Rates of molecular evolution predict intraspecies genetic diversity

Jiaqi Wu[1]*, Takahiro Yonezawa[2], Hirohisa Kishino[3]**

[1] School of Life Science and Technology, Tokyo Institute of Technology, Meguro Ward, Tokyo 152-8550, Japan

[2] Faculty of Agriculture, Tokyo University of Agriculture, Atsugi City, Kanagawa 243-0034, Japan

[3] Graduate School of Agricultural and Life Sciences, The University of Tokyo, Bunkyo Ward, Tokyo 113-8657, Japan

Correspondence:

* Jiaqi Wu (wujiaqi06@gmail.com)

** Hirohisa Kishino (kishino@lbm.ab.a.u-tokyo.ac.jp)

Abstract

Because the evolution of modern humans from early mammals has been a continuous process, human genetic diversity at the molecular level should reflect mammalian genetic diversity. To test this prediction, we contrasted the proportion of segregating sites, *q*, in human data from the 1000 Genome Project with phylogenetic trees of genes in 96 mammalian genomes ranging from human to platypus. The evolvability of a gene in the human population was found to be positively correlated with the long-term molecular evolutionary rate (r = 0.592). Further analysis of human singleton and *de novo* mutations suggested that natural selection begins shaping intraspecies genetic diversity soon after a mutation's initial appearance. By introducing the rate of molecular evolution as a predictor of *q*, species-specific selection on genes can be detected as deviations from predicted values. We analyzed the average deviations of 10,173 disease-related gene sets and detected 33 gene sets likely to be under human species-specific selection. Among them, nine gene sets were related, with highly significant and positive *t*-values, to substance-related disorders, thus indicating that the accelerated evolution of the natural reward pathway is the most representative adaptation in humans. Eight gene sets were significantly related, with negative *t*-values, to infectious diseases. This result may reflect co-evolution between pathogens and humans or recent bottlenecks in the human population caused by pandemic infectious diseases. Our findings demonstrate the potential usefulness of this population genomics approach that incorporates the rates of long-term molecular evolution as predictors.

Introduction

As pointed out by R. A. Fisher, the majority of mutations in a population are deleted by chance[1]. According to Motoo Kimura's neutral theory of molecular evolution, most mutations are either deleterious or selectively neutral[2; 3]. Hence, the molecular evolutionary rate of a gene is the product of the mutation rate and the proportion of neutral mutations[4]. In contrast, intraspecies genetic diversity is generated by mutations occurring along genealogies in a population that have neither reached fixation nor been lost. The shape of a genealogy depends on population history and effective population size, and the lifespan of a mutation depends on the selection intensity as well[5; 6]. Here, we note that all genes in a genome share the same population history. Among-gene variation in genetic diversity thus largely reflects among-gene variation in the proportion of neutral mutations, as lineages originating from deleterious mutants are too short to substantially contribute to present genetic diversity. Because among-gene variation in molecular evolutionary rates is also determined by the among-gene variation in the proportion of neutral mutations[7], we arrive at the hypothesis that long-term molecular evolutionary rates affect intraspecies genetic diversity.

To test the above hypothesis, we analyzed and compared the genetic diversity of humans with the mammalian molecular evolutionary rates of each gene, focusing on protein coding regions. The different evolutionary time scales of mammals and humans, including examples of gene trees and intraspecies genetic diversities, are shown in Figure 1. Indices of genetic diversity capture different age groups of surviving mutations. Depending on their unique population histories, the effect of long-term molecular evolutionary rates on intraspecies genetic diversity may be different among species. The proportion of segregating sites (q) is the most direct estimate of human intraspecies genetic diversity. Singletons are the collection of recent mutations along the terminal edges of a genealogy. The allele frequency of singletons in autosomes is 1/5,008, which is less than 0.002% of the number of sites in 1000 Genome Project data[8-10]. *De novo* mutations are obtained by comparing genomes of direct offspring with those of their parents, with the differences uncovered being the most close reflection of germ cell mutations[11]. By comparing the effects of molecular evolutionary rates on these different indices, we may gain some sense of the longevity of deleterious or slightly deleterious mutations that have been selected against[3].

In each population, natural selection works on phenotypes as well as the sets of genes that generate them. Depending on the surrounding environment, the targets of natural selection may vary among species. To identify species-specific targets, we searched for

significant deviations from the genetic diversities predicted by long-term molecular evolutionary rates. Genes with an unexpectedly large genetic diversity may be under reduced functional constraints, whereas those with an unexpectedly small genetic diversity may be subject to enhanced ones.

Materials and Methods

The probability of intraspecies polymorphism and measurement of molecular evolutionary rates

Genetic polymorphism within a species is generated by past mutations that have not reached fixation nor been deleted. The probability q that a site is polymorphic is formulated as

$$q = \int_{t=0}^{\infty} dt \int_{-\infty}^{0} g(s)ds \Pr(T > t|s, N_e(u), 0 \leq u \leq t) 2N_e(t)\nu \ . \qquad (1)$$

Here, ν is the mutation rate, s and $g(s)$ are the selection intensity and its distribution, and $N_e(u)$ is the effective population size at time u before present. We note that the probability of a positive value for s can be assumed to be negligible. T, which is the survival time of a mutation against either being fixed in the population or deleted, depends on the selection intensity and population history. Ignoring the contribution of slightly deleterious mutations for simplicity, we considered the scenario in which genetic polymorphism is largely generated by neutral mutations. In this scenario, (1) becomes

$$q = \nu p F, \qquad (2)$$

where

$$F = \int_{t=0}^{\infty} \Pr(T > t|s = 0, N_e(u), 0 \leq u \leq t) 2N_e(t)dt \ .$$

The probability that a mutation is selectively neutral, p, expresses the extent of the independence of the mutation from functional constraints. While p varies among genes, F does not. As a result, among-gene variation of the proportion of segregating sites reflects either the variation of mutation rates or the variation of functional constraints.

Interspecies molecular evolution, in contrast, is based on mutations that have become fixed in populations. Assuming the above scenario, the neutral theory of molecular evolution expresses the molecular evolutionary rate as

$$r = 2N_e(t)\nu \times p \times \frac{1}{2N_e(t)} = \nu p. \qquad (3)$$

Although intraspecies genetic variation is many orders of magnitude lower than that generated during the evolution of all mammals, equations (2) and (3) predict that the among-gene variation of the proportion of segregating sites is correlated with the among-gene variation of the molecular evolutionary rate.

The among-gene variation of the molecular evolutionary rate may be measured from the variation in total branch lengths of gene trees, which expresses the number of substitutions across the whole tree. At the same time, mean branch lengths of gene trees should well reflect the variation in evolutionary rates among genes, even in the presence of missing values, unless the resultant sub-trees of the (unobserved) gene trees are seriously skewed. Gene trees sometimes include extremely long terminal branches, even after careful inspection to control the quality of the alignment. Because mean branch length is sensitive to extremely long outlier branches, we also calculated the median of branch lengths as a robust estimate of among-gene variation in molecular evolutionary rates.

