Functional and Evolutionary Genetics of a Wild Baboon Population

by

Jenny Tung

Department of Biology
Duke University

Date:_____
Approved:

_____
Susan C. Alberts, Co-Supervisor

_____
Gregory A. Wray, Co-Supervisor

_____
David B. Goldstein

_____
Paul Magwene

_____
Mohamed A.F. Noor

_____
Carole Ober

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor
of Philosophy in the Department of
Biology in the Graduate School
of Duke University

2010

ABSTRACT

Functional and Evolutionary Genetics of a Wild Baboon Population

by

Jenny Tung

Department of Biology
Duke University

Date:_____
Approved:

_____
Susan C. Alberts, Co-Supervisor

_____
Gregory A. Wray, Co-Supervisor

_____
David B. Goldstein

_____
Paul Magwene

_____
Mohamed A.F. Noor

_____
Carole Ober

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the Department of
Biology in the Graduate School
of Duke University

2010

Abstract

Although evolution results from differential reproduction and survival at the level of the individual, most research in evolutionary genetics is concerned with comparisons made at the level of divergent populations or species. This is particularly true in work focused on the evolutionary genetics of natural populations. While this level of inquiry is extremely valuable, in order to develop a complete understanding of the evolutionary process we also need to understand how traits evolve within populations, on the level of differences between individuals, and in the context of natural ecological and environmental variation. A major difficulty confronting such work stems from the difficulty of assessing interindividual phenotypic variation and its sources within natural populations. This level of inquiry is, however, the main focus for many long-term field studies. Here, I take advantage of one such field study, centered on the wild baboon population of the Amboseli basin, Kenya, to investigate the possibilities for integrating functional, population, and evolutionary genetic approaches with behavioral, ecological, and environmental data. First, I describe patterns of hybridization and admixture in the Amboseli population, a potentially important component of population structure. Second, I combine field sampling, laboratory measurements of gene expression, and a computational approach to examine the possibility of using allele-specific gene expression as a tool to study functional regulatory variation in natural populations. Finally, I outline an example of how these and other methods can be used to understand the relationship between genetic variation and naturally occurring infection by a malaria-like parasite, *Hepatocystis*, also in the Amboseli baboons. The results of this work emphasize that developing genetic approaches for nonmodel genetic systems is becoming increasingly feasible, thus opening the door to pursuing such studies

in behavioral and ecological model systems that provide a broader framework for genetic results. Integrating behavioral, ecological, and genetic perspectives will allow us to better appreciate the interplay between these different factors, and thus achieve a better understanding of the raw material upon which selection acts.

Contents

List of Tables

List of Figures

Acknowledgements

I owe thanks to many people for their assistance, collaboration, support, and wisdom while I completed this work. First, many thanks to the staff of the Amboseli Baboon Research Project: Bernard, Gideon, Chris, Nkii, Moonyoi, Longida, Bro, Vivian, and Tim. Thanks especially to Raphael Mututua, Serah Sayialel, and Kinyua Warutere, who not only aided in the collection of samples used directly in these projects, but also have been instrumental in collecting behavioral and environmental field data for many years. They have been wonderful teachers. Thanks also to Mercy Akinyi, the veterinarian who supported us during our sample collection efforts, and to Tom Kariuki and the Institute of Primate Research in Karen for seconding her to us. Acknowledgements are also due to the Kenya Wildlife Service, the National Museums of Kenya, the Office of the President, Republic of Kenya, and the members of the Amboseli-Longido pastoralist communities for permission to conduct research in Amboseli.

Thanks for aid in obtaining samples from outside of Amboseli are due to the Integrated Primate Biomaterial and Information Resource and the Coriell Institute, as well as Robert Sapolsky for providing DNA from Masai Mara anubis baboons and Jeff Rogers for providing DNA from Mikumi yellow baboons.

Although the chapters contained within this dissertation are "mine", collaboration has played a deeply important role during my graduate training, and all of the work represented here benefited form close collaboration with others. Because their names cannot go on the title page, I would like to acknowledge here the contributions of Marie Charpentier, David Garfield, Samson Mutura, Olivier Fedrigo, Ralph Haygood, Alex Primus, and Andrew Bouley, among others mentioned on this page. My colleagues in the Alberts and Wray labs at Duke University, and in Jeanne Altmann's lab at

Princeton, have been vital for informing and challenging my viewpoints on baboons, evolution, behavior, and genetics and genomics. Thanks also to David Lowry, for consistently excellent feedback and conversation about evolutionary genetics throughout the last six years. And thanks to Sayan Mukherjee, who has been the embodiment of support in so many different ways, and to whom I owe a great deal for getting me over the statistical and computational roadblocks in my head.

Outside of the world of science, I owe thanks to my parents, Wae-hai and Ming Tung, and my sister, Wenny Tung Katzenstein, for their support and appreciation of my work, even though no one's kids that *they* know go off to Africa every summer to look at monkeys.

Finally, thanks to my co-advisors, Susan Alberts and Greg Wray, and my committee members, David Goldstein, Paul Magwene, Mohamed Noor, and Carole Ober, for shepherding me through this process. Although not formally in this group, many thanks also to Jeanne Altmann, for the long-term vision she has brought to bear in shaping the Amboseli Baboon project and for her support in allowing me to work in the field and with these data. Susan deserves an infinite amount of credit for her keen editorial eye, the many things she has taught me about the scientific process, and the general inspiration she offers as a successful female scientist. Greg reminds me to try always to see the bigger picture, and why that is such an important component of being a scientist.

1. Introduction

Although evolution results from differential reproduction and survival at the level of the individual, most research in evolutionary genetics is concerned with comparisons made at the level of divergent populations or species. This is particularly true in work focused on the evolutionary genetics of natural populations, which are often aimed at understanding differences between populations or species that may have been the result of natural selection (Nachman *et al.* 2003; Abzhanov *et al.* 2004; Shapiro *et al.* 2004; Colosimo *et al.* 2005; Gompel *et al.* 2005; Hoekstra *et al.* 2005; Abzhanov *et al.* 2006; Steiner *et al.* 2007; Jeong *et al.* 2008). Such work is extremely valuable for explaining the genetic basis of differences that are already fixed. However, in order to develop a complete understanding of the evolutionary process, we also need to understand how traits evolve within populations, on the level of differences between individuals, and in the context of natural ecological and environmental variation.

Assessing interindividual phenotypic variation in natural populations is challenging in the best of cases, and may be impossible for some genetic model systems. This level of inquiry is, however, the main focus for many long-term field studies. Several such studies explicitly focus on how ecological and environmental variation influences variation in adaptive traits (e.g., Grant 1986; Altmann and Alberts 2003; Altmann and Alberts 2003; Kruuk and Hill 2008). These efforts have provided some of the best estimates of both reproductive fitness in the wild and the environmental selective pressures that influence adaptively important traits. Data and analyses of this type therefore provide a rich foundation for genetic studies in the same populations. For example, recent quantitative genetic work on red deer and Soay sheep in the Scottish isles, both subjects of long-term field observation, have yielded unique new insights into

how sexual antagonism (Foerster *et al.* 2007) and changing environmental conditions (Wilson *et al.* 2006) constrain the evolution of adaptive traits within wild populations.

Genetic studies and field-based ecological and behavioral studies yield complementary levels of insight into the evolutionary process. When taken together, they can provide a more complete picture of evolution in natural populations than when considered independently. For instance, while population history can be traced through genetic data, these patterns themselves are the results of individual behavior, movement across the landscape, and reproductive success—characteristics that have historically been of great interest to field researchers. Hence, the combination of behavioral and ecological data with genetic data provides an opportunity to explain patterns of population dynamics embedded in genetic data using known patterns of individual behavior and life history, and to estimate how these individual-level effects have changed over time (Pope 1992; Nussey *et al.* 2005; Archie *et al.* 2008; Tung *et al.* 2008). For example, Pope (1992) used classical F statistics to dissect patterns of population genetic diversity in Venezuelan red howler monkeys (*Alouatta seniculus*). Because the demography and dispersal patterns of her study subjects had been intensively monitored for more than ten years, she was able to interpret these results in the light of known levels of reproductive skew, female philopatry, and male dispersal. Indeed, the measured genetic differentiation from the marker data matched very well with predictions made from the behavioral data: howler monkeys exhibited high levels of genetic differentiation between social groups, but low levels of inbreeding within groups, due to a social structure that combined small sets of related females with one or two unrelated males (Pope 1992). These outbred, substructured groups may act as reservoirs for allelic variation, making howler monkey populations and other, similarly structured populations more successful at retaining genetic variation than predicted by their relatively small census sizes.

Similarly, having access to genetic data as well as to behavioral and ecological data makes it possible to examine both the genetic and environmental contributions to traits of interest, as well as context-dependence between the two. Indeed, while studies conducted in the lab can reveal whether genetic variation has functional potential in controlled environments, complementary studies on natural populations are needed to establish whether this variation is ecologically and evolutionarily relevant. For example, a recent study in red wolves demonstrates that gene expression levels of hundreds of transcripts significantly differ between free-ranging animals and captive animals, particularly for genes involved in the stress response (Kennerly *et al.* 2008). Such results strongly imply that individuals sampled under natural conditions (especially when these conditions are well described) need to be included in functional genetic studies. Studies focused on natural populations may also be important for capturing functional effects that never manifest themselves in captivity, including genetic effects that are only measurable within certain kinds of environments (i.e., gene-environment interactions). A recent study in yeast, for example, estimated that as many as 47% of gene expression phenotypes exhibited evidence for GEI across two different environments (Smith and Kruglyak 2008). Additionally, specific cases of gene-environment interaction for organism-level complex traits have been well documented across many species, including *Drosophila* (Leroi *et al.* 1994; Gurganus *et al.* 1998; Leips and Mackay 2000), *C. elegans* (Shook and Johnson 1999), humans (reviewed in Hunter 2005), and non-human primates (Barr *et al.* 2003; Barr *et al.* 2004; Newman *et al.* 2005). The extensive behavioral and ecological data available for long-term field populations may provide important clues to the types of environmental effects that play a role in GEIs in nature.

Finally, field data can complement sequence-based signatures of selection by confirming the fitness effects of a given trait and providing a mechanism for selection on that trait. For example, most species of New World monkeys exhibit "allelic

trichromacy" conferred by a set of X-linked polymorphisms for dichromatic (i.e., red-green color blind) or trichromatic vision (Surridge *et al.* 2003). The phenotypic consequences of this variation for color vision are well established, and genetic evidence strongly suggests that these polymorphisms have been maintained by natural selection (Surridge and Mundy 2002; Hiwatashi *et al.* 2009). However, the actual selective advantages that produced these patterns have been difficult to understand (Melin *et al.* 2008). Recent field studies have shed important new light on this question, demonstrating for the first time that dichromats and trichromats experience different kinds of foraging advantages in the wild (Melin *et al.* 2007; Vogel *et al.* 2007; Melin *et al.* 2008). Results of this kind are exciting because they embody one of the first cases of a tantalizing possibility: that by combining field data and genetic data, we might one day understand not only the genetic and environmental architecture of a trait, but also its adaptive relevance within natural populations.

For most systems and for most traits, moving towards this direction will require significant groundwork in the coming years. My thesis work has been focused on laying this groundwork for one such system, the well-studied baboon population of the Amboseli basin, Kenya. In this introduction, I describe the background context in which this work has taken place, including the motivation to focus on this system. I then summarize how each of the subsequent four chapters is situated within this overall framework. Finally, I provide my perspective on the work as a whole, taking the opportunity to consider possible directions that genetic work on both this system and other similar populations may take in the near future.

1.1 Genetic research on wild nonhuman primates: motivations and challenges

Our closest living relatives, the nonhuman primates, are perennial subjects of public and scientific fascination because they occupy a unique place in evolutionary

biology and ecology. The striking similarities we share with other primates make them important models for human physiology, behavior, and health. At the same time, variation among primate species provides a rich basis for comparative work. Such work is critical for both explaining the common threads that tie primates together and the differences that make specific branches of the primate tree, including the human lineage, unique.

Three separate motivations position primates as good systems for integrating genetic data with behavioral and ecological work. First, detailed observational field studies have a long history within nonhuman primates. Many of these studies have focused on variation between known individuals within a larger population. As a result of these efforts, extremely fine-grained information on the behavior, life history, and environmental milieu of each individual are now available for some species (and for a few species, these kinds of information are available for multiple populations). These data provide an excellent context for genetic studies. Second, genetic and genomic resources for primates are increasing at a rapid rate (Enard and Paabo 2004; Siepel 2009). With the upcoming release of a set of new primate genomes, primates will soon be one of the most genomic data-dense clades among all animals, especially considering the relatively recent divergence times in this group (compared with, for instance, drosophilids, for which the sequenced species last shared a common ancestor at a time comparable to the radiation of mammals: Clark *et al.* 2007). These resources have already improved, and will continue to improve, the feasibility of genetic studies. Finally, as alluded to above, because nonhuman primates are our closest living relatives, conducting such studies within natural primate populations will provide important keys to understanding the evolutionary history of our own species. This will hold particularly true for closely related species, such as the other great apes, and for species that share a

similar ancestral ecology to humans, such as baboons and other African savannah primates (Jolly 2001).

The development of tools for conducting genetic work on nonhuman primate populations is still in progress. Thus, with a few notable exceptions (reviewed in Vigilant 2009; Vigilant and Guschanski 2009), genetic studies of population history and functional genetic variation have largely progressed along parallel tracks from behavioral and ecological studies of wild primates. Within functional genetics, most nonhuman primate research has been based on either comparisons between aligned sequenced genomes (Bustamante *et al.* 2005; Nielsen *et al.* 2005; Pollard *et al.* 2006; Prabhakar *et al.* 2006; Haygood *et al.* 2007; Prabhakar *et al.* 2008) or on comparisons of gene expression profiles from a small number of captive individuals in a small number of species (Enard *et al.* 2002; Khaitovich *et al.* 2005; Gilad *et al.* 2006; Blekhman *et al.* 2008; Babbitt *et al.* 2010). In contrast, most genetic work on nonhuman primate populations has been focused on noninvasive sampling and genotyping techniques aimed at assigning parentage and estimating relatedness, an approach usually aimed at answering outstanding questions in behavioral ecology, not in genetics (see for example Vigilant *et al.* 2001; Buchan *et al.* 2003; Alberts *et al.* 2006; Archie *et al.* 2006). More recently, population genetic and genomic studies have become more closely aligned with the dynamics of natural populations. Genetic data have recently been leveraged to estimate historical patterns of hybridization, gene flow, and population contraction and expansion for gorillas (Yu *et al.* 2004; Thalmann *et al.* 2007), chimpanzees (Yu *et al.* 2003; Won and Hey 2005; Becquet *et al.* 2007; Becquet and Przeworski 2007), and macaques (Hernandez *et al.* 2007; Stevison and Kohn 2009), providing a novel perspective on the long-term evolution of these species. However, even for species in which substantial genomic information is already available, existing information on population genetic variation largely does not overlap with those populations that are best characterized

from an ecological and behavioral point of view. Hence, the contribution of wild primates to this work often extends no further than the contribution of the DNA samples themselves.

In order to exploit the potential of natural primate populations for genetic work, then, several obstacles need to be overcome. First, for field populations of interest, a basic understanding of population structure and its sources should be developed. Population structure defines the distribution of genetic variation in the population, and therefore influences its evolutionary capacity. For primates, population structure may be highly dynamic (for example, due to change in rates of hybridization and admixture: Tung *et al.* 2008). Many primates exhibit socially complex and ecologically flexible behavioral patterns that govern social grouping patterns, dispersal between groups, and rates of reproductive skew within groups (Altmann *et al.* 1996). These factors will influence both the analysis and results of population genetic and functional genetic work. Second, methods that are applicable to studying functional genetics in natural populations need to be developed. Studying the genetics of natural primate populations is challenging, as conventional approaches involving controlled crosses or forward genetics cannot be applied. Neither are the resources and data in place to conduct the very large scale mapping and resequencing studies currently being conducted in humans, the most closely related "model system." Finding appropriate technical and analytical methods for meeting this challenge will therefore be crucial in order to develop these model ecological and behavioral field systems into ecological and evolutionary genetic and genomic systems as well.

The work presented in this thesis addresses these challenges using a combination of field-based, lab-based, and computational methods. It also provides a window into the types of findings that can result from applying these methods to well-studied field populations. I have structured this work into four chapters. In chapter 2, I describe one

aspect of recent efforts to understand population structure and its sources in the Amboseli baboon population. Chapter 2 focuses on the contribution of admixture between the dominant species in Amboseli, the yellow baboon (*Papio cynocephalus*) and its congener, the anubis baboon (*P. anubis*). In chapters 3 and 4, I investigate the possibility of using measurements of allele-specific gene expression (i.e., allelic imbalance) as a tool to study functional regulatory variation in natural populations. Finally, in chapter 5, I outline an example of how these and other methods can be used to understand the relationship between genetic variation and phenotypic variation in an ecologically relevant complex trait.

1.2 The Amboseli baboon population

Most of the work presented here, with the exception of chapter 4, explicitly focuses on a natural population of baboons (*Papio cynocephalus*) that ranges in the Amboseli basin of southern Kenya. This population represents one of the most extensive field studies of nonhuman primates, and indeed of any animal system, to date (Altmann and Altmann 1970; Altmann *et al.* 1996; Buchan *et al.* 2003; Silk *et al.* 2003; Alberts *et al.* 2006); see also www.princeton.edu/~baboon). Initial work on the population commenced in the early 1960's, and continuous monitoring of individual baboons began in 1971. The resulting data set includes over 1500 unique, individually recognized animals across five to six baboon generations. As a direct result of this work, we now know a great deal about the behavior, physiology, and life history of these animals, including many of the social and abiotic environmental effects that shape lifetime fitness within the population (Alberts and Altmann 1995; Alberts and Altmann 1995; Alberts *et al.* 2003; Altmann and Alberts 2003; Altmann and Alberts 2003; Silk *et al.* 2003; Alberts *et al.* 2006; Beehner *et al.* 2006; Charpentier *et al.* 2008; Charpentier *et al.* 2008).

Currently, five social groups of baboons are intensively monitored within Amboseli; a few additional "non-study" groups are observed on a less regular basis. Social groups (synonymous with "breeding groups" in the population genetics literature: (Chesser 1991; Sugg *et al.* 1996) are the major unit of organization among baboons, and comprise a mixed set of adult males, adult females, and juveniles and infants of both sexes. Whereas males disperse from their natal groups upon maturity, females remain in the same social group throughout their lives, barring relatively rare group fission events (Van Horn *et al.* 2007). Both male baboons and female baboons are organized in hierarchical, sex-specific linear dominance hierarchies. Females inherit their dominance ranks from their mothers, making the matriline the major unit of organization for females within groups (Hausfater *et al.* 1982). In contrast, males directly compete to improve their dominance ranks. A male's dominance rank therefore depends on his health and competitive fighting abilities, and tends to increase after maturation, to reach its peak during his prime, and then to decrease afterwards as he ages (Packer *et al.* 2000; Alberts *et al.* 2003). Male rank plays an important role in determining reproductive opportunities with estrus females (Alberts *et al.* 2003; Alberts *et al.* 2006; Charpentier *et al.* In prep). Female philopatry, male dispersal, and (primarily male-mediated) reproductive skew therefore all influence the distribution of phenotypic variation and genetic variation in the Amboseli population.

Prior genetic work in the Amboseli baboon population has largely focused on making paternity assignments or estimating relatedness using genotypes obtained from noninvasive sampling (Alberts *et al.* 2003; Buchan *et al.* 2003; Alberts *et al.* 2006). These efforts have provided important resources for the work presented here, particularly for the estimates of hybridization and admixture presented in Chapter 2, and for outlining other sources of population structure within the Amboseli population (not presented here). Several earlier population genetic and functional genetic studies have also been

conducted on the baboons. Storz *et al.* used a panel of highly variable microsatellite markers to obtain estimates of current and ancestral effective population sizes, N_e, and of change in population size over time during the Pleistocene (Storz *et al.* 2002; Storz *et al.* 2002). Loisel *et al.* (Loisel *et al.* 2006; Loisel 2007) conducted the first functional genetic work on the Amboseli baboons, focusing on *cis*-regulatory genetic variation and coding sequence variation in the *DQA1* and *DQB* genes of the baboon major histocompatibility complex. Their results demonstrated high levels of genetic diversity within the population at these two loci, and long-term *trans*-specific selection on functionally variable regions in the *DQA1 cis*-regulatory region. These earlier analyses made significant contributions to the present work through logistical contributions to DNA sample curation (Loisel *et al.* 2006; Loisel 2007), by supplying useful estimates for population genetic parameters (particularly Storz *et al.* 2002), by helping to lay early emphasis on the possibilities for *cis*-regulatory sequence and gene expression analysis in this population.

1.3 Chapter 2: hybridization and admixture in the Amboseli baboons

Hybridization between closely related primate species has long been recorded in captivity, and observations of hybridization in the wild have now also been made for all of the major primate lineages (Arnold and Meyer 2006). Indeed, hybrid or partially hybrid origins have been suggested for at least three extant primates: the bear macaque (*Macaca arctoides*) (Tosi *et al.* 2000), the kipunji (*Rungwecebus kipunji*) (Burrell *et al.* 2009; Zinner *et al.* 2009), and most provocatively, for either humans or chimpanzees (Patterson *et al.* 2006), but see McVicker *et al.* 2009; Presgraves and Yi 2009). These claims have varying degrees of support, but highlight the possibility that hybridization may facilitate creative evolutionary processes within primates, as has already been demonstrated in other taxa, most notably sunflowers (Rieseberg 1997; Rieseberg *et al.*

2003) and butterflies (Mavarez *et al.* 2006). Contributions of hybridization could include introgression of advantageous alleles across taxon boundaries or assimilation of incipient species by secondary contact. These scenarios predict that admixture would make a measurable contribution to population structure within hybridizing species, and especially within hybrid zones.

The Amboseli basin is located along the border between the range of yellow baboons, which extends to the east and to the south, and the range of anubis baboons, which extends to the north and to the west (see Figure 1 in Zinner *et al.* 2009). Interestingly, at the outset of long-term fieldwork in Amboseli, no evidence of admixture was detected. The recent wave of hybridization in Amboseli, characterized by immigration of anubis baboon males into the basin, began in the early 1980's (Samuels and Altmann 1986). Anubis baboons and yellow baboons are morphologically distinct as well as geographically distinct, and may also exhibit important behavioral and ecological niche differences (Jolly 1993). Hence, this population provides an opportunity to study the effects of hybridization on phenotypic variation over time, as well as the behavioral and demographic patterns that influence hybridization events themselves. Additionally, it raises a question about the degree to which admixture between yellow baboons and anubis baboons contributes to overall population structure within the Amboseli basin. Establishing an objective genetic identification system for hybrid admixture would greatly aid in these analyses.

Prior work on hybrids in Amboseli produced a morphological hybrid score index based on a series of seven morphological characteristics (all of which could be scored through noninvasive observation) that generally differ between yellow and anubis baboons (Alberts and Altmann 2001). Chapter 2 builds on this work by developing and validating a complementary system for assessing admixture using highly variable microsatellite markers. These markers exhibit significant levels of differentiation between

yellow baboons and nearby anubis baboon populations, much higher than between yellow baboon populations at comparable distances (St George *et al.* 1998). The advantages of this approach lay in its freedom from observer bias and from the use of morphological characteristics that may reflect ancestry at one or a few loci in the genome, but which may not provide good genome-wide estimates. Additionally, because the resulting "genetic hybrid score" is based on a Bayesian mixture modeling approach (Pritchard *et al.* 2000; Falush *et al.* 2003), it provides direct estimates of uncertainty in the hybrid score (based on the posterior distribution for each individual's estimate) that could not be obtained using morphological traits.

As described in chapter 2, I was able to assign genetic hybrid scores to 450 individual baboons born in Amboseli between the late 1960's and the early 2000's (Tung *et al.* 2008). The distribution of these scores revealed how hybridization has altered the composition of the Amboseli population over the last several decades; in particular, individuals that exhibit a detectable level of anubis admixture have rapidly increased during this time. Admixture levels correlate with life history timing (more anubis-like animals reach social and physical maturation earlier, especially males: (Charpentier *et al.* 2008) and influence mating behavior in Amboseli (anubis-like males have a general advantage is obtaining mates, although assortative mating by genetic background also occurs: Charpentier *et al.*, in prep). The rapid increase in hybrids within Amboseli may therefore reflect underlying selective effects on introgressed "anubis" alleles. Hence, although the original motivation for this work was to tease out the contribution of admixture to (neutral) population structure, it has since made an important contribution to understanding the genetics of adaptively important traits in Amboseli as well.

In an interesting twist, although the results described in Chapter 2 reveal that hybridization has changed the population genetic landscape of Amboseli over the last few decades, they also demonstrate that admixture is in fact not the main source of

population structure within Amboseli. Indeed, in order for the genetic hybrid scores to accurately reflect admixture proportions, I had to constrain the mixture model analysis to make assignments for animals of unknown background using genotype data from anubis baboons of known provenance. When using only unlabeled data, neither a mixture model approach nor a principal components approach for identifying population structure strongly reflect patterns of hybridization and admixture (Figure 1). Rather, the social group is the dominant source of genetic structure in this population, although this effect varies depending on recent rates of reproductive skew, dispersal, and group fission events. This result roughly confirms theoretical predictions that emphasize that, in socially complex animals like baboons, behavioral, life history, and demographic factors will strongly inform patterns of population structure (Chesser 1991; Sugg *et al*. 1996). Additionally, because these factors themselves can change over time, population structure will tend to be a dynamic phenomenon. In the long run, then, the contribution of anubis admixture to population structure may be more observable through its influence on adaptively important phenotypic variation than through neutral processes alone. Indeed, admixture has already altered the distribution and underlying sources of variation in maturation timing and mating behavior in this population (Charpentier *et al*. 2008; Charpentier *et al*. In prep).

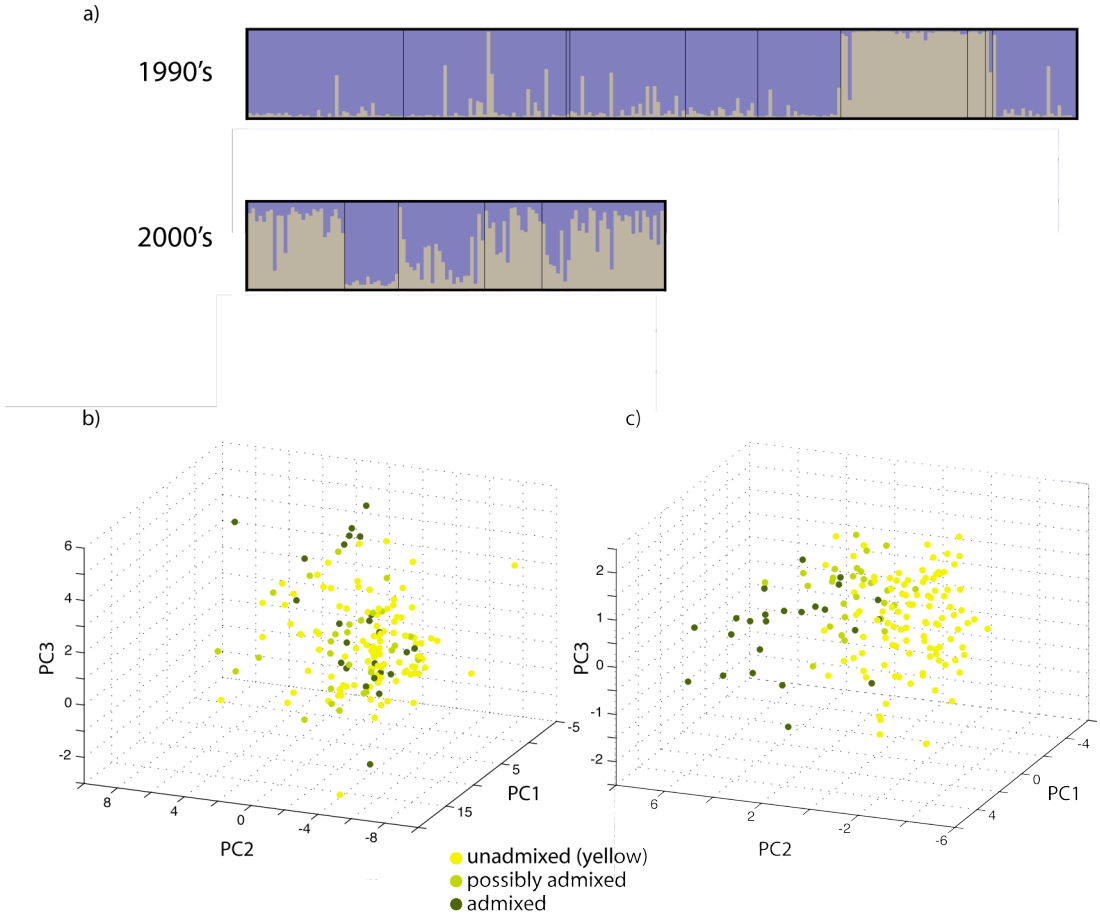

Figure 1: The dominant aspect of population structure in the Amboseli baboons arises from social grouping patterns, not anubis-yellow admixture. a) Output from K = 2 runs in Structure for individuals born in the population in the 1990's (top; n = 229) and 2000's (bottom; n = 116), based on 14 polymorphic microsatellite loci. Each colored vertical line represents one individual; thin black lines divide different social groups. The dominant pattern of structure in the 1990's comes from low rates of immigration and emigration from a few social groups colored primarily in gray; the dominant pattern of structure in the 2000's comes from recent group fission of a single group into two smaller groups (colored primarily in purple), roughly along matrilineal lines. Hybrid individuals occurred in all social groups at this time. b) Results from a principal components analysis of 232 SNPs (n = 155 individuals) also do not separate hybrids from yellow baboons. c) However, the SNP data do contain information about hybrid background, based on a PCA-like analysis (sliced inverse regression) of the same data set that focuses explicitly on ancestry-informative genetic variation.

1.4 Chapters 3 and 4: measuring and predicting allele-specific gene expression in primates

Functional genetic studies conducted in natural populations test the relevance of genetic variation to phenotype within a natural context. In well-studied long-term populations like the Amboseli baboons, they also permit individual-specific environmental effects to be included in functional analyses. Such work contributes to our understanding of how traits evolve in the context of complex social structures that influence both the distribution and propagation of functional variation, and also the manner in which it is expressed. Functional data are useful additions to genotype-phenotype work in natural populations because sample sizes for these systems are generally relatively small. Thus, functional data offer an additional layer of support to suggestive associations between genotype and phenotype. Unfortunately, most established functional genetic methods applied in laboratory settings cannot be translated directly to natural populations because of the need for invasive manipulations and/or controlled crosses.

One approach for extending functional genetic work to such populations is to focus on the contributions of regulatory genetic variation, which has a natural functional read-out, gene expression. *Cis*-regulatory genetic variation, which refers to regulatory sequence that influences only the copy of the gene on the same physical chromosome, is of particular interest because such variation can be more readily localized than its *trans*-regulatory counterpart (which influences both alleles of the gene and can reside anywhere in the genome: Yan *et al.* 2002; Wittkopp *et al.* 2004; a commonly cited example is functional variation within the coding sequence for an upstream transcription factor). Methods for measuring the proximal functional consequences of *cis*-regulatory variation, gene expression, are well established and generalize well across systems. Additionally, it has long been proposed that regulatory genetic variation plays an

important role in the evolution of adaptively relevant traits, and explicitly in the evolution of our own species (King and Wilson 1975; Wray 2007). These hypotheses have garnered some empirical support in recent years, with findings that illustrate a direct connection between regulatory genetic variation, gene expression, and evolutionarily important trait variation (Abzhanov *et al.* 2004; Shapiro *et al.* 2004; Colosimo *et al.* 2005; Gompel *et al.* 2005; Abzhanov *et al.* 2006; Prud'homme *et al.* 2006; Steiner *et al.* 2007; Tishkoff *et al.* 2007; Jeong *et al.* 2008). Thus, functional genetic work conducted on gene expression traits may, in some cases, serve as a connection to downstream, organism-level phenotypes. Indeed, several of the success stories in genetic work on ecologically well-characterized populations have revealed an important role for regulatory genetic variation. For example, beak shape in Darwin's finches appears to be in part determined by allelic variation that influences gene expression in two genes, calmodulin and *BMP4* (Abzhanov *et al.* 2004; Abzhanov *et al.* 2006).

Because gene expression is influenced by environmental effects as well as genetic effects, and because the environment largely cannot be controlled in field studies, direct measurements of total gene expression can be challenging to interpret. Additionally, the proximity of a field site to a laboratory in which genetic analyses can be conducted, as well as the availability of adequate facilities for temperature control, can influence the storage and transport of genetic samples. These factors may in turn influence the quality of the resulting expression profile. One approach for resolving these issues is to measure allele-specific gene expression (also known as allelic imbalance) in place of total gene expression. Allele-specific gene expression (ASGE) measures the relative contribution to total expression of one allele of a gene versus the alternative allele of the same gene, within individuals (Cowles *et al.* 2002; Yan *et al.* 2002; Bray *et al.* 2003; Lo *et al.* 2003; Pastinen and Hudson 2004; Wittkopp *et al.* 2004; Cheung *et al.* 2005; de Meaux *et al.* 2005; de Meaux *et al.* 2006; Pant *et al.* 2006; Milani *et al.* 2007; Campbell *et al.* 2008;

Gruber and Long 2008; Serre *et al.* 2008; Wittkopp *et al.* 2008; Zhang and Borevitz 2009). Thus, unlike most types of gene expression measurements, ASGE assays unambiguously indicate a causal basis for gene expression variation that lies in *cis* to the gene. This is because a *cis*-acting effect influences only the copy of the gene on the same physical chromosome, whereas effects that influence both alleles of the gene, such as environmental or genetic background effects, are *trans*-acting. In comparisons between alleles of a gene within individuals, the *trans* genetic and *trans* environmental backgrounds are held constant, while the *cis*-regulatory genetic context for an allele is free to vary. ASGE assays therefore provide methodological controls for otherwise uncontrolled environmental, *trans*-genetic, and sampling related variance. Additionally, ASGE can be treated as a phenotype itself and linked to *cis*-regulatory genetic variation, or measured across known environments in order to identify gene-environment interactions (de Meaux *et al.* 2005; Tao *et al.* 2006; Zhu *et al.* 2006; Milani *et al.* 2007; Babbitt *et al.* 2009; Tung *et al.* 2009; von Korff *et al.* 2009).

ASGE measurements have been used as a tool to disentangle the global contributions of *cis*- and *trans*-acting factors to gene regulation (Yan *et al.* 2002; Morley *et al.* 2004; Pastinen and Hudson 2004; Zhang and Borevitz 2009), to investigate the relationship between *cis*-regulatory variation and genetic divergence within and between species (Wittkopp *et al.* 2004; Gruber and Long 2008; Wittkopp *et al.* 2008), and to test specific hypotheses about functional genetic variation that influences a specific locus of interest (de Meaux *et al.* 2005; de Meaux *et al.* 2006; Tao *et al.* 2006; Zhu *et al.* 2006; Babbitt *et al.* 2009; Tung *et al.* 2009). With the exception of studies in humans, however, most work on ASGE has been conducted on model systems, or at least systems in which controlled crosses can be made. Even in the human literature, the largest studies have been conducted using RNA derived from cell lines rather than *in vivo*, from normal human tissue (Cheung *et al.* 2005; Serre *et al.* 2008). As a result, although ASGE assays

have some natural advantages for work on natural systems, their utility for this kind of research has not been explored. In chapter 3, I describe an ASGE study in the Amboseli baboons that explores the applicability and robustness of this approach for measuring allelic imbalance in samples obtained directly from members of a wild population. I show that, using ASGE as a starting point, it is possible to both identify the presence of allelic imbalance and, in some cases, identify the probable underlying *cis*-regulatory variants. In addition, I explore the possibility of using these measurements in conjunction with observational field data to dissect how environmental variation interacts with *cis*-regulatory effects.

The results of these analyses are promising. However, a limitation of this approach lies in the fact that many genes will either never exhibit significant allelic imbalance within a population, suggesting that no segregating *cis*-regulatory variation influences expression of those loci, or will exhibit allelic imbalance only rarely, making identification of putative causal variants difficult. As exemplified by the work described in chapter 3, most investigators working within natural populations are likely to be most interested in common allelic imbalance, which reflects *cis*-regulatory variation that is common enough to produce allelic imbalance in multiple individuals. Because both resources and sample quantities are often limited, it would therefore be desirable to focus one's measurements on loci that are likely to exhibit this pattern. In Chapter 4, I investigate whether it is possible to predict which loci are likely to exhibit common ASGE versus those that never exhibit ASGE. Using two published data sets in humans (Cheung *et al.* 2005; Serre *et al.* 2008), and given sequence information, polymorphism data and divergence data in and around a set of several hundred genes, I used a machine learning approach (Cortes and Vapnik 1995; Joachims 2005; Joachims 2006) to attempt to classify these genes into one of these two categories. The results suggest that a coarse level of prediction appears to be feasible, although a substantial amount of

noise remains unaccounted for by the model. Interestingly, the variables that contribute to the model's predictive accuracy in turn correlate with gene density around the focal gene, evenness of gene expression across different human tissues, and an estimate of negative selection on the *cis*-regulatory region (Haygood *et al*. 2007). These results suggest possible biological biases that influence which genes are likely to functionally vary within populations.

On the whole, the findings reported in Chapters 3 and 4 suggest that measurements of allelic imbalance may be of some utility in studies of natural populations. One application is towards understanding the general architecture of gene expression in these populations, using expression as a model for other types of complex traits. For example, variation in allele-specific expression across environments is indicative of gene-environment interaction (de Meaux *et al*. 2005; Zhu *et al*. 2006; von Korff *et al*. 2009). Because the expression of many genes can be measured using the techniques described here, allele-specific measurements could be used to better understand, for example, the relative contributions of different kinds of environmental variation to these interactions. This possibility is particularly intriguing given the ability to measure allele-specific expression variation using new high-throughput sequencing techniques, methods for which are currently in development (Degner *et al*. 2009; Heap *et al*. 2010). For targeted studies of specific phenotypes, however, allele-specific expression measurements are likely to be of greatest use in combination with other types of data that implicate the involvement of a specific gene or set of genes. This means that, while the results reported here represent a practical first step towards thinking about functional genetics and genotype-phenotype relationships in natural populations, other, complementary methods will also need to be applied. Some possibilities, such as expression quantitative trait locus mapping, admixture mapping, or linkage analyses, have already been well developed for other systems. However, adapting these methods

for use in field populations will not necessarily be straightforward, and will require careful development in each individual case.

1.5 Chapter 5: evolution of a malaria resistance gene in wild baboons

The results of chapters 2 and 3 make available both basic information about population structure within the Amboseli baboons, and a methodology for accessing the functional effects of regulatory variation within these animals. These tools represent a useful starting point for further investigations of population history and gene expression variation in this population. An outstanding question, however, is whether they also can make a contribution to understanding traits expressed on the organism level, which include most phenotypes of known adaptive significance from field studies.

In the Amboseli baboons, as for most other field studies of natural populations, genome-scale work on the genotype-phenotype relationship is not yet possible. However, the close relationship between baboons and humans makes it possible to leverage known information in humans to investigate a trait of potential evolutionary relevance in the baboons. Because blood samples can be gathered from these animals, but generally not other tissues, I became particularly interested in studying genes expressed in blood, many of which are relevant to immunity and disease. Chapter 2 reflects some of this work. In chapter 4, I present additional pieces of data for one specific gene, the Duffy antigen receptor for chemokines (*FY*; also known as *DARC*).

In humans, genetic variation in the *cis*-regulatory region of the *FY* gene has been well studied with respect to its effects on malaria infection. *FY* encodes a cell surface chemokine receptor that is expressed on the erythrocyte surface, as well as in several other tissues in the body. In erythrocytes, this receptor is the entry point for *Plasmodium vivax*, one of the most deadly forms of the four *Plasmodium* species that infect humans.

