Frequent non-allelic gene conversion on the human lineage and its effect on the divergence of gene duplicates – Supplementary Information

Arbel Harpak, Xun Lan, Ziyue Gao and Jonathan K. Pritchard

November 2, 2017

Contents

1 Supplementary Methods

1.1 Gene families data

To avoid complex gene families, where Non-Allelic Gene Conversion (NAGC) could occur between multiple members within the family, we focused our analyses on a set of 1,444 reciprocal best-matched protein-coding gene pairs in the human reference genome (build 37) identified by Lan and Pritchard [1]. We obtained the orthologs of these genes in four other primates (chimpanzee, gorilla, orangutan, macaque) and in mouse from the same study (**Table S1**). We required the orthologs to have at least 80% of the coding sequences aligned and at least 50% of the coding sequences identical to the human genes. For the task of identifying converted tracts, we applied further filtering on the input data. We used the software *MrBayes* [2] to estimate gene family genealogies with the set of exons of our genes as input. Note that here–unlike the rest of the analysis–we used exonic sequences rather that intronic. We filtered the gene families used as data for the Hidden-Markov Model (HMM) to gene families in which the most probable genealogy supports a duplication prior to the split of the two focal species (human and one of four non-human primates) were kept.

Species	Genome assembly	Gene annotation
Human	Ensembl GRCh37	release 73
Chimpanzee	Ensembl CHIMP2.1.4	release 70
Gorilla	Ensembl gorGor3	release 73
Orangutan	Ensembl PPYG2	release 73
Macaque	Ensembl Mmul_1	release 70
Mouse	Ensembl GRCm38	release 70

Table S1: A list of genome assemblies and gene annotations used.

1.2 Identifying converted regions

1.2.1 Hidden Markov Model (HMM)

NAGC can change the local genealogy of gene families (**Fig. 1B**). We designed an HMM to identify genealogy changes underlying variation patterns in the gene family sequences. We used a subset of the data, namely introns from small gene families with duplicates in two species (either human/chimpanzee or human/macaque) as input. Each intron family is composed of 4 sequences– two for each species. After filtering, 39 gene families (each consisting of one or more introns; 26 for human/chimpanzee and 13 for human/macaque) were included as input.

Although the application of the HMM are mostly standard, we briefly describe them here for completeness. One noteworthy feature is that the parameters the HMM are not the emission and transition probabilities themselves but instead parameters that determine these probabilities through an evolutionary model. Another feature of note is the partial sharing of parameters across introns and across gene families which we describe below.

Each intron consists of 4 orthologous sequences (two for each species). For each species, each nucleotide can be in one of three hidden states: unconverted (00), converted using gene 1 as template (10), and converted using gene 2 as template (01). We assume that all NAGC events involve only the two genes at hand and that at most one NAGC event occurred at each nucleotide. The full state space for a nucleotide is a combination of the two independent species-specific states. Therefore, the HMM has 9 hidden states, S=$\{$0000, 0010, 0001, 1000, 1010, 1001, 0100, 0110, 0101$\}$. Observations $O = O_y$ consist of introns y from $Q = 39$ gene families. Each intron y in each gene family q has four homologous sequences with total length, l_y. The parameters of the HMM are as follows:

π_i, the probability of the first nucleotide of an intron being in state i.

ν, the probability of the $t + 1$ nucleotide being in a converted state (10 or 01) given that the nucleotide t is in the unconverted state (00).

α, the probability of the $t + 1$ nucleotide being in a converted state (10 or 01) given that

nucleotide $t+1$ is in a converted state (10 or 01).

r_{0q}, the probability of substitution per nucleotide from duplication to speciation for gene family q.

r_{1q}, the probability of substitution per nucleotide from speciation to conversion for gene family q.

r_{2q}, the probability of substitution per nucleotide from conversion to present for gene family q.

Note that first three parameters are shared across all intronic sequences of all Q genes, while the last three are shared between introns of a gene, but not across genes. The likelihood function for $\Theta = (\boldsymbol{\pi}, \alpha, \nu, R_0 = (r_{01}, r_{02}, \ldots, r_{0Q}), R_1 = (r_{11}, r_{12}, \ldots, r_{1Q}), R_2 = (r_{21}, r_{22}, \ldots, r_{2Q}))$ is defined as follows:

$$\mathcal{L}(\Theta) = P(O|\Theta) = \prod_{q=1}^{Q} \prod_{y \in Y_q} P(O_y|\Theta) = \prod_{q=1}^{Q} \prod_{y \in Y_q} P(O_y|\pi, \alpha, \nu, r_{0q}, r_{1q}, r_{2q}),$$

where Y_q is the set of introns in gene q. The transition matrix for a single species is

$$\mathbf{A}' = \begin{array}{c} \begin{array}{ccc} 00 & 10 & 01 \end{array} \\ \begin{bmatrix} 1-\nu & \nu/2 & \nu/2 \\ 1-\alpha & \alpha & 0 \\ 1-\alpha & 0 & \alpha \end{bmatrix} \end{array} \begin{array}{c} 00 \\ 10 \\ 01 \end{array}$$

and the full transition matrix (i.e., for the state space of two species) is derived by considering the independent evolution of orthologs following speciation,

$$\mathbf{A}'' = \begin{array}{c} \begin{array}{ccccccccc} 0000 & 0010 & 0001 & 1000 & 1010 & 1001 & 0100 & 0110 & 0101 \end{array} \\ \begin{bmatrix} (1-\nu)\cdot(1-\nu) & (1-\nu)\cdot\nu/2 & (1-\nu)\cdot\nu/2 & \nu/2\cdot(1-\nu) & \nu/2\cdot\nu/2 & \nu/2\cdot\nu/2 & \nu/2\cdot(1-\nu) & \nu/2\cdot\nu/2 & \nu/2\cdot\nu/2 \\ (1-\nu)\cdot(1-\alpha) & (1-\nu)\cdot\alpha & (1-\nu)\cdot 0 & \nu/2\cdot(1-\alpha) & \nu/2\cdot\alpha & \nu/2\cdot 0 & \nu/2\cdot(1-\alpha) & \nu/2\cdot\alpha & \nu/2\cdot 0 \\ (1-\nu)\cdot(1-\alpha) & (1-\nu)\cdot 0 & (1-\nu)\cdot\alpha & \nu/2\cdot(1-\alpha) & \nu/2\cdot 0 & \nu/2\cdot\alpha & \nu/2\cdot(1-\alpha) & \nu/2\cdot 0 & \nu/2\cdot\alpha \\ (1-\alpha)\cdot(1-\nu) & (1-\alpha)\cdot\nu/2 & (1-\alpha)\cdot\nu/2 & \alpha\cdot(1-\nu) & \alpha\cdot\nu/2 & \alpha\cdot\nu/2 & 0\cdot(1-\nu) & 0\cdot\nu/2 & 0\cdot\nu/2 \\ (1-\alpha)\cdot(1-\alpha) & (1-\alpha)\cdot\alpha & (1-\alpha)\cdot 0 & \alpha\cdot(1-\alpha) & \alpha\cdot\alpha & \alpha\cdot 0 & 0\cdot(1-\alpha) & 0\cdot\alpha & 0\cdot 0 \\ (1-\alpha)\cdot(1-\alpha) & (1-\alpha)\cdot 0 & (1-\alpha)\cdot\alpha & \alpha\cdot(1-\alpha) & \alpha\cdot 0 & \alpha\cdot\alpha & 0\cdot(1-\alpha) & 0\cdot 0 & 0\cdot\alpha \\ (1-\alpha)\cdot(1-\nu) & (1-\alpha)\cdot\nu/2 & (1-\alpha)\cdot\nu/2 & 0\cdot(1-\nu) & 0\cdot\nu/2 & 0\cdot\nu/2 & \alpha\cdot(1-\nu) & \alpha\cdot\nu/2 & \alpha\cdot\nu/2 \\ (1-\alpha)\cdot(1-\alpha) & (1-\alpha)\cdot\alpha & (1-\alpha)\cdot 0 & 0\cdot(1-\alpha) & 0\cdot\alpha & 0\cdot 0 & \alpha\cdot(1-\alpha) & \alpha\cdot\alpha & \alpha\cdot 0 \\ (1-\alpha)\cdot(1-\alpha) & (1-\alpha)\cdot 0 & (1-\alpha)\cdot\alpha & 0\cdot(1-\alpha) & 0\cdot 0 & 0\cdot\alpha & \alpha\cdot(1-\alpha) & \alpha\cdot 0 & \alpha\cdot\alpha \end{bmatrix} \end{array} \begin{array}{c} 0000 \\ 0010 \\ 0001 \\ 1000 \\ 1010 \\ 1001 \\ 0100 \\ 0110 \\ 0101 \end{array}.$$

The alleles at the four homologous sites some nucleotide position t are assumed to derive from the same allele corresponding to the ancestral state of the sequences at the time of gene duplication. Each observation consists of four alleles corresponding to species 1 gene 1, species 1 gene 2, species 2 gene 1 and species 2 gene 2. The observation (observed state) space is

$V = \{AAAA, AAAG, AAAC, \ldots, TTTT\}$ with size $|V| = 256$ ($= 4^4$). The emission matrix B is a 256 (observations) by 9 (states) matrix. The time between duplication and the present is split into three parts: (1) from duplication to speciation, with substitution probability r_{0q} during this time; (2) from speciation to NAGC, with substitution probability r_{1q}; (3) from NAGC to the present, with substitution probability r_{2q}. We consider all of the possible evolutionary paths that could lead to the observed state. For example, the set of paths w for the observation AACC, $w \in \{w_{\rightarrow AACC}\}$ includes a path starting from an ancestral state A, followed by gene duplication (AA), speciation (AAAA), point substitution (AAAC) and NAGC (AACC), a path starting from the ancestral state C, followed by gene duplication (CC), speciation (CCCC), point substitution (ACCC) and NAGC (AACC), a path starting from an ancestral nucleotide C, followed by gene duplication (CC), speciation (CCCC), point substitution (ACCC), and point substitution again (AACC), and more.

We use an Expectation Maximization (EM) algorithm [3] implemented in the R package *Hmm.discnp* [4] to estimate the parameters Θ.

