bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Identifying Lineage-specific Targets of Darwinian Selection by a Bayesian Analysis of Genomic Polymorphisms and Divergence from Multiple Species

Shilei Zhao[a,c,1], Tao Zhang[b,c,1], Qi Liu[a,c], Yongming Liu[a,c], Hao Wu[a,c], Bing Su[b,c], Peng Shi[b,c], Hua Chen[a,c,*]

[a] *CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*
[b] *State Key Laboratory of Evolution and Genetic Resources, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China*
[c] *University of Chinese Academy of Sciences, Beijing 100049, China*

Abstract

We present a method that jointly analyzes the polymorphism and divergence sites in genomic sequences of multiple species to identify the genes under positive or negative selection and pinpoints the occurrence time of selection to a specific lineage of the species phylogeny. This method integrates population genetics models using the Bayesian Poisson random field framework and combines information over all gene loci to boost the power to detect selection. The method provides posterior distributions of the fitness effects of each gene along with parameters associated with the evolutionary history, including the species divergence times and effective population sizes of external species. A simulation is performed, and the results demonstrate that our method provides accurate estimates of these population genetic parameters.

The proposed method is applied to genomic sequences of humans, chimpanzees, gorillas and orangutans, and a spatial and temporal map is constructed of the natural selection that occurred during the evolutionary history of the four Hominidae species. In addition to FOXP2 and other known genes, we identify a new list of lineage-specific targets of Darwinian selection. The positively selected genes in the human lineage are enriched in pathways of gene expression regulation, immune system, metabolism etc. Interestingly, some pathways, such as gene expression, are significantly enriched with positively selected genes, whereas other

*Corresponding author: chenh@big.ac.cn
[1]These two authors contributed equally.

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

pathways, such as metabolism, are enriched with both positively and negatively selected genes. Our analysis provides insights into Darwinian evolution in the coding regions of humans and great apes and thus serves as a basis for further molecular and functional studies.

Keywords: positive selection, adaptive evolution, dN/dS ratio, MK test, Bayesian, Poisson random field.

1. Introduction

Comparing the genomic sequences of multiple species is useful for studying evolutionary mechanism and identifying functionally important genes (Fay et al., 2002; Smith and Eyre-Walker, 2002; Clark et al., 2003; Rogers and Gibbs, 2014). Codon comparison across different species is perhaps the most widely used approach in comparative genomics (Li et al., 1985; Hughes and Nei, 1988; Yang, 1998). This method estimates the ratio of the number of replacement sites to the number of synonymous sites for different species (dN/dS ratio) to detect the genes under recurrent directional selection. The method was further developed to identify ratio changes on a specific branch along the phylogenetic tree (Yang, 1998; Yang and Nielsen, 2002; Yang, 2007; Zhang et al., 2005). These methods were designed to analyze single gene locus with one sequence per species without making use of population genetic samples, and they have been extensively used to identify targets of natural selection in various species (e.g., Clark et al. (2003)).

The McDonald-Kreitman test (MK test) is another method that uses codon replacement to detect Darwinian selection. The MK test compares coding sequences from two species and requires that at least one of the two species has multiple sequences. Polymorphism sites (within each species) and divergence sites (between species) of coding sequences are identified in the sample and further classified as synonymous or replacement (non-synonymous) sites. A contingency table is constructed based on the above four site types, and the chi-square test is used to examine the equality of the within-species ratios and the between-species ratios of replacements over synonymous sites. The underlying assumptions of the MK test are that (1) synonymous sites are under neutrality while non-synonymous sites are potentially under positive or negative selection; and (2) if one of the two species is under long-term recurrent

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

selection since their divergence, then the ratio of the non-synonymous site number over the synonymous site number for the between-species divergence sites will be significantly greater than that of the within-species sites, and this situation is reversed for negative selection (McDonald and Kreitman, 1991).

Comparing within-species polymorphism to between-species divergence improves the power of detecting selection and enables the ability to control for the inter-locus heterogeneity of mutation rates, and it can also distinguish between positive selection and the relaxation of negative selection (Wyckoff et al., 2000). However, the MK test analyzes each gene individually and has limited power to detect loci under weak or moderate selection, especially when the numbers of these sites are small. Moreover, the chi-square test is not model based and cannot provide the strength and direction of selection acting on mutations in the genome. To tackle these problems, Bustamante et al. (2001) extended the MK test to embrace the population genetic model using the Poisson random field theory developed by Sawyer and Hartl (1992). The McDonald-Kreitman Poisson random field method (MKPRF) created by Bustamante et al. (2001) assumes that the number of sites in the MK table follows Poisson distributions, and their means are parameterized as functions of the population history, mutation rates and selection intensity. The method then uses a Bayesian approach to combine information across different gene loci and obtain the posterior distributions of selection intensity for non-synonymous sites of each gene.

Bustamante et al. (2005) applied the MKPRF Bayesian approach to analyze $11,624$ coding regions of 39 human and 1 chimpanzee sequences, and they identified 304 genes under rapid evolution since the divergence of the two species. However, the MKPRF approach applies only to directionless comparisons of two species; therefore, the method cannot determine whether selection occurred in the human or the chimpanzee lineage.

With the development of sequencing technology, abundant population genomic data are now available for multiple species. However, few efficient methods are available that can simultaneously analyze genomic polymorphisms and divergence from multiple species. The method developed in this paper is designed to satisfy a growing request for such methods, and it is an extension of the MKPRF method that simultaneously analyzes polymorphism sites and divergence sites from multiple species (aka, high-dimensional MKPRF, HDMKPRF).

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Based on the joint pattern of divergence and polymorphism sites from multiple species, the proposed method not only identifies genes under selection in any of the analyzed species but also pinpoints the occurrence time of selection to a specific lineage of the species phylogeny. Compared with codon-based phylogenetic methods, such as PAML (phylogenetic analysis by maximum likelihood), the new method gains power by jointly analyzing both between-species divergence and within-species polymorphism sites and combining the information from all gene loci with a Bayesian approach. Furthermore, the method provides estimates of phylogenetic and population genetic parameters, such as the divergence times of species, mutation rates of each gene loci and effective population sizes of different species. The method constructs the spatial and temporal landscape of natural selection during the evolutionary history of the species. We applied the method to the genomic sequences of four species of Hominidae: humans, chimpanzees, gorillas and orangutans. We identified a set of genes under rapid adaptation in the human and chimpanzee lineages as well as genes under weak purifying selection. The positively selected genes are enriched in pathways related to gene expression regulation, immunity, metabolism, neurological diseases, etc, thereby providing a natural selection map for further investigation of the molecular mechanism in Hominidae evolution.

Methods

Poisson random field model (PRF) for three species

In the scenario involving three species (Figure 1), we analyze n_1, n_2 and n_3 aligned sequences for a coding region from the three species. If we assume an infinite-sites mutation model with no introgression among species, the polymorphism and divergence sites can be classified into 14 types according to their joint patterns across the three species (see Table 1). For example, polymorphism sites that are segregating within species 1 and fixed in species 2 and 3 are denoted as $P_{1\sim(2,3)}$; divergence sites with one allele that is fixed in species 1 and another that is fixed in species 2 and 3 are denoted as $D_{1\sim(2,3)}$ etc. Following the MK test, a chi-square test with six degrees of freedom can be immediately constructed based on Table 1 to test the deviation of entries of the contingency table from randomness. In the

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

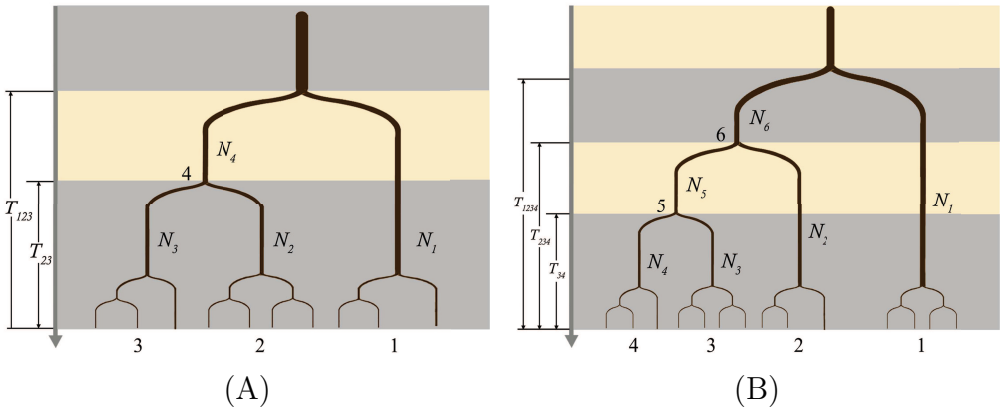


Figure 1: An illustration of the genealogies for three species and four species and the parameters.

		Synonymous	Replacement
$1 \sim (2,3)$	Divergence $D_{1\sim(2,3)}$	$\theta_{s,1}(T_{12} - \frac{1}{2}T_{23} + M(n_1))$	$\theta_{r,1}\frac{2\gamma_1}{1-e^{(-2\gamma_1)}}(T_{12} - \frac{1}{2}T_{23} + G(n_1,\gamma_1))$
	Polymorphism $P_{1\sim(2,3)}$	$\theta_{s,1}L(n_1)$	$\theta_{r,1}\frac{2\gamma_1}{1-e^{(-2\gamma_1)}}F(n_1,\gamma_1)$
$2 \sim (1,3)$	Divergence $D_{2\sim(1,3)}$	$\theta_{s,2}(\frac{1}{2}v_3T_{23} + M(n_2))$	$\theta_{r,2}\frac{2\gamma_2}{1-e^{(-2\gamma_2)}}(\frac{1}{2}v_2T_{23} + G(n_2,\gamma_2))$
	Polymorphism $P_{2\sim(1,3)}$	$\theta_{s,2}L(n_2)$	$\theta_{r,2}\frac{2\gamma_2}{1-e^{(-2\gamma_2)}}F(n_2,\gamma_2)$
$3 \sim (1,2)$	Divergence $D_{3\sim(1,2)}$	$\theta_{s,3}(\frac{1}{2}v_3T_{23} + M(n_3))$	$\theta_{r,3}\frac{2\gamma_3}{1-e^{(-2\gamma_3)}}(\frac{1}{2}v_3T_{23} + G(n_3,\gamma_3))$
	Polymorphism $P_{3\sim(1,2)}$	$\theta_{s,3}L(n_3)$	$\theta_{r,3}\frac{2\gamma_3}{1-e^{(-2\gamma_3)}}F(n_3,\gamma_3)$
$(2,3) \sim 1$	Polymorphism $P_{(2,3)\sim1}$	$\theta_{s,4}H(n_2,n_3,N_2,N_3,T_{23})$	$\theta_{r,4}\frac{2\gamma_4}{1-e^{-2\gamma_4}}I(n_2,n_3,\gamma_2,\gamma_3,\gamma_4)$

Table 1: Three-species McDonald-Kreitman table.

following sections, we set up the model using the more efficient Bayesian Poisson random field framework.

In the Poisson random field modeling assumption, each of the 14 entries in Table 1 follows a Poisson distribution, and the means are parameterized with population genetic models (Sawyer and Hartl, 1992). The population genetic parameters $\mathbf{\Gamma}$ include: the divergence time between species 1 and the common ancestor of species 2 and 3 (T_{123}), the divergence time between species 2 and 3 (T_{23}), and the effective haploid population sizes of the three species (N_1, N_2, and N_3). For each gene locus i, mutation rate μ_s^i is observed for synonymous sites, mutation rate μ_r^i is observed for replacement sites, and selection intensities γ_1^i, γ_2^i, and γ_3^i are observed in the three species.

Synonymous polymorphism sites in species 3, which are denoted by $P^s_{3\sim(1,2)}$, are neutral mutations that have occurred in species 3 since the divergence of species 2 and 3. If the divergence time is sufficiently large, then the population allele frequency x of $P^s_{3\sim(1,2)}$ follows

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

a stationary distribution (Sawyer and Hartl, 1992)

$$f(x) = \frac{2N\mu_s}{x}, \tag{1}$$

where $N = N_3$. Thus, the expected number in a sample with n_3 sequences is:

$$\begin{aligned} \mathbb{E}P^s_{3\sim(1,2)} &= \int_0^1 2N_3\mu_s \frac{1 - x^{n_3} - (1-x)^{n_3}}{x} dx \\ &= 2N_3\mu_s \sum_{i=1}^{n_3-1} \frac{1}{i} = \theta_{s,3} L(n_3), \end{aligned} \tag{2}$$

with $L(n) = \sum_1^{n-1} \frac{1}{i}$.

The synonymous divergence sites, $D^s_{3\sim(1,2)}$, include those sites segregating in the population but fixed in the sample:

$$\int_0^1 2N_3\mu_s x^{n_3} \frac{dx}{x} \tag{3}$$

$$= 2N_3\mu_s \frac{1}{n_3} = \theta_{s,3} M(n_3). \tag{4}$$

The synonymous divergence sites also include those sites fixed in the population of species 3: $\mu_s T_{23}$. If we scale time T_{23} in units of N_1, then

$$\mu_s T_{23} = 1/2\theta_3 * \frac{N_1 T_{23}}{N_3} = \frac{1}{2}\theta_{s,3} T_{23} v_3, \tag{5}$$

and $v_3 = \frac{N_1}{N_3}$.

For non-synonymous (replacement) sites, we know that the stationary population allele frequency under selection is as follows:

$$g(x) = \frac{2N\mu_r}{x(1-x)} \frac{1 - e^{-2\gamma(1-x)}}{1 - e^{-2\gamma}}, \tag{6}$$

where x is the population allele frequency and $\gamma = Ns$, with s representing the selection

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

intensity (Sawyer and Hartl, 1992). The fixation rate is:

$$\mu_r \frac{2\gamma}{1-e^{-2\gamma}}, \tag{7}$$

Thus, the number of replacement sites along branch leading to species 3 is:

$$\frac{1}{2}\theta_{r,3} v_3 T_{23} \mu_r \frac{2\gamma_3}{1-e^{-2\gamma_3}}, \tag{8}$$

where T_{23} is in units of N_1 generations.

