nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-018-0778-x

Direct estimation of mutations in great apes reconciles phylogenetic dating

Søren Besenbacher[1*], Christina Hvilsom[2], Tomas Marques-Bonet[3], Thomas Mailund[4] and Mikkel Heide Schierup[4,5*]

The human mutation rate per generation estimated from trio sequencing has revealed an almost linear relationship with the age of the father and the age of the mother, with fathers contributing about three times as many mutations per year as mothers. The yearly trio-based mutation rate estimate of around 0.43×10^{-9} is markedly lower than previous indirect estimates of about 1×10^{-9} per year from phylogenetic comparisons of the great apes calibrated by fossil evidence. This suggests either a slowdown in the accumulation of mutations per year in the human lineage over the past 10 million years or an inaccurate interpretation of the fossil record. Here we inferred de novo mutations in chimpanzee, gorilla, and orangutan parent-offspring trios. Extrapolating the relationship between the mutation rate and the age of parents from humans to these other great apes, we estimated that each species has higher mutation rates per year by factors of 1.50 ± 0.10, 1.51 ± 0.23, and 1.42 ± 0.22 for chimpanzee, gorilla, and orangutan, respectively, and by a factor of 1.48 ± 0.08 for the three species combined. These estimates suggest an appreciable slowdown in the yearly mutation rate in the human lineage that is likely to be recent as genome comparisons almost adhere to a molecular clock. If the nonhuman rates rather than the human rate are extrapolated over the phylogeny of the great apes, we estimate divergence and speciation times that are much more in line with the fossil record and the biogeography.

The current mutation rate in humans has been extensively studied through the sequencing of thousands of parent-offspring trios[1–3]. The consensus is that the rate increases almost linearly with the age of the parents, with a higher rate of mutations from the father (2.51 mutations per year) than from the mother (0.78 mutations per year), yielding a mutation rate of 0.43×10^{-9} per base pair per year[2]. The mutation rate per year is essential for the calibration of the molecular clock to enable dating of evolutionary events in human ancestry, such as human population divergences, human-Neanderthal divergence and the split time with ancestors of the other great apes, such as chimpanzees, gorillas, and orangutans[4–6]. If we extrapolate the present yearly human mutation rate estimated from trios as discussed above, we predict the average time of divergence across the genome from chimpanzees of more than 15 million years ago, from gorillas about 19 million years ago, and from orangutans about 35 million years ago. These estimates are difficult to reconcile with the fossil record, which puts the upper bound of human-orangutan speciation to around 20 million years ago[5,7–10]. It is possible that the mutation rate could have decreased over time in the lineage that leads to humans. However, the phylogeny of the great apes almost supports a molecular clock that has branches for chimpanzee, gorilla and orangutan that are only 2–3%, 6–7%, and 11%, respectively, longer than the human branch[7]. A decrease in the human mutation rate per year would, therefore, either have to be very recent or have to have decreased independently in the lineages that lead to each of the great apes, perhaps because of the generally longer generation time[7,11,12]. One way to distinguish among these alternatives is to determine the present mutation rate in other great apes from sequencing of parent-offspring trios. The first such study using a pedigree with six chimpanzee trios[13] reported a mutation rate per generation that was very similar to that in humans and a more male-biased contribution to mutations. However, the trios used had parents that were younger (average age of fathers, 18.9 years; average age of mothers, 18.5 years) than the human generation time of 29 years; therefore, the rate per year in these trios are expected to be higher than in humans. A recent study using deeper sequencing for a single trio suggests a higher rate than in humans by about 30% per year[14].

Here we extended the trio sequencing in great apes by one additional extended trio of chimpanzees, two extended trios of gorillas, and one trio of orangutan and reanalyse the data from the previous study of six chimpanzee trios. We then used the detailed knowledge of the parental effects on mutation rates in humans to turn generation-based mutation rates in the other great apes into an estimate of the yearly mutation rate, which is relevant for phylogenetic dating. We report a mutation rate estimate of each great ape species that significantly exceeded the rate in humans with a point estimate of about 50%. If we parsimoniously assume a recent slowdown in the yearly human rate, then extrapolation of the nonhuman rates to the phylogeny of the great apes is well-aligned with fossil evidence.

Results

We called mutations in extended trios for chimpanzees, gorillas, and orangutans sequenced to high coverage and combine our results with a reanalysis of the raw data from a previous study[13]. Reads were mapped against the reference genomes of each of the three species. Finding the optimal quality threshold for estimating the mutation rate is a balance between avoiding false-positive calls while analysing as much of the genome as possible[15]. Our approach was to use a range of quality thresholds and estimate the callable proportion of the genome probabilistically for each of these (see Methods and Supplementary Figs. 1,2). We conservatively chose

[1]Department of Molecular Medicine (MOMA), Aarhus, Denmark. [2]Copenhagen Zoo, Frederiksberg, Denmark. [3]Institut Biologia Evolutiva, Universitat Pompeu Fabra/CSIC, Barcelona, Spain. [4]Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark. [5]Department of Bioscience, Aarhus University, Aarhus, Denmark. *e-mail: besenbacher@clin.au.dk; mheide@birc.au.dk