Humans homozygous for the derived allele (a C) at a single functional site in the *FY cis*-regulatory region do not express *FY* in their red blood cells, thus conferring a high degree of protection from *P. vivax* infection (Miller *et al.* 1976; Tournamille *et al.* 1995). This protective effect has been implicated in the near-fixation of this variant in some malaria-endemic parts of the world, although patterns of genetic variation around *FY* on a worldwide scale are complex (Hamblin and Di Rienzo 2000; Hamblin *et al.* 2002), possibly due to the pleiotropic effects of variation in this gene (He *et al.* 2008; Reich *et al.* 2009). Heterozygotes for the functional *cis*-regulatory variant are afforded a lower level of protection (Zimmerman *et al.* 1999), suggesting that expression level of the gene may correlate with rates of red blood cell infection in a quantitative manner. Indeed, levels of antibodies to *P. vivax* in the blood are also correlated with genotype at this site (Herrera *et al.* 2005).

The unusual evolutionary history of this locus in humans led me to investigate the pattern of genetic variation in its baboon homologue, and to explore the possibility that it might also explain phenotypic variation in parasite infection in the Amboseli baboons. This effort also gave me the opportunity to apply some of the tools outlined elsewhere in this work. While *P. vivax* does not infect baboons in the wild, a closely related parasite, *Hepatocystis kochi*, is common in Amboseli (Myers and Kuntz 1965; Tung *et al.* 2009). *Hepatocystis* is nested within the primate clade of the *Plasmodium* genus (Perkins and Schall 2002), and like *Plasmodium* is a vector-borne pathogen that infects the red blood cells of its mammalian hosts. In chapter 5, I used a PCR-based assay for *Hepatocystis* to phenotype infection status for 190 known individuals from Amboseli. Using a combination of allele-specific gene expression data, which test for functional *cis*-regulatory effects *in vivo*, and a set of experiments in cell culture, I show that, as in humans, *cis*-regulatory genetic variation influences *FY* gene expression in baboons. Susceptibility to *Hepatocystis* infection also correlates with variation in the

same region, which may be evolving non-neutrally in East African baboon populations. Taken together, these results suggest striking parallels, but not identical patterns, in the relationship between *FY* gene regulation and blood parasite infection in humans and baboons. Given that baboons and humans share a similar ancestral ecology (Jolly 2001), they highlight how work on natural nonhuman primate systems may help shed light on the environmental selection pressures that shaped the evolution of our own species.

1.6 Conclusions

One of the most exciting aspects of genetic research on primates lies in the fact that both the quality and quantity of field-based observational data and the scope of genetic and genomic resources are maturing to the point where the two approaches can be profitably combined. My hope is that this kind of integrative research will foster new interdisciplinary collaborations between geneticists and field biologists, and that the results of these collaborations will provide a new perspective for evolutionary genetic research. Such a research agenda will take time to develop. However, the results reported here are encouraging. In particular, they emphasize two more general results.

First, developing genetic approaches for unconventional nonmodel systems is becoming increasingly feasible. Indeed, it is likely to become even more so when the use of new methods and technologies become practical for these kinds of systems: high-throughput sequencing technologies, for example, not only allow whole-genome gene expression to be interrogated, but also uncover genome-wide patterns of genetic variation and epigenetic variation (Gilad *et al.* 2009). This is promising, because as exemplified by the results of Chapter 5, several complementary layers of evidence together make the strongest cases. In the short term, the field studies that will be most able to take advantage of these approaches will be those that are not limited to noninvasive sampling and those that focus on systems that are genomically well

characterized or have a well characterized near relative. For example, work on natural variation in the deer mouse genus *Peromyscus* has benefited from its close similarity to the classical laboratory mouse model, *Mus* (e.g., Turner and Hoekstra 2008; Linnen *et al.* 2009). Encouragingly, many nonhuman primates also exhibit high levels of genetic similarity to species that are already sequenced (although New World monkeys and prosimians are less well represented). Along with further methodological development for these systems, another challenge will be therefore lie in extending genetic methods for studying populations that lay outside these confines. In particular, making the new generation of genomic tools available for systems in which only noninvasive samples can be collected will be crucial.

Second, even at the level of laying initial groundwork and conducting early studies, the availability of extensive behavioral, ecological, and demographic data for a study system yields rapid payoffs. For instance, these kinds of data allowed the admixture analysis described in Chapter 2 to be placed within a broader context, making it clear that, although admixture does not make a major contribution to population structure measured within a static period of time, it does influence fitness-related traits in significant and detectable manner. By combining genetic estimates of admixture and gene flow with observational field data, the evolutionary contributions of hybridization to this population and to the larger set of East African baboon populations will become much better understood. With regard to functional work, field data can help define the phenotypes that may be of interest, as in the case of *Hepatocystis* infection, and can suggest ecological and environmental factors that may also play a role. The environmental effects incorporated in the allele-specific gene expression analysis reported in Chapter 3, for example, were suggested by prior analyses in the same population that demonstrated their phenotypic importance. While on one hand, these factors introduce additional complexity, this complexity reflects the

situation experienced by individual animals in the field. Appreciating the interplay between social behavior, ecological and environmental variation, and genetic variation will therefore allow us to better understand the raw material upon which selection acts.

2. Genetic evidence reveals temporal change in hybridization patterns in a wild baboon population[1]

2.1 Background

Naturally occurring interspecific hybrids have been of long-standing interest in evolutionary biology because of their importance in helping to understand the processes of introgression, speciation, and reproductive isolation (Mayr 1942; Anderson and Stebbins 1954; Arnold 1992; Arnold and Hodges 1995; Barton 2001). Depending on the adaptive consequences of hybridization, hybrids can reveal strong selective boundaries between species when hybrids are selected against, or can illustrate how increased heterozygosity and genetic diversity may lead to a hybrid fitness advantage. Additionally, hybridization is itself an important mechanism of evolutionary change. Through the introduction of new genetic variation and new allelic combinations, hybridization may influence the evolutionary trajectory of the hybrid population, the parental populations, or both (Anderson and Stebbins 1954; Lewontin and Birch 1966; Arnold 1992; Rieseberg 1997; Rieseberg *et al.* 2003).

The evolutionary consequences of hybridization are related to the frequency of interspecific mating, the genetic distance between parental species, and the fitness effects of hybridity. Studies on hybridization, particularly within hybrid zones, have largely focused on this last component, especially on the classification of hybrids as either more or less fit than one or both of their parental species. Hybrids with relatively high fitness suggest hybrid advantage or hybrid superiority; this is often associated with hybridization that occurs in specialized ecological circumstances (e.g., temporal or clinal ecological transitions). In contrast, hybrids with relatively low fitness suggest that

---

[1] The contents of this chapter have been previously published in: J Tung, MJE Charpentier, DA Garfield, J Altmann, and SC Alberts (2008). *Genetic evidence reveals temporal change in hybridization patterns in a wild baboon population.* Molecular Ecology 17: 1998 – 2011.

selection against hybrids in a "tension zone" helps to maintain species boundaries and counteracts the effect of regular gene flow (Barton and Hewitt 1985; Grant and Grant 1992; Barton 2001); reviewed in Arnold and Hodges 1995). Alternatively, if hybrid fitness is equivalent to that of parental species and independent of ecological context, hybrids may represent a snapshot of species fusion in process (Rhymer and Simberloff 1996; Salzburger *et al.* 2002). This framework provides three mutually exclusive predictions about the consequences of hybridization, differentiated by the direction of relative fitness differences between hybrids and the parental species (Moore 1977; Arnold and Hodges 1995). However, while these predictions suggest that the conditions surrounding hybridization and the fitness consequences of hybridization are static, the rate and consequences of hybridization within a population may in fact fluctuate over time.

Here, we describe a dynamically changing pattern of hybridity in a wild population of savanna baboons from the Amboseli basin of southern Kenya, a known baboon hybrid zone (Maples and McKern 1967; Samuels and Altmann 1986; Alberts and Altmann 2001). The focal population has been under continuous observation on a near-daily basis since 1971, resulting in a data set representing up to six generations of individually known animals. DNA samples are available for a large number of these individuals (Altmann *et al.* 1996; Alberts *et al.* 2006; Loisel *et al.* 2006). The Amboseli baboon population is comprised primarily of yellow baboons (*Papio cynocephalus*). It represents one of the type examples of the widespread "ibean" morphotype of yellow baboons (Jolly 1993), which shares more morphological similarities with anubis baboons than do the two other yellow baboon morphotypes (the "typical" and "kinda" morphotypes), possibly because of anubis admixture that has occurred in the ibean lineage over the course of evolutionary history (Jolly 1993). In addition, hybrids are found in the population due to the occasional immigration of anubis (olive) baboons (*P.*

anubis) from outside the basin (Alberts and Altmann 2001). Specifically, six anubis males have immigrated into study groups in the basin over the course of the study, and one small (ca.18) mixed-sex group of anubis baboons also entered the basin in the early 1980's (Samuels and Altmann 1986). Hybrids now occur in both study groups and in non-study groups in the basin, and they have resulted not only from these anubis immigrations, but also from the movement and successful reproduction of hybrid males between and within study and non-study groups. The status of Amboseli as a hybrid zone is consistent with the geographical distribution of baboon species: this population is situated on the boundary between the ranges of yellow and anubis baboons, with yellow baboons roughly to the south and east and anubis baboons to the north and west (Jolly 1993; Newman *et al.* 2004). These two species represent two of the five commonly recognized baboon species (or subspecies: see discussion in Jolly 1993) within the genus *Papio* (also including *P. hamadryas*, *P. papio*, and *P. ursinus*), all of which exhibit moderate geographical separation, are readily distinguished morphologically, and represent a range of distinct patterns of social structure and behavior (Jolly 1993; Jolly 2001; Henzi and Barrett 2003; Newman *et al.* 2004). Nevertheless, all baboon species can interbreed with their neighboring congeners to produce viable, fertile hybrid offspring, and several naturally occurring hybrid zones have been described near the geographical boundaries between species (Maples and McKern 1967; Nagel 1973; Phillips-Conroy and Jolly 1986; Alberts and Altmann 2001; Jolly and Phillips-Conroy 2007). Hybrid anubis-yellow baboons have also been documented in captivity (Ackermann *et al.* 2006).

Intriguingly, in both the well-described anubis-hamadryas hybrid zone in Ethiopia and in the Amboseli anubis-yellow hybrid zone, morphological estimates of hybridity indicate that patterns of hybridization and introgression have changed over time (Phillips-Conroy and Jolly 1986; Alberts and Altmann 2001). In Ethiopia, the

original pattern described by Nagel (1973) based on work in the late 1960's was characterized by spatially distinct anubis, hybrid, and hamadryas groups, with hybrids confined to a narrow intermediate zone between the two parent species. Between the late 1960's and 1973, the anubis-hamadryas hybrid zone expanded and gave way to a graded clinal pattern, suggesting that hybrids enjoyed success in backcrossing into both parent populations (Phillips-Conroy and Jolly 1986). Hybrids have also been reproductively successful within Amboseli. Based on morphological estimates of hybridity, the frequency of hybrid births in Amboseli increased from the 1960's and 1970's, when no anubis and few possible hybrid baboons were observed, to the 1990's, when hybrids made up an estimated 10% of births (Alberts and Altmann 2001). These changes may reflect either increasing anubis baboon gene flow into the predominantly yellow baboon-occupied basin, the selective outcome of fitness differences between hybrid baboons and yellow baboons, or a combination of both. Analyses based on morphological scoring of hybrids indicated that hybrid males tend to undergo natal dispersal earlier in life than do yellow males (Alberts and Altmann 2001). Dispersal represents a major life history marker for male baboons, and variation in the timing of this event is correlated with the timing of other important social and reproductive milestones, including age at physical maturation and age at first mate guarding episode, a proxy for first reproduction (Alberts and Altmann 1995; Charpentier *et al.* 2008). Therefore, if the benefits of earlier dispersal are not offset by costs later in life, earlier dispersal may result in a selective advantage. We hypothesized that a selective advantage would therefore accrue to hybrids, mediated by early maturation and dispersal in males, and that this advantage would be reflected in changes in the frequency of hybrid individuals in the population. This possibility motivates a more in-depth, genetically based analysis of hybridity within the Amboseli population.

Towards that end, here we extend our previous analysis of hybridization patterns in Amboseli by assessing multilocus microsatellite genotypes for evidence of admixture in this population. This genetic analysis of hybridity is an important extension of our previous morphological analysis of hybridity. First, genetic marker-based analyses do not depend on observer-defined phenotypes (e.g., pelage color or body size) identified *a priori* to differentiate the parental species. Second, relying on specific phenotypes can be misleading because phenotypic differences may reflect variation at only one or a few loci, whereas hybridization is a genome-wide phenomenon. In cases involving dominant and recessive variants, the degree of hybridization inferred from the trait is particularly vulnerable to overestimation or underestimation because heterozygotes may not express the mean parental phenotype. Third, credible intervals can readily be assigned to genetic marker-based hybridity estimates, permitting interpretation of these estimates in the light of quantitative uncertainty. Finally, genetic assignments of hybridity lend themselves directly to analyses of admixture-mediated changes in population genetic structure, which can help address questions about the possible fitness consequences of hybridity and introgression. We also compare our results to previously collated morphological hybridity estimates. Such comparisons help identify any systematic biases that differentiate the morphological and genetic hybrid scoring methods, and particularly increase confidence in those assignments for which morphological and genetic estimates are congruent.

We assigned genetic hybrid scores with data from 14 unlinked microsatellites typed in 450 Amboseli baboons born from 1968 – 2004, using the Bayesian clustering algorithm implemented in the program Structure 2.0 (Pritchard *et al.* 2000; Falush *et al.* 2003). These hybrid scores estimate the proportion of each individual's genome derived from *P. anubis* ancestry. Similar approaches have been previously applied towards the

identification of introgression in European wildcats (Beaumont *et al.* 2001; Pierpaoli *et al.* 2003; Lecis *et al.* 2006), characterization of hybrid zone dynamics in Baltic fish (Beaumont *et al.* 2001; Nielsen *et al.* 2003; Pierpaoli *et al.* 2003; Nielsen *et al.* 2004; Lecis *et al.* 2006), and confirmation of the genetic integrity of endangered species, such as the black-faced impala (Lorenzen and Siegismund 2004). We assessed the robustness of our results by checking for consistency of the hybrid score assignment in families using pedigree data, and through simulations that tested the sensitivity of our results to different conditions.

Our analyses suggest that, even with a modest number of genetic markers, we have good power to identify the signature of hybridity within individual baboons. Using these data, we describe how hybridization patterns within the Amboseli population – both changes in the abundance of hybrids and in the distribution of hybrid scores – have changed over time. We evaluate these results in the light of known and inferred patterns of anubis immigration into this population, and speculate on the resulting implications for the evolutionary dynamics of this hybrid zone.

2.2 Materials and methods

2.2.1 Samples and genotyping

We assigned genetic admixture scores to 450 Amboseli baboons born between 1968 and 2004. All subjects were born in or immigrated into groups subject to long-term monitoring by the Amboseli Baboon Research Project, with continuous observation starting in 1971 and continuing to the present (Altmann and Alberts 2003; Alberts *et al.* 2006). Those born into study groups had birth dates known to within a few days. Birth dates for immigrants were estimated using morphological and behavioral evidence and a set of criteria calibrated to baboons of known age, such as pelage condition and canine wear (Alberts and Altmann 1995; Alberts *et al.* 2003). One of the study groups began

feeding at a refuse pit associated with a tourist lodge in the 1980's (Altmann and Muruthi 1988; Muruthi *et al.* 1991; Altmann and Alberts 2005). Because this alternative foraging pattern influenced immigration and emigration in this group, we excluded from our analysis all individuals born into this group after 1979.

In order to capture change in hybridity within the population over time, we partitioned the total dataset into four non-overlapping data partitions, or "cohorts," corresponding to individuals born in the late 1960's or 1970's (the "1960's/1970's" data partition: n = 31), the 1980's ("1980's:" n = 117), the 1990's ("1990's:" n = 187), and the 2000's ("2000's:" n = 115). The ten-year span of these partitions is somewhat arbitrary, but is a convenient method of dividing the dataset and allows for one to two generations (~6 years in this population) to pass between reevaluations of the data.

As part of previous analyses of paternity and relatedness, all 450 Amboseli baboons included in this analysis were genotyped at 14 polymorphic microsatellite loci. Genomic DNA was available for all individuals based on either extractions from blood samples obtained during infrequent dartings or from noninvasively collected faecal samples (Buchan *et al.* 2005; Alberts *et al.* 2006). The methods used for genotype assignments and data on the performance of the 14 microsatellite primer pairs have been reported elsewhere (Buchan *et al.* 2005; Alberts *et al.* 2006). Importantly for these analyses, no two loci were located on the same chromosome, ensuring that there was no physical linkage between any of the 14 markers (Rogers *et al.* 2000). Infrequent PCR failure and inconsistent genotyping results, which may occur during noninvasive genotyping, led to missing data for 2.48% of the total Amboseli genotyping dataset (i.e., for a small subset of individuals at a few loci).

In order to help generate estimates of yellow-anubis hybridity, we also produced genotypes from the same set of microsatellite markers for a total of 13 *P. anubis* individuals. Three of these were anubis males that immigrated into Amboseli study

groups from anubis source populations; they were designated as anubis based on their morphology and coat color as assessed by experienced observers. The other ten *P. anubis* samples were from Masai Mara National Reserve, Kenya, about 250 km to the northwest of Amboseli and far from the range of yellow baboons, as well as from any hybrid zone (Jolly 1993; Kingdon 1997). All Masai Mara samples were collected in August 2004 and were obtained as extracted DNA from the Integrated Primate Biomaterial and Information Resource, IPBIR (courtesy of R. Sapolsky; sample numbers are provided in Table 1). Due to PCR failure or inconsistent genotyping, 2.14% of the total set of individual-by-locus genotypes for Masai Mara individuals were missing in this analysis.

Table 1: Masai Mara sample information. Ten *Papio anubis* samples were obtained as extracted DNA from the Integrated Primate Biomaterial and Information Resource (IPBIR), courtesy of R. Sapolsky. All samples originated from the Masai Mara National Reserve, Kenya, and were originally sampled in August 2004.

IPBIR Repository #	Date of original sampling	Local Identification
BP00232	12 August 2004	York
BP00234	13 August 2004	Manda
BP00236	14 August 2004	Oscar
BP00237	16 August 2004	Stefano
BP00242	19 August 2004	Rocket
BP00243	21 August 2004	Leakey
BP00244	22 August 2004	Puck
BP00245	22 August 2004	Julius
BP00246	23 August 2004	Facko
BP00247	25 August 2004	Duke

Summary statistics on heterozygosity at the 14 microsatellite markers for the Amboseli population (including the three anubis males that immigrated into Amboseli) and for the Masai Mara population are provided in Table 2. Twelve of 14 loci conformed to expected levels of heterozygosity in Amboseli, but two loci (Table 2, shown in bold) showed significantly elevated levels of heterozygosity in Amboseli. These two loci showed elevated levels of heterozygosity in all four temporal subsets, with one

exception for each locus (data not shown). In contrast, all loci in Masai Mara conformed with expected levels of heterozygosity. This result – higher than expected levels of heterozygosity at a modest number of loci in Amboseli but not the Masai Mara – is consistent with the expectation of some hybridization in Amboseli, but a pure anubis population in Masai Mara.

Table 2: Summary statistics for microsatellite genotyping data. Bonferroni corrected *p*-values (within populations) are provided corresponding to the probability of obtaining the observed levels of heterozygosity under the assumption of Hardy-Weinberg equilibrium.

Population	Locus	No. alleles	Obs. Het.	Exp. Het.	*p*
Amboseli	AGAT006	10	0.866	0.830	0.988
	D1s1656	10	0.852	0.804	0.176
	D2s1326	**9**	**0.828**	**0.816**	**0.00322**
	D3s1768	10	0.824	0.814	1.00
	D4s243	7	0.805	0.814	0.855
	D5s1457	8	0.811	0.792	1.00
	D6s501	15	0.806	0.800	1.00
	D7s503	11	0.841	0.808	0.291
	D8s1106	8	0.759	0.778	1.00
	D10s611	12	0.817	0.818	1.00
	D11s2002	**8**	**0.865**	**0.831**	**0.00140**
	D13s159B	8	0.839	0.803	1.00
	D14s306	8	0.785	0.771	1.00
	D18s851	8	0.784	0.738	1.00
Masai Mara	AGAT006	4	0.600	0.595	1.00
	D1s1656	5	0.778	0.739	1.00
	D2s1326	5	0.667	0.797	1.00
	D3s1768	6	0.700	0.826	1.00
	D4s243	7	0.800	0.863	1.00
	D5s1457	6	0.900	0.837	1.00
	D6s501	7	0.800	0.863	1.00
	D7s503	6	0.800	0.763	1.00
	D8s1106	7	1.000	0.811	1.00
	D10s611	6	0.500	0.742	1.00
	D11s2002	6	0.700	0.805	0.574
	D13s159B	3	0.875	0.660	1.00
	D14s306	7	0.900	0.726	1.00
	D18s851	5	0.900	0.795	1.00

2.2.2 Assignment of genetic hybrid scores

For the 450 Amboseli baboons included in this study, we generated an estimate of the proportion of each individual's genome attributable to anubis (as opposed to yellow) baboon heritage (i.e., a "genetic hybrid score") using the admixture analysis implemented in the program Structure 2.0 (Pritchard *et al.* 2000; Falush *et al.* 2003). Structure uses a Bayesian model-based clustering algorithm to estimate the allele frequency distributions for each marker locus for each source population, K, that contributes to the admixed population. The program probabilistically assigns each genotyped allele for each individual to one of these populations. The result is an estimate of the amount of genetic material contributed from each source population to each individual. Importantly, this method allows individual-specific admixture estimates to be produced even when most alleles are shared between source populations, and does not require prior specification of allele frequencies in these populations. Rather, it draws on genotype data for all individuals in the dataset (n = 463 total individuals, including 13 individuals of known genetic background) in order to assign estimates of admixture. The degree to which the assignments maximize linkage equilibrium and Hardy-Weinberg equilibrium within populations determines the likelihood of a particular set of assignments.

We ran Structure under the *F* model, which allows allele frequency spectra between the source populations to be correlated and allows admixture within individuals (Falush *et al.* 2003). All individuals were analyzed in a pooled analysis because temporal variation across the sample sets accounted for a very small component of overall genetic variation in the sample (data not shown). We flagged the 13 anubis baboons (10 Masai Mara baboons and 3 anubis immigrants into Amboseli) as members of a single identified population and flagged the 450 Amboseli baboons as

unassigned to a population. We set the total number of populations, K, equal to 2. Thus, the 450 Amboseli individuals could have been assigned to the anubis cluster (made up of the Masai Mara baboons and the 3 anubis immigrants), to an alternative cluster distinct from the anubis cluster, which we interpret as characteristic of a yellow baboon genetic make-up, or as hybrids between the two clusters. Ideally we would have drawn more of the anubis sample from the population of origin for anubis immigrants into Amboseli, but sampling constraints prevented us from pursuing this strategy both because the identity of this population is uncertain, and because sampling from unhabituated baboons is logistically difficult. Hence, we also examined the effects of our small anubis sample size using three different sets of simulations, described below.

Each analysis was run with a burnin length of 100,000 MCMC iterations and a run length of 1,000,000 iterations. We altered several of the default parameters in Structure in order to reflect known biological aspects of the Amboseli baboon system. First, we set the migration prior ν to 0.015 and the GENSBACK parameter to 2, corresponding to the probability that a given individual was himself an immigrant between populations or that he or she had an immigrant ancestor in the last 2 generations. This prior was based on estimates from field observations of the number of anubis immigrants into Amboseli study groups during the last 35 years, relative to the total number of immigrant males in the same time period. We also allowed α, the parameter for admixture, to vary between populations. This allows the total contribution of each of the two source populations (anubis and yellow) to the overall dataset to be asymmetrical. All other parameters were set to the defaults recommended in Pritchard *et al.* 2000 or Falush *et al.* 2003), and/or the documentation for the Structure program.

To obtain a final genetic hybrid score for each individual, we reran the entire analysis three times and averaged the proportion of each Amboseli individual's genome assigned to the anubis population over these three runs.

2.2.3 Comparison with morphological hybrid scores

Morphological hybrid scores were assigned prior to genetic analysis, based on observation and scoring of seven phenotypic characteristics that distinguish anubis and yellow baboons: coat color, body shape, hair length, head shape, tail length and thickness, tail bend, and muzzle skin appearance (Alberts and Altmann 2001). Three to four experienced observers independently assigned separate morphological hybrid scores, which were then averaged into one composite hybrid score (Alberts and Altmann 2001). We re-scaled these morphological hybrid scores to correspond to the scale of the genetic hybrid scores, with 0 representing pure yellow and 1 representing pure anubis. Interobserver agreement and agreement between morphological scores assigned at different life stages were both high (data not shown). In all comparisons of morphological hybrid scores and genetic hybrid scores, we used the average of the composite scores assigned during adulthood as the morphological point estimate of hybridity. For those individuals that had yet to reach adulthood by the end of data collection or that died before reaching adulthood, we used the average of composite scores assigned as juveniles instead (n = 84). In all, morphological hybrid scores were available for 315 of the 450 Amboseli baboons used in the genetic analysis.

In order to assess agreement between the morphological scores and the genetic hybrid scores assigned in this study, we calculated the Pearson correlation between the two scores for the same individuals (n = 315). However, because the Amboseli population is predominantly yellow, a large proportion of these scores fall at or near 0. We therefore calculated a p value for this correlation using a nonparametric approach.

We randomly permuted the genetic hybrid scores 10,000 times against constant morphological hybrid scores, using the R statistical package (Team 2007). We ranked the observed correlation coefficient, *r*, from the actual data among the 10,000 *r* values calculated from these permutations. The significance of the observed correlation was defined as the proportion of larger *r* values observed in the permutation tests.

2.2.4 Assessment of the consistency of genetic hybrid scores using pedigree data

In order to test whether our method of assigning genetic hybridity was consistent across individuals, we correlated the genetic hybrid score for individuals with the midpoint value for their parents, when all three baboons were included in the study (n = 272 offspring-parent triads, including the offspring-parent triads that included anubis immigrants). Because Structure infers the population of origin for each allele copy for each individual (not the probability of population of origin for each allele across individuals) and has no prior information on pedigree relatedness, a strong correlation between the parental mean and the offspring hybrid score is not a necessary outcome of the program, and should occur only when it is performing consistently for the whole dataset. For comparison, we conducted the same analysis using morphological hybrid scores (n = 151 offspring-parent triads). Because of the large number of 0 or near zero values in both datasets, significance values for these analyses were assigned by repeated permutations of parental midpoint values on constant individual hybrid scores; this approach was identical to the method we used to assign significance in the morphological hybrid score-genetic hybrid score comparison. A high correlation between parents and offspring for the genetic hybrid scores would indicate that assignments were made in a consistent manner; it would not independently validate these scores, because paternity assignments were made using the same microsatellite loci used for the hybridity analysis. However, this method does act as an independent measure of the

validity of the morphological scores, because the morphological hybrid scores were assigned using a completely different dataset than that used to generate pedigrees.

2.2.5 Assessment of the robustness and replicability of genetic hybrid scores using simulation

We used three sets of simulations to assess whether the genetic hybrid scores we assigned were robust to replication and/or different estimates of the allele frequency spectra for the known anubis baboons. Because we use a relatively small number of known anubis in the analysis, the estimated allele frequencies based solely on the known anubis will approximate the "true" allele frequency spectra in the anubis source population, but are almost certainly inexact. The purpose of these simulations is to test whether the genetic hybrid score assignments remain stable within a realistic range of uncertainty surrounding these allele frequency estimates.

2.2.5.1 Simulation 1: Replicability of hybrid scores given observed allele frequencies

We generated 100 simulated baboon datasets that were of the same size (450 individuals) as the empirical dataset and that exhibited a similar distribution of genetic hybrid scores as inferred from the observed genetic data. We asked how well the inferred hybrid score for a simulated individual matched with the known degree of hybridity for the same individual, given (1) the observed degree of genetic differentiation between yellow and anubis baboons, (2) the number of markers used in this study and the observed allelic diversity for each marker, and (3) the number of individuals in the dataset. This analysis assessed the replicability of our results, given the same model of allele frequency distributions for both source populations.

First, we created a "yellow" pool of alleles from the genotypes of 120 Amboseli individuals with the lowest genetic hybrid scores (range = 0.029 – 0.058 in the empirical dataset), and an "anubis" pool of alleles from the ten Masai Mara baboons and the 3

anubis immigrants into Amboseli, for all 14 markers. These pools of alleles were used to create all 450 simulated individuals in the 100 simulated datasets in Simulation 1. Second, to create each individual within a simulated dataset, we randomly drew a genetic hybrid score from the 450 actual (i.e., not simulated) genetic hybrid scores in the empirical dataset. Third, simulated genotypes were created for each of these simulated individuals by sampling twice from the anubis pool of alleles at a probability equal to the value of the genetic hybrid score previously assigned for that individual, or else from the yellow pool of alleles, for each of the 14 marker loci. For example, if a genetic hybrid score of 0.60 was randomly drawn from the observed dataset, then for each of the 2 alleles at each of the 14 marker loci, the simulated individual would have a 60% probability of being assigned an allele from the anubis pool of alleles, and a 40% chance of being assigned an allele from the yellow baboon pool of alleles. The proportion of the total genotypes drawn from the anubis baboon pool of alleles following this step represented the 'known' hybrid score for that simulated individual. All sampling was conducted with replacement, so that for each draw of a genetic hybrid score for each new individual in the simulated population, and for each draw of an allele from the anubis or yellow baboon pools, the original probabilities still obtained. We repeated this procedure 100 times to create 100 simulated datasets, each containing 450 individuals. The resulting datasets were run in Structure using the parameter set chosen for the original assignment of genetic hybrid scores to produce the 'inferred' genetic hybrid scores for the simulated individuals. In order to assess the accuracy of hybrid score assignment, we evaluated the difference between this inferred genetic hybrid score and the known genetic hybrid score for each simulated individual (n = 45,000).

2.2.5.2 Simulation 2: Sensitivity of hybrid scores to incorrect estimates of anubis allele frequency distributions

Assignment of individual genetic hybrid scores depends in part upon the inferred allele frequency distributions for the 14 marker loci in the two source populations; these are drawn from a large sample in the case of the Amboseli population, but from a small sample in the case of anubis baboons. This small sample size could potentially affect the accuracy of our inferences due to incorrect estimation of the anubis allele frequency distributions for the marker loci.

In Simulation 2, we investigated this possibility by randomly simulating 10 individuals from the genotype pool of the 13 anubis baboons, as in simulation 1. We used this subset of 10 individuals as the full set of anubis baboons in the analysis, in combination with the actual empirical genotype data for the 450 Amboseli baboons, which we designated of unknown ancestry for the simulation. Each run therefore drew on true genotype data for 450 unknown Amboseli individuals and simulated data for 10 anubis baboons instead of 13 known anubis baboons. We then produced genetic hybrid scores in Structure using the same parameter set chosen for the original assignment of genetic hybrid scores. This subsampling routine created modest run-to-run fluctuations in allele frequencies within the pool of anubis. We also repeated this set of simulations using a sample of only five anubis baboons, which created much larger fluctuations in the allele frequencies for the anubis. We repeated both sets of simulations 100 times each and then analyzed the difference between the hybrid scores assigned to individuals in the subsampled, simulated datasets and the hybrid scores assigned to the corresponding individuals in the actual dataset. The results of these simulations provide an estimate of the threshold at which small sample sizes of anubis will cause large errors in the inferred allele frequency distributions for the anubis source population, which would also affect hybrid score assignment.

2.2.5.3 Simulation 3: Sensitivity of hybrid scores to the detection of rare alleles

A small sample size of anubis baboons may also impact genetic hybrid score assignment due to a failure to sample rare alleles in the anubis population. In such cases, allele frequency distributions will not be greatly affected, but rare alleles that are actually shared between both anubis and yellow populations will then look like private alleles found only in yellow baboon populations. Individuals that carried these alleles would therefore be assigned genetic hybrid scores that are biased towards lower values.

In Simulation 3, we asked how the small sample size of anubis individuals may have impacted our results due to a failure to sample rare anubis alleles. First, we randomly selected one of the 14 marker loci. Then, we randomly removed one of the three rarest alleles for that locus from the dataset. Designation of rare alleles was based on observed allele frequencies among the pool of 13 anubis baboons. We readjusted the frequencies of the other alleles upwards to compensate for the missing allele by uniformly allocating the number of times the missing allele was originally observed across the remaining set of alleles. This process simulated the resulting genotype data if we had failed to sample one rare allele at one of the 14 marker loci. We then produced genetic hybrid scores in Structure using this altered dataset. After 100 iterations of this simulation, we asked how well the resulting hybrid scores correlated with the hybrid scores produced in the full analysis. We repeated the same procedure in two additional sets of simulations, in which we simulated a failure to sample one rare allele at each of five marker loci and one rare allele at all of the marker loci, respectively. If the differences between the results of these simulations and the results from the whole dataset are small, then the genetic hybrid scores we have assigned are robust to missing rare alleles within the anubis dataset. This test is in fact conservative, because the definition of "rare allele" we use here encompasses alleles that actually were sampled in

the anubis dataset, and are therefore unlikely to be extremely rare among true anubis populations.

2.2.6 Analysis of temporal changes in hybridization patterns

We used three metrics to assess potential changes in hybridization patterns over time. First, we asked about increase, decrease, or stability in the percentage of hybrid baboons born in the population from the late 1960's/1970's to the present. We defined hybrid individuals as all Amboseli baboons in the dataset for which the lower bound of the 90% credible interval for their genetic hybrid score was greater than or equal to 0.05. This cutoff is a conservative threshold that assures that we have counted as hybrids only the individuals for which a genetic hybrid score of 0, corresponding to a pure yellow genomic composition, could be ruled out with high confidence. We calculated the percentage of hybrids born into the population separately for the 1960's/1970's, 1980's, 1990's, and 2000's data partitions.

Second, we asked whether the degree of hybridization among hybrids showed any trend up or down over time. We defined degree of hybridization as the average of genetic hybrid scores in a data partition, considering only hybrids. Changes in the degree of hybridization reveal information about introgression, gene flow, and the success of anubis and/or hybrid baboons in reproducing within the Amboseli population. For example, if all hybrids in every data partition had hybrid scores around 0.50, with no change over time, we would infer that although anubis baboons could successfully mate within Amboseli, F1 hybrids generally suffered from poor reproductive success. In contrast, if the degree of hybridization among hybrids decreased over time but the number of hybrids (as revealed by the categorical analysis described above) did not, we would infer that backcrosses and hybrid-hybrid matings were common in the population due to hybrids reproducing in the population.

Third, we examined the frequency distribution of hybrid scores of individuals born in different decades to ask whether the frequency distribution of hybrid scores shifted down over time; this analysis included both hybrid and non-hybrid individuals (i.e., the full set of individuals in the study). We conducted two-sample one-tailed Kolmogorov-Smirnov tests comparing the distribution of hybrid scores among the cohorts represented by each pair of temporal data partitions. A significant result would indicate that a random draw from the more recent cohort would be significantly more likely to correspond to a lower hybrid score than would a random draw from the earlier cohort. This third metric is closely related to the above analysis of changes in hybridization among hybrids, but also tests whether the patterns of change among individuals with high genetic hybrid scores (those for whom anubis ancestry can be inferred with very high confidence) are reflected in the hybrid dynamics in the population as a whole. Because we did not identify any hybrids in the 1960's/1970's dataset, we excluded those individuals from this component of our analysis.

2.3 Results

2.3.1 Genetic hybrid score assignments in Structure

Our analysis generated a mean hybrid score (± 90% credible interval) for each of the 450 Amboseli baboons. Individual hybrid scores showed very close run-to-run agreement (mean standard deviation across runs for the same individual = 0.0025). Figure 2 shows the cumulative distribution of genetic hybrid scores for all Amboseli individuals. 90% credible intervals were largest for baboons with mean hybrid scores in the midrange values, as has also been the case for similar analyses of admixture in other systems (Beaumont *et al.* 2001; Pierpaoli *et al.* 2003). 99 of 450 individuals were deemed to have anubis ancestry based on their genetic hybrid scores, using the criterion of credible intervals with a lower bound > 0.05.

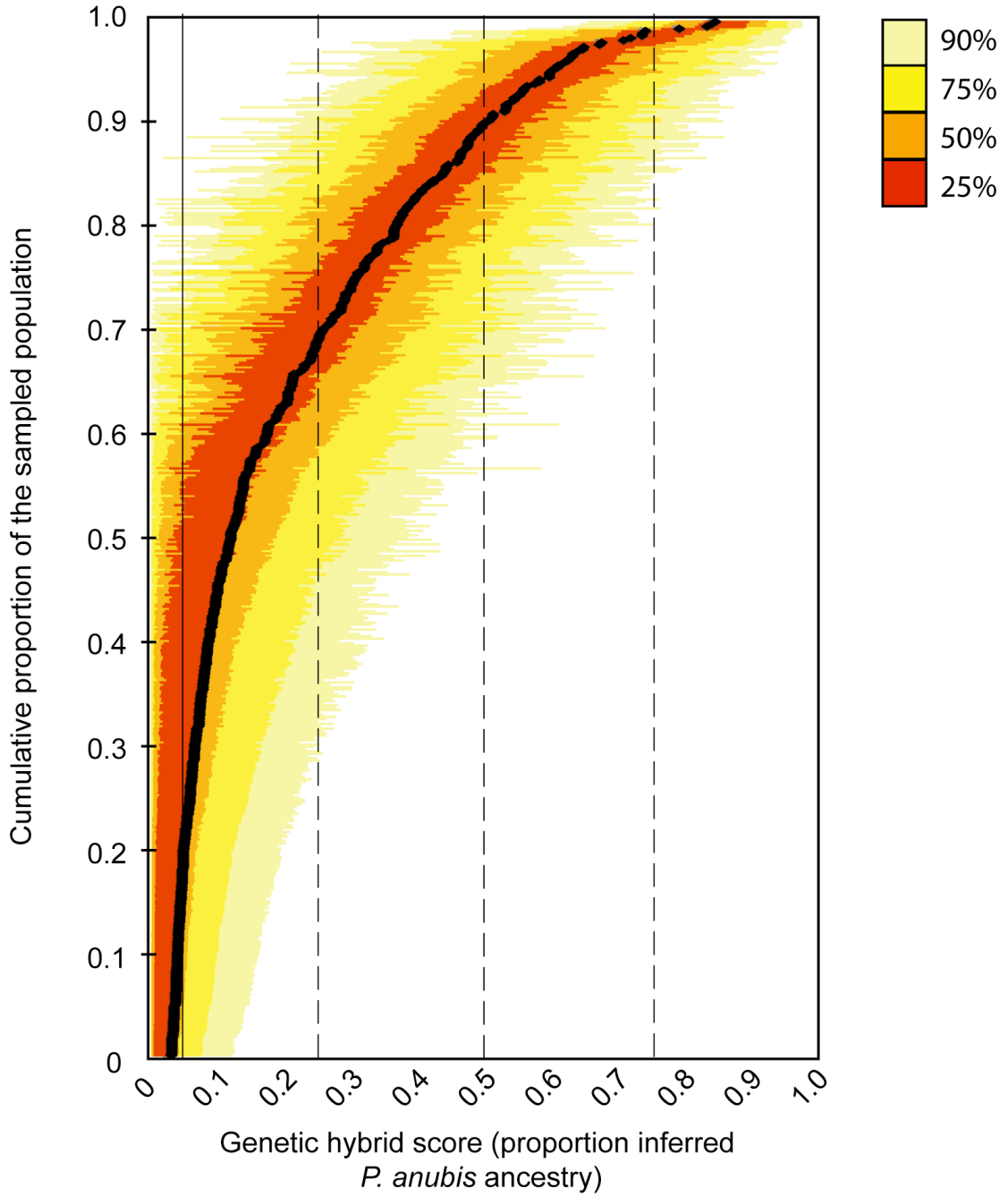

Figure 2: Genetic hybrid scores (i.e., percent anubis ancestry) for each of the 450 individuals in the analysis, averaged over three Structure runs and shown as the cumulative proportion of the sampled population. Each black point represents the mean hybrid score for one individual. Individuals are ordered along the y axis from lowest (least anubis ancestry) to highest (most anubis ancestry) genetic hybrid scores. Flanking lines show 25%, 50%, 75%, and 90% credible intervals. Individuals with lower 90% credible intervals > 0.05 (boundary indicated by solid vertical line) were considered hybrids for the purposes of this analysis.