The number of replacement sites fixed in the sample n_3 is as follows:

$$\begin{aligned} & \int_0^1 x^{n_3} \frac{2N_3\mu_r}{x(1-x)} \frac{1-e^{-2\gamma_3(1-x)}}{1-e^{-2\gamma_3}} dx \\ = \; & 2N_3\mu_r \frac{2\gamma_3}{1-e^{-2\gamma_3}} G(n_3, \gamma_3) \end{aligned} \tag{9}$$

with

$$G(n,\gamma) = \int_0^1 x^{n-1} \frac{1-e^{-2\gamma(1-x)}}{2\gamma(1-x)} dx, \tag{10}$$

Similarly, the expected number of segregating replacement sites, $P^r_{s\sim(1,2)}$, is:

$$2N_3\mu_r \frac{2\gamma_3}{1-e^{-2\gamma_3}} F(n_3, \gamma_3), \tag{11}$$

with

$$F(n,\gamma) = \int_0^1 \frac{1-x^n-(1-x)^n}{x(1-x)} \frac{1-e^{-2\gamma(1-x)}}{2\gamma} dx. \tag{12}$$

Through similar logic, we can obtain the expected number of divergence and polymorphism sites for $P_{2\sim(1,3)}$, $D_{2\sim(1,3)}$, $P_{1\sim(2,3)}$ and $D_{1\sim(2,3)}$ (see Table 1 for details).

In the aforementioned paragraphs, we assume that the divergence time T_{23} is sufficiently large; thus, the chance of observing polymorphism sites shared by species 2 and 3, $P_{(2,3)\sim1}$, has a very low probability. However, for closely-related species with $T_{23} \leq$ TMRCA of n_2, the expected number of neutral polymorphic sites segregating in species 2 and 3, $P_{(2,3)\sim1}$, cannot be ignored and should be calculated from the joint allele frequency of species 2 and

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

3 $f(y, z|T_{23}, N_2, N_3)$ (see AppendixA for details):

$$\begin{aligned} & \int_0^1 \int_0^1 (1 - y^{n_2} - (1-y)^{n_2})(1 - z^{n_3} - (1-z)^{n_3}) \\ & \times f(y, z|T_{23}, N_2, N_3, N_4, \mu_s) dy dz \\ = \; & \theta_{s,4} H(n_2, n_3, N_2, N_3, T_{23}), \end{aligned} \tag{13}$$

where $\theta_{s,4} = 2N_4\mu_s$ stands for scaled mutation rate in species 4 and y and z are the allele frequencies in species 2 and 3, respectively. Additionally,

$$\begin{aligned} H(n_2, n_3, N_2, N_3, T_{23}) = & \int_0^1 \int_0^1 \int_0^1 (1 - y^{n_2} - (1-y)^{n_2})(1 - z^{n_3} - (1-z)^{n_3}) \\ & \times \frac{1}{x} \phi(y|x, T_{23}, N_2) \phi(z|x, T_{23}, N_3) dx dy dz, \end{aligned} \tag{14}$$

with $\phi(y|x, T, N)$ representing the transient allele frequency distribution conditional on an initial frequency x, population size N and time T (see Equation 25 for the detailed form).

The expected number of selected polymorphic sites segregating in species 2 and 3 is

$$\begin{aligned} & \int_0^1 \int_0^1 (1 - y^{n_2} - (1-y)^{n_2})(1 - z^{n_3} - (1-z)^{n_3}) \\ & \times g(y, z|T_{23}, N_2, N_3, \gamma_4) dy dz \\ = \; & \theta_{r,4} \frac{2\gamma_4}{1 - e^{-2\gamma_4}} I(n_2, n_3, \gamma_2, \gamma_3, \gamma_4), \end{aligned} \tag{15}$$

where γ_2, γ_3, and γ_4 are the selection coefficients in species 2, 3 and, 4 respectively. And,

$$\begin{aligned} I(n_2, n_3, \gamma_2, \gamma_3, \gamma_4) = & \int_0^1 \int_0^1 \int_0^1 \frac{(1 - y^{n_2} - (1-y)^{n_2})(1 - z^{n_3} - (1-z)^{n_3})}{2x(1-x)\gamma_4} \\ & \times \psi(y|x, T_{23}, N_2, \gamma_2) \psi(z|x, T_{23}, N_3, \gamma_3) dx dy dz, \end{aligned} \tag{16}$$

with $\psi(y|x, T, N, \gamma)$ representing the transient allele frequency distribution conditional on an initial frequency x, population size N, time T and selection intensity γ (see Appendix A for details).

Accordingly, when under the assumption $T_{23} \leq$ TMRCA of n_2, $P^s_{3\sim(1,2)}$ and other entries in Table 1 should be derived in a new form, which can be found in Appendix A.

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Assuming a Poisson random field model, the joint probability of the data given parameter $\Gamma = \{\theta^l_{s,i}, \theta^l_{r,i}, T_{123}, T_{23}, \gamma^l_i, i \in \{1,2,3\}, 1 \leq l \leq L\}$ is the product of the individual entries of Table 1 for all the L gene loci:

$$\begin{aligned} &\Pr(\mathbf{P}, \mathbf{D}|\gamma, \theta, \mathbf{T}) \\ &= \prod_{l=1}^{L} \prod_{c\in\{s,r\}} \prod_{i,j,k\in\{1,2,3\}} \Pr(D^l_{c,i\sim(j,k)}|\mathbf{T}, \gamma^l_c, \theta^l_c) \Pr(P^l_{c,i\sim(j,k)}|\mathbf{T}, \gamma^l_c, \theta^l_c) \Pr(P^l_{c,(2,3)\sim 1}|\mathbf{T}, \gamma^l_c, \theta^l_c), \end{aligned} \tag{17}$$

where $\Pr(\cdot)$ denotes the Poisson distributions.

Poisson random field model (PRF) for four species

The phylogeny of four species is shown in Figure 1 (B). Similar to the three-species case, the four-species data contains branch-specific divergence and polymorphism sites (e.g., $D_{1\sim(2,3,4)}$, $P_{1\sim(2,3,4)}$, etc.). In addition, there are some unique site patterns corresponding to internal branches connecting species 5 and 6. Mutations that occur on this branch could have generated multiple site patterns in the modern four-species samples. These patterns could be created by polymorphism sites shared by species 3 and 4, which is denoted by $P_{(3,4)\sim(1,2)}$; by sites with one allele type fixed in species 3 and 4, and the other allele type fixed in species 1 and 2, which is denoted by $D_{(3,4)\sim(1,2)}$; or by sites with one allele type that is fixed in species 3 and another fixed in species 1, 2 and 4 (or perhaps the reverse). Since we assume no migration or introgression between species and an infinite-sites mutation model, common mutations shared between species 3 and 4 can only be descended from ancestral mutations existing in N_5. Similar to $P_{s,(2,3)\sim 1}$ in the three-species scenario, the expected number of neutral polymorphic sites segregating in species 3 and 4 is as follows:

$$\begin{aligned} &\int_0^1 \int_0^1 (1 - y^{n_3} - (1-y)^{n_3})(1 - z^{n_4} - (1-z)^{n_4}) \\ &\times f(y, z|T_{34}, N_3, N_4) dy dz \\ = \;& \theta_{s,5} H(n_3, n_4, N_3, N_4, T_{34}), \end{aligned} \tag{18}$$

where y and z now represent the allele frequencies in species 3 and 4, respectively (see Appendix A for details). Similarly, the expected number of polymorphic replacement sites

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

segregating in species 3 and 4 is

$$\begin{aligned} & \int_0^1 \int_0^1 (1-y^{n_3}-(1-y)^{n_3})(1-z^{n_4}-(1-z)^{n_4}) \\ & \times g(y,z|T_{34},N_3,N_4,\gamma_5)dydz \\ = \; & \theta_{r,5}\frac{2\gamma_5}{1-e^{-2\gamma_5}} I(n_3,n_4,\gamma_3,\gamma_4,\gamma_5). \end{aligned} \tag{19}$$

If the divergence time between species 3 and 4 is sufficiently large, then these common ancestral polymorphic sites are mostly lost or fixed, the values of $H(n_3,n_4)$ and $I(n_3,n_4)$ become negligible, and $P_{(3,4)\sim(1,2)}$ collapses into $P_{3\sim(1,2,4)}$, $P_{4\sim(1,2,3)}$ and $D_{(3,4)\sim(1,2)}$. Thus, although $P_{(3,4)\sim(1,2)}$ exists, the number could be only represent a small proportion and provides limited information for inference. This is consistent with the genomic data we observed in the four Hominidae species (see the result section), with $P_{(n_3,n_4)\sim(1,2)}$ only representing 0.2896% of the total number of segregating sites $(197,878)$.

The expected number of fixed synonymous sites $D^s_{(3,4)\sim(1,2)}$ includes two components:

$$\begin{aligned} & \int_0^1 \int_0^1 y^{n_3} z^{n_4} \times f(y,z|T,N_3,N_4)dydz \\ = \; & \theta_{s,5} J(n_3,n_4,N_3,N_4,T_{34}), \end{aligned} \tag{20}$$

and

$$\frac{1}{2}\theta_{s,5}(T_{234}-T_{34})v_5, \tag{21}$$

with $v_5 = N_1/N_5$.

Similarly, the expected number of fixed nonsynonymous sites $D^r_{(3,4)\sim(1,2)}$ includes two parts:

$$\begin{aligned} & \int_0^1 \int_0^1 y^{n_3} z^{n_4} \times g(y,z|T,N_3,N_4,\gamma_5)dydz \\ = \; & \theta_{r,5}\frac{2\gamma_5}{1-e^{-2\gamma_5}} K(n_3,n_4,\gamma_3,\gamma_4,\gamma_5). \end{aligned} \tag{22}$$

and

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

		Synonymous	Replacement
$1 \sim (2,3,4)$	Divergence $D_{1\sim(2,3,4)}$	$\theta_{s,1}(T_{1234} - \frac{1}{2}T_{234} + M(n_1))$	$\theta_{r,1}\frac{2\gamma_1}{1-e^{-2\gamma_1}}(T_{1234} - \frac{1}{2}T_{234} + G(n_1,\gamma_1))$
	Polymorphism $P_{1\sim(2,3,4)}$	$\theta_{s,1}L(n_1)$	$\theta_{r,1}\frac{2\gamma_1}{1-e^{-2\gamma_1}}F(n_1,\gamma_1)$
$2 \sim (1,3,4)$	Divergence $D_{2\sim(1,3,4)}$	$\theta_{s,2}(\frac{1}{2}v_2T_{234} + M(n_2))$	$\theta_{r,2}\frac{2\gamma_2}{1-e^{-2\gamma_2}}(\frac{1}{2}v_2T_{234} + G(n_2,\gamma_2))$
	Polymorphism $P_{2\sim(1,3,4)}$	$\theta_{s,2}L(n_2)$	$\theta_{r,2}\frac{2\gamma_2}{1-e^{-2\gamma_2}}F(n_2,\gamma_2)$
$3 \sim (1,2,4)$	Divergence $D_{3\sim(1,2,4)}$	$\theta_{s,3}(\frac{1}{2}v_3T_{34} + M(n_3))$	$\theta_{r,3}\frac{2\gamma_3}{1-e^{-2\gamma_3}}(\frac{1}{2}v_3T_{34} + G(n_3,\gamma_3))$
	Polymorphism $P_{3\sim(1,2,4)}$	$\theta_{s,3}L(n_3)$	$\theta_{r,3}\frac{2\gamma_3}{1-e^{-2\gamma_3}}F(n_3,\gamma_3)$
$4 \sim (1,2,3)$	Divergence $D_{4\sim(1,2,3)}$	$\theta_{s,4}(\frac{1}{2}v_4T_{34} + M(n_4))$	$\theta_{r,4}\frac{2\gamma_4}{1-e^{-2\gamma_4}}(\frac{1}{2}v_4T_{34} + G(n_4,\gamma_4))$
	Polymorphism $P_{4\sim(1,2,3)}$	$\theta_{s,4}L(n_4)$	$\theta_{s,4}\frac{2\gamma_4}{1-e^{-2\gamma_4}}F(n_4,\gamma_4)$
$(3,4) \sim (1,2)$	Divergence $D_{(3,4)\sim(1,2)}$	$\theta_{s,5}(\frac{1}{2}v_5(T_{234} - T_{34})$ $+J(n_3,n_4,N_3,N_4,T_{34}))$	$\theta_{s,5}\frac{2\gamma_5}{1-e^{-2\gamma_5}}(\frac{1}{2}v_5(T_{234} - T_{34})$ $+K(n_3,n_4,\gamma_3,\gamma_4,\gamma_5))$
	Polymorphism $P_{(3,4)\sim(1,2)}$	$\theta_{s,5}H(n_3,n_4,N_3,N_4,T_{34})$	$\theta_{r,5}\frac{2\gamma_5}{1-e^{-2\gamma_5}}I(n_3,n_4,\gamma_3,\gamma_4,\gamma_5)$

Table 2: Four-species McDonald-Kreitman table.

$$\frac{1}{2}\theta_{r,5}v_5(T_{234} - T_{34})\frac{2\gamma_5}{1-e^{-2\gamma_5}}. \tag{23}$$

The first part contains a small proportion when the divergence time T_{34} is large.