E-step. We define, $\xi_{y,t}(i,j)$ as the probability of nucleotide t of intron y being in state i and nucleotide $t+1$ being in state j, given the observed sequence O and model parameters Θ. The probability of nucleotide t in intron y being in state i given the parameters and the observations is

$$\gamma_{y,t}(i) = P(s_{q,t} = i|O, \Theta) = \sum_{j=1}^{|V|} \xi_{y,t}(i,j).$$

In the E-step we compute $\xi_{q,t}(i,j)_{i,j}$ and $\gamma_{y,t}(i)_i$ to derive the following key summary statistics:

$$\xi(i,j) = \sum_{q=1}^{Q} \sum_{y \in Y_q} \sum_{t=1}^{l_y - 1} \xi_{y,t}(i,j)$$

is the expected number of transitions from state i to state j given the observed sequence O and Θ, and

$$\gamma(i) = \sum_{q=1}^{Q} \sum_{y \in Y_q} \sum_{t=1}^{l_y} \gamma_{y,t}(i),$$

is the expected number of nucleotides in state i given the observed sequence O. We use the shorthand

$$\xi(c,u) = \xi(10,00) + \xi(01,00).$$

for the expected number of transitions from the converted to the unconverted state and

$$\xi(u,c) = \xi(00,10) + \xi_t(00,01).$$

for the expected number of transitions from the unconverted to the converted state. Similarly, the expected number of nucleotides in the converted state is

$$\gamma(c) = \gamma(10) + \gamma(01),$$

and the expected number of nucleotides in the unconverted state is

$$\gamma(u) = \gamma_t(00).$$

M-step. In each iteration of the EM algorithm, we update the model parameters Θ_{st+1} based on the current model parameters Θ_{st+1}. The global parameters setting the transition matrix are:

$$\pi^{st+1} := \frac{\sum_{q=1}^{Q} \sum_{y \in Y_q} \gamma_{y,1}}{\sum_{q=1}^{Q} |Y_q|},$$

$$\nu^{st+1} := \frac{\xi(u,c)}{\gamma(u)},$$

$$\alpha^{st+1} := 1 - \frac{\xi(c,u)}{\gamma(c)}.$$

The updated gene-specific parameters are:

$$r_{0q}^{st+1} := \frac{\sum_{y \in Y_q} \sum_{t=1}^{l_y - 1} \sum_{j=1}^{N} \sum_{w \in \{w_{\to O_t}\}} \gamma_{y,t}(j) P(w | S = j, \Theta^{st}) D_0(w)}{2L},$$

$$r_{1q}^{st+1} := \frac{\sum_{y \in Y_q} \sum_{t=1}^{l_y-1} \sum_{j=1}^{N} \sum_{w \in \{w_{\rightarrow O_t}\}} \gamma_{y,t}(j) P(w|S=j,\Theta^{st}) D_1(w)}{4L},$$

and

$$r_{2q}^{st+1} := \frac{\sum_{y \in Y_q} \sum_{t=1}^{l_y-1} \sum_{j=1}^{N} \sum_{w \in \{w_{\rightarrow O_t}\}} \gamma_{y,t}(j) P(w|S=j,\Theta^{st}) D_2(w)}{4L},$$

where $P(w|S=j,\Theta^{st})$ is the probability of the evolutionary path w given hidden state j, and parameters Θ^{st}, $D_0(w) \in \{0,1,2\}$ is the number of changed nucleotides from the time of duplication to the time of speciation in the path w, $D_1(w) \in \{0,1,2,3,4\}$ is the number of changed nucleotides from the time of speciation to the time of conversion and $D_2(w) \in \{0,1,2,3,4\}$ is the number of changed nucleotides from the time of conversion in w.

The criterion of convergence for the EM algorithm is set to be

$$|\frac{log(P(O|\Theta_{st+1})) - log(P(O|\Theta_{st}))}{log(P(O|\Theta_{st}))}| < 10^{-5}.$$

1.2.2 Stochastic simulations of duplicates' sequences evolution

We developed stochastic, discrete-generation simulations of the divergence dynamics of a pair of duplicate genes. These simulations were used to evaluate the performance of the HMM (**section 1.2.1**) and to estimate GC bias in NAGC (**section 1.3**). Simulations begin with a duplication event of a $10,000bp$ genes with alleles A,T,C or G. The initial sequence is drawn according to a multinomial distribution with probability 0.3 for weak nucleotides (A and T) and probability 0.2 for strong nucleotides (G and C). The two gene duplicates then evolve along a tree with fixed branch lengths (**Table S4**). Unless otherwise specified, point mutations occur at a rate of 1.55×10^{-8} per bp per generations and NAGC to point mutation rate ratio is set to be 20.1—the same as estimated in the two-site model. Not all NAGC events that are "suggested" (at the rate of NAGC)

are "accepted"; this depends on the NAGC regime: in the global threshold regime, the event occurs if the paralog divergence does not exceed 4%; in the local threshold model, NAGC occurs only if the $100bp$ upstream of the (uniformly-drawn) initiation site are identical in the two paralogs; in the continuous NAGC regime, all suggested NAGC events are accepted. If a NAGC event is accepted, the tract length is drawn from a Geometric distribution with a mean of 250bp. Point mutations have a strong to weak bias: conditional on a mutation in a weak nucleotide, it will mutate to one of the strong nucleotides with probability 1/6 each—whereas a mutation in a strong nucleotide would lead to one of the weak nucleotide with probability 1/3 each. The two paralogs are equally likely to be drawn as a template gene (unless otherwise specified, as in **section 1.3.2**).

1.2.3 Performance of the HMM and comparison with GENECONV

To evaluate the performance of our HMM, we used 100 iterations of the stochastic simulations described in **Section 1.2.2**. We focused on evaluating the preformance on recent duplicates, by simulating the evolution of duplications occurring at the time of the human-macaque split, and examining the resultant sequences in the two paralogs in human and chimpanzee. We assumed that the global threshold model with value 0.04. We further compared the performance on simulations with *GENECONV* [5], a popular software for the detection of converted tracts. We used the following *Unix* command line to run *GENECONV* on our simulation results:

$$geneconv < simulation\_output\_filename > /w123\ /a\ /b2\ -nolog\ -Skip\_indels$$

We computed a confusion matrix for each of the two methods by labeling the state each nucleotide in each of the iterations. A nucleotide is labeled as converted in a species if and only if it was converted at any point during the evolution of the paralogs on the corresponding species' lineage (starting from the duplication event).

Our method substantially outperforms *GENECONV* in overall accuracy and in sensitivity, but performs less well than *GENECONV* in terms of specificity (**Table S2**). As noted by Mansai and Innan [6] in their systematic comparison of methods to detect NAGC, methods based on local

deviations from the true (null or unconverted) gene tree—like our HMM—increase in power with NAGC rate, while the power of model-free methods—like *GENECONV*—often diminishes when NAGC is frequent. *GENECONV* tends to perform best when a small fraction of the sequence is converted. Since our simulations were generated with the high rate inferred using the two-site model, our comparison agrees with Mansai and Innan's conclusions—despite our use of a different simulation scheme and different evaluation metrics.

Our HMM has the appeal of being rooted in a population-genetic model, and it performs well in the parameter regime most relevant for primate evolution. However, it may fare less well on new data generated with different parameters. We thus reiterate Mansai and Innan's recommendation to use multiple detection methods in conjunction when evaluating the footprints of NAGC [7].

	GENECONV	Our HMM
Sensitivity	18.5%	72.4%
Specificity	98.9%	75.4%
Accuracy	42.3%	73.3%

Table S2: Performance of GENECONV and our HMM on simulated data. The confusion table used to compute the three metrics was computed using each nucleotide in the sequence as a data point.

1.2.4 Performance of HMM on different input species

The choice of the two input species for the HMM can affect the inference of converted tracts because of the distinct evolutionary histories and, potentially, because of limitations of the HMM or quality of the sequence assemblies and alignments. We evaluated the dependence on the input species using the two following comparisons:

macaque and human vs. macaque and chimpanzee: The same 7 gene families are identified as having a converted tract within them. 26/28 (92.8%) of the tracts are at least partially overlapping. Out of 12703 nucleotides inferred as converted in either of the inputs, 12513 nucleotides (98.5%) were inferred as converted in both. This agreement is encouraging but somewhat expected

because of the highly similar evolutionary histories and resulting sequences of human and chimpanzee. In the analysis in the main text we did not include the macaque and chimpanzee input—but only pairs of input species where one of the species is human.

macaque and human vs. chimpanzee and human: only one gene family was inferred to have recent conversions in both sets of inputs. Within this gene family, there is good agreement across these inputs and others on the positions of the converted tracts (**Fig. S1**). All other tracts identified using macaque and human were not identified with chimpanzee and human. The depletion of converted tracts found using chimpanzee and human is likely due to the largely shared evolutionary history and resulting sequence of human and chimpanzee since the split from macaque.

1.3 Estimating GC bias in NAGC

1.3.1 Estimating GC bias from real data

To test whether NAGC is GC-biased, we used sites that are identical across paralogous genes (within the same species) but different between the two species (purple sites in **Fig. 1C**) that were identified as converted using our HMM. The alleles in the unconverted species provide information of the ancestral state of that site. For example, if a site is G in both genes in the species in which NAGC occurred, and is A in the other species, then we estimate that the site experienced a weak (w)→strong (s) conversion. We observed that 61% (51 out of 83) of w↔s substitutions are in the w→s direction.

To test whether this proportion deviates from that expected with no GC biased NAGC, we next develop an estimator and a test for GC bias using the counts of purple sites in converted tracts. Denote by $Z^{(i)}_{w\to s}$ ($Z^{(i)}_{s\to w}$) be an indicator (dummy) random variable equaling 1 if nucleotide i is a purple site with a weak to strong (strong to weak) mutation, and 0 otherwise. Then

$$E\big[Z^{(i)}_{w\to s}\big] = (1 - s_{NAGC}) \cdot \tilde{c} \cdot u_{w\to s} \cdot (0.5 + \delta)$$

and

$$E\big[Z^{(i)}_{s\to w}\big] = s_{NAGC} \cdot \tilde{c} \cdot u_{s\to w} \cdot (0.5 - \delta),$$

where s_{NAGC} is the GC content in converted tracts, $\tilde{c}$ is the probability that the nucleotide was converted since the split of the two input species, $u_{w\to s}$ ($u_{s\to w}$) is the probability that a weak (strong) nucleotide is substituted to a strong (weak) nucleotide before the conversion occurred and δ is the GC bias, meaning that $0.5+\delta$ is the probability a weak/strong heteroduplex is resolved in favor of the strong allele. We can use our previously derived estimates of s_{NAGC} and s_{Null} (the GC content in non-converted tracts); since the sample sizes for these estimates are very large, we treat these as fixed parameters.

To estimate the substitution rates $u_{w\to s}$ and $u_{s\to w}$ we looked at unconverted sites where only one out of the four genes carries an allele different from the rest; the most parsimonious scenario for this pattern is that only one substitution (arising from a point mutation) occurred. 53.0% (2390 out of 4513) of A/T↔G/C sites are A/T→G/C. We can estimate the ratio of substitution rates by

$$\frac{u_{w\to s}}{u_{s\to w}} = \frac{(1-s_{Null})}{s_{Null}} \cdot \frac{Y_{w\to s}}{Y_{s\to w}},$$

where $Y_{w\to s}$ ($Y_{s\to w}$) is the number of substitutions—in one of the four sequences, as outlined above—from a weak (strong) nucleotide to a strong (weak) nucleotide. Again, we treat this ratio as fixed because of the large sample size used for estimation. Finally, we define the following statistic:

$$\hat{\delta} = \frac{\frac{1}{(1-s_{NAGC})\cdot u_{w\to s}} \sum_i Z^{(i)}_{w\to s}}{\frac{1}{(1-s_{NAGC})\cdot u_{w\to s}} \sum_i Z^{(i)}_{w\to s} + \frac{1}{s_{NAGC}\cdot u_{s\to w}} \sum_i Z^{(i)}_{s\to w}} - \frac{1}{2}.$$

Note that since–assuming that the denominator is never zero for a large sample–the expectation of a ratio asymptotes to the ratio of expectations, $\hat{\delta}$ is a consistent estimator of δ, i.e.

$$E\left[\hat{\delta}\right] \to \delta$$

for large samples. In the next section we evaluate the accuracy of $\hat{\delta}$ in simulated data with a sample size comparable to our data, and various models of GC bias.

