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

Arbel Harpak[1,*,+], Xun Lan[2,3,*], Ziyue Gao[2,3] and Jonathan K. Pritchard[1,2,3,+]

[1] Department of Biology, Stanford University, Stanford, CA

[2] Department of Genetics, Stanford University, Stanford, CA

[3] Howard Hughes Medical Institute, Stanford University, Stanford, CA

[*] These authors contributed equally to this work

[+] Correspondence should be addressed to A.H. (arbelh@stanford.edu) or J.K.P. (pritch@stanford.edu)

Abstract

Gene conversion is the unidirectional transfer of genetic sequence from a "donor" region to an "acceptor". In non-allelic gene conversion (NAGC), the donor and the acceptor are at distinct genetic loci. Despite the role NAGC plays in various genetic diseases and the concerted evolution of many gene families, the parameters that govern NAGC are not well-characterized. Here, we survey duplicate gene families and identify converted tracts in 46% of them. These conversions reflect a significant GC-bias of NAGC. We develop a population-genetic model that exploits information from a long evolutionary history and use it to estimate the parameters that govern NAGC in humans: a mean conversion tract length of 250bp and a probability of 2.5×10^{-7} *per generation for a nucleotide to be converted (an order of magnitude higher than point mutations). Despite this seemingly high rate, we show that NAGC has only a small average effect on the sequence divergence of duplicates. This work improves our understanding of NAGC mechanism and the role that it plays in the evolution of gene duplicates.*

Background

As a result of recombination, distinct alleles that originate from the two homologous chromosomes may end up on the two strands of the same chromosome. This mismatch ("heteroduplex") is then repaired by synthesizing a DNA segment to overwrite the sequence on one strand, using the other strand as a template. This process is called gene conversion.

Although gene conversion is not an error but rather a natural part of recombination, it can result in the non-reciprocal transfer of alleles from one sequence to another, and can therefore be thought of as a "copy and paste" mutation. Gene conversion typically occurs between allelic regions (allelic gene conversion, AGC) [40]. However, *non-allelic* gene conversion (NAGC) between distinct genetic loci can also occur when the paralogous sequences are accidently aligned during recombination because they are highly similar [9]—as is often the case with young tandem gene duplicates [24].

NAGC is implicated as a driver of over twenty diseases [5, 9, 8]. The transfer of alleles between tandemly duplicated genes—or psuedogenes—can cause nonsynonymous mutations [18, 60], frameshifting [45] or aberrant splicing [35]—resulting in functional impairment of the acceptor gene. A recent study showed that alleles introduced by NAGC are found in 1% of genes associated with inherited diseases [8].

NAGC is also considered to be a dominant force restricting the evolution of gene duplicates [42, 14]. It was noticed half a century ago that duplicated genes can be highly similar within one species, even when they differ greatly from their orthologs in other species [51, 50, 7, 33]. This phenomenon has been termed "concerted evolution" [64]. NAGC is an immediate suspect for driving concerted evolution, because it homogenizes paralogous sequences by overturning differences that accumulate through other mutational mechanisms [51, 50, 42, 44]. Another possible driver of concerted evolution is natural selection. Both directional (purifying or positive) and balancing selection may restrict sequence evolution to be similar in paralogs [57, 52, 24, 14, 53, 36, 17]. Importantly, if NAGC is indeed slowing down sequence divergence, it puts in question the fidelity of molecular clocks for gene duplicates. In order to develop expectations for sequence and function evolution in duplicates, we must characterize NAGC and its interplay with other mutations.

In attempting to link NAGC mutations to sequence evolution, two questions arise: (i) what is the rate of NAGC? and (ii) what is the distribution of the tract length? These questions have been mostly probed in non-human organisms with mutation accumulation experiments limited to single genes—typically artificially inserted DNA sequences [28, 38]. The mean tract length has been estimated fairly consistently across organisms and experiments to be a few hundred base pairs [37]. However, estimates of the rate of NAGC vary by as much as eight orders of magnitude [63, 61, 54, 28, 34]—presumably due to key determinants of the rate that vary across experiments, such as genomic location, sequence similarity of the duplicate sequences and the distance between them, and experimental variability [38]. Alternatively, evolutionary-based approaches [22, 47] tend to be less variable: NAGC has been estimated to be 10-100 times faster than point mutation rate in *Saccharomyces cerevisiae* [55], in *Drosophila melanogaster* [58, 1] and in humans [23, 46, 6, 21]. These estimates are typically based on single loci (but see [12]). Recent family studies [62, 16] have estimated the rate of AGC to be 5.9×10^{-6} per bp per generation. This is likely an upper bound on the rate of NAGC, since NAGC requires a misalignment of homologous chromosomes, while AGC does not.

Here, we estimate the parameters governing NAGC with a novel sequence evolution model. Our method is not based on direct empirical observations, but it leverages substantially more information than previous experimental and computational methods: we use data from a large set of segmental duplicates in multiple species, and exploit information from a long evolutionary history. We estimate that the rate of NAGC in newborn duplicates is an order of magnitude higher than point mutation rate in humans. Surprisingly, we show that this high rate does not necessarily imply that NAGC distorts molecular clocks.

Results

To investigate NAGC in duplicate sequences across primates, we used a set of gene duplicate pairs in humans that we had assembled previously [32]. We focused on young pairs where we estimate that the duplication occured after the human-mouse split, and identified their orthologs in the reference genomes of chimpanzee, gorilla, orangutan, macaque and mouse. We required that each gene pair have both orthologs in at least one non-human primate and exactly one ortholog in mouse. Since our inference methods will implicitly assume neutral sequence evolution, we focused our analysis on intronic sequence at least 50bp away from intron-exon junctions. After applying these filters, our data consisted of $97,055$bp of sequence in 169 intronic regions from 75 gene families (**Methods**).

We examined divergence patterns (the partition of alleles in gene copies across primates) in these gene families. We noticed that some divergence patterns are rare and clustered in specific regions. We hypothesized that NAGC might be driving this clustering. To illustrate this, consider a family of two duplicates in human and macaque which resulted from a duplication followed by a speciation event—as illustrated in **Fig. 1B** ("Null tree"). Under this genealogy, we expect certain divergence patterns across the four genes to occur more frequently than others. For example, the grey sites in **Fig. 1C** can be parsimoniously explained by one substitution under the null genealogy. They should therefore be much more common than purple sites, as purple sites require at least two mutations. However, if we consider sites in which a NAGC event occurred after speciation (**Fig. 1A** and "NAGC tree" in **Fig. 1B**), our expectation for variation patterns changes: now, purple sites are much more likely than grey sites.

