Adaptive History of the Chimpanzee Subspecies in the Genomic Era

Jessica Nye

TESI DOCTORAL UPF 2018

Directors de la Tesi

Jaume Bertranpetit i Hafid Laayouni

CIÈNCIES EXPERIMENTALS I DE LA SALUT

upf. Universitat Pompeu Fabra Barcelona

Acknowledgements

I would like to thank my advisors Dr. Jaume Bertranpetit and Dr. Hafid Laayouni for guiding me during my doctoral research. I would like to thank my family and friends for supporting me during my many years of education. To my fellow lab mates, I am grateful for the theoretical discussions we had over the last four years.

This work was supported by FI Agaur grant FI-DGR 2015.

Abstract

Chimpanzees are our closest living genetic cousins. The species consists of four subspecies, each with a unique demographic history. In order to better understand their species, a detailed map of signatures of selection will highlight important genetic differences between the four subspecies. Here, we interrogate the genome using 15 tests of selection. We simulate neutral and selective scenarios taking their unique demographic history into account in the model. We combine all these elements using a machine learning approach to highlight regions of interest in each subspecies. We specifically investigate the regions of the genome that have been selected after introgression with bonobo. We investigate the haplotype changes that have occurred since the divergence with humans in a few specially selected genes. From this, we find that the evolutionary history of each of the four subspecies is unique. That subtle differences in their demographic history and environment have greatly shaped their genetic diversity. In general, we observe signatures in selection in phenotypes involving immunity, muscle function, reproduction, and DNA repair. This dissertation highlights the important genes and regions that have resulted in the divergence between subspecies of chimpanzee.

Resum

Els ximpanzés són els nostres cosins genètics més propers. L'espècie consta de quatre subespècies, cadascuna amb una història demogràfica única. Per comprendre millor la diversitat dels ximpanzés hem elaborat un mapa detallat d'empremtes de selecció presents en el genoma que ressaltarà importants diferències evolutives entre les quatre subespècies. En aquest treball, interroguem el genoma amb 15 tests de selecció diferents i simulem escenaris neutres i selectius tenint en compte la història demogràfica única de cada subespècie. A través d'un procés de *machine learning*, hem combinat els resultats de tots aquests tests utilitzant un enfocament d'aprenentatge automàtic per reconèixer regions d'interès en cada subespècie, que són les que mostren clarament els rastres de la selecció positiva. Hem investigat específicament les regions del genoma que han estat seleccionades després de la introgressió amb el bonobo, donat que la selecció després de la introgressió és una àrea nova molt interessant. Pel conjunt del genoma, hem investigat els canvis que s'han produït des de la divergència amb els humans en gens especialment sotmesos a selecció positiva, obtenint un panorama de general adaptatiu en cada subespècie. A partir d'aquí, trobem que la història evolutiva de cadascuna de les quatre subespècies és única. Les subtils diferències entre subespècies en la seva història demogràfica i el seu entorn han donat forma a la seva diversitat genètica adaptativa. En general, observem empremtes de selecció en gens que informen fenotips que inclouen immunitat, funció muscular, reproducció i reparació d'ADN. Aquesta tesi destaca els gens i regions importants que han donat lloc a la divergència adaptativa entre subespècies de ximpanzé.

Preface

Humans have had an interest in the great apes for many decades. However, despite decades long observational studies, their genetic diversity is only just beginning to be understood. Due to the lack of a fossil record and difficult logistics of sample collection, few genome-wide multi-species studies exist in the great apes.

Furthermore, the computational power needed to adequately compare such large datasets have only become readily available in recent decades. Simulations of a neutral expectation are needed so that observations can be more precisely calculated. While it has been possible to simulate simple scenarios, only recently possible to reasonably simulate an accurate demographic history, or to implement theoretical selective scenarios in which to compare to real data. The increase in computational power has also led to the development of machine learning approaches. A machine learning approach allows for the combination of any number of characteristics in order to draw a single conclusion.

This doctoral dissertation takes advantage of a newly available genetic data consisting of genome-wide sequences of chimpanzees and bonobos. It takes advantage of available computational power to simulate differing scenarios of evolutionary history of the chimpanzee. It combines these two aspects using a machine learning approach in order to mine the genome. Without these recent advances in both genetic data collection and computational tools, this type of study would be impossible.

List of Publications

Nye J, Laayouni H, Kuhlwilm M, Mondal M, Marques-Bonet T, Bertranpetit J. Selection in the Introgressed Regions of the Chimpanzee Genome. Genome Biology and Evolution. 2018. 10(4):1132-1138

Nye J, Mondal M, Bertranpetit J, Laayouni H. A Fully Integrated Scan of Selection in the Chimpanzee Genome. Journal of Evolutionary Biology. Submitted.

Villegas P, Nye J, Bertranpetit J, Laayouni H. Screening for positive selection signals on the X chromosome in human populations. Genome Biology and Evolution. Submitted.

Table of Contents

1. Introduction

1.1 A Brief History of Population Genetics

When Charles Darwin published his theory of Evolution by Natural Selection in 1859, he posited that if a variable trait were heritable and some version of the trait donates an advantage, then the beneficial copy will increase in frequency over time. When it was published, no one was aware of the source of a trait's variability. After the rediscovery of Mendel's 1866 work on heritability in pea plants, research into discrete heritable units began at the turn of the 20th century.

The field of genetics and genomics has advanced greatly in the past hundred years, to the point where we understand much of the coding variation in a genome. Variation in a genome is understood to be from three major sources. Single nucleotide polymorphisms (SNPs) are the changes to the genetic code due to mutations. Next, small insertions and deletions or microsatellites may donate small neighborhood changes. Lastly, larger sources of variability can come from chromosomal inversions and copy number variation which can have larger impacts on entire regions of a chromosome. There have been many studies into untangling the biological impact of all these types of variation within coding regions; however, a large gap in our knowledge of genetics is into the impact of variation in noncoding regions.

In modern population genetics, more variables have been added to Darwin's theory. The Hardy-Weinberg equation was published separately in 1908, and demonstrates that allele frequencies will remain at a constant equilibrium over time in the absence of evolutionary forces. While this equation still is much used to this day, it contains several unrealistic assumptions, such as random mating or infinite population size.

The next major iteration in population genetic theory came when Fisher in 1930 and Wright in 1931 published models that included finite population sizes, allowing the evolutionary force of drift to be defined. According to what is now known as the Wright-Fisher model, the size of the population is directly proportional to the chance that an allele will become fixed or disappear from the population, purely due to sampling. Later, in 1955 and 1956, Kimura added the two-allele case to the Wright-Fisher model and posited, in 1983, that almost all variation in the genome is neutral; producing what is now known as the Neutral Theory of Evolution. Finally, in 1973 Ohta published the Nearly Neutral Theory of Evolution. Her theory extends on the neutral theory and states that the most powerful evolutionary force is genetic drift, and that nearly all mutational changes to a genome will be deleterious, and consequently will be removed from the genome. Meaning that beneficial mutations, and therefore positive selection, are exceedingly rare.

We can define three major types of selective forces. Negative selection (also called background, purifying, or stabilizing) is the most common type of selection. As predicted by Ohta, new mutations are likely to be harmful, and as a result selection will try to remove the deleterious allele from the population. This type of selection is powerless when the deleterious allele is recessive because only homozygous individuals will suffer a fitness cost, while heterozygotes can harbor the deleterious allele allowing it to remain in the population.

Balancing selection (also called overdominance) serves to keep diversity in a population. This can occur in one of several ways. In the case of fluctuating selection, made famous by Darwin himself, an oscillating environment can lead to a drastic change of the fitness maxima from one generation to the next (Haldane and Jayakar 1963). Frequency dependent selection results when fitness is contingent on its concentration within a population (Pritchard and

Di Rienzo 2010). Frequency dependent selection can favor either novel (negative frequency dependence) or common (positive frequency dependence) haplotypes. Finally, heterozygote advantage occurs when heterozygotes are more fit than either the homozygous individuals, forcing both copies of alleles to remain present at intermediate frequencies (Borzan et al. 2014).

Positive selection (also called adaptive or Darwinian) causes an advantageous allele to increase in frequency. This type of selection serves to increase beneficial alleles and decreases diversity around the selected allele. Positive selection is the most commonly studied selective force. This is due to the patterns it leaves in the genome, making it the simplest to study. When a de novo beneficial allele is selected, it raises in frequency within the population (Figure 1.1B). The derived alleles that surround this beneficial allele also rise in frequency with a phenomenon called genetic hitchhiking (Maynard-Smith and Haigh 1974). This happens because recombination is not able to break up the beneficial allele from its close neighbors, causing them to be genetically linked, forming a haplotype (Figure 1.1C). Over time, after the beneficial allele reaches fixation, selection will relax, allowing new singletons to appear and for recombination to break up the haplotype. Differences between population frequencies will differ between populations that have had a selection event, and a population that did not (Figure 1.1D). Based on these patterns of allele frequencies, haplotype blocks, and population differences, many tests for selective sweeps have been formulated.

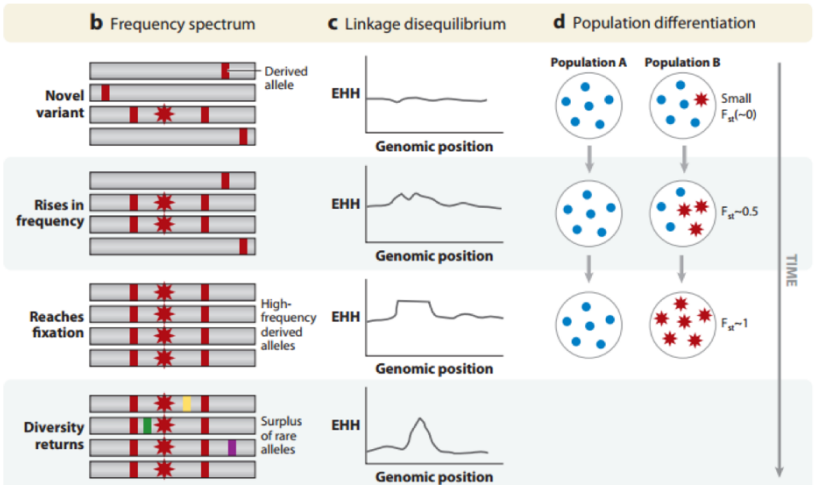


Figure 1.1: From Vitti et al. 2013, the three patterns used to detect a positive selective sweep. B) The frequency of the beneficial allele (star) will rise in frequency along with hitchhiking derived alleles. C) A haplotype will form around the selected allele. D) Differences between frequencies will be observed between populations.

1.2 Tests of Selection

a) Site frequency spectrum-based tests

The site frequency spectrum (SFS) is simply the distribution of all counts of alleles in the genome, or a given set of loci. First SNPs are separated by their state (i.e., whether they are ancestral or derived, reference or non-reference). Under neutral conditions, a large proportion of derived singletons are expected, with an exponential decay toward very few derived alleles at high frequency (Figure 1.2A). However, when selection drives derived alleles to fixation, an increased number of derived alleles can be observed at high frequencies (Figure 1.2B). Here, I discuss four different tests for a selective sweep based on the principles of SFS.

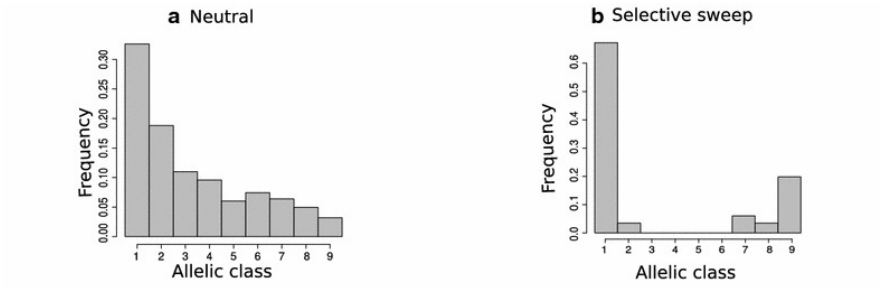


Figure 1.2: From Pavlos and Alachiotis 2017. A) The expected site frequency spectrum under a neutral scenario. B) Site frequency spectrum deviation expected under a selective sweep.

Tajima's D is the classical test for a selective sweep based on SFS. It was first published in 1989, and is simple to calculate but not to interpret. It is the standardized difference between the average number of pairwise differences in a sequence and its number of segregating sites. Under neutrality, the value is expected to equal zero. If the value is above zero this indicates that there are more haplotypes than you expect in the sample. This may be due to either balancing selection or to a recent population bottleneck. If the value of D is below zero this indicates there are fewer haplotypes than expected. This can be a byproduct of either a selective sweep or a population expansion.

Fu and Li formulated two statistics D and F (1993) which are an expansion of Tajima's D. Both of their statistics are deviations and have essentially the same interpretation as Tajima's D. Fu and Li's D is the deviation between derived segregating sites and derived singletons; while Fu and Li's F is the deviation between the mean pairwise difference and derived singletons. One difficulty in interpreting these three statistics, is determining what value is a significant deviation. Because all these statistics can be skewed by changes in population demography over time, it is important to measure the entire genome before conclusions are drawn from them.

Lastly, R_2 is a more recently derived statistic (Ramos-Onsins and Rozas 2002) that was motivated to avoid the problems that Tajima's and Fu and Li's statistics have with demography by using the SFS to detect signatures of past population growth. This is done by calculating the difference of singleton number and the average nucleotide differences. Calculating this will assess whether or not the region of interest has the correct proportion of singletons as expected by neutrality.

b) Linkage disequilibrium-based tests

Linkage was first described by Morgan in 1915, when he discovered that eye color seemed to be correlated with the sex of fruit flies. Today we now understand that recombination between sister chromatids is an important driver of genetic diversity by allowing for novel combinations of haplotypes to occur in an individual. As discussed in the section above, during a selective sweep, a single haplotype is usually found around a selected allele. The block of linkage disequilibrium (LD) will only decay with time, and as selection is relaxed, recombination can break up the haplotype block (Figure 1.1C). I will discuss eight statistics based on LD.

First, the decay of the haplotype is calculated as the extended haplotype homozygosity (EHH). To calculate EHH, at each focal SNP, the probability is calculated that the SNPs on either side of the focal point are homozygous (identical by state), and therefore a member of the same haplotype block (Sabeti et al. 2002). The EHH average is a weighted average for all core haplotypes. An extension of EHH is the test integrated haplotype score (iHS), which is the integral of EHH (Voight et al. 2006). Values deviating from zero indicate selection. Each of these tests are useful only for ongoing selective sweeps, because they rely on defining the focal SNP, if the

beneficial allele has reached fixation, however, it is no longer a SNP in the population.

Fu's F assumes a certain amount of diversity under the infinite sites model of mutations (Kimura 1969). Based on this assumption, Fu's F calculates the probability of finding the observed diversity under neutral conditions. The interpretation is similar to Tajima's D, but Fu posits that this statistic is more sensitive to genetic hitchhiking than Tajima's D (Fu 1997).

At the same time that Fu published his F statistic, Kelly (1997) proposed the Z_{nS} statistic which calculates that average allelic correlation coefficient over all pairwise comparisons in a fixed region of the genome. An extension of Kelly's statistic is Z_A (Rozas et al. 2001) which compares the same average allelic correlation with only adjacent pairwise comparisons in a fixed region. Rozas in the same paper compared the two Z statistics to formulate ZZ which is the difference between the two (Rozas et al. 2001). This family of statistics are powerful when the selection event in question is older ~50,000 years ago (Ramírez-Soriano et al. 2008).

The last two LD-based statistics discussed here were formulated by Wall. The first is Wall's B (1999) which counts the number of adjacent segregating sites that are from the same haplotype. While Wall's Q (2000) adds the number of pairs which aren't from the same haplotype to Wall's B. The motivation in developing these statistics was to deal with confounding results due to population substructure and balancing selection present in the demographic history of the organism.

c) Population differentiation-based tests

A third way to search for signatures of selection is to use either SFS or LD-based statistics, but to compare measurements not to a neutral expectation, but to another population. Here, I will discuss two statistics, one based on SFS and the other on LD.

First, ΔDAF is a comparison of the derived allele frequencies between two populations. Where large differences appear, this can be evidence that selection drove derived alleles in the genomic region to high frequencies (Hofer et al. 2009).

Second, XP-EHH is the comparison of the EHH in the same region between two populations. If one population has a large haplotype that is absent in another, this is evidence of selection somewhere in that haplotype block (Sabeti et al. 2007).

1.3 Demography

a) Population history

As mentioned above, a population's demographic history can greatly impact the interpretation of statistical outcomes; making it imperative in any adequate scan of a genome to account for these factors. Each important parameter is discussed below.

The first principle is the effective population size (N_e). This was first defined by Wright in 1931. It is not the census number of a population, but the proportionate amount of breeding individuals that would result in the same population diversity and inbreeding. The effective population is an indicator of how strong drift will act on the population, the smaller it is, the stronger the force of drift will be. A recent study has demonstrated that the number of selective sweeps in a population is proportional to N_e (Nam et al.

2017), meaning that populations with larger a N_e will have more signals of selection in their genome.

Next, drastic changes in population size, either expansions or contractions, will affect both the SFS and LD patterns. In the case of a population expansion, you would find an increase in the number of haplotypes, but that each haplotype is similar to others, and would cluster together in a star-like phylogeny. This is the case of the diversity observed among out of Africa humans. For a population bottleneck, you would expect to see both a low number of haplotypes and low nucleotide diversity. Both these changes are especially problematic when inferring selection with SFS-based tests, because demographic changes may be ill-attributed as selective events.

Gene flow between populations will introduce new diversity into a population. Migration and the exchange of genes can occur between populations of the same species, a subspecies, or a separate species. In all these cases the influx of new haplotypes to a population will increase the observed diversity. While these new haplotypes may be the key to acquiring new adaptations, and may undergo a selective sweep, the new genetic material could also resemble a modest bottleneck. Furthermore, gene flow between separate groups can make it difficult to estimate the time of divergence between the two populations.

Lastly, when estimating the demographic past of a population, great care must be taken into determining the mutation and recombination rates, and generation time. Mutations in the germ line are considered to be the major driver of generating genetic diversity, therefore over or under estimation of this rate can lead to spurious results as to the age of a selective sweep, or of the divergence between populations (Fu and Huai 2003). Recombination rate is overall tricky because it is not constant across a genome. Although it is common to use an average recombination rate when modeling

demography, the rate in a hotspot can be as much as ten times the average rate of the genome (Kong et al. 2002; Jeffreys et al. 2005). Finally, the underlying population genetic theory on which much of the statistical repertoire is based assumes discrete generations. While this assumption is impractical for real scenarios, the modeling of the average age of reproduction of any species is part of demographic modeling and will impact the real-world estimates of population divergences, selective sweeps, and population sizes because shorter generation times lead to larger populations.

b) Simulating demography

A best-practice method when searching for signatures of selection is to model the specific demographic history of your population of interest. This allows more precise assessment of a statistic's outcome (Voight et al. 2006; Pickrell et al. 2009). Two major branches of simulations exist: the coalescent and forward simulators.

The coalescent model was first described by Kingman in 1982. He posited, based on the Wright-Fisher model, that all individuals of a population are related through their most recent common ancestor (MRCA) on a genealogical tree. If you sample a derived allele at present time, you can predict its frequency based on how many branches are related to the MRCA of that derived allele. After Kingman's theory was released, and with the help of computational advancement, today modern coalescent simulators are able to accommodate all the demographic parameters described above (Hudson 1983; Kaplan et al. 1988). The benefit of a simulator based on coalescence is that it is fast because it starts at present time and simulates to the coalescent event using trees instead of physically sampling the virtual chromosomes.

Forward simulators have only been developed recently because computational power has increased. The forward simulator starts at the initial time and physically samples at every generation to create the next generation (Slatkin 2001; Peng and Kimmel 2005). These simulators are ideal because they are able to implement any desired demographic and selective scenario, and all genealogical information is preserved. However, due to computation, only relatively small genetic fragments are able to be simulated and this method is often prohibitively slow.

A composite simulator uses both the coalescent and the forward simulator. This method relies on the coalescent to do the heavy lifting, and only utilizes the forward simulator when a complex demographic problem, like when a selective event needs to be implemented. The composite simulator selects the benefits from both methods while avoiding the problems of either computation or the inability to implement complex scenarios (Ewing and Hermission 2010).

1.4 Machine Learning

Not only have forward simulators been made possible as the computational power of the average computer has risen, but so have methods that allow the computer to derive patterns from complex collections of data. Machine learning methods make it possible to combine data with different statistical distributions and differing levels of precision in order to draw a single conclusion on a suite of traits. Machine learning started with decision trees. More recently, complex subsampling methods have improved the accuracy of machine learning. Still today new methods are being developed, like Bayesian expansions or neural networks.

Decision trees are a hierarchical table that differentiates between groups, by first learning from a training set, then confirming its accuracy with an evaluation set, and lastly, classifying unknown data (Kass 1980; Breiman et al. 1984). This type of machine learning algorithm is prone to several problems. First, it runs the risk of becoming over parameterized if many of the input parameters are correlated. To avoid this, the user must decide which parameters to feed to the algorithm. Second, the structure of the table needs to be provided. In other words, the user must decide in which order groups should be discerned from each other. This can be problematic, because one of the theoretical benefits of using a machine learning approach is to let data indicate which characteristics best differentiate between groups, which are not always obvious to the user.

An improvement on decision trees is the method of bagging. This is an ensemble method that uses bootstrap sampling to improve the fit of a regression model on tree data (Krzywinski and Altman 2017). Essentially, the training set is sampled with replacement and a regression model is recalculated at each iteration. Depending on how noisy the data are, after a sufficient number of sampling events, an optimal regression model is built from the data. The validation set is not a separate set of samples, but the samples that were left out of the bag (Breiman 1996).

The random forest algorithm enhances the variable selection of the bagging method. Here, random trees are formed from the training set by constructing nodes of trees with only a random subset of the provided variables. By selecting only partial data, the variability between trees is maximized, avoiding the likelihood of over parameterization that can be problematic for tree-based approaches. After a large random forest is grown, bagging progresses as normal (Breiman 2001).

Overall, the goal of this thesis is to interrogate the selective past of the chimpanzee species. I utilize a diverse array of statistics based on population genetics theory, so that the strengths of each test may be combined. The computational power that is currently available can allow for a project that involves calculations of many statistics genome-wide on real data, simulations of neutral data (accounting for demography), and simulations of a multitude of selective scenarios. Using all these data, a random forest algorithm highlights the regions that likely had selective events in the genomes of the four subspecies of chimpanzee. I then investigate the specific differences within regions of the genome that were introgressed from bonobo.

1.5 Chimpanzees

Apes are our closest living genetic cousins, and as such we have had a long-standing interest in studying and understanding their proclivities. In depth field research began with the chimpanzees by Jane Goodall in 1960 and with the gorillas by Dian Fossey in 1966. Despite more than a half century of research, many phenotypes exhibited in the wild are still unknown, and the genetic diversity of these creatures is only beginning to be elucidated.

Data from the great ape genetic diversity panel has indicated that the great apes began to diverge around 10 million years ago (mya), when orangutans split first (Figure 1.3). Some 5 million years later, the gorillas diverged, followed by humans splitting from bonobo and chimpanzee around 3 mya (Prado-Martinez et al. 2013). The branch that contains bonobos and chimpanzee has been difficult to discern. This is due in part to a lack of samples or fossil record, but mainly due to the large amount of gene flow between populations.

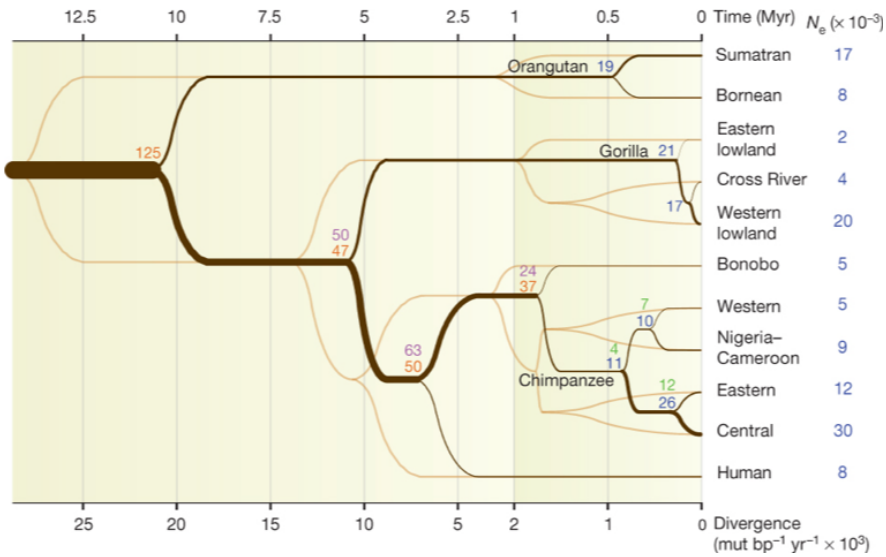


Figure 1.3: From Prado-Martinez et al. 2013. The demographic history of the great apes featuring split times and effective population size. The dark lines represent the split times while the light lines indicate the divergence time.

In order to classify bonobos and chimpanzees into distinct groups, researchers used taxonomy, to differentiate between bonobo and chimpanzee, and genetics, mainly mitochondrial, to differentiate between the chimpanzee subspecies (Coolidge 1933; Morin et al. 1992). However, before genetic data was abundant enough to adequately elucidate the fine details that separate the chimpanzees, their geographic landscape helped to define them (Figure 1.4).

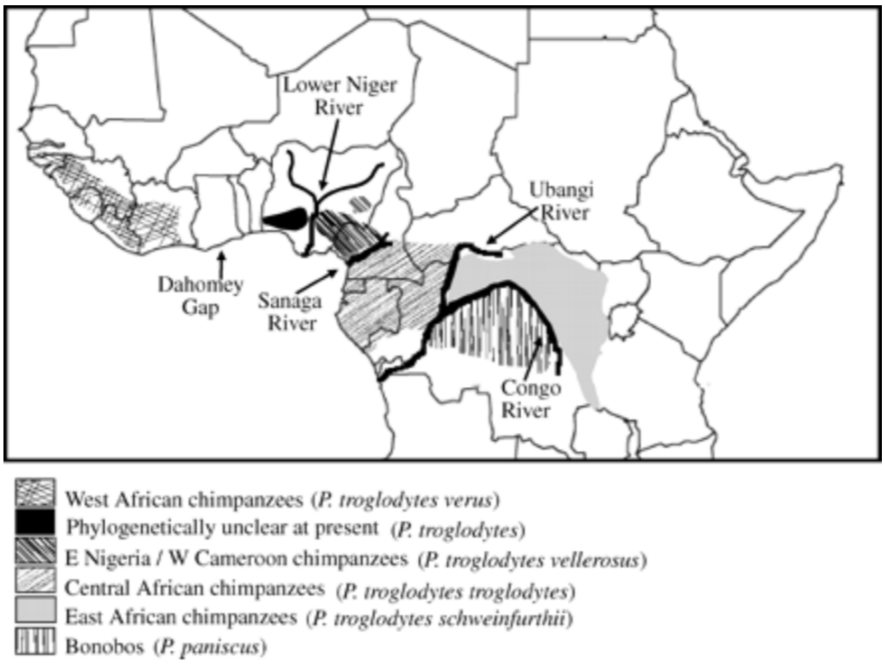


Figure 1.4: From Won and Hey 2005. Distribution of bonobo and chimpanzee across Africa.

It is clear from observations in the field that large bodies of water serve as barriers to migration for these apes. The bonobos (*Pan paniscus*) live south of the Congo River in the Democratic Republic of Congo (DRC). The chimpanzees live in both western and central Africa. For the population in the west, the large expanse of inhospitable habitat between western and central Africa, known as the Dahomey Gap, has likely led to the genetic isolation of this population. For the chimpanzees in central Africa, two additional rivers are likely boarders to migration. The most eastern population lives north of the Congo River and south of the Ubangi River. The population in the center of Africa is bounded by the Ubangi to the east and the Sanaga River to the north. These three populations were officially recognized as distinct subspecies in 2001 (Groves), specifically the western chimpanzee is called *Pan troglodytes verus*, eastern chimpanzee is *P.t. schweinfurthii*, and central chimpanzee is *P.t. troglodytes*. However, chimpanzees are also known to live

north of the Sanaga River, although the geography suggested this may be a distinct subspecies from the rest, no phenotypic characteristics seemed to justify its separation. After much debate (Gondor et al. 1997; Pilbrow 2006), in 2006 a fourth subspecies was defined through genetics, now called *P.t. ellioti*.

Still today, there is much debate over what the definition of a species really is (Hey 2006). The most widely accepted definition was proposed by Ernst Mayr in 1963 which defines a species by a population of interbreeding organisms that are reproductively isolated from one another. The isolation time required to cause reproductive isolation is around 9-12N_e generations (Hudson and Coyne 2002). If a population is still sharing genes, it is difficult, by this definition to call it a separate species. It also makes discerning its demographic history much more difficult. The sharing of genetic information would suggest that their divergence was much closer in time than it may actually be (Won and Hey 2005).

Using the data from the great ape genetic diversity panel, the demographic history of the *Pan* clade has become more clear (Figure 1.5). Bonobo split from chimpanzee between 1.5 and 2 mya. The ancestors of *P.t. verus* and *P.t. ellioti* separated from *P.t. troglodytes* and *P.t. schweinfurthii* 500 thousand years ago (kya). *P.t. verus* and *P.t. ellioti* became their own subspecies around 250 kya, while *P.t. schweinfurthii* and *P.t. troglodytes* diverged more recently, a over 100 kya (de Manuel et al. 2016). Extensive gene flow is observed between chimpanzee subspecies, making the estimated split times highly dependent on the amount of gene flow accounted for in the model.

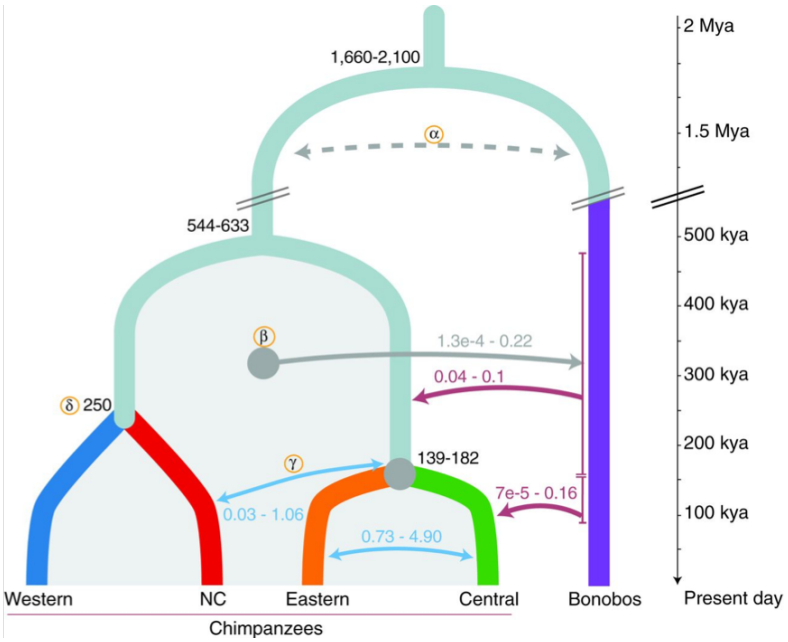


Figure 1.5: From de Manuel et al. 2016. The demography of the *Pan* clade that shows split times and gene flow between the five populations.

This high amount of genetic sharing is not entirely surprising, as it has well been established that chimpanzee females leave their birth communities. This behavior both maintains a higher genetic diversity and causes population substructure due to the fact that males are more closely related than females in each community (Morin et al. 1994). The spread of the female dispersal is also wide, with distances between 600 and 900 km. This fact is likely responsible for maintaining low morphological differentiation between groups (Shea and Coolidge 1988). It is also likely to be one of the sources of gene flow between subspecies, because it is part of the natural behavior for females to travel far outside their group to reproduce. Gene flow has also been predicted between chimpanzees and bonobos, in at least two waves during their evolutionary history; first with the ancestor of the *P.t. schweinfurthii* and *P.t. troglodytes* lineage, and a second wave with *P.t. troglodytes* (de Manuel et al. 2016).

The effective population size of the *Pan* clade has varied greatly throughout time. It is likely that the population was maximized around 3 mya after humans diverged. The population size is likely to have declined steadily, until the split of bonobo around 1 mya. After that split, the chimpanzee population size increased once again (Prado-Martinez et al. 2013). However, all population sizes have been subject to large bottlenecks in the last 100 thousand years, likely due to human destruction of habitat. The subspecies *P.t. troglodytes* has the highest effective population size among chimpanzees, followed by *P.t. schweinfurthii*, *P.t. ellioti*, and *P.t. verus* having the smallest, close to that of African humans. The amount of genetic diversity in each population follows the same pattern, which suggests that *P.t. verus* is more susceptible to the effects of genetic drift, while *P.t. troglodytes* will be more efficient at fixing beneficial alleles and removing deleterious ones (Charlesworth 2009).

Humans share around 98.7% (as reviewed by Kuhlwilm et al. 2016) of their genome with chimpanzees and bonobos. As such, we should understand much of the content of their genome. However, it is not excessively obvious how such little difference genetically could produce radically different beings. It is therefore imperative that we thoroughly interrogate these differences, which would lead to better understanding of our own genetic attributes.

There have been studies into natural selection in chimpanzee populations, especially in the context of contrasting with humans. These studies indicate that signatures of positive selection are more likely to be unique to individual subspecies. Specifically, a previous study found evidence of selection in genes related to immunity, neurological functions, diet, and anatomy (Cagan et al. 2016). When comparing to humans, it is clear that a high divergence of brain and testis gene expression is observed (Khaitovich et al. 2004; Khaitovich et al. 2005; Nowick et al. 2009). It is also apparent that ubiquitous genes are more similar between humans and

chimpanzees than specialized genes (Khaitovich et al. 2005). This suggests that specific adaptations have driven the differentiation of the two species, rather than an organism-wide change. Before in-depth studies into human-chimpanzee differentiation were possible, one hypothesis was that expression patterns were the driver of differentiation due to the fact that our genetic material is so similar. However, evidence suggests that in fact these two processes are evolving at similar rates (Khaitovich et al. 2004), in other words evolution does not favor changes to a gene over its regulation.

Besides the paper by Cagan (et al. 2016), most research into chimpanzee evolution has treated the species as a whole. However, this species is made up of four distinct subspecies which span differing habitats (*P.t. verus* lives in the arid savanna while the other three live primarily in the lush forests of central Africa). The variation of these habitats provides different access to quality and quantity of resources, differing exposures to pathogens, and likely driving different behaviors.

A study that used observations made between 2001 and 2006 showed that chimpanzees in southern Senegal (*P.t. verus*) shelter in caves during the hottest times of the year (Pruetz 2007). This behavior appears to be driven by temperatures, because they only were observed accessing the caves in May and June, and during the hottest times of day. Similar cave behaviors have been reported for baboons (Brain and Mitchell 1999; Barrett et al. 2004) and white-headed langurs (Huang et al. 2003), however to my knowledge this population is the only among the great apes, besides hominins, to exhibit temperature-related cave dwelling behaviors. Another striking behavioral difference observed in *P.t. verus* is their method of hunting with tools (Pruetz and Bertolani 2007). Although meat only makes up a small part of their diet, the female and juvenile chimpanzees in this subspecies hunt with the assistance of a spear-like tool (Pruetz et al. 2015). This behavior has not been observed in the other three subspecies (Pruetz and Bertolani 2007). These

observations are interesting because they are likely the result of an interplay between both behavior and available resources in their unique habitat.

An important difference between subspecies of chimpanzee is their exposure to pathogens through their environment. For instance, chimpanzees (*P.t. schweinfurthii*) have been observed in Gombe National Park, Tanzania to build fresh nests consisting of leaves and branches, which they sleep in solitarily, and only for one night (Morin et al. 1994). A recent paper thoroughly cataloged all the different pathogens that were collected from chimpanzee nests in Tanzania. They found that the microorganisms and parasites present were highly correlated to those found in the surrounding environment and changed seasonally (Thoemmes et al. 2018). Indeed, these results suggest, although no equivalent study for the other populations exist, that the microbial repertoire in the nests of chimpanzee should vary across environments. There is evidence, that indeed different subspecies present different levels of infections in the wild, likely a result to differing exposures.

It has been well established that half of the four subspecies are infected with simian immunodeficiency virus (SIV). With *P.t. troglodytes* having lower infection rates than *P.t. schweinfurthii* (Sharp et al. 2005; Li et al. 2012; Locatelli et al. 2016). Several recent studies have cataloged infection rates in differing populations of chimpanzee, including rhinovirus (Scully et al. 2018) and adenovirus (Dadáková et al. 2018) in *P.t. schweinfurthii* and cytomegalovirus in *P.t. verus* (Anoh et al. 2018). Although complete data on current infection rates in other populations do not exist, it is unlikely that infection rates are ubiquitous over such a long distance and between the different subspecies.

Understanding how genetic diversity affects infectious diseases is imperative for human health. Researching chimpanzee offers an opportunity to understand disease correlation with additional

diversity in a similar animal, who, in almost all cases are susceptible to the same diseases. Furthermore, outbreaks of diseases are relatively common that originated in apes and transmitted to humans. For example, an outbreak of human monkeypox is currently plaguing western and central Africa (Durski et al. 2018). The ancestor of human immunodeficiency virus (HIV) groups M and N is from SIV that originated in chimpanzees (Keele et al. 2006; Van Heuverswyn et al. 2007). Likewise, humans are able to infect chimpanzees with disease. Since field research of chimpanzees began in the 1960's, 23 documented cases of human to chimpanzee transmission have been documented (Dunay et al. 2018), including a devastating outbreak of polio in 1966. This is important to study because the chimpanzees in central Africa are endangered and in the west are critically endangered. Their populations are projected to keep declining in numbers. Therefore, great care should be taken in order to avoid passing pathogens that may lead to further population decimation.

Overall, we understand that this species has four distinct genetic groups. However, it is not clear to which extent their phenotypes differ in the wild, and how that relates to their genetic diversity. A dataset consisting of an adequate number of full genome sequences of each of the four subspecies has only recently been made available (de Manuel et al. 2016). Using these data, the most important genetic differences need to be highlighted. Once these differences are defined, detailed phenotypic information, which is more time consuming and costly to collect, can be obtained. A fully integrated scan and comparison of these four subspecies is an important first step into unraveling the relationship between phenotype and genotype in the chimpanzee.

2. A fully integrated scan of selection in the Chimpanzee genome

Jessica Nye[1], Mayukh Mondal[2], Jaume Bertranpetit[1], and Hafid Laayouni[1,3]

[1]Institut de Biologia Evolutiva (UPF-CSIC), Universitat Pompeu Fabra, Doctor Aiguader 88, Barcelona 08003, Catalonia, Spain.

[2]Estonian Biocentre, Tartu 51010, Estonia.

[3]Bioinformatics Studies, ESCI-UPF, Pg. Pujades 1, 08003 Barcelona, Spain

Journal of Evolutionary Biology, Submitted

2.1 Introduction

Although chimpanzees are the closest genetic relative to humans, sharing much of our genetic material, we still understand little about their evolutionary history. In recent years a complete picture of their demographic history has been elucidated (Kaessmann et al. 1999; Fischer et al. 2004; Prado-Martinez et al. 2013). Chimpanzees and bonobos diverged around a million years ago. The two major branches of the chimpanzee lineage began to split from each other some 500 thousand years ago (kya). Today we identify four subspecies of chimpanzee, and understand that *Pan troglodytes verus* and *P.t. ellioti* are more closely related and diverged from each other first, followed by *P.t. troglodytes* and *P.t. schweinfurthii* (diverging around 150kya). Although the four subspecies are genetically and geographically distinct, it is clear there has been extensive gene flow among chimpanzee subspecies and with bonobo (de Manuel et al. 2016).

The demographic history of a species is an important key to understanding their evolution. The size of a population's breeding pool can indicate how many selective events are likely to have taken place (Nam et al. 2017) and how strongly the effects of genetic drift may in fact be (Ohta and Kimura 1969). Furthermore, introgression between species and gene flow from a close subspecies may be a source for beneficial genetic material (as reviewed by Arnold and Martin 2009). Previous studies on the impact of interspecies introgression indicates that the pressure to create a viable hybrid is likely to lead to population changes in fertility phenotypes (Sankararaman et al. 2014; Nye et al. 2018).

Beyond demography, unique selective events are likely to have impacted the genomes of chimpanzee based on their habitats. The four subspecies live in two distinct regions of Africa. Western chimpanzees (*P.t. verus*) live in Ivory Coast and Guinea. While the

other three subspecies live in central Africa. Specifically, Nigeria-Cameroon chimpanzees (*P.t. ellioti*) lives in its namesake countries, eastern chimpanzees (*P.t. schweinfurthii*) inhabits seven countries but primarily in the Democratic Republic of Congo, and central chimpanzees (*P.t. troglodytes*) inhabits five countries but primarily Gabon. The selective pressures that these populations face are likely due to their unique habitats. In differing regions these populations will experience exposure to divergent pathogens, differing quantity and quality of resources, and, most importantly, for a social animal, separate cultures. All these factors together, are likely to be responsible for some of the genetic differences we observe between chimpanzees.

When a beneficial change occurs in a genotype that leads to some phenotypic advantage, that new allele may be selected for. After many generations, the region undergoing a selective sweep will contain reduced variation accompanied by an extended homozygosity as compared to neutral regions or other populations. These features can be interpreted as evidence of selection at a certain point in history by using statistical tests based on allele frequencies, patterns of linkage disequilibrium, and comparisons between populations. Here, we present the first comprehensive scan of the chimpanzee genome that integrates varied selective simulations which encompass complete and ongoing selection occurring between present time and some 60 kya, These simulations are interrogated by 15 statistical tests and using a machine-learning approach we draw conclusion in order to better understand the unique evolutionary history of our genetic cousins, and unveiling their adaptive history through the unique or shared signals of positive selection.

2.2 New Approaches

We provide a comprehensive genome-wide map that includes the signatures of positive selection. These data can be viewed and used in the form of a genome browser, available at http://hsb.upf.edu/chimp_browser/index.html; following the criteria and configuration of a published human dataset (Pybus et al. 2014; 2015). The UCSC-style format facilitates the integration with the rich UCSC browser tracks, a search allows easy access to results for specific genes or genomic regions, and all the information can be conveniently downloaded using the UCSC Table function. We expect this to be a valuable resource for a wide range of future analyses. As such, it provides a broad picture of the action of positive selection in each genomic region in all four chimpanzee subspecies, mapped to Pantro2.1.4.

2.3 Materials and Methods

a) Genome Sequences

Full genome sequences of four subspecies of Chimpanzee were obtained from the Great Ape Genome Project (Prado-Martinez et al. 2014; de Manuel et al. 2016) as vcf files aligned to Chimpanzee genome release Pantro4. The sample sizes are 18 Pan troglodytes troglodytes, 19 Pan troglodytes schweinfurthii, 10 Pan troglodytes ellioti, and 12 Pan troglodytes verus. The ancestral states for each single nucleotide polymorphisms (SNP) were extracted from the 1000 genomes data (1000 Genome Project Consortium 2010) using the human-gorilla-chimpanzee alignment. The data were pruned to exclude missing sites, missing ancestral information, and insertions and deletions for a total of 1,022,493 SNPs. The data were phased using shapeit (Howie et al. 2009; for additional information see Supplementary information section 2.7).

b) Demographic Model

We used a demographic model adapted from (de Manuel et al. 2016). Our model includes all four subspecies of chimpanzee and bonobo with all meaningful admixture events. For the fit of our model compared to real data see Supplementary Figure S2.1 and Supplementary Table S2.1. The formulation of the demographic model is explained in detail in the Supplementary Information section 2.7. The estimates for chimpanzee effective population are 39,925 for *P.t. troglodytes*, 12,829 for *P.t. schweinfurthii*, 12,364 for *P.t. ellioti*, and 10,742 for *P.t. verus*. With *P.t. troglodytes* and *P.t. schweinfurthii* diverging from each other 136,350 years ago (ya), *P.t. ellioti* and *P.t. verus* diverging 498,462 ya, and the ancestors of the two major branches of chimpanzee split 512,050 ya.

c) Simulations

We used the coalescent simulator msms (Ewing and Hermisson 2010). For the neutral simulations we simulated 2,000 replicates of 600,000 bp sections. We matched the sample sizes of each subspecies, resulting in equal samples sizes for the length of the genome.

