Human germline mutation and the erratic molecular clock

Priya Moorjani[+,*,#], Ziyue Gao[&,*,#] and Molly Przeworski[+,%,#]

[+] Dept. of Biological Sciences, Columbia University
[&] Howard Hughes Medical Institute & Dept. of Genetics, Stanford University
[%] Dept. of Systems Biology, Columbia University
* These authors contributed equally.
[#] To whom correspondence should be addressed: pm2730@columbia.edu, ziyuegao@stanford.edu and mp3284@columbia.edu

Abstract

Our understanding of the chronology of human evolution relies on the "molecular clock" provided by the steady accumulation of substitutions on an evolutionary lineage. This understanding has been called into question by recent analyses of human pedigrees, suggesting that mutations accrue more slowly than previously believed. Translating mutation rates estimated from pedigrees into substitution rates is not as straightforward as it may seem, however. In this Unsolved Mystery, we dissect the steps involved, emphasizing that dating evolutionary events requires not "a mutation rate," but a precise characterization of how germline mutations accumulate in development, in males and females—knowledge that remains elusive.

Introduction

One of the most fundamental discoveries in evolutionary biology is the "molecular clock": the observation that changes to the genome along an evolutionary lineage accumulate steadily with time [1-3] and the subsequent development of a theory—the Neutral theory—that explains why [4,5]. We now understand that neutral genetic changes (i.e., changes with no fitness effects) fix in the population at the rate at which they arise, irrespective of demographic history or natural selection at linked sites [4,6]. Thus, the accumulation of neutral substitutions over generations provides a record of the time elapsed on an evolutionary lineage. It is this molecular clock that allows researchers to date evolutionary events.

Conversely, the existence of a molecular clock allows the number of substitutions on an evolutionary lineage to be translated into a yearly mutation rate, given an independent estimate of when that evolutionary lineage branched off [2,7-9]. For example, interpreting the fossil record as reflecting a 30 million year (My) divergence time between humans (apes) and rhesus macaques (Old World Monkeys (OWM)) and using the average nucleotide divergence of ~6.2% between the two species [10] suggests an average yearly mutation rate of 10^{-9} per base pair (bp). Until maybe five years ago, single nucleotide substitutions were the main source of data from which to learn about mutation rates, and analyses of substitution patterns consistently suggested rates around 10^{-9} per bp per year for primates [9,11-13].

Recent findings in human genetics therefore threw a spanner in the works when they suggested de novo point mutation rates estimated from human pedigrees to be less than half what was previously believed, or approximately 0.4×10^{-9} per bp per year [14-18]. Because sequencing pedigrees is a much more direct and in principle definitive approach to learn about mutation, these new rate estimates have been widely adopted. They have led to a reappraisal of the chronology of human evolution, suggesting in particular that populations split longer ago than previously believed (e.g., [17,19]). Extrapolating farther back in time becomes problematic however, as pedigree-based

estimates imply split times with other primates that are older than compatible with the fossil record, at least as currently interpreted [20-24] (D. Pilbeam, personal communication). One possible solution, suggested by Scally and Durbin (2012) [17] as well as others, is that yearly mutation rates have decreased towards the present, consistent with the "hominoid rate slowdown" observed in phylogenetic data [25-27].

As we discuss, changes in the yearly mutation rate over the course of human evolution are not only plausible, but follow from first principles. The expected number of de novo mutations inherited by a child depends on paternal (and possibly maternal) ages at puberty and reproduction [26,28-32], traits that differ markedly among extant primates [21,33,34]. Because these traits evolve, there is no fixed mutation rate per generation, and almost certainly no fixed mutation rate per year. An important implication is that the use of mutations as a molecular clock requires a precise characterization of how germline mutations accumulate in development, in males and females, and across species. This knowledge is still elusive and, as a result, it remains unclear how to calibrate the human molecular clock. For recent time depths, a complementary approach from the study of ancient DNA samples may offer a solution.

The puzzle

Heritable mutations stem from accidental changes to the genome that occur in the development of the germline and production of egg and sperm. A natural definition of the germline mutation rate "per generation" is therefore the rate at which differences arise between the genome of a newly formed zygote and the gametes that it eventually produces. While this quantity cannot be readily measured, it has recently become possible to estimate something highly related, the number of mutations seen in the genome of an offspring's soma but absent from the parents' [14] (henceforth μ_G). At least eleven whole genome studies have applied this approach, resequencing parents and offspring, usually in trios. They reported estimates of μ_G on the order of 10^{-8} per bp (Table 1).

Although the trio studies were primarily conducted to identify de novo disease mutations, they also inform our understanding of the chronology of primate evolution. Assuming that changes to the genome are neutral, the expected number of substitutions that accumulate on a lineage, d, equals μt, where μ is the mutation rate and t is the expected length of the lineage (e.g., the human lineage since it diverged from chimpanzee). Thus, given an estimate of μ and orthologous sequences from more than one species, an estimate of t can be obtained from d/μ. In practice, researchers are interested in an estimate of t in years, not generations, and therefore require an estimate of the yearly mutation rate, $\hat{\mu}_y$. To obtain it, common practice has been to divide the $\hat{\mu}_G$ obtained from sequencing of parents and children by the average age of parents in the study (at conception). Doing so suggests $\hat{\mu}_y$= 0.3-0.5x10^{-9} per bp (Table 1).

Taken at face value, this mutation rate suggests a split time between human populations of over 100,000 years [17,19] and a human-chimpanzee divergence time of 12-19 Mya (for a human-chimpanzee average nucleotide divergence of 1.2%) [10,17,20]. These estimates are older than previously believed, but may well be correct. Less sensible are the divergence times that are obtained for humans and orangutans or humans and OWMs. As an illustration, using whole genome divergence estimates [10] and a yearly mutation rate of 0.5×10^{-9} per bp suggests a human-orangutan divergence time of 31 Mya and human-OWM divergence time of 62 Mya. These estimates are implausibly old, implying a human-orangutan divergence well into the Oligocene and OWM-hominoid divergence well into or beyond the Eocene [35,36] (D. Pilbeam, personal communication). Thus, the yearly mutation rates obtained from pedigrees seem to suggest dates that are too old to be readily reconciled with the current understanding of the fossil record.

Another way of viewing the same problem is to compare values of $\hat{\mu}_y$ obtained from resequencing pedigrees to those obtained from divergence levels among primates, given estimates of divergence times *t* based on the fossil record. Such estimates of *t* are highly indirect, in part because the fossil record is sparse and in part because relying on fossils with derived traits provides only a lower bound for when that lineage branched off [22,23]. A further complication is that for closely related species, notably humans and chimpanzees, *t* reflects the time since the species split as well as the contribution of ancestral polymorphism, which can be substantial, in particular for great apes [20,37]. Thus, this approach is mired in uncertainty. Nonetheless, until recently, the consensus in the field has been to use $\hat{t}$ values of 6-7.5 million years ago (Mya) for humans and chimpanzees [9,38], 15-20 Mya for humans and orangutans [39] and 25-35 Mya for humans and OWMs [26,39,40]. Assuming these values and solving $\hat{\mu}_y = \hat{d}/2\hat{t}$ suggests a mutation rate of 10^{-9} per year, more than two-fold higher than what is obtained from pedigree-based estimates. In other words, accepted divergence times suggest that the molecular clock ticks faster than do pedigree studies of mutation.