Results

Simulation

We tested the performance of the method using simulation data. We simulated $20,000$ genes for the four-species scenario in Figure 1, and the population history parameters were set to approximate the evolutionary history of humans, chimpanzees, gorillas and orangutans inferred in previous studies (Prado-Martinez et al., 2013), and the sample includes $n_4 = 10$, $n_3 = 24$, $n_2 = 20$, and $n_1 = 20$ haplotypes from the four species, respectively. We set the effective haploid population sizes at $N_2 = N_1$, $N_3 = 1.2N_1$, and $N_4 = 0.8N_1$. The divergence times were $T_{34} = 4$, $T_{234} = 6$ and $T_{1234} = 12$ in units of $2N_1$. The scaled mutation rate for synonymous sites for each gene locus, $\theta_s = 2N_1\mu_s$, and the scaled mutation rate for replacement sites, $\theta_r = 2N_1\mu_r$, were chosen from several fixed values 1, 2, ..., 5. Among the $20,000$ genes, 1400 were under selection in species 4 or 5 (the common ancestor of humans and chimpanzees), and the other $18,600$ genes were neutral. The selection intensities $\gamma_i, i = 4, 5$ of every 100 genes were chosen from fixed values in the range of $(-6, -4, -2, 0, 2, 4, 6)$. Given

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the values of these parameters, simulated data were generated from Poisson distributions, and the means were calculated based on the formulae in Table 2. We then applied the method to the simulated data, and the maximum a posteriori (MAP) estimates of the parameters were recorded. The above simulations were repeated for 100 times. The boxplots of the inferred selection intensity, scaled mutation rates and divergence times are shown in Figure 2. For the global parameters, such as the divergence times T_{1234}, T_{234} and T_{34}, the inferred values are accurate and unbiased since sufficient information for these parameters is derived from the Bayesian joint analysis of all $20,000$ gene loci. The other locus-specific parameters, including the selection intensity and mutation rates of different branches, are generally also unbiased. However, we note that for selection intensity, when under strong negative selection, the inferred values become biased toward smaller values, because few or zero divergence and polymorphism sites are observed when the gene is under strong negative selection, which provides limited information for parameter inference (Bustamante et al., 2005). Overall, we evaluated the performance of HDMKPRF with simulated data, and the boxplots shown in Figure 2 demonstrate that the inferred parameter values well match the true values.

Selection in Hominidae

We applied the HDMKPRF method to multiple genomic sequences of humans, chimpanzees, gorillas and orangutans from Prado-Martinez et al. (2013). The data were generated via NGS technology with an average sequencing depth of 25, and details of the SNP calling pipeline and filtering criteria can be found in the original paper (Prado-Martinez et al., 2013). After excluding several individuals based on further criteria described in Cagan et al. (2016), the final dataset in our analysis includes *Pongo pygmaeus* (5), *Gorrilla gorilla* (12), *Pan troglodytes ellioti* (9) and *Homo sapiens* (9). We aligned the sequences of $23,362$ genes for these samples, from which $5,429$ genes with no protein coding information were excluded from the analysis. For the remaining $17,933$ genes ($17,234$ from autosomes and 699 from X chromosome), we used ANNOVAR (Wang et al., 2010) to annotate the SNPs in the coding regions. Synonymous and replacement polymorphism and divergence sites were identified to construct the four-species MK tables for each gene (see entries of Table 2). In total, 133 genes with no divergence and segregating sites in all lineages were excluded from the

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

four-species MK tables.

We applied our method to the HDMK tables of the 17,800 genes. After 200,000 burn-in steps, we then ran the Markov chain Monte Carlo process for 200,000 steps to achieve the posterior distributions of the parameters (see Appendix 2 for details). The maximum a posterior (MAP) estimate of the divergence time between orangurans and the common ancestor of humans, chimpanzees and gorillas (T_{1234}) is 11.6444 (posterior interval: (11.4125, 11.9126)), between gorillas and the common ancestor of humans and chimpanzees (T_{234}) is 4.9340 (4.6461, 5.2420), between humans and chimpanzees (T_{34}) is 3.7038 (3.6529, 3.7784), with all values in units of $2N_1$. The MAP estimates of the effective population sizes of gorillas, chimpanzees and humans are: $N_2 = 0.9665$ $(0.9458, 0.9878)$, $N_3 = 1.1610$ $(1.1351, 1.1877)$, and $N_4 = 0.8306$ $(0.8118, 0.8501)$, with all values in units of N_1. Our estimates of the demographic parameters are consistent overall with previous studies (Prado-Martinez et al., 2013).

By using orangutans and gorillas as outgroups, we identified 27,144 fixed synonymous sites and 19,123 fixed non-synonymous sites in the human lineage ($D_{4\sim(1,2,3)}$ in Table 2). The average genomic synonymous and non-synonymous divergence are 4.5197×10^{-4} and 3.1842×10^{-4} (per nucleotide site). We also identified 22,926 synonymous and 21,243 non-synonymous segregating sites in humans ($P_{4\sim(1,2,3)}$ in Table 2). The average synonymous and non-synonymous densities (per nucleotide site) are 3.8174×10^{-4} and 3.5372×10^{-4}. The ratio of non-synonymous to synonymous divergence sites is smaller than the ratio of non-synonymous to synonymous polymorphisms sites, which is consistent with the fact that the majority of amino acid variations in the genome are deleterious.

Table 3: Top 30 positively selected genes in the three lineages

Gene name	Function	Dn	Ds	Pn	Ps	Selection coefficient	$p(\gamma > 0)$
Human lineage							
SLC5A1	Primary mediator of dietary glucose and galactose	7	1	0	0	11.15 (4.07, 20.88)	1
FAM166B	Unknown	4	2	0	1	10.77 (3.18, 21.48)	0.999925
PDXDC1	Pyridoxal phosphate binding and carboxy-lyase activity	5	2	0	2	10.71 (3.26, 20.84)	0.999975

continued on next page

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

continued from previous page

Gene name	Function	Dn	Ds	Pn	Ps	Selection coefficient	$p(\gamma > 0)$
LPIN2	Controlling the metabolism of fatty acids	3	9	0	0	10.56 (2.75, 21.68)	0.999775
PKN2	Regulation of transcription activation signaling processes	3	1	0	3	10.49 (2.66, 21.32)	0.99985
C20orf30	Involved in trafficking and recycling of synaptic vesicles	3	1	0	1	10.44 (2.66, 21.37)	0.99965
LOC646498	Unknown	4	1	0	0	10.40 (3.09, 20.62)	1
ZNF605	Transcriptional regulation	8	8	0	1	10.32 (2.86, 21.19)	0.999975
SLC26A3	Transporting chloride ions across the cell membrane	9	7	1	0	10.28 (4.05, 19.16)	1
CASP10	Activation cascade of caspases responsible for apoptosis	7	1	0	2	10.26 (3.49, 19.77)	1
PRKD1	A serine/threonine protein kinase	3	2	0	1	10.13 (2.55, 20.78)	0.999775
ATXN3L	Unknown (Paralog of ATXN3)	5	3	0	1	10.10 (3.06, 20.03)	0.999975
NCR1	Cytotoxicity-activating receptor that may contribute to the increased efficiency of activated natural killer (NK) cells	4	0	0	1	10.07 (2.88, 20.18)	0.999825
CDH24	Mediating strong cell-cell adhesion	5	4	0	1	9.94 (2.84, 20.20)	0.999975
NHEJ1	DNA repair factor	4	0	1	0	9.86 (2.77, 20.00)	0.99985
LAMP2	Chaperone-mediated autophagy	3	1	0	1	9.79 (2.38, 20.27)	0.9995
TBCCD1	Regulation of centrosome and Golgi apparatus positioning	3	1	0	0	9.79 (2.49, 20.23)	0.999675
DAGLB	Required for axonal growth during development and for retrograde synaptic signaling at mature synapses	3	2	0	3	9.75 (2.25, 20.54)	0.9995
TKTL1	Oxidoreductase activity and transketolase activity	7	2	0	0	9.72 (3.35, 19.04)	1
ZDHHC4	Cytochrome-c oxidase activity	5	3	1	0	9.71 (2.67, 19.75)	0.999875
SLC25A25	Calcium-dependent mitochondrial solute carrier	4	1	0	2	9.70 (2.03, 20.76)	0.999125
SLC6A12	Transporting betaine and GABA	6	2	0	2	9.70 (2.36, 20.39)	0.9999
SYT10	Protein-protein interactions at synapses	2	4	0	3	9.65 (1.70, 20.86)	0.997175
BBC3	Essential mediator of p53/TP53-dependent and p53/TP53-independent apoptosis	2	0	0	0	9.59 (1.57, 20.81)	0.99725
FOXP2	Involved in neural mechanisms mediating the development of speech and language	2	1	0	1	9.58 (1.53, 20.80)	0.996625
P4HTM	Catalyzing the post-translational formation of 4-hydroxyproline	3	1	0	1	9.57 (2.21, 20.10)	0.999325
RALY	RNA-binding protein that acts as a transcriptional cofactor for cholesterol biosynthetic genes in the liver	2	1	0	0	9.56 (1.77, 20.47)	0.997825
EPM2AIP1	Interacting with EPM2A, related to adolescent progressive myoclonus epilepsy	3	1	0	1	9.54 (1.91, 20.61)	0.99885
MRPL14	Encoding a protein of the mitochondrial ribosome	2	0	0	2	9.53 (1.40, 20.99)	0.996475
UBE2CBP	Ubiquitin-conjugating enzyme E2C-binding protein	6	1	1	0	9.52 (2.63, 19.52)	0.99995

continued on next page

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

continued from previous page

Gene name	Function	Dn	Ds	Pn	Ps	Selection coefficient	$p(\gamma > 0)$
Chimpanzee lineage							
SCML1	Maintaining the transcriptionally repressive state of homeotic genes throughout development	18	1	0	2	14.99 (6.81, 25.33)	1
ZNF280C	Transcription factor	16	3	0	0	14.85 (6.80, 24.90)	1
TKTL1	Oxidoreductase activity and transketolase activity	10	0	0	3	12.55 (4.99, 22.48)	1
AIMP2	Required for assembly and stability of the aminoacyl-tRNA synthase complex	6	3	0	1	11.93 (3.78, 22.91)	1
PNMA5	Paraneoplastic Ma antigen protein	8	3	0	1	11.69 (4.57, 21.28)	1
CCDC154	Unknown	9	12	0	3	11.56 (3.47, 22.60)	1
TXLNG	Involving in intracellular vesicle traffic	8	1	0	1	11.55 (3.31, 22.91)	1
OTUD6A	Deubiquitinating enzyme	7	0	0	2	11.47 (4.04, 21.64)	1
GSDMD	Regulation of epithelial proliferation	5	5	0	1	11.33 (3.04, 22.30)	0.9999
TIPARP	T-cell function	4	0	0	0	11.20 (3.25, 21.88)	0.99995
THEM4	Acyl-CoA thioesterase	4	1	0	0	11.15 (3.41, 21.88)	0.999925
DOT1L	Histone methyltransferase	4	8	0	6	11.11 (2.68, 22.61)	0.9998
ELF4	Transcription factor	4	5	0	1	11.03 (2.86, 22.05)	0.99975
VRK3	Serine/threonine protein kinases	7	2	0	1	11.01 (3.68, 21.18)	1
DDX53	ATP-dependent RNA unwinding	16	1	0	1	10.97 (4.02, 20.46)	1
[1] APOL2	Apolipoprotein L gene family that may affect the movement of lipids or allow the binding of lipids to organelles	8	3	0	1	10.95 (3.15, 22.09)	0.999975
HIST1H4G	Histone H4;Core component of nucleosome	4	0	0	0	10.92 (3.45, 21.23)	0.999975
TAC4	Regulating peripheral endocrine and paracrine functions	4	1	0	0	10.90 (3.39, 21.41)	1
DCAF12L1	Involving in cell cycle progression, signal transduction, apoptosis, and gene regulation	8	0	1	0	10.84 (3.19, 21.63)	0.99995
YY2	Transcription factor	6	2	0	1	10.77 (3.81, 20.39)	0.999975
SPC25	Involving in kinetochore-microtubule interaction and spindle checkpoint activity	3	0	0	0	10.70 (2.53, 22.04)	0.99955
USP18	Ubiquitin-specific proteases	3	0	0	0	10.67 (2.74, 21.85)	0.999625
PTGDR	Guanine nucleotide-binding protein (G protein)-coupled receptor (GPCR)	4	1	0	2	10.66 (2.76, 21.72)	0.9997
GYPC	Regulating the mechanical stability of red cells	4	2	0	0	10.61 (2.66, 21.70)	0.999775
RSPH1	Male meiotic metaphase chromosome-associated acidic protein	4	2	1	1	10.55 (2.80, 21.50)	0.999875
VGLL1	Specific coactivator for the mammalian TEFs	3	2	0	0	10.51 (2.63, 21.44)	0.99965
PRDM9	Zinc finger protein with histone methyltransferase activity	4	3	0	0	10.51 (2.57, 21.46)	0.999725
LCN12	Binding all-trans retinoic acid and may act as a retinoid carrier protein within the epididymis	4	3	0	0	10.50 (2.57, 21.85)	0.9996

continued on next page

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

continued from previous page

Gene name	Function	Dn	Ds	Pn	Ps	Selection coefficient	$p(\gamma > 0)$
CUTA	Part of a complex of membrane proteins attached to acetylcholinesterase (AChE)	3	0	0	0	10.49 (2.05, 22.26)	0.9988
ASPHD1	Dioxygenase	3	0	0	0	10.46 (2.21, 22.00)	0.99885
Human-Chimpanzee lineage							
TAS2R8	Taste receptors	14	9	0	0	18.14 (7.42, 30.03)	1
ULBP3	Binding and activating the KLRK1/NKG2D receptor, mediating natural killer cell cytotoxicity	6	1	0	0	17.77 (8.17, 29.24)	1
BEND2	Participating in protein and DNA interactions	10	0	0	0	16.71 (8.46, 26.66)	1
TFDP3	The DP family of transcription factors	11	4	0	0	16.37 (7.72, 26.82)	1
ASB9	The ankyrin repeat and suppressor of cytokine signaling (SOCS) box protein family	4	2	0	0	15.62 (6.48, 26.95)	1
DDX25	A gonadotropin-regulated and developmentally expressed testicular RNA helicase	4	1	0	0	15.57 (6.44, 27.04)	1
CTSF	Cysteine protease	5	0	0	0	15.54 (6.49, 26.58)	1
YY2	Transcription factor	5	1	0	0	15.40 (6.57, 26.03)	1
PASD1	Transcription factor	8	1	0	0	15.00 (7.12, 24.75)	1
C16orf78	Unknown	7	0	0	0	14.85 (4.66, 26.82)	1
FAM122C	Unknown	5	1	0	0	14.37 (5.45, 25.31)	1
ZNF182	Involving in cell proliferation, differentiation, and apoptosis	3	1	0	0	14.30 (5.10, 25.86)	1
OR4D11	Olfactory receptor	8	1	0	0	14.22 (5.51, 24.97)	1
OR9Q1	Olfactory receptor	5	2	0	0	14.21 (5.96, 24.57)	1
COX7A2	Cytochrome c oxidase	3	0	0	0	14.00 (5.00, 25.14)	1
C12orf54	Unknown	3	0	0	0	13.71 (4.78, 25.15)	1
OR4X2	Olfactory receptor	6	1	0	0	13.31 (4.64, 24.11)	1
OR2C1	Olfactory receptor	5	4	0	0	13.23 (4.52, 23.91)	1
PNMA5	Development of paraneoplastic disorders resulting from an immune response directed against them	4	1	0	0	13.20 (5.04, 23.55)	1
ESX1	Involving in placental development and spermatogenesis	4	0	0	0	13.05 (4.73, 23.55)	1
RBMX2	Unknown	3	2	0	0	13.03 (4.44, 24.18)	1
B3GALT4	Involved in GM1/GD1B/GA1 ganglioside biosynthesis	2	2	0	0	12.76 (3.69, 24.68)	0.99985
ZNF649	Transcriptional repressor	4	1	0	0	12.66 (4.57, 23.00)	1
DDX26B	Unknown	3	2	0	0	12.59 (3.98, 23.66)	1
FAM46D	Non-canonical poly(A) RNA polymerase	4	0	0	0	12.55 (4.59, 22.85)	1
UBA7	The E1 ubiquitin-activating enzyme family that is involved in the conjugation of the ubiquitin-like interferon-stimulated gene 15 protein	5	4	0	0	12.46 (4.66, 22.69)	1

continued on next page

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

continued from previous page

Gene name	Function	Dn	Ds	Pn	Ps	Selection coefficient	$p(\gamma > 0)$
HUWE1	E3 ubiquitin ligase	4	4	0	0	12.35 (3.39, 23.69)	0.99995
ZMYND10	Involving in assembly of the dynein motor	3	0	0	0	12.24 (3.91, 23.12)	0.999975
IL13RA2	Playing role in internalization of IL13	3	0	0	0	12.20 (4.04, 23.01)	0.99995
C10orf131	Unknown	3	0	0	0	12.08 (3.66, 23.18)	0.999975

