Philosophical Society* 81:435–441.

Morin, Phillip A, Karen K Martien, and Barbara L Taylor. 2009. Assessing statistical power of SNPs for population structure and conservation studies. *Mol Ecol Resour* 9(1):66–73.

Moxon, E.R., P.B. Rainey, M.A. Nowak, and R.E. Lenski. 1994. Adaptive evolution of highly mutable loci in pathogenic bacteria. *Current Biology* 4:24–33.

Mularoni, Loris, Alice Ledda, Macarena Toll-Riera, and M. Mar Albà. 2010. Natural selection drives the accumulation of amino acid tandem repeats in human proteins. *Genome Res* 20(6):745–754.

Nachman, M.W., and S.L. Crowell. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* 156:297–304.

Nair, S., B. Miller, M. Barends, A. Jaidee, J. Patel, M. Mayxay, P. Newton, F. Nosten, M. T. Ferdig, and T. J. C. Anderson. 2008. Adaptive copy number evolution in malaria parasites. *PLoS Genetics* 4:e1000243.

Nair, S., J. T. Williams, A. Brockman, L. Paiphun, M. Mayxay, P. N. Newton, J. P. Guthmann, F. M. Smithuis, T. T. Hien, N. J. White, F. Nosten, and T. J. Anderson. 2003. A selective sweep driven by pyrimethamine treament in southeast Asian malaria parasites. *Molecular Biology and Evolution* 20:1526–1536.

Navascués, M., O. J. Hardy, and C. Burgarella. 2009. Characterization of demograhpic expansions from pairwise comparisons of linked microsatellite haplotypes. *Genetics* 181: 1013–1019.

Naylor, L. H., and E. M. Clark. 1990. d(TG)n-d(CA)n sequences upstream of the rat prolactin gene form Z-DNA and inhibit gene transcription. *Nucleic Acides Research* 18: 1595–1601.

Nelissen, I., K. Vandenbroeck, P. Fiten, J. Hillert, T. Olsson, M. G. Marrosu, and G. Opdenakker. 2000. Polymorphism analysis suggests that the gelatinase B gene is not a susceptibility factor for multiple sclerosis. *J Neuroimmunol* 105(1):58–63.

Nevo, E., A. Beharav, R. C. Meyer, C. A. Hackett, B. P. Forster, J. R. Russell, and W. Powell. 2005. Genomic microsatellite adaptive divergence of wild barley by microclimatic stress in Evolution Canyon, Israel. *Biological Journal of the Linnean Society* 84:205–224.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, and C. Bustamante. 2005. Genomic scans for selective sweeps using SNP data. *Genome Research* 15:1566–1675.

Normanno, Nicola, Antonella De Luca, Caterina Bianco, Luigi Strizzi, Mario Mancino, Monica R Maiello, Adele Carotenuto, Gianfranco De Feo, Francesco Caponigro, and David S Salomon. 2006. Epidermal growth factor receptor (*EGFR*) signaling in cancer. *Gene* 366(1):2–16.

Ohno, S. 1970. So much 'junk' DNA in our genome. *Evolution of Genetic Systems* 23: 366–370.

Ohta, T., and M. Kimura. 1973. Model of mutation appropriate to estimate number of electrophoretically detectable alleles in a finite population. *Genetical Research* 22:201–204.

Oki, E., S. Oda, Y. Maehara, and K. Sugimachi. 1999. Mutated gene-specific phenotypes of dinucleotide repeat instability in human colorectal carcinoma cell lines deficient in DNA mismatch repair. *Oncogene* 18(12):2143–2147.

Okladnova, O., Y. V. Syagailo, M. Tranitz, P. Riederer, G. Stöber, R. Mössner, and K. P. Lesch. 1999. Functional characterization of the human *PAX3* gene regulatory region. *Genomics* 57(1):110–119.

Okladnova, O., Y. V. Syagailo, M. Tranitz, G. Stöber, P. Riederer, R. Mössner, and K. P. Lesch. 1998. A promoter-associated polymorphic repeat modulates *PAX-6* expression in human brain. *Biochem Biophys Res Commun* 248(2):402–405.

Oleksyk, T. K., M. W. Smith, and S. J. O'Brien. 2010. Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society, Series B* 365:185–205.

Oliveira, E.J., J.G. Padua, M.I. Zucchi, R. Vencovsky, and M.L.C. Vieira. 2006. Origin, evolution and genome distribution of microsatellites. *Genetics and molecular biology* 29: 294–307.

Orr, Harry T, and Huda Y Zoghbi. 2007. Trinucleotide repeat disorders. *Annu Rev Neurosci* 30:575–621.

Pandolfo, M. 2008. Friedreich ataxia. *Archives of Neurology* 65:1296–1303.

Payseur, B. A., P. Jing, and R. J. Haasl. 2011. A genomic portrait of human microsatellite variation. *Molecular Biology and Evolution* 28:303–312.

Pennings, P. S., and J. Hermisson. 2006a. Soft sweeps II – molecular population genetics of adptation from recurrent mutation or migration. *Molecular Biology and Evolution* 23: 1076–1084.

———. 2006b. Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS Genetics* 2:1998–2012.

Perry, George H, Nathaniel J Dominy, Katrina G Claw, Arthur S Lee, Heike Fiegler, Richard Redon, John Werner, Fernando A Villanea, Joanna L Mountain, Rajeev Misra, Nigel P Carter, Charles Lee, and Anne C Stone. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet* 39(10):1256–1260.

Peters, D. G., A. Kassam, P. L. St Jean, H. Yonas, and R. E. Ferrell. 1999. Functional polymorphism in the matrix metalloproteinase-9 promoter as a potential risk factor for intracranial aneurysm. *Stroke* 30(12):2612–2616.

Pianese, L., F. Cavalcanti, G. De Michele, A. Filla, G. Campanella, O. Calabrese, I. Castaldo, A. Monticelli, and S. Cocozza. 1997. The effect of parental gender on the GAA dynamic mutation in the *FRDA* gene. *American Journal of Human Genetics* 60:460–463.

Pritchard, J. K., and M. W. Feldman. 1996. Statistics for microsatellite variation based on coalescence. *Theoretical Population Biology* 50:325–344.

Pritchard, J. K., J. K. Pickrell, and G. Coop. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology* 20:R208–R215.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution* 16:1791–1798.

Przeworski, M., G. Coop, and J. D. Wall. 2005. The signature of positive selection on standing genetic variation. *Evolution* 59:2312–2323.

Ramachandran, Sohini, Omkar Deshpande, Charles C Roseman, Noah A Rosenberg, Marcus W Feldman, and L. Luca Cavalli-Sforza. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in africa. *Proc Natl Acad Sci U S A* 102(44):15942–15947.

Rieckmann, Thorsten, Lei Zhuang, Christa E Flück, and Beat Trueb. 2009. Characterization of the first *FGFRL1* mutation identified in a craniosynostosis patient. *Biochim Biophys Acta* 1792(2):112–121.

Riley, D.E., and J.N. Krieger. 2009. Embryonic nervous system genes predominate in searches for dinucleotide simple sequence repeats flanked by conserved sequences. *Gene* 429:74–79.

Risch, N., D. de Leon, L. Ozelius, P. Kramer, L. Almasy, B. Singer, S. Fahn, X. Breakefield, and S. Bressman. 1995. Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. *Nature Genetics* 9:152–159.

Roach, Jared C, Gustavo Glusman, Arian F A Smit, Chad D Huff, Robert Hubley, Paul T Shannon, Lee Rowen, Krishna P Pant, Nathan Goodman, Michael Bamshad, Jay Shendure, Radoje Drmanac, Lynn B Jorde, Leroy Hood, and David J Galas. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328(5978):636–639.

Robert, Christian P, Jean-Marie Cornuet, Jean-Michel Marin, and Natesh S Pillai. 2011. Lack of confidence in approximate Bayesian computation model choice. *Proc Natl Acad Sci U S A* 108(37):15112–15117.

Rockman, M. V., M. W. Hahn, N. Soranzo, F. Zimprich, D. B. Goldstein, and G. A. Wray. 2005. Ancient and recent positive selection transformed opioid *cis*-regulation in humans. *PLoS Biology* 3:e387.

Rockman, M. V., and G. A. Wray. 2002. Abundant raw material for *cis*-regulatory evolution in humans. *Molecular Biology and Evolution* 19:1991–2004.

Rosenberg, N. A., and M. Jakobsson. 2008. The relationship between homozygosity and the frequency of the most frequent allele. *Genetics* 179:2027–2036.

Rothenburg, S., F. Koch-Nolte, A. Rich, and F. Haag. 2001. A polymoprhic dinucleotide repeat in the rate nucleolin gene forms Z-DNA and inhibits promoter activity. *Proceedings of the National Academy of Sciences, USA* 98:8985–8990.

Rozanska, M., K. Sobczak, A. Jasinska, M. Napierala, D. Kaczynska, A. Czerny, M. Koziel, P. Kozlowski, M. Olejniczak, and W.J. Krzyzosiak. 2007. CAG and CTG repeat polymorphism in exons of human genes shows distinct features at expandable loci. *Human Mutation* 28:451–458.

Rubinsztein, D. C., W. Amos, J. Leggo, S. Goodburn, S. Jain, S. H. Li, R. L. Margolis, C. A. Ross, and M. A. Ferguson-Smith. 1995. Microsatellite evolution – evidence for directionality and variation in rate between species. *Nature Genetics* 10:337–343.

Russell, D. W., M. Smith, D. Cox, V. M. Williamson, and E. T. Young. 1983. Dna sequences of two yeast promoter-up mutants. *Nature* 304(5927):652–654.