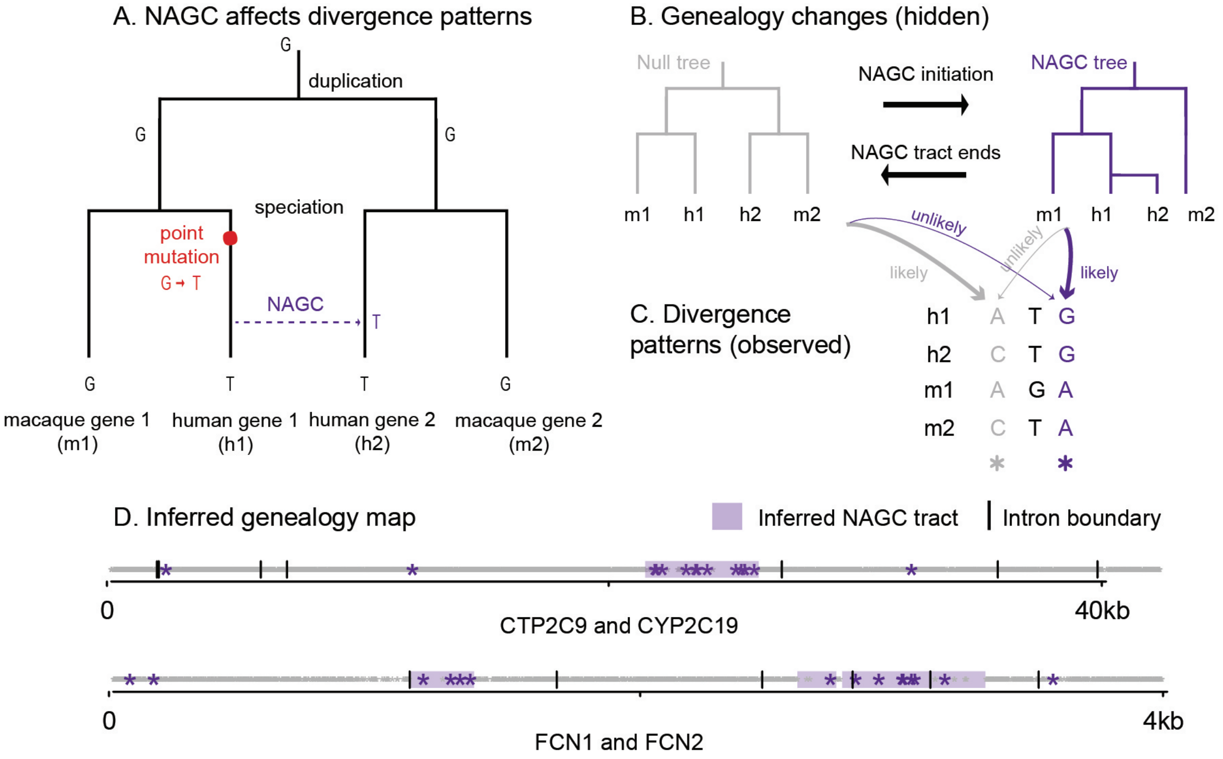

Figure 1: Non-allelic gene conversion (NAGC) alters divergence patterns. **(A)** NAGC can drive otherwise rare divergence patterns, like the sharing of alleles across paralogs but not orthologs. **(B)** An example of a local change in genealogy, caused by NAGC. **(C)** examples of divergence patterns in a small multigene family. Some divergence patterns—such as the one highlighted in purple—were both rare and spatially clustered. We hypothesized that underlying these changes are local changes in genealogy, caused by NAGC. **(D)** State of local genealogy (null by white, NAGC by purple tracts) inferred by our Hidden Markov Model (HMM) based on observed divergence patterns (stars) in two gene families. For simplicity, only the most informative patterns (purple and grey sites, as exemplified in panel C) are plotted.

Mapping recent NAGC events

We developed a Hidden Markov Model which exploits the fact that observed local changes in divergence patterns may point to hidden local changes in the genealogy of a gene family (**Fig 1B,C**). In our model, genealogy switches occur along the sequence at some rate; the likelihood of a given divergence pattern at a site then depends only on its own genealogy and nucleotide substitution rates (**Methods**). We applied the model to a subset of the gene families that we described above: families of four genes consisting of two duplicates in human

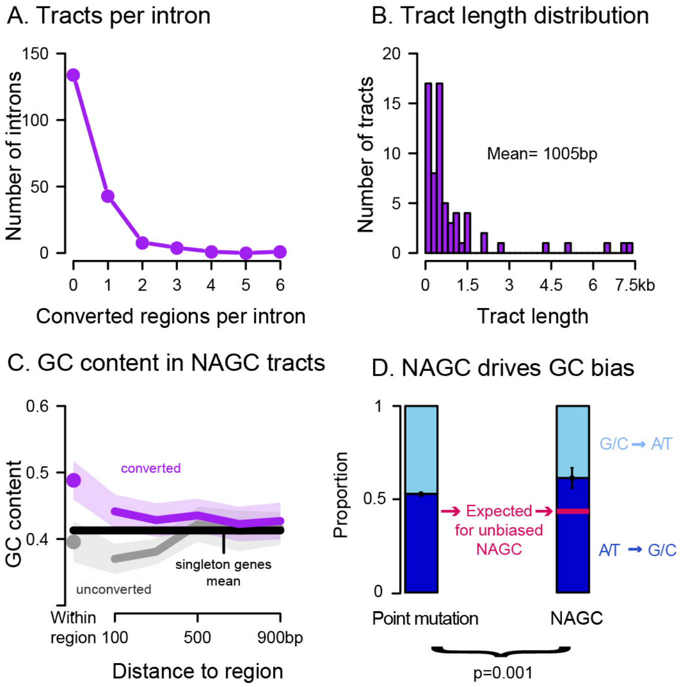

Figure 2: Properties of HMM-inferred converted tracts. **(C)** The purple dot shows the average GC content in converted regions. The grey dot shows the average for random unconverted regions, matched in length and within the same gene as the converted regions. The lines show GC content for symmetric 200bp bins centered at the respective regions (excluding the focal region itself). Shaded regions show 95% confidence intervals. Black line shows the intronic average for human genes with no identified paralogs. **(D)** In purple sites (**Fig. 1C**) that are most likely to be a direct result of NAGC (right bar), AT→GC substitutions through NAGC are significantly more common than GC→AT substitutions. The left bar shows the estimated proportion of AT→GC substitutions through point mutations and AGC in unconverted regions, which we used to derive the expected proportion for unbiased NAGC (pink line) after accounting for their different GC content.

and another primate—either chimpanzee or macaque. We required that the overall intronic divergence patterns are most compatible with a duplication event preceding speciation, using the software *MrBayes* [20].

Applying our HMM, we identified putatively converted tracts in 18/39 (46%) of the gene families considered, affecting 13.2% of intronic sequence (**Fig 2A, File S2**)—roughly 8% higher than previous estimates [25, 12, 10]. **Fig. 1D** shows an example of the maximum likelihood genealogy maps for two gene families (see complete list of identified tracts in the **Methods**). The average length of the detected converted tracts is 1005bp (**Fig. 2B**).