We tested the null hypothesis that w/s heteroduplex are symmetrically repaired in NAGC using the exact binomial test with the null hypothesis

$$H_0 : \delta = 0,$$

and reject the null (exact Binomial test $p = 7.5 \cdot 10^{-7}$, **Fig. 2D**). We conclude that NAGC is GC-biased and turn to estimate the extent of the bias using Monte Carlo simulations.

1.3.2 GC bias simulations

To estimate the extent of GC bias using the $\hat{\delta}$ statistic we described in **Section 1.3**, we used a Monte Carlo approach: we added a GC bias component to the choice of template in our simulations and identified the GC bias value δ that agrees with our empiric $\hat{\delta}$ value.

There is evidence for multiple error correction mechanisms by which G/C alleles are preferentially chosen as templates in NAGC [8, 9]. We considered three different models for the choice of template following a random draw of the initiation site and the tract length:

Tract model: the gene with higher GC content in the tract is chosen as the template with probability $\frac{1}{2} + \delta$.

First mismatch model: if divergent sites exist in the tract, and the first (most upstream) is a G/C in one of the genes and A/T in the other, then the gene with the G/C allele serves as the template with probability $\frac{1}{2} + \delta$. Otherwise, the genes have equal probability to be the template.

Base excision model: in every divergent site, the template is chosen independently. If the allele is G/C in one gene and A/T in the other, then the allele with the strong allele serves as the template at the site with probability $\frac{1}{2} + \delta$. Otherwise, the alleles have equal probability to be the template.

Note that although we denote GC bias as δ in all three models, they are not straightforwardly comparable across models. For a given value of δ, we expect that the strongest bias for the base excision model, followed by the tract model and then the first mismatch model. We therefore evaluated the relationship between our HMM-based estimator $\hat{\delta}$ and the underlying δ in each of the GC bias models. We performed 200 simulations for each of the three GC bias models and $\delta \in \{0.01 \cdot i; i \in \{0, 1, ..., 10\} \bigcup \{0.05 \cdot i; i \in \{3, ..., 10\}\}$. We applied our HMM to infer converted tracts in the simulated sequences and compute $\hat{\delta}$. The number of purple mutations in inferred

converted regions from 200 simulations was roughly double the number of purple mutations in inferred conversions in the real data.

For all three models, our $\hat{\delta}$ exhibits some bias with respect to the underlying δ: a large bias for the first mismatch and tract model and a small bias for the base excision model. Further, the base excision model is the only model that attains the observed value of $\hat{\delta}$ in the real data (0.173) at $\delta = 0.21$ (**Fig. S3A**).

Interestingly, all three GC-bias models predict that GC-biased NAGC hardly changes GC content during primate evolution (**Fig. S3B**). We hypothesize that the observed difference in GC content between converted and unconverted regions (**Fig. 2C**) is due to the correlation of AGC rates with NAGC rates along the genome, as the genomic factors that drive AGC and NAGC likely overlap substantially.

1.3.3 Comparison with Assis and Kondrashov 2011

Assis and Kondrashov [10] tested whether GC bias exists in NAGC in the primate and in the *Drosophila* clades. They used a parsimony-based approach on single nucleotide divergence patterns and conclude that there is no evidence for GC-bias. Since Assis and Kondrashov's conclusion seemingly contradics our conclusion that NAGC is GC biased in primates, we discuss the potential reasons for disagreement here.

Assis and Kondrashov's approach examines pairs of paralogs shared by two sister species and an outgroup. They identify single nucleotide divergence patterns that are parsimoniously explained with either NAGC or point mutations (**Table S3** patterns A,E) and subtract the counts of patterns consistent with point mutations alone (**Table S3** patterns B,D). To control for nucleotide composition, they divide this difference by the nucleotide content of the presumed-ancestral allele in the outgroup. They then compare this quotient for weak $\rightarrow$ strong and strong $\rightarrow$ weak substitutions using the exact Binomial test.

While intuitive and simple to apply, Assis and Kondrashov's approach is susceptible to some

Pattern	Outgroup		Species 1		Species 2		Parsimonious mechanism
	Gene 1	Gene 2	Gene 1	Gene 2	Gene 1	Gene 2	
A	a	b	a	b	a	a	i. NAGC with gene 2 as the template ii. mutation in gene 2
B	a	b	a	b	a	c	mutation in gene 2
C	a	b	a	b	c	c	i. mutation followed by NAGC ii. mutation in gene 2
D	a	b	a	b	c	d	mutation in both genes
E	a	a	a	a	c	c	i. mutation followed by NAGC ii. mutation in both genes
F	a	a	a	a	a	c	mutation in gene 2
G	a	a	a	a	d	c	mutation in both genes

Table S3: Classification of single nucleotide divergence patterns in Assis and Kondrashov's parsimony-based approach. a,b,c and d represent distinct alleles.

subtle biases. The expected number of substitutions through point mutations and NAGC, and the sequence content (denominator) should be computed separately for each pattern to take into account the asymmetry in mutation rates (e.g. a strong $\rightarrow$ weak mutation is about two fold more likely than a weak $\rightarrow$ strong mutation) [11]. In Kondrashov and Assis' analysis, all of the patterns are clumped together, and the division by the sequence content in the ancestor is sometimes inappropriate. For example, in cases of pattern A that they consider, the presumed ancestral state is one strong and one weak allele in the two paralogs; therefore a division by the content of the presumed template is not required. More importantly, we show below that even after modifying the method to overcome these biases, the method is underpowered to detect GC bias in NAGC.

To overcome the potential biases in the original method, we modified Assis and Kondrashov's method and defined two similar statistics that are based on single nucleotide divergence patterns in two sister species and an outgroup; we also limit the scope of these statistics to sites at which one sister species and the outgroup have identical alleles—while the other sister species has a different allele in at least one of paralogs. We identify cases that involve weak (w) $\leftrightarrow$ strong (s) substitutions and stratify them by the direction of the substitution. Denote by n_x^d the number of cases of pattern $x \in \{A, B, C, D, E, F, G\}$ and direction $d \in \{s \rightarrow w, w \rightarrow s\}$. Our modified statistics are defined as

$$R_1 = \frac{n_A^{w\rightarrow s} - n_B^{w\rightarrow s}}{(n_A^{w\rightarrow s} - n_B^{w\rightarrow s}) + (n_A^{s\rightarrow w} - n_B^{s\rightarrow w})},$$

and

$$R_2 = \frac{\frac{n_E^{w\to s} - n_G^{w\to s}}{n_F^{w\to s}}}{\left(\frac{n_E^{w\to s} - n_G^{w\to s}}{n_F^{w\to s}}\right) + \left(\frac{n_E^{s\to w} - n_G^{s\to w}}{n_F^{s\to w}}\right)}.$$

R_1 estimates the fraction of $s \leftrightarrow w$ NAGC events acting on a pre-existing w/s difference between paralogs in which the strong allele was the template. The number of pattern A substitutions expected under point mutations alone is exactly $n_B^{w\to s}$—so the difference $n_A^{w\to s} - n_B^{w\to s}$ estimates the number of $w \to s$ NAGC events. $R_1 > 0.5$ suggests a GC bias towards strong alleles. Similarly, R_2 estimates the fraction of $s \leftrightarrow w$ NAGC events that were preceded by a point substitution in which the strong allele was the template. The rationale of the subtraction is the same as for R_1. The division by $n_F^{w\to s}$ or $n_F^{s\to w}$ is required here because point substitution is a prerequisite for patterns E and G, and these occur at different rates in $w \to s$ and $s \to w$ substitutions; pattern F is presumed to arise from one point mutation from the same background, and is therefore the appropriate normalization here.

We first evaluated the sensitivity of the modified Assis and Kondrashov test statistics to GC bias. We used 200 duplicate sequence evolution simulations for each of the combinations of GC bias models and δ values as specified in **Section 1.3.2**. As expected, both R_1 and R_2 roughly equal 0.5 when there is no GC bias and increase with the magnitude of GC bias. However, both statistics are noisy even with $50,000$ sites with informative patterns (**Table S3**)—a sample size comparable to that of Assis and Kondrashov's original study. Further, R_1 and R_2 often fall below 0.5 even in simulations with $\delta > 0$.

We next evaluated R_1 on a subset of the real intronic data. Namely, we used all available sequence from gene families that had both paralogs in both human, chimpanzee and orangutan. There were no pattern E or pattern G cases, so we were not able to examine R_2. To our surprise, the number of pattern A $w \to s$ substitutions was smaller than the number of pattern B $w \to s$ substitutions, leading to a negative R_1 value. R_1 is expected to be a ratio, bounded between 0 and 1. It is assumed that pattern A occurs more often than pattern B as pattern A arises through

both NAGC and point mutations while pattern B arises only through point mutations that occur at the same rate as point mutations leading to pattern A. The negative R_1 value could be due to a combination of mutation rate variation and the small sample size at our disposal—only 1211 sites with informative patterns. In both the Assis and Kondrashov modeling approach and in our simulations, it is assumed that the only relevant mutation rate variation is the difference between $s \to w$ and $w \to s$ mutations. However, substantial mutation rate variation exists in primates [11, 12, 13]. These results suggest that—despite its intuitive appeal—Assis and Kondrashov's approach could have limited power to test for the presence of NAGC, and is too sensitive to mutation rate variation with the sample size of this study.