Safarinejad, Mohammad Reza, Nayyer Shafiei, and Shiva Safarinejad. 2011. Association of the (TAAAA)n repeat and Asp327Asn polymorphisms in the sex hormone-binding globulin (*SHBG*) gene with idiopathic male infertility and relation to serum SHBG concentrations. *J Steroid Biochem Mol Biol* 123(1-2):37–45.

Sakimura, K., T. Kutsuwada, I. Ito, T. Manabe, C. Takayama, E. Kushiya, T. Yagi, S. Aizawa, Y. Inoue, and H. Sugiyama. 1995. Reduced hippocampal LTP and spatial learning in mice lacking NMDA receptor epsilon 1 subunit. *Nature* 373(6510):151–155.

Sandman, K., and J. N. Reeve. 1999. Archaeal nucleosome positioning by CTG repeats. *Journal of Bacteriology* 181:1035–1038.

Sawaya, Sterling M, Dustin Lennon, Emmanuel Buschiazzo, Neil Gemmell, and Vladimir N Minin. 2012. Measuring microsatellite conservation in mammalian evolution with a phylogenetic birth-death model. *Genome Biol Evol* 4(6):636–647.

Schaffner, S. F., C. Foo, S. Gabriel, D. Reich, M. J. Daly, and D. Altshuler. 2005. Calibrating a coalscent simulation of human genome sequence variation. *Genome Research* 15:1576–1583.

Schlötterer, C. 2002. A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics* 160:753–763.

Schlötterer, C., B. Amos, and D. Tautz. 1991. Conservation of polymorphic simple sequence loci in cetacean species. *Nature* 354(6348):63–65.

Schlötterer, C., R. Ritter, B. Harr, and G. Brem. 1998. High mutation rate of a long microsatellite allele in *Drosophila melanogaster* provides evidence for allele-specific mutation rates. *Molecular Biology and Evolution* 15:1269–1274.

Schlötterer, C., and D. Tautz. 1992. Slippage synthesis of simple sequence DNA. *Nucleic Acids Res* 20(2):211–215.

Schug, M. D., T. F. Mackay, and C. F. Aquadro. 1997. Low mutation rates of microsatellite loci in *Drosophila melanogaster*. *Nat Genet* 15(1):99–102.

Scotti, I., G. G. Vendramin, L. S. Matteotti, C. Scarponi, M. Sari-Gorla, and G. Binelli. 2000. Postglacial recolonization routes for *Picea abies* k. in Italy as suggested by the analysis of sequence-characterized amplified region (SCAR) markers. *Mol Ecol* 9(6):699–708.

Seyfert, A. L., M. E. A. Cristescu, L. Frisse, S. Schaack, W. K. Thomas, and M. Lynch. 2008. The rate and spectrum of microsatellite mutation in *Caenorhabditis elegans* and *Daphnia pulex*. *Genetics* 178:2113–2121.

Sheng, M., J. Cummings, L. A. Roldan, Y. N. Jan, and L. Y. Jan. 1994. Changing subunit composition of heteromeric NMDA receptors during development of rat cortex. *Nature* 368(6467):144–147.

Shimajiri, S., N. Arima, A. Tanimoto, Y. Murata, T. Hamada, K. Y. Wang, and Y. Sasaguri. 1999. Shortened microsatellite d(CA)21 sequence down-regulates promoter activity of matrix metalloproteinase 9 gene. *FEBS Lett* 455(1-2):70–74.

Sibly, R. M., J. C. Whittaker, and M. Talbot. 2001. A maximum-likelihood approach to fitting equilibrium models of microsatellite evolution. *Molecular Biology and Evolution* 18: 413–417.

Slatkin, M. 1995a. Hitchhiking and associative overdominance at a microsatellite locus. *Molecular Biology and Evolution* 12:473–480.

———. 1995b. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139(1):457–462.

Sobczak, K., M. de Mezer, G. Michlewski, J. Krol, and W. J. Krzyzosiak. 2003. RNA structure of trinucleotide repeats associated with human neurological diseases. *Nucleic Acids Research* 31:5469–5482.

Sousa, V. C., M. A. Beaumont, P. Fernandes, M. M. Coelho, and L. Chikhi. 2012. Population divergence with or without admixture: selecting models using an ABC approach. *Heredity (Edinb)* 108(5):521–530.

Spritz, R. A. 1981. Duplication/deletion polymorphism 5′ - to the human beta globin gene. *Nucleic Acids Res* 9(19):5037–5047.

Stallings, R. L., A. F. Ford, D. Nelson, D. C. Torney, C. E. Hildebrand, and R. K. Moyzis. 1991. Evolution and distribution of (GT)n repetitive sequences in mammalian genomes. *Genomics* 10(3):807–815.

Stankiewicz, P., and J. R. Lupski. 2010. Structural variation in the human genome and its role in disease. *Annual Review of Medicine* 61:437–455.

den Steen, Philippe E Van, Bénédicte Dubois, Inge Nelissen, Pauline M Rudd, Raymond A Dwek, and Ghislain Opdenakker. 2002. Biochemistry and molecular biology of gelatinase B or matrix metalloproteinase-9 (*MMP-9*). *Crit Rev Biochem Mol Biol* 37(6):375–536.

Strasburg, J. L., N. A. Sherman, K. M. Wright, L. C. Moyle, J. H. Willis, and L. H. Rieseberg. 2012. What can patterns of differentiation across plant genomes tell us about adaptation and speciation? *Philosophical Transactions of the Royal Society, Series B* 367:364–373.

Sun, J. X., A. Helgason, G. Masson, S. S. Ebeneserdottir, H. Li, S. Mallick, S. Gnerre, N. Patterson, A. Kong, D. Reich, and K. Stefansson. 2012. A direct characterization of human mutation based on microsatellites. *Nature Genetics* [Epub ahead of print].

Sun, J. X., J. C. Mullikin, N. Patterson, and D. E. Reich. 2009. Microsatellites are molecular clocks that support accurate inferences about history. *Molecular Biology and Evolution* 26: 1017–1027.

Tajima, F. 1989. Statistical method for testing the neutral muatation hypothesis by dna polymorphism. *Genetics* 123:585–595.

Tang, Y. P., E. Shimizu, G. R. Dube, C. Rampon, G. A. Kerchner, M. Zhuo, G. Liu, and J. Z. Tsien. 1999. Genetic enhancement of learning and memory in mice. *Nature* 401(6748): 63–69.

Tassabehji, M., V. E. Newton, K. Leverton, K. Turnbull, E. Seemanova, J. Kunze, K. Sperling, T. Strachan, and A. P. Read. 1994. *PAX3* gene structure and mutations: close analogies between Waardenburg syndrome and the Splotch mouse. *Hum Mol Genet* 3(7):1069–1074.

Tautz, D. 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Res* 17(16):6463–6471.

Tautz, D., M. Trick, and G. A. Dover. 1986. Cryptic simplicity in DNA is a major source of genetic variation. *Nature* 322(6080):652–656.

Teshima, Kosuke M, Graham Coop, and Molly Przeworski. 2006. How reliable are empirical genomic scans for selective sweeps? *Genome Res* 16(6):702–712.

Tian, B., R. J. White, T. Xia, S. Welle, D. H. Turner, M. B. Mathews, and C. A. Thornton. 2000. Expanded CUG repeat RNAs form hairpins that activate the double-stranded RNA-dependent protein kinase PKR. *RNA* 6:79–87.

Tishkoff, S. A., R. Varkonyi, N. Cahinhinan, S. Abbes, G. Argyropoulos, G. Destro-Bisol, A. Drousiotou, B. Dangerfield, G. Lefranc, J. Loiselet, A. Piro, M. Stoneking, A. Tagarelli, G. Tagarelli, E. H. Touma, S. M. Williams, and A. G. Clar. 2001. Haplotype diversity and linkage disequilibrium at human *G6PD*: recent origin of alleles that confer malarial resistance. *Science* 293:455–462.

Tomita, N., R. Fujita, D. Kurihara, H. Shindo, R. D. Wells, and M. Shimizu. 2002. Effects of triplet repeat sequences on nucleosome positioning and gene expression in yeast minichromosomes. *Nucelic Acids Research* Suppl. 2:231–232.

Trifonov, E.N. 1989. The multiple codes of nucleotide sequences. *Bulletin of Mathematical Biology* 51:417–432.

———. 2004. Tuning function of tandemly repeating sequences: a molecular device for fast adaptation. In *Evolutionary theory and processes: papers in honour of eviatar nevo*, ed. S.P. Wasser. Kluwer Academic Publishers.

Trueb, Beat. 2011. Biology of *FGFRL1*, the fifth fibroblast growth factor receptor. *Cell Mol Life Sci* 68(6):951–964.

Trueb, Beat, Lei Zhuang, Sara Taeschler, and Markus Wiedemann. 2003. Characterization of *FGFRL1*, a novel fibroblast growth factor (FGF) receptor preferentially expressed in skeletal tissues. *J Biol Chem* 278(36):33857–33865.

Turner, Thomas L, Elizabeth C Bourne, Eric J Von Wettberg, Tina T Hu, and Sergey V Nuzhdin. 2010. Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. *Nat Genet* 42(3):260–263.

Turner, Thomas L, Eric J von Wettberg, and Sergey V Nuzhdin. 2008. Genomic analysis of differentiation between soil types reveals candidate genes for local adaptation in *Arabidopsis lyrata*. *PLoS One* 3(9):e3183.

Tzeng, Ching-Wei D, Andrey Frolov, Natalya Frolova, Nirag C Jhala, J. Harrison Howard, Selwyn M Vickers, Donald J Buchsbaum, Martin J Heslin, and J. Pablo Arnoletti. 2007. Pancreatic cancer epidermal growth factor receptor (*EGFR*) intron 1 polymorphism influences postoperative patient survival and in vitro erlotinib response. *Ann Surg Oncol* 14(7): 2150–2158.