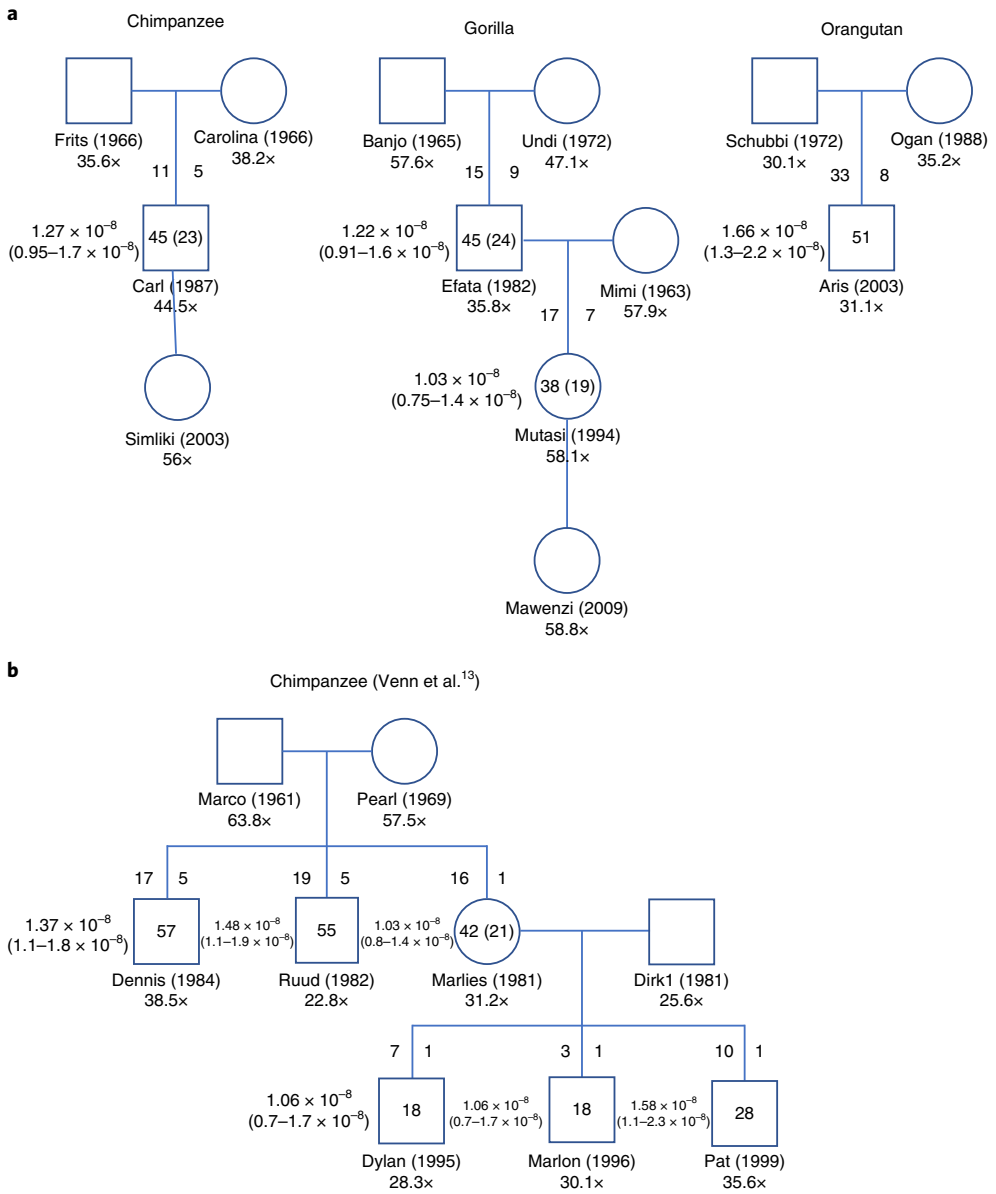

Fig. 1 | Numbers, rates, and transmission of de novo mutations. a, The sequenced pedigrees of chimpanzee, gorilla, and orangutan with names, birth years, and sequence coverage. All rates are based on genotype qualities ≥65 (see Methods). The number of mutations observed within individuals and the number of transmitted mutations are indicated in parentheses. To the left of the transmission line is the number of mutations inferred to have arisen in the father, on the right the number of mutations inferred to have arisen in the mother. Next to the individuals is the inferred rate and 95% confidence interval (in parentheses). **b**, Results from the reanalysis of the chimpanzee pedigree sequenced in a previous study[13].

phred scaled genotype quality Q65 as the threshold at which the estimated mutation rate does not change by increasing the quality (see Supplementary Fig. 1 for values for other quality thresholds). At Q65, we called a total of 397 mutations (Supplementary Table 1). The number of mutations called for each of the three species, together with our reanalysis of the previously published sequencing dataset from six chimpanzee trios of an extended pedigree[13] is shown in Fig. 1. The estimated rates take the number of callable sites into account (presented in Supplementary Table 1) and are shown for each trio in Fig. 1. The number of mutations that could be assigned to parental origin using read-backed phasing is also shown. The distributions of mutations along the genome in the trios show a slight clustering as reported for humans (Supplementary Fig. 3) and there are no significant differences in their composition (Supplementary Fig. 4) among species. Our mutation rate pipeline is essentially the same as the one we previously used in humans, and which we showed has a very low false-positive rate[1,15]. However, it is conceivable that the lower quality of the reference genomes that are currently available for the great apes might increase the chance of false positives. We therefore separately estimated the rates for the repetitive and nonrepetitive parts of the reference genomes and found both rates to be very similar (Fig. 2a). We also found that the number of transmitted mutations to the grandchild is not significantly different from 50% (Supplementary Fig. 5). Finally, we conservatively removed all mutations (a total of 12) that were located in segmental duplications in the genomes of the great apes. For the final set of mutations, BAM files were visually inspected for multiple mutation events, which affects at most 12 mutations (three in Aris, three in Efata, and six in Pat). These could be enriched in alignment artefacts, but were kept in the analysis. If these were all false positives,