1.4 Two-site model

1.4.1 Per-generation transition matrix

In the main text, we outline the derivation of per-generation transition probabilities between states in the two-site model. The corresponding transition probability matrix is:

$$\mathbf{P} = \begin{array}{c} \\ 0000 \\ 0001 \\ 0010 \\ 0011 \\ 0100 \\ 0101 \\ 0110 \\ 0111 \\ 1000 \\ 1001 \\ 1010 \\ 1011 \\ 1100 \\ 1101 \\ 1110 \\ 1111 \end{array} \begin{array}{c} \begin{array}{cccccccccccccccc} 0000 & 0001 & 0010 & 0011 & 0100 & 0101 & 0110 & 0111 & 1000 & 1001 & 1010 & 1011 & 1100 & 1101 & 1110 & 1111 \end{array} \\ \begin{pmatrix} 1-r_1 & \mu & \mu & 0 & \mu & 0 & 0 & 0 & \mu & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \mu/3+c & 1-r_2 & 0 & \mu/3+c & 0 & \mu & 0 & 0 & 0 & \mu & 0 & 0 & 0 & 0 & 0 & 0 \\ \mu/3+c & 0 & 1-r_3 & \mu/3+c & 0 & 0 & \mu & 0 & 0 & 0 & \mu & 0 & 0 & 0 & 0 & 0 \\ 0 & \mu & \mu & 1-r_4 & 0 & 0 & 0 & \mu & 0 & 0 & 0 & \mu & 0 & 0 & 0 & 0 \\ \mu/3+c & 0 & 0 & 0 & 1-r_5 & \mu & \mu & 0 & 0 & 0 & 0 & 0 & \mu/3+c & 0 & 0 & 0 \\ cg(d) & \mu/3+c(1-g(d)) & 0 & 0 & \mu/3+c(1-g(d)) & 1-r_6 & 0 & \mu/3+c(1-g(d)) & 0 & 0 & 0 & 0 & 0 & \mu/3+c(1-g(d)) & 0 & cg(d) \\ 0 & 0 & \mu/3+c(1-g(d)) & cg(d) & \mu/3+c(1-g(d)) & 0 & 1-r_7 & \mu/3+c(1-g(d)) & 0 & 0 & 0 & 0 & cg(d) & 0 & \mu/3+c(1-g(d)) & 0 \\ 0 & 0 & 0 & \mu/3+c & 0 & \mu & \mu & 1-r_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu/3+c \\ \mu/3+c & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1-r_9 & \mu & \mu & 0 & \mu/3+c & 0 & 0 & 0 \\ 0 & \mu/3+c(1-g(d)) & 0 & cg(d) & 0 & 0 & 0 & 0 & \mu/3+c(1-g(d)) & 1-r_{10} & 0 & \mu/3+c(1-g(d)) & cg(d) & \mu/3+c(1-g(d)) & 0 & 0 \\ cg(d) & 0 & \mu/3+c(1-g(d)) & 0 & 0 & 0 & 0 & 0 & \mu/3+c(1-g(d)) & 0 & 1-r_{11} & \mu/3+c(1-g(d)) & 0 & 0 & \mu/3+c(1-g(d)) & cg(d) \\ 0 & 0 & 0 & \mu/3+c & 0 & 0 & 0 & 0 & 0 & \mu & \mu & 1-r_{12} & 0 & 0 & 0 & \mu/3+c \\ 0 & 0 & 0 & 0 & \mu & 0 & 0 & 0 & \mu & 0 & 0 & 0 & 1-r_{13} & \mu & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu & 0 & 0 & 0 & \mu & 0 & 0 & \mu/3+c & 1-r_{14} & 0 & \mu/3+c \\ 0 & 0 & 0 & 0 & 0 & 0 & \mu & 0 & 0 & 0 & \mu & 0 & \mu/3+c & 0 & 1-r_{15} & \mu/3+c \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu & 0 & 0 & 0 & \mu & 0 & \mu & \mu & 1-r_{16} \end{pmatrix} \end{array},$$

where

$$r_i = \sum_{j \neq i} \mathbf{P}_{ij}.$$

1.4.2 Small effect of recombination on NAGC fixation probabilities

While elsewhere we assume that mutations (both point mutations and NAGC at a single site) fix at a rate equal to the mutation rate, we pause to examine this assumption for the case of a NAGC mutation including both focal sites—because the two derived alleles might decouple before one of them fixes. The probability of fixation in both sites conditional on fixation in one of them is

$$g(d) = g_{init}(d)q(d),$$

where $q(d)$ is the probability that the second derived allele remains linked during the fixation at the first site. We make a few simplifying assumptions in evaluating $q(d)$: The fixation time is assumed to be $4N_e$ generations where N_e is the (constant) effective population size. If at least one recombination event occurs, we approximate the probability of decoupling by the mean allele frequency of the first allele during fixation, $\frac{1}{2}$. Denoting the per bp per generation recombination rate by r, we get:

$$q(d) = 1 - \frac{1}{2}[1 - (1 - r)^{4N_e d}],$$

and

$$g(d) = (1 - \frac{1}{\lambda})^d \frac{1 - (1 - r)^{4N_e d}}{2}.$$

Plugging in $r = 10^{-8}$ [14] and $N_e = 10^4$, we found that the probability of decoupling is high only for distances d where g_{init} is already very small. Consequently, difference between g_{init} and g are small throughout (**Fig. S4**). We therefore use the approximation

$$g \approx g_{init}$$

in our implementation of this model.

1.4.3 Estimation in the two-site model

Our model describes the evolution of two sites in paralogs along primate evolution. Each of the nodes in the primate tree (Fig. 3B) consists of observed states—corresponding to primate references that include all four orthologous nucleotides—and hidden nodes corresponding to the state

in most recent common ancestors (MRCAs) of these species. To fully determine the likelihood we must also set a prior on the state in the MRCA of all species with an observed state ("data root"). We explain the choice of prior in **Section 1.4.4**.

We compute the full log likelihood for each datum (a set of 4-bit states for 2-5 primates) with transition probability matrices $\mathbf{P}^*_{\mathbf{edge\ ij}}$. To do so in a computationally efficient way, we apply Felsenstein's pruning algorithm [15]. We then compute the composite likelihood by summing log likelihoods over all of the data (all pairs of sites in each of the introns). We then evaluate composite likelihoods over a grid of values—the cross product of mean tract lengths $\lambda \in \{10^{z/5}; z \in \{5, 6, ..., 20\}\}$ and rates $c \in \{0\} \cup \{10^{-k/10}; k \in \{50, 51, ..., 80\}\}$—and identify the parameter values that maximize the composite likelihood.

1.4.4 Setting a prior on the "data root"

In our two-site model described in the main text, we compute the full likelihood for each datum (a set of observations in two sites in two duplicate genes, across several primates). To compute this likelihood we need a prior on the state at what we refer to as the "data root", i.e. the internal node corresponding to the MRCA of human and the most distant primate relative of human for which we have two paralogs. Here, we describe how we set this prior.

We use the information that only one ortholog is found in mouse (and possibly some of the primates). Namely, we assume that the duplication occurred on the branch ending at the data root r and take it to be uniformly distributed along this branch (**Fig. S6**). We denote by T_{single} the length of the branch between the mouse node and the duplication event. The prior on the data root is set to be

$$\pi_0' \mathbf{P}_{\mathbf{single}}{}^{T_{single}} \mathbf{P}^{t_{mouse,r} - T_{single}},$$

where $r \in \{2, 3, 4\}$ is the data root internal node (**Fig. S6**),

$$\pi_0 := e_{0000}$$

is set to be the mouse gene state, π_0' denotes the transpose of π_0, and $\mathbf{P_{single}}$ is a transition matrix corresponding to a single gene evolution without gene conversion,

$$
\mathbf{P_{single}} = \begin{array}{c} \\ 0000 \\ 0001 \\ 0010 \\ 0011 \\ 0100 \\ 0101 \\ 0110 \\ 0111 \\ 1000 \\ 1001 \\ 1010 \\ 1011 \\ 1100 \\ 1101 \\ 1110 \\ 1111 \end{array}
\begin{array}{c} \begin{array}{cccccccccccccccc} 0000 & 0001 & 0010 & 0011 & 0100 & 0101 & 0110 & 0111 & 1000 & 1001 & 1010 & 1011 & 1100 & 1101 & 1110 & 1111 \end{array} \\
\left(\begin{array}{cccccccccccccccc}
1-2\mu & 0 & 0 & \mu & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\mu & 0 & 0 & 1-2\mu & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\mu & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1-2\mu & 0 & 0 & \mu \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \mu & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu & 0 & 0 & 1-2\mu
\end{array} \right) \end{array} .
$$

1.4.5 Use of lowly-diverged genes for parameter estimation

In our NAGC parameter estimation procedure, we perform our estimation using introns from genes with $ds \leq 5\%$. We evaluated the sensitivity of our estimate of NAGC rate to the choice of the ds threshold. While different ds thresholds between $0-10\%$ change the point rate estimate somewhat, the rate estimate remains on the order of 10^{-7} conversions per bp per generation (**Fig. S7**).

Although the choice of ds threshold seems to be of limited effect, this strategy may be inappropriate if substantial variation in NAGC rate exists (beyond the variation associated with sequence divergence). The reason is that in such a case, genes experiencing frequent NAGC would be enriched among lowly-diverged genes—leading to an upward-biased rate estimate.

To mimic the scenario of using a ds threshold in the face of NAGC rate heterogeneity, we performed stochastic simulations of genes duplicated at the time of the human-mouse split (**Table S4**) and examined the fraction of genes with sequence divergence below different thresholds for each of the rates. We performed the simulations as described in **section 1.2.2** with the "continuous

NAGC" regime. As expected, lower-rate genes are more likely to be included for estimation than higher-rate genes (**Fig. S8B**).

We next asked whether NAGC rate variation could have driven a large upward bias in our rate estimation (which was limited to $ds \leq 0.05$ genes) through this effect. We thus simulated divergence trajectories resulting from randomly sampled rates and evaluated the average rate of the simulations with divergence below 0.05. For simplicity, we refer to this rate as the "estimated rate", implicitly assuming perfect estimation in order to focus on the effect of the use of a ds threshold rather than on our two-site estimation method. We model the distribution of NAGC rate as a truncated Normal. Namely,

$$C_i = max(0, E_i)$$

where C_i is the NAGC rate for simulation i, and

$$E_i \sim N(c, \sigma^2)$$

for some mode $c > 0$. We find that for moderate rate variation ($\sigma/c \leq 2$), the estimated rate is only slightly upward biased. For $\sigma/c = 10$ and $c = 3.2 \times 10^{-8}$), the estimated rate is 1.2×10^{-7}, comparable with our MLE for the two-site model (**Fig. S8C**). We conclude that the strategy of limiting estimation to lowly-diverged genes incurs a bias in the presence of NAGC rate variation; but the bias is expected to be negligible—as long as the standard deviation of the rate is smaller or comparable to the mean rate.

1.4.6 The effect of physical distance between duplicates

The physical distance between two paralogous sequences has been shown to affect NAGC rate [6, 16]. This decrease is expected because the likelihood of a paralog being used as a template (rather than the homologous sequence) should decrease for sequences that are physically far apart. To test the evidence for this effect with our data, we used the two-site model to estimate MLEs

for each intron separately and matched these estimates with the physical distance between the transcription start sites of the human genes. We found that NAGC rate MLEs and the physical distance are somewhat negatively correlated, but the statistical support for this trend in our data is weak (Spearman $\rho = -0.22, p = 0.10$, **Fig. S9**).

1.5 NAGC slowdown and synonymous sequence divergence

In **Fig. 4** we show predictions for the dynamics of mean neutral sequence divergence between duplicates. We show both theoretical predictions and data for ds between human gene duplicates. Below, we explain how we derive both.

1.5.1 Estimating the duplication time interval for human duplicates

We attained a list of human tandem gene duplicate pairs and their synonymous sequence divergence (ds) from [1]. For each pair, we also considered the sharing of both paralogs in other species, including chimpanzee, gorilla, orangutan, macaque, mouse, opposum and chicken. Specifically, we noted species most distantly-related to humans for which [1] identify orthologs of both human paralogs ("distant sharer", **File S2**). We wished to get an estimate of the age of duplication that is independent of sequence divergence between the human duplicates. We therefore estimated that the duplication occurred on the branch ending at the human-distant sharer split. For example, if the most-distant sharer is macaque, then we estimate that the duplication occurred sometime between the human-mouse split and the human-macaque split. Note that the low quality of genome assemblies can result in unidentified orthologs, which would in turn down-bias the duplication interval estimate. The derived interval estimates are shown as grey lines between estimated split times (see below) in **Fig. 4**.