Neutral theory extended to multiple-gene molecular evolution

The well-accepted neutral theory of molecular evolution asserts that selectively advantageous mutations are rare among those mutations leading to inter-species variation[2]. As a result, the rate of molecular evolution is the product of the mutation rate, ν, and the proportion of neutral mutations, p[4]. By extending this relationship to the rates of multiple-gene molecular evolution, we obtained the following formula for the branch lengths of gene trees[7]:

$$b_{(i)}^{k} = c \times \bar{p}_i \times (t_k \times \nu_k) \times \bar{p}_{ik}.$$

Here, $\bar{p}_i$ represents the relative variability of gene i ($i = 1, \ldots, N$), and t_k and ν_k are the evolutionary time span and mutation rate along branch k ($k = 1, \ldots, M$). By applying a two-way ANOVA-type Poisson regression to the scaled branch lengths $\widetilde{N}_{(i)}^{k} = b_{(i)}^{k} \times L_i$ (where L_i is the length of gene i), namely,

$$E\left[\widetilde{N}_{(i)}^{k}\right] = \mathrm{c} \times L_i \times \alpha_i \times \beta_k$$

$$\text{where } \sum_{i=1}^{N} L_i \times \alpha_i = \sum_{k=1}^{M} \beta_k = 1,$$

we can estimate $\bar{p}_i$ and $t_k \times \nu_k$ based on gene effect $\hat{\alpha}_i$and branch effect $\hat{\beta}_k$. The interaction $\bar{p}_{ik}$, which captures the among-branch variation of functional constraints on a gene, is estimated as the ratio of $\widetilde{N}_{(i)}^{k}$ to the predicted value. When genomes include a complete set of

orthologous genes for all species, the maximum likelihood estimators of the gene effect and the branch effect can be obtained as follows:

$$\hat{\alpha}_i = \frac{\sum_{k=1}^{M} b_{(i)}^k}{\sum_{i=1}^{N} M \sum_{k=1}^{M} b_{(i)}^k \times L_i} \qquad (*)$$

$$\hat{\beta}_k = \frac{\sum_{i=1}^{N} b_{(i)}^k \times L_i}{\sum_{i=1}^{N} M \sum_{k=1}^{M} b_{(i)}^k \times L_i}$$

Summary statistics describing variation in long-term evolutionary rates among genes in a genome

In this study, we investigated the possibility that intraspecies population genetic diversity is affected by the between-species long-term molecular evolutionary rate. Specifically, we contrasted the rate of segregation in humans with the rate of molecular evolution in mammals. As can be seen in equation (*), the among-gene variation of the molecular evolutionary rate can be measured from the total branch lengths of gene trees.

In actual practice, however, gene trees often include missing taxa, partly because of incomplete sequencing but mainly because of gene loss over the course of mammalian evolution. As a result, the among-gene variation of total branch lengths reflects not only the among-gene variation of molecular evolutionary rate but also variation in taxon sampling. To reduce the effect of differential taxon sampling, we used mean branch lengths:

$$\bar{b}_{(i)} := \frac{1}{M} \sum_{k=1}^{M} b_{(i)}^k.$$

The variation in mean branch length should well reflect the variation in evolutionary rates among genes, even in the presence of missing values, unless the resultant sub-trees of the (unobserved) complete gene trees are seriously skewed.

The generated gene trees sometimes included extremely long terminal branches. Long branches in a gene tree may be ascribed to three phenomena: 1) sequencing or annotation errors, 2) bias caused by the model used for tree inference, or 3) lineage-specific selection. Even after careful inspection to control the alignment quality, however, we could not exclude the possibility of incomplete modeling of molecular evolution or lineage-specific selection in some cases. Because mean branch length is sensitive to extremely long outlier branches, we

also calculated the median of the branch lengths as a robust estimate of the among-gene variation in molecular evolutionary rates:

$$r_{(i)} \propto \tilde{b}_{(i)} := \frac{b_{(i)[\#b_{(i)}\div 2]} + b_{(i)[\#b_{(i)}\div 2+1]}}{2}.$$

With regard to protein evolution at the amino acid level, more than half of the branches of some extremely conserved genes were of zero length; this resulted in a median branch length of 0, which would be problematic in the succeeding regression analysis. We therefore removed the top five longest branches from each gene tree and obtained the truncated mean of branches, $\hat{\bar{b}}_{(i)}$, as an estimate of $\bar{r}_{(i)}$. Both $\hat{\bar{b}}_{(i)}$ and $\tilde{b}_{(i)}$ were used in this study.

Observed and corrected correlations between intraspecies genetic diversity and long-term molecular evolutionary rates

To examine the effects of long-term molecular evolutionary rates on intraspecies genetic diversity, we calculated the correlation between the rate of molecular evolution $B = \{b_{(1)}, b_{(2)}, \dots, b_{(i)}, \dots\}$ and the number of segregating sites $K = \{K_{(1)}, K_{(2)}, \dots, K_{(i)}, \dots\}$. Because $\bar{b}_{(i)}$ and $K_{(i)}$ were accompanied by the random noise of estimation and sampling, we corrected the observed correlation by taking into account the sampling variances of each $\bar{b}_{(i)}$ and $K_{(i)}$.

To see the effect of sampling variance, we denote the observed values of B and K as $\hat{B}$ and $\hat{K}$, where $\mathrm{E}_s[\hat{B}|B] = B$, $\mathrm{E}_s[\hat{K}|K] = K$, and the subscript "S" represents the sampling mean and variance. In addition, we use the subscript "G" to represent the among-gene mean and variance. To the first order, the expectation of the observed correlation between $\hat{B}$ and $\hat{K}$, $r_G(\hat{B}, \hat{K})$, is contrasted with $r_G(B, K)$ as follows:

$$\mathrm{E}_s[r_G(\hat{B}, \hat{K})] \sim \frac{\mathrm{E}_s[\mathrm{COV}_G(\hat{B}, \hat{K})]}{\sqrt{\mathrm{E}_s[\mathrm{V}_G(\hat{B})]\mathrm{E}_s[\mathrm{V}_G(\hat{K})]}}$$

$$= \frac{r_G(B, K)}{\sqrt{\left(1 + \frac{\mathrm{E}_G\left[\mathrm{V}_S[\hat{B}|B]\right]}{\mathrm{V}_G(B)}\right)\left(1 + \frac{\mathrm{E}_G\left[\mathrm{V}_S[\hat{K}|K]\right]}{\mathrm{V}_G(K)}\right)}} .$$

By evaluating the sampling variances of $\hat{B}$ and $\hat{K}$, we obtain the bias-corrected correlation as follows:

$$r_G(\hat{B},\hat{K}) := r_G(\hat{B},\hat{K}) \times \sqrt{\left(1+\frac{\mathrm{E}_G\left[\hat{\mathrm{V}}_S[\hat{B}|B]\right]}{\mathrm{V}_G(\hat{B})}\right)\left(1+\frac{\mathrm{E}_G\left[\hat{\mathrm{V}}_S[\hat{K}|K]\right]}{\mathrm{V}_G(\hat{K})}\right)}.$$

Sampling variance of mean branch lengths

We evaluated the sampling variance of branch lengths with reference to the Poisson random variables for the numbers of nucleotide or amino acid substitutions. Given that the branch length of gene *i* of branch k , $\hat{b}_{(i)}^{k}$, is the expected number of substitutions per site, we assume that the stochastic variance of $\hat{b}_{(i)}^{k} l_{(i)}$ is approximated by the variance of a Poisson random variable with mean $b_{(i)}^{k} l_{(i)}$, where $l_{(i)}$ is the sequence length of gene *i*:

$$\hat{V}\left(\hat{b}_{(i)}^{k} l_{(i)}\right) = \hat{b}_{(i)}^{k} l_{(i)}.$$

The variance of the mean branch lengths of gene *i* is therefore evaluated as

$$\hat{V}\left[\bar{\hat{b}}_{(i)}\right] = \hat{V}\left[\frac{1}{M}\sum_{k=1}^{M}\hat{b}_{(i)}^{k}\right] = \frac{1}{M^2}\sum_{k=1}^{M}\hat{V}\left(\hat{b}_{(i)}^{k}\right) = \frac{1}{M^2}\sum_{k=1}^{M}\frac{\hat{b}_{(i)}^{k}}{l_{(i)}} = \frac{1}{Ml_{(i)}}\bar{b}_{(i)}\,.$$

Bootstrap resampling of columns in the alignment would allow estimation of the variance of median and mean branch lengths. As an alternative approach, we developed a simple simulation method to evaluate the stochastic variance due to the randomness of substitutions. Given a gene tree, we generated the number of substitutions for each branch from the Poisson distribution, with a mean set to the product of the branch length and the gene length. We obtained the sample of the mean (and median) branch length by dividing the mean (and median) of the generated number of substitutions by the gene length.