2.3.2 Agreement between genetic hybrid scores and morphological hybrid scores

In general, we observed good agreement between morphological hybrid scores and the genetic hybrid scores assigned in this study (n = 315, $r = 0.484$, $p < 0.0001$). Permutation tests yielded $p < 0.0001$ in 10,000 permutations, demonstrating that the observed correlation was not a product of the structure of the dataset but actually reflected significant concordance between these two metrics. However, the cumulative distribution of genetic hybrid scores was right-shifted (towards more anubis ancestry) relative to the cumulative distribution of morphological scores (compare Figure 2 with Figure 3 in Alberts and Altmann 2001). Discrepancies in the scores originated primarily from individuals who were assigned higher genetic hybrid scores than morphological hybrid scores. This bias is clear when the two metrics are compared in subsets. Individuals with low genetic scores (< 0.25) almost invariably had morphological scores that were also lower than 0.25 and similar to the genetic scores (only 5 of 200 animals in this category violated this pattern). However, individuals with genetic hybrid scores above 0.25 generally had morphological scores that were lower (more yellow) than their genetic scores: specifically, 68 of the 77 individuals with genetic scores between 0.25 and 0.5 and for whom we had both scores had morphological hybrid scores lower than their genetic scores. Only 9 had morphological scores higher than their genetic scores. Similarly, 28 of the 36 individuals with genetic hybrid scores between 0.5 and 0.75 had morphological hybrid scores lower than their genetic scores, whereas only 8 had morphological hybrid scores higher than their genetic scores.

These comparisons suggest that differences between the two metrics were not random, but were caused almost entirely by cases in which genetic estimates indicated some anubis admixture, but morphological assessments did not. In other words, the individuals inferred as predominantly yellow by the genetic analysis were almost always

assessed as predominantly yellow in morphological analyses, and individuals assessed as hybrids in the morphological analyses were almost always assessed as hybrids in the genetic analyses, but individuals assessed as hybrids in the genetic analyses were frequently assigned morphological scores that suggested lower levels of anubis ancestry.

2.3.3 Consistency within the dataset

Comparisons of individual genetic hybrid scores with the midpoint values of the parents showed that the assignment of genetic hybrid scores was extremely consistent with predictions from previously constructed pedigrees, such that the distance between the scores of parents and the scores of offspring were in agreement with Mendelian inheritance at the 14 microsatellite markers (n = 272, $r = 0.905$; $p < 0.0001$). As an example, Figure 3 shows the genetic hybrid scores of offspring of several different types of crosses that we observed in the study population, including yellow x anubis crosses, both types of backcrosses, and hybrid-hybrid crosses.

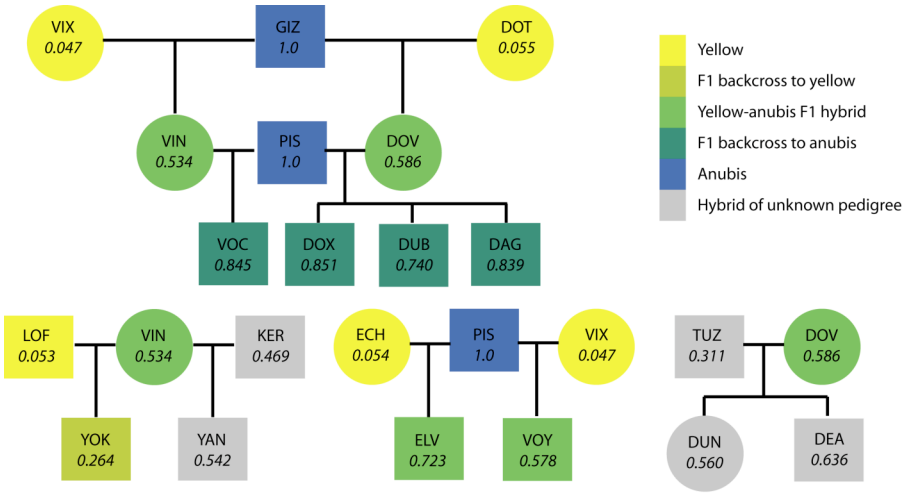

Figure 3: Pedigrees showing a subset of the hybrid crosses and backcrosses that we have observed in the Amboseli population. All genotyped offspring and some grandoffspring of two anubis male immigrants (GIZ and PIS) are shown, as well as crosses between other hybrids in the population. Note that four of PIS' offspring are GIZ's grandoffspring. The genetic hybrid score for each individual is shown in italics below the three-letter ID. Circles represent females and squares represent males; yellow, F1 hybrid, and backcrossed individuals (based on pedigree relationships and genetic hybrid score) are represented as different colors.

These results suggest that an individual's genetic hybrid score is a good representation of genome wide hybridization. The same analysis conducted on a distinct dataset, the morphological hybrid scores, yielded $r = 0.588$ (n = 151, $p < 0.0001$). Parent-offspring resemblance in hybrid scores based on morphological traits, although high, is apparently not as consistent a metric as one based on genetic markers. Both measures indicate a general ability to assign hybrid scores across a wide range of degrees of admixture.

2.3.4 Simulation results

The results of all three simulations are summarized in Figure 4. Together, they showed that the genetic hybrid scores we assigned were 1) highly repeatable given the parameters of the observed data (Figure 4a); 2) robust to modest errors in measuring allele frequencies (but less so to the more extreme errors that would result if, for example, we had sampled only 5 individuals) (Figure 3b and c); and 3) robust to cases in which rare alleles were not sampled (Figure 3d, e, and f). In particular, the results of Simulation 2 suggest that increasing the number of anubis individuals in the analysis tends to stabilize the point estimates of genetic hybridity, but that we have achieved much of this stability already by sampling 13 individuals. Interestingly, Simulation 3 suggests that rare alleles in the anubis population provide very little information about hybridity in the Amboseli population.

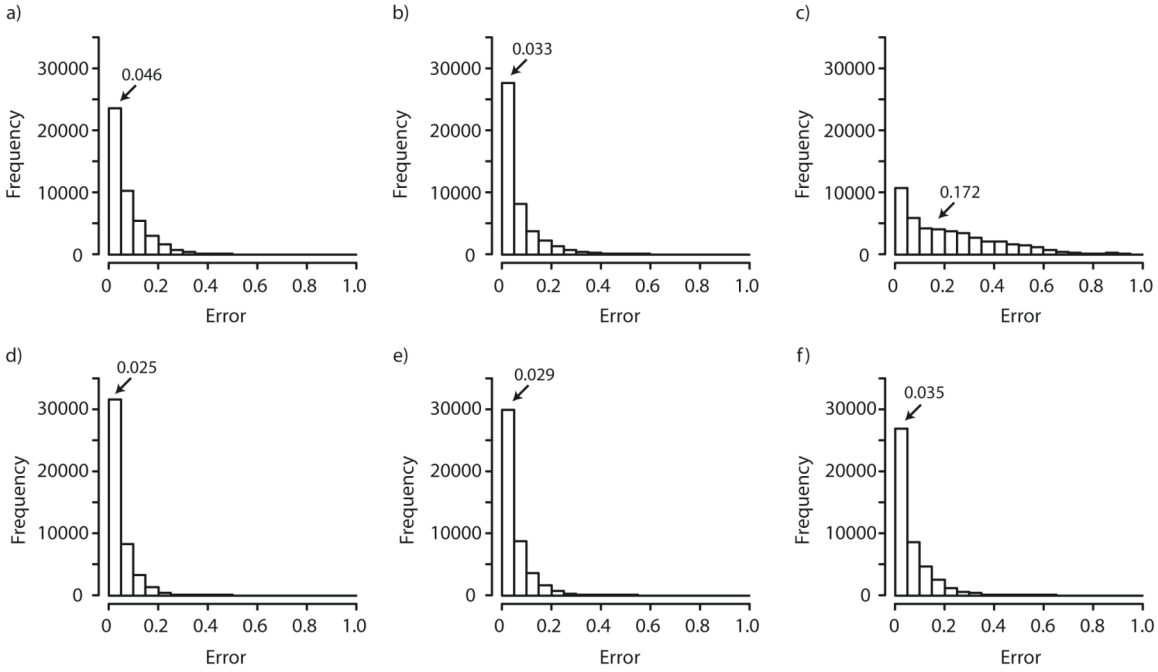

Figure 4: Simulation results. Arrows show the value of the median error for each simulation. a) Results from Simulation 1, showing the distribution of the margin of error between inferred genetic hybrid scores and known, simulated admixture proportions for each of 450 individuals in 100 simulated datasets (n = 45,000). Margins of error were calculated as the absolute value of the difference between the inferred hybrid score from Structure runs and the actual simulated degree of anubis ancestry for each individual. b) and c) Results from Simulation 2, showing the distribution of the absolute value of the difference between inferred genetic hybrid scores for runs in which 10 anubis individuals were included and for runs in which 5 anubis individuals were included, respectively, and corresponding hybrid scores assigned to the same individuals in the full analysis. d), e), and f) Results from Simulation 3, showing the distribution of the absolute value of the difference between inferred genetic hybrid scores for runs in which a rare allele was not sampled at 1, 5, and all 14 loci, respectively, and corresponding hybrid scores assigned to the same individuals in the full analysis. All Structure runs were conducted using the same parameters we applied to the observed data.

2.3.5 Changes in patterns of hybridization over time

The percentage of individuals born into the Amboseli population with hybrid ancestry increased in the study groups over the time period we considered (Figure 5a). Whereas none of the baboons in the sample born from 1968 – 1979 had anubis ancestry based on our criterion, 12.8% of the genotyped individuals born during the 1980's (15 of 117 animals), 25.1% of animals born during the 1990's (47 of 187), and 31.3% of those

born from 2000 – 2004 (36 of 115) had anubis ancestry based on our criterion. This suggests that the proportion of individuals with anubis ancestry increased in the Amboseli population throughout the study, but that the rate of increase slowed after the year 2000.

In contrast to the increase in the percentage of hybrids born over time, our results suggest that the mean genetic hybrid score assigned to hybrid individuals actually decreased in the population over time (Figure 5b). The mean hybrid score among hybrids decreased by 0.055 between animals born in the 1980's (0.522 ± 0.099 SD) and those born in the 2000's (0.467 ± 0.131 SD), and the variance in hybrid scores within data partitions increased. This suggests that the hybrids we detected were increasingly offspring of backcrosses and crosses between hybrids (see Figure 3 for examples), and not first generation F1 hybrids.

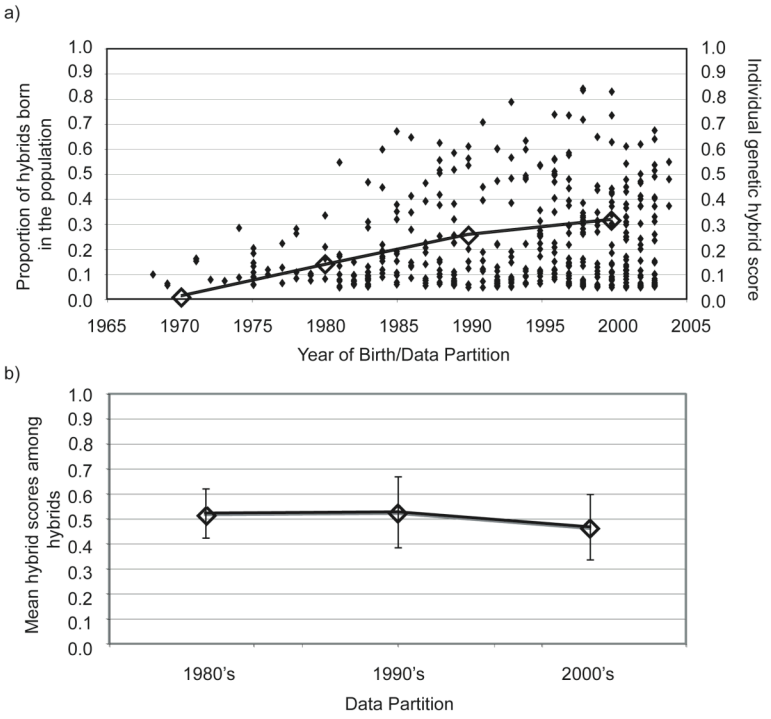

Figure 5: Patterns of admixture over time. a) Broken line showing the proportion of hybrid baboons born in the Amboseli population during different decades, based on genetic hybrid scores (left y-axis; hybrids are defined as individuals for whom the lower bound of the 90% credible interval on the genetic hybrid score is > 0.05), and scatterplot of the genetic hybrid scores of all 450 Amboseli individuals used in the analysis (right y-axis), plotted against year of birth. The number of hybrid individuals has increased over time, as has variation in the amount of anubis admixture among hybrids. b) Mean genetic hybrid score among hybrids born in Amboseli over the same period of time.

This trend was further supported by the results of pairwise Kolmogorov-Smirnov tests (Table 3) comparing the frequency distributions of all hybrid scores (including both hybrids and non-hybrids) across the sequential datasets. We observed a subtle but significant decrease in the distribution of hybrid scores among individuals born in the 1980's and the 2000's, but no significant differences between sequential decades, which was unsurprising given that the decrease in average hybrid score among hybrids was also only detectable on this scale. The overall pattern suggests that, while the representation of hybrids in the population has increased, hybrids born today are likely

to have a smaller proportion of anubis ancestry relative to hybrids born 15 – 20 years ago. This trend has changed the distribution of hybrid scores in the population as a whole.

Table 3: D values for pairwise one-tailed Kolmogorov-Smirnov tests comparing the distribution of hybrid scores across temporal datasets within Amboseli. Significance values are given in parentheses; significant values ($p < 0.05$) are indicated in bold. Comparisons against the 1960's/1970's were not conducted due to the small sample size of individuals in that data partition, which included a single anubis individual and no apparent hybrids.

	1980's	1990's	2000's
1980's	*		
1990's	0.130 (0.089)	*	
2000's	**0.196 (0.012)**	0.082 (0.386)	*

2.4 Discussion

2.4.1 Robustness in the genetic hybrid score assignments

Using the clustering method implemented in Structure, we were able to assign estimates of anubis ancestry to 450 individuals in the Amboseli baboon population, and to identify 99 of the 450 individuals as highly probable hybrids. Although we used only a modest number of markers and a small number of anubis individuals for this analysis, the results of the simulations suggest that our scores are robust and reliable measures of genetic hybridity. Additionally, the agreement we observed within parent-offspring triads shows that the genetic hybrid score assignments are consistent with expectations from pedigree data. Most importantly, although the 90% credible intervals surrounding many of the genetic hybrid score estimates are large (but comparable to those in other similar studies: see Beaumont *et al.* 2001; Pierpaoli *et al.* 2003), the lack of strong genetic differentiation between temporal datasets suggests that year of birth and error in genetic hybrid score assignment are unlikely to be correlated. Thus, while this uncertainty adds noise to the dataset, it is unlikely to have created or significantly altered the trends over time we have identified.

We also observed good agreement between the genetic hybrid scores assigned here and previous morphologically based estimates of hybridity, especially given the complete independence of the two metrics and the subjectivity inherent in assigning morphological hybrid scores. Our results are comparable to the results of Beaumont *et al* (2001), who also compared morphological methods and genetic methods of assessing hybridity in wildcat-domestic cat hybrids. They report a "strong correlation" between these methods based on a significant Spearman rank correlation (Spearman's rho = 0.372, $p < 0.01$); this is similar to the significant correlation we report of r = 0.484. Because genetic and morphological scores represent two completely independent methods of assessing yellow-anubis ancestry, this result provides strong support for the assertion that we are accurately identifying hybridity in our study population. Much of the discrepancy between the two scores occurred when genetic estimates indicated a hybrid background but morphological scores suggested these individuals were yellow. Such results are similar to those of Pierpaoli *et al.* (2003) in wildcats and Noren *et al.* (2005) in foxes, in that they also identified probable hybrids that show no clear morphological signature of hybridity ("cryptic hybrids").

These differences may reflect a greater sensitivity to detecting hybridity using genetic markers than with phenotypic traits in some cases. However, because of the modest number of loci used here and the conservative threshold we used to classify hybrids, it is also likely that genetic assignments will produce some false negatives and false positives. Increased confidence in a genetic hybrid score can be conferred when independent assessments of hybridity, such as morphologically based hybrid scores, corroborate the genetic hybrid score. Overall, the results of our simulations, comparisons with morphological scores, and pedigree analysis suggest that, in general, the majority of individual estimates do not strongly differ from the "true" proportion of anubis ancestry for those individuals.

2.4.2 Dynamic patterns of hybridization among the Amboseli baboons

The results of our analyses suggest that patterns of hybridization are changing in Amboseli over time. Specifically, the number of hybrids born into the population increased from the late 1960's and 1970's through the early 2000's, although the rate of increase seems to have slowed in the final decade of the analysis. While individuals with anubis ancestry were rare in the 1960's and 1970's, hybrids were born with increasing frequency in Amboseli beginning in the 1980's, and individuals with some degree of anubis ancestry comprised more than one quarter of the baboons born into the study population by the 2000's. At the same time, the level of anubis ancestry among hybrids born into the population appears to have gradually decreased within this period, reflecting gradual introgression of "anubis-like" genetic material into the still predominantly yellow baboon population. This pattern was also apparent in the shift of the distribution of genetic hybrid scores towards lower values over the last two and a half decades.

These results indicate that the pattern of hybridization in Amboseli has been dynamic over time. This argues for the importance of observing hybrid zones over multiple time points, either by repeated sampling or by assigning individuals observed together to different age cohorts (Albert *et al.* 2006), in order to capture the magnitude and direction of these changes (see also Verardi *et al.* 2006). Understanding the dynamics of hybridization is critical because hybridization can alter population genetic patterns over time, thus impacting related evolutionary processes such as adaptation and speciation (Moore 1977; Arnold 1992; Rieseberg *et al.* 2003).

The most commonly described hybrid zone patterns do not appear to pertain to Amboseli. The rapid increase over time in the abundance of hybrids within Amboseli, despite the low level of observed *P. anubis* immigration, suggests that hybrids were not

selected against, as would be the case if hybrids exhibited reduced fitness relative to the yellow baboon parental species. Instead, hybrid individuals have clearly reproduced successfully within the Amboseli population, and have done so over multiple generations to create descendant crosses and backcrosses, as the broad distribution of the genetic hybrid scores indicates (Figure 2). In fact, both previous and current analyses suggest that hybrid males in this population mature and disperse at an earlier age than yellow males (Alberts and Altmann 2001; Charpentier *et al.* 2008) and this may confer a selective advantage on hybrid males (see discussion in Charpentier *et al.* 2008).

However, the possibility of hybrid advantage runs counter to our observation that the distribution of hybrid scores has shifted downwards through time. A simple pattern of consistent anubis immigration and subsequent anubis and hybrid advantage over yellow baboons would lead one to predict higher rather than lower average anubis ancestry in the population over time, while an alternative pattern of hybrid superiority over both parental types would predict the maintenance of a steady intermediate level of anubis ancestry. These two conflicting pieces of evidence lead us to hypothesize that the dynamics of the Amboseli hybrid zone are driven by both non-selective processes – specifically, stochasticity in the immigration rate of anubis males into Amboseli – and selective processes – specifically, an advantage experienced by hybrids relative to yellow males. Such an advantage, coupled with a low rate of anubis immigration, would account for the increase over time in the number of hybrids in the population and the simultaneous decrease in anubis ancestry among hybrids, as well as the unchanged genetic distance between Amboseli and the Masai Mara anubis population during the study period.

2.4.3 Nonselective processes

With respect to stochasticity in the rate of anubis immigration, we envision a scenario in which chance plays a large role in whether anubis males immigrate into Amboseli. All of the nearest possible anubis source populations are moderately far from Amboseli, and are separated from it by physical obstacles, particularly a large stretch of waterless land inhospitable to baboons. The severity of these physical obstacles will presumably fluctuate over time due to local changes in habitat or weather, resulting in a low rate of anubis male immigration that varies stochastically over time. These barriers create a moderate degree of geographically-mediated prezygotic isolation between these populations.

Existing data on the Amboseli baboon population indicate that anubis male baboons have immigrated into the population at a mean rate of about once every six years (Alberts and Altmann 2001). How far do they have to travel to do this, and is this within the typical range of dispersal distances for male baboons? If the necessary travel distance is on the extreme end of dispersal distances, then this would account for the low and variable rate of anubis immigration into Amboseli. Although most male baboons disperse to neighboring baboon groups during both natal and secondary dispersal (Samuels and Altmann 1986; Alberts and Altmann 1995), males occasionally disperse much farther: currently, several males natal to the Amboseli study groups are resident in groups up to 30 km from their natal home range (Alberts & Altmann, unpublished data). (Rogers and Kidd 1996) used Wright's isolation by distance model (Wright 1946) to estimate that two-thirds of male yellow baboons in the Mikumi region of Tanzania dispersed less than 15 – 22 km from their natal groups, based on the effective population size of the Mikumi population and estimates of population density. We applied Wright's model to the Amboseli data in a similar manner, using an effective

population size of 1037 – 3456 (estimated from genotype data presented in (Storz *et al.* 2002) and a population density of 1.5 baboons/km^2 in the late 1980's and early 1990's (Samuels and Altmann 1991). The resulting estimate suggests that about two-thirds of Amboseli males dispersed less than 10.5 – 19.1 km from their natal groups during this time. If members of the source *P. anubis* population show a similar pattern of dispersal, and if the proportion of anubis immigrants into the Amboseli population is approximately 0.025 (~6 immigrant males over the 30-year study period were anubis), then these individuals would potentially have to travel some 20.6 – 37.5 km to reach the Amboseli basin. As noted above, the degree to which the physical environment in this stretch of land operates as a barrier to crossing this distance would introduce an additional degree of stochasticity to these events.

2.4.4 Selective processes

However, those anubis males that immigrate successfully into Amboseli do successfully reproduce (Samuels and Altmann 1991; Alberts and Altmann 2001) and the number of hybrids in the population has in fact increased over time. These observations support the hypothesis, posed above, that selective processes might be acting alongside stochastically varying rates of gene flow to influence the dynamics of this hybrid zone. Specifically, we hypothesize that early hybrid male maturation relative to yellow males reflects a selective advantage that has contributed to the increase over time in the number of animals with anubis ancestry (see discussion in Charpentier *et al.* 2008).

The hypothesis that hybrids are advantaged relative to yellow baboons is also supported by the geographic patterning of genetic variation in *Papio* in the wild (Newman *et al.* 2004; Wildman *et al.* 2004), which has led C.J. Jolly (personal communication) to argue that anubis baboons represent an "invasive" phenotype

relative to other members of the genus *Papio*. According to this hypothesis, the anubis phenotype is engaged in a gradual process of range expansion driven by dispersal of anubis males into other *Papio* populations. The patterns of earlier dispersal and earlier maturation observed among hybrid males (Alberts and Altmann 2001; Charpentier *et al.* 2008) may represent one mechanism by which this invasive tendency is manifested.

2.5 Conclusions

We report changing patterns of hybridization in the Amboseli baboon population over the past three and a half decades. Specifically, we observed an increased abundance of hybrids during this time, coupled with a shift in the population to a decreased level of hybrid ancestry among hybrid individuals. These patterns emphasize the utility of long-term observations on hybrid zone dynamics; we would not have been able to identify these trends using samples from any single point in time. By utilizing longitudinal data, we not only identified the presence of trends over time, but also began to identify the evolutionary and demographic influences that have shaped the particular hybrid zone dynamics within this population. Our results suggest that some selective advantage among hybrids may combine with low gene flow and stochastic variance in dispersal to produce the patterns we have observed. These hypotheses are amenable to additional testing, using both empirical data on life history markers and reproductive success within the study population, and detailed theoretical population genetic models.

3. Allele specific gene expression in wild nonhuman primates

3.1 Background

The relationship between genetic variation and phenotypic variation is a fundamental component of evolutionary change. Because many of the phenotypes of greatest evolutionary and ecological interest are complex traits, dissecting the genotype-phenotype relationship for these traits continues to be a substantial challenge. This is especially true in studies of nonmodel systems in the wild, for which inbred lines cannot be constructed and for which extensive genomic resources are not yet available. Nevertheless, some of the most ecologically and evolutionarily well characterized systems on the phenotypic level fall in this category (e.g., Grant 1986; Clutton-Brock 1989; Clutton-Brock and Pemberton 2004; Kruuk and Hill 2008). In these cases, prior knowledge about trait variation and its fitness impact in the wild would doubly reward efforts to link genetic variation to phenotypic variation. For instance, while the finding that variation in the calmodulin and *BMP4* genes influences beak shape in Darwin's finches (*Geospiza* sp.) was itself a major contribution to evolutionary genetics, its significance was greatly expanded by the existence of long-term observational work on the relationship between beak morphology, feeding behavior, and ecological niche differentiation (Grant 1986).

However, studying the functional genetics of natural populations remains challenging. Many of the tools that have helped reveal functional genetic variation in the laboratory (knock-out and knock-in models, gene silencing, transgenics) are not applicable to organisms in the wild. Additionally, gathering large sample sizes can be challenging, particularly for endangered or long-lived organisms, and controlled breeding

experiments may not be possible or desirable, especially if the goal of the study is to evaluate genotype-phenotype relationships in a natural ecological context.

Allele-specific gene expression (ASGE) assays, also known as allelic imbalance assays, directly assay genome function and can be broadly applied to both laboratory model systems (Wittkopp *et al.* 2004; de Meaux *et al.* 2005; de Meaux *et al.* 2006; Campbell *et al.* 2008; Gruber and Long 2008; Wittkopp *et al.* 2008; Zhang and Borevitz 2009) and nonmodel systems (Yan *et al.* 2002; Morley *et al.* 2004; Pastinen and Hudson 2004; Cheung *et al.* 2005; Guo *et al.* 2005; Schaart *et al.* 2005; Serre *et al.* 2008; Tung *et al.* 2009; von Korff *et al.* 2009; Heap *et al.* 2010). These assays measure the relative contribution of the two alleles of the same gene to total gene expression within the same individual. When one allele drives significantly higher expression than the other allele, that gene shows evidence of ASGE. Thus, unlike most types of gene expression measurements, ASGE assays unambiguously indicate a causal basis for gene expression variation that lies in *cis* to the gene (a *cis*-acting effect influences only the copy of the gene on the same physical chromosome; effects that influence both alleles of the gene, such as environmental or genetic background effects, are *trans*-acting). This is because, in comparisons between alleles of a gene within individuals, the *trans* genetic and *trans* environmental backgrounds are held constant.

The requirements for developing allele-specific expression assays are minimal: in principle, any gene that harbors a single nucleotide polymorphism (SNP) in its transcribed sequence can be assayed for ASGE. In addition, ASGE can be measured from samples obtained directly from organisms in the field (Tung *et al.* 2009). Thus, unlike *in vitro* assessments of functional genetic variation, the functional relevance of this variation to organisms under natural environmental conditions is indisputable. Both theoretical arguments (King and Wilson 1975; Carroll 2005; Wray 2007) and empirical data (Tournamille *et al.* 1995; Steiner *et al.* 2007; Jeong *et al.* 2008; Hofmann *et al.* 2009;

Linnen *et al.* 2009; Tung *et al.* 2009; Wittkopp *et al.* 2009) indicate that *cis*-regulatory genetic variation can be an important factor in shaping downstream phenotypes, including fitness-related traits. Using ASGE assays to identify the presence of functional *cis*-regulatory variants can therefore serve as an important first step in connecting genotype and phenotype. The approach is particularly powerful because, in combination with regulatory sequence data, ASGE measurements can also be used to identify the causal functional variants themselves (de Meaux *et al.* 2005; Milani *et al.* 2007; Serre *et al.* 2008; Babbitt *et al.* 2009; Tung *et al.* 2009).

ASGE measurements have been used as a tool to disentangle the global contributions of *cis*- and *trans*-acting factors to gene regulation (Yan *et al.* 2002; Morley *et al.* 2004; Pastinen and Hudson 2004; Zhang and Borevitz 2009), to investigate the relationship between *cis*-regulatory variation and genetic divergence within and between species (Wittkopp *et al.* 2004; Gruber and Long 2008; Wittkopp *et al.* 2008), and to test specific hypotheses about functional genetic variation that influences a given locus (de Meaux *et al.* 2005; Tao *et al.* 2006; Zhu *et al.* 2006; Babbitt *et al.* 2009; Linnen *et al.* 2009; Tung *et al.* 2009; Wittkopp *et al.* 2009). With the exception of studies in humans, most of this work has been conducted on model systems, or at least systems in which controlled crosses can be made. ASGE assays have generally not been brought to bear in studies of natural populations, despite several advantages of this approach. First, ASGE measurements control for *trans*-acting variation, which is particularly important when working with systems for which inbred lines cannot be constructed, and where sampling conditions in the field cannot be standardized. Second, ASGE studies require relatively small sample sizes. For example, Milani *et al* (2007) used a sample size of only 13 cell lines to identify functionally variable sites in eight cancer-related genes (Milani *et al.* 2007). Even some of the most comprehensive studies in humans have relied on modest sample sizes on the order of 30 – 100 individuals (Cheung *et al.* 2008; Serre *et al.* 2008).

This is an important advantage when studying organisms in natural field conditions, for which sampling RNA can be challenging.

Here, we present evidence that, using allele-specific expression measurements, we can reliably detect allelic imbalance in samples taken directly from individuals in a natural population. Our results suggest that this strategy is a useful method for investigating functional *cis*-regulatory genetic variation and its reaction norms in the field as well as in the laboratory, as demonstrated by the concrete examples arising from this work.

Specifically, we focused on a wild population of savanna baboons (*Papio cynocephalus*) that have been monitored continuously since 1971 as the focus of a long-term study in the Amboseli basin of southern Kenya (Altmann and Altmann 1970; Altmann *et al.* 1996; Buchan *et al.* 2003; Alberts *et al.* 2006). This research has produced a large body of knowledge on environmental and phenotypic variation in this population (Altmann and Alberts 2003; Silk *et al.* 2003; Beehner *et al.* 2006; Charpentier *et al.* 2008; Charpentier *et al.* 2008), making it an ideal candidate system for integrating genetic data into an existing ecological framework. We analyzed expression data on ten genes expressed in whole blood, and reanalyzed data on one gene (*FY*) that we previously characterized in another study (Tung *et al.* 2009), for a total gene set of eleven genes. These genes were chosen because they are all known to harbor functional genetic variation in humans (McKusick-Nathans Institute of Genetic Medicine and National Center for Biotechnology Information 2010), suggesting that they might also be functionally variable in baboons.

We were able to (1) validate the use of allele-specific expression measurements on samples obtained under field conditions; (2) estimate the proportion of genes that exhibit common functional *cis*-regulatory variation in the Amboseli population; (3) identify specific variants that associate with allele-specific expression for two of these

genes by combining ASGE data with *cis*-regulatory sequence data; and (4) test for gene-environment interactions (GEIs) involving *cis*-regulatory variation by combining ASGE results for these two genes with behavioral and ecological data on the same individuals. Together, our results indicate that measuring allele-specific expression differences can serve as a practical method for exploring functional regulatory genetic variation in wild populations, even when sampling conditions cannot be highly controlled, and even with modest sample sizes.

3.2 Materials and methods

3.2.1 Study subjects

The Amboseli basin is a semi-arid short-grass savanna in southern Kenya, bordering Tanzania on the south. The Amboseli baboon population consists of primarily yellow baboons (*Papio cynocephalus*) with some hybrid admixture from immigration of anubis baboons (*Papio anubis*) from outside the basin (Samuels and Altmann 1986; Alberts and Altmann 2001; Tung *et al.* 2008). Five study groups composed of individually recognized animals are currently monitored on a near-daily basis within the larger population: life history, behavioral, and physiological data are recorded for all individuals, maternal pedigrees are available for all natal individuals, and paternal pedigrees are available for many (Buchan *et al.* 2003; Alberts *et al.* 2006). The individuals used in this study were 101 adult baboons (55 females and 46 males) representing all five main study groups and one group that is monitored for demographic information only, on a monthly basis. Samples were collected between 2005 and 2009.

All study subjects were anesthetized with a Telazol-laden dart using a handheld blowgun. Darting occurred in the morning (0700 to 1200), when animals descended from known sleeping sites. In order to minimize disruption to the study groups, darting only occurred when no individuals within the group would observe the actual dart delivery,

and we darted no more than two animals per day, no more than three days a week. Anesthetized baboons were quickly removed to a processing site distant from the rest of the group. We collected RNA samples by drawing two 2.5 mL samples of whole blood into PaxGene Vacutainer tubes (BD Vacutainer), which protect RNA from environmental degradation and prevent further transcription post-draw. We also collected blood samples for DNA extraction. Upon regaining consciousness, study subjects were placed into a covered holding cage until fully recovered from the effects of the anesthetic (~ 3 – 4 hours). They were then released in the vicinity of their group. All subjects rejoined their social groups quickly upon release and without incident.

Blood samples were stored for no more than 3 days in an evaporatively cooled charcoal structure at Amboseli, which maintains a temperature at about 10 °C below ambient. They were then shipped to Nairobi, where they were either preserved frozen at -20 °C until they could be hand couriered to the United States, or, in a few cases, immediately extracted at the Institute of Primate Research in Nairobi (see Section 3.2.4). RNA extractions were conducted using the PaxGene RNA kit (Qiagen) and RNA was reverse transcribed into cDNA (High Capacity cDNA Archive Kit; Applied Biosystems) for subsequent pyrosequencing. DNA samples were extracted for each study subject using the DNEasy DNA Extraction kit (Qiagen).

3.2.2 Candidate gene assay development

All eleven candidate loci used in this study are well studied in humans with respect to disease risk and progression, and all contain segregating genetic variants in human populations that have been associated with disease-related phenotypes, many of which are *cis*-regulatory (McKusick-Nathans Institute of Genetic Medicine and National Center for Biotechnology Information 2010). Additionally, either intraspecific sequence data or interspecific comparisons in humans or nonhuman primates have suggested

interesting selective patterns for several of these loci (Hamblin and Di Rienzo 2000; Bamshad *et al.* 2002; Hamblin *et al.* 2002; Hughes *et al.* 2005).

Table 4: Genes included in this study.

Gene	Gene name	Role	n	ASGE range[1]	p-value[2]
CCL5	chemokine (CC motif) ligand 5	pro-inflammatory chemokine	36	0.201 – 3.32	< 0.0001
CCR5	chemokine (CC motif) receptor 5	membrane-bound chemokine receptor; T-cell entry point for HIV	25	-0.960 – 0.518	0.1651
CD14	monocyte differentiation antigen CD14	monocyte cell surface marker; recognizes bacterial lipopolysaccharide	7	0.112 – 0.425	0.0923
CXCR4	chemokine (CXC motif) receptor 4	membrane-bound chemokine receptor; T cell entry point for HIV	50	-0.420 – 0.418	0.0001
FY	Duffy antigen receptor for chemokines	non-specific chemokine receptor; erythrocyte receptor for *Plasmodium vivax* malaria	38	-0.002 – 2.13	< 0.0001
IL10	interleukin 10	anti-inflammatory cytokine	31	-0.491 – 0.108	< 0.0001
IL1B	interleukin 1-beta	pro-inflammatory cytokine	33	-0.111 – 0.104	0.7103
IL6	interleukin 6	pro-inflammatory/anti-inflammatory cytokine	13	-0.741 – 0.001	0.0011
LTA	lymphotoxin alpha	lymphocytic cytokine involved in the inflammatory and antiviral response	36	-0.429 – 0.463	0.5798
TAP2	transporter, ATP binding cassette, major histocompatibility complex, 2	MHC cluster gene involved in antigen presentation to T cells	15	-0.481 – 0.402	0.4512
TNF	tumor necrosis factor	pro-inflammatory cytokine, also involved in apoptosis	8	-0.125 – 0.039	0.0781

[1]Range refers to the range of mean $\log_2$-transformed corrected ASGE values for each individual, across all replicate measurements. CCL5, CXCR4, FY, and IL10 reflect samples collected from 2005 – 2009; all other genes include samples from 2005 – 2008.
[2]Uncorrected p-values for common ASGE were derived from 10000 random permutations of the data, as described in the Methods section.

All methods of measuring ASGE depend on the presence of one or more segregating SNP variants in the transcribed region of a target gene. Allele-specific assays are applied to individuals who are heterozygous for this variant, because by discriminating between the two alleles at the transcribed SNP, the two gene transcripts (and, by proxy, their linked *cis*-regulatory regions) can also be differentiated. We identified common transcribed SNPs segregating in the Amboseli baboon population by sequencing transcribed regions of the eleven candidate loci in an ascertainment panel of 10 – 12 unrelated baboons. We focused specifically on identifying intermediate frequency SNPs. These SNPs are the most useful variants for constructing ASGE assays because multiple individuals are likely to be heterozygous at these sites. Thus, we designed ASGE assays only around transcribed SNPs with an estimated minor allele frequency of at least 10%. We identified suitable transcribed SNPs for all eleven loci, and designed one assay each for all genes except for *FY*, for which we designed two assays as reported in Tung *et al.* (2009).

We then genotyped these SNPs using pyrosequencing (using the PyroMark Q96 MD instrument and PyroGold reagents, Biotage) or direct sequencing (using an ABI 3730xl sequencer and Big Dye Terminator reagents, version 3.1, Applied Biosystems) for all individuals sampled from 2005 – 2007. 1.05% of the genotypes are missing in this dataset due to failed genotyping or sequencing reactions. For those genes that suggested interesting patterns of allelic imbalance based on this subset of the overall sample set, we also genotyped and assayed individuals sampled in 2008 – 2009 (0.09% missing genotypes in this total set). At least 7 heterozygotes (range: 7 – 37 heterozygotes per gene, mean: 25.1) were assayed for each of the genes in this study (Table 4).

All sequences were visually inspected for ascertainment of variable sites in the population and identification of heterozygous individuals using Sequencher 4.8 (GeneCodes). Pyrosequencing genotypes were assigned by calculation of relative peak

heights at the variable site and/or by automated assignment using PyroMark MD software.

3.2.3 ASGE measurements via pyrosequencing

Allele-specific expression assays were conducted using pyrosequencing on a PyroMark Q96 MD instrument. Briefly, pyrosequencing is a genotyping/cycle sequencing approach that produces light emissions upon the successful addition of a complementary base to a sequencing template. When the template contains a heterozygous SNP, as in ASGE assays, the amount of light produced upon addition of one complement versus the alternative complement at that SNP reflects the relative prevalence of the two templates (e.g., Yan *et al.* 2002; Wittkopp *et al.* 2004). We conducted pyrosequencing-based ASGE assays for individuals heterozygous at the transcribed assay SNP for each candidate locus. For each gene-individual combination, we ran four replicates produced from four independent initial PCR reactions on each of two replicate plates. Thus, a total of eight measurements were obtained for each individual for each candidate gene. For *IL6*, greater technical variance in the assay led us to measure each individual twelve times (across three plates). Each measurement corresponds to the ratio of expression of one allele of the gene versus the alternative allele of the same gene. For example, if two alleles could be discriminated based on a C/T transcribed SNP, allele-specific differences in expression would be represented as the signal for the allele carrying the "C" variant divided by the signal for the allele carrying the "T" variant.

PCR assays sometimes preferentially amplify one allele of a gene over the alternative allele. In order to control for this source of technical bias, we ran corresponding assays using genomic DNA extracted from the same individuals in parallel with the expression assays (as in Wittkopp *et al.* 2004; Tung *et al.* 2009). Two

independent replicates were run on each plate, for a total of four genomic DNA controls per individual per gene (six for *IL6*). Both alleles of a gene should be equally represented in genomic DNA; thus, any differences identified from the genomic controls can be used to generate a correction factor for technical bias equal to $1/b$, where b represents the ratio measured in the genomic DNA. Except where noted, we analyzed ASGE by multiplying the average correction factor over the two genomic DNA controls per individual per gene in a plate with all parallel measurements of gene expression for the same individual-gene combination on the same plate. These corrected ratios were then $\log_2$-transformed for downstream analyses.