We approximate split times with divergence times. This leads to an upward estimate of the split time, which is likely substantial for chimpanzee and gorilla but small for the rest of the species. To estimate divergence times, we use sequence divergence in singleton (non-duplicated) genes between each species and humans. For each species i, we take the average ps [17] value

computed for singleton genes. We denote this average by ps_i. We take human-chimpanzee and human-gorilla divergence time estimates from Moorjani et al. ([18]). We then perform simple linear regression with no intercept (forcing the fitted line to go through the origin) regressing $0.5 \cdot ds_{chimpanzee}$ and $0.5 \cdot ds_{gorilla}$ to these divergence times to estimate the synonymous site substitution rate μ'. Note that this substitution rate is different from the intronic mutation rate used in the two-site model. We then plug μ' to estimate the rest of the split times $\{t_i | i \in \{orangutan, macaque, mouse, opposum, chicken\}\}$ by [17]:

$$t_i = \frac{-3/4 \cdot log(1 - 4/3 * ps_i)}{2 \cdot \mu'}.$$

The mean divergence times estimated by this procedure are shown in **Table S4**.

Species	Estimated divergence time (My)
Chimpanzee	12.1
Gorilla	15.1
Orangutan	32.6
Macaque	48.7
Mouse	359.9
Opossum	817.7
Chicken	1269.3

Table S4: Estimated divergence times between human and other species.

1.5.2 Theoretical single-site sequence evolution models

We compute the mean divergence between duplicate sequences under different models of NAGC. We use single-site models to evolve a length-two row vector describing the probability of identity of the two duplicates at a random site. The first entry is the probability that the paralogous sites are identical by state and the second entry is the probability that they are diverged. For each model $j \in \{1, 2, 3, 4\}$, the state $v^{(t)}$ at time $t > 0$ (in years) is

$$v^{(t-1)}\mathbf{A_j},$$

where $v^{(0)} = e'_{00}$.

model 1, no NAGC: In this model, NAGC does not act at all and the evolution follows the Jukes-Cantor mutation model [19],

$$\mathbf{A_1} = \begin{pmatrix} 1-2\mu' & 2\mu' \\ 2\mu'/3 & 1-2\mu'/3 \end{pmatrix},$$

where μ' is set as explained in section **1.5.1**.

model 2, continuous NAGC: In this model, NAGC acts continuously at rate c determined by the ratio of c to μ' inferred from introns in the two-site model,

$$\mathbf{A_2} = \begin{pmatrix} 1-2\mu' & 2\mu' \\ 2c+2\mu'/3 & 1-(2c+2\mu'/3) \end{pmatrix}.$$

model 3, global threshold: In this model, NAGC acts only if the mean sequence divergence is lower than some threshold γ,

$$\mathbf{A_3}(v^{(t-1)}) = \mathbb{1}\{v^{(t-1)} < \gamma\}\mathbf{A_1} + \mathbb{1}\{v^{(t-1)} \geq \gamma\}\mathbf{A_2}.$$

model 4, local threshold: In this model, the evolution is a weighted mean of NAGC acting and not acting, where the weights are set by the probability that a random sequence of m sites are identical between the genes, given the mean sequence evolution $v^{(t-1)}$. This probability, $g(v^{(t-1)})$ is set to be

$$g(v^{(t-1)}) = (v_1^{(t-1)})^m,$$

where $v_1^{(t-1)}$ is the first entry of $v^{(t-1)}$. We set $m = 400$ [20]. The transition matrix in this model is

$$\mathbf{A_4}(v^{(t-1)}) = g(v^{(t-1)})\mathbf{A_1} + (1 - g(v^{(t-1)}))\mathbf{A_2}.$$

In addition to the above-stated parameter values which were used in **Fig. 4**, we investigated the change in expected divergence trajectories for other μ', c, γ and m values. The expected divergence trajectories are insensitive to the point mutation rate (**Fig. S10**) and to local thresholds on the order

of 100bp, that largely resemble the "continuous NAGC" regime (**Fig. S12**). The global threshold trajectory is highly sensitive to the NAGC rate and the global threshold, exhibiting a phase shift between closely tracking the "continuous NAGC" regime and closely tracking the "no NAGC" regime (**Fig. S11**).

2 Supplementary Figures

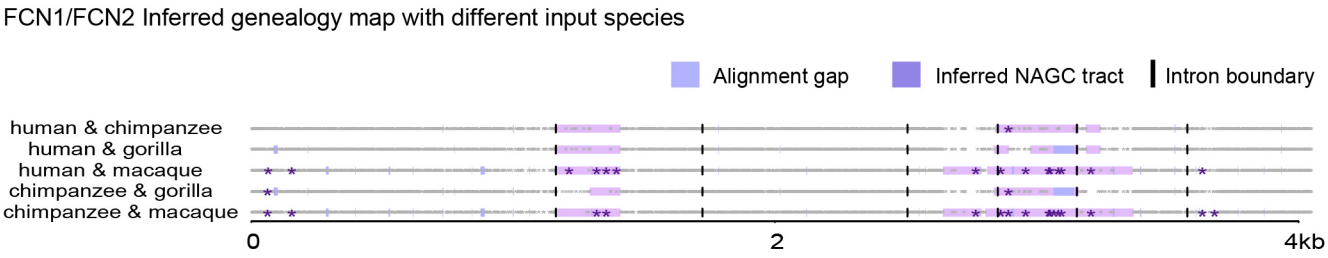

Figure S1: Genealogy map for one gene family, FCN1/FCN2 (null genealogy marked by white, NAGC by purple tracts) inferred using different pairs of input species. Genealogies were inferred based on observed divergence patterns (stars). For simplicity, only the most informative patterns (purple and grey sites, as in **Fig. 1C**) are plotted. In this gene family, the inferred conversions co-localize for all input pairs examined, though the exact position and length of the tract vary slightly. In other gene families, converted tracts were identified using some input species pairs but not with others.

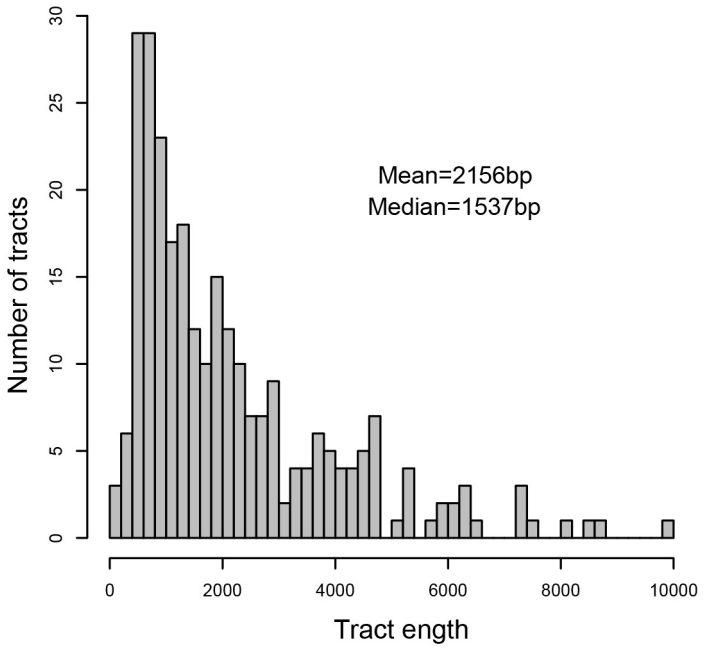

Figure S2: Length distribution of tracts identified by the HMM from simulated sequences. The data are the aggregate of 100 simulations of the evolution of a pair of paralogs of length $10,000$bp in human and chimpanzee, for a duplication that occurred at the time of the human-macaque split. Although the mean tract length used in the simulations was 250bp, the average length of inferred tracts is 2156bp, because many inferred NAGC tracts result from multiple NAGC events occurring in close proximity.

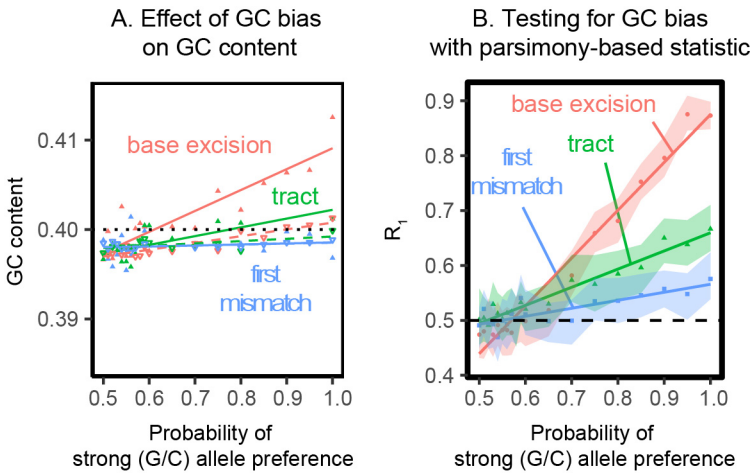

Figure S3: **(A)** During primate evolution, GC biased NAGC alone can only slightly alter nucleotide composition. Triangles show averages of 200 simulation iterations each for paralogous 10,000bp sequences duplicated at the time of the human-macaque split, sampled in human. Filled, top-facing (hollow. bottom-facing) triangles show the content in (outside) regions identified as converted by our HMM. The x-axis shows the probability to prefer GC alleles $(0.5 + \delta)$ which is manifested differently in each of the three models shown. Solid lines show linear fits to filled triangles, while dashed lines show linear fits to hollow triangles. The dotted black line shows the GC content at the begining of the simulations. **(B)** R_1 is a statistic based on parsimony in single nucleotide divergence patterns and does not require the prior identification of converted tracts. Points show R_1 values computed on $50,000$ informative patterns from the same simulations as in (A), and shaded regions show 95% confidence intervals using sampling without replacement from the same simulations. The dashed line at $R_1 = 0.5$ shows the expectation under the null of no GC bias.

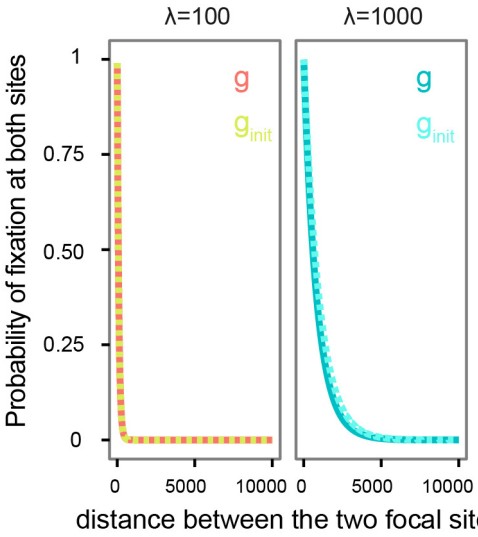

Figure S4: The probability of a NAGC mutation fixing at both sites, conditional on fixation in one of them. Shown is the probability as a function of the distance between focal sites for two mean tract lengths (λ) values. g denotes this probability when accounting for the possibility of decoupling of the sites through recombination, while g_{init} ignores it. Since the differences between the two are very small, we approximate g by g_{init}.