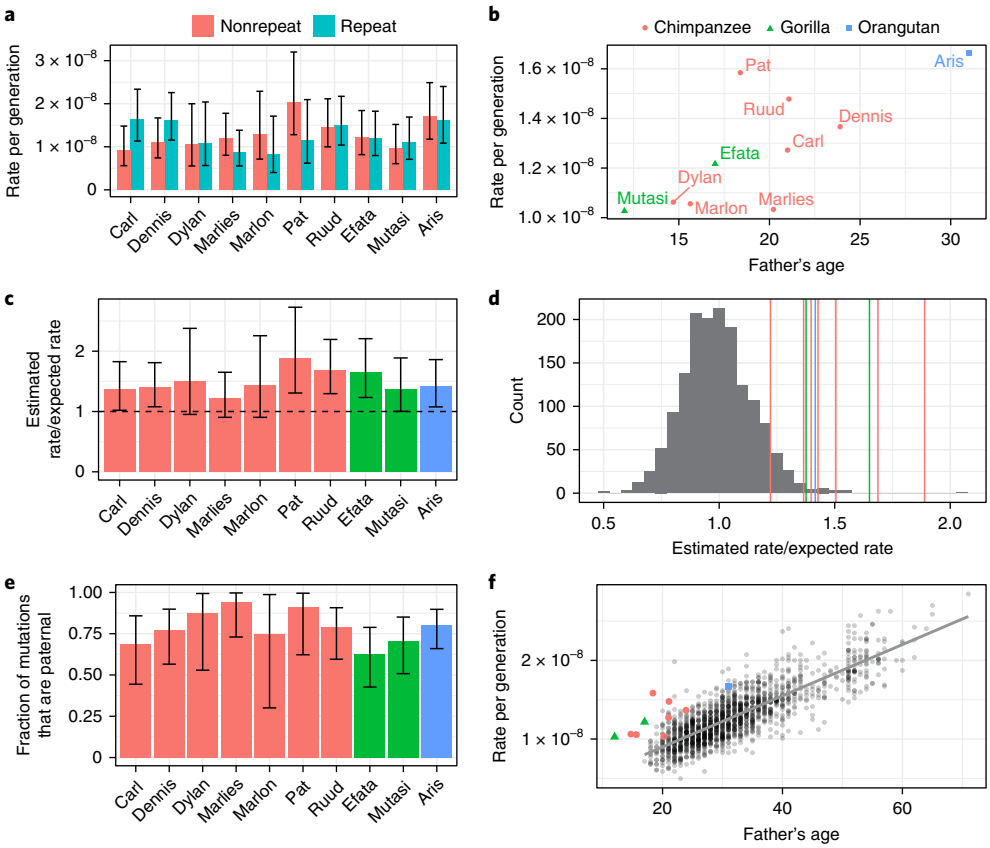

Fig. 2 | Properties of de novo mutations. a, The estimated per generation rate for each pedigree divided into the nonrepetitive and repetitive parts (using repeatmasker) of the genome (bars show 95% confidence intervals). The numbers can be found in Supplementary Tables 1 and 2. No significant differences between the regions were found for any of the individuals. **b**, The estimated mutation rate as a function of paternal age, significant positive slope for all species $P = 0.014$ (95% confidence interval, 8.694361×10^{-11} to 5.738461×10^{-10}), but not for chimpanzees alone ($P = 0.28$). **c**, The estimated rates compared to the expected rates extrapolating the human relationship (95% confidence intervals are shown). **d**, The histogram shows the distribution of estimated/expected rates for the 1,548 human trios from a previously published study[2] and the vertical lines show the estimated/expected rates for the analysed great ape trios. **e**, The fraction of the mutations for which the parent-of-origin could be assigned to the father (95% confidence intervals are shown). **f**, The estimated mutation rate as a function of paternal age for human trios (grey points, with linear regression) and for the great ape trios investigated in the present study.

they would affect our estimates for Pat by 20%, but not the combined estimate for chimpanzees by more than 3%. The estimated mutation rate correlated with paternal age across species (Fig. 2b).

To perform phylogenetic dating, we turned these generation estimates of the mutation rate with parents of known age into a yearly rate. In order to do this, we assumed linear relationships with parental ages and a fixed contribution at birth. We extrapolated the relationship with maternal and paternal age that has been calculated for a large number of human trios[2] to the other great ape species. This relationship in humans is estimated[2] as: mutation rate $= 1.77 \times 10^{-9} + 7.26 \times 10^{-11} \times$ maternal age (years) $+ 2.87 \times 10^{-10} \times$ paternal age (years). There is a relatively small, constant per-generation contribution (approximately 14% of the total if we assume a human generation time for males and females of 29 years) and the yearly rate in humans therefore depends little on changes in generation time. Extrapolations of this relationship to each offspring in the trios of great apes from the age of its parents at the time of birth thus provide us with an estimate of the mutation rate that is expected for this combination of parental ages. We can then compare this rate to the observed rate to determine whether the yearly contributions from parents are the same for the different species of great ape (see Table 1). Figure 2c and Table 1 show that under this strong assumption (discussed below), the estimated yearly rates for chimpanzee, gorilla and orangutan are each significantly higher than in humans, but not significantly different from each other (see Supplementary Fig. 2 for the yearly rates as a function of quality thresholds). Figure 2d shows these ratios compared to the range observed in 1,548 human trios[2], which is shown as a function of paternal age in Fig. 2f.

The mutation rate in humans is male-biased, possibly as a result of a higher number of cell divisions in the male germline than in the female germline and/or less efficient repair of double-strand breaks in the male germline. In all three great ape species, we also found that the majority of mutations passed on from the father, with the male bias highest in chimpanzee as reported previously[13], but the differences among the species are not statistically significant from each other, neither different from humans (Fig. 2e and Table 1). This supports the validity of extrapolating the human relationships with parental ages. Among the mutations that can be assigned to the parental origin (50–80% for the different trios), the number of mutations passed on increase significantly with paternal age for species combined and for chimpanzee separately but not with maternal age (Supplementary Fig. 6).