The stochastic variance of intraspecies genetic diversity

If one assumes that mutations in a population are neutral and that the population is at equilibrium, the variance in the number of segregating sites, *K*, is

$$Var(K) = E(K) + \left(E(K)\right)^2 \frac{b_n}{{a_n}^2}$$

where $a_n = \sum_{i=1}^{n-1}\frac{1}{i}$ and $b_n = \sum_{i=1}^{n-1}\frac{1}{i^2}$. $\theta^2 b_n$, while *n* is the number of chromosomes in a sample from the population[12]. The first term reflects the randomness of mutations, and the

second term measures the stochasticity of the genealogy. When the population is not at equilibrium but has its own history, the variance becomes

$$Var(K) = E(K) + \left(E(K)\right)^2 \frac{V(T_c)}{E(T_c)^2}.$$

Here, $E(T_c)$ and $V(T_c)$ are the mean and variance of waiting times of coalescence events in genealogies[13]. Unfortunately, detailed information on genealogies is not available in most cases. We note, however, that the above variance formula is consistent with the variance formula of a negative binomial distribution. We therefore applied negative-binomial regression with a log link:

$$\log[E(K)] = \log(L) + \alpha + \beta \log(BL) + \gamma D_X$$

where $L = \{l_{(1)}, l_{(2)}, \dots, l_{(i)}, \dots\}$ is the length of genes and BL is the mean branch length. D_X is a dummy variable for genes on the X chromosome that incorporates differences in effective population size between an autosome and the X chromosome. We carried out this regression using the function `glm.nb` in the R package MASS[14]. The variance of K can be obtained by

$$V(K) = E(K) + \frac{[E(K)]^2}{\hat{\theta}}$$

where $\hat{\theta}$ is the shape parameter of the negative binomial distribution.

We also applied the model

$$E(K) = L \times [\alpha + \beta(BL)] \times \exp(\gamma D_X).$$

Maximum likelihood inference was conducted by applying the function `optim`. A positive value of α implies the intraspecies survival of mutations destined for deletion from the population on a macroevolutionary time scale.

Species specificity: the significant deviation of trait-associated gene sets

Intraspecies genetic diversity is affected by long-term molecular evolutionary rates. Species-specific enhanced/reduced functional constraints may be observed as residuals of the prediction of the above negative binomial regression. For each gene in a trait-associated gene set, we calculated the *p*-value of the observed number of segregating sites based on the predicted value and the estimated scale parameter. To obtain the significance of the gene set, we standardized the *p*-values by transforming them to the percentiles of their corresponding standard normal distribution. The *p*-value of the sample mean of these transformed values

228 was obtained using a *t*-distribution as a reference. The *t*-values measure the deviation from
229 the predicted genetic diversities of genes in gene sets. Positive *t*-values imply the existence of
230 reduced functional constraints or diversifying selection in a human lineage, whereas negative
231 values indicate the presence of enhanced functional constraints or purifying selection. Given
232 the value of false discovery rate, significant genes were selected by the Benjamini and
233 Hochberg procedure[15].

234 **Multiple alignment of mammalian single-copy genes**

235 We downloaded 96 complete mammalian genomes from GenBank and used a custom
236 Perl script to extract protein-coding sequences of each species. We generated a multi-
237 sequence file of all 6,366 single-copy genes in Class Mammalia listed in Wu et al. (2017)[7].
238 Alignments at the codon level were performed in Prank v.170427[16]. Gaps and sites with less
239 than 70% coverage among species as well as sequences with less than 30% coverage among
240 gene loci were removed. Out of the 6,366 single-copy genes, 6,137 were shared by more than
241 70 species and were used in this study. The total aligned sequence length of the 6,137 genes
242 was 12,665,862 bp corresponding to 4,221,954 amino acid residues.

243 **Inference of gene trees**

244 We estimated the maximum likelihood tree for each gene using IQ-TREE software[17],
245 which automatically performed model selection and determined the best data partitions. The
246 best evolutionary model for each gene was independently selected based on the Bayesian
247 information criterion and used for inference of the nucleotide tree. For the protein trees, we
248 used the MAM + F + γ model[7; 18], where MAM is the amino acid substitution matrix fitted to
249 the data[7]. All gene trees were calculated using 1,000 bootstrap replicates.

250 **Population data**

251 We downloaded single-nucleotide polymorphism (SNP) data for 2,504 humans
252 (*Homo sapiens*) from the 1000 Genome Project (Phase 3 data)[8-10]. On the basis of exon
253 annotations of reference genomes, we assigned each SNP to a gene. Only SNPs assigned to
254 coding regions of a gene were used in this study. Finally, we obtained 902,040 SNPs for
255 humans.

256 Information on *de novo* mutations in the human genome was collected from Jonsson
257 et al.[11] Human ancestral alleles were collected from Keightley and Jackson[19]. Individual-

258 derived alleles of each gene were collected by comparing SNPs of each individual with
259 human ancestral sequences.

260 **Improvement of exon annotations**

261 Because our study required precise identification of the codon positions of each SNP,
262 we further improved the exon annotations, especially the information on start codons in the
263 1000 Genome Project human annotation file.

264 We first extracted exon sequences from chromosomes. Next, we translated the exon
265 sequences into protein sequences using different start codons, i.e., with the codon starting
266 from either the first, second, or third nucleotide. We then used the translated protein
267 sequences as queries in blast searches (blastp-short in blast+ [20]) against all human proteins in
268 the 1000 Genome Project database. The blast results were further screened on the basis of
269 identity and the results of comparisons between query and match lengths. We defined
270 matches with an identity of 1 and a length equal to the query length as successful hits. Using
271 this approach, the start-codon annotations of all exons were further improved. We obtained
272 251,702 reliable exons in 17,921 genes from human reference genome hg19 and extracted
273 synonymous and non-synonymous SNPs based on the improved exon annotations in the
274 reference genome. Due to some gene includes missing information on its ancestral alleles, we
275 used 16, 260 human genes in this work.

276 **Data availability**

277 We combined all data into a single tab-delimited text file using a custom Perl script. Because
278 of inconsistencies in annotations and missing data, only 5,560 genes were used in the
279 correlation analysis. Source data and gene trees are available at
280 https://github.com/wujiaqi06/Long-Term-Rate.