3.2.4 Robustness of pyrosequencing-based ASGE results for samples collected in the field

RNA is less stable than DNA, and RNA profiles have been known to change post-sampling, depending on the quality of storage conditions and the timeliness of follow-up analyses. ASGE measurements are less likely to be vulnerable to these problems than total gene expression measurements because they focus on the relative expression of the two alleles of a gene, not the total absolute expression of the gene. Additionally, comparisons are made within individuals, so both alleles are exposed to the same environmental conditions *in vivo*, during sampling, and during post-sampling transport. These qualities should make ASGE measurements well suited to studies for which sampling must be conducted under field conditions. To test this hypothesis, we investigated whether field sampling protocols compromise the quality of ASGE results through experimental validation.

Specifically, we conducted a comparison of ASGE measurements for three of the genes (*CCL5*, *CXCR4*, and *TAP2*) in our gene set under three different storage conditions, ranging from ideal conditions to substandard conditions. The genes chosen for this analysis include a gene for which strong common ASGE was detected (*CCL5*), a gene for

which weak common ASGE was detected (*CXCR4*), and a gene for which no signature of common ASGE was detected (*TAP2*). This allowed us to assess whether sampling condition either exaggerates or reduces ASGE estimates for genes that show ASGE, or whether it can introduce false positives for genes that do not show ASGE.

We darted 8 individuals in Amboseli in March 2009 and collected blood in PaxGene RNA tubes for later RNA extraction. We then shipped the blood samples via a half hour flight to Nairobi, Kenya, 20 – 24 hours after collection, the earliest transport time possible. Each sample was taken directly to the Institute of Primate Research (IPR), which maintains a molecular biology lab equipped for RNA extraction. At IPR, each sample was subdivided into three parts, so that the first subsample could be extracted at IPR, and the second and third subsamples could be transported to the United States following our normal protocols for comparison.

Subsample set 1 reflected ideal conditions, in that the samples were never frozen and were extracted at the earliest possible time point, the day after samples were collected in the field. Indeed, waiting 24 hours post-sampling has been recommended for clinical samples collected in PaxGene tubes in order to improve cell lysis and increase overall RNA yield (Wang *et al*. 2004); our own experience with samples collected from captive animals is in agreement (unpublished data).

Subsample set 2 followed our standard protocols, as outlined above. These samples were frozen in Nairobi, beginning approximately 24 hours after collection, until they could be hand-couriered to the United States on ice. They were then kept at 4° C for less than 3 days prior to RNA extraction.

Subsample set 3 reflected poorer conditions than in our standard protocol, and also poorer conditions than recommended by the PaxGene tube manufacturers. These samples were frozen in Nairobi, hand-couriered to the US on ice, and then kept at 4° C for 10 full days prior to RNA extraction, twice the amount of time recommended by the

PaxGene manufacturers. By including this treatment in our validation experiment, we tested whether we could induce biased allele-specific gene expression measurements by using substandard RNA storage protocols.

Six additional animals were darted and sampled in December 2008 or January 2009 for other purposes. RNA samples in PaxGene tubes for these animals had been stored at -20° C in Nairobi. We also divided these samples in three parts (subsample set 1: extracted at IPR; subsample set 2: extracted in the United States < 3 days after arrival; subsample set 3: extracted in the United States 10 days after arrival) and included them in a second tier of comparisons, after our analyses showed no significant differences in sample quality between these samples and those collected in March 2009.

We genotyped the test subjects for the validation experiment at all three test loci. We then evaluated ASGE for assay SNP heterozygotes for these genes. These assays followed our normal protocols, except that in each case, each individual was represented by three separate cDNA samples: one extracted in subsample set 1 (ideal conditions), one extracted in subsample set 2 (standard conditions), and one extracted in subsample set 3 (poor conditions). We repeated each set of assays across two independent plates. We then used general linear mixed models to model the $\log_2$-transformed value for ASGE, treating sampling condition as a fixed effect and plate identity and individual as random effects. Models were fit using the *lmer* function in the *lme4* package in R; *p*-values were assigned using the *pvals.fnc* function in the *languagesR* package in R (Team 2007). If measurements of allele-specific expression were robust to field sampling conditions, then we expected to see no significant effect of sampling condition within this model. Because measurements for all individuals at *CCL5* suggested ASGE, we also assessed the correlation between *CCL5* measurements for the same individuals across the three conditions.

3.2.5 Assessment of allele-specific gene expression for each locus

In order to identify functional *cis*-regulatory variants in the Amboseli population, we focused on genes that commonly exhibited ASGE within our sample (i.e., those for which ASGE occurred in multiple individuals, not in one or a few individuals; rare cases of ASGE are more likely to reflect rare genetic variants that would be difficult to associate with *cis*-regulatory genetic variation). We assessed common ASGE by comparing the measurements made on cDNA for a given gene to the control genomic DNA measurements for the same gene, using raw $\log_2$-transformed values for both sets of measurements.

We evaluated the significance of common ASGE for each gene by randomly permuting the labels (cDNA or gDNA) over these values. Because more cDNA measurements were made for each individual than gDNA measurements, we first randomly subsampled the number of cDNA measurements for each individual to equal the number of gDNA measurements. We repeated this subsampling routine 10,000 times. We then calculated a p-value for each subsampled data set using a two-tailed nonparametric Wilcoxon summed ranks test, which tested whether the cDNA values were significantly different than the values for the gDNA set. We took the mean of this set of p-values to be the nominal p-value for the gene. This value was then compared to a distribution of p-values derived from random permutations, which provided a null distribution on p-values for each gene.

3.2.6 Sequencing of gene regulatory regions

In order to identify genetic variants associated with ASGE, we focused on the four genes that exhibited the strongest evidence for common allelic imbalance. We sequenced 0.65 – 0.82 kilobases upstream of the transcription start site for the set of individuals assayed for each of these genes. For *IL10*, we also sequenced 0.72 kilobases

in the 3′ untranslated region and 3′ flanking region because our analyses suggested that the upstream sequence did not explain observed ASGE patterns and because the assay SNP used for this gene is located in its last exon (which also contains the 3′ UTR). These regions were identified based on close sequence similarity to the annotated promoter/*cis*-regulatory region in humans and macaques. Variable sites were identified by visual examination of the resulting sequence traces, and genotype assignments were produced for each individual-gene combination based on the sequence data.

3.2.7 Association between ASGE data and regulatory variants

ASGE is caused by *cis*-regulatory genetic variants that functionally differ in their abilities to drive gene expression. Because ASGE reflects the ratio of gene expression between alleles within individuals, only individuals that are heterozygous at these variants will therefore exhibit significant ASGE. This leads to the expectation that heterozygotes at a functional *cis*-regulatory variant will exhibit more extreme values of ASGE than homozygotes for the same variant. We used this expectation to test for an association between the *cis*-regulatory variants detected in the sequenced regulatory regions of the four commonly imbalanced genes, and the pyrosequencing data. For these genes, we expanded the original dataset (individuals darted in 2005 – 2007) to include an additional set of 32 individuals darted in 2008 – 2009, as indicated above.

We used general linear mixed models to analyze variation in allelic imbalance in all assayed individuals. Genotypes were coded as heterozygous or homozygous at known *cis*-regulatory variable sites (excluding singletons in the sample, which are too rare to account for common ASGE and also impossible to analyze in this context) and treated as fixed effects within the model. For perfectly linked *cis*-regulatory sites, we analyzed genotype at only one representative site. Year of sampling was treated as a random effect. Parameter estimates for all model effects were conducted using the *lme4*

package in R 2.8.1 (Team 2007). For two genes, *CXCR4* and *IL10*, the distribution of ASGE included a large number of both negative and positive values (i.e., imbalance was detected at the assay SNP in both directions; Table 4). This effect may result from incomplete linkage between an assay SNP and a causal *cis*-regulatory SNP, leading to a case in which heterozygotes at the regulatory site exhibit both more negative and more positive values of ASGE than homozygotes. To avoid misidentifying the genes that exhibited increased variance as genes for which *cis*-regulatory genotype has no effect, we therefore used "unsigned" ASGE values (i.e., the absolute value of the $\log_2$-transformed ASGE values; see for example Babbitt *et al*. 2009) for these two genes.

We assigned *p*-values for each model effect using random permutations of the allelic imbalance measurements for a given individual against individual identity (as in Tung *et al*. 2009). We then used a backwards model selection procedure to sequentially eliminate the model effect with the highest *p*-value, until all *p*-values were below 0.10. For each gene, several individuals in the data set were close relatives; to ensure that genetic correlations between these individuals did not produce a false signal of association, we also analyzed the data after eliminating individuals in the data set so that no close relatives were included. In each case, eliminating different sets of individuals could produce this outcome; however, in no case did elimination of close relatives qualitatively change the results.

3.2.8 GEI effects on gene expression

A significant correlation between ASGE and environmental effects suggests the presence of a gene-environment interaction in which the *trans*-acting environment modifies the effect of the *cis*-regulatory variant(s) (de Meaux *et al*. 2005; von Korff *et al*. 2009). Understanding GEIs that influence evolution in the wild may prove to be a particularly important contribution of evolutionary genetic work on natural populations.

To explore this possibility, we tested for the presence of GEIs involving *cis*-regulatory variation in the two genes for which significant common ASGE was detected and could be associated with a known *cis*-regulatory genotype. We focused on an environmental effect of known importance in the Amboseli population: the social rank of an individual's mother, at the time of that individual's conception (i.e., maternal dominance rank), which is known to exert long-term effects on maturation timing and stress hormone profiles in this population (Alberts and Altmann 1995; Charpentier *et al*. 2008; Onyango *et al*. 2008).

To investigate the possibility of GEIs that influence ASGE, we analyzed the residuals from the previous model of ASGE on genotype in the context of a general linear model. We stratified the data by genotypic class at the associated SNP (heterozygous or homozygous) based on the expectation that, if an environmental effect modifies the effect of a functional *cis*-regulatory variant, a relationship between ASGE and the environment should be observed in heterozygotes for this variant, but not in homozygotes. This expectation arises because an environmental interaction with *cis*-regulatory variation should not be observable via ASGE measurements if the two alleles for an individual are not functionally differentiated (i.e., homozygous).

P-values for this analysis were assigned by running the same analysis after permuting the response variable (residuals of a model taking into account year of sampling and the genotype effect) with respect to the explanatory environmental variable. A null distribution of effects was obtained from 1000 random permutations, and the probability of observing an effect size greater than the estimated effect from the unpermuted data was taken as the p-value for the test (equivalent to a two-tailed test).

3.3 Results

3.3.1 Allele-specific gene expression measurements are robust to field sampling conditions

We were able to analyze variation in ASGE measurements in three individuals darted in the main validation set for *CCL5* and *CXCR4*, and two individuals darted in the main validation set for *TAP2*. Sampling condition was not a significant effect within the model for any of the three genes (*CCL5*: n = 9 sets of individual by treatment measurements, $p = 0.837$; *CXCR4*: n = 9, $p = 0.677$; *TAP2*: n = 6, $p = 0.194$); nor was sampling condition significant when including measurements made for the five individuals darted in January 2009, which increased the sample size for *CCL5* to 7 baboons and for *CXCR4* to 6 baboons (*CCL5*: n = 21 sets of individual by treatment measurements, $p = 0.501$; *CXCR4*: n = 18, $p = 0.941$). Additionally, for *CCL5*, the correlation across sampling conditions was high and significant in all three pairwise comparisons (Table 5). A summary of the results for all three genes is depicted in Figure 6.

Table 5: Correlations between *CCL5* measurements obtained under different sample storage conditions. Values of *r* are given for each cell with associated *p*-values from 1000 random permutations in parentheses.

	Subsample 1	*Subsample 2*	*Subsample 3*
Subsample 1	*	0.986 (0.001)	0.998 (0.007)
Subsample 2	*	*	0.993 (0.007)
Subsample 3	*	*	*

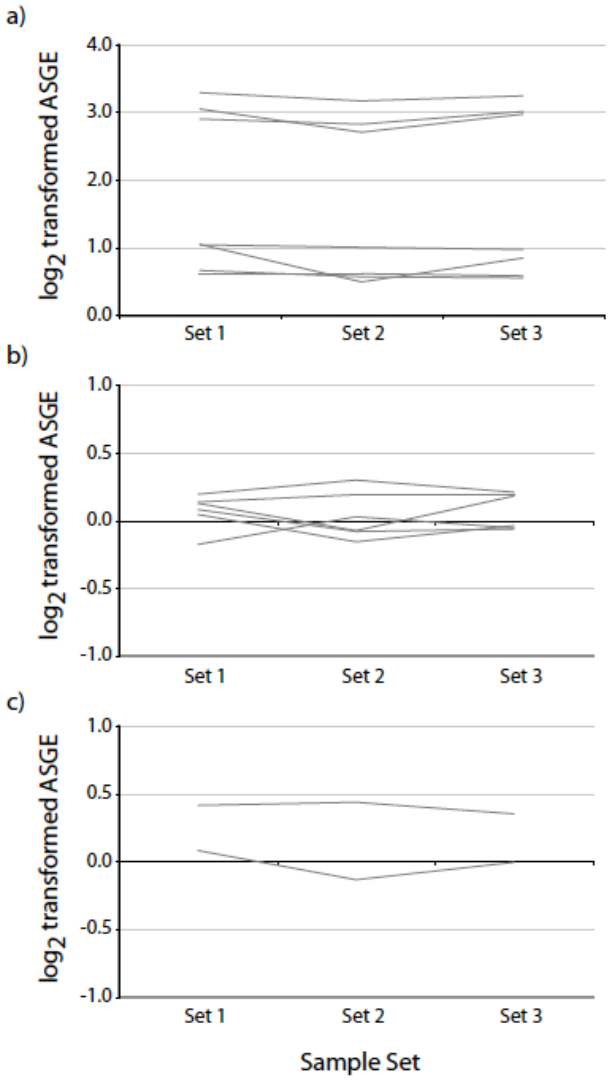

Figure 6: ASGE measurements are consistent across sample handling treatments for (a) *CCL5* (n = 7 individuals), a gene that exhibits common, large-scale ASGE; (b) *CXCR4* (n = 6 individuals), a gene that exhibits significant ASGE at smaller scales; and (c) *TAP2* (n = 2 individuals), for which we detected no significant common ASGE. Each line represents a sample from a single individual, subdivided into the three subsample sets, where Set 1 refers to the subsample set treated under ideal conditions; Set 2 refers to the subsample set treated under our standard protocol; and Set 3 refers to the subsample set treated under substandard conditions, as described in the Methods. Points on each line correspond to the mean $\log_2$-transformed ASGE value for that subsample for that individual.

3.3.2 Allele-specific gene expression is common in the Amboseli baboons

Of the eleven loci included in this study, five (45.4%) exhibited significant evidence for common allele-specific gene expression in the Amboseli baboon population; four of the five (36.4%) remained significant at $p < 0.01$ following Bonferroni correction for multiple testing (*CCL5*, *CXCR4*, *FY*, *IL10*; Table 4, Figure 7).

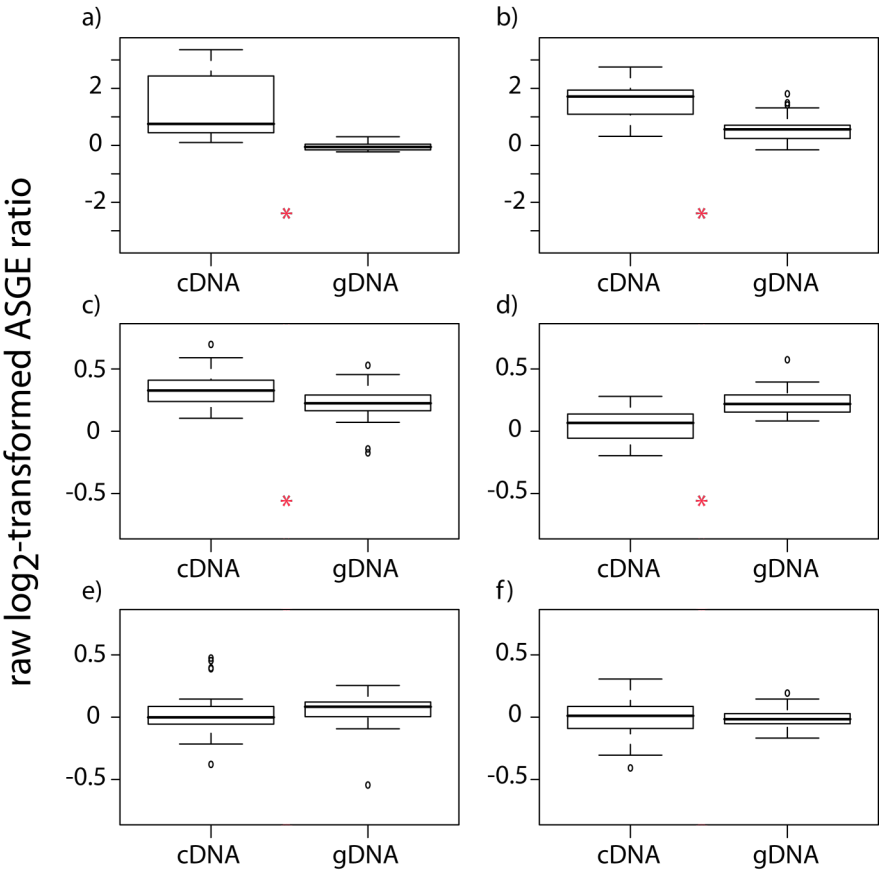

Figure 7: Example ASGE ratios for cDNA and genomic DNA (gDNA) for six genes. a) *CCL5*; b) *FY*; c) *CXCR4*; and d) *IL10* illustrate significant differences in $\log_2$-transformed ASGE ratios between cDNA samples and gDNA samples (indicated by the red asterisk); whereas, for comparison, e) *CCR5* and f) *LTA* illustrate cases of statistically indistinguishable cDNA and gDNA measurements.

Among the four genes with the greatest support for ASGE, we detected a range of effect sizes. For example, cDNA measurements almost never overlapped with genomic DNA measurements for any individual for the gene *CCL5* (Figure 7). The

average corrected, $\log_2$-transformed ASGE measurement for an individual assayed at *CCL5* was 1.38 (range: 0.201 – 3.32), corresponding to a foldchange difference in gene expression for the higher expressing allele of 2.60. In contrast, the average corrected ASGE measurement for an individual assayed at *CXCR4* was 0.112 (range: -0.420 – 0.418; 1.08 foldchange difference between alleles).

3.3.3 Associations between ASGE measurements and cis-regulatory genetic variation

Of the four genes we investigated further (*CCL5*, *CXCR4*, *IL10*, and *FY*), two genes exhibited an association between heterozygosity/homozygosity at a putative *cis*-regulatory genetic variant and magnitude of ASGE, such that heterozygotes exhibited more extreme ASGE than homozygotes (Figure 8). As in the case of ASGE itself, we observed considerable variation in effect sizes for these loci. For *CCL5*, genotype at the associated variant explains 66.5% of the variance in the overall set of ASGE measurements, after taking into account the effects of year of sampling ($p < 0.0001$, n = 14 heterozygotes and 22 homozygotes). Additionally, we observed no overlap between the range of ASGE for heterozygotes at this site (range of mean per individual $\log_2$-transformed ratios: 1.77 – 3.32) and the range of ASGE for homozygotes at this site (range of mean per individual $\log_2$-transformed ratios: 0.201 – 1.52). In contrast, the variant associated with ASGE for *FY* explained a more modest proportion of the overall ASGE variance, 22% ($p = 0.0002$, n = 18 heterozygotes and 20 homozygotes), and the ranges for ASGE measurements overlapped between heterozygotes and homozygotes (heterozygotes: 0.203 – 2.131; homozygotes: -0.002 – 1.46). In both cases, the SNPs we identified were the closest SNPs in each set to the transcription start site. Also in both cases, our results suggest that additional functional variants and/or *cis*-by-*trans* regulatory interactions also influence expression of these genes. In particular, for *CCL5*, even the homozygotes for the associated *cis*-regulatory variant exhibited strong signals of

ASGE, although the magnitude of allelic imbalance for these animals was attenuated relative to heterozygotes.

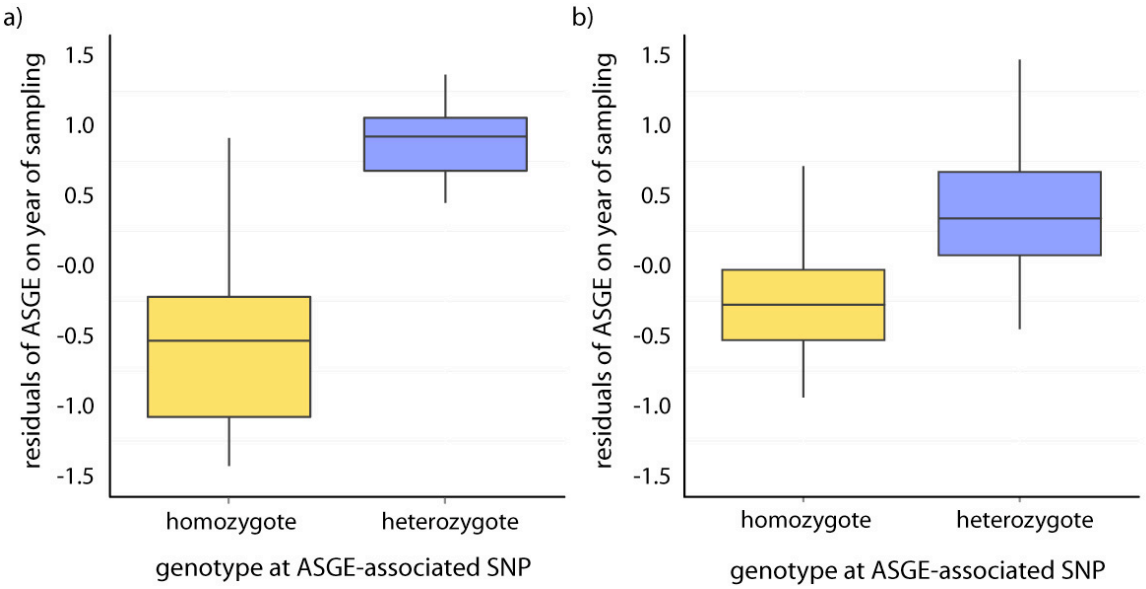

Figure 8: Heterozygotes at ASGE-associated SNPs exhibit more extreme levels of ASGE than homozygotes at a) *CCL5* ($p < 0.0001$) and b) *FY* ($p = 0.0002$).

We were unable to identify an association between SNP genotypes in the immediate *cis*-regulatory region and ASGE levels at *CXCR4* and *IL10*. Both of these genes exhibited more modest levels of ASGE than *CCL5* and *FY* (Table 4) and were characterized by variation in ASGE that encompassed both positive $\log_2$-transformed ratios and negative $\log_2$-transformed ratios, probably reducing the power to detect an association. Additionally, we surveyed only a small region of sequence in which *cis*-regulatory variants may occur. For *IL10*, we analyzed both variation upstream of the gene and downstream of the end of the protein-coding region. We did identify a SNP in the 3' untranslated region of this gene with weak evidence of a possible role ($p = 0.08$), but only two heterozygotes at this SNP were contained within our data set, eliminating the possibility of a well-powered analysis and strongly suggesting that this variant, even if it is indicative of a real effect, cannot account for the variation observed in the full sample.

3.3.4 GEI analysis

We identified an effect of maternal rank at the time of an individual's conception on gene expression of *CCL5*, such that high maternal rank is correlated with more pronounced ASGE in individuals that are heterozygous for the putative functional *cis*-regulatory site, after controlling for the direct effect of genotype on ASGE and year of sampling ($p < 0.001$; Figure 9). No such effect was detected for homozygotes at the *cis*-regulatory site ($p = 0.464$), as expected if maternal rank interacts with genetic variation captured by this site. Because maternal rank can sometimes predict the adult rank of individuals later in life, we also checked whether the maternal rank effect we identified was a proxy for the rank of the individuals themselves at time of sampling. Indeed, maternal rank was significantly correlated with rank at time of sampling for individuals assayed at *CCL5* (Spearman's rho = 0.492, $p = 0.008$, n = 28 individuals because the ranks at time of sampling for individuals that are no longer members of the five intensively observed study groups are not known). However, we found no evidence that an individual's own rank at sampling influences ASGE in *CCL5 cis*-regulatory variant heterozygotes ($p = 0.918$). Together, these results suggest that maternal dominance rank at the time of an individual's conception exerts a long-term effect on *CCL5* expression, effectively modifying the functional impact of *cis*-regulatory genetic variation. In contrast, we found no evidence for GEI involving maternal rank on *FY* expression ($p = 0.766$).

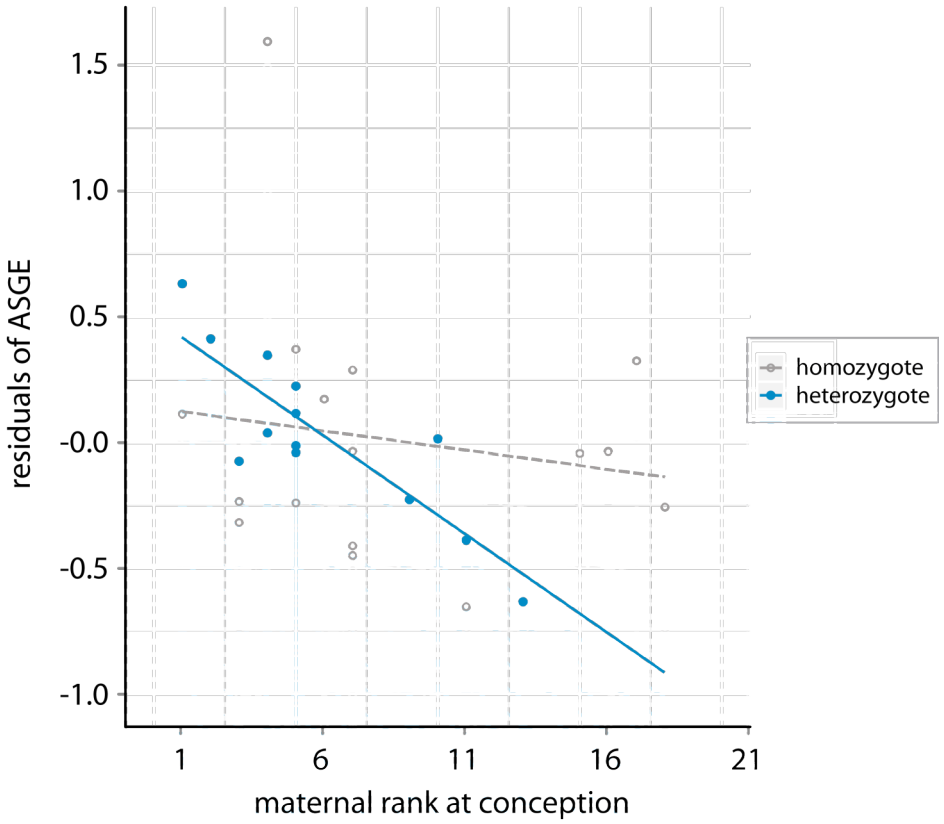

Figure 9: Maternal rank at conception influences allelic imbalance in heterozygotes at the *CCL5* putative functional *cis*-regulatory site ($p < 0.001$), but not homozygotes ($p = 0.464$), after taking into account year of sampling and the direct effect of the ASGE-associated SNP. High ranking individuals have low rank numbers (rank 1 is highest); low ranking individuals have high rank numbers.

3.4 Discussion

3.4.1 ASGE in the Amboseli baboon population

The presence of common ASGE was well supported for four of the eleven genes included in this study. Although it is difficult to compare rates of ASGE across studies, given different kinds and numbers of samples, different measurement platforms, and different statistical methods, this frequency falls well within the rather broad range of previous estimates for different taxa given in the literature (from 5% - 70% in humans, mice, *Drosophila*, and *Arabidopsis*: Yan *et al.* 2002; Pant *et al.* 2006; Milani *et al.* 2007;

Campbell *et al.* 2008; Gruber and Long 2008; Serre *et al.* 2008; Zhang and Borevitz 2009; Heap *et al.* 2010).

The four genes that exhibited evidence for common ASGE varied in both the range of ASGE detected across individuals and in our ability to map them onto genetic variants. Notably, the two genes that we were able to link to *cis*-regulatory genetic variants were those that exhibited the greatest magnitude of ASGE, and the most consistent direction of imbalance. These results imply, perhaps unsurprisingly, that genes that exhibit more pronounced functional differentiation between alleles and tighter linkage between functional *cis*-regulatory variation and the transcribed region of a gene will be the most tractable subjects for ASGE studies in natural populations. Additionally, our results suggest that prior functional studies in model and/or laboratory organisms can inform the choice of candidate genes in natural populations, and may in fact lend themselves to work on parallel genetic evolution across species, as exemplified by work incorporating ASGE measurements species in *Drosophila* (Wittkopp *et al.* 2009), mice (Linnen *et al.* 2009), and primates (Tung *et al.* 2009).

Our failure to associate ASGE with *cis*-regulatory variation at two loci that exhibit common allelic imbalance, *CXCR4* and *IL10*, highlights the fact that, in some cases, associating ASGE with putative causal regulatory variants will be difficult. This is particularly true for species like baboons, for which relatively little is known about segregating genetic variation. Indeed, in rare cases, *cis*-regulatory variants actually occur many kilobases away from the transcribed sequence (reviewed in Wray *et al.* 2003), outside the scope of surveys of sequence close to the gene. Thus, even if a distant causal variant is genotyped, it is much less likely to be in strong linkage disequilibrium with a transcribed assay SNP. Either or both of these conditions may have held in the cases of *CXCR4* and *IL10*. However, recent evidence from humans suggests that most functional *cis*-regulatory variants probably do lie close to either the transcription start site or the

transcription end site (Veyrieras *et al.* 2008). Given that we surveyed only a small amount of *cis*-regulatory sequence for the genes in this study, such findings are promising for future studies that adopt a similar approach but incorporate more detailed genotype data. In particular, resequencing tens to hundreds of kilobases of contiguous sequence will become increasingly feasible in the near future, with the development of improved sequence capture and high throughput sequencing methods.

Indeed, an expanded search space for functional *cis*-regulatory variants would likely also benefit genes like *CCL5* and *FY*. Although we identified significant associations between ASGE at these genes and *cis*-regulatory variation, in both cases substantial variation in ASGE measurements remained unexplained. In the case of *CCL5*, all homozygotes at the associated *cis*-regulatory SNP also exhibited strong evidence of allelic imbalance, although it was reduced relative to heterozygotes at this site. In agreement with results from humans (Tao *et al.* 2006) and *Drosophila* (McGregor *et al.* 2007; Gruber and Long 2008), then, our findings suggest that ASGE is a complex trait influenced by multiple genetic variants, and that *cis*-regulatory interactions with *trans*-acting environmental or regulatory genetic effects may also play a role. More extensive characterization of additional regulatory variants around these genes would aid in identifying some of these effects.

3.4.2 ASGE measurements and genetic studies of natural populations

Our results lead us to conclude that measuring allele-specific gene expression, whether via pyrosequencing or other methods (e.g., Heap *et al.* 2010), is a practical method for identifying functional *cis*-regulatory genetic variation in nonmodel systems, including in organisms sampled from natural populations. One of the major advantages of this method is that allele-specific measurements appear to be largely robust to sampling and transport conditions in the field. Even samples that were pushed beyond

the boundaries of our normal protocol yielded results very similar to samples from the same individuals that were maintained under ideal conditions. Additionally, identification of genes for which ASGE appears to be common sets up a natural strategy for pinpointing the underlying causal variants. First, because ASGE implies a *cis*-acting mechanism, the search space for such variants is constrained to regions that are likely to be linked to the gene in question (i.e., sequence in or close to the transcribed sequence used in the assay). Second, heterozygotes for a functional variant are expected to exhibit a greater magnitude of ASGE than homozygotes, providing a statistical framework in which to test the correlation between genotype and expression. Third, independent, well established tests for functional regulatory variation that influences gene expression, including electrophoretic mobility shift assays and *in vitro* cell culture assays, can be applied for further validation if good candidates are identified (e.g., Kurreeman *et al.* 2004; Tao *et al.* 2006; Zhu *et al.* 2006; Babbitt *et al.* 2009; Tung *et al.* 2009). Finally, ASGE measurements can be readily scaled from one gene to dozens of genes (and possibly thousands of genes, if using next-generation high-throughput sequencing methods: dealing with such data will likely require a sequenced genome, however (Degner *et al.* 2009). While this scale is modest relative to methods for high-throughput gene expression, it means that studies that utilize ASGE do not require initial high cost investments that may be challenging for field investigators.

As for all methods, however, using ASGE-based approaches will have several limitations. Most obviously, ASGE measurements are specifically relevant to regulatory variation, not structural variation. For traits for which genetic variation in protein-coding sequence makes the primary contribution to phenotypic variation (for example, in some cases of pelage coloration: Nachman *et al.* 2003; Hoekstra *et al.* 2005), ASGE assays will have little utility. An additional major challenge for studies in the wild is that ASGE measurements require RNA samples. While immense progress has been made in

noninvasive sampling of DNA, collecting RNA requires live tissue samples, which can be difficult to gather for some species. However, because the sample sizes required for measuring ASGE are modest, opportunistic sampling of tissues from recently dead individuals may sometimes be able to address this issue. For difficult systems, studying functional regulatory variation in tissues that are easier to collect, like skin and blood, will represent the most feasible approaches for the foreseeable future.

Where applicable, though, ASGE assays have potential to be of value in several regards for studies of wild populations. For example, ASGE-based surveys of many genes can be used to assess levels of functional genetic diversity within populations, the raw material for evolution. The four genes in this study that exhibited the best evidence for common ASGE in the Amboseli population (*CCL5*, *CXCR4*, *FY*, and *IL10*), for example, are linked by their functional roles in the primate immune system. All four genes are involved in cytokine or chemokine signaling and play a part in mediating the inflammatory response. Controlling inflammation is a crucial component of the immune response, and likely important for wild baboons, which are subject to a wide array of both pathogen infections and physical insults; we have noted enlarged and inflamed lymph nodes on many of the baboons sampled in this population during darting efforts (unpublished data). Functional differences that vary the expression levels of these genes may therefore prove important in fine-tuning this response, and suggest that substantial genetic variation is segregating in the Amboseli baboons that could impact phenotypic evolution in this population.

Indeed, for some genes, evidence from other types of analyses may indicate a relationship between genetic variation at those loci and organism-level traits of interest. In these cases, ASGE can be used to test whether the basis for this link is functional *cis*-regulatory variation, suggesting that gene expression variation is the mechanistic link between genotype and phenotype. This approach has already been applied to work on

morphological trait variation within and between *Drosophila* species (Wittkopp *et al.* 2009) and between subspecies of deer mice (Linnen *et al.* 2009), both of which used ASGE measurements to follow up a possible link between color patterning and candidate genes suggested by other data. In the case of this study, several of the genes for which we identified common ASGE in the baboons have also been associated with organism-level susceptibility to pathogens in humans, including possible links with HIV/AIDS (Liu *et al.* 1999) and with malaria (Tournamille *et al.* 1995). This kind of prior information may inform downstream strategies for exploring whether functional *cis*-regulatory variation in these genes also influences similar traits in nonhuman primates (e.g., Tung *et al.* 2009).

Finally, ASGE analyses can be used to help reveal how *cis*-regulatory effects can be modified by environmental variation, a phenomenon that may play an important role in shaping the genetic architecture of complex traits in natural populations (Qvarnstrom 1999; Wilson *et al.* 2006). This approach is likely to be most informative where environmental effects that play an important role in the population of interest have already been identified. Our results for *CCL5* provide an example of such a case: we found evidence that the dominance rank of an individual's mother, at the time of that individual's conception, appears to exert long-term effects on the magnitude of ASGE for this gene. Long-term effects of maternal dominance rank on phenotypic variation have already been documented for this population (Onyango *et al.* 2008), and early life effects involving social status and access to resources are well known in humans and other animals (Ravelli *et al.* 1976; Altmann 1991; Lindstrom 1999; Qvarnstrom 1999; Godfrey and Barker 2000; Reifsnyder *et al.* 2000; Barker 2002; Barker *et al.* 2002; Weaver *et al.* 2004; Hoffjan *et al.* 2005; Meaney and Szyf 2005; St Clair *et al.* 2005). To our knowledge, however, these data represent the first evidence of a GEI involving an early life effect in wild primates. With respect to this specific case, they suggest the possibility

that early social environment in baboons may help shape the baboon immune system over the long-term, such that the same environmental exposure produces different consequences for different individuals. This result echoes findings in humans that early life environment, including exposure to other children, influences the risk of asthma and allergy in a context-dependent manner (Hoffjan *et al.* 2005; Ober and Thompson 2005); indeed, *CCL5* attracts and stimulates histamine release in basophils, an important component of the pro-inflammatory allergic response (Laing and Secombes 2004). With respect to the broader phenomenon of GEIs, our results also suggest a strategy for further work aimed at revealing how common such effects may be, and whether GEIs are enriched for specific types of environmental exposures or for particular classes of genes. For such studies, the availability of fine-grained environmental data from field studies will be invaluable.

3.5 Conclusions

Natural populations have long served as important models for behavior, demography, evolution, and speciation. Recently, interest has been building in bringing complementary genetic perspectives to bear on the same systems. Indeed, the availability of nongenetic data for these systems can provide important insight into the evolutionary significance of genetic analyses. Here, we explore one method of integrating functional genetic work into research on natural populations by extending allele-specific gene expression studies from the laboratory to the field. In agreement with studies in humans, mice, and other model systems, we found that functional *cis*-regulatory variation is common in wild baboons, but that the degree of ASGE detectable for different genes and the power to detect associated genetic variants varies across loci. Our results also demonstrate that measuring allele-specific gene expression is a viable method for identifying functional *cis*-regulatory variation and exploring gene-

environment interactions in natural populations, including those like the Amboseli baboons for which genetic manipulations and controlled breeding are impossible. Importantly, ASGE measurements appear to be largely robust to RNA sampling in the field, which makes this strategy generalizable to a wide variety of systems and types of studies. Because *cis*-regulatory variants can make important contributions to downstream organism-level traits, measuring ASGE may therefore open a window onto further functional genetic studies in wild populations.

4. Genomic features that predict allelic imbalance in humans suggest patterns of constraint on gene expression variation[1]

4.1 Background

A growing number of studies illustrate that variation in non-coding regions of the genome has important consequences for organismal phenotypic variation, including traits of adaptive importance (Boffelli *et al.* 2004; Wray 2007). As first suggested over thirty years ago (King and Wilson 1975), many of these relationships are mediated by effects on gene regulation. Hence, many studies now focus on regulatory DNA and its proximate molecular phenotype, gene expression, as a strategy for identifying relevant variation in organism-level morphological, physiological, and behavioral traits. Variation in gene expression is predictive of phenotypic traits both globally, as demonstrated by genome-wide expression profiling studies (Golub *et al.* 1999; West *et al.* 2001; Whitfield *et al.* 2003), and on an individual gene basis, as shown by studies connecting *cis*-regulatory genetic variation in specific genes to variation in adaptively important traits (Tournamille *et al.* 1995; Shapiro *et al.* 2004; Colosimo *et al.* 2005; Gompel *et al.* 2005; Prud'homme *et al.* 2006; Tishkoff *et al.* 2007; Jeong *et al.* 2008).

While identification of either genetic variation or gene expression variation alone is now straightforward, establishing a causal relationship between them remains challenging. For example, genetic effects on a gene's expression may be located in *cis* to the gene (such that they influence only the linked allele of the gene, in a nearby region of the same physical chromosome) or in *trans* to the gene (such that they influence both alleles of the gene, regardless of linkage), a distinction that has both practical and

[1] The contents of this chapter have been previously published as: J Tung, O Fedrigo, R Haygood, S Mukherjee, and GA Wray (2009). *Genomic features that predict allelic imbalance in humans suggest patterns of constraint on gene expression variation*. Molecular Biology & Evolution 26: 2047-2059.

biological implications. From a practical perspective, the distinction between *cis* and *trans* is important for establishing the likely physical location of the causal variant: *cis* acting variants tend to lie close to the gene of interest, whereas *trans* acting effects can reside almost anywhere in the genome (e.g., Morley *et al.* 2004; Cheung *et al.* 2005). From a biological perspective, the functional and evolutionary significance of *cis* and *trans* effects may differ. For instance, recent work has suggested that *cis*-acting effects tend to act more additively than *trans*-acting effects (Lemos *et al.* 2008); that *cis*-effects tend to be more pronounced in explaining interspecific differences than intraspecific differences, while the reverse may be true for *trans* effects (Wittkopp *et al.* 2004; Wittkopp *et al.* 2008); and that *cis*-effects may have more restricted consequences than *trans*-effects, thus mitigating adaptive conflicts arising from pleiotropy across tissues (Blekhman *et al.* 2008; Campbell *et al.* 2008), splice variants (Campbell *et al.* 2008), and/or environmental contexts (de Meaux *et al.* 2005; Zhu *et al.* 2006).