Genes under selection in the human lineage

The histogram in Figure 3(A) shows the posterior distribution of the selection intensity γ_4 for genes under selection in the human lineage. Among these genes, 1211 genes with a 95% confidence interval above 0 were identified as targets under positive selection and 1349 genes with a 95% confidence interval (CI) below 0 are under negative selection. Genes under positive selection in the human-lineage should attract particular attention because they potentially confer to the emergence of human-specific phenotypes and functionality. In the following sections, we focused our analysis on genes showing evidence of positive selection.

We first investigated the biological functions of those genes under positive selection by performing gene enrichment analyses with the package DAVID (Huang et al., 2009). Gene sets were defined using the KOBAS pathway and disease categories along with Gene Ontology (GO) categories (see Table 4).

Several top pathways in the human lineage are related to the immune systems ($P < 2.16 \times 10^{-12}$, Table 4), which is similar to the findings of previous studies (Bustamante et al., 2005; Clark et al., 2003; Nielsen et al., 2005). Several genes, such as NCR1, are among the top genes with the highest selection intensity (Table 3): $\gamma_{NCR1} = 10.07\ (2.88, 20.18)$. Interestingly, these genes are under strong negative selection in the chimpanzee lineage and the human-chimpanzee common ancestor lineage (HC lineage), indicating that accelerated evolution of these genes is likely caused by human-specific resistance to pathogens. These genes belong to different parts of the immune system, i.e., CASP10, DEFB110, HSP90AA1, and INSR, and others are found in the innate immune system; NCR1, UBE2CBP, TRAIP, and SEC24C and other 51 genes are from the adaptive immune system; and another 14 genes, including SAMHD1, NUP107, and UBA7, are related to interferon signaling (see Table 4).

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Another extremely significantly enriched pathway is gene expression ($P < 2.79 \times 10^{-7}$). This gene set includes 84 genes that fall into four main categories:

(1) RNA polymerase II transcription. This category includes many zinc-finger genes, including ZNF549, ZNF510, ZNF500, ZNF684, and ZNF750, along with others. (2) Transcriptional regulation by small RNAs, including NUP88, NUP133 and NUP107. (3) Regulation of TP53 activity, including AURKB, DNA2, TPX2, RMI1, RBBP8 (through phosphorylation), KAT6A, and EHMT1 (through acetylation or methylation). (4) Regulation of histone methylation, including PHF1, BCOR, CTCFL, MLL2, and BRCA1. Among these genes, JARID2, MLL, and BRCA1 are only under strong lineage-specific positive selection in humans.

The large number of genes from gene expression regulation pathways under positive selection indicates that evolutionary changes at the level of gene regulation played an essential role in the divergence of humans and chimpanzees, which is consistent with the hypothesis of King and Wilson (1975). The evolutionary roles of these genes in forming human-specific phenotypes have yet to be properly recognized, and our results could provide inspiration for further investigations.

Our study also identified genes involved in spermatogenesis under positive selection (Nielsen et al., 2005). High selection intensities were inferred according to the MK table patterns of these genes, including SLC26A3, SEPT4, PRM1, NPHP1 and OVOL1. Among these genes, PRM1 is known for its potential influence on sperm morphology and ability to fertilize eggs (Rooney and Zhang, 1999; Wyckoff et al., 2000; Nielsen et al., 2005).

Metabolism is another large gene category under selection ($P < 7.83 \times 10^{-5}$) and includes the metabolism of proteins, lipids and lipoproteins, purines, and carbohydrates. Certain genes, such as SLC5A1, show an extremely strong selection signal ($\gamma = 11.15$ $(4.07, 20.88)$) in humans, with 7 nonsynomymous divergence sites and 1 synonymous divergence site combined with zero within-species polymorphism sites. This gene is a sodium/glucose cotransporter that mediates the transport of glucose and related substances across cellular membranes, and it is involved in the absorption of glucose. Interestingly, Pontremoli et al. (2015) identified positive selection on the regulatory elements of SLC2A5 and SLCA2, which are fructose or glucose transporters and functionally similar to SLC5A1. However, their coding regions are

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

under neutrality or negative selection in our analysis, which implies that the evolution of carbohydrate consumption pathways may involve both coding regions and regulatory regions. Strong selection on these genes may reflect the importance of starch metabolism and dietary transitions during human evolution. In addition, other genes, such as PKN2, are involved in pathways regulating glucose and lipid metabolism in different tissues, such as skeletal muscle (Ruby et al., 2017).

We also analyze the association between positively selected genes and common diseases (Table 4). Immune system diseases and metabolic diseases, such as diabetes and obesity, are significantly enriched for genes under positive selection in the human lineage. This finding is consistent with the pathway enrichment results. Interestingly, our findings indicate that positive selection of certain genes is correlated with the development of schizophrenia (corrected $P = 0.00011$). The adaptive evolution of human cognitive abilities was hypothesized to be achieved by pushing the limits of metabolic capabilities, which created side effect such as schizophrenia (Khaitovich et al., 2008).

We then investigated the posterior mean and CIs of genes in these significantly enriched gene categories in further detail. In Figure 4, we present the results for four pathways showing distinct patterns. The genes in each pathway were ranked according to their posterior mean of the selection intensity, with bars representing the 95% CIs. Red and blue colors denote CIs completely above or below the line $\gamma = 0$, respectively, which correspond to the identification of being positively or negatively selected. Pathways for gene expression and schizophrenia are over-dominantly enriched for positively selected genes. These pathways are putatively under accelerated evolution and thus have played essential roles in the emergence of new functionality. In contrast, certain pathways, such as olfactory signaling, are enriched for negative selection, indicating their conservative evolution in the human lineage. Other pathways, such as metabolism, are significantly enriched for both positive selection and negative selection, thereby showing manifold functionality in evolution.

In addition to performing a categorical analysis of pathways and genes, we also investigated the top 30 genes under positive selection in humans, chimpanzees, and the common ancestor of humans and chimpanzees (see Table 3). The genes were ranked according to their MAP selection intensity. All positively and negatively selected genes are also shown in

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figures 5 and 6 as a selection map of the human and chimpanzee genomes. The genes identified as targets of positive or negative selection are marked on each chromosome according to their physical positions and labelled with red and blue, respectively.

Among the genes from the top list, FOXP2 is well known for its putative functionality in the evolution of language (Enard et al., 2002). Our analysis demonstrates that FOXP2 has been under strong positive selection in humans since the divergence of humans and chimpanzees and presents a selection intensity of $\gamma = 9.58$ (95% CI. $(1.53, 20.80)$). FOXP2 is quite conservative in chimpanzees ($\gamma = -2.70$ $(-16.49, 11.43)$) and the common ancestor of humans and chimpanzees ($\gamma = -1.31$ $(-16.13, 13.41)$, which indicates that positive selection on FOXP2 contributes to the emergence of human-specific phenotypes and functionality.

Interestingly, a number of genes on the list are related to neurological systems. RBFOX2 is a conserved RNA binding protein serving as a key regulator of alternative splicing in the nervous system and may play an important role in neuromuscular functions (Raj and Blencowe, 2015). SYT10 is found in pathways of protein-protein interactions at synapses and transmission across chemical synapses. This gene was identified to be important for epileptogenesis in previous studies (Glavan et al., 2009; Woitecki et al., 2016). ATXN3L is a paralog of ATXN3 that has undergone limited study, although GO annotations related to the gene include ubiquitin protein ligase binding and obsolete ubiquitin thiolesterase activity. Mutations of this gene were identified to be correlated with neurodegenerative disease (do Carmo Costa and Paulson, 2012). SLC6A12 is a transporter of betaine and GABA, which may have a role in the regulation of GABAergic transmission in the brain through the reuptake of GABA into presynaptic terminals. EPM2AIP1 interacts with EPM2A, which produces a protein called laforin, and is related to epilepsy (Lafora disease).

Table 4: Kobas enrichment of positively selected genes in the Human lineage

Term	Input(Background)	P-Value	Corrected P-Value
Pathway			
Immune system	94(1583)	3.522e-14	2.158e-12
Generic transcription pathway	56(822)	5.851e-11	2.991e-09
Adaptive immune system	51(787)	2.088e-09	8.629e-08
Gene expression	84(1719)	7.44e-09	2.79e-07
Signal transduction	103(2448)	2.028e-07	6.026e-06

continued on next page

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

continued from previous page

Term	Input(Background)	P-Value	Corrected P-Value
Complement and coagulation cascades	12(79)	1.938e-06	4.836e-05
Metabolic pathways	59(1243)	3.231e-06	7.828e-05
Regulation of TP53 activity	16(152)	3.4e-06	8.179e-05
Innate immune system	41(769)	8.341e-06	0.000185
Chromosome maintenance	13(110)	8.957e-06	0.0001956
Lysine degradation	9(52)	1.47e-05	0.0003077
Purine metabolism	16(176)	1.856e-05	0.000381
Complement cascade	8(44)	3.211e-05	0.0006189
Regulation of TP53 activity through phosphorylation	11(91)	3.586e-05	0.0006804
Metabolism of proteins	59(1378)	5.648e-05	0.00101
Metabolism	77(1975)	8.72e-05	0.001458
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	12(120)	8.737e-05	0.001459
Hemostasis	32(605)	9.097e-05	0.001518
Cell cycle	32(607)	9.635e-05	0.001595
Transcriptional regulation by TP53	22(355)	0.0001427	0.002278
Extension of telomeres	6(30)	0.0002005	0.003111
Activation of the pre-replicative complex	6(32)	0.0002728	0.004053
GPCR downstream signaling	42(940)	0.0002819	0.004168
Disease			
Immune system diseases	24(243)	4.066e-08	1.353e-06
Primary immunodeficiency	17(138)	2.302e-07	6.725e-06
Other congenital disorders	31(462)	1.518e-06	3.905e-05
Schizophrenia	29(440)	4.525e-06	0.0001067
Obesity-related traits	39(696)	4.68e-06	0.0001095
Diabetes	8(63)	0.0003039	0.004448
GO: biological process			
Spermatogenesis	42(449)	0.0006161	-
Male gamete generation	42(450)	0.000643	-
Lymphocyte mediated immunity	23(202)	0.001257	-
Regulation of viral life cycle	20(169)	0.001817	-
Protein activation cascade	12(74)	0.001868	-
Lymphocyte activation involved in immune response	17(133)	0.002029	-
Negative regulation of multi-organism process	18(147)	0.002289	-
Modification of morphology or physiology of other organism involved in symbiotic interaction	13(88)	0.002511	-
Sexual reproduction	56(697)	0.00262	-
B cell mediated immunity	14(101)	0.002853	
Cellular aromatic compound metabolic process	334(5508)	0.002969	-

continued on next page

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

continued from previous page

Term	Input(Background)	P-Value	Corrected P-Value
Complement activation	9(47)	0.003233	-
Multi-organism reproductive process	65(846)	0.003266	-
Adaptive immune response based on somatic recombination of immune receptors	22(207)	0.003785	-
Negative regulation of viral life cycle	12(81)	0.00386	-
Nucleic acid metabolic process	291(4752)	0.003885	-
Nucleobase-containing compound metabolic process	324(5348)	0.003908	-
Regulation of transcription, DNA-templated	210(3311)	0.004105	-
Heterocycle metabolic process	330(5465)	0.004212	-
Cellular nitrogen compound metabolic process	359(6000)	0.004583	-
Complement activation, classical pathway	7(30)	0.004587	-
Organic cyclic compound metabolic process	341(5675)	0.004848	-

Terms with Benjamini corrected P-value<0.005 are shown in table

Genes under selection in the chimpanzee lineage

Among the top 30 signals under strong positive selection in chimpanzees (Table 3), SCML1 is the one with the highest inferred selection intensity at 14.99 (6.81, 25.33). SCML1 was identified as a target of repeated positive selection by Nielsen et al. (2005) because it has 15 nonsynonymous and one synonymous substitutions between humans and chimpanzees, although zero polymorphisms are observed in humans. SCML1 is an expression repressor of the Hox genes and important for the developmental differences between humans and chimpanzees. Wu and Su (2008) also identified strong positive selection on SCML1 in multiple primate species and noticed that the gene is expressed in testes, implying its role in testis development and spermatogenesis.