These discrepancies have led to considerable discussion of whether our understanding of primate evolution is incorrect. In this Unsolved Mystery, we argue that in some ways this question is premature. Indeed, while our current understanding of the fossil record could be inaccurate, there is underappreciated complexity in the conversion of mutation rates from pedigrees into mutation rates per year, and their translation into substitution rates, that remains to be resolved. We try to unpack this complexity, by discussing each step in turn: (1) what it is we are truly estimating from resequencing pedigrees; (2) what we have learned to date and what we have yet to understand; (3) and how to translate the mutation rates into evolutionary dates (Fig 1).

Step1: What exactly is being estimated from human pedigrees?

Human pedigree studies have relied primarily on blood samples from trios, estimating the total number of mutations present in ~50% of reads from the child but absent in both parents. A mutation *rate* is obtained by dividing this count by the number of base pairs for which there was complete power to detect de novo mutations, or equivalently, dividing it by the genome length, adjusting for the power to detect mutations at a typical position in the genome.

Because the mutation rate is so low (~10^{-8} per bp per generation), it is challenging to reliably identify de novo mutations using sequencing technologies with error rates on the order of 10^{-2} per read, and given the presence of cryptic copy number variation, alignment uncertainty and other confounders [41,42]. Detection pipelines therefore have high false discovery rates, and a stringent set of filters on sequence complexity, read depth, and allelic balance (i.e., requiring close to 50% read depth) have to be applied to weed out spurious mutations [43]. This aggressive filtering process substantially increases specificity but decreases the number of sites at which mutations can be detected, so power has to be carefully assessed for any given set of filters.

An additional complication is "mosaicism", that is the presence of two or more genotypes in a given population of cells. This term is not well defined in that, strictly speaking, all mutations in the germline are mosaic unless they arise in the last cell division before the formation of egg or sperm. In practice, the term often refers to mutations that occurred in the development of the zygote, rather than in the production of egg and sperm. If carried by germline cells, such mosaic mutations should be included when estimating the germline mutation rate per generation (Fig 2). In practice, however, neither the parents nor the offspring is sampled as a zygote; instead, blood samples are used. In these somatic samples, some of the mutations that are detected will have arisen during somatic tissue development and renewal, and should not be considered as a germline mutation (Fig 2; [44]; M. Georges, personal communication). At the same time, blood samples are only a small subset of somatic cells in an individual, so even when ~50% of reads support the presence of a mutation, it is unclear whether this mutation is truly constitutional (i.e., heterozygous in all cells) of the individual, or only present in the blood. Furthermore, when a mutation is supported by few reads, it is hard to distinguish sequencing errors from low-level mosaicism. Aiming to guard against false positives, standard pipelines require allelic balance in the child and exclude any potential mutations present at appreciable read depths in the parent. This procedure leads to the inclusion of some mutations that arose during the development of the child (especially at early stages) and the exclusion of a fraction of true germline mutations in the parents (Fig 2). Whether it inflates or deflates the estimated number of mutations per generation depends on the precise filters, the sequencing error rate and the mutation rate per cell division in various stages of development. On balance, it appears that mutation rates during early embryogenesis are likely to be somewhat underestimated due to the inability to detect mutations that

arise in early cell divisions in the parents; by how much remains unclear ([45,46]; M. Georges, personal communication).

In addition to these technical considerations, there are conceptual subtleties in interpreting the mutation rate estimates from intergenerational studies. As expected a priori and from older studies of disease incidences in children [47,48], all large pedigree studies published to date have reported a linear effect of the age of the father on the number of de novo mutations inherited by a child (Table 1). Because spermatogenesis occurs continuously after the onset of puberty, the number of replication-driven mutations inherited by a child is expected to depend on paternal age—more precisely, on the age at which the father enters puberty, his rate of spermatogonial stem cell divisions and age at reproduction (Fig 1). Therefore, the observation that the number of mutations increases linearly with paternal age is consistent with a fixed rate of cell division after puberty and a constant rate of mutation per cell division during spermatogenesis. In contrast, oocytogenesis is completed by the birth of the future mother, so the number of replication-driven mutations inherited by an offspring should be independent of maternal age (Fig 1). For the subset of mutations that do not stem from mistakes during replication—mutations that arise from DNA damage and are poorly repaired for example—there may be a dependence on maternal age as well, if damage accumulates in oocytes [31]. Interestingly, recent studies report for the first time that a maternal age effect is also present, potentially supporting the existence of a non-replicative source of germline mutations [49,50]. In any case, what is clear is that the number of de novo mutations in a child is a function of the age of the father at conception and possibly that of the mother, so values obtained from pedigree studies are estimates of mutation rate *at given mean paternal* (and maybe maternal) ages of the sampled families.

Another complication is that distinct types of mutations may differ in their accrual rates with age, depending on their source and repair rates over ontogenesis [31,51]. For instance, transitions at methylated CpG sites are thought to occur primarily by spontaneous deamination; beyond this example, the DNA molecule is known to be subject to a large number of chemical assaults from normal cellular metabolism and additional environmental agents [52,53]. While the relative contribution of germline mutations from different sources is unclear, their accrual rates with parental age are unlikely to be identical (even if the differences are hard to detect in small samples [45]). Therefore, the mutation rate estimated from pedigree studies is the composite of distinct mutational processes that have distinct dependencies on age and sex, making the time-dependency of the overall mutation rate harder to interpret (Fig 1).

With these considerations in mind, what have we learned to date? All large-scale pedigree studies report similar mutation rates per generation, a strong male bias in mutation, and a paternal age effect. On closer inspection, however, their parameter estimates are not consistent. To illustrate this point, we report the estimated mutation

rate at paternal age of 30 years, which differ by as much as 40% (Table 1). Given the relatively small sample sizes, some uncertainty is expected from sampling error alone. However, differences in sequencing technology and choice of filters are also likely to be playing a role. As one illustration, the fraction of mutations that involve transitions from CpG sites differs significantly among studies, from 11% to >20% (Chi-Square test, $p < 10^{-8}$). At least some of this variation appears to be due to whether the studies excluded mutations seen in dbSNP [54] as spurious. As databases become larger, this step increasingly leads to the exclusion of true mutations [55], with a disproportionate effect on CpG transitions, which are more mutable [56].

