Life history effects on neutral polymorphism and divergence rates, in autosomes and sex chromosomes

Guy Amster

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

ABSTRACT

Life history effects on neutral polymorphism and divergence rates, in autosomes and sex chromosomes

Guy Amster

Much of modern population genetics revolves around neutral genetic differences among individuals, populations, and species. In this dissertation, I study how sex-specific life history traits affects neutral diversity levels within populations (polymorphism) and between species (divergence) on autosomes and sex chromosomes. In chapter 1, I consider the effects of sex specific life histories, and particularly generation times, on substitution rates along the great ape phylogeny. Using a model that approximates features of the mutational process in most mammals, and fitting the model on data from pedigree-studies in humans, I predict the effects of life history traits on specific great ape lineages. As I show, my model can account for a number of seemingly disparate observations: notably, the puzzlingly low X-to-autosome ratios of substitution rates in humans and chimpanzees and differences in rates of autosomal substitutions among great ape lineages. The model further suggests how to translate pedigree-based estimates of human mutation rates into split times among extant apes, given sex-specific life histories. In so doing, it largely bridges the gap reported traditional fossil-based estimates of mutation rates, and recent pedigree-based estimates. In chapter 2, I consider the effects of sex- and age- dependent mortalities, fecundities, reproductive variances and mutation rates on polymorphism levels in humans. Using a coalescence framework, I provide closed formulas for the expected polymorphism rate,

accounting for life history effects. These formulas generalize and simplify previous models. Applying the model to humans, my results suggest that the effects of life history – and of sex differences in generation times in particular – attenuate how changes in historical population sizes affect X to autosome polymorphism ratios. Applying these results to observations across human populations, I find that life history effects and demographic histories can largely explain the reduction in X to autosome polymorphism ratios outside Africa. More generally, my work elucidates the major role of sex-specific life history traits – and male and female generation times in particular – in shaping patterns of neutral genetic diversity within and between species.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I wish to thank my PhD advisor, Guy Sella, for his guidance through the last five years. Throughout this period, his involvement, his willingness to consider my ideas, his uncompromising drive for quality and his high standards for scientific communication have been a driving factor in my work. On a personal level, his support and encouragement have made my graduate studies pleasant and productive. His high standards for scientific research and for scientific communication will continue to set an example for me to follow.

I am deeply grateful to Molly Przeworski, who was extremely generous with her advice and support. I learned a lot from collaborating with her and with members of her lab. I was lucky to share a temporary office with her lab members in 2014, and I still regret having to move to our permanent offices.

I am also grateful to my committee members, Peter Andolfatto, Itsik Pe'er and Joe Pickrell, for their advice and encouragement. I was lucky to receive feedback and advice from Brian Charlesworth, Graham Coop, Magnus Nordborg, `David Pilbeam ,David Reich and Aylwyn Scally. I wish to thank them for their many insightful advices on my research projects.

I would like to thank Ziyue Gao, Priya Moorjani and David Murphy for inviting me to collaborate on their research projects and for sharing their insights and their scientific passion. I enjoyed collaborating with Ziyue, Priya and Molly on our recent manuscript, and I take pride in my small contribution to that research project. I wish to thank the three of them for putting up with the constant flow of last-minute statistical reservations prior to the first submission.

I would also like to thank Ipsita Agarwal, Eduardo Amorim, Zachary Baker, Chen Chen, Eyal Elyashiv, Zachary Fuller, Arbel Harpak, Laura Hayward, Amir Kermani, Will Milligan,

Hakhamanesh Mostafavi, Molly Schumer, Yuval Simons, Sonal Singhal, Amy Williams, Felix Wu and Minyoung Wyman for their friendship and for many helpful discussions.

Introduction

Much of modern population genetics revolves around neutral genetic differences among individuals, populations, and species. Importantly, patterns of neutral polymorphism and divergence reflect a wealth of information about evolutionary processes and have been used extensively to make inferences about them. Since genetic diversity accumulates slowly over millions of years, these inferences provide an opportunity to learn about processes in the distant past that cannot be observed directly. Neutral divergence among species reflects the accumulation of genetic mutations since their split from a common ancestor and is our primary source of information about the chronology and evolutionary relationship among all living species, over time scales ranging from millions to billions of years. Neutral polymorphism reflects more recent evolutionary history and has been our main source of information about recent demographic history as well as basic processes, such as mutation and recombination. Understanding the determinants of neutral polymorphism and divergence is therefore of focal importance in population genetics.

A fundamental result in molecular evolution is -the rate at which neutral substitutions accumulate on a lineage equals the rate at which mutations arise (1). Thus, divergence patterns allow inferences on the average mutation rate over phylogenetic timescales. This result is sometimes phrased in terms of "the molecular clock", as it implies that sequence divergence could be normalized to units of years, given an estimate for the mutation rate. In the opposite direction, given an estimate for the split times of two species (for example, based on fossil evidence), one could infer the mutation rate from their neutral divergence levels. Most estimates of species split times over relatively short evolutionary times scales rely on the molecular clock. In particular,

until recently, estimates of the chronology of great ape evolution were based on estimating the mutation rate in the few splits where fossil evidence is reliable and then rely on this estimate of the mutation rate to date the entire phylogeny.

However, these estimates have recently been put into question over the last decade, since it became technologically feasible to estimate the mutation rate directly by sequencing two-generations or three-generation families and obtaining direct counts of the number of de-novo mutations per-generation. Surprisingly, in humans, these new estimates are ~2-fold lower than the previous, fossil-based estimates (e.g., (2)). This dramatic reduction in estimates of human mutation rates seriously challenged previous estimates of the split times between us and our closest living relatives. For instance, they suggest that the split of humans and chimpanzees from a common ancestor occurred ~10 MYA instead of ~4 MYA, i.e., more than twofold than the previous estimate. Moreover, the new split time estimates appear to be at odds with the catarrhine fossil record (3).

In chapter 1, published in PNAS (4), I suggest a plausible solution to this puzzle and a revision to the estimates for split times throughout the great ape phylogeny. To this end, I model how the molecular clock is affected by sex-specific life history traits, particularly generation times in males and females, relying on a model that approximates known features of the mutational process in mammals. Using this model, I provide a quantitative prediction for the effect of sex-specific life history traits on divergence rates. A surprising result of the model is that yearly mutation rates in humans are fairly insensitive to changes in the mean generation time but are strongly affected by relative changes in generation times between males and females. I then rely on plausible values of life history traits to derive revised estimates for corresponding split times. For example, considering that estimates of generation times are consistently higher in males (e.g., in hunter

gatherer societies and chimpanzees groups (5, 6)), that the age of puberty in males is considerably lower in chimpanzees and plausibly lower in Homo erectus (e.g., from Nariokotome boy dated ~1.5 MYA (7)) than in modern humans, and relying on the pedigree estimates of mutation rates, the model suggests humans and chimpanzees might have split from a common ancestor as recently as ~6.6 MYA, compared to a naïve pedigree based estimate of ~10 MYA (3). Moreover, my split time estimates for the great-ape phylogeny are consistent with the fossil record and thus reconcile them with the pedigree-based estimates of the mutation rate. More generally, these results suggest that life-history traits can have a substantial effect on the molecular clock and thus on phylogenetic estimates.

The effects of life history on the molecular clock can also explain several other intriguing patterns observed in the molecular phylogeny of great apes and more generally, mammals. Notably, they can explain the puzzlingly low X-to-autosome (X:A) ratios of substitution rates observed in the human and chimpanzee lineages and the differences in ratios among great apes lineages (8, 9). In particular, this provides an alternative, and arguably more natural, explanation for the low X:A divergence ratio on the human and chimpanzee lineages than the earlier suggestion that attributed this observation to 'complex speciation' (8). I also find that life history can explain why autosomal substitution rates on the gorilla lineage are substantially greater than rates on the human and chimpanzee lineages (10). Finally, considering these effects can help to reconcile recent evidence that changes in generation times had only minor effects on mutation rates in humans (11) with classic studies suggesting they have had a major effects on the rate of molecular evolution in the mammalian phylogeny (12, 13). Overall, these finding help resolve several recent and long-standing puzzles and highlight the importance of life history traits in shaping patterns of neutral genetic divergence.

Life history may have an even more pronounced effect on neutral polymorphism levels. The time to the most recent ancestor separating alleles from two different species is largely determined by the species' split times (neglecting the effects of ancestral polymorphism), and life history therefore affects divergence primarily by affecting the yearly rate of mutation on the lineage. In contrast, the time to the most recent common ancestor of two alleles from the same population can be affected by life history traits. Thus life history affects polymorphism level through both genealogical and mutation processes.

Comparing relative levels of neutral polymorphism on the X and autosomes is of particular interest in this regard. Because autosomes spend an equal number of generations in both sexes while the X spends twice as many of generations in females, relative diversity levels on the X and autosomes are likely affected by differences in male and female life history and mutation processes. Notably, males and females of many species differ in their age-dependent mortality and fecundity, as well as in the variance of their number of offspring. These differences would differentially affect polymorphism levels on X and autosomes through both mutational and genealogical processes.

For these reasons and others, there has been a longstanding interest in comparing polymorphism levels on X and autosomes (tracing back to Haldane; (14)).

In particular, over the past decade, when sequence data became available for a variety of different populations, there has been considerable interest in these comparisons. Specifically, neutral X:A polymorphism ratios have been shown to be consistently lower outside of Africa. The reduction outside of Africa could be partially explained by the demographic history of human populations in the last 200,000 years (i.e., population bottlenecks, expansions, and migrations), but previous works suggest that demography can explain at most ~50% of the reduction in ratios.

Motivated by these observations, in chapter 2, I model the effects of life history on neutral polymorphism on autosomes and sex chromosomes. The model is quite general and incorporates the effects of sex- and age- dependent mortalities, fecundities, reproductive variances and mutation rates. Despite its generality, I present an analytic solution for the model, which is formulated in terms of the neutral coalescent. The model and its solution generalize and simplify previous theoretical results by Hill (15), Charlesworth (16) and others. In particular, I show that the effects of life history on the X:A ratio can be summarized in terms of separate ratios of male-to-female number of newborns, generation times, reproductive variances, and mutation rates. This formulation provides a straightforward understanding of the separate effects of each of these factors.

I then apply my results to humans, relying on estimates of life-history parameters and of the observed dependence of mutation rates on age and sex. I find that life history effects, and the effects of male and female generation times in particular, may account for much of the observed variation in X:A ratios of polymorphism levels across human populations. Finally, I explain why the standard approach used to separate the genealogical and mutational effects on the X:A polymorphism ratio (i.e., by normalizing by divergence to an outgroup) introduces considerable bias and outline alternative approaches for moving forward.

Chapter 2 has been posted in bioRxiv (17) and is now in revision for PNAS. Following the editor's suggestion, I decided to split the paper into two. One will focus on the theory (which now appears mostly in in appendix 2 of this thesis), and I plan to submit it to Genetics. The other, which will be resubmitted to PNAS, is focused on the application of the results to explaining the patterns observed in humans and includes a number of new analyses that I am now finalizing. First, I show that estimates of male mutation bias based on pedigree studies are about two-fold greater than

those based on X:A ratios of divergence to an outgroup (the standard approach to date), largely due to generation time effects on divergence that substantially bias the latter. I therefore rely on pedigree-based estimates and find that human X:A ratios of effective population sizes far from genes (where the effect of selection at linked sites is minimal) are considerably greater than previously appreciated. Next, I show how the effects of life history – of sex differences in generation times in particular – attenuate how historical changes in population size affect X:A polymorphism ratios. I therefore rely on inferences about mutation rates and the demographic history of human populations to infer a range of sex specific generation times and reproductive variances that can explain observed polymorphism ratios across human populations jointly. The inferred parameters appear to fall within the range of estimates in extant hunter gatherer societies. Thus, the effects of life history – of sex specific generation times in particular – and their interaction with demographic history, can go a long way toward explaining the X:A polymorphism ratios observed across different human populations.

Chapter 1 - Life history effects on the molecular clock of autosomes and sex chromosomes

1.1 Abstract

One of the foundational results in molecular evolution is that the rate at which neutral substitutions accumulate on a lineage equals the rate at which mutations arise. Traits that affect rates of mutation therefore also affect the phylogenetic "molecular clock". We consider the effects of sex-specific generation times and mutation rates in species with two sexes. In particular, we focus on the effects that the age of onset of male puberty and rates of spermatogenesis have likely had in hominids (great apes), considering a model that approximates features of the mutational process in mammals, birds and some other vertebrates. As we show, this model can account for a number of seemingly disparate observations: notably, the puzzlingly low X-to-autosome ratios of substitution rates in humans and chimpanzees and differences in rates of autosomal substitutions among hominine lineages (i.e., humans, chimpanzees and gorillas). Importantly, the model further suggests how to translate pedigree-based estimates of human mutation rates into split times among extant hominoids (apes), given sex-specific life histories. In so doing, it largely bridges the gap reported between estimates of split times based on fossil and molecular evidence, in particular suggesting that the human-chimpanzee split may have occurred as recently as 6.6 MYA. The model also implies that the generation time effect should be stronger in short-lived species, explaining why the generation time has a major influence on yearly substitution rates in mammals but only a subtle one in human pedigrees.

1.2 Significance Statement

Recent estimates of mutation rates obtained by sequencing human pedigrees have challenged conceptions about split times between humans and our closest living relatives. In particular, estimates of human split times from chimpanzees and gorillas based on the new mutation rate estimates are more than twofold shorter than previously believed, seemingly at odds with the fossil record. Here, we show that accounting for the effects of sex-specific life histories on mutation rates along the hominid phylogeny largely bridges this apparent gap and leads to more accurate split time estimates. Doing so can also explain other intriguing phylogenetic patterns in hominid and mammalian evolution.

1.3 Introduction

Most of our inferences about species split times on short phylogenetic time scales rely on the neutral molecular clock. According to the Neutral Theory, the number of substitutions K that accumulate in a lineage over T years (e.g., since the split from another species) is $K = (\bar{u}/\bar{G})T$, where $\bar{u}$ and $\bar{G}$ are the average mutation rate per generation and average generation time, respectively (1). Inferring split times therefore requires estimates of the yearly mutation rate $\bar{u}/\bar{G}$ on the lineage in question. Such estimates generally derive from securely dated fossils on other lineages or from measurements of mutation rates in extant species (2, 11, 18, 19). Using these estimates for dating thus necessitates an understanding of the way that yearly mutation rates may change over time.

Neutral substitution patterns in mammals offer some insights. Variation in yearly mutation rates on phylogenetic time scales can be assessed by comparing the number of neutral substitutions along two branches leading from a common ancestor to extant species. These comparisons show

marked variation in yearly rates on autosomes. For example, there are 50% fewer substitutions on the human branch compared with rodents (12) and 26% fewer compared to baboons, with more moderate differences among hominine lineages (8, 10, 12, 20, 21). The average yearly rates are also negatively correlated with generation times (and their correlates) in extant mammals, leading to the notion of a generation time effect on the molecular clock (12, 13, 22).

Neutral substitutions rates vary not only among taxa but also between sex chromosomes and autosomes. For brevity, we consider the relative rates on X and autosomes, but these considerations extend naturally to Y (or ZW). Because autosomes spend the same number of generations in both sexes while the X spends twice as many generations in females, rates of neutral substitutions on autosomes reflect a greater relative contribution of male mutations than on the X. In a wide range of taxa, neutral substitutions rates on autosomes are greater than on the X (or lower than on the Z), suggesting a male biased contribution to yearly mutation rates (23). Moreover, observed X-to-autosome ratios are extremely variable, ranging between 0.76-0.9 in hominids and up to 1.0 in surveyed mammals, indicating that the degree of male bias itself varies greatly on phylogenetic time scales (9, 23).

Our current understanding of mutation can help tie these observations together (11). Pedigree studies in humans and chimpanzees establish that most mutations are paternal in origin and that the paternal but not the maternal contribution increases strongly with age (2, 24). This has long been thought to be true because germ-cell division is arrested before birth in females but proceeds continuously post-puberty in males (11, 14, 25, 26). The same reasoning may extend to mammals, birds and other vertebrate taxa in which oogenesis ceases before birth or hatching (27-29). These considerations suggest that maternal and paternal generation times should affect the molecular clock differently. They also imply that the age of puberty in males and the rate of spermatogenic

germ cell divisions should affect yearly mutation rates (11). The variation observed among closely related extant species indicates that these parameters change over phylogenetic time scales. Here we ask how such changes would affect the molecular clock on X and autosomes.

1.3 Model

The molecular clock with two sexes. We model the accumulation of neutral substitutions allowing for different generation times and mutation rates in males and females (cf. (16)). If, for example, the paternal generation time, G_M (defined as the average time between births), is longer than the maternal, G_F, then autosomes would spend a greater proportion of time (but not of generations) in males than in females (i.e., $G_M/(G_F + G_M)$ in males). Therefore, the yearly mutation rate in males, μ_M/G_M (where μ_M is the rate per generation in males), would have a greater relative contribution to the autosomal mutation rate. Taking the corresponding weighted average for the expected number of substitutions on an autosomal lineage over T years yields

$$K_A = (\mu_A/G_A)T, \qquad [1.1]$$

where $G_A = \frac{1}{2}(G_F + G_M)$ and $\mu_A = \frac{1}{2}(\mu_F + \mu_M)$ are the expected sex-averaged generation time and mutation rate per-generation on an autosomal lineage. By the same token, on the X

$$K_X = (\mu_X/G_X)T, \qquad [1.2]$$

where in this case $G_X = \frac{2}{3}G_F + \frac{1}{3}G_M$ and $\mu_X = \frac{2}{3}\mu_F + \frac{1}{3}\mu_M$ are the expected sex-averaged generation time and mutation rate per-generation on a X lineage (cf. Appendix section A1.1 for rigorous derivations and the corresponding Eqs. for Y and Z).

Sex and age dependent mutation rates in hominines. We model male and female mutation rates per generation, μ_M and μ_F, based on the current understanding of the process of accumulation of germline mutations (11, 26). Namely, we assume that mutations accumulate linearly with the number of germ-cell divisions, and the rate per division varies at different stages of development (cf. 11). This is a natural assumption for replicative mutations and has recently been suggested to apply to non-replicative mutations that are efficiently repaired (30; cf. Discussion for other kinds of mutations).

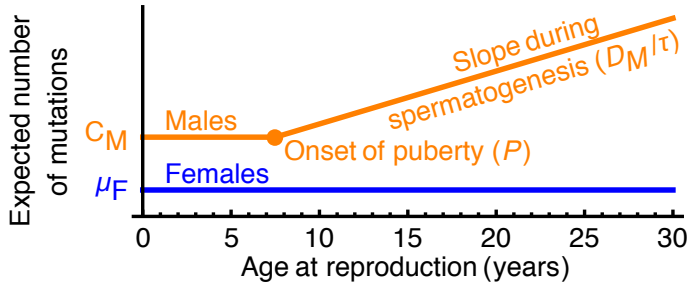

Fig. 1.1. The mutational model.

In females, all oogonial mitotic divisions occur before birth, so the number of mutations should have no dependence on the age of reproduction (28). We therefore model the female per-generation mutation rate as a constant.

In males, cell divisions in the germline exhibit two main phases: pre-puberty, starting from the zygote through the proliferation of germ cells in the growing testis, and post-puberty, with continuous divisions in the adult testis during spermatogenesis. While age of puberty and testis mass varies considerably among hominids (9), the number of germ cell divisions before puberty should increase only logarithmically with mass and is similar in species exhibiting large differences in age of puberty (such as mice and humans; 31). We therefore approximate the

expected number of mutations pre-puberty as constant (C_M). Post-puberty, one germ cell division is thought to occur at each spermatogenic cycle (i.e., the seminiferous epithelial cycle; 32), and the length of the cycle (τ) varies among hominines (Table 1.1). We therefore assume a mutation rate per-year of D_M/τ in adult males, where D_M is the expected number of mutations per spermatogenic division. We then model the average mutation rate per generation in males by

$$\mu_M = C_M + (D_M/\tau)(G_M - I - P), \qquad [1.3]$$

where P is the expected age of puberty and G_M-I is the expected generation time excluding gestation time (I). Both the constancy with age in females and the piecewise linearity in males (Fig. 1.1) are consistent with the observations of pedigree studies in humans and chimpanzees (2, 24). Moreover, a similar model (e.g., accounting for intermittent spermatogenesis; 27) may provide a reasonable approximation for vertebrate species in which oogenesis ceases before birth or hatching (e.g., mammals, birds, elasmobranchs, cyclostomes and a few teleosts; 28, 29).

In Appendix sections A1.1 and A1.2 we consider how the molecular clock is affected by variation of mutational parameters (e.g., age of puberty and generation times) both within a population and over phylogenetic time. Specifically, we show that the expected number of substitutions is insensitive to the variance of the number of mutations or to the variance of the age at reproduction, thus justifying our consideration of only the means.

Parameter values and ranges. To study how changes in life history traits and spermatogenesis should affect the rate of the molecular clock, we would like to assign realistic values to the parameters of the mutational model. Lacking evidence to the contrary, we assume that the parameters associated with the rates of mutation per germ-cell division at different developmental stages remained constant throughout the hominine phylogeny. We infer these parameters from the relationship between mutation rates and paternal ages in the largest human pedigree study

published to date (2), which yields: $C_M = 6.13 \times 10^{-9}$, $D_M = 3.33 \times 10^{-11}$ and $\mu_F = 5.42 \times 10^{-9}$ per base pair (see Appendix section A1.4 for details).

In contrast, life history and spermatogenesis parameters (G_M, G_F, P and τ) are known to vary among species and even populations (cf. Table 1.1 and Tables A1.3-A1.7 and references therein). We use the variation among extant species to guide our choice of plausible ranges for these parameters on the phylogeny. These parameter ranges should be treated only as a rough guide, because of the considerable uncertainty associated with estimates in extant species (cf. discussion in Appendix section A1.4) and the uncertainty about the extent to which they reflect the variation along the phylogeny.

Populations	τ (days)	*P* (years)	G_F (years)	G_M (years)
Human:	16	13		
Hunter-gatherers			26.9*	33.8*
Chimpanzee:	14	7.5		
Western			26.3	24.3
Eastern			24.8	24
Average+			25.2	24.1
Gorilla:	-	7		
Mountain gorillas			18.2	20.4

Table 1.1: Estimates of spermatogenesis and life history parameters in hominines (cf. Appendix section A1.4 for details). * Revised from (5) (cf. Appendix section A1.4). + Weighted by sample size (6).

1.4 Results

The autosomal molecular clock. We first consider the effects of the average and ratio of male and female generation times. In general, the impact of changing the ratio or average of generation times will depend on the way in which mutation rates vary with sex and age (see Appendix section A1.3 for more details). If, for example, mutation rates increase more rapidly with paternal than maternal age, as is expected for mammals, then increasing the ratio of male-to-female generation times necessarily increases the mutation rate per year.

Predicting the effect of increasing the average generation time on yearly mutation rates requires additional assumptions. If we consider our mutational model (regardless of parameter values) and assume a constant proportion of the male generation time spent pre- and post-puberty (Fig. A1.1), we find that an increase in the average generation time will decrease the rate of the molecular clock (see Appendix section A1.3), consistent with the "generation time effect" observed in mammals (12, 13, 22). If instead, we assume that age of puberty remains constant while the average generation time increases, which better describes the observations in extant hominids (cf. Table 1.1 and Table A1.3), then the effect on yearly mutation rates can go either way, depending on parameter values (i.e., on C_M, D_M and μ_F; cf. Appendix section A1.3 and (11, 30)). Intriguingly, the parameters estimates from human pedigree studies are close to the boundary where the effect of the average generation time changes direction and is therefore negligible (Fig. 1.2A).