For the selective scenarios, the program msms requires that all selection be simulated during time invariant portions of the demography model. This means that our selection scenarios are required to occur between the current time and any migration events or population size changes in the demographic model. For this reason, we chose to simulate selection at the following generation time points 600, 900, 1200, 1500, 1800, 2100, and 2400 (generation time is 25 years). To do this we used the tag -e. We allowed for

hard selection (where selection begins on a singleton in the populations) using the tag -SAA. We allowed for the selection coefficient to vary between 0.05 and 0.55. We selected 1,000 successfully selected simulations where the final allele frequency (FAF) was fixed (complete sweep) for each subspecies and 1,000 successful selected simulations where the FAF was between 0.6 and less than 1 (incomplete) for each subspecies using the tag -oTrace. Resulting in a total of 7,000 hard incomplete and 7,000 hard complete, or 14,000 simulations per subspecies of length 600,000 bp.

d) Statistical Tests

The suite of statistical tests of this study combine statistics based on SFS, linkage disequilibrium (LD), descriptive statistics, and population differentiation (Table 2.1, adapted from Pybus et al. 2014). We chose 15 statistics based on the results from Pybus et al. which employed a machine-learning approach to search and classify selection in the human genome using a combination of signals from various statistics (2014). All statistics were calculated genome-wide for both real and simulated datasets. The window-based statistics were calculated in windows of 30kb with a 3kb sliding window. Windows were dropped if there were less than 5 SNPs in the window, to avoid the possibility of poor coverage in that area. The window and SNP-based statistics were combined by selecting the median value of each SNP-based statistic per window. All statistics were calculated using scripts provided by Pybus et al. (2014). The genome-wide real data and the neutral simulations for each statistic are presented in Supplemental Figures S2.2-S2.22, and for simulated neutral and selective scenarios in Supplemental Figures S2.23-S2.43.

e) Random Forest Algorithm

We employed a machine-learning approach in order to differentiate between regions of the genome that are neutral and regions that have evidence of a selective sweep based on the results of a large number of selection tests, specifically a random forest method. Briefly, we chose this method because it is an improvement on previous machine-learning approaches such as decision trees (which run the risk of over-parameterization when using correlated input statistics) (Kass 1980; Breiman et al. 1984) or bagging (which runs the risk of basing the regression model on too many similar trees) (Krzywinski and Altman 2017). Our input data consists of many correlated statistics (Supplementary Figure S2.44), all of which have benefits and disadvantages. The random forest algorithm is an extension of bagging, in that it constructs an entire forest of trees of random structure at each node. This ensures that the ultimate regression model is based on a sufficient mixture of the data which avoids both over parameterization and bias being built into the underlying model (Breiman 2001), for more information about our chosen model, see Supplementary information. We used the R package RandomForest (Liaw and Wiener 2002). The RandomForest model was trained with the 15 calculated statistics for both the neutral simulations and the selective simulations. Because we modeled extensive selection scenarios occurring between present time and 60 thousand years ago, with the machine-learning approach, we are able to confidently determine that regions have signatures of selection at some point in the recent past. We ran forests of 5,000 random trees. According to our OOB error estimates, ~500 trees are sufficient (Supplementary Figure S2.45). The output is a maximum likelihood prediction that each window is either categorized as under selection or neutral. In order to be as conservative as possible, we accepted only the most robust signals of putative positive selection ($Pr(selection) \geq 0.95$).

f) Genes Under Selection

To extract the genes within windows under selection, we used intersectBed using the .gtf file from ensemble release 90. We then used the desktop version of the ensemble variant effect predictor (release 93; McLaren et al. 2016) to extract possible functional variants under selection. We extract all sites with predicted impact high, moderate, and low. We extract allele frequencies for each variant site for all four subspecies in order to assess differences among them. Gene ontology analysis was performed using the program Gowinda (Kofler and Schlötterer 2012) which uses simulations to provide a robust p-value for a presented list of SNPs. We used the options --mode gene, --min-genes 1, and --simulations 1000000 to obtain the most robust results.

2.4 Results

a) Regions identified using the random forest approach

While training the random forest algorithm, we observe a low out of bag (OOB) error rate, as predicted from our training set, for each subspecies (*P.t. ellioti* = 2.62%, *P.t. schweinfurthii* = 1.26%, *P.t. troglodytes* = 1.09%, and *P.t. verus* = 2.94%). Indicating that we are able to discern between selective and neutral regions (see Supplementary Figures S2.46-S2.49 and Supplementary Tables S2.2-S2.5 for subspecies-specific and test-specific performances and Supplementary Tables S2.6 for detailed OOB rates).

After applying the regression model formulated by the training of the random forest, we receive a prediction for each window in the

genome for each subspecies of chimpanzee. In order to remain confident that we are extracting true signatures of selection, we select only regions whose maximum likelihood probability of belonging to the selection group is ≥0.95. We identify the greatest number of regions under selection in the subspecies *P.t. troglodytes*, and similar numbers for the other three subspecies (Table 2.2). This is due in part to the underlying demography of these lineages, where *P.t. troglodytes* has an effective population size (N_e) nearly four times larger than the other three (see Supplementary Information section 2.7 for further demography details). Previous research has demonstrated that the number of selective sweeps scales with effective population sizes (Nam et al. 2017). Our study in part confirms these findings; however, we see a less dramatic increase in the amount of regions under selection than previously reported (i.e., closer to 2.5 times greater as compared to 4 times greater as is predicted by their N_e).

Unfortunately, as chimpanzee is not a model organism, it is difficult to interpret signals that fall outside coding regions. The annotation of the genome does not allow a precise interpretation of selection signals in noncoding or regulatory regions. We find a similar proportion of regions predicted as under putative positive selection that contain no genes for three of the four subspecies ~30% (Table 2.2). The subspecies that has an increase of signals falling outside coding regions is *P.t. verus* (42%). It is interesting to note that this subspecies has a far lower level of genetic diversity than the other three. Overall, the proportion of regions that contain genes is much higher than we find in studies of selection in the human genome (Pybus et al. 2014; 2015), similar results were previously found (Cagan et al. 2016). We expect that some of these signals are responsible for affecting protein coding genes through regulation; however, we are only able to focus on protein coding genes.

After extracting the genes located within each region (Table 2.2) from ensemble release 90, we find 823 genes as being targets of

selection for *P.t. troglodytes*, while the remaining three subspecies have around 300 genes each. We first compared our results with a previous scan of selection in chimpanzee (Cagan et al. 2016) and our study is able to confirm signatures of selection in 49 genes (Supplementary Table S2.7).

When we compare significant regions across subspecies (Figure 2.1), we find that the target of selection in the species as a whole is not common. In fact, we find only one gene, *FHOD1*, as significant in the scan for all four subspecies (Table 2.3). A further 13 genes overlap with three of the four subspecies while 152 genes appear in scans for two subspecies (Table 2.3). This leaves the majority of genes as unique to the individual populations (total of 1,512 genes). Altogether, this observation indicates the selective pressures exerted on each of the four subspecies have been unique to each population since their divergence some 500,000 years ago. We found similar results into selection of the introgressed regions between chimpanzee and bonobos (Nye et al. 2018).

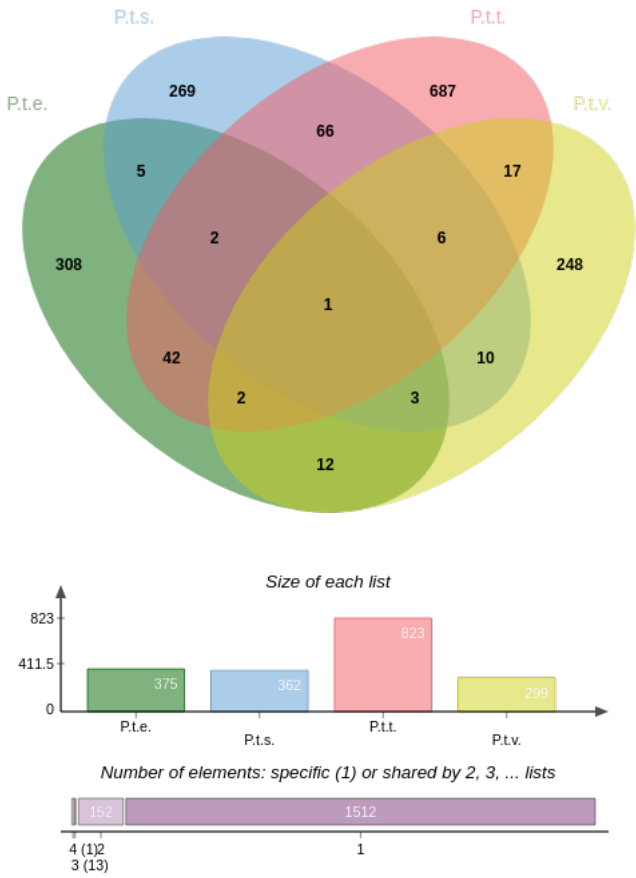


Figure 2.6: Venn diagram of genes that overlap between subspecies of chimpanzee. P.t.e. is *Pan troglodytes ellioti*, P.t.s. is *Pan troglodytes schweinfurthii*, P.t.t. is *Pan troglodytes troglodytes*, and P.t.v. is *Pan troglodytes verus*.

b) Genes commonly under selection in more than two subspecies of Chimpanzee

As mentioned above, we identify a region containing one gene as significant for all four subspecies. The gene *FHOD1* (formin homology 2 domain containing protein 1) is a member of the formin family and regulates the cytoplasmic actin for spindle movement

and is required for cell organization and limb development (Pan et al. 2017). Although expressed in many tissues, its role is in muscles where it plays a key role in the coordination of microtubules and allows for cell elongation. It has been suggested that nuclear-cytoskeletal linkages regulate a feedback loop that tunes internal stiffness of cells to match that of their soft microenvironment, through inside to outside pathways involving the actin cytoskeleton and the formin *FHOD1* (Schwartz et al. 2017). No information about the phenotypic effect of genetic variation in this gene exists, it is nonetheless likely that its activity is related to muscle activity in all subspecies of chimpanzee.

Interestingly, we find that five of the thirteen genes that overlap with three subspecies (half of which do not include the subspecies *P.t. ellioti*, Table 2.3) have some phenotypic role in the immune system. The gene *E2F4* has been linked to infection by the adenovirus (Pelka et al. 2011). *KPNA6* is bound by the PB2 subunit of the influenza A virus, as is required for its viral replication in the lung (Resa-Infante et al. 2014; Pumroy et al. 2015), like humans, chimpanzees are commonly infected with the influenza virus (Buitendijk *et al.* 2014), and likely experience selective pressures due to its infection. The gene *RASGRP4* is required for the natural killer (NK) cell response to lipopolysaccharide, a common Gram-negative bacterial membrane component (Zhou et al. 2016). The gene *KCNK6* is required for inflammasome-induced inflammation (Di et al. 2018). *ELMO3*, although primarily involved in cytoskeletal rearrangements required for phagocytosis of apoptotic cells and cell motility, and is found in the pathways of Shigellosis and Bacterial invasion of epithelial cells (Belinky et al. 2015).

We also find one gene, *RYR1*, which was described as having signatures of both purifying and positive selection (McKay and Griswold 2014) and has an important function in modulating the physiology of muscle contraction, and is mainly expressed in the skeletal muscle. *PRDM16* specifies the brown fat lineage from a

progenitor that expresses myoblast markers and is fundamental in muscle differentiation (Seale el al. 2008). Lastly, *SLC9B1* encodes a sodium/hydrogen exchanger and transmembrane protein that is a testis-specific solute carrier and is involved in sperm motility and fertility (Holmes et al. 2016). The remaining five genes which overlap three subspecies (*SDSL*, *TMEM205*, *LRRC29*, *TLK2*, and *TMEM150B*), we find no noteworthy function. The most common traits that are shared between multiple subspecies of chimpanzee, and therefore are an ancient adaptation, are immunity and muscle function, with a single signature in a fertility-related trait.

c) Genes commonly under selection in two subspecies of Chimpanzee

To fully interrogate the genes under selection in two subspecies (and below, genes unique to each subspecies), we use the variant effect predictor (McLaren et al. 2016) which categorizes SNPs into four main types of variants based on the impact they are likely to have: modifiers (i.e., intronic, splice, UTR, or 3' variants), low impact (i.e., synonymous variants), moderate impact (i.e., missense variants), and high impact (i.e., start gain or loss, stop gain or loss, and splice acceptors or donors). We disregard modifiers and prioritize genes by the impact of the predicted variant and its allele frequency. From all genes that appear on two subspecies lists we find that 21 (out of a total 152) have a shared segregating variant that is absent from the other two subspecies (see Supplementary Table S2.8), and of those, the allele frequency is high (≥15%) in both subspecies for 7 of the genes.

We find a missense substitution that is fixed in *P.t. schweinfurthii* and nearly fixed (89%) in *P.t. troglodytes* in the gene *PPL*. This gene is in an unstable region of the genome that contains many repeats (Aho et al. 1999). The gene is a keratinocyte that is part of the plakin family. Its function is essential for keratin differentiation

and is highly expressed in both skin and nasal respiratory epithelium.

The actin assembly gene *RAB35* has a shared synonymous substitution that is nearly fixed (92%) in *P.t. schweinfurthii* and intermediate (44%) in *P.t. troglodytes*. It regulates intracellular membrane trafficking, from the formation of transport vesicles to their fusion with membranes. Interestingly, a recent study found that the function of this gene is among the first to be disrupted upon infection with cytomegalovirus (CMVs; Karleuša et al. 2018). CMVs extensively rearrange the cellular membrane system to develop an assembly compartment. This virus has been well documented among nonhuman primates for decades (Leendertz et al. 2009).

The gene *REV3L* is the catalytic subunit of DNA polymerase zeta. It is an essential gene for life, resulting in inviable mice when knocked down (Van Sloun et al. 2002), but its primary function is in translesion synthesis of damaged DNA. We find eight functional variants, specifically 4 missense, 3 synonymous, and 1 splice variant substitutions shared in *P.t. schweinfurthii* and *P.t. troglodytes* and absent from the other two subspecies.

Arginine methyltransferases are responsible for posttranslational modifications of proteins, which may lead to the stability of proteins. This gene family is important, but not that well understood. The gene *PRMT7* is a type II methyltransferase, and is less conserved through evolutionary history compared to other members of the family (Bachand 2007). We find a synonymous substitution shared between *P.t. schweinfurthii* and *P.t. troglodytes* (26% and 47%, respectively).

GYPA may be one of the most interesting genes under selection, although their shared missense and synonymous changes is not segregating at high frequencies (*P.t. schweinfurthii* 5% and *P.t.*

troglodytes 11%). This gene is known to be at the base of malaria resistance in humans and forms the MNs blood group. This gene was also shown to be one of the most highly diverged between humans, chimpanzees, and gorillas in the entire genome (Baum et al. 2002; Wang et al. 2003).

Lastly, *OAS2* is a gene that encodes a member of the 2-5A synthetase family which includes essential proteins involved in innate immune response to viral infection. This gene has shared signals between *P.t. ellioti* and *P.t. troglodytes* (synonymous change 15% and 3%, respectively). In fact, the variants of this gene have been associated with differing progression of HIV infection (Hosseini et al. 2015; Bakhteeva et al. 2016).

d) Genes uniquely under selection in a specific subspecies

Using the same variant scan as above, we focus on genes where each subspecies has a unique variant as compared with the other three subspecies. Unlike the section above, due to the large number of interesting variants, we concentrate here on only high and moderate impact variants. For *P.t. ellioti*, we observe 74 genes with a unique variant (Supplementary Table S2.9), 20 of those variants are segregating at least 15%, and 5 of those above 30%. For *P.t. schweinfurthii*, we observe 130 genes with a unique variant (Supplementary Table S2.10), 7 of those variants are segregating at least 15%, and 3 of which above 30%. For *P.t. troglodytes*, we observe 342 genes with a unique variant (Supplementary Table S2.11), 18 of those variants are segregating at least 15%, and 5 of which above 30%. For *P.t. verus*, we observe 79 genes with a unique variant (Supplementary Table S2.12), 11 of those variants are segregating at least 15%, and 3 of which above 30%. In all these cases it is extremely likely that the variants have functional implications. Although we see no significant enrichment in gene

ontology from our results as a whole, like the shared signals discussed above, we see a general trend in reproduction and muscle function. At the individual level, we also see genes involved in DNA repair and replication.

Reproduction

We find three unique missense substitutions for *P.t. ellioti* (15%, 15%, and 5%) in the gene *CREM*. This is a key transcription factor which is needed for the maturation of spermatozoa. Amino acid changes to this gene have been implicated in a study of male infertility. Interestingly, the human *CREM* is more closely related to gorilla than chimpanzee (Christensen et al. 2006). In combination with our current study, this indicates selection within this gene in a branch of the chimpanzee lineage.

For *P.t. troglodytes*, we find two genes with unique variants segregating at a high frequency with phenotypes relating to male fertility. The gene *IQCH* is an immunogenic antigen present of the surface of sperm cells (Xiang et al. 2017). We observe 4 missense (highest frequency=39%), 1 synonymous, and 1 splice variant substitutions. The gene *RAD9B* has 1 missense (36%) and 1 synonymous substitution. Its function is as a check point, mediating DNA damage, and is highly expressed in testis (Lyndaker et al. 2013).

DNA repair and replication

The gene that contains the highest frequency unique variant for *P.t. ellioti* is *CEP85*. We find two missense substitutions (70% and 10%). This gene was recently found to regulate the gene *NEK2*, which initiates spindle separation during meiosis (Chen et al. 2015), an imperative step of the interphase stage.

Two genes for *P.t. troglodytes* have functions DNA damage repair (additionally to *RAD9B* above). The gene *POLL* has 1 missense substitution segregating at 19% and is responsible for strand synthesis in gapped DNA. It is highly expressed in ovary and in testis, suggesting involvement with meiosis (Kobayashi et al. 2002). The gene *ATRIP* has 2 missense (highest segregating at 17%) and 1 synonymous substitution. This protein binds single stranded DNA following damage to mediate repair (Wold 1997).

Muscle

The giant muscle protein *TTN* is the largest protein in the body, with the chimpanzee variant containing over 32,000 amino acids. We find a large number of unique variants in *P.t. schweinfurthii*, specifically 61 missense, 29 synonymous, and 4 splice variant substitutions (see Supplementary Table S2.10). This protein connects the Z disk (which anchors the actin filament) to the M line (which anchors myosin) in muscle cells. Every time a muscle is flexed this giant protein unfolds and refolds to allow for muscle function (Manteca et al. 2017). This protein is responsible for the twitch reflex of muscles (Rief et al. 1997). Interestingly, the super strength of chimpanzees, as compared to humans, has been attributed to this reflex (O'Neill et al. 2017). We would like to note here that we find unique variants in *P.t. troglodytes* and *P.t. verus* segregating at high frequencies and that windows containing this gene are trending toward significant scores of selection in these two subspecies as well.

Other Genes Under Selection

We find a missense substitution segregating at 26% in *P.t. schweinfurthii* for the *MTHFD2L* gene. It is part of folate synthesis metabolism which coverts 5,10-methylene-tetrahydrofolate to 10-formyl-THF (Bolusani et al. 2011). Deficiencies and excesses of folate have been associated with facial deformities and an

imbalance with thyroid hormone levels (Prescott and Malcolm 2002; Colleran et al. 2003; Sittig et al. 2012) and thyroid hormone levels themselves are associated with proper palette formation during development (Bronchain et al. 2017). One source of exposure to tetrahydrofolate is through pesticides. We would like to note here that this subspecies has been observed in the wild to have facial and cranial deformities (Krief et al. 2017). These deformities may indicate that exposure to excess toxins in their environment may create excess burden on folate metabolism which in turn may have driven recent selection in this gene.

We find 1 missense change (23%) in *P.t. verus* in the receptor for the anthrax toxin (*ANTXR2*). Most interestingly, it was recently reported that this specific subspecies is being decimated by anthrax in the wild. The authors predict the total extirpation of its population in the next 150 years due to anthrax poisoning (Hoffmann et al. 2017). However, our results indicated that selective pressure on the receptor has elicited a genetic response, indicating that this population may adapt and with any luck survive longer than 150 years.

The gene *GRIN3A* is a member of a family of the GluR glutamate receptors which are functionally active in the central nervous system. This family is attributed to be a key for learning and memory (as reviewed by Riedel et al. 2003). This particular gene, *GRIN3A*, is one of the most diverged in the family when comparing human and chimpanzee sequences (Goto et al. 2009). Interestingly, we find a missense substitution segregating at 32% in *P.t. verus* that has not previously been reported. It is also important to note (although unlikely to be due to changes in only one gene), that this subspecies in particular has been observed to exhibit unique behaviors as compared to other chimpanzees, including rock collecting and throwing (Kühl et al. 2016), spear hunting (Pruetz et al. 2007), and algae fishing (unpublished data).

2.5 Discussion

We have created a community resource that allows researchers to investigate signals of selection at the level of subspecies in chimpanzee. This tool will be of use to researchers investigating chimpanzee phenotypes, clinicians investigating disease differences among human and ape species, and evolutionary biologists interested in speciation, among others. These results indicate that positive selection is an important driver for differentiation between populations and eventual speciation. We find that out of all the genes with signals of positive selection, the vast majority (81%) are unique to a certain subspecies. This indicates that many interesting differences between the subspecies appear to be mainly driven by environmental differences. Although three of the four subspecies live in seemingly ecologically similar habitats, their genomes indicate that exposures to subtle differences have resulted in differential adaptations.

We observe a clear impact of the demographic history of individual subspecies on the signals of selection. The population that has more than twice the number of signals, *P.t. troglodytes*, as compared with the other three subspecies has the highest effective size. This is an expected result because genetic drift is weaker in populations with large effective population sizes, and as a result selection can more easily drive beneficial alleles to high frequencies (Nam et al. 2017). We also observe a higher proportion of signals outside coding regions for *P.t. verus*.

Signals that are shared between subspecies are most often between *P.t. troglodytes* and *P.t. schweinfurthii*. This is expected due to their demographic past, as these two subspecies are more closely related and diverged more than 300 ky after the other two subspecies. Shared signals in other pairs of subspecies are low; because they do

not share a common evolutionary history. It is clear, however that the species as a whole has had selective pressures in phenotypes involved in immunity, DNA repair, and muscle function.

We find that there has been significant pressure from pathogens especially due to viruses and bacteria. We find four genes (*E2F4*, *KPNA6*, *OAS2*, and *RAB35*) that have some function in protection from viral pathogens, two (*RASGRP4* and *ELMO3*) that protect against bacterial pathogens, one that is involved with the plasmodia parasite (*GYPA*), and one involved in innate immunity (*KCNK6*). It has long been established that chimpanzees are susceptible to almost all the same pathogens as humans, and have been observed to experience epidemics of disease, much in the same way that humans do (Dunay et al. 2018). Although we expect to find signals of selection in genes involved with immunity, understanding the cause and effect of these signals is of utmost importance for both improving human health and for designing protective strategies for these endangered creatures. Especially because disease passing between humans and animals is becoming a more common occurrence as human populations continue to grow and encroach on animal habitats.

It is less obvious why we find signals in DNA repair and muscle function. We observe signals of positive selection in five genes (*REV3L*, *CEP85*, *POLL*, *ATRIP*, and *RAD9B*) which have some function in DNA repair, and four (*FHOD1*, *RYR1*, *PRDM16*, and *TTN*) with muscle function. In comparison with humans, chimpanzees are superiorly strong. Clearly some aspect of their daily lives requires strong muscles; however much more phenotypic research is needed to understand what is driving this adaptation.

This is a common theme when investigating aspects of a non-model organism. It can be extremely difficult to relate phenotype to genotype because much of the information is missing. For instance, while many studies have investigated some aspect of phenotype in

this species, comparative studies that encompass all four subspecies are nonexistent. This is due to the fact that their habitats are expansive, making detailed observations difficult if not impossible. They reside, in many cases, in areas that make conducting a study difficult politically and safely. Furthermore, the subspecies *P.t. ellioti* was first described a little over a decade ago (Ely et al. 2005), making the body of research of this subspecies small. There are also aspects that make this type of study difficult genetically. Chimpanzees have a much lower amount of linkage disequilibrium than humans, which results in the statistics based on LD less powerful. The annotation of the genome also makes it difficult to interpret signals that lie outside coding regions.

Nonetheless, our comprehensive scan of the chimpanzee genome coupled with the extensive simulations of selection scenarios occurring in the last 60 thousand years allow for robust categorization of regions of the genome as under selection. This work confirms and expands the previous scan of selection in the chimpanzee genome (Cagan et al. 2016). Our results are available for use by the community in a simple to use genome browser. A detailed map of interesting genetic differences between these subspecies is an important tool to use when building a better understanding the genotype-phenotype map of chimpanzees.

2.6 Tables

Table 2.1: Selection tests used.

Principle	Method	Reference
Site frequency spectrum	Tajima's D	Tajima, 1989
	Fu and Li's D	Fu and Li, 1993
	Fu and Li's F	Fu and Li, 1993
	R_2	Ramos-Onsins and Rozas, 2002
Linkage Disequilibrium	iHS	Voight *et al.*, 2006
	$EHH_{Average}$	Sabeti *et al.*, 2002
	Wall's B	Wall, 1999
	Wall's Q	Wall, 2000
	Fu's F	Fu, 1997
	Z_a	Rozas *et al.*, 2001
	Z_{nS}	Kelly, 1997
	ZZ	Rozas *et al.*, 2001
Population differentiation	ΔDAF	Hofer *et al.*, 2009
	XP-EHH	Sabeti *et al.*, 2007
Descriptive statistics	π (nucleotide diversity)	Nei and Li, 1979

Table 2.2: Regions under putative positive selection, number of genes, proportion without genes, and relative genome proportion by subspecies.

Subspecies	Regions under selection	Relative genome proportion	Total number of genes	Number of unique genes	Proportion of regions without genes
P.t. ellioti	339	0.51%	375	308	29.5%
P.t. schweinfurthii	347	0.71%	362	269	33.4%
P.t. troglodytes	766	1.31%	823	687	29.9%
P.t. verus	380	0.57%	380	248	42.1%

Table 2.3: Number of overlapping regions and genes between subspecies.

Subspecies	Overlapping regions	Overlapping genes
ellioti-schweinfurthii	15	5
ellioti-troglodytes	31	42
ellioti-verus	11	12
schweinfurthii-troglodytes	69	66
schweinfurthii-verus	10	10
troglodytes-verus	22	17
ellioti-schweinfurthii-troglodytes	6	2
ellioti-schweinfurthii-verus	2	3
ellioti-troglodytes-verus	2	2
schweinfurthii-troglodytes-verus	5	6
ellioti-schweinfurthii-troglodytes-verus	1	1

2.7 Supplementary Information

a) Genome Phasing

The genomes were phased using the program shapeit (Howie et al. 2009). We performed the phasing in three steps. First, we performed the statistical phasing using an input reference map. We downloaded the reference map from Auton (et al. 2009). This chimpanzee genetic map was mapped to genome release Pantro2. We lifted the map from Pantro2 to Pantro3 and from Pantro3 to Pantro4. We extrapolated missing sites using the approx function in R. The following commands were used to phase the data in shapeit:

```
--effective-size 10000 \
--window 0.1 \
--burn 10 \
--prune 10 \
--main 50 \
--thread 2 \
```

We next used shapeit to convert the output graph file to a phased file using the --convert function. In the third step we used shapeit to convert to a vcf file using the --convert function while also removing all badly phased haplotypes.

b) Demography

A demographic model for the Pan clade was measured specifically for these data (de Manuel et al. 2016). The model is split into two scenarios in which one model features bonobo, *P.t. troglodytes*, *P.t. schweinfurthii*, and *P.t. ellioti*; while the other describes bonobo, *P.t. troglodytes*, *P.t. schweinfurthii*, and *P.t. verus*. For the purpose

of this study, we needed to use a single model that contains all the four chimpanzee subspecies. The two models are as follows, in msms code:

Model 1 (1=bonobo, 2=P.t. troglodytes, 3=P.t. schweinfurthii, and 4=P.t. ellioti):

-N 10000 -t 1.44 -r 1.2 3000 -I 4 0 38 36 20 0.0 -n 1 0.0742 -n 2 0.3181 -n 3 0.3092 -n 4 0.0386 -m 1 2 0 -m 1 3 0 -m 1 4 0 -m 2 1 0 -m 2 3 1.8181960943074 -m 2 4 0 -m 3 1 0 -m 3 2 2.02290154800773 -m 3 4 0 -m 4 1 0 -m 4 2 0 -m 4 3 0 -en 0.001 1 1.83290809268 -en 0.001 2 1.161030985567 -en 0.001 3 3.66914400056 -en 0.001 4 1.23640124358 -em 0.020875 1 2 0 -em 0.020875 1 3 0 -em 0.020875 1 4 0 -em 0.020875 2 1 0 -em 0.020875 2 3 1.8181960943074 -em 0.020875 2 4 1.12888460726286 -em 0.020875 3 1 0 -em 0.020875 3 2 2.02290154800773 -em 0.020875 3 4 0.514005225416364 -em 0.020875 4 1 0 -em 0.020875 4 2 0.61034918826118 -em 0.020875 4 3 2.77081002950074 -em 0.042025 1 2 0 -em 0.042025 1 3 0.0447270935214584 -em 0.042025 1 4 0.00204350937063846 -em 0.042025 2 1 0 -em 0.042025 2 3 1.8181960943074 -em 0.042025 2 4 1.12888460726286 -em 0.042025 3 1 0.0340892941439601 -em 0.042025 3 2 2.02290154800773 -em 0.042025 3 4 0.514005225416364 -em 0.042025 4 1 0.00878072013784504 -em 0.042025 4 2 0.61034918826118 -em 0.042025 4 3 2.77081002950074 -en 0.104325 2 0.0402577179646081 -en 0.104325 3 0.192594746352967 -en 0.106325 3 8.73162876459514 -ej 0.106325 2 3 -em 0.106325 1 2 0 -em 0.106325 1 3 0.0177338314347154 -em 0.106325 1 4 0.00204350937063846 -em 0.106325 2 1 0 -em 0.106325 2 3 0 -em 0.106325 2 4 0 -em 0.106325 3 1 0.00723425109237692 -em 0.106325 3 2 0 -em 0.106325 3 4 0.193855714034029 -em 0.106325 4 1 0.00878072013784504 -em 0.106325 4 2 0 -em 0.106325 4 3 0.00771007640703268 -en 0.41955 1 0.158405393915496 -en 0.42155 1 0.299481445247702 -en 0.473075 4 0.0306317427630759 -en 0.475075 4 2.79429564470655 -en 0.480625 4 0.0872103733618782 -em 0.480625 1 2 0 -em 0.480625 1 3 0.0177338314347154 -em 0.480625 1 4 0.00204350937063846 -em 0.480625 2 1 0 -em 0.480625 2 3 0 -em 0.480625 2 4 0 -em 0.480625 3 1 0.00723425109237692 -em 0.480625 3 2 0 -em 0.480625 3 4 0.193855714034029 -em 0.480625 4 1 0.00878072013784504 -em 0.480625 4 2 0 -em 0.480625 4 3 0.00771007640703268 -en 0.482625 3 1.66920782430592 -ej 0.482625 4 3 -em 0.482625 1 2 0 -em 0.482625 1 3 0.241282075772286 -em 0.482625 1 4 0 -em 0.482625 2 1 0 -em 0.482625 2 3 0 -em 0.482625 2 4 0 -em 0.482625 3 1 0.0101771164248256 -em 0.482625 3 2 0 -em 0.482625 3 4 0 -em 0.482625 4 1 0 -em 0.482625 4 2 0 -em 0.482625 4 3 0 -en 1.5988 3 0.00336130452736601 -en 1.6008 3 1.47105091660349 -ej 1.6008 1 3 -em 1.6008 1 2 0 -em 1.6008 1 3 0 -em 1.6008 1 4 0 -em 1.6008 2 1 0 -em 1.6008 2 3

0 -em 1.6008 2 4 0 -em 1.6008 3 1 0 -em 1.6008 3 2 0 -em 1.6008 3 4 0 -em 1.6008 4 1 0 -em 1.6008 4 2 0 -em 1.6008 4 3 0

Model 2 (1=bonobo, 2=P.t. troglodytes, 3=P.t. schweinfurthii, and 4=P.t. verus):

-N 10000 -t 1.44 -r 1.2 3000 -I 4 0 38 36 24 0.0 -n 1 0.1636 -n 2 0.2168 -n 3 0.30865 -n 4 0.081 -m 1 2 0 -m 1 3 0 -m 1 4 0 -m 2 1 0 -m 2 3 1.26647232666457 -m 2 4 0 -m 3 1 0 -m 3 2 2.26397610393913 -m 3 4 0 -m 4 1 0 -m 4 2 0 -m 4 3 0 -en 0.001125 1 1.57097089776 -en 0.001125 2 1.404710323488 -en 0.001125 3 4.3158739382 -en 0.001125 4 1.0742217327 -em 0.11655 1 2 0 -em 0.11655 1 3 0 -em 0.11655 1 4 0 -em 0.11655 2 1 0 -em 0.11655 2 3 1.26647232666457 -em 0.11655 2 4 0.79048180986804 -em 0.11655 3 1 0 -em 0.11655 3 2 2.26397610393913 -em 0.11655 3 4 0.00440335360192708 -em 0.11655 4 1 0 -em 0.11655 4 2 0.00666829030333492 -em 0.11655 4 3 0.108325487900296 -em 0.1205 1 2 0 -em 0.1205 1 3 0.0404882853109968 -em 0.1205 1 4 0.0060307968152018 -em 0.1205 2 1 0 -em 0.1205 2 3 1.26647232666457 -em 0.1205 2 4 0.79048180986804 -em 0.1205 3 1 0.0688517134577088 -em 0.1205 3 2 2.26397610393913 -em 0.1205 3 4 0.00440335360192708 -em 0.1205 4 1 0.0115754432490942 -em 0.1205 4 2 0.00666829030333492 -em 0.1205 4 3 0.108325487900296 -en 0.168375 2 0.172841376512123 -en 0.168375 3 0.157675036183995 -en 0.1706 3 8.35342687509545 -ej 0.1706 2 3 -em 0.1706 1 2 0 -em 0.1706 1 3 0.077739053851208 -em 0.1706 1 4 0.0060307968152018 -em 0.1706 2 1 0 -em 0.1706 2 3 0 -em 0.1706 2 4 0 -em 0.1706 3 1 0.0101745515682769 -em 0.1706 3 2 0 -em 0.1706 3 4 0.203210242128393 -em 0.1706 4 1 0.0115754432490942 -em 0.1706 4 2 0 -em 0.1706 4 3 1.30532712910343 -en 0.21195 4 0.143861004427068 -en 0.214175 4 0.229328737963366 -en 0.51205 4 0.163429675293312 -em 0.51205 1 2 0 -em 0.51205 1 3 0.077739053851208 -em 0.51205 1 4 0.0060307968152018 -em 0.51205 2 1 0 -em 0.51205 2 3 0 -em 0.51205 2 4 0 -em 0.51205 3 1 0.0101745515682769 -em 0.51205 3 2 0 -em 0.51205 3 4 0.203210242128393 -em 0.51205 4 1 0.0115754432490942 -em 0.51205 4 2 0 -em 0.51205 4 3 1.30532712910343 -en 0.5143 3 1.69172360590542 -ej 0.5143 4 3 -em 0.5143 1 2 0 -em 0.5143 1 3 0.00187056444537476 -em 0.5143 1 4 0 -em 0.5143 2 1 0 -em 0.5143 2 3 0 -em 0.5143 2 4 0 -em 0.5143 3 1 0.00144484023958745 -em 0.5143 3 2 0 -em 0.5143 3 4 0 -em 0.5143 4 1 0 -em 0.5143 4 2 0 -em 0.5143 4 3 0 -en 0.57125 1 0.00125055964958021 -en 0.573475 1 2.47905585034123 -en 1.343775 3 0.00527072359478981 -em 1.343775 1 2 0 -em 1.343775 1 3 0.00187056444537476 -em 1.343775 1 4 0 -em 1.343775 2 1 0 -em 1.343775 2 3 0 -em 1.343775 2 4 0 -em 1.343775 3 1 0.00144484023958745 -em 1.343775 3 2 0 -em 1.343775 3 4 0 -em 1.343775 4 1 0 -em 1.343775 4 2 0 -em 1.343775 4 3 0 -en 1.346025 3 1.5950192271541 -ej 1.346025 1 3 -em 1.346025 1 2 0 -em

1.346025 1 3 0 -em 1.346025 1 4 0 -em 1.346025 2 1 0 -em 1.346025 2 3 0 -em 1.346025 2 4 0 -em 1.346025 3 1 0 -em 1.346025 3 2 0 -em 1.346025 3 4 0 -em 1.346025 4 1 0 -em 1.346025 4 2 0 -em 1.346025 4 3 0

In order to create a single model, we averaged duplicated parameters to create model 3.

Model 3 (1=bonobo, 2=P.t. troglodytes, 3=P.t. schweinfurthii, 3=P.t. ellioti, and 4=P.t. verus):

-N 10000 -t 1.44 -r 1.2 3000 -I 5 0 38 36 20 24 0.0 -n 1 0.1189 -n 2 0.26745 -n 3 0.308925 -n 4 0.0386 -n 5 0.081 -m 1 2 0 -m 1 3 0 -m 1 4 0 -m 1 5 0 -m 2 1 0 -m 2 3 1.542334 -m 2 4 0 -m 2 5 0 -m 3 1 0 -m 3 2 2.143439 -m 3 4 0 -m 3 5 0 -m 4 1 0 -m 4 2 0 -m 4 3 0 -m 4 5 0 -m 5 1 0 -m 5 2 0 -m 5 3 0 -m 5 4 0 -en 0.0010625 1 1.701939 -en 0.0010625 2 1.282871 -en 0.0010625 3 3.992509 -en 0.0010625 4 1.23640124358 -en 0.0010625 5 1.0742217327 -em 0.0687125 1 2 0 -em 0.0687125 1 3 0 -em 0.0687125 1 4 0 -em 0.0687125 1 5 0 -em 0.0687125 2 1 0 -em 0.0687125 2 3 1.542334 -em 0.0687125 2 4 1.12888460726286 -em 0.0687125 2 5 0.79048180986804 -em 0.0687125 3 1 0 -em 0.0687125 3 2 2.143439 -em 0.0687125 3 4 0.514005225416364 -em 0.0687125 3 5 0.00440335360192708 -em 0.0687125 4 1 0 -em 0.0687125 4 2 0.61034918826118 -em 0.0687125 4 3 2.77081002950074 -em 0.0687125 4 5 0 -em 0.0687125 5 1 0 -em 0.0687125 5 2 0.00666829030333492 -em 0.0687125 5 3 0.108325487900296 -em 0.0687125 5 4 0 -em 0.0812625 1 2 0 -em 0.0812625 1 3 0.04260769 -em 0.0812625 1 4 0.00204350937063846 -em 0.0812625 1 5 0.0060307968152018 -em 0.0812625 2 1 0 -em 0.0812625 2 3 1.542334 -em 0.0812625 2 4 1.12888460726286 -em 0.0812625 2 5 0.79048180986804 -em 0.0812625 3 1 0.0514705 -em 0.0812625 3 2 2.143439 -em 0.0812625 3 4 0.514005225416364 -em 0.0812625 3 5 0.00440335360192708 -em 0.0812625 4 1 0.00878072013784504 -em 0.0812625 4 2 0.61034918826118 -em 0.0812625 4 3 2.77081002950074 -em 0.0812625 4 5 0 -em 0.0812625 5 1 0.0115754432490942 -em 0.0812625 5 2 0.00666829030333492 -em 0.0812625 5 3 0.108325487900296 -em 0.0812625 5 4 0 -en 0.13635 2 0.1065495 -en 0.13635 3 0.1751349 -em 0.1384625 1 2 0 -em 0.1384625 1 3 0.04773644 -em 0.1384625 1 4 0.00204350937063846 -em 0.1384625 1 5 0.0060307968152018 -em 0.1384625 2 1 0 -em 0.1384625 2 3 0 -em 0.1384625 2 4 0 -em 0.1384625 2 5 0 -em 0.1384625 3 1 0.008704401 -em 0.1384625 3 2 0 -em 0.1384625 3 4 0.193855714034029 -em 0.1384625 3 5 0.203210242128393 -em 0.1384625 4 1 0.00878072013784504 -em 0.1384625 4 2 0 -em 0.1384625 4 3 0.00771007640703268 -em 0.1384625 4 5 0 -em 0.1384625 5 1 0.0115754432490942 -em 0.1384625 5 2 0 -em 0.1384625 5 3 1.30532712910343 -em 0.1384625 5 4 0 -en 0.1384625 3 8.542528 -ej

0.1384625 2 3 -en 0.21195 5 0.143861004427068 -en 0.214175 5 0.229328737963366 -en 0.41955 1 0.158405393915496 -en 0.42155 1 0.299481445247702 -en 0.473075 4 0.0306317427630759 -en 0.475075 4 2.79429564470655 -en 0.480625 4 0.0872103733618782 -em 0.4963375 1 3 0.04773644 -em 0.4963375 1 4 0.00204350937063846 -em 0.4963375 1 5 0.0060307968152018 -em 0.4963375 3 1 0.008704401 -em 0.4963375 3 4 0.193855714034029 -em 0.4963375 3 5 0.203210242128393 -em 0.4963375 4 1 0.00878072013784504 -em 0.4963375 4 3 0.00771007640703268 -em 0.4963375 4 5 0 -em 0.4963375 5 1 0.0115754432490942 -em 0.4963375 5 3 1.30532712910343 -em 0.4963375 5 4 0 -en 0.4984625 3 1.680466 -ej 0.4984625 5 4 -em 0.51205 1 3 0.00187056444537476 -em 0.51205 1 4 0 -em 0.51205 3 1 0.00144484023958745 -em 0.51205 3 4 0 -em 0.51205 4 1 0 -em 0.51205 4 3 0 -en 0.51205 4 0.163429675293312 -ej 0.51205 4 3 -em 1.515208 1 3 0 -em 1.515208 3 1 0 -en 1.515208 3 1.023144 -ej 1.515208 1 3

In order to assess the fit of our combined model to the real sequenced data, we used the program dadi-1.7.0 (Gutenkunst et al. 2009). To calculate allele frequencies for the four subspecies of chimpanzee, the whole genomes of these individuals were extracted from the vcf file, we removed sites where no ancestral information was available to create an unfolded frequency spectrum file, closely following the method described in Gutenkunst et al. (2009). To calculated allele frequencies for each model, we created 1,000,000 simulations of 1 theta region (close to 3,000 bp) in order to obtain the average site frequency spectrum (SFS) for a 3,000 bp region. We then extrapolated allele counts for the derived allele to reach 3,000,000,000 bp, to match the size of real genome data (allele frequency does not change with the length of genome, enabling us to extrapolate from short simulations). We next removed fixed sites from both the real and simulated data, as these sites can be biased due to the coverage of the sequenced data and again compared to the simulations to see how the middle of the allele frequency spectrum fits without the potential bias of fixed sites. We compared the SFS of the real data with the simulations from each model and calculated the largest deviation between the two SFSs. This allows us to visualize how each demographic model correctly captures the diversity found in these four populations. Our target was to see if by

changing parameters our prediction of allele frequency is changed substantially.