One method of discriminating between *cis*-acting effects and *trans*-acting effects involves measuring gene expression in an allele-specific manner, generally known as assaying allele-specific gene expression or "allelic imbalance" (Cowles *et al.* 2002; Yan *et al.* 2002; Bray *et al.* 2003; Lo *et al.* 2003; Pastinen and Hudson 2004; Wittkopp *et al.* 2004; de Meaux *et al.* 2005; de Meaux *et al.* 2006; Pant *et al.* 2006; Milani *et al.* 2007; Campbell *et al.* 2008; Cheung *et al.* 2008; Gruber and Long 2008; Serre *et al.* 2008; Wittkopp *et al.* 2008; Tung *et al.* 2009). Allelic imbalance describes the relative ability of the two alleles of a *cis*-regulatory region to drive expression of a linked gene within individuals: a gene is "imbalanced" when one allele drives significantly higher expression than the alternative allele. Because both alleles experience identical *trans*-acting genetic and environmental backgrounds, deviations from the null expectation (equal contribution of both alleles to total expression) unambiguously identify *cis*-acting genetic

effects (although *cis* x *trans* interaction effects can also be detected: Wittkopp *et al.* 2004; Wittkopp *et al.* 2008).

Allelic imbalance has been well documented in many systems, including human, mouse, and *Drosophila* (Cowles *et al.* 2002; Yan *et al.* 2002; Wittkopp *et al.* 2004; Campbell *et al.* 2008; Gruber and Long 2008). However, studies that have evaluated allelic imbalance in large, population-based sets of individuals suggest that *common* allelic imbalance, as opposed to imbalance that sporadically occurs in one or only a few individuals, affects only about 10 - 20% of expressed genes (Milani *et al.* 2007; Serre *et al.* 2008; Verlaan *et al.* 2009). In other words, genes that harbor functional *cis*-regulatory variation common enough to produce allelic imbalance in multiple individuals in a population (or that harbor many distinct functional *cis*-regulatory variants) are the minority, at least in humans. Given that surveying allelic imbalance in a large number of genes *de novo* is cost- or sample-prohibitive for many populations, identification of patterns that predict which genes are likely to be commonly imbalanced could therefore serve as a useful tool. Such patterns might also shed light on the molecular basis for *cis*-regulatory variation by identifying what types of genomic characteristics co-segregate with common imbalance, and what evolutionary processes produce these characteristics.

Towards that end, we applied a machine learning approach, the support vector machine (SVM) (Cortes and Vapnik 1995), to fit a predictive model for data generated in a published study of allelic imbalance in humans (Serre *et al.* 2008). Serre and colleagues (2008) validated a novel, high-throughput method of assaying allelic imbalance that produced measurements for several hundred genes, in one of the most comprehensive studies of allelic imbalance to date. Because the original study subjects were members of the HapMap CEU analysis panel, we were able to combine polymorphism data with human genome sequence data and with divergence data from human-chimpanzee comparisons to fit the model. We found that a signal of common

allelic imbalance can be extracted from these data, and that this signal predicts common imbalance with a modest, but potentially useful, level of accuracy. Further, our results were consistent when applied to a second dataset of imbalanced genes in humans identified using different methods (Cheung *et al.* 2008), suggesting that the model captures aspects of some broader biological phenomena. Hence, we explored the biological basis for the predictive ability of our model by investigating the sources of variance in the main component that contributes to the model's predictive accuracy. We found a strong explanatory effect of gene density in this analysis, suggesting that genes that reside in gene-dense regions are less likely to exhibit allelic imbalance than genes in less dense regions of the genome. Our results suggest that the important features we identified are proxies for evolutionary constraint, such that genes that exhibit common imbalance are significantly more likely to evolve under relaxed selective constraint than genes that do not exhibit imbalance.

4.2 Materials and methods

4.2.1 Allelic imbalance training set

We stratified genes into one of two mutually exclusive classes based on the dataset of Serre *et al*. (2008): genes that exhibited common allelic imbalance (the 'AI' class) and genes that never exhibited allelic imbalance (the 'non-AI' class). We chose to use the data presented in Serre *et al*. (2008) rather than other published surveys of imbalance for three reasons. First, this study surveyed allelic imbalance in a large number of genes (n = 643 that exhibited expression levels above background noise). Second, a relatively large number of individuals (n = 83) were included in the study, meaning that the authors were able to impose a more stringent cut-off criterion: for any given gene in the final dataset, allelic imbalance measurements were made on multiple heterozygotes (at least three individuals). Because this sampling scheme provided an actual

distribution of allelic imbalance for each gene, we were therefore able to distinguish commonly imbalanced genes from those that exhibit imbalance as a result of a rare mutation. We defined AI genes following the methods of the authors: these genes were characterized by high mean allelic imbalance across individuals or higher variance in imbalance measurements than under null expectations. Non-AI genes included those loci for which the mean imbalance across individuals was exactly 0 (equal expression of both alleles), and for which the variance across individuals was not significantly greater than expected by chance. Finally, the subjects in the Serre *et al.* (2008) study were members of the CEPH pedigrees included in the HapMap CEU panel, allowing us to include polymorphism-based features in the predictive models we developed. Our initial focus on the Serre *et al.* (2008) dataset also allowed us to conduct further validation of our model using data from a different published dataset, as described below.

We restricted our analysis to autosomal genes in order to avoid the confounding effects of X inactivation. In order to maintain consistency in our definition of coding regions, flanking regions, and exon/intron boundaries, we further restricted the dataset to those genes that have a current consensus annotation curated by the Consensus CoDing region Project (CCDS) for Build 36.3 of the human genome (http://www.ncbi.nlm.nih.gov/projects/CCDS/). For genes with multiple entries in the CCDS database, we always chose the annotation that maximized the size (end base pair – start base pair) of the gene in question. After filtering for autosomal CCDS-curated genes, our final training set included 103 AI genes (16% of the original 643 gene dataset) and 184 non-AI genes. We extracted genome sequence data for each gene from human genome build 36 (hsap18: http://genome.ucsc.edu/, Kent *et al.* 2002), based on the CCDS annotations for exon-intron boundaries and coding region start and stop sites.

4.2.2 Feature extraction

We modeled allelic imbalance using three sets of features: genome sequence, polymorphism data for the CEPH samples, and divergence data based on differences between the human genome and the chimpanzee genome. All feature extraction was handled using publicly available software appropriate to the different types of features, and/or custom Ruby code. The full list of features is provided in the supplementary materials for Tung *et al.* 2009.

The sequence features set included data on the presence, distribution, and abundance of four feature subsets: (1) repeat families; (2) 5-mer sequence motifs; (3) CpG islands; and (4) gene composition (i.e., exon/intron content). Except where noted below, or where not applicable, features were extracted for several different partitions of sequence around the gene: the annotated conserved coding sequence (from start of translation to end of translation), the 2 kb flanking regions, the 5 kb flanking regions, and the 10 kb flanking regions (see Supplementary Materials Table S2 for Tung et al. 2009). Repeat features were identified using RepeatMasker v 3.2.0 (Smit *et al.* 1996-2004). Five-mer sequence motifs and CpG features were identified using the compseq and newcpgreport programs, respectively, in the EMBOSS v 5.0.0 software package (Rice *et al.* 2000). Due to the large number of possible five-mers, we restricted the sequence feature set for the full model to five-mers in the flanking regions of the gene (5 kb upstream and downstream of the coding region) based on preliminary analyses that suggested that five-mers in the coding region contained relatively little information about allelic imbalance. Number and proportion of exon content for gene coding regions were extracted directly from the genome sequence data and the CCDS annotations for each gene.

Polymorphism features were identified using publicly available data on the CEU/CEPH samples for HapMap release 18 (http://www.hapmap.org). These features included data on the abundance, distribution, and proportion of different types of polymorphisms (i.e., all six possible mutations, transitions/transversions), and a d_n/d_s-like calculation of the relative number of nonsynonymous changes to synonymous changes within each gene.

The divergence features set was generated by aligning probable homologues for each locus of interest between human and chimpanzee (panTro2), and calculating the abundance, distribution, and proportion of different types of divergent sites between the two species (including unalignable sites and gaps). Probable homologues were identified using the LiftOver tool from the UCSC Genome Browser (http://genome.ucsc.edu/: Kent *et al.* 2002), and alignments were conducted using the program TBA v 12 (Blanchette *et al.* 2004). For flanking regions, the position of the chimpanzee homologue (relative to the chimpanzee gene coding sequence) is not always identical to the position of the original sequence in humans (relative to the human gene coding sequence). For example, the 5 kb upstream sequence for human for a given gene might not be precisely equivalent to the 5 kb upstream sequence for the gene homologue in chimpanzee, even when the extracted sequence itself is the correct homologue for the original human 5' sequence.

Missing features (from unalignable regions across species or from "intronic" regions of single exon genes) were imputed by the following procedure: 1) we calculated the sum of the squared difference for all features between the gene containing missing data and every other gene in the dataset; 2) we identified the five genes that were most similar to the gene containing missing data, based on the sum of squares metric; and 3) we assigned a value for the missing feature equal to the mean of the values for the five most similar genes for the same feature. Any features that resulted in a value of 0 for all

genes were removed from the dataset for downstream numerical stability. The final full feature set consisted of 2,269 features. Values for all features were scaled on the interval [0,1] based on dividing the value for each feature by the maximum value for that feature in the entire dataset.

4.2.3 Wilcoxon summed ranks tests

We applied a nonparametric Wilcoxon summed ranks test to each feature in the feature set. This analysis tested whether the values of the feature for genes in the AI class tended to be significantly different from values of the same feature for genes in the non-AI class. Under the null hypothesis of no difference between the two classes for any of the features we examined, the p-values for this series of tests should be uniformly distributed along the interval [0,1]. We compared the actual distribution of p-values to this expectation using a Kolmogorov-Smirnov test.

2.2.4 Support vector machine (SVM) classification and recursive feature selection

All SVM model fitting was conducted using SVM^{perf} (Joachims 2005; Joachims 2006):

$$\min_{w,b} \ \tfrac{1}{2} w^T w + \tfrac{C}{n} \sum_{i=1}^{n} \max\left[1 - y_i\left(w^T x_i + b\right), 0\right],$$

where $\left(x_i, y_i\right)_{i=1}^{n}$ are the n samples and $(y_i)_{i=1}^{n}$ represents the labels of the samples and $(x_i)_{i=1}^{n}$ represents the feature values. The parameters of the model are w, a column vector with each element corresponding to a feature weight, and b, the offset or intercept. The cost function was set to minimize overall error rate (the '-l 2' option in SVM^{perf}). The regularization parameter C, was set to 0.05 for the full feature set, based on exploratory analyses. In analyses of the smaller feature subsets (i.e., sequence features alone, polymorphism features alone, etc.), $C = 1$. All final analyses were

conducted using a linear kernel; exploratory analyses using a radial basis kernel function did not improve model performance. Generalization error was estimated by leave-one-out cross-validation. Specifically, we sequentially removed one gene from the dataset, fit the model on the remaining n – 1 genes, and then used the resulting model to predict allelic imbalance class for the gene that was initially removed. We asked about the concordance between the model prediction and the true value for each gene over the whole dataset, producing a measure of overall error and recall and precision for both the AI class and the non-AI class.

Recursive feature selection was also conducted in a leave-one-out framework. We removed one gene from the dataset and used n – 1 genes to fit sequential SVMs, where the results of each sequential model were used to calculate the weights for each feature and used to remove (1) first, the 300 least informative features until fewer than 1000 features remained in the model; (2) second, the 100 least informative features until fewer than 100 features remained; and (3) finally, the 20 least informative features until fewer than 20 features remained. At each step, we asked whether the model accurately predicted the allelic imbalance class of the gene that was initially removed. We repeated this procedure over all 287 genes in the dataset, resulting in a 287 x 15 matrix, where the columns represent progressively smaller model sizes (2269, 1969, 1669, 1369, 1069, 769, 669, 569, 469, 369, 269, 169, 69, 49, and 29 features respectively), and each cell takes the value 0 or 1, where 0 reflects correct prediction of the imbalance state for that gene, and 1 reflects an incorrect prediction. We used this information to evaluate the relationship between the number of features in the model and predictive accuracy.

4.2.5 Non-negative matrix factorization (NMF)

We ranked all features by frequency of occurrence in the 469-feature model over the 287 different iterations of recursive feature selection. We identified the 500 features

that occurred most often in the 469-feature model. We then factored this set of 500 features into k factors using NMF (Brunet *et al.* 2004). The reason for using NMF rather than spectral based methods (e.g., singular value decomposition) is that factors computed via NMF tend to be sparser and more localized (i.e., fewer non-zero features are contained in each factor) than those computed via spectral methods. The input to NMF was the data matrix G with element G_{ij} corresponding to the j-th feature in the i-th sample (gene). The algorithm factors G into two matrices F and M with the property that

$$G \approx FM, \text{ and } F_{ij}, M_{ij} \geq 0,$$

where F is a matrix of n rows and k columns and M is a matrix of k rows and p columns, where n equals the number of genes, p equals the number of features, and k equals the number of factors. Methods for choosing the number of factors k and for the least squares implementation to solve for F and M followed (Brunet *et al.* 2004). For our data, we obtained k = 4 factors (Figure 10).

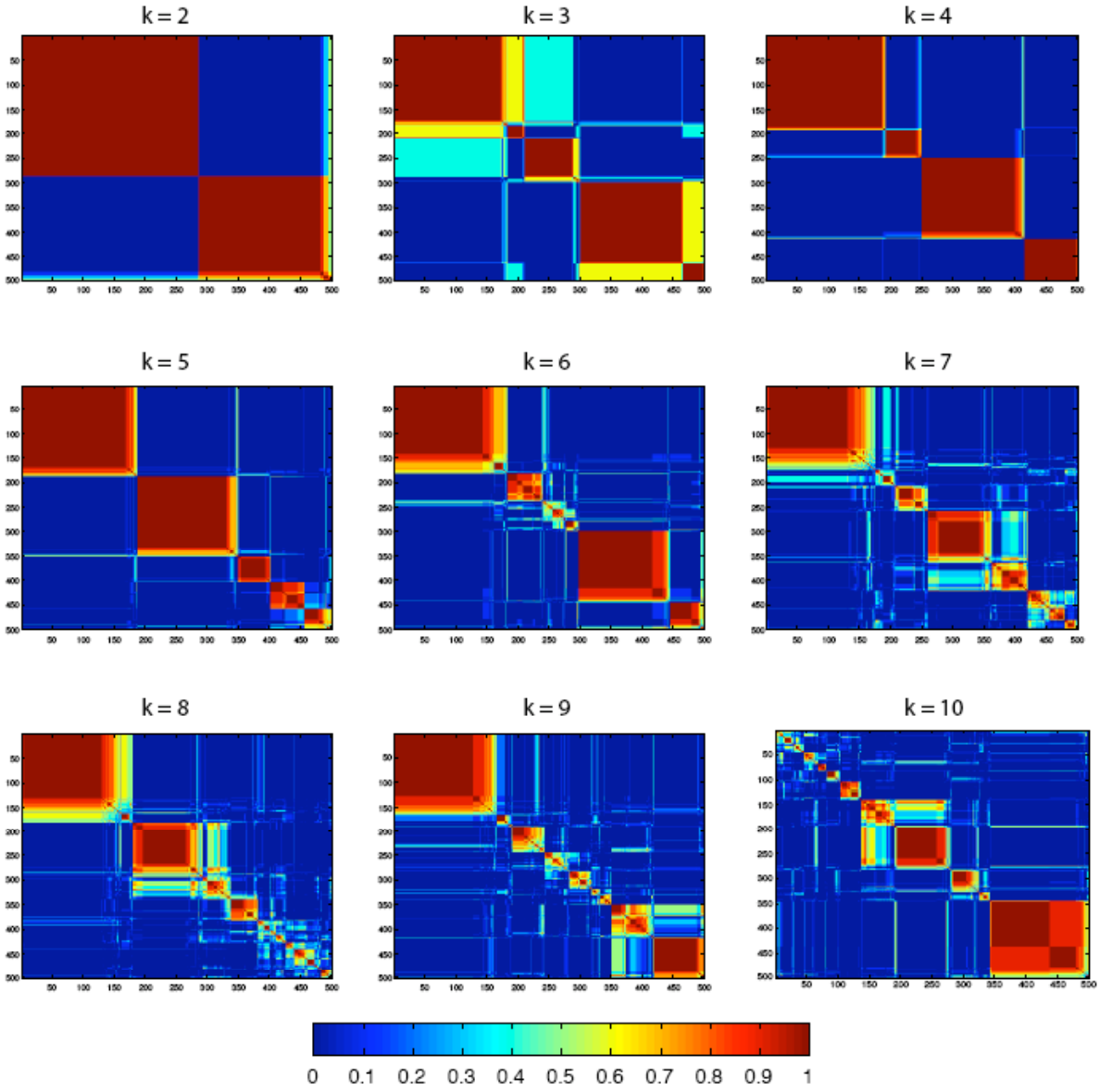

Figure 10: NMF consensus clustering matrices for varying number of clusters k. The matrix displays the frequency that the 500 features are found to be in the same cluster when the NMF algorithm is applied with different random initialization values.

We then tested whether each of the four factors that resulted from the NMF analysis individually associated with imbalance status by conducting a Wilcoxon summed ranks test comparing the distribution of factor values between the AI class and

the non-AI class. Only one factor significantly explained variation in allelic imbalance class: we termed this factor the "AI factor."

4.2.6 Validation using an external dataset

If the overall model and the AI factor identified within the model reflect general biological characteristics associated with allelic imbalance, then the results obtained on the Serre *et al.* (2008) dataset should also generalize well to unseen data (i.e., data that was not involved in the original model fit). Cheung *et al.* (2008) used genotyping microarrays to measure allelic imbalance in 21 sets of monozygotic twins and 10 members of the HapMap CEU panel. They identified 163 SNPs that revealed significant allelic imbalance in genes in their sample, after restricting this set to those SNPs assayed in at least 5 individuals (counting members of a monozygotic twin set only once). This dataset is therefore similar to that of Serre *et al.* (2008) in that it captures common imbalance in Caucasian populations. However, Cheung *et al.* (2008) used a different technology to measure gene expression (microarrays instead of the Illumina genotyping platform), and different statistical thresholds to call imbalance. Hence, cross-validation of our model on the Cheung *et al.* (2008) data represents a conservative test of the generalizability of our results.

From the Cheung *et al.* (2008) dataset, we were able to obtain data for 122 commonly imbalanced genes that were not included in the gene sets derived from Serre *et al.* (2008) (the list of genes that did not exhibit allelic imbalance were not provided in the Supplementary Materials for their paper). We then tested two hypotheses. First, we reasoned that the probability of observing allelic imbalance estimated by our model should be significantly greater for genes in the Cheung *et al.* (2008) dataset than genes in the original non-AI gene set, but should be no different from genes in the original AI gene set (where predictions for the AI and non-AI genes were obtained from leave-one-out

cross-validation, as described above). We therefore compared the predictions for the Cheung dataset with predictions from the non-AI set and predictions from the AI set using Wilcoxon summed ranks tests. Second, if the relationship between the AI factor and allelic imbalance generalizes well, we hypothesized that the values of the AI factor for the Cheung dataset would be enriched for high values compared to those for a set of genes for which AI status is unknown. We tested this hypothesis by comparing the Cheung dataset with the 3,908 genes used in the AI factor annotation analyses described below, again using Wilcoxon summed ranks tests.

4.2.7 Annotation of the AI factor

In order to annotate the AI factor, we cross-referenced it to publicly available datasets on gene expression, negative and positive selection on gene regulatory regions, and gene density.

To measure evenness of gene expression around the genome, we summarized data available from 73 non-cancerous human tissues in the Novartis Gene Expression Atlas (Su *et al.* 2004) following the method of Haygood *et al.* (in review). To compute evenness scores, we analyzed gene expression in the 73 non-cancerous human tissues included in the Novartis Gene Expression Atlas (Su *et al.* 2004). We first took the mean of gene expression for each gene over multiple (replicate) arrays per tissue, and then extracted the maximum expression over the multiple probes available per gene to obtain a single value for the expression of each gene in each tissue. Each gene was therefore associated with 73 values for gene expression (one per tissue). We then regarded the 73 values as a 73-dimensional vector and considered the angle between this vector and a vector representing perfectly even expression (i.e., a vector for which all components took the same value). The evenness score of the gene is the squared cosine of this angle,

which ranges from 1 (for perfectly even expression) to $1/73$ (for expression exclusive to a single tissue).

To measure negative selection and positive selection, we used the product of the estimate of the fraction of sites under selection in the 5 kb region upstream of a gene (f_1 in the case of negative selection and f_3 in the case of positive selection) and the estimate of the strength of selection on the same region ($1 - \xi_1$ for negative selection, ξ_3 for positive selection). Estimates of f and ξ were available for three discrete regulatory regions around each gene: the 5 kb upstream of the gene (Haygood *et al.* 2007), the 5′ untranslated region, and the 3′ untranslated region (G.A. Wray., unpublished data); we used the average over these three estimates in the analysis, excluding missing data for 5′ or 3′ UTR regions when no UTR scores were available. In this analysis, ξ is analogous to ω in a branch-specific d_n/d_s test, so that $\xi = 1$ is indicative of neutral evolution, very small values of ξ are indicative of strong negative selection, and $\xi >> 1$ is indicative of strong positive selection. Because ξ_1 is evaluated between 0 and 1, in our analysis the product of $(1 - \xi_1)$ and f also ranges between 0 and 1, where 0 corresponds to the least evolutionary constrained and 1 to the most evolutionary constrained.

To measure gene density in the region around a focal gene, we used the entries in the CCDS database to count the number of genes within 100 kb upstream and 100 kb downstream of the coding region of the focal gene. If the length of a gene spanned the 100 kb cutoff, we included it in this count.

We were able to extract the value of the AI factor and values for evenness, negative and positive selection, and gene density for 3,908 genes in the human genome. We then modeled variation in the AI factor according to the following linear model:

$$y = \beta_0 + e + s_1 + s_3 + g + \varepsilon$$

where y represents the value of the AI factor; β_0 represents the model intercept; e represents the evenness score; s_1 represents the product of f_1 and 1 - ξ_1 (i.e., the magnitude of negative selection); s_3 represents the product of f_3 and ξ_3 (i.e., the magnitude of positive selection); g represents gene density; and ε represents model error.

Model fitting was conducted using the *lm* function in R (Team 2007). *P*-values for each effect are taken directly from the model fit based on the estimated effect size and standard error around the estimate. R^2 values for single effects were calculated as the percentage of variation explained in the residuals of y regressed on all other model effects by the given single effect.

4.3 Results

4.3.1 Prediction of commonly imbalanced genes

Using the full set of sequence, polymorphism, and divergence based features (2,269 features), we were able to fit a predictive model for allelic imbalance that accurately classified 68.3% of the 287 genes in the dataset (103 of which exhibit common allelic imbalance in the Serre *et al.*, 2008, dataset). This level of classification accuracy corresponds to an area under the curve (AUC) value of 0.66. In agreement with this result, when we conducted Wilcoxon summed ranks tests comparing the distribution of values for commonly imbalanced genes versus non-imbalanced genes for each feature, the resulting distribution of p-values was strongly skewed towards low p-values, in contrast to the null expectation of a uniform distribution of p-values (as would be observed if no signal of imbalance was contained within our feature set: $p << 10^{-16}$; Figure 11).

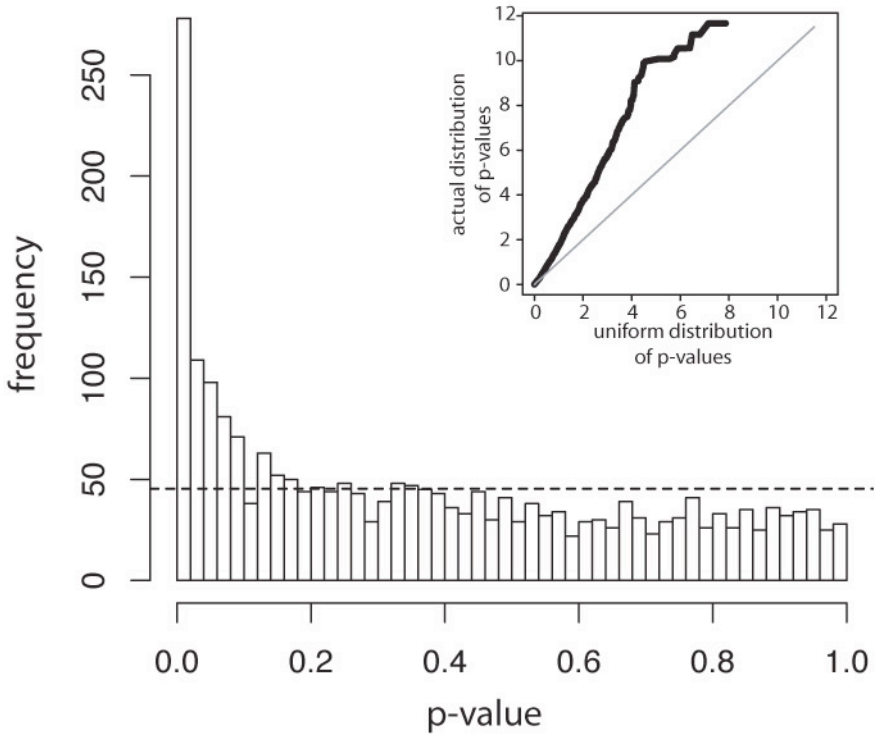

Figure 11: The distribution of *p*-values from Wilcoxon summed-ranks test on each feature. Each test compared the value of the feature for genes in the AI class and genes in the non-AI class; a low *p*-value indicates that the AI class and the non-AI class were significantly differentiated by values of the feature. The dashed line gives the expected uniform distribution of *p*-values for a case in which no such signal could be detected in the feature set. The distribution is strongly skewed towards the left (low *p*-values), demonstrating that a signal of allelic imbalance status is embedded within the original 2,269 feature set (comparison between the observed and the expected distribution: $p << 1 \times 10^{-16}$ from a Kolmogorov-Smirnov test). The inset shows a Q-Q plot of the same results (with *p*-values depicted as –log(*p*)), with the cumulative distribution function for a uniform distribution on the x-axis and the cumulative distribution function of the *p*-values for all features on the y-axis.

The estimated generalization error of this model, 31.7%, was obtained using cross-validation, a method that controls for model overfitting. Specifically, we removed one gene from the dataset, fit the model on the remaining data, and asked whether the prediction from the resulting model for the missing gene matched the actual class for that gene (either common allelic imbalance, hereafter referred to as "AI," or non-imbalanced, hereafter "non-AI"). Our results indicated that accurate classification of genes in the AI

class was much more difficult than classification of genes in the non-AI class. Recall for the non-AI class (proportion of true members of the class that were correctly identified by the model) was 88.0%, compared to only 33.0% for the AI class. Similarly, precision for the non-AI class (proportion of those genes identified by the model as members of a class that are truly members of the class) was also higher than precision for the AI class (non-AI: 70.1%; AI: 60.7%). We could obtain more equivalent results for the two classes if we allowed the generalization error to increase slightly (corresponding to decreasing the value of the regularization term): for example, as overall error increased to 35.5%, non-AI recall and AI recall values were 75.5% and 44.7% respectively. In either case, model prediction worked reasonably well—we were able to correctly predict the status of over 2/3 of genes in the dataset—but together these results suggest that the AI class is fundamentally more heterogeneous with respect to our feature set than the non-AI class (Table 6).

Table 6: Classification accuracy and precision and recall by class for the full feature set and the six possible feature subsets. The regularization parameter *c* was set to 1 in all cases except for the full feature set, where c = 0.05. Results for the full feature set with c = 1 are also shown for comparison.

	Full (c = 0.05[a])	Full (c = 1)	Seq	Poly	Div	Seq + Poly	Seq + Div	Poly + Div
Overall accuracy	68.3%	64.5%	62.7%	62.0%	65.8%	64.1%	62.7%	58.9%
AI precision	60.7%	50.5%	47.8%	45.6%	54.1%	50%	47.8%	41.0%
AI recall	33.0%	44.7%	42.7%	30.1%	32.0%	46.6%	42.7%	33.0%
Non-AI precision	70.1%	70.9%	69.7%	67.1%	69.0%	71.2%	69.7%	66.2%
Non-AI recall	88.0%	75.5%	73.9%	79.9%	84.8%	73.9%	73.9%	73.4%

One possible source of this heterogeneity is inclusion of genes that exhibit AI due to imprinting instead of due to *cis*-regulatory genetic variation. However, only 4 of the genes included in the 287 genes used to fit the model are known, provisionally known, or computationally predicted to be imprinted in humans (based on the curated set available at www.geneimprint.com). One of these genes never exhibited detectable allelic

imbalance in the Serre *et al.* (2008) dataset, suggesting that imprinting for at least this gene is specific to other tissues. In this case, removal of those genes from the analysis produced model predictions that were highly correlated with the full data set ($p < 2.2 \times 10^{-16}$, Spearman's rho = 0.993), and did not appreciably alter the model's predictive accuracy (generalization error was 31.1% when the four genes were moved). Hence, we retained all 287 genes for the downstream analyses.

Although classification in this analysis is binary, model predictions are made as continuous real numbers, where positive predictions correspond to an assignment to the AI class and negative predictions correspond to an assignment in the non-AI class. The more extreme a predicted value, the greater the certainty behind that prediction, given the fit model. This certainty can be directly expressed as a probability by passing the predicted value through a logit link function. Genes that received a more extreme predicted value, corresponding to a higher probability of common imbalance on the positive end and a lower probability of common imbalance on the negative end, tended to be classified more accurately than genes with a value closer to 0 (Figure 12).

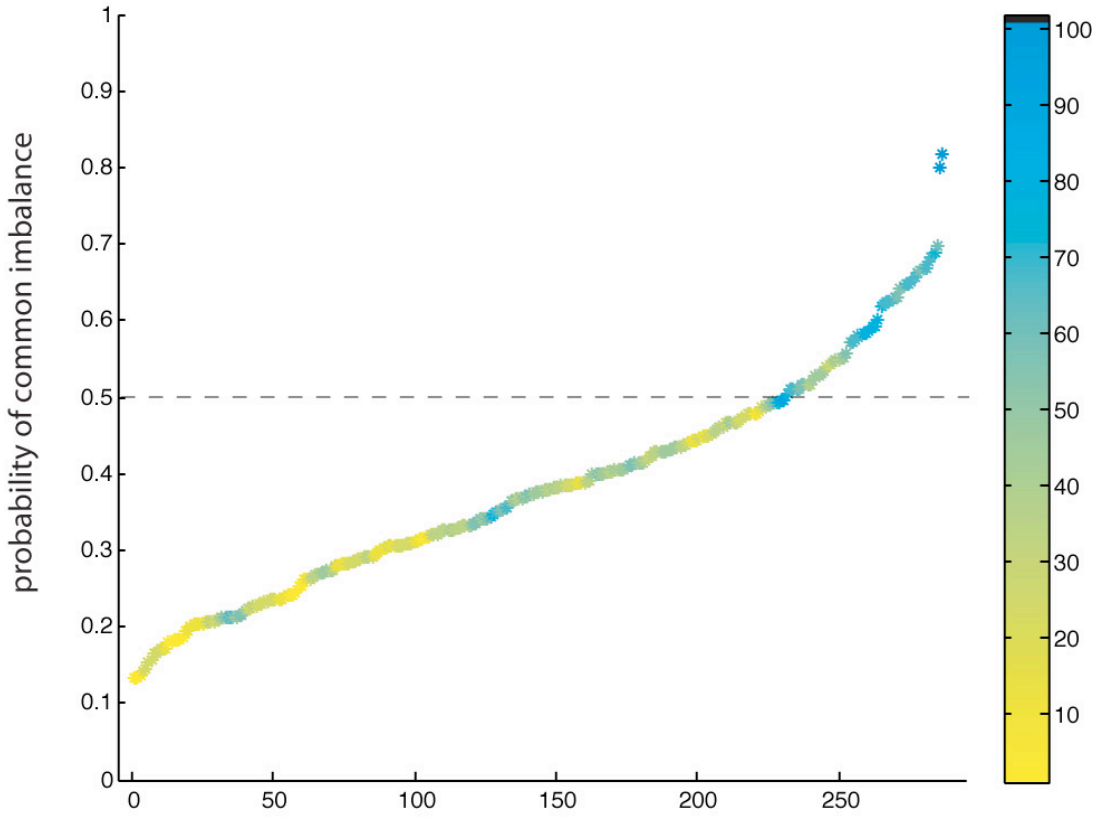

Figure 12: Genes with more extreme predicted values are more likely to be predicted correctly. Predictions from the full model were passed through a logit link function to produce a predicted probability of common imbalance at each gene. All 287 probability values are plotted, ranked from lowest probability of common imbalance to highest probability of common imbalance. True imbalance class is reflected by the color bar: yellow represents non-AI genes and blue represents AI genes. The color for each dot represents the degree to which model predictions were correct for a window size of eight genes around a given gene, in the list ordered by probability. Non-AI genes are predicted as commonly imbalanced with lower probability (lower left of the figure); AI genes are predicted as commonly imbalanced with higher probability (upper right of the figure). For comparison, perfect prediction would produce yellow dots below probability = 0.5 and blue dots above probability = 0.5, with a small region of green dots at the transition point around probability = 0.5.

Characteristics of the full feature set were generally recapitulated when using only one feature subset or only two feature subsets (of the three classes of features: sequence, polymorphism, and divergence; see Table 6). Interestingly, predictions generated from the polymorphism data set alone and the divergence data set alone were significantly correlated with each other (Spearman's $\rho = 0.248$, $p = 2.18 \times 10^{-5}$),

suggesting that the information about imbalance contained within these two data sets was somewhat redundant; in contrast, neither of these sets of predictions were correlated with predictions from the sequence data set alone (sequence predictions versus polymorphism predictions: Spearman's $\rho = 0.051$, $p = 0.389$; sequence predictions versus divergence predictions: Spearman's $\rho = 0.085$, $p = 0.150$). All three single subset models performed approximately as well, and, as was the case for the full dataset, more extreme predicted values tended to reflect more accurate classification of the gene.

4.3.2 Dimension reduction in the feature set

In order to reduce the dimensionality of the full model, we recursively eliminated features that provided the least predictive power from the model. The predictive accuracy of the model remained stable as the number of features in the model decreased from the full feature set (n = 2,269) to approximately 500 features, but dropped rapidly as the number of features grew smaller than 500 (Figure 13).

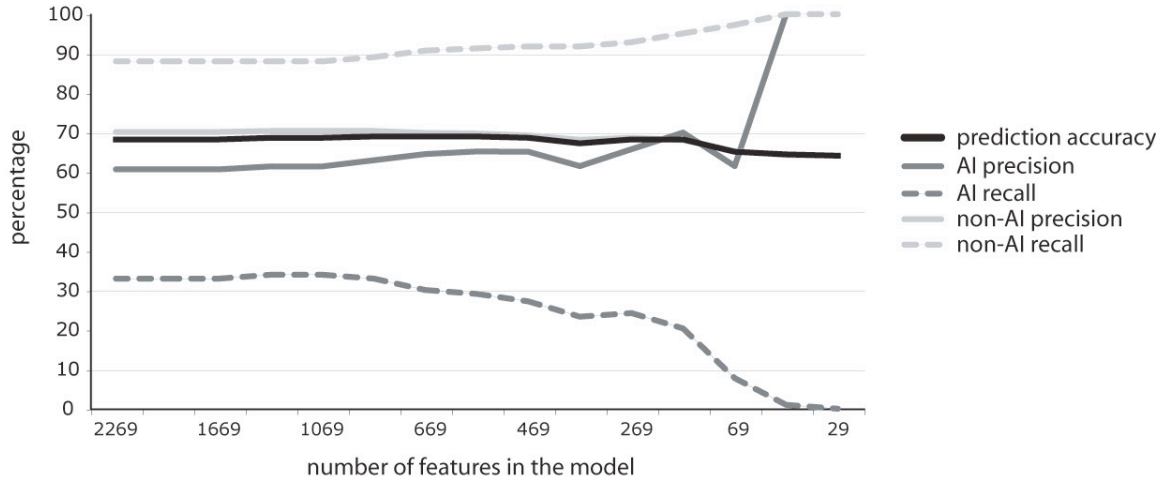

Figure 13: Results of recursive feature elimination. Predictive accuracy of the SVM decreases as the number of features in the model drops below about 500. A rapid drop in AI recall, such that true AI genes are consistently predicted as non-AI genes, predominantly drives this effect (the corresponding rise in AI precision is due to the very small number of genes still predicted as AI at small model sizes).

This result suggested that the signal of allelic imbalance in our feature set is diffuse, making interpretation of the effect of individual features in the model difficult. In order to analyze these features, we used non-negative matrix factorization (NMF: Brunet *et al.* 2004, reviewed in Devarajan 2008), a method that is analogous to principle components analysis but typically produces much sparser factors. We extracted four factors that summarize the 500 top features in the model (Figure 10). Each factor represents a weighted linear combination of the individual features. Most of the features in our model contributed to several or all of the resulting factors, indicating that the four factors were not completely orthogonal to each other, and none of them could be readily interpreted as, for example, a "polymorphism" factor or a "repeat" factor. However, we found that only one of these factors, which we refer to as the "AI factor," significantly differentiates between the AI class and non-AI class of genes (Wilcoxon summed ranks test: $p = 3.87 \times 10^{-5}$). Specifically, a higher value of the AI factor corresponds to an increased probability that the associated gene will be subject to common imbalance.

4.3.3 Validation using an external dataset

Model predictions for genes that exhibited significant allelic imbalance in Cheung *et al.* (2008) were significantly different from the non-AI genes extracted from the Serre *et al.* (2008) dataset (one-tailed Wilcoxon summed ranks test, $p = 4.70 \times 10^{-6}$) but were not significantly different from the AI genes from the Serre *et al.* (2008) dataset ($p = 0.506$). In other words, commonly imbalanced genes identified through two different methods were indistinguishable through our model, but both of these gene sets were predicted as more likely to be imbalanced than a third set of genes known to exhibit no common imbalance (Figure 14).

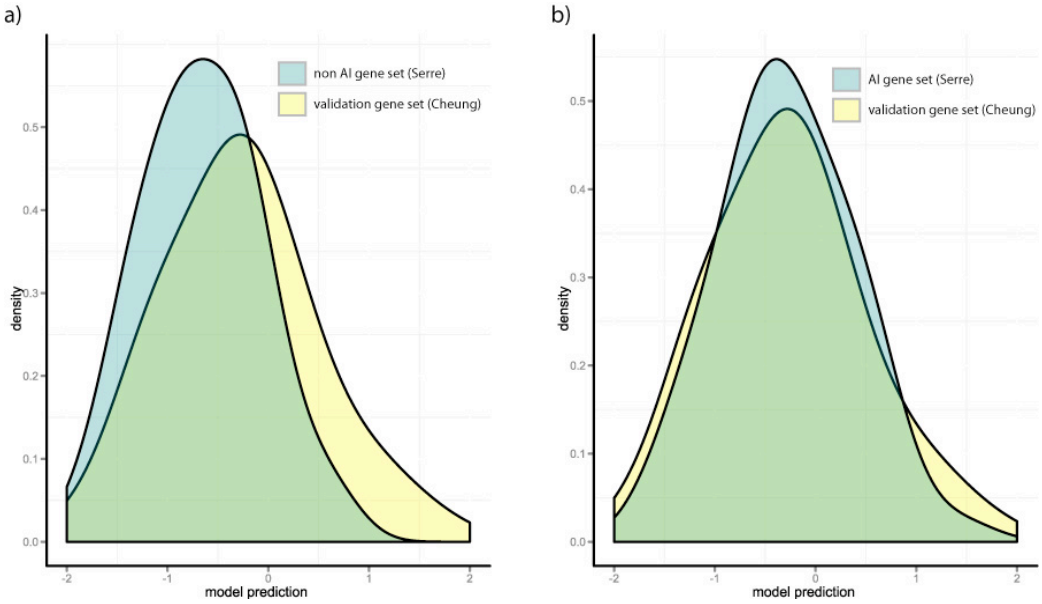

Figure 14: Raw predictions from the full model for genes that exhibit allelic imbalance in the Cheung et al. (2008) dataset are significantly different from predictions for the non-AI gene set (p = 4.70 x 10^{-6}) derived from Serre et al. (2008), but not significantly different from predictions for the AI gene set from Serre et al. (p = 0.506).