PNMA5 shows a strong signal for positive selection in chimpanzees ($\gamma = 11.69$ (4.57, 21.28)). The exact molecular function of PNMA5 is still unclear, although previous research indicated that PNMA5 is highly expressed in the neocortex of the brain and may be involved in primate brain evolution. Interestingly, we found that the gene is also under strong positive selection in the human-chimpanzee common ancestor lineage but is under neutral evolution in humans, implying that it once played an important role in the brain evolution of ancient primates but that its effect likely terminated in the human lineage. AIMP2 is another gene

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

known to be related to human neurodegenerative diseases, such as Parkinson's disease.

Genes under selection in the common ancestor of humans and chimpanzees

TAS2R8 shows the strongest positive selection signal in the common ancestor of humans and chimpanzees, although it is under strong negative selection in human and under neutrality in chimpanzee ($\gamma_{HC} = 18.142$ $(7.4216, 30.0282)$; $\gamma_{human} = -4.32549$ $(-15.0859, 2.04698)$; and $\gamma_{chimp} = 2.13833$ $(-0.79666, 7.64481)$). TAS2R8 is one of the bitter taste receptor genes and may reflect a change in diet or toxin avoidance during different stages of primate evolution.

Several genes from the top list are olfactory receptors, such as OR4D11, OR9Q1, OR4X2, OR2C1, and OR4F6. Olfactory receptor genes are a large group of genes that show a strong tendency for positive selection (Clark et al., 2003), and our results demonstrate that these olfactory receptor genes have been important for sensory perception ever since the the time of the common ancestor of humans and chimpanzees. Among these genes, some are still actively evolving and under positive selection in humans, such as QR9R1, whereas the rest are conservative and under neutrality or negative selection in humans and/or chimpanzees.

A number of positively selected genes in the common ancestor lineage of humans and chimpanzees are related to transcription regulation, including TFDP3, YY2, PASD1, ZNF182, ESX1, ZNF649, and ZMYND10. In particular, some of these genes are involved in spermatogenesis. ZMYND15, for example, encodes a histone deacetylase-dependent transcriptional repressor that is important for spermatogenesis and male infertility (Yan et al., 2010). ZMYND10 is a zinc finger gene functioning in assembly of the dynein motor, and it is highly expressed in sperm and important for sperm movement.

ASB9, PNMA5, ICAM1 and IL13RA2 are found in the immune system. ICAM1 is a receptor that mediates the binding of *Plasmodium falciparum* to erythrocytes. Mutations on ICAM1 were under very recent positive selection in human populations (Kun et al., 1999). We found that the gene is under strong selection in the common ancestor of human and chimpanzee ($\gamma = 9.81(3.22, 19.27)$), which is consistent with the long history of malaria as a pathogen of primates and may reflect the gene's role in resistance to malaria since the early stages of primate evolution. We also noticed that ICAM1 is under selection in

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

both chimpanzees ($\gamma = 8.71(3.15, 16.95)$ and humans ($\gamma = 4.95(1.19, 11.25)$). Humans and chimpanzees are affected by different malaria species (*P. falciparum* and *P. reichenowi*, Martin et al. (2005)). ICAM1 is likely undergoing parallel evolution for resistance to the two parasites. Further investigation of the divergence sites among the three species may provide insights into the mechanism of malaria resistance.

Overall, in the above study, we discussed the genes and pathways under strong positive selection in humans, chimpanzees and their common ancestral lineage. Nielsen et al. (2005) used a likelihood-ratio test to compare the dN/dS ratios for coding regions of one human and one chimpanzee genome and provided a list of the 50 top genes under positive selection, and we found that 12 of their genes are also identified as significant in our analysis. However, only 7 out of the 12 genes are under positive selection in the human lineage; three are under positive selection in chimpanzees (CD72, SLC22A4 and DFFA), and the other two genes (RBM23 and C16orf3) are under selection only in the common ancestor of human and chimpanzee. Some other genes, such as SPATA3, are not significantly under positive selection. The MK table entry for SPATA3 in the human lineage is $(1, 1, 3, 1)$, indicating that positive selection on SPATA3 is not likely. The difference between our results and those of Nielsen et al. (2005) demonstrates that by extending the MK-like or dN/dS tests to multiple species, our method can gain information for precisely distinguishing selection that occurred in the human lineage, in the chimpanzee lineage, or in the common ancestor of the two species, thereby providing additional insights into understanding the selection process.

Discussion

We present the computational method HDMKPRF for detecting lineage-specific selection by jointly analyzing polymorphism and divergence data from multiple species. HDMKPRF is an extension of the MKPRF method and can be used from two species to multiple species (Bustamante et al., 2001). Our method has several advantages over existing approaches. The pairwise comparison of two species in both the MK test and MKPRF is directionless, and it only identifies genes that are divergent between the two species or populations but provides no further conclusions as to the lineage under selection (Akey et al., 2002; Bustamante et al., 2005; Clark et al., 2003; Chen et al., 2010). However, the HDMKPRF method can

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

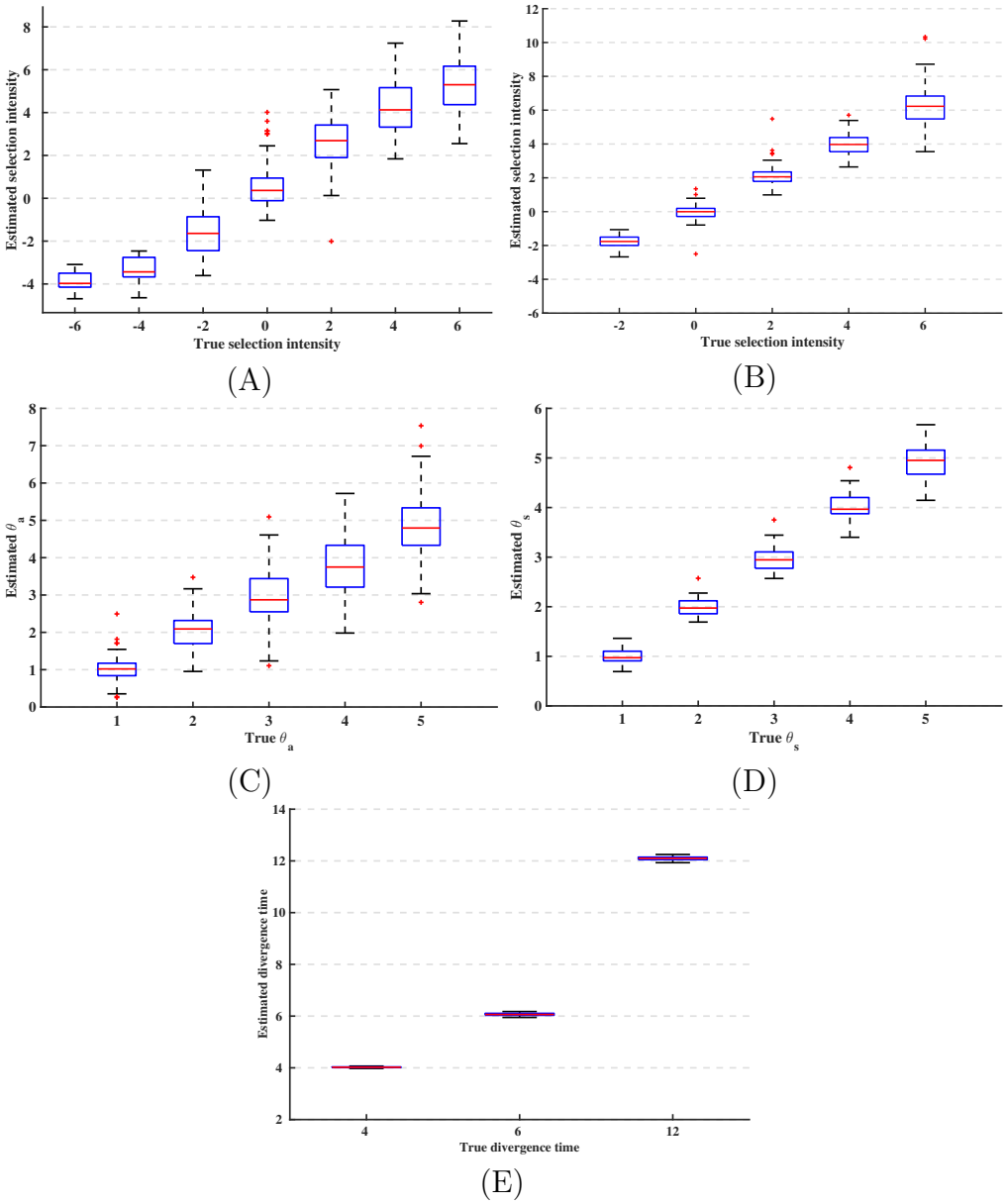


Figure 2: Comparison of the true and inferred selection coefficients for the four species in the simulation of $20,000$ gene loci: (A) selection coefficient γ_4; (B)selection coefficient γ_5; (C) scaled mutation rate; (D) scaled synonymous mutation rate; and (E) divergence times T_{34}, T_{234} and T_{1234}.

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

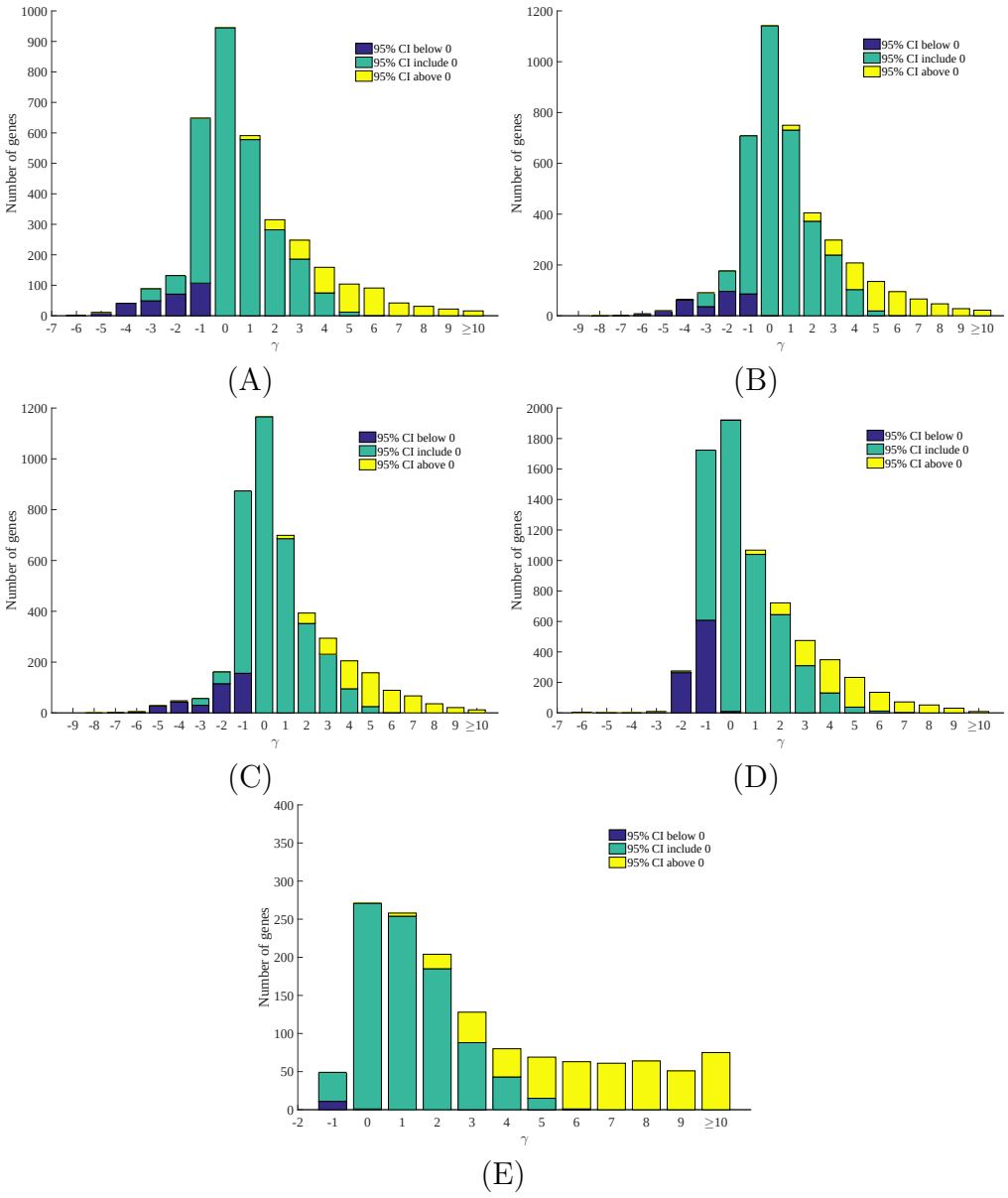
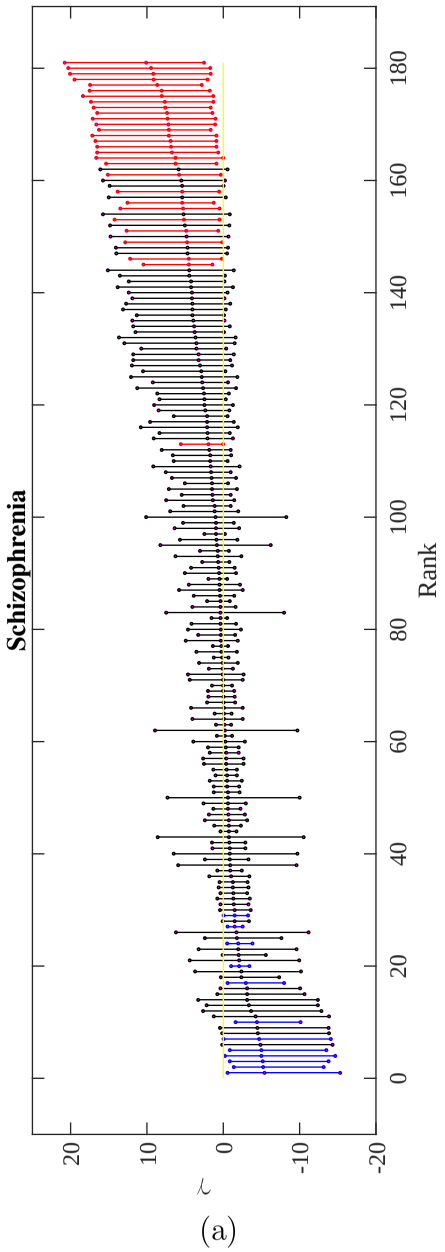
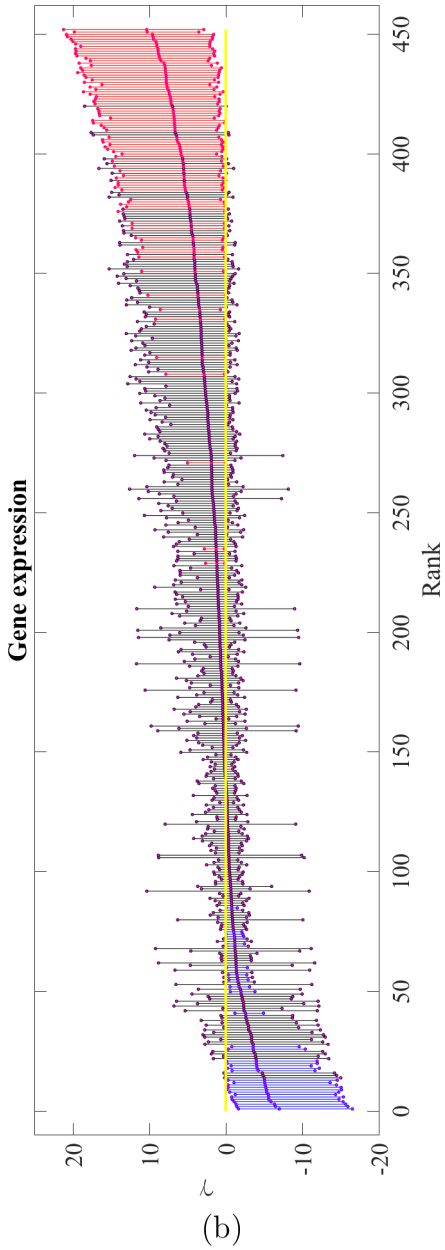


Figure 3: Posterior distribution of γ for humans and chimpanzees.