Valdes, A. M., M. Slatkin, and N. B. Freimer. 1993. Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* 133:737–749.

Vanbillemont, Griet, Veerle Bogaert, Dirk De Bacquer, Bruno Lapauw, Stefan Goemaere, Kaatje Toye, Kristel Van Steen, Youri Taes, and Jean-Marc Kaufman. 2009. Polymorphisms of the *SHBG* gene contribute to the interindividual variation of sex steroid hormone blood levels in young, middle-aged and elderly men. *Clin Endocrinol (Oxf)* 70(2):303–310.

Vigouroux, Y., J. S. Jaqueth, Y. Matsuoka, O. S. Smith, W. D. Beavis, J. S. Smith, and J. Doebley. 2002. Rate and pattern of mutation at microsatellite loci in maize. *Molecular Biology and Evolution* 19:1251–1260.

Vinces, M.D., M. Legendre, M. Caldara, M. Hagihara, and K.J. Verstrepen. 2009. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* 324:1213–1216.

Wang, Xintao, Junpei Saito, Takashi Ishida, and Mitsuru Munakata. 2006. Polymorphism of *egfr* intron 1 is associated with susceptibility and severity of asthma. *J Asthma* 43(9): 711–715.

Warpeha, K. M., W. Xu, L. Liu, I. G. Charles, C. C. Patterson, F. Ah-Fat, S. Harding, P. M. Hart, U. Chakravarthy, and A. E. Hughes. 1999. Genotyping and functional analysis of a polymorphic (CCTTT)(n) repeat of *NOS2A* in diabetic retinopathy. *FASEB J* 13(13): 1825–1832.

Watterson, G. A. 1975. Number of segregating sites in genetic models without recombination. *Theoretical Population Biology* 7:255–276.

Weber, J. L., and P. E. May. 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am J Hum Genet* 44(3):388–396.

Weber, J.L., and C. Wong. 1993. Mutation of human short tandem repeats. *Human Molecular Genetics* 2:1123–1128.

Wegmann, D., C. Leuenberger, and L. Excoffier. 2009. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics* 182: 1207–1218.

Weiser, J. N., J. M. Love, and E. R. Moxon. 1989. The molecular mechanism of phase variation of *H. influenzae* lipopolysaccharide. *Cell* 59(4):657–665.

Weiser, J. N., D. J. Maskell, P. D. Butler, A. A. Lindberg, and E. R. Moxon. 1990. Characterization of repetitive sequences controlling phase variation of *Haemophilus influenzae* lipopolysaccharide. *J Bacteriol* 172(6):3304–3309.

White, P. C., A. K. Agarwal, B. S. Nunez, G. Giacchetti, F. Mantero, and P. M. Stewart. 2000. Genotype-phenotype correlations of mutations and polymorphisms in *HSD11B2*, the gene encoding the kidney isozyme of 11-beta-hydroxysteroid dehydrogenase. *Endocrine Research* 26:771–780.

Wiehe, T. 1998. The effect of selective sweeps on the variance of the allele distribution of a linked multiallele locus: hitchhiking of microsatellites. *Theoretical Population Biology* 53: 272–283.

Wierdl, M., M. Dominska, and T. D. Petes. 1997. Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics* 146:769–779.

Wild, Sarah, Gojka Roglic, Anders Green, Richard Sicree, and Hilary King. 2004. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care* 27(5):1047–1053.

Wilkie, Andrew O M. 2005. Bad bones, absent smell, selfish testes: the pleiotropic consequences of human FGF receptor mutations. *Cytokine Growth Factor Rev* 16(2):187–203.

Willems, R., A. Paul, H. G. van der Heide, A. R. ter Avest, and F. R. Mooi. 1990. Fimbrial phase variation in *Bordetella pertussis*: a novel mechanism for transcriptional regulation. *EMBO J* 9(9):2803–2809.

Williams, T. A., P. Mulatero, F. Filigheddu, C. Troffa, A. Milan, G. Argiolas, P. P. Parpaglia, F. Veglio, and N. Glorioso. 2005. Role of *HSD11B2* polymorphisms in essential hypertension and the diuretic response to thiazides. *Kidney International* 67:631–637.

Wren, J.D., E. Forgacs, J.W. Fondon, A. Pertsemlidis, S.Y. Cheng, T. Gallardo, R.S. Williams, R.V. Shohet, J.D. Minna, and H.R. Garner. 2000. Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. *American Journal of Human Genetics* 67:345–356.

Wright, S. 1937. The distribution of gene frequencies in populations. *Proceedings of the National Academy of Sciences, USA* 23:307–320.

Xita, N., A. Tsatsoulis, I. Stavrou, and I. Georgiou. 2005. Association of *SHBG* gene polymorphism with menarche. *Mol Hum Reprod* 11(6):459–462.

Xita, Nectaria, Agathocles Tsatsoulis, Anthi Chatzikyriakidou, and Ioannis Georgiou. 2003. Association of the (TAAAA)n repeat polymorphism in the sex hormone-binding globulin (*SHBG*) gene with polycystic ovary syndrome and relation to SHBG serum levels. *J Clin Endocrinol Metab* 88(12):5976–5980.

Xu, X., M. Peng, Z. Fang, and X. Xu. 2000. The direction of microsatellite mutations is dependent upon allele length. *Nautre Genetics* 24:396–399.

Yamada, N., M. Yamaya, S. Okinaga, K. Nakayama, K. Sekizawa, S. Shibahara, and H. Sasaki. 2000. Microsatellite polymorphism in the heme oxygenase-1 gene promoter is associated with susceptibility to emphysema. *Am J Hum Genet* 66(1):187–195.

Yogev, D., R. Rosengarten, R. Watson-McKown, and K. S. Wise. 1991. Molecular basis of mycoplasma surface antigenic variation: a novel set of divergent genes undergo spontaneous mutation of periodic coding regions and 5′ regulatory sequences. *EMBO J* 10(13): 4069–4079.

Yoon, S., G. Tromp, S. Vongpunsawad, A. Ronkainen, T. Juvonen, and H. Kuivaniemi. 1999. Genetic analysis of *MMP3*, *MMP9*, and *PAI-1* in finnish patients with abdominal aortic or intracranial aneurysms. *Biochem Biophys Res Commun* 265(2):563–568.

Young, E. T., J. S. Sloan, and K. Van Riper. 2000. Trinucleotide repeats are clustered in regulatory genes in *Saccharomyces cerevisiae*. *Genetics* 154(3):1053–1068.

Zeng, K., Y. X. Fu, S. Shi, and C. I. Wu. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174:1431–1439.

Zhang, B., S. Dhillon, I. Geary, W. M. Howell, F. Iannotti, I. N. Day, and S. Ye. 2001. Polymorphisms in matrix metalloproteinase-1, -3, -9, and -12 genes in relation to subarachnoid hemorrhage. *Stroke* 32(9):2198–2202.

Zhivotovsky, L. A., and M. W. Feldman. 1995. Microsatellite variability and genetic distances. *Proceedings of the National Academy of Sciences USA* 92:11549–11552.

Zhuang, Lei, Andrei V Karotki, Philip Bruecker, and Beat Trueb. 2009. Comparison of the receptor *FGFRL1* from sea urchins and humans illustrates evolution of a zinc binding motif in the intracellular domain. *BMC Biochem* 10:33.

Zhuang, Lei, Peter Villiger, and Beat Trueb. 2011. Interaction of the receptor *FGFRL1* with the negative regulator Spred1. *Cell Signal* 23(9):1496–1504.

Zuckerkandl, E. 1974. Possible role of inert heterochromatin in cell-differentiation: action of and competition for locking molecules. *Biochimie* 56:937–954.

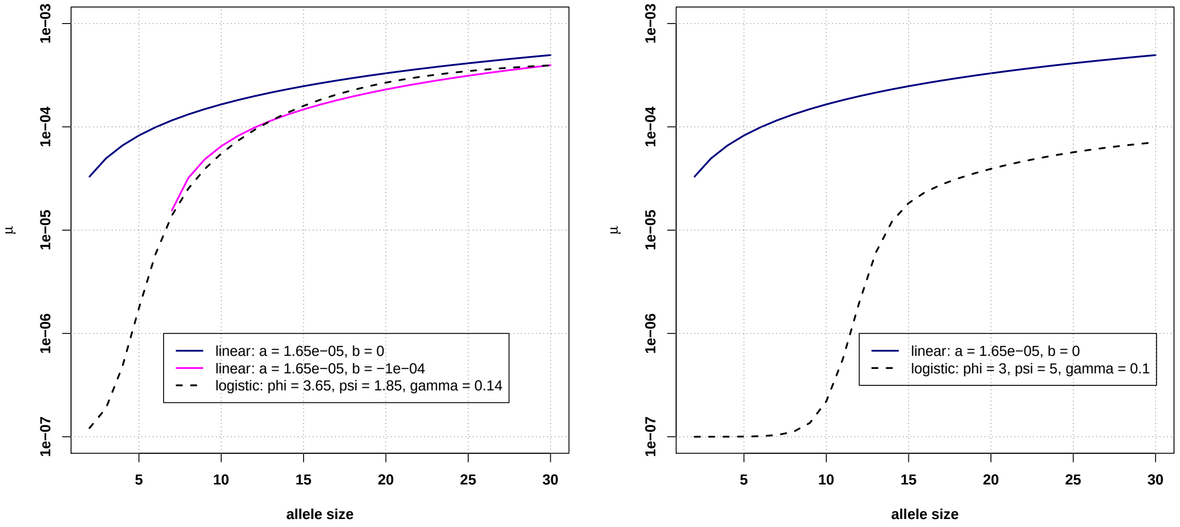

Figure A-2-1:Comparing logistic (dashed lines) and linear (solid lines) models of allele-specific mutation rate. The logistic model (dashed line) allows low but appreciable mutation rates for the smallest allele sizes, while the linear model can lead to $\mu = 0$ for certain negative mutation rates (left, magenta line). Also note that for larger allele sizes, the logistic model can largely recapitulate the mutational dynamics of the linear model. (right) Another illustration of the fact that the logistic model provides a larger dynamic range along the axis of allele size.