Among studies, there is also three-fold variation in the estimated strength of the paternal age effect (Table 1), which remains significant after accounting for the fraction of the genome surveyed for mutation (Fig 3). In principle, differences in the paternal age effect among studies could reflect true biological differences. In this regard, we note a recent report that fathers differ significantly in their paternal age effects (Fig 3), pointing to possible inter-male differences in rates of spermatogonial stem cell divisions or in mutation rates per cell division [45]. If such inter-male differences exist, however, then it becomes problematic to estimate the strength of the paternal age effect by pooling data from multiple families. When a single line is fit to data from fathers who have distinct paternal age effects, the resulting slope will likely not be the average slope and, in principle, could even lie outside the range of the true slope for each family (Fig 4). Whether, in practice, this phenomenon contributes to variation in the estimated paternal age effect across studies remains to be seen. To this end, an important step will be to conduct studies of larger nuclear families.

In summary, while pedigree-based approaches are more direct and in principle straight-forward, they have not yet provided a definitive answer about the mutation rate at any given paternal and maternal ages, let alone a precise characterization of how mutations of different sources accumulate over ontogeny in males and females.

Step 2: How to obtain a yearly de novo mutation rate?

Even if the germline mutation rate per generation, μ_G, were known exactly, strong assumptions are required to translate the per generation mutation rates of the sampled families into a yearly rate. Common practice has been to obtain a yearly mutation rate by dividing the mutation rate estimated from all the children by the average age of their parents (i.e., setting $\hat{\mu}_y = \hat{\mu}_G/\bar{G}$, where $\bar{G}$ is the mean age of parents in the study, at conception).

This practice is only valid if μ_G increases in strict proportion to $\bar{G}$. Yet modeling suggests that this condition will only hold under very specific conditions about the development and renewal of germ cells, and (limited) available data suggest that it is not met in human or chimpanzee [18,31]. If this condition does not hold, at least two problems

emerge: first, the value of $\hat{\mu}_y$ is only comparable across studies if the same $\bar{G}$ is used, rather than whatever happens to be the average parental age in the study. Second, because the expectation of a ratio is not the ratio of expectations, the estimate of $\hat{\mu}_y$ will be biased. Also complicating matters are possible differences among fathers within a study in their age at puberty [57]. These differences will lead to distinct relationships between μ_G and G for each couple, introducing an additional source of bias in estimating a yearly mutation rate from $\hat{\mu}_y = \hat{\mu}_G/\bar{G}$. Moreover, if μ_y is not independent of G, then it becomes important to ask whether $\bar{G}$ for the (predominantly European) samples is representative of the human species, when it is known that ages at reproduction differ substantially across populations [58].

Step 3: How to relate μ to the substitution rate expected over evolutionary time?

Changes in life history and reproductive traits. Mammalian species vary over three-fold in substitution rates, indicating that the yearly mutation rates change over time [30,59,60]. In primates, in particular, 35-65% variation is seen in substitution rates across apes and monkeys [10,11]. The cause of variation in substitution rates was long hypothesized to be a "generation time effect", whereby younger mean ages of reproduction—i.e., shorter generation times—lead to more cell divisions per unit time and hence higher rates of replication-driven mutations [28,30]. Support for this claim comes from phylogenetic analyses of mammals, in which reproductive span is a predictor of mutation rates per year [26,28,30]. As we have discussed, a dependence of yearly mutation rates on generation times is also expected from what is known of mammalian sperm and egg production. Thus, to accurately convert mutation rates per generation into expected substitution rates per year, changes in the generation time over evolution need to be taken into account.

Doing so requires knowledge of numerous parameters that are currently uncertain or simply unknown. A solvable problem is that the conversion depends on the precise dependence of μ_G on parental ages [18], about which there remains considerable uncertainty (Fig 3). A thornier issue is that the yearly substitution rate depends not only on the sex-averaged generation time, but also on the mean ages at reproduction for males and females separately. The reason being that in males, the germline mutation rate depends more strongly on reproductive age than it does in females; thus, for the same average parental age, de novo mutation rates are much lower in a child born to a young father and an old mother than in a child born to an old father and a young mother. As a result, changing the ratio of male to female generation times can have substantial effects on the yearly mutation rate, even when the average remains fixed: for example, a range of ratios from 0.92 to 1.26, as observed in extant hominines, could lead to up to 10% difference in μ_y, and thus introduce uncertainty in phylogenetic dating [32].

Beyond the effect of generation times, the yearly mutation rate will vary with any heritable change in life history traits (e.g., the age at puberty) and germ line developmental process (e.g., the number of cell divisions in each development stage). We know that among extant primates, the onset of puberty differs substantially, from ~1 year in marmosets to 6-13 years in apes [33], as does the length of spermatogonial stem cell divisions [61]. Thus, life history traits can and have evolved across primates. This evolution introduces additional uncertainty in the yearly mutation rate expected at any point in the past [32]. Moreover, these factors influence μ_y in intertwined ways, so it is important to consider their co-evolution [10,32].

Changes in the mutation process. Thus far, we have only discussed sources of changes in the yearly mutation rate due to development and life history, but another layer of evolution occurs at cellular level, in terms of mutational processes of DNA [62-64]. Could the rates of replication error, DNA damage or DNA repair have evolved over millions or even thousands of years? One study, for example, compared the spectra of rare segregating variants among populations, and found that Europeans, but not Africans or Asians, had an increased rate of a specific mutation type (TCC -> TTC), which is highly enriched among somatic mutations in melanoma [64]. This observation raises the possibility of recent evolutionary changes in the mutation process itself.

While a change in mutation rates of a specific mutation type is parsimoniously explained by a change in the damage or repair rates, modeling suggests that, even in the absence of such changes, life history traits alone could shift the relative contributions of mutations of different sources [31]. As one example, CpG transitions appear to be more clock-like across species than do other types of mutations (possibly due to a weaker dependence on life history traits) [10,59] and accordingly, the proportion of substitutions due to CpG transitions varies across species [10]. As another example, an increase in paternal age leads not only to an increase in the total germline mutation rate but also to an increase in the proportion of mutations in genic regions (0.26% per year) [65], which should lead to shifts in the mutation spectrum. More generally, it remains highly unclear how much of the differences in mutation rates across populations or species can be attributed to changes in life history and behavior, in the development and renewal of germ cells, in genetic modifiers of mutation (such as enzymes involved in DNA replication and repair) or in the environment (e.g., in the concentration of external mutagens).

Selection and biased gene conversion. Lastly, even if we were able to obtain a reliable estimate of the average yearly mutation rate for some time period, the equation of mutation and substitution rates is only valid under neutrality. The substitution rate of a population can be factored into two components: the rate at which mutations arise in the population and the probability that a mutation is eventually fixed in the population. When changes are neutral, larger populations experience a greater input of mutations, but exactly counterbalancing this effect is a smaller probability of fixation for each

mutation. When positive or negative selection is operating, however, the probability of fixation deviates from the neutral expectation, so the substitution rates at sites under selection are not expected to equal the mutation rate. To minimize this problem, researchers have estimated divergence in regions that are less likely to be targets of direct selection (e.g., pseudogenes, fourfold-degenerate sites or genomes with conserved regions excluded) [7,9,10]. Nonetheless, this filtering process is likely imperfect. In this regard, the de novo mutation rate estimated from pedigrees provides an upper bound for the substitution rate (as it includes deleterious mutations that will not reach fixation). Thus, if anything, the ticking of the molecular clock should be *slower* than expected from pedigree-based mutation rates.