We find, when comparing the genome-wide SFS, our model differs from the previously published models very negligibly. Supplementary Table S2.1, shows the greatest count difference between the simulated model and the real data. In other words, out of the full spectrum SFS the previously published models differ by 1100 and 1000 singletons, respectively, whereas our model differs between 1200 and 1100 out of a total of slightly more than 10^6 singletons. The difference between the pruned spectrum and the real data is even smaller (Supplementary Table S2.1). Indeed, the pruned spectrum may be a better assessment of the model's fit because singletons are more prone to be a result of sequencing error. For the pruned spectrum the previously published models differ by 79 and 240 respectively, while Model 3 differs by 160. We would like to point out, that for *P.t. verus*, Model 3 improves the fit, which is apparent when looking at these data graphically (Supplemental Figure S2.1). We conclude that our combined model is robustly similar to the separated models i.e., an error rate of 0.11% as compared to 0.12% for the full spectrum.

Lastly, in order to simulate different selection scenarios using a coalescent simulator (see below), we needed to assess whether or not the recent bottleneck present in all the above models has a significant impact on the current SFS of the populations. We removed the recent bottleneck present in model 3 to form model 4.

Model 4 (1=bonobo, 2=P.t. troglodytes, 3=P.t. schweinfurthii, 3=P.t. ellioti, and 4=P.t. verus):

-N 10000 -ms 118 1000000 -t 1.44 -r 1.2 3000 -I 5 0 38 36 20 24 0.0 -n 1 1.701939 -n 2 1.282871 -n 3 3.992509 -n 4 1.23640124358 -n 5 1.0742217327 -em 0.0687125 1 2 0 -em 0.0687125 1 3 0 -em 0.0687125 1 4 0 -em 0.0687125 1 5 0 -em 0.0687125 2 1 0 -em 0.0687125 2 3 1.542334 -em 0.0687125 2 4 1.12888460726286 -em 0.0687125 2 5 0.79048180986804 -em 0.0687125 3 1 0 -em 0.0687125 3 2 2.143439 -em 0.0687125 3 4 0.514005225416364 -em 0.0687125 3 5 0.00440335360192708 -em 0.0687125 4 1 0 -em 0.0687125 4 2

0.61034918826118 -em 0.0687125 4 3 2.77081002950074 -em 0.0687125 4 5 0 -em 0.0687125 5 1 0 -em 0.0687125 5 2 0.00666829030333492 -em 0.0687125 5 3 0.108325487900296 -em 0.0687125 5 4 0 -em 0.0812625 1 2 0 -em 0.0812625 1 3 0.04260769 -em 0.0812625 1 4 0.00204350937063846 -em 0.0812625 1 5 0.0060307968152018 -em 0.0812625 2 1 0 -em 0.0812625 2 3 1.542334 -em 0.0812625 2 4 1.12888460726286 -em 0.0812625 2 5 0.79048180986804 -em 0.0812625 3 1 0.0514705 -em 0.0812625 3 2 2.143439 -em 0.0812625 3 4 0.514005225416364 -em 0.0812625 3 5 0.00440335360192708 -em 0.0812625 4 1 0.00878072013784504 -em 0.0812625 4 2 0.61034918826118 -em 0.0812625 4 3 2.77081002950074 -em 0.0812625 4 5 0 -em 0.0812625 5 1 0.0115754432490942 -em 0.0812625 5 2 0.00666829030333492 -em 0.0812625 5 3 0.108325487900296 -em 0.0812625 5 4 0 -en 0.13635 2 0.1065495 -en 0.13635 3 0.1751349 -em 0.1384625 1 2 0 -em 0.1384625 1 3 0.04773644 -em 0.1384625 1 4 0.00204350937063846 -em 0.1384625 1 5 0.0060307968152018 -em 0.1384625 2 1 0 -em 0.1384625 2 3 0 -em 0.1384625 2 4 0 -em 0.1384625 2 5 0 -em 0.1384625 3 1 0.008704401 -em 0.1384625 3 2 0 -em 0.1384625 3 4 0.193855714034029 -em 0.1384625 3 5 0.203210242128393 -em 0.1384625 4 1 0.00878072013784504 -em 0.1384625 4 2 0 -em 0.1384625 4 3 0.00771007640703268 -em 0.1384625 4 5 0 -em 0.1384625 5 1 0.0115754432490942 -em 0.1384625 5 2 0 -em 0.1384625 5 3 1.30532712910343 -em 0.1384625 5 4 0 -en 0.1384625 3 8.542528 -ej 0.1384625 2 3 -en 0.21195 5 0.143861004427068 -en 0.214175 5 0.229328737963366 -en 0.41955 1 0.158405393915496 -en 0.42155 1 0.299481445247702 -en 0.473075 4 0.0306317427630759 -en 0.475075 4 2.79429564470655 -en 0.480625 4 0.0872103733618782 -em 0.4963375 1 3 0.04773644 -em 0.4963375 1 4 0.00204350937063846 -em 0.4963375 1 5 0.0060307968152018 -em 0.4963375 3 1 0.008704401 -em 0.4963375 3 4 0.193855714034029 -em 0.4963375 3 5 0.203210242128393 -em 0.4963375 4 1 0.00878072013784504 -em 0.4963375 4 3 0.00771007640703268 -em 0.4963375 4 5 0 -em 0.4963375 5 1 0.0115754432490942 -em 0.4963375 5 3 1.30532712910343 -em 0.4963375 5 4 0 -en 0.4984625 3 1.680466 -ej 0.4984625 5 4 -em 0.51205 1 3 0.00187056444537476 -em 0.51205 1 4 0 -em 0.51205 3 1 0.00144484023958745 -em 0.51205 3 4 0 -em 0.51205 4 1 0 -em 0.51205 4 3 0 -en 0.51205 4 0.163429675293312 -ej 0.51205 4 3 -em 1.515208 1 3 0 -em 1.515208 3 1 0 -en 1.515208 3 1.023144 -ej 1.515208 1 3

When comparing the SFS between a model with and without a recent bottleneck, we see (Supplementary Table S2.1) a higher deviation of singletons. This deviation however is still negligible when considering there are more than 10^6 total singletons genome-wide. We conclude that we are able to simulate selection without

the recent bottleneck. For a comparison of the fit of the models to the real data, we present the deviations in Supplemental Figure S2.1.

The real-world estimates represented by Model 4 of the five populations are that *P.t. troglodytes* and *P.t. schweinfurthii* diverge from each other 136,350 years ago (ya), *P.t. ellioti* and *P.t. verus* diverge 498,462 ya, the ancestors of the two major branches of chimpanzee split 512,050 ya, and the ancestor of the chimpanzee diverges from bonobo 1,515,208 ya. Disregarding the bottleneck 1,063 ya, we estimate the present-day real-world population sizes at 17,019 for bonobo, 39,925 for *P.t. troglodytes*, 12,829 for *P.t. schweinfurthii*, 12,364 for *P.t. ellioti*, and 10,742 for *P.t. verus.*

c) Random Forest Algorithm

We employed a random forest method. This method is an extension of the classic machine learning approach, the decision tree. In a decision tree, the user provides the structure of a tree and decides which variables should be discerned at each node. This method can be biased by both the user-derived tree structure, and, most importantly, runs the risk of overfitting when the training set includes correlated variables. In our case, the 15 statistics (19 when cross-population tests are counted separately) are highly correlated (Supplemental Figure S2.44). Random forests avoid the risk of over-fitting by creating decision trees of random structures. This way, overfitting is avoided due to the Strong Law of Large Numbers (Breiman 2001).

We tested the complexity of random tree structures. Because we have simulations every 300 generations between 600 and 2400 generations ago as well as both complete and incomplete selection scenarios, we wanted to see if we could truly differentiate the type and time of selection. We tested 3 tree structures: the 5-node tree

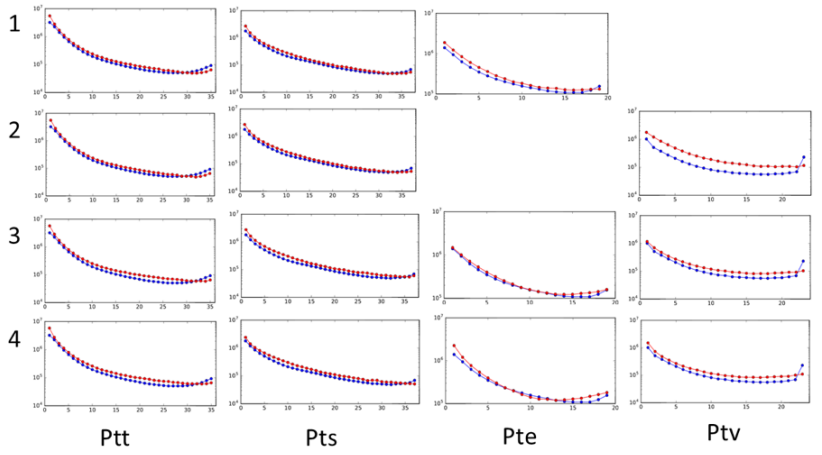
(neutral, recent incomplete (600 generations ago and FAF between 0.6 and less than 1), recent complete (600 generations ago and FAF equal to 1), ancient incomplete (2400 generations ago and FAF between 0.6 and less than 1), and ancient complete (2400 generations ago and FAF equal to 1)), 3-node tree (neutral, recent (600, 900, and 1200 generations ago and both incomplete and complete sweeps), and ancient (1800, 2100, and 2400 generations ago and both incomplete and complete sweeps)), and a 2-node tree (neutral and selection (600, 900, 1200, 1500, 1800, 2100, and 2400 generations ago both complete and incomplete)).

The importance of each statistic for the differing tree topology regressions are presented in Supplemental Figures S2.46-S2.49. The mean decrease in accuracy is the global variable importance for each statistic over all out-of-bag predictions after training. The mean decrease in gini is a measure of the gain of purity for each statistic (Breiman 2001). Although we observe variation across subspecies and tree structures, the most important statistics in general are pi, XPEHH, and Tajima's D. We present the error rates (Supplementary Table S2.6) for each of the three tree structures, where the OOB error rate is the error rate of the training set that was not bagged during bootstrapping, false positives are neutral cases falsely classified as selection, false negatives are selection cases falsely classified as neutral, and the confusion rate is when one type of selection is misclassified as another (i.e. recent incomplete for recent complete).

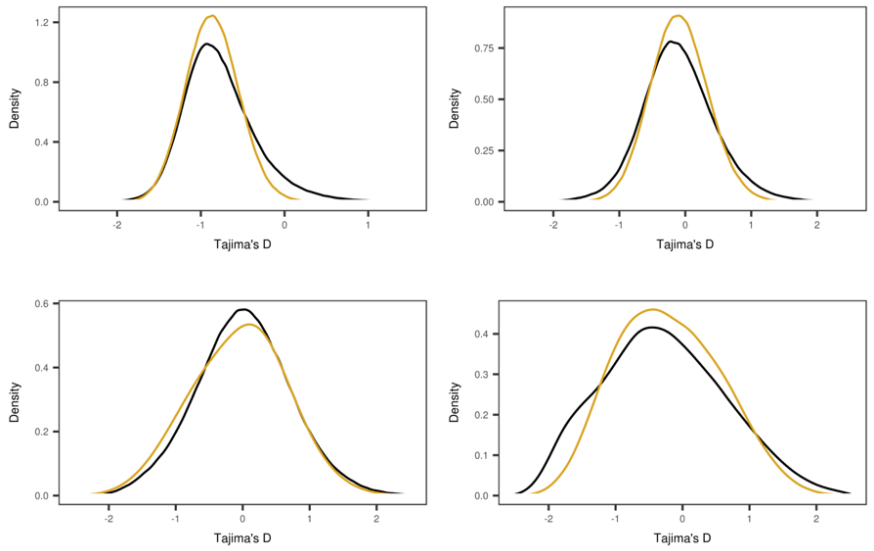
It is clear from the error rates that although the lowest false positive rate is measured in the 5-node tree, the confusion rates for both the 5-node and 3-node trees are high. This indicates that the real difference is between the neutral and selective scenarios and not between the types of selection. This is likely a by-product of the fact that the times of selection are arbitrary in the sense that they are set at times where our demographic model is time-invariant instead of at times based on biological significance. For this reason,

coupled with the fact that the 2-node tree is trained with the maximum amount of selective scenarios, we present results from the 2-node tree.

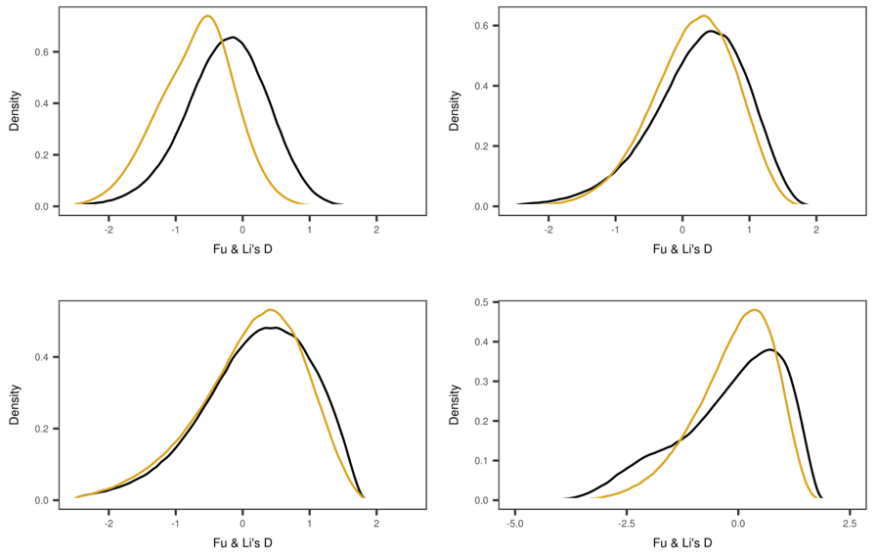
2.8 Supplementary Figures



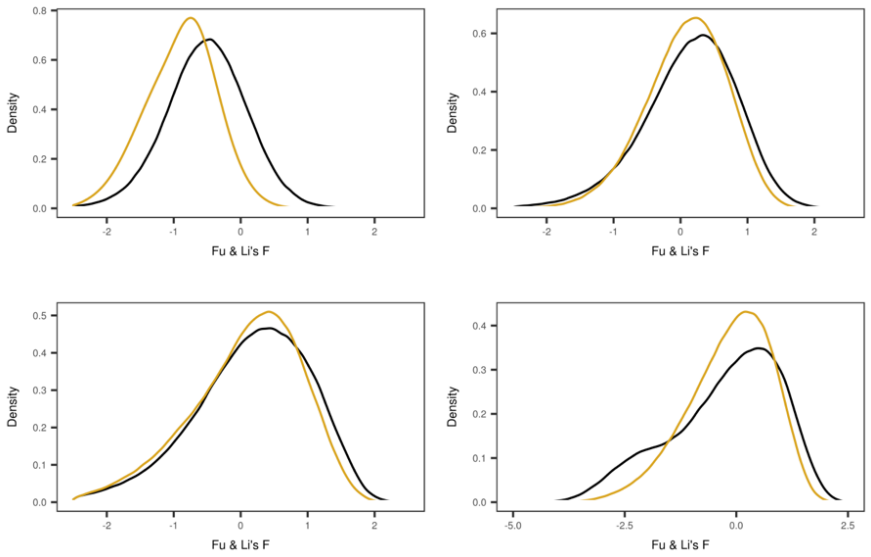
Supplementary Figure S2.1: Comparison between full genome site frequency spectrum of each subspecies (blue) with neutral simulations (red) for each of the four demographic models (Model 1 – Model 4). P.t.e. is *Pan troglodytes ellioti*, P.t.s. is *Pan troglodytes schweinfurthii*, P.t.t. is *Pan troglodytes troglodytes*, and P.t.v. is *Pan troglodytes verus*.



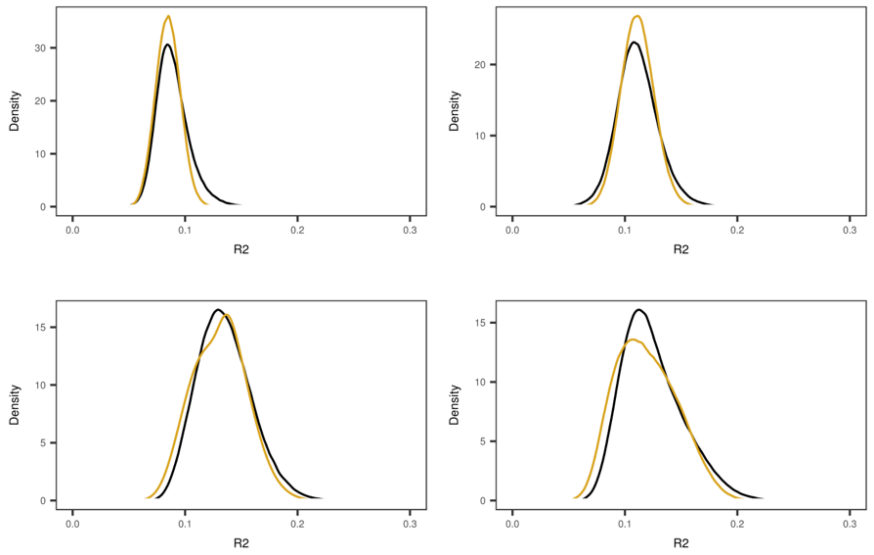
Supplementary Figure S2.2: Genome-wide distribution of Tajima's D (black) as compared with the neutral simulations (yellow). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



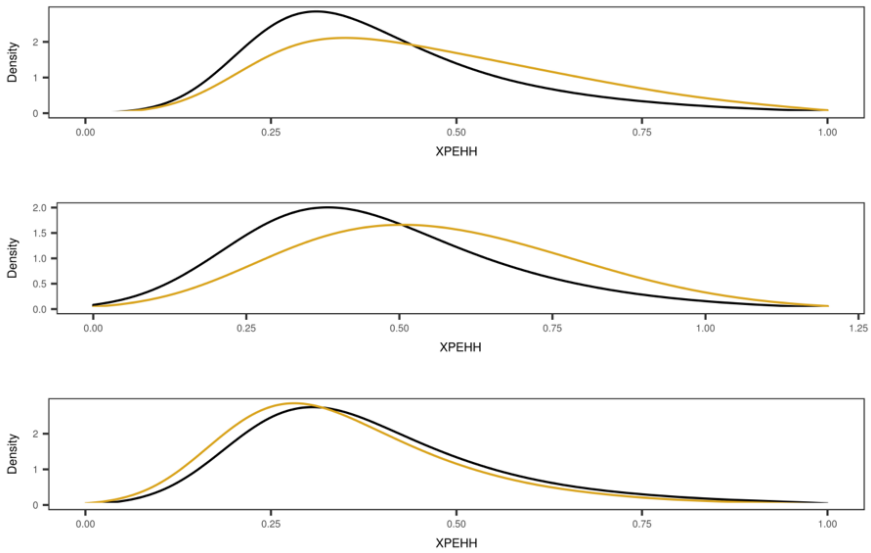
Supplementary Figure S2.3: Genome-wide distribution of Fu and Li's D (black) as compared with the neutral simulations (yellow). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



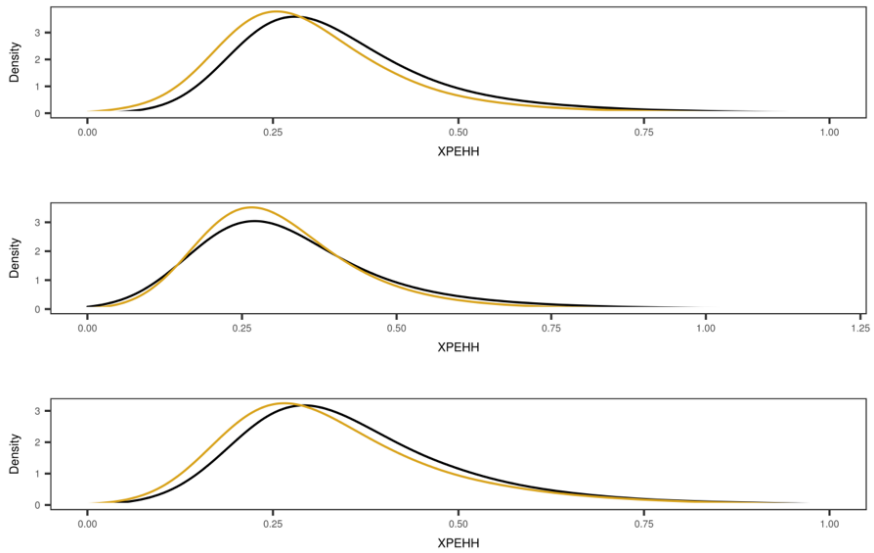
Supplementary Figure S2.4: Genome-wide distribution of Fu and Li's F (black) as compared with the neutral simulations (yellow). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



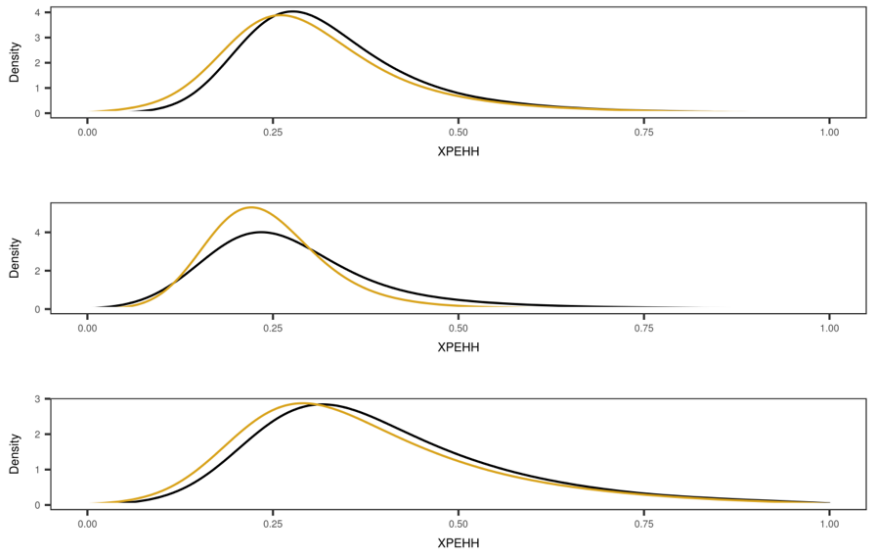
Supplementary Figure S2.5: Genome-wide distribution of R_2 (black) as compared with the neutral simulations (yellow). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



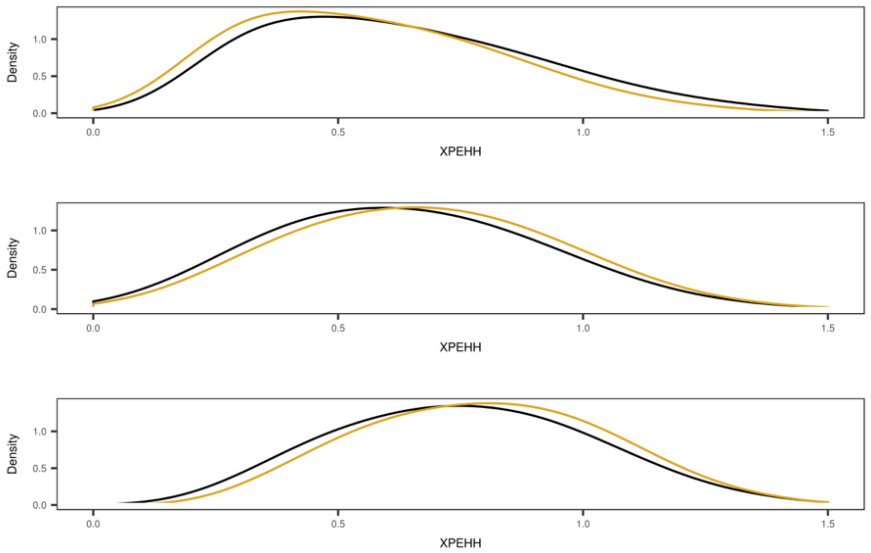
Supplementary Figure S2.6: Genome-wide distribution of XPEHH of *P.t. ellioti* (black) as compared with the neutral simulations (yellow). The subspecies are top *P.t. troglodytes*, middle *P.t. schweinfurthii*, and bottom *P.t. verus.*



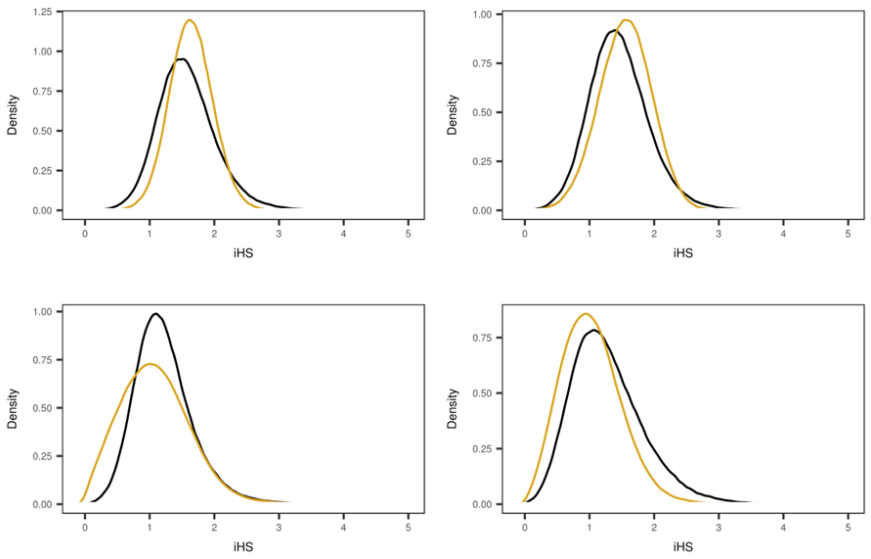
Supplementary Figure S2.7: Genome-wide distribution of XPEHH of *P.t. schweinfurthii* (black) as compared with the neutral simulations (yellow). The subspecies are top *P.t. troglodytes*, middle *P.t. ellioti*, and bottom *P.t. verus.*



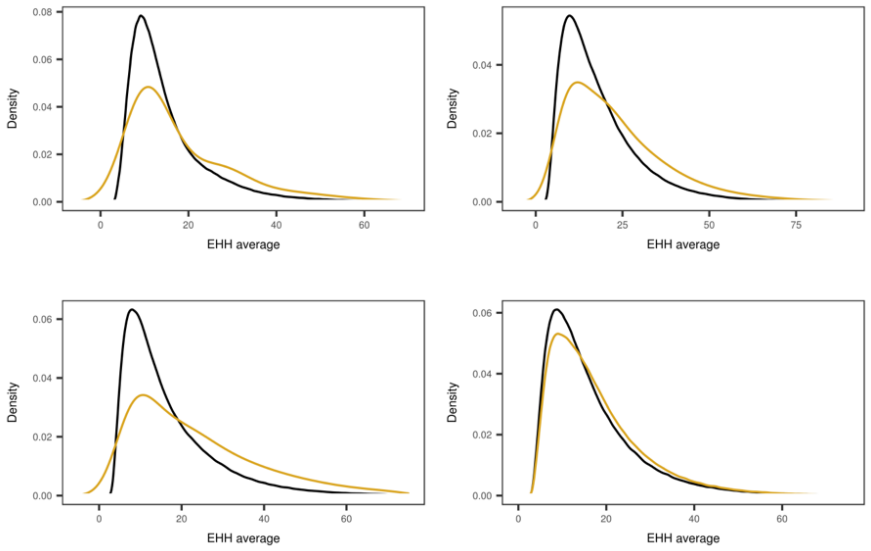
Supplementary Figure S2.8: Genome-wide distribution of XPEHH of *P.t. troglodytes* (black) as compared with the neutral simulations (yellow). The subspecies are top *P.t. schweinfurthii*, middle *P.t. ellioti*, and bottom *P.t. verus.*



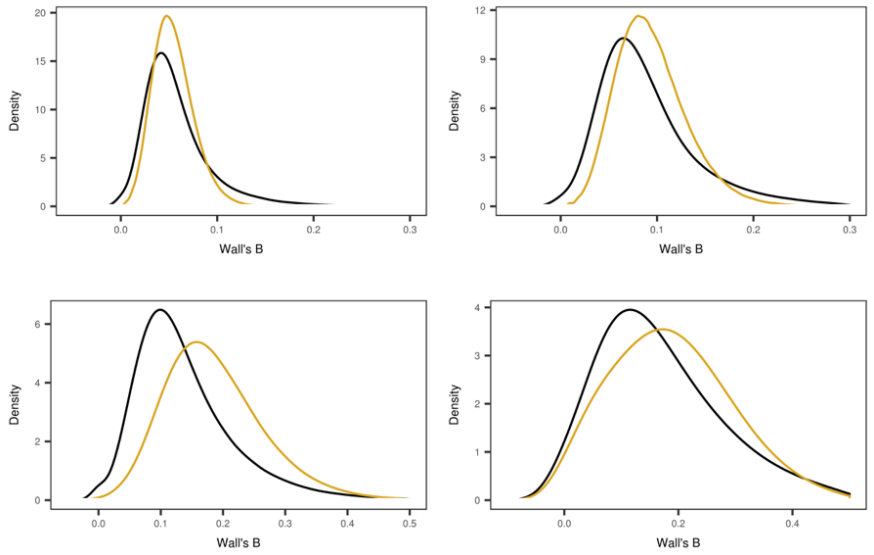
Supplementary Figure S2.9: Genome-wide distribution of XPEHH of *P.t. verus* (black) as compared with the neutral simulations (yellow). The subspecies are top *P.t. schweinfurthii*, middle *P.t. ellioti*, and bottom *P.t. troglodytes.*



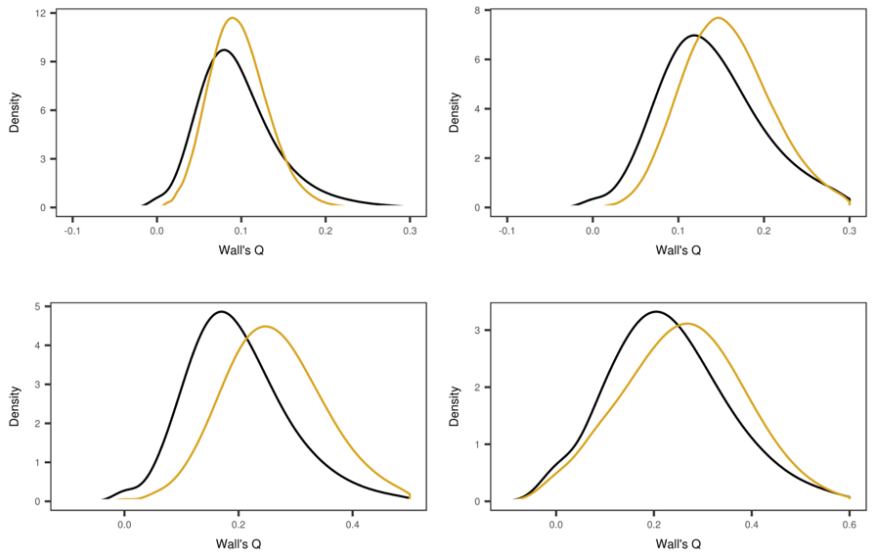
Supplementary Figure S2.10: Genome-wide distribution of iHS (black) as compared with the neutral simulations (yellow). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



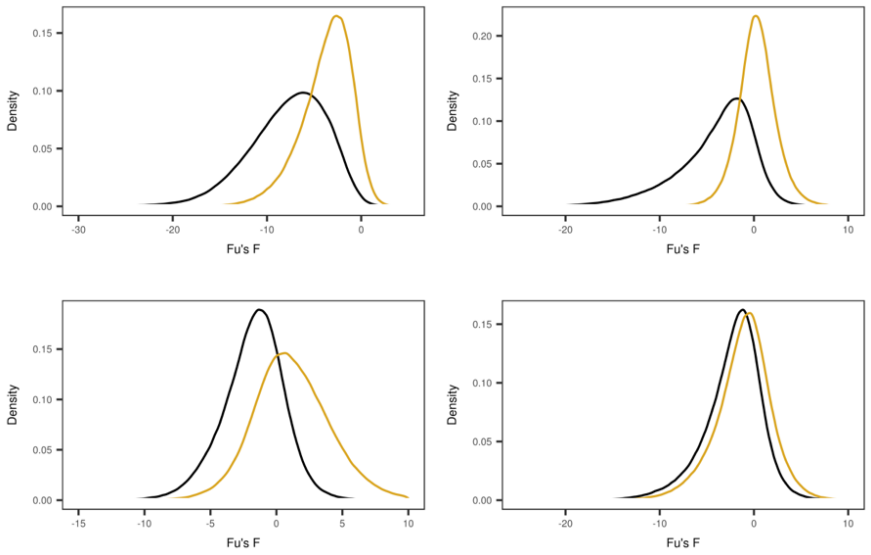
Supplementary Figure S2.11: Genome-wide distribution of $EHH_{Average}$ (black) as compared with the neutral simulations (yellow). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



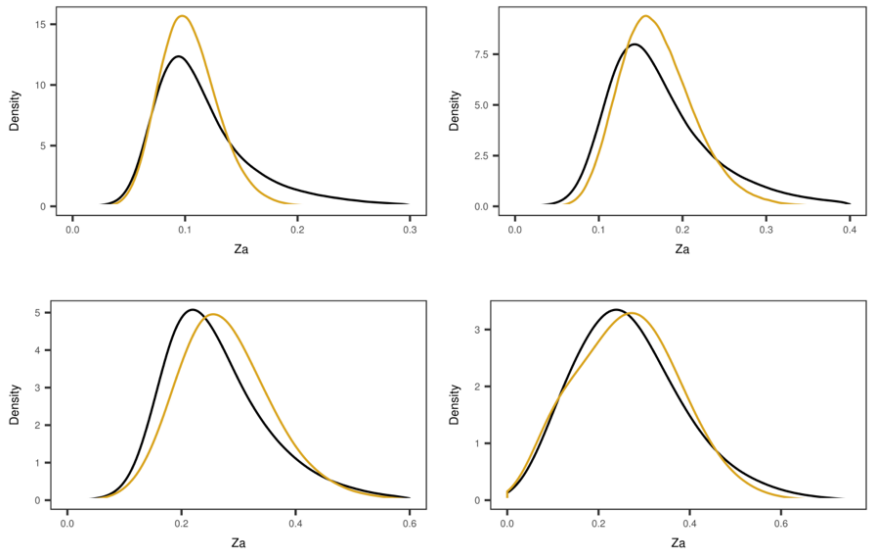
Supplementary Figure S2.12: Genome-wide distribution of Wall's B (black) as compared with the neutral simulations (yellow). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



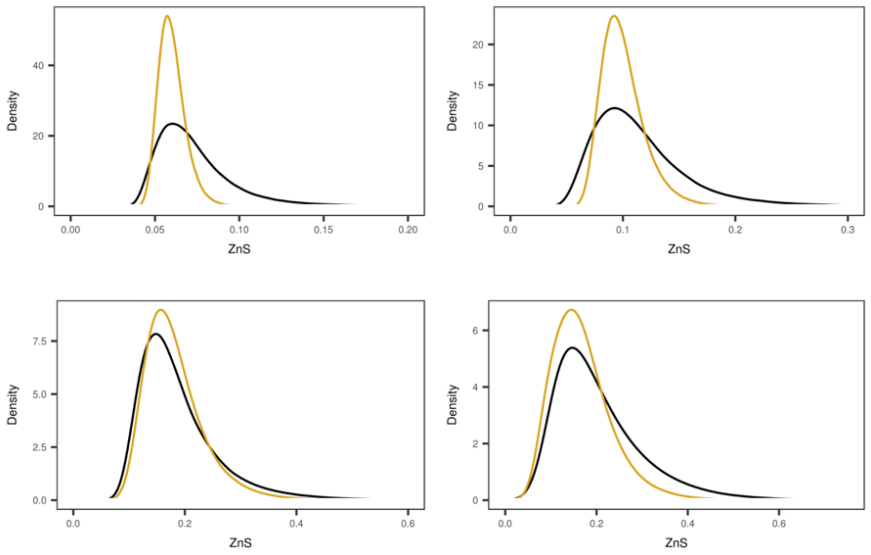
Supplementary Figure S2.13: Genome-wide distribution of Wall's Q (black) as compared with the neutral simulations (yellow). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



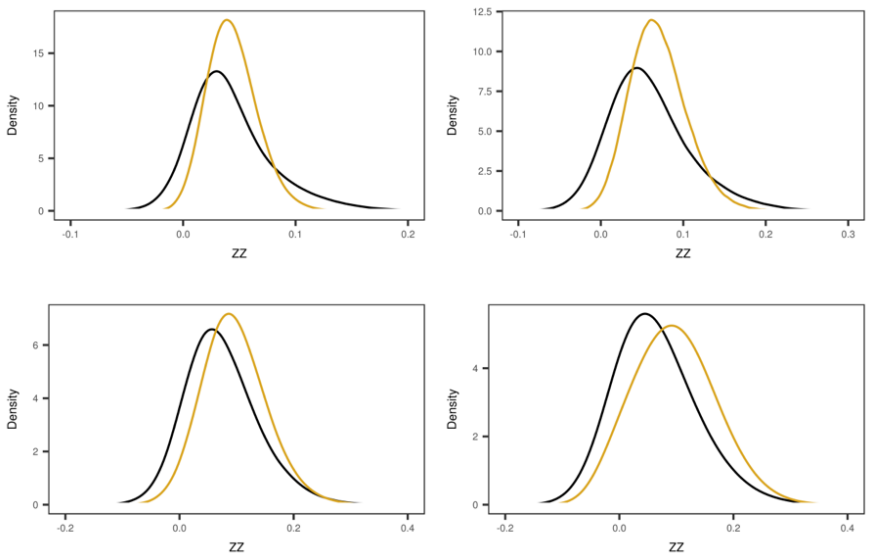
Supplementary Figure S2.14: Genome-wide distribution of Fu's F (black) as compared with the neutral simulations (yellow). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



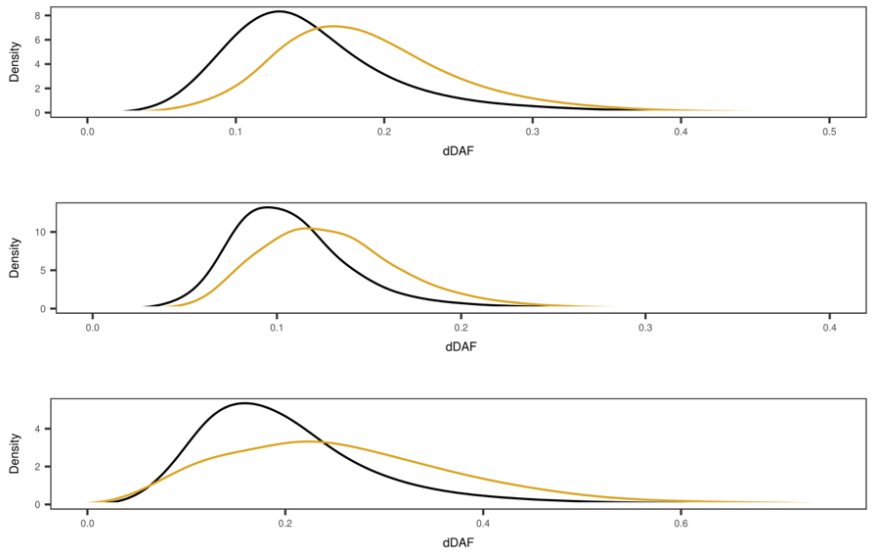
Supplementary Figure S2.15: Genome-wide distribution of Z_a (black) as compared with the neutral simulations (yellow). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



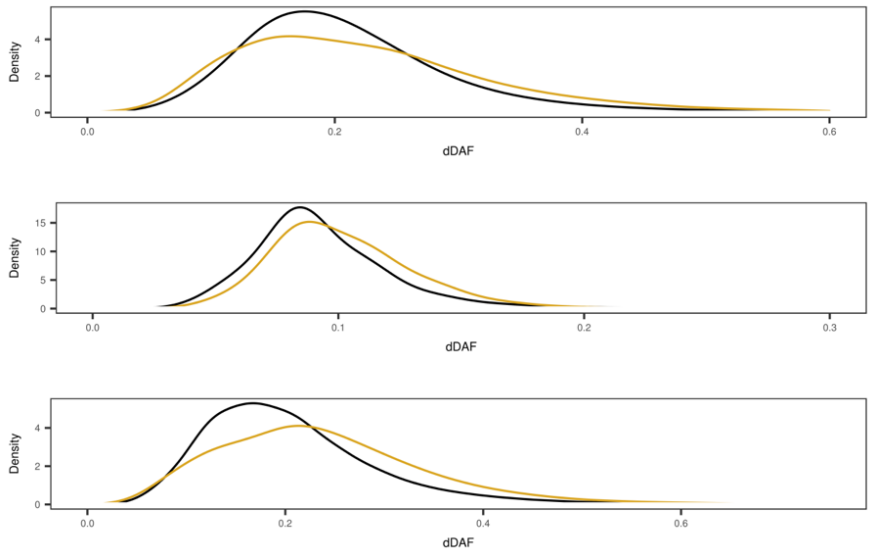
Supplementary Figure S2.16: Genome-wide distribution of Z_{nS} (black) as compared with the neutral simulations (yellow). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



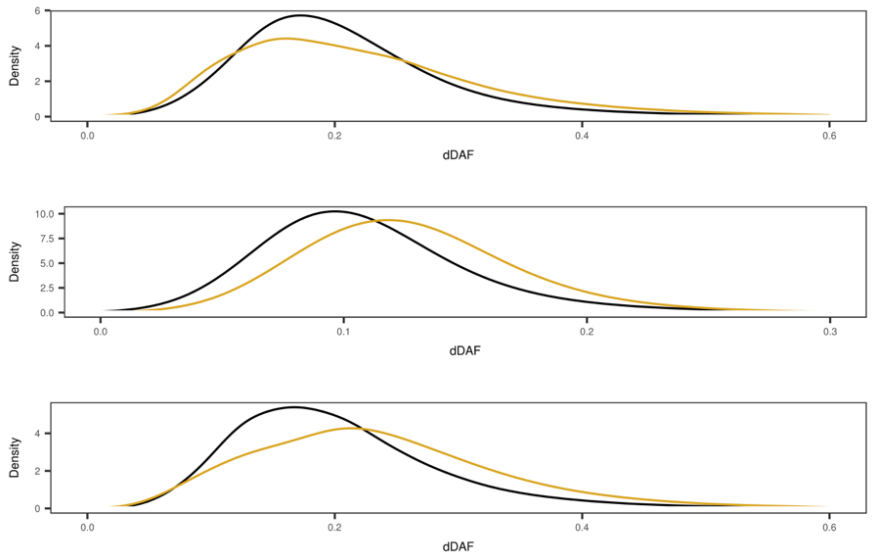
Supplementary Figure S2.17: Genome-wide distribution of ZZ (black) as compared with the neutral simulations (yellow). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



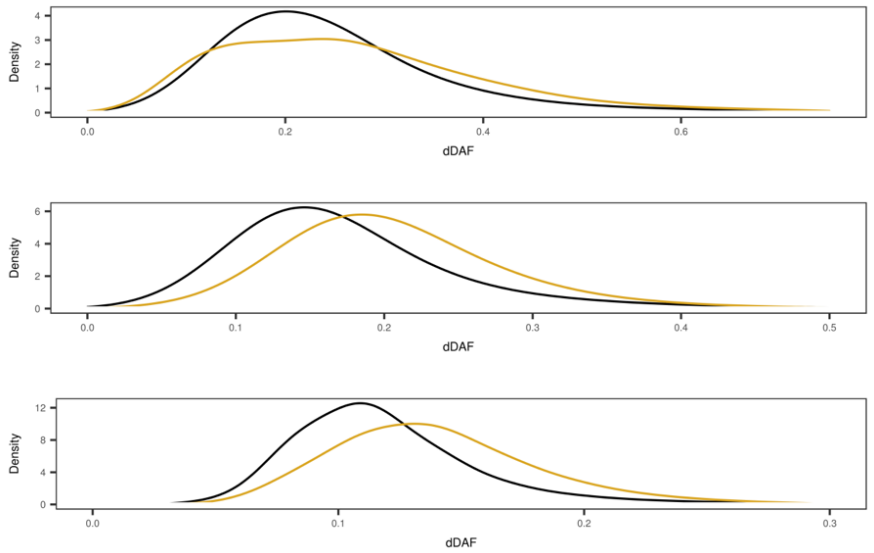
Supplementary Figure S2.18: Genome-wide distribution of ΔDAF of *P.t. ellioti* (black) as compared with the neutral simulations (yellow). The subspecies are top *P.t. troglodytes*, middle *P.t. schweinfurthii*, and bottom *P.t. verus.*



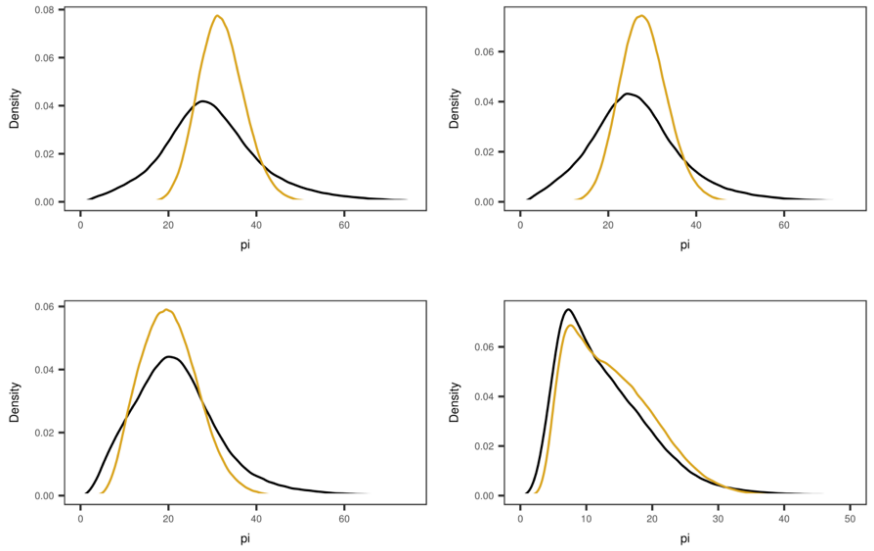
Supplementary Figure S2.19: Genome-wide distribution of ΔDAF of *P.t. schweinfurthii* (black) as compared with the neutral simulations (yellow). The subspecies are top *P.t. troglodytes*, middle *P.t. ellioti*, and bottom *P.t. verus.*



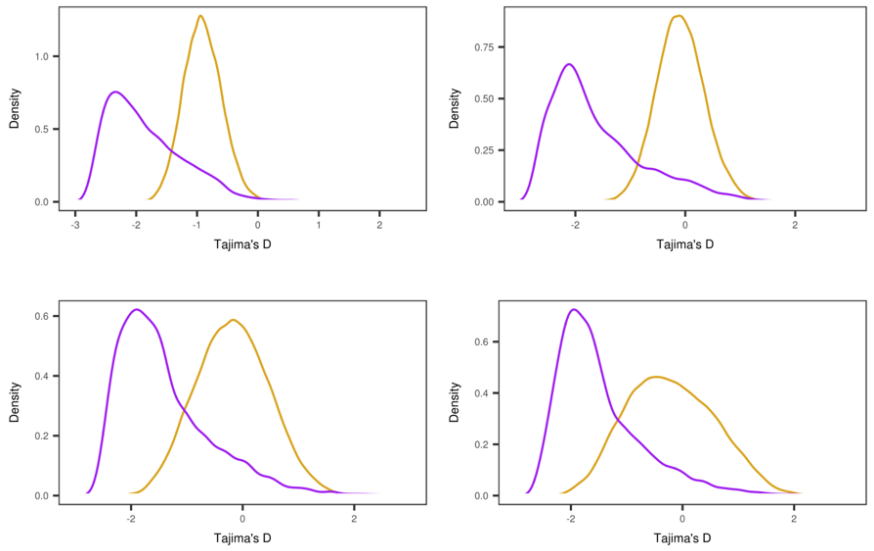
Supplementary Figure S2.20: Genome-wide distribution of ΔDAF of *P.t. schweinfurthii* (black) as compared with the neutral simulations (yellow). The subspecies are top *P.t. troglodytes*, middle *P.t. ellioti*, and bottom *P.t. verus.*



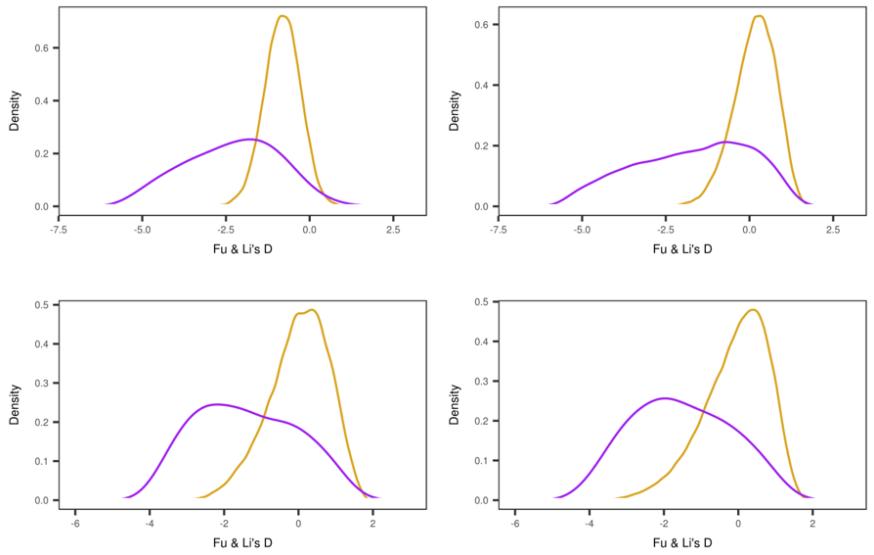
Supplementary Figure S2.21: Genome-wide distribution of ΔDAF of *P.t. troglodytes* (black) as compared with the neutral simulations (yellow). The subspecies are top *P.t. schweinfurthii*, middle *P.t. ellioti*, and bottom *P.t. verus*.