When an AT/GC heteroduplex DNA arises during AGC, it is preferentially repaired towards GC alleles [13, 43]. We sought to examine whether the same bias can be observed for NAGC [13, 2]. We found that converted regions have a high GC content: 48.9%, compared with 39.6% in matched unconverted regions ($p = 4 \times 10^{-5}$, two-sided t-test and see **Fig. 2C**). However, this difference in base composition could either be a driver and/or a result of NAGC. To test whether NAGC is a driver of high GC content, we focused on sites that carry the strongest evidence of nucleotide substitution by NAGC—these are the sites with the "purple" divergence pattern as before (**Fig. 1C**). Using a simple parsimony-based model, we inferred the directionality of such substitutions involving both weak (AT) and strong (GC) nucleotides. We found that 61% of these changes were weak to strong changes, compared with an expectation of 44% through point mutations and GC-biased AGC alone (exact binomial test $p = 1 \times 10^{-3}$ and see **Methods**; **Fig. 2D**). This difference supports a GC bias driven directly by NAGC, and is in broad agreement with the GC bias estimated for AGC [62, 16].

The power of our HMM is likely limited to recent conversions, where local divergence patterns show clear disagreement with the global intron-wide patterns; it is therefore applicable only in cases where NAGC is not so pervasive that it would have a global effect on divergence patterns [37, 4]. Next, we describe a method that allowed us to estimate NAGC parameters without making this implicit assumption.

NAGC is an order of magnitude faster than the point mutation rate

To estimate the rate and the tract length distribution of NAGC, we developed a two-site model of sequence evolution with mutation and NAGC (**Methods**). This model is inspired by the rationale that guided Hudson [19] and McVean et al. [39] in estimating recombination rates. In short, mutation acts to increase—while NAGC acts to decrease—sequence divergence between paralogs. When the two sites under consideration are close-by (with respect to NAGC mean tract length), NAGC events affecting one site are likely to incorporate the other (**Fig. 3A**). For each pair of sites in each intron in our data, we computed the likelihood of the observed alleles in all available species, over a grid of NAGC rate and mean tract length values. We then attained maximum *composite* likelihood estimates (MLE) over all pairs of sites (ignoring the dependence between pairs).

We first estimated MLEs for each intron separately, and matched these estimates with ds [33] in exons of the respective gene. We found that NAGC rate estimates decrease as ds increases (Spearman $p = 1 \times 10^{-5}$, **Fig. 3C**). This trend is likely due to a slowdown in NAGC rate, or complete stop thereof, as the duplicates diverge in sequence. Since our model assumes a constant NAGC rate, we concluded that the model would be most applicable to lowly diverged genes and therefore limited our parameter estimation to introns with $ds < 5\%$.

We define NAGC rate as the probability that a random nucleotide is converted per basepair per generation. We estimate this rate to be 2.5×10^{-7} ($[0.8 \times 10^{-7}, 5.0 \times 10^{-7}]$ 95% nonparametric bootstrap CI, **Fig. 3D**). This estimate accords with previous estimates based on smaller sample sizes using polymorphism data [22, 38] and is an order of magnitude slower than AGC rate [62, 16]. We simultaneously estimated a mean NAGC tract length of 250bp ($[63, 1000]$ nonparametric bootstrap CI)—consistent with estimates for AGC [26, 62]) and with a meta-analysis of many NAGC mutation accumulation experiments and NAGC-driven diseases [38].

Live fast, stay young? the effect of NAGC on neutral sequence divergence

We next consider the implications of our results on the divergence dynamics of orthologs post their duplication. In light of the high rate we infer, the question arises: if mutations that increase sequence divergence are much slower than NAGC [30, 49]—which acts to eliminate divergence—should we expect gene duplicates never to diverge in sequence?

We considered several models of sequence divergence (**Methods**). First, we considered a