281 **Results**

282 **Limits on gene evolvability at the molecular level**

283 After collecting human ancestral alleles from the data of Keightley and Jackson[19], we
284 counted the number of derived alleles of 16, 260 genes in 2,504 human individuals and found
285 strong evidence for a limit on evolvability at the molecular level. Figure 2A, in which each
286 point represents a gene and the mean and variance of this number among individuals is
287 indicated, shows the differences between mean derived alleles and the proportion of
288 segregating sites of each gene. In this figure, the number of derived alleles is obviously much

289 smaller than the proportion of segregating sites. To control variability due to differences in
290 gene lengths, we plotted all values after dividing them by gene lengths. The among-
291 individual variance was extremely close to the mean as long as the mean was less than $1 \times$
292 10^{-4} per coding-region site. In sharp contrast, the variance became extremely small when the
293 mean was larger than 1×10^{-3} per site.

294 The equality of the variance to the mean suggests a Poisson distribution, which is
295 consistent with neutral mutations on the star-like genealogy generated by the expanding
296 human population[12]. The low level of the variance, however, means that the number of
297 evolvable sites is strictly controlled, possibly by epistasis. A value of 1×10^{-4} per site
298 corresponds to an average of 10 kb between neighboring sites with derived alleles.

299 **Bridging micro- and macroevolution at the molecular level**

300 We obtained 5,560 single-copy genes across 96 mammalian species, ranging from platypus to
301 human, and calculated median branch lengths of nucleotide gene trees as an estimate of
302 among-gene variation in mammalian molecular evolutionary rates. Because such rate
303 variation arose during 185 million years of mammalian evolution[21], we refer to these rates as
304 long-term rates. In addition, we collected genomes of 2,504 humans from the 1000 Genome
305 Project (phase 3 data[8-10]) and calculated the proportion of segregating sites (q) as the among-
306 gene variation in human genetic polymorphism (Figure 2C). The time to the most recent
307 common ancestor of modern humans (*Homo sapiens*) is only 202,000 years[22], which
308 corresponds to approximately 0.1% of the mammalian time scale (Figure 1). As expected,
309 strong linear relationships were found between mammalian long-term rates and human q
310 (Figure 2C, cor = 0.592, $p < 2 \times 10^{-16}$).

311 Because uncertainties in branch lengths, the randomness of mutations, and the
312 stochasticity of genealogies deflate the values of net correlations, differences among species
313 in these correlations are difficult to evaluate (see Methods). Here, we note that the variance in
314 the proportion of segregating sites (q) is generally in the form of the sum of two terms. The
315 first term reflects random mutations and can be assumed to follow a Poisson distribution. The
316 second term, the stochasticity of genealogies, is proportional to the square of the mean[6; 13].
317 This variance formula is consistent with that of a negative binomial distribution, which takes
318 into account over-dispersion relative to the Poisson distribution. Because the coefficient of
319 the second term is determined by factors such as population histories, it can be assumed to be
320 constant among genes. We therefore conducted a negative binomial regression of human q on

321 mammalian long-term rates, with gene length included as an offset (Figure 2C, red curve).
322 The variances in human q were obtained using the estimated shape parameter of the negative
323 binomial distribution. With the estimated variances, we corrected the correlation between
324 mammalian long-term rates and human q. The bias-corrected correlation was as high as 0.665
325 (Figure 1C, $p < 2 \times 10^{-16}$), which indicates that the intraspecies genetic diversity of a gene is
326 thus largely influenced by the long-term evolutionary rate of the gene. Variation in functional
327 constraints on genes is the likely source of this causal relation[23]. The estimated shape
328 parameter in the negative binomial distribution was as large as 125,785, thus indicating that
329 over-dispersion of human q can be ignored, and human q largely followed a Poisson
330 distribution.

331 **Duration of the bridge between micro- and macroevolution**

332 Functional constraints vary among lineages[7], and molecular evolutionary rates accordingly
333 vary as well. Consequently, intraspecies genetic diversity may be more strongly correlated
334 with molecular evolutionary rates in subtrees of closely related species. The size of the most
335 influential subtree may depend on the scale of variability. We thus also wished to address the
336 question of how long the "long-term" molecular evolutionary rate of a gene affects
337 intraspecies diversity.

338 Figure 2D shows the raw correlation between the human proportion of segregating
339 sites (q) and molecular evolutionary rate variations inferred from different taxonomic
340 hierarchies, ranging from the human/chimpanzee divergence (human terminal branches) to
341 Family Hominidae (great apes) to superorder Euarchontoglires. Significant correlations
342 between human q and rate variations were found for all clades, with higher taxonomic
343 hierarchies showing stronger correlations. Furthermore, among-gene rate variations of
344 rodents and Laurasiatheria, neither of which includes humans, were highly correlated with
345 human q (cor = 0.453 and 0.552, respectively; both $p < 2 \times 10^{-16}$). The rate variation of the
346 Euarchontoglires had the highest raw correlation with human q (cor = 0.600, $p < 2 \times 10^{-16}$).
347 Taken together, these results indicate that long-term molecular evolutionary rates may have
348 been established at the early stage of the origin of mammals, thus acting as the main force
349 impacting the intraspecies genetic diversity of mammalian species.

350 **The immediate effect of natural selection**

351 Among-gene variation of intraspecies genetic diversity is caused by among-gene variation of
352 mutation rate and among-gene variation of functional constraints. To examine whether the

353 long-term molecular evolutionary rate of a gene is correlated with its mutation rate, we
354 regressed the rate of point mutations per generation (proportion of human *de novo* mutations)
355 on mammalian long-term rates. We found no correlation between the proportion of *de novo*
356 mutations and mammalian long-term rates (cor = 0.00, Figure 2E, Table 1). In contrast, the
357 proportion of singletons in 1000 Genome Project human data was found to be affected by
358 mammalian long-term rates (cor = 0.467; Figure 2F, Table 1). The expected number of
359 singletons was estimated as

360 $$\mathrm{E[number\ of\ singletons]/(gene\ length)} = 0.116 \times (\mathrm{long-term\ rates})^{0.418}.$$

361 The effect of long-term molecular evolutionary rate was highly significant ($p < 2 \times 10^{-16}$).

362 Singletons, which reflect relatively recent changes, may help shape complex traits[24;
363 25]. Selection, caused by long-term evolutionary force, has in turn influenced human genetic
364 diversity over a short time scale, probably soon after the occurrence of the initial mutation.

365 **Human species-specific evolution**

366 Summarizing the above results, the intraspecies genetic diversity of a gene depends on the
367 functional constraints on that gene, which are reflected by its long-term molecular
368 evolutionary rate. The next issue we addressed was species specificity. Species-specific
369 enhanced/reduced functional constraints may be ascertained by examining deviations in the
370 proportion of segregating sites (q) from predictions based on long-term molecular
371 evolutionary rates. We standardized the deviation by transforming the negative binomial
372 regression to a standard normal distribution. First, we obtained the lower tail p-value of a
373 gene by contrasting the observed q-value with a negative binomial distribution based on the
374 predicted value and the estimated shape parameter. This p-value was mapped to the quantile
375 of the standard normal distribution. A positive value for the deviation indicates a reduction in
376 functional constraints or the presence of diversifying selection, whereas a negative value
377 reflects enhancement of functional constraints or the occurrence of purifying selection.

378 To examine species-specific constraints on traits, we analyzed the average deviation
379 of disease-associated genes. In this manner, we hoped to identify the human species-
380 specificity of biological traits responsible for various diseases. We first downloaded disease-
381 associated gene sets from the DisGeNET database[26]. To obtain a robust result based on
382 multiple genes, we focused on gene sets that included at least two genes contained in our own
383 dataset. Consequently, we analyzed 10,173 disease-related gene sets (Table 2, S1–S2). To
384 obtain the average deviation of a gene set from predicted genetic diversities, we converted the

385 *p*-values of genes in the gene set to quantiles of the standard normal distribution. The average
386 deviation was measured as the *t*-value of the quantiles (see Methods). Using a false discovery
387 rate of 0.10, we identified 33 disease-associated gene sets in humans.