Within the ranges for the average and ratio of generation times estimated in hominines, changes to the ratio have a much greater impact (Fig. 1.2A). Replacing the standard (though often implicit) assumption of equal generation times in males and females with estimates of their ratio in current hunter-gatherer populations increases yearly mutation rates by ~10%. Further considering that the ratios in extant hominines range between 0.92 for western chimpanzees and 1.26 for hunter-

gatherers (cf. Appendix section A1.4) suggests that yearly rates could vary between 4% lower to 10% higher than the rate with a ratio of 1 (unless noted otherwise, when we vary a single parameter, other parameters are assigned their estimated values in humans). By comparison, varying the average generation time between 19 years, for gorillas, and 30.4 years, for hunter-gatherer populations, impacts rates by less than 2% (see also (30)).

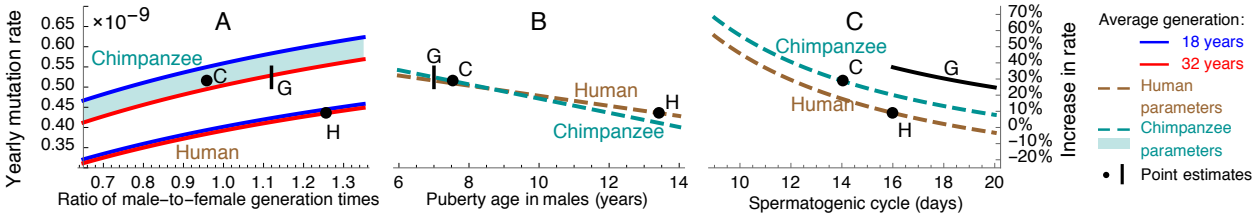

Fig. 1.2. Predicted yearly mutation rates on autosomes as a function of the ratio of male-to-female generation times (A), male age of puberty (B) and spermatogenic cycle length (C). Rates are measured relative to the estimate of 0.4×10^{-9} per bp per year reported by Kong et al. (2), whose data we used to fit our mutational model (cf. Model section). In each panel, we vary one parameter, while fixing others to their estimated values in extant humans (brown) or chimpanzees (green) (see Table 1.1). Note that the point estimates for humans and chimpanzees (black points) do not coincide with those reported in pedigree studies, because ours account for the predicted effects of life history and spermatogenesis parameters. The estimated range for gorillas (black line), where the spermatogenic cycle length has not been measured, corresponds to cycles between 16 and 20 days.

Both an earlier onset of male puberty and an increased rate of spermatogenesis increase the yearly mutation rates in males, resulting in an increased rate on autosomes (Figs. 2B and C; (11, 30)).

Varying their values within the range known for hominines markedly affects yearly rates, by ~18% for the male puberty age and ~8% for the spermatogenic cycle length.

Considered jointly, these factors should generate differences in the branch lengths leading to extant hominines. Assuming parameter estimates for extant humans and chimpanzees, our model suggests that the human branch would be 15% shorter. Making the more plausible assumption that the yearly mutation rate in the ancestor was between extant estimates and that the rates on each lineage changed gradually, we would still expect the human branch to be somewhat shorter. The human branch has indeed been estimated to be 0.4%-2.9% shorter (8, 10, 20); these estimates include the contribution of ancestral polymorphism in common to both branches, suggesting that the branch specific difference is larger.

Our ability to make similar predictions about the gorilla branch is hindered by the lack of estimates for the spermatogenic cycle length. Nonetheless, spermatogenic cycle lengths are negatively correlated with relative testis mass in mammals (Fig. A1.3), and the variation in both has been attributed to differences in the intensity of sperm competition in different mating systems (33, 34). In agreement with this hypothesis, the relative testis mass and rate of spermatogenesis in chimpanzees (0.27% of body weight and 14 days), which have a promiscuous mating system, are greater than in humans (0.07% of body weight and 16 days), which are largely monogamous, and the relative testis mass in polygynous gorillas (one male controls reproductive access to many females) are smaller than in both (0.02% of body weight) (9, 34). This suggests that the rate of spermatogenesis in gorillas is lower than in humans. Varying the length of the spermatogenic cycle between 16 days (its value for humans) and 20 days and assuming current estimates for life history yields predicted branch lengths 14% to 26% longer than the human branch, respectively. These predictions accord with current estimates (again, without accounting for the ancestral

contribution), which suggest that the gorilla branch is between 6.1%-11% longer than the human branch (8, 10, 20).

Lastly, we consider how accounting for life history affects estimates of the split time between humans and chimpanzees. Current estimates of this parameter are obtained by dividing the neutral divergence on the human lineage (~0.4% per bp after subtracting the estimated contribution of ancestral polymorphism (35)) by estimates of the yearly mutation rate. This approach yields a split time of ~10 MYA ((3) and Appendix section A1.5), given pedigree-based estimates of the mutation rate of $\sim 0.4 \times 10^{-9}$ per bp per year (2).

These estimates, however, do not account for most of the effects that we have considered (cf. Discussion). Notably, rather than assuming an equal generation times in males and females, one may want to consider the ratio estimated in extant hunter-gatherers. Moreover, the earlier age of puberty in chimpanzees and possible earlier puberty in *Homo erectus* (e.g., from Nariokotome boy dated ~1.5 MYA (7)) suggest that the onset of puberty may have occurred at younger ages during most of the human lineage. Similarly, the longer spermatogenesis cycle in humans than in chimpanzees (and all other surveyed primates; Fig. A1.2) indicates that it may have been shorter along the human lineage. While it is difficult to translate these considerations into point estimates, they all suggest a higher yearly mutation rate (Fig. 1.2) and a split time that is more recent than suggested by the standard pedigree-based estimates. As an illustration, if the average mutation rate on the human lineage were between the point estimates in extant humans and chimpanzees then our model would suggest a split time between 7.7-9.1 MYA (Fig. 1.2 and Appendix section A1.5). If we further allow individual life history parameters to vary between their values in extant humans and chimpanzees then the lower bound on the split time is further reduced to ~6.6 MYA (see Fig. 1.5).

The relative rates of the molecular clock on X and autosomes. Differences in neutral rates of substitutions on sex chromosomes and autosome have generally been attributed to differences in mutation rates per generation between sexes. In fact, X-to-autosome ratios have been widely used to infer the male mutation bias, $\alpha = \mu_M/\mu_F$, in lineages of mammals, birds, flies, fish and plants (13, 23). These inferences rely on Miyata's formula (36)

$$K_X/K_A = f(\mu_M/\mu_F), \qquad [1.4]$$

where $f(x) = \frac{2x+4}{3x+3}$. Critically, Miyata's formula assumes equal male and female generation times. When this assumption is relaxed (using Eqs. 1.1 and 1.2), the X-to-autosome ratio also depends on the ratio of male-to-female generation times:

$$K_X/K_A = f(\mu_M/\mu_F)/f(G_M/G_F) \qquad [1.5]$$

(the corresponding relationships for other sex chromosomes are provided in Appendix section A1.1). This suggests a theoretical range between ½ and 2 compared to a range between 2/3 and 4/3 based on Miyata's formula. We note that this predicted ratio applies to the accumulation of substitutions after species split but not to the contribution of ancestral polymorphism, because the average TMRCA also differs between X and autosomes.

What is made clear by this derivation is that ignoring differences in generation times could introduce substantial biases in estimates of the male mutation bias, α (Fig. 1.3). Notably, when α is large, a greater increase in α is required in order to explain an incremental increase in the X-to-autosome ratio (because of the form of *f*; Fig. 1.3A). In this case, even moderate differences in the ratio of generation times translate into large biases in estimates of α (Figs. 1.3B and A1.3). As differences in generation times among sexes are common (6), this consideration suggests that current estimates of the male mutation bias might suffer from substantial bias and uncertainty (Fig.

1.3B and A1.3), and that information about the ratio of generation times (e.g., from extant species) is key to assessing this uncertainty.

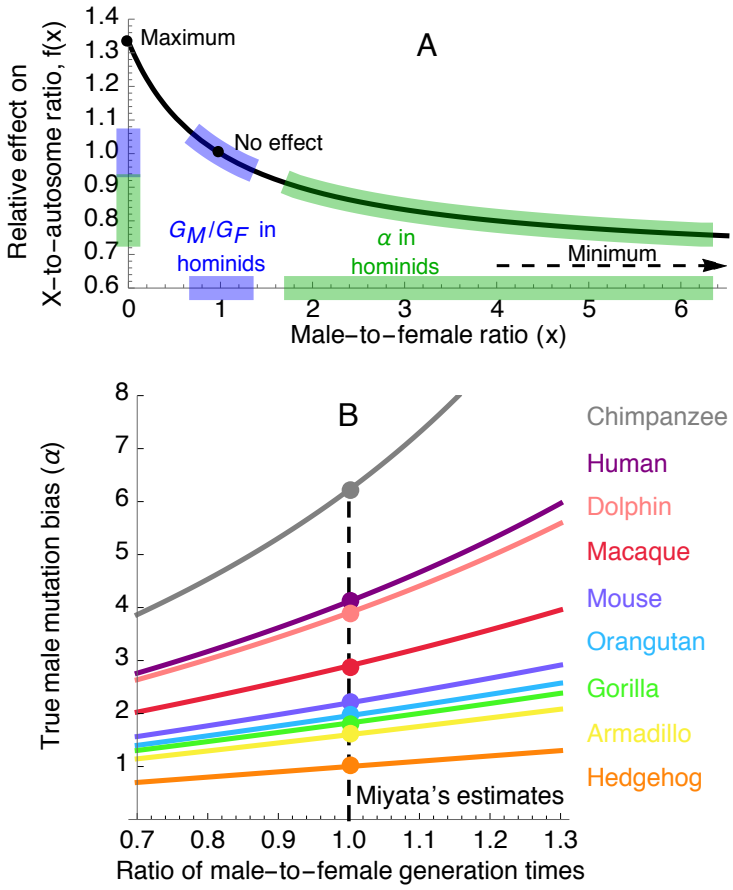

Fig. 1.3. The effects of the male-to-female ratio of generation times on estimates of the male mutation bias, α. (A) The function f mediating the effects of the ratios of male-to-female generation times and mutation rates on the X-to-autosome ratio of substitution rates. (B) The potential biases in estimates of α in 9 mammalian species, as a function of the ratio of male-to-female generation times. The curve for each species is based on the X-to-autosome divergence ratio reported along its lineage (9; these ratios likely underestimate the X-to-autosome ratios of substitutions due to the contribution of ancestral polymorphism, 23).

X-to-Autosomes ratios in hominines. We use our model and parameter estimates to investigate how life history and spermatogenesis affect the X-to autosome ratio in hominines. Increasing the paternal generation time has opposing effects on the ratio (Eq. 1.5), as it increases both the male mutation bias (α; Fig. A1.4) and the ratio of generation times (G_M/G_F). In contrast, changing the maternal generation time affects only the ratio of generation times, resulting in a greater effect on the X-to-autosome ratio (Fig. 1.4A). In turn, both earlier male puberty and an increased rate of spermatogenesis increase mutation rates in males, resulting in a lower ratio (Figs. 1.4B and 1.4C). Based on the known parameter ranges in hominines, the X-to-autosome ratio varies by ~6.9% due to maternal age, whereas the variation due to each of the other parameters is smaller (~1.2-2.7%).

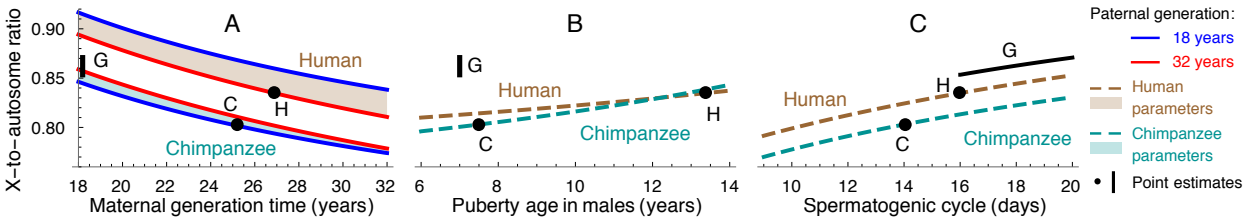

Fig. 1.4. Predicted X-to-autosome ratios of substitution rates as a function of male and female generation times (A), male age of puberty (B) and rate of spermatogenesis (C). Other details are the same as in Fig. 1.2.

In hominines, X-to-autosome ratios have garnered considerable attention, in part because of what seems like puzzlingly low ratios (8, 9, 37). Estimates of the ratios of neutral divergence are ~0.76 for the chimpanzee lineage, ~0.8 for the human lineage and ~0.9 for gorillas (9). These estimates, however, do not correspond directly to the ratios that we are considering. Notably, they include the contributions of ancestral polymorphism, which are greater on autosomes (8, 37), suggesting that the ratio after the species splits is closer to 1. Also, they do not control for differences in the

average mutability of base pairs between X and autosomes (e.g., due to local effects of base composition (38)). In that regard, the standard practice of dividing by divergence to an outgroup only complicates the interpretation by compounding estimates of the X-to-autosome ratios in hominines with the ratio on the outgroup lineage. For all these reasons, we consider existing estimates to be rough.

These caveats notwithstanding, we ask whether the observed values can be explained by considering our mutational model. Using parameter estimates from extant species and assuming a spermatogenic cycle between 16 and 20 days in gorillas, the model predicts an X-to-autosome ratio of 0.80 on the chimpanzee lineage, 0.83 on the human lineage and between 0.85 and 0.87 on the gorilla lineage. These estimates recover the ordering of ratios among lineages as well as the rough magnitude of the reduction below 1. Further allowing individual mutational parameters to vary within their ranges in extant species can explain both a smaller ratio in chimpanzees and a greater difference between chimpanzees and gorillas (Table A1.9). Thus, it is plausible that observed X-to-autosome ratios could be explained by differences in rates of spermatogenesis (as first suggested by (9)) and life history, without recourse to elaborate demographic scenarios. We note, however, that these factors cannot explain the greater variation in divergence levels on the X compared to autosomes, which has been suggested to reflect the footprints of linked selection in the ancestral population or during speciation in the presence of gene flow (8, 37).

1.5 Discussion

Pedigree-based estimates of the mutation rate in humans should allow for improved inferences about the timing of species' splits and other evolutionary events (e.g., Out-of-Africa migrations). To translate estimates from pedigrees into yearly mutation rates, the standard approach is to divide

the sex-averaged mutation rate per generation in a study by estimates of the sex-averaged generation time along the lineage under consideration. This calculation assumes that the mutation rate per generation remains constant, implicitly reflecting an extreme interpretation of the generation time effect (11). It also ignores other life history traits and rates of spermatogenesis that differ between pedigree studies and the lineage of interest, and whose effects are likely to be substantially greater than those of the sex-averaged generation time. Thus, translating the results of pedigree studies into yearly rates is not as straightforward as it seems, and requires consideration of these life history factors.

A more appropriate approach is to fit the results of pedigree studies to a mutational model that reflects the dependency of yearly rates on age and sex. The fitted model can then be used to estimate yearly mutation rates and the uncertainty associated with them, based on estimates of life history parameters for the lineage under consideration (e.g., based on estimates for extant species). We illustrate this approach with a simple mutational model for hominines and show that it leads to substantial changes in estimates of yearly rates and corresponding split times. Importantly, it revises split times for the human-chimpanzee split downwards, from ~10 MYA ((3) and Appendix section A1.5) to as low as ~6.6 MYA (see Fig. 1.5 and Table A1.9).

Our mutational model will surely be refined as we learn more about germline mutations. For example, molecular evolutionary patterns suggest that rates of certain types of mutations (notably, transitions at CpG sites; 21, 38) may track some combination of absolute time and number of cell divisions (11). If so, maternal age could affect the accrual of these mutations (11). In addition, there is still uncertainty about the number of cell divisions in the male germline during spermatogenesis. In particular, it has been suggested that A_{pale} cells—the stem cells from which sperm are generated—are replenished by, or experience turnover with, A_{dark} spermatogonial stem

cells throughout adulthood, resulting in fewer cell divisions in the germline between puberty and reproduction (32, 39, 40). If this hypothesis is true, it would suggest that the mutation rate per-spermatogenic cell division (D_M) is higher than our estimates and closer to the (considerably higher) rates per-division in females, in males pre-puberty (11) and in other taxa (cf. Table A1.2 and (41)). In terms of the hominine mutational model, such a revision would introduce an additional parameter for turnover, which could lead to greater variation among species in male mutation rates post-puberty (40). These refinements notwithstanding, our mutational model already suggests several predictions that align with observations. For instance, our predicted mutation rates in chimpanzees fall within confidence intervals of a recent pedigree study (24). We also predict the reduction in branch length leading to humans compared to gorillas and provide a plausible explanation for the X-to-autosome ratios of substitutions rates observed in hominines.

Our results also bear on the recently invigorated discussion about split times in the catarrhine (i.e., old world monkeys and apes) phylogeny (3). For instance, assuming that current pedigree-based estimates reflect yearly rates along the hominid phylogeny places the human-orangutan split at ~40 MYA, much earlier than all hominid or even hominoid fossils (cf. Appendix section A1.5 and (3, 35)), putting results from human genetics at odds with those from paleontology. Attempts to reconcile pedigree and fossil-based evidence appear to be moving from both ends. From one end, it has been suggested that mutation rates may have experienced a slow down in the hominoid phylogeny towards the present (cf. 3, 21, 35). From the other, the new mutation rate estimates triggered a reevaluation of the (already controversial) phylogenetic positioning of key catarrhine fossils, and specifically those used to place bounds on the hominine-orangutan and hominoid-cercopithecoid (i.e., old world monkeys) split times (42). More realistic models of the molecular clock can inform this discussion by suggesting the plausible extent of the mutational slowdown on

different phylogenetic timescales and, as a result, by better delineating which fossil positions would need to be revised for their ages to be consistent with pedigree studies.

As an illustration, we use our mutational model, with estimates of life history and spermatogenesis parameters from extant species, to suggest plausible ranges for yearly mutation rates on the hominoid phylogeny (Fig. 5; Table A1.9 and Appendix section A1.5). We allow each of the parameters on branches following a split to vary independently within the range observed in descendant species; in the few cases in which we lack estimates in extant species, we also rely on information from an outgroup (see Appendix section A1.5). Under these assumptions, ranges for individual parameters and the resulting yearly mutation rates can only become larger when we go farther back in time. Interestingly, only the upper bound on the range of mutation rates increases as we go farther back in time, supporting the notion of a slowdown. Moreover, the inferred ranges could, in theory, reconcile the apparent discrepancy between yearly rates of 0.4×10^{-9} per bp observed in pedigree studies, and yearly rates between $0.65 - 1.1 \times 10^{-9}$ per bp inferred from the fossil record (Fig. 1.5 and Table A1.11). The assumption that life history and spermatogenesis parameters vary independently might be overly permissive, however, because combinations of these traits could be under stabilizing selection.

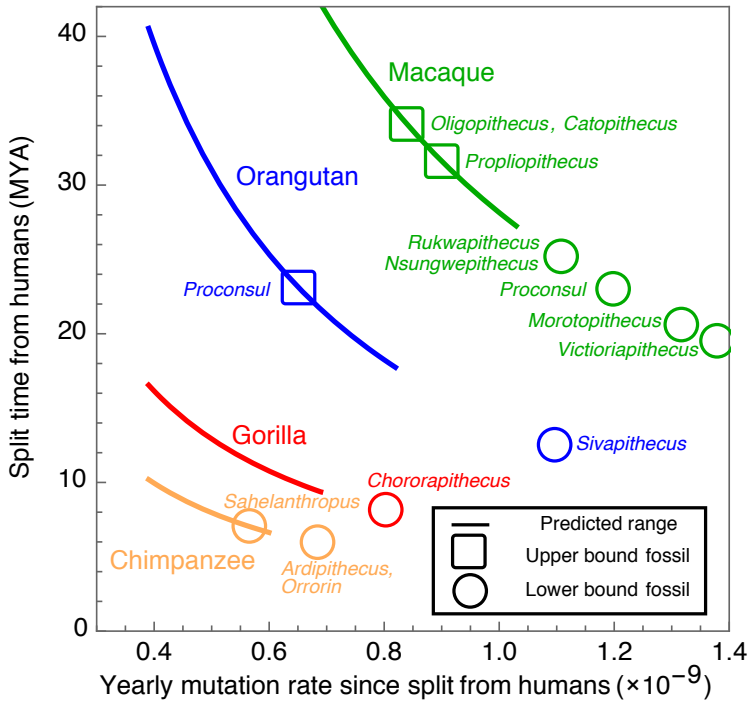

Fig. 1.5. Estimated ranges of split times from humans (y-axis) and the yearly mutation rates since the split (x-axis). The inference is detailed in Appendix section 1.5. Upper (squares) and lower (circles) bounds on spit times are based on the hypothesized phylogenetic positioning of fossils (cf. Appendix section 1;42).

Moving to broader phylogenetic contexts, how can we reconcile our results suggesting that the (sex-averaged) generation time has a moderate effect in hominines (Fig. 1.2; also cf. (11, 30)) with phylogenetic studies showing that it is a major predictor of the rates of neutral substitutions in mammals (12, 13, 22)? Assuming that the number of mutations tracks the number of cell divisions in the germline, a possible answer may arise from the way the number of divisions relates to the generation time. Comparing mice, with a generation time of ~9 months, and humans, with a generation time of ~30 years, the estimated numbers of cell divisions in the germline are: 25 and 31 in females, 27 and 34 pre-puberty in males, and 35 and 390 during male spermatogenesis,

respectively ((31) and Table A1.1). If we assume more generally that only the number of cell divisions during spermatogenesis increases rapidly with the generation time, then many of the mutations in hominines occur during spermatogenesis because of their exceptionally long generation times. In species with short generation times, the relative contribution of spermatogenesis would be much smaller and therefore the rate per-generation would be roughly constant; equivalently, in such species, the yearly mutation rate would be roughly inversely proportional to the generation time (43).

To draw out these implications, we extrapolate our hominine mutational model to a wider range of life history and spermatogenesis parameter values roughly corresponding to mammals (Fig. 1.6), acknowledging that such an extrapolation provides only a qualitative depiction as underlying mutational parameters may vary among mammals ((41) and Table A1.2). We find that the sex-averaged generation time dominates the variation in yearly mutation rates when species with shorter generation times are included in the comparison (Fig. 1.6A). The average generation time also affects the expected X-to-autosome ratios (with shorter generation times corresponding to lower α and a ratio closer to 1; (43)), though its effect is expected to be much more moderate and comparable with those of other life history traits (Fig. 1.6B and Fig. A1.5). Both of these predictions accord with observations in mammals (12, 13, 22).

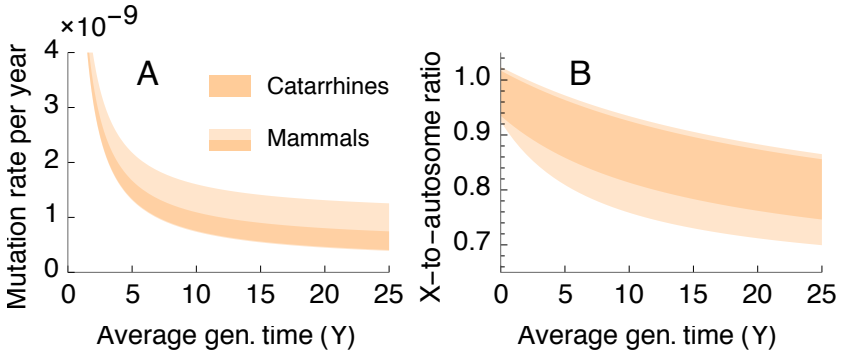

Fig. 1.6. The generation time effect on the molecular clock in a broader phylogenetic context. The yearly mutation rates (A) and X-to-autosome ratios (B) as a function of the sex averaged generation time, based on our mutational model. Other parameter ranges roughly correspond to catarrhines and mammals (Table A1.4).