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



(a)



(b)

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

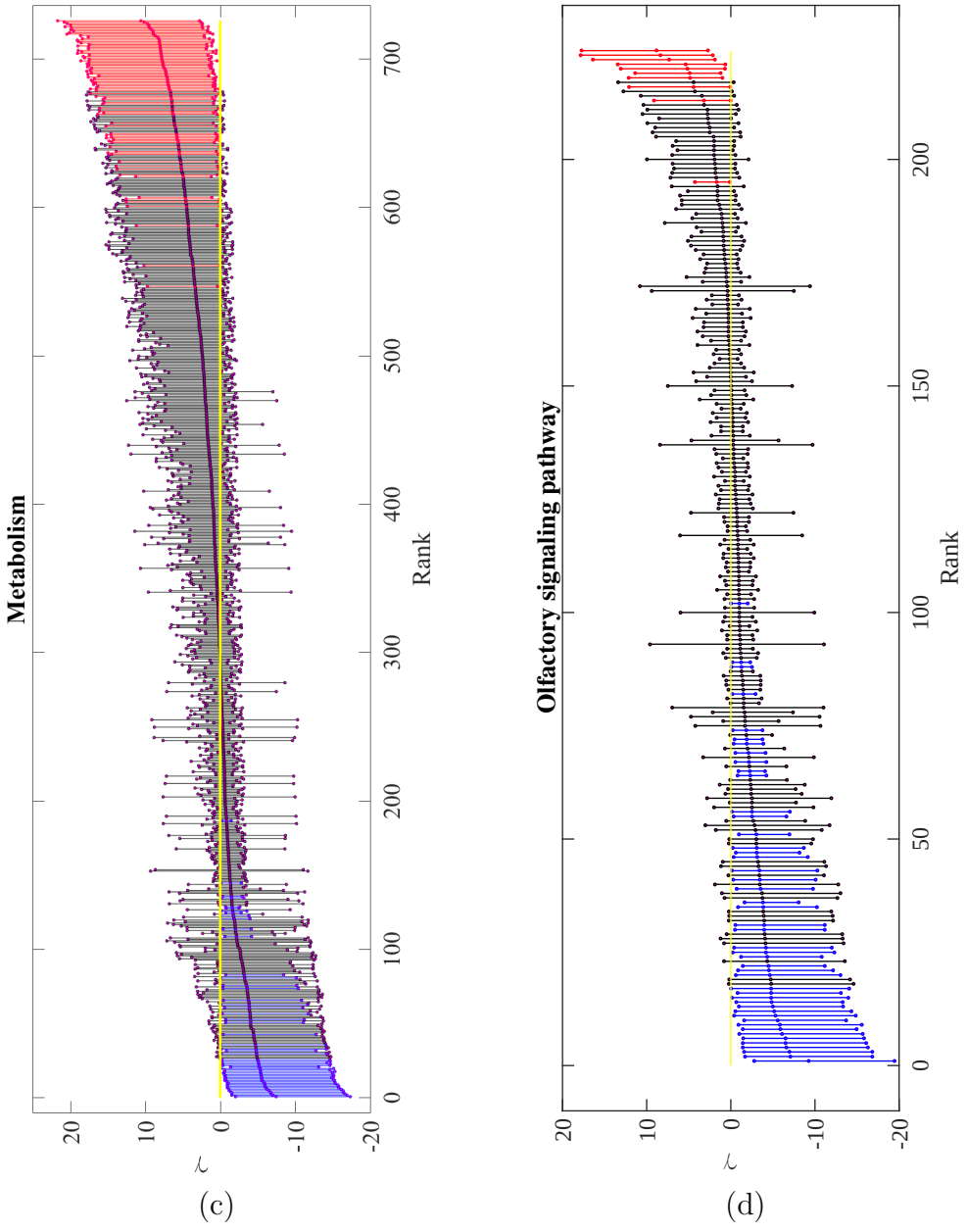


Figure 4: Distribution of γ for four pathways or disease-related gene sets.

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

pinpoint the occurrence of selection to a specific lineage, including the internal lineages. As we demonstrated in the analysis of the Hominidae data, HDMKPRF identifies specific targets of selection in human and chimpanzee lineages as well as the common ancestor of humans and chimpanzees.

Similar to the MK test and MKPRF, the HDMKPRF method analyzes both polymorphism and divergence sites, which increases the power for detecting selection and helps distinguish positive selection from the relaxation of negative selection (Wyckoff et al., 2000). By using the Bayesian approach to combine information from all gene loci, the method is more powerful than single-locus methods, such as the MK test and dN/dS ratio tests. We apply the HDMKPRF method to the genomic sequences of four species of Hominidae and identify gene loci under lineage-specific positive and negative selection. Cagan et al. (2016) recently analyzed the same data set using the MK test and only identified a limited number of genes under positive selection since the divergence of humans and chimapnzees. A comparison of our results with former studies demonstrates that the HDMKPRF method outperforms alternative methods with higher power and provides additional insights into the distribution of selection effects and the temporal and spatial occurrence of Darwinian selection over evolutionary history.

With the development of sequencing technologies, genomic population data for multiple species are abundant, thus necessitating methods that can efficiently analyze both within-species and between-species data. The HDMKPRF method presented in this paper satisfies such a need, and we expect that its application potential will be extensive in comparative genomic studies.

Appendices

A1. Segregating site pattern $P_{(2,3)\sim 1}$, $P_{(3,4)\sim(1,2)}$ and $D_{(3,4)\sim(1,2)}$

For the three-species scenario, we assume that the allele frequency x of the common ancestor of species 2 and 3 (species 4) at time T_{23} follows the stationary distribution under neutrality for synonymous sites ($f(x)$, Equation 1) and distribution under selection for replacement sites ($g(x)$, Equation 6). After the population split and with evolution over time, the joint allele frequency distribution of species 2 (y) and species 3 (z) of synonymous sites

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

is as follows:

$$\begin{aligned} & f(y,z|T_{23},N_2,N_3) \\ = & \int_0^1 \phi(y|x,T_{23},N_2)\phi(z|x,T_{23},N_3)f(x)dx \\ = & 2N_4\mu_s \int_0^1 \frac{1}{x}\phi(y|x,T_{23},N_2)\phi(z|x,T_{23},N_3)dx, \end{aligned} \tag{24}$$

where $\phi(y|x,T,N)$ represents the transient allele frequency distribution y given its initial frequency x at time T and population size N (Chen et al., 2007). Kimura (1955a) found that for neutral evolution

$$\begin{aligned} \phi(y|x,T,N) = & \sum_{i=1}^{\infty} \frac{(2i+1)(1-(1-2x)^2)}{i(i+1)} C_{i-1}^{3/2}(1-2x) \\ & \cdot C_{i-1}^{3/2}(1-2y)e^{-1/2i(i+1)T}, \end{aligned} \tag{25}$$

where $C_i^{3/2}(x)$ is the Gegenbauer polynomial with $\lambda = 3/2$.

Kimura (1955b) provides the transient distribution for alleles under selection, which is in complicated form and is difficult to calculate. Williamson et al. (2005) and Evans et al. (2007) adopted the Crank-Nicolson finite difference method to approximate the transient distribution as the solution of a forward diffusion equation, and several recent studies provided analytical solutions using perturbation methods (Schraiber, 2014; Živković et al., 2015). Similar to the neutral case, the joint allele frequency distribution of nonsynonymous sites under selection in species 2 and 3 is:

$$\begin{aligned} & g(y,z|T_{23},N_2,N_3,\gamma) \\ = & \int_0^1 \psi(y|x,T_{23},N_2,\gamma_2)\psi(z|x,T_{23},N_3,\gamma_3)g(x)dx \\ = & 2N_4\mu_r \frac{2\gamma_4}{1-e^{-2\gamma_4}} \int_0^1 \frac{\psi(y|x,T_{23},N_2,\gamma_2)\psi(z|x,T_{23},N_3,\gamma_3)}{x(1-x)\times 2\gamma_4}dx, \end{aligned} \tag{26}$$

where $\psi(y|x,T,N,\gamma)$ represents the transient allele frequency distribution for y with an initial allele frequency x, time T, population size N and selection intensity γ.

The expected number of polymorphism synonymous sites segregating in species 2 and 3

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

is

$$\begin{aligned} & \int_0^1 \int_0^1 (1-y^{n_2}-(1-y)^{n_2})(1-z^{n_3}-(1-z)^{n_3}) \\ & \times f(y,z|T_{23},N_2,N_3)dydz \\ = \; & \theta_{s,4} H(n_2,n_3). \end{aligned} \tag{27}$$

The expected number of polymorphism replacement sites segregating in species 2 and 3 is as follows:

$$\begin{aligned} & \int_0^1 \int_0^1 (1-y^{n_2}-(1-y)^{n_2})(1-z^{n_3}-(1-z)^{n_3}) \\ & \times g(y,z|T_{23},N_2,N_3,\gamma_4)dydz \\ = \; & \theta_{r,4} \frac{2\gamma_4}{1-e^{-2\gamma_4}} I(n_2,n_3,\gamma_2,\gamma_3,\gamma_4). \end{aligned} \tag{28}$$

The other entries of the three-species MK table are now different from that of Table 1. For example, the expected number of $P_{2\sim(1,3)}$ includes two components. The first part consists of sites fixed in sample 3 but still segregating in sample 2:

$$\begin{aligned} & \int_0^1 \int_0^1 (1-y^{n_2}-(1-y)^{n_2})((1-z)^{n_3}) \\ & \times f(y,z|T_{23},N_2,N_3)dydz. \end{aligned} \tag{29}$$

The second part consists of the new mutations occurring in species 2 since T_{23}:

$$\int_0^1 (1-y^{n_2}-(1-y)^{n_2}) \times \frac{\theta_{s,2}}{2} \int_0^{T_{23}} \phi(y|\frac{1}{N_2},t,N_2)dtdy, \tag{30}$$

where $\frac{\theta_{s,2}}{2}$ is the number of new neutral mutations that enter the population every generation with the initial frequency of $1/N_2$.

Note that the above formula for $P_{s,2\sim(1,3)}$ is different from that in Table 1 and is applicable in different situations. An empirical criterion for choosing the two fomulae is based on the distribution of TMRCA of n_2. According to Griffiths (1984), the TMRCA asymptotically follows a normal distribution, with the mean and variance determined by N_2 and the

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

population history (see Griffiths (1984) and Chen and Chen (2013) for detailed formulae); thus $Pr(\text{TMRCA} < T_{23})$. When T_{23} is sufficiently large and $Pr(\text{TMRCA} < T_{23}) > 0.95$, we adopt the formulae in Table 1; otherwise, we use Equation 29 and 30.

For the four-species scenario, we can obtain the expected values for $P_{s,(3,4)\sim(1,2)}$ and $P_{r,(3,4)\sim(1,2)}$ via a similar method used for $P_{s,(2,3)\sim1}$ and $P_{r,(2,3)\sim1}$. Synonymous divergence sites $D_{(3,4)\sim(1,2)}$ in the four-species scenario include two components: the sites fixed on the branch between species 5 and 6 and the sites fixed in the sample of n_3 and n_4. Following Equation 5, the first component is as follows:

$$\frac{1}{2}\theta_{s,5}(T_{234} - T_{34})v_5, \tag{31}$$

with $v_5 = N_1/N_5$. The second component includes those sites fixed in the sample of n_3 and n_4:

$$\begin{aligned} & \int_0^1 \int_0^1 y^{n_3} z^{n_4} \times f(y, z|T, N_3, N_4) dy dz \\ = \; & \theta_{s,5} J(n_3, n_4). \end{aligned} \tag{32}$$

Similarly for replacement sites, the two components are:

$$\frac{1}{2}\theta_5 v_5 (T_{234} - T_{34}) \frac{2\gamma_5}{1 - e^{-2\gamma_5}}, \tag{33}$$

and

$$\begin{aligned} & \int_0^1 \int_0^1 y^{n_3} z^{n_4} \times g(y, z|T, N_3, N_4, \gamma_5) dy dz \\ = \; & \theta_{r,5} \frac{2\gamma_5}{1 - e^{-2\gamma_5}} K(n_3, n_4, \gamma_3, \gamma_4, \gamma_5). \end{aligned} \tag{34}$$

Therefore, we have $P_{(3,4)\sim(1,2)}$ and $D_{(3,4)\sim(1,2)}$.