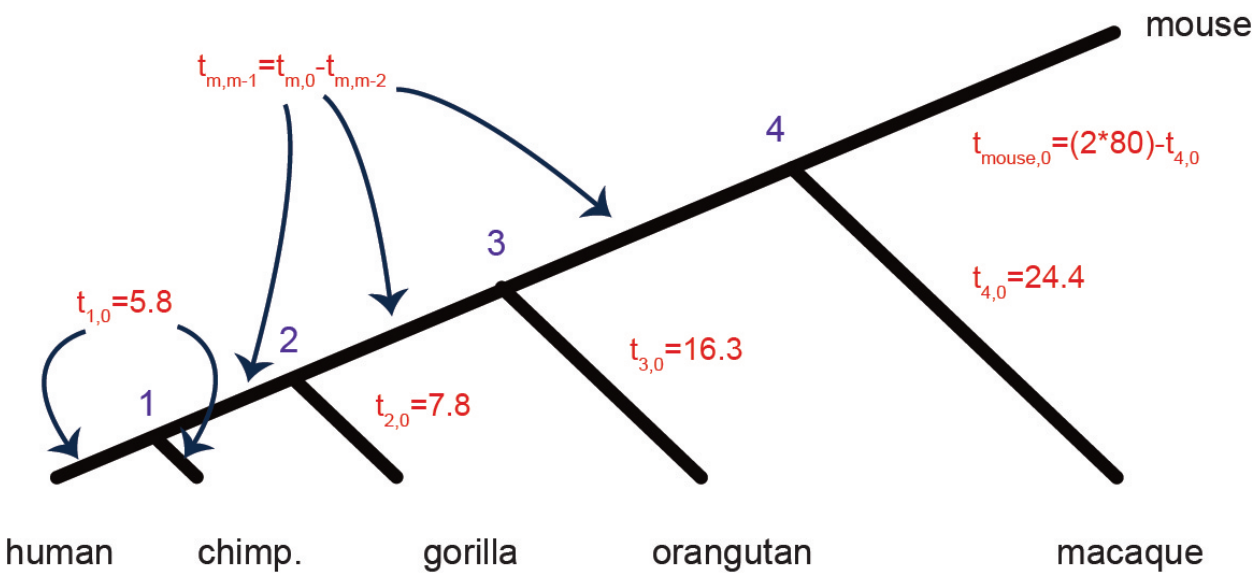

Figure S5: Split times parameterization for the two-site model. Times (in red) are given in millions of years, and translated to generations by dividing by a fixed generation time of 25 years.

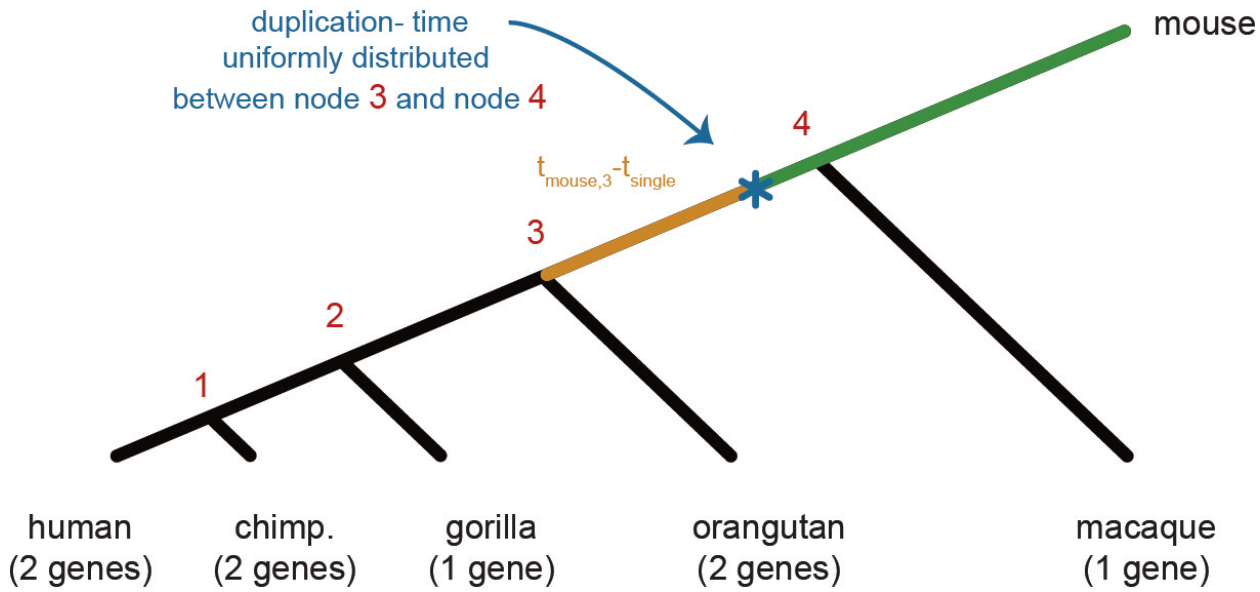

Figure S6: Setting the prior on the data root.

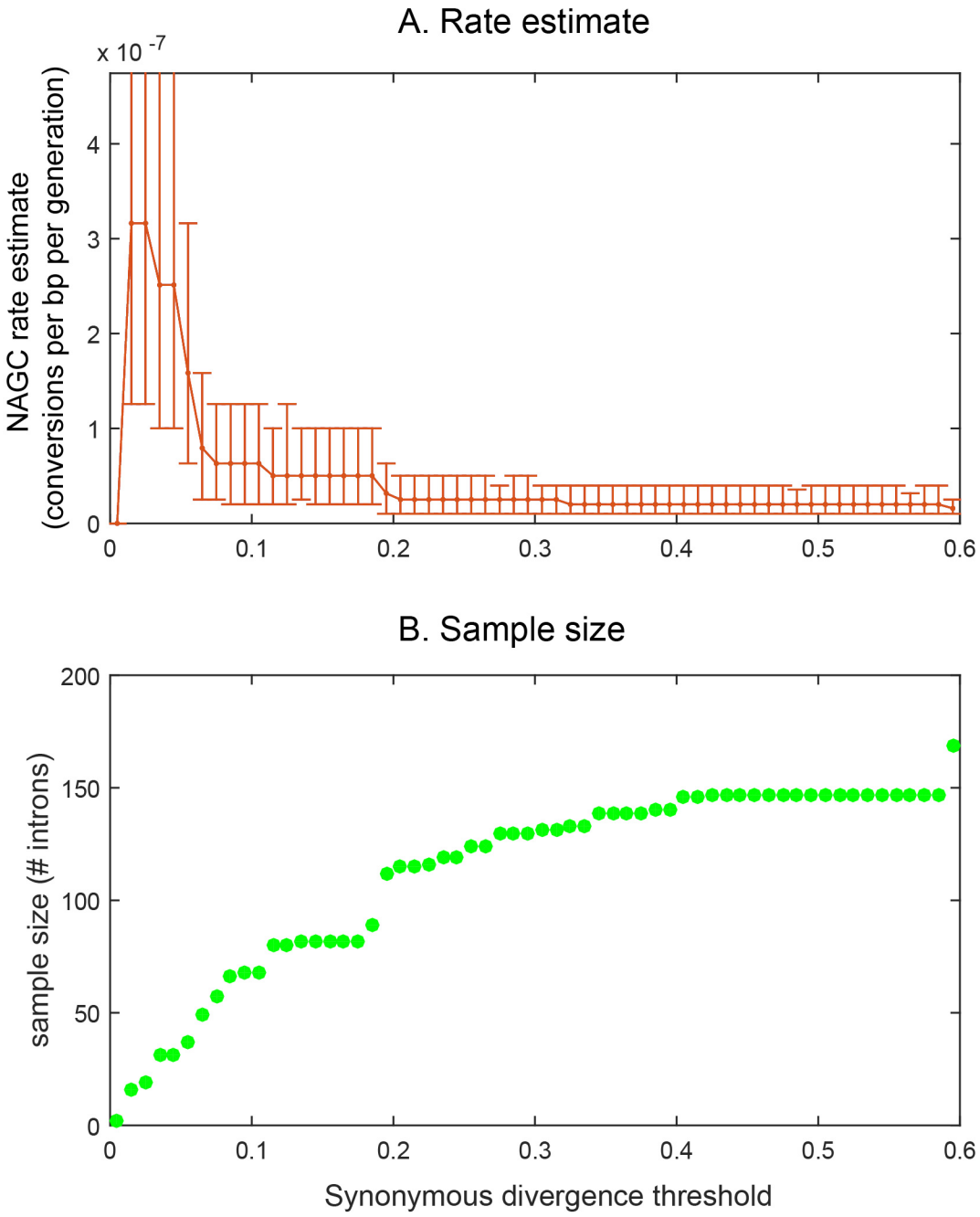

Figure S7: **(A)** The solid line shows NAGC rate estimate across ds upper thresholds (on exons of the same gene). Error bars show 95% non-parametric bootstrap confidence intervals. **(B)** The number of introns that passed all filters and do not exceed each ds upper threshold.