Chapter 2 - Life history effects on neutral polymorphism levels of autosomes and sex chromosomes

2.1 Abstract

In human and other hominid (great apes) populations, estimates of the relative levels of neutral polymorphism on the X and autosomes differ from each other and from the naive theoretical expectation of ¾. These differences have garnered considerable attention over the past decade, with studies highlighting the potential importance of several factors, including historical changes in population size and linked selection near genes. Here, we examine a more realistic neutral model than has been considered to date, which incorporates sex- and age-dependent mortalities, fecundities, reproductive variances and mutation rates, and ask whether such a model can account for diversity levels observed far from genes. To this end, we derive analytical expressions for the X to autosome ratio of polymorphism levels, which incorporate all of these factors and clarify their effects. In particular, our model shows that the genealogical effects of life history can be reduced to ratios of sex-specific generation times and reproductive variances. Applying our results to hominids by relying on estimated life-history parameters and approximate relationships of mutation rates to age and sex, we find that life history effects, and the effects of male and female generation times in particular, may account for much of the observed variation in X to autosome ratios of polymorphism levels across populations and species.

2.2 Significance Statement

All else being equal, the ratio of diversity levels on X and autosomes at selectively neutral sites should mirror the ratio of the number of chromosomes in the population and thus equal ¾. In reality, the ratios observed in humans and other species differ substantially from ¾ and from each other. Because autosomes spend an equal number of generations in both sexes while the X spends twice as many generations in females, these departures from the naïve expectation may reflect differences between male and female life histories and mutation processes. We derive theoretical predictions for these effects, and show that they likely played an important part in shaping X to autosome ratios in populations of humans and closely related species.

2.3 Introduction

Neutral polymorphism patterns on X (or Z) and autosomes reflect a combination of evolutionary forces. Everything else being equal, the ratio of X to autosome polymorphism levels should be ¾, because the number of X-chromosomes in a population is ¾ that of autosomes. However, autosomes spend an equal number of generations in diploid form in both sexes, whereas the X spends two thirds of the number of generations in diploid form in females and a third in haploid form in males. As a result, the ratio of X to autosome polymorphism levels can also be shaped by differences in male and female life history and mutation processes as well as by differences in the effects of linked selection (44).

The differential effects of these factors on X and autosome has been discussed for many species (45-47), and in particular, they have been the focus of considerable recent interest in humans and other hominids (48-56). In particular, it has been noted that linked selection could affect X and

autosomes differently because of differences in recombination rates, the density of selected regions, and the efficacy and mode of selection. Notably, the hemizygosity of the X in males leads to a more rapid fixation of recessive beneficial alleles and to a more rapid purging of recessive deleterious one (57, 58). Accounting for these effects and for differences in recombination rates suggests that in humans—in mammals more generally—the effects of linked selection should be stronger on the X ((59), but see (47)). To evaluate these effects empirically, several studies have examined how polymorphism levels on the X and autosomes vary with genetic distance from putatively selected regions, e.g., from coding and conserved non-coding regions (49, 53-55, 60, 61). In most hominids, including humans, such comparisons confirm that linked selection reduces X to autosome ratios. They further suggest that the effects are minimal sufficiently far from genes (49, 60), potentially allowing the effects of other factors shaping X to autosome ratios to be examined in isolation.

Even far from genes, however, the X to autosome ratios in humans and other hominids appear to differ from the naive expectation (49, 53, 55). To control for the effects of higher mutation rates in males and variation in mutation rates along the genome, polymorphism levels on X and autosomes are typically divided by divergence to an outgroup. In regions far from genes, the normalized X to autosome ratios range between ¾ and 1 among human populations, generally decreasing with the distance from Africa (48, 49, 53, 54). Ratios exceeding ¾ have also been observed in most other hominids ((55), but see (56)).

Proposed explanations for these departures from ¾ and the variation among populations and species fall into two main categories: those based on population history, such as historical changes in population size, and those based on life history, such as differences between male and female reproductive variances (i.e., the variance in the number of offspring that individuals have

throughout their life). If we assume that the effective population size on the X is generally smaller than on autosomes, then changes in population size will have a different impact on polymorphism levels on X and autosomes (62-65). Notably, population bottlenecks that occurred sufficiently recently, such as the Out of Africa bottleneck in human evolution, will have decreased the X to autosome ratio, because a greater proportion of X-linked loci will have coalesced during the bottleneck (65). While what is known of human population history (including the Out of Africa bottleneck) is indeed expected to lead to lower X to autosome ratios in non-Africans, estimates of the expected reduction in diversity levels fall short of explaining its full extent (54). Differences in population history between males and females may have also contributed to the differences among human populations. For example, Keinan and Reich (50) suggested that an increase in the ratio of males to female effective population sizes (or generation times) during an Out-of-Africa bottleneck contributed to the lower X to autosome ratios in non-Africans. While the possibility of inferring such historical differences from extant polymorphism patterns is intriguing, and there are comparable, well established sex-differences in the more recent past (e.g., (66)), differences in life history may offer a more straight-forward explanation of the observations.

Notably, sex differences in life history traits that are readily observed today could have substantially affected X to autosome ratios. Among traits, perhaps the most straightforward effect arises from the higher reproductive variance in males than in females (e.g., due to sexual selection), which causes higher coalescence rates on autosomes and an increased X to autosome ratio (48, 67). Theoretical studies suggest that the resulting increased ratio is bound by $^{9}/_{8}$ (67). In reality the effect is likely much smaller, yet differences in male and female variance among extant human populations and hominids (68) suggest that it could have contributed to observed differences in X to autosome ratios, as well as account for why ratios often exceed $^{3}/_{4}$.

Differences between male and female generation times could have affected X to autosome ratios in more intricate but substantial ways. Mutation rates in mammals increase with paternal age (and to a lesser extent with maternal age), notably in humans (2, 24, 26, 69). Ignoring genealogical effects (see below), we would therefore expect increased paternal generation times to decrease the X to autosome ratios of polymorphism and divergence. In that regard, given that male and female generation times vary considerably among populations and species (5, 6), the common practice of dividing polymorphism estimates by divergence levels might not fully control for male mutational biases and could result in X to autosomes ratios that deviate from ¾. Sex differences in generation times also affect the genealogical process, with longer generation times in males than in females resulting in slower coalescence rates on the X compared to autosomes, and therefore in increased polymorphism ratios.

The overall effect of these factors has not been considered jointly. We do so here, by studying how these life history traits and mutational differences between sexes affect neutral polymorphism levels on X and autosomes. Earlier work by Felsenstein (70) modeled the effects of age structure on the effective population size of haploid populations, and that work was later extended by Hill (15), Johnson (71), Charlesworth (16, 67) and Orive (72), who incorporated the effects of age structure in diploid populations, as well as the effects of sex- (but not age-) dependent rates of mutation. Where our work departs from those is in building on coalescent models of age-structured populations developed by Sagitov and Jagers (73) and Pollack (74), which we describe below.

2.4 Results

The haploid model. To illustrate how we treat the coalescence process in an age-structured population, we first consider the haploid model proposed by Felsenstein (70). We assume a haploid

panmictic population that is divided into age classes of one year (for convenience), and denote the (constant) number of individuals of age $a \geq 1$ by M_a, where $M_{a+1} \leq M_a$ due to mortality (see Table 2.1 for a summary of notations). We further assume that each of the M_1 newborns is independently descended from a random parent, with the parent's age chosen from a distribution $A = (p_a)_{a=1}^{\infty}$ with expectation G (the generation time). The parent is then chosen with uniform probability within the age class. We begin with the case without endogenous reproductive variance, but incorporate this effect below.

Notation	Definition
$\boldsymbol{p_a}$	Probability that a newborn descends from a parent of age a
$\boldsymbol{M_a}$	Number of individuals of age a
$\boldsymbol{G}$	Average generation time
$\boldsymbol{M}$	Effective age-class size
$\boldsymbol{\vec{r}}$	Relative reproductive success, where component r_a is the relative reproductive success at age a
$\boldsymbol{W_{i,j}}$	Expected value of $r_i \cdot r_j$ conditioned on survival to age j ($j \geq i$)
$\boldsymbol{W}$	Weighted average of $W_{i,j}$

Table 2.1: Notation for the haploid model.

This model was solved by Sagitov and Jagers (73) in a coalescent framework, and here we provide an intuitive account of the solution. First, we consider the rate of coalescence for a sample of two alleles, where to this end, we trace their lineages backward in time. For the alleles to coalesce in a

given age class a, one of them would have to be a newborn in the previous generation; this occurs with probability p_a/G per year. The other allele would have to be present in the same age class, which occurs with probability $\sum_{j\geq a} p_a\,/G$, because it could have been born to a parent of age $j \geq a$, $j - a + 1$ generations ago (the rigorous derivation establishes that this is in fact the stationary age distribution along a lineage). Both alleles would also have to be in the same individual in age class a, which occurs with probability $1/M_a$. Taken together, we find that the probability of coalescence per-year in age class a is

$$2 \cdot \left(\frac{p_a}{G}\right) \cdot \left(\frac{\sum_{j\geq a} p_j}{G}\right)\left(\frac{1}{M_a}\right) - \left(\frac{p_a}{G}\right)^2 \left(\frac{1}{M_a}\right),$$

where we multiplied by 2, because either allele could have been the newborn in the previous generation, but subtracted the probability that they both were, because this event should be counted only once. The probability of coalescence per generation and corresponding effective population size follow from multiplying these probabilities by the generation time, summing them over age classes, and rearranging terms:

$$\frac{1}{N_e} = \frac{1}{G}\sum_a \frac{w_a}{M_a}, \qquad [2.1]$$

where $w_a = p_a^2 + 2\sum_{j>a} p_a p_j$. Note that the (a, j)-term in w_a is proportional to the probability that coalescence in age class a occurs in an individual that fathered a newborn carrying one of the alleles at age a, and a newborn carrying the other allele at age j. Moreover, the w_a terms add up to one, allowing us to define the effective age class size as the weighted harmonic mean

$$\frac{1}{M} = \sum_a \frac{w_a}{M_a}. \qquad [2.2]$$

The effective population size can then be viewed as the product of the effective age-class size and the effective number of age-classes, which is simply the expected generation time G, i.e.,

$$N_e = G \cdot M. \qquad [2.3]$$

Next, we extend the model to incorporate endogenous reproductive variance (as opposed to the variance introduced by stochastic mortality and birth). To this end, we assume that each newborn is assigned a vector $\vec{r}$ describing its age-dependent, relative reproductive success, such that its probability of being chosen as a parent among the individuals of age class a is r_a/M_a (i.e., r_a corresponds to the expected, rather than actual, reproductive success of the individual at age a). We further assume that the proportion of individuals with a given vector $\vec{r}$ that reach age a, $f_a(\vec{r})$, can vary with age, in effect allowing for dependencies between age-specific mortality and reproductive success. Thus, the model is set up to allow for dependencies between reproductive success, fecundity and longevity, examples of which have been observed in natural populations (e.g., between the age of first reproduction and longevity (75, 76), or between reproductive success and longevity (76, 77)).

The model thus extended can be solved along the same lines as described above (see Appendix section A2.1.2). Specifically, the coalescence rate per generation and corresponding effective population size take a similar form:

$$\frac{1}{N_e} = \frac{1}{G}\sum_a \frac{w_a}{M_a}, \qquad [2.4]$$

but in this case

$$w_a = p_a^2 W_{a,a} + 2\sum_{j>a} p_a p_j W_{a,j}, \qquad [2.5]$$

where $W_{a,j} \equiv E_{\vec{r}}\left(r_a \cdot r_j \middle|\, \text{individual reaches age} \geq j\right)$ for $j \geq a$. As in the simpler case, the (a, j)-term in w_a is proportional to the probability that coalescence in age class a occurs in an individual that fathered a newborn carrying one of the alleles at age a, and a newborn carrying the other allele at age j; but in this case, the coefficients $W_{a,j}$ factor in the effect of endogenous

reproductive variance. In contrast to the simpler case, the w_a terms do not necessarily add up to 1. We therefore introduce a normalization by $W = \sum_a w_a$, and define the effective age class size as

$$\frac{1}{M} = \frac{1}{G}\sum_a \frac{w_a/W}{M_a}. \quad [2.6]$$

In these terms, the effective population size takes the form

$$N_e = G \cdot M/W. \quad [2.7]$$

To provide some intuition, we consider the special case in which relative reproductive success is independent of age and of mortality rates. Namely, each newborn is assigned a relative reproductive success r at birth, and its probability of being chosen as a parent among the individuals of age class a is r/M_a. The distribution of r values then has expectation 1 (by construction) and we denote its variance by V_r ($V_r = 0$ when there is no endogenous reproductive variance). In this case, the coalescence rates in any given age class are increased by a factor $1 + V_r$, and therefore the effective population size is

$$N_e = \frac{1}{1+V_r} G \cdot M. \quad [2.8]$$

Thus, $W = 1 + V_r$ and it can be interpreted as the reproductive variance caused by the Poisson sampling of parents, which contributes 1, and by the endogenous reproductive variance, which contributes V_r. In turn, the contribution of age-structure to reproductive variance is reflected in $G \cdot M$.

More generally, the results of the full model can be recast in terms of the total reproductive variance. First consider a haploid Wright-Fisher process, i.e., with non-overlapping generations, with endogenous reproductive variance, modeled similarly to the example considered above. In this case, Wright (78) showed that the effective population size is

$$N_e = N/V, \quad [2.9]$$

where N is the census population size and $V = 1 + V_r$ is the total reproductive variance. Second, in the case with age-structure but without endogenous reproductive variance, Hill showed that the effective population size can also be written as

$$N_e = G \cdot M_1/V, \qquad [2.10]$$

where $G \cdot M_1$ is the number of newborns per generation, and V is the reproductive variance introduced by age-structure (15). Comparing Eqs. 2.7 and 2.10, we can express this variance as $V = M_1/M$ (≥ 1). This expression makes intuitive sense, as this would be the reproductive variance in the case considered by Wright, if the next generation were randomly chosen from a reproductive pool including only M out of M_1 individuals in the population. In Appendix section A2.1.3, we show that Eq. 2.10 also holds for the general model with age-structure and endogenous reproductive variance. In this case, the total reproductive variance is

$$V = (M_1/M) \cdot W, \qquad [2.11]$$

where the first term in this product, M_1/M, is the variance introduced by stochasticity in mortality and birth, and the second term, W, reflects the contribution of endogenous reproductive variance. Thus, we conclude (i.e., in Eq. 2.10) that all the effects of age-structure and endogenous reproductive variance (as well as any dependence between them) on the effective population size can be summarized in terms of the generation time, G, number of newborns per year, M_1, and total reproductive variance, V.

The model for X and autosomes. The diploid model with two sexes is more elaborate, but it is defined and solved along the same lines that we have described for the haploid model. In Appendix section A2.2.2, we describe the model formally, and here, for brevity, we only describe how it differs from the haploid model. Notably, in this case, we allow for age-dependent mortality,

fecundity, and endogenous reproductive variance to vary between the sexes. We further assume equal sex ratios at birth (although we also consider the general case in Appendix section A2.2), and accommodate X and autosomal modes of inheritance (i.e., X linked loci in males always descend from females, whereas in all other cases the sex of parents is randomly chosen with probability ½). In Appendix section A2.2.3, we solve this model for the stationary distributions of sex and age and corresponding coalescence rates, by extending Pollack's results for age-structured populations with two sexes (74) to account for endogenous reproductive variance. Here we describe the main results relying on the intuition gained from the haploid model.

By analogy with the haploid case (Eq. 2.10), we can express the effective population sizes for X and autosomes in terms of the reproductive variance of males and females. To this end, we denote the number of newborns per year, including both sexes, by M_1 (see Table 2.2 for a summary of notations). We further define the generation times for X and autosome linked loci as averages of the male and female generation times, G_M and G_F, weighted by the proportions of generations that they spend in each sex, i.e.,

$$G_X = \frac{2}{3}G_F + \frac{1}{3}G_M \text{ and } G_A = \frac{1}{2}(G_M + G_F). \qquad [2.12]$$

To draw the analogy with the haploid case, we consider the effective population size in terms of alleles rather than individuals (they are equivalent for haploids). Specifically, we define the reproductive success of an allele as the number of an individual's offspring carrying that allele, and denote the reproductive variance associated with X and autosome linked alleles by V_X^* and V_A^*. In these terms, the effective population sizes for X and autosomes are analogous to Eq. 2.10:

$$\frac{3}{2}\cdot N_e^X = G_X\left(\frac{3}{2}\cdot M_1\right)/V_X^* \text{ and } 2\cdot N_e^A = G_A(2\cdot M_1)/V_A^*, \qquad [2.13]$$

where the factor of 3/2 for X and 2 for autosomes follow from considering the effective size of alleles rather than individuals on the left-hand side, and the number of newborn alleles rather than individuals on the right-hand side. Expressing these results in terms of the reproductive variance of males, V_M, and of females, V_F, in Appendix section A2.2.5, we show that

$$V_X^* = \frac{1}{4}(2 + \frac{1}{3}V_M + \frac{2}{3}V_F) \text{ and } V_A^* = \frac{1}{4}(2 + \frac{1}{2}V_M + \frac{1}{2}V_F), \quad [2.14]$$

where the weights reflect the proportion of generations spent in males and females and the additive factor 2 results from ploidy. Substituting these expressions into Eq. 2.13 we find that

$$N_e^X = \frac{4G_X M_1}{2+\frac{1}{3}V_M+\frac{2}{3}V_F} \text{ and } N_e^A = \frac{4G_A M_1}{2+\frac{1}{2}V_M+\frac{1}{2}V_F}. \quad [2.15]$$

Notation	Definition
$\boldsymbol{M_1}$	Number of male and female newborns per-year
$\boldsymbol{G_M, G_F}$	Average generation times in males and females
$\boldsymbol{G_X, G_A}$	Average generation times for X and autosomes
$\boldsymbol{V_X^*, V_A^*}$	Reproductive variance for X and autosome linked alleles
$\boldsymbol{V_M, V_F}$	Reproductive variance of males and females
$\boldsymbol{\mu_M, \mu_F}$	Average mutation rate per generation in males and females
$\boldsymbol{\mu_X, \mu_A}$	Average mutation rate on the X and autosomes

Table 2.2: Notation for the diploid model with two sexes.

Polymorphism levels on X and autosomes. Having solved for the effective population sizes, we now turn to polymorphism levels. To this end, we assume that the expected mutation rates in males and females depend linearly on age (2, 11), and that their expectations for generation times G_M

and G_F are μ_M and μ_F respectively (see Appendix section A2.3 for general dependency). The expected mutation rates per generation on X and autosome linked lineages are weighted averages of these expected rates,

$$\mu_A = \frac{1}{2}(\mu_M + \mu_F) \text{ and } \mu_X = \frac{2}{3}\mu_F + \frac{1}{3}\mu_M. \quad [2.16]$$

The expected heterozygosity on X and autosomes then follows from the standard forms:

$$E(\pi_A) = 4N_e^A\mu_A \text{ and } E(\pi_X) = 3N_e^X\mu_X. \quad [2.17]$$

Note that these expressions are usually derived assuming that the genealogical and mutational processes are independent (79). This assumption is violated here, because both the coalescence and mutation rates depend on the ages along a lineage. In Appendix section A2.3, we show that the standard forms hold nonetheless.

We can now combine our results to derive an expression for the X to autosome ratio of polymorphism levels. From Eqs. 2.15 and 2.17 and rearranging terms, we find that the ratio of expected heterozygosity is

$$\frac{E(\pi_X)}{E(\pi_A)} = \frac{3}{4} \cdot \frac{f(\mu_M/\mu_F) \cdot f(G_M/G_F)}{f\left(\frac{2+V_M}{2+V_F}\right)}, \quad [2.18]$$

where $f(x) = \frac{2x+4}{3x+3}$ (see Eq. S86 for the case in which the sex ratio at birth differs from 1). When the mutation rates, generation times, and reproductive variance are identical in both sexes, this expression reduces to the naïve neutral expectation of ¾. When these factors differ between sexes, Eq. 2.18 provides a simple expression for the effect of each factor. Notably, the effects of age-structure and endogenous reproductive variance reduce to effects of male to female ratios of generation times and of reproductive variances (more precisely $(V_M + 2)/(V_F + 2)$).

Comparing ratios of polymorphism and divergence. We can now ask how changes in these factors affect X to autosome ratios of polymorphism and divergence. More precisely, considering the numbers of substitutions rather than divergence (i.e., ignoring multiple hits) and excluding the contribution of ancestral, the equivalent expression to Eq. 2.18 is

$$\frac{E(K_X)}{E(K_A)} = \frac{f(\mu_M/\mu_F)}{f(G_M/G_F)} \qquad [2.19]$$

(4, 16). While a higher ratio of male to female mutation rates, $\alpha = \mu_M/\mu_F$, has the same effect on X to autosome ratios of polymorphism and divergence, a higher ratio of male to female generation times (G_M/G_F) has opposite effects, decreasing the ratio for polymorphism but increasing it for divergence (Fig. 2.1). The difference arises because one way in which the male to female ratio of generation times affects polymorphism levels is by changing the relative rates of coalescence on X and autosomes, whereas there is no such effect on divergence levels, for which coalescence times on both X and autosomes equal the species split time (again, neglecting the contribution of ancestral polymorphism). By the same token, reproductive variances affect only the coalescence rates, and thus polymorphism but not divergence levels. Importantly, the different effects of life history factors on polymorphism and divergence complicate the interpretation of X to autosome polymorphism ratios that are divided by divergence to an outgroup (e.g., (48, 49, 53, 80)) (see Discussion).

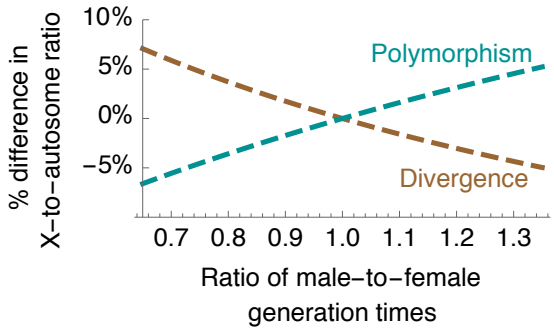

Fig. 2.1: Generation time effects on X to autosome divergence and polymorphism ratios.

X to autosome polymorphism ratios in humans and other hominids. Thus far, we considered general results for the effects of life history on the X to autosome ratio of polymorphism levels, and now we turn to the effects in human populations and in other hominid species. To this end, we rely on a model that approximates the dependence of hominid mutation rates on sex and age (11, 26). We assume that mutations accumulate linearly with the number of germ-cell divisions, and that the rate per division varies at different stages of development (11). In females, the number of cell divisions does not depend on age because all oogonial mitotic divisions occur before birth (81). In males, germ-cell division exhibits two main phases, with an approximately constant number until puberty, followed by a linear rate of division in adult testis during spermatogenesis (31, 32). We previously (4) fitted the corresponding mutation model to the relationship between the number of mutations and the parents' sex and age observed in human pedigrees (2), to obtain the following approximations for female and male mutation rates per bp per generation:

$$\mu_F = 5.42 \cdot 10^{-9} \text{ and } \mu_M = 6.13 \cdot 10^{-9} + 3.33 \cdot 10^{-10}(G_M - P)/\tau, \qquad [2.20]$$

where P is the puberty age in males and τ is the duration of the seminiferous epithelial cycle (during spermatogenesis). Recent pedigree studies indicate that maternal age also affect mutation rates (e.g. (69)), more generally suggesting that some mutations are in fact spontaneous rather than

replication driven (11, 30). At present, however, the studies are not sufficiently precise for us to incorporate maternal age and spontaneous effects in our mutational model. Once they are, it should be straight forward to study how they affect the results that we report below.