Our yearly estimates for nonhuman great apes suggest that the human rate has recently decreased significantly and this should have direct consequences for the phylogenetic dating of the speciation time and genomic divergence of great apes. Here we defined speciation time as the time of cessation of gene flow and noted that it was smaller than the genomic divergence time because of

Table 1 | Basic statistics for de novo mutation calling

Child	Species	Father age (years)	Mother age (years)	Callable base pairs	Total number of mutations	CpG	Non-CpG strong	Weak	Observed rate	Expected rate	Relative rate
Carl	Chimpanzee	21	21	3,536,629,178	45	8	24	13	1.27×10^{-8}	9.32×10^{-9}	1.36
Pat	Chimpanzee	18.39	18.48	1,767,339,554	28	2	15	11	1.58×10^{-8}	8.39×10^{-9}	1.89
Dennis	Chimpanzee	23.9	15.89	4,170,326,188	57	11	26	20	1.37×10^{-8}	9.78×10^{-9}	1.40
Ruud	Chimpanzee	21.07	13.06	3,720,595,013	55	18	21	16	1.48×10^{-8}	8.77×10^{-9}	1.69
Marlies	Chimpanzee	20.22	12.22	4,064,760,290	42	9	20	13	1.03×10^{-8}	8.46×10^{-9}	1.22
Dylan	Chimpanzee	14.7	14.79	1,693,813,311	18	6	8	4	1.06×10^{-8}	7.06×10^{-9}	1.50
Marlon	Chimpanzee	15.63	15.72	1,704,746,791	18	4	7	7	1.06×10^{-8}	7.40×10^{-9}	1.43
Efata	Gorilla	17	10	3,698,229,861	45	6	19	20	1.22×10^{-8}	7.38×10^{-9}	1.65
Mutasi	Gorilla	12	31	3,700,254,515	38	7	18	13	1.03×10^{-8}	7.46×10^{-9}	1.38
Aris	Orangutan	31	15	3,065,633,840	51	12	13	26	1.66×10^{-8}	1.18×10^{-8}	1.42

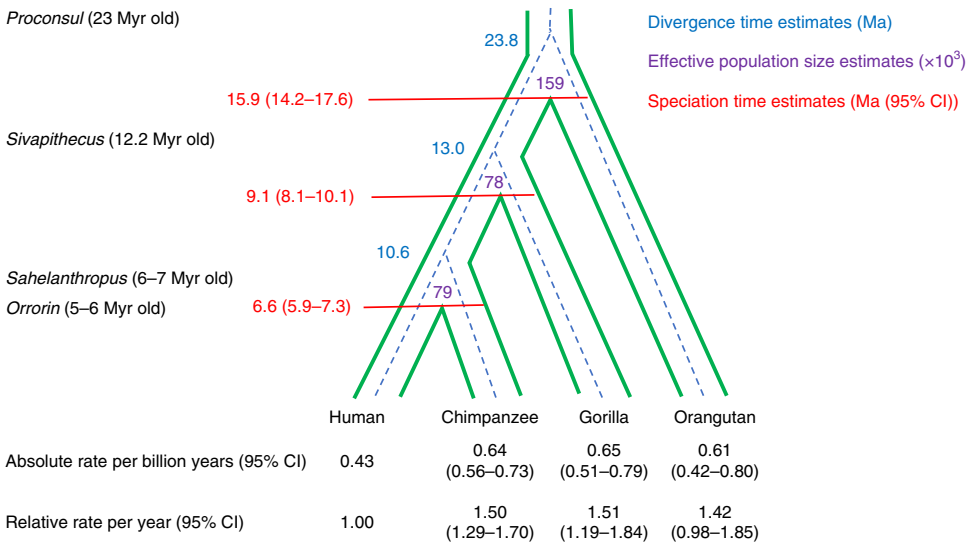

Fig. 3 | Estimates of the genomic divergence time and species separation times of the great apes. The estimated mutation rate per year for nonhuman great apes combined is extrapolated over the phylogeny to yield divergence time estimates (blue numbers) that are more recent than estimates based solely on the current estimate of the human mutation rate (15, 19, and 35 million years, respectively). Estimated ancestral effective population sizes (purple numbers, in thousands) are used to estimate speciation separation time (speciation time, red numbers (95% confidence interval (CI))). Absolute and relative rates per year and 95% confidence intervals were calculated based on extrapolations of the human rate, combined over individuals for each species. Ma, million years ago; Myr, million years. See Methods for details.

ancestral polymorphisms. We first estimated the divergence time of each species by combining the estimates from all nonhuman great apes, propagating their standard errors and including the reported deviations from the molecular clock (Fig. 3 and Methods). We then used previous estimates of ancestral effective population sizes[10], calibrated to the new rates, to estimate the speciation times (Fig. 3).

The phylogeny of the great apes is very close to adhering to a molecular clock in which the chimpanzee branch is slightly longer than the human branch (by 2%) and the gorilla branch is longer by about 6%[7]. For this to be consistent with our estimated rates, a slowdown in mutation rate in the human lineage must be recent. If a decrease of 33% occurred instantly on the human branch, the point estimate would be around 400,000 years to explain a 2% shorter human than chimpanzee branch. It is therefore possible that even the dating of the human-Neanderthal divergence is affected. This is in accordance with mutation rate estimates using the recombination clock being higher than trio based human rates by about 20%[6] and consistent with the branch-shortening estimates obtained from a 45,000-year-old human fossil[16].