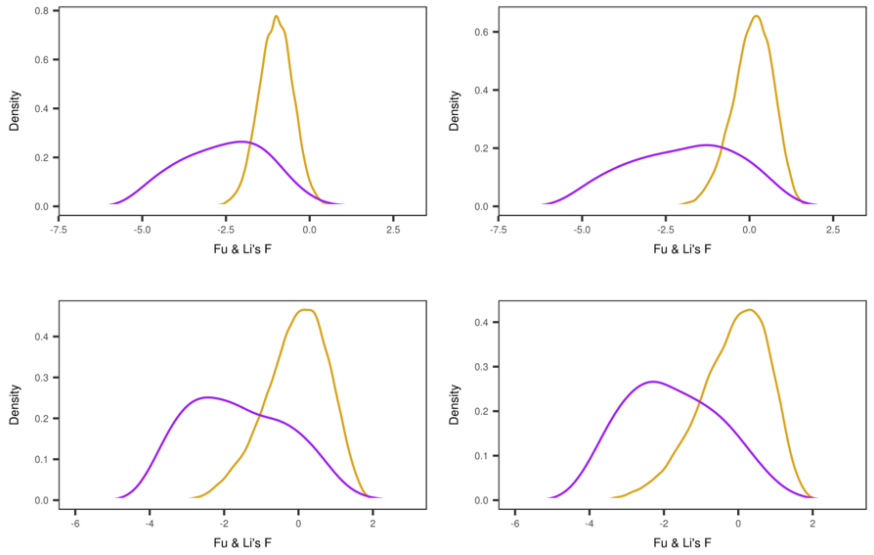
Supplementary Figure S2.22: Genome-wide distribution of ΔDAF of *P.t. verus* (black) as compared with the neutral simulations (yellow). The subspecies are top *P.t. schweinfurthii*, middle *P.t. ellioti*, and bottom *P.t. troglodytes.*



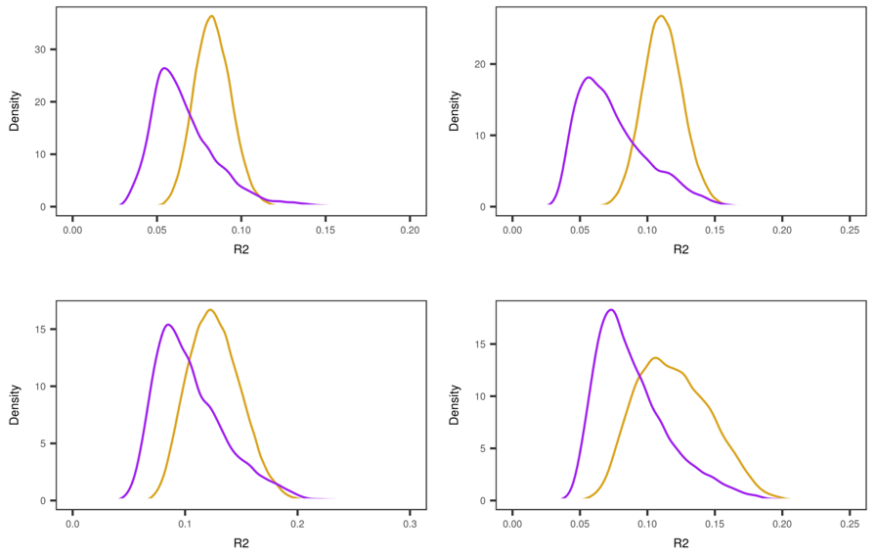
Supplementary Figure S2.23: Genome-wide distribution of Tajima's D in simulated scenarios for both the neutral distribution (yellow) and selective scenarios (purple). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



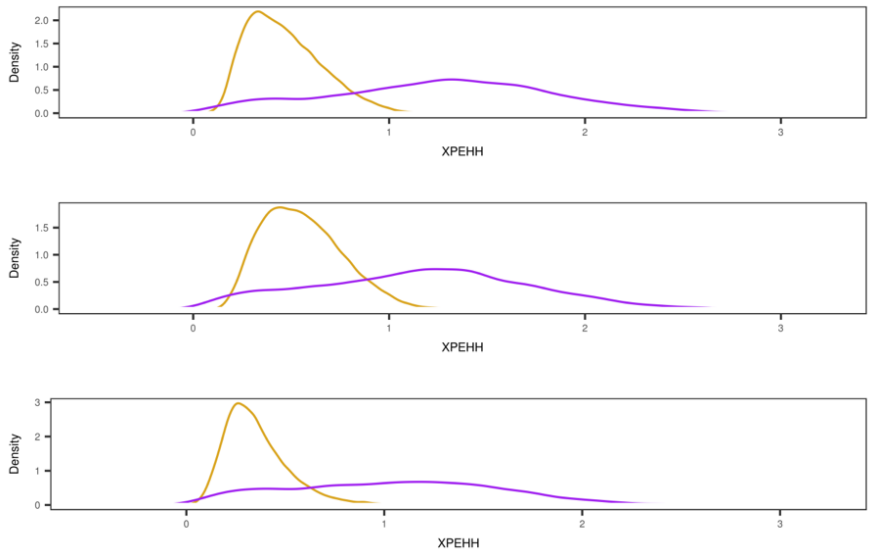
Supplementary Figure S2.24: Genome-wide distribution of Fu and Li's D in simulated scenarios for both the neutral distribution (yellow) and selective scenarios (purple). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



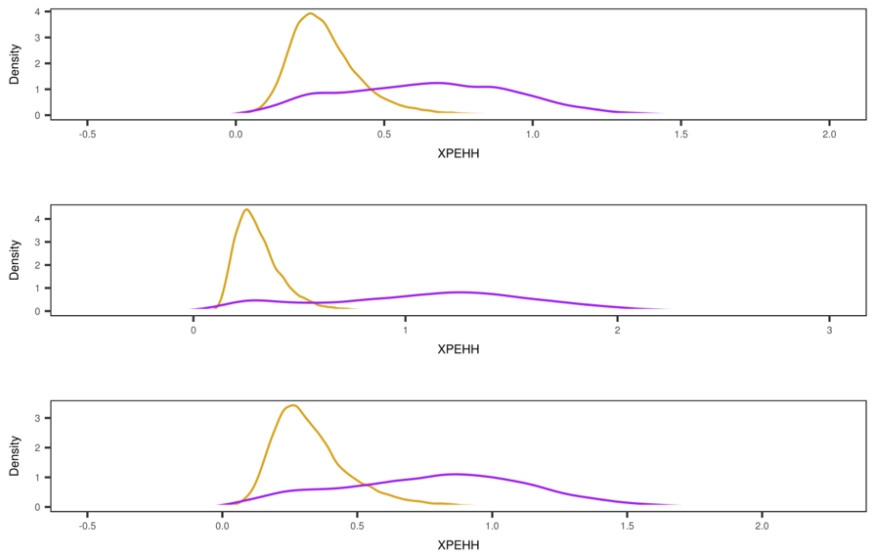
Supplementary Figure S2.25: Genome-wide distribution of Fu and Li's F in simulated scenarios for both the neutral distribution (yellow) and selective scenarios (purple). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



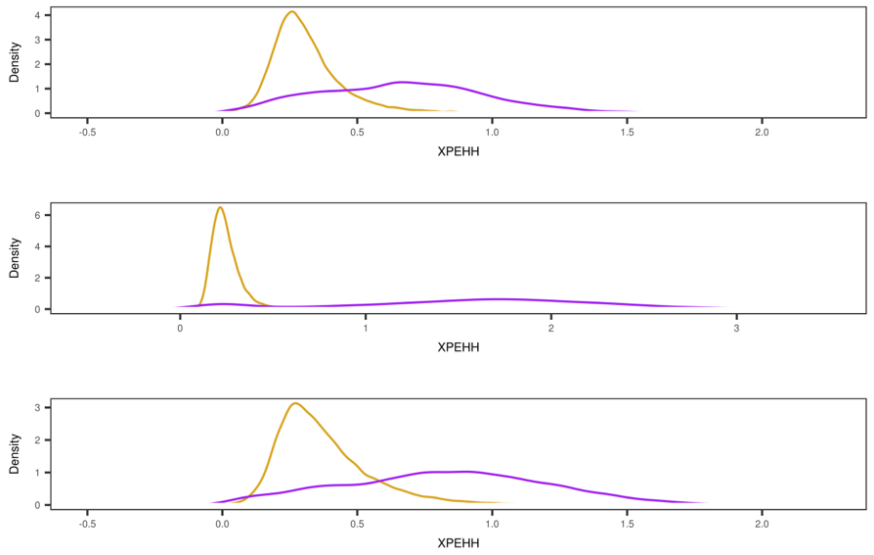
Supplementary Figure S2.26: Genome-wide distribution of R_2 in simulated scenarios for both the neutral distribution (yellow) and selective scenarios (purple). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



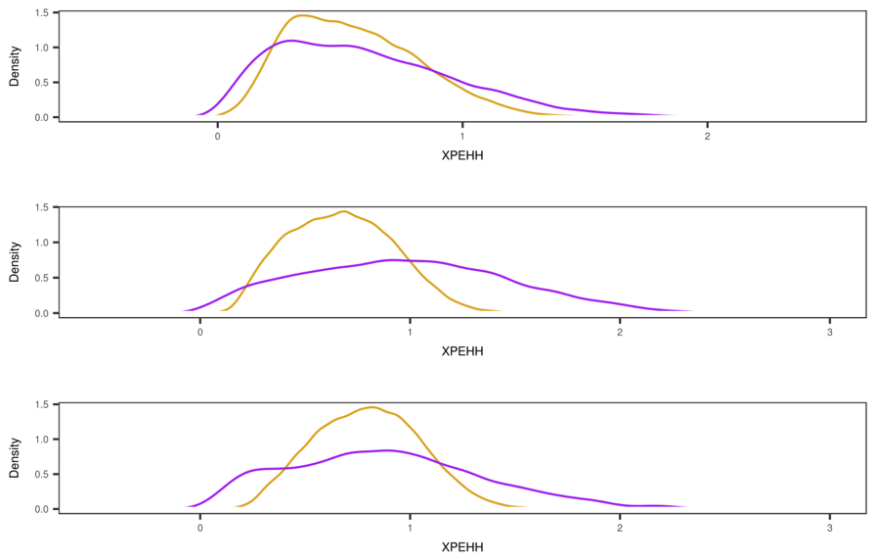
Supplementary Figure S2. 27: Genome-wide distribution of XPEHH for *P.t. ellioti* across all simulated scenarios, the neutral distribution is in yellow and selective scenarios are in purple. The subspecies are top *P.t. troglodytes*, middle *P.t. schweinfurthii*, and bottom *P.t. verus*.



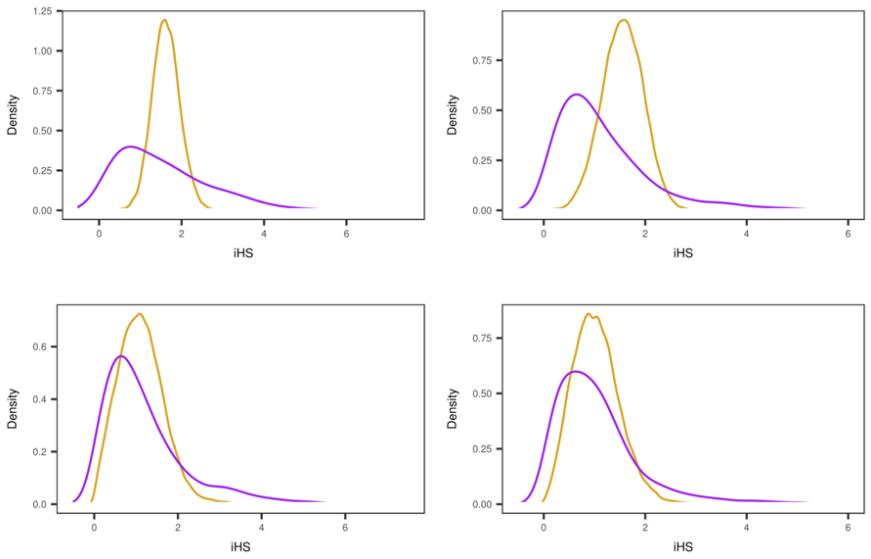
Supplementary Figure S2.28: Genome-wide distribution of XPEHH for *P.t. schweinfurthii* across all simulated scenarios, the neutral distribution is in yellow and selective scenarios are in purple. The subspecies are top *P.t. troglodytes*, middle *P.t. ellioti*, and bottom *P.t. verus*.



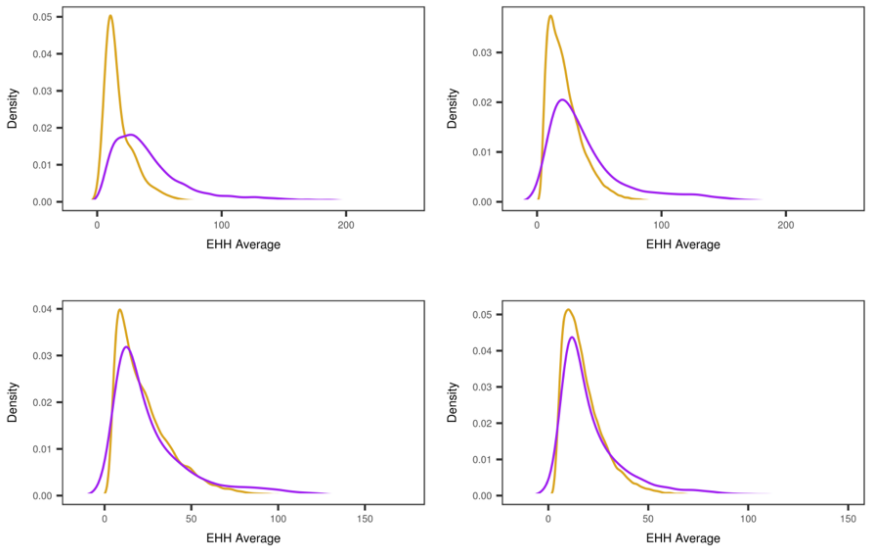
Supplementary Figure S2.29: Genome-wide distribution of XPEHH for *P.t. troglodytes* across all simulated scenarios, the neutral distribution is in yellow and selective scenarios are in purple. The subspecies are top *P.t. schweinfurthii*, middle *P.t. ellioti*, and bottom *P.t. verus*.



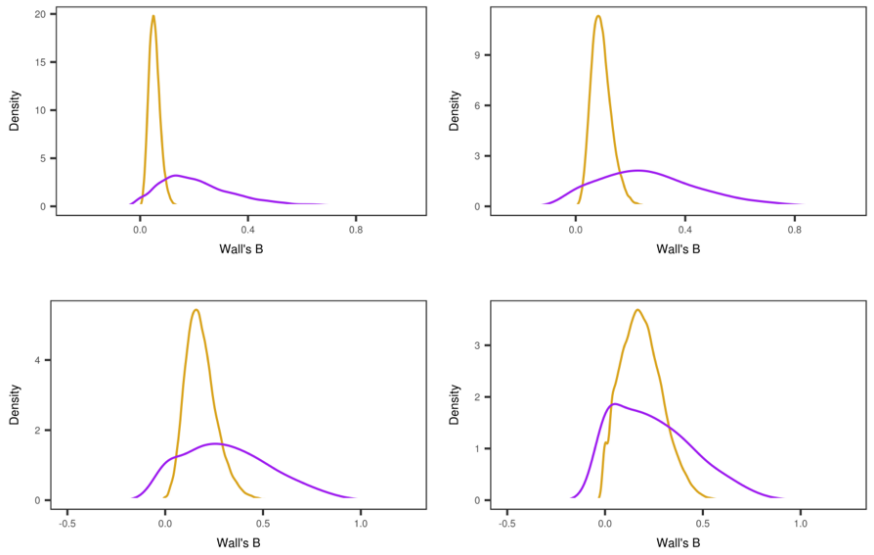
Supplementary Figure S2.30: Genome-wide distribution of XPEHH for *P.t. verus* across all simulated scenarios, the neutral distribution is in yellow and selective scenarios are in purple. The subspecies are top *P.t. schweinfurthii*, middle *P.t. ellioti*, and bottom *P.t. troglodytes*.



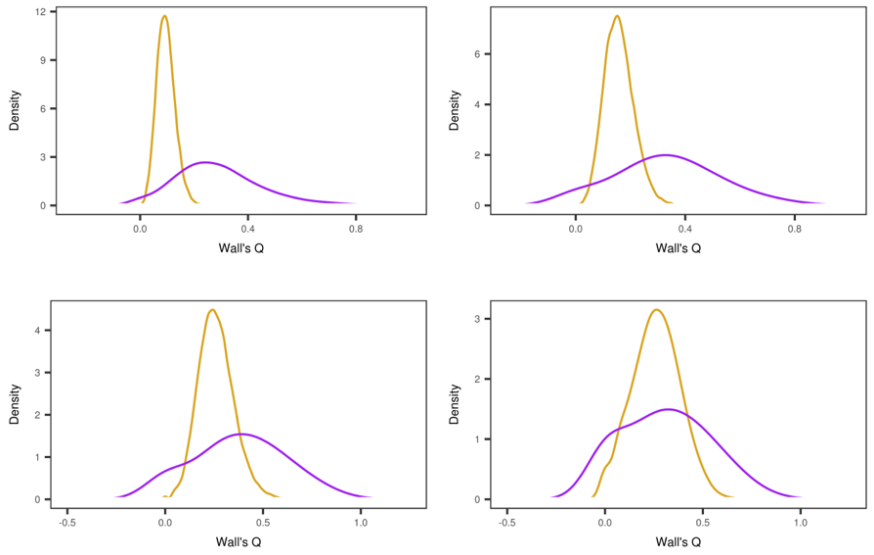
Supplementary Figure S2.31: Genome-wide distribution of iHS in simulated scenarios for both the neutral distribution (yellow) and selective scenarios (purple). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



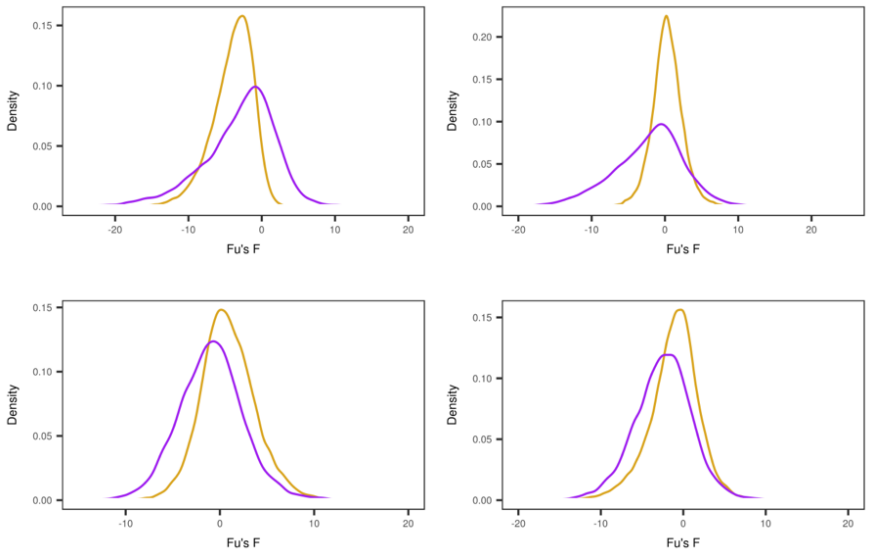
Supplementary Figure S2.32: Genome-wide distribution of $EHH_{Average}$ in simulated scenarios for both the neutral distribution (yellow) and selective scenarios (purple). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



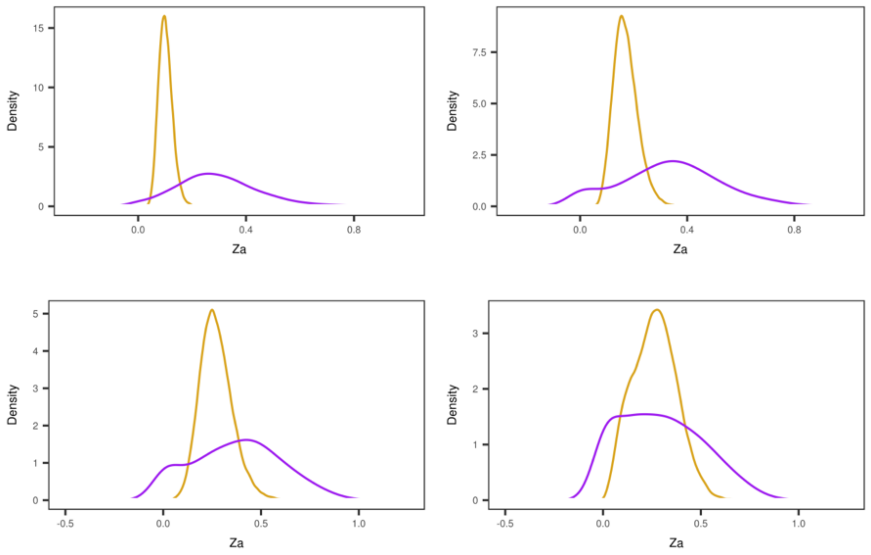
Supplementary Figure S2.33: Genome-wide distribution of Wall's B in simulated scenarios for both the neutral distribution (yellow) and selective scenarios (purple). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



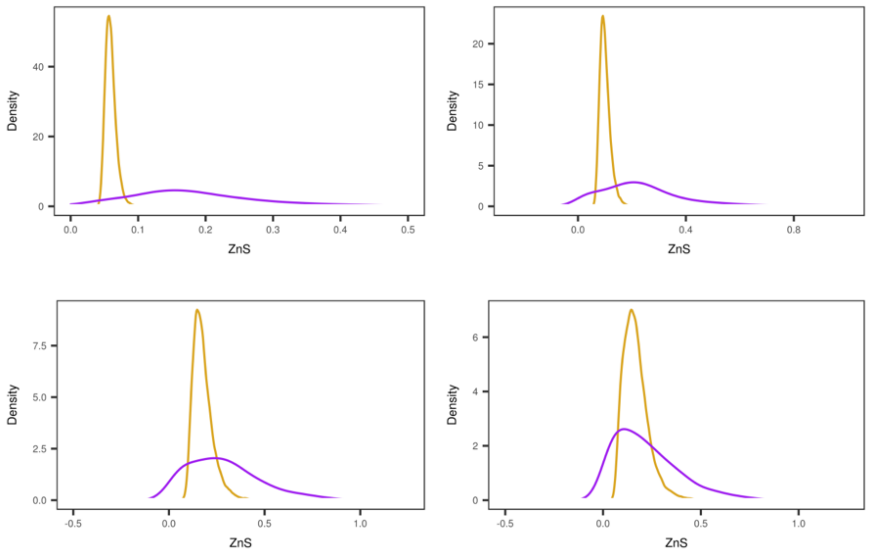
Supplementary Figure S2.34: Genome-wide distribution of Wall's Q in simulated scenarios for both the neutral distribution (yellow) and selective scenarios (purple). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



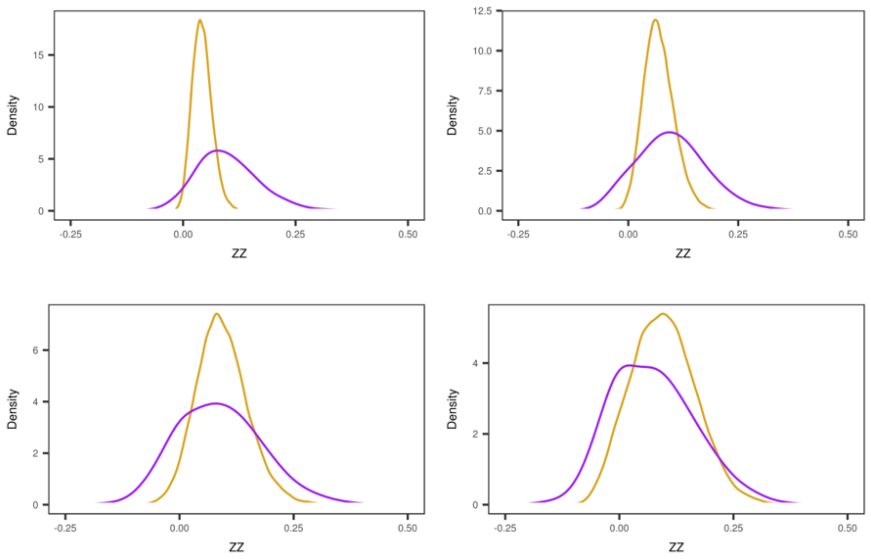
Supplementary Figure S2.35: Genome-wide distribution of Fu's F in simulated scenarios for both the neutral distribution (yellow) and selective scenarios (purple). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



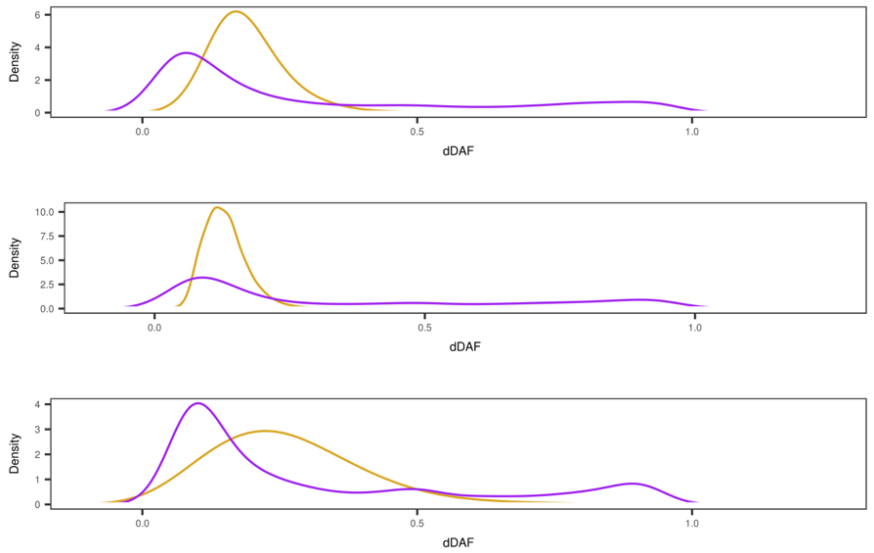
Supplementary Figure S2.36: Genome-wide distribution of Z_a in simulated scenarios for both the neutral distribution (yellow) and selective scenarios (purple). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



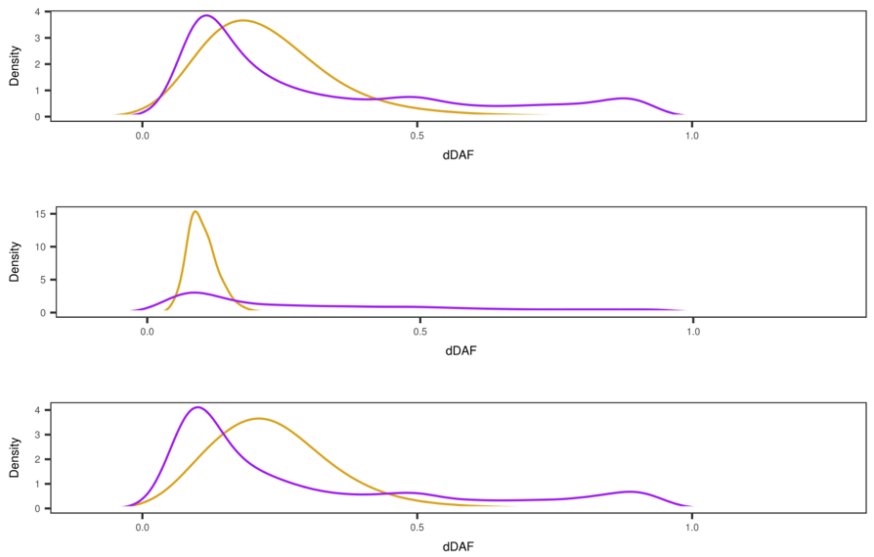
Supplementary Figure S2.37: Genome-wide distribution of Z_{nS} in simulated scenarios for both the neutral distribution (yellow) and selective scenarios (purple). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



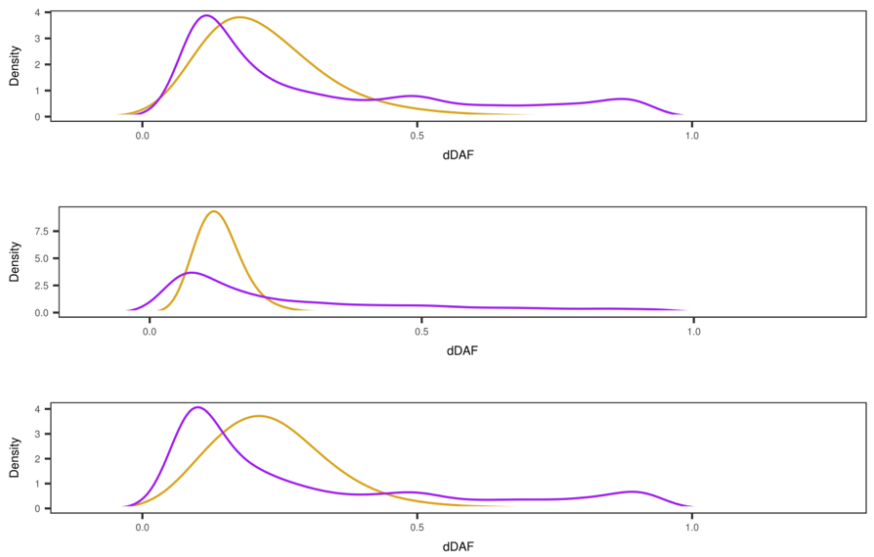
Supplementary Figure S2.38: Genome-wide distribution of ZZ in simulated scenarios for both the neutral distribution (yellow) and selective scenarios (purple). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



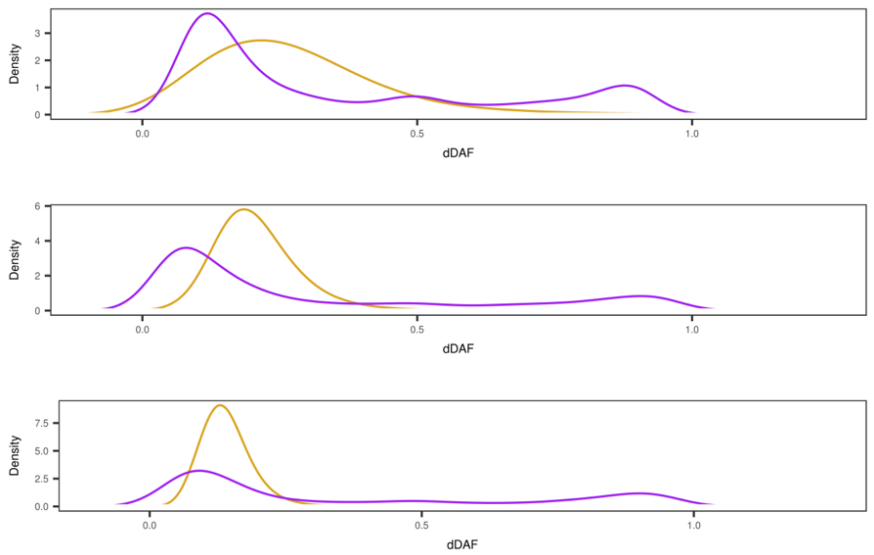
Supplementary Figure S2.39: Genome-wide distribution of ΔDAF for *P.t. ellioti* across all simulated scenarios, the neutral distribution is in yellow and selective scenarios are in purple. The subspecies are top *P.t. troglodytes*, middle *P.t. schweinfurthii*, and bottom *P.t. verus*.



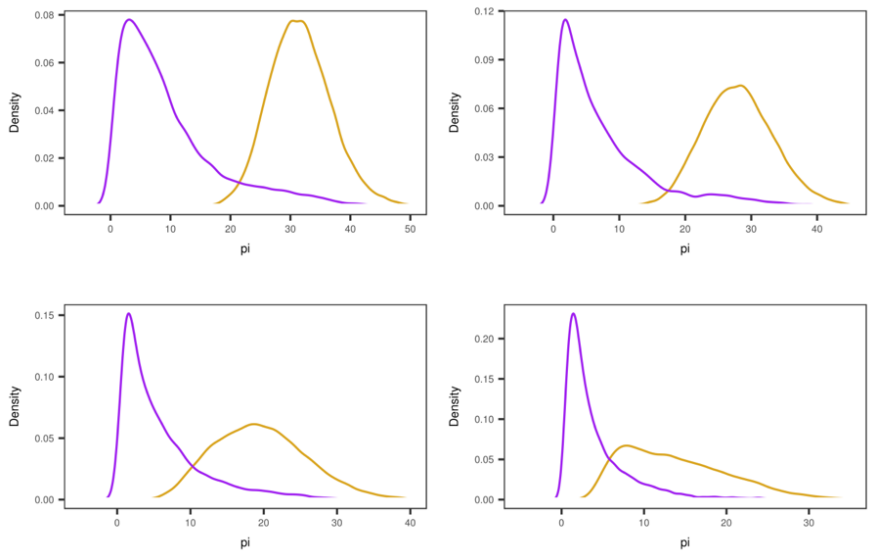
Supplementary Figure S2.40: Genome-wide distribution of ΔDAF for *P.t. schweinfurthii* across all simulated scenarios, the neutral distribution is in yellow and selective scenarios are in purple. The subspecies are top *P.t. troglodytes*, middle *P.t. ellioti*, and bottom *P.t. verus*.



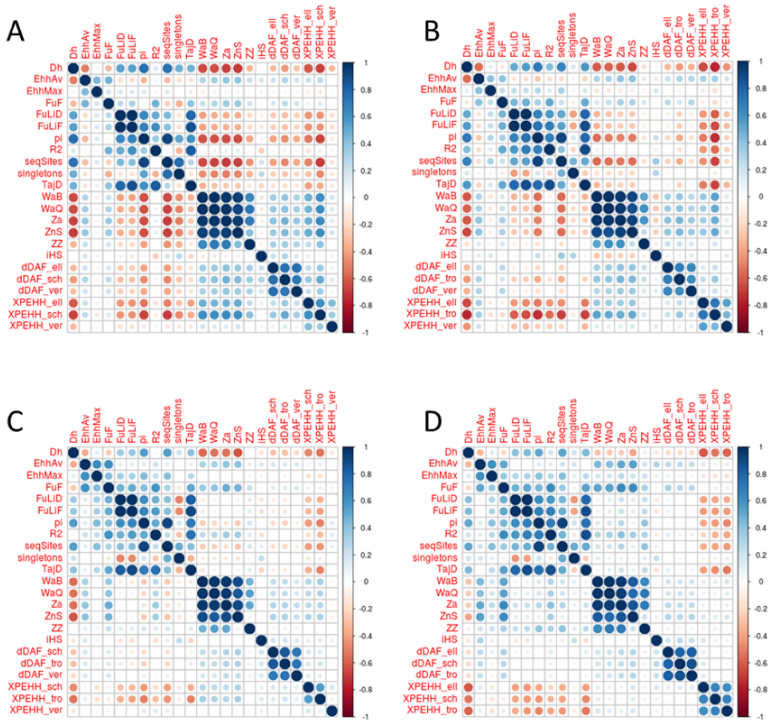
Supplementary Figure S2.41: Genome-wide distribution of ΔDAF for *P.t. troglodytes* across all simulated scenarios, the neutral distribution is in yellow and selective scenarios are in purple. The subspecies are top *P.t. schweinfurthii*, middle *P.t. ellioti*, and bottom *P.t. verus*.



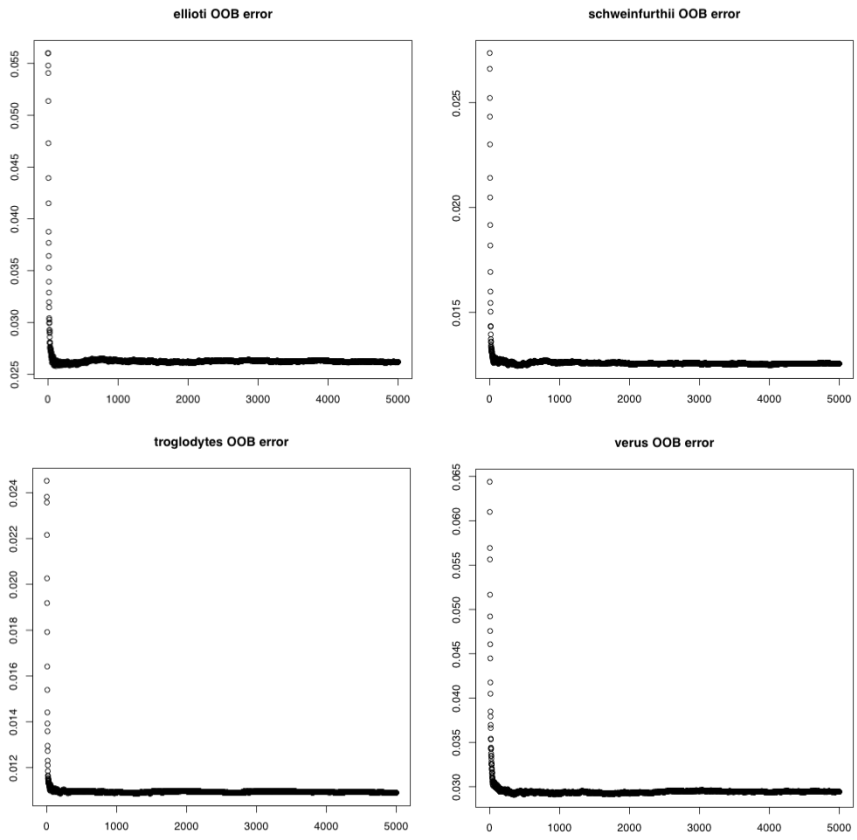
Supplementary Figure S2.42: Genome-wide distribution of ΔDAF for *P.t. verus* across all simulated scenarios, the neutral distribution is in yellow and selective scenarios are in purple. The subspecies are top *P.t. schweinfurthii*, middle *P.t. ellioti*, and bottom *P.t. troglodytes*.