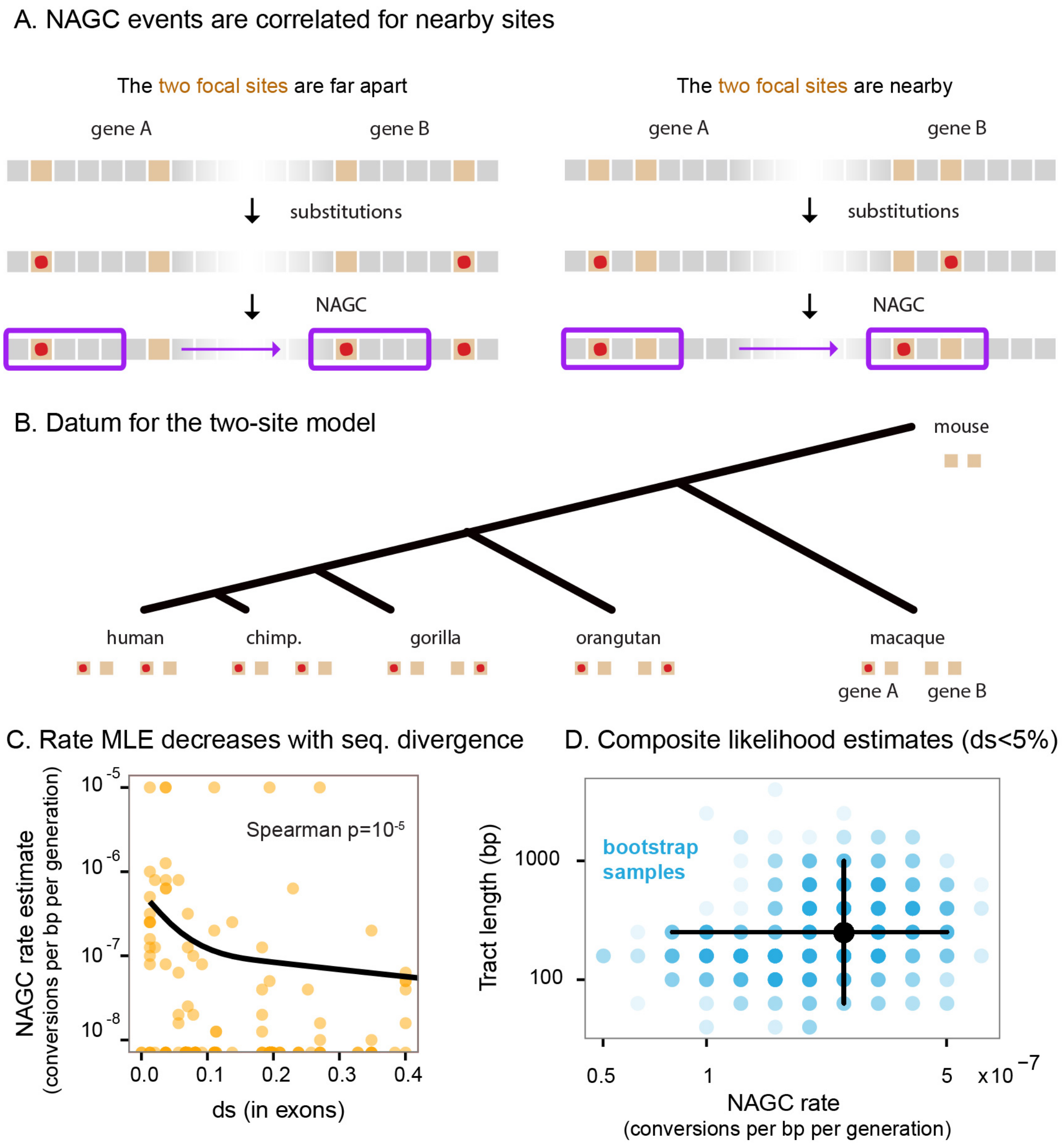

Figure 3: Estimation of NAGC parameters. **(A)** The two-site sequence evolution model exploits the correlated effect of NAGC on nearby sites (near with respect to the mean tract length). In this illustration, orange squares represent focal sites. Point substitutions are shown by the red points, and a converted tract is shown by the purple rectangle. **(B)** Illustration of a single datum on which we compute the full likelihood, composed of two sites in two duplicates across multiple species (except for the mouse outgroup for which only one ortholog exists). **(C)** Maximum composite likelihood (MLE) rate estimates for each intron (orange points). MLEs of zero are plotted at the bottom. Solid line shows a natural cubic spline fit. The rate decreases with sequence divergence (ds). We therefore only use lowly-diverged genes ($ds \leq 5\%$) to get point estimates of the baseline rate. **(D)** Composite likelihood estimates. The black point is centered at our point estimate for $ds \leq 5\%$ genes. Blue points show non-parametric bootstrap estimates. The corresponding 95% marginal confidence intervals are shown by black lines.

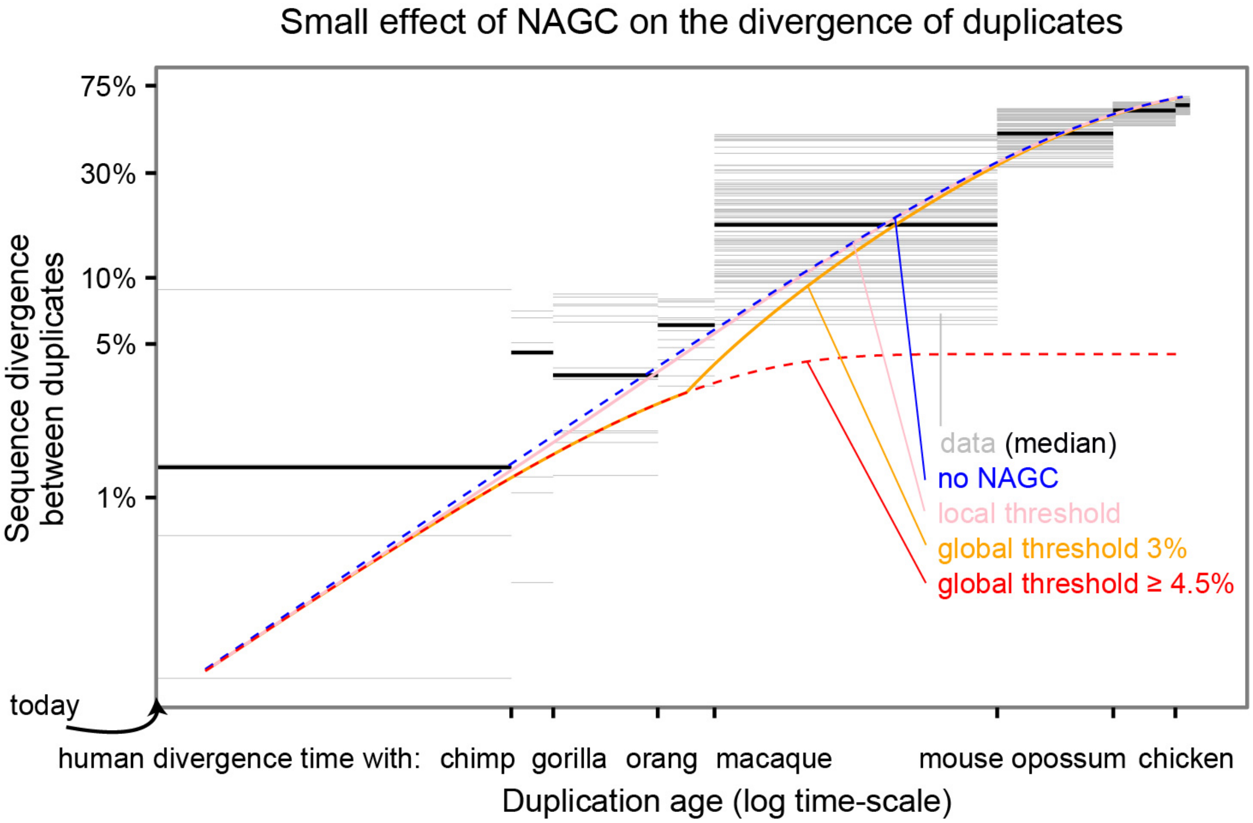

Figure 4: The effect of NAGC on the divergence of duplicates. The figure shows both data from human duplicate pairs and theoretical predictions of different NAGC models. The blue line shows expected divergence in the absence of NAGC, and the red line shows the expectation with NAGC acting continuously. The pink, orange and red lines show the mean sequence divergence for models in which NAGC initation is contingent on sequence similarity between the paralogs. The grey horizontal lines correspond to human duplicate pairs. The duplication time for each pair is inferred by examining the non-human species that carry orthologs for both of the human paralogs. Y-axis shows ps between the two human paralogs.

model where NAGC acts at the constant rate that we estimated throughout the duplicates' evolution ("continuous NAGC"). In this case, divergence is expected to plateau around 4.5%, and concerted evolution continues for a long time (red line in **Fig. 4**; in practice there will eventually be an "escape" through a chance rapid accumulation of multiple mutations [56, 14]). However, NAGC is hypothesized to be contingent on high sequence similarity between paralogs.

We therefore considered two alternative models of NAGC dynamics. First, a model in which NAGC acts only while the sequence divergence between the paralogs is below some threshold ("global threshold"). Second, a model in which the initiation of NAGC at a site is contingent on perfect sequence homology at a short 400bp flanking region upstream to the site ("local threshold", [27, 38, 9]).

The local threshold model yielded a similar average trajectory to that in the absence of NAGC. A global threshold of as low as 4.5% may lead to an extended period of concerted evolution as in the continuous NAGC model. A global threshold of $< 4.5\%$ results in a different trajectory. For example, with a global threshold of 3%, duplicates born at the time of the primates most recent common ancestor (MRCA) would diverge at 3.9% of their sequence, as compared to 5.7% in the absence of NAGC (**Fig. 4**).

Lastly, we asked what these results mean for the validity of molecular clocks for gene duplicates. We examined the explanatory power of different theoretical models for synonymous divergence in human duplicates. We wished to get an estimate of the age of duplication that is independent of ds between the human duplicates; we therefore used the extent of sharing of both paralogs in different species as a measure of the duplication time. For example, if a duplicate pair was found in human, gorilla and orangutan—but only one ortholog was found in macaque—we estimated that the duplication occurred at the time interval between the human-macaque split and the human-orangutan split. Except for the continuous NAGC model, all models displayed similar broad agreement with the data (**Fig. 4**).

The small effect of NAGC on divergence levels is intuitive in retrospect: for identical sequences, NAGC has no effect. Once differences start to accumulate, there is only a small window of opportunity for NAGC to act before the paralogous sequences escape from its hold. This suggests that neutral sequence divergence (e.g. ds) may be an appropriate molecular clock even in the presence of NAGC (as also suggested by [12, 11, 32]).