Additionally, genes from the external Cheung *et al.* dataset were significantly enriched for high values of the AI factor (which correspond to a higher likelihood of common allelic imbalance) compared to a background distribution of the AI factor derived from 3,908 genes of unknown status (p = 6.23 x 10^{-11}; Figure 15). This result suggests that the AI factor, and hence annotations of the AI factor, retains explanatory power for genes not included in the original dataset derived from Serre *et al.* (2008).

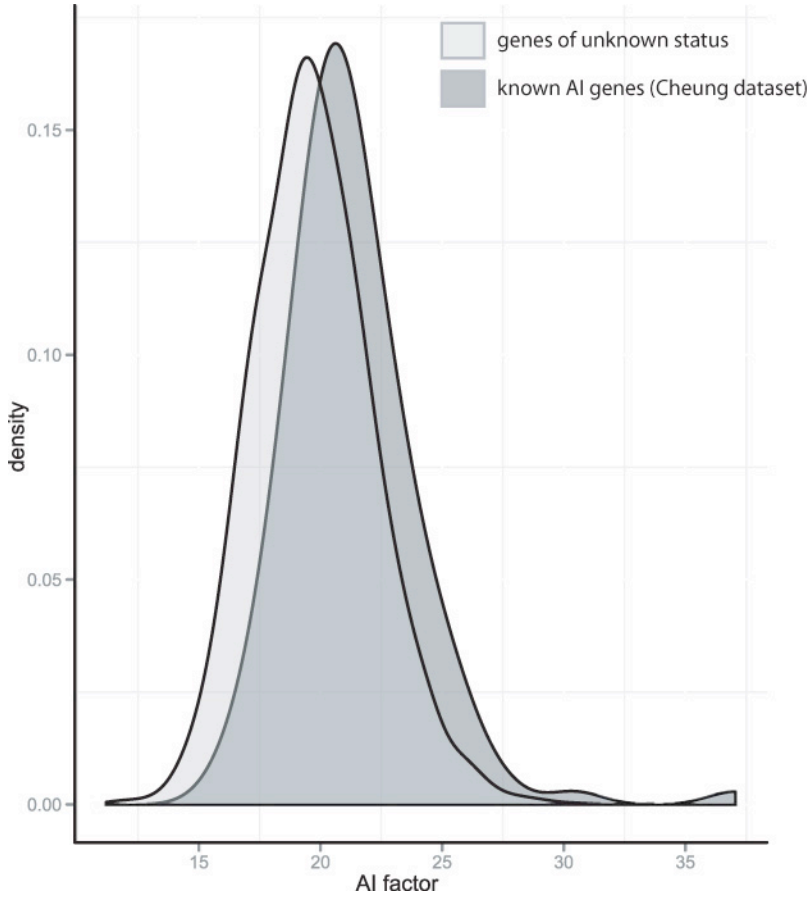

Figure 15: Smoothed distributions of genes that exhibit common allelic imbalance included in a second validation dataset (Cheung *et al.* 2008) and 3,908 genes from the AI factor annotation analyses chosen without respect to allelic imbalance. The genes known to be imbalanced are enriched for higher values of the AI factor (p = 6.23 x 10^{-11}).

4.3.4 Annotating the "AI factor"

Our results made it difficult to explain the predictive ability of our model through direct assessment of the features within the model: too many features were required for the model to perform well, and these features do not neatly reduce into orthogonal factors. In order to better understand why the feature set we identified contains information about allelic imbalance, we attempted to annotate the "AI factor" using external datasets. Specifically, we incorporated estimates of natural selection on gene regulatory regions from the work of Haygood *et al.* (2007); a metric of tissue

specificity in gene expression derived from the Novartis Gene Expression Atlas (Su *et al.* 2004), which we refer to as expression "evenness;" and a metric of gene density around the focal genes based on annotations from the Consensus CoDing Sequence Project (http://www.ncbi.nlm.nih.gov/projects/CCDS/). This approach allowed us to investigate the possible biological significance underlying the model using a much larger dataset, because the AI factor can be extracted for genes that lack allelic imbalance measurements in the original dataset. Our aim was to understand why the AI factor, which is derived entirely from sequence, polymorphism, and divergence data, had explanatory power with regards to allelic imbalance at all.

We hypothesized that the regulatory regions of genes that exhibit common allelic imbalance evolve under less selective constraint than the regulatory regions of genes that do not exhibit common imbalance, and that this relationship could be captured by analyzing sources of variance in the AI factor. If so, an increased estimate of negative selection on a gene's likely regulatory region might be correlated with a decrease in the value of the AI factor. As a corollary to this hypothesis, we did not expect to observe a relationship between the AI factor and estimates of positive selection, which were available for the same genes.

Given that statistical tests of natural selection have somewhat low power, we also attempted to model variation in the AI factor using two other variables that have been connected with gene regulation in the literature. We asked whether the number of neighboring genes in the region surrounding a focal gene or the degree of tissue specificity in the expression of a focal gene explain variation in the AI factor. Neighboring genes tend to exhibit more correlated patterns of expression than sets of randomly distributed genes (Kruglyak and Tang 2000; Lercher *et al.* 2002; Gierman *et al.* 2007). Thus, if *cis*-regulatory mutations potentially disrupt a neighborhood of genes instead of one or a few genes, genes in gene-dense regions may exhibit significantly lower AI factors due to

stronger negative selection in these regions of the genome. Similarly, when genes are broadly expressed, regulatory genetic changes may be subjected to increased evolutionary constraint due to deleterious effects introduced by pleiotropy. If so, genes that are broadly and evenly expressed in human tissues may also be associated with lower levels of the AI factor.

We modeled variation in the AI factor for 3,908 genes in the human genome for which estimates of negative and positive selection, tissue specificity (expression "evenness"), and positional information on nearby gene density were available. The overall model was highly significant and explained an appreciable amount of variation in the AI factor ($p < 2.2 \times 10^{-16}$; $R^2 = 0.178$ for the full model). Within the full model, we identified significant effects of the average strength of negative selection, the density of neighboring genes, and the evenness of gene expression across tissues, but not the strength of positive selection on the upstream region.

Specifically, genes subject to greater evolutionary constraint (i.e., a higher magnitude of negative selection in their putative regulatory regions) were also characterized by smaller AI factors, although this effect was very small ($p = 2.72 \times 10^{-9}$, $R^2 = 0.009$). Similarly, we also observed a very small, but significant effect of tissue specificity on the AI factor ($p = 9.08 \times 10^{-8}$, $R^2 = 0.007$): genes that are more evenly expressed across tissues exhibit on average smaller AI factors, corresponding to a lower likelihood of common allelic imbalance, than genes that are expressed much more strongly in one or a few tissues than in others. By contrast, the magnitude of positive selection did not explain a significant amount of variation in the value of the AI factor ($p = 0.062$).

We found that the density of neighboring genes had by far the strongest explanatory effect ($p << 1 \times 10^{-16}$; $R^2 = 0.159$), accounting for more than an order of magnitude more of the overall variance in the AI factor than estimated for the direct

effect of negative selection. Thus, as the number of neighbors within a 100 kb flanking region on either side of the gene (200 kb of total sequence) increased, the AI factor decreased (Figure 16). The interpretation of this result in the light of allelic imbalance is that genes in gene-rich regions of the genome are somewhat less likely to exhibit common imbalance than genes with fewer neighbors.

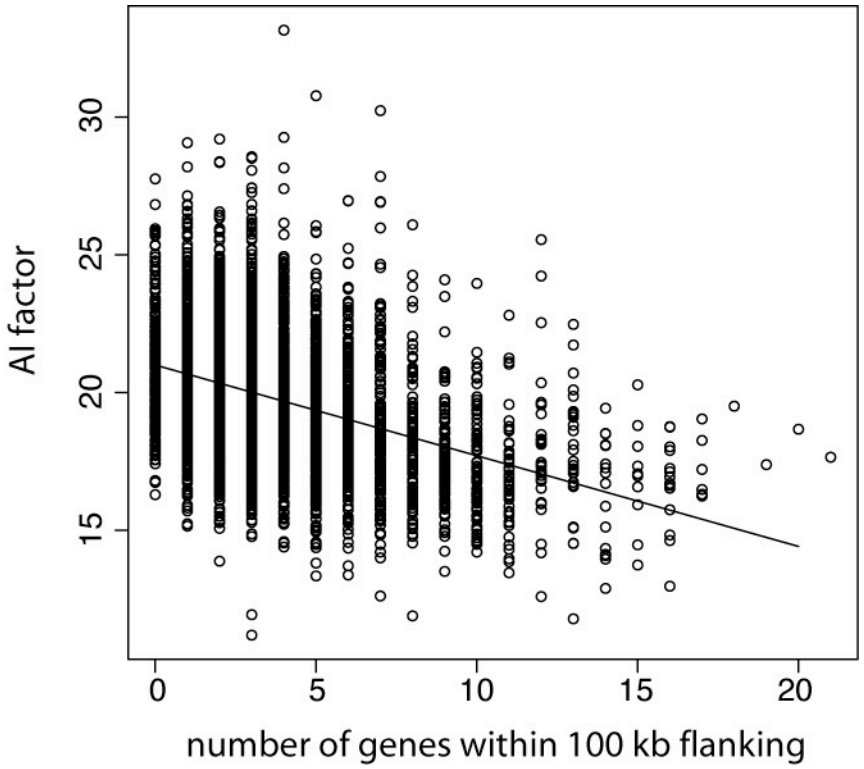

Figure 16: Genes that reside in more gene-dense neighborhoods exhibit lower values of the AI factor ($p << 1 \times 10^{-16}$; $R^2 = 0.159$). The line running through the graph shows the estimated slope for the number of genes within 100 kb flanking when this effect is estimated by itself (i.e., not within the full linear model).

4.4 Discussion

4.4.1 Prediction of common allelic imbalance

Our results indicate that the signature of allelic imbalance is detectable in the human genome, and that this signature can, at least diffusely, be captured using support vector machine (SVM) models of features extracted from sequence, polymorphism, and

divergence data. While the classification accuracy of our model exhibits a detectable improvement over random assignment to the AI or non-AI class (the set of genes that commonly exhibit allelic imbalance and the set of genes that do not exhibit allelic imbalance, respectively), the level of overall accuracy we were able to achieve is modest relative to that observed for other biological phenomena. For example, Wang *et al.* (2006) were able to differentiate between X-inactivated genes and genes that escape from X-inactivation in humans with over 80% accuracy (Wang *et al.* 2006), and Luedi *et al.* (2005) were able to distinguish between imprinted and non-imprinted genes in mice with about 94% accuracy (Luedi *et al.* 2005). Both of these studies used approaches similar to those we applied here, including some overlap in feature types (although neither included data on polymorphism or divergence). However, while SINE and LINE repeat elements were important features in both Luedi *et al.* (2005) and Wang *et al.* (2006), they were not strongly highlighted in our analysis: although repetitive elements did appear in the 500 features identified through recursive feature elimination, none of them were weighted very heavily in the AI factor (see Supplementary Materials for Tung *et al.*, 2009). These comparisons suggest that allelic imbalance is a more difficult phenotype to classify, at least using readily available genomic features.

The likely reason for this comparative difficulty is that allelic imbalance is a complex quantitative trait (e.g., Tao *et al.* 2006), although we dichotomized it for the purposes of this study. Gene expression has a multifactorial basis, including both genetic and environmental effects, and also can vary temporally and spatially across different tissues. Indeed, our results are comparable to those from the handful of studies that have attempted to analyze other complex traits in a predictive framework (Khoury *et al.* 2008; Lango *et al.* 2008; van Hoek *et al.* 2008; Jakobsdottir *et al.* 2009; Liu *et al.* 2009); but see Lee *et al.* 2008). For example, recent whole genome association studies have identified multiple susceptibility loci for type 2 diabetes, and the replicability and strong

statistical support for these loci have made type 2 diabetes one of the relative success stories of the genome-wide approach (Prokopenko *et al.* 2008). However, when assessed in a predictive context, these loci exhibit only modest predictive ability for the disease: area under the curve (AUC), a metric that summarizes the trade-off between true positive and false positive rates (random prediction is 0.50; perfect prediction is 1.0; values below 0.50 reflect prediction that is worse than random), was estimated at 0.60 in two different studies (Lango *et al.* 2008; van Hoek *et al.* 2008). By way of comparison, AUC for our dataset was 0.66, even though, unlike the diabetes studies, we did not have prior information about specific variants that were highly associated with the trait.

Additionally, our results suggest that there is heterogeneity within classes: for example, genes in the AI class include genes for which allelic imbalance is substantial as well as genes for which allelic imbalance is modest (but detectable and replicable in multiple individuals). In contrast, prediction for non-AI genes appears to be easier. These findings are also in agreement with other attempts to predict complex traits. For example, Liu *et al.* (2009) attempted to predict eye color using up to 24 SNPs previously implicated in eye color differences. While they were able to achieve prediction of brown eyes and blue eyes at AUC levels of 0.88 – 0.93, prediction of "intermediate" colored eyes ranged from 0.63 – 0.73, suggesting that this phenotypic class is more difficult to accurately predict than the other two classes. We were not able to detect a robust effect of magnitude of imbalance on classification accuracy (data not shown). However, magnitude of imbalance is difficult to take into account because all allelic imbalance datasets thus far focus on a relatively small set of individuals (n = 83 in Serre *et al.*, 2008), and, even within these datasets, appreciable variation is observed among individuals that exhibit imbalance, suggesting that imbalance magnitude may be context-dependent on *trans* genetic variation or environmental factors.

Finally, the genetic variation that produces allelic imbalance arises and is maintained by a complex combination of mutation, recombination, selection, and demographic history. For example, because allelic imbalance is only detectable within individuals that are heterozygous at a transcribed site, the allele frequency spectrum for the causal *cis*-regulatory polymorphism, along with population structure, is a critical component of the frequency with which allelic imbalance will be detected. Unlike for phenomena like X-inactivation or imprinting, then, no gene will always exhibit allelic imbalance, even if these polymorphisms are very common. Further, the rate at which allelic imbalance arises may vary due to differences in the underlying mutation rate, and the frequency with which it is expressed may differ across tissues (Campbell *et al.* 2008) and across environments (de Meaux *et al.* 2005; Zhu *et al.* 2006). Across populations, variation in the extent of *cis*-regulatory polymorphism between genes may be due to differences in the occurrence and strength of balancing selection (e.g. at MHC loci: Loisel *et al.* 2006; Tan *et al.* 2006) or, as suggested by our results, could reflect variation in selective constraint on gene expression profiles. Additional genome-wide measurements of allelic imbalance in more of these contexts would increase the accuracy of the labels we used here, and likely improve the classification ability of the resulting models.

For the preceding reasons, it is perhaps surprising that the features used here are predictive of common allelic imbalance at all, especially given that, unlike other predictive studies (Lango *et al.* 2008; van Hoek *et al.* 2008; Liu *et al.* 2009), we could not filter our feature set for features that were *a priori* known to be involved in producing allelic imbalance for these genes. Predictive models derived from machine learning have been frequently used in molecular and cancer genetics (Mukherjee *et al.* 1999; Brown *et al.* 2000; Guyon *et al.* 2002; Zhang *et al.* 2003), and have been applied to a handful of problems in ecology (Guo *et al.* 2005; Drake *et al.* 2006). To our knowledge, however, they have rarely been used to interrogate differences in the degree of variation in specific

molecular phenotypes, as we have done here. Our results suggest that this general approach may have some applicability to these kinds of problems, and may therefore be useful as an additional tool for investigating problems in biological fields specifically interested in variation, including genetic epidemiology and evolution (e.g., Roettger *et al.* 2009). Given that the features used in these models are becoming available for more and more systems, including non-traditional, non-model systems, they could be of particular use when informed prediction is an important step to take prior to conducting empirical measurements.

4.4.2 Selective constraints on gene expression

The initial SVM model fitting for allelic imbalance did not rely on careful hypothesis generation or modeling of the process by which imbalance arises. However, understanding the biological meaning behind its predictive ability demands that such methods be applied. We attempted to do so here by annotating a factor that contains many of the features responsible for our model's predictive ability, and that is itself significantly correlated with allelic imbalance class, using additional publicly available data and the results of prior work incorporating a formal modeling perspective.

These analyses allowed us to test hypotheses to account for the apparent nonrandom distribution of allelic imbalance around the genome. We reasoned that, if gene expression is frequently under negative selection in the primate lineage, as has been suggested by others (Khaitovich *et al.* 2005; Gilad *et al.* 2006; Gilad *et al.* 2006), genes that exhibit common imbalance may be those that are evolving under less evolutionary constraint than genes that do not exhibit common imbalance. This possibility has also been suggested by Campbell and colleagues (2008) to explain the observation that genes that are imbalanced in humans also tend to be imbalanced in mice, despite the substantial evolutionary time separating these two species (Campbell *et al.* 2008).

Alternatively, if natural selection has little to do with imbalance, then the distribution of commonly imbalanced genes around the genome may have more to do with variation in local mutation rates. Currently, genome-wide datasets that estimate the strength of selection and evolutionary constraint on gene regulatory regions are available at the resolution of single genes; in contrast, fine-scale estimates of mutation rate variation across the genome are not yet available. Hence, we focused largely on the currently more tractable hypothesis that variation in allelic imbalance across the genome is related to evolutionary constraint. Specifically, we examined the relationship between the "AI factor," a linear combination of variables that predicts allelic imbalance, and three other effects that are directly or indirectly related to evolutionary constraint. We found that the value of the AI factor increases (corresponding to a higher probability of common imbalance in the gene) with decreased negative selection on the upstream regulatory region of a gene, decreased evenness of expression across human tissues, and decreased density of genes in the region surrounding the focal gene.

By far, the strongest effect we identified was that of density of genes around the focal gene: genes in gene-dense regions are associated with lower values of the AI factor, corresponding to a lower likelihood of common allelic imbalance. At least two mechanisms can account for this observation. First, the presence of nearby genes evolving under negative selection could reduce the proportion of nearby sites that are likely to harbor common segregating genetic variation. Under this scenario, negative selection on neighboring genes (even if only in the coding regions) means that fewer variants with potential *cis*-regulatory effects on the focal gene will reach frequencies high enough to produce common allelic imbalance. Second, functional *cis*-regulatory variants that arise in gene-dense regions could be more likely to produce deleterious pleiotropic effects on gene expression. Genes that cluster together in the same physical location tend to exhibit correlated patterns of gene expression (Kruglyak and Tang 2000; Lercher *et al.*

2002; Gierman *et al.* 2007). If these effects are due to shared *cis*-regulatory sequence or to shared patterns of chromatin condensation, changes in the expression of one gene may ramify outwards to also affect neighboring loci. Consequently, functional *cis*-regulatory variation that arises in gene-dense regions might alter the expression of not one, but several (or many) linked genes, and therefore be subject to greater constraint than *cis*-regulatory variants near physically isolated genes. Although both of these mechanisms invoke patterns of evolutionary constraint, only the second requires negative selection on the gene expression profile itself. Given that they are not mutually exclusive, however, it is possible that the combination of both mechanisms acting together accounts for the strong signal of gene density on the AI factor.

Pleiotropy may also influence the observed relationship between allelic imbalance and evenness of expression. Genes that are more evenly expressed across tissues in the human body have, on average, lower values of the AI factor, although this effect is very small. One of the main arguments in favor of the importance of *cis*-regulatory variants in complex trait evolution is that changes in *cis*-acting gene regulation can evade pleiotropic constraints by altering gene expression in a tissue- or condition-dependent manner (Wray 2007; Blekhman *et al.* 2008; Smith and Kruglyak 2008). Recent evidence strongly suggests that tissue-specific changes in expression have been important during human evolution (Blekhman *et al.* 2008; Kosiol *et al.* 2008); for example, a selectively advantageous change in the *DARC cis*-regulatory region abolishes expression of the gene on red blood cells, conferring strong protection against infection by malarial parasites, but does not interfere with *DARC* expression elsewhere in the body (Tournamille *et al.* 1995). However, tissue-specificity may be more difficult to achieve when a gene is truly evenly expressed across many tissues. Hence, mutations influencing these genes may be subject to a slightly increased level of constraint, in this case due to pleiotropy across tissues as opposed to pleiotropy across genes.

As in the case of evenness of expression, the relationship between negative selection and the AI factor was weak but in the direction predicted by our hypothesis. Genes subject to greater negative selection, as measured by the comparison between the rate of evolution in the region upstream of the gene and the rate of evolution in downstream introns (Haygood *et al.* 2007), tend to have lower values for the AI factor, suggesting that these genes are less likely to exhibit common allelic imbalance. Perhaps surprisingly, if gene density is a proxy for evolutionary constraint, the relationship between imbalance and this direct measure of negative selection was much weaker than the relationship between allelic imbalance and gene density. This discrepancy may be due to the limited scope of the measure of negative selection (functional *cis*-regulatory elements can further upstream, downstream, or within a gene, so we therefore averaged over the three regions for which data were available; however, it is possible that these regions accumulate functional *cis*-regulatory differences at different rates, and with different downstream effects), the inherent lack of power in estimating the strength of selection, and/or differences between patterns of selection on the gene expression phenotype itself and patterns of selection on the associated *cis*-regulatory sequence.

Overall, our results suggest that evolutionary constraint plays an important role in determining whether a gene is likely to accumulate functional *cis*-regulatory variation at moderate to high frequencies within human populations. The role of mutational biases in this process remains an important outstanding question, however. Our results do not preclude the possibility that genes that are more likely to exhibit common imbalance might also fall, with some greater probability, in mutational "warmspots." What then are the relative contributions of mutation and selection to allelic imbalance within human populations? Measures of GC content (except in the coding sequence itself) were not included in the set of 500 features that were most predictive of common allelic imbalance in our analyses. Given that the mutation rate at CpG dinucleotides is

estimated to be over an order of magnitude higher than background (Nachman and Crowell 2000), this result circumstantially suggests that mutational bias might not play as important of a role as selective constraint in determining the distribution of allelic imbalance. In the next several years, we anticipate that next-generation sequencing technologies will produce much more fine-scaled estimates of mutation rate across the genome than are currently available. At that point, it will be worth revisiting the relative role of selection and mutation in determining segregating functional *cis*-regulatory variation in human populations.

Taken together, our analyses support the hypothesis that the nonrandom distribution of common allelic imbalance in the human genome, as demonstrated by the ability to classify and predict which genes are subject to common imbalance, is the product of weak negative selection. Specifically, commonly imbalanced genes tend to be subjected to less evolutionary constraint than genes that are never (or rarely) imbalanced. We were able to detect this effect only by analyzing a large number of genes, most of which were not actually included in the initial dataset on allelic imbalance. This result suggests that the machine learning-based approach we applied here might be useful not only for exploratory analyses, but also for producing a proxy for a phenotype of interest (here, the AI factor) that can be used to expand the size of the dataset to be analyzed. It also suggests that negative selection on gene expression, as has been documented in both primates (Gilad *et al.* 2006; Gilad *et al.* 2006) and model systems (Rifkin *et al.* 2003; Denver *et al.* 2005), may translate into negative selection on functional *cis*-regulatory variants. As in the case of other molecular characteristics with evolutionary implications, such as codon usage bias (Akashi 1995; dos Reis and Wernisch 2009) or mutation to spurious transcription factor binding sites (Hahn *et al.* 2003), the effect of negative selection on allelic imbalance appears to be weak.

Further work needs to be done in order to understand whether the results we describe are typical of functional genetic changes in gene expression in general or are specific to *cis*-regulatory genetic effects, and whether the predictive models developed here extend to other taxa. Additionally, the greater difficulty we encountered in classifying genes in the AI class than genes in the non-AI class suggests that the category of genes subject to common allelic imbalance is somewhat heterogeneous. Further exploration may reveal possible sources of this heterogeneity. It would be interesting if genes that exhibit imbalance in a context-dependent manner (those sensitive to developmental timing or tissue-dependent effects, or those influenced by epistasis and/or gene-environment interactions) behave quantitatively or qualitatively differently from genes for which the architecture of allelic imbalance is more simple. Functional regulatory effects make important contributions to organism level phenotypic variation of both medical and evolutionary import. Understanding how these effects are distributed across the genome, and in particular when and in what genes they may persist, is therefore critical to developing a better understanding of how trait variation arises within populations.

5. Evolution of a malaria resistance gene in wild primates[1]

5.1 Introduction

The ecology, behaviour and genetics of our closest living relatives, the nonhuman primates, should help us to understand the evolution of our own lineage. Although a large amount of data has been amassed on primate ecology and behaviour, much less is known about the functional and evolutionary genetic aspects of primate biology, especially in wild primates. As a result, even in well-studied populations in which nongenetic factors that influence adaptively important characteristics have been identified, we have almost no understanding of the underlying genetic basis for such traits. Here, we report on the functional consequences of genetic variation at the malaria-related *FY* (*DARC*) gene in a well-studied population of yellow baboons (*Papio cynocephalus*) living in Amboseli National Park in Kenya. *FY* codes for a chemokine receptor normally expressed on the erythrocyte surface that is the known entry point for the malarial parasite *Plasmodium vivax* (Miller *et al.* 1975; Miller *et al.* 1976; Barnwell *et al.* 1989). We identified variation in the *cis*-regulatory region of the baboon *FY* gene that was associated with phenotypic variation in susceptibility to *Hepatocystis*, a malaria-like pathogen that is common in baboons (Myers and Kuntz 1965; Garnham 1966). Genetic variation in this region also influenced gene expression *in vivo* in wild individuals, a result we confirmed using *in vitro* reporter gene assays. The patterns of genetic variation in and around this locus were also suggestive of non-neutral evolution, raising the possibility that the evolution of the *FY cis*-regulatory region in baboons has exhibited both mechanistic and selective parallelisms with the homologous region in humans

[1] The contents of this chapter have been previously published as: J Tung, A Primus, AJ Bouley, TF Severson, SC Alberts, and GA Wray (2009). *Evolution of a malaria resistance gene in wild primates*. Nature 460: 388 – 392.

(Hamblin and Di Rienzo 2000; Hamblin *et al.* 2002; Sabeti *et al.* 2006). Together, our results represent the first reported association and functional characterization linking genetic variation and a complex trait in a natural population of nonhuman primates.

5.2 Background

In humans, a transition from the wild-type T variant to a C variant at a single polymorphic site in the *FY cis*-regulatory region causally abolishes all expression of this gene in erythrocytic precursors. As a result, C homozygotes at this site are strongly protected from infection by *P. vivax* (Tournamille *et al.* 1995), and a lower level of protection is also conferred on C/T heterozygotes (Zimmerman *et al.* 1999; Michon *et al.* 2001). The C variant has apparently arisen independently at least twice in geographically distinct human populations (in Africa and in Papua New Guinea: Miller *et al.* 1976; Zimmerman *et al.* 1999), and has been driven to high frequencies on at least two haplotypic backgrounds within Africa (Hamblin and Di Rienzo 2000). Additionally, the pattern of variation in the *cis*-regulatory region as a whole strongly indicates a historical pattern of natural selection in different populations around the world, probably as the product of directional selection in some populations (for example, local positive selection), and a complex mix of selection and demographic history in others (Hamblin and Di Rienzo 2000; Hamblin *et al.* 2002; Sabeti *et al.* 2006). The unusual evolutionary history of this locus led us to investigate the pattern of genetic variation in its baboon homologue, and to explore the possibility that it might also explain phenotypic variation in parasite infection in a wild primate population, the well-studied baboon population of the Amboseli basin in East Africa (Buchan *et al.* 2003; Alberts *et al.* 2006; Tung *et al.* 2008).

Baboons are not generally infected by *Plasmodium* in the wild, but are vulnerable to infection by several closely related haematoprotozoans (Myers and Kuntz 1965;

Garnham 1966) including *Hepatocystis kochi*, a blood parasite nested within the paraphyletic *Plasmodium* genus (Perkins and Schall 2002). *Hepatocystis* parasites do not produce the cyclical fever spikes typical of malaria in humans, but do produce anaemia and visible merocyst formation, followed by scarring on the liver (Garnham 1966). The similarities between *P. vivax* and *Hepatocysis* therefore prompted us to investigate whether regulatory genetic variation linked to the *FY* gene influences incidence of *Hepatocystis* infection in Amboseli, expression of the gene, or both, in a manner parallel to that observed in humans.

5.3 Materials and methods

5.3.1 DNA and RNA sampling

Blood samples for DNA extraction and *Hepatocystis* screening were collected from 190 Amboseli baboons between 1989 and 2008 (Altmann *et al.* 1996). DNA was extracted using standard methods. DNA for some individuals was whole genome amplified (Qiagen Repli-G Kit). Blood samples for RNA extraction were collected in PaxGene RNA tubes from 101 adult Amboseli baboons darted between 2004 and 2008. RNA was extracted using the PaxGene RNA Blood kit (Qiagen), and reverse transcribed into complementary DNA (ABI High Capacity cDNA Archive Kit).

5.3.2 Sequencing

We amplified and sequenced the region homologous to the annotated *FY cis*-regulatory region in humans in 174 individuals and sequenced or genotyped the two pyrosequencing assay SNPs in 150 individuals. To assess congruence between the *in vitro* and *in vivo* gene expression results for the C/T SNP, we inferred haplotype phasing using PHASE 2.1.1 (Stephens *et al.* 2001).

5.3.3 Hepatocystis screen and association with *FY*

We screened for *Hepatocystis* in all 190 baboons using *Hepatocystis* mtDNA specific primers, which produced a band of approximately 251 base pairs on an agarose gel in the presence of *Hepatocystis*. High rates of infection were found in groups sampled in 2004-8 as well as in groups sampled in the late 1980's and early 1990's, indicating that our ability to detect *Hepatocystis* infection did not markedly decrease with the age of the sample. However, to rule out the possibility that a failure to amplify *Hepatocystis* was due to poor quality DNA, we eliminated from subsequent analyses any individuals for whom we were not able to generate high quality genomic DNA sequence from other regions, including the *FY cis*-regulatory region.

We also used *Plasmodium*-genus specific primers (Rougemont *et al.* 2004) as a secondary confirmation of infection for 103 individuals (*Hepatocystis* is phylogenetically nested within the *Plasmodium* species that infect primates: (Perkins and Schall 2002). All individuals included in this study had concordant results with both the *Hepatocystis* mtDNA primers and the *Plasmodium* genus-specific primers.

We then fitted the following generalized linear mixed model for 150 individuals, using a binomial error structure:

$$P(y_{ij} = 1 \mid G_{ij}) = \text{logit}\left(\beta G_{ij} + \sum_{u=1}^{5} D_{iu} v_u + S_j + b + \varepsilon \right)$$

where y is *Hepatocystis* infection status ($y = 1$ corresponds to infected; $y = 0$ corresponds to uninfected), individuals are indexed by i, and study group is indexed by j. β is a fixed effect of genotype, G_{ij}; v_u is the fixed effect of the projection D_{iu} on the uth principal component of population structure; S_j is a random effect of study group; b is the intercept; and ε represents model error. We used two approaches to control for possible population structure in our sample. First, we included social group as a random

effect when modeling *Hepatocystis* infection on genotype (infection rate was clearly structured by social group: see Figure 17). In baboons, sex-biased dispersal and philopatry predict that social group ('breeding group') will be the most important unit of population structure. We have also used this approach to take account of social group in previous studies of the Amboseli baboon population (e.g., Charpentier *et al.* 2008). Second, we applied a principle components-based analysis to identify the major axes of population structure using genotype data from 47 unlinked loci (33 SNPs and 14 microsatellites) from around the baboon genome. Estimates of population structure were obtained following the method of (Price *et al.* 2006), using custom MATLAB code. Missing genotype data (9% of the overall data matrix) were imputed using two different methods: local least squares regression (Kim *et al.* 2006) and k nearest neighbors, with k = 3 (Troyanskaya *et al.* 2001); exploratory analyses with k = 1 – 7 produced very similar results. Results based on the k nearest neighbors approach are reported in the main text, but the results were qualitatively identical regardless of imputation method. In our final model, we incorporated projections from the first five eigenvectors obtained through PCA (these explained approximately 60% of the overall genetic variation in population).

The final analysis was run using the *lmer* function in the R package *lme4*, version 0.99875-9 (Team 2007). We evaluated the significance of β_g, the SNP effect, as evidence for association between infection and *cis*-regulatory variation.

5.3.4 Pyrosequencing

We designed pyrosequencing assays based on two variable SNPs in the transcribed region of the *FY* gene. 38 individuals for which cDNA samples were available were also heterozygous at one or both of these sites. For each of these individuals, we performed six to eight pyrosequencing reactions across two plates (mean number of measurements per individual = 7.05, range = 3–8, excluding failed reactions).

The resulting values were expressed as the $\log_2$ transformed ratio of the expression of transcripts carrying one versus the other allele at one of the assay SNPs (values based on the alternative assay SNP were converted on the basis of linkage between the two sites).

We identified an effect of one of the upstream *cis*-regulatory sites by modelling variation in allelic imbalance using the following general linear mixed model:

$$y_{ij} = \beta G_{ij} + Y_j + b + \varepsilon$$

where y is allelic imbalance, indexed by individual i and year of sampling j; β_g is a fixed effect of homozygous or heterozygous genotype, G_{ij}; Y_j is a random effect of year of sampling (2006, 2007 or 2008; one individual was sampled in 2005 and grouped with the 2006 samples); b is the intercept; and ε is the model error. We assessed the significance of β using a permutation test: all measurements for an individual were grouped as a block and permuted 1,000 times over individual identity. We assigned a P value to the original SNP parameter estimate by ranking it among the corresponding estimates for the permuted data sets. The estimate of variance explained by the C/T site is based on modelling the residuals of allelic imbalance on year of sampling, using the C/T site alone.

Three pairs of individuals in the allelic imbalance analysis were related at $r = 0.5$; removing any set of three individuals so that no individuals were closely related did not qualitatively change our results. Simulations based on data from both related and unrelated individuals showed that dyads related at $r < 0.5$ were only about 10 – 14% more likely than random dyads to share genotypes, and unrelated dyads were about 1% less likely than random dyads to share genotypes (reflecting the outbred nature of the population and the accuracy of the pedigree data).

5.3.5 Transfection assays

Human erythroleukaemic cells (HEL 92.1.7) were maintained using the ATCC protocol. The wells of 24-well cell culture plates were seeded with 2×10^5 cells in 500 μl media, transfected, incubated for 48 h, and lysed. Cells were co-transfected with experimental constructs or empty firefly luciferase vector as control (pGL4.10; 180 ng per well) and the CMV Renilla normalization construct (pGL4.75, 20 ng per well; Promega) using Fugene 6 (Roche). Expression levels were measured with a dual-luciferase reporter assay (DLR1000 assay kit, Promega) and reported as relative ratios of luminescence (firefly:Renilla). Eight replicate wells were transfected for each experimental and control vector within an assay, with the assay repeated three times ($n = 24$ total measurements per construct).

We compared the measurements for each pair of constructs (A versus G for the *Hepatocystis*-associated SNP and C versus T for the SNP associated with allelic imbalance) separately using the following model:

$$y_{ijk} = \beta C_j + E_k + b + \varepsilon$$

where y is the relative ratio of luminescence for replicate i of construct j in experiment k; β is a fixed effect of construct, C_j; E_k is a random effect of experiment; b is the intercept; and ε is model error.

5.3.5 Signature of selection

5.3.5.1 F_{st}-based comparisons

For comparison of the F_{st} values, we genotyped up to 36 polymorphic microsatellite loci in ten baboons from the Masai Mara Reserve, Kenya; 20 baboons from Mikumi National Park, Tanzania; and 12 baboons from Amboseli National Park. One locus is a polymorphic microsatellite located ~3.7 kb upstream of the sequenced *FY cis*-regulatory region (Seixas *et al.* 2002); the other 35 loci reside in putatively neutral sites

dispersed around the baboon genome (Buchan *et al.* 2003; Alberts *et al.* 2006). We calculated F_{st} values for each of these loci independently using Arlequin 3.1 (Excoffier *et al.* 2005), and compared the F_{st} value for the *FY*-linked microsatellite locus with the F_{st} values for the 35 neutral loci. Specifically, we modeled the distribution of F_{st} values for the 35 neutral markers as a gamma distribution, and asked about the likelihood of observing the value of F_{st} for the *FY*-linked marker or a more extreme value, given this model. However, because our inference was based on a modest number of markers, we formally tested the stability of our *p*-value estimate given uncertainty in the model (i.e., under other parameterizations of the gamma, including some parameter settings that might also be highly consistent with the data, but could potentially provide weaker support for the hypothesis of non-neutral evolution).

We therefore calculated the *p*-values for 10,000 other possible combinations of parameters for the gamma distribution and weighted these values by the likelihood of these parameter combinations, given the data. We sampled the values of the two parameters independently, in each case from a uniform distribution bounded by two standard deviations above or below the maximum likelihood estimate, where the standard deviations were based on the estimated marginal distribution for that parameter. This approach is equivalent to sampling from the posterior of a probability distribution, an approach that gives both the expectation of the true *p*-value (the mean of the posterior probability distribution), and the variance of the *p*-value across all the alternative parameterizations. We used random subsamplings of the data from n = 10 to n = 35 to examine how the variance decreases with increasing n, and averaged the mean and Var(*p*) over multiple random subsamples of the same size.

Analyses were conducted using custom scripts in Ruby and R, and a modification of freely available code for estimating maximum likelihood parameters for gamma distributions (Wessa 2008).

5.3.5.2 Tajima's D comparisons

For comparison of Tajima's *D*, we resequenced 21 other regions in and around genes in the same or a subset of the individuals resequenced at the *FY cis*-regulatory region (Table 7). We calculated Tajima's *D* for all loci using the program DnaSP version 4.9 (Rozas *et al.* 2003), assuming no recombination.

Table 7: Loci used for comparisons of Tajima's D. Short segments were resequenced within the transcribed region of each locus or 5' of the transcription start site in the putative *cis*-regulatory region for each locus given above. Haplotype was inferred using the program PHASE v. 2.1.1. In all cases, the individuals that were resequenced were included in the main set of individuals sequenced at the *FY cis*-regulatory region. N gives the number of alleles (2 per individual) in the final sample, S is the number of segregating sites identified in each region, pi gives the mean pairwise distance between alleles, theta is Watterson's estimate of theta for nucleotide diversity, and D is the estimated value of Tajima's D for each locus.

Locus	N	ungapped length	S	pi	theta/site	D
FY	344	636	6	0.00245	0.00147	1.260
CCL5	318	648	4	0.00036	0.00097	-1.046
CCR5	320	633	8	0.00130	0.00199	-0.722
CD58	304	161	2	0.00088	0.00197	-0.713
CD59	314	623	7	0.00217	0.00178	0.450
CXCR4	202	583	4	0.00070	0.00117	-0.709
CYP1A1	330	580	6	0.00166	0.00162	0.042
CYP1B1	302	682	15	0.00131	0.00350	-1.527
ESR1	242	415	1	0.00032	0.00040	-0.201
IFNGR1	332	488	5	0.00115	0.00161	-0.512
IL1A	342	514	6	0.00123	0.00182	-0.614
IL4R	332	461	5	0.00257	0.00170	0.924
IL6	286	602	6	0.00070	0.00160	-1.093
IL10	162	716	3	0.00015	0.00074	-1.312
IL12B1	322	589	6	0.00204	0.00160	0.515
IL19	338	272	5	0.00259	0.00287	-0.177
LTA	242	550	3	0.00070	0.00090	-0.339
MEFV	326	524	11	0.00381	0.00330	0.350
MPO	330	376	4	0.00006	0.00167	-1.600
MSR1	336	281	8	0.00901	0.00445	2.122
PHF11	332	318	4	0.00281	0.00197	0.711
TAP2	204	584	14	0.00478	0.00407	0.437

5.4 Results

5.4.1 *Hepatocystis* prevalence and association with the *FY cis*-regulatory region

We tested for the presence of *Hepatocystis* parasites by screening DNA samples extracted from baboon blood for 190 individuals in the Amboseli baboon population. We found a high incidence of *Hepatocystis* infection in the Amboseli population (61.9%), although rates of infection varied substantially between different social groups and over time (Figure 17), possibly because of differences in home range and hence exposure to the vector, a biting midge (Garnham *et al.* 1961).