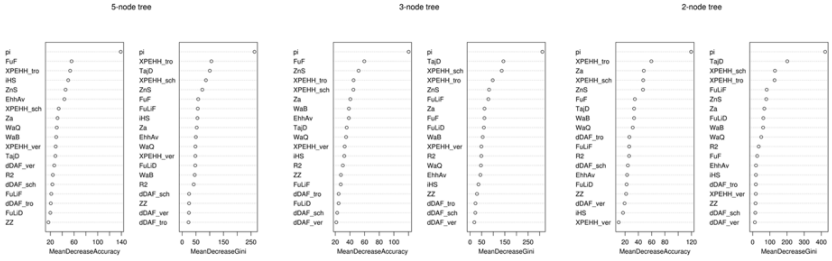
Supplementary Figure S2.43: Genome-wide distribution of π in simulated scenarios for both the neutral distribution (yellow) and selective scenarios (purple). The subspecies are top left *P.t. troglodytes*, top right *P.t. schweinfurthii*, bottom left *P.t. ellioti*, and bottom right *P.t. verus*.



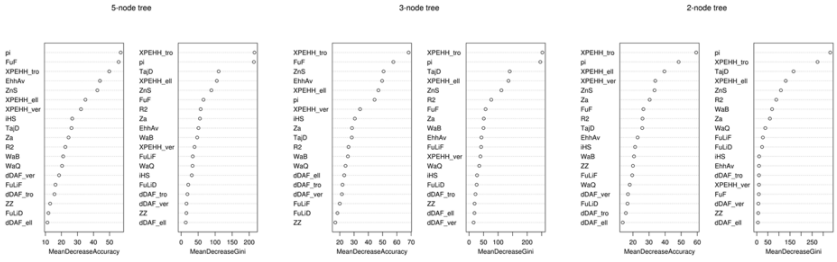
Supplementary Figure S2.44: The correlations between all measured statistics. **A)** *P.t. troglodytes*, **B)** *P.t. schweinfurthii*, **C)** *P.t. ellioti*, and **D)** *P.t. verus*.



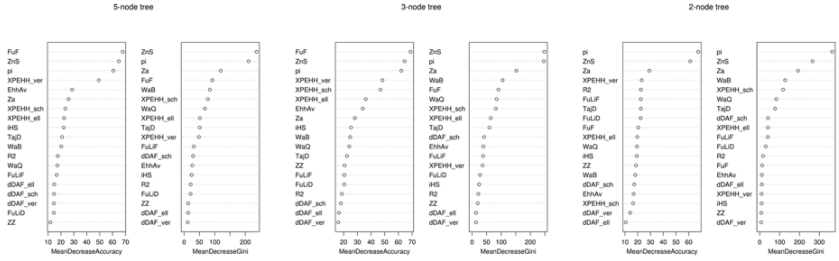
Supplementary Figure S2.45: Change in out of bag (OOB) error rate by the number of trees in the random forest algorithm.



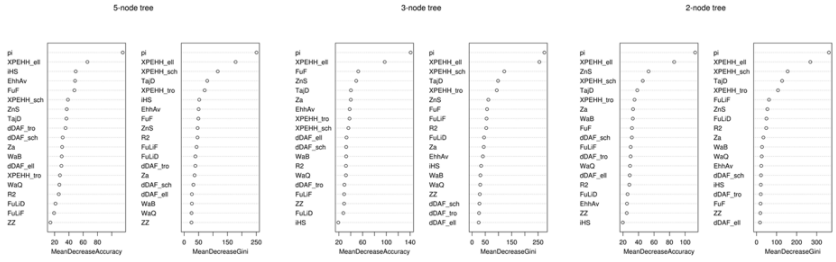
Supplementary Figure S2. 46: Mean decrease in accuracy and mean decrease in ginning for each tested tree topology for *P.t. ellioti*.



Supplementary Figure S2. 47: Mean decrease in accuracy and mean decrease in ginning for each tested tree topology for *P.t. schweinfurthii*.



Supplementary Figure S2. 48: Mean decrease in accuracy and mean decrease in ginning for each tested tree topology for *P.t. troglodytes*.



Supplementary Figure S2. 49: Mean decrease in accuracy and mean decrease in ginning for each tested tree topology for *P.t. verus*.

2.9 Supplementary Tables

Supplementary Table S2. 1: Differences between genome-wide site frequency spectrum counts for each of the four considered demographic models.

Model	*schweinfurthii-troglodytes-ellioti* full spectrum	*schweinfurthii-troglodytes-ellioti* pruned spectrum	*schweinfurthii-troglodytes-verus* full spectrum	*schweinfurthii-troglodytes-verus* pruned spectrum
Model 1	± 1100	± 79	NA	NA
Model 2	NA	NA	± 1000	± 240
Model 3	± 1200	± 160	± 1100	± 160
Model 4	± 1500	± 180	± 1500	± 180

Supplementary Table S2. 2: Test-specific mean decrease in accuracy based on the three tree topology models considered for *P.t. ellioti*.

Test	2-Node Tree	3-Node Tree	5-Node Tree
EHH average	144.57	178.87	146.45
FuF	149.89	273.22	171.46
FuLiD	79.88	78.70	64.19
FuLiF	90.25	88.82	65.52
pi	408.15	433.51	399.19
R_2	82.60	101.44	76.05
Tajima's D	113.45	111.58	80.04
Wall's B	116.87	128.08	84.29
Wall's Q	117.90	131.87	76.73
Z_a	139.14	144.34	90.68
Z_{nS}	191.08	185.27	135.17
ZZ	103.68	111.27	52.15
iHS	132.92	179.98	190.00
dDAF vs. sch	90.23	117.92	80.82
dDAF vs. tro	92.94	123.17	77.29
dDAF vs. ver	128.16	144.70	106.99
XPEHH vs. sch	241.17	245.73	106.69
XPEHH vs. tro	217.69	233.60	137.58
XPEHH vs. ver	93.16	188.45	133.51

Supplementary Table S2. 3: Test-specific mean decrease in accuracy based on the three tree topology models considered for *P.t. schweinfurthii*.

Test	2-Node Tree	3-Node Tree	5-Node Tree
EHH average	97.44	176.61	128.53
FuF	102.14	270.09	173.74
FuLiD	71.67	76.84	40.11
FuLiF	72.98	79.09	43.22
pi	142.34	159.91	168.83
R_2	80.76	86.70	57.85
Tajima's D	70.05	87.83	63.01
Wall's B	63.34	74.45	52.34
Wall's Q	70.55	81.48	50.19
Z_a	90.18	93.19	59.48
Z_{nS}	126.51	152.41	136.58
ZZ	81.32	76.02	38.38
iHS	83.26	121.73	112.62
dDAF vs. ell	59.91	71.33	50.07
dDAF vs. tro	53.05	71.09	60.57
dDAF vs. ver	64.44	80.39	60.65
XPEHH vs. ell	123.70	154.92	85.58
XPEHH vs. tro	361.14	398.08	115.09
XPEHH vs. ver	167.95	226.54	96.70

Supplementary Table S2. 4: Test-specific mean decrease in accuracy based on the three tree topology models considered for *P.t. troglodytes*.

Test	2-Node Tree	3-Node Tree	5-Node Tree
EHH average	92.50	128.14	82.72
FuF	106.22	299.90	207.46
FuLiD	73.79	84.05	45.00
FuLiF	77.43	84.16	48.64
pi	183.35	218.15	189.52
R_2	75.79	81.82	42.62
Tajima's D	63.12	73.32	47.91
Wall's B	62.70	76.53	48.03
Wall's Q	61.58	77.24	47.46
Z_a	82.87	88.91	64.92
Z_{nS}	139.46	137.78	133.19
ZZ	73.10	78.80	33.16
iHS	49.80	69.94	71.01
dDAF vs. ell	37.88	37.97	32.02
dDAF vs. sch	64.08	55.96	41.87
dDAF vs. ver	37.83	42.37	39.72
XPEHH vs. ell	136.86	180.03	63.45
XPEHH vs. sch	198.75	219.32	55.95
XPEHH vs. ver	152.73	255.51	148.13

Supplementary Table S2. 5: Test-specific mean decrease in accuracy based on the three tree topology models considered for *P.t. verus*.

Test	2-Node Tree	3-Node Tree	5-Node Tree
EHH average	141.28	192.80	168.80
FuF	141.08	221.59	153.68
FuLiD	85.21	82.82	55.32
FuLiF	91.35	86.69	52.81
pi	469.25	488.85	306.42
R_2	78.22	90.63	66.99
Tajima's D	122.42	121.12	95.06
Wall's B	135.25	136.85	87.24
Wall's Q	131.12	140.10	87.44
Z_a	137.19	151.94	98.41
Z_{nS}	209.13	198.73	128.36
ZZ	99.20	111.44	63.90
iHS	121.20	151.32	168.28
dDAF vs. ell	134.24	167.38	112.01
dDAF vs. sch	112.91	130.47	102.88
dDAF vs. tro	118.36	151.13	99.62
XPEHH vs. ell	251.70	264.58	217.99
XPEHH vs. sch	234.33	234.44	108.48
XPEHH vs. tro	129.52	143.38	98.50

Supplementary Table S2. 6: Out of bag (OOB) error rates for each tree topology considered. Confusion is the rate of miss classifying one type of selection for another.

Tree Topography	OOB	False Positive	False Negative	Confusion
Pan troglodytes ellioti				
5-node tree	4.59%	0.30%	15.67%	17.56%
3-node tree	5.00%	0.39%	15.46%	15.11%
2-node tree	2.62%	0.65%	12.00%	NA
Pan troglodytes schweinfurthii				
5-node tree	2.78%	0.09%	2.95%	21.11%
3-node tree	3.35%	0.08%	7.70%	13.80%
2-node tree	1.26%	0.14%	6.83%	NA
Pan troglodytes troglodytes				
5-node tree	2.19%	0.02%	1.17%	15.50%
3-node tree	2.56%	0.03%	6.33%	10.31%
2-node tree	1.09%	0.05%	6.05%	NA
Pan troglodytes verus				
5-node tree	4.72%	0.38%	16.17%	17.5%
3-node tree	5.10%	0.50%	16.24%	14.44%
2-node tree	2.94%	0.78%	13.25%	NA

Supplementary Table S2. 7: Genes that are confirmed as being under putative positive selection from Cagan et al. 2016 with this current study.

Genes				
SGCB	DLGAP4	SCEL	TSPAN18	UGDH
LDB2	GRIN3A	GPR160	TEX29	FAF1
ADK	MACROD2	FHOD3	FAM135B	PCBP3
STAG1	GABRB1	KCNIP3	ABLIM2	CHCHD6
CDKAL1	FER1L5	FER	FGD4	ZRANB1
STK33	FAM117B	GYPA	GLG1	CDH19
SGMS2	CRCT1	GRID2	MROH8	BRD2
MCC	PLCL2	KATNA1	ZNF670	TGS1
RAI14	OR11A1	EHBP1L1	MCCC1	ZNF213
KIAA0586	FYCO1	SLC26A6	CDC25C	

Supplementary Table S2. 8: Possible functional variants unique to the overlapping subspecies whose list they appear on.

Subspecies	SNP Location	Variant	Gene	*ellioti* Frequency	*schweinfurthii* Frequency	*troglodytes* Frequency	*verus* Frequency	AA Change
sch-tro	chr10_38270902_C/T	synonymous	*ANKRD30A*	0.00	0.03	0.03	0.00	-
sch-tro	chr9_72947958_T/A	missense	*FANCC*	0.00	0.13	0.03	0.00	H/L
sch-tro	chr12_112995207_C/T	missense	*GIT2*	0.00	0.47	0.06	0.00	V/I
sch-tro	chr14_51198821_A/G	missense	*GPHN*	0.00	0.11	0.08	0.00	N/S
sch-tro	chr4_147775813_G/A	synonymous	*GYPA*	0.00	0.05	0.17	0.00	-
sch-tro	chr4_147777997_C/T	missense	*GYPA*	0.00	0.05	0.11	0.00	G/E
ell-tro	chr6_131274890_C/T	synonymous	*KIAA0408*	0.20	0.00	0.08	0.00	-
sch-tro	chr5_20992906_T/G	missense	*KIAA0825*	0.00	0.03	0.08	0.00	I/M
sch-tro	chr5_21039833_C/T	missense	*KIAA0825*	0.00	0.16	0.11	0.00	P/L
sch-tro	chr5_21041808_T/C	synonymous	*KIAA0825*	0.00	0.50	0.06	0.00	-
sch-tro	chr5_21041823_G/A	synonymous	*KIAA0825*	0.00	0.08	0.06	0.00	-
sch-tro	chr5_21073281_C/T	missense	*KIAA0825*	0.00	0.26	0.17	0.00	R/C
sch-tro	chr1_127844744_G/A	synonymous	*MTMR11*	0.00	0.74	0.14	0.00	-
ell-tro	chr12_116112266_C/T	synonymous	*OAS2*	0.15	0.00	0.03	0.00	-
sch-tro	chr10_74168799_G/A	splice donor	*P4HA1*	0.00	0.34	0.03	0.00	-
sch-tro	chr16_5202180_C/T	missense	*PPL*	0.00	1.00	0.11	0.00	A/T
ell-tro	chr8_48734293_A/G	synonymous	*PRKDC*	0.10	0.00	0.03	0.00	-
sch-tro	chr16_59244767_G/A	synonymous	*PRMT7*	0.00	0.26	0.47	0.00	-
ell-tro	chr9_71828120_A/C	missense	*PTPDC1*	0.05	0.00	0.028	0.00	K/T
sch-tro	chr12_123336870_C/T	splice variant	*RAB35*	0.00	0.92	0.44	0.00	-

sch-tro	chr6_114969474_T/C	splice variant	*REV3L*	0.00	0.03	0.06	0.00	-
sch-tro	chr6_114969809_G/A	synonymous	*REV3L*	0.00	0.03	0.03	0.00	-
sch-tro	chr6_114975947_G/A	missense	*REV3L*	0.00	0.03	0.06	0.00	T/M
sch-tro	chr6_114977020_C/T	synonymous	*REV3L*	0.00	0.79	0.44	0.00	-
sch-tro	chr6_114977830_T/A	missense	*REV3L*	0.00	0.03	0.19	0.00	E/D
sch-tro	chr6_114978064_T/G	synonymous	*REV3L*	0.00	0.79	0.47	0.00	-
sch-tro	chr6_114978465_A/G	missense	*REV3L*	0.00	0.03	0.06	0.00	S/P
sch-tro	chr6_114980123_T/C	missense	*REV3L*	0.00	0.11	0.03	0.00	D/G
sch-tro	chr11_66153355_G/A	missense	*SF3B2*	0.00	0.11	0.11	0.00	V/I
sch-tro	chr11_66153973_C/T	synonymous	*SF3B2*	0.00	0.11	0.11	0.00	-
sch-tro	chr1_127844744_G/A	synonymous	*SF3B4*	0.00	0.74	0.86	0.00	-
ell-tro	chr8_48583567_C/T	missense	*SPIDR*	0.05	0.00	0.14	0.00	S/F
sch-tro	chr16_59844707_G/A	missense	*TANGO6*	0.00	0.32	0.19	0.00	E/K
ell-tro	chr12_125380614_G/A	missense	*WDR66*	0.15	0.00	0.03	0.00	D/N

Supplementary Table S2. 9: Possible functional variants unique to *P.t. ellioti.*

SNP Location	Variant	Gene	Frequency	Change
chr2B_110135924_G/A	synonymous	*ABCB6*	0.05	-
chr14_58814170_G/T	synonymous	*ABCD4*	0.50	-
chr20_25545090_C/G	splice variant	*ABHD12*	0.05	-
chr5_24783679_T/C	missense	*ADGRV1*	0.20	K/E
chr5_24783691_T/C	missense	*ADGRV1*	0.10	M/V
chr5_24783761_G/A	synonymous	*ADGRV1*	0.20	-
chr5_24820333_T/C	missense	*ADGRV1*	0.20	I/V
chr5_24830510_C/T	missense	*ADGRV1*	0.05	V/I
chr5_24852847_G/A	synonymous	*ADGRV1*	0.15	-
chr5_24856166_C/T	missense	*ADGRV1*	0.15	R/H
chr5_24889681_T/C	missense	*ADGRV1*	0.05	H/R
chr5_24889789_C/T	splice variant	*ADGRV1*	0.05	-
chr5_24898840_T/C	missense	*ADGRV1*	0.35	N/S
chr5_24900972_A/C	synonymous	*ADGRV1*	0.30	-
chr5_24906795_G/A	missense	*ADGRV1*	0.10	S/L
chr5_24917457_G/A	missense	*ADGRV1*	0.30	R/W
chr5_24931333_A/G	synonymous	*ADGRV1*	0.05	-
chr5_24931680_C/G	missense	*ADGRV1*	0.05	D/H
chr5_24931703_T/C	missense	*ADGRV1*	0.15	H/R
chr5_24935708_T/C	missense	*ADGRV1*	0.15	M/V
chr5_24959672_A/T	missense	*ADGRV1*	0.05	I/N
chr5_24975623_G/A	synonymous	*ADGRV1*	0.05	-
chr5_24984083_G/A	missense	*ADGRV1*	0.10	P/L
chr15_66939293_A/G	missense	*AGBL1*	0.20	M/V
chr15_66941068_A/G	missense	*AGBL1*	0.10	K/R
chr1_35232339_C/T	synonymous	*AGO1*	0.10	-
chr20_25151706_G/T	missense	*APMAP*	0.15	T/K
chr20_25166620_T/C	missense	*APMAP*	0.10	I/V
chr1_128766774_T/C	missense	*ARNT*	0.05	T/A
chr1_128795863_C/T	missense	*ARNT*	0.05	R/Q
chr2B_110146117_G/A	synonymous	*ATG9A*	0.05	-
chr1_92746781_G/T	missense	*C1H1orf146*	0.15	A/S
chr1_7016889_G/A	missense	*CAMTA1*	0.10	S/N
chr1_7024624_G/A	missense	*CAMTA1*	0.05	R/Q
chr1_7030754_G/A	splice variant	*CAMTA1*	0.15	-
chr12_41009753_T/C	missense	*CCDC65*	0.20	K/E
chr12_41010198_C/T	missense	*CCDC65*	0.05	E/K

chr22_15315377_G/A	synonymous	*CCDC117*	0.25	-
chr9_37025985_C/T	synonymous	*CCIN*	0.15	-
chr7_78010703_G/A	missense	*CCL26*	0.20	P/S
chr10_36551408_C/T	missense	*CCNY*	0.20	R/C
chr10_36551516_G/C	missense	*CCNY*	0.10	A/P
chr10_36554608_T/A	stop gain	*CCNY*	0.10	L/STOP
chr10_36638270_C/T	synonymous	*CCNY*	0.20	N
chr6_21233582_G/A	missense	*CDKAL1*	0.15	V/I
chr6_21465404_G/A	missense	*CDKAL1*	0.10	G/S
chr1_25359355_G/C	missense	*CEP85*	0.10	E/Q
chr1_25367592_G/A	missense	*CEP85*	0.70	V/M
chr1_132037450_G/A	synonymous	*CREB3L4*	0.10	-
chr10_36217405_A/G	missense	*CREM*	0.15	D/G
chr10_36247454_G/T	missense	*CREM*	0.15	Q/H
chr10_36250411_C/T	missense	*CREM*	0.05	R/C
chr1_32872874_C/T	missense	*CSMD2*	0.10	R/H
chr1_32894515_C/T	splice variant	*CSMD2*	0.05	-
chr1_33074741_G/A	synonymous	*CSMD2*	0.05	-
chr1_33182895_C/T	synonymous	*CSMD2*	0.15	-
chr1_33182940_G/A	synonymous	*CSMD2*	0.10	-
chr1_33367280_G/A	synonymous	*CSMD2*	0.05	-
chr1_40420570_A/G	missense	*CTPS1*	0.10	I/V
chr6_43540464_C/T	missense	*CUL7*	0.15	V/M
chr15_31877468_G/A	synonymous	*CYP19A1*	0.05	-
chr15_32135566_G/C	missense	*DMXL2*	0.05	A/G
chr15_32156404_T/C	synonymous	*DMXL2*	0.20	-
chr15_32186970_G/A	missense	*DMXL2*	0.05	T/M
chr15_32191822_T/G	synonymous	*DMXL2*	0.05	-
chr11_65407963_A/G	synonymous	*DPF2*	0.05	-
chr11_65408048_G/A	splice variant	*DPF2*	0.15	-
chr19_17336433_G/A	missense	*EPS15L1*	0.10	S/L
chr19_17374554_G/A	synonymous	*EPS15L1*	0.05	-
chr3_45975485_C/A	splice variant	*EXOSC7*	0.05	-
chr12_40308474_G/T	synonymous	*FAM186B*	0.15	-
chr12_40309718_G/A	missense	*FAM186B*	0.10	R/Q
chr6_99880634_T/C	synonymous	*FHL5*	0.10	-
chr6_99885233_G/C	splice variant	*FHL5*	0.80	-
chr6_99885402_A/G	synonymous	*FHL5*	0.30	-

chr12_40254198_C/T	synonymous	*FMNL3*	0.95	-
chr12_40259152_G/C	synonymous	*FMNL3*	0.10	-
chr1_92774111_T/G	missense	*GLMN*	0.05	E/A
chr1_133915175_T/C	synonymous	*GON4L*	0.10	-
chr1_133961794_A/G	missense	*GON4L*	0.05	L/P
chr4_95900296_C/T	synonymous	*GRID2*	0.05	-
chr4_96100742_T/C	synonymous	*GRID2*	0.15	-
chr10_121201398_C/T	synonymous	*GRK5*	0.05	-
chr5_146018495_T/A	missense	*GRXCR2*	0.15	D/V
chr12_36292104_A/G	missense	*MAP3K12*	0.10	Y/C
chr17_28443344_A/G	missense	*NEK8*	0.05	V/A
chr9_34202449_A/G	synonymous	*NOL6*	0.05	-
chr9_34206145_A/C	splice variant	*NOL6*	0.30	-
chr1_132069208_G/A	synonymous	*NUP210L*	0.15	-
chr1_132211590_A/G	missense	*NUP210L*	0.15	V/A
chr12_41435461_T/C	missense	*OR8S1*	0.05	S/G
chr7_103649959_G/T	missense	*PCOLCE*	0.15	R/L
chr11_117627506_G/A	synonymous	*PCSK7*	0.40	-
chr17_28249246_T/C	synonymous	*PHF12*	0.05	-
chr17_12065174_C/T	missense	*PLEKHM1*	0.05	R/W
chr12_109323893_C/T	missense	*POLR3B*	0.05	T/I
chr17_17995578_C/T	missense	*PPP1R1B*	0.05	D/N
chr1_132050698_G/A	synonymous	*RAB13*	0.10	-
chr1_152813539_C/T	splice variant	*RABGAP1L*	0.15	-
chr13_33497930_T/C	missense	*RCBTB2*	0.05	K/R
chr13_33510025_T/C	splice variant	*RCBTB2*	0.30	-
chr2B_88254673_C/T	synonymous	*RFTN2*	0.05	-
chr11_117655468_G/A	synonymous	*RNF214*	0.10	-
chr7_87829729_G/A	synonymous	*SEMA3D*	0.10	-
chr3_74265815_G/C	synonymous	*SHQ1*	0.15	-
chr3_164029277_A/G	synonymous	*SMC4*	0.05	-
chr9_36652038_C/G	missense	*SPAG8*	0.05	V/L
chr22_10759473_T/A	synonymous	*SPECC1L*	0.55	-
chr15_31376741_T/A	splice variant	*SPPL2A*	0.25	-
chr2B_110167475_G/A	missense	*STK16*	0.15	V/M
chr17_77717000_C/T	synonymous	*TMC6*	0.30	-
chr14_40818530_C/T	missense	*TMEM260*	0.45	H/Y
chr14_40845849_C/T	missense	*TMEM260*	0.05	A/V

chr15_32457248_G/A	missense	*TMOD2*	0.15	V/I
chr15_32546698_G/T	synonymous	*TMOD3*	0.05	-
chr15_32546708_C/T	missense	*TMOD3*	0.05	P/S
chr8_142998891_G/A	splice variant	*TRAPPC9*	0.10	-
chr15_31239802_A/G	synonymous	*TRPM7*	0.30	-
chr9_34991037_C/T	missense	*UBAP1*	0.30	A/V
chr9_34999681_C/T	synonymous	*UBAP1*	0.45	-
chr6_99813408_T/C	synonymous	*UFL1*	0.05	-
chr3_183621855_G/A	synonymous	*USP13*	0.30	-
chr8_81865717_A/G	missense	*ZBTB10*	0.10	M/V
chr8_81865746_C/G	synonymous	*ZBTB10*	0.05	-
chr7_102484497_C/T	synonymous	*ZKSCAN5*	0.05	-
chr7_102484539_C/T	synonymous	*ZKSCAN5*	0.20	-
chr7_102543555_T/G	missense	*ZNF655*	0.05	C/G
chr7_102444864_C/T	missense	*ZNF789*	0.25	R/W

Supplementary Table S2. 10: Possible functional variants unique to *P.t. schweinfurthii*.

SNP Location	Variant	Gene	Frequency	Change
chr17_7601461_G/C	missense	*ACAP1*	0.03	E/Q
chr4_101788679_T/C	synonymous	*ADH4*	0.03	-
chr4_101793842_G/C	missense	*ADH4*	0.08	P/A
chr4_101798583_A/G	synonymous	*ADH4*	0.03	-
chr4_101805291_G/A	synonymous	*ADH4*	0.03	-
chr15_45758266_G/A	splice variant	*ANKDD1A*	0.03	-
chr16_5031224_G/A	synonymous	*ANKS3*	0.08	-
chr14_7314913_C/T	missense	*AP1G2*	0.03	R/H
chr4_41298248_G/T	missense	*APBB2*	0.05	P/T
chr4_41312622_C/T	missense	*APBB2*	0.05	E/K
chr9_94499694_T/C	missense	*ASTN2*	0.03	T/A
chr16_15095057_G/A	synonymous	*BFAR*	0.03	-
chr16_15112966_T/A	synonymous	*BFAR*	0.03	-
chr17_67272713_C/T	missense	*BPTF*	0.03	R/W
chr17_67309633_A/G	synonymous	*BPTF*	0.05	-
chr17_67317473_A/T	missense	*BPTF*	0.03	D/V
chr17_67317512_A/G	missense	*BPTF*	0.13	K/R
chr17_67317973_G/A	missense	*BPTF*	0.03	V/I
chr17_67318131_A/C	missense	*BPTF*	0.03	K/N
chr17_67337783_G/A	missense	*BPTF*	0.03	R/Q
chr17_67351842_C/T	synonymous	*BPTF*	0.03	-
chr9_72446702_A/G	missense	*C9H9orf3*	0.08	R/G
chr15_22946587_A/G	synonymous	*CAPN3*	0.05	-
chr15_22948844_G/A	missense	*CAPN3*	0.03	G/D
chr5_73388911_T/C	synonymous	*CARD6*	0.05	-
chr5_73389043_C/A	missense	*CARD6*	0.21	K/N
chr15_25005530_G/A	synonymous	*CASC4*	0.03	-
chr3_125456493_C/T	synonymous	*CASR*	0.5	-
chr14_43801308_C/T	synonymous	*CCDC175*	0.11	-
chr20_36916956_G/A	missense	*CEP250*	0.03	R/H
chr20_36919742_G/A	synonymous	*CEP250*	0.05	-
chr10_74464045_C/T	missense	*CFAP70*	0.08	A/T
chr10_74465613_A/G	synonymous	*CFAP70*	0.03	-
chr10_74468356_C/T	missense	*CFAP70*	0.03	V/I
chr10_74468383_T/A	missense	*CFAP70*	0.03	M/L
chr10_74484368_T/C	missense	*CFAP70*	0.03	N/D
chr1_139551966_C/T	missense	*CFAP126*	0.03	D/N
chr16_60045797_C/T	missense	*CHTF8*	0.03	R/K

chr15_46059345_C/G	missense	*CILP*	0.05	A/P
chr15_46010405_A/G	splice variant	*CLPX*	0.05	-
chr1_130454300_C/T	missense	*CRNN*	0.08	V/M
chr4_110891129_C/T	synonymous	*CYP2U1*	0.16	-
chr14_43631526_A/C	missense	*DAAM1*	0.03	E/D
chr13_56711167_T/C	synonymous	*DACH1*	0.08	-
chr1_133804769_C/T	missense	*DAP3*	0.08	T/I
chr1_133813123_T/A	missense	*DAP3*	0.05	F/Y
chr14_37256309_C/T	synonymous	*DDHD1*	0.11	-
chr2B_52103884_T/C	splice variant	*DPP4*	0.03	-
chr2B_52122072_A/G	splice variant	*DPP5*	0.03	-
chr2B_52131049_A/G	splice variant	*DPP6*	0.18	-
chr17_7575110_C/T	synonymous	*EIF5A*	0.08	-
chr17_7575124_G/A	splice variant	*EIF5A*	0.03	-
chr4_142359600_C/G	splice variant	*ELF2*	0.05	-
chr8_140678104_A/G	synonymous	*FAM135B*	0.08	-
chr8_140694196_G/T	synonymous	*FAM135B*	0.08	-
chr8_140898082_C/A	synonymous	*FAM135B*	0.05	-
chr3_125575131_G/A	missense	*FAM162A*	0.03	S/N
chr14_18947459_A/G	synonymous	*SRP54*	0.03	-
chr14_19025578_T/C	missense	*SRP54*	0.05	F/L
chr12_58173462_T/C	synonymous	*FGD4*	0.03	-
chr12_58191526_T/C	synonymous	*FGD4*	0.03	-
chr6_78047424_C/A	missense	*FILIP1*	0.03	V/L
chr4_83499387_C/T	missense	*FRYL*	0.08	S/F
chr4_83572402_C/A	missense	*FRYL*	0.03	L/I
chr4_83577888_A/G	missense	*FRYL*	0.63	I/V
chr12_123356295_C/T	synonymous	*GCN1*	0.21	-
chr12_123356387_G/C	missense	*GCN1*	0.03	L/V
chr12_123398266_T/G	synonymous	*GCN1*	0.08	-
chr12_123403898_C/T	missense	*GCN1*	0.03	R/H
chr3_33743853_C/T	missense	*GLB1*	0.05	R/Q
chr3_33749659_T/C	missense	*GLB1*	0.03	T/A
chr3_33797069_T/C	synonymous	*GLB1*	0.03	-
chr3_33801004_C/T	missense	*GLB1*	0.08	R/Q
chr16_65406433_T/C	splice	*GLG1*	0.05	-

	variant			
chr16_65440641_C/T	missense	*GLG1*	0.03	R/H
chr17_7578154_G/A	splice variant	*GPS2*	0.05	-
chr12_115309542_C/T	synonymous	*HECTD4*	0.05	-
chr12_115309638_G/A	synonymous	*HECTD4*	0.03	-
chr12_115321532_G/A	synonymous	*HECTD4*	0.03	-
chr12_115356322_G/A	synonymous	*HECTD4*	0.05	-
chr12_115383081_T/C	synonymous	*HECTD4*	0.03	-
chr12_115386644_C/T	missense	*HECTD4*	0.08	R/Q
chr12_115423184_C/T	missense	*HECTD4*	0.03	V/M
chr6_26421971_G/A	missense	*HIST1H1C*	0.05	A/V
chr6_26473946_G/C	missense	*HIST1H1C*	0.03	L/V
chr19_8893546_A/G	missense	*HNRNPM*	0.08	I/V
chr3_128052589_A/G	missense	*ITGB5*	0.03	I/T
chr14_43751771_C/T	missense	*JKAMP*	0.13	P/S
chr14_7322592_G/A	missense	*JPH4*	0.08	T/I
chr18_44254021_T/C	synonymous	*KATNAL2*	0.08	-
chr1_175241580_T/C	synonymous	*KCNT2*	0.08	-
chr13_31664370_A/G	missense	*LRCH1*	0.11	N/S
chr13_31722589_C/G	missense	*LRCH1*	0.03	L/V
chr4_78746768_C/T	synonymous	*LRRC66*	0.05	-
chr4_78747283_G/T	missense	*LRRC66*	0.03	C/F
chr7_95006983_G/A	splice variant	*LRRD1*	0.03	-
chr15_24111294_T/G	missense	*MAP1A*	0.05	S/R
chr15_24116805_G/C	missense	*MAP1A*	0.03	Q/H
chr15_24117175_G/A	synonymous	*MAP1A*	0.05	-
chr2B_100486862_T/G	missense	*MAP2*	0.05	F/V
chr2B_100487576_G/A	missense	*MAP2*	0.05	V/I
chr3_186986734_C/T	missense	*MCCC1*	0.03	A/T
chr3_187006867_A/G	synonymous	*MCCC1*	0.03	-
chr6_38052488_G/A	missense	*MDGA1*	0.03	T/M
chr16_65643213_T/C	missense	*MLKL*	0.08	T/A
chr16_65658872_T/C	missense	*MLKL*	0.13	I/V
chr16_65663109_A/G	synonymous	*MLKL*	0.18	-
chr10_98814423_G/C	missense	*MMS19*	0.03	Q/E
chr4_56328123_C/G	missense	*MTHFD2L*	0.26	S/T
chr20_36187677_A/G	synonymous	*NCOA6*	0.03	-
chr20_36191898_T/C	missense	*NCOA6*	0.05	N/S
chr20_36221152_T/A	splice variant	*NCOA6*	0.03	-

chr4_142344154_G/T	synonymous	*NOCT*	0.03	-
chr4_142344550_C/T	synonymous	*NOCT*	0.08	-
chr3_13645938_G/A	synonymous	*NUP210*	0.03	-
chr3_13688799_T/C	synonymous	*NUP210*	0.03	-
chr3_13693780_C/T	synonymous	*NUP210*	0.03	-
chr6_28190339_C/T	stop gain	*OR2B6*	0.03	Q/STOP
chr6_29679103_C/T	missense	*OR10C1*	0.24	R/W
chr6_29679766_C/A	missense	*OR10C1*	0.03	P/T
chr6_29665582_A/G	missense	*OR11A1*	0.03	V/A
chr11_73872056_C/T	synonymous	*PAAF1*	0.03	-
chr11_73872157_A/T	missense	*PAAF1*	0.03	E/V
chr11_73899560_C/T	splice variant	*PAAF1*	0.32	-
chr9_37851226_G/A	missense	*PAX5*	0.03	P/L
chr2B_68040626_T/C	synonymous	*PDE11A*	0.03	-
chr2B_68306939_C/T	missense	*PDE11A*	0.08	A/T
chr15_24352196_G/A	missense	*PDIA3*	0.05	S/N
chr9_71403366_G/A	synonymous	*PHF2*	0.03	-
chr22_7251982_T/C	missense	*PI4KA*	0.03	N/S
chr22_7368611_A/G	synonymous	*PI4KA*	0.03	-
chr22_7368673_T/C	missense	*PI4KA*	0.03	I/V
chr22_7387541_T/C	missense	*PI4KA*	0.03	H/R
chr10_23357429_G/A	synonymous	*PIP4K2A*	0.11	-
chr10_23384536_G/A	missense	*PIP4K2A*	0.03	T/I
chr10_23388886_T/C	missense	*PIP4K2A*	0.03	Y/C
chr10_23407364_G/A	synonymous	*PIP4K2A*	0.03	-
chr17_18849506_C/T	missense	*PIP4K2B*	0.03	T/M
chr17_18849567_C/A	synonymous	*PIP4K2B*	0.03	-
chr17_18860382_C/T	synonymous	*PIP4K2B*	0.03	-
chr11_63609038_C/T	synonymous	*PLA2G16*	0.08	-
chr15_24121709_G/A	synonymous	*PPIP5K1*	0.11	-
chr14_19054775_A/G	synonymous	*PPP2R3C*	0.05	-
chr14_19248491_C/T	missense	*PSMA6*	0.18	L/F
chr14_19261982_T/G	splice variant	*PSMA6*	0.13	-
chr7_80404619_G/A	synonymous	*PTPN12*	0.03	-
chr7_80404647_A/G	missense	*PTPN12*	0.03	I/V
chr14_19515542_C/T	synonymous	*RALGAPA1*	0.03	-
chr14_19697475_C/T	missense	*RALGAPA1*	0.11	A/T
chr14_19700774_C/T	synonymous	*RALGAPA1*	0.08	-
chr14_19724361_G/A	synonymous	*RALGAPA1*	0.03	-
chr14_19736634_T/C	synonymous	*RALGAPA1*	0.03	-

chr16_24024717_T/C	splice variant	*RBBP6*	0.03	-
chr16_24030429_A/G	synonymous	*RBBP6*	0.29	-
chr16_24031099_A/T	missense	*RBBP6*	0.03	I/F
chr16_65612581_G/A	missense	*RFWD3*	0.05	A/V
chr16_29662034_A/C	missense	*RNF40*	0.03	Q/P
chr14_44016879_G/A	synonymous	*RTN1*	0.05	-
chr11_63743824_G/A	missense	*RTN3*	0.03	A/T
chr1_32152114_C/T	missense	*S100PBP*	0.03	R/W
chr1_32152384_C/T	missense	*S100PBP*	0.08	P/S
chr19_49322299_T/C	synonymous	*SAE1*	0.08	-
chr11_17789368_C/T	missense	*SERGEF*	0.03	G/S
chr4_110841551_C/A	missense	*SGMS2*	0.03	P/T
chr4_110841556_A/G	synonymous	*SGMS2*	0.03	-
chr4_110841736_T/C	synonymous	*SGMS2*	0.03	-
chr6_149987127_A/G	synonymous	*SHPRH*	0.03	-
chr11_117281179_A/G	synonymous	*SIK3*	0.08	-
chr15_56400223_T/C	splice variant	*SIN3A*	0.05	-
chr15_56404724_C/T	synonymous	*SIN3A*	0.05	-
chr18_44462697_A/G	missense	*SKOR2*	0.13	F/S
chr22_27422641_C/T	synonymous	*SLC25A17*	0.13	-
chr15_30791367_A/T	missense	*SLC27A2*	0.05	D/V
chr1_133225353_A/T	missense	*SLC50A1*	0.11	T/S
chr12_63344394_G/C	missense	*SMCO2*	0.03	L/V
chr16_60205048_A/G	synonymous	*SNTB2*	0.03	-
chr1_129592272_A/C	splice variant	*SNX27*	0.05	-
chr4_78660276_C/T	missense	*SPATA18*	0.03	R/H
chr15_45807249_C/T	synonymous	*SPG21*	0.05	-
chr2A_74359891_G/A	synonymous	*SPR*	0.03	-
chr13_34958393_G/T	synonymous	*SPRYD7*	0.05	-
chr11_8294524_G/A	synonymous	*STK33*	0.03	-
chr11_8298982_T/C	synonymous	*STK33*	0.03	-
chr3_124055942_G/C	missense	*STXBP5L*	0.32	V/L
chr3_124421489_G/A	synonymous	*STXBP5L*	0.03	-
chr3_124573923_G/A	missense	*STXBP5L*	0.03	M/I
chr7_195512_C/T	synonymous	*SUN1*	0.03	-
chr7_195545_A/G	synonymous	*SUN1*	0.05	-
chr7_226058_C/T	missense	*SUN1*	0.03	R/H
chr19_51567116_C/G	missense	*TEAD2*	0.03	E/D
chr16_60290412_C/G	synonymous	*TERF2*	0.16	-

chr16_60304977_G/A	synonymous	*TERF2*	0.03	-
chr1_44891412_T/C	missense	*TESK2*	0.03	I/V
chr1_44891878_C/T	synonymous	*TESK2*	0.05	-
chr1_133286072_A/G	synonymous	*THBS3*	0.03	-
chr16_57337858_C/A	synonymous	*TK2*	0.08	-
chr7_133495486_C/A	synonymous	*TNPO3*	0.08	-
chr7_133524767_C/A	missense	*TNPO3*	0.03	D/Y
chr7_133524776_C/T	missense	*TNPO3*	0.03	A/T
chr3_26168132_G/T	missense	*TOP2B*	0.03	P/H
chr3_26194145_A/G	synonymous	*TOP2B*	0.03	-
chr15_23995837_A/C	synonymous	*TP53BP1*	0.03	-
chr15_24019163_T/G	missense	*TP53BP1*	0.03	Q/H
chr15_24043772_T/C	missense	*TP53BP1*	0.03	H/R
chr15_24079005_G/A	synonymous	*TP53BP1*	0.03	-
chr9_94720378_C/T	synonymous	*TRIM32*	0.05	-
chr1_133268182_C/T	synonymous	*TRIM46*	0.05	-
chr15_23327138_G/A	missense	*TTBK2*	0.05	S/F
chr15_23328004_A/C	synonymous	*TTBK2*	0.03	-
chr15_23376640_T/C	missense	*TTBK2*	0.03	M/V
chr2B_68830908_T/C	missense	*TTN*	0.03	M/V
chr2B_68833355_C/A	missense	*TTN*	0.37	A/S
chr2B_68834461_T/A	missense	*TTN*	0.03	K/M
chr2B_68835143_G/A	missense	*TTN*	0.03	P/S
chr2B_68835429_A/G	synonymous	*TTN*	0.05	-
chr2B_68835610_T/A	missense	*TTN*	0.50	E/V
chr2B_68835615_T/C	synonymous	*TTN*	0.05	-
chr2B_68836692_G/A	synonymous	*TTN*	0.05	-
chr2B_68837208_A/T	synonymous	*TTN*	0.03	-
chr2B_68837219_G/T	missense	*TTN*	0.03	L/I
chr2B_68837567_A/C	missense	*TTN*	0.03	F/V
chr2B_68837647_G/T	missense	*TTN*	0.03	S/Y
chr2B_68840460_C/G	missense	*TTN*	0.05	V/L
chr2B_68843944_C/T	missense	*TTN*	0.03	V/I
chr2B_68846860_C/T	missense	*TTN*	0.05	V/M
chr2B_68849338_C/T	missense	*TTN*	0.03	E/K
chr2B_68850106_C/T	missense	*TTN*	0.08	V/I
chr2B_68854765_G/A	synonymous	*TTN*	0.05	-
chr2B_68862238_G/T	missense	*TTN*	0.03	T/N
chr2B_68862770_C/T	missense	*TTN*	0.03	V/I
chr2B_68863064_C/T	missense	*TTN*	0.03	E/K
chr2B_68863405_T/G	missense	*TTN*	0.03	N/T
chr2B_68864135_C/T	missense	*TTN*	0.03	G/S

chr2B_68864599_C/T	missense	*TTN*	0.03	R/H
chr2B_68866402_G/A	missense	*TTN*	0.03	T/M
chr2B_68868102_T/G	missense	*TTN*	0.05	E/D
chr2B_68868301_A/G	missense	*TTN*	0.03	I/T
chr2B_68870296_A/G	missense	*TTN*	0.08	I/T
chr2B_68873138_C/T	missense	*TTN*	0.03	V/I
chr2B_68873886_C/T	synonymous	*TTN*	0.08	-
chr2B_68876757_T/G	synonymous	*TTN*	0.05	-
chr2B_68881598_G/A	missense	*TTN*	0.03	A/V
chr2B_68884865_A/G	splice variant	*TTN*	0.03	-
chr2B_68889241_G/A	missense	*TTN*	0.03	A/V
chr2B_68892657_A/G	missense	*TTN*	0.05	V/A
chr2B_68893118_A/T	missense	*TTN*	0.03	D/E
chr2B_68893146_G/A	missense	*TTN*	0.03	T/M
chr2B_68893753_C/T	missense	*TTN*	0.03	D/N
chr2B_68893900_C/T	missense	*TTN*	0.03	V/I
chr2B_68894482_G/A	missense	*TTN*	0.03	P/L
chr2B_68902461_G/T	missense	*TTN*	0.03	P/Q
chr2B_68904036_C/T	missense	*TTN*	0.03	G/D
chr2B_68910009_G/C	missense	*TTN*	0.05	D/E
chr2B_68915297_G/A	synonymous	*TTN*	0.03	-
chr2B_68915414_C/G	missense	*TTN*	0.05	E/D
chr2B_68919835_G/C	missense	*TTN*	0.03	L/V
chr2B_68923781_A/G	missense	*TTN*	0.05	I/T
chr2B_68932745_C/T	missense	*TTN*	0.05	R/Q
chr2B_68935170_G/C	synonymous	*TTN*	0.03	-
chr2B_68936521_T/C	synonymous	*TTN*	0.05	-
chr2B_68937658_C/T	splice variant	*TTN*	0.05	-
chr2B_68939035_G/C	synonymous	*TTN*	0.05	-
chr2B_68944787_T/C	missense	*TTN*	0.03	K/R
chr2B_68953220_G/A	missense	*TTN*	0.08	A/V
chr2B_68979843_C/T	missense	*TTN*	0.13	V/I
chr2B_68988099_C/G	missense	*TTN*	0.03	K/N
chr2B_68991052_C/G	missense	*TTN*	0.16	V/L
chr2B_68993124_T/G	synonymous	*TTN*	0.03	-
chr2B_68996577_C/A	missense	*TTN*	0.11	V/L
chr2B_69004610_T/G	splice variant	*TTN*	0.03	-
chr2B_69011782_A/G	missense	*TTN*	0.05	Y/H
chr2B_69019093_A/G	splice	*TTN*	0.03	-

	variant			
chr2B_69022230_G/A	synonymous	*TTN*	0.05	-
chr2B_69022397_G/A	missense	*TTN*	0.08	H/Y
chr2B_69022758_A/G	synonymous	*TTN*	0.21	-
chr2B_69022764_C/A	synonymous	*TTN*	0.05	-
chr2B_69025242_A/G	synonymous	*TTN*	0.18	-
chr2B_69025931_T/C	synonymous	*TTN*	0.05	-
chr2B_69026814_G/A	synonymous	*TTN*	0.03	-
chr2B_69030221_T/C	synonymous	*TTN*	0.03	-
chr2B_69036533_G/T	synonymous	*TTN*	0.03	-
chr2B_69047460_G/A	missense	*TTN*	0.03	A/V
chr2B_69048478_T/G	missense	*TTN*	0.24	T/P
chr2B_69053183_C/T	missense	*TTN*	0.03	R/Q
chr2B_69053349_C/T	missense	*TTN*	0.03	E/K
chr2B_69054926_G/C	missense	*TTN*	0.03	S/C
chr2B_69056605_C/T	synonymous	*TTN*	0.03	-
chr2B_69057097_C/T	missense	*TTN*	0.03	M/I
chr2B_69057650_G/A	missense	*TTN*	0.05	T/M
chr2B_69057765_G/A	missense	*TTN*	0.03	R/C
chr2B_69059171_C/T	missense	*TTN*	0.05	R/H
chr2B_69063562_G/A	synonymous	*TTN*	0.08	-
chr2B_69063654_G/A	missense	*TTN*	0.03	H/Y
chr2B_69075505_G/A	splice variant	*TTN*	0.18	-
chr2B_69076325_T/C	synonymous	*TTN*	0.05	-
chr2B_69077403_C/T	missense	*TTN*	0.05	E/K
chr2B_69078797_G/A	missense	*TTN*	0.05	L/F
chr2B_69081213_C/T	missense	*TTN*	0.03	V/M
chr2B_69084542_A/G	synonymous	*TTN*	0.03	-
chr2B_69088680_G/A	synonymous	*TTN*	0.13	-
chr2B_69089825_C/T	synonymous	*TTN*	0.03	-
chr2B_69090442_G/A	synonymous	*TTN*	0.03	-
chr2B_69093561_C/T	synonymous	*TTN*	0.03	-
chr2B_69098249_T/C	synonymous	*TTN*	0.08	-
chr2B_69108090_C/T	missense	*TTN*	0.05	V/M
chr2B_69109597_C/T	missense	*TTN*	0.03	V/M
chr3_127963790_C/T	missense	*UMPS*	0.08	A/V
chr3_127963797_G/A	synonymous	*UMPS*	0.03	-
chr6_148454127_C/T	missense	*UTRN*	0.05	T/I
chr6_148461991_G/A	synonymous	*UTRN*	0.05	-
chr6_148463697_G/A	synonymous	*UTRN*	0.03	-
chr6_148477507_C/T	missense	*UTRN*	0.11	S/L

chr6_148486579_A/G	synonymous	*UTRN*	0.11	-
chr6_148486792_T/C	synonymous	*UTRN*	0.05	-
chr6_148486985_G/A	missense	*UTRN*	0.08	G/S
chr6_148502601_A/G	missense	*UTRN*	0.11	K/R
chr6_148516614_G/C	missense	*UTRN*	0.03	G/R
chr6_148518937_G/A	synonymous	*UTRN*	0.05	-
chr6_148543967_C/T	missense	*UTRN*	0.05	R/C
chr6_148584236_A/G	synonymous	*UTRN*	0.05	-
chr6_148787193_C/G	synonymous	*UTRN*	0.16	-
chr6_148865886_C/T	missense	*UTRN*	0.05	S/L
chr6_148873174_G/T	missense	*UTRN*	0.03	W/C
chr2A_97086620_C/T	synonymous	*VWA3B*	0.03	-
chr2A_97118999_G/A	missense	*VWA3B*	0.03	A/T
chr3_125587385_G/C	synonymous	*WDR5B*	0.03	-
chr19_39849537_A/G	missense	*WDR87*	0.03	L/S
chr19_39850477_A/G	missense	*WDR87*	0.03	S/P
chr19_39851468_C/T	synonymous	*WDR87*	0.03	-
chr19_39853509_G/A	missense	*WDR87*	0.03	S/L
chr19_39854995_G/C	missense	*WDR87*	0.08	H/D
chr9_92380957_C/T	synonymous	*WHRN*	0.05	-
chr9_92380990_G/C	missense	*WHRN*	0.03	D/E
chr9_92403321_C/T	synonymous	*WHRN*	0.05	-
chr1_32104426_C/T	synonymous	*YARS*	0.05	-
chr1_32109523_C/T	missense	*YARS*	0.03	R/Q
chr22_18544536_G/A	missense	*YWHAH*	0.03	C/Y
chr16_3387268_G/A	synonymous	*ZNF205*	0.03	-
chr15_23948393_C/T	missense	*ZSCAN29*	0.05	G/D
chr15_23955840_G/A	missense	*ZSCAN29*	0.05	S/L