Given our mutational model and relying on prior knowledge about life history parameters, we consider how each of the factors detailed in Eq. 2.18 affects deviation of the X to autosome ratio from the naïve expectation of ¾, and how they affect variation in the ratio among populations and species. We begin with the effect of sex differences in reproductive variance (Fig. 2.2). This effect has been noted by many (e.g., (67, 82)) and follows intuitively from the fact that a higher male variance results in higher coalescence rates on autosomes than on the X, thus increasing π_X/π_A. In theory, we know that higher reproductive variance in males can increase π_X/π_A by as much as 50% (with an upper bound of $^9/_8$ compared to $^3/_4$), but in practice, the effect is expected to be much smaller. Consider, for example, a model in which only a fraction p of the males can reproduce and the number of offspring for each follows a Poisson distribution with mean $2/p$, while all females reproduce with equal probabilities following a Poisson with mean 2, resulting in a ratio $(V_M + 2)/(V_F + 2) = p^{-1}$. In this case, assuming that only $p = 0.3$ of males can reproduce would result in an X to autosome polymorphism ratio that exceeds ¾ by only 22% (i.e., much lower than the theoretical limit). Unfortunately, little is known about the plausible range of reproductive variances in humans. In five hunter-gatherer groups in which this variance was measured in small samples, it was found to be 1.7-4.2 folds higher in males (68), corresponding to an increase of 6-20% relative to ¾, which translates into a relative difference of ~13% among populations.

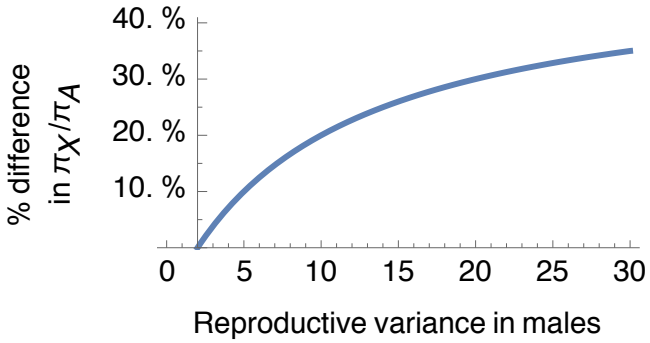

Fig. 2.2: The effect of male reproductive variance on the ratio of X to autosome polymorphism levels, assuming the female reproductive variance equals 2.

Next, we consider the effect of the ratio of male to female generation times (Fig. 2.3A). In seven hunter-gatherer groups in which these were measured, the mean generation times vary between 25 and 33 years and the ratio of male to female generation times between 1.03 and 1.37 (4, 5). Accounting for both mutational and genealogical effects, this variation corresponds to an 18%-24% reduction, and an ~8% relative difference in X to autosome polymorphism ratios. Intriguingly, the relative effect is on the order of the variation in observed X to autosome ratios among human populations, which are 6%-9% lower in Europeans and 8%-9% lower in Asians compared to Africans (54). However, the effect of generation times on the X to autosome polymorphism ratio should be averaged over the period that gave rise to extant polymorphism levels. Thus, while these estimates suggest that higher ratios of male to female generation times in non-Africans may have contributed substantially to observed differences, a more precise account of this contribution should factor in the shared evolutionary history among populations, as well as subsequent lineage-specific variation in generation times.

Differences in ratios of male to female generation times, have also been reported among groups of wild chimpanzees, albeit based on small samples (6). Specifically, mean generation times have been estimated to vary between 22 and 27 years and the ratio of male to female generation times

between 0.8 and 1.2, corresponding to a 14%-24% reduction and a ~13% relative difference in X to autosome polymorphism ratios. Assuming such differences in generation times have persisted during the millions of years over which neutral genetic variation accumulated, they would have therefore contributed substantially to the variation in polymorphism ratios among hominid populations and species.

As we already noted, paternal and maternal generation times affect the X to autosome ratios of polymorphism and divergence differently. Based on our mutational model, we can now ask about the combined mutational and genealogical effect of paternal and maternal generation times in hominids. Changing the paternal or maternal generation times has a genealogical effect of the same magnitude (but not direction; see Eq. 2.18), but changing the paternal generation time has a greater effect on the mutation rate per generation than changing the maternal one. Thus, overall, changes in paternal generation time have a greater effect on the X to autosome polymorphism ratio than the maternal one (Fig. 2.3A). In contrast, as we have shown previously ((4)), changes in maternal generation times have a greater effect on the divergence ratio (Fig. 2.3B), because the paternal age has opposing effects on the mutation rate per generation and on the number of generations along a lineage.

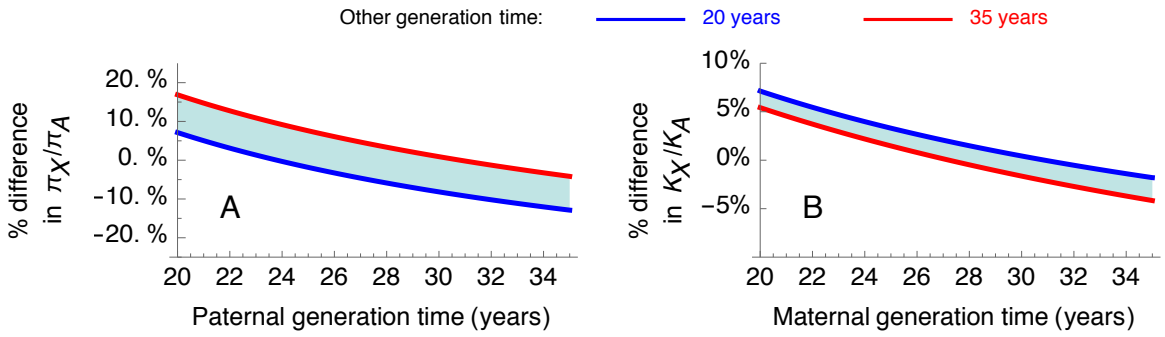

Fig. 2.3: The effects of maternal and paternal generation times on X to autosome ratios of polymorphism (A) and divergence (B) in humans. Note that the range on the y-axis differs between (A) and (B), reflecting the added genealogical effect on polymorphism but not on divergence.

Lastly, the degree to which mutation is male-biased depends on the age of puberty in males and on the seminiferous epithelial cycle length, suggesting that changes to these parameters would also affect the X to autosome polymorphism ratio (Fig. 2.4). For example, varying the age of puberty between 13.5 years, the value estimated for humans (83), and 7.5 years, the value estimated for chimpanzees (84), while holding other life history parameter at their values in humans, decreases X to autosome polymorphism ratios by 3%. Similarly, varying the seminiferous epithelial cycle length from 16 days (humans (85)) to 14 days (chimpanzees (86)) decreases polymorphism ratios by 1%. Differences in these parameters between humans and more remotely related species may be greater, resulting in a more substantial effect on polymorphism and divergence ratios (4). For example, using ages of puberty in cercopithecoids, estimated to be between 3.5-6 years (87), leads to a reduction of 3-4% in polymorphism ratios compared to humans, and assuming the seminiferous epithelial cycle lengths in cercopithecoids is between 9.5-11.6 days (88, 89) leads to a reduction of 3-5%.

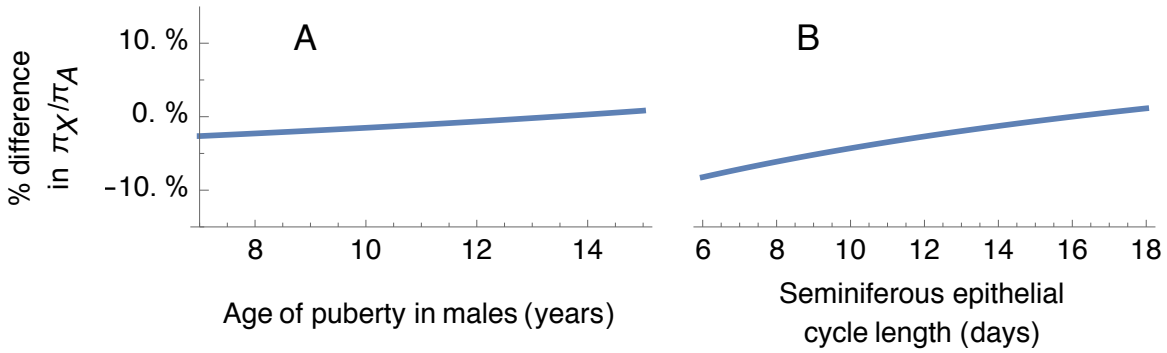

Fig. 2.4: The effects of male puberty age (A) and seminiferous epithelial cycle length (B) on the polymorphism ratio. Other parameters are fixed for their estimates in humans, i.e., generation times of $G_M = 33.8$ and $G_F = 26.9$ years (estimated in hunter-gatherers (4, 5)), male age of puberty of P = 13.5 years (83), and seminiferous epithelial cycle length of τ = 16/365 years (85).

2.5 Discussion

We have derived general expressions for the effects of life history factors on X and autosome polymorphism levels, and explored their plausible effects on levels observed in humans and closely related species. The expression for the ratio (Eq. 2.18) clarifies the effect of each factor, and in particular, shows that the genealogical effect of life history factors reduces to the effect of the ratios of male to female generation times, G_M/G_F, and of reproductive variances, $(V_M + 2)/(V_F + 2)$. These results apply to any species with X and autosomes, and can be readily generalized to species with Z and autosomes. We focused on hominid species, because more is known about the factors affecting their polymorphism levels. Specifically, by considering a hominid model for the effects of life history on mutation rates, we have shown that plausible variation in G_M/G_F within human populations and among closely related species can have a substantial effect on levels of polymorphism on the X compared to the autosomes. It therefore

seems plausible that evolutionary changes in life history contributed substantially to the variation in X to autosome ratios among human populations and hominid species, and to their seemingly ubiquitous deviation from the naïve expectation of ¾.

This generation time effect acts jointly with myriad other factors that have been highlighted previously (e.g., 48, 50, 54, 67). Notably, the ratio of male to female reproductive variances, $(V_M + 2)/(V_F + 2)$, which tends to be greater than 1 and varies considerably among human populations and closely related species, is likely to have been an important explanatory factor (48, 67). Another likely contributor is the different responses of polymorphism levels on X and autosomes to changes in population size, and to bottlenecks in particular (65). Notably, changes in population size (most notably, the Out-of-Africa bottleneck) have been estimated to contribute 36%-47% of the observed reduction in X to autosome ratios in Asians and 38%-60% in Europeans compared to African populations, where the ranges correspond primarily to the use of different demographic models (54). Importantly, however, these estimates assume that the X to autosomes ratio of effective population sizes in the absence of bottlenecks is ¾, while consideration of sex-specific generation times and reproductive variances point to a ratio closer to 1, and thus suggest a weaker effect of population size changes. Finally, X to autosome ratios have been shown to vary substantially along the genome due to variation in the effects of linked selection, e.g., with an estimated average reduction of 25-35% near compared to far from genes (54). Taken together, these factors are likely to explain most of the variation in ratios observed among human populations and along the genome. In considering variation among hominid species, differences in other factors affecting the male bias of mutation rates, $\alpha = \mu_M/\mu_F$, such as the male age of puberty and the length of the spermatogenetic cycle (11), may have also contributed substantially.

Ultimately, we would like to quantify the contributions of each factor to observed X to autosome polymorphism ratios. At present, however, a number of practical obstacles stand in the way. Most importantly, despite numerous papers characterizing X to autosome ratios, we still lack reliable estimates of their absolute values, and thus of their deviations from the naïve expectation of ¾. The main reason being that current estimates are based on dividing polymorphism levels by divergence to an outgroup (e.g., to macaques (49, 53), chimpanzees (48, 49), orangutans (48, 49), gorillas (49), marmosets (49), olive baboons (80) or, when non-human species are considered, to humans (80)), in an attempt to control for differences in mutation rates on X and autosomes. The "normalized" ratios are assumed to reflect only genealogical differences between X and autosomes, but this assumption is wrong. Notably, changes to life history traits should lead to changes in the X to autosome ratio of divergence over evolutionary time (4), and indeed this ratio has been found to vary substantially with the choice of outgroup (9, 13).

Our results make the problematic nature of this normalization procedure quite clear, by allowing us to write down an explicit form for the normalized ratio:

$$\frac{\pi_X/K_X}{\pi_A/K_A} = \left[\frac{f(\mu_M/\mu_F)}{f(\mu_M^*/\mu_F^*)} \cdot f(G_M^*/G_F^*)\right] \cdot \frac{3}{4}\frac{f(G_M/G_F)}{f((2+V_M)/(2+V_F))}, \qquad [2.21]$$

where parameters corresponding to the outgroup lineage are denoted by "*". The second term (on the right-hand side) in this expression reflects the genealogical component of the X to autosome ratio, and will henceforth be referred to as "the genealogical ratio", whereas the first (in brackets) includes the mutational effect on the ratio and the terms introduced by the normalization. For the normalization to fulfill its purpose of canceling out the mutational effects, the first term must equal 1. As noted, however, the male bias in mutation rates ($\alpha = \mu_M/\mu_F$) has evolved substantially over phylogenetic time scales (9, 23), and therefore the mutational terms $f(\mu_M/\mu_F)/f(\mu_M^*/\mu_F^*)$ do not cancel out. Even if they did, the different dependencies of polymorphism and divergence ratios on

the ratio of male to female generation times introduces the additional term $f(G_M^*/G_F^*)$ into the "normalized" polymorphism ratios. If we assume that generation times in males are greater than in females, as observed in extant hunter-gatherers (5), gorillas and some chimpanzee groups (6), this term would introduce a downward bias (e.g., up to 6%, for estimated generation times in different hunter-gatherer groups). The combined effect of male mutation bias and generation times on the normalization in hominoids, i.e., the term $f(G_M^*/G_F^*)/f(\mu_M^*/\mu_F^*)$, is dominated by the maternal generation time G_F^* (4). Varying G_F^* along the lineage to the outgroup from 18 to 30, while fixing other parameters to human values, results in a 10% increase in the normalized ratio. Taken together, these factors make existing, "normalized" estimates of the X to autosome ratio of polymorphism levels (e.g., (54)) difficult to interpret. Specifically, they cannot be interpreted as estimates of the genealogical X to autosome ratios, and are thus not comparable with the naïve expectation of ¾.

These difficulties suggest that it may be preferable to consider X to autosome polymorphism ratios without normalization, and find alternative means to account for differences in the mutability of X and autosomes. We can begin by separating polymorphism ratios by types of mutation and genomic contexts, e.g., look at the ratio of C to T mutations in an ancestral ACA context (38). If we knew the sex specific mutation rates corresponding to these types and contexts during the genealogical history of extant samples, then we could divide out the mutational effects on these polymorphism ratios (i.e., $f(\mu_M/\mu_F)$) to obtain estimates of the genealogical X to autosome ratio. In principle, we can learn about these mutation rates from the extant mutational spectrum, and specifically from the dependency of the rates of different types of mutations on sex and life history parameters (see below). While we currently lack a sufficiently accurate characterization of the sex and age specific mutational spectrum to do so, it may very well be within reach in the near future,

at least in humans, owing to the increasing power of pedigree studies of mutation (69). Meanwhile, focusing on specific categories of mutations, the rates of which have been fairly stable over phylogenetic timescales (e.g., CpG transitions and ACA->ATA (38, 90)), may provide initial approximations for the genealogical X to autosome ratio.

One can also apply new approaches to learn about the relative contribution of the different factors that shaped the genealogical ratio. Notably, as different types of mutations have distinct dependencies on life history parameters but X to autosome ratios corresponding to these types share the same genealogical history, one can make inferences about life history parameters (and the corresponding mutational spectrum) by requiring the same genealogical X to autosome ratio for different types of mutations (Gao, Moorjani et al., in preparation). Specifically, this approach may allow one to learn about the effects of the ratio of male to female generation times on the genealogical ratio. In addition, one could imagine using pairwise sequentially Markovian coalescent (PSMC) based inferences about historical effective population sizes on X and autosomes, to learn about the effects of bottlenecks, e.g., the OoA bottleneck, on the genealogical ratio directly rather than through simulations (cf. (54, 91)). Moreover, it may be possible to extend PSMC to use data from X and autosomes jointly in order to infer historical changes in female and male effective population sizes over time, and examine, for example, if they exhibited dramatic changes during the Out of Africa bottleneck (i.e., to examine the hypothesis of (50)).

More generally, all the factors that are thought to affect neutral polymorphism levels on X and autosomes can now be integrated into a coalescent framework, providing a natural way to quantify how their effects combine, and an efficient model from which to simulate. Our results describe how life history factors affect the effective population sizes on X and autosomes in the Kingman coalescent. The integration of other factors then follows from their known descriptions in terms of

the coalescent, starting from these effective population sizes. For example, the effects of changes in population size and generation times can thus be integrated as appropriate changes to the effective population sizes on X and autosomes backwards in time. Moreover, the effects of linked selection on relative diversity levels along the genome also follow from their description in terms of the coalescent (e.g., (92-94)), starting from these effective population sizes. Lastly, the effects of life history on mutation can be incorporated in terms of mutational models akin to our hominid model. With the theoretical framework in place, and the data needed to infer the effects of each factor either already present or around the corner, a comprehensive, quantitative understanding of diversity levels on X and autosomes in hominoids and other taxa should be within reach.

References

1. Kimura M (1983) *The neutral theory of molecular evolution* (Cambridge Univ. Press, London).
2. Kong A*, et al.* (2012) Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488(7412):471-475.
3. Scally A & Durbin R (2012) Revising the human mutation rate: implications for understanding human evolution. *Nat Rev Genet* 13(10):745-753.
4. 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(6):1588-1593.
5. 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(2):415-423.
6. Langergraber KE*, et al.* (2012) Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc Natl Acad Sci U S A* 109(39):15716-15721.
7. Graves RR, Lupo AC, McCarthy RC, Wescott DJ, & Cunningham DL (2010) Just how strapping was KNM-WT 15000? *Journal of Human Evolution* 59(5):542-554.
8. Patterson N, Richter DJ, Gnerre S, Lander ES, & Reich D (2006) Genetic evidence for complex speciation of humans and chimpanzees. *Nature* 441(7097):1103-1108.
9. Presgraves DC & Yi SV (2009) Doubts about complex speciation between humans and chimpanzees. *Trends in Ecology & Evolution* 24(10):533-540.
10. Elango N, Thomas JW, Program NCS, & Yi SV (2006) Variable molecular clocks in hominoids. *Proc Natl Acad Sci U S A* 103(5):1370-1375.
11. Segurel L, Wyman MJ, & Przeworski M (2014) Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet* 15:47-70.
12. 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(6):1741-1745.
13. Sayres MAW, 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(10):2800-2815.
14. Haldane J (1946) The mutation rate of the gene for haemophilia, and its segregation ratios in males and females. *Ann Eugen* 13(1):262-271.
15. Hill WG (1972) Effective size of populations with overlapping generations. *Theor Popul Biol* 3(3):278-289.
16. Charlesworth B (1980) *Evolution in age-structured populations* (Cambridge University Press, Cambridge).
17. Amster G & Sella G (2017) Life history effects on neutral polymorphism levels of autosomes and sex chromosomes. *bioRxiv*.
18. Kondrashov AS (2003) Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Human mutation* 21(1):12-27.
19. Nachman MW & Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics* 156(1):297-304.

20. Moorjani P, Amorim CEG, Arndt PF, & Przeworski M (2015) Quantifying variation in the molecular clock among primates. *Manuscript in preperation*.
21. Kim SH, Elango N, Warden C, Vigoda E, & Yi SV (2006) Heterogeneous genomic molecular clocks in primates. *PLoS Genet* 2(10):e163.
22. Ohta T (1993) An examination of the generation-time effect on molecular evolution. *Proc Natl Acad Sci U S A* 90(22):10676-10680.
23. Sayres MAW & Makova KD (2011) Genome analyses substantiate male mutation bias in many species. *Bioessays* 33(12):938-945.
24. Venn O*, et al.* (2014) Strong male bias drives germline mutation in chimpanzees. *Science* 344(6189):1272-1275.
25. Penrose LS (1955) Parental age and mutation. *Lancet* 269(6885):312-313.
26. Crow JF (2000) The origins, patterns and implications of human spontaneous mutation. *Nat Rev Genet* 1(1):40-47.
27. Pudney J (1995) Spermatogenesis in nonmammalian vertebrates. *Microsc Res Tech* 32(6):459-497.
28. Franchi L, Mandl AM, & Zuckerman S (1962) The development of the ovary and the process of oogenesis. *The ovary* 1:1-88.
29. Nieuwkoop PD & Sutasurya LA (1979) *Primordial germ cells in the chordates: embryogenesis and phylogenesis* (Cambridge Univ. Press, London).
30. Gao Z, Wyman MJ, Sella G, & Przeworski M (2015) Interpreting the dependence of mutation rates on age and time. *ArXiv e-prints.* 1507:6890.
31. Drost JB & Lee WR (1995) Biological basis of germline mutation: comparisons of spontaneous germline mutation rates among drosophila, mouse, and human. *Environmental and molecular mutagenesis* 25(S2):48-64.
32. Ehmcke J, Wistuba J, & Schlatt S (2006) Spermatogonial stem cells: questions, models and perspectives. *Hum Reprod Update* 12(3):275-282.
33. Ramm SA & Stockley P (2010) Sperm competition and sperm length influence the rate of mammalian spermatogenesis. *Biol Lett* 6(2):219-221.
34. Martin RD (2007) The evolution of human reproduction: a primatological perspective. *American Journal of Physical Anthropology* 134(S45):59-84.
35. Scally A*, et al.* (2012) Insights into hominid evolution from the gorilla genome sequence. *Nature* 483(7388):169-175.
36. Miyata T, Hayashida H, Kuma K, Mitsuyasu K, & Yasunaga T (1987) Male-driven molecular evolution: a model and nucleotide sequence analysis. *Cold Spring Harbor symposia on quantitative biology*, pp 863-867.
37. Dutheil JY, Munch K, Nam K, Mailund T, & Schierup MH (2015) Strong Selective Sweeps on the X Chromosome in the Human-Chimpanzee Ancestor Explain Its Low Divergence. *PLoS Genet* 11(8): e1005451.
38. 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(39):13994-14001.
39. Forster P*, et al.* (2015) Elevated germline mutation rate in teenage fathers. *Proc Biol Sci* 282(1803):20142898.
40. Scally A (2015) Mutation rates and the evolution of germline structure. *Proc. R. Soc. B, Manuscript in preparation*.