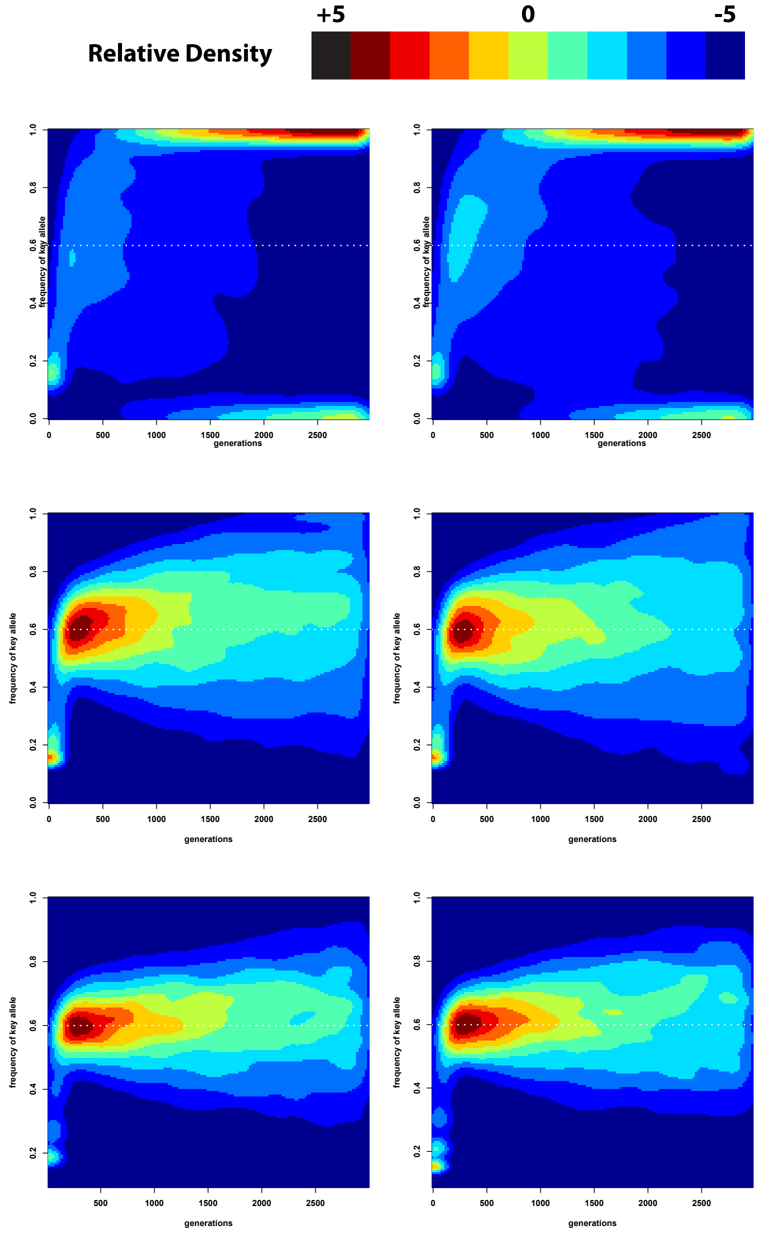

Figure A-2-2: Individual-based (left column) and recursion (right column) simulations produce near identical distributions of key allele frequency and time under identical starting conditions. Mutation, selection, and drift were modeled. Each contour map shows the joint distribution (across 1000 independent simulations) of the frequency of the key allele (also the most-fit allele) and the number of generations since selection began. Top row: $N_e = 1000$, middle row: N_e = 5000, bottom row: N_e = 10000. In the absence of drift the frequency of the key allele at mutation-selection equilibrium was 0.6008.

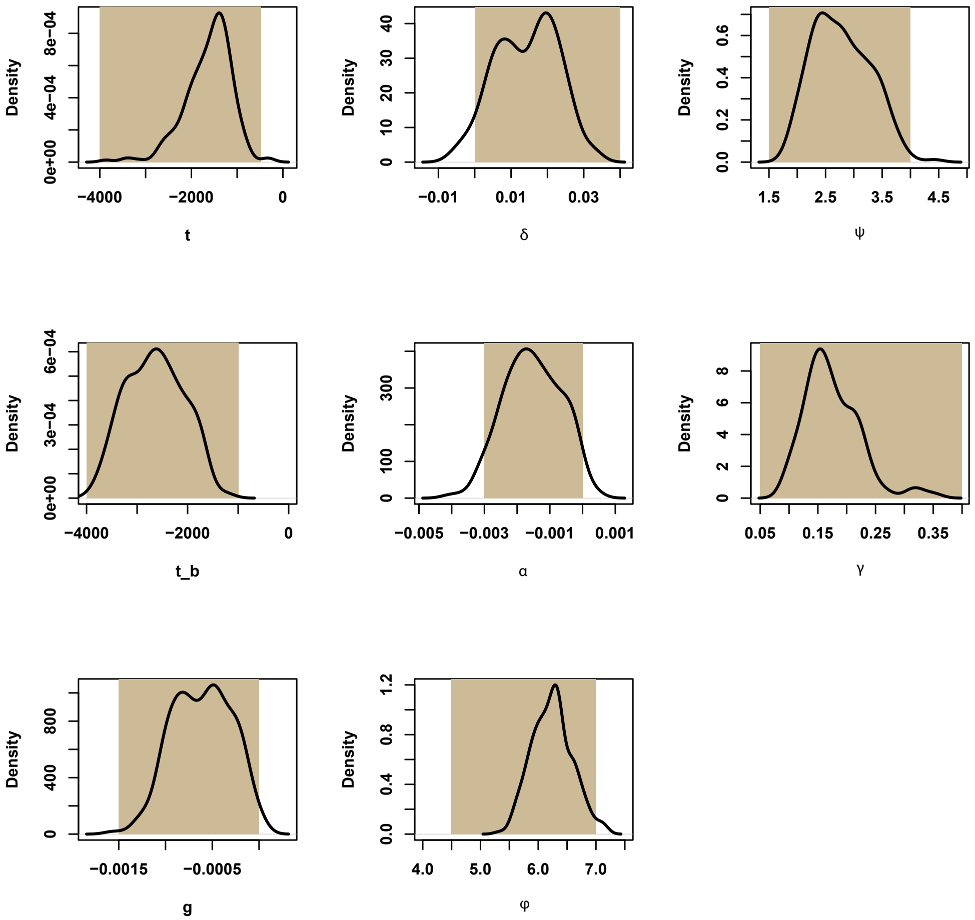

Figure A-2-3: Prior and posterior distributions for estimated parameters surrounding evolution of the microsatellite that causes Friedreich's ataxia. Tan boxes indicate the range of uniform priors for each parameter. Solid black lines are the regression-adjusted posterior densities. Because posterior distributions are regression adjusted, some parameters show small densities outside the prior range.

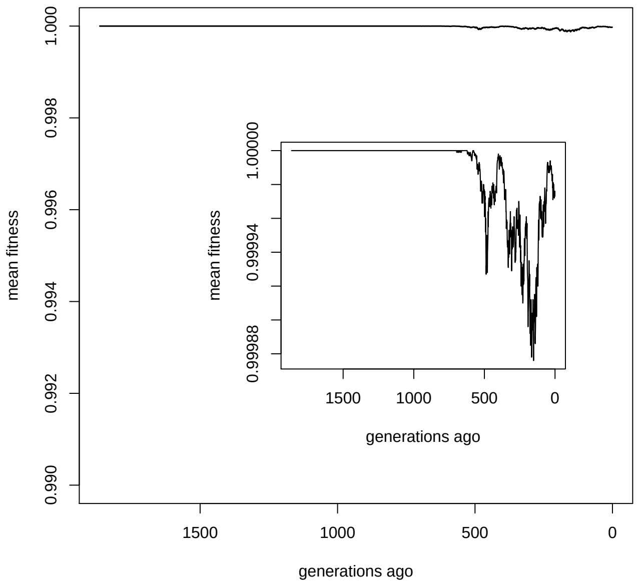

Figure A-2-4: Genetic load associated with emergence of E class alleles is minimal. Mean fitness as a function of time since emergence of the founding LN allele is shown. Minor deflections are noticeable on the main graph, but it is necessary to magnify the y-axis in order to observe these tiny deflections of mean fitness (inset). These data are drawn from a single simulation using the median posterior estimates of the relevant parameters (Table 1).