Discussion

We report evidence for an apparent very recent slowdown in the yearly mutation rate in the human lineage. This conclusion rests on the assumption that we can extrapolate the relative contribution from parents and the almost linear relationship with age of both parents that is observed in humans to other great apes. We estimated a similar proportion of mutations coming from each parent as in humans supporting the first assumption. Our sequenced trios had slightly younger parents (average male parent 19.5 years, female parent 16.7 years) than the estimated generation time in nature (25 years for chimpanzees, 19 years for gorillas). Thus, if early mutations are more prevalent because of earlier puberty (see below), then we might slightly overestimate the yearly mutation rate for nonhuman great apes from parents that are younger than in nature by about 4%

using the previously published model[11]. Furthermore, recent studies suggest a superlinear increase for older mothers, in particular, suggesting that we might underestimate the yearly rate[17–19]. It is currently debated which proportion of the new mutations are a result of replication errors in the germline and how many are a result of repair of nonreplicative damage, particularly during meiosis and in the first few cell diversions in the zygote[17,19,20]. The latter can also be time-dependent if repairs become less efficient with age, particular in the first few zygotic cell divisions[17,19]. Another reason for our higher estimated rate may be that we have a larger number of false-positive mutation calls than in human studies, perhaps because the nonhuman reference genomes are of lower quality. However, because we used the same pipeline, found no difference between repetitive and nonrepetitive rates and no enrichment in complex parts of the genome, we think this probably did not cause much of an effect. We note that in most studies it is the denominator (that is, the callable fraction of the genome that we estimate probabilistically) in the rate estimation that causes the most significant uncertainty.

On the assumption that our estimated yearly rates are correct, we can analyse the causes of the very recent human slowdown in mutation rate. It may be partly explained by humans having a later onset of puberty and a longer generation time (human estimate, 29 years; chimpanzee, 24 years; gorilla, 19 years; orangutan, 25 years)[2,11,21]. Later onset of puberty suggests that there are more years with very few mutations accumulating in males. However, although these life-history characteristics can explain the differences in branch lengths in the phylogeny of the great apes[11], they do not explain a reduction in the mutation rate by about 1/3 in human trios compared to great ape trios. For example, assuming a chimpanzee time to puberty of 8 years compared to 13.5 years in humans would only lead to a rate increase in chimpanzee by 14% according to the previously published models[11,21].

Estimates of the mutation rate in humans further back in time using the recombination clock have been reported to be 0.55 ± 0.05 per base pair per billion years[6]. This rate falls almost exactly in between the trio-based rate in humans today and our estimated rates. It is possible that a large part of the slowdown in humans is very recent and perhaps caused by lifestyle or environmental differences in the predominantly Caucasian populations for which trios have been investigated for de novo mutations. If a significant proportion of de novo mutations are because of postzygotic repair of double-stranded lesions that occur in sperm or egg cells, then recent changes to environmental exposures may also contribute to the observed slowdown[17]. Sequencing of a large number of great ape parent-offspring trios and human populations other than Caucasians, as well as the molecular investigation of the repair processes and their dependency on parental age, are needed to resolve this.

The estimates of speciation time of great apes based on the mutation rate estimate that assumes a human slowdown push all speciation events closer to the present time. The human-chimpanzee separation time of 6.6 million years (95% confidence interval, 5.9–7.3 million years) that we estimated is consistent with *Ardipithecus*[22] being on the human line, *Orrorin* (5–6 million years old[23]) being very early on the human lineage and *Sahelanthropus* (6–7 million years old) being right at the split between human and chimpanzee. A species separation time of 15.9 million years ago (95% confidence interval,14.2–17.6 million years ago) with orangutan is well within the time interval between the age of *Sivapithecus* (12.2 million years old[24]) and *Proconsul* (23 million years old), which are generally assumed to be the lower and upper bounds for the human–orangutan divergence, respectively (see Fig. 3).

Methods

Samples. All blood samples were taken during routine health checks and Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) permits were obtained from countries outside the European Union.

Sequencing. Genomic DNA was extracted directly from EDTA whole-blood samples using a DNeasy Blood and Tissue Kit (Qiagen), following the manufacturer's instructions. Then, 2 μg DNA was used for the construction of PCR-free libraries with an average insert size of 250 base pairs. The libraries were sequenced on Illumina HiSeq X instruments by Novogene using standard chemistry for paired-end sequencing of 2×150 base pairs to coverage between 30× and 56× (see Fig. 1).

Reanalysis of previously published chimpanzee data. We downloaded the fastQ files from ftp://ftp.well.ox.ac.uk/panPed and analysed the data from the previously published study[13] alongside our own sequencing data using the same pipelines.

Mapping. Reads were aligned to the following reference genomes of the great apes as indicated: chimpanzee, Pan_tro 3.0 (UCSC: panTro5); gorilla, GorGor4.1 (UCSC: gorGor4); orangutan WUGSC2.0.2 (UCSC: ponAbe2) (BWA version 0.7.15). The average mapped coverage of the new data is shown in Fig. 1 and Supplementary Table 1, and 90.8% of the aligned genome covered by at least 15 reads.

Variant calling. The realigned and base quality score recalibrated BAM files were used as input for multi-sample genotyping for each species separately using the HaplotypeCaller of the Genome Analysis Toolkit version 3.8.

Identification of de novo single-nucleotide variation. To limit the number of false positives, we only consider a Mendelian violation as a possible de novo mutation if both parents in the family in question are homozygotes for the reference allele and if the variant is not called in any of the other families. We apply these filters when we look for de novo mutations.

A site filter was used to analyse the reads from all individuals to filter away bad sites that are not true variants. The site filter uses the following parameters: (1) FS: Fisher's exact test on strand bias. (2) ReadPosRankSum: rank-sum test on the position of the alternative allele in the reads. (3) BaseQualityRankSum. (4) MappingQualityRankSum.