Supplementary Table S2. 11: Possible functional variants unique to *P.t. troglodytes.*

SNP Location	Variant	Gene	Frequency	Change
chr12_36497920_A/C	missense	*AAAS*	0.25	N/T
chr12_36498612_C/T	missense	*AAAS*	0.06	P/L
chr6_43958468_G/A	missense	*ABCC10*	0.03	C/Y
chr6_31855682_G/T	missense	*ABHD16A*	0.03	Q/K
chr2B_94046600_C/T	splice variant	*ABI2*	0.08	-
chr22_28183124_G/A	synonymous	*ACO2*	0.11	-
chr12_37929412_C/T	missense	*ACVRL1*	0.06	M/I
chr4_58263517_C/T	synonymous	*ADAMTS3*	0.06	-
chr4_58269182_A/C	synonymous	*ADAMTS3*	0.03	-
chr4_58271898_C/T	synonymous	*ADAMTS3*	0.06	-
chr4_58299621_C/G	missense	*ADAMTS3*	0.03	P/A
chr4_58299939_G/A	missense	*ADAMTS3*	0.06	A/T
chr1_128491428_C/T	synonymous	*ADAMTSL4*	0.03	-
chr3_103400304_C/A	missense	*ADGRG7*	0.03	Q/K
chr3_103400483_G/C	missense	*ADGRG7*	0.03	R/S
chr10_75421895_G/A	missense	*ADK*	0.03	E/K
chr10_75421907_C/T	splice variant	*ADK*	0.06	-
chr10_75876818_T/C	splice variant	*ADK*	0.06	-
chr11_47937187_G/C	missense	*AGBL2*	0.06	P/R
chr11_47943882_C/T	synonymous	*AGBL2*	0.06	-
chr11_47981527_A/G	synonymous	*AGBL2*	0.03	-
chr11_47987248_T/C	missense	*AGBL2*	0.03	D/G
chr7_103595323_G/A	synonymous	*AGFG2*	0.08	-
chr7_103596106_C/T	missense	*AGFG2*	0.03	A/V
chr3_53318543_C/T	missense	*ALAS1*	0.03	R/C
chr3_53320494_T/C	synonymous	*ALAS1*	0.03	-
chr17_36717293_C/T	synonymous	*ALDH3A1*	0.06	-
chr17_36718014_C/T	missense	*ALDH3A1*	0.03	R/W
chr17_36718049_C/T	synonymous	*ALDH3A1*	0.03	-
chr17_36801699_C/T	missense	*ALDH3A2*	0.03	R/Q
chr12_33592721_C/T	synonymous	*ANKRD52*	0.08	-
chr15_63606017_C/G	synonymous	*AP3B2*	0.11	-
chr12_33484751_A/G	missense	*APOF*	0.03	N/D
chr1_135067513_C/T	missense	*ARHGEF11*	0.03	G/R
chr1_135068497_T/C	missense	*ARHGEF11*	0.06	H/R
chr1_135105434_G/A	synonymous	*ARHGEF11*	0.06	-
chr12_104238814_T/C	synonymous	*ARL1*	0.06	-
chr2B_122351949_A/G	missense	*ARMC9*	0.06	N/S
chr2B_122371623_T/C	synonymous	*ARMC9*	0.08	-

chr2B_122380965_A/G	missense	*ARMC9*	0.22	I/V
chr12_15021188_T/C	synonymous	*ATF7IP*	0.03	-
chr12_15022053_A/G	missense	*ATF7IP*	0.06	T/A
chr12_15022319_T/C	synonymous	*ATF7IP*	0.03	-
chr12_15022526_A/G	synonymous	*ATF7IP*	0.06	-
chr12_15022567_C/T	missense	*ATF7IP*	0.06	S/L
chr12_15096638_A/G	synonymous	*ATF7IP*	0.11	-
chr11_46874662_C/T	splice variant	*ATG13*	0.03	-
chr12_113377443_T/C	synonymous	*ATP2A2*	0.03	-
chr16_58331152_G/A	missense	*ATP6V0D1*	0.03	T/M
chr3_49500585_T/G	missense	*ATRIP*	0.17	D/E
chr3_49500670_G/C	missense	*ATRIP*	0.06	E/Q
chr3_49511316_C/T	synonymous	*ATRIP*	0.14	-
chr9_70437307_G/A	synonymous	*BICD2*	0.03	-
chr12_38548605_G/T	missense	*BIN2*	0.03	A/S
chr2A_75617555_A/G	synonymous	*BOLA3*	0.06	-
chr7_145389514_C/T	synonymous	*BRAF*	0.08	-
chr7_145394205_C/T	synonymous	*BRAF*	0.06	-
chr7_145412669_C/T	synonymous	*BRAF*	0.06	-
chr6_33297046_C/T	synonymous	*BRD2*	0.06	-
chr6_33300716_A/G	synonymous	*BRD2*	0.06	-
chr11_62661343_G/A	synonymous	*BSCL2*	0.06	-
chr1_31682001_C/T	missense	*BSDC1*	0.03	G/R
chr6_26754368_G/A	missense	*BTN2A2*	0.03	A/T
chr6_26754518_T/C	missense	*BTN2A2*	0.03	C/R
chr6_26759917_G/A	synonymous	*BTN2A2*	0.03	-
chr6_26759931_T/C	synonymous	*BTN2A2*	0.03	-
chr6_26761685_C/T	synonymous	*BTN2A2*	0.06	-
chr6_26762008_T/C	missense	*BTN2A2*	0.06	I/T
chr6_26815208_T/A	synonymous	*BTN3A3*	0.03	-
chr6_26821367_C/T	missense	*BTN3A3*	0.03	P/L
chr15_20639586_G/A	missense	*BUB1B*	0.03	A/T
chr15_20639606_C/T	synonymous	*BUB1B*	0.03	-
chr15_20674656_C/A	missense	*BUB1B*	0.03	F/L
chr1_132283994_C/T	synonymous	*C1H1orf43*	0.03	-
chr1_132283995_G/T	missense	*C1H1orf43*	0.06	T/K
chr1_132284007_C/T	missense	*C1H1orf43*	0.03	R/Q
chr1_132285925_G/C	missense	*C1H1orf43*	0.11	Q/E
chr1_150814535_A/G	synonymous	*C1H1orf105*	0.06	-
chr2A_75284610_C/T	synonymous	*C2AH2orf78*	0.06	-
chr2A_75284702_C/T	missense	*C2AH2orf78*	0.03	A/V
chr2A_75284896_C/T	missense	*C2AH2orf78*	0.03	P/S

chr2A_75284930_G/A	missense	*C2AH2orf78*	0.03	R/Q
chr2A_75285199_C/G	missense	*C2AH2orf78*	0.03	P/A
chr2A_75285518_C/T	missense	*C2AH2orf78*	0.03	S/F
chr5_77079232_T/C	missense	*C5H5orf42*	0.03	C/R
chr5_77120428_A/G	missense	*C5H5orf42*	0.03	D/G
chr5_77120684_A/G	synonymous	*C5H5orf42*	0.03	-
chr5_77133626_C/T	synonymous	*C5H5orf42*	0.14	-
chr5_77133762_G/A	missense	*C5H5orf42*	0.03	V/I
chr5_77134398_C/T	missense	*C5H5orf42*	0.06	S/F
chr5_77134667_A/G	missense	*C5H5orf42*	0.03	R/G
chr5_77134813_T/C	synonymous	*C5H5orf42*	0.06	-
chr5_77134969_A/C	missense	*C5H5orf42*	0.06	R/S
chr5_77161162_C/T	missense	*C5H5orf42*	0.03	T/I
chr5_77164789_C/T	missense	*C5H5orf42*	0.22	T/I
chr5_77178134_G/T	missense	*C5H5orf42*	0.03	A/S
chr5_77178147_A/G	missense	*C5H5orf42*	0.06	N/S
chr7_103213547_G/A	missense	*C7H7orf43*	0.03	P/S
chr7_103213734_G/C	missense	*C7H7orf43*	0.06	A/G
chr7_103214032_G/A	synonymous	*C7H7orf43*	0.06	-
chr10_103391931_A/G	synonymous	*C10H10orf76*	0.03	-
chr10_103466390_C/T	missense	*C10H10orf76*	0.03	V/M
chr10_103477376_G/A	synonymous	*C10H10orf76*	0.03	-
chr11_47212864_A/G	missense	*C11H11orf49*	0.03	K/R
chr11_108697178_T/A	synonymous	*C11H11orf65*	0.03	-
chr11_108707394_C/A	synonymous	*C11H11orf65*	0.03	-
chr11_66895806_T/G	missense	*C11H11orf80*	0.03	F/C
chr12_36499939_T/C	missense	*C12H12orf10*	0.06	D/G
chr12_36502763_G/A	missense	*C12H12orf10*	0.06	P/L
chr14_44720034_A/G	synonymous	*C14H14orf39*	0.03	-
chr14_44744648_A/G	synonymous	*C14H14orf39*	0.06	-
chr14_44766945_G/A	synonymous	*C14H14orf39*	0.03	-
chr11_3034939_G/A	missense	*CARS*	0.03	R/W
chr5_134939872_C/G	missense	*CATSPER3*	0.06	D/E
chr3_43645065_G/A	missense	*CCDC13*	0.08	L/F
chr3_43645097_C/T	missense	*CCDC13*	0.03	R/Q
chr3_43658009_T/G	synonymous	*CCDC13*	0.08	-
chr15_55326436_G/A	synonymous	*CCDC33*	0.08	-
chr3_50328265_T/A	synonymous	*CCDC36*	0.03	-
chr7_55500205_C/T	synonymous	*CCT6A*	0.03	-
chr7_55503462_T/G	synonymous	*CCT6A*	0.03	-
chr6_14247047_C/T	synonymous	*CD83*	0.03	-
chr5_138335746_A/T	missense	*CDC25C*	0.08	D/E

chr16_59713807_C/T	synonymous	*CDH1*	0.08	-
chr16_59713810_G/A	synonymous	*CDH1*	0.03	-
chr16_59714265_T/C	synonymous	*CDH1*	0.03	-
chr16_59724305_G/A	synonymous	*CDH1*	0.06	-
chr16_59725547_A/G	synonymous	*CDH1*	0.06	-
chr18_64011754_G/A	synonymous	*CDH19*	0.06	-
chr18_64014202_T/C	synonymous	*CDH19*	0.03	-
chr18_64047288_C/T	missense	*CDH19*	0.03	V/I
chr17_18109580_G/A	synonymous	*CDK12*	0.03	-
chr17_18165816_G/A	synonymous	*CDK12*	0.03	-
chr19_43542999_A/G	splice variant	*CEACAM4*	0.03	-
chr19_43549941_G/A	missense	*CEACAM4*	0.03	P/S
chr12_38499118_G/A	missense	*CELA1*	0.03	A/T
chr5_48820420_G/C	missense	*CENPK*	0.03	E/Q
chr18_47528230_T/G	missense	*CFAP53*	0.06	K/Q
chr18_47528375_G/A	synonymous	*CFAP53*	0.06	-
chr18_47528411_C/T	synonymous	*CFAP53*	0.03	-
chr18_47551426_G/C	missense	*CFAP53*	0.03	L/V
chr18_47551441_T/C	missense	*CFAP53*	0.03	I/V
chr18_47552286_T/C	missense	*CFAP53*	0.03	Q/R
chr2B_109923315_C/T	missense	*CFAP65*	0.03	E/K
chr2B_109932409_T/C	synonymous	*CFAP65*	0.14	-
chr2B_109938948_C/G	missense	*CFAP65*	0.11	G/A
chr2B_109943327_T/G	splice variant	*CFAP65*	0.03	-
chr2B_109954799_G/A	synonymous	*CFAP65*	0.06	-
chr2B_109954995_A/G	missense	*CFAP65*	0.06	M/T
chr2A_54985268_T/A	stop loss	*CHAC2*	0.03	STOP/K
chr8_61800970_G/A	synonymous	*CHD7*	0.08	-
chr8_61893468_G/A	synonymous	*CHD7*	0.08	-
chr11_67884243_G/A	missense	*CHKA*	0.03	S/F
chr11_46974859_T/C	missense	*CKAP5*	0.03	T/A
chr11_46980754_T/C	synonymous	*CKAP5*	0.03	-
chr11_46981719_T/C	splice variant	*CKAP5*	0.08	-
chr11_47001284_G/C	synonymous	*CKAP5*	0.03	-
chr15_55624433_C/T	synonymous	*CLK3*	0.14	-
chr3_49635554_C/T	synonymous	*COL7A1*	0.06	-
chr15_29709972_T/G	splice variant	*COPS2*	0.14	-
chr15_29720299_T/C	synonymous	*COPS2*	0.08	-
chr4_84269668_G/A	missense	*CORIN*	0.03	V/I
chr4_84312894_C/T	synonymous	*CORIN*	0.03	-
chr4_84312895_G/A	missense	*CORIN*	0.03	A/T
chr4_84400705_G/A	synonymous	*CORIN*	0.11	-

chr4_84404213_C/A	stop gain	*CORIN*	0.03	C/STOP
chr4_84409963_C/T	synonymous	*CORIN*	0.08	-
chr4_84442380_T/C	splice variant	*CORIN*	0.03	-
chr15_45006745_T/C	missense	*CSNK1G1*	0.06	M/V
chr15_45043740_T/C	missense	*CSNK1G1*	0.14	I/V
chr10_126906103_C/T	synonymous	*CTBP2*	0.03	-
chr5_138840853_C/T	missense	*CTNNA1*	0.06	P/S
chr5_138843846_G/A	synonymous	*CTNNA1*	0.03	-
chr5_138950022_C/T	synonymous	*CTNNA1*	0.11	-
chr10_96219776_G/A	missense	*CYP2C9*	0.03	G/D
chr10_96219787_A/G	missense	*CYP2C9*	0.06	I/V
chr10_96043309_A/G	missense	*CYP2C19*	0.06	K/E
chr10_96124050_C/T	missense	*CYP2C19*	0.03	R/W
chr7_102638242_G/A	synonymous	*CYP3A5*	0.06	-
chr7_102647484_G/A	splice variant	*CYP3A5*	0.17	-
chr5_140207141_C/T	missense	*CYSTM1*	0.03	P/L
chr3_52500468_C/T	synonymous	*DCAF1*	0.06	-
chr3_52502469_G/A	synonymous	*DCAF1*	0.03	-
chr8_37995876_T/C	synonymous	*DDHD2*	0.03	-
chr8_38009090_A/G	missense	*DDHD2*	0.03	T/A
chr8_38014754_C/T	missense	*DDHD2*	0.03	P/L
chr8_38016447_T/C	synonymous	*DDHD2*	0.11	-
chr8_38022872_T/C	splice variant	*DDHD2*	0.03	-
chr16_58911660_G/A	synonymous	*DDX28*	0.11	-
chr16_58911815_C/T	missense	*DDX28*	0.11	A/T
chr11_625974_G/A	synonymous	*DEAF1*	0.11	-
chr9_101728553_G/A	synonymous	*DENND1A*	0.06	-
chr17_7999105_C/T	missense	*DNAH2*	0.06	R/W
chr17_7999149_C/T	synonymous	*DNAH2*	0.03	-
chr17_8020410_C/T	missense	*DNAH2*	0.03	R/W
chr17_8046967_A/G	missense	*DNAH2*	0.03	N/D
chr17_8047056_T/C	synonymous	*DNAH2*	0.03	-
chr17_8056805_T/G	synonymous	*DNAH2*	0.03	-
chr17_8066750_G/A	synonymous	*DNAH2*	0.08	-
chr17_8073944_G/A	synonymous	*DNAH2*	0.08	-
chr17_8102697_C/T	synonymous	*DNAH2*	0.17	-
chr17_8104183_C/A	synonymous	*DNAH2*	0.03	-
chr17_8110801_G/A	synonymous	*DNAH2*	0.03	-
chr17_8117209_T/A	synonymous	*DNAH2*	0.06	-
chr10_74437198_G/A	splice variant	*DNAJC9*	0.08	-
chr15_46324972_T/G	missense	*DPP8*	0.03	K/T
chr15_46325475_T/C	missense	*DPP8*	0.11	Q/R

chr15_46347944_T/C	missense	*DPP8*	0.03	T/A
chr22_18009067_C/G	splice variant	*DRG1*	0.06	-
chr17_39506846_G/C	splice variant	*DRG2*	0.11	-
chr16_58967874_G/A	missense	*DUS2*	0.03	V/I
chr16_58967908_T/C	missense	*DUS2*	0.03	I/T
chr2A_75251009_G/T	synonymous	*DUSP11*	0.11	-
chr2A_75254815_C/T	missense	*DUSP11*	0.17	R/Q
chr16_57548401_G/A	synonymous	*DYNC1LI2*	0.03	-
chr16_57550806_G/A	stop gain	*DYNC1LI2*	0.06	R/STOP
chr16_57552478_G/A	splice donor	*DYNC1LI2*	0.03	-
chr13_34668371_T/A	synonymous	*EBPL*	0.03	-
chr3_131471956_C/T	synonymous	*EEFSEC*	0.06	-
chr22_30306648_G/T	synonymous	*EFCAB6*	0.06	-
chr22_30434075_A/C	missense	*EFCAB6*	0.06	N/K
chr22_30434083_C/G	missense	*EFCAB6*	0.06	E/Q
chr22_30478800_T/C	missense	*EFCAB6*	0.03	E/G
chr22_30499211_C/T	synonymous	*EFCAB6*	0.11	-
chr22_30523092_C/T	missense	*EFCAB6*	0.06	R/H
chr22_30533288_C/T	synonymous	*EFCAB6*	0.06	-
chr22_30565246_T/C	splice variant	*EFCAB6*	0.06	-
chr15_63305922_T/C	missense	*EFL1*	0.06	M/V
chr15_63384885_G/T	missense	*EFL1*	0.06	P/T
chr15_63384911_G/C	missense	*EFL1*	0.03	A/G
chr11_65645341_C/T	missense	*EHBP1L1*	0.08	T/M
chr11_65645402_G/A	synonymous	*EHBP1L1*	0.03	-
chr11_65645409_G/A	missense	*EHBP1L1*	0.06	D/N
chr11_65645790_T/C	synonymous	*EHBP1L1*	0.08	-
chr2A_64177525_C/T	missense	*EHBP1*	0.03	P/L
chr2A_64192110_A/G	synonymous	*EHBP1*	0.03	-
chr2A_64263902_G/A	missense	*EHBP1*	0.08	R/K
chr2A_64269342_A/G	synonymous	*EHBP1*	0.03	-
chr22_18045543_A/G	synonymous	*EIF4ENIF1*	0.06	-
chr1_49779015_G/T	synonymous	*ELAVL4*	0.03	-
chr1_49779180_C/G	synonymous	*ELAVL4*	0.03	-
chr11_31719412_C/T	synonymous	*ELP4*	0.03	-
chr11_31771647_T/G	synonymous	*ELP4*	0.03	-
chr11_31843481_C/T	stop gain	*ELP4*	0.03	Q/STOP
chr22_27789809_T/A	splice variant	*EP300*	0.03	-
chr22_27816201_G/C	missense	*EP300*	0.03	M/I
chr22_27823412_T/C	synonymous	*EP300*	0.06	-
chr22_27834603_T/A	synonymous	*EP300*	0.08	-
chr3_100019320_A/G	synonymous	*EPHA6*	0.08	-

chr3_100102997_C/T	missense	*EPHA6*	0.06	P/L
chr3_100103120_G/A	splice variant	*EPHA6*	0.03	-
chr3_100210594_C/A	missense	*EPHA6*	0.06	A/D
chr3_100210637_G/A	synonymous	*EPHA6*	0.08	-
chr3_100210670_G/A	synonymous	*EPHA6*	0.06	-
chr3_100269326_C/T	missense	*EPHA6*	0.06	L/F
chr2A_55012174_T/C	synonymous	*ERLEC1*	0.08	-
chr2A_55012389_A/G	missense	*ERLEC1*	0.08	K/R
chr2A_55023728_G/A	splice variant	*ERLEC1*	0.25	-
chr12_36516520_A/G	synonymous	*ESPL1*	0.31	-
chr12_36519341_A/G	missense	*ESPL1*	0.25	V/A
chr12_36528493_C/T	missense	*ESPL1*	0.03	R/Q
chr2A_73843416_T/C	splice variant	*EXOC6B*	0.03	-
chr2A_73978840_T/C	synonymous	*EXOC6B*	0.14	-
chr5_138347525_C/T	missense	*FAM53C*	0.03	R/C
chr2B_93438395_T/C	synonymous	*FAM117B*	0.03	-
chr10_74383182_C/T	synonymous	*FAM149B1*	0.03	-
chr10_74383183_G/A	missense	*FAM149B1*	0.03	A/T
chr10_74401026_C/G	synonymous	*FAM149B1*	0.03	-
chr10_74421124_A/G	synonymous	*FAM149B1*	0.03	-
chr10_74426223_G/A	missense	*FAM149B1*	0.03	R/Q
chr10_74430316_T/C	splice variant	*FAM149B1*	0.08	-
chr10_74430370_C/G	missense	*FAM149B1*	0.33	R/G
chr1_54360062_C/T	missense	*FAM151A*	0.03	V/M
chr15_29917717_T/C	missense	*FAM227B*	0.03	T/A
chr15_30168747_G/T	missense	*FAM227B*	0.03	A/E
chr15_30206994_G/A	synonymous	*FAM227B*	0.06	-
chr2B_132678619_G/T	missense	*FARP2*	0.03	Q/H
chr2B_132678634_C/T	synonymous	*FARP2*	0.03	-
chr2B_132678635_G/A	missense	*FARP2*	0.03	V/I
chr1_15438706_T/C	synonymous	*FBXO42*	0.06	-
chr1_15440468_G/A	synonymous	*FBXO42*	0.17	-
chr2A_95790582_C/T	synonymous	*FER1L5*	0.06	-
chr2A_95792352_C/T	missense	*FER1L5*	0.03	P/L
chr2A_95799026_C/T	synonymous	*FER1L5*	0.33	-
chr2A_95802025_C/T	synonymous	*FER1L5*	0.03	-
chr2A_95831290_G/A	synonymous	*FER1L5*	0.08	-
chr11_48001268_T/C	missense	*FNBP4*	0.03	Q/R
chr11_48029774_T/C	synonymous	*FNBP4*	0.06	-
chr5_131614144_C/A	missense	*FNIP1*	0.03	K/N
chr5_131614456_C/G	missense	*FNIP1*	0.06	Q/H
chr5_131614698_A/T	missense	*FNIP1*	0.08	S/T

chr5_131614756_G/A	synonymous	*FNIP1*	0.06	-
chr7_118982780_A/G	splice variant	*FOXP2*	0.03	-
chr7_118995776_A/C	synonymous	*FOXP2*	0.06	-
chr7_119041024_C/A	synonymous	*FOXP2*	0.06	-
chr16_30095687_G/T	missense	*FUS*	0.03	G/C
chr16_30095785_C/T	splice variant	*FUS*	0.08	-
chr3_46976096_C/T	synonymous	*FYCO1*	0.06	-
chr1_129057790_A/G	missense	*GABPB2*	0.03	K/E
chr15_29777097_G/A	synonymous	*GALK2*	0.08	-
chr15_29866108_G/A	missense	*GALK2*	0.03	R/Q
chr15_29913843_C/T	stop gain	*GALK2*	0.06	R/STOP
chr11_62596642_A/C	missense	*GANAB*	0.03	L/V
chr11_62598100_A/G	synonymous	*GANAB*	0.08	-
chr11_62600973_A/G	synonymous	*GANAB*	0.08	-
chr22_24360079_C/T	synonymous	*GCAT*	0.06	-
chr2A_70732870_G/A	synonymous	*GFPT1*	0.06	-
chr2A_70732982_G/T	splice variant	*GFPT1*	0.03	-
chr2A_70744845_A/C	splice variant	*GFPT1*	0.11	-
chr2A_70765698_G/A	splice variant	*GFPT1*	0.03	-
chr17_20882709_T/G	missense	*GGNBP2*	0.06	K/N
chr16_5147818_T/C	synonymous	*GLYR1*	0.03	-
chr3_159484709_A/G	missense	*GMPS*	0.03	K/R
chr16_1545108_C/A	synonymous	*GNPTG*	0.08	-
chr16_1545174_A/G	synonymous	*GNPTG*	0.08	-
chr3_173793717_G/A	missense	*GPR160*	0.11	R/H
chr3_173793754_A/T	synonymous	*GPR160*	0.14	-
chr3_173794336_A/T	missense	*GPR160*	0.03	L/F
chr14_65854406_A/G	synonymous	*GTF2A1*	0.03	-
chr15_46411564_C/T	missense	*HACD3*	0.03	P/L
chr15_46415341_G/A	missense	*HACD3*	0.03	E/K
chr11_46840332_A/G	synonymous	*HARBI1*	0.03	-
chr1_132345057_C/T	missense	*HAX1*	0.03	P/S
chr1_132346851_T/C	synonymous	*HAX1*	0.03	-
chr14_58024014_G/A	synonymous	*HEATR4*	0.06	-
chr10_95936808_G/C	missense	*HELLS*	0.06	E/Q
chr10_95964525_A/G	missense	*HELLS*	0.03	I/M
chr10_95967211_A/G	synonymous	*HELLS*	0.03	-
chr1_121921439_T/C	synonymous	*HIPK1*	0.17	-
chr1_121930568_G/A	missense	*HIPK1*	0.03	P/L
chr1_121948932_C/T	synonymous	*HIPK1*	0.06	-
chr1_121949343_G/A	synonymous	*HIPK1*	0.03	-
chr1_121949583_A/G	synonymous	*HIPK1*	0.08	-

chr1_22987885_C/T	synonymous	*HMGCL*	0.03	-
chr16_4785638_G/A	missense	*HMOX2*	0.06	E/K
chr11_62703276_G/A	synonymous	*HNRNPUL2*	0.11	-
chr11_62703767_C/T	missense	*HNRNPUL2*	0.03	E/K
chr1_128666336_G/A	synonymous	*HORMAD1*	0.08	-
chr1_128672023_A/G	synonymous	*HORMAD1*	0.06	-
chr4_96818909_T/C	synonymous	*HPGDS*	0.17	-
chr12_113172935_C/T	synonymous	*IFT81*	0.17	-
chr12_113172974_G/A	synonymous	*IFT81*	0.03	-
chr12_113250405_T/C	splice variant	*IFT81*	0.03	-
chr8_42037434_G/C	splice donor	*IKBKB*	0.06	-
chr12_33507274_G/A	splice variant	*IL23A*	0.06	-
chr3_50091109_G/A	splice variant	*IMPDH2*	0.06	-
chr11_77975583_G/A	splice variant	*INTS4*	0.03	-
chr15_48287247_C/T	missense	*IQCH*	0.03	T/I
chr15_48300586_C/G	missense	*IQCH*	0.39	Q/E
chr15_48327799_A/C	missense	*IQCH*	0.03	L/F
chr15_48372551_T/C	splice variant	*IQCH*	0.06	-
chr15_48372593_G/A	synonymous	*IQCH*	0.03	-
chr15_48431590_C/T	missense	*IQCH*	0.06	T/M
chr2A_95434208_C/T	synonymous	*ITPRIPL1*	0.03	-
chr10_64044302_A/G	synonymous	*JMJD1C*	0.03	-
chr10_64052701_G/T	missense	*JMJD1C*	0.06	L/I
chr10_64060589_G/T	synonymous	*JMJD1C*	0.06	-
chr10_64066922_A/T	synonymous	*JMJD1C*	0.06	-
chr10_64084878_G/A	synonymous	*JMJD1C*	0.06	-
chr10_64090937_C/T	missense	*JMJD1C*	0.11	V/I
chr10_64091559_T/C	synonymous	*JMJD1C*	0.06	-
chr10_64256606_T/C	missense	*JMJD1C*	0.03	Q/R
chr11_108802565_C/T	missense	*KDELC2*	0.03	G/D
chr11_67329279_A/T	missense	*KDM2A*	0.03	D/V
chr11_67345386_A/C	synonymous	*KDM2A*	0.08	-
chr5_138385789_G/A	missense	*KDM3B*	0.03	R/H
chr5_138397257_C/T	synonymous	*KDM3B*	0.03	-
chr5_138424892_G/A	synonymous	*KDM3B*	0.03	-
chr5_138433490_G/C	synonymous	*KDM3B*	0.03	-
chr8_138006437_G/A	synonymous	*KHDRBS3*	0.14	-
chr8_138033493_G/A	synonymous	*KHDRBS3*	0.03	-
chr14_42678997_A/G	synonymous	*KIAA0586*	0.06	-
chr14_42681402_G/T	missense	*KIAA0586*	0.08	V/L
chr14_42693485_A/G	synonymous	*KIAA0586*	0.08	-
chr14_42699067_T/C	synonymous	*KIAA0586*	0.03	-

chr14_42718622_A/G	missense	*KIAA0586*	0.03	Q/R
chr14_42721410_C/T	missense	*KIAA0586*	0.08	P/S
chr14_42727865_G/A	synonymous	*KIAA0586*	0.08	-
chr14_42739472_T/C	missense	*KIAA0586*	0.06	L/S
chr14_42759473_G/A	missense	*KIAA0586*	0.06	V/I
chr14_42789869_G/A	synonymous	*KIAA0586*	0.03	-
chr14_42797703_G/A	missense	*KIAA0586*	0.03	G/R
chr12_32247730_G/A	splice variant	*KIF5A*	0.19	-
chr3_187619351_G/A	synonymous	*KLHL24*	0.06	-
chr3_50193691_C/T	missense	*LAMB2*	0.03	D/N
chr4_115713899_T/C	splice variant	*LARP7*	0.06	-
chr4_115716550_C/T	splice variant	*LARP7*	0.11	-
chr22_17855547_G/A	synonymous	*LIMK2*	0.08	-
chr22_11785889_G/A	synonymous	*LRP5L*	0.06	-
chr8_106441681_G/A	synonymous	*LRP12*	0.03	-
chr8_106441760_C/T	missense	*LRP12*	0.03	R/Q
chr8_106442257_G/A	synonymous	*LRP12*	0.06	-
chr8_106482842_T/C	missense	*LRP12*	0.06	I/V
chr16_58248582_C/G	splice variant	*LRRC36*	0.03	-
chr16_58262509_C/T	splice variant	*LRRC36*	0.06	-
chr1_135057698_C/T	synonymous	*LRRC71*	0.03	-
chr3_37774996_G/A	splice variant	*LRRFIP2*	0.03	-
chr3_37787292_G/A	missense	*LRRFIP2*	0.14	T/M
chr8_75283480_A/T	synonymous	*LY96*	0.17	-
chr11_47534542_C/T	missense	*MADD*	0.08	P/L
chr11_47538943_A/C	synonymous	*MADD*	0.03	-
chr11_47540164_C/T	synonymous	*MADD*	0.03	-
chr1_122206803_C/T	synonymous	*MAGI3*	0.03	-
chr1_122206827_T/C	synonymous	*MAGI3*	0.03	-
chr1_122324120_C/T	missense	*MAGI3*	0.03	G/R
chr5_112627840_A/G	synonymous	*MCC*	0.19	-
chr5_112645560_G/T	synonymous	*MCC*	0.11	-
chr1_128513257_G/C	splice variant	*MCL1*	0.03	-
chr3_130818161_C/T	synonymous	*MCM2*	0.03	-
chr10_74017158_G/A	missense	*MCU*	0.25	E/K
chr2A_64952754_C/T	missense	*MDH1*	0.39	P/L
chr2A_64955666_G/A	missense	*MDH1*	0.06	A/T
chr2A_64962965_G/A	missense	*MDH1*	0.03	D/N
chr7_78290259_G/A	synonymous	*MDH2*	0.03	-
chr7_78292722_C/T	synonymous	*MDH2*	0.06	-
chr17_18205276_A/C	missense	*MED1*	0.08	N/H
chr17_18220420_A/C	missense	*MED1*	0.03	E/A

chr17_18222497_C/T	synonymous	*MED1*	0.69	-
chr17_18222527_A/G	synonymous	*MED1*	0.03	-
chr19_44282078_C/T	synonymous	*MEGF8*	0.03	-
chr12_36552940_G/A	synonymous	*MFSD5*	0.14	-
chr3_37728046_T/C	splice variant	*MLH1*	0.03	-
chr2A_75650418_G/A	synonymous	*MOB1A*	0.08	-
chr18_6079015_C/T	synonymous	*MPPE1*	0.03	-
chr18_6079037_C/T	missense	*MPPE1*	0.03	R/C
chr18_6079064_T/C	synonymous	*MPPE1*	0.03	-
chr18_6082563_G/A	synonymous	*MPPE1*	0.06	-
chr11_73783668_C/T	start loss	*MRPL48*	0.08	T/M
chr11_73847580_G/T	missense	*MRPL48*	0.03	L/F
chr2A_94141214_G/A	synonymous	*MRPS5*	0.06	-
chr2A_94143915_G/A	synonymous	*MRPS5*	0.03	-
chr2A_94154277_A/G	synonymous	*MRPS5*	0.14	-
chr10_74441948_C/T	missense	*MRPS16*	0.03	R/Q
chr18_47134453_T/C	splice variant	*MYO5B*	0.03	-
chr18_47258151_C/T	synonymous	*MYO5B*	0.03	-
chr11_63978746_G/A	synonymous	*NAA40*	0.06	-
chr2B_93740103_C/T	missense	*NBEAL1*	0.06	T/M
chr2B_93761033_C/T	missense	*NBEAL1*	0.06	P/S
chr2B_93761113_T/C	synonymous	*NBEAL1*	0.08	-
chr2B_93789917_G/A	missense	*NBEAL1*	0.03	V/I
chr2B_93804444_T/C	synonymous	*NBEAL1*	0.03	-
chr2B_93804545_G/A	missense	*NBEAL1*	0.06	R/H
chr2B_93813823_G/A	missense	*NBEAL1*	0.06	R/K
chr2B_93820512_G/A	missense	*NBEAL1*	0.03	G/S
chr2B_93842905_C/A	missense	*NBEAL1*	0.03	Q/K
chr2B_93848407_T/C	synonymous	*NBEAL1*	0.03	-
chr2B_93848452_G/A	missense	*NBEAL1*	0.03	V/I
chr2B_93880660_C/G	missense	*NBEAL1*	0.06	Q/E
chr13_19935560_G/A	synonymous	*NBEA*	0.03	-
chr13_19935689_T/C	synonymous	*NBEA*	0.03	-
chr13_20105406_C/T	synonymous	*NBEA*	0.14	-
chr13_20207201_G/A	synonymous	*NBEA*	0.03	-
chr2A_95447796_G/A	missense	*NCAPH*	0.08	E/K
chr12_40087737_G/T	splice variant	*NCKAP5L*	0.06	-
chr20_47763753_G/A	synonymous	*NCOA5*	0.03	-
chr6_32137010_C/A	synonymous	*NELFE*	0.06	-
chr2A_70807102_A/G	synonymous	*NFU1*	0.06	-
chr2A_70812601_G/A	splice variant	*NFU1*	0.03	-
chr5_77239379_A/G	missense	*NIPBL*	0.03	M/T

chr5_77276569_G/A	splice variant	*NIPBL*	0.03	-
chr5_77281689_G/A	synonymous	*NIPBL*	0.03	-
chr5_77284599_C/T	synonymous	*NIPBL*	0.06	-
chr5_77315768_C/T	synonymous	*NIPBL*	0.03	-
chr5_77316621_A/G	missense	*NIPBL*	0.03	I/T
chr5_77316772_T/C	missense	*NIPBL*	0.03	S/G
chr5_77325612_T/C	synonymous	*NIPBL*	0.03	-
chr3_43543755_G/A	missense	*NKTR*	0.06	R/Q
chr3_43544922_A/G	missense	*NKTR*	0.06	N/S
chr3_43545412_A/G	synonymous	*NKTR*	0.03	-
chr3_43545466_T/A	missense	*NKTR*	0.03	D/E
chr3_43545507_G/C	missense	*NKTR*	0.06	S/T
chr3_43548450_C/T	synonymous	*NKTR*	0.17	-
chr3_43549847_C/T	synonymous	*NKTR*	0.17	-
chr11_119640801_G/A	synonymous	*NLRX1*	0.17	-
chr11_119640824_C/T	missense	*NLRX1*	0.03	A/V
chr17_12444047_G/C	missense	*NMT1*	0.08	P/A
chr10_103614368_G/A	missense	*NOLC1*	0.03	R/Q
chr11_47510047_G/A	missense	*NR1H3*	0.03	V/I
chr11_47511189_T/C	synonymous	*NR1H3*	0.03	-
chr10_104570716_A/G	synonymous	*NT5C2*	0.06	-
chr6_119865901_C/T	missense	*NT5DC1*	0.06	T/M
chr6_119866726_C/T	synonymous	*NT5DC1*	0.06	-
chr11_48072906_A/G	synonymous	*NUP160*	0.06	-
chr11_48087204_A/G	splice variant	*NUP160*	0.08	-
chr11_48126471_G/A	synonymous	*NUP160*	0.08	-
chr9_70134703_T/C	missense	*OMD*	0.06	I/V
chr9_70135899_G/A	synonymous	*OMD*	0.03	-
chr9_70136323_T/C	missense	*OMD*	0.03	E/G
chr9_70136424_A/T	synonymous	*OMD*	0.03	-
chr17_3168916_T/G	synonymous	*OR1D2*	0.06	-
chr17_3169408_G/C	synonymous	*OR1D2*	0.06	-
chr12_32093431_C/G	synonymous	*OS9*	0.03	-
chr12_32094259_C/T	missense	*OS9*	0.03	E/K
chr12_32094330_C/T	missense	*OS9*	0.14	G/E
chr11_17389625_A/G	missense	*OTOG*	0.06	H/R
chr11_17390640_A/G	synonymous	*OTOG*	0.03	-
chr11_17401612_C/T	splice variant	*OTOG*	0.06	-
chr11_17425527_T/C	missense	*OTOG*	0.03	I/T
chr11_17425588_C/T	synonymous	*OTOG*	0.03	-
chr11_17425621_C/T	synonymous	*OTOG*	0.03	-
chr11_17434671_G/T	missense	*OTOG*	0.06	A/S