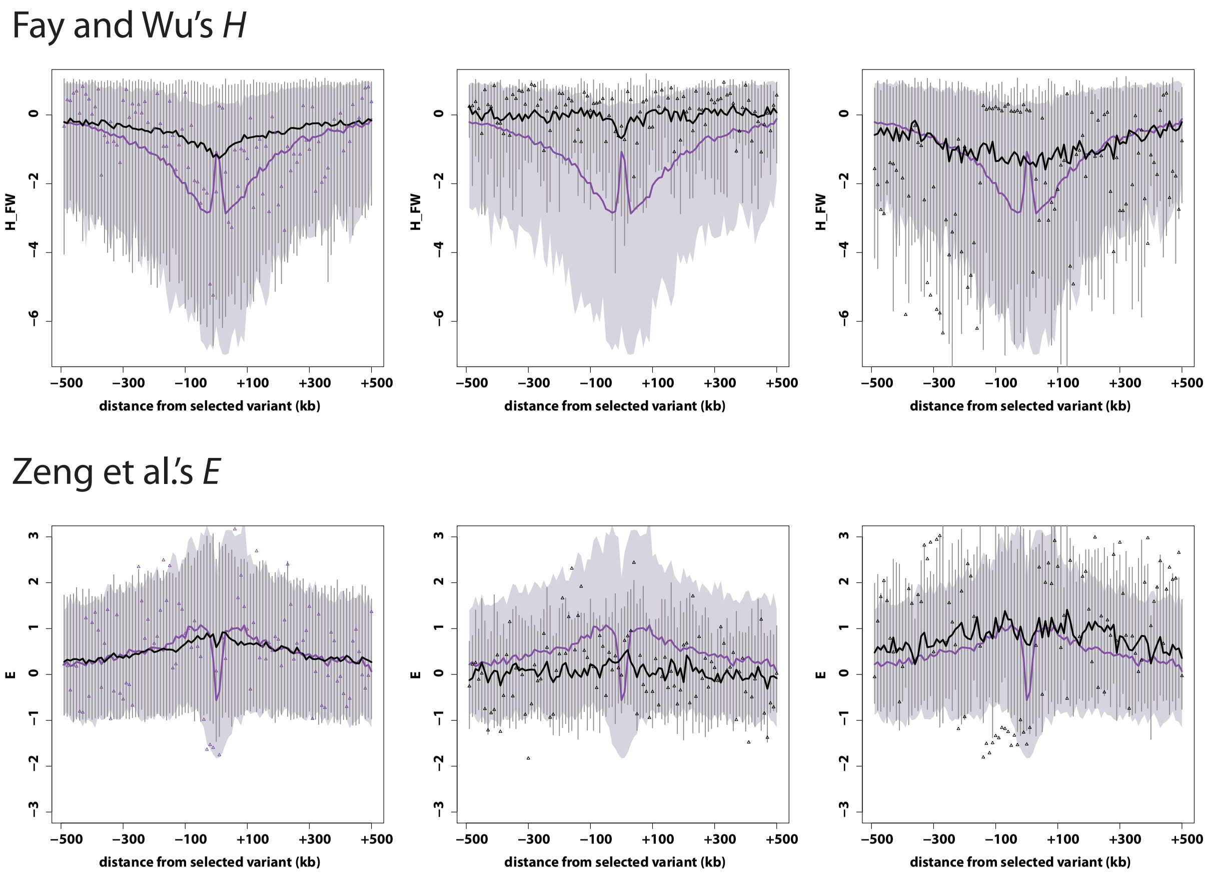

Figure A-3-1: The spatial footprint of hard sweeps and microsatellite selection measured by Fay and Wu's HFW (top row) and Zeng et al.'s E (bottom row). The first panel of each row summarizes each statistic across 500 simulation replicates of a hard sweep (s = 0:05, h = 0:5) or microsatellite selection (ϕ = 5, g = -0:05). Lines show the mean value of each statistic for hard sweeps (black) or microsatellite selection (purple). The 5%-95% interquantile range is shown as a light purple cloud (hard sweeps) or vertical bars (microsatellite selection). The middle column only includes replicates of microsatellite selection where Δ_{msat} was among lowest 10% of those recorded. The righthand column only includes replicates of microsatellite selection where Δ_{msat} was among the highest 10% of those recorded. All panels also show results for a single, representative replicate, where the value of the statistic for each windows is represented by a triangle (purple=hard sweep; black=microsatellite selection

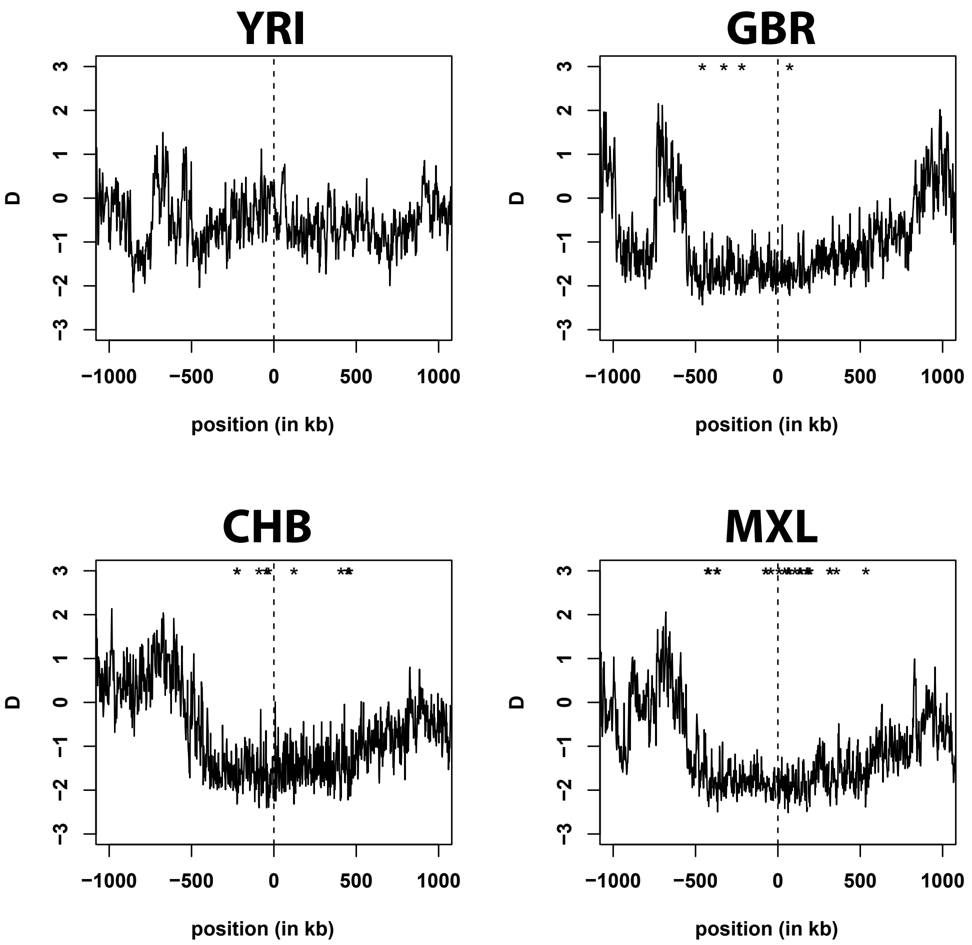

Figure A-3-2: Values of D for the 2Mb region flanking the studied HSD11B2 microsatellite, which is located at position 0 of each plot. Plotted values are for 10kb sliding windows (4kb jumps). Windows with significant values of D are marked by asterisks towards the top of each plot. From the top left and moving clockwise, the populations shown are are YRI, GBR, MXL, and CHB.

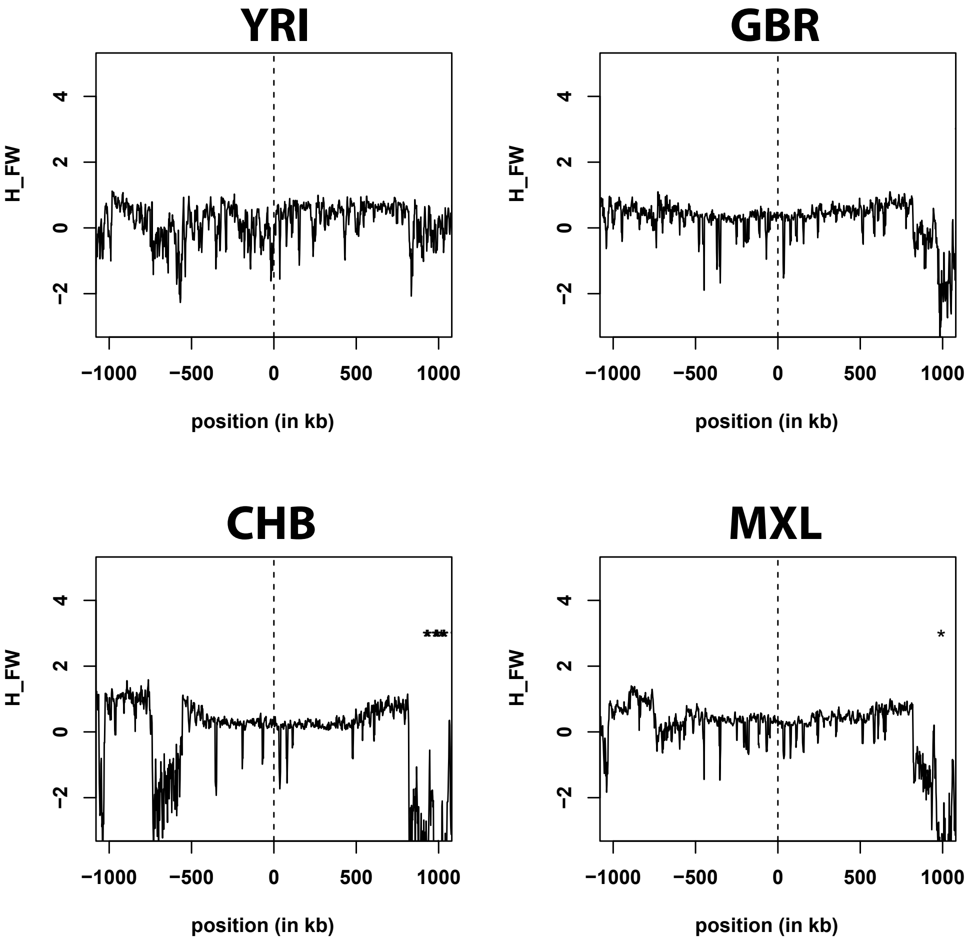

Figure A-3-3: Values of H_{FW} for the 2Mb region flanking the studied HSD11B2 microsatellite, which is located at position 0 of each plot. Plotted values are for 10kb sliding windows (4kb jumps). Windows with significant values of H_{FW} are marked by asterisks towards the top of each plot. From the top left and moving clockwise, the populations shown are are YRI, GBR, MXL, and CHB.