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A2. MCMC steps for parameter optimization for the 4-species McDonald-Kreitman tables

In the Bayesian Poisson random field model of four species, the parameters include

$$\Gamma = \{\theta^l_{s,i}, \theta^l_{r,i}, T_{1234}, T_{234}, T_{34}, \gamma_{i,l}, i \in \{1,2,3,4,5\}, 1 \leq l \leq L\}, \tag{35}$$

where i is the index of lineages (species, see Figure 1B), l is the index of genes. D and P are four-species data containing lineage-specific divergence and polymorphism sites ($1 \sim (2,3,4)$ etc.). Since the posterior distributions of Γ are analytically intractable, we apply the Markov chain Monte Carlo method (MCMC) to achieve them. The details of MCMC steps are as follows.

Initialization

The initial values of divergence time T_{1234}, T_{234} and T_{34} are chosen according to prior knowledge or with arbitrary positive values satisfying $T_{1234} > T_{234} > T_{34}$. The selection parameters $\gamma_i, i = 1,2,3,4,5$ are generated from the normal distribution $N(0,8)$. We denote $v_i, i = 2,3,4,5$ as the ratio of effective population sizes N_1/N_i. In the four-species McDonald-Kreitman table, the number of polymorphism synonymous mutations occurring in gene l in lineage i, with $i = 1,2,3,4$ (as is shown in Figure 1B), is $P^l_{s,i} = \theta^l_{s,i} L(n_i) = N_i \mu_{s,l} L(n_i)$. Therefore, for $i = 2,3,4$, v_i can be estimated as follows:

$$v_i \approx \frac{L(n_i) \sum_l P^l_{s,1}}{L(n_1) \sum_l P^l_{s,i}}. \tag{36}$$

MCMC iteration

Locus-specific mutation parameters for gene l in lineage $1 \sim (2,3,4)$ are updated via Gibbs sampling. For each gene l, given values of $\gamma_{i,l}, i = 1,2,3,4,5$, a value of $\theta^l_{r,1}$ ($\theta^l_{r,1} = 2N_1\mu_{r,l}$) is generated from a gamma distribution

$$\theta^l_{r,1} \sim \Gamma(par_{1,l}, par_{2,l}) \tag{37}$$

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

with

$$par_{1,l} = \alpha + \sum_{i=1}^{5} D_{r,i}^{l} + \sum_{i=1}^{5} P_{r,i}^{l} \tag{38}$$

and

$$\begin{aligned} par_{2,l} = \beta &+ \frac{2\gamma_{1,l}}{1-e^{-2\gamma_{1,l}}}(2T_{1234} - T_{234} + G(n_1,\gamma_{1,l}) + F(n_1,\gamma_{1,l})) \\ &+ \frac{2\gamma_{2,l}}{1-e^{-2\gamma_{2,l}}}(T_{234} + \frac{1}{v_2}(G(n_2,\gamma_{2,l}) + F(n_2,\gamma_{2,l}))) \\ &+ \frac{2\gamma_{3,l}}{1-e^{-2\gamma_{3,l}}}(T_{34} + \frac{1}{v_3}(G(n_3,\gamma_{3,l}) + F(n_3,\gamma_{3,l}))) \\ &+ \frac{2\gamma_{4,l}}{1-e^{-2\gamma_{4,l}}}(T_{34} + \frac{1}{v_4}(G(n_4,\gamma_{4,l}) + F(n_4,\gamma_{4,l}))) \\ + \frac{2\gamma_{5,l}}{1-e^{-2\gamma_{5,l}}}&((T_{234} - T_{34}) + \frac{1}{v_5}(K(n_3,n_4,\gamma_{3,l},\gamma_{4,l},\gamma_{5,l}) + I(n_3,n_4,\gamma_{3,l},\gamma_{4,l},\gamma_{5,l}))). \end{aligned} \tag{39}$$

Since $D_{(3,4)\sim(1,2)}$ and $P_{(3,4)\sim(1,2)}$ are usually of low information with high volatility in our example, we approximate the posterior distribution of $\theta_{r,1}^{l}$:

$$\theta_{r,1}^{l} \sim \Gamma(\widehat{par_{1,l}}, \widehat{par_{2,l}}) \tag{40}$$

with

$$\widehat{par_{1,l}} = \alpha + \sum_{i=1}^{4} D_{r,i}^{l} + \sum_{i=1}^{4} P_{r,i}^{l} \tag{41}$$

and

$$\begin{aligned} \widehat{par_{2,l}} = \beta &+ \frac{2\gamma_{1,l}}{1-e^{-2\gamma_{1,l}}}(2T_{1234} - T_{234} + G(n_1,\gamma_{1,l}) + F(n_1,\gamma_{1,l})) \\ &+ \frac{2\gamma_{2,l}}{1-e^{-2\gamma_{2,l}}}(T_{234} + \frac{1}{v_2}(G(n_2,\gamma_{2,l}) + F(n_2,\gamma_{2,l}))) \\ &+ \frac{2\gamma_{3,l}}{1-e^{-2\gamma_{3,l}}}(T_{34} + \frac{1}{v_3}(G(n_3,\gamma_{3,l}) + F(n_3,\gamma_{3,l}))) \\ &+ \frac{2\gamma_{4,l}}{1-e^{-2\gamma_{4,l}}}(T_{34} + \frac{1}{v_4}(G(n_4,\gamma_{4,l}) + F(n_4,\gamma_{4,l}))) \end{aligned} \tag{42}$$

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Similarly, a value for $\theta^l_{s,1}$ was generated from the gamma distribution

$$\theta^l_{s,1} \sim \Gamma(par_{3,l}, par_{4,l}) \tag{43}$$

with

$$par_{3,l} = \alpha + \sum_{i=1}^{5} D^l_{s,i} + \sum_{i=1}^{5} P^l_{s,i}, \tag{44}$$

and

$$\begin{aligned} par_{4,l} = \beta &+ (2T_{1234} - T_{234} + M(n_1) + L(n_1)) \\ &+ (T_{234} + \frac{1}{v_2}(M(n_2) + L(n_2))) \\ &+ (T_{34} + \frac{1}{v_3}(M(n_3) + L(n_3))) \\ &+ (T_{34} + \frac{1}{v_4}(M(n_4) + L(n_4))) \\ &+ (T_{234} - T_{34} + \frac{1}{v_5}(J(n_3, n_4) + H(n_3, n_4))). \end{aligned} \tag{45}$$

We also approximate the gamma distribution:

$$\theta^l_{s,1} \sim \Gamma(\widehat{par_{3,l}}, \widehat{par_{4,l}}), \tag{46}$$

where

$$\widehat{par_{3,l}} = \alpha + \sum_{i=1}^{4} D^l_{s,i} + \sum_{i=1}^{4} P^l_{s,i} \tag{47}$$

and

$$\begin{aligned} \widehat{par_{4,l}} = \beta &+ (2T_{1234} - T_{234} + M(n_1) + L(n_1)) \\ &+ (T_{234} + \frac{1}{v_2}(M(n_2) + L(n_2))) \\ &+ (T_{34} + \frac{1}{v_3}(M(n_3) + L(n_3))) \\ &+ (T_{34} + \frac{1}{v_4}(M(n_4) + L(n_4))) \end{aligned} \tag{48}$$

Note that in the above gamma distributions, α and β are uninformative small values close to 0.

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Once $\theta^l_{s,1}$ and $\theta^l_{r,1}$ are updated, $\theta^l_{s,i}, \theta^l_{r,i}, i = 2, 3, 4$ can be calculated by $\theta^l_{s,i} = N_i\mu_{s,l} = \theta^l_{s,1}/v_i$ and $\theta^l_{r,i} = \theta^l_{r,1}/v_i$.

Then, we update the selection parameters using Metropolis sampling as shown in Algorithem 1.

Algorithm 1: Updating selection parameters by Metropolis sampling

For selection coefficient $\gamma_{i,l}$ of gene l of lineage i
  Sampling $\gamma^{'}_{i,l}$, $\gamma^{'}_{i,l} \sim Uniform[\gamma_{i,l} - \varepsilon, \gamma_{i,l} + \varepsilon]$
    If $p(\gamma^{'}_{i,l}|D, P) > p(\gamma_{i,l}|D, P)$
      $\gamma_{i,l} = \gamma^{'}_{i,l}$
    Else
      Sampling $u, u \sim Uniform[0, 1]$
      If $u < \frac{p(\gamma^{'}_{i,l}|D,P)}{p(\gamma_{i,l}|D,P)}$
        $\gamma_{i,l} = \gamma^{'}_{i,l}$
      Else
        $\gamma_{i,l} = \gamma_{i,l}$
      End
    End
End

In Algorithem 1, $p(\gamma_{i,l}|D, P)$ is the posterior probability for $\gamma_{i,l}$.

$$p(\gamma_{i,l}|D, P) = p(\gamma_{i,l})L(\gamma_{i,l}|D, P) \tag{49}$$

with

$$p(\gamma_{i,l}) \sim Normal(\mu, \sigma) \tag{50}$$

and

$$L(\gamma|D, P) = p(D, P|\gamma_{i,l}). \tag{51}$$

Considering that selection coefficient $\gamma_{i,l}$ only affects the non-synonymous mutation numbers, we have

$$p(D, P|\gamma_{i,l}) = p(D^l_{r,i}, P^l_{r,i}|\gamma_{i,l}). \tag{52}$$

Since the expected K and I in the four-species McDonald-Kreitman table are difficult to calculate and only have a weak influence on the estimation of $\gamma_{5,l}$, we approximate the likelihood of $\gamma_{5,l}$ as follows:

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$p(D, P|\gamma_{5,l}) = p(P^l_{r,5}|\gamma_{5,l}), \tag{53}$$

$E(P^l_{r,5})$ is simplified as follows:

$$\begin{aligned} E(P^l_{r,5}) &\approx \theta^l_{r,5}\frac{2\gamma_{5,l}}{1-e^{-2\gamma_{5,l}}}(v_5(T_{234}-T_{34})) \\ &= \theta^l_{r,1}\frac{2\gamma_{5,l}}{1-e^{-2\gamma_{5,l}}}(T_{234}-T_{34}) \end{aligned} \tag{54}$$

Once all the selection and mutation parameters have been updated, the divergence times T_{1234}, T_{234} and T_{34} are updated by Metropolis sampling in a manner analogous to the updating of the γ values. The steps for updating divergence time are shown in Algorithm 2 below.

Algorithm 2: Updating divergence time by Metropolis sampling

For divergence times T in $\{T_{1234}, T_{234}, T_{34}\}$ in turn
  Sampling $T^{'}$, $T^{'} \sim Uniform[T-\varepsilon, T+\varepsilon]$
  If $p(T^{'}|D,P) > p(T|D,P)$
    $T = T^{'}$
  Else
    Sampling $u, u \sim Uniform[0,1]$
    If $u < \frac{p(T^{'}|D,P)}{p(T|D,P)}$
      $T = T^{'}$
    Else
      $T = T$
    End
  End
End

In Algorithm 2, divergence time T_{1234} affects divergence sites $D_{1\sim(2,3,4)}$; T_{234} affects $D_{1\sim(2,3,4)}$, $D_{2\sim(1,3,4)}$ and $D_{(3,4)\sim(1,2)}$; and T_{34} affects $D_{3\sim(1,2,4)}$, $D_{4\sim(1,2,3)}$ and $D_{(3,4)\sim(1,2)}$:

$$p(T_{1234}|D,P) = p(T_{1234}|D_{1\sim(2,3,4)}) \tag{55}$$

$$p(T_{234}|D,P) = p(T_{1234}|D_{1\sim(2,3,4)}, D_{2\sim(1,3,4)}, D_{(3,4)\sim(1,2)}) \tag{56}$$

$$p(T_{34}|D,P) = p(T_{1234}|D_{3\sim(1,2,4)}, D_{4\sim(1,2,3)}, D_{(3,4)\sim(1,2)}) \tag{57}$$

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Note that in Algorithm 2, T has to satisfy $T_{34} < T_{234} < T_{1234}$. In our algorithm, we only use the synonymous divergence mutant numbers in the MH sampling step to update T. The simulation results show that this processing method can improve the performance of divergence time estimation.

The total number of MCMC iterations was 400,000, and the first 200,000 iterations were treated as burn-in and disregarded; subsequently, we sampled data points every 10 iterations, which yielded a total of 40,000 sample points.

Acknowledgements

This project was supported by the "Strategic Priority Research Program" of the Chinese Academy of Sciences (Grant No. XDB13020400), the National Natural Science Foundation of China (Grant No. 91731302, 31571370 and 91631106), and the "Hundred Talents Program" of the Chinese Academy of Sciences.

References

Akey, J. M., Zhang, G., Zhang, K., Jin, L., and Shriver, M. D., 2002. Interrogating a high-density snp map for signatures of natural selection. *Genome Res.*, **12**:1805–1814.

Bustamante, C., Wakeley, J., Sawyer, S., and Hartl, D., 2001. Directional selection and the site-frequency spectrum. *Genetics*, **159**(4):1779–1788.

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

Cagan, A., Theunert, C., Laayouni, H., Santpere, G., Pybus, M., Casals, F., Prüfer, K., Navarro, A., Marques-Bonet, T., Bertranpetit, J., *et al.*, 2016. Natural selection in the great apes. *Mol Biol Evol*, **33**(12):3268–3283.

Chen, H. and Chen, K., 2013. Asymptotic distributions of coalescence times and ancestral lineage numbers for populations with temporally varying size. *Genetics*, **194**(3):721–736.

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Chen, H., Green, R. E., Pääbo, S., and Slatkin, M., 2007. The joint allele-frequency spectrum in closely related species. *Genetics*, **177**(1):387–398.

Chen, H., Patterson, N., and Reich, D., 2010. Population differentiation as a test for selective sweeps. *Genome Res.*, **20**(3):393–402.

Clark, A. G., Glanowski, S., Nielsen, R., Thomas, P. D., Kejariwal, A., Todd, M. A., Tanenbaum, D. M., Civello, D., Lu, F., Murphy, B., *et al.*, 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, **302**(5652):1960–1963.

do Carmo Costa, M. and Paulson, H. L., 2012. Toward understanding machado–joseph disease. *Prog Neurobiol*, **97**(2):239–257.