Discussion

In this work, we identify recently converted regions in humans and other primates, and estimate the parameters that govern NAGC. Previously, it has been somewhat ambiguous whether concerted evolution observations were due to natural selection, pervasive NAGC, or a combination of the two [53, 36, 24]. Today, equipped with genomic data, we can revisit the pervasiveness of concerted evolution; the data in **Fig. 4** suggests that in humans, duplicates' divergence levels are roughly as expected from the accumulation of point mutations alone. When we plugged in our estimates for NAGC rate, most mechanistic models of NAGC also predicted a small effect on neutral sequence divergence. This result suggests that neutral sequence divergence may be an appropriate molecular clock even in the presence of NAGC.

One important topic left for future investigation is the variation of NAGC parameters. Our model assumes constant action of NAGC through time and across the genome in order to get a robust estimate of the mean parameters. However, substantial variation likely exists across gene pairs due to factors such as recombination rate, genomic position, physical distance between paralogs and sequence similarity.

Our estimates for the parameters that govern the mutational mechanism alone could guide future studies of the forces guiding the evolution of gene duplicates. Together with contemporary efforts to measure the effects of genomic factors on gene conversion, our results may clarify the potential of NAGC to drive disease, improve our dating of molecular events and further our understanding of the evolution of gene duplicates.

Acknowledgements

This work was funded by NIH grants HG008140 and MH101825 and by the Howard Hughes Medical Institute (HHMI). AH and ZG were supported in part by fellowships from the Stanford Center for Computational, Evolutionary and Human Genomics (CEHG). We thank

Eilon Sharon, Doc Edge and Kelley Harris for helpful discussions and comments on the manuscript.

Methods

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 [32]. We obtained the orthologs of these genes in four other primates (chimpanzee, gorilla, orangutan, macaque) and in mouse from the same study (**Table 1**). 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 both of the inference methods that follow (one for the task of identifying converted tracts and the other for estimating NAGC parameters) we applied further filtering on the input data. We used the software *MrBayes* [20] to estimate gene family genealogies with the set of exons of our genes as input (note that only here we used exonic sequences rather that intronic). In the Hidden-Markov Model (HMM) used for identifying converted tracts, only gene families in which the most probable genealogy supports a duplication prior to the split of the two focal species (either human/chimpanzee or human/macaque) were kept. In the two-sites model used for estimating NAGC parameters, we require only that the duplication happened after the primates-mouse split.

Identifying converted regions using a 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.

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 1: A list of genome assemblies and gene annotations used.

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 one NAGC event at most 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

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

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,

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

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* [59] 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

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

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:

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

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:

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

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

and

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

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

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

Estimating GC bias in NAGC

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 an A/T→G/C conversion. We observed that $f_{obs} = 61\%$ (51 out of 83) of A/T↔G/C substitutions are in the A/T→G/C direction.

To evaluate the deviation of this proportion from that expected with no GC biased NAGC, we estimated f_0, the expected fraction of A/T→G/C substitutions out of A/T↔G/C sites using unconverted regions. We looked at sites where only one out of the four genes carries an allele different from the rest, the most parsimonious scenario 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. However, GC content was lower in unconverted regions (39.6%) than unconverted regions (48.9%). Adjusting for this difference in GC content,

$$\frac{f_0}{1-f_0} = \frac{\frac{1-(GC\ content\ in\ converted)}{GC\ content\ in\ converted}}{\frac{1-(GC\ content\ in\ unconverted)}{GC\ content\ in\ unconverted}} \cdot \frac{AT \rightarrow GC\ count\ in\ unconverted}{AT \leftarrow GC\ count\ in\ unconverted} = \frac{\frac{1-39.6\%}{39.6\%}}{\frac{1-48.9\%}{48.9\%}} \cdot \frac{2390}{4513-2390}.$$

Note that this expectation encapsulates both mutation rate and GC bias in AGC. Thus, if NAGC is not GC-biased, the expected fraction of A/T→G/C out of A/T↔G/C "purple" sites is

$$f_0 = 0.435.$$

We tested the null hypothesis that NAGC is unbiased,

$$H_0 : f_{obs} = f_0$$

using the exact binomial test and found that these proportions are significantly different ($p = 0.001$, **Fig. 2D**).

Two-site model

Transition matrix

We consider the evolution of two biallelic sites in two duplicate genes as a discrete homogeneous Markov Process. We describe these four sites with a 4-bit vector ("state vector"). The state $l_A l_B r_A r_B \in \{0,1\}^4$ corresponds to allele l_A at the "left" site in copy A, allele l_B at the "left" site in copy B, allele r_A at the "right" site in copy A and allele r_B at the "right" site in copy B. Note that the labels 0 and 1 are defined with respect to each site separately—the state 0000 does not mean that the the left and right site necessarily have the same allele. We first derive the (per generation) transition probability matrix. There are two possible events that may result in a transition: point mutations which occur at a rate of $\mu = 1.2 \times 10^{-8}$ per generation and NAGC. The probability of a site being converted per generation is c. We consider these mutational events to be rare and ignore terms of the order $O(\mu^2)$, $O(c^2)$ and $O(\mu c)$. For example, consider the per-generation transition probability from 0110 to 0100, for two sites that are d bp apart. This transition can happen either through point mutation at the right site of copy A, or by NAGC from copy A to copy B involving the left site but not the right. The transition probability is therefore

$$P(0110 \to 0100) = \mu + c(1 - g(d)) + O(\mu^2) + O(c^2) + O(\mu c),$$

where $g(d)$ is the probability of a conversion event including one of the sites given that it includes the other. Similarly, we can derive the full transition probability matrix $\mathbf{P}$:

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

where

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

Note that this parameterization ignores possible mutations to (third and fourth) unobserved alleles.

We next derive $g(d)$. Following previous work [37], we model the tract length as geometrically distributed with mean λ. It follows that the probability of a conversion including one site conditional on it includes the other is

$$g_{init}(d) = (1 - \frac{1}{\lambda})^d,$$

by the memorylessness of the geometric distribution. 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 during fixation. 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 $r = 10^{-8}$ [31] 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. 3–Figure Supplement 1**). We therefore use the approximation

$$g \approx g_{init}$$

in our implementation of this model.

Lastly, we turn to compute transition probabilities along evolutionary timescales. Each datum consists of state vectors (corresponding to two biallelic sites in two paralogs) encoding the alleles in the human reference genome and 1-4 other primate reference genomes. The mouse 2-bit state (two sites in one gene) will only be used to set a prior on the root of the tree (see separate section below). We assume a constant tree—namely, a constant topology and constant edge lengths $\{t_{ij}\}$ as defined in **Fig. 3–Figure Supplement 2**. We used estimates for primate split times from [48], and assumed a constant generation time of 25 years. Each node corresponds to a state. We assume that—for both mutation types—substitution occurs at a rate equal to the mutation rate. Therefore, the transition probability matrix $\mathbf{P}^*_{(\mathbf{edge\ ij})}$ for the edge between node i and node j is

$$\mathbf{P}^*_{(\mathbf{edge\ ij})} = \mathbf{P}^{t_{ij}}.$$

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 a following section.