41. Lynch M (2010) Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A* 107(3):961-968.
42. Jensen Seaman MI & Hooper Boyd KA (2013) Molecular Clocks: Determining the Age of the Human–Chimpanzee Divergence. *eLS (John Wiley & Sons: Chichester) doi: 10.1002/9780470015902.a0020813.pub2*.
43. Li W-H, Ellsworth DL, Krushkal J, Chang BH-J, & Hewett-Emmett D (1996) Rates of nucleotide substitution in primates and rodents and the generation–time effect hypothesis. *Molecular phylogenetics and evolution* 5(1):182-187.
44. Webster TH & Wilson Sayres MA (2016) Genomic signatures of sex-biased demography: progress and prospects. *Curr Opin Genet Dev* 41:62-71.
45. Ellegren H (2009) The different levels of genetic diversity in sex chromosomes and autosomes. *Trends Genet* 25(6):278-284.
46. Leffler EM*, et al.* (2012) Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol* 10(9):e1001388.
47. Aquadro CF, Begun DJ, & Kindahl EC (1994) Selection, recombination, and DNA polymorphism in Drosophila. *Non-neutral evolution: Theories and Molecular Data*, ed B G (Springer, Dordrecht), pp 46-56.
48. Hammer MF, Mendez FL, Cox MP, Woerner AE, & Wall JD (2008) Sex-biased evolutionary forces shape genomic patterns of human diversity. *PLoS Genet* 4(9):e1000202.
49. Hammer MF*, et al.* (2010) The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. *Nature Genetics* 42(10):830-831.
50. Keinan A & Reich D (2010) Can a sex-biased human demography account for the reduced effective population size of chromosome X in non-Africans? *Molecular biology and evolution* 27(10):2312-2321.
51. Veeramah KR, Gutenkunst RN, Woerner AE, Watkins JC, & Hammer MF (2014) Evidence for increased levels of positive and negative selection on the X chromosome versus autosomes in humans. *Mol Biol Evol* 31(9):2267-2282.
52. Keinan A, Mullikin JC, Patterson N, & Reich D (2009) Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nat Genet* 41(1):66-70.
53. Gottipati S, Arbiza L, Siepel A, Clark AG, & Keinan A (2011) Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing. *Nat Genet* 43(8):741-743.
54. Arbiza L, Gottipati S, Siepel A, & Keinan A (2014) Contrasting X-linked and autosomal diversity across 14 human populations. *The American Journal of Human Genetics* 94(6):827-844.
55. Prado-Martinez J*, et al.* (2013) Great ape genetic diversity and population history. *Nature* 499(7459):471-475.
56. Nam K*, et al.* (2015) Extreme selective sweeps independently targeted the X chromosomes of the great apes. *Proceedings of the National Academy of Sciences* 112(20):6413-6418.
57. Charlesworth B, Coyne JA, & Barton NH (1987) The Relative Rates of Evolution of Sex-Chromosomes and Autosomes. *American Naturalist* 130(1):113-146.
58. Haldane JBS (1924) A mathematical theory of natural and artificial selection. . *Trans. Camb. Philos. Soc.* Part I.(23):19-41.

59. Charlesworth B (2012) The Effects of Deleterious Mutations on Evolution at Linked Sites. *Genetics* 190(1):5-22.
60. McVicker G, Gordon D, Davis C, & Green P (2009) Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet* 5(5):e1000471.
61. Hernandez RD*, et al.* (2011) Classic Selective Sweeps Were Rare in Recent Human Evolution. *Science* 331(6019):920-924.
62. Hey J & Harris E (1999) Population bottlenecks and patterns of human polymorphism. *Mol Biol Evol* 16(10):1423-1426.
63. Fay JC & Wu CI (1999) A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. *Molecular Biology and Evolution* 16(7):1003-1005.
64. Wall JD, Andolfatto P, & Przeworski M (2002) Testing models of selection and demography in Drosophila simulans. *Genetics* 162(1):203-216.
65. Pool JE & Nielsen R (2007) Population size changes reshape genomic patterns of diversity. *Evolution* 61(12):3001-3006.
66. Bryc K*, et al.* (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci U S A* 107(2):786-791.
67. Charlesworth B (2001) The effect of life-history and mode of inheritance on neutral genetic variability. *Genet Res* 77(2):153-166.
68. Betzig L (2012) Means, variances, and ranges in reproductive success: comparative evidence. *Evolution and Human Behavior* 33(4):309-317.
69. Wong WS*, et al.* (2016) New observations on maternal age effect on germline de novo mutations. *Nat Commun* 7:10486.
70. Felsenstein J (1971) Inbreeding and variance effective numbers in populations with overlapping generations. *Genetics* 68(4):581-597.
71. Johnson DL (1977) Inbreeding in populations with overlapping generations. *Genetics* 87(3):581-591.
72. Orive ME (1993) Effective population size in organisms with complex life-histories. *Theoretical population biology* 44(3):316-340.
73. Sagitov S & Jagers P (2005) The coalescent effective size of age-structured populations. *Annals of Applied Probability* 15(3):1778-1797.
74. Pollak E (2011) Coalescent theory for age-structured random mating populations with two sexes. *Math Biosci* 233(2):126-134.
75. Pettay JE, Kruuk LE, Jokela J, & Lummaa V (2005) Heritability and genetic constraints of life-history trait evolution in preindustrial humans. *Proc Natl Acad Sci U S A* 102(8):2838-2843.
76. Westendorp RG & Kirkwood TB (1998) Human longevity at the cost of reproductive success. *Nature* 396(6713):743.
77. Thomas F, Teriokhin A, Renaud F, De Meeûs T, & Guégan J-F (2000) Human longevity at the cost of reproductive success: evidence from global data. *Journal of Evolutionary Biology* 13(3):409-414.
78. Wright S (1939) Statistical genetics in relation to evolution, Vol. 802. *Hermann et Cie.: Paris*.
79. Hudson RR (1990) Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology* 7(1):44.

80. Evans BJ, Zeng K, Esselstyn JA, Charlesworth B, & Melnick DJ (2014) Reduced representation genome sequencing suggests low diversity on the sex chromosomes of tonkean macaque monkeys. *Mol Biol Evol* 31(9):2425-2440.
81. Franchi L, Mandl AM, & Zuckerman S (1962) The development of the ovary and the process of oogenesis. *The Ovary*, ed Zuckerman S (Academic Press, London), Vol 1, pp 1-88.
82. Charlesworth B & Charlesworth D (2010) *Elements of evolutionary genetics* (Roberts And Company Publishers, Greenwood Village, Colorado).
83. Nielsen CT*, et al.* (1986) Onset of the Release of Spermatozia (Supermarche) in Boys in Relation to Age, Testicular Growth, Pubic Hair, and Height. *J Clin Endocrinol Metab* 62(3):532-535.
84. Marson J, Meuris S, Cooper R, & Jouannet P (1991) Puberty in the male chimpanzee: progressive maturation of semen characteristics. *Biology of reproduction* 44(3):448-455.
85. Heller CG & Clermont Y (1963) Spermatogenesis in man: an estimate of its duration. *Science* 140(3563):184-186.
86. Smithwick EB, Young LG, & Gould KG (1996) Duration of spermatogenesis and relative frequency of each stage in the seminiferous epithelial cycle of the chimpanzee. *Tissue & Cell* 28(3):357-366.
87. Dixson AF (2009) *Sexual selection and the origins of human mating systems* (Oxford Univ, Press, Oxford).
88. Barr AB (1973) Timing of spermatogenesis in four nonhuman primate species. *Fertil Steril* 24(5):381-389.
89. Clermont Y & Antar M (1973) Duration of the cycle of the seminiferous epithelium and the spermatogonial renewal in the monkey Macaca arctoides. *American Journal of Anatomy* 136(2):153-165.
90. Moorjani P, Amorim CEG, Arndt PF, & Przeworski M (2016) Variation in the molecular clock of primates. *Proceedings of the National Academy of Sciences* 113(38):10607-10612.
91. Li H & Durbin R (2011) Inference of human population history from individual whole-genome sequences. *Nature* 475(7357):493-496.
92. Hudson RR & Kaplan NL (1995) Deleterious background selection with recombination. *Genetics* 141(4):1605-1617.
93. Nordborg M, Charlesworth B, & Charlesworth D (1996) The effect of recombination on background selection. *Genet Res* 67(2):159-174.
94. Kaplan NL, Hudson RR, & Langley CH (1989) The Hitchhiking Effect Revisited. *Genetics* 123(4):887-899.

Appendix 1 – Supplementary Information for Chapter 1

A1.1. The expected substitution rate with sex and age dependent mutation rates

If mutation rates depend on age and sex then the expected rate of substitutions will depend on the distributions of male and female reproductive ages in a population. Here we provide an intuition for as well as formal derivations of the rate of substitutions under general assumptions about these distributions. We show that, assuming the linear mutational model described in chapter 1, the substitution rate depends on the average reproduction ages in males and females, but is unaffected by the variance of these ages in the population. We also consider the rate of substitutions for the case in which these distributions and the mutation rates per cell division change over time.

Intuition. We begin with an informal argument for the expected rate of substitutions on autosomes (also cf. Charlesworth (1), *pp.* 92-93). We denote the expected mutation rates per-generation in males and females by μ_M and μ_F and the expected generation times by G_M and G_F, respectively. The yearly substitution rates in males and females are therefore μ_M/G_M and μ_F/G_F, respectively. The expected yearly substitution rate on autosomes can be calculated as a weighted average over yearly rates in males and females, where the weights are the proportions of time an autosomal locus spends in each sex: i.e., $G_M/(G_M + G_F)$ for males and $G_F/(G_M + G_F)$ for females. The expected yearly rate on autosomes is therefore $(\mu_M + \mu_F)/(G_M + G_F)$. Equivalently, consider a branch of X generations. The expected number of substitutions along the branch at an autosomal locus is $\frac{1}{2}(\mu_M + \mu_F)X$, whereas the expected length of the branch in years is $\frac{1}{2}(G_M + G_F)X$. The expected yearly mutation rate autosomal is therefore $(\mu_M + \mu_F)/(G_M + G_F)$.

Assumptions and notation. We consider the genealogical process along a lineage backward in time. We denote the distributions of paternal and maternal ages at offspring birth by A_M and A_F and their expectations by G_M and G_F, respectively. For brevity, we consider these distributions to be discrete (i.e., in years). The number of mutations that a parent of sex *s* and age *a* bequeaths to its newborn, $X_{s,a}$, is a random variable with expectation $\mu_{s,a}$ (averaging over any other factors that could affect mutation rates, e.g., age of puberty). We later show that the variance (or higher moments) of $X_{s,a}$ do not affect the expected substitution rates, and therefore we require no further assumptions about the distribution of $X_{s,a}$. Since mutation rates can vary with age, the expected mutation rate per-generation depends on the distribution of reproduction ages. We define the paternal and maternal mutation rates per-generation as the appropriate weighted averages, $\mu_M = E_{A_M}(\mu_{M,a})$ and $\mu_F = E_{A_F}(\mu_{F,a})$. Thus, if we sample a random newborn, μ_M and μ_F are the expected numbers of mutations it inherits from his father and mother. Note that for a linear mutational model (i.e., $\mu_{s,a} = C_s + D_s a$ where C_s and D_s are constants), μ_M and μ_F depend only on the expected (but not the variance of) ages of reproduction (i.e., $\mu_s = E_{A_s}(\mu_{s,a}) = C_s + D_s G_s$). We first consider the case in which all distributions remain constant in time.

Autosomal substitution rate. The number of substitutions K_A over *T* years is affected by the stochasticity in both the genealogical history on the lineage and the mutational process. First we consider the genealogical history. Tracking an autosomal site backwards in time, at generation *i* we have a parent of sex s_i and age a_i, where s_i is either male or female with probability ½ and $(a_i|s_i) \sim A_{s_i}$. The number of de-novo substitutions at generation *i*, y_i, is a random variable with conditional expectation μ_{s_i,a_i}. We assume that the variables $h_i = (a_i, s_i)$ are independent between

generations (the derivation can be readily generalized so long as the autocorrelation time between ages in subsequent generations is short relative to the lineage length). It follows that the number of generations in a lineage of *T* years is a random variable defined as

$$n_G(T) = max\{n | \sum_{i=1}^{n} a_i \leq T\}$$

and that

$$K_A = \sum_{i=1}^{n_G(T)} y_i.$$

K_A is a renewal-reward process (with time *T*), so

$$E(K_A) = \frac{T}{E(a_i)} E(y_i) + O(1)$$

(2). The expected autosomal generation time is $G_A = E(a_i) = \frac{1}{2}(G_M + G_F)$ and the expected autosomal mutation rate per-generation is $\mu_A = E(y_i) = E\left(\mu_{s_i,a_i}\right) = \frac{1}{2}(\mu_M + \mu_F)$. The $O(1)$ term is negligible as long as $T >> G_A$ (and is at most μ_{Aut}), so

$$E(K_A) = (\mu_A / G_A)T. \qquad \text{[A1.1]}$$

Substitution rates on the X. The derivation for substitution rates on the X requires a small adjustment, because in this case the sexes in consecutive generations are dependent variables (as males necessarily inherit their X from the maternal side). We therefore take each step backwards in time to extend back to the next female ancestor instead of the next generation. We denote the length of the i^{th} step in years by l_i and the number of de-novo substitutions in that step by y_i. If the X-linked allele is directly inherited from a female (with probability ½), then $l_i \sim A_F$ and $E(y_i | F, l_i) = \mu_{F,a_i}$. Otherwise, the X is inherited from a male, and only then from a female,

yielding $l_i = a_f + a_m$ and $E(y_i|M, a_f, a_m) = \mu_{F,a_f} + \mu_{M,a_m}$ where $a_f \sim A_F$ and $a_m \sim A_M$. The average length of a step (in years) is $E(l_i) = G_F + \frac{1}{2}G_M = \frac{3}{2}G_X$ (where $G_X = \frac{1}{3}G_M + \frac{2}{3}G_F$ is the sex-averaged generation time) and the average number of de-novo substitutions in a step is $E(y_i) = \mu_F + \frac{1}{2}\mu_M = \frac{3}{2}\mu_X$ (where $\mu_X = \frac{1}{3}\mu_M + \frac{2}{3}\mu_F$ is the sex-averaged mutation rate per-generation on the X).

The steps $h_i = (\mathrm{l_i}, \mathrm{y_i})$ are i.i.d and K_X is a renewal-reward process (with time T) defined by $n_G(T) = max\{n|\sum_{i=1}^{n} a_i \leq T\}$ and $K_X = \sum_{i=1}^{G(T)} y_i$. It follows that

$$E(K_X) = \frac{T}{E(l_i)}E(y_i) + O(1) = \frac{\mu_X}{G_X}T + O(1),$$

Again, neglecting the $O(1)$ term we attain

$$E(K_X) = \frac{\mu_X}{G_X}T. \qquad \text{[A1.2]}$$

Substitution rates on the Y. By the same token,

$$E(K_Y) = \frac{\mu_M}{G_M}T. \qquad \text{[A1.3]}$$

Relative substitution rates on sex chromosomes and autosomes. From Eqs. A1.1- A1.3, we attain

$$\frac{E(K_X)}{E(K_A)} = \frac{f(\mu_M/\mu_F)}{f(G_M/G_F)}, \quad \frac{E(K_Y)}{E(K_A)} = \frac{h(\mu_M/\mu_F)}{h(G_M/G_F)} \text{ and } \frac{E(K_Y)}{E(K_X)} = \frac{q(\mu_M/\mu_F)}{q(G_M/G_F)}, \qquad \text{[A1.4]}$$

where $f(x) = \frac{2x+4}{3x+3}$, $h(x) = \frac{2x}{x+1}$ and $q(x) = \frac{3x}{x+2}$. In the case where both sexes have the same expected generation times, Eqs. A1.4 reduce to Miyata's formulae (3). Note that Eqs. A1.4 imply

that the X-to-autosome ratio is bound between ½ and 2 whereas the Y-to-autosome and Y-to-X ratios do not have a theoretical bound.

Corresponding results for the ZW system. By the same token,

$$E(K_A) = (\mu_A/G_A)T \ , E(K_Z) = \frac{\mu_Z}{G_Z}T \text{ and } E(K_W) = \frac{\mu_F}{G_F}T, \qquad \text{[A1.5]}$$

where $G_Z = \frac{2}{3}G_M + \frac{1}{3}G_F$ is the sex-averaged generation time, $\mu_Z = \frac{2}{3}\mu_M + \frac{1}{3}\mu_F$ is the sex-averaged mutation rate per-generation on the Z and the other parameters are defined as in the XY system. It follows that

$$\frac{E(K_Z)}{E(K_A)} = \frac{f^*(\mu_M/\mu_F)}{f^*(G_M/G_F)}, \ \frac{E(K_W)}{E(K_A)} = \frac{h^*(\mu_M/\mu_F)}{h^*(G_M/G_F)} \text{ and } \frac{E(K_W)}{E(K_Z)} = \frac{q^*(\mu_M/\mu_F)}{q^*(G_M/G_F)}, \qquad \text{[A1.6]}$$

where $f^*(x) = \frac{4x+2}{3x+3}$, $h^*(x) = \frac{2}{x+1}$ and $q^*(x) = \frac{3}{2x+1}$. In the case where both sexes have the same expected generation times, Eqs. A1.6 reduce to Miyata's formulae for the ZW system (3). Note that Eqs. A1.6 imply that the Z-to-autosome ratio is bound between ½ and 2 whereas the W-to-autosome and W-to-Z ratios do not have a theoretical bound.

Changing substitution rates. The sex-specific distributions of reproduction age and the dependency of mutation rates on age and sex likely change over phylogenetic timescales. To derive the substitution rates in this case, we assume that the distributions of reproduction ages $A_M(t)$ and $A_F(t)$ and the expected mutation rates $\mu_{s,a}(t)$ are time-dependent. Under standard smoothness requirements on these functions, we can divide the lineage into a large number of segments with almost constant parameters, where in the limit

$$K_A = \int_0^T \frac{\mu_A(t)}{G_A(t)} dt \text{ , } K_X = \int_0^T \frac{\mu_X(t)}{G_X(t)} dt \text{ and } K_Y = \int_0^T \frac{\mu_M(t)}{G_M(t)} dt. \quad \text{[A1.7]}$$

Here, $\mu_A(t)$, $G_A(t)$, etc. denote the expectations at time *t*.

A1.2. Linear mutation model: averaging

Our linear model relates mutation rates with life history and spermatogenetic parameters. Yet each of these parameters can vary in the population and along a lineage. Here we consider how to average over these parameters to obtain the expected rate of substitutions.

A1.2.1. Averaging over the population

In the previous section we have shown that the dependency of mutation rates on age and sex affects the substitution rates through the expectations $\mu_M = E_M(\mu_{M,a})$ and $\mu_F = E_F(\mu_{F,a})$ (where the expectation is over the sex-specific distributions of ages at offspring birth, A_M and A_F; see Section A1.1). According to our model, the mutation rate per generation in females is constant, but the rate in males depends on age of reproduction, age of puberty and the spermatogenic cycle length, all of which vary in the population. For a specific father *f*, we assume that the expected number of de-novo mutations is

$$\mu_{M,a_B}(a_P, I, l) = C_M + D_M \frac{(a_B - a_P - I)}{l},$$

where age at puberty and at offspring birth, a_P and a_B, and the spermatogenic cycle length l are values specific to an individual. We assume that μ_F, C_M, D_M and the gestation time I are constant in the population, and first consider a simplified scenario in which l is constant in all germ cells within an individual.

It follows from the derivations in the previous section that the average per-generation male mutation rate in the population is

$$\mu_M = C_M + D_M E(\frac{a_B - a_P - I}{l}),$$

where the expectation is over the population of fathers to newborns and each male is weighted by the number of children he has and the ages at which he has them, thereby excluding individuals that never become parents (see definition of μ_M in the previous section). If we assume that the spermatogenic cycle length is independent from both age of puberty and age of reproduction, it follows that

$$\mu_M = C_M + D_M E(\frac{1}{l})(G_M - E(a_P)).$$

Hence, Eq. 1.4 holds if $P = E(a_P)$, the average age of puberty in fathers, and $\tau = 1/E(\frac{1}{l})$, the harmonic average of the spermatogenic cycle length in fathers.

These averages differ from those reported in the literature in several ways. Notably, they are taken over the population of fathers as opposed to the population average. For example, if fathers with earlier puberty tend to have more offspring then P would be lower than the population average. A similar argument might apply to τ (e.g., if males with shorter spermatogenic cycles tend to produce more offspring). In addition, the definition of τ as a harmonic mean differs from the simple averages reported in the literature.

If the length of the spermatogenic cycle varies among germ cell lineages or changes with age within an individual then an individual's average length *l* should account for the variability. Similar considerations to those that applied to the population suggest that one should use a harmonic

average over germ cell lineages and changes with age in these lineages. At present, the data required for correctly averaging over individuals, let alone within individuals is not available.

A1.2.2. Averaging over a lineage

Next we derive the appropriate averages when life history and spermatogenesis parameters vary along lineages. To this end, we substitute the linear mutation model into Eq. A1.7

$$K_A = \int_0^T \frac{\mu_A(t)}{G_A(t)}dt = \frac{C_M+\mu_F}{2}\int_0^T \frac{1}{G_A(t)}dt + D_M \int_0^T \frac{1}{\tau(t)} \cdot \frac{\frac{G_M(t)}{G_F(t)}}{1+\frac{G_M(t)}{G_F(t)}}[1-\frac{P(t)}{G_M(t)}]dt,$$

where the parameter values at a given time reflect the population averages we considered above. It follows that the sex averaged generation time along a lineage should be defined as the harmonic mean $\overline{G_A} = \left[T^{-1}\int_0^T \frac{1}{G_A(t)}\right]^{-1}$. Assuming that $\frac{G_M(t)}{G_F(t)}$, $\frac{P(t)}{G_M(t)}$ and $\tau(t)$ evolve independently, it also follows that the average ratio of male puberty age to generation time is $T^{-1}\int_0^T \frac{P(t)}{G_M(t)}$ and the average spermatogenic cycle length is given by the harmonic mean $\bar{\tau} = \left[T^{-1}\int_0^T \frac{1}{\tau(t)}\right]^{-1}$. Assuming independence also implies that the ratio of male-to-female generation times affects substitution rates through the term $T^{-1}\int_0^T \frac{\frac{G_M(t)}{G_F(t)}}{1+\frac{G_M(t)}{G_F(t)}}dt$. Using the average male-to-female ratio of generation times $r = T^{-1}\int_0^T \frac{G_M(t)}{G_F(t)}$, this term can be approximated by $\frac{r}{1+r}$, where the accuracy of the approximation satisfies

$$\left| T^{-1}\int_0^T \frac{\frac{G_M(t)}{G_F(t)}}{1+\frac{G_M(t)}{G_F(t)}}dt - \frac{T^{-1}\int_0^T \frac{G_M(t)}{G_F(t)}}{1+T^{-1}\int_0^T \frac{G_M(t)}{G_F(t)}} \right| \leq \frac{V}{(r+1)^2},$$

and therefore depends on the variability of the ratio over time $V = T^{-1}\int_0^T \left[\frac{G_M(t)}{G_F(t)} - r\right]^2 dt$.

A1.3. Linear mutation model: dependency on parameters

Here we consider whether changing one parameter while leaving the others constant increases or decreases the rate of substitutions on autosomes. It is apparent from Eqs. 1.4 and A1.1 that the autosomal substitution rate monotonically decreases with both the male age of puberty (P) and the spermatogenic cycle length (τ).

In order to consider the effects of the average and ratio of generation times, G_A and G_M/G_F, we reorganize Eqs. 1.4 and A1.1 to obtain the yearly substitution rate $K_{year} = K_A/T$,

$$E(K_{year}) = \frac{C_M + C_F - D_M(P+I)/\tau}{2G_A} + \frac{D_M}{\tau} \cdot \frac{G_M/G_F}{1+G_M/G_F}. \qquad [A1.8]$$

In this form it is apparent that the substitution rate increases monotonically with the ratio of male-to-female generation times (G_M/G_F); this would be the case for any linear mutation model in which mutation rates increase more rapidly with paternal than maternal age (which follows from $\frac{D_M}{\tau} > 0$ in our case) and monotonically decreases otherwise.

While the generation time effect in mammals would suggest that substitution rates increase when the average generation (G_A) decreases, Segurel et al. (4) and Gao et al. (5) used a model similar to ours to show that this is not necessarily the case. Specifically, they show that rates would decrease only if $C_M + C_F > D_M(P+I)/\tau$ (using our notation) and increase otherwise. This derivation, however, is predicated on the assumption that puberty age and gestation time remain constant when the generation time changes.