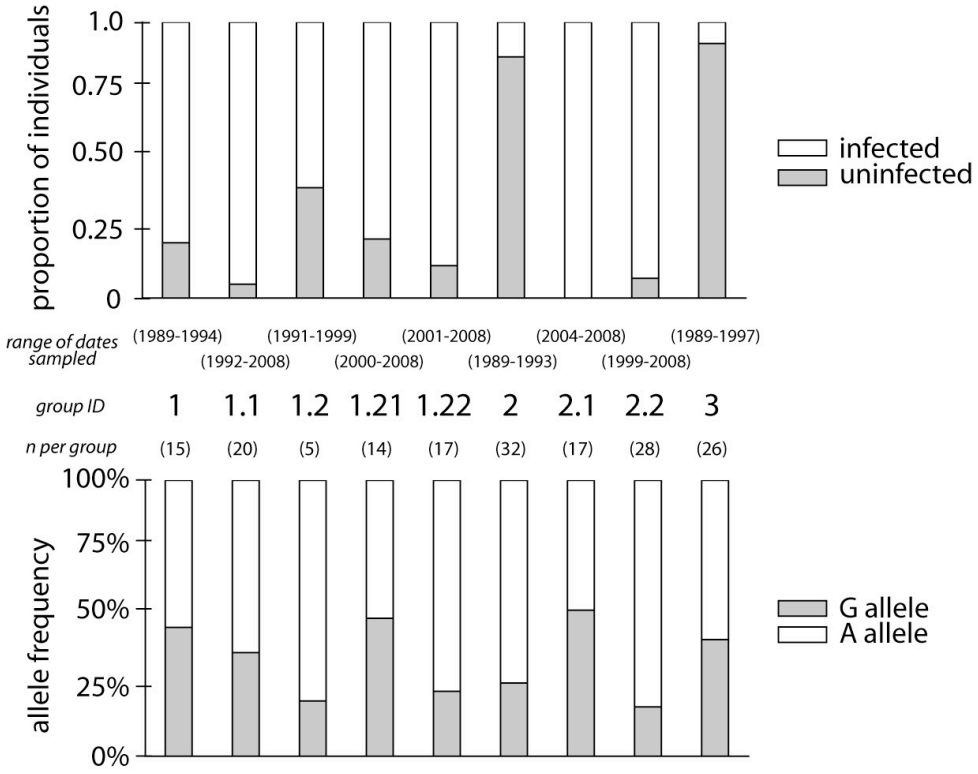

Figure 17: Differences by study group. The top bar graph shows differences in the proportion of individuals infected in each of the 9 study groups included in this study; the bottom bar graph shows the allele frequencies for the *Hepatocystis* associated *FY cis*-regulatory SNP. The x-axis label (study group) is given in the space between the two graphs, and is the same for both of them. Also given in parentheses are the range of years from which samples were obtained for each group (above the group ID), and the number of individuals sampled in each group (below the group ID).

In 174 of 190 baboons that we screened for *Hepatocystis*, we also sequenced the region of baboon DNA homologous to the annotated human *FY cis*-regulatory region. We identified six single nucleotide polymorphisms (SNPs) in the baboon *FY cis*-regulatory region (Figure 18; the malaria-associated SNP documented in humans was invariant in the baboons). *Hepatocystis* infection was significantly associated with an A/G variable site in the *FY cis*-regulatory region, in a model that took social group (a significant source of variance in infection) and genetic background into account. The risk of infection decreased as the number of G alleles an individual carried increased ($p < 0.012$, n = 174; Figure 19).

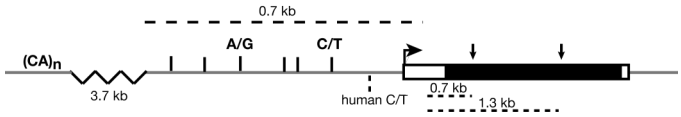

Figure 18: Schematic of the baboon *FY* gene (not to scale). Boxed regions: regions of the gene present in mature mRNA (open boxes: untranslated regions; black boxes: protein coding sequence). Gray lines: untranscribed regions. Bent arrow: start of transcription. Downward arrows: two baboon SNPs used as markers for the pyrosequencing assays. Vertical black bars: *cis*-regulatory SNPs in baboons, with the *Hepatocystis*-associated SNP labeled as "A/G" and the allelic imbalance-associated SNP labeled as "C/T." Dashed vertical bar: location of the functional SNP known in humans. Dashed horizontal lines provide relative distances and/or sizes of features. Location of the $(CA)_n$ microsatellite used in the F_{st} analyses is also shown upstream.

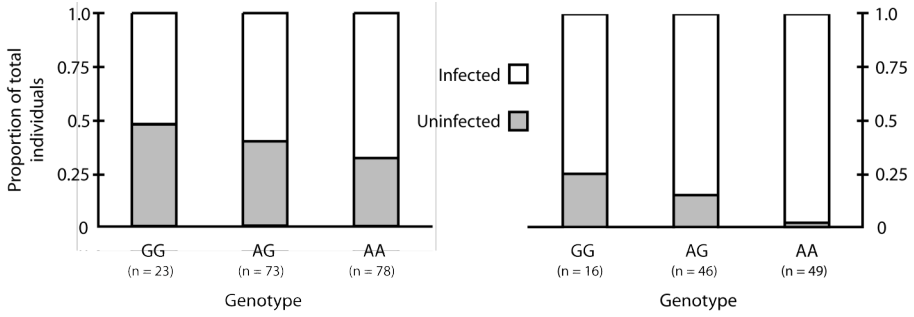

Figure 19: Genotype at the *FY cis*-regulatory A/G SNP is associated with *Hepatocystis* infection. The proportion of uninfected individuals is shown in grey, and the proportion of infected individuals is shown in white. Left side shows results for the entire sample set (n = 174; $p < 0.012$); right side shows results only for members of the six groups with high prevalence (> 75%) of *Hepatocystis* infection (n = 111; $p < 0.004$). Numbers below each genotype show the number of individuals for the given genotype.

5.4.2 Genetic variation in the *FY cis*-regulatory region influences gene expression

We also investigated whether *FY cis*-regulatory variation in baboons causally influences gene expression, as the C/T variant does in humans. We collected 101 samples of RNA-preserved blood from adults in six baboon social groups between 2004 and 2008, and used these samples to measure allele-specific expression at the *FY* locus using pyrosequencing. Specifically, we investigated whether the level of *FY* expression driven by one *cis*-regulatory *FY* allele differed from the level of *FY* expression driven by the other *cis*-regulatory *FY* allele, within the same individual. Because allele-specific expression compares the relative amounts of gene expression within individuals, it controls for effects on gene expression operating in *trans*, such as those produced by genetic background or by environmental main effects (Yan *et al.* 2002; Wittkopp *et al.* 2004). If the two alleles within an individual drive expression differently (allelic imbalance), that individual is likely to harbour a functional *cis*-regulatory variant that influences gene expression.

We measured allele-specific expression in 38 individuals (all the individuals among the 101 RNA-sampled baboons that were heterozygous at a transcribed pyrosequencing assay SNP: see Figure 18). Average $\log_2$ fold-change differences in expression between alleles within heterozygous individuals ranged from –0.002 (no difference between alleles) to 2.13 (substantial difference between alleles). This suggested that one or more common functional *cis*-regulatory variants influenced expression of the baboon *FY* gene in the Amboseli population. We predicted that if a *cis*-regulatory variant contributes to variation in gene expression, then individuals heterozygous at the *cis*-regulatory site would show significantly higher levels of allelic imbalance than individuals homozygous for the same variant.

Genotypes at four of the six SNPs in the baboon *FY cis*-regulatory region were sufficiently variable to test for an association with allelic imbalance. Genotype at the SNP closest to the start of transcription, a C/T transition, was significantly associated with allelic imbalance in the predicted direction: heterozygotes exhibited higher levels of allelic imbalance than homozygotes ($P < 0.002$, $n = 38$; Figure 20). However, this site explained only 22.0% of the overall variance in the allelic imbalance samples, after taking into account year of sampling. These results suggested that this C/T SNP functionally influences gene expression of the *FY* gene within the baboon population, but that additional *cis*-regulatory variants probably also play a part.

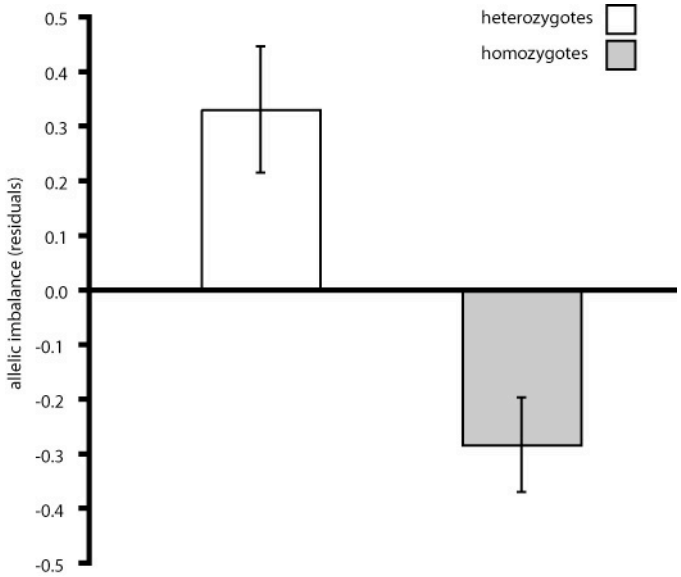

Figure 20: Allelic imbalance associates with *FY cis*-regulatory genotype. Height of the bars shows the mean for each genotypic class (heterozygotes: 0.329 ± 0.116 SEM; homozygotes: -0.285 ± 0.086 SEM). The y-axis gives the residuals of $\log_2$-transformed allelic imbalance on year of sampling (n = 38). Heterozygotes at the C/T SNP exhibit high values of allelic imbalance relative to homozygotes at the C/T SNP after controlling for the effect of year of sampling. High values indicate that the two alleles of the *FY* gene are transcribed at different levels within individuals, and suggest a functional *cis*-regulatory role for this SNP.

We next investigated the A/G SNP in the *FY cis*-regulatory region that we had associated with *Hepatocystis* infection risk. Of the 38 individuals for whom we measured allele-specific expression, 37 were heterozygous for this A/G SNP. Hence, we could not

compare allelic imbalance levels between heterozygotes and homozygotes at this site. We therefore tested this variant for possible functional effects using an *in vitro* approach in cell culture. We also used this framework to test further whether the C/T variant that was associated with allelic imbalance causally influenced *FY* expression. For each of these two regulatory SNPs (A/G and C/T), we built two plasmid constructs consisting of the *FY cis*-regulatory region linked to the firefly *luciferase* gene, such that the two constructs differed only at the variable site. We then tested the ability of these constructs to drive gene expression in a human erythroleukaemic (HEL) cell line. For the C/T SNP, the T allele construct drove significantly higher levels of expression than the alternative C allele construct ($P < 0.0001$; Figure 21a). Similarly, for the A/G SNP, the G allele construct drove significantly higher levels of expression than the alternative A allele construct ($P < 0.0001$; Figure 21b). These results suggest that both SNPs have the capacity to drive differential expression of the *FY* gene. In the case of the C/T SNP, for which both *in vivo* and *in vitro* analyses were possible, the T allele was associated with higher levels of expression in both experiments. Unlike the human case, in which one regulatory SNP results in null expression, all the baboon haplotypes we tested drove robust expression of the gene.

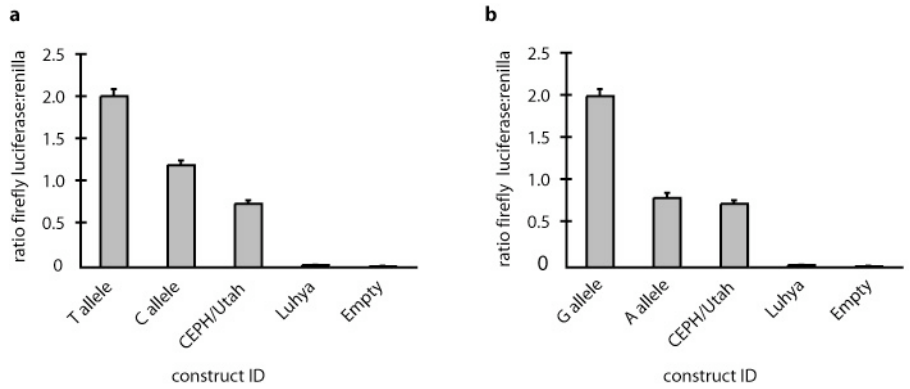

Figure 21: *FY cis*-regulatory variation drives differential expression *in vitro*. (a) The C/T SNP identified through the allelic imbalance measurements and (b) the A/G SNP that associates with *Hepatocystis* infection (right) drive differential gene expression compared to the alternative allele of the same SNP in cell culture. Values on the y-axis give the relative ratio of firefly luciferase luminescence to a control renilla luciferase reporter. Human constructs from a normal expressing individual (haplotype from the CEPH/Utah HapMap panel) and a null expressing individual (Luhya) and an empty vector are shown for comparison. Error bars show SEM for each construct.

5.4.3 The *FY cis*-regulatory region may have been a target of natural selection

Baboons, like humans, may also exhibit evidence of non-neutral evolution at the *FY cis*-regulatory region. We detected an increased level of population differentiation among East African baboon populations around *FY*, by comparing a *FY*-linked microsatellite with 35 neutral microsatellites ($F_{st} = 0.31$, $P < 0.029$; range of F_{st}, a metric describing genetic divergence between populations based on allele frequency differences at variable sites, for the neutral markers was 0.008–0.346, Figure 22). We also detected a higher value for the Tajima's D statistic ($D = 1.26$) in this region relative to nine of nine other resequenced putative *cis*-regulatory regions in the Amboseli population and 11 of 12 resequenced transcribed regions (range of D for all other loci was –1.60 to 2.12). The only locus with a higher value of D, a transcribed portion of the gene *MSR1*, exhibited an even more extreme value than that identified for the *MHC DQA1* promoter in baboons

(Loisel 2007), which is known to evolve under strong *trans*-specific balancing selection (Loisel *et al.* 2006).

Interestingly, a sliding window analysis showed that the peak values of *D* corresponded well with the *Hepatocystis* and allelic-imbalance-associated SNPs (Figure 22). Given that rates of *Hepatocystis* infection appear to vary across different populations (30% in the Masai Mara Reserve, Kenya, $n = 10$; 90% in Mikumi National Park, Tanzania, $n = 20$), these results suggest that the baboon *FY* cis-regulatory region may be subject to a complex selective history similar to the case described in humans (Hamblin *et al.* 2002; Seixas *et al.* 2002) in which differing levels of pathogen pressure across populations are associated with high levels of population differentiation around the *FY* gene, and varying signatures of selection within populations.

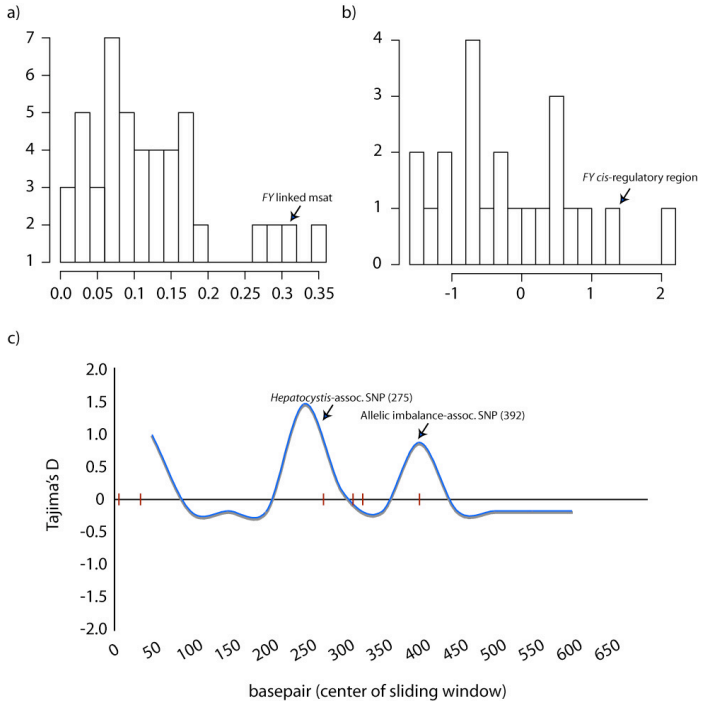

Figure 22: Comparison of genetic variation in and around the *FY cis*-regulatory region in relationship to other loci. Between population variation is shown in a) the distribution of F_{st} values between the Amboseli, Mikumi, and Masai Mara baboon populations for the *FY* linked microsatellite and 35 putatively neutral microsatellites around the genome. Within population variation in Amboseli is shown in b) the distribution of Tajima's D values within Amboseli from 22 loci in the baboon genome. Tajima's D was calculated using resequencing data for each locus and implemented using the DNAsp v. 4.9, assuming no recombination; and c) results of a sliding window analysis of Tajima's D for the *FY cis*-regulatory region within Amboseli, where window size = 100 bp and window interval = 50 bp. The locations of SNPs in the region are shown in red.

5.5 Discussion

Together, these data indicate that the *FY cis*-regulatory region is associated with parasite infection in a wild population of baboons, and that functional sequence variants within this region causally influence the level of expression of the *FY* gene. As in humans, variation in gene expression at the *FY* locus may therefore be important in parasite susceptibility, either through altering the direct access of *Hepatocystis* to baboon erythrocytes or, as has recently been demonstrated in humans, by altering a more general

property of the immune system, such as relative white blood cell counts (Reich *et al.* 2009).

These results suggest that the genetic basis of phenotypic variation in different primate species can exhibit a remarkable degree of parallelism. In this case, not only are these similarities present on the molecular level or on the level of trait association, as shown by previous work (Loisel *et al.* 2006; Wooding *et al.* 2006), but they also extend to the mechanism that links molecular and phenotypic variation (which is probably gene expression).

In spite of the parallelisms that we have documented for baboon and human *FY*, the functional variants we have identified in baboons are not homologous to the known functional variant in humans, which reveals that phenotypic variation in different primate species may show similar, but not precisely convergent, patterns of evolution. Indeed, while in humans the *FY*–malaria relationship is Mendelian, both *FY* expression and infection by *Hepatocystis* in baboons are clearly complex traits: even individuals homozygous for the *Hepatocystis* 'resistance' variant (the G allele at the A/G SNP) suffer from parasitism, albeit at a lower rate (52.2% of GG homozygotes were infected, versus 67.9% of AA homozygotes, across all study groups: see Figure 19). Additionally, the *in vitro* cell culture experiments suggest that the G allele of this variant actually drives higher expression of *FY* than the alternative A allele, even though the G allele is associated with a lower risk of *Hepatocystis* infection. The relationship between *FY* gene expression and *Hepatocystis* in baboons is therefore clearly different from that in humans, perhaps owing to balancing the cost of infection by other blood parasites, some of which are not known to co-occur with, and might be excluded by, *Hepatocystis* (Moore and Kuntz 1975). Alternatively, while the *in vitro* data on the A/G variant strongly suggest that this site has the capacity to influence *FY* gene expression, the direction and magnitude of its effects may differ in its natural cellular context.

This possibility is supported by the differences in magnitude of the effect of the C/T *cis*-regulatory variant in the *in vitro* transfection assays and the *in vivo* allelic imbalance measurements. *In vivo* gene expression measurements are complicated by variation in genetic background and in the environment, both of which can modify functional *cis*-regulatory effects (Brem *et al*. 2005; Smith and Kruglyak 2008). Indeed, our results show that even baboons that are homozygotes at the C/T site sometimes exhibit allelic imbalance in *FY* expression, suggesting that other, unidentified functional *cis*-regulatory variants are also segregating in the population. In contrast, in the *in vitro* comparisons, only a single *cis*-regulatory site differed between the experimental constructs, thus controlling for both environment and genetic backgrounds. Using both approaches in tandem can be synergistic: while *in vitro* experiments can help pin down specific functional sites, *in vivo* results demonstrate that these effects are relevant to the biology of individuals in the wild.

Thus, although identifying the genetic basis for phenotypic variation in wild primates poses substantial challenges, we present this study as a model to motivate additional evolutionary genetic research on natural primate populations. This work is essential if we hope to integrate an evolutionary and functional genetic perspective into the rich tradition of organismal research on these species. Our results demonstrate that patterns of variation in nonhuman primates can provide unique insights into the influence of ecological and environmental factors on genetic and trait variation in humans. Integrative research on nonhuman primates should also help us develop a better understanding of the evolution of our own species.

Works Cited

1. Abzhanov, A., W.P. Kuo, C. Hartmann, B.R. Grant, P.R. Grant and C.J. Tabin (2006). The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches. *Nature* 442(7102): 563-567.

2. Abzhanov, A., M. Protas, B.R. Grant, P.R. Grant and C.J. Tabin (2004). Bmp4 and morphological variation of beaks in Darwin's finches. *Science* 305(5689): 1462-1465.

3. Ackermann, R.R., J. Rogers and J.M. Cheverud (2006). Identifying the morphological signatures of hybridization in primate and human evolution. *Journal of Human Evolution* 51: 632-645.

4. Akashi, H. (1995). Inferring weak selection from patterns of polymorphism and divergence at silent sites in *Drosophila* DNA. *Genetics* 139(2): 1067-1076.

5. Albert, V., B. Jonsson and L. Bernatchez (2006). Natural hybrids in Atlantic eels (*Anguilla anguilla, A. rostrata*): evidence for successful reproduction and fluctuating abundance in space and time. *Molecular Ecology* 15(7): 1903-1916.

6. Alberts, S.C. and J. Altmann (1995). Balancing costs and opportunities: dispersal in male baboons. *American Naturalist* 145(2): 279-306.

7. Alberts, S.C. and J. Altmann (1995). Preparation and activation: determinants of age at reproductive maturity in male baboons. *Behavioral Ecology and Sociobiology* 36: 397-406.

8. Alberts, S.C. and J. Altmann (2001). Immigration and hybridization patterns of yellow and anubis baboons in and around Amboseli, Kenya. *American Journal of Primatology* 53(4): 139-154.

9. Alberts, S.C., J.C. Buchan and J. Altmann (2006). Sexual selection in wild baboons: from mating opportunities to paternity success. *Animal Behaviour* 72: 1177-1196.

10. Alberts, S.C., H.E. Watts and J. Altmann (2003). Queuing and queue-jumping: long-term patterns of reproductive skew in male savannah baboons, *Papio cynocephalus*. *Animal Behaviour* 65: 821-840.

11. Altmann, J. and S.C. Alberts (2003). Intraspecific variability in fertility and offspring survival in a nonhuman primate: behavioral control of ecological and social sources. Offspring: The Biodemography of Fertility and Family Behavior. K. Wachter and R. Bulatao. Washington, D.C., National Academy Press: 140-169.

12. Altmann, J. and S.C. Alberts (2003). Variability in reproductive success viewed from a life-history perspective in baboons. *American Journal of Human Biology* 15: 401-409.

13. Altmann, J. and S.C. Alberts (2005). Growth rates in a wild primate population: ecological influences and maternal effects. *Behavioral Ecology and Sociobiology* 57(5): 490-501.

14. Altmann, J., S.C. Alberts, S.A. Haines, J. Dubach, P. Muruthi, T. Coote, E. Geffen, D.J. Cheesman, R.S. Mututua, S.N. Saiyalel, *et al.* (1996). Behavior predicts genetic structure in a wild primate group. *Proceedings of the National Academy of Sciences of the United States of America* 93: 5797-5801.

15. Altmann, J. and P. Muruthi (1988). Differences in daily life between semiprovisioned and wild-feeding Baboons. *American Journal of Primatology* 15(3): 213-221.

16. Altmann, S. and J. Altmann (1970). Baboon Ecology. Basel, S. Karger.

17. Altmann, S.A. (1991). Diets of yearling female primates (*Papio cynocephalus*) predict lifetime fitness. *Proceedings of the National Academy of Sciences of the United States of America* 88: 420-423.

18. Anderson, E. and G.L. Stebbins (1954). Hybridization as an evolutionary stimulus. *Evolution* 8(4): 378-388.

19. Archie, E.A., J.E. Maldonado, J.A. Hollister-Smith, J.H. Poole, C.J. Moss, R.C. Fleischer and S.C. Alberts (2008). Fine-scale population genetic structure in a fission-fusion society. *Molecular Ecology* 17(11): 2666-2679.

20. Archie, E.A., T.A. Morrison, C.A.H. Foley, C.J. Moss and S.C. Alberts (2006). Dominance rank relationships among wild female African elephants, *Loxodonta africana*. *Animal Behaviour* 71: 117-127.

21. Arnold, M.L. (1992). Natural hybridization as an evolutionary process. *Annual Review of Ecology and Systematics* 23: 237-261.

22. Arnold, M.L. and S.A. Hodges (1995). Are natural hybrids fit or unfit relative to their parents? *Trends in Ecology & Evolution* 10(2): 67-71.

23. Arnold, M.L. and A. Meyer (2006). Natural hybridization in primates: One evolutionary mechanism. *Zoology* 109(4): 261-276.

24. Babbitt, C.C., O. Fedrigo, A.D. Pfefferle, J. Horvath, T.S. Furey and G.A. Wray (2010). Both noncoding and protein-coding RNAs contribute to gene expression evolution in the primate brain. *Genome Biology and Evolution* doi:10.1093/gbe/evq002.

25. Babbitt, C.C., J.S. Silverman, R. Haygood, J.M. Reininga, M.V. Rockman and G.A. Wray (2009). Multiple functional variants in *cis* modulate *PDYN* expression. *Molecular Biology and Evolution* doi:10.1093/molbev/msp276.

26. Bamshad, M.J., S. Mummidi, E. Gonzalez, S.S. Ahuja, D.M. Dunn, W.S. Watkins, S. Wooding, A.C. Stone, L.B. Jorde, R.B. Weiss, *et al.* (2002). A strong signature of

balancing selection in the 5' cis-regulatory region of *CCR5*. *Proc Natl Acad Sci U S A* 99(16): 10539-10544.

27. Barker, D.J.P. (2002). Fetal programming of coronary heart disease. *Trends in Endocrinology and Metabolism* 13(9): 364-368.

28. Barker, D.J.P., J.G. Eriksson, T. Forsen and C. Osmond (2002). Fetal origins of adult disease: strength of effects and biological basis. *International Journal of Epidemiology* 31(6): 1235-1239.

29. Barnwell, J.W., M.E. Nichols and P. Rubinstein (1989). *In vitro* evaluation of the role of the Duffy Blood Group in erythrocyte invasion by *Plasmodium vivax*. *Journal of Experimental Medicine* 169(5): 1795-1802.

30. Barr, C.S., T.K. Newman, M.L. Becker, M. Champoux, K.P. Lesch, S.J. Suomi, D. Goldman and J.D. Higley (2003). Serotonin transporter gene variation is associated with alcohol sensitivity in rhesus macaques exposed to early-life stress. *Alcoholism: Clinical and Experimental Research* 27(5): 812-817.

31. Barr, C.S., T.K. Newman, S. Lindell, C. Shannon, M. Champoux, K.P. Lesch, S.J. Suomi, D. Goldman and J.D. Higley (2004). Interaction between serotonin transporter gene variation and rearing condition in alcohol preference and consumption in female primates. *Archives of General Psychiatry* 61(11): 1146-1152.

32. Barton, N.H. (2001). The role of hybridization in evolution. *Molecular Ecology* 10(3): 551-568.

33. Barton, N.H. and G.M. Hewitt (1985). Analysis of hybrid zones. *Annual Review of Ecology and Systematics* 16: 113-148.

34. Beaumont, M., E.M. Barratt, D. Gottelli, A.C. Kitchener, M.J. Daniels, J.K. Pritchard and M.W. Bruford (2001). Genetic diversity and introgression in the Scottish wildcat. *Molecular Ecology* 10(2): 319-336.

35. Becquet, C., N. Patterson, A.C. Stone, M. Przeworski and D. Reich (2007). Genetic structure of chimpanzee populations. *Plos Genetics* 3(4): e66.

36. Becquet, C. and M. Przeworski (2007). A new approach to estimate parameters of speciation models with application to apes. *Genome Research* 17(10): 1505-1519.

37. Beehner, J.C., D. Onderdonk, S.C. Alberts and J. Altmann (2006). The ecology of conception and pregnancy failure in wild baboons. *Behavioral Ecology* 17: 741-750.

38. Blanchette, M., W.J. Kent, C. Riemer, L. Elnitski, A.F.A. Smit, K.M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E.D. Green, *et al.* (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research* 14(4): 708-715.

39. Blekhman, R., A. Oshlack, A.E. Chabot, G.K. Smyth and Y. Gilad (2008). Gene regulation in primates evolves under tissue-specific selection pressures. *Plos Genetics* 4(11): e1000271.

40. Boffelli, D., M.A. Nobrega and E.M. Rubin (2004). Comparative genomics at the vertebrate extremes. *Nature Reviews Genetics* 5(6): 456-465.

41. Bray, N.J., P.R. Buckland, M.J. Owen and M.C. O'Donovan (2003). Cis-acting variation in the expression of a high proportion of genes in human brain. *Human Genetics* 113(2): 149-153.

42. Brem, R.B., J.D. Storey, J. Whittle and L. Kruglyak (2005). Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* 436(7051): 701-703.

43. Brown, M.P.S., W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares and D. Haussler (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America* 97(1): 262-267.

44. Brunet, J.-P., P. Tamayo, T. Golub and J.P. Mesirov (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America* 101: 4164-4169.

45. Buchan, J.C., S.C. Alberts, J.B. Silk and J. Altmann (2003). True paternal care in a multi-male primate society. *Nature* 425: 179-181.

46. Buchan, J.C., E.A. Archie, R.C. Van Horn, C.J. Moss and S.C. Alberts (2005). Locus effects and sources of error in noninvasive genotyping. *Molecular Ecology Notes* 5(3): 680-683.

47. Burrell, A.S., C.J. Jolly, A.J. Tosi and T.R. Disotell (2009). Mitochondrial evidence for the hybrid origin of the kipunji, *Rungwecebus kipunji* (Primates: Papionini). *Molecular Phylogenetics and Evolution* 51(2): 340-348.

48. Bustamante, C.D., A. Fledel-Alon, S. Williamson, R. Nielsen, M.T. Hubisz, S. Glanowski, D.M. Tanenbaum, T.J. White, J.J. Sninsky, R.D. Hernandez, *et al.* (2005). Natural selection on protein-coding genes in the human genome. *Nature* 437(7062): 1153-1157.

49. Campbell, C.D., A. Kirby, J. Nemesh, M.J. Daly and J.N. Hirschhorn (2008). A survey of allelic imbalance in F1 mice. *Genome Research* 18: 555-563.

50. Carroll, S.B. (2005). Endless Forms Most Beautiful: the New Science of Evo-Devo. New York, W.W. Norton & Company.

51. Charpentier, M.J.E., J. Tung, J. Altmann and S.C. Alberts (2008). Age at maturity in wild baboons: genetic, environmental, and demographic influences. *Molecular Ecology* 17(8): 2026-2040.

52. Charpentier, M.J.E., J. Tung, J. Altmann and S.C. Alberts (In prep). Genetic and nongenetic predictors of mating behavior in wild baboons.

53. Charpentier, M.J.E., R.C. Van Horn, J. Altmann and S.C. Alberts (2008). Paternal effects on offspring fitness in a multi-male primate society. *Proceedings of the National Academy of Sciences of the United States of America* 105: 1988-1992.

54. Chesser, R.K. (1991). Influence of gene flow and breeding tactics on gene diversity within populations. *Genetics* 129(2): 573-583.

55. Cheung, V.G., A. Bruzel, J.T. Burdick, M. Morley, J.L. Devlin and R.S. Spielman (2008). Monozygotic twins reveal germline contribution to allelic expression differences. *American Journal of Human Genetics* 82(6): 1357-1360.

56. Cheung, V.G., R.S. Spielman, K.G. Ewens, T.M. Weber, M. Morley and J.T. Burdick (2005). Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 437(7063): 1365-1369.

57. Clark, A.G., M.B. Eisen, D.R. Smith, C.M. Bergman, B. Oliver, T.A. Markow, T.C. Kaufman, M. Kellis, W. Gelbart, V.N. Iyer, *et al.* (2007). Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450(7167): 203-218.

58. Clutton-Brock, T.H. (1989). Red Deer in the Highlands. Oxford, England, Oxford University Press.

59. Clutton-Brock, T.H. and J. Pemberton, Eds. (2004). Soay Sheep: Dynamics and Selection in an Island Population. Cambridge, Cambridge University Press.

60. Colosimo, P.F., K.E. Hosemann, S. Balabhadra, G. Villarreal, Jr., M. Dickson, J. Grimwood, J. Schmutz, R.M. Myers, D. Schluter and D.M. Kingsley (2005). Widespread parallel evolution in sticklebacks by repeated fixation of *Ectodysplasin* alleles. *Science* 307(5717): 1928-1933.

61. Cortes, C. and V. Vapnik (1995). Support-Vector Networks. *Machine Learning* 20(3): 273-297.

62. Cowles, C.R., J.N. Hirschhorn, D. Altshuler and E.S. Lander (2002). Detection of regulatory variation in mouse genes. *Nature Genetics* 32(3): 432-437.

63. de Meaux, J., U. Goebel, A. Pop and T. Mitchell-Olds (2005). Allele-specific assay reveals functional variation in the chalcone synthase promoter of *Arabidiopsis thaliana* that is compatible with neutral evolution. *Plant Cell* 17(3): 676-690.

64. de Meaux, J., A. Pop and T. Mitchell-Olds (2006). *Cis*-regulatory evolution of chalcone-synthase expression in the genus Arabidopsis. *Genetics* 174: 2181-2202.

65. Degner, J.F., J.C. Marioni, A.A. Pai, J.K. Pickrell, E. Nkadori, Y. Gilad and J.K. Pritchard (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* 25(24): 3207-3212.

66. Denver, D.R., K. Morris, J.T. Streelman, S.K. Kim, M. Lynch and W.K. Thomas (2005). The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans. *Nature Genetics* 37(5): 544-548.

67. Devarajan, K. (2008). Nonnegative matrix factorization: an analytical and interpretive tool in computation biology. *Plos Computational Biology* 4: e1000029.

68. dos Reis, M. and L. Wernisch (2009). Estimating translational selection in eukaryotic genomes. *Molecular Biology and Evolution* 26: 451-461.

69. Drake, J.M., C. Randin and A. Guisan (2006). Modelling ecological niches with support vector machines. *Journal of Applied Ecology* 43(3): 424-432.

70. Enard, W., P. Khaitovich, J. Klose, S. Zollner, F. Heissig, P. Giavalisco, K. Nieselt-Struwe, E. Muchmore, A. Varki, R. Ravid, *et al.* (2002). Intra- and interspecific variation in primate gene expression patterns. *Science* 296(5566): 340-343.

71. Enard, W. and S. Paabo (2004). Comparative primate genomics. *Annual Review of Genomics and Human Genetics* 5: 351-378.

72. Excoffier, L., G. Laval and S. Schneider (2005). Arlequin ver. 3.0: an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* 1: 47-50.

73. Falush, D., M. Stephens and J.K. Pritchard (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* 164(4): 1567-1587.

74. Foerster, K., T. Coulson, B.C. Sheldon, J.M. Pemberton, T.H. Clutton-Brock and L.E. Kruuk (2007). Sexually antagonistic genetic variation for fitness in red deer. *Nature* 447(7148): 1107-1110.

75. Garnham, P.C.C. (1966). Malaria parasites and other haemosporidia. Oxford, Blackwell.

76. Garnham, P.C.C., R.B. Heisch, D.M. Minter, J.D. Phipps and M. Ikata (1961). *Culicoides adersi* Ingram and Macfie, 1923, a presumed vector of *Hepatocystis* (= *Plasmodium*) *kochi* (Laveran, 1899). *Nature* 190: 739-741.

77. Gierman, H.J., M.H.G. Indemans, J. Koster, S. Goetze, J. Seppen, D. Geerts, R. van Driel and R. Versteeg (2007). Domain-wide regulation of gene expression in the human genome. *Genome Research* 17(9): 1286-1295.

78. Gilad, Y., A. Oshlack and S.A. Rifkin (2006). Natural selection on gene expression. *Trends in Genetics* 22(8): 456-461.

79. Gilad, Y., A. Oshlack, G.K. Smyth, T.P. Speed and K.P. White (2006). Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* 440(7081): 242-245.

80. Gilad, Y., J.K. Pritchard and K. Thornton (2009). Characterizing natural variation using next-generation sequencing technologies. *Trends in Genetics* 25(10): 463-471.

81. Godfrey, K.M. and D.J.P. Barker (2000). Fetal nutrition and adult disease. *American Journal of Clinical Nutrition* 71(5): 1344S-1352S.

82. Golub, T.R., D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, *et al*. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286(5439): 531-537.

83. Gompel, N., B. Prud'homme, P.J. Wittkopp, V.A. Kassner and S.B. Carroll (2005). Chance caught on the wing: *cis*-regulatory evolution and the origin of pigment patterns in Drosophila. *Nature* 433(7025): 481-487.

84. Grant, P.R. (1986). Ecology and Evolution of Darwin's Finches. Princeton, NJ, Princeton University Press.

85. Grant, P.R. and B.R. Grant (1992). Hybridization of bird species. *Science* 256: 193-197.

86. Gruber, J.D. and A.D. Long (2008). *Cis*-regulatory variation is typically poly-allelic in *Drosophila*. *Genetics* 181: 661-670.

87. Guo, Q.H., M. Kelly and C.H. Graham (2005). Support vector machines for predicting distribution of sudden oak death in California. *Ecological Modelling* 182(1): 75-90.

88. Gurganus, M.C., J.D. Fry, S.V. Nuzhdin, E.G. Pasyukova, R.F. Lyman and T.F. Mackay (1998). Genotype-environment interaction at quantitative trait loci affecting sensory bristle number in Drosophila melanogaster. *Genetics* 149(4): 1883-1898.

89. Guyon, I., J. Weston, S. Barnhill and V. Vapnik (2002). Gene selection for cancer classification using support vector machines. *Machine Learning* 46(1-3): 389-422.

90. Hahn, M.W., J.E. Stajich and G.A. Wray (2003). The effects of selection against spurious transcription factor binding sites. *Molecular Biology and Evolution* 20(6): 901-906.

91. Hamblin, M.T. and A. Di Rienzo (2000). Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *American Journal of Human Genetics* 66(5): 1669-1679.

92. Hamblin, M.T., E.E. Thompson and A. Di Rienzo (2002). Complex signatures of natural selection at the Duffy blood group locus. *American Journal of Human Genetics* 70: 369-383.

93. Hausfater, G., J. Altmann and S. Altmann (1982). Long-term consistency of dominance relations among female baboons (*Papio cynocephalus*). *Science* 217(4561): 752-755.

94. Haygood, R., C.C. Babbitt, O. Fedrigo and G.A. Wray (in review). Most developmental adaptation during human evolution occurred via changes in noncoding DNA.

95. Haygood, R., O. Fedrigo, B. Hanson, K.-D. Yokoyama and G.A. Wray (2007). Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nature Genetics* 39: 1140-1144.