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.

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 have called "data prior", 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 internal branch ending at the data root r and take it to be uniformly distributed along this internal branch (**Fig. 3–Figure Supplement 3**). 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

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

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

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

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,

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

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.

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 [32]. 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 [32] 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 leading to 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 ([33]) 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. ([41]). We then perform simple linear regression with no intercept (forcing the fitted line to go through the origin) regressing $2 \cdot ds_{chimpanzee}$ and $2 \cdot ds_{gorilla}$ 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 ([33]):

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

The mean divergence times estimated by this procedure are shown in **Table 2**.

Theoretical single-site sequence evolution models

We compute the mean divergence between duplicate sequences under different models of NAGC. We use a single-site models to evolve a length-two probability vector corresponding

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 2: Estimated divergence times between human and other species.

to 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 ([29]),

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

where μ is set as explained in the section **Estimating the duplication time interval for human duplicates**.

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,

$$\mathbf{A_3} = \mathbb{1}\{v_t < \gamma\}\mathbf{A_1} + \mathbb{1}\{v_t \geq \gamma\}\mathbf{A_2}.$$

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_{i-1}. This probability $g(t)$ is set by

$$g(t) = exp(-m \cdot v_{t-1}),$$

where we set$m = 400$ [9]. The transition matrix in this model is

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

List of Supplementary Files

Supplementary File 1 - Converted regions identified by the HMM

Supplementary File 2 - Divergence levels between duplicates

429 Supplementary Figures

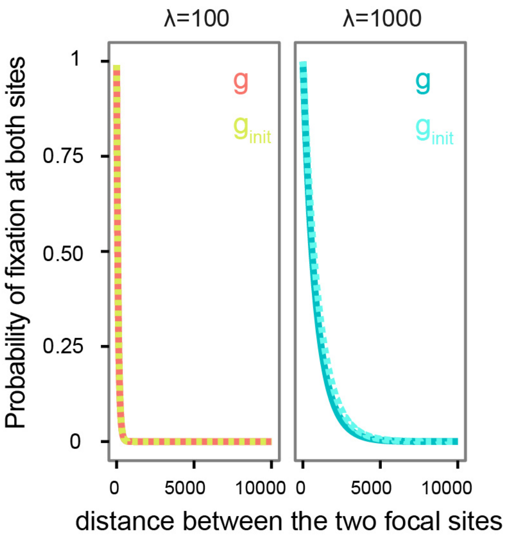

Figure 3–Figure Supplement 1: The probability of a NAGC mutation fixing at both sites, conditional on fixation in one of them. Shown is the probability as a function of the distance between focal sites for two mean tract lengths (λ) values. g denotes this probability when accounting for the possibility of decoupling of the sites through recombination, while g_{init} ignores it. However, the differences between the two are very small and we therefore approximate g by g_{init}.

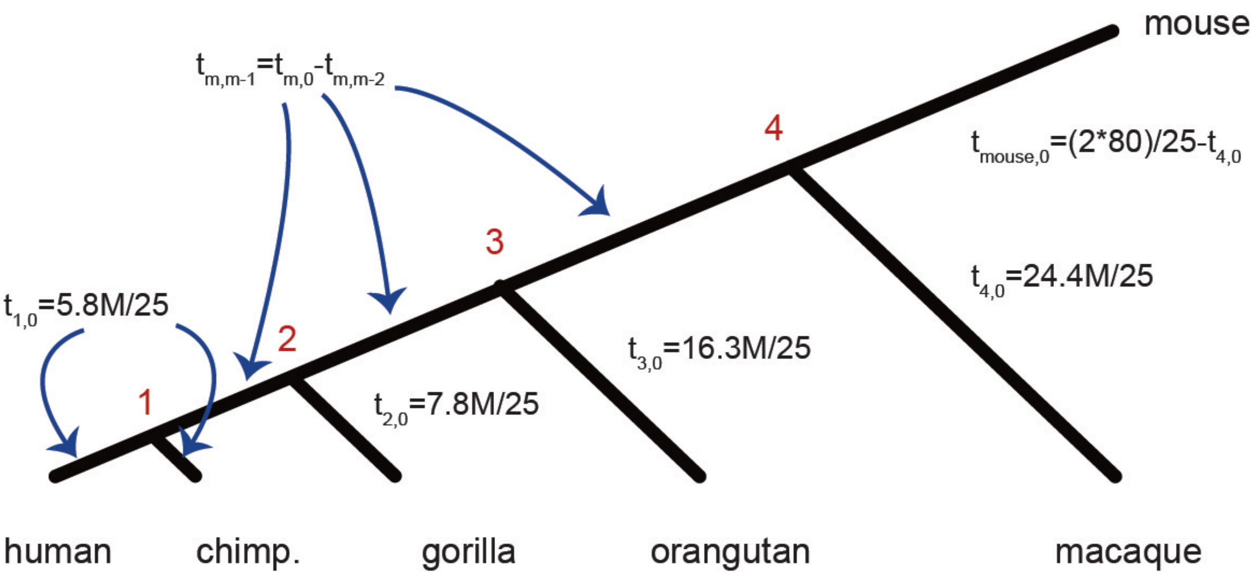

Figure 3–Figure Supplement 2: Split times parameterization for the two-site model

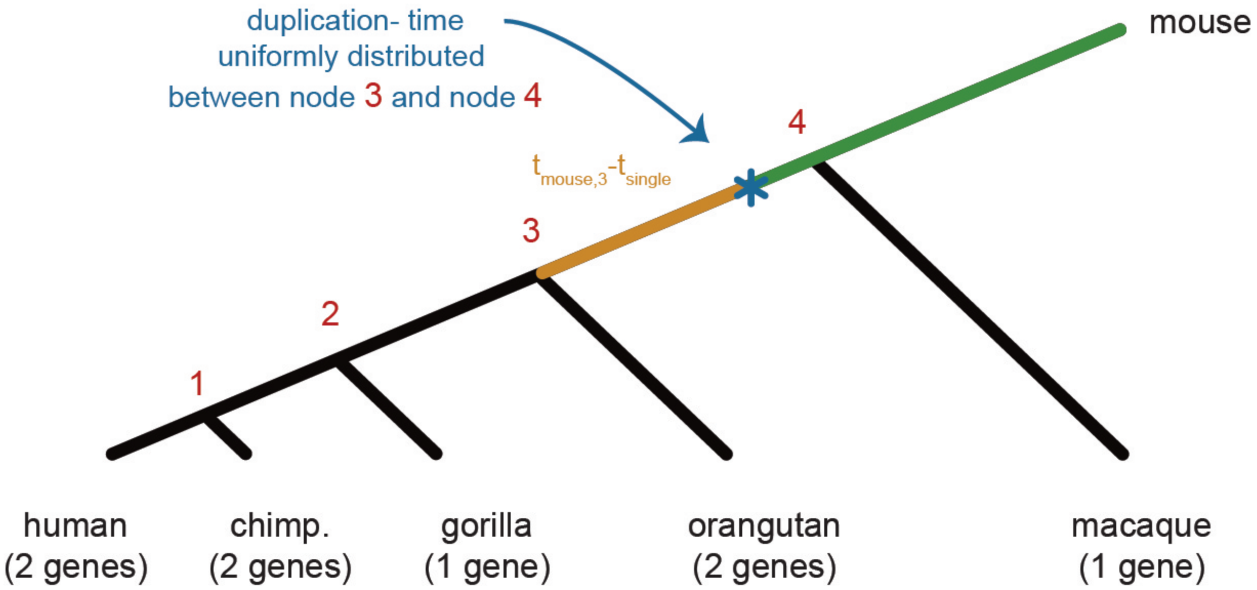

Figure 3–Figure Supplement 3: Setting the prior on the data root.