The fixation probability can also be affected by mechanisms other than natural selection. The best-known instance is GC-biased gene conversion (BCG), a process that preferentially resolves mismatches in heteroduplex DNA arising from meiotic recombination in favor of strong alleles (C or G) over weak alleles (A or T). This asymmetry leads to the preferential fixation of mutations from A/T to C/G and decreases the fixation rate of C/G to A/T mutations. Clear evidence for this process is seen both in mammalian substitution patterns and in human pedigree data [66,67]. How big an effect this phenomenon has had on skewing substitution patterns is hard to quantify, as it depends on the demographic history as well as local recombination rates, which are evolving rapidly, but a recent estimate suggests that BGC tracts contain 1.2% of human-chimpanzee single nucleotide substitutions [68]. Thus, it seems to have been a relatively minor force in increasing substitution rates to GC beyond what is expected from mutation rates, and a similar effect in different primate lineages [10].

Next steps

To obtain a molecular clock from pedigree data is not as straightforward as it may seem. The main reason being that there is no such thing as a mutation rate per generation—all that exists is a mean mutation rate for a given set of paternal and maternal life history traits, including ages at puberty and reproduction. These traits are variable among closely related primates [33,34], and heritable variation is seen even among humans [57,58]. Therefore, primate species are expected to differ substantially in both the per generation mutation rate and the yearly mutation rate (e.g., see Table S9 in [32]).

Indeed, phylogenetic analyses show that, over millions of years, substitution rates vary >60% among distantly related primates [10]. The observed variation in substitution rates observed across primate lineages appears to be smaller than that predicted from life history traits in extant species, however [10,32]. A likely reason is that, throughout much of their evolutionary past, the lineages had similar life histories. Direct surveys of de novo mutation rates in non-human primates are therefore needed.

So far, the only direct estimate of mutation rate in a non-human primate is based on one three-generation pedigree of chimpanzees [69]. The point estimate of the mutation rate at age 30 is higher in chimpanzees than in humans (Table 1), qualitatively consistent with an earlier onset of puberty and faster rate of spermatogenesis [33,61]. Given the differences in detection pipelines, random sampling error and potential intra-species variation, however, these results are still tentative. Both inter- and intra- species variation in mutation rates need to be further characterized in primates.

If mutation rates turn out to vary substantially across species, it will be interesting to examine whether they are well predicted by typical ages at puberty and reproduction, without the need to invoke other changes, such as differences in per cell division mutation rates. This finding would imply that, over evolutionary timescales, the yearly mutation rate is less variable than mutation rate per generation, contrary to what is usually assumed (e.g., [28,70]).

If, on the other hand, despite clear differences in life history traits, the per generation mutation rate across primate species turns out to be relatively constant, it would follow that strong stabilizing selection or developmental constraint must have shaped the evolution of mutation rates. It would also follow that species with longer generation times will have lower yearly mutation rates, providing stronger support for the "generation time effect" than can be obtained from phylogenetic evidence [26,28,30].

That yearly mutation rates are expected to be unsteady poses difficulties for the use of the molecular clock to date evolutionary events. One solution is to explicitly model the changes in life-history traits over the course of primate evolution and to study their impact on substitution rates. To this end, Amster and Sella (2016) [32] proposed a model that estimates divergence and split times across species, accounting for differences in sex-specific life history and reproductive traits. A next step will be to extend their model to consider replicative and non-replicative mutations separately. In addition, as more reliable estimates of mutational parameters become available from pedigree studies of humans and non-human primates, models will need to be revised to account for differences in cell division rates and possible differences in repair rates. Unfortunately, however, some uncertainty will remain due to lack of knowledge about life history traits in ancestral lineages.

An alternative might be to use only CpG transitions for dating. This solution is based on the observation that CpG transitions accumulate in a quasi-clocklike manner across primates [10,59], as well as across human populations [64]. Puzzlingly, however, in human pedigree data, there is no detectable difference between the effects of paternal age on CpG transitions and other types of mutations [16,45], suggesting that CpG transitions are no more clock-like. In that regard, it will be highly relevant to compare accrual rates of CpG transitions in pedigree studies from multiple primate species.

In addition to the use of pedigree studies, two other types of approaches have been introduced recently to learn about mutation rates. The first is a set of ingenious methods that use population genetic modeling to estimate mutation rates based on segments of the genome inherited from a distant common ancestor [71,72]. Unfortunately, these methods rely on detailed demographic assumptions or on extremely precise estimates of fine-scale meiotic recombination rates (and to obtain yearly rates, on estimates of generation times), so the resulting estimates of the mutation rates remain quite uncertain.

The second approach is to use precisely-dated ancient DNA samples to estimate average yearly mutation rates over different evolutionary periods. In this method, the divergence from an extant sample (e.g., human) to an outgroup (e.g., chimpanzee) is compared to what is seen between a precisely-dated ancient genome and the outgroup. The "missing divergence" then provides an estimate of the average mutation rate per year over that timescale. Applied to archaic human samples from the past 50,000 years, this approach suggests yearly rates around 0.5×10^{-9} per bp [73]. The study of many such ancient samples distributed over the past tens of thousands of years could potentially serve as "spike ins" for the molecular clock, allowing one to adjust for distensions and contractions over different time periods—at least when studying individuals with ancestry from similar populations.

Together, these approaches will both inform us about how to reliably set the molecular clock and provide a first direct look at the evolution of mutation rates over different time scales.

Acknowledgments

We are grateful to Kay Prüfer and David Reich for organizing "The Human Mutation Rate Meeting" at the Max Planck Institute for Evolutionary Anthropology in February 2015, and to all the participants for many enlightening discussions. We also thank Augustine Kong, Raheleh Rahbari, Ryan Yuen, Laurent Francioli, Paul de Bakker and Wendy Wong for providing data used in generating Table 1 and Fig 3, and David Pilbeam for helpful discussions about the fossil record.

Fig 1. The many steps involved in the conversion of mutation rate estimates from pedigree studies into yearly substitution rates.