388 **Positive emotion shapes modern humans**

389 Out of the 33 detected human gene sets, 9 were related to substance-related disorders, and
390 their relationships were all highly significant (Table 2, Figure 3B). Substance-related
391 disorders are mainly defined by a reliance and/or abuse of a certain substance. The
392 underlying mechanism of these disorders stems from the function of the "substance" (e.g.,
393 drug) on the natural reward pathway of the human motivational-emotional system. In nature,
394 reward pathways generate positive emotions that motivate organisms to search for food,
395 water, mates, and other beneficial environmental resources and to avoid predators[27]. Humans
396 have the most complex societal culture of all living organisms and enjoy the largest array of
397 "natural rewards", including science, art, literature, and music. Motivation is a basic driver of
398 human behavior, and positive emotion is at its core[28]. The positive *t*-values associated with
399 substance-related disorders imply that their human specificity among mammals is due to
400 evolution of the natural reward pathway in the motivational-emotional system. We also
401 conducted a gene set analysis using human singletons (Table S1). The detected diseases were
402 mostly substance-related disorders, which indicates that such selection has occurred quite
403 recently.

404 We extracted a subset comprising non-synonymous SNPs and conducted a gene set
405 test. Because changes in amino acids in protein sequences are subject to stronger selection
406 pressure, the detected gene sets were almost all associated with negative *t*-values (Table S2);
407 however, substance-related disorders with a positive *t*-value ($t = 4.286$) were also found in
408 the significant selected gene sets. We conclude that genes related to substance-related
409 disorders are under positive selection in modern humans and are the most representative
410 example of species-specific selection in our own species. Finally, gene sets related to
411 profound mental retardation were also uncovered with a negative *t*-value (Table 2, $t =$
412 -3.993).

413 **Selection on cancer-associated gene sets**

414 The cancer gene sets detected in humans had negative *t*-values. The associated cancers were
415 all blood cancers; for example, multiple myeloma, B-cell lymphomas, and chronic
416 lymphocytic leukemia were listed as well as adult T-cell lymphoma/leukemia, a blood cancer

417 caused by human T-lymphotropic virus 1 (HTLV-1) infection. These cancers may have
418 severe effects on the human population; thus, the associated gene sets are under strong
419 purifying selection. These results are consistent with those obtained using non-synonymous
420 SNPs.

421 **Echoes of infectious diseases in the human genome**

422 The human outliers also included eight gene sets with negative *t*-values associated with
423 infectious diseases. These diseases were influenza, chronic hepatitis B, *Helicobacter pylori*
424 infection, candidiasis, lupus vulgaris, adult T-cell lymphoma/leukaemia, Saint Louis
425 encephalitis, and gingivitis (Table 2). The unexpectedly low genetic diversity of these gene
426 sets has two possible causes related to ancient infectious disease—either a pathogen co-
427 evolved with humans, or an infectious disease affected human populations worldwide,
428 leading to a selection bottleneck. *Helicobacter pylori* is known to have coevolved with
429 human populations and has caused adaptations of disease-associated genes[29]. Influenza has
430 infected almost the entire human population and has caused several severe pandemics within
431 the last 300 years[30]. In regard to the other infectious diseases mentioned above, their
432 relationship to human evolution has rarely been studied.

433 **Discussion**

434 Intraspecies genetic diversity is influenced by long-term molecular evolutionary rates, which
435 are largely the products of mutation rate and the proportion of neutral mutations. Because we
436 saw no correlation between the rate of *de novo* mutations and long-term evolutionary rates,
437 the among-gene variation in genetic diversity may reflect variable selective constraints on
438 genes and is therefore correlated at the protein level with mammalian molecular evolutionary
439 rates. The correlation between the human proportion of segregating sites (q) and mammalian
440 molecular evolutionary rates at the protein level was 0.493, which was not as strong as the
441 correlation between human q and mammalian nucleotide molecular evolutionary rates (cor =
442 0.592). Protein-level long-term rates had an unexpectedly lower predictive power than DNA-
443 level long-term rates, which is possibly because the former encompassed smaller numbers of
444 evolutionary events.

445 The intraspecies diversity level of a gene can be decomposed into the proportion of
446 segregating sites and the pattern of allele frequencies of the segregating sites. The long-term
447 molecular evolutionary rate largely affects the proportion of segregating sites, the latter

determining evolvability at the molecular level. The lifespan of deleterious mutations is short, and intraspecies evolvability is mostly determined by the proportion of neutral mutations.

In summary, we have demonstrated that the molecular evolutionary rate, which represents interspecies genomic diversity, sets the baseline for intraspecies genetic diversity. This effect has been largely ignored in population genomics studies, in part because of an insufficient amount of genomic data. The effect of natural selection on intraspecies genetic diversity has been implicitly assumed to be small because recent mutations, which may be destined for deletion from a population over a long-term evolutionary time scale, have a better chance of survival to the present. The evolvability of mutations over an evolutionary time scale, however, seems to be determined shortly after their origin. The amount of intraspecies genetic diversity can be predicted from interspecific genetic diversity, and deviations from predicted values represent the species specificity of enhanced/reduced functional constraints. An in-depth inspection of these deviations is in progress.

Acknowledgements

This study was supported by a Grant-in-Aid for Scientific Research (B) (19H04070) and a Grant-in-Aid for JSPS Research Fellows (18F18385) from the Japan Society for the Promotion of Science. We deeply thank Jeffrey L. Thorne of North Carolina State University, Zhenqiu Liu of Fudan University and Nikaido Masato of the Tokyo Institute of Technology for their valuable comments and suggestions on our manuscript. We thank Barbara Goodson, from Edanz Group (www.edanzediting.com/ac), for editing the English text of a draft of this manuscript.

Contributions

J.W., T.Y., and H.K. designed the research. J.W. and H.K. formulated the methods and performed the data analysis. J.W., T.Y., and H.K. wrote the paper.

Declaration of interests

The authors declare no competing interests.

Correspondence and requests for materials

Correspondence and requests for materials should be addressed to wujiaqi06@gmail.com

Tables

Table 1. Negative binomial regression of the proportion of human *de novo* mutations and singletons on mammalian molecular evolutionary rates of genes. The regression model is $\log\big(\mathrm{E}[\mathrm{counts}]/(\mathrm{gene\ length})\big) = \alpha + \beta \log(\mathrm{long\ term\ rates})$, that is, $\mathrm{E}[\mathrm{counts}]/(\mathrm{gene\ length}) = e^{\alpha}(\mathrm{long\ term\ rates})^{\beta}$.