References

[1] Arguello, J. R., Chen, Y., Yang, S., Wang, W., and Long, M. Origination of an x-linked testes chimeric gene by illegitimate recombination in drosophila. *PLoS Genetics 2*, 5 (2006), e77.

[2] Assis, R., and Kondrashov, A. S. Nonallelic gene conversion is not GC-biased in drosophila or primates. *Molecular Biology and Evolution* (2011), 304.

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

[4] Betrán, E., Rozas, J., Navarro, A., and Barbadilla, A. The estimation of the number and the length distribution of gene conversion tracts from population dna sequence data. *Genetics 146*, 1 (1997), 89–99.

[5] Bischoff, J., Chiang, A., Scheetz, T., Stone, E., Casavant, T., Sheffield, V., and Braun, T. Genome-wide identification of pseudogenes capable of disease-causing gene conversion. *Human Mutation 27*, 6 (2006), 545ˆ552.

[6] Bosch, E., Hurles, M. E., Navarro, A., and Jobling, M. A. Dynamics of a human interparalog gene conversion hotspot. *Genome Research 14*, 5 (2004), 835–844.

[7] Brown, D. D., and Sugimoto, K. 5s dnas of xenopus laevis and xenopus mulleri: Evolution of a gene family. *Journal of Molecular Biology 78*, 3 (1973), 397–415.

[8] Casola, C., Zekonyte, U., Phillips, A. D., Cooper, D. N., and Hahn, M. W. Interlocus gene conversion events introduce deleterious mutations into at least 1% of human genes associated with inherited disease. *Genome Research 22*, 3 (2012), 429–435.

[9] Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C., and Patrinos, G. P. Gene conversion: mechanisms, evolution and human disease. *Nature Reviews Genetics 8*, 10 (2007), 762–775.

[10] Dennis, M. Y., Harshman, L., Nelson, B. J., Penn, O., Cantsilieris, S., Huddleston, J., Antonacci, F., Penewit, K., Denman, L., Raja, A., et al. The evolution and population diversity of human-specific segmental duplications. *Nature Ecology & Evolution 1* (2016), 0069.

[11] Dumont, B. L. Interlocus gene conversion explains at least 2 . 7 % of single nucleotide variants in human segmental duplications. *BMC Genomics* (2015), 1–11.

[12] Dumont, B. L., and Eichler, E. E. Signals of historical interlocus gene conversion in human segmental duplications. *PLoS One 8*, 10 (2013), e75949.

[13] Duret, L., and Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics 10* (2009), 285–311.

[14] Fawcett, J. A., and Innan, H. Neutral and non-neutral evolution of duplicated genes with gene conversion. *Genes 2*, 1 (2011), 191–209.

[15] Felsenstein, J., and Nov, N. Evolutionary trees from gene frequencies and quantitative characters : finding maximum likelihood estimates. *Evolution 35*, 6 (1981), 1229–1242.

[16] Halldorsson, B. V., Hardarson, M. T., Kehr, B., Styrkarsdottir, U., Gylfason, A., Thorleifsson, G., Zink, F., Jonasdottir, A., Jonasdottir, A., Sulem, P., Masson, G., Thorsteinsdottir, U., Helgason, A., Kong, A., Gudbjartsson, D. F., and Stefansson, K. The rate of meiotic gene conversion varies by sex and age. *Nature Genetics* (2016).

[17] Hanikenne, M., Kroymann, J., Trampczynska, A., Bernal, M., Motte, P., Clemens, S., and Krämer, U. Hard selective sweep and ectopic gene conversion in a gene cluster affording environmental adaptation. *PLoS Genetics 9*, 8 (2013), e1003707.

[18] Heinen, S., Sanchez-Corral, P., Jackson, M. S., Strain, L., Goodship, J. A., Kemp, E. J., Skerka, C., Jokiranta, T. S., Meyers, K., Wagner, E., et al. De novo gene conversion in the RCA gene cluster (1q32) causes mutations in complement factor h associated with atypical hemolytic uremic syndrome. *Human mutation 27*, 3 (2006), 292–293.

[19] Hudson, R. R. Two-locus sampling distributions and their application. *Genetics 159*, 4 (2001), 1805–1817.

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

[21] Hurles, M. E. Gene conversion homogenizes the cmt1a paralogous repeats. *BMC Genomics 2*, 1 (2001), 11.

[22] Innan, H. The coalescent and infinite-site model of a small multigene family. *Genetics 163*, 2 (2003), 803–810.

[23] Innan, H. A two-locus gene conversion model with selection and its application to the human rhce and rhd genes. *Proceedings of the National Academy of Sciences 100*, 15 (2003), 8793–8798.

[24] Innan, H., and Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics 11*, 2 (2010), 97–108.

[25] Jackson, M. S., Oliver, K., Loveland, J., Humphray, S., Dunham, I., Rocchi, M., Viggiano, L., Park, J. P., Hurles, M. E., and Santibanez-Koref,

M. Evidence for widespread reticulate evolution within human duplicons. *The American Journal of Human Genetics 77*, 5 (2005), 824–840.

[26] Jeffreys, A. J., and May, C. A. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nature Genetics 36*, 2 (2004), 151–156.

[27] Jinks-Robertson, S., Michelitch, M., and Ramcharan, S. Substrate length requirements for efficient mitotic recombination in saccharomyces cerevisiae. *Molecular and Cellular Biology 13*, 7 (1993), 3937–3950.

[28] Jinks-Robertson, S., and Petes, T. D. High-frequency meiotic gene conversion between repeated genes on nonhomologous chromosomes in yeast. *Proceedings of the National Academy of Sciences 82*, 10 (1985), 3350–3354.

[29] Jukes, T. H., and Cantor, C. R. Evolution of protein molecules. *Mammalian Protein Metabolism 3*, 21 (1969), 132.

[30] Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., et al. Rate of de novo mutations and the importance of father/'s age to disease risk. *Nature 488*, 7412 (2012), 471–475.

[31] Kong, A., Thorleifsson, G., Gudbjartsson, D. F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G. B., Jonasdottir, A., Gylfason, A., Kristinsson, K. T., et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature 467*, 7319 (2010), 1099–1103.