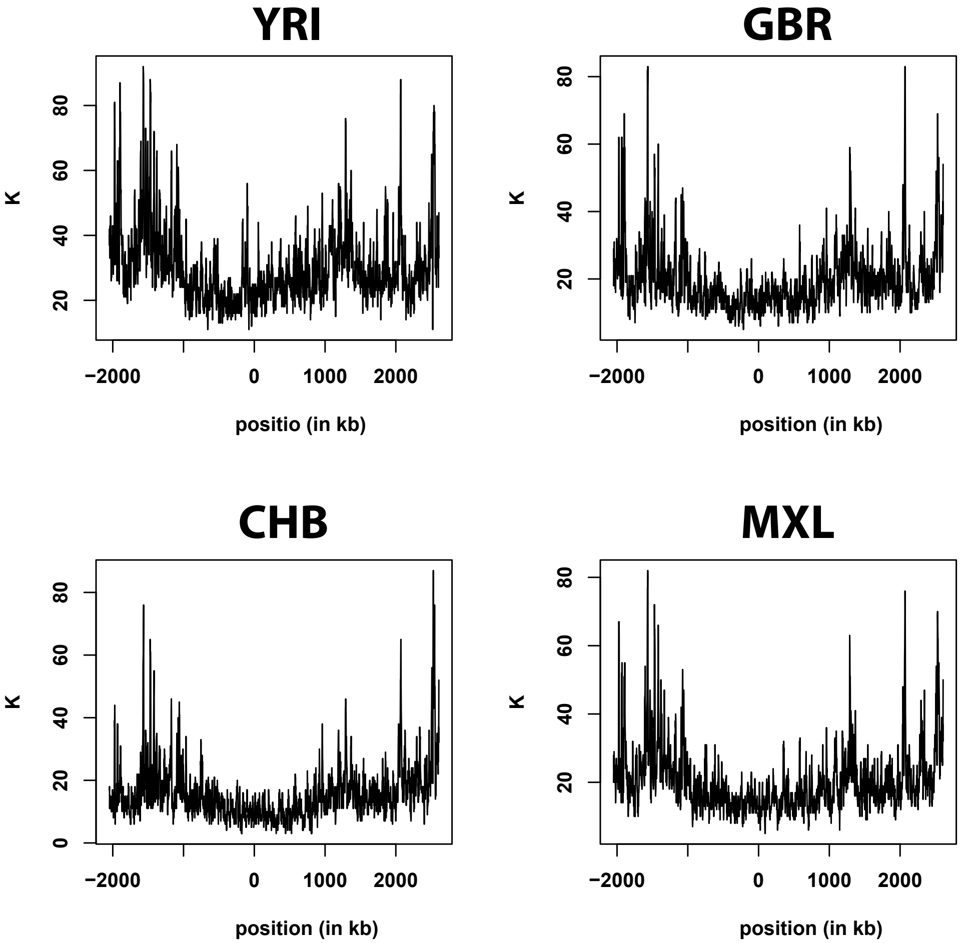

Figure A-3-4: An extended view of K for populations YRI, GBR, MXL, and CHB (clockwise from top left panel). Position is in units of 1kb.

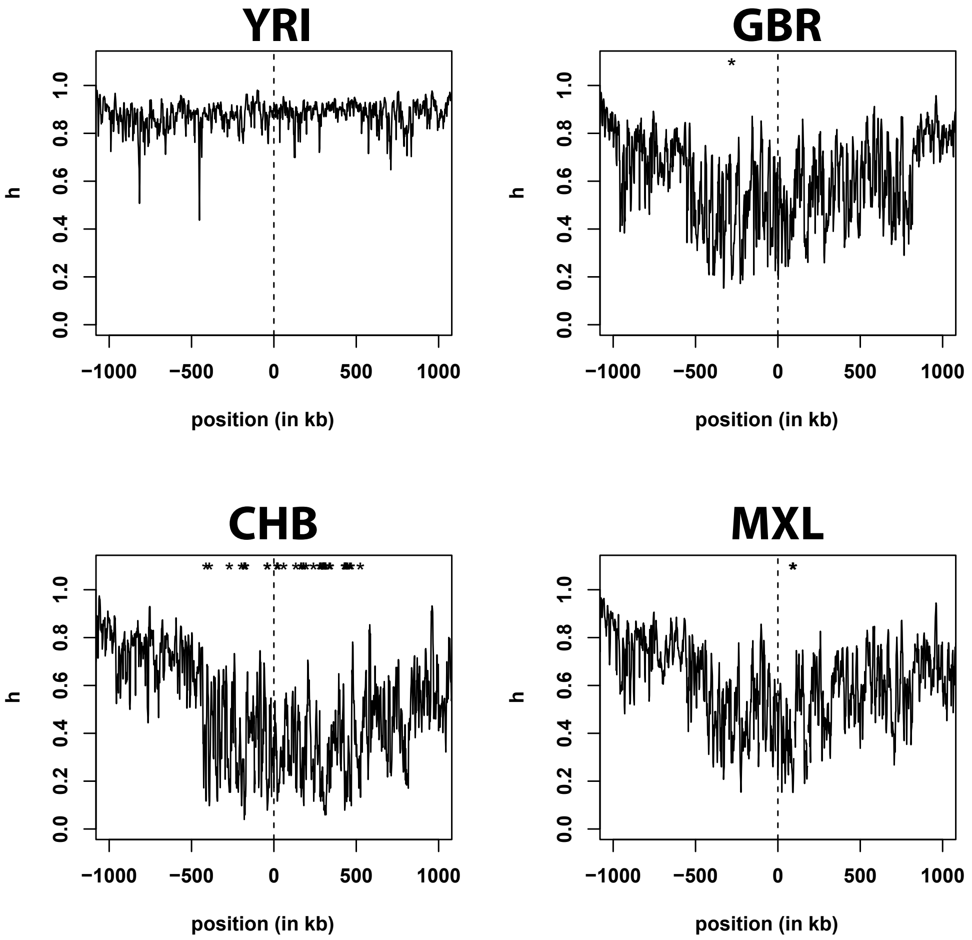

Figure A-3-5: Values of H for the 2Mb region flanking the studied HSD11B2 microsatellite, which is located at position 0 of each plot. Plotted values are for 10kb sliding windows (4kb jumps). Windows with significant values of H are marked by asterisks towards the top of each plot. From the top left and moving clockwise, the populations shown are are YRI, GBR, MXL, and CHB.

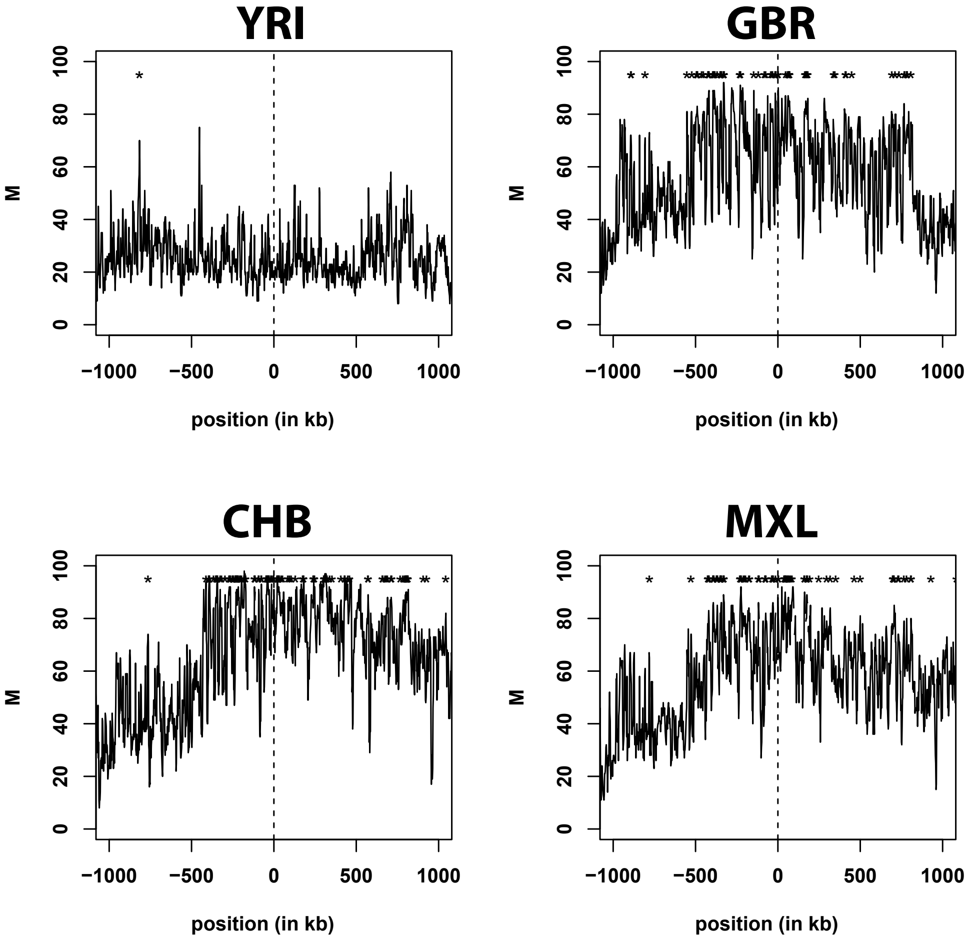

Figure A-3-6: Values of M for the 2Mb region flanking the studied HSD11B2 microsatellite, which is located at position 0 of each plot. Plotted values are for 10kb sliding windows (4kb jumps). Windows with significant values of M are marked by asterisks towards the top of each plot. From the top left and moving clockwise, the populations shown are are YRI, GBR, MXL, and CHB.