If, instead, we assume that the proportion of a generation before puberty and in gestation, $(P + I)/G_M$, remains constant, then substitution rates necessarily increase when the average generation (G_A) decreases. This becomes apparent when we reorganize Eq. A1.8 as follows

$$E(K_{year}) = \frac{C_M+C_F}{2G_A} + \frac{D_M}{\tau}\left(1 - \frac{P}{G_M} - \frac{I}{G_M}\right)\frac{G_M/G_F}{1+G_M/G_F}.$$

Using this parameterization does not affect our conclusions with respect to the effects of parameters other than G_A and P.

A1.4. Linear mutation model: parameter estimates

A1.4.1. Inference from human pedigrees

We infer the parameters associated with rates of mutation per germ-cell division from the Kong et al. human pedigree study (6). The Kong et al. study is based on 78 trios and is one of the largest published to date (6). We choose not to combine data across multiple studies to circumvent the problems that arise from the use of different methodologies, e.g., for controlling for false positives and negatives ((4); but see (5)).

We infer the parameters μ_F, C_M and D_M. To that end, we rely on the Kong et al. fit of a linear model to the relationship between mutation rates and paternal ages. Further assuming an average male puberty age of 13.4 years (cf. section A1.4; the puberty age in Iceland aligns with the age in other western European populations (7)) yields an estimate of

$$\frac{(C_M+\mu_F)}{2} = 5.77 \times 10^{-9} \text{ per bp,} \qquad \text{[A1.9]}$$

corresponding to the mutation rate averaged over females and pre-pubertal males, and an average mutation rate of $1.245 \cdot 10^{-8}$ per bp in males during spermatogenesis (taken together these

estimates correspond to a sex averaged mutation rate of $1.2 \cdot 10^{-8}$ per bp per generation). Given that the average age of reproduction in the sample is 29.7 years, corresponding to 374 stem cell divisions (cf. Table A1.1), the per-division rate in spermatogenesis is thus

$$D_M = 3.33 \times 10^{-11} \text{ per bp per division.} \quad \text{[A1.10]}$$

We rely on the smaller phased sample of five trios to separate μ_F and C_M, yielding

$$\mu_F = 5.42 \times 10^{-9} \text{ per bp and } C_M = 6.13 \times 10^{-9} \text{ per bp.} \quad \text{[A1.11]}$$

Since autosomal mutation rates depend on $(\mu_F + C_M)$ but are invariant to the separation between the male and female component, this implies that our inferences are more accurate on autosomes than on the X.

A1.4.2. Mutation rates per cell division in humans and other taxa

We rely on estimates of the number of germ cell divisions at different developmental stages (Table A1.1) to translate the pedigree based estimates into estimates of the mutation rate per cell division (Table A1.2). This inference reveals that the per-division rates in females and pre-pubertal males are similar but the rates during spermatogenesis are five-folds lower (and 2-20 times lower than estimates in other species; see Table A1.2 and (5)). The lower rates during spermatogenesis could reflect real differences in replication fidelity or the accumulation of other kinds of mutations (cf. (4, 5)) but they may also reflect errors in estimates of the numbers of cell divisions during spermatogenesis, e.g., associated with the possibility of stem cell turnover in testis (see chapter 1).

	Human	Mouse
Females	31[8] / 33[9]	25[8] /27[9]
Males until puberty	34[8] / 40[9]	27[8] / 29[9]
Spermatogenesis (stem cell divisions)	1 every 16 days from puberty until 74 days before reproduction[8]	1 every 8.6 days from puberty (~6 days) until 43 days before reproduction[8]
Spermatogenesis (non-stem cell divisions)	7[8] / 5[9]	9[8] / 10[9]

Table A1.1. Number of germ cell divisions at different developmental stages in humans and mice. Estimates from Drost and Lee (8) are used throughout the paper.

		Per-division mutation rate ($\times 10^{-10}$ per-base)
H. sapiens[a]	Females	1.75
	Pre-pubertal males	1.80
	Spermatogenesis	0.33
	Average	0.54
M. musculus[b]	Females	0.78
	Males	1.49
	Average	1.25
D. melanogaster[10]		1.3
A. thaliana[10]		1.6
C. elegans[10]		6.5
S. cerevisiae[10]		3.3
E. coli[10]		2.6

Table A1.2. Comparison of estimates of per-division mutation rates in different taxa. Average per-division rates at different developmental stages are calculated assuming a paternal generation time of 30 years in humans and 6 months in mice. a) Using data from Kong et al. (6) (cf. Eqs. A1.10-A1.11 and division rates from Table A1.1). b) Using data from Adewoye et al. (11) assuming an

average paternal reproduction age of 2 months in their control sample, an average per-generation rate of $3.75 \cdot 10^{-9}$ per bp, a male bias of $\alpha = 20/7$ and dividing by the estimated numbers of germ cell divisions (Table A1.1).

A1.4.3. Life history and spermatogenesis parameters in extant catarrhines and mammals

We use the variation in life history and spermatogenesis parameters (i.e., G_M, G_F, P and τ) among extant populations and species to guide our choice of plausible ranges for these parameters over different phylogenetic timescales. An extended list of parameter estimates based on our survey of the literature, including life history, physiological and mutational studies, is provided in Table A1.3. As we note in chapter 1, there is considerable uncertainty about many of these estimates, for reasons including limited data availability and differences in methodology and operational definitions among studies. We provide a brief account of the issues regarding each parameter (excluding gestation time and rates of spermatogenesis for which the estimates are satisfactory for our purposes), as well as additional details when those appear relevant. We also consider modified estimates of the generation time in humans (cf. discussion regarding generation times).

	Generation times (years)	Puberty age in males (years)	Gestation time (days)	Spermatogenic cycle length (days)
Human	Hunter-gatherer societies: average generation times between 24.5-33 years (mean 30.3) and male-to-female ratios between 1.03-1.37 (mean 1.26) [a] (cf. Tables A1.7, A1.8). Agrarian and industrialized societies: average generation times between 29-32.4 years and male-to-female ratios between 1.11-1.19[b] (cf. Table A1.5).	13.4 (S) [(12)], 11.6 (TG) [(13)].	280[c]	16[(14)]
Chimpanzee	In wild communities, observed average generation times between 22.5-28.9 years (mean 24.6) and male-to-female ratios between 0.79-1.22 (mean 0.96)[d] (cf. Table A1.6).	7.5 (S)[(15)], 6.7 (TG) [(15)], 8 (H)[(16, 17)]	229[c]	14[(18)]
Bonobo	Maternal generation time is likely shorter than in Chimpanzees: Puberty is 3 years earlier in female Bonobos (16) and captive female Bonobos give first birth earlier than captive female Chimpanzees (19).	Same age as the male chimpanzee (H)[(16)].	232[c]	-
Gorilla	Two wild mountain gorilla communities: average generation time of 19.3 years, male-to-female ratio of 1.12 (20).	6-8[(21)], 6-8(H)[(22)], 7-10(H)[(23)]	256 [c]	-
Orangutan	Likely higher than in chimpanzees: in Sumatran orangutans age at first	6.4\7.7 (S)[(25)],	249 [c]	-

	reproduction and interbirth interval are higher than in wild chimpanzees whereas mortality rates are lower (24).	6-7[21], 5-7(H) [22], 5-9(TD)[26]		
Gibbon	Life history resembles great apes more than cercopithecoids of similar weight; mean age at first reproduction is 11 years, comparable to Gorillas (27).	Less than 6 years[21]	Between 207 and 243 days in different species[c].	-
Old World Monkeys	Estimated between 10-12 years in most species (28). Maternal generation time of 12.6 years in wild *M. fascicularis*[e]. Maternal generation time of 11.8 years in wild Amboseli *P. cynocephalus*[f].	Between 3.5 and 6 years in different species[21] .	Between 132 and 211 days in 38 species for which data is available [c].	*M. mulatta* 9.5[29] \ 10.5[30], *M. fascicularis* 10.16[31], *C. aethiops* 10.2[21], *P. cynocephalus* 10.2[21], *P. anubis* 11[32], *M. arctoides* 11.6[33].

Table A1.3. Life history and spermatogenesis parameters in extant catarrhines. We use the following notation for the different methodologies used to infer male puberty ages: 'S' – spermarche, 'TG' - onset of testicular growth, 'H' - hormonal changes and 'TD' - testis descent. References: a) Matsumura and Forster (34) and revised estimates from Fenner (35) (cf. Tables A1.7, A1.8 and section A1.4). b) Table A1.5 and references therein. c) *AnAge: The animal ageing and longevity database*, build 13 (36). d) Table A1.6 and references therein; mean values from Langergraber et al. (20). e) Approximated based on reported survival and birth rates (37). f) Approximated based on reported birth (38) and survival (39) rates.

	Ratio of male-to-female generation times (G_M/G_F)	Ratio of gestation time to paternal generation time (I/G_M)	Ratio of male age of puberty to paternal generation time (P/G_M)	Spermatogenic cycle length (τ)
Catarrhines	0.75 – 1.25	2.5% - 6%	25% - 40%	9.5 – 16 days
Mammals	0.7 – 1.3	0% - 6%	0% - 40%	6.7 – 17 days

Table A1.4. Spermatogenic cycle lengths and co-variation of life-history parameters in catarrhines and mammals. Reported ranges correspond to observed values in different species or populations, and reflect our assumptions in Figs. 6 and 5S. Life history parameter ranges in catarrhines are summarized from table A1.3. Life history values in mammals roughly correspond the majority of species ((40); Fig. A1.1). Spermatogenic cycle lengths correspond to 48 surveyed mammals ((41); Fig. A1.2 and Table A1.3).

Male age of puberty and spermatogenic cycle length. Published estimates for male puberty ages are based on a variety of criteria (specified in Table A1.1), and even studies that apply the same criterion often vary in methodology. The criteria include the time of testicular descent or enlargement, depending on whether testis are scrotal or inguinal before the onset of puberty, changes in hormonal levels and the presence of spermatozoa in ejaculates or urine samples. As the onset of puberty and its manifestations are gradual, even studies using the same criterion often apply different thresholds (42). For studies based on enlargement, we therefore report the age at the onset of growth (when it is available). In addition, most available studies in primates, excluding humans, are limited to individuals in captivity. Since spermarche occurs shortly after the onset of stem cell divisions (74 days in humans (8)), age of spermarche is the most relevant criterion for the purposes of our study (assuming that the onset of spermatogenesis is discrete). To the best of

our knowledge, however, reliable estimates based on the age of spermarche exist only for humans and chimpanzees (excluding an estimate in orangutan based on a sample of two individuals; (25)). If the onset of spermatogenesis is gradual, then current measurements (based on first occurrence of sperm in urine or ejaculate samples) underestimate the average age in which a germ cell starts dividing (the most appropriate criteria for our model). The observed increase in daily sperm production over several years until reaching adult levels in humans supports the notion of a gradual onset, although other explanations are also possible (43, 44).

Generation times. To the best of our knowledge, estimates of sex-specific generation times exist only for hominines. Moreover, while there are indications of considerable variation among populations within species, this variation is often poorly quantified. In hunter-gatherers (Tables A1.7 and A1.8), there appear to be substantial differences between societies but confidence intervals for individual societies have not been reported impeding our ability to assess their statistical significance. Estimates for agrarian and industrialized societies rely on considerably more data and are generally better (Table A1.5), but they are probably less relevant for our purposes. In chimpanzees (Table A1.6), estimates by Langergraber et al. (20) show statistically significant variation between communities, with male-to-female generation times ratios ranging from 0.79-1.22 (20). In gorillas, estimates are available for only two communities (20).

	Maternal generation time (years)	Paternal generation time (years)	Average generation time (years)	Ratio of male-to-female generation times
Icelanders, 1698-1742[45]	28.72	31.93	30.3	1.11
Icelanders, 1848-1892[45]	28.12	31.13	29.6	1.11
Icelanders, 274 patrilines[46]	-	34	-	-
Developed nations average[35]	27.3	30.8	29.1	1.13
Less-developed nations average[35]	28.3	31.8	30.1	1.12
French Canadians, late 16th century – present[35]	28.9 / 30.3	34.5 / 34.4	31.7 / 32.4	1.19 / 1.14

Table A1.5. Generation times in agrarian and industrialized societies.

	Maternal generation time (years)	Paternal generation time (years)	Average generation time (years)	Ratio of male-to-female generation times
Western chimpanzee (Taï, 3 sites) [20]	23-31.7	23.1-26.1	23-28.9	0.82-1
Western chimpanzee (Taï) [34]	19	-	-	-
Eastern chimpanzee (5 communities) [20]	23.3-26.1	19.9-28.4	22.5-26.4	0.79-1.22
Eastern chimpanzee (Mahale) [34]	24	-	-	-
Eastern chimpanzee (Gombe) [47]	19.6	-	-	-

Table A1.6. Generation times in wild chimpanzee populations. Ranges from Langergraber et al. (20) are for mean values in different communities.

Generation times in hunter-gatherers. We consider estimates of generation times in hunter-gatherers in greater detail, because they play an important role in our analyses; moreover, similar estimates would likely play an important role in future work. Fenner's meta-analysis (35) relies on mean maternal ages at first and last birth, a_f and a_l, in 11 hunter-gatherers societies (Table A1.7). Importantly, the mean ages at last birth do not include mothers that died within their reproductive lifespan. To the best of our understanding, he derives estimates of the maternal generation time assuming that: 1) the probability of surviving to age a is described by the arithmetic relationship

$$w(a) = \frac{1-m_f(a-a_f)}{a_l-a_f+1}, \qquad \text{[A1.12]}$$

where m_f is the yearly mortality rate in females, which is assumed to be 0.6%; 2) conditional on surviving, maternal ages at birth are distributed uniformly between the mean ages of first and last birth.

Ideally, one would like to estimate the mean generation time directly from an unbiased sample of ages of birth, thus circumventing the aforementioned assumptions. It is easy to imagine how these assumptions could be violated, e.g., by relationships between the number of births and mortality or the age of last birth, but harder to foresee how these violations would affect estimates of the mean maternal generation time. However, even equipped with just the data employed by Fenner (2005), we believe that the estimates can be improved. In particular, we suggest using a geometric model for survival, i.e.,

$$w(a) \propto \left(1 - m_f\right)^{a-a_f}. \qquad \text{[A1.13]}$$

Importantly, the probabilities of survival should be normalized to one. These revisions yield estimates that are 4%-8% higher than the original ones, with an average maternal generation time of 26.9 instead of 25.6 (see Table A1.7; assuming the same yearly mortality rate in females).

Hunter-gatherer society		Mean Maternal age at first birth	Mean Maternal age at last birth	Maternal generation time	Maternal generation time (revised)
Africa	Dobe Ju/'hoansi (1963–1973)	21.40	34.40	26.71	27.80
Australia	Anbarra	15.90	35.00	23.79	25.25
	Arnhem Land (polygamous)	19.20	34.30	25.41	26.62
	Arnhem Land (monogamous)	19.30	34.10	25.39	26.58
	All	18.13	34.47	24.86	26.15
Asia	Batek (Palawan)	18.00	26.30	21.56	22.11
North America	Mistassini Cree (1828)	21.90	39.00	28.72	30.29
	Kutchin (pre-1900)	22.80	35.00	27.76	28.81
	Kutchin (post-1900)	19.80	39.00	27.50	29.20
	All	21.50	37.67	27.99	29.44
South America	Northern Ache (born before 1959)	19.50	42.10	28.43	30.52
	Northern Ache (reservation period)	17.70	38.50	26.11	27.86
	Yanomama (Mucajai)	16.80	39.90	26.10	28.06
	All	18.00	40.17	26.88	28.81
Average over 5 continents		19.41	34.60	25.60	26.86

Table A1.7. Maternal generation times in hunter-gatherer societies. Mean ages at first and last births taken from Fenner (35).

Fenner (35) estimates the paternal generation time indirectly, using the difference between paternal and maternal ages at first marriage to obtain the paternal ages at first and last birth. He then estimates the paternal generation time based on the model described above, assuming a yearly mortality rate of 0.9% in males. Applying our corrections, we obtain an average paternal age of 33.79 instead of 31.5 years.

Fenner's data suggests that the male-to-female ratio of generation times is generally greater than 1, but also highly variable across societies. Notably, the sex-differences in age at first marriage are highly variable, ranging between -1.5 years in the Bella-Coola and 26 years in the Tiwi. While confidence intervals are not provided, given that in only one out of 157 societies reported females married at an older age than males, the observation that paternal generation times are generally higher is obviously significant. In the few societies where data is available to assess both maternal and paternal generation times, estimates of the ratio of male-to-female generation times vary between 1.03-1.37 (Table A1.8).

Hunter-gatherer society		Mean age difference at first marriage (M – F, years)	Maternal generation time (years)	Paternal generation time (years)	Revised maternal generation time (years)	Revised paternal generation time (years)
Australia	Anbarra[35]	11.5	23.79	32.71	25.25	34.75
Asia	Batek (Palawan) [35]	4.8	21.56	25.88	22.11	26.89
South America	Northern Ache (born before 1959) [35]	5	28.43	31.74	30.52	35.38
	Northern Ache (reservation period) [35]	5	26.11	29.65	27.86	32.74
North America	Mistassini Cree (1828) [35]	3	28.72	30.63	30.29	33.20
	Kutchin (pre-1900) [35]	1	27.76	28.13	28.81	29.77
	Kutchin (post-1900) [35]	1	27.50	27.47	29.20	30.09
	North Greenland polar Eskimos [34]		27	32		

Table A1.8. Estimates of maternal and paternal generation times in societies for which raw data is available. Measurements of maternal generation times and mean age differences are from the same societies but not necessarily from the same time period.

A1.5. Dating species split times

A1.5.1. Inferences

We use our mutational model alongside estimates of life history and spermatogenesis parameters in extant species to suggest plausible ranges for yearly mutation rates on the great apes phylogeny. Specifically, we consider four splits on the human lineage: with chimpanzee, gorilla, orangutan and cercopithecoids. For each split, we assign plausible parameter ranges primarily based on

estimates in the extant species that descend from that split. The parameter ranges are translated into ranges of yearly rates based on our mutational model, which, in turn, are converted into ranges for split times based on sequence divergence.

Parameterization. We make the simplifying assumption that life history traits can evolve independently within the parameter ranges assigned to a given split. This assumption, however, might be implausible for certain combinations of parameter values; for example, a combination of the male puberty age of 3.5 years observed in macaques with the paternal generation time of 33.8 years observed in humans. To avoid (at least some of) these unrealistic combinations, we parameterize the model in terms of: the relative age of male puberty (P/G_M) (instead of the absolute age (P)), the sex-averaged generation time (G) and the ratio of male-to-female generation times (G_M/G_F). We do not change the parameterization of the spermatogenic cycle length (τ) and gestation time (I) (using relative gestation times has negligible effects on our results).

Parameter ranges. Our assumptions about the parameter ranges associated with each split are detailed in Table A1.9. In cases where we have estimates for a parameter in each of the extant species that descend from the split, we use the ranges spanned by these estimates (Table A1.3). When we have estimates for several populations of a given species, we include these estimates as well.

	Assumptions					Predictions			
	Mean generation time	Ratio of male-to-female gen. time	Ratio of male puberty age to male generation time	Gestation time (days)	Spermatogenic cycle length (days)	Autosomal yearly mutation rate ($\times 10^{-9}$ per bp)	Genomic divergence times (MYA)	Split times (MYA) *	Ratio of X-to-autosome substitution rates
Chimpanzee	23 – 30	0.79 – 1.26	0.26 – 0.39	229-280	14 - 16	0.39 – 0.6	11.4 – 17.6	6.6 – 10.2	0.77 – 0.86
Gorilla	19 - 30	0.79 – 1.26	0.26 – 0.39	229-280	**12.5** – 16	0.39-0.69	12.7 – 22.4	9.4 – 16.6	0.76 – 0.87
Orangutan	**15**-30	0.79 – 1.26	0.26 – 0.39	229-280	**11** - 16	0.39-0.82	20.7 – 43.6	3MY after genome divergence date	0.75 – 0.89
Cercopithecoids	11-30	0.79 – 1.26	0.26 – 0.39	132-280	9.5 - 16	0.39-1.03	30.2 – 79.9		0.74 – 0.92

Table A1.9. Parameter ranges and corresponding predictions for the different splits. Parameter ranges that require additional assumptions (i.e., for which we lack estimates in one or more species) are underlined. * Values used in Fig. 1.5.

Other cases, in which we lack estimates for one or more extant species, and our corresponding assumptions, are the following:

- As far as we can ascertain, paternal generation times have only been estimated in hominines. We therefore lack estimates for the ratio of male-to-female generation times

(G_M/G_F) and relative male puberty age (P/G_M) in orangutans and cercopithecoids. For these splits, we conservatively assume the parameter ranges spanned by the estimates in hominines.

- We lack an estimate for the average generation time in orangutans. We assume a lower bound on the average generation time following the split with orangutans that is intermediate between the lower bound for hominines and cercopithecoids (which generally have lower generation times than hominines). While our bound of 15 years is likely lower than the average generation time in extant orangutans, it may be a plausible value for this lineage, given the observed increase in body mass in great apes over the past 15-20 million years (48) and the correlation between body mass and generation time ((40, 49); but note that few putative Miocene hominoids are not smaller than extant hominids).
- We also lack estimates for spermatogenic cycle lengths in gorillas and orangutans. We assume that the rates along the corresponding lineages were between the low lengths observed in cercopithecoids and the high lengths observed in humans and chimpanzees.

Estimating split times. To obtain split times we divide estimates of divergence since the split by yearly mutation rates. To this end we rely on the average genome-wide divergence reported by Scally et al. (50) (Table A1.10). For the chimpanzee and gorilla splits (from humans), these estimates have been corrected for the contribution of ancestral polymorphism (inferred from patterns of ILS). For the macaque and orangutan splits from humans, we follow Scally and Durbin (49) in assuming a (relatively high) ancestral coalescence time of 3MY (cf. 51).

	Chimpanzee	Gorilla	Orangutan	Macaque
Average divergence	1.37%	1.75%	3.4%	6.23%
Divergence post-split	0.796%	1.292%	-	-
Ancestral contribution	0.697%	0.376%	-	-

Table A1.10. The genome-wide average divergence from humans. Estimates correspond to the Scally et al. (2012) 'FULL' dataset. Estimates of the contribution of ancestral polymorphism and divergence after the splits were obtained using CoalHMM (52, 53). The estimates of divergence post-splits and levels of ancestral polymorphisms do not add up to the total divergence, because of CoalHMM's assumption of an ultra-metric hominine tree.

A1.5.2. Fossil-based estimates

Given the position of a fossil on a phylogeny, e.g., the branch on which it occurred, the fossil's dating provides lower or upper bounds on the timing of splits along the phylogeny. With estimates of sequence divergence corresponding to these splits, these bounds can be translated into bounds on the average yearly mutation rate along the corresponding branches. In practice, such inferences about split times and mutation rates in catarrhines are severely limited by the paucity of fossils and uncertainties about their position in the fossil record. Here we briefly review the fossil evidence used in our analyses of split times on the hominoid phylogeny (Fig. 1.5).

Several catarrhine fossils have been suggested to flank the splits between orangutan and hominines and between hominoids and cercopithecoids, although most of these positions are debated. The association of fossils with splits among hominines is regarded as less reliable (54), and evidence is lacking in particular for the last common ancestor of humans and chimpanzees as well as for the early gorillins (i.e., species related to the gorilla lineage following the split with humans and

chimpanzees). In general, lower bounds on split times are considered to be more reliable than upper bounds, because fossils that share derived characters with extant species on a given branch are more likely to have occurred after that branch has split off whereas fossils that share derived characters with the descendants of the branch that preceded the split might have occurred after it and gone extinct (55). For that reason, conservative maximal bounds for the hominine-orangutan and hominoid-cercopithecoid splits should be inferred from the earliest (rather than latest) putative ancestor (cf. (56, 57)); i.e., from the earliest catarrhine fossil for the hominoid-cercopithecoid split and the earliest hominoid fossil for the hominine-orangutan split. For all these reasons, the common practice is to infer mutation rates by considering the hominine-orangutan and hominoid-cercopithecoid splits and distinguish between estimates derived from lower and upper bounds (Table A1.11).