Enard, W., Przeworski, M., Fisher, S., Lai, C., Wiebe, V., Kitano, T., Monaco, A., and Pääbo, S., 2002. Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*, **418**(6900):869–872.

Evans, S., Shvets, Y., and Slatkin, M., 2007. Non-equilibrium theory of the allele frequency spectrum. *Theor. Popul. Biol.*, **71**(1):109–119.

Fay, J. C., Wyckoff, G. J., and Wu, C.-I., 2002. Testing the neutral theory of molecular evolution with genomic data from drosophila. *Nature*, **415**(6875):1024–1026.

Glavan, G., Schliebs, R., and Živin, M., 2009. Synaptotagmins in neurodegeneration. *The Anatomical Record*, **292**(12):1849–1862.

Griffiths, R. C., 1984. Asymptotic line of descent distributions. *J. Math. Biol.*, **21**(1):67–75.

Huang, D. W., Sherman, B. T., and Lempicki, R. A., 2009. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat Protoc*, **4**(1):44–57.

Hughes, A. L. and Nei, M., 1988. Pattern of nucleotide substitution at major histocompatibility complex class i loci reveals overdominant selection. *Nature*, **335**(6186):167–170.

Khaitovich, P., Lockstone, H. E., Wayland, M. T., Tsang, T. M., Jayatilaka, S. D., Guo, A. J., Zhou, J., Somel, M., Harris, L. W., Holmes, E., *et al.*, 2008. Metabolic changes in schizophrenia and human brain evolution. *Genome Biol*, **9**(8):R124.

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Kimura, M., 1955a. Solution of a process of random genetic drift with a continuous model. *Proc Natl Acad Sci U S A.*, **41**(3):144–50.

Kimura, M., 1955b. Stochastic processes and distribution of gene frequencies under natural selection. *Cold Spring Harb Symp Quant Biol*, (20):33–53.

King, M.-C. and Wilson, A. C., 1975. Evolution at two levels in humans and chimpanzees. *Science*, **188**(4184):107–116.

Kun, J., Klabunde, J., Lell, B., Luckner, D., Alpers, M., May, J., Meyer, C., and Kremsner, P. G., 1999. Association of the icam-1kilifi mutation with protection against severe malaria in lambaréné, gabon. *Am J Trop Med Hyg*, **61**(5):776–779.

Li, W.-H., Wu, C.-I., and Luo, C.-C., 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol*, **2**(2):150–174.

Martin, M. J., Rayner, J. C., Gagneux, P., Barnwell, J. W., and Varki, A., 2005. Evolution of human-chimpanzee differences in malaria susceptibility: relationship to human genetic loss of n-glycolylneuraminic acid. *Proc Natl Acad Sci U.S.A.*, **102**(36):12819–12824.

McDonald, J. H. and Kreitman, M., 1991. Adaptive protein evolution at the adh locus in drosophila. *Nature*, **351**(6328):652.

Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., Fledel-Alon, A., Tanenbaum, D. M., Civello, D., White, T. J., *et al.*, 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol*, **3**(6):e170.

Pontremoli, C., Mozzi, A., Forni, D., Cagliani, R., Pozzoli, U., Menozzi, G., Vertemara, J., Bresolin, N., Clerici, M., and Sironi, M., *et al.*, 2015. Natural selection at the brush-border: adaptations to carbohydrate diets in humans and other mammals. *Genome Biol Evol*, **7**(9):2569–2584.

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., OConnor, T. D., Santpere, G., *et al.*, 2013. Great ape genetic diversity and population history. *Nature*, **499**(7459):471.

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Raj, B. and Blencowe, B. J., 2015. Alternative splicing in the mammalian nervous system: recent insights into mechanisms and functional roles. *Neuron*, **87**(1):14–27.

Rogers, J. and Gibbs, R. A., 2014. Comparative primate genomics: emerging patterns of genome content and dynamics. *Nat Rev Genet*, **15**(5):347.

Rooney, A. P. and Zhang, J., 1999. Rapid evolution of a primate sperm protein: relaxation of functional constraint or positive darwinian selection? *Mol Biol Evol*, **16**(5):706–710.

Ruby, M. A., Riedl, I., Massart, J., Åhlin, M., and Zierath, J. R., 2017. Protein kinase n2 regulates amp kinase signaling and insulin responsiveness of glucose metabolism in skeletal muscle. *Am J Physiol Endocrinol Metab.*, **313**(4):E483–E491.

Sawyer, S. A. and Hartl, D. L., 1992. Population genetics of polymorphism and divergence. *Genetics*, **132**(4):1161–1176.

Schraiber, J. G., 2014. A path integral formulation of the wright–fisher process with genic selection. *Theor Popul Biol*, **92**:30–35.

Smith, N. G. and Eyre-Walker, A., 2002. Adaptive protein evolution in drosophila. *Nature*, **415**(6875):1022–1024.

Wang, K., Li, M., and Hakonarson, H., 2010. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, **38**(16):e164–e164.

Williamson, S. H., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R., and Bustamante, C. D., 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc Natl Acad Sci U.S.A.*, **102**(22):7882–7887.

Woitecki, A. M., Müller, J. A., van Loo, K. M., Sowade, R. F., Becker, A. J., and Schoch, S., 2016. Identification of synaptotagmin 10 as effector of npas4-mediated protection from excitotoxic neurodegeneration. *J Neurosci*, **36**(9):2561–2570.

Wu, H. and Su, B., 2008. Adaptive evolution of scml1 in primates, a gene involved in male reproduction. *BMC Evol Bio*, **8**(1):192.

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Wyckoff, G. J., Wang, W., Chung, W., et al., 2000. Rapid evolution of male reproductive genes in the descent of man. *Nature*, **403**(6767):304.

Yan, W., Si, Y., Slaymaker, S., Li, J., Zheng, H., Young, D. L., Aslanian, A., Saunders, L., Verdin, E., and Charo, I. F., *et al.*, 2010. Zmynd15 encodes a histone deacetylase-dependent transcriptional repressor essential for spermiogenesis and male fertility. *J Biol Chem.*, **285**(41):31418–31426.

Yang, Z., 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol*, **15**(5):568–573.

Yang, Z., 2007. Paml 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, **24**(8):1586–1591.

Yang, Z. and Nielsen, R., 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*, **19**(6):908–917.

Zhang, J., Nielsen, R., and Yang, Z., 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*, **22**(12):2472–2479.

Živković, D., Steinrücken, M., Song, Y. S., and Stephan, W., 2015. Transition densities and sample frequency spectra of diffusion processes with selection and variable population size. *Genetics*, **200**:601–617.

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

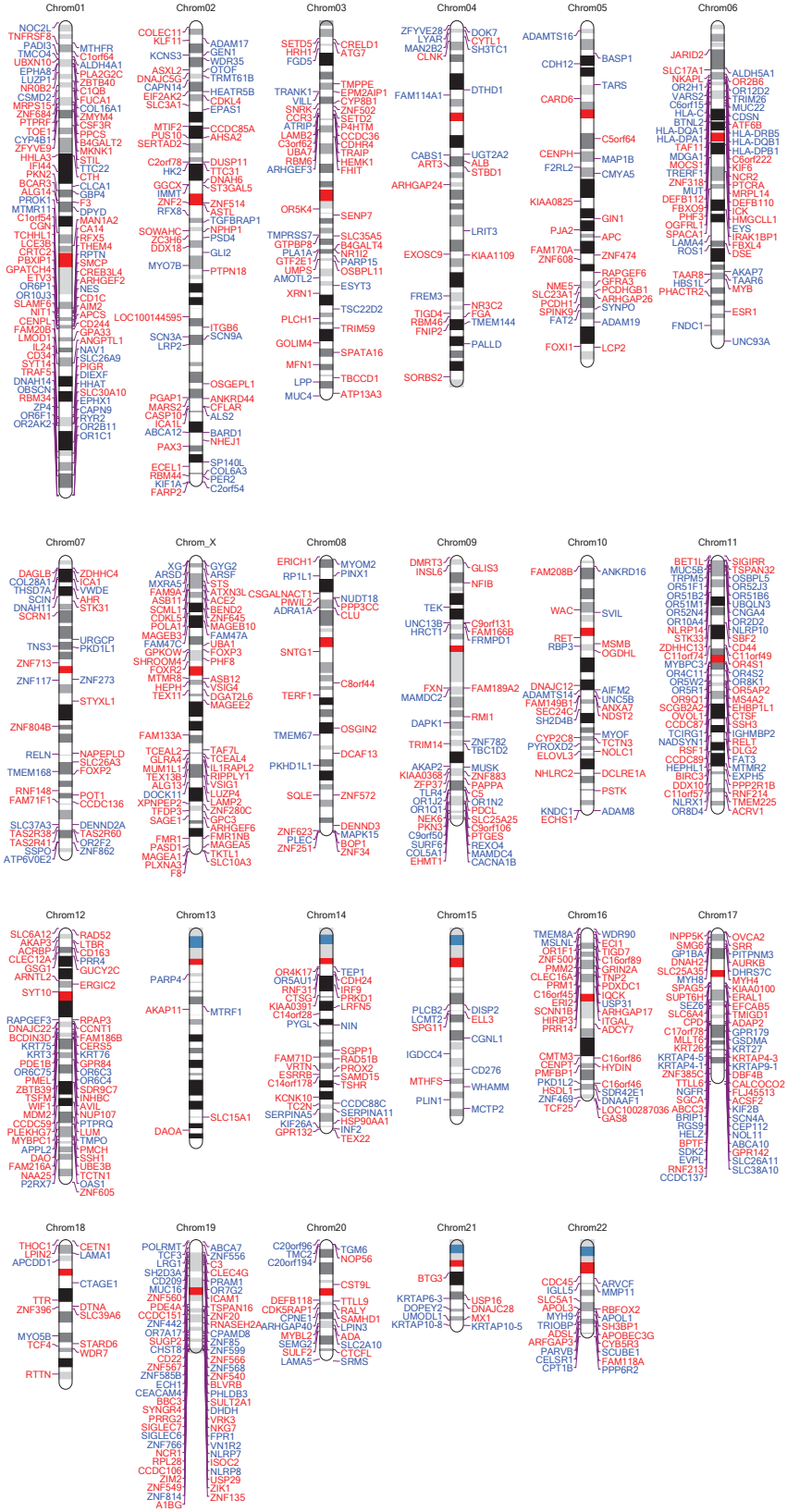


Figure 5: A map of Darwin selection of the human genome.

bioRxiv preprint doi: https://doi.org/10.1101/367482; this version posted July 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

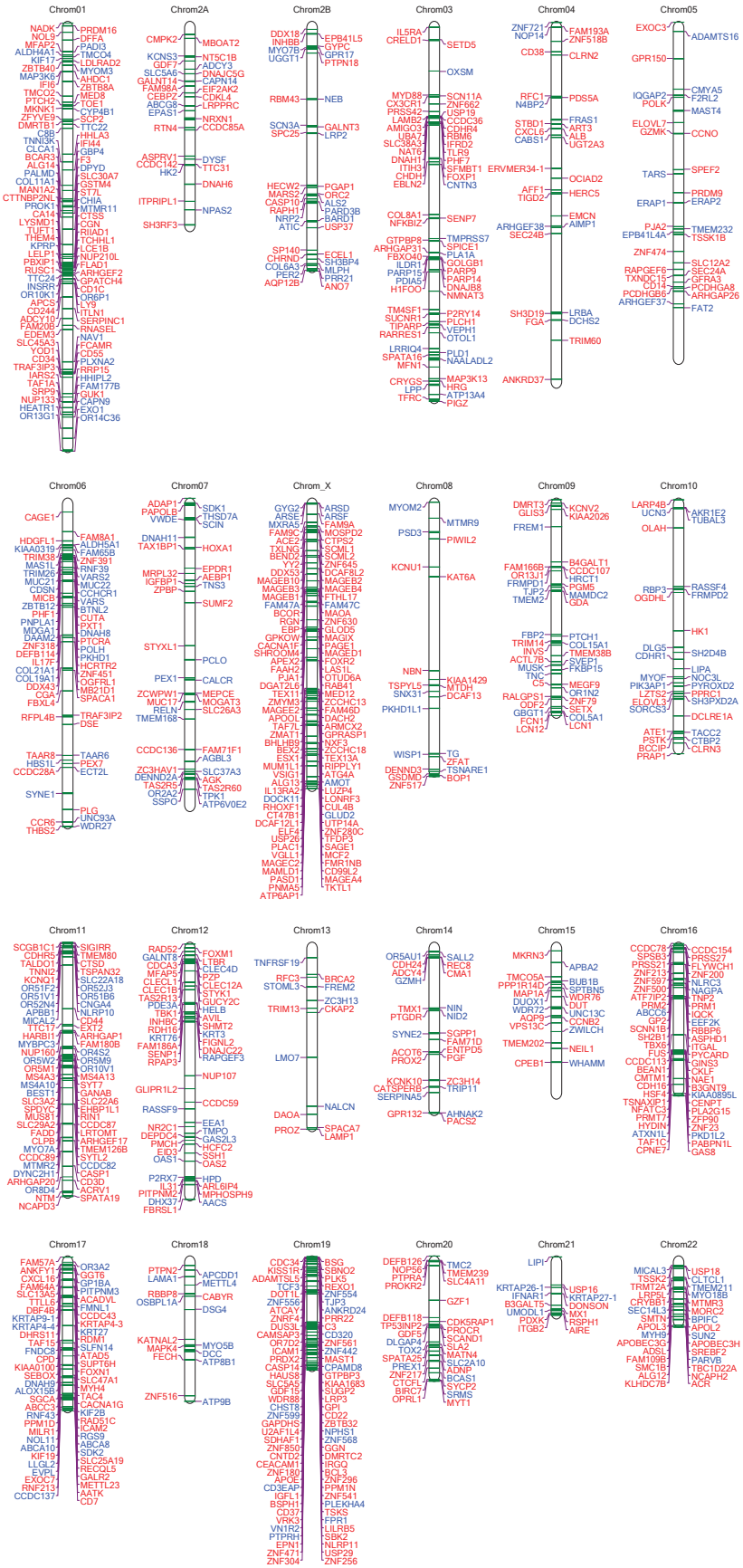


Figure 6: A map of Darwin selection of the chimpanzee genome.