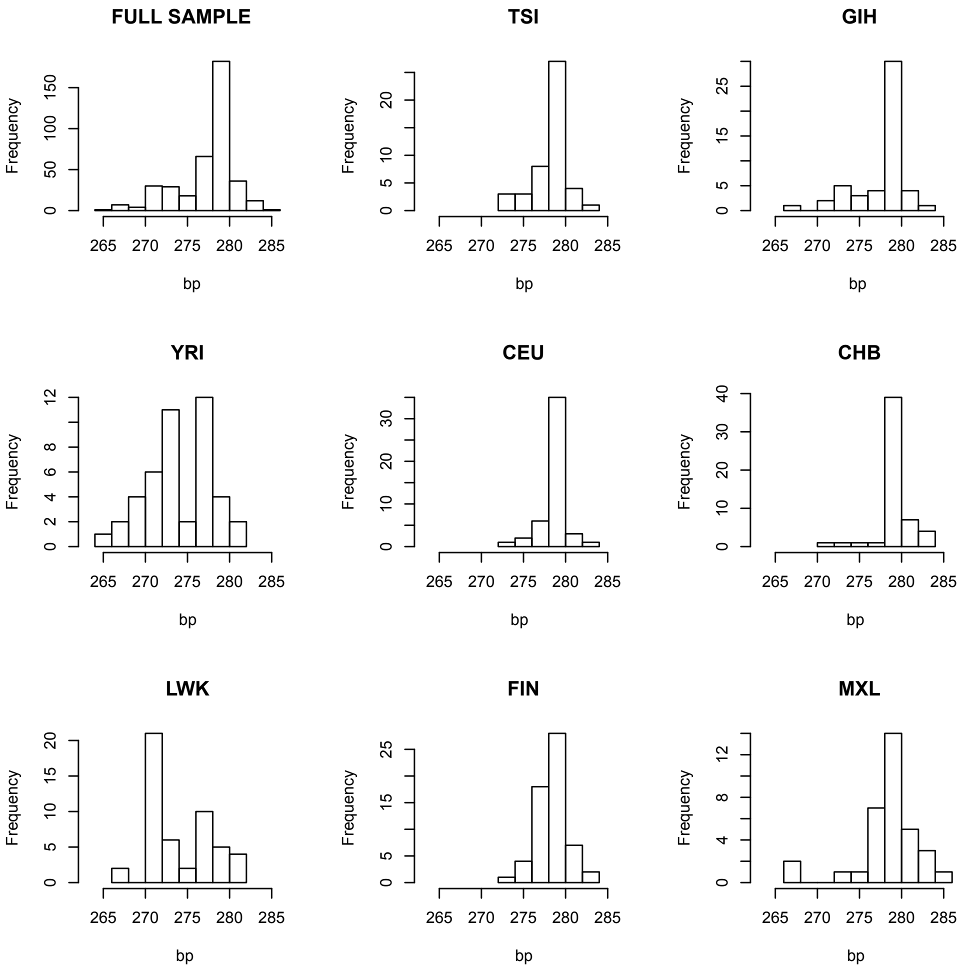

Figure A-3-7: Observed allele frequency distributions for a CA repeat in intron 1 of human gene *HSD11B2*. Raw allele counts can be found in Table 1 of the main text. Top left panel shows the allele frequency distribution for the entire sample of participants in the 1000 Genomes Project from eight populations. The other eight panels show allele frequency distributions for individual populations. Raw base pair sizes are shown. The common CAx18 allele is 279bp long.

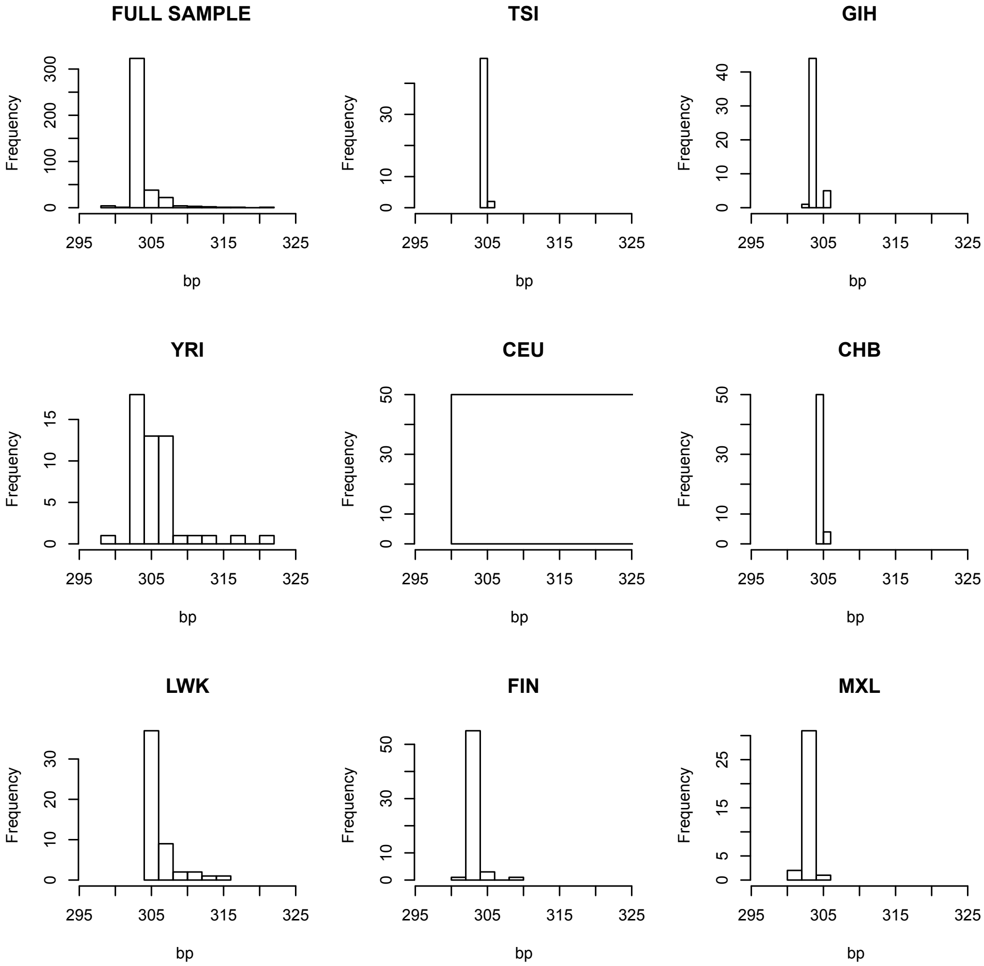

Figure A-3-8: Observed allele frequency distributions for a CA repeat 18kb downstream of human gene *SLC24A5*. Top left panel shows the allele frequency distribution for the entire sample of participants in the 1000 Genomes Project from eight populations. The other eight panels show allele frequency distributions for individual populations. Raw base pair sizes are shown. The reference sequence length for this microsatellite is 16.

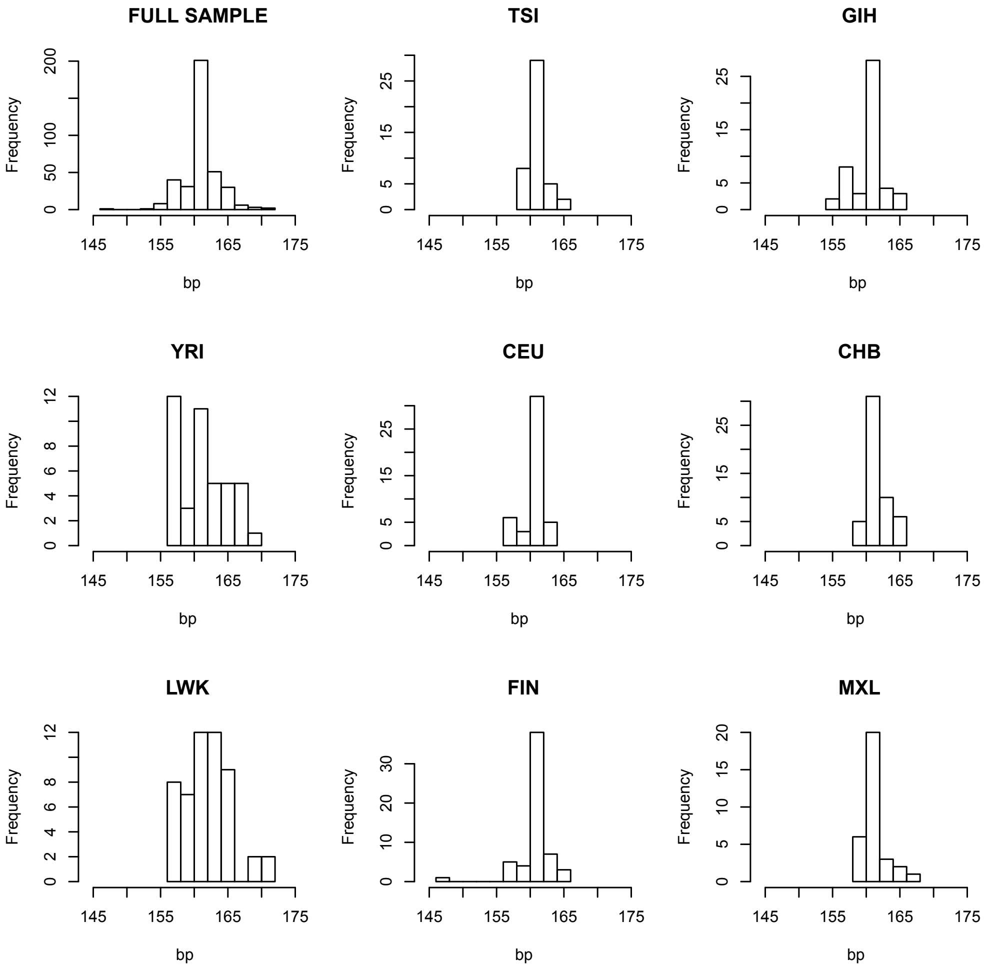

Figure A-3-9: Observed allele frequency distributions for a CA repeat 1.4kb downstream of human gene *CYP3A5*. Top left panel shows the allele frequency distribution for the entire sample of participants in the 1000 Genomes Project from eight populations. The other eight panels show allele frequency distributions for individual populations. Raw base pair sizes are shown. The reference sequence length for this microsatellite is 20.