chr11_17461459_C/G	missense	*OTOG*	0.14	T/S
chr2A_62055191_C/T	synonymous	*PAPOLG*	0.03	-
chr2A_62055287_A/G	synonymous	*PAPOLG*	0.08	-
chr2A_62063913_T/C	synonymous	*PAPOLG*	0.14	-
chr15_53262560_T/C	synonymous	*PARP6*	0.03	-
chr15_53274949_G/A	missense	*PARP6*	0.03	-
chr15_53275839_G/A	synonymous	*PARP6*	0.03	-
chr15_46117947_C/T	missense	*PARP16*	0.03	V/I
chr15_46121713_G/A	synonymous	*PARP16*	0.03	-
chr15_46124919_C/T	splice variant	*PARP16*	0.03	-
chr15_46125114_T/C	synonymous	*PARP16*	0.03	-
chr11_31850863_G/A	synonymous	*PAX6*	0.03	-
chr5_141296129_A/G	missense	*PCDHB15*	0.08	T/A
chr5_141296401_T/C	synonymous	*PCDHB15*	0.03	-
chr7_102329806_C/T	missense	*PDAP1*	0.03	V/I
chr5_325798_G/A	synonymous	*PDCD6*	0.08	-
chr10_104889322_G/A	missense	*PDCD11*	0.03	V/I
chr10_104889342_T/C	synonymous	*PDCD11*	0.03	-
chr10_104893113_A/G	synonymous	*PDCD11*	0.03	-
chr10_104906648_C/T	synonymous	*PDCD11*	0.03	-
chr10_104915146_C/T	splice variant	*PDCD11*	0.17	-
chr10_104917959_G/A	missense	*PDCD11*	0.03	E/K
chr10_104927647_A/G	splice variant	*PDCD11*	0.06	-
chr11_14676822_G/A	synonymous	*PDE3B*	0.03	-
chr11_14676942_A/G	synonymous	*PDE3B*	0.06	-
chr11_14718856_T/C	synonymous	*PDE3B*	0.03	-
chr16_2830444_C/T	synonymous	*PDPK1*	0.03	-
chr16_2835896_C/T	missense	*PDPK1*	0.06	H/Y
chr16_2836234_C/T	synonymous	*PDPK1*	0.08	-
chr11_119658406_G/A	missense	*PDZD3*	0.08	R/Q
chr1_135040167_G/A	synonymous	*PEAR1*	0.06	-
chr6_81822568_A/T	synonymous	*PHIP*	0.14	-
chr6_81822919_T/C	synonymous	*PHIP*	0.25	-
chr6_81851866_G/C	synonymous	*PHIP*	0.14	-
chr7_111140894_G/A	synonymous	*PIK3CG*	0.11	-
chr7_111140973_G/A	missense	*PIK3CG*	0.06	A/T
chr7_111141026_A/G	synonymous	*PIK3CG*	0.08	-
chr7_111144612_A/G	synonymous	*PIK3CG*	0.03	-
chr7_111155580_C/T	synonymous	*PIK3CG*	0.03	-
chr7_111176607_C/T	synonymous	*PIK3CG*	0.17	-
chr1_129204043_C/T	missense	*PIP5K1A*	0.06	R/W
chr1_22951187_A/C	synonymous	*PITHD1*	0.06	-

chr1_22958469_C/T	synonymous	*PITHD1*	0.06	-
chr12_15111891_C/T	synonymous	*PLBD1*	0.06	-
chr12_15137047_C/T	missense	*PLBD1*	0.03	G/S
chr12_15141264_A/G	missense	*PLBD1*	0.03	I/T
chr12_15142591_G/A	splice variant	*PLBD1*	0.03	-
chr12_15153616_A/G	synonymous	*PLBD1*	0.03	-
chr12_15153730_A/G	synonymous	*PLBD1*	0.06	-
chr12_116501079_A/G	missense	*PLBD2*	0.03	-
chr3_149983083_C/T	missense	*PLSCR1*	0.03	A/T
chr3_149988062_C/T	synonymous	*PLSCR1*	0.06	-
chr3_149988063_G/A	missense	*PLSCR1*	0.03	S/L
chr3_149994679_A/G	synonymous	*PLSCR1*	0.11	-
chr3_149919997_C/T	synonymous	*PLSCR2*	0.03	-
chr3_149920033_T/G	missense	*PLSCR2*	0.03	K/N
chr3_149921425_A/G	synonymous	*PLSCR2*	0.03	-
chr10_103036152_T/G	missense	*POLL*	0.19	E/A or K/Q
chr22_28197652_G/A	splice variant	*POLR3H*	0.11	-
chr2A_27956879_T/A	missense	*PPM1G*	0.03	S/C
chr2A_27961388_C/T	synonymous	*PPM1G*	0.03	-
chr3_53358787_G/A	splice acceptor	*PPM1M*	0.03	-
chr10_103595535_C/T	synonymous	*PPRC1*	0.03	-
chr10_103595935_A/G	missense	*PPRC1*	0.03	T/A
chr5_48789743_A/G	synonymous	*PPWD1*	0.06	-
chr5_48796361_A/C	splice variant	*PPWD1*	0.03	-
chr5_48798866_T/C	synonymous	*PPWD1*	0.03	-
chr20_5174364_G/C	missense	*PROKR2*	0.06	A/G
chr20_5176929_C/T	missense	*PROKR2*	0.08	A/T
chr19_37704912_G/C	missense	*PROSER3*	0.08	R/T
chr6_31792991_C/T	missense	*PRRC2A*	0.03	P/S
chr7_146450994_G/T	synonymous	*PRSS37*	0.03	-
chr7_146451089_C/T	missense	*PRSS37*	0.03	V/I
chr5_139852725_C/T	missense	*PRSD2*	0.11	A/V
chr19_45093109_A/C	missense	*PSG2*	0.06	L/R
chr19_45098038_T/A	missense	*PSG2*	0.03	Q/L
chr7_55456991_C/T	missense	*PSPH*	0.03	R/H
chr17_75877749_G/A	missense	*QRICH2*	0.03	P/S
chr17_75888343_T/C	missense	*QRICH2*	0.03	K/R
chr2A_74556284_C/A	missense	*RAB11FIP5*	0.11	E/D
chr12_113553619_T/C	synonymous	*RAD9B*	0.17	-
chr12_113579679_C/G	missense	*RAD9B*	0.36	P/R

chr5_79553442_G/C	missense	*RAI14*	0.03	A/G
chr5_79554269_C/A	missense	*RAI14*	0.14	E/D
chr5_79554572_T/G	missense	*RAI14*	0.06	E/D
chr5_79574103_G/A	synonymous	*RAI14*	0.03	-
chr5_79707403_G/A	missense	*RAI14*	0.08	A/V
chr5_79708670_G/A	synonymous	*RAI14*	0.03	-
chr3_51414225_G/A	missense	*RASSF1*	0.03	R/C
chr11_66763927_G/A	splice donor	*RBM4B*	0.06	-
chr3_51191313_G/A	synonymous	*RBM5*	0.17	-
chr3_51201686_C/G	synonymous	*RBM5*	0.17	-
chr7_155721212_C/A	stop gain	*RHEB*	0.11	G/STOP
chr7_155735274_A/G	synonymous	*RHEB*	0.03	-
chr9_37219781_A/G	synonymous	*RNF38*	0.03	-
chr9_37221874_T/C	synonymous	*RNF38*	0.08	-
chr1_104309633_C/T	missense	*RNPC3*	0.06	R/C
chr1_104315260_G/A	synonymous	*RNPC3*	0.11	-
chr1_24453273_T/G	missense	*RSRP1*	0.08	E/D
chr6_33523815_T/C	missense	*RXRB*	0.03	I/V
chr2B_89857604_G/A	synonymous	*SATB2*	0.03	-
chr2B_89934456_G/A	splice variant	*SATB2*	0.03	-
chr2B_89966960_G/A	synonymous	*SATB2*	0.19	-
chr15_55850053_A/T	missense	*SCAMP2*	0.03	L/M
chr13_62888702_G/A	splice variant	*SCEL*	0.25	-
chr13_62909256_C/T	missense	*SCEL*	0.03	R/C
chr13_62954637_G/A	missense	*SCEL*	0.03	D/N
chr13_62954648_G/C	missense	*SCEL*	0.14	E/D
chr13_62965412_A/G	missense	*SCEL*	0.06	D/G
chr17_63449528_G/A	synonymous	*SCN4A*	0.08	-
chr6_35563545_G/A	synonymous	*SCUBE3*	0.06	-
chr6_35575727_G/C	missense	*SCUBE3*	0.03	E/D
chr11_65598747_C/T	synonymous	*SCYL1*	0.03	-
chr11_61283102_T/G	splice variant	*SDHAF2*	0.14	-
chr11_61291134_G/C	missense	*SDHAF2*	0.03	Q/H
chr5_238797_C/T	synonymous	*SDHA*	0.08	-
chr3_173664565_A/G	synonymous	*SEC62*	0.03	-
chr1_129105666_C/T	missense	*SEMA6C*	0.11	D/N
chr22_18126764_G/A	missense	*SFI1*	0.06	R/Q
chr22_18154930_C/T	missense	*SFI1*	0.03	R/W
chr22_18196175_C/T	missense	*SFI1*	0.03	R/C
chr22_18199906_G/C	missense	*SFI1*	0.06	E/D
chr2B_91171491_C/A	missense	*SGO2*	0.06	A/D
chr2B_91172570_T/C	synonymous	*SGO2*	0.06	-

chr2B_91172756_G/C	missense	*SGO2*	0.03	G/R
chr2B_91173439_C/G	missense	*SGO2*	0.03	I/M
chr2B_91173679_A/G	synonymous	*SGO2*	0.14	-
chr2B_91175454_G/A	missense	*SGO2*	0.03	R/K
chr16_37003676_G/A	missense	*SHCBP1*	0.06	T/I
chr16_37024827_C/T	missense	*SHCBP1*	0.03	G/D
chr16_37025222_G/A	splice variant	*SHCBP1*	0.11	-
chr16_37025274_G/A	missense	*SHCBP1*	0.03	T/M
chr16_37025585_T/C	missense	*SHCBP1*	0.03	N/S
chr3_49518894_C/T	splice variant	*SHISA5*	0.06	-
chr5_77624619_C/T	missense	*SLC1A3*	0.03	R/Q
chr5_77627442_G/A	synonymous	*SLC1A3*	0.06	-
chr17_37404083_T/G	missense	*SLC5A10*	0.06	I/L
chr6_114774605_C/G	missense	*SLC16A10*	0.36	F/L
chr6_114816056_G/A	missense	*SLC16A10*	0.03	V/I
chr1_147732205_C/A	missense	*SLC19A2*	0.03	L/F
chr1_147736371_A/G	missense	*SLC19A2*	0.11	M/T
chr11_63152784_G/A	synonymous	*SLC22A24*	0.06	-
chr3_49916900_G/A	splice variant	*SLC25A20*	0.08	-
chr6_33527871_G/A	missense	*SLC39A7*	0.03	R/H
chr5_101599513_T/C	missense	*SLCO4C1*	0.06	K/E
chr5_101603411_G/A	stop gain	*SLCO4C1*	0.03	Q/STOP
chr5_101609986_C/T	synonymous	*SLCO4C1*	0.03	-
chr5_101619856_C/T	synonymous	*SLCO4C1*	0.03	-
chr5_101633455_G/C	missense	*SLCO4C1*	0.06	T/S
chr5_101654089_G/A	missense	*SLCO4C1*	0.06	A/V
chr7_101967080_T/G	synonymous	*SMURF1*	0.03	-
chr7_101972631_G/A	synonymous	*SMURF1*	0.03	-
chr7_101985366_G/A	synonymous	*SMURF1*	0.14	-
chr15_44995156_C/T	missense	*SNX22*	0.11	P/L
chr1_109817330_G/A	synonymous	*SORT1*	0.11	-
chr12_36406046_A/C	splice variant	*SP1*	0.03	-
chr12_36406069_T/G	missense	*SP1*	0.03	I/L
chr12_40382142_C/T	missense	*SPATS2*	0.06	G/E
chr12_40388999_T/G	synonymous	*SPATS2*	0.03	-
chr17_27477230_C/G	missense	*SSH2*	0.03	P/A
chr17_27477358_A/G	synonymous	*SSH2*	0.03	-
chr17_27481516_G/A	missense	*SSH2*	0.03	R/K
chr17_27481629_G/A	missense	*SSH2*	0.03	V/M
chr17_27482106_C/T	missense	*SSH2*	0.03	P/S
chr17_27482215_G/A	missense	*SSH2*	0.17	R/Q
chr17_27482495_G/A	synonymous	*SSH2*	0.25	-

chr7_103258913_G/A	synonymous	*STAG3*	0.03	-
chr2A_75314098_A/C	synonymous	*STAMBP*	0.08	-
chr12_33490932_T/C	missense	*STAT2*	0.11	I/T
chr12_33492071_A/G	synonymous	*STAT2*	0.08	-
chr12_33495485_A/G	synonymous	*STAT2*	0.03	-
chr12_33496522_C/T	synonymous	*STAT2*	0.08	-
chr12_33497170_G/C	synonymous	*STAT2*	0.36	-
chr12_33500153_T/C	synonymous	*STAT2*	0.06	-
chr20_50866067_C/T	synonymous	*STAU1*	0.03	-
chr20_50883970_G/A	synonymous	*STAU1*	0.36	-
chr20_50914948_A/G	synonymous	*STAU1*	0.03	-
chr11_64220388_A/G	missense	*STIP1*	0.06	D/G
chr11_64227560_A/G	synonymous	*STIP1*	0.08	-
chr7_55518560_C/T	synonymous	*SUMF2*	0.03	-
chr2B_92826710_G/A	missense	*SUMO1*	0.08	T/M
chr20_34371905_C/G	missense	*SUN5*	0.03	S/T
chr6_45529373_G/A	synonymous	*SUPT3H*	0.11	-
chr6_45773371_G/A	splice variant	*SUPT3H*	0.22	-
chr1_127821050_G/A	synonymous	*SV2A*	0.06	-
chr11_67146495_G/A	synonymous	*SYT12*	0.03	-
chr10_104871348_A/G	synonymous	*TAF5*	0.03	-
chr10_104871369_C/T	synonymous	*TAF5*	0.11	-
chr18_52590733_G/C	synonymous	*TCF4*	0.06	-
chr18_52631826_T/G	synonymous	*TCF4*	0.14	-
chr15_38012093_T/C	synonymous	*TCF12*	0.11	-
chr15_38076528_A/G	missense	*TCF12*	0.06	T/A
chr6_35458229_G/A	missense	*TCP11*	0.03	T/M
chr16_57573236_G/C	missense	*TERB1*	0.03	Q/E
chr16_57584625_A/G	missense	*TERB1*	0.03	Y/H
chr16_57596244_C/T	synonymous	*TERB1*	0.03	-
chr16_57607600_C/A	synonymous	*TERB1*	0.03	-
chr11_68611332_G/T	synonymous	*TESMIN*	0.06	-
chr13_97234465_C/T	synonymous	*TEX29*	0.08	-
chr1_129837309_C/T	synonymous	*THEM5*	0.08	-
chr14_42662679_T/A	splice variant	*TIMM9*	0.06	-
chr3_53337177_C/T	missense	*TLR9*	0.06	R/Q
chr1_18691976_G/C	missense	*TMCO4*	0.08	L/V
chr1_18737280_C/T	missense	*TMCO4*	0.03	R/Q
chr1_18757995_G/T	missense	*TMCO4*	0.03	T/K
chr1_18758066_G/A	synonymous	*TMCO4*	0.08	-
chr1_18775010_C/T	synonymous	*TMCO4*	0.08	-
chr1_31388620_C/T	synonymous	*TMEM39B*	0.22	-

chr12_115105233_T/C	synonymous	*TMEM116*	0.03	-
chr19_50515815_G/A	synonymous	*TMEM143*	0.08	-
chr17_40695804_G/A	missense	*TNFRSF13B*	0.22	R/Q
chr17_39314975_T/C	missense	*TOP3A*	0.03	I/T
chr17_39324699_C/T	synonymous	*TOP3A*	0.11	-
chr17_39324762_A/G	synonymous	*TOP3A*	0.03	-
chr6_115178682_C/T	synonymous	*TRAF3IP2*	0.06	-
chr6_115197155_A/C	missense	*TRAF3IP2*	0.03	F/V
chr6_115197163_A/G	missense	*TRAF3IP2*	0.06	F/S
chr5_48711472_T/C	synonymous	*TRAPPC13*	0.03	-
chr5_48712790_T/C	synonymous	*TRAPPC13*	0.03	-
chr5_48722723_T/C	splice variant	*TRAPPC13*	0.06	-
chr5_48764139_C/T	synonymous	*TRIM23*	0.06	-
chr22_24247992_T/C	missense	*TRIOBP*	0.17	F/S
chr22_24269350_G/A	missense	*TRIOBP*	0.06	G/R
chr22_24269452_C/T	missense	*TRIOBP*	0.03	R/W
chr22_24313417_C/G	missense	*TRIOBP*	0.03	L/V
chr7_101822800_T/C	missense	*TRRAP*	0.03	F/L
chr7_101842748_C/T	synonymous	*TRRAP*	0.06	-
chr7_101850507_C/T	synonymous	*TRRAP*	0.03	-
chr7_101852456_A/G	missense	*TRRAP*	0.03	N/S
chr7_101856647_C/A	splice variant	*TRRAP*	0.03	-
chr7_101856791_C/T	synonymous	*TRRAP*	0.03	-
chr7_101860870_T/C	synonymous	*TRRAP*	0.08	-
chr7_101887616_C/T	synonymous	*TRRAP*	0.03	-
chr7_101890994_G/A	synonymous	*TRRAP*	0.14	-
chr7_101892308_G/A	splice variant	*TRRAP*	0.03	-
chr7_101902307_C/T	synonymous	*TRRAP*	0.03	-
chr7_101904113_A/G	synonymous	*TRRAP*	0.06	-
chr7_101910399_G/A	splice variant	*TRRAP*	0.03	-
chr7_101915958_C/T	synonymous	*TRRAP*	0.11	-
chr11_62719624_A/G	missense	*TTC9C*	0.03	N/S
chr17_59207316_T/C	missense	*TUBD1*	0.25	I/V
chr1_132320645_G/A	splice variant	*UBAP2L*	0.06	-
chr6_29784276_A/C	missense	*UBD*	0.03	S/A
chr2A_65210551_C/T	synonymous	*UGP2*	0.03	-
chr2A_65211777_G/A	synonymous	*UGP2*	0.06	-
chr2A_65240161_G/A	synonymous	*UGP2*	0.06	-
chr2A_65247356_G/A	synonymous	*UGP2*	0.44	-
chr6_35165954_G/C	missense	*UHRF1BP1*	0.03	R/T
chr6_35189140_T/C	synonymous	*UHRF1BP1*	0.03	-
chr6_35190358_G/A	missense	*UHRF1BP1*	0.03	A/T

chr6_35190619_G/T	missense	*UHRF1BP1*	0.08	A/S
chr6_35202446_C/T	synonymous	*UHRF1BP1*	0.19	-
chr17_36656612_G/A	synonymous	*ULK2*	0.03	-
chr10_74716813_C/T	missense	*USP54*	0.06	A/T
chr10_74717184_A/G	missense	*USP54*	0.06	I/T
chr10_74736688_T/G	synonymous	*USP54*	0.03	-
chr10_74786935_C/G	splice variant	*USP54*	0.11	-
chr10_75281454_C/T	synonymous	*VCL*	0.06	-
chr10_75298901_C/T	synonymous	*VCL*	0.03	-
chr6_33594003_C/T	missense	*VPS52*	0.03	A/T
chr2A_65270388_A/T	splice variant	*VPS54*	0.06	-
chr2A_65277448_G/A	synonymous	*VPS54*	0.06	-
chr2A_65277548_A/C	missense	*VPS54*	0.06	M/R
chr2A_65290269_G/C	missense	*VPS54*	0.03	H/Q
chr2A_65300893_A/T	synonymous	*VPS54*	0.03	-
chr1_129144150_T/C	synonymous	*VPS72*	0.56	-
chr2A_24494537_G/A	synonymous	*WDCP*	0.03	-
chr2A_24501781_A/G	synonymous	*WDCP*	0.03	-
chr2A_24502282_G/A	synonymous	*WDCP*	0.06	-
chr2A_64845008_T/C	splice variant	*WDPCP*	0.03	-
chr5_76545871_G/A	synonymous	*WDR70*	0.11	-
chr8_38038655_C/T	synonymous	*WHSC1L1*	0.03	-
chr8_38042533_T/C	synonymous	*WHSC1L1*	0.06	-
chr8_38051478_C/G	missense	*WHSC1L1*	0.03	E/Q
chr8_38101077_C/A	missense	*WHSC1L1*	0.08	R/S
chr8_38110044_C/A	missense	*WHSC1L1*	0.03	R/L
chr11_62732974_A/G	synonymous	*ZBTB3*	0.56	-
chr11_62733221_C/T	missense	*ZBTB3*	0.03	R/Q
chr3_117493114_G/A	missense	*ZBTB20*	0.03	R/W
chr3_43567439_C/T	synonymous	*ZBTB47*	0.17	-
chr10_39218662_A/G	missense	*ZNF33A*	0.08	Q/R
chr10_39220040_C/T	synonymous	*ZNF33A*	0.03	-
chr10_39220351_C/T	missense	*ZNF33A*	0.06	T/I
chr6_35623847_G/A	missense	*ZNF76*	0.03	R/H
chr10_38990719_G/A	missense	*ZNF248*	0.08	S/L
chr10_38990761_T/C	missense	*ZNF248*	0.22	Y/C
chr20_35129549_C/G	synonymous	*ZNF341*	0.08	-
chr20_35154274_C/T	synonymous	*ZNF341*	0.03	-
chr15_45517574_C/A	missense	*ZNF609*	0.03	P/T
chr15_45518830_G/A	synonymous	*ZNF609*	0.28	-
chr15_45519236_G/A	missense	*ZNF609*	0.03	A/T
chr15_45521249_T/C	synonymous	*ZNF609*	0.11	-

chr7_64642929_C/G	synonymous	*ZNF680*	0.03	-
chr7_64643755_T/C	missense	*ZNF680*	0.11	E/G
chr7_64644113_T/G	synonymous	*ZNF680*	0.03	-
chr7_64648152_A/G	synonymous	*ZNF680*	0.03	-
chr19_39457500_C/T	missense	*ZNF793*	0.06	P/L
chr19_39457564_T/C	synonymous	*ZNF793*	0.06	-
chr6_28455567_C/G	synonymous	*ZSCAN9*	0.03	-
chr6_28461117_A/G	missense	*ZSCAN9*	0.03	E/G
chr20_47575973_C/T	missense	*ZSWIM3*	0.03	T/M
chr20_47575976_G/A	missense	*ZSWIM3*	0.06	R/Q
chr20_47576321_A/T	missense	*ZSWIM3*	0.03	E/V
chr20_47576587_G/A	missense	*ZSWIM3*	0.03	D/N
chr20_47577190_G/A	missense	*ZSWIM3*	0.03	A/T
chr20_47577613_G/A	missense	*ZSWIM3*	0.06	D/N

Supplementary Table S2. 12: Possible functional variants unique to *P.t. verus.*

SNP Location	Variant	Gene	Frequency	Change
chr4_50448161_C/T	missense	*ANTXR2*	0.23	L/F
chr19_47014034_A/G	splice acceptor	*APOC1*	0.05	-
chr4_74061326_C/G	missense	*ARL9*	0.05	D/H
chr18_76934278_A/G	missense	*ATP9B*	0.05	M/V
chr7_32495242_G/A	missense	*BBS9*	0.09	V/I
chr20_52652010_G/A	missense	*BCAS4*	0.18	E/K or R/Q
chr1_92473022_G/A	synonymous	*BRDT*	0.05	-
chr2A_108832994_C/T	splice variant	*BUB1*	0.05	-
chr9_89638118_A/G	synonymous	*C9H9orf84*	0.32	-
chr9_89659531_T/C	missense	*C9H9orf84*	0.32	D/G
chr9_89687977_G/A	missense	*C9H9orf84*	0.05	P/S
chr16_79196170_T/G	synonymous	*CA5A*	0.05	-
chr16_79246402_G/A	missense	*CA5A*	0.14	S/L
chr21_22600487_C/T	synonymous	*CBR3*	0.05	-
chr1_42041016_G/A	missense	*CCDC30*	0.14	E/K
chr1_42088992_G/A	missense	*CCDC30*	0.09	R/H
chr1_42096492_C/T	missense	*CCDC30*	0.05	R/C
chr19_12048847_G/C	missense	*CCDC151*	0.09	S/C
chr19_8134631_C/T	missense	*CD209*	0.05	V/I
chr5_134248663_T/C	synonymous	*CDKL3*	0.05	-
chr19_34815734_G/A	stop gain	*CEP89*	0.05	R/STOP
chr19_34852467_T/G	synonymous	*CEP89*	0.14	-
chr1_158507809_G/C	synonymous	*CEP350*	0.05	-
chr1_158513989_C/T	missense	*CEP350*	0.09	P/L
chr1_158514303_C/T	synonymous	*CEP350*	0.27	-
chr1_158539907_G/A	missense	*CEP350*	0.09	R/Q
chr1_158570182_T/C	splice variant	*CEP350*	0.14	-
chr1_158610071_A/C	synonymous	*CEP350*	0.09	-
chr1_158610533_G/A	synonymous	*CEP350*	0.05	-
chr1_158612668_G/A	missense	*CEP350*	0.05	V/I
chr1_158615825_A/G	synonymous	*CEP350*	0.09	-
chr16_57760934_C/T	synonymous	*CES2*	0.14	-
chr5_58831063_T/C	synonymous	*DDX4*	0.05	-
chr21_22683098_C/T	missense	*DOPEY2*	0.09	R/C
chr21_22714308_C/T	synonymous	*DOPEY2*	0.05	-
chr1_20034290_T/C	synonymous	*EIF4G3*	0.05	-
chr5_72314399_G/C	synonymous	*FBXO4*	0.05	P/A
chr1_139797036_C/T	missense	*FCRLA*	0.05	R/C

chr5_108296781_T/C	missense	*FER*	0.05	I/T
chr18_33892745_A/C	missense	*FHOD3*	0.05	K/T
chr3_59310399_G/A	synonymous	*FLNB*	0.18	-
chr13_34203354_C/T	synonymous	*FNDC3A*	0.09	-
chr2B_7414665_G/A	missense	*GLI2*	0.09	G/S
chr2B_7414802_G/C	missense	*GLI2*	0.05	K/N
chr9_79431273_C/T	missense	*GRIN3A*	0.32	G/D
chr3_128256679_G/A	missense	*HEG1*	0.18	P/S
chr3_128257346_C/T	synonymous	*HEG1*	0.18	-
chr3_128272660_C/T	missense	*HEG1*	0.09	G/S
chr16_1796611_T/A	synonymous	*IFT140*	0.09	-
chr3_38521566_A/G	missense	*ITGA9*	0.14	Y/C
chr17_11759340_A/G	synonymous	*KANSL1*	0.14	-
chr17_11862623_C/A	missense	*KANSL1*	0.59	G/V
chr16_26743988_C/T	missense	*KDM8*	0.18	A/V
chr9_61217419_T/A	synonymous	*KIF27*	0.09	-
chr12_40876351_C/G	missense	*KMT2D*	0.23	P/R
chr12_40887894_G/A	splice variant	*KMT2D*	0.05	-
chr12_40901889_G/A	splice variant	*KMT2D*	0.05	-
chr8_21455664_G/A	splice variant	*LGI3*	0.09	-
chr20_14238592_G/A	synonymous	*MACROD2*	0.14	-
chr20_15909742_G/A	missense	*MACROD2*	0.14	D/N
chr6_30894946_C/G	missense	*MDC1*	0.05	G/A
chr6_30903861_C/T	synonymous	*MDC1*	0.05	-
chr6_30904140_T/C	synonymous	*MDC1*	0.05	-
chr6_30904800_G/A	synonymous	*MDC1*	0.14	-
chr6_30905779_G/A	missense	*MDC1*	0.09	P/L
chr7_121225366_T/C	synonymous	*MET*	0.05	-
chr1_1815413_C/T	missense	*MMEL1*	0.18	D/N
chr1_1815414_G/A	synonymous	*MMEL1*	0.05	-
chr20_38654056_C/T	missense	*MROH8*	0.09	V/M
chr20_38708162_A/C	synonymous	*MROH8*	0.05	-
chr11_77242776_C/T	synonymous	*MYO7A*	0.05	-
chr11_77257303_A/G	splice variant	*MYO7A*	0.05	-
chr6_32041142_G/A	synonymous	*NEU1*	0.05	-
chr2A_74678882_G/C	missense	*NOTO*	0.23	E/D
chr3_197978308_T/C	synonymous	*OPA1*	0.14	-
chr3_197989211_A/G	synonymous	*OPA1*	0.09	-
chr3_197991566_A/G	splice variant	*OPA1*	0.14	-
chr3_197999282_T/C	synonymous	*OPA1*	0.23	-
chr6_42201491_G/T	missense	*PGC*	0.05	-
chr10_99012982_C/G	missense	*PI4K2A*	0.05	D/E

chr16_63109446_C/T	missense	*PMFBP1*	0.05	C/Y
chr16_63109811_A/C	missense	*PMFBP1*	0.05	L/V
chr16_63134927_C/T	synonymous	*PMFBP1*	0.05	-
chr20_50385722_C/T	synonymous	*PREX1*	0.09	-
chr20_50395149_G/A	synonymous	*PREX1*	0.09	-
chr19_12055817_C/T	synonymous	*PRKCSH*	0.32	-
chr1_30164742_A/G	synonymous	*PUM1*	0.18	-
chr4_101076025_A/G	missense	*RAP1GDS1*	0.05	K/R
chr12_40864975_C/T	splice variant	*RHEBL1*	0.14	-
chr20_38739168_C/T	synonymous	*RPN2*	0.05	-
chr2B_114534936_C/T	synonymous	*SCG2*	0.05	-
chr2B_129197557_T/C	synonymous	*SCLY*	0.05	-
chr11_70895734_C/T	synonymous	*SHANK2*	0.09	-
chr3_14748421_C/T	missense	*SLC6A6*	0.05	T/I
chr8_92904594_C/T	synonymous	*SLC26A7*	0.05	-
chr12_107828341_T/C	synonymous	*SLC41A2*	0.05	-
chr12_107828400_T/A	missense	*SLC41A2*	0.23	S/C
chr12_107828634_T/C	missense	*SLC41A2*	0.18	T/A
chr12_107828827_A/G	splice variant	*SLC41A2*	0.05	-
chr6_32045369_C/T	splice variant	*SLC44A4*	0.27	-
chr17_58519530_A/G	missense	*SMG8*	0.09	S/G
chr19_42585963_C/T	synonymous	*SPTBN4*	0.05	-
chr15_23249085_C/T	missense	*STARD9*	0.05	A/V
chr15_23253964_A/C	missense	*STARD9*	0.14	K/T
chr15_23254632_G/A	missense	*STARD9*	0.05	A/T
chr15_23258305_G/C	missense	*STARD9*	0.05	R/T
chr15_23258905_A/C	missense	*STARD9*	0.09	H/P
chr15_23262247_C/T	synonymous	*STARD9*	0.23	-
chr15_23293072_C/T	synonymous	*STARD9*	0.14	-
chr7_148406887_G/A	synonymous	*TCAF1*	0.14	-
chr7_148412491_G/A	synonymous	*TCAF1*	0.14	-
chr7_148427193_G/A	synonymous	*TCAF1*	0.23	-
chr5_146661256_C/T	synonymous	*TCERG1*	0.09	-
chr15_23865813_C/T	missense	*TGM7*	0.09	R/Q
chr8_56784192_C/T	synonymous	*TGS1*	0.27	-
chr8_56824331_G/A	synonymous	*TGS1*	0.05	-
chr19_3898349_G/A	synonymous	*TJP3*	0.05	-
chr7_4499355_A/G	synonymous	*TNRC18*	0.09	-
chr22_8373334_G/A	synonymous	*TOP3B*	0.05	-
chr22_8389864_C/T	synonymous	*TOP3B*	0.05	-
chr9_47792305_G/A	synonymous	*TRPM3*	0.05	-
chr15_56887434_T/G	splice variant	*UBE2Q2*	0.14	-

chr17_78449198_G/A	synonymous	*USP36*	0.09	-
chr1_25454932_A/G	splice variant	*ZNF683*	0.05	-
chr1_25458872_G/A	synonymous	*ZNF683*	0.05	-
chr1_39960728_G/A	missense	*ZNF684*	0.09	A/T

3. Selection in the Introgressed Regions of the Chimpanzee Genome

Nye J, Laayouni H, Kuhlwilm M, Mondal M, Marques-Bonet T, Bertranpetit J. Selection in the Introgressed Regions of the Chimpanzee Genome. Genome Biol Evol. 2018 Apr 1;10(4):1132–8. DOI: 10.1093/gbe/evy077

4. Discussion

4.1 Discussion

a) General Dissertation Review

This dissertation is made possible by the advancements in both sequencing data and computational power. A study of this kind is difficult to attempt in a non-model organism; and, these data had their own difficulties. A comprehensive genetic scan requires an adequate sample size. The great ape genetic diversity panel (Prado-Martinez et al. 2013; de Manuel et al. 2016) is the first publicly available data set for apes that has fully sequenced genomes for at least 10 individuals. Calculating 15 statistics (19 if cross population statistics are counted separately) requires access of high computational resources. Moreover, before these statistics could be calculated, the ancestral information of each SNV and the phase of the genetic data were needed; this information was not available at the beginning of the work. Another missing component was a detailed demographic model of all five members of the pan clade. Demographic modeling is complex because many parameters need to be estimated. Much work was invested in elucidating such a model, which is presented in section 2.7b. The demographic model was the basis for the simulations of neutral and selective scenarios, in which I simulated 2 selective scenarios (complete and ongoing selection) at 7 time periods (between 60 kya and current time) for each of the 4 subspecies. The exhaustive selection scenarios resulted in 56,000 simulations of 600,000 bp. After calculating the same statistics in each simulation, I used these data to train a random forest algorithm. The benefit of using a machine learning approach is that you are able to combine many different statistics that each have unique benefits and allow a single conclusion to be drawn from many sampling distributions.

b) Impacts of Demography

From this work, I am able to draw general conclusions about the chimpanzee. The evolutionary divergence of the ape species has been complex. The vast majority of genetic material is orthologous between branches (98.7%, between humans and the Pan genus; Kuhlwilm et al. 2016). This number is debatable, as the quality of the ape reference genomes are lower than that of humans. Regardless, despite a massive amount of similarity, the extant creatures we observe today have immensely differing phenotypes. These differences are rooted in their distinct evolutionary past.

After interrogating each chimpanzee subspecies genome with a battlement of statistics, it has become clear that the demographic history of Pan is a strong driver for differentiation between the subspecies. Out of all the great ape species, the Pan clade has had the largest effective population. This large size was observed, by pairwise sequential Markovian analysis (PSMC; Li and Durbin 2011), to have begun after the divergence of humans, reaching its maximum around 3 million years ago. There was a drastic decline in effective population size when bonobo diverged from the ancestor of chimpanzee, after which time, the chimpanzee population boomed once again, and began to increase around 1 million years ago. The subspecies *P.t. troglodytes* has the largest effective population size of all four, and it reached its personal pinnacle between 200 and 300 thousand years ago. In the most recent ten thousand years, all the populations have experienced profound bottlenecks (Prado-Martinez et al. 2013).

The differences between subspecies in effective population can be observed in the results from this study. *P.t. trolgodytes* returned the largest proportion of regions as under putative positive selection. This is most likely due to its larger size. A large effective population is more immune to genetic drift. It can more easily

respond to the pressures of selection by increasing a beneficial allele's frequency. The other three subspecies have similar effective population sizes (~$\frac{1}{4}$ the size of *P.t. troglodytes*) and returned a similar proportion (~0.5%) of regions under selection. As previously predicted (Nam et al. 2017), we should observe a proportion of signals that scales with N_e. Instead, we observe only ~2.5 times the size, as compared with the other three subspecies. This may be due to the conservative cut off. This study was designed to capture only those signals that are most robust.

The genetic diversity among ape species also is observed to vary. Out of Africa humans, eastern lowland gorillas, bonobos, and western chimpanzees have the lowest amount of genetic diversity. In contrast the highest genetic diversity is observed in western lowland gorilla, bonobos, and central chimpanzees (Prado-Martinez et al. 2013). It is curious to note the highest and lowest harbingers of genetic diversity are members of the same species. Besides the relative proportion of regions under selection, which is likely due to effective population size differences, it is difficult to assess how this dearth of diversity in *P.t. verus* has impacted its population. Genetic diversity can be tricky to relate to selection, as the process of selection itself serves to reduce diversity. However, due to the similar proportions of regions under selection as the other two subspecies, whose genetic diversity is similar to *P.t. troglodytes*, it appears that the impact of genetic diversity on patterns of selective sweeps as a whole, has little affect. *P.t. verus* returned more signals of selection outside coding regions. But it is unclear how genetic diversity affects the target of selection. Further studies are needed to address this issue.

The Pan clade has experienced large amounts of asymmetric gene flow between subspecies, primarily amongst the three subspecies in central Africa, and between species, with bonobo (de Manuel et al. 2016). The large amount of sharing has led to some signatures of selection in genes common between subspecies, but mostly between

P.t. schweinfurthii and *P.t. troglodytes*. This is expected because these two subspecies diverged most recently, meaning that they share thousands of years of adaption that the other two subspecies do not.

c) Targets of Selection

The introgression events from bonobo have driven selection in genes involved with immunity, but primarily in reproduction. This pattern of genes under selection in reproductive traits during introgression scenarios has been observed before (Arnold and Martin 2009). When two highly diverged species share genes, it is likely that the produced hybrid is not viable. However, when introgression makes a viable offspring, selection is often at work in order to avoid the many reproductive barriers that one can imagine two differing organisms face when mating, both pre- and postzygotic. Interestingly, while we identified genes as under selection for male-related fertility in both the genome-wide and introgressed scans, we observed female-related fertility only in the introgressed scans. As discussed in Chapter 3, the female cycles of bonobo and chimpanzee are very different in both the aspect of the monthly cycle and the life-long proportion of fertility (Furuici 1987; Behringer et al. 2014; Ryu et al. 2015). This suggests that while male-related traits are under constant selective strain in chimpanzee, that the female fertility traits are more prohibitive during an introgression event.

Although general categories of phenotypes are observed to be common in the selective history of chimpanzees (i.e., reproduction, immunity, muscle function, neurology, and DNA damage repair), the general pattern is that selection has targeted each individual subspecies in differing ways. Based on the genes that are unique to each subspecies, it is clear that the same general selective pressures

exist. However, each population has had genetic ingenuity, resulting in changes to differing genes or regions. This trend for each subspecies to have differing selective profiles is key when referring to the literature. Much of the comparative selective studies treat chimpanzee as a single group and draw conclusions about the species based on a single subspecies or even a mixture. However, these data would suggest that this is an improper way to research chimpanzee.

The large proportion of immunity-related genes is striking in both the genome-wide and introgression scans. We know that apes are susceptible to most of the same diseases as humans, and that passing diseases between the two species are well-established. (Keele et al. 2006; Van Heuverswyn et al. 2007; Dunay et al. 2018). Data on functional candidates that protect against destructive pathogens could be of great use to researchers in human health. We encountered genes related with infection from influenza, adenovirus, cytomegalovirus, HIV, and malaria, among others; all of which are of utmost importance in healthcare today. Furthermore, as the chimpanzee species is in danger of extinction due to population decline. Evidence suggests that passing diseases between ape and human are on the rise (Dunay et al. 2018) due to humans encroaching on their natural habitat. Fully understanding which pathogens are detrimental to these populations may help mitigate their decline. For example, we detect a strong signal for selection in the anthrax receptor in *P.t. verus*. This population has the lowest population numbers and is critically endangered. It has been decimated recently, and is predicted to be extinct from the wild in the next 150 years. This projection is based on death rates due to anthrax poisoning (Hoffmann et al. 2017). However, our data suggests that selection is acting on the receptor, indicating that there may be hope for some genetic protection.

Interpretation of selection in genetic signals is tricky. When selection favors a particular SNP, that variant will raise if frequency

in the population. Once the variant is fixed, it may be difficult to locate amongst all other fixed variants. It is therefore likely that the selection scan cannot identify the exact favorable SNP. In this work we attempted to isolate possible functional candidates by selecting divergent SNPs that have predicted functional variants. This, however is by no means a complete synopsis of the act of selection.

d) Further Considerations

A previous study (Ruiz-Orera et al. 2015) identified 780 novel genes in the chimpanzee lineage, identifying a number of genes that were not named or studied previously. Although a small fraction of my results, these genes may be important for selection within chimpanzee, as they are specific to its species. However, due to the limited information, they are still in need of further research.

The regions that are observed as under selection in this scan of selection that do not contain genes are likely important components to the evolution of the species. As discussed above, the majority of genetic material is common among apes. Therefore, it is likely that noncoding, regulatory, and epigenetic factors are important in shaping the observed phenotypic diversity. The understanding and significance of noncoding variation is not well understood. It is important to note that signals in regions devoid of genes are based on the underlying genetic diversity in these regions which deviate from neutral expectations and therefore likely impart important impacts on the phenotype of the organism. At this time, it is difficult to interrogate noncoding variants for their phenotypic significance in any organism, but especially for non-model organisms, which have lower quality reference genomes.

Another difficult hurdle when working with non-model organisms is trying to relate genotype to phenotype. This is a difficult task in

any organism, as phenotype is highly multivariate and difficult to measure adequately. In the case of chimpanzee, the phenotype of each subspecies is incomplete. The studies that exist are generally observational and of only one subspecies. This is to be expected as chimpanzee live in a large habitat across Africa, which is a prohibitively large area to observe. Furthermore, observational studies can run the risk of bias, due to their inherent design. Unfortunately, the literature of the subspecies *P.t. ellioti* is limited due to their late subspecies classification. Likewise, all studies conducted before all four subspecies were cataloged are subject to subspecies ambiguity, because *P.t. ellioti* individuals may have been misidentified.

One major goal of this dissertation was to investigate the subspecies as individuals. At the start, I expected for many of the signals of the selection to overlap. However, the data indicated that the unique demography and environmental differences have been key to shaping each population. As such, I attempted to isolate possible functional differences by interrogating unique SNPs that impact the amino acid code.

4.2 Community Resource

This work serves to improve the quality of resources for this non-model organism. All the data calculated genome-wide for all four subspecies can be easily accessed through the interactive web browser (Figure 4.1) located at:

http://hsb.upf.edu/chimp_browser/index.html

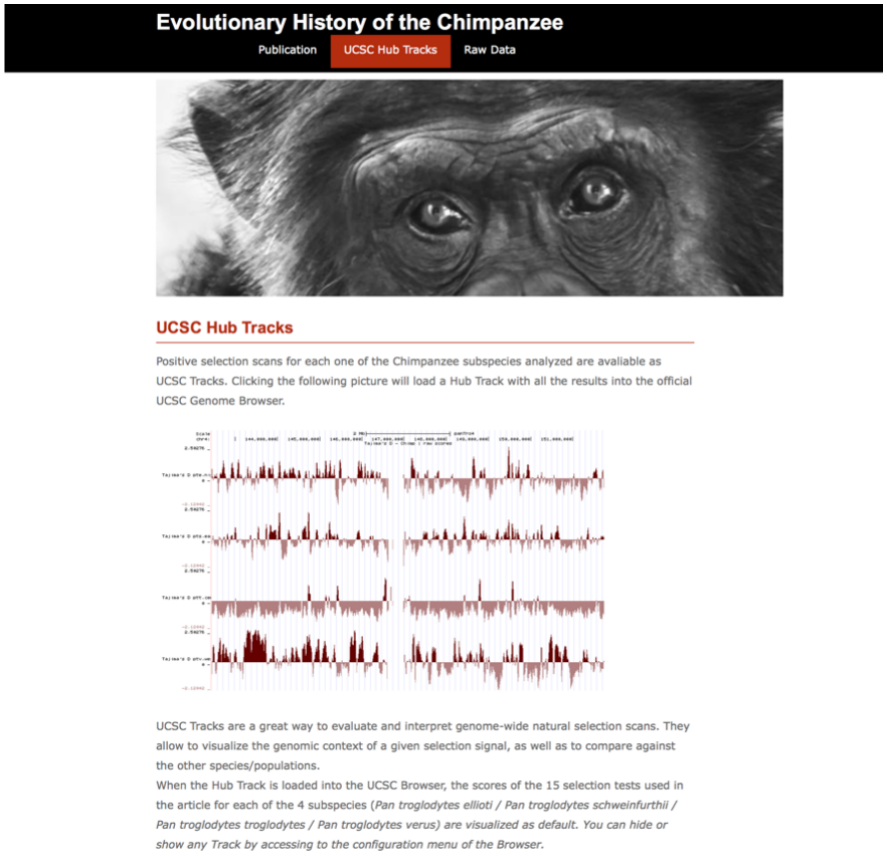


Figure 4.1: Portal of available data, located at http://hsb.upf.edu/chimp_browser/ucsc_tracks.html

I hope that the work performed in this dissertation can be used by members of the scientific community that are interested in chimpanzee genetics, ape evolution, and speciation, among others. These data can be downloaded from UCSC genome browser using the table function, and will also be available for download straight from my website after publication of Chapter 2, currently submitted to the Journal of Evolutionary Biology. The interactive UCSC browser allows for any of the 19 calculated statistics to be compared across all 4 subspecies, and each track can be selectively turned off or on. Furthermore, the probability of each region under selection, calculated through the random forest algorithm is

available. Therefore, regions that are trending toward significance, but not discussed in this dissertation are available. Likewise, regions that are devoid of genes, and regions that contain chimp-specific genes can easily be interrogated using these tools.