	De novo mutation			Singleton		
	Estimate	SE	*p*-value	Estimate	SE	*p*-value
Intercept	–10.403	0.544	$< 2 \times 10^{-16}$	–2.158	0.046	$< 2 \times 10^{-16}$
Log(long-term rates)	–0.084	0.110	0.443	0.418	0.009	$< 2 \times 10^{-16}$

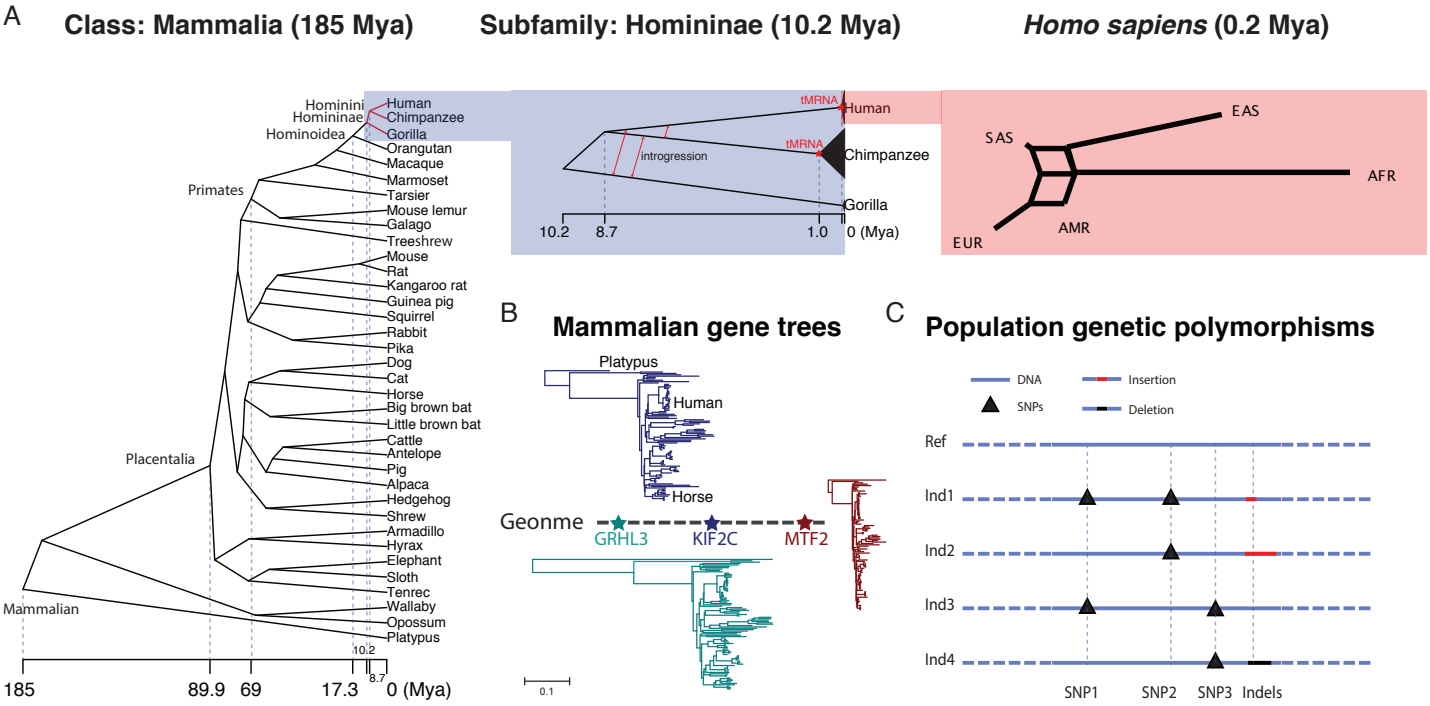
485 **Table 2. Significant disease-associated human gene sets with a false discovery rate of**
486 **10%.**

	t-value*	*p*-value	*q*-value	Disease Type
Substance-related disorders	5.974	1.18×10^{-7}	0.0012	Mental disease
Organic mental disorders, substance-induced	5.826	2.51×10^{-7}	0.0013	Mental disease
Prescription drug abuse	5.826	2.51×10^{-7}	0.0009	Mental disease
Multiple myeloma	−5.060	6.43×10^{-7}	0.0016	Blood cancer
Drug use disorders	5.462	9.52×10^{-7}	0.0019	Mental disease
Substance dependence	5.218	1.91×10^{-6}	0.0032	Mental disease
Drug habituation	4.995	4.47×10^{-6}	0.0065	Mental disease
Influenza	−4.607	7.98×10^{-6}	0.0101	Infectious disease
Substance use disorders	4.556	2.15×10^{-5}	0.0243	Mental disease
Substance abuse problem	4.520	2.46×10^{-5}	0.0250	Mental disease
Hepatitis B, chronic	−4.373	5.49×10^{-5}	0.0508	Infectious disease
Thyroid-associated opthalmopathies	−4.568	7.39×10^{-5}	0.0626	Autoimmune disease
Calcification of coronary artery	4.087	7.72×10^{-5}	0.0604	Blood vessel disease
Pneumonia	−4.029	8.97×10^{-5}	0.0652	Infectious disease
Infection	−3.995	1.03×10^{-4}	0.0699	Infectious disease
Thrombocytopenia due to platelet alloimmunization	−4.996	1.10×10^{-4}	0.0702	Blood disease
B-Cell Lymphomas	−3.922	1.16×10^{-4}	0.0695	Blood cancer
Drug abuse	4.000	1.25×10^{-4}	0.0707	Mental disease
Infection caused by *Helicobacter pylori*	−4.005	1.33×10^{-4}	0.0709	Infectious disease
Candidiasis	−5.516	1.33×10^{-4}	0.0675	Infectious disease
Lupus erythematosus	−3.992	1.38×10^{-4}	0.0666	Autoimmune disease
Profound mental retardation	−3.993	1.42×10^{-4}	0.0655	Mental disease
Joubert syndrome 21	−4469.702	1.42×10^{-4}	0.0630	Developmental disorders
Chronic lymphocytic leukemia	−3.825	1.57×10^{-4}	0.0664	Blood cancer
Lupus vulgaris	−3.968	1.64×10^{-4}	0.0668	Infectious disease
Lupus erythematosus, discoid	−3.893	1.95×10^{-4}	0.0762	Skin disease
Autoimmune thrombocytopenia	−4.386	1.98×10^{-4}	0.0745	Autoimmune disease
Dysplastic nevus	−4.476	2.31×10^{-4}	0.0841	Developmental disorders
Fanconi anemia	−3.794	2.46×10^{-4}	0.0864	Genetic disease
Immunosuppression	−4.080	2.57×10^{-4}	0.0873	-
Adult T-Cell lymphoma/leukemia	−3.740	2.58×10^{-4}	0.0847	Blood cancer
Encephalitis, St. Louis	−4.104	3.01×10^{-4}	0.0957	Infectious disease
Gingivitis	−4.645	3.17×10^{-4}	0.0978	Infectious disease

487 * see Methods.