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

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

[34] Lichten, M., and Haber, J. Position effects in ectopic and allelic mitotic recombination in saccharomyces cerevisiae. *Genetics 123*, 2 (1989), 261–268.

[35] Lorson, C. L., Hahnen, E., Androphy, E. J., and Wirth, B. A single nucleotide in the smn gene regulates splicing and is responsible for spinal muscular atrophy. *Proceedings of the National Academy of Sciences 96*, 11 (1999), 6307–6311.

[36] Mano, S., and Innan, H. The evolutionary rate of duplicated genes under concerted evolution. *Genetics 180*, 1 (2008), 493–505.

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

[38] Mansai, S. P., Kado, T., and Innan, H. The rate and tract length of gene conversion between duplicated genes. *Genes 2*, 2 (2011), 313–331.

[39] McVean, G., Awadalla, P., and Fearnhead, P. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics 160*, 3 (2002), 1231–1241.

[40] Mitchell, M. B. Aberrant recombination of pyridoxine mutants of neurospora. *Proceedings of the National Academy of Sciences 41*, 4 (1955), 215–220.

[41] Moorjani, P., Amorim, C. E. G., Arndt, P. F., and Przeworski, M. Variation in the molecular clock of primates. *Proceedings of the National Academy of Sciences 113*, 38 (2016), 10607–10612.

[42] Nei, M. *Molecular evolutionary genetics.* Columbia university press, 1987.

[43] Odenthal-Hesse, L., Berg, I. L., Veselis, A., Jeffreys, A. J., and May, C. A. Transmission distortion affecting human noncrossover but not crossover recombination: a hidden source of meiotic drive. *PLoS Genetics 10*, 2 (2014), e1004106.

[44] Ohta, T. How gene families evolve. *Theoretical Population Biology 37*, 1 (1990), 213–219.

[45] Roesler, J., Curnutte, J. T., Rae, J., Barrett, D., Patino, P., Chanock, S. J., and Goerlach, A. Recombination events between the p47-phoxgene and its highly homologous pseudogenes are the main cause of autosomal recessive chronic granulomatous disease. *Blood 95*, 6 (2000), 2150–2156.

[46] Rozen, S., Skaletsky, H., Marszalek, J. D., Minx, P. J., Cordum, H. S., Waterston, R. H., Wilson, R. K., and Page, D. C. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature 423*, 6942 (2003), 873–876.

[47] Sawyer, S. Statistical tests for detecting gene conversion. *Molecular Biology and Evolution 6*, 5 (1989), 526–538.

[48] Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T., McCarthy, S., Montgomery, S. H., Schwalie, P. C., Tang, Y. A., Ward, M. C., Xue, Y., Yngvadottir, B., Alkan, C., Andersen, L. N., Ayub, Q., Ball, E. V., Beal, K., Bradley, B. J., Chen, Y., Clee, C. M., Fitzgerald, S., Graves, T. A., Gu, Y., Heath, P., Heger, A., Karakoc, E., Kolb-Kokocinski, A., Laird, G. K., Lunter, G., Meader, S., Mort, M., Mullikin, J. C., Munch, K., O'Connor, T. D., Phillips, A. D., Prado-Martinez, J., Rogers, A. S., Sajjadian, S., Schmidt, D., Shaw, K., Simpson, J. T., Sten-

son, P. D., Turner, D. J., Vigilant, L., Vilella, A. J., Whitener, W., Zhu, B., Cooper, D. N., de Jong, P., Dermitzakis, E. T., Eichler, E. E., Flicek, P., Goldman, N., Mundy, N. I., Ning, Z., Odom, D. T., Ponting, C. P., Quail, M. A., Ryder, O. A., Searle, S. M., Warren, W. C., Wilson, R. K., Schierup, M. H., Rogers, J., Tyler-Smith, C., and Durbin, R. Insights into hominid evolution from the gorilla genome sequence. *Nature 483*, 7388 (2012), 169–75.

[49] Ségurel, L., Wyman, M. J., and Przeworski, M. Determinants of mutation rate variation in the human germline. *Annual Review of Genomics and Human Genetics 15* (2014), 47–70.

[50] Smith, G. P. Unequal crossover and the evolution of multigene families. In *Cold Spring Harbor Symposia on Quantitative Biology* (1974), vol. 38, Cold Spring Harbor Laboratory Press, pp. 507–513.

[51] Smith, G. P., Hood, L., and Fitch, W. M. Antibody diversity. *Annual Review of Biochemistry 40*, 1 (1971), 969–1012.

[52] Storz, J. F., Baze, M., Waite, J. L., Hoffmann, F. G., Opazo, J. C., and Hayes, J. P. Complex signatures of selection and gene conversion in the duplicated globin genes of house mice. *Genetics 177*, 1 (2007), 481–500.

[53] Sugino, R. P., and Innan, H. Selection for more of the same product as a force to enhance concerted evolution of duplicated genes. *Trends in Genetics 22*, 12 (2006), 642–644.

[54] Taghian, D. G., and Nickoloff, J. A. Chromosomal double-strand breaks induce gene conversion at high frequency in mammalian cells. *Molecular and Cellular Biology 17*, 11 (1997), 6386–6393.

[55] Takuno, S., and Innan, H. Selection to maintain paralogous amino acid differences under the pressure of gene conversion in the heat-shock protein genes in yeast. *Molecular Biology and Evolution 26*, 12 (2009), 2655–2659.

[56] Teshima, K. M., and Innan, H. The effect of gene conversion on the divergence between duplicated genes. *Genetics 166*, 3 (2004), 1553–1560.

[57] Teshima, K. M., and Innan, H. Neofunctionalization of Duplicated Genes Under the Pressure of Gene Conversion. *Genetics 1398*, March (2007), 1385–1398.

[58] Thornton, K., and Long, M. Excess of amino acid substitutions relative to polymorphism between x-linked duplications in drosophila melanogaster. *Molecular Biology and Evolution 22*, 2 (2005), 273–284.

[59] Turner, R., and Liu., L. *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.

[60] Watnick, T. J., Gandolph, M. A., Weber, H., Neumann, H. P., and Germino, G. G. Gene conversion is a likely cause of mutation in pkd1. *Human Molecular Genetics 7*, 8 (1998), 1239–1243.

[61] Whelden Cho, J., Khalsa, G. J., and Nickoloff, J. A. Gene-conversion tract directionality is influenced by the chromosome environment. *Current Genetics 34*, 4 (1998), 269–279.

[62] Williams, A. L., Genovese, G., Dyer, T., Altemose, N., Truax, K., Jun, G., Patterson, N., Myers, S. R., Curran, J. E., Duggirala, R., Blangero, J., Reich, D., and Przeworski, M. Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *Elife 4* (2015), e04637.

[63] Yang, D., and Waldman, A. S. Fine-resolution analysis of products of intrachromosomal homeologous recombination in mammalian cells. *Molecular and Cellular Biology 17*, 7 (1997), 3614–3628.

[64] Zimmer, E., Martin, S., Beverley, S., Kan, Y., and Wilson, A. C. Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. *Proceedings of the National Academy of Sciences 77*, 4 (1980), 2158–2162.