96. He, W., S. Neil, H. Kulkarni, E. Wright, B.K. Agan, V.C. Marconi, M.J. Dolan, R.A. Weiss and S.K. Ahuja (2008). Duffy antigen receptor for chemokines mediates trans-infection of HIV-1 from red blood cells to target cells and affects HIV-AIDS susceptibility. *Cell Host & Microbe* 4(1): 52-62.

97. Heap, G.A., J.H. Yang, K. Downes, B.C. Healy, K.A. Hunt, N. Bockett, L. Franke, P.C. Dubois, C.A. Mein, R.J. Dobson, *et al.* (2010). Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Human Molecular Genetics* 19(1): 122-134.

98. Henzi, P. and L. Barrett (2003). Evolutionary ecology, sexual conflict, and behavioral differentiation among baboon populations. *Evolutionary Anthropology* 12(5): 217-230.

99. Hernandez, R.D., M.J. Hubisz, D.A. Wheeler, D.G. Smith, B. Ferguson, J. Rogers, L. Nazareth, A. Indap, T. Bourquin, J. McPherson, *et al.* (2007). Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques. *Science* 316(5822): 240-243.

100. Herrera, S., A. Gomez, O. Vera, J. Vergara, A. Valderrama-Aguirre, A. Maestre, F. Mendez, R. Wang, C.E. Chitnis, S.S. Yazdani, *et al.* (2005). Antibody response to Plasmodium vivax antigens in Fy-negative individuals from the Colombian Pacific coast. *American Journal of Tropical Medical Hygiene* 73(5 Suppl): 44-49.

101. Hiwatashi, T., Y. Okabe, T. Tsutsui, C. Hiramatsu, A.D. Melin, H. Oota, C.M. Schaffner, F. Aureli, L.M. Fedigan, H. Innan, *et al.* (2009). An explicit signature of balancing selection for color vision variation in New World monkeys. *Molecular Biology and Evolution*.

102. Hoekstra, H.E., J.G. Krenz and M.W. Nachman (2005). Local adaptation in the rock pocket mouse (*Chaetodipus intermedius*): natural selection and phylogenetic history of populations. *Heredity* 94(2): 217-228.

103. Hoffjan, S., D. Nicolae, I. Ostrovnaya, K. Roberg, M. Evans, D.B. Mirel, L. Steiner, K. Walker, P. Shult, R.E. Gangnon, *et al.* (2005). Gene-environment interaction effects on the development of immune responses in the 1st year of life. *American Journal of Human Genetics* 76(4): 696-704.

104. Hofmann, C.M., K.E. O'Quin, N.J. Marshall, T.W. Cronin, O. Seehausen and K.L. Carleton (2009). The eyes have it: regulatory and structural changes both underlie cichlid visual pigment diversity. *PLoS Biol* 7(12): e1000266.

105. Hughes, A.L., B. Packer, R. Welch, S.J. Chanock and M. Yeager (2005). High level of functional polymorphism indicates a unique role of natural selection at human immune system loci. *Immunogenetics* 57(11): 821-827.

106. Hunter, D.J. (2005). Gene-environment interactions in human diseases. *Nature Review Genetics* 6(4): 287-298.

107. Jakobsdottir, J., M.B. Gorin, Y.P. Conley, R.E. Ferrell and D.E. Weeks (2009). Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. *PLoS Genet* 5(2): e1000337.

108. Jeong, S., M. Rebeiz, P. Andolfatto, T. Werner, J. True and S.B. Carroll (2008). The evolution of gene regulation underlies a morphological difference between two Drosophila sister species. *Cell* 132(5): 783-793.

109. Joachims, T. (2005). A support vector method for multivariate performance measures. Proceedings of the International Conference on Machine Learning.

110. Joachims, T. (2006). Training linear SVMs in linear time. Proceedings of the ACM Conference on Knowledge Discovery and Data Mining.

111. Jolly, C.J. (1993). Species, subspecies, and baboon systematics. Species, Species Concepts, and Primate Evolution. W.H. Kimbel and L.B. Martin. New York, Plenum Press: 67-107.

112. Jolly, C.J. (2001). A proper study for mankind: analogies from the papionin monkeys and their implications for human evolution. *Yearbook of Physical Anthropology* 116(S33): 177-204.

113. Jolly, C.J. and J.E. Phillips-Conroy (2007). Ecology, history and society as determinants of hybrid zone structure in baboons. *American Journal of Physical Anthropology* 132(S44): 138.

114. Kennerly, E., A. Ballmann, S. Martin, R. Wolfinger, S. Gregory, M. Stoskopf and G. Gibson (2008). A gene expression signature of confinement in peripheral blood of red wolves (*Canis rufus*). *Molecular Ecology* 17(11): 2782-2791.

115. Kent, W.J., C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler and D. Haussler (2002). The human genome browser at UCSC. *Genome Research* 6: 996-1006.

116. Khaitovich, P., I. Hellmann, W. Enard, K. Nowick, M. Leinweber, H. Franz, G. Weiss, M. Lachmann and S. Paabo (2005). Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* 309(5742): 1850-1854.

117. Khoury, M.J., R. Valdez and A. Albright (2008). Public health genomics approach to type 2 diabetes. *Diabetes* 57(11): 2911-2914.

118. Kim, H., G.H. Golub and H. Park (2006). Missing value estimation for DNA microarray gene expression data: local least squares imputation (vol 21, pg 187, 2005). *Bioinformatics* 22(11): 1410-1411.

119. King, M. and A. Wilson (1975). Evolution at two levels in humans and chimpanzees. *Science* 188: 107-116.

120. Kingdon, J. (1997). The Kingdon Field Guide to African Mammals. San Diego, CA, Plenum Press.

121. Kosiol, C., T. Vinar, R.R. da Fonseca, M.J. Hubisz, C.D. Bustamante, R. Nielsen and A. Siepel (2008). Patterns of positive selection in six mammalian genomes. *Plos Genetics* 4(8): e1000144.

122. Kruglyak, S. and H.X. Tang (2000). Regulation of adjacent yeast genes. *Trends in Genetics* 16(3): 109-111.

123. Kruuk, L.E. and W.G. Hill (2008). Introduction. Evolutionary dynamics of wild populations: the use of long-term pedigree data. *Proceedings of the Royal Society B* 275(1635): 593-596.

124. Kurreeman, F.A., J.J. Schonkeren, B.T. Heijmans, R.E. Toes and T.W. Huizinga (2004). Transcription of the *IL10* gene reveals allele-specific regulation at the mRNA level. *Human Molecular Genetics* 13(16): 1755-1762.

125. Laing, K.J. and C.J. Secombes (2004). Chemokines. *Developmental and Comparative Immunology* 28(5): 443-460.

126. Lango, H., C.N. Palmer, A.D. Morris, E. Zeggini, A.T. Hattersley, M.I. McCarthy, T.M. Frayling and M.N. Weedon (2008). Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. *Diabetes* 57(11): 3129-3135.

127. Lecis, R., M. Pierpaoli, Z.S. Biro, L. Szemethy, B. Ragni, F. Vercillo and E. Randi (2006). Bayesian analyses of admixture in wild and domestic cats (*Felis silvestris*) using linked microsatellite loci. *Molecular Ecology* 15(1): 119-131.

128. Lee, S.H., J.H. van der Werf, B.J. Hayes, M.E. Goddard and P.M. Visscher (2008). Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genetics* 4(10): e1000231.

129. Leips, J. and T.F. Mackay (2000). Quantitative trait loci for life span in Drosophila melanogaster: interactions with genetic background and larval density. *Genetics* 155(4): 1773-1788.

130. Lemos, B., L.O. Araripe, P. Fontanillas and D.L. Hartl (2008). Dominance and the evolutionary accumulation of *cis*- and *trans*-effects on gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 105: 14471-14476.

131. Lercher, M.J., A.O. Urrutia and L.D. Hurst (2002). Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature Genetics* 31(2): 180-183.

132. Leroi, A.M., A.K. Chippindale and M.R. Rose (1994). Long-term laboratory evolution of a genetic life-history trade-off in *Drosophila melanogaster* 1. The role of genotype-by-environment interaction. *Evolution* 48(4): 1244-1257.

133. Lewontin, R.C. and L.C. Birch (1966). Hybridization as a source of variation for adaptation to new environments. *Evolution* 20(3): 315-336.

134. Lindstrom, J. (1999). Early development and fitness in birds and mammals. *Trends in Ecology & Evolution* 14(9): 343-348.

135. Linnen, C.R., E.P. Kingsley, J.D. Jensen and H.E. Hoekstra (2009). On the origin and spread of an adaptive allele in deer mice. *Science* 325(5944): 1095-1098.

136. Liu, F., K. van Duijn, J.R. Vingerling, A. Hofman, A.G. Uitterlinden, A.C. Janssens and M. Kayser (2009). Eye color and the prediction of complex phenotypes from genotypes. *Current Biology* 19(5): R192-193.

137. Liu, H., D. Chao, E.E. Nakayama, H. Taguchi, M. Goto, X. Xin, J.K. Takamatsu, H. Saito, Y. Ishikawa, T. Akaza, *et al.* (1999). Polymorphism in RANTES chemokine promoter affects HIV-1 disease progression. *Proceedings of the National Academy of Sciences of the United States of America* 96(8): 4581-4585.

138. Lo, H.S., Z.N. Wang, Y. Hu, H.H. Yang, S. Gere, K.H. Buetow and M.P. Lee (2003). Allelic variation in gene expression is common in the human genome. *Genome Research* 13(8): 1855-1862.

139. Loisel, D. (2007). Evolutionary genetics of immune system genes in a wild primate population. Department of Biology. Durham, NC, Duke University. PhD Dissertation.

140. Loisel, D.A., M.V. Rockman, G.A. Wray, J. Altmann and S.C. Alberts (2006). Ancient polymorphism and functional variation in the primate *MHC-DQA1* 5 ' *cis*-regulatory region. *Proceedings of the National Academy of Sciences of the United States of America* 103(44): 16331-16336.

141. Lorenzen, E.D. and H.R. Siegismund (2004). No suggestion of hybridization between the vulnerable black-faced impala (*Aepyceros melampus petersi*) and the common impala (*A.m. melampus*) in Etosha National Park, Namibia. *Molecular Ecology* 13(10): 3007-3019.

142. Luedi, P.P., A.J. Hartemink and R.L. Jirtle (2005). Genome-wide prediction of imprinted murine genes. *Genome Research* 15(6): 875-884.

143. Maples, W. and T. McKern (1967). A preliminary report on classification of the Kenya baboon. The Baboon in Medical Research. H. Vagtborg. Austin, TX, University of Texas Press. 2: 13-22.

144. Mavarez, J., C.A. Salazar, E. Bermingham, C. Salcedo, C.D. Jiggins and M. Linares (2006). Speciation by hybridization in Heliconius butterflies. *Nature* 441(7095): 868-871.

145. Mayr, E. (1942). Systematics and the Origin of Species. New York, Columbia University.

146. McGregor, A.P., V. Orgogozo, I. Delon, J. Zanet, D.G. Srinivasan, F. Payre and D.L. Stern (2007). Morphological evolution through multiple *cis*-regulatory mutations at a single gene. *Nature* 448(7153): 587-590.

147. McKusick-Nathans Institute of Genetic Medicine, J.H.U. and N.L.o.M. National Center for Biotechnology Information (2010). Online Mendelian Inheritance in Man (OMIM).

148. McVicker, G., D. Gordon, C. Davis and P. Green (2009). Widespread genomic signatures of natural selection in hominid evolution. *Plos Genetics* 5(5): e1000471.

149. Meaney, M.J. and M. Szyf (2005). Maternal care as a model for experience-dependent chromatin plasticity? *Trends in Neurosciences* 28(9): 456-463.

150. Melin, A.D., L.M. Fedigan, C. Hiramatsu and S. Kawamura (2008). Polymorphic color vision in white-faced capuchins (*Cebus capucinus*): Is there foraging niche divergence among phenotypes? *Behavioral Ecology and Sociobiology* 62(5): 659-670.

151. Melin, A.D., L.M. Fedigan, C. Hiramatsu, C.L. Sendall and S. Kawamura (2007). Effects of colour vision phenotype on insect capture by a free-ranging population of white-faced capuchins, *Cebus capucinus*. *Animal Behaviour* 73: 205-214.

152. Michon, P., I. Woolley, E.M. Wood, W. Kastens, P.A. Zimmerman and J.H. Adams (2001). Duffy-null promoter heterozygosity reduces *DARC* expression and abrogates adhesion of the *P vivax* ligand required for blood-stage infection. *FEBS Letters* 495(1-2): 111-114.

153. Milani, L., M. Gupta, M. Andersen, S. Dhar, M. Fryknas, A. Isaksson, R. Larsson and A. Syvanen (2007). Allelic imbalance in gene expression as a guide to *cis*-acting regulatory single nucleotide polymorphisms in cancer cells. *Nucleic Acids Research* 35: e34.

154. Miller, L.H., S.J. Mason, D.F. Clyde and M.H. Mcginniss (1976). Resistance factor to *Plasmodium vivax* in blacks - Duffy-Blood-Group genotype, Fyfy. *New England Journal of Medicine* 295(6): 302-304.

155. Miller, L.H., S.J. Mason, J.A. Dvorak, M.H. Mcginniss and I.K. Rothman (1975). Erythrocyte receptors for *Plasmodium knowlesi* malaria - Duffy Blood-Group determinants. *Science* 189(4202): 561-563.

156. Moore, J.A. and R.E. Kuntz (1975). *Entopolyploides macaci* Mayer, 1934 in the African baboon (*Papio cynocephalus* L. 1766). *Journal of Medical Primatology* 4: 1-7.

157. Moore, W.S. (1977). An evaluation of narrow hybrid zones in vertebrates. *The Quarterly Review of Biology* 52(3): 263-277.

158. Morley, M., C.M. Molony, T.M. Weber, J.L. Devlin, K.G. Ewens, R.S. Spielman and V.G. Cheung (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* 430(7001): 743-747.

159. Mukherjee, S., P. Tamayo, D. Slonim, A. Verri, T. Golub, J.P. Mesirov and T. Poggio (1999). Support vector machine classification of microarray data, Massachusetts Institute of Technology.

160. Muruthi, P., J. Altmann and S. Altmann (1991). Resource base, parity, and reproductive condition affect females feeding time and nutrient intake within and between groups of a baboon population. *Oecologia* 87(4): 467-472.

161. Myers, B.J. and R.E. Kuntz (1965). A checklist of parasites reported for the baboon. *Primates* 6(2): 137-194.

162. Nachman, M.W. and S.L. Crowell (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics* 156(1): 297-304.

163. Nachman, M.W., H.E. Hoekstra and S.L. D'Agostino (2003). The genetic basis of adaptive melanism in pocket mice. *Proceedings of the National Academy of Sciences of the United States of America* 100(9): 5268-5273.

164. Nagel, U. (1973). A comparison of anubis baboons, hamadryas baboons and their hybrids at a species border in Ethiopia. *Folia Primatologica* 19(2): 104-165.

165. Newman, T.K., C.J. Jolly and J. Rogers (2004). Mitochondrial phylogeny and systematics of baboons (*Papio*). *American Journal of Physical Anthropology* 124(1): 17-27.

166. Newman, T.K., Y.V. Syagailo, C.S. Barr, J.R. Wendland, M. Champoux, M. Graessle, S.J. Suomi, J.D. Higley and K.P. Lesch (2005). Monoamine oxidase A gene promoter variation and rearing experience influences aggressive behavior in rhesus monkeys. *Biological Psychiatry* 57(2): 167-172.

167. Nielsen, E.E., M.M. Hansen, D.E. Ruzzante, D. Meldrup and P. Gronkjaer (2003). Evidence of a hybrid-zone in Atlantic cod (*Gadus morhua*) in the Baltic and the Danish Belt Sea revealed by individual admixture analysis. *Molecular Ecology* 12(6): 1497-1508.

168. Nielsen, E.E., P.H. Nielsen, D. Meldrup and M.M. Hansen (2004). Genetic population structure of turbot (*Scophthalmus maximus* L.) supports the presence of multiple hybrid zones for marine fishes in the transition zone between the Baltic Sea and the North Sea. *Molecular Ecology* 13(3): 585-595.

169. Nielsen, R., C. Bustamante, A.G. Clark, S. Glanowski, T.B. Sackton, M.J. Hubisz, A. Fledel-Alon, D.M. Tanenbaum, D. Civello, T.J. White, *et al.* (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. *Plos Biology* 3(6): 976-985.

170. Noren, K., L. Dalen, K. Kvaloy and A. Angerbjorn (2005). Detection of farm fox and hybrid genotypes among wild arctic foxes in Scandinavia. *Conservation Genetics* 6(6): 885-894.

171. Nussey, D.H., D.W. Coltman, T. Coulson, L.E.B. Kruuk, A. Donald, S.J. Morris, T.H. Clutton-Brocks and J. Pemberton (2005). Rapidly declining fine-scale spatial genetic structure in female red deer. *Molecular Ecology* 14(11): 3395-3405.

172. Ober, C. and E.E. Thompson (2005). Rethinking genetic models of asthma: the role of environmental modifiers. *Current Opinion in Immunology* 17(6): 670-678.

173. Onyango, P.O., L.R. Gesquiere, E.O. Wango, S.C. Alberts and J. Altmann (2008). Persistence of maternal effects in baboons: mother's dominance rank at son's conception predicts stress hormone levels in subadult males. *Hormones & Behavior* 54(2): 319-324.

174. Packer, C., D.A. Collins and L.E. Eberly (2000). Problems with primate sex ratios. *Philosophical Transactions of the Royal Society of London B Biological Sciences* 355(1403): 1627-1635.

175. Pant, P.V.K., H. Tao, E.J. Beilharz, D.G. Ballinger, D.R. Cox and K.A. Frazer (2006). Analysis of allelic differential expression in human white blood cells. *Genome Research* 16(3): 331-339.

176. Pastinen, T. and T.J. Hudson (2004). *Cis*-acting regulatory variation in the human genome. *Science* 306(5296): 647-650.

177. Patterson, N., D.J. Richter, S. Gnerre, E.S. Lander and D. Reich (2006). Genetic evidence for complex speciation of humans and chimpanzees. *Nature* 441(7097): 1103-1108.

178. Perkins, S.L. and J.J. Schall (2002). A molecular phylogeny of malarial parasites recovered from cytochrome b gene sequences. *Journal of Parasitology* 88(5): 972-978.

179. Phillips-Conroy, J.E. and C.J. Jolly (1986). Changes in the structure of the baboon hybrid zone in the Awash National Park, Ethiopia. *American Journal of Physical Anthropology* 71(3): 337-350.

180. Pierpaoli, M., Z.S. Biro, M. Herrmann, K. Hupe, M. Fernandes, B. Ragni, L. Szemethy and E. Randi (2003). Genetic distinction of wildcat (*Felis silvestris*) populations in Europe, and hybridization with domestic cats in Hungary. *Molecular Ecology* 12(10): 2585-2598.

181. Pollard, K.S., S.R. Salama, N. Lambert, M.A. Lambot, S. Coppens, J.S. Pedersen, S. Katzman, B. King, C. Onodera, A. Siepel, *et al.* (2006). An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443(7108): 167-172.

182. Pope, T.R. (1992). The influence of dispersal patterns and mating system on genetic differentiation within and between populations of the red howler monkey (*Alouatta seniculus*). *Evolution* 46(4): 1112-1128.

183. Prabhakar, S., J.P. Noonan, S. Paabo and E.M. Rubin (2006). Accelerated evolution of conserved noncoding sequences in humans. *Science* 314(5800): 786-786.

184. Prabhakar, S., A. Visel, J.A. Akiyama, M. Shoukry, K.D. Lewis, A. Holt, I. Plajzer-Frick, H. Morrison, D.R. FitzPatrick, V. Afzal, *et al.* (2008). Human-specific gain of function in a developmental enhancer. *Science* 321(5894): 1346-1350.

185. Presgraves, D.C. and S.V. Yi (2009). Doubts about complex speciation between humans and chimpanzees. *Trends in Ecology & Evolution* 24(10): 533-540.

186. Price, A.L., N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick and D. Reich (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38(8): 904-909.

187. Pritchard, J.K., M. Stephens and P. Donnelly (2000). Inference of population structure using multilocus genotype data. *Genetics* 155(2): 945-959.

188. Prokopenko, I., M.I. McCarthy and C.M. Lindgren (2008). Type 2 diabetes: new genes, new understanding. *Trends Genet* 24(12): 613-621.

189. Prud'homme, B., N. Gompel, A. Rokas, V.A. Kassner, T.M. Williams, S.D. Yeh, J.R. True and S.B. Carroll (2006). Repeated morphological evolution through *cis*-regulatory changes in a pleiotropic gene. *Nature* 440(7087): 1050-1053.

190. Qvarnstrom, A. (1999). Genotype-by-environment interactions in the determination of the size of a secondary sexual character in the collared flycatcher (*Ficedula albicollis*). *Evolution* 53(5): 1564-1572.

191. Ravelli, G.P., Z.A. Stein and M.W. Susser (1976). Obesity in young men after famine exposure in utero and early infancy. *New England Journal of Medicine* 295(7): 349-353.

192. Reich, D., M.A. Nalls, W.H.L. Kao, E.L. Akylbekova, A. Tandon, N.J. Patterson, J. Mullikin, W. Hsueh, C. Chen, J. Coresh, *et al.* (2009). Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy Antigen Receptor for Chemokines gene. *PLoS Genetics* 5(1): e1000360.

193. Reifsnyder, P.C., G. Churchill and E.H. Leiter (2000). Maternal environment and genotype interact to establish diabesity in mice. *Genome Research* 10(10): 1568-1578.

194. Rhymer, J.M. and D. Simberloff (1996). Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics* 27: 83-109.

195. Rice, P., I. Longden and A. Bleasby (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* 16: 276-277.

196. Rieseberg, L.H. (1997). Hybrid origins of plant species. *Annual Review of Ecology and Systematics* 28: 359-389.

197. Rieseberg, L.H., O. Raymond, D.M. Rosenthal, Z. Lai, K. Livingstone, T. Nakazato, J.L. Durphy, A.E. Schwarzbach, L. Donovan and C. Lexer (2003). Major ecological transitions in wild sunflowers facilitated by hybridization. *Science* 301: 1211-1216.

198. Rifkin, S.A., J. Kim and K.P. White (2003). Evolution of gene expression in the Drosophila melanogaster subgroup. *Nature Genetics* 33(2): 138-144.

199. Roettger, M., W. Martin and T. Dagan (2009). A machine-learning approach reveals that alignment properties alone can accurately predict inference of lateral gene transfer from discordant phylogenies. *Molecular Biology and Evolution* 26(9): 1931-1939.

200. Rogers, J. and K.K. Kidd (1996). Nucleotide polymorphism, effective population size, and dispersal distances in the yellow baboons (*Papio hamadryas cynocephalus*) of Mikumi National Park, Tanzania. *American Journal of Primatology* 38(2): 157-168.

201. Rogers, J., M.C. Mahaney, S.M. Witte, S. Nair, D. Newman, S. Wedel, L.A. Rodriguez, K.S. Rice, S.H. Slifer, A. Perelygin, *et al.* (2000). A genetic linkage map of the baboon (*Papio hamadryas*) genome based on human microsatellite polymorphisms. *Genomics* 67(3): 237-247.

202. Rougemont, M., M. Van Saanen, R. Sahli, H.P. Hinrikson, J. Bille and K. Jaton (2004). Detection of four Plasmodium species in blood from humans by 18S rRNA gene subunit-based and species-specific real-time PCR assays. *Journal of Clinical Microbiology* 42(12): 5636-5643.

203. Rozas, J., J.C. Sanchez-Del Barrio, X. Messeguer and R. Rozas (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19: 2496-2497.

204. Sabeti, P.C., S.F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T.S. Mikkelsen, D. Altshuler and E.S. Lander (2006). Positive natural selection in the human lineage. *Science* 312(5780): 1614-1620.

205. Salzburger, W., S. Baric and C. Sturmbauer (2002). Speciation via introgressive hybridization in East African cichlids? *Molecular Ecology* 11(3): 619-625.

206. Samuels, A. and J. Altmann (1986). Immigration of a *Papio anubis* male into a group of *Papio cynocephalus* baboons and evidence for an Anubis-Cynocephalus hybrid zone in Amboseli, Kenya. *International Journal of Primatology* 7(2): 131-138.

207. Samuels, A. and J. Altmann (1991). Baboons of the Amboseli Basin - demographic stability and change. *International Journal of Primatology* 12(1): 1-19.

208. Schaart, J.G., L. Mehli and H.J. Schouten (2005). Quantification of allele-specific expression of a gene encoding strawberry polygalacturonase-inhibiting protein (PGIP) using Pyrosequencing. *Plant Journal* 41(3): 493-500.

209. Seixas, S., N. Ferrand and J. Rocha (2002). Microsatellite variation and evolution of the human Duffy blood group polymorphism. *Molecular Biology and Evolution* 19(10): 1802-1806.

210. Serre, D., S. Gurd, B. Ge, R. Sladek, D. Sinnett, E. Harmsen, M. Bibikova, E. Chudin, D.L. Barker, T. Dickinson, *et al.* (2008). Differential allelic expression in

the human genome: a robust approach to identify genetic and epigenetic *cis*-acting mechanisms regulating gene expression. *PLoS Genetics* 4: e1000006.

211. Shapiro, M.D., M.E. Marks, C.L. Peichel, B.K. Blackman, K.S. Neregn, B. Jonsson, D. Schluter and D.M. Kingsley (2004). Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* 428: 717-723.

212. Shook, D.R. and T.E. Johnson (1999). Quantitative trait loci affecting survival and fertility-related traits in Caenorhabditis elegans show genotype-environment interactions, pleiotropy and epistasis. *Genetics* 153(3): 1233-1243.

213. Siepel, A. (2009). Phylogenomics of primates and their ancestral populations. *Genome Research* 19(11): 1929-1941.

214. Silk, J.B., S.C. Alberts and J. Altmann (2003). Social bonds of female baboons enhance infant survival. *Science* 302: 1231-1234.

215. Smit, A.F.A., R. Hubley and P. Green. (1996-2004). "RepeatMasker Open-3.0." from www.repeatmasker.org.

216. Smith, E.N. and L. Kruglyak (2008). Gene-environment interaction in yeast gene expression. *PLoS Biology* 6: e83.

217. St Clair, D., M.Q. Xu, P. Wang, Y.Q. Yu, Y.R. Fang, F. Zhang, X.Y. Zheng, N.F. Gu, G.Y. Feng, P. Sham, *et al.* (2005). Rates of adult schizophrenia following prenatal exposure to the Chinese famine of 1959-1961. *Journal of the American Medical Association* 294(5): 557-562.

218. St George, D., S.M. Witte, T.R. Turner, M.L. Weiss, J. Phillips-Conroy, E.O. Smith and J. Rogers (1998). Microsatellite variation in two populations of free-ranging yellow baboons (Papio hamadryas cynocephalus). *International Journal of Primatology* 19(2): 273-285.

219. Steiner, C.C., J.N. Weber and H.E. Hoekstra (2007). Adaptive variation in beach mice produced by two interacting pigmentation genes. *PLoS Biol* 5(9): e219.

220. Stephens, M., N.J. Smith and P. Donnelly (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* 68(4): 978-989.

221. Stevison, L.S. and M.H. Kohn (2009). Divergence population genetic analysis of hybridization between rhesus and cynomolgus macaques. *Molecular Ecology* 18(11): 2457-2475.

222. Storz, J.F., M.A. Beaumont and S.C. Alberts (2002). Genetic evidence for long-term population decline in a savannah-dwelling primate: Inferences from a hierarchical Bayesian model. *Molecular Biology and Evolution* 19(11): 1981-1990.

223. Storz, J.F., U. Ramakrishnan and S.C. Alberts (2002). Genetic effective size of a wild primate population: Influence of current and historical demography. *Evolution* 56(4): 817-829.

224. Su, A.I., T. Wiltshire, S. Batalov, H. Lapp, K.A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, *et al.* (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* 101(16): 6062-6067.

225. Sugg, D.W., R.K. Chesser, F.S. Dobson and J.L. Hoogland (1996). Population genetics meets behavioral ecology. *Trends in Ecology & Evolution* 11(8): 338-342.

226. Surridge, A.K. and N.I. Mundy (2002). Trans-specific evolution of opsin alleles and the maintenance of trichromatic colour vision in Callitrichine primates. *Molecular Ecology* 11(10): 2157-2169.

227. Surridge, A.K., D. Osorio and N.I. Mundy (2003). Evolution and selection of trichromatic vision in primates. *Trends in Ecology & Evolution* 18(4): 198-205.

228. Tan, Z., J. Fan, A.M. Shon, M.G. Schwartz, B. Camoretti-Mercado and C. Ober (2006). A polymorphism in the HLA-G 3'-UTR influences targeting of mir-148 and is associated with asthma. *Journal of Allergy and Clinical Immunology* 117(2): S141-S141.

229. Tao, H., D.R. Cox and K.A. Frazer (2006). Allele-specific *KRT1* expression is a complex trait. *PLoS Genetics* 2(6): 848-858.

230. Team, R.D.C. (2007). R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing.

231. Thalmann, O.H., A.H. Fischer, F.H. Lankester, S.H. Paabo and L.H. Vigilant (2007). The complex evolutionary history of gorillas: Insights from genomic data. *Molecular Biology and Evolution* 24(1): 146-158.

232. Tishkoff, S.A., F.A. Reed, A. Ranciaro, B.F. Voight, C.C. Babbitt, J.S. Silverman, K. Powell, H.M. Mortensen, J.B. Hirbo, M. Osman, *et al.* (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics* 39(1): 31-40.

233. Tosi, A.J., J.C. Morales and D.J. Melnick (2000). Comparison of Y chromosome and mtDNA phylogenies leads to unique inferences of macaque evolutionary history. *Molecular Phylogenetics and Evolution* 17(2): 133-144.

234. Tournamille, C., Y. Colin, J. Cartron and C. Levankim (1995). Disruption of a GATA motif in the Duffy gene promoter abolishes erthyroid gene-expression in Duffy negative individuals. *Nature Genetics* 10 (2): 224-228.

235. Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R.B. Altman (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics* 17(6): 520-525.

236. Tung, J., M.J.E. Charpentier, D.A. Garfield, J. Altmann and S.C. Alberts (2008). Genetic evidence reveals temporal change in hybridization patterns in a wild baboon population. *Molecular Ecology* 17: 1998-2011.

237. Tung, J., O. Fedrigo, R. Haygood, S. Mukherjee and G.A. Wray (2009). Genomic features that predict allelic imbalance in humans suggest patterns of constraint on gene expression variation. *Molecular Biology and Evolution* 26(9): 2047-2059.

238. Tung, J., A. Primus, A.J. Bouley, T. Severson, S.C. Alberts and G.A. Wray (2009). Evolution of a malaria resistance gene in wild primates. *Nature* 460: 388-392.

239. Turner, L.M. and H.E. Hoekstra (2008). Reproductive protein evolution within and between species: maintenance of divergent ZP3 alleles in Peromyscus. *Molecular Ecology* 17(11): 2616-2628.

240. van Hoek, M., A. Dehghan, J.C. Witteman, C.M. van Duijn, A.G. Uitterlinden, B.A. Oostra, A. Hofman, E.J. Sijbrands and A.C. Janssens (2008). Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study. *Diabetes* 57(11): 3122-3128.

241. Van Horn, R.C., J.C. Buchan, J. Altmann and S.C. Alberts (2007). Divided destinies: group choice by female savannah baboons during group fission. *Behavioral Ecology and Sociobiology* 61: 1823-1837.

242. Verardi, A., V. Lucchini and E. Randi (2006). Detecting introgressive hybridization between free-ranging domestic dogs and wild wolves (*Canis lupus*) by admixture linkage disequilibrium analysis. *Molecular Ecology* 15(10): 2845-2855.

243. Verlaan, D.J., B. Ge, E. Grundberg, R. Hoberman, K.C.L. Lam, V. Koka, J. Dias, S. Gurd, N.W. Martin, H. Mallmin, *et al.* (2009). Targeted screening of *cis*-regulatory variation in human haplotypes. *Genome Research* 19: 118-127.

244. Veyrieras, J.B., S. Kudaravalli, S.Y. Kim, E.T. Dermitzakis, Y. Gilad, M. Stephens and J.K. Pritchard (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet* 4(10): e1000214.

245. Vigilant, L. (2009). Elucidating population histories using genomic DNA sequences. *Current Anthropology* 50(2): 201-212.

246. Vigilant, L. and K. Guschanski (2009). Using genetics to understand the dynamics of wild primate populations. *Primates* 50(2): 105-120.

247. Vigilant, L., M. Hofreiter, H. Siedel and C. Boesch (2001). Paternity and relatedness in wild chimpanzee communities. *Proceedings of the National Academy of Sciences of the United States of America* 98(23): 12890-12895.

248. Vogel, E.R., M. Neitz and N.J. Dominy (2007). Effect of color vision phenotype on the foraging of wild white-faced capuchins, *Cebus capucinus*. *Behavioral Ecology* 18(2): 292-297.

249. von Korff, M., S. Radovic, W. Choumane, K. Stamati, S.M. Udupa, S. Grando, S. Ceccarelli, I. Mackay, W. Powell, M. Baum, *et al.* (2009). Asymmetric allele-specific expression in relation to developmental variation and drought stress in barley hybrids. *Plant Journal* 59(1): 14-26.

250. Wang, J., J.F. Robinson, H.M. Khan, D.E. Carter, J. McKinney, B.A. Miskie and R.A. Hegele (2004). Optimizing RNA extraction yield from whole blood for microarray gene expression analysis. *Clinical Biochemistry* 37(9): 741-744.

251. Wang, Z., H.F. Willard, S. Mukherjee and T.S. Furey (2006). Evidence of influence of genomic DNA sequence on human X chromosome inactivation. *Plos Computational Biology* 2(9): 979-988.

252. Weaver, I.C.G., N. Cervoni, F.A. Champagne, A.C. D'Alessio, S. Sharma, J.R. Seckl, S. Dymov, M. Szyf and M.J. Meaney (2004). Epigenetic programming by maternal behavior. *Nature Neuroscience* 7(8): 847-854.

253. Wessa, P. (2008). Free Statistics Software, Office for Research Development and Education.

254. West, M., C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J.A. Olson, J.R. Marks and J.R. Nevins (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 98(20): 11462-11467.

255. Whitfield, C.W., A.M. Cziko and G.E. Robinson (2003). Gene expression profiles in the brain predict behavior in individual honey bees. *Science* 302(5643): 296-299.

256. Wildman, D.E., T.J. Bergman, A. al-Aghbari, K.N. Sterner, T.K. Newman, J.E. Phillips-Conroy, C.J. Jolly and T.R. Disotell (2004). Mitochondrial evidence for the origin of hamadryas baboons. *Molecular Phylogenetics and Evolution* 32(1): 287-296.

257. Wilson, A.J., J.M. Pemberton, J.G. Pilkington, D.W. Coltman, D.V. Mifsud, T.H. Clutton-Brock and L.E. Kruuk (2006). Environmental coupling of selection and heritability limits evolution. *PLoS Biol* 4(7): e216.

258. Wittkopp, P., B. Haerum and A. Clark (2004). Evolutionary changes in cis and trans gene regulation. *Nature* 430(6995): 85-88.

259. Wittkopp, P., B. Haerum and A.G. Clark (2008). Regulatory changes underlying expression differences within and between *Drosophila* species. *Nature Genetics* 40: 346-350.

260. Wittkopp, P.J., E.E. Stewart, L.L. Arnold, A.H. Neidert, B.K. Haerum, E.M. Thompson, S. Akhras, G. Smith-Winberry and L. Shefner (2009). Intraspecific polymorphism to interspecific divergence: genetics of pigmentation in Drosophila. *Science* 326(5952): 540-544.

261. Won, Y.J. and J. Hey (2005). Divergence population genetics of chimpanzees. *Molecular Biology and Evolution* 22(2): 297-307.

262. Wooding, S., B. Bufe, C. Grassi, M.T. Howard, A.C. Stone, M. Vazquez, D.M. Dunn, W. Meyerhof, R.B. Weiss and M.J. Bamshad (2006). Independent evolution of bitter-taste sensitivity in humans and chimpanzees. *Nature* 440(7086): 930-934.

263. Wray, G.A. (2007). The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics* 8(3): 206-216.

264. Wray, G.A., M.W. Hahn, E. Abouheif, J.P. Balhoff, M. Pizer, M.V. Rockman and L.A. Romano (2003). The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution* 20(9): 1377-1419.

265. Wright, S. (1946). Isolation by distance under diverse systems of mating. *Genetics* 31: 39-59.

266. Yan, H., W. Yuan, V.E. Velculescu, B. Vogelstein and K.W. Kinzler (2002). Allelic variation in human gene expression. *Science* 297(5584): 1143.

267. Yu, N., M.I. Jensen-Seaman, L. Chemnick, J.R. Kidd, A.S. Deinard, O. Ryder, K.K. Kidd and W.H. Li (2003). Low nucleotide diversity in chimpanzees and bonobos. *Genetics* 164(4): 1511-1518.

268. Yu, N., M.I. Jensen-Seaman, L. Chemnick, O. Ryder and W.H. Li (2004). Nucleotide diversity in gorillas. *Genetics* 166(3): 1375-1383.

269. Zhang, X. and J.O. Borevitz (2009). Global analysis of allele-specific expression in Arabidopsis thaliana. *Genetics* 182(4): 943-954.

270. Zhang, X.H.F., K.A. Heller, L. Hefter, C.S. Leslie and L.A. Chasin (2003). Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. *Genome Research* 13(12): 2637-2650.

271. Zhu, C.Y., J. Odeberg, A. Hamsten and P. Eriksson (2006). Allele-specific *MMP-3* transcription under in vivo conditions. *Biochemical and Biophysical Research Communications* 348(3): 1150-1156.

272. Zimmerman, P.A., I. Woolley, G.L. Masinde, S.M. Miller, D.T. McNamara, F. Hazlett, C.S. Mgone, M.P. Alpers, B. Genton, B.A. Boatin, *et al.* (1999). Emergence of FY*A(null) in a *Plasmodium vivax*-endemic region of Papua New Guinea. *Proceedings of the National Academy of Sciences of the United States of America* 96(24): 13973-13977.

273. Zinner, D., M.L. Arnold and C. Roos (2009). Is the new primate genus *Rungwecebus* a baboon? *PLoS One* 4(3): e4859.

274. Zinner, D., L.F. Groeneveld, C. Keller and C. Roos (2009). Mitochondrial phylogeography of baboons (Papio spp.) - Indication for introgressive hybridization? *BMC Evolutionary Biology* 9: 83.

Biography

Jenny Tung was born on January 13, 1982 in Seaford, Delaware, where she spent the majority of her childhood. She graduated from high school in Marietta, Georgia, and completed her undergraduate Bachelor of Science degree in Biology from Duke University in 2003. She spent a year in Raleighvallen National Park, Suriname, observing brown capuchin monkeys, before returning to Duke for her graduate work.

During graduate school, Jenny has received the following fellowships: a National Science Foundation Graduate Research Fellowship, a James B. Duke Graduate Fellowship, a Katherine Goodman Stern Dissertation Year Fellowship, and a Duke University Primate Genomics Initiative Graduate Student Fellowship. She has been the recipient of grant funding from the Duke University Graduate School, the Duke University chapter of Sigma Xi, the National Science Foundation, the American Society of Primatologists, the Society for Molecular Biology and Evolution, and the Patricia William Mwangaza Foundation.

She has also been an author on the following scientific articles: "Parallel effects of genetic variation on ACE activity in baboons and humans" (American Journal of Physical Anthropology 134: 1); "Age at maturity in wild baboons: genetic, demographic, and environmental influences" (Molecular Ecology 17: 2026); "Genetic evidence reveals dynamic patterns of hybridization in a wild baboon population" (Molecular Ecology 17: 1998); "Seeing red: behavioral evidence of trichromatic color vision in strepsirrhine primates" (Behavioral Ecology 20: 1); "Evolution of a malaria resistance gene in wild primates" (Nature 460: 388); "Genomic features that predict allelic imbalance in humans suggest patterns of constraint on gene expression variation" (Molecular Biology and Evolution 26: 2047); and "Evolution of traits deduced from genome comparisons" (The Encyclopedia of Life Sciences, doi:10.1002/9780470015902).