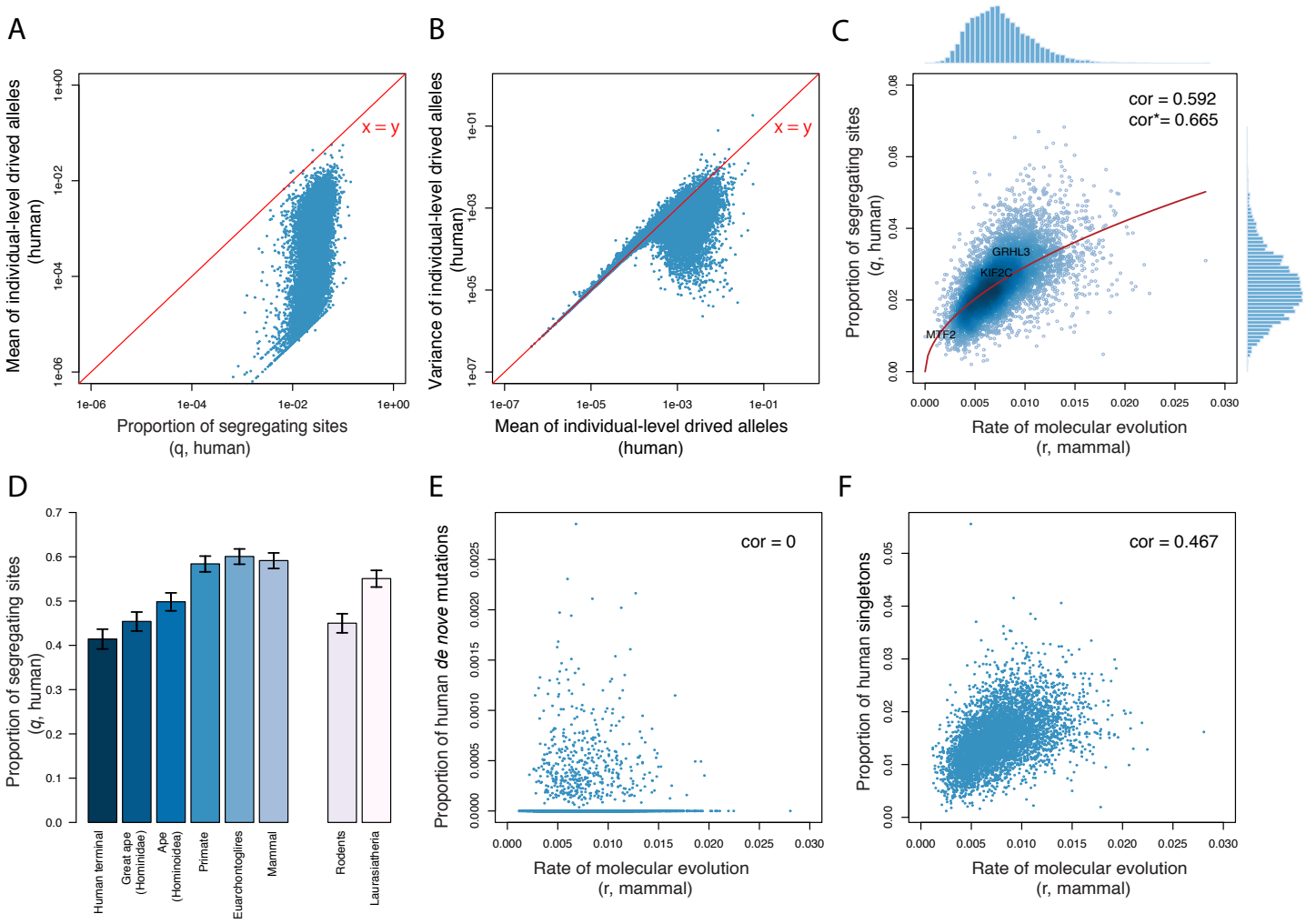
488

Figures


A
Class: Mammalia (185 Mya)
Subfamily: Homininae (10.2 Mya)
Homo sapiens (0.2 Mya)
Hominini
Hominine
Hominoidea
Human
Chimpanzee
Gorilla
Orangutan
Macaque
Marmoset
Tarsier
Mouse lemur
Galago
Treeshrew
Mouse
Rat
Kangaroo rat
Guinea pig
Squirrel
Rabbit
Pika
Dog
Cat
Horse
Big brown bat
Little brown bat
Cattle
Antelope
Pig
Alpaca
Hedgehog
Shrew
Armadillo
Hyrax
Elephant
Sloth
Tenrec
Wallaby
Opossum
Platypus
Primates
Placentalia
Mammalian
185
89.9
69
17.3
10.2
8.7
0 (Mya)
tMRNA
introgression
10.2
8.7
1.0
0 (Mya)
SAS
EAS
AFR
AMR
EUR
B
Mammalian gene trees
Platypus
Human
Horse
Geonme
GRHL3
KIF2C
MTF2
0.1
C
Population genetic polymorphisms
DNA
Insertion
SNPs
Deletion
Ref
Ind1
Ind2
Ind3
Ind4
SNP1
SNP2
SNP3
Indels

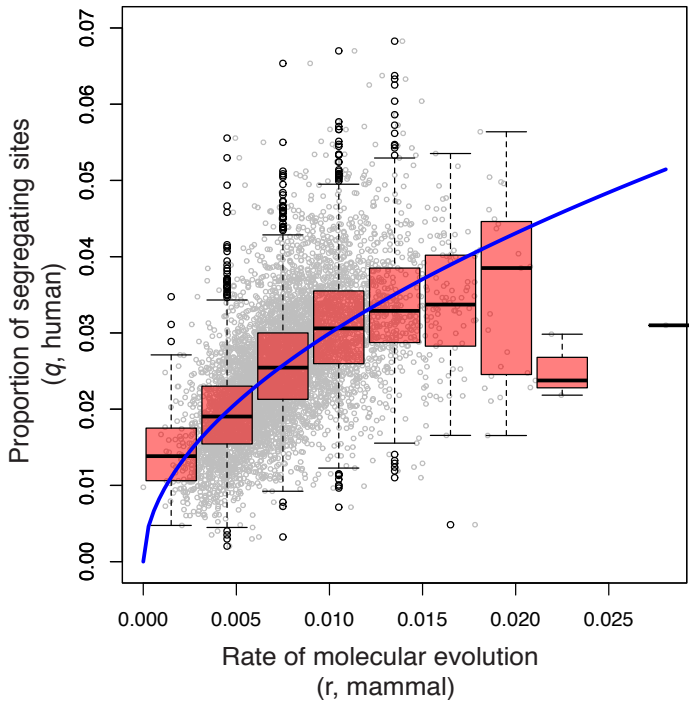
491 **Figure 1. Time scales of mammalian and human evolution. A.** The differing evolutionary
492 time scales of Class Mammalia, Subfamily Homininae, and *Homo sapiens*. The time tree for
493 mammals is from dos Reis (2012)[21]. **B.** Examples of mammalian gene trees. **C.** Examples of
494 population-level genetic diversities, including indels and single nucleotide polymorphisms
495 (SNPs).

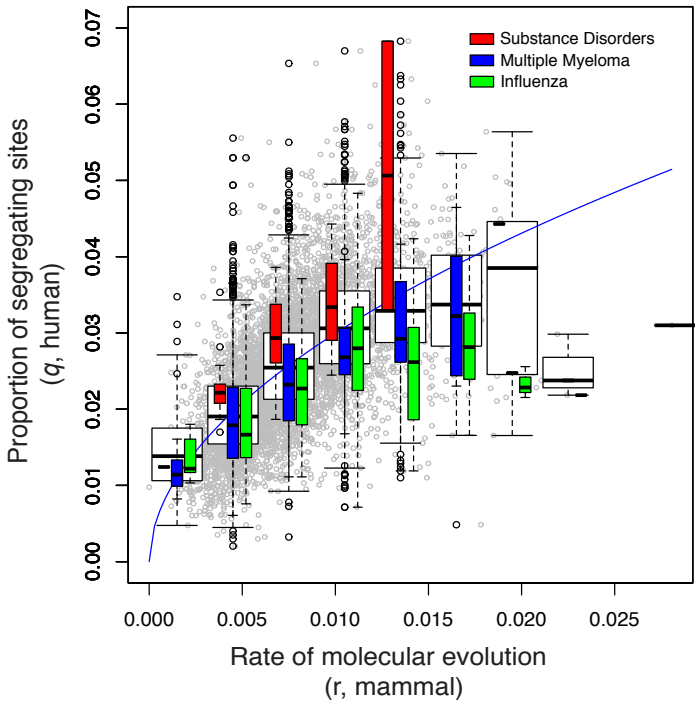
496


A
Mean of individual-level drived alleles (human)
Proportion of segregating sites (q, human)
x = y
B
Variance of individual-level drived alleles (human)
Mean of individual-level drived alleles (human)
x = y
C
Proportion of segregating sites (q, human)
Rate of molecular evolution (r, mammal)
cor = 0.592
cor*= 0.665
GRHL3
KIF2C
MTF2
D
Proportion of segregating sites (q, human)
Human terminal
Great ape (Hominidae)
Ape (Hominoidea)
Primate
Euarchontoglires
Mammal
Rodents
Laurasiatheria
E
Proportion of human de nove mutations
Rate of molecular evolution (r, mammal)
cor = 0
F
Proportion of human singletons
Rate of molecular evolution (r, mammal)
cor = 0.467