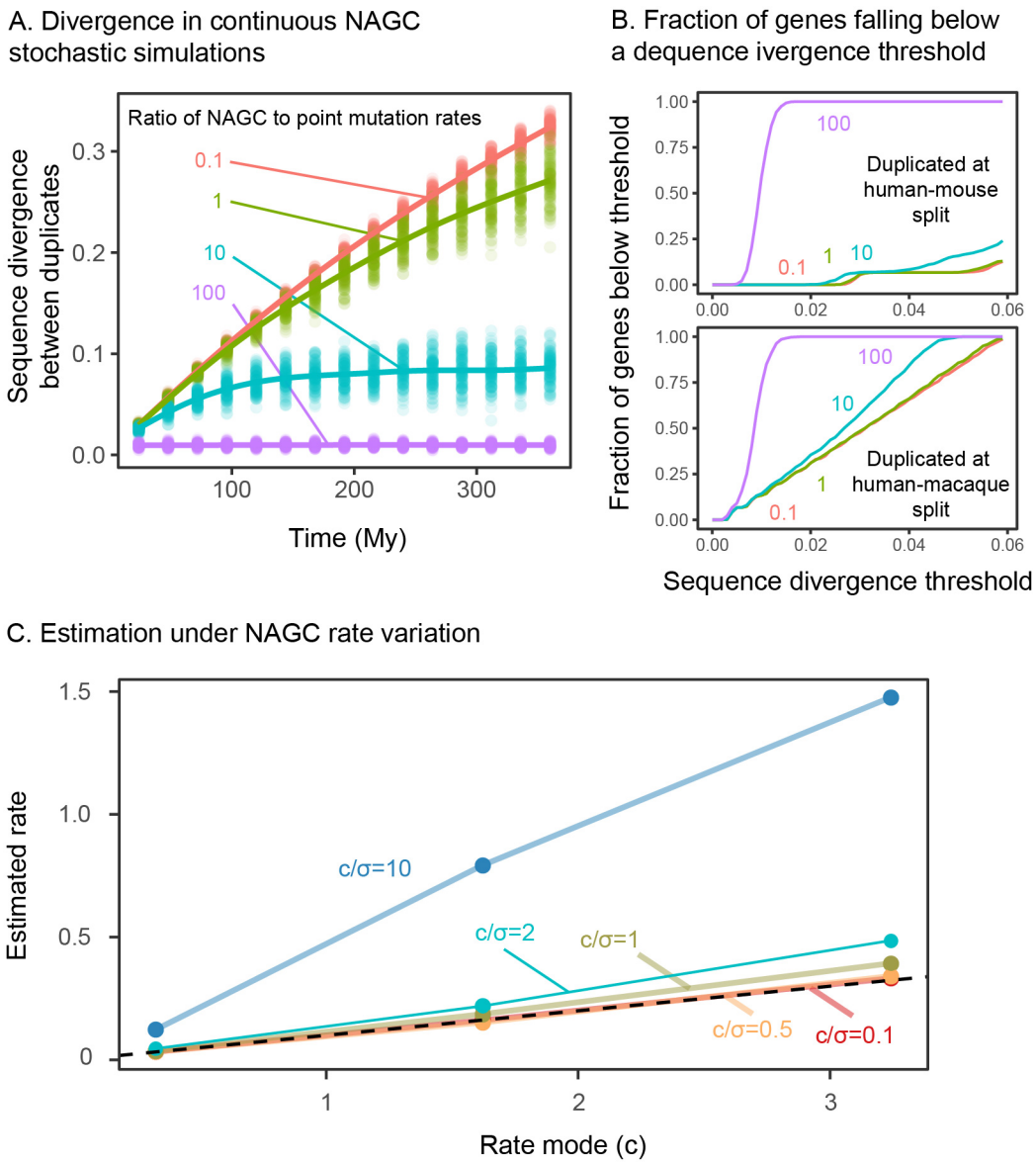

Figure S8: **(A)** Sequence divergence levels from 100 simulations for each of four NAGC rates. Points show divergence levels in stochastic simulations. Lines show natural cubic spline fits. **(B)** Fraction of simulations in (A) that do not exceed sequence divergence thresholds. The upper and lower panel correspond to 359.9 and 48.7 million years of evolution, respectively. **(C)** With randomly distributed NAGC rates, genes with higher rates are likelier to be included for estimation, leading to an upward bias. Points show the mean rate in genes included for estimation ($ds \leq 0.05$; "estimated rate") vs. the mode of the underlying truncated Normal rate distribution, out of 100 simulated human genes that duplicated at the time of the human-macaque split (48.7Mya). Colors correspond to different coefficients of variation (CV) of the (untruncated) rate distribution. The black dashed line shows the unbiased 1:1 relationship.

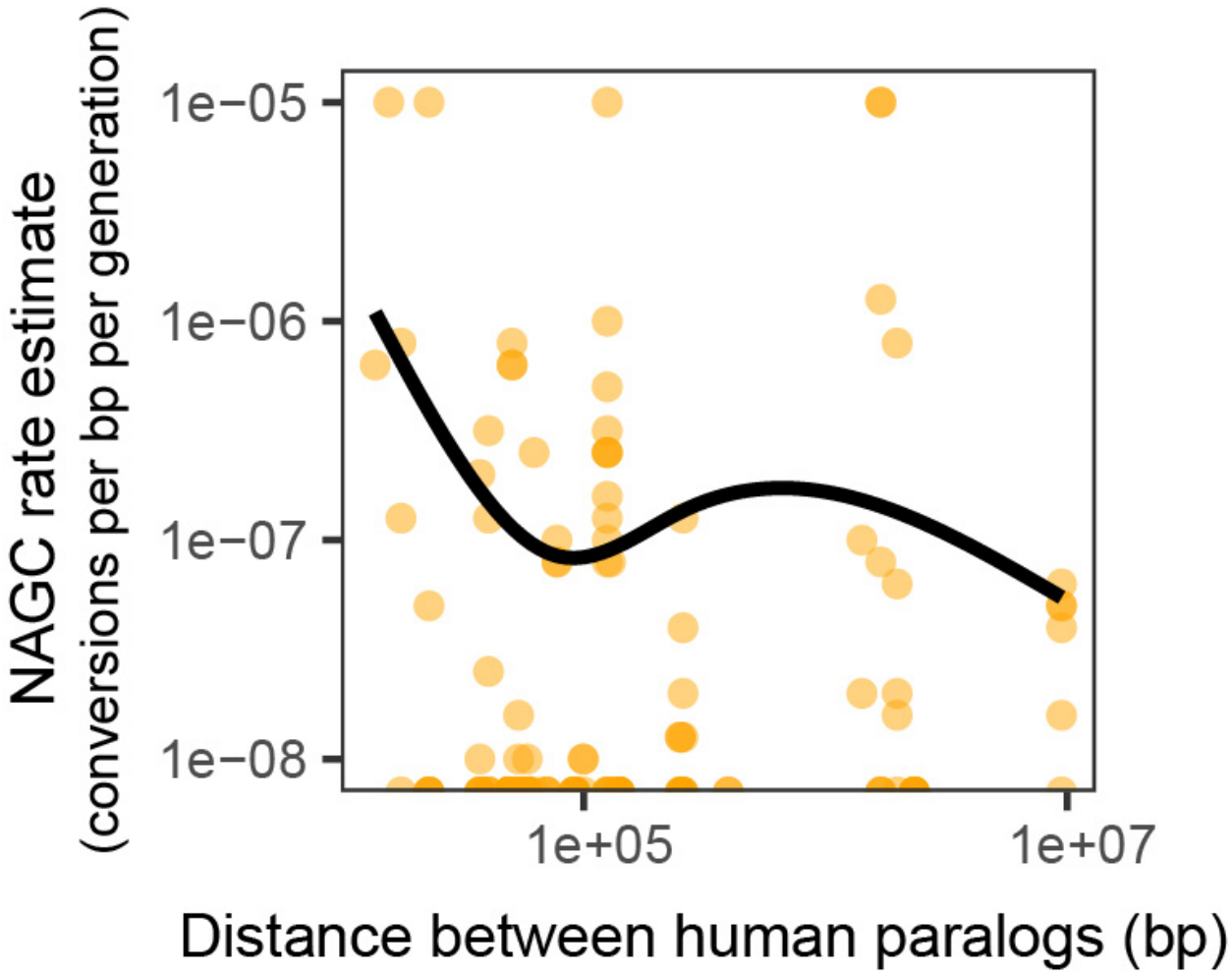

Figure S9: NAGC rate decreases with physical distance between duplicate sequences. Yellow points show maximum composite likelihood (MLE) rate estimates for each intron (orange points). MLEs of zero are plotted at the bottom. The solid line shows a natural cubic spline fit.

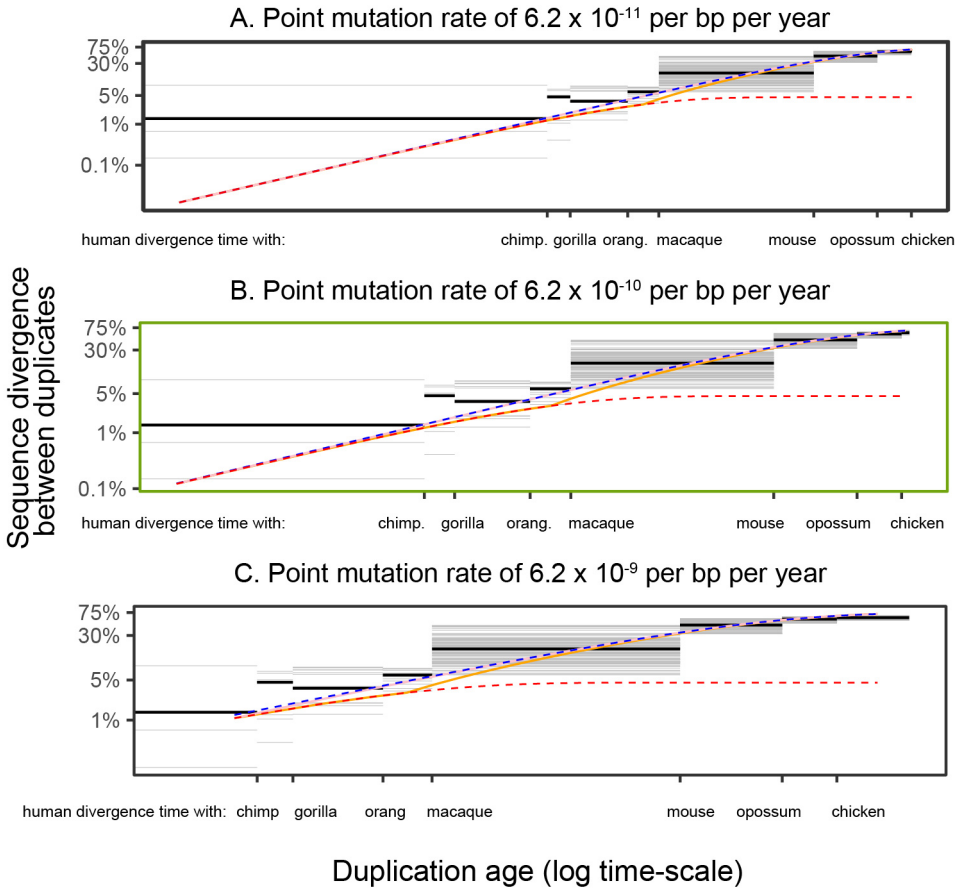

Figure S10: Expected divergence trajectories for different point mutation rates. All other parameters remain the same as in the main text. Split times of human and other species were recalculated with the given mutation rates as described in section 1.5.1. The panel with green border lines shows the trajectories for the mutation rate estimated using the human-chimpanzee and human-gorilla split times estimated by Moorjani et al. [18] that was used in Fig. 4 of the main text.