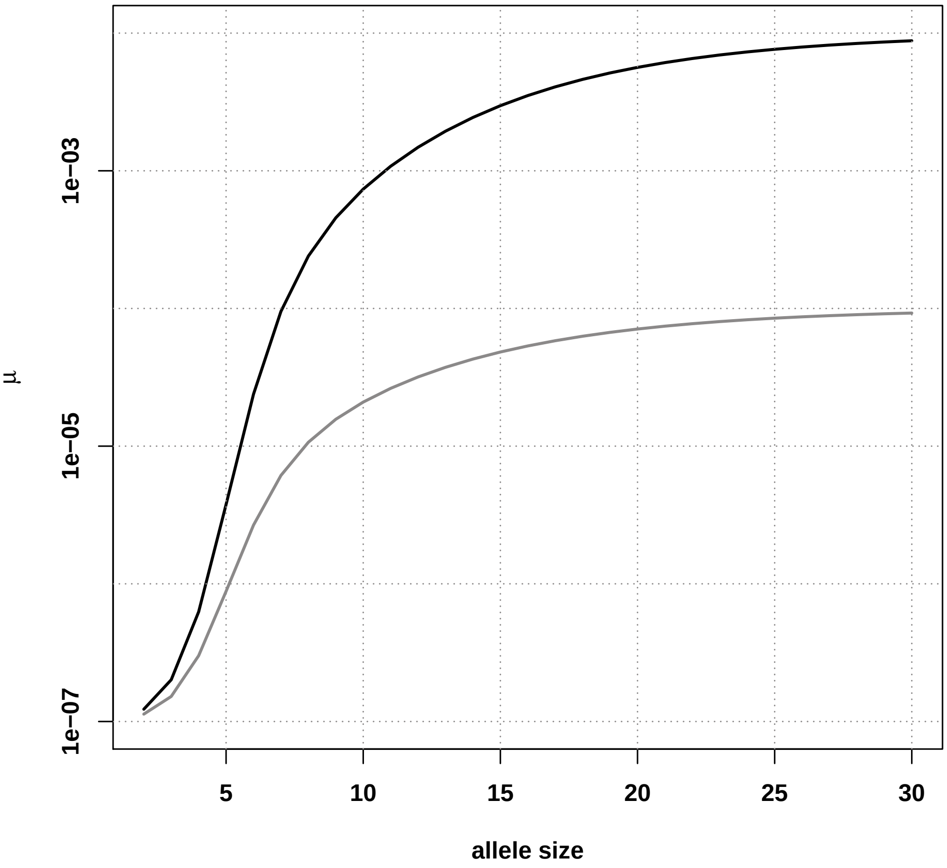

Figure A-3-10: The two mutation rates curves used in simulations of microsatellite selection. The black line corresponds to simulations where phi = 5, while the gray line corresponds to simulations where ϕ = 3.

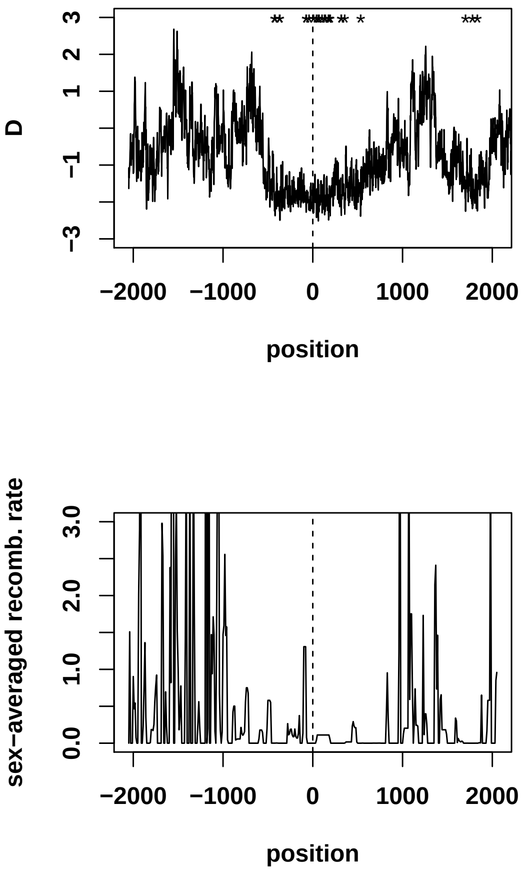

Figure A-3-11:Comparison of Tajima's D and sex-averaged recombination rate surrounding the *HSD11B2* microsatellite, which is at position 0. (top panel) Values of D for the MXL population sample are shown, with significant 10kb windows marked by asterisks. Position is in units of 1kb. (bottom panel) Sex-averaged recombination rate for the same region, estimated for 10kb windows (Kong et al., 2010). Note that values for some windows outside the central 2Mb are greater than 3. The y-axis is reduced so that variation within the central region can be seen.

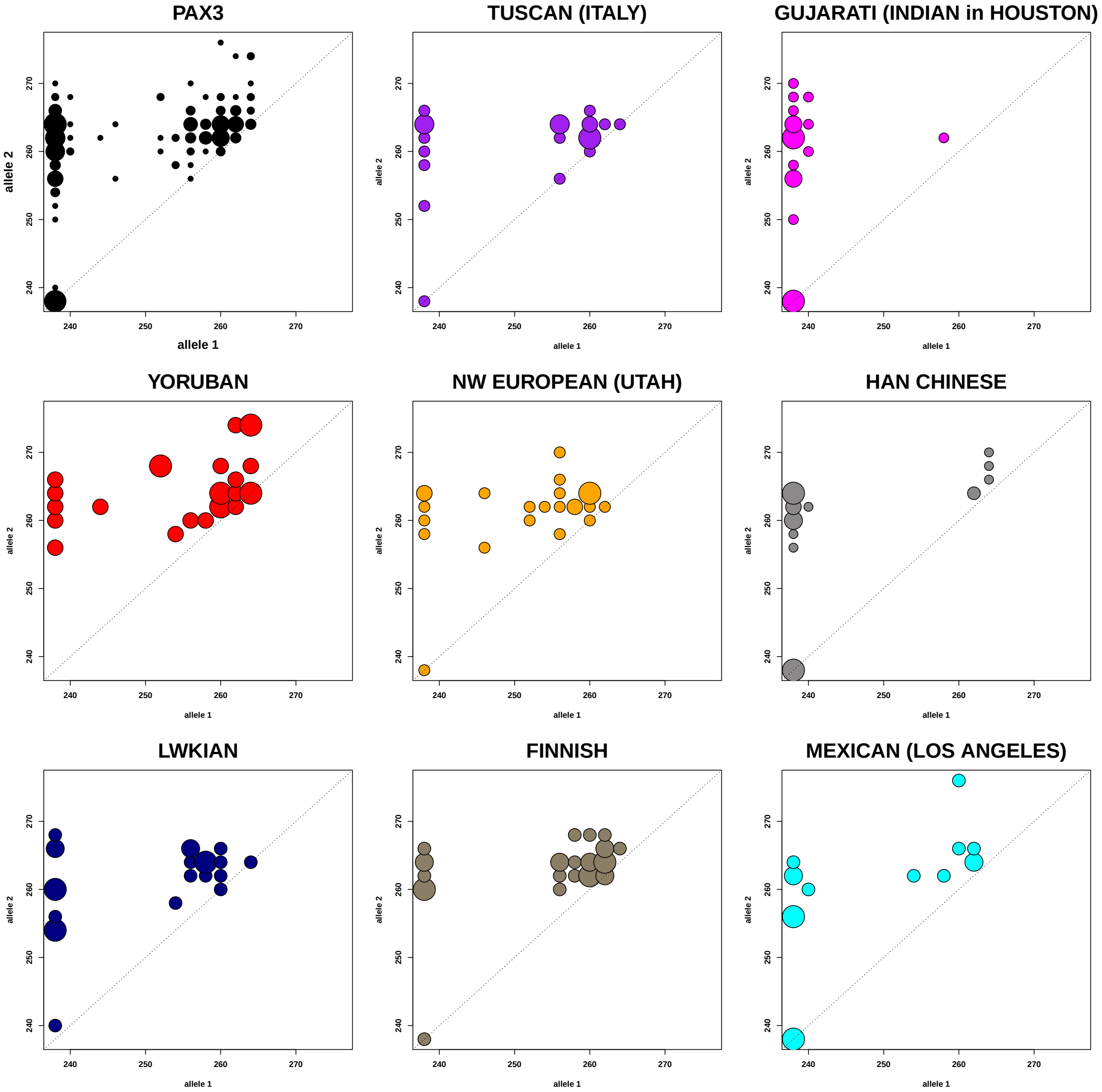

Figure A-4-1: Distribution of haplotypes of the *PAX3* promoter microsatellite. Allele sizes are given in base pairs. Size 238bp = 13x. The size of each bubble is proportional to the frequency of the haplotype in the sample. The upper-leftmost panel shows the distribution across all eight sampled populations.