The data collected in this dissertation brings the community of researchers one step closer to untangling the complex interactions between all aspects of genetics that ultimately give rise to the chimpanzee phenotype.

6. References

1000 Genomes Project Consortium. 2005. A haplotype map of the human genome. *Nature*. 437(7063):1299-1320.

Abe H, Yamada N, Kamata K, Kuwaki T, Shimada M, Osuga J, Shionoiri F, Yahagi N, Kadowaki, T, Tamemoto, H, et al. 1998. Hypertension, hypertriglyceridemia, and impaired endothelium-dependent vascular relaxation in mice lacking insulin receptor substrate-1. *Journal of Clinical Investigation*. 101(8):1784-1788.

Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, Babrzadeh F, Gharizadeh B, Luo M, Plummer FA, et al. 2011. The Shaping of Modern Human Immune Systems by Multiregional Admixture with Archaic Humans. *Science*. 334(6052):89-94.

Aho S, Rothenberger K, Tan EML, Ryoo YW, Cho BH, McLean WHI, Uitto J. 1999. Human periplakin: Genomic organization in a clonally unstable region of chromosome 16p with an abundance of repetitive sequence elements. *Genomics*. 56(2):160-168.

Anderson MJ, Dixson AF. 2002. Sperm competition - Motility and the midpiece in primates. *Nature*. 416(6880):496-496.

Anoh AE, Murthy S, Akoua-Koffi C, Couacy-Hymann E, Leendertz FH, Calvignac-Spencer S, Ehlers B. 2018. Cytomegaloviruses in a Community of Wild Nonhuman Primates in Tai National Park, Cote D'Ivoire. *Viruses*. 10(1):e11.

Arnold ML, Martin NH. 2009. Adaptation by introgression. *Journal of Biology*. 8(9):82.

Aruga J, Mikoshiba K. 2003. Identification and characterization of Slitrk, a novel neuronal transmembrane protein family controlling neurite outgrowth. *Molecular and Cellular Neuroscience*. 24(1):117-129.

Auton A, Fledel-Alon A, Pfeifer S, Venn O, Segurel L, Street T, Leffler EM, Bowden R, Aneas I, Broxholme J, et al. 2012. A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. *Science*. 336(6078):193-198.

Bachand F. 2007. Protein arginine methyltransferases: From unicellular eukaryotes to humans. *Eukaryotic Cell*. 6(6):889-898.

Bakhteeva LB, Khaibullina SF, Lombardi VK, Anokhin VA, Rizvanov AA, Boichuk SV, Nagimova FI. 2016. Significance of Polymorphism in 2 ', 5 '-Oligoadenylate Synthetase Genes in HIV Infection. *Modern Technologies in Medicine*. 8(1):99-104.

Bamshad MJ, Mummidi S, Gonzalez E, Ahuja SS, Dunn DM, Watkins WS, Wooding S, Stone AC, Jorde LB, Weiss RB, et al. 2002. A strong signature of balancing selection in the 5 ' cis-regulatory region of CCR5. *Proceedings of the National Academy of Sciences of the United States of America*. 99(16):10539-10544.

Barrett L, Gaynor D, Rendall D, Mitchell D, Henzi SP. 2004. Habitual cave use and thermoregulation in chacma baboons (Papio hamadryas ursinus). *Journal of Human Evolution*. 46(2):215-222.

Basel-Vanagaite L, Dallapiccola B, Ramirez-Solis R, Segref A, Thiele H, Edwards A, Arends MJ, Miro X, White JK, Desir J, et al. 2012.

Deficiency for the Ubiquitin Ligase UBE3B in a Blepharophimosis-Ptosis-Intellectual-Disability Syndrome. *American Journal of Human Genetics*. 91(6):998-1010.
Bauer CK, Wulfsen I, Schafer R, Glassmeier G, Wimmers S, Flitsch J, Ludecke DK, Schwarz JR. 2003. HERG K+ currents in human prolactin-secreting adenoma cells. *European Journal of Physiology*. 445(5):589-600.
Baum J, Ward RH, Conway DJ. 2002. Natural selection on the erythrocyte surface. *Molecular Biology and Evolution*. 19(3):223-229.
Baumann P, Cech TR. 2001. Pot1, the putative telomere end-binding protein in fission yeast and humans. *Science*. 292(5519):1171-1175.
Behringer V, Deschner T, Deimel C, Stevens JMG, Hohmann G. 2014. Age-related changes in urinary testosterone levels suggest differences in puberty onset and divergent life history strategies in bonobos and chimpanzees. *Hormones and Behavior*. 66(3):525-533.
Belinky F, Nativ N, Stelzer G, Zimmerman S, Iny Stein T, Safran M, Lancet D. 2015. PathCards: multi-source consolidation of human biological pathways. *Database: the journal of biological databases and curation*. 2015.
Bolusani S, Young BA, Cole NA, Tibbetts AS, Momb J, Bryant JD, Solmonson A, Appling DR. 2011. Mammalian MTHFD2L Encodes a Mitochondrial Methylenetetrahydrofolate Dehydrogenase Isozyme Expressed in Adult Tissues. *Journal of Biological Chemistry*. 286(7):5166-5174.
Borzan V, Tomasevic B, Kurbel S. 2014. Hypothesis: Possible respiratory advantages for heterozygote carriers of cystic fibrosis linked mutations during dusty climate of last glaciation. *Journal of Theoretical Biology*. 363:164-168.
Brain C, Mitchell D. 1999. Body temperature changes in free-ranging baboons (Papio hamadryas ursinus) in the Namib Desert, Namibia. *International Journal of Primatology*. 20(4):585-598.
Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature*. 478(7369):343-348.
Breiman L, Friedman JH, Olshen RA, Stone CJ. 1984. Classification and regression trees. Wadsworth Inc.:Belmont, CA.
Breiman L. 1996. Bagging predictors. *Machine Learning*. 24(2):123-140.
Breiman L. 2001. Random forests. *Machine Learning*. 45(1):5-32.
Bronchain OJ, Chesneau A, Monsoro-Burq AH, Jolivet P, Paillard E, Scanlan TS, Demeneix BA, Sachs LM, Pollet N. 2017. Implication of thyroid hormone signaling in neural crest cells migration: Evidence from thyroid hormone receptor beta knockdown and NH3 antagonist studies. *Molecular and Cellular Endocrinology*. 439(C):233-246.
Buitendijk H, Fagrouch Z, Niphuis H, Bogers WM, Warren KS, Verschoor EJ. 2014. Retrospective Serology Study of Respiratory Virus Infections in Captive Great Apes. *Viruses*. 6(3):1442-1453.
Cagan A, Theunert C, Laayouni H, Santpere G, Pybus M, Casals F, Prufer K, Navarro A, Marques-Bonet T, Bertranpetit J, et al. 2016. Natural

Selection in the Great Apes. *Molecular Biology and Evolution*. 33(12):3268-3283.
Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*. 10(3)195-205.
Charlesworth D, Morgan MT, Charlesworth B. 1993. Mutation Accumulation in Finite Populations. *Journal of Heredity*. 84(5):321-325.
Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. *PLOS Genetics*. 2(4):379-384.
Chen C, Tian F, Lu L, Wang Y, Xiao Z, Yu C, Yu X. 2015. Characterization of Cep85-a new antagonist of Nek2A that is involved in the regulation of centrosome disjunction. *Journal of Cell Science*. 128(17):3290-3303.
Christensen GL, Wooding SP, Ivanov IP, Atkins JF, Carrell DT. 2006. Sequencing and haplotype analysis of the Activator of CREM in the Testis (ACT) gene in populations of fertile and infertile males. *Molecular Human Reproduction*. 12(4):257-262.
Chung WS, Clarke LE, Wang GX, Stafford BK, Sher A, Chakraborty C, Joung J, Foo LC, Thompson A, Chen C, et al. 2013. Astrocytes mediate synapse elimination through MEGF10 and MERTK pathways. *Nature*. 504(7480):394-400.
Colleran KM, Ratliff DM, Burge MR. 2003. Potential association of thyrotoxicosis with vitamin B and folate deficiencies, resulting in risk for hyperhomocysteinemia and subsequent thromboembolic events. *Endocrine Practice*. 9(4):290-295.
Coolidge HJ Jr. 1933. Pan paniscus. Pigmy chimpanzee from south of the Congo River. *American Journal of Physical Anthropology*. 18(1)1-59.
Dadakova E, Brozova K, Piel AK, Stewart FA, Modry D, Celer V, Hrazdilova K. 2018. Adenovirus infection in savanna chimpanzees (Pan troglodytes schweinfurthii) in the Issa Valley, Tanzania. *Archives of Virology*. 163(1):191-196.
Dannemann M, Andres AM, Kelso J. 2016. Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. *American Journal of Human Genetics*. 98(2):399.
Darwin C. 1859. On the Origin of Species by Means of Natural Selection, Or, the Preservation of Favoured Races in the Struggle for Life. London.
de Luca C, Kowalski TJ, Zhang YY, Elmquist JK, Lee C, Kilimann MW, Ludwig T, Liu SM, Chua SC. 2005. Complete rescue of obesity, diabetes, and infertility in db/db mice by neuron-specific LEPR-B transgenes. *Journal of Clinical Investigation*. 115(12):3484-3493.
de Manuel M, Kuhlwilm M, Frandsen P, Sousa VC, Desai T, Prado-Martinez J, Hernandez-Rodriguez J, Dupanloup I, Lao O, Hallast P, et al. 2016. Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science*. 354(6311):477-481.
Depaulis F, Mousset S, Veuille M. 2003. Power of neutrality tests to detect bottlenecks and hitchhiking. *Journal of Molecular Evolution*. 57:S190-S200.
Deschamps M, Laval G, Fagny M, Itan Y, Abel L, Casanova JL, Patin E, Quintana-Murci L. 2016. Genomic Signatures of Selective Pressures

and Introgression from Archaic Hominins at Human Innate Immunity Genes. *American Journal of Human Genetics*. 98(1):5-21.
Di A, Xiong S, Ye Z, Malireddi RKS, Kometani S, Zhong M, Mittal M, Hong Z, Kanneganti TD, Rehman J, et al. 2018. The TWIK2 Potassium Efflux Channel in Macrophages Mediates NLRP3 Inflammasome-Induced Inflammation. *Immunity*. 49(1):56-65.
Dunay E, Apakupakul K, Leard S, Palmer JL, Deem SL. 2018. Pathogen Transmission from Humans to Great Apes is a Growing Threat to Primate Conservation. *Ecohealth*. 15(1):148-162.
Durrleman S, Pennec X, Trouve A, Ayache N, Braga J. 2012. Comparison of the endocranial ontogenies between chimpanzees and bonobos via temporal regression and spatiotemporal registration. *Journal of Human Evolution*. 62(1):74-88.
Durski KN, McCollum AM, Nakazawa Y, Petersen BW, Reynolds MG, Briand S, Djingarey MH, Olson V, Damon IK, Khalakdina A. 2018. Emergence of Monkeypox - West and Central Africa, 1970-2017. *Morbidity and Mortality Weekly Report*. 67(10):306-310.
Ely JJ, Dye B, Frels WI, Fritz J, Gagneux P, Khun HH, Switzer WM, Lee DR. 2005. Subspecies composition and founder contribution of the captive US chimpanzee (Pan troglodytes) population. *American Journal of Primatology*. 67(2):223-241.
Ewing G, Hermisson J. 2010. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*. 26(16)2064-2065.
Feng GP, Krejci E, Molgo J, Cunningham JM, Massoulie J, Sanes JR. 1999. Genetic analysis of collagen Q: Roles in acetylcholinesterase and butyrylcholinesterase assembly and in synaptic structure and function. *Journal of Cell Biology*. 144(6):1349-1360.
Fischer A, Wiebe V, Paabo S, Przeworski M. 2004. Evidence for a complex demographic history of chimpanzees. *Molecular Biology and Evolution*. 21(5):799-808.
Fisher R. 1930. The Genetical Theory of Natural Selection. Clarendon Press:Oxford.
Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*. 147(2):915-925.
Fu YX, Huai HY. 2003. Estimating mutation rate: How to count mutations? *Genetics*. 164(2):797-805.
Fu YX, Li WH. 1993. Statistical Tests of Neutality of Mutations. *Genetics*. 133(3):693-709.
Furuichi T. 1987. Sexual Swelling, Receptivity, and Grouping of Wild Pygmy Chimpanzee Females at Wamba, Zaire. *Primates*. 28(3):309-318.
Gonder MK, Oates JF, Disotell TR, Forstner MRJ, Morales JC, Melnick DJ. 1997. A new west African chimpanzee subspecies? *Nature*. 388(6640):337.
Goto H, Watanabe K, Araragi N, Kageyama R, Tanaka K, Kuroki Y, Toyoda A, Hattori M, Sakaki Y, Fujiyama A, et al. 2009. The identification and functional implications of human-specific "fixed" amino acid

substitutions in the glutamate receptor family. *BMC Evolutionary Biology*. 9:224.
Goverdhan SV, Howell MW, Mullins RF, Osmond C, Hodgkins PR, Self J, Avery K, Lotery AJ. 2005. Association of HLA class I and class II polymorphisms with age-related macular degeneration. *Investigative Ophthalmology & Visual Science*. 46(5):1726-1734.
Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MHY, et al. 2010. A Draft Sequence of the Neandertal Genome. *Science*. 328(5979):710-722.
Gruber T, Clay Z. 2016. A Comparison Between Bonobos and Chimpanzees: A Review and Update. *Evolutionary Anthropology*. 25(5):239-252.
GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. *Nature*. 550(7675):204-213.
Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genetics*. 5(10):e1000695.
Haldane JBS, Jayakar SD. 1963. Polymorphism due to selection of varying direction. *Journal of Genetics*. 58(2):237-242.
Hardy GH. 1908. Mendelian proportions in a mixed population. *Science*. 28(706):49-50.
Hare B, Wobber V, Wrangham R. 2012. The self-domestication hypothesis: evolution of bonobo psychology is due to selection against aggression. *Animal Behaviour*. 83(3):573-585.
Hedrick PW, Thomson G. 1983. Evidence for Balancing Selection at HLA. *Genetics*. 104(3):449-456.
Hedrick PW. 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology*. 22(18):4606-4618.
Hey J. 2006. Recent advances in assessing gene flow between diverging populations and species. *Current Opinion in Genetics & Development*. 16(6):592-596.
Hofer T, Ray N, Wegmann D, Excoffier L. 2009. Large Allele Frequency Differences between Human Continental Groups are more Likely to have Occurred by Drift During range Expansions than by Selection. *Annals of Human Genetics*. 73:95-108.
Hoffmann C, Zimmermann F, Biek R, Kuehl H, Nowak K, Mundry R, Agbor A, Angedakin S, Arandjelovic M, Blankenburg A, et al. 2017. Persistent anthrax as a major driver of wildlife mortality in a tropical rainforest. *Nature*. 548(7665):82-86.
Holmes RS, Spradling-Reeves KD, Cox LA. 2016. Evolution of Vertebrate Solute Carrier Family 9B Genes and Proteins (SLC9B): Evidence for a Marsupial Origin for Testis Specific SLC9B1 from an Ancestral Vertebrate SLC9B2 Gene. *Journal of Phylogenetics & Evolutionary Biology*. 4(3):167.
Hosseini I, Gama L, Mac Gabhann F. 2015. Multiplexed Component Analysis to Identify Genes Contributing to the Immune Response during Acute SIV Infection. *PLOS One*. 10(5):e0126843.

Howie BN, Donnelly P, Marchini J. 2009. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genetics*. 5(6):e1000529.
Hu W, Gauthier L, Baibakov B, Jimenez-Movilla M, Dean J. 2010. FIGLA, a Basic Helix-Loop-Helix Transcription Factor, Balances Sexually Dimorphic Gene Expression in Postnatal Oocytes. *Molecular and Cellular Biology*. 30(14):3661-3671.
Huang CM, Wei FW, Li M, Li YB, Sun RY. 2003. Sleeping cave selection, activity pattern and time budget of white-headed langurs. *International Journal of Primatology*. 24(4):813-824.
Huang YH, Zhu C, Kondo Y, Anderson AC, Gandhi A, Russell A, Dougan SK, Petersen BS, Melum E, Pertel T, et al. 2015. CEACAM1 regulates TIM-3-mediated tolerance and exhaustion. *Nature*. 517(7534):386-390.
Hubert RS, Vivanco I, Chen E, Rastegar S, Leong,K, Mitchell SC, Madraswala R, Zhou YH, Kuo J, Raitano AB, et al. 1999. STEAP: A prostate-specific cell-surface antigen highly expressed in human prostate tumors. *Proceedings of the National Academy of Sciences of the United States of America*. 96(25):14523-14528.
Hudson RR. 1983. Properties of a Neutral Allele Model with Intragenic Recombination. *Theoretical Population Biology*. 23(2):183-201.
Hudson RR, Coyne JA. 2002. Mathematical consequences of the genealogical species concept. *Evolution*. 56(8):1557-1565.
Huerta-Sanchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, Liang Y, Yi X, He M, Somel M, et al. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*. 512(7513):194-197.
Jeffreys AJ, Neumann R, Panayi M, Myers S, Donnelly P. 2005. Human recombination hot spots hidden in regions of strong marker association. *Nature Genetics*. 37(6):601-606.
Juric I, Aeschbacher S, Coop G. 2016. The Strength of Selection against Neanderthal Introgression. *PLOS Genetics*. 12(11):e1006340.
Kaessmann H, Wiebe V, Paabo S. 1999. Extensive nuclear DNA sequence diversity among chimpanzees. *Science*. 286(5442):1159-1162.
Kaplan NL, Darden T.; Hudson RR. 1988. The Coalescent Process in Models with Selection. *Genetics*. 120(3):819-829.
Karleusa L, Mahmutefendic H, Tomas MI, Zagorac GB, Lucin P. 2018. Landmarks of endosomal remodeling in the early phase of cytomegalovirus infection. *Virology*. 515:108-122.
Kass GV. 1980. An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*. 20(2):119-127.
Keele BF, Van Heuverswyn F, Li Y, Bailes E, Takehisa J, Santiago ML, Bibollet-Ruche F, Chen Y, Wain LV, Liegeois F, et al. 2006. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science*. 313(5786):523-526.
Kelly JK. 1997. A test of neutrality based on interlocus associations. *Genetics*. 146(3):1197-1206.
Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Paabo S. 2005. Parallel patterns of evolution in the

genomes and transcriptomes of humans and chimpanzees. *Science*. 309(5742):1850-1854.
Khaitovich P, Muetzel B, She XW, Lachmann M, Hellmann I, Dietzsch J, Steigele S, Do HH, Weiss G, Enard W, et al. 2004. Regional patterns of gene expression in human and chimpanzee brains. *Genome Research*. 14(8):1462-1473.
Kimura M. 1955. Stochastic Processes and Distribution of Gene Frequencies under Natural Selection. *Cold Spring Harbor Symposia on Quantitative Biology*. 20:33-53.
Kimura M. 1956. A Model of a Genetic System which leads to Closer Linkage by Natural-Selection. *Evolution*. 10(3):278-287.
Kimura M. 1968. Evolutionary Rate at Molecular Level. *Nature*. 217(5129):624-626.
Kimura M. 1969. Number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*. 61(4):893-903.
Kimura M. 1983. The neutral theory of molecular evolution. Cambridge University Press:New York.
Kingman JFC. 1982. The coalescent. *Stochastic Processes and their Applications*. 13(3):235-248.
Kobayashi Y, Watanabe M, Okada Y, Sawa H, Takai H, Nakanishi M, Kawase Y, Suzuki H, Nagashima K, Ikeda K, et al. 2002. Hydrocephalus, situs inversus, chronic sinusitis, and male infertility in DNA polymerase lambda-deficient mice: Possible implication for the pathogenesis of immotile cilia syndrome. *Molecular and Cellular Biology*. 22(8):2769-2776.
Kofler R, Schloetterer C. 2012. Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. *Bioinformatics*. 28(15):2084-2085.
Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, et al. 2002. A high-resolution recombination map of the human genome. *Nature Genetics*. 31(3):241-247.
Krief S, Berny P, Gumisiriza F, Gross R, Demeneix B, Fini JB, Chapman CA, Chapman LJ, Seguya A, Wasswa J. 2017. Agricultural expansion as risk to endangered wildlife: Pesticide exposure in wild chimpanzees and baboons displaying facial dysplasia. *Science of the Total Environment*. 598:647-656.
Krzywinski M, Altman N. 2017. Classification and regression trees. *Nature Methods*. 14(8):755-756.
Kuhl HS, Kalan AK, Arandjelovic M, Aubert F, D'Auvergne L, Goedmakers A, Jones S, Kehoe L, Regnaut S, Tickle A, et al. 2016. Chimpanzee accumulative stone throwing. *Scientific Reports*. 6:22219-22219.
Kuhlwilm M, de Manuel M, Nater A, Greminger MP, Krutzen M, Marques-Bonet T. 2016. Evolution and demography of the great apes. *Current Opinion in Genetics & Development*. 41:124-129.
Kuhlwilm M, Gronau I, Hubisz MJ, de Filippo C, Prado-Martinez J, Kircher M, Fu QM, Burbano HA, Lalueza-Fox C, de la Rasilla M, et al. 2016. Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature*. 530(7591):429-433.

Leendertz FH, Deckers M, Schempp W, Lankester F, Boesch C, Mugisha L, Dolan A, Gatherer D, McGeoch DI, Ehlers B. 2009. Novel cytomegaloviruses in free-ranging and captive great apes: phylogenetic evidence for bidirectional horizontal transmission. *Journal of General Virology*. 90:2386-2394.
Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature*. 47(7357):493-U84.
Li Y, Ndjango JB, Learn GH, Ramirez MA, Keele BF, Bibollet-Ruche F, Liu W, Easlick JL, Decker JM, Rudicell RS, et al. 2012. Eastern Chimpanzees, but Not Bonobos, Represent a Simian Immunodeficiency Virus Reservoir. *Journal of Virology*. 86(19):10776-10791.
Liaw A, Wiener M. 2002. Classification and Regression by randomForest. *R News*. 2:18-22.
Locatelli S, Harrigan RJ, Clee PRS, Mitchell MW, McKean KA, Smith TB, Gonder MK. 2016. Why Are Nigeria-Cameroon Chimpanzees (Pan troglodytes ellioti) Free of SIVcpz Infection? *PLOS One*. 11(8):e0160788.
Lyndaker AM, Vasileva A, Wolgemuth DJ, Weiss RS, Lieberman HB. 2013. Clamping down on mammalian meiosis. *Cell Cycle*. 12(19):3135-3145.
Mallon AM, Iyer V, Melvin D, Morgan H, Parkinson H, Brown SDM, Flicek P, Skarnes WC. 2012. Accessing data from the International Mouse Phenotyping Consortium: state of the art and future plans. *Mammalian Genome*. 23(9-10):641-652.
Mannowetz N, Wandernoth P, Wennemuth G. 2012. Basigin interacts with both MCT1 and MCT2 in murine spermatozoa. *Journal of Cellular Physiology*. 227(5):2154-2162.
Manteca A, Schonfelder J, Alonso-Caballero A, Fertin MJ, Barruetabena N, Faria BF, Herrero-Galan E, Alegre-Cebollada J, De Sancho D, Perez-Jimenez R. 2017. Mechanochemical evolution of the giant muscle protein titin as inferred from resurrected proteins. *Nature Structural & Molecular Biology*. 24(8):652-657.
Maynard-Smith J, Haigh J. 1974. Hitch-Hicking Effect of a Favorable Gene. *Genetics Research*. 23(1):23-35.
McKay PB, Griswold CK. 2014. A comparative study indicates both positive and purifying selection within ryanodine receptor (RyR) genes, as well as correlated evolution. *Journal of Experimental Zoology*. 321(3):151-163.
McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. *Genome Biology*. 17(1):122.
McVicker G, Gordon D, Davis C, Green P. 2009. Widespread Genomic Signatures of Natural Selection in Hominid Evolution. *PLOS Genetics*. 5(5):e1000471.
Mendel G. 1866. Experiments in Plant Hybridization. Verhandlungen des Naturforschenden Vereines in Brunn:4.
Mondal M, Casals F, Xu T, Dall'Olio GM, Pybus M, Netea MG, Comas D, Laayouni H, Li Q, Majumder PP, et al. 2016. Genomic analysis of

Andamanese provides insights into ancient human migration into Asia and adaptation. *Nature Genetics*. 48(9):1066-1070.
Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, Burns E, Ostrer H, Price AL, Reich D. 2011. The History of African Gene Flow into Southern Europeans, Levantines, and Jews. *PLOS Genetics*. 7(4):e1001373.
Morgan TH, Sturtevant AH, Muller HJ, Bridges CB. 1915. The mechanism of mendelian heredity. Henry Holt and Company:New York.
Morin PA, Moore JJ, Chakraborty R, Jin L, Goodall J, Woodruff DS. 1994. Kin Selection, Social-Structure, Gene Flow, and the Evolution of Chimpanzees. *Science*. 265(5176):1193-1201.
Morin PA, Moore JJ, Woodruff DS. 1992. Identification of Chimpanzee Subspecies with DNA from Hair and Allele-Specific Probes. *Proceedings of the Royal Society B-Biological Sciences*. 249(1326):293-297.
Nam K, Munch K, Mailund T, Nater A, Greminger MP, Krutzen M, Marques-Bonet T, Schierup MH. 2017. Evidence that the rate of strong selective sweeps increases with population size in the great apes. *Proceedings of the National Academy of Sciences of the United States of America*. 114(7):1613-1618.
Nei M. 1987. Molecular Evolutionary Genetics. Columbia University Press:New York.
Nei M, Li WH. 1979. Mathematical-Model for studying Genetic-Variation in Variation in terms of Restriction Endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*. 76(10):5269-5273.
Nelson EA, Walker SR, Li W, Liu XS, Frank DA. 2006. Identification of human STAT5-dependent gene regulatory elements based on interspecies homology. *Journal of Biological Chemistry*. 281(36):26216-26224.
Nowick K, Gernat T, Almaas E, Stubbs L. 2009. Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proceedings of the National Academy of Sciences of the United States of America*. 106(52):22358-22363.
Nye J, Laayouni H, Kuhlwilm M, Mondal M, Marques-Bonet T, Bertranpetit J. 2018. Selection in the Introgressed Regions of the Chimpanzee Genome. *Genome Biology and Evolution*. 10(4):1132-1138.
O'Neill MC, Umberger BR, Holowka NB, Larson SG, Reiser PJ. 2017. Chimpanzee super strength and human skeletal muscle evolution. *Proceedings of the National Academy of Sciences of the United States of America*. 114(28):7343-7348.
Ohta T. 1973. Slightly Deleterious Mutant Substitution in Evolution. *Nature*. 246(5428):96-98.
Ohta T, Kimura M. 1969. Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics*. 63(1):229-238.
Pan MH, Wang F, Lu Y, Tang F, Duan X, Zhang Y, Xiong B, Sun SC. 2017. FHOD1 regulates cytoplasmic actin-based spindle migration for mouse oocyte asymmetric cell division. *Journal of cellular physiology*. 233(3):2270-2278.

Park TJ, Haigo SL, Wallingford JB. 2006. Ciliogenesis defects in embryos lacking inturned or fuzzy function are associated with failure of planar cell polarity and Hedgehog signaling. *Nature Genetics*. 38(3):303-311.
Pelka P, Miller MS, Cecchini M, Yousef AF, Bowdish DM, Dick F, Whyte P, Mymryk JS. 2011. Adenovirus E1A Directly Targets the E2F/DP-1 Complex. *Journal of Virology*. 85(17):8841-8851.
Peng B, Kimmel, M. 2005. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*. 21(18):3686-3687.
Perez-Perez A, Sanchez-Jimenez F, Maymo J, Duenas JL, Varone C, Sanchez-Margalet V. 2015. Role of leptin in female reproduction. *Clinical Chemistry and Laboratory Medicine*. 53(1):15-28.
Pers TH, Karjalainen JM, Chan Y, Westra HJ, Wood AR, Yang J, Lui JC, Vedantam S, Gustafsson S, Esko T, et al. 2015. Biological interpretation of genome-wide association studies using predicted gene functions. *Nature Communications*. 6:5890.
Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*. 19(5):826-837.
Pilbrow V. 2006. Population systematics of chimpanzees using molar morphometrics. *Journal of Human Evolution*. 51(6):646-662.
Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, et al. 2013. Great ape genetic diversity and population history. *Nature*. 499(7459):471-475.
Prescott NJ, Malcolm S. 2002. Folate and the face: Evaluating the evidence for the influence of folate genes on craniofacial development. *Cleft Palate-Craniofacial Journal*. 39(3):327-331.
Pritchard JK, Di Rienzo A. 2010. Adaptation - not by sweeps alone. *Nature Reviews Genetics*. 11(10):665-667.
Pruetz JD. 2007. Evidence of cave use by savanna chimpanzees (Pan troglodytes verus) at Fongoli, Senegal: implications for thermoregulatory behavior. *Primates*. 48(4):316-319.
Pruetz JD, Bertolani P, Ontl KB, Lindshield S, Shelley M, Wessling EG. 2015. New evidence on the tool-assisted hunting exhibited by chimpanzees (Pan troglodytes verus) in a savannah habitat at Fongoli, Senegal. *Royal Society Open Science*. 2(4).
Prufer K, Muetzel B, Do HH, Weiss G, Khaitovich P, Rahm E, Paabo S, Lachmann M, Enard W. 2007. FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics*. 8:41.
Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*. 505(7481):43-49.
Pumroy RA, Ke S, Hart DJ, Zachariae U, Cingolani G. 2015. Molecular Determinants for Nuclear Import of Influenza A PB2 by Importin alpha Isoforms 3 and 7. *Structure*. 23(2):374-384.

Pybus M, Dall'Olio GM, Luisi P, Uzkudun M, Carreno-Torres A, Pavlidis P, Laayouni H, Bertranpetit J, Engelken J. 2014. 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Research*. 42(D1):D903-D909.

Pybus M, Luisi P, Dall'Olio GM, Uzkudun M, Laayouni H, Bertranpetit J, Engelken J. 2015. Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. *Bioinformatics*. 31(24):3946-3952.

Ramirez-Soriano A, Ramos-Onsins SE, Rozas J, Calafell F, Navarro A. 2008. Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. *Genetics*. 179(1):555-567.

Ramos-Onsins SE, Rozas J. 2002. Statistical properties of new neutrality tests against population growth. *Molecular Biology and Evolution*. 19(12):2092-2100.

Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, Pugach I, Ko AMS, Ko YC, Jinam TA, Phipps ME, et al. 2011. Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania. *American Journal of Human Genetics*. 89(4):516-528.

Resa-Infante P, Thieme R, Ernst T, Arck PC, Ittrich H, Reimer R, Gabriel G. 2014. Importin-alpha 7 Is Required for Enhanced Influenza A Virus Replication in the Alveolar Epithelium and Severe Lung Damage in Mice. *Journal of Virology*. 88(14):8166-8179.

Riedel G, Platt B, Micheau J. 2003. Glutamate receptor function in learning and memory. *Behavioural Brain Research*. 140(1-2):1-47.

Rief M, Gautel M, Oesterhelt F, Fernandez JM, Gaub HE. 1997. Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science*. 276(5315):1109-1112.

Rozas J, Gullaud M, Blandin G, Aguade M. 2001. DNA variation at the rp49 gene region of Drosophila simulans: Evolutionary inferences from an unusual haplotype structure. *Genetics*. 158(3):1147-1155.

Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabido E, Kondova I, Bontrop R Marques-Bonet T, Alba MM. 2015. Origins of De Novo Genes in Human and Chimpanzee. *PLOS Genetics.* 11(12):e1005721.

Ryu H, Hill DA, Furuichi T. 2015. Prolonged maximal sexual swelling in wild bonobos facilitates affiliative interactions between females. *Behaviour*. 152(3-4):285-311.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature*. 419(6909):832-837.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. 2007. Int HapMap, C., Genome-wide detection and characterization of positive selection in human populations. *Nature*. 449(7164):913-918.

Sagane K, Hayakawa K, Kai J, Hirohashi T, Takahashi E, Miyamoto N, Ino M, Oki T, Yamazaki K, Nagasu T. 2005. Ataxia and peripheral nerve

hypomyelination in ADAM22-deficient mice. *BMC Neuroscience*. 6:33.
Saifullah MK, Fox DA, Sarkar S, Abidi SMA, Endres J, Piktel J, Haqqi TM, Singer NG. 2004. Expression and characterization of a novel CD6 ligand in cells derived from joint and epithelial tissues. *Journal of Immunology*. 173(10):6125-6133.
Sankararaman S, Mallick S, Patterson N, Reich D. 2016. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Current Biology*. 26(9):1241-1247.
Sankararaman S, Mallick S, Dannemann M, Pruefer K, Kelso J, Paeaebo S, Patterson N, Reich D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*. 507(7492):354-357.
Saveanu L, Carroll O, Weimershaus M, Guermonprez P, Firat E, Lindo V, Greer F, Davoust J, Kratzer R, Keller SR, et al. 2009. IRAP Identifies an Endosomal Compartment Required for MHC Class I Cross-Presentation. *Science*. 325(5937):213-217.
Schultz DW, Klein ML, Humpert AJ, Luzier CW, Persun V, Schain M, Mahan A, Runckel C, Cassera M, Vittal V, et al. 2003. Analysis of the ARMD1 locus: evidence that a mutation in HEMICENTIN-1 is associated with age-related macular degeneration in a large family. *Human Molecular Genetics*. 12(24):3315-3323.
Schwartz C, Fischer M, Mamchaoui K, Bigot A, Lok T, Verdier C, Duperray A, Michel R, Holt I, Voit T, et al. 2017. Lamins and nesprin-1 mediate inside-out mechanical coupling in muscle cell precursors through FHOD1. *Scientific Reports*. 7(1):1253.
Scully EJ, Basnet S, Wrangham RW, Muller MN, Otali E, Hyeroba D, Grindle KA, Pappas TE, Thompson ME, Machanda Z, et al. 2018. Lethal Respiratory Disease Associated with Human Rhinovirus C in Wild Chimpanzees, Uganda, 2013. *Emerging Infectious Diseases*. 24(2):267-274.
Seale P, Bjork B, Yang W, Kajimura S, Chin S, Kuang S, Scime A, Devarakonda S, Conroe HM, Erdjument-Bromage H, et al. 2008. PRDM16 controls a brown fat/skeletal muscle switch. *Nature*. 454(7207):961-967.
Sharp PM, Shaw GM, Hahn BH. 2005. Simian immunodeficiency virus infection of chimpanzees. *Journal of Virology*. 79(7):3891-3902.
Shea BT, Coolidge HJ. 1988. Craniometric Differentiation and Systematics in the genus Pan. *Journal of Human Evolution*. 17(7):671-685.
Shimizu T, Tashiro-Yamaji J, Hayashi M, Inoue Y, Ibata M, Kubota T, Tanigawa N, Yoshida R. 2010. HLA-B62 as a possible ligand for the human homologue of mouse macrophage MHC receptor 2 (MMR2) on monocytes. *Gene*. 454(1-2):31-38.
Singh N, Bhat VK, Tiwari A, Kodaganur SG, Tontanahal SJ, Sarda A, Malini KV, Kumar A. 2017. A homozygous mutation in TRIM36 causes autosomal recessive anencephaly in an Indian family. *Human Molecular Genetics*. 26(6):1104-1114.
Sittig LJ, Herzing LBK, Xie H, Batra KK, Shukla PK, Redei EE. 2012. Excess folate during adolescence suppresses thyroid function with permanent

deficits in motivation and spatial memory. *Genes Brain and Behavior*. 11(2):193-200.
Slatkin M. 2001. Simulating genealogies of selected alleles in a population of variable size. *Genetical Research*. 78(1):49-57.
Smith FH, Jankovic I, Karavanic I. 2005. The assimilation model, modern human origins in Europe, and the extinction of Neandertals. *Quaternary International*. 137:7-19.
Song Y, Endepols S, Klemann N, Richter D, Matuschka FR, Shih CH, Nachman MW, Kohn MH. 2011. Adaptive Introgression of Anticoagulant Rodent Poison Resistance by Hybridization between Old World Mice. *Current Biology*. 21(15):1296-1301.
Sowden JC, Holt JKL, Meins A, Smith HK, Bhattacharya SS. 2001. Expression of Drosophila omb-related T-box genes in the developing human and mouse neural retina. *Investigative Ophthalmology & Visual Science*. 42(13):3095-3102.
Starr DA, Saffery R, Li ZX, Simpson AE, Choo KHA, Yen TJ, Goldberg ML. 2000. HZwint-1,a novel human kinetochore component that interacts with HZW10. *Journal of Cell Science*. 113(11):1939-1950.
Strehl S, Glatt K, Liu QM, Glatt H, Lalande M. 1998. Characterization of two novel protocadherins (PCDH8 and PCDH9) localized on human chromosome 13 and mouse chromosome 14. *Genomics*. 53(1):81-89.
Sun C, Huo DZ, Southard C, Nemesure B, Hennis A, Leske MC, Wu SY, Witonsky DB, Olopade OI, Di Rienzo A. 2011. A signature of balancing selection in the region upstream to the human UGT2B4 gene and implications for breast cancer risk. *Human Genetics*. 130(6):767-775.
Tajima F. 1989. Statistical-Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics*. 123(3):585-595.
Tassabehji M, Hammond P, Karmiloff-Smith A, Thompson P, Thorgeirsson SS, Durkin ME, Popescu NC, Hutton T, Metcalfe K, Rucka A, et al. 2005. GTF2IRD1 in craniofacial development of humans and mice. *Science*. 310(5751):1184-1187.
Thavachelvam K, Gad HH, Ibsen MS, Despres P, Hokland M, Hartmann R, Kristiansen H. 2015. Rapid Uptake and Inhibition of Viral Propagation by Extracellular OAS1. *Journal of Interferon and Cytokine Research*. 35(5):359-366.
Thoemmes MS, Stewart FA, Hernandez-Aguilar RA, Bertone MA, Baltzegar DA, Borski RJ, Cohen N, Coyle KP, Piel AK, Dunn RR. 2018. Ecology of sleeping: the microbial and arthropod associates of chimpanzee beds. *Royal Society Open Science*. 5(5):180382.
Torii M, Hashimoto-Torii K, Levitt P, Rakic P. 2009. Integration of neuronal clones in the radial cortical columns by EphA and ephrin-A signalling. *Nature*. 461(7263):524-528.
Van Heuverswyn F, Peeters M. 2007. The origins of HIV and implications for the global epidemic. *Current infectious disease reports*. 9(4):338-346.
Van Sloun PPH, Varlet I, Sonneveld E, Boei J, Romeijn RJ, Eeken JCJ, De Wind N. 2002. Involvement of mouse Rev3 in tolerance of endogenous and exogenous DNA damage. *Molecular and Cellular Biology*. 22(7):2159-2169.

Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. 2016. SIFT missense predictions for genomes. *Nature Protocols*. 11(1):1-9.
Vernot B, Tucci S, Kelso J, Schraiber JG, Wolf AB, Gittelman RM, Dannemann M, Grote S, McCoy RC, Norton H, et al. 2016. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science*. 352(6282):235-239.
Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting Natural Selection in Genomic Data. *Annual Review of Genetics*. 47:97-120.
Voight BF, Kudaravalli S, Wen XQ, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLOS Biology*. 4(3):446-458.
Wall JD. 1999. Recombination and the power of statistical tests of neutrality. *Genetical Research*. 74(1):65-79.
Wall JD. 2000. A comparison of estimators of the population recombination rate. *Molecular Biology and Evolution*. 17(1):156-163.
Wang HY, Tang H, Shen CKJ, Wu CI. 2003. Rapidly evolving genes in human. I. The glycophorins and their possible role in evading malaria parasites. *Molecular Biology and Evolution*. 20(11):1795-1804.
Weinberg W. 1908. ‹ber den Nachweis der Vererbung beim Menschen. Jahresh. Ver. Vaterl. Naturkd. *Wurttemb*. 64:368-382.
Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population-Structure. *Evolution*. 38(6):1358-1370.
Whitney KD, Randell RA, Rieseberg LH. 2006. Adaptive introgression of herbivore resistance traits in the weedy sunflower Helianthus annuus. *American Naturalist*. 167(6)794-807.
Wilde A, Zheng YX. 1999. Stimulation of microtubule aster formation and spindle assembly by the small GTPase Ran. *Science*. 284(5418):1359-1362.
Wilson JN, Rockett K, Keating B, Jallow M, Pinder M, Sisay-Joof F, Newport M, Kwiatkowski D. 2006. A hallmark of balancing selection is present at the promoter region of interleukin 10. *Genes and Immunity*. 7(8):680-683.
Wobber V, Lipson S, Hare B, Wrangham R, Ellison P. 2012. Species differences in the ontogeny of testosterone production between chimpanzees and bonobos. *American Journal of Physical Anthropology*. 147:305-306.
Wold MS. 1997. Replication protein A: A heterotrimeric, single-stranded DNA-binding protein required for eukaryotic DNA metabolism. *Annual Review of Biochemistry*. 66:61-92.
Wolff T, O'Neill RE, Palese P. 1998. NS1-binding protein (NS1-BP): a novel human protein that interacts with the influenza A virus nonstructural NS1 protein is relocalized in the nuclei of infected cells. *Journal of Virology*. 72(9):7170-7180.
Won YJ, Hey J. 2005. Divergence population genetics of chimpanzees. *Molecular Biology and Evolution*. 22(2):297-307.
Wright S. 1931. Evolution in Mendelian populations. Genetics. 16(2):0097-0159.
Xiang D, Liang S, Wang H, Ma J, Chen C, He S, Liu P, Jiang W, Yuan X, Li X, et al. 2017. Application of co-immunoprecipitation coupled LC-MS/MS for identification of sperm immunogenic membrane antigens. *International Journal of Clinical and Experimental Pathology*. 10(4):4198-4209.

Yang MA, Malaspinas AS, Durand EY, Slatkin M. 2012. Ancient Structure in Africa Unlikely to Explain Neanderthal and Non-African Genetic Similarity. *Molecular Biology and Evolution*. 29(10):2987-2995.
Zhou S, Tanaka K, O'Keeffe M, Qi M, El-Assaad F, Weaver JC, Chen G, Weatherall C, Wang Y, Giannakopoulos B, et al. 2016. CD117(+) Dendritic and Mast Cells Are Dependent on RasGRP4 to Function as Accessory Cells for Optimal Natural Killer Cell-Mediated Responses to Lipopolysaccharide. *PLOS One*. 11(3):e0151638.