Maximal bounds for the hominoid-cercopithecoid split time rely on the earliest catarrhine species following the catarrhine-platyrrhine split. The earliest widely accepted stem-catarrhines are propliopithecids (*Propliopithecus*, earliest appearance 31-31.5 MYA and the later *Aegyptopithecus* (58)). *Catopithecus* and *Oligopithecus* (earliest appearance ~34 MYA (58)) are generally considered to be stem-anthropoids, although they have also been argued to be the earliest stem-catarrhines (48). Lower bounds for the hominoid-cercopithecoid split rely on either the earliest cercopithecoid or the earliest hominoid. The earliest possible cercopithecoid is the recently discovered *Nsungwepithecus* (~25 MYA) (59), but it remains to been seen whether its positioning becomes widely accepted (57). The next candidate on the cercopithecoid lineage is *Victoriapithecus* (~19.5 MYA), which is unanimously recognized as a cercopithecoid but provides a much lower bound. While there are earlier hominoid candidates, none is unanimously accepted. The putative hominoids *Rukwapithecus* (~25 MYA), *Kamoyapithecus* (~25 MYA), *Ekembo*

(earliest appearance ~23 MYA; for clarity we use the previous classification, *Proconsul,* in Fig. 1.5 (60)) and *Morotopithecus* (~20.6 MYA) have also been argued to be stem-catarrhines, and the dating of *Ekembo* and *Morotopithecus* has been challenged with revised estimates of ~21 and ~17 MYA respectively (cf. (57, 59-61)). Other putative hominoids do not provide useful bounds since they are younger than *Victoriapithecus* (e.g., Harrison suggests *Kenyapithecus*, ~14 MYA, as the earliest unanimously accepted hominoid (61)). In summary, the narrowest (and thus speculative) range inferred for the hominoid-cercopithecoid split time appears to be 25-31.5 MYA, whereas a more conservative range appears to be 19.5-34 MYA (cf. (56, 57, 61) and Table A1.11).

Maximal bounds for the hominine-orangutan split time rely on the earliest hominoid species following the hominoid-cercopithecoid split (see above), with possible upper bounds as high as ~25 MYA. We note that such upper bounds (e.g., 23 MYA as suggested by (56)) are substantially higher than the those assumed by most molecular studies (e.g., (62) and (50)). Lower bounds rely on either the earliest pongine (i.e., species related to the orangutan lineage following the split with hominines) or the earliest hominine. The earliest fossil widely interpreted as pongine is *Sivapithecus* (earliest appearance 12.5 MYA), but its classification as pongine still raises some challenges and the Orangutan-like facial-palatal features of *Sivapithecus* are not found until ~9 MYA (cf. (63); D. Pilbeam Personal Communication). Other possible pongines include *Khoratpithecus* (earliest appearance ~13 MYA, but the identification of early samples has been contested (57)), *Ankarapithecus* (10 MYA), *Lufengpithecus* (9 MYA) and *Indopithecus* (6.5 MYA; sometimes classified as *Gigantopithecus*). None of these is unanimously accepted. There is even less certainty regarding putative early hominines, with early candidates including *Dryopithecus* (12.5 MYA), and (with even less certainty) *Pierolapithecus* (12.5 MYA) and *Anoiapithecus* (12.5 MYA); Later candidates include *Rudapithecus* (10 MYA), *Hispanopithecus*

(10 MYA) and *Ouranopithecus* (9.5 MYA) (57). In summary, lower bounds for the hominine-orangutan split are still under debate, where the most widely accepted based on *Sivapithecus,* supports a lower bound of 12.5 MYA.

Fossil based estimates of yearly mutation rates in conjunction with sequence divergence have been used to date split times that lack reliable fossil dating, where these estimates typically assume that mutation rates have remained constant on different branches of the phylogeny (55) (see Table A1.11; (49, 50, 62)). By comparison, the naïve pedigree based estimate of $0.4 \cdot 10^{-9}$ per bp per year (6) places the Human-Chimpanzee split 9.95 MYA and the Human-Gorilla split 16.2 MYA (see Table A1.10).

Species	Dating (MYA)	Putative position (relative to splits from the human lineage)	Inferred mutation rate ($x10^{-9}$ per year)	Implications for other split times from the human lineage (MYA)			
				Macaque	Orangutan	Gorilla	Chimpanzee
Catopithecus or *Oligopithecus*	34	Predating split from macaque	> 0.84	-	< 17.2	< 7.7	< 4.7
Propliopithecus	31.5	Predating split from macaque	> 0.90	-	< 15.8	< 7.2	< 4.4
Rukwapithecus, Kamoyapithecus or *Nsungwepithecus*	25	Following split with macaque	< 1.11	-	> 12.3	> 5.8	> 3.6
Ekembo	23	Following split with macaque	< 1.20	-	> 11.2	> 5.4	> 3.3
Victoriapithecus	19.5	Following split with macaque	< 1.38	-	> 9.3	> 4.7	> 2.9
Ekembo	23	Predating split with orangutan	> 0.65	< 44.6	-	< 9.9	< 6.1
Sivapithecus or *Dryopithecus*	12.5	Following split with orangutan	< 1.10	> 25.4	-	> 6.9	> 3.6
Pedigree studies (6)			0.4	74.9	39.5	16.2	9.95

Table A1.11. Fossil evidence associated with catarrhine species splits and the corresponding bounds on yearly mutation rates, based on divergence (Table A1.10; also see Fig. 1.5 and (49)).

While the positions of hominine fossils are more contentious, it is still interesting to see how their hypothesized positions match-up with our bounds on split times. To this end, we consider the following fossils (Fig. 1.5): *Chororapithecus* (~8 MYA (64)), is suggested to be related to the gorilla lineage following the split with humans (65), although this hypothesis has been contested (57, 61, 65); *Ardipithecus*[SEP](5.6-5.8 MYA), *Orrorin* (5.8 MYA) and *Sahelanthropus* (6.5-7.5 MYA), hypothesized to be hominins (i.e., following the split with chimpanzees). If we assume *Chororapithecus* is related to the gorilla lineage, it would imply an average yearly mutation rate of at most 0.81×10^{-9} per-bp on the human and gorilla branches. Placing the other fossils on the hominin lineage would imply average yearly rates of at most 0.7×10^{-9} (*Ardipithecus*), 0.69×10^{-9} (*Orrorin*) and 0.57×10^{-9} (*Sahelanthropus*) per bp on the human and chimpanzee branches.

A1.6. Notation

Notation	Short definition
G_F	Average generation time in females (years)
G_M	Average generation time in males (years)
G_A	Average sex-averaged generation time on autosomes (years)
G_X	Average sex-averaged generation time on the X (years)
μ_F	Average mutation rate in females (per bp per generation)
μ_M	Average mutation rate in males (per bp per generation)
μ_A	Average autosomal mutation rate (per bp per generation)
μ_X	Average mutation rate on the X (per bp per generation)
I	Gestation time (years)
P	Average age of puberty in males (years)
τ	Spermatogenic cycle length (days)
C_M	Mutation rate per pb in pre-pubertal males
D_M	Mutation rate per pb per spermatogenic stem cell division
T	Lineage length (years)
K_A	Average rate of autosomal substitutions over a lineage (per bp)
K_X	Average rate of substitutions on the X over a lineage (per bp)

Table A1.12. Notation for Appendix 1.

A1.7. Supporting Figures

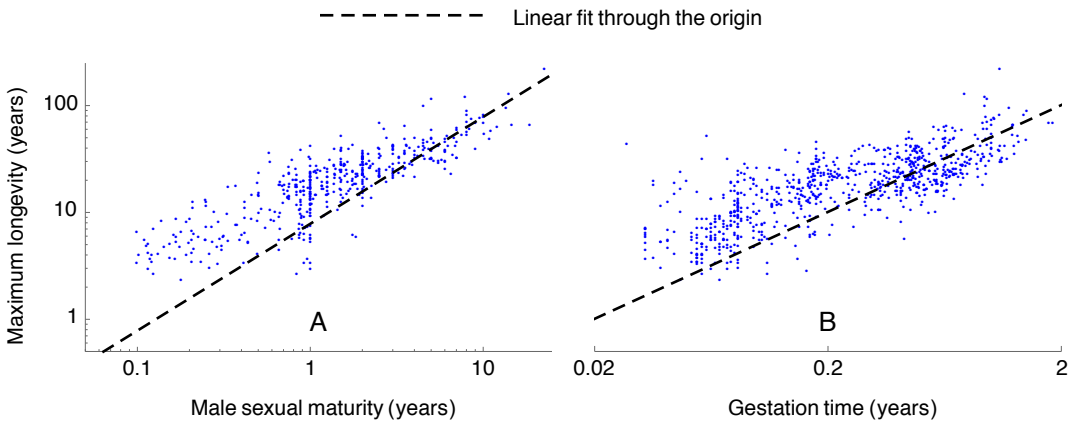

Fig. A1.1. Maximum longevity vs. age of male sexual maturity for 482 species (A) and gestation time for 806 species (B) in mammals. The correlations accords with our assumption that age of puberty and gestation time scale with generation time (cf. Fig. .6, Fig. A1.5 and section A1.3). While the graphs are shown in logarithmic scales, the fit was performed on the linear scale. The parameter estimates are: male sexual maturity = 13% of maximum longevity and gestation time = 2% of maximum longevity. While maximum longevity and age at maturity are reasonable proxies for the generation time and age at puberty, they are obviously expected to be larger. Data from *AnAge: The animal ageing and longevity database,* build 13 (36).

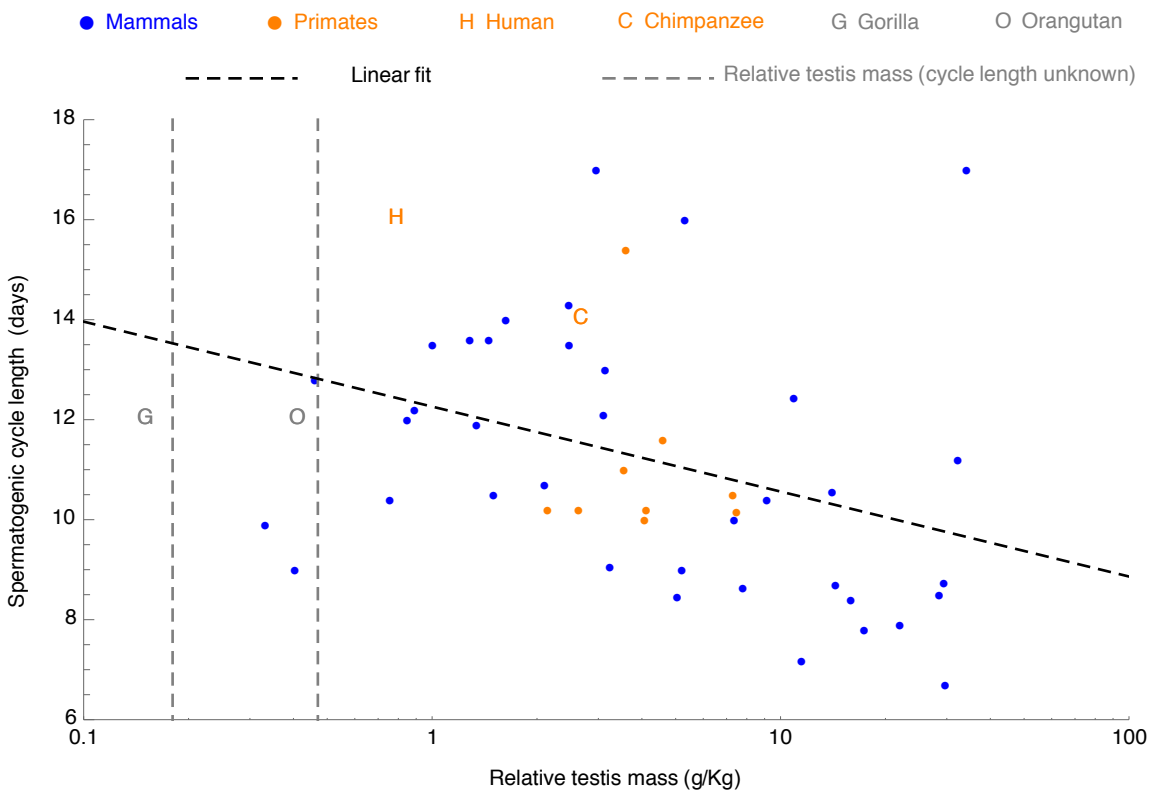

Fig. A1.2. Spermatogenic cycle length vs. relative testis mass in 48 mammals. The dashed gray lines denote the relative testis mass in orangutans and gorillas, in which the spermatogenic cycle length has not been measured. The dashed black line denotes the linear regression of spermatogenesis rates against relative testis mass $(cycle\ length\ (days) = 12.3 - 0.74 \times \log\ (relative\ mass\ (g/Kg))$. Data from (41, 66).

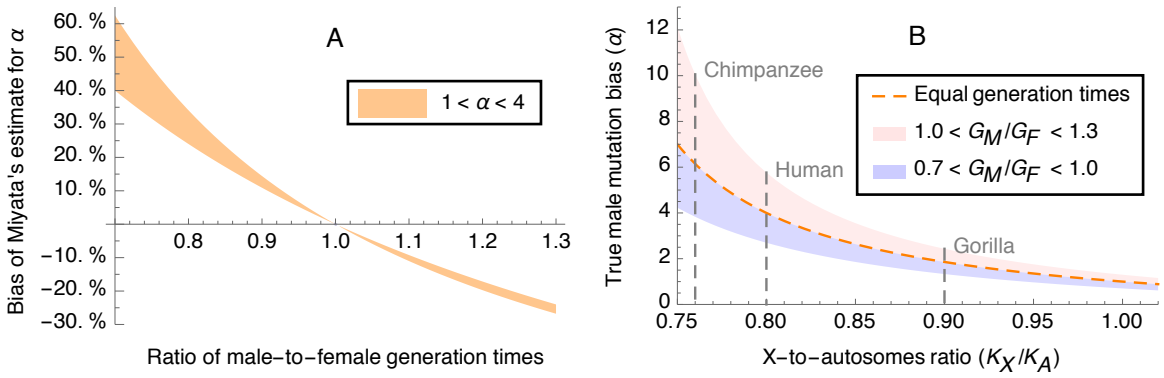

Fig. A1.3. (A) Bias in estimates of α based on Miyata's formula as a function of the male-to-female ratio of generation times. The orange range corresponds to values of α ranging between 1 and 4, where the absolute value of the bias increases with α. Note that the bias can be substantial even when the sex-bias in mutation rates is small (i.e., when α is near 1). (B) True male mutational bias (α) as a function of the X-to-autosome substitution ratio, under a range of male-to-female ratios of generation times. The marked values for hominines are rough estimates of the X-to-autosome substitution ratio (as they include the contribution of ancestral polymorphism).

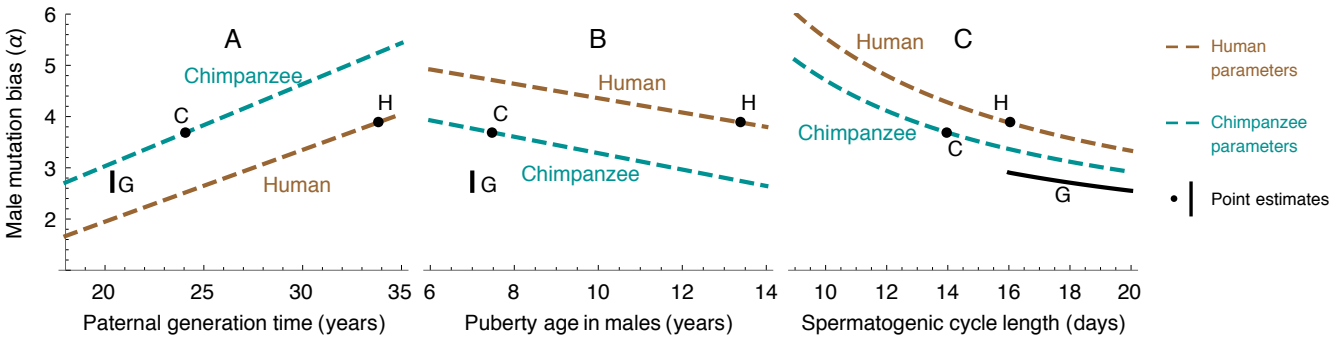

Fig. A1.4. Predicted male mutational bias as a function of paternal generation time (A), male age of puberty (B) and rate of spermatogenesis (C). Other details are the same as in Fig. 1.2.

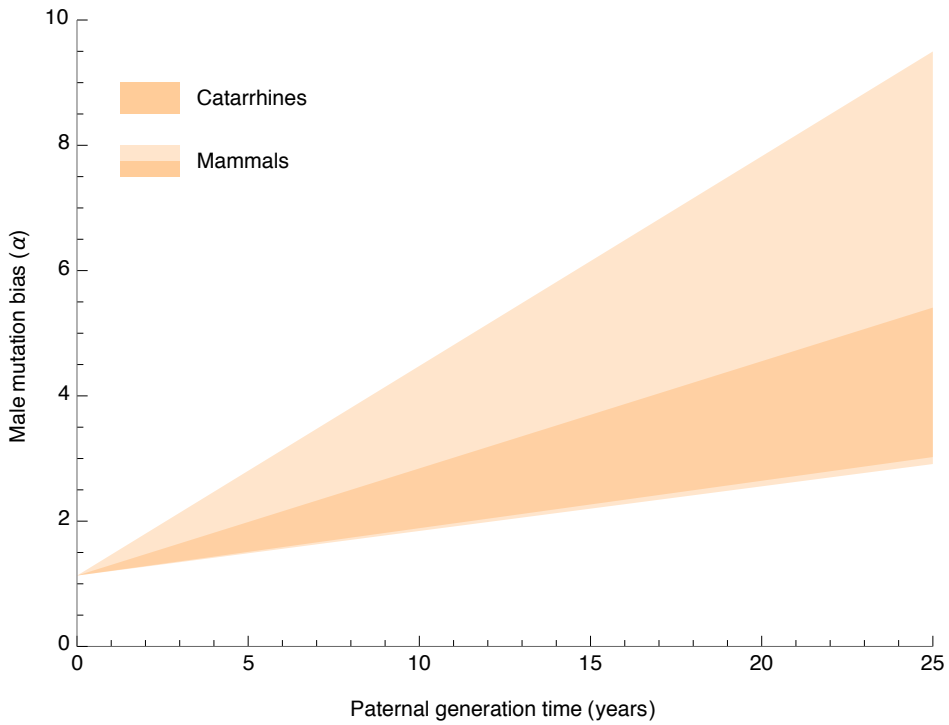

Fig. A1.5. Male mutation bias (α) as function of paternal generation time in a broader phylogenetic context, based on our mutational model. Other parameters roughly correspond to catarrhines and mammals (Table A1.4).

A1.8. References

1. Charlesworth B (1980) *Evolution in age-structured populations* (Cambridge University Press, Cambridge).
2. Cox DR (1962) *Renewal Theory* (Spottiswoode Ballantyne and Co. Ltd., London and Colchester).
3. Miyata T, Hayashida H, Kuma K, Mitsuyasu K, & Yasunaga T (1987) Male-driven molecular evolution: a model and nucleotide sequence analysis. *Cold Spring Harbor symposia on quantitative biology*, pp 863-867.
4. Segurel L, Wyman MJ, & Przeworski M (2014) Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet* 15:47-70.
5. Gao Z, Wyman MJ, Sella G, & Przeworski M (2015) Interpreting the dependence of mutation rates on age and time. *ArXiv e-prints.* 1507:6890.
6. Kong A*, et al.* (2012) Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488(7412):471-475.
7. Thorsson A, Dagbjartsson A, Palsson G, & Arnorsson V (1999) Puberty in Icelandic boys. *Laeknabladid* 86(10):665-659.
8. Drost JB & Lee WR (1995) Biological basis of germline mutation: comparisons of spontaneous germline mutation rates among drosophila, mouse, and human. *Environmental and molecular mutagenesis* 25(S2):48-64.
9. Li W-H, Ellsworth DL, Krushkal J, Chang BH-J, & Hewett-Emmett D (1996) Rates of nucleotide substitution in primates and rodents and the generation–time effect hypothesis. *Molecular phylogenetics and evolution* 5(1):182-187.
10. Lynch M (2010) Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A* 107(3):961-968.
11. Adewoye AB, Lindsay SJ, Dubrova YE, & Hurles ME (2015) The genome-wide effects of ionizing radiation on mutation induction in the mammalian germline. *Nat Commun* 6:6684.
12. Nielsen CT*, et al.* (1986) Onset of the Release of Spermatozia (Supermarche) in Boys in Relation to Age, Testicular Growth, Pubic Hair, and Height. *J Clin Endocrinol Metab* 62(3):532-535.
13. Marshall WA & Tanner JM (1970) Variations in the pattern of pubertal changes in boys. *Arch Dis Child* 45(239):13-23.
14. Heller CG & Clermont Y (1963) Spermatogenesis in man: an estimate of its duration. *Science* 140(3563):184-186.
15. Marson J, Meuris S, Cooper R, & Jouannet P (1991) Puberty in the male chimpanzee: progressive maturation of semen characteristics. *Biology of reproduction* 44(3):448-455.
16. Behringer V, Deschner T, Deimel C, Stevens JM, & Hohmann G (2014) Age-related changes in urinary testosterone levels suggest differences in puberty onset and divergent life history strategies in bonobos and chimpanzees. *Horm Behav* 66(3):525-533.
17. Martin D, Swenson R, & Collins D (1977) Correlation of serum testosterone levels with age in male chimpanzees. *Steroids* 29(4):471-481.

18. Smithwick EB, Young LG, & Gould KG (1996) Duration of spermatogenesis and relative frequency of each stage in the seminiferous epithelial cycle of the chimpanzee. *Tissue & Cell* 28(3):357-366.
19. Parish AR (1996) Female relationships in bonobos (Pan paniscus). *Hu Nat* 7(1):61-96.
20. Langergraber KE*, et al.* (2012) Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc Natl Acad Sci U S A* 109(39):15716-15721.
21. Dixson AF (2009) *Sexual selection and the origins of human mating systems* (Oxford Univ, Press, Oxford).
22. Kingsley SR (1988) Physiological development of male orang-utans and gorillas. *Orang-utan Biology*, ed Schwartz JH (Oxford Univ. Press, New York).
23. Robbins MM & Czekala NM (1997) A preliminary investigation of urinary testosterone and cortisol levels in wild male mountain gorillas. *American Journal of Primatology* 43(1):51-64.
24. Wich S*, et al.* (2009) Orangutan life history variation. *Orangutans: Geographic Variation in Behavioral Ecology and Conservation*, eds Wich S, Utami Atmoko S, Mitra Setia T, & van Schaik C (Oxford Univ Press, Oxford).
25. Dixson A, Knight J, Moore H, & Carman M (1982) Observations on sexual development in male Orang-utans. *Int Zoo Yearb* 22(1):222-227.
26. Fooden J & Izor RJ (1983) Growth curves, dental emergence norms, and supplementary morphological observations in known‐age captive orangutans. *American Journal of Primatology* 5(4):285-301.
27. Reichard UH & Barelli C (2008) Life history and reproductive strategies of Khao Yai Hylobates lar: Implications for social evolution in apes. *Int J Primatol* 29(4):823-844.
28. Molur S*, et al.* (2003) *Status of South Asian Primates: Conservation Assessment and Management Plan (C.A.M.P.) workshop report* (Zoo Outreach Organization and CBSG South Asia, Coimbatore, India).
29. Barr AB (1973) Timing of spermatogenesis in four nonhuman primate species. *Fertil Steril* 24(5):381-389.
30. De Rooij D, van Alphen M, & van de Kant H (1986) Duration of the cycle of the seminiferous epithelium and its stages in the rhesus monkey (Macaca mulatta). *Biology of reproduction* 35(3):587-591.
31. Aslam H*, et al.* (1999) The cycle duration of the seminiferous epithelium remains unaltered during GnRH antagonist-induced testicular involution in rats and monkeys. *J Endocrinol* 161(2):281-288.
32. Chowdhury AK & Steinberger E (1976) A study of germ cell morphology and duration of spermatogenic cycle in the baboon, Papio anubis. *Anat Rec* 185(2):155-169.
33. Clermont Y & Antar M (1973) Duration of the cycle of the seminiferous epithelium and the spermatogonial renewal in the monkey Macaca arctoides. *American Journal of Anatomy* 136(2):153-165.
34. Matsumura S & Forster P (2008) Generation time and effective population size in Polar Eskimos. *Proc Biol Sci* 275(1642):1501-1508.
35. 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(2):415-423.