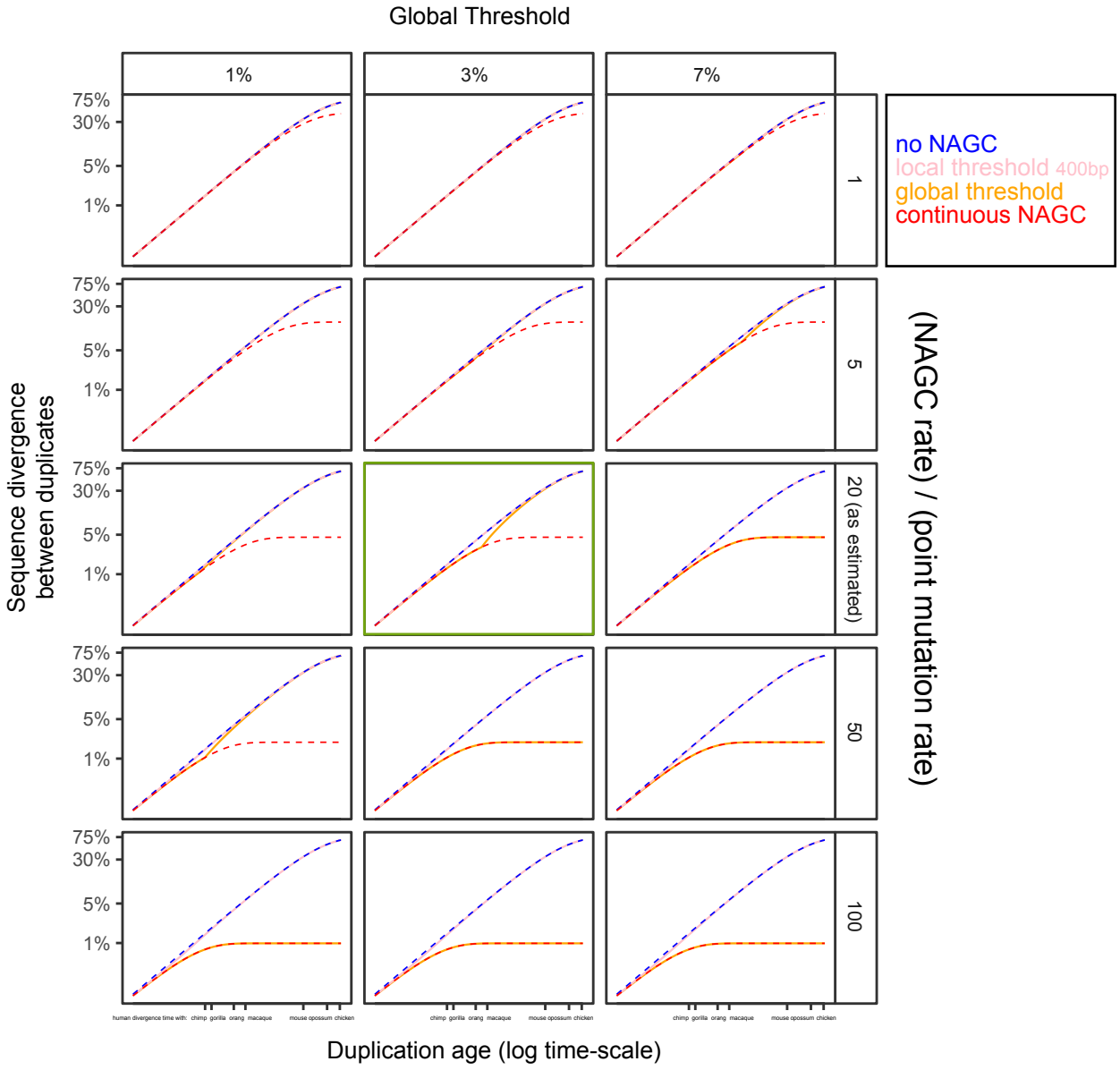

Figure S11: Expected divergence trajectories for different NAGC rates and global thresholds. The local threshold and the mutation rate remain the same as in the main text throughout. The panel with green border lines shows the trajectories for the NAGC rate and global threshold that was used in Fig. 4 of the main text.

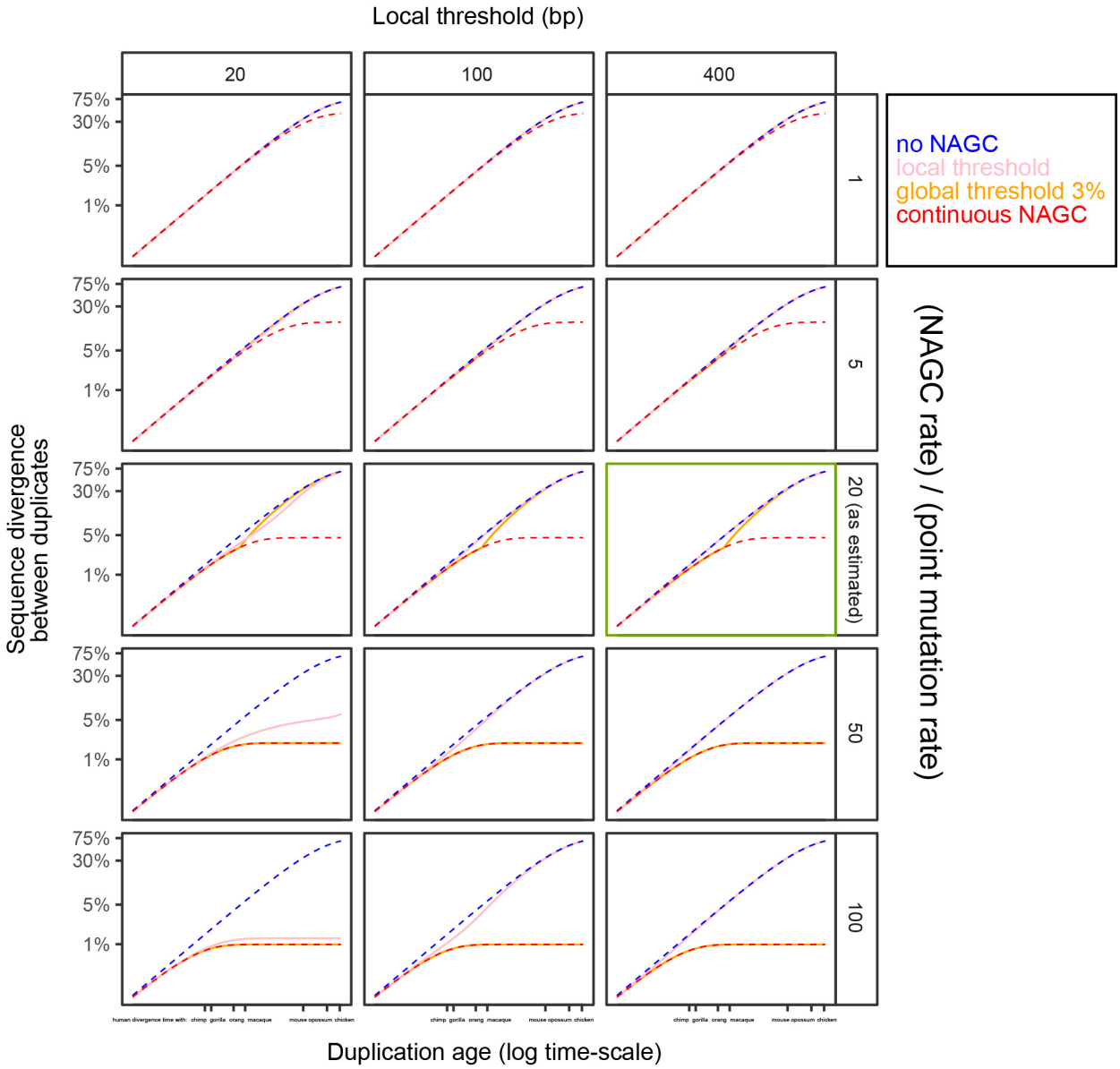

Figure S12: Expected divergence trajectories for different NAGC rates and local thresholds (width of region with perfect sequence homology required for NAGC to occur). The global threshold and the mutation rate remain the same as in the main text throughout. The panel with green border lines shows the trajectories for the NAGC rate and local threshold that was used in Fig. 4 of the main text.

3 List of Supplementary Files

Supplementary File 1 - Converted intronic regions identified by the HMM using human and macaque

Supplementary File 2 - Converted intronic regions identified by the HMM using human and orangutan

Supplementary File 3 - Converted intronic regions identified by the HMM using human and gorilla

Supplementary File 4 - Converted intronic regions identified by the HMM using human and chimpanzee

Supplementary File 5 - Divergence levels between human duplicates

References

[1] Xun Lan and Jonathan K Pritchard. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science*, 352(6288):1009–1013, 2016.

[2] John P. Huelsenbeck, Fredrik Ronquist, et al. Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, 2001.

[3] Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[4] Rolf Turner and Limin Liu. *hmm.discnp: Hidden Markov models with discrete non-parametric observation distributions.*, 2014. R package version 0.2-3, http://CRAN.R-project.org/package=hmm.discnp.

[5] SA Sawyer. GENECONV: A computer package for the statistical detection of gene conversion. *http://www.math.wustl.edu/˜sawyer*, 1999.

[6] Sayaka P Mansai and Hideki Innan. The power of the methods for detecting interlocus gene conversion. *Genetics*, 184(2):517–527, 2010.

[7] Sayaka P Mansai, Tomoyuki Kado, and Hideki Innan. The rate and tract length of gene conversion between duplicated genes. *Genes*, 2(2):313–331, 2011.

[8] Yann Lesecque, Dominique Mouchiroud, and Laurent Duret. GC-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. *Molecular biology and evolution*, 30(6):1409–1419, 2013.

[9] Barbara Arbeithuber, Andrea J Betancourt, Thomas Ebner, and Irene Tiemann-Boege. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proceedings of the National Academy of Sciences*, 112(7):2109–2114, 2015.

[10] Raquel Assis and Alexey S Kondrashov. Nonallelic gene conversion is not GC-biased in drosophila or primates. *Molecular Biology and Evolution*, page 304, 2011.

[11] Augustine Kong, Michael L Frigge, Gisli Masson, Soren Besenbacher, Patrick Sulem, Gisli Magnusson, Sigurjon A Gudjonsson, Asgeir Sigurdsson, Aslaug Jonasdottir, Adalbjorg Jonasdottir, et al. Rate of de novo mutations and the importance of father/'s age to disease risk. *Nature*, 488(7412):471–475, 2012.

[12] Michael W Nachman and Susan L Crowell. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1):297–304, 2000.

[13] Arbel Harpak, Anand Bhaskar, and Jonathan K Pritchard. Mutation rate variation is a primary determinant of the distribution of allele frequencies in humans. *PLoS Genetics*, 12(12):e1006489, 2016.

[14] Augustine Kong, Gudmar Thorleifsson, Daniel F Gudbjartsson, Gisli Masson, Asgeir Sigurdsson, Aslaug Jonasdottir, G Bragi Walters, Adalbjorg Jonasdottir, Arnaldur Gylfason, Kari Th Kristinsson, et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467(7319):1099–1103, 2010.

[15] Joseph Felsenstein and No Nov. Evolutionary trees from gene frequencies and quantitative characters : finding maximum likelihood estimates. *Evolution*, 35(6):1229–1242, 1981.

[16] Ezra Schildkraut, Cheryl A Miller, and Jac A Nickoloff. Gene conversion and deletion frequencies during double-strand break repair in human cells are controlled by the distance between direct repeats. *Nucleic Acids Research*, 33(5):1574–1580, 2005.

[17] W-H Li and Dan Graur. *Fundamentals of molecular evolution*. Sinauer Associates, 1991.

[18] Priya Moorjani, Carlos Eduardo G Amorim, Peter F Arndt, and Molly Przeworski. Variation in the molecular clock of primates. *Proceedings of the National Academy of Sciences*, 113(38):10607–10612, 2016.

[19] Thomas H Jukes and Charles R Cantor. Evolution of protein molecules. *Mammalian Protein Metabolism*, 3(21):132, 1969.

[20] Jian-Min Chen, David N Cooper, Nadia Chuzhanova, Claude Férec, and George P Patrinos. Gene conversion: mechanisms, evolution and human disease. *Nature Reviews Genetics*, 8(10):762–775, 2007.