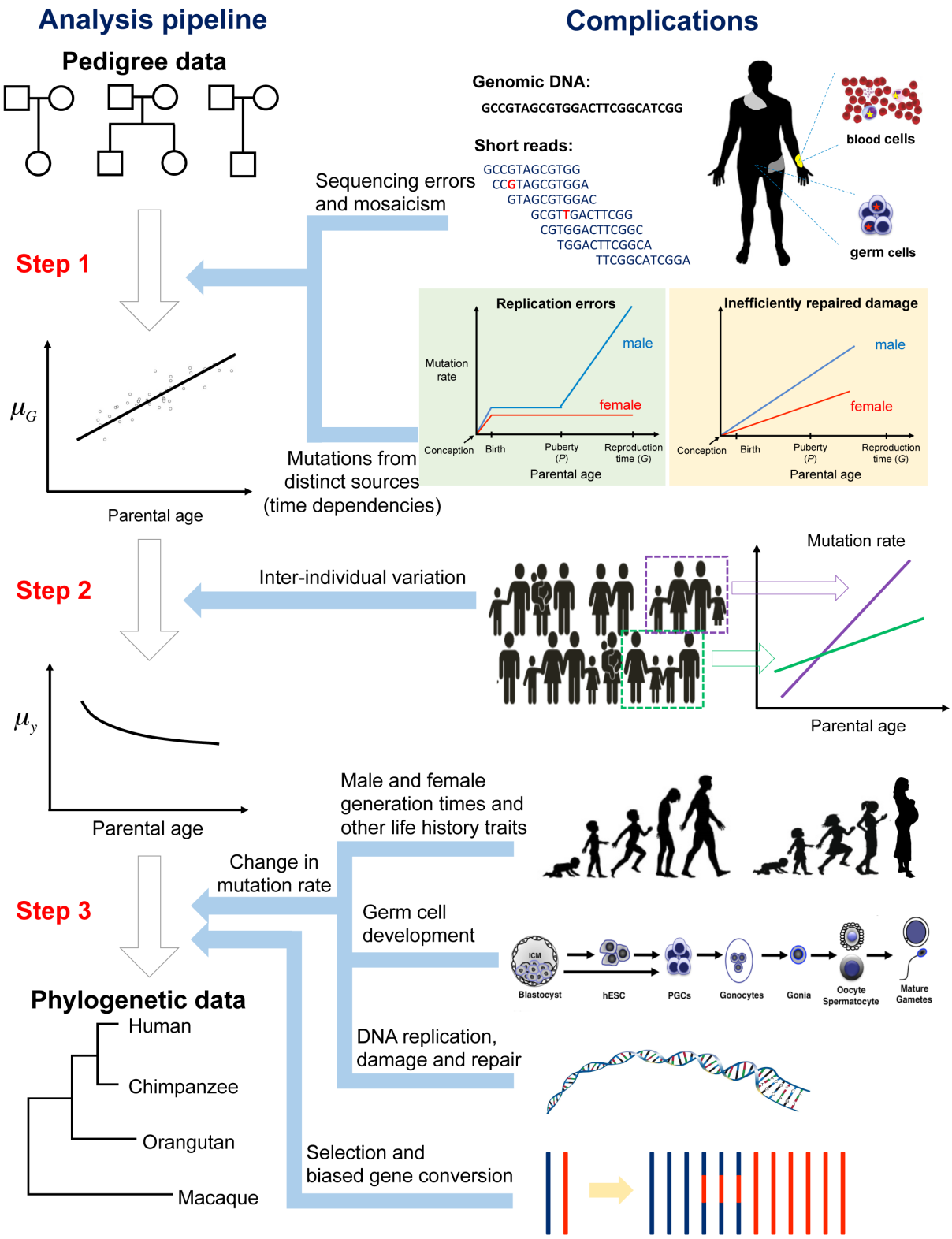

Fig 2. Schematic illustration of mutations occurring during embryonic development and gametogenesis. Stars represent mutations that arise in different stages of embryogenesis and gametogenesis of the parents and the offspring; filled stars are mutations that arise in the parents and hollow stars those that occur in the offspring. Shown below each individual are the expected frequencies of mutations in his or her blood sample. Red, brown and green stars are heritable and should be included in an estimate of germ line mutation rates, whereas blue stars are somatic mutations present in blood samples only, which should be excluded.

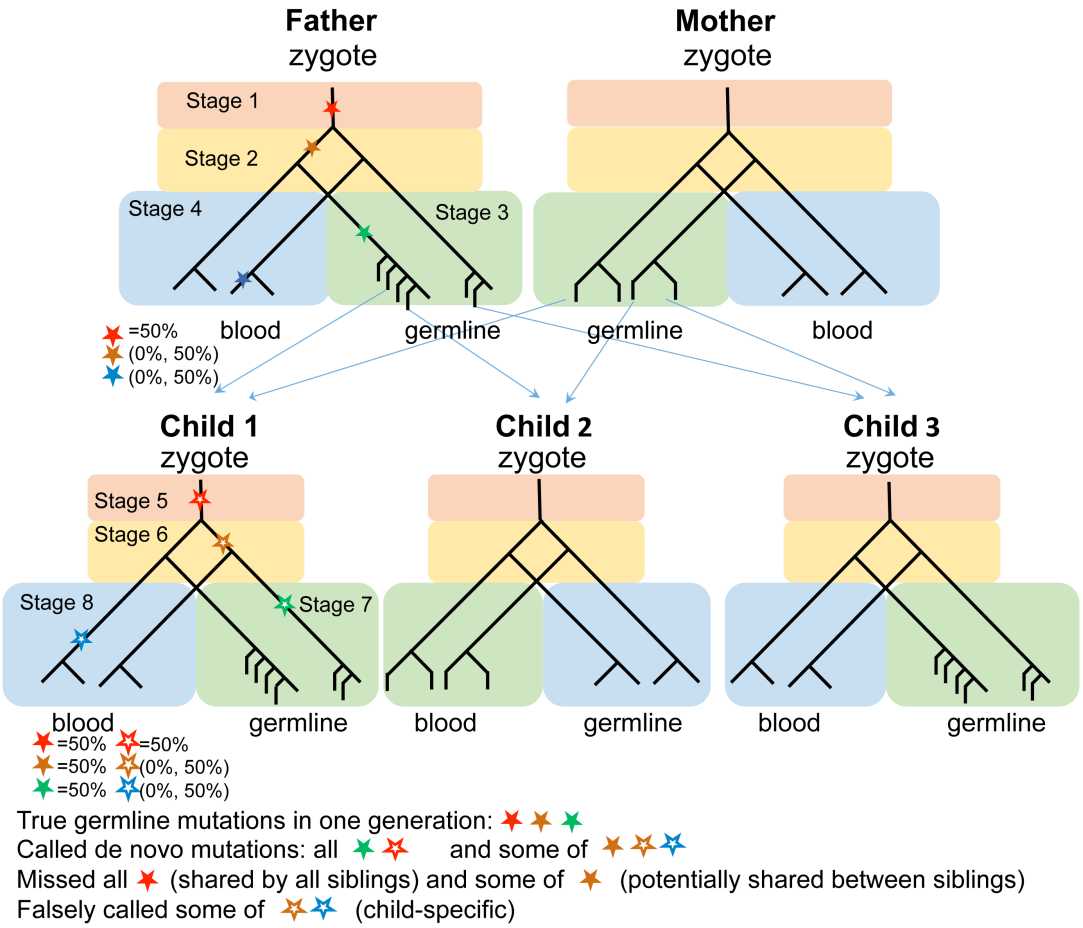

Fig 3. Variation in the estimated paternal age effect, for the autosomes. We plot the de novo mutation rate as a function of the paternal age at conception of the child. The rate was obtained from the reported counts of de novo mutations divided by the fraction of the genome assayed in each study (shown in the title of each subplot, along with the mean coverage per individual). The solid line denotes the fitted slope (i.e., the increase in the mutation rate for each additional year of father's age). For Rahbari et al. 2015 [45], we used the corrected counts of de novo mutations, which are extrapolated to a genome length of 3 Gb.

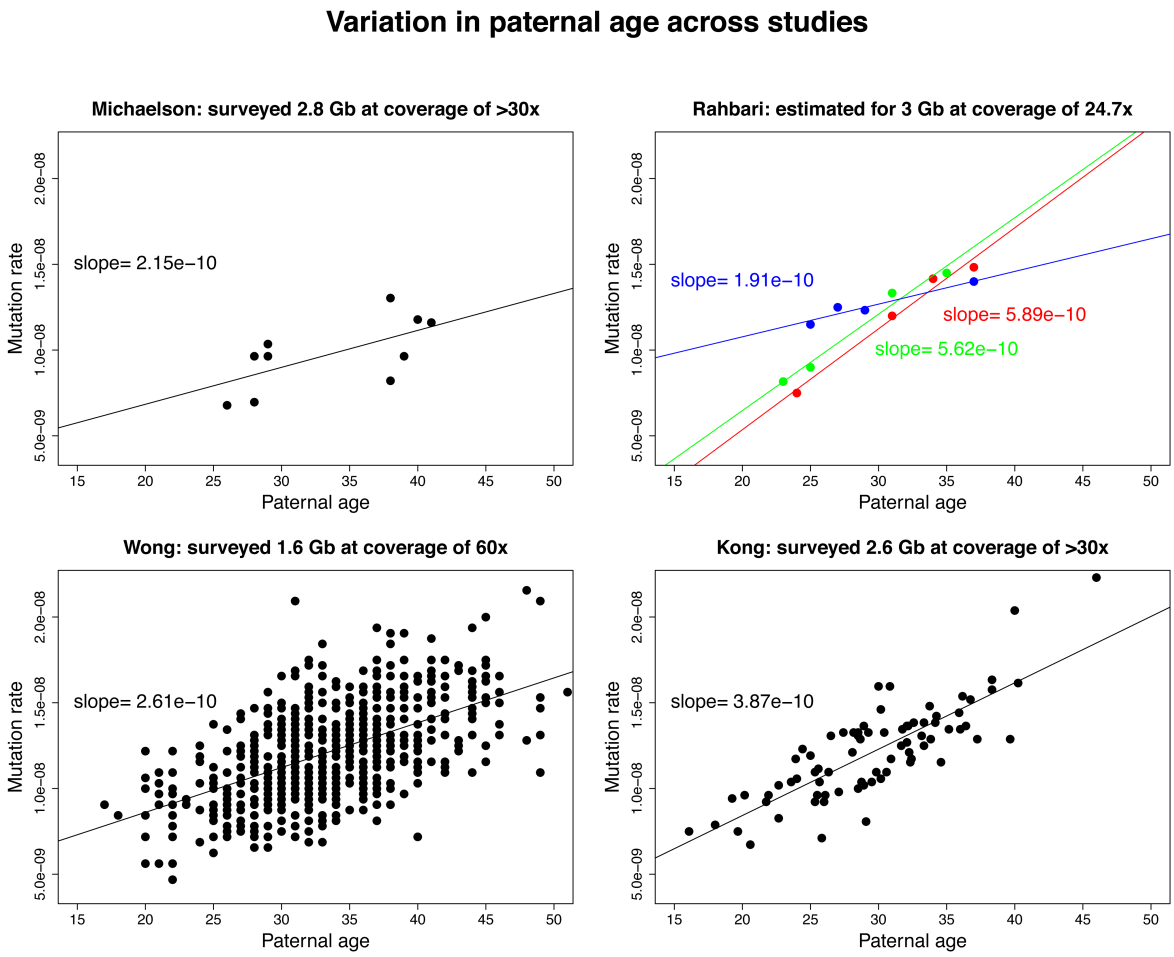

Fig 4. A potential estimation bias in the paternal age effect in the presence of inter-individual variation. If males vary in the strength of their paternal age effect, then estimates obtained by pooling data across families, especially trios, can be misleading. To illustrate this point, we subsampled trios from the larger families analyzed by Rahbari et al. (2015) [45]. Colored dots represent the full data, and black stars highlight one offspring sampled from each family. The colored lines are the fit of a linear regression for each family separately, whereas the black dotted line is the fit to subsampled data (i.e., the slopes indicate the estimated strengths of paternal age effect for each family and for the pooled data for the subsample, respectively). In both scenarios A and B, the average paternal ages are around 30 and the average mutation rates per generation are similar. However, the estimated strength of paternal age effect differs substantially from the mean paternal age effect—in principle, it could even lie outside the range of true slopes across fathers.

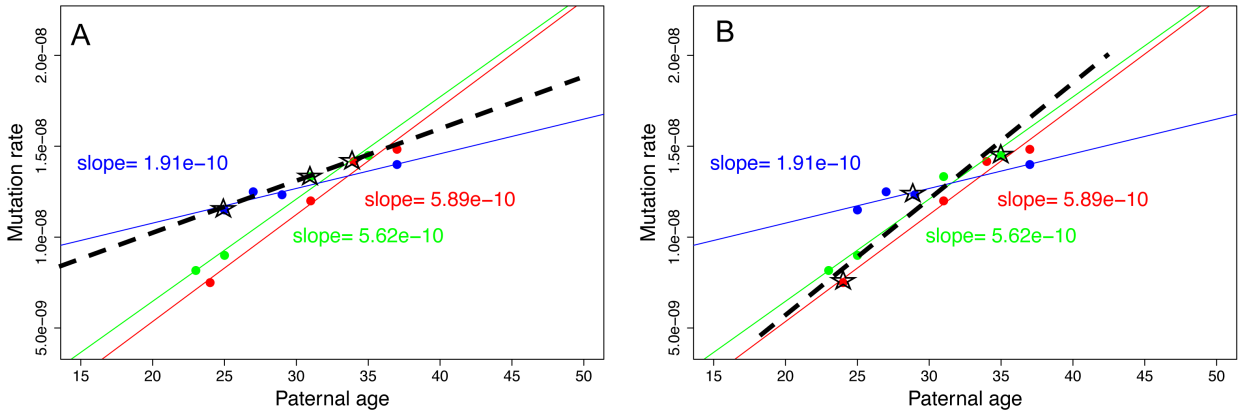

Table 1. Estimates of mutation rates from pedigree studies

Study	Reported mutation rate per bp per generation (x 10^{-8})	Mean paternal age in study (in years)	Mutation rate at paternal age of 30 years (x 10^{-8})[†]	Paternal age effect reported as the increase in number of mutations for each year of fathers age	Callable genome in Gb (reported FNR in %)	Sample size	Mean sequence coverage	Fraction of CpG transitions[‡]
Chimpanzee pedigree study:								
Venn 2014[#] [69]	1.20	24.3[a]	1.51	3.00	2.4 (13.4[b])	6	34.4[b]	0.239 (0.183 - 0.296)
Human pedigree studies:								
Roach 2010[§,c] [14]	1.10 (0.68 - 1.70)	--	--	--	1.8 (5.0)	2	61.3[b]	0.178 (0.037 - 0.320)
Conrad 2011 (CEU) [74]	1.17 (0.88 - 1.62)	--	--	--	2.5 (5.0)	1	29.3	0.146 (0.046 - 0.246)
Conrad 2011 (YRI) [74]	0.97 (0.67 - 1.34)	--	--	--	2.5 (3.5)	1	29.2	0.114 (0.009 - 0.220)
Campbell 2012 [75]	0.96 (0.82–1.09)	26.3	--	--	2.2 (1.7)[b]	5	13.0	0.165 (0.110 - 0.220)
Kong 2012 [16]	1.20	29.7	1.21	2.01	2.6 (2.0)	78	30.0	0.173 (0.163 - 0.184)
Michaelson 2012 [76]	1.00	33.6	0.93	1.02	2.8[e] (9.5)	10	30.0	0.128 (0.099 - 0.156)
Jiang 2013 [77]	--	34.4[e]	--	1.50	--	32	30.0	0.162 (0.146 - 0.177)
Francioli 2014 [65]	--	29.4	--	1.20[*]	2.1 (31.1)	250	13.0	0.165 (0.158 - 0.172)
Besenbacher 2015 [56]	1.27 (1.16 - 1.38)	28.4	1.3[d]	2.00	--	10	50.0	0.201 (0.166 - 0.236)
Rahbari 2015[c] [45]	1.28 (1.13 - 1.43)	29.8	1.29	2.87 (1.46 - 3.65)[f]	2.5 (--)	12	24.7	0.210 (0.180 - 0.240)
Yuen 2015[§,g,c] [78]	1.18[h]	34.1	1.08	1.19[*]	2.5 (8.0[i])	140	56.0	0.159 (0.151 - 0.167)
Wong 2016[§] [49]	1.05	33.4	0.95	0.92	1.6 (13.0)	693	60.0	0.131 (0.127 - 0.135)

-- Not available

§ - Denotes studies that used Complete Genomics technology for sequencing, as most were based on Illumina sequencing.

† - Estimated assuming linearity and using the reported paternal age effect, accounting for the length of the callable genome.

* A paternal age effect is not reported in paper but estimated by Poisson regression on counts of autosomal *de novo* mutations.

‡- CpG fraction based on autosomal mutations and binomial 95% CI shown. When possible, we relied on validated mutations. However, in some studies, only a small fraction of mutations were validated and hence we used all putative de novo mutations.

– This study includes one multi-generational pedigree

a - This age is based on an estimate of the mean male chimpanzee ages at reproduction in the wild, not the age of the actual individual studied.

b - Reported average, even though the estimate varies across individuals or families in the study.

c - Includes siblings.

d - Based on visual inspection of slope reported in the study.

e - Not reported in the article, based on personal communication.

f - Based on validated de novo mutation counts, and extrapolated to a surveyed genome size of 3 Gb.

g - All estimates are based on the subset of the 140 non-LCL samples.

h - Not reported in the article; based on the number of de novo mutations reported in the study and an estimated denominator (2.5 Gb, personal communication).

i - Average of the estimates based on two cohorts.

References:

1. Zuckerkandl E, Pauling L (1962) Molecular disease, evolution, and genic heterogeneity. In: Kasha M, Pullman B, editors. Horizons in Biochemistry Academic Press, New York,. pp. 189–225.
2. Kumar S (2005) Molecular clocks: four decades of evolution. Nat Rev Genet 6: 654-662.
3. Bromham L, Penny D (2003) The modern molecular clock. Nat Rev Genet 4: 216-224.
4. Kimura M (1983) The neutral theory of molecular evolution.: Cambridge University Press.
5. Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism New York: Academic Press. pp. pp. 21–123.
6. Birky CW, Jr., Walsh JB (1988) Effects of linkage on rates of molecular evolution. Proc Natl Acad Sci U S A 85: 6414-6418.
7. Kumar S, Subramanian S (2002) Mutation rates in mammalian genomes. Proceedings of the National Academy of Sciences 99: 803-808.
8. Kondrashov FA, Kondrashov AS (2010) Measurements of spontaneous rates of mutations in the recent past and the near future. Philosophical Transactions of the Royal Society of London B: Biological Sciences 365: 1169-1176.
9. Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. Genetics 156: 297-304.
10. Moorjani P, Amorim CEG, Arndt PF, Przeworski M (2016) Variation in the molecular clock of primates. bioRxiv: 036434.
11. Yi S, Ellsworth DL, Li W-H (2002) Slow molecular clocks in Old World monkeys, apes, and humans. Molecular Biology and Evolution 19: 2191-2198.
12. Steiper ME, Young NM (2006) Primate molecular divergence dates. Mol Phylogenet Evol 41: 384-394.
13. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69-87.
14. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328: 636-639.
15. Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, et al. (2011) Variation in genome-wide mutation rates within and between human families. Nat Genet 43: 712-714.
16. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, et al. (2012) Rate of de novo mutations and the importance of father's age to disease risk. Nature 488: 471-475.
17. Scally A, Durbin R (2012) Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet.
18. Ségurel L, Wyman M, Przeworski M (2014) Determinants of mutation rate variation in the human germline. Annual Review of Human Genetics 15:47-70.

19. Schiffels S, Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. Nature genetics 46: 919-925.
20. Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, et al. (2013) Great ape genetic diversity and population history. Nature 499: 471-475.
21. Langergraber KE, Prüfer K, Rowney C, Boesch C, Crockford C, et al. (2012) Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. Proceedings of the National Academy of Sciences 109: 15716-15721.
22. Steiper ME, Young NM (2008) Timing primate evolution: lessons from the discordance between molecular and paleontological estimates. Evolutionary Anthropology: Issues, News, and Reviews 17: 179-188.
23. Jensen-Seaman MI, Hooper-Boyd KA (2013) Molecular Clocks: Determining the Age of the Human–Chimpanzee Divergence. eLS.
24. Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, et al. (2012) Insights into hominid evolution from the gorilla genome sequence. Nature 483: 169-175.
25. Goodman M (1962) Evolution of the immunologic species specificity of human serum proteins. Human Biology 34: 104-150.
26. Li W-H, Tanimura M (1987) The molecular clock runs more slowly in man than in apes and monkeys. Nature 326: 93-96.
27. Goodman M (1961) The role of immunochemical differences in the phyletic development of human behavior. Human Biology 33: 131-162.
28. Laird CD, McCONAUGHY BL, McCARTHY BJ (1969) Rate of fixation of nucleotide substitutions in evolution.
29. Kohne DE, Chiscon J, Hoyer B (1972) Evolution of primate DNA sequences. Journal of Human Evolution 1: 627-644.
30. Wu C-I, Li W-H (1985) Evidence for higher rates of nucleotide substitution in rodents than in man. Proceedings of the National Academy of Sciences 82: 1741-1745.
31. Gao Z, Wyman MJ, Sella G, Przeworski M (2016) Interpreting the Dependence of Mutation Rates on Age and Time. PLoS Biol 14: e1002355.
32. Amster G, Sella G (2016) Life history effects on the molecular clock of autosomes and sex chromosomes. Proc Natl Acad Sci U S A 113: 1588-1593.
33. Dixson AF (2009) Sexual selection and the origins of human mating systems: OUP Oxford.
34. Gage TB (1998) The comparative demography of primates: with some comments on the evolution of life histories. Annual Review of Anthropology: 197-221.
35. Begun DR (2015) Fossil Record of Miocene Hominoids. Handbook of Paleoanthropology: Springer. pp. 1261-1332.
36. Hartwig WC (2002) The primate fossil record: Cambridge University Press.
37. Mailund T, Munch K, Schierup MH (2014) Lineage sorting in apes. Annual review of genetics 48: 519-535.
38. Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D (2006) Genetic evidence for complex speciation of humans and chimpanzees. Nature 441: 1103-1108.

39. Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, et al. (2011) A molecular phylogeny of living primates. PLoS Genet 7: e1001342.
40. Steiper ME, Young NM, Sukarna TY (2004) Genomic data support the hominoid slowdown and an Early Oligocene estimate for the hominoid–cercopithecoid divergence. Proceedings of the National Academy of Sciences of the United States of America 101: 17021-17026.
41. Li H (2014) Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 30: 2843-2851.
42. Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics 12: 443-451.
43. Beal MA, Glenn TC, Somers CM (2012) Whole genome sequencing for quantifying germline mutation frequency in humans and model species: Cautious optimism. Mutation Research/Reviews in Mutation Research 750: 96-106.
44. Scally A (2015) Mutation rates and the evolution of germline structure. bioRxiv: 034298.
45. Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, et al. (2015) Timing, rates and spectra of human germline mutation. Nature Genetics.
46. Acuna-Hidalgo R, Bo T, Kwint MP, van de Vorst M, Pinelli M, et al. (2015) Post-zygotic point mutations are an underrecognized source of de novo genomic variation. The American Journal of Human Genetics 97: 67-74.
47. Haldane JB (1935) The rate of spontaneous mutation of a human gene. Journal of Genetics 31: 317-326.
48. Crow JF (2006) Age and sex effects on human mutation rates: an old problem with new complexities. J Radiat Res 47 Suppl B: B75-82.
49. Wong WS, Solomon BD, Bodian DL, Kothiyal P, Eley G, et al. (2016) New observations on maternal age effect on germline de novo mutations. Nature communications 7.
50. McRae JF, Clayton S, Fitzgerald TW, Kaplanis J, Prigmore E, et al. (2016) Prevalence, phenotype and architecture of developmental disorders caused by de novo mutation. bioRxiv: 049056.
51. Alexandrov LB, Stratton MR (2014) Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Curr Opin Genet Dev 24: 52-60.
52. Lodish H (2008) Molecular cell biology: Macmillan.
53. Lindahl T, Wood RD (1999) Quality control by DNA repair. Science 286: 1897-1905.
54. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic acids research 29: 308-311.
55. Lek M, Karczewski K, Minikel E, Samocha K, Banks E, et al. (2015) Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv: 030338.
56. Besenbacher S, Liu S, Izarzugaza JM, Grove J, Belling K, et al. (2015) Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. Nature communications 6.

57. Day FR, Helgason H, Chasman DI, Rose LM, Loh P-R, et al. (2016) Physical and neurobehavioral determinants of reproductive onset and success. Nature genetics.
58. Fenner JN (2005) Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. American journal of physical anthropology 128: 415-423.
59. Hwang DG, Green P (2004) Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci U S A 101: 13994-14001.
60. Wilson Sayres MA, Venditti C, Pagel M, Makova KD (2011) Do variations in substitution rates and male mutation bias correlate with life-history traits? A study of 32 mammalian genomes. Evolution 65: 2800-2815.
61. Ramm SA, Stockley P (2010) Sperm competition and sperm length influence the rate of mammalian spermatogenesis. Biol Lett 6: 219-221.
62. Britten RJ (1986) Rates of DNA sequence evolution differ between taxonomic groups. Science 231: 1393-1398.
63. Thomas GW, Hahn MW (2014) The human mutation rate is increasing, even as it slows. Molecular biology and evolution 31: 253-257.
64. Harris K (2015) Evidence for recent, population-specific evolution of the human mutation rate. Proceedings of the National Academy of Sciences 112: 3439-3444.
65. Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, et al. (2015) Genome-wide patterns and properties of de novo mutations in humans. Nature genetics 47: 822-826.
66. Duret L, Galtier N (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. Annual review of genomics and human genetics 10: 285-311.
67. Williams AL, Genovese G, Dyer T, Altemose N, Truax K, et al. (2015) Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. Elife 4: e04637.
68. Capra JA, Hubisz MJ, Kostka D, Pollard KS, Siepel A (2013) A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. PLoS Genet 9: e1003684.
69. Venn O, Turner I, Mathieson I, de Groot N, Bontrop R, et al. (2014) Strong male bias drives germline mutation in chimpanzees. Science 344: 1272-1275.
70. Lehtonen J, Lanfear R (2014) Generation time, life history and the substitution rate of neutral mutations. Biology letters 10: 20140801.
71. Lipson M, Loh P-R, Sankararaman S, Patterson N, Berger B, et al. (2015) Calibrating the human mutation rate via ancestral recombination density in diploid genomes. PLoS Genet 11: e1005550.
72. Palamara PF, Francioli LC, Wilton PR, Genovese G, Gusev A, et al. (2015) Leveraging Distant Relatedness to Quantify Human Mutation and Gene-Conversion Rates. The American Journal of Human Genetics 97: 775-789.

73. Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, et al. (2014) Genome sequence of a 45,000-year-old modern human from western Siberia. Nature 514: 445-449.
74. Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, et al. (2011) Variation in genome-wide mutation rates within and between human families. Nat Genet 43: 712-714.
75. Campbell CD, Chong JX, Malig M, Ko A, Dumont BL, et al. (2012) Estimating the human mutation rate using autozygosity in a founder population. Nature genetics 44: 1277-1281.
76. Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, et al. (2012) Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell 151: 1431-1442.
77. Jiang Y-h, Yuen RK, Jin X, Wang M, Chen N, et al. (2013) Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. The American Journal of Human Genetics 93: 249-263.
78. Yuen RK, Thiruvahindrapuram B, Merico D, Walker S, Tammimies K, et al. (2015) Whole-genome sequencing of quartet families with autism spectrum disorder. Nature medicine 21: 185-191.