Then, individual filters were used to analyse the reads and genotype calls of a single individual to discard a possible de novo call if we were not sure that all of the individuals in the family in question were called correctly. We use two different kinds of individual filters. First, a homozygote reference filter was applied to the parents to check that we have confidence that they are homozygous for the reference allele. The filter uses the following parameters: (1) GQ: genotype quality of the individual. (2) DP: number of reads for this individual at this site. (3) AD2: number of times the alternative allele is seen in this individual. (4) lowQ_AD2: the number of low-quality reads (not used in the calling) that contained the alternative allele. (5) RC: the number of reads in this individual relative to the average coverage of this individual. Second, a heterozygote filter was applied to the child to ensure that the child is heterozygous at this site. The filter uses the following parameters: (1) GQ: genotype quality of the individual. (2) DP: the number of reads for this individual at this site. (3) AlleleBalance: the fraction of the reads in the individual that contains the alternative allele. (4) minStrandCount: the minimum number of counts of the alternative variant on each strand. (5) RC: the number of reads in this individual relative to the average coverage of this individual.

Estimating callable sites for de novo mutations. We calculate the denominator of the rate estimate following the same strategy as previously described[15].

To achieve a better estimate of the rate of de novo mutations in a trio, we base the denominator of the rate estimate on the probability at each site that we can call de novo mutation rather than merely counting a site as either callable or uncallable. The probability of calling site x as a de novo mutation given that it is a true de novo mutation in family f, we name the callability and denote it by $C_f^{\text{de novo}}(x)$. The callability can be estimated independently for each family based on the sequencing depth for the family members at the site, and the expected number of callable sites in a given family is then the sum of the callability of all sites in that family.

Because the site filter is based on statistical tests that follow a known distribution, we can estimate how many good sites we expect to be filtered away by this filter by looking at the null distribution of the tests and assuming that the two tests are independent. We denote by α_{site} the fraction of good sites that we expect to be filtered away.

The mutation rate of a family f can then be estimated as:

$$\text{rate}_f = \frac{|\text{de novo mutations in } f \text{ after filtering}|}{2(1-\alpha_{\text{site}}) \sum_{x \in \text{sites}} C_f^{\text{de novo}}(x)}$$

Now let Z be a genotype (heterozygote (hetero), homozygote, reference allele (HomRef) or homozygote, alternative allele (HomAlt)) and consider for an individual i the probability of calling it as Z at position x (and not filtering it away) given that the individual truly is Z at x. We denote this conditional probability by $C_i^Z(x)$, and it signifies the ability to give a true call of Z at x. This will be a function of sequencing quality at x (not least the depth). If we assume that the ability to truly call each member of a family is independent, then the callability of a site in

a given family can be calculated as the probability of calling each individual correctly after filtering:

$$C_f^{\text{de novo}}(x) = C_c^{\text{Hetero}}(x) \times C_p^{\text{HomRef}}(x) \times C_m^{\text{HomRef}}(x)$$

Where c, p, and m indicate the child, father and mother of family f.

Assuming that $C_c^{\text{Hetero}}(x)$ is independent of the parental genotypes as long as they are conducive to a heterozygous offspring, we can estimate it by considering only variants for which one parent has a homozygous reference allele with high confidence, and the other parent is homozygous for the alternative allele. At such sites, the child should always be a heterozygote (barring de novo events). Using only these sites we can estimate:

$$C_c^{\text{Hetero}}(x) = \frac{|S_{d(c,x)}^{\text{Hetero}}|}{|S_{d(c,x)}^{\text{All}}|}$$

where $d(c, x)$ is the depth at x for child c, S_d^{All} are variant–family pairs (f', x') for which the child c' has depth d at variant x' and one of the parents is HomRef for the variant and the other parent is HomAlt, after applying the sites filter and a conservative filter on the genotype quality of the parents and S_d^{Hetero} is the subset of S_d^{All} for which the child is called as heterozygous and pass the heterozygote filter.

Similarly, we can calculate:

$$C_i^{\text{HomRef}}(x) = \frac{|M_{d(i,x)}^{\text{HomRef}}|}{|M_{d(i,x)}^{\text{All}}|}$$

where i is either m or p and M_d^{All} indicates variant–family pairs (f', x') for which the child c' has depth d, and both parents in each family in question are HomRef for the variant, the variant is present in at least one of the other families after applying the sites filter and a conservative filter on the genotype quality of the parents.

M_d^{HomRef} is the subset of M_d^{All} for which the children are called as homozygous for the reference allele and pass the homozygote filter.

Minimizing false-positive de novo mutation calls. Although the estimation of callability, as described above, reduces the effect of false negatives on the estimated mutation rate, it is still necessary to set the cut-offs in the filters so high that only very few or no false positives get into the set of estimated de novo mutations. We can fit the filter criteria by looking at the effect of different criteria on the rate estimate and the effect on how large a fraction of the called de novo variants are present in unrelated individuals from the same species (Supplementary Fig. 1).

On the basis of these considerations we set the filter values at: (1) GQ ≥ 65 (for both the homozygote and heterozygote filters); (2) DP ≥ 10 (for both the homozygote and heterozygote filters); (3) RC < 1.9 (for both the homozygote and heterozygote filters); (4) AD2 = 0 (for the homozygote filter); (5) lowQ_AD2 = 1; (6) AlleleBalance > 0.3; (7) minStrandCount = 1.

The AlleleBalance filter was set on the basis of the distribution of AlleleBalance in the children after applying the other filters.

Segmental duplications in each of the great ape reference genomes were identified using the tool ASGART[25] with default parameters.

Parent of origin assignment of de novo mutations. We estimate the paternal origin of each de novo variant using the previously described strategy[1].

For each variant, X, we use $o(X)$ to denote the parental origin of the alternative allele. The reads might provide conflicting evidence and to find the most likely parental origin, we calculated a likelihood ratio comparing probability that the alternative allele is on the paternal chromosome ($o(X) = 1$) to the probability that the alternative allele is on the maternal chromosome ($o(X) = 0$):

$$\begin{aligned} \text{LR}_X &= \frac{P(o(X)=1|\text{data})}{P(o(X)=0|\text{data})} \\ &= \prod_{\substack{Y \in X \text{variant less than} \\ 30 \text{ kb from } X \\ \text{with POO assigned}}} \frac{\sum_{\beta \in \{0,1\}} P(o(Y)=\beta) \times P(\text{data}|o(X)=1, o(Y)=\beta)}{\sum_{\beta \in \{0,1\}} P(o(Y)=\beta) \times P(\text{data}|o(X)=0, o(Y)=\beta)} \end{aligned}$$

If LR_X is above one, it indicates that the alternative allele of variant X is on the paternal chromosome and if LR_X is below one, it indicates that it is on the maternal chromosome. The data that is informative about the parent of origin (POO) are the reads that cover both X and Y:

$$P(\text{data}|o(X), o(Y)) = \prod_{\substack{r_{XY} \in \text{ reads} \\ \text{covering } X \text{ and } Y}} P(r_{XY}|o(X), o(Y))$$

The probability that a read supports the correct phasing is 1 if the read is mapped correctly and 0.5 if the read is not mapped correctly. We calculated the conditional probability of the read as:

$$P(r_{XY} = (i,j)|o(X) = \alpha, o(Y) = \beta) = \begin{cases} P(r_{XY} \text{ correct}) & (i = j \wedge \alpha = \beta) \vee (i \neq j \wedge \alpha \neq \beta) \\ \quad + (1 - p(r_{XY} \text{ correct}))/2 & \\ (1 - p(r_{XY} \text{ correct}))/2 & (i \neq j \wedge \alpha = \beta) \vee (i = j \wedge \alpha \neq \beta) \end{cases}$$

where $P(r_{XY} \text{ correct})$ is the probability that r_{XY} is mapped correctly (estimated from the phred score in the BAM file) and the values of i and j are either the reference or alternative allele, depending on whether the read contains the reference allele or the alternative allele at position X and Y. For inherited variants for which the parental origin could be assigned by just looking at the genotypes of the family members, $P(o(Y) = 1)$ is calculated using the phred-scaled genotype probabilities of the three family members. If the parent of origin of variant Y has been assigned using read information, we calculate $P(o(Y) = 1)$ from the estimated LR: $P(o(Y) = 1) = \text{LR}_Y/(\text{LR}_Y + 1)$. The assignment of parental origin is carried out iteratively until no additional variants can be assigned.

Estimation of the yearly mutation rate. The mutation rate estimates for each of the trios were converted into yearly estimates by extrapolating from the relationship between mutation rate and maternal and paternal parent age observed in humans. In humans, the best estimate is that the mutation rate in a child depends on the parental ages as follows:

$$\text{Mutation rate} = 1.77 \times 10^{-9} + 7.26 \times 10^{-11} \times \text{maternal age} + 2.87 \times 10^{-10} \times \text{paternal age}$$

From this relationship, we calculate the expected mutation rate for each great ape trio taking parental ages into account and assuming that the overall mutation rate per year is the same as in humans. The relative mutation rate for each trio is then calculated as the observed rate divided by the expected rate from the relationship in humans with 95% confidence intervals.

Phylogenetic dating. We produced estimates of genomic divergence rates using the average yearly, nonhuman mutation rate over the ten trios in the following way. From a previous study[10], we obtained the following average genomic divergences between human and chimpanzee of 0.0137; human and gorilla of 0.0175; and human and orangutan of 0.034.

The phylogeny of the great apes deviates slightly from a molecular clock according to a previous study[7], with the chimpanzee branch being 2% longer than the human branch, the gorilla branch 6% longer than the human branch since their common ancestry and the orangutan branch 11% longer than the human branch since their common ancestry. Using these numbers and focusing on the human branch, the branch lengths from human to the common ancestor with the chimpanzee becomes 0.006713, with gorilla 0.008225 and with orangutan 0.01513.

Using the estimated average yearly mutation rate from the nonhuman trios (0.634 per billion years (95% confidence interval, 0.565–0.704)), this corresponds to 10.88 million years for the average human–chimpanzee genomic divergence time, 12.96 million years for the average human–gorilla genomic divergence time and 23.85 million years for the average human–orangutan genomic divergence time.

To turn the divergence numbers into estimates of species separation time (here equal to speciation time), we used the ancestral effective population sizes reported previously[10] scaled to the nonhuman mutation rate assumed in the common ancestors yielding: human-chimpanzee ancestral effective population size = 79,435; human-gorilla ancestral effective population size = 77,544; human-orangutan ancestral effective population size = 158,871.

Since the expected coalescence time in the common ancestors is $2N$, the separation times are calculated as:

$$T_{\text{species}} = T_{\text{divergence}} - 2N_{\text{anc}} u_{\text{anc}} G$$

Where G is the generation time, here assumed to be 25 years (approximate average of generation times in extant species: humans, 29 years; chimpanzee, 24 years; gorilla, 19 years; orangutan, 25 years). This yields the following estimates: human-chimpanzee speciation time = 6.61 million years (95% confidence interval, 5.89–7.33); human-gorilla speciation time = 9.09 million years (95% confidence interval, 8.09–10.08); human-orangutan speciation time = 15.90 million years (95% confidence interval, 14.16–17.64). Numbers are summarized in Fig. 3.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All sequence data have been deposited in the European Nucleotide Archive (ENA) under accession number PRJEB29710. The identified de novo mutations are included as Supplementary Table 1. All scripts and code used to generate the results are available at https://github.com/besenbacher/GreatApeMutationRate2018.

Received: 18 June 2018; Accepted: 5 December 2018;
Published online: 21 January 2019

References

1. Maretty, L. et al. Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. **548**, 87–91 (2017).
2. Jónsson, H. et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).
3. Kong, A. et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
4. Prüfer, K. et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655–658 (2017).
5. Scally, A. The mutation rate in human evolution and demographic inference. *Curr. Opin. Genet. Dev.* **41**, 36–43 (2016).
6. Lipson, M. et al. Calibrating the human mutation rate via ancestral recombination density in diploid genomes. *PLoS Genet.* **11**, e1005550 (2015).
7. Moorjani, P., Amorim, C. E., Arndt, P. F. & Przeworski, M. Variation in the molecular clock of primates. *Proc. Natl Acad. Sci. USA* **113**, 10607–10612 (2016).
8. Moorjani, P., Gao, Z. & Przeworski, M. Human germline mutation and the erratic evolutionary clock. *PLoS Biol.* **14**, e2000744 (2016).
9. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.* **13**, 745–753 (2012).
10. Scally, A. et al. Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169–175 (2012).
11. Amster, G. & Sella, G. Life history effects on the molecular clock of autosomes and sex chromosomes. *Proc. Natl Acad. Sci. USA* **113**, 1588–1593 (2016).
12. Scally, A. Mutation rates and the evolution of germline structure. *Phil. Trans. R. Soc. Lond. B* **371**, 20150137 (2016).
13. Venn, O. et al. Strong male bias drives germline mutation in chimpanzees. *Science* **344**, 1272–1275 (2014).
14. Tatsumoto, S. et al. Direct estimation of de novo mutation rates in a chimpanzee parent–offspring trio by ultra-deep whole genome sequencing. *Sci. Rep.* **7**, 13561 (2017).
15. Besenbacher, S. et al. Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat. Commun.* **6**, 5969 (2015).
16. Moorjani, P. et al. A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years. *Proc. Natl Acad. Sci. USA* **113**, 5652–5657 (2016).
17. Gao, Z. et al. Overlooked roles of DNA damage and maternal age in generating human germline mutations. Preprint at https://www.biorxiv.org/content/early/2018/10/11/327098 (2018).
18. Goldmann, J. M. et al. Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. *Nat. Genet.* **50**, 487–492 (2018).
19. Lindsay, S. J., Rahbari, R., Kaplanis, J., Keane, T. & Hurles, M. Striking differences in patterns of germline mutation between mice and humans. Preprint at https://www.biorxiv.org/content/early/2018/05/23/082297 (2018).
20. Jónsson, H. et al. Multiple transmissions of de novo mutations in families. *Nat. Genet.* **50**, 1674–1680 (2018).
21. Thomas, G. W. C. et al. Reproductive longevity predicts mutation rates in primates. *Curr. Biol.* **28**, 3193–3197 (2018).
22. White, T. D. et al. *Ardipithecus ramidus* and the paleobiology of early hominids. *Science* **326**, 64–86 (2009).
23. Richmond, B. G. & Jungers, W. L. *Orrorin tugenensis* femoral morphology and the evolution of hominin bipedalism. *Science* **319**, 1662–1665(2008).
24. Pilbeam, D., Rose, M. D., Barry, J. C. & Shah, S. M. New *Sivapithecus* humeri from Pakistan and the relationship of *Sivapithecus* and *Pongo*. *Nature* **348**, 237–239 (1990).
25. Delehelle, F., Cussat-Blanc, S., Alliot, J.-M., Luga, H. & Balaresque, P. ASGART: fast and parallel genome scale segmental duplications mapping. *Bioinformatics* **34**, 2708–2714 (2018).

Acknowledgements

We thank M. Przeworski and P. Moorjani for comments to a previous version of the manuscript and A. Brandstrup for laboratory expertise. The study was supported by grant number 6108-00385A from the Danish Council for Independent Research | Natural Sciences (to M.H.S.).

Author contributions

S.B., T.M., T.M.-B., C.H., and M.H.S. designed the study. C.H. contributed reagents. S.B., T.M., and M.H.S. performed analysis. S.B. and M.H.S. wrote the paper with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-018-0778-x.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.B. or M.H.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

nature research

Corresponding author(s): Søren Besenbacher, Mikkel Heide Schierup

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	not relevant
Data analysis	For analysis we used the following software bwa version 0.7.15 GATK version 3.8 samtools version 1.5.0 asgart version 1.1.0. Custom python and R scripts are available to editors/reviewers on request and will be available on github before publication

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All sequence data will be submitted to EGA.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Bloods samples from four individuals of a chimpanzee three generation family, six individuals of a four generation gorilla family and three individuals from a two generation orangutan family were sequenced by Illumina sequencing with the aim of discovering de novo mutations and estimate the de novo mutation rate
Research sample	All blood samples were taken during routine health checks and Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) permits were obtained from countries outside the EU.
Sampling strategy	The sample size was limited by available samples
Data collection	The data collection procedure was carried out by Christina Hvilsom at the Copenhagen Zoo
Timing and spatial scale	Blood samples were from the Copenhagen Zoo biobank
Data exclusions	No data was excluded
Reproducibility	The sequencing data will be freely available as will the computer code generated for analyses so the study is reproducible from the primary sequence data
Randomization	Not relevant
Blinding	Not relevant

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	The study did not involve laboratory animals
Wild animals	The study did not involve wild animals. Only zoo animals were included in the study.
Field-collected samples	The study did not involve samples collected in the field.