36. Tacutu R*, et al.* (2013) Human Ageing Genomic Resources: integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Res* 41(Database issue):D1027-1033.
37. van Noordwijk MA & van Schaik CP (1999) The effects of dominance rank and group size on female lifetime reproductive success in wild long-tailed macaques, Macaca fascicularis. *Primates* 40(1):105-130.
38. Altmann J, Gesquiere L, Galbany J, Onyango PO, & Alberts SC (2010) Life history context of reproductive aging in a wild primate model. *Ann N Y Acad Sci* 1204(1):127-138.
39. Bronikowski AM*, et al.* (2002) The aging baboon: comparative demography in a non-human primate. *Proc Natl Acad Sci U S A* 99(14):9591-9595.
40. De Magalhães JP, Costa J, & Church GM (2007) An analysis of the relationship between metabolism, developmental schedules, and longevity using phylogenetic independent contrasts. *J Gerontol A Biol Sci Med Sci* 62(2):149-160.
41. Ramm SA & Stockley P (2010) Sperm competition and sperm length influence the rate of mammalian spermatogenesis. *Biol Lett* 6(2):219-221.
42. Gesquiere LR*, et al.* (2005) Coming of age: steroid hormones of wild immature baboons (Papio cynocephalus). *American journal of primatology* 67(1):83-100.
43. Amann RP (2008) The cycle of the seminiferous epithelium in humans: a need to revisit? *J Androl* 29(5):469-487.
44. Courot M, Hochereau-de Reviers M-T, & Ortavant R (1970) Spermatogenesis. *The testis* 1:339-432.
45. Helgason A, Hrafnkelsson B, Gulcher JR, Ward R, & Stefansson K (2003) A populationwide coalescent analysis of Icelandic matrilineal and patrilineal genealogies: evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes. *Am J Hum Genet* 72(6):1370-1388.
46. Helgason A*, et al.* (2015) The Y-chromosome point mutation rate in humans. *Nat Genet* 47(5):453-457.
47. Teleki G, Hunt EE, & Pfifferling J-H (1976) Demographic observations (1963–1973) on the chimpanzees of Gombe National Park, Tanzania. *Journal of Human Evolution* 5(6):559-598.
48. Hartwig WC (2002) *The primate fossil record* (Cambridge Univ. Press).
49. Scally A & Durbin R (2012) Revising the human mutation rate: implications for understanding human evolution. *Nat Rev Genet* 13(10):745-753.
50. Scally A*, et al.* (2012) Insights into hominid evolution from the gorilla genome sequence. *Nature* 483(7388):169-175.
51. Mailund T, Munch K, & Schierup MH (2014) Lineage sorting in apes. *Annual review of genetics* 48:519-535.
52. Hobolth A, Christensen OF, Mailund T, & Schierup MH (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet* 3(2):e7.
53. Dutheil JY*, et al.* (2009) Ancestral population genomics: the coalescent hidden Markov model approach. *Genetics* 183(1):259-274.
54. Wood B & Harrison T (2011) The evolutionary context of the first hominins. *Nature* 470(7334):347-352.

55. Benton MJ & Donoghue PC (2007) Paleontological evidence to date the tree of life. *Mol Biol Evol* 24(1):26-53.
56. Jensen Seaman MI & Hooper Boyd KA (2013) Molecular Clocks: Determining the Age of the Human–Chimpanzee Divergence. *eLS (John Wiley & Sons: Chichester) doi: 10.1002/9780470015902.a0020813.pub2*.
57. Begun DR (2015) Fossil Record of Miocene Hominoids. *Handbook of Paleoanthropology*, eds Henke W & Tattersall I (Springer, Heidelberg), 2nd Ed Vol 2, pp 1261-1332.
58. Seiffert ER (2006) Revised age estimates for the later Paleogene mammal faunas of Egypt and Oman. *Proc Natl Acad Sci U S A* 103(13):5000-5005.
59. Stevens NJ*, et al.* (2013) Palaeontological evidence for an Oligocene divergence between Old World monkeys and apes. *Nature* 497(7451):611-614.
60. McNulty KP, Begun DR, Kelley J, Manthi FK, & Mbua EN (2015) A systematic revision of Proconsul with the description of a new genus of early Miocene hominoid. *J Hum Evol* 84:42-61.
61. Terry H (2010) Dendropithecoidea, Proconsuloidea, and Hominoidea. *Cenozoic Mammals of Africa*, eds Werdelin L & Sanders WJ (University of California Press, Berkeley), pp 429-469.
62. Patterson N, Richter DJ, Gnerre S, Lander ES, & Reich D (2006) Genetic evidence for complex speciation of humans and chimpanzees. *Nature* 441(7097):1103-1108.
63. Morgan ME*, et al.* (2015) A partial hominoid innominate from the Miocene of Pakistan: description and preliminary analyses. *Proc Natl Acad Sci U S A* 112(1):82-87.
64. Suwa G*, et al.* (2015) Newly discovered cercopithecid, equid and other mammalian fossils from the Chorora Formation, Ethiopia. *J. Archaeol. Sci.* 123(1):19-39.
65. Suwa G, Kono RT, Katoh S, Asfaw B, & Beyene Y (2007) A new species of great ape from the late Miocene epoch in Ethiopia. *Nature* 448(7156):921-924.
66. Presgraves DC & Yi SV (2009) Doubts about complex speciation between humans and chimpanzees. *Trends in Ecology & Evolution* 24(10):533-540.

Appendix 2 – Supplementary Information for Chapter 2

A2.1 Haploid Model

Here we rigorously solve the haploid model with age-structure and endogenous reproductive variance, and relate our results with those of previous work that considered special cases of our model. In Section A2.1.2 we consider the implications of our assumptions about endogenous reproductive variance. In Section A2.1.3 we solve for the joint stationary distribution of the age and relative reproductive success associated with an allele, going backwards in time. Based on this distribution, we derive the stationary per-generation coalescence rate for a sample of two alleles, to obtain Eq. 2.7 in chapter 2:

$$N_e = \frac{M \cdot G}{W}. \tag{A2.1}$$

In Section A2.1.4, we recast these results in terms of total reproductive variance, showing that the relationship derived by Hill for the case with age-structure alone (1):

$$N_e = G \cdot M_1/V, \tag{A2.2}$$

applies to the extended model with endogenous reproductive variance, and derive Eq. 2.11 in for the total reproductive variance in this case, i.e.,

$$V = W \cdot (M_1/M). \tag{A2.3}$$

In section A2.1.5 we consider the case without endogenous reproductive variance, to show that the our results for the effective population size (Eq. A2.1) reduce to the formula obtained by Felsenstein (2), and to consider a simple example of how age-structure affects the effective population size.

A2.1.1 Requirements on f_a

When we introduced the haploid model with endogenous reproductive variance, we assumed that each newborn is assigned a relative reproductive success vector $\vec{r}$, and denoted the (constant) proportion of individuals with a given vector $\vec{r}$ in age class a by $f_a(\vec{r})$ (see Table A2.1 for summary of notation). Here we describe the requirements on the probability mass function f_a that these assumptions entail. First, given that the probability of being born to a parent of age a is p_a, and to a specific parent of age a and with reproductive success $\vec{r}$ is $p_a \cdot \frac{r_a}{M_a}$, we require that $E_{f_a}(r_a) = \sum_{\vec{r}} f_a(\vec{r}) \cdot r_a = 1$ for any age a. Second, given that the number of individuals with a given $\vec{r}$ can only decrease with age (due to mortality), we further require that $M_a \cdot f_a(\vec{r}) \geq M_{a+1} \cdot f_{a+1}(\vec{r})$. Third, requiring that the number of individuals of a given age a and with a given $\vec{r}$ is constant and equal to $M_a \cdot f_a(\vec{r})$, implies that this number needs to be an integer. Notably, if we would like to model the distribution of relative reproductive success using a given (continuous or discrete) distributions $\widetilde{f_a}(\vec{r})$, which satisfies the first two requirements, we would need to discretize $\widetilde{f_a}$ to obtain a probability mass function $\tilde{f}_a'$ such that $M_a \cdot \tilde{f}_a'(\vec{r})$ is always an integer. However, if we assume that the relative sizes of the age-class, i.e., the ratios M_i/M_j, are constant, and increase the total population sizes, the discretized functions $\tilde{f}_a'$ will approach $\widetilde{f_a}$, and the value of the $W_{i,j} = E_{\tilde{f}_j'}(r_i \cdot r_j) = E(r_i \cdot r_j | \text{survival to age } j)$ terms, which summarize the effect of endogenous reproductive variance on the effective population size, will approach $E_{\widetilde{f_j}}(r_i \cdot r_j)$. We implicitly assumed this limit when we considered the special case in which relative reproductive success is independent of age and of mortality rates. More generally, while the assumption that for any age a, $M_a \cdot f_a(\vec{r})$ is an integer, might appear highly restrictive, these restrictions are relaxed under the standard coalescent assumption that the population size is sufficiently large.

Notation	Definition	Remarks
p_a	Probability that a newborn descends from a parent of age a	$\sum_a p_a = 1$
q_a	Probability that a newborn descends from a parent of age $\geq a$	$q_a = \sum_{i \geq a} p_i$
G	Average generation time	$G = \sum_a a \cdot p_a$
M_a	Number of individuals of age a (M_1 is the number of newborns per-year)	$M_{a+1} \leq M_a$
$\vec{r}$	Relative reproductive success, where component r_a is the relative reproductive success at age a	
$f_a(\vec{r})$	The frequency of individuals with relative reproductive success $\vec{r}$ among individuals of age a	
$g_a(\vec{r})$	Given an individual I of age a and a newborn n, $g_a(\vec{r})$ is the probability that I has relative reproductive success $\vec{r}$, conditioned on n being descended from I	$g_a(\vec{r}) = r_a f_a(\vec{r})$
$\epsilon(a, \vec{r})$	Joint stationary probability of age a and relative reproductive success $\vec{r}$ along a lineage, going backwards in time	$\epsilon(a, \vec{r}) = \frac{1}{G} \sum_{j \geq a} p_j g_j(\vec{r})$
ϵ_a	Marginal stationary distribution of age a	$\epsilon_a = \frac{q_a}{G}$
M	Effective age-class size	See Eq. A2.16
$W_{i,j}$	Average value of $r_i \cdot r_j$ among individuals of sex s and age a	Defined for $i \leq j$
W	Weighted average of the $W_{i,j}$	See Eq. A2.15
X, X_a	An individual's number of offspring, throughout its life or at age a, respectively	
V	Reproductive variance (i.e., $V = Var(X)$)	See Eq. A2.28
S_a	The event of surviving to age $\geq a$	

Table A2.1: Notation for the haploid model, with parameters of the model in red.

A2.1.2 Solution backwards in time

Here, we extend the derivations of Sagitov and Jagers (3) to account for endogenous reproductive variance. Tracing an allele backward in time, the age a_t and relative reproductive success $\vec{r}_t$ of the individual I_t who carries the allele t years in the past defines a Markov chain $(a_t, \vec{r}_t)$. To define the transition probabilities of the chain, we distinguish between two cases. First, if the individual carrying the allele is not a newborn, i.e., $a_t > 1$, then at time t+1 that individual will be one year

younger and its relative reproductive success $\vec{r}$ will remain unchanged, i.e., $(a_{t+1}, \vec{r}_{t+1}) = (a_t - 1, \vec{r}_t)$ with probability one. Second, if individual carrying the allele is a newborn, i.e., $a_t = 1$, then a_{t+1} equals a with probability p_a. The probability mass function of $\vec{r}_{t+1}$ conditional on a_{t+1}, follows from Bayes' theorem, further conditioning on the fact that the parent, $I_{t+1} = I$, necessarily reproduced successfully

$$P(\vec{r}_I = \vec{r} | I_{t+1} = I, a_{t+1} = a)$$

$$= \frac{P(I_{t+1}=I|\vec{r}_{t+1}=\vec{r}, a_{t+1}=a) \cdot P(\vec{r}_I=\vec{r}|\, a_{t+1}=a)}{P(I_{t+1}=I)} = \frac{(r_a/M_a) \cdot f_a(\vec{r})}{\sum_{\vec{k}} (r_a/M_a) \cdot f_a(\vec{k})} = r_a \cdot f_a(\vec{r}). \quad \text{(A2.4)}$$

We denote this probability by $g_a(\vec{r}) \equiv r_a \cdot f_a(\vec{r})$, and conclude that

$$P((a_{t+1}, \vec{r}_{t+1}) = (a, \vec{r}) | a_t = 1) = p_a \cdot g_a(\vec{r}). \quad \text{(A2.5)}$$

g_a is a proper probability mass function since $\sum_{\vec{r}} g_a(\vec{r}) = \sum_{\vec{r}} r_a \cdot f_a(\vec{r}) = 1$. Moreover, the parent's expected value of r_a is $E_{\vec{r} \sim g_a}(r_a) = E_{\vec{r} \sim f_a}(r_a^2) = 1 + V_{\vec{r} \sim f_a}(r_a) \geq 1$. The latter inequality makes intuitive sense, as it implies that the allele is more likely to be descended from an individual that has higher than average relative reproductive success in its age class.

We rely on the transition probabilities to derive and solve a recursion for the stationary probability $\epsilon(a, \vec{r})$ of age, a and relative reproductive successes, $\vec{r}$, of the individuals carrying the allele. Namely,

$$\epsilon(a, \vec{r}) = \epsilon(a + 1, \vec{r}) + \left(\sum_{\vec{k}} \epsilon\left(1, \vec{k}\right)\right) \cdot p_a \cdot g_a(\vec{r}), \quad \text{(A2.6)}$$

where the first term corresponds to aging within the same individual and the second corresponds to parenting a newborn. In order to solve these recursions we first consider the marginal stationary distribution of age, $\epsilon_a = \sum_{\vec{r}} \epsilon(a, \vec{r})$. To this end, we sum the recursions over $\vec{r}$ to obtain recursions on the marginal distribution,

$$\epsilon_a = \epsilon_{a+1} + \epsilon_1 \cdot p_a, \quad \text{(A2.7)}$$

where we also require that $\sum_a \epsilon_a = 1$. This recursion was solved by Sagitov and Jagers (3) for the case without endogenous reproductive variance, yielding

$$\epsilon_a = q_a/G, \quad \text{(A2.8)}$$

where $q_a \equiv \sum_{j \geq a} p_j$. Substituting this expression into Eq. A2.6, the recursions simplify to

$$\epsilon(a, \vec{r}) = \epsilon(a+1, \vec{r}) + \frac{1}{G} \cdot p_a \cdot g_a(\vec{r}), \quad \text{(A2.9)}$$

where we further require that $\sum_{a,\vec{r}} \epsilon(a, \vec{r}) = 1$. The solution of these recursions is

$$\epsilon(a, \vec{r}) = \frac{1}{G} \sum_{j \geq a} p_j g_j(\vec{r}). \quad \text{(A2.10)}$$

The marginal stationary probability mass function of $\vec{r}$ is $\sum_a \epsilon(a, \vec{r}) = \frac{1}{G} \sum_j \left(j \cdot p_j\right) \cdot g_j(\vec{r})$, which is a proper probability mass function because $\frac{1}{G} \sum_j j \cdot p_j = 1$, and $\sum_{\vec{r}} g_a(\vec{r}) = 1$ for any age *a*.

We rely on the stationary distribution to derive the probability of coalescence of two alleles, along the same lines as detailed in chapter 2 for the case without endogenous reproductive variance. For the coalescence to occur at time *t* in the past, one of the alleles (A) would descent from the other (B) or both would descent from the same parental allele at that time (this is contrary to the case of non-overlapping generations, in which coalescence necessarily occurs when both alleles descent from the same parental allele in the previous generation). Specifically, if allele B is in an individual of age *a* and relative reproductive success $\vec{r}$ at time *t* (with probability $\epsilon(a, \vec{r})$), then allele A must be in a newborn at time *t*-1 (with probability ϵ_1) having descended from the same individual carrying allele B (with probability $p_a \cdot \frac{r_a}{M_a}$). Summing over the individual's possible ages and reproductive success vectors, we obtain the probability

$$\sum_{a,\vec{r}} \epsilon(a, \vec{r}) \cdot \epsilon_1 \cdot p_a \cdot \frac{r_a}{M_a} = \frac{1}{G^2} \sum_a \frac{\sum_{j \geq a} p_a p_j \sum_{\vec{r}} r_a g_j(\vec{r})}{M_a} = \frac{1}{G^2} \sum_a \frac{\sum_{j \geq a} p_a p_j W_{a,j}}{M_a}, \quad \text{(A2.11)}$$

where for $j \geq i$,

$$W_{i,j} = \sum_{\vec{r}} r_i g_j(\vec{r}) = \sum_{\vec{r}} r_i r_j f_j(\vec{r}) = E_{\vec{r} \sim f_j}(r_i \cdot r_j) \quad \text{(A2.12)}$$

is the expectation of $(r_i \cdot r_j)$ conditional on surviving to age $\geq j$. Further allowing for either allele or both to be the newborn (using the inclusion-exclusion principal to subtract the probability $\epsilon_1^2 \sum_a \frac{p_a^2 W_{a,a}}{M_a}$ that both alleles were in a newborn prior to coalescence), and measuring the coalescence rate in generations (rather than years), we obtain the per-generation coalescence rate and corresponding effective population size:

$$\frac{1}{N_e} = \frac{1}{G} \sum_a \frac{p_a^2 W_{a,a} + 2 \sum_{j>a} p_a p_j W_{a,j}}{M_a}. \quad \text{(A2.13)}$$

Eq. A2.13 can be rearranged to obtain Eq. 2.7 of chapter 2. To this end, we define

$$w_i = (p_i^2 W_{i,i} + 2 \sum_{j>i} p_i p_j W_{i,j})/W, \quad \text{(A2.14)}$$

where

$$W = \sum_i p_i^2 W_{i,i} + 2 \sum_{i<j} p_i p_j W_{i,j} \quad \text{(A2.15)}$$

is a weighted average of the $W_{i,j}$. Noting that $\sum_a w_a = 1$, we then define the effective age class size as a weighted harmonic average of the age class sizes,

$$\frac{1}{M} = \sum_a \frac{w_a}{M_a}. \quad \text{(A2.16)}$$

Substituting this expression into Eq. A2.13, we obtain Eq. 2.7 of chapter 2:

$$\frac{1}{N_e} = W/(M \cdot G). \quad \text{(A2.17)}$$

A2.1.3 Reproductive variance

To recast our results for N_e in terms of the total reproductive variance V, we first consider the case with non-overlapping generations in a haploid population of constant size, i.e., with Wright-Fisher sampling. We denote the number of offspring of the i^{th} individual by k_i and the census size by N. The expected number of offspring is 1, i.e., $\frac{1}{N}\sum_i k_i = 1$, and we denote the variance in number of offspring, which we also refer to as the reproductive variance, by $V = \frac{1}{N}\sum_i (k_i - 1)^2$. In the standard neutral model, without endogenous reproductive variance, $V = 1$. Since the probability that two distinct gametes descend from the same ancestor in the previous generation is $\sum_i \frac{k_i}{N} \cdot \frac{k_i - 1}{N-1} = \frac{V}{N-1}$, we find that the effective population size is

$$N_e = \frac{N-1}{V} \cong \frac{N}{V}, \tag{A2.18}$$

which is the expression derived by Wright (4) and presented in Eq. 2.9 of chapter 2.

To extend Eq. A2.18 to the case with overlapping generations, we consider the first two moments of an individual's number of offspring, X, throughout its lifetime. First, we note that an individual's number of offspring can be expressed as a sum over the number at each age, i.e., $X = \sum_a X_a$, where X_a is the number of offspring at age a; $X_a = 0$ if the individual does not survive to that age. In these terms, the first two moments are

$$E(X) = \sum_a E(X_a) \text{ and } E(X^2) = \sum_a E(X_a^2) + 2\sum_{i<j} E(X_i \cdot X_j). \tag{A2.19}$$

Denoting the event of surviving to age $\geq a$ by S_a, we note that

$$E\left(X_a^i\right) = Pr(S_a) \cdot E\left(X_a^i \middle| S_a\right) = \frac{M_a}{M_1} \cdot E\left(X_a^i \middle| S_a\right), \tag{A2.20}$$

The latter term, $E\left(X_a^i \middle| S_a\right)$, can be simplified further by conditioning on $\vec{r}$. Since the probability mass function of $\vec{r}$ conditional on S_a is f_a,

$$E\left(X_a^i \middle| S_a\right) = E_{\vec{r}\sim f_a} E\left(X_a^i \middle| S_a, \vec{r}\right). \quad \text{(A2.21)}$$

Moreover, the distribution of X_a conditional on S_a and $\vec{r}$ is simply $(X_a|\vec{r}, S_a) \sim Bin(M_1, p_a \cdot r_a/M_a)$, and therefore

$$E(X_a|S_a) = E_{\vec{r}\sim f_a}\left(\frac{M_1 r_a p_a}{M_a}\right) = \frac{M_1 p_a}{M_a}$$

and

$$E(X_a^2|S_a) = E_{\vec{r}\sim f_a}\left(M_1 \frac{r_a p_a}{M_a} + 2\binom{M_1}{2}\left(\frac{r_a p_a}{M_a}\right)^2\right) = \frac{M_1 p_a}{M_a} + 2\binom{M_1}{2}\left(\frac{p_a}{M_a}\right)^2 W_{a,a}. \quad \text{(A2.22)}$$

Substituting these expression into Eq. A2.20, we find that

$$E(X_a) = p_a \text{ and } E(X_a^2) = p_a + \frac{M_1 - 1}{M_a} p_a^2 \cdot W_{a,a}. \quad \text{(A2.23)}$$

To calculate the remaining terms in Eq. A2.19, $E(X_i \cdot X_j)$ for $j > i$, we note that conditioning on S_j, and on $\vec{r}|S_j$,

$$E\left(X_i \cdot X_j\right) = Pr\left(S_j\right) \cdot E\left(X_i \cdot X_j \middle| S_j\right) = \frac{M_j}{M_1} \cdot E_{\vec{r}\sim f_j} E\left(X_i \cdot X_j \middle| S_j, \vec{r}\right). \quad \text{(A2.24)}$$

The latter term is easily calculated, since conditional on S_j and $\vec{r}$, X_i and X_j are independent binomial variables, with $\left(X_i \middle| \vec{r}, S_j\right) \sim Bin(M_1, p_i \cdot r_i/M_i)$ and $\left(X_j \middle| \vec{r}, S_