497 **Figure 2. Limits on evolvability at the molecular level. A.** Differences between mean
498 derived alleles and the proportion of segregating sites of 17,591 genes across 2,504 human
499 individuals. **B.** Mean and variance of derived alleles of 17,591 genes across 2,504 human
500 individuals. **C.** Correlation between the human proportion of segregating sites (q) and the rate
501 of mammalian molecular evolution (r) of 5,560 genes. **D.** Correlation between the human
502 proportion of segregating sites (q) and the rate of molecular evolution (r) based on gene trees
503 of different taxonomic clades. **E.** Correlation between the proportion of human *de novo*
504 mutations and the rate of mammalian molecular evolution (r) of 5,560 genes. **F.** Correlation
505 between the proportion of human singletons and the rate of mammalian molecular evolution
506 (r) of 5,560 genes.

507


Proportion of segregating sites
(q, human)
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.000
0.005
0.010
0.015
0.020
0.025
Rate of molecular evolution
(r, mammal)


Substance Disorders
Multiple Myeloma
Influenza
Proportion of segregating sites
(q, human)
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.000
0.005
0.010
0.015
0.020
0.025
Rate of molecular evolution
(r, mammal)

508 **Figure 3. Disease-associated human gene sets.** **A**. Scatter plot, boxplot, and regression
509 model of the human proportion of segregating sites (q) vs. mammalian molecular
510 evolutionary rates of 5,560 genes. **B.** Examples of significant disease-associated gene sets in
511 humans revealed by analysis of 10,173 disease-associated gene sets. Genes related to the
512 following diseases in humans are shown as examples: substance-related disorders (n = 64),
513 multiple myeloma (n = 395), and influenza (n = 588).

514

515

516 **Supplemental table captions**

517 **Table S1. Significant disease-associated gene sets of human singletons with a false**
518 **discovery rate of 10%.**

519 **Table S2. Significant disease-associated gene sets of human non-synonymous SNPs with**
520 **a false discovery rate of 10%.**

521

522

523

References

1. Fisher, R.A. (1930). The Genetical Theory of Natural Selection (London: Oxford University Press).
2. Kimura, M. (1968). Evolutionary rate at the molecular level. Nature 217, 624–626.
3. Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. Nature 246, 96-98.
4. Kimura, M. (1977). Preponderance of Synonymous Changes as Evidence for Neutral Theory of Molecular Evolution. Nature 267, 275–276.
5. Kimura, M., and Ohta, T. (1969). The average number of generations until extinction of an individual mutant gene in a finite population. Genetics 63, 701–709.
6. Kimura, M., and Ohta, T. (1969). The average number of generations until fixation of a mutant gene in a finite population. Genetics 61, 763–771.
7. Wu, J., Yonezawa, T., and Kishino, H. (2017). Rates of Molecular Evolution Suggest Natural History of Life History Traits and a Post-K-Pg Nocturnal Bottleneck of Placentals. Curr Biol 27, 3025–3033 e3025.
8. Genomes Project, C., Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature 467, 1061-1073.
9. Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature 526, 68-74.
10. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature 526, 75-81.
11. Jonsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., Hardarson, M.T., Hjorleifsson, K.E., Eggertsson, H.P., Gudjonsson, S.A., et al. (2017). Parental influence on human germline de novo mutations in 1,548 trios from Iceland. Nature 549, 519-522.
12. Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7, 256-276.
13. Živković, D., and Wiehe, T. (2008). Second-order moments of segregating sites under variable population size. Genetics 180, 341–357.
14. Venables, W.N., and Ripley, B.D. (2002). Modern Applied Statistics with S. Fourth Edition.(Springer Publishing Company, Incorporated).

557 15. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical
558 and powerful approach to multiple testing. J Roy Stat Soc B Met 57, 289–300.
559 16. Loytynoja, A. (2014). Phylogeny-aware alignment with PRANK. Methods Mol Biol
560 1079, 155-170.
561 17. Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast
562 and effective stochastic algorithm for estimating maximum-likelihood phylogenies.
563 Mol Biol Evol 32, 268-274.
564 18. Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses.
565 Trends Ecol Evol 11, 367–372.
566 19. Keightley, P.D., and Jackson, B.C. (2018). Inferring the Probability of the Derived vs. the
567 Ancestral Allelic State at a Polymorphic Site. Genetics 209, 897-906.
568 20. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and
569 Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics
570 10, 421.
571 21. dos Reis, M., Inoue, J., Hasegawa, M., Asher, R.J., Donoghue, P.C., and Yang, Z. (2012).
572 Phylogenomic datasets provide both precision and accuracy in estimating the
573 timescale of placental mammal phylogeny. Proc Biol Sci 279, 3491–3500.
574 22. Hallast, P., Maisano Delser, P., Batini, C., Zadik, D., Rocchi, M., Schempp, W., Tyler-
575 Smith, C., and Jobling, M.A. (2016). Great ape Y Chromosome and mitochondrial
576 DNA phylogenies reflect subspecies structure and patterns of mating and dispersal.
577 Genome Res 26, 427–439.
578 23. Havrilla, J.M., Pedersen, B.S., Layer, R.M., and Quinlan, A.R. (2019). A map of
579 constrained coding regions in the human genome. Nat Genet 51, 88-95.
580 24. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S.,
581 McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of
582 rare coding variation from deep sequencing of human exomes. Science 337, 64-69.
583 25. Field, Y., Boyle, E.A., Telis, N., Gao, Z., Gaulton, K.J., Golan, D., Yengo, L., Rocheleau,
584 G., Froguel, P., McCarthy, M.I., et al. (2016). Detection of human adaptation during
585 the past 2000 years. Science 354, 760–764.
586 26. Pinero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M.,
587 Sanz, F., and Furlong, L.I. (2015). DisGeNET: a discovery platform for the dynamical
588 exploration of human diseases and their genes. Database (Oxford) 2015, bav028.
589 27. Durrant, R., Adamson, S., Todd, F., and Sellman, D. (2009). Drug use and addiction:
590 evolutionary perspective. Aust N Z J Psychiatry 43, 1049–1056.

591 28. Chiew, K.S., and Braver, T.S. (2014). Dissociable influences of reward motivation and
592 positive emotion on cognitive control. Cogn Affect Behav Neurosci 14, 509–529.
593 29. Atherton, J.C., and Blaser, M.J. (2009). Coadaptation of *Helicobacter pylori* and humans:
594 ancient history, modern implications. J Clin Invest 119, 2475–2487.
595 30. Nicholls, H. (2006). Pandemic influenza: the inside story. PLoS Biol 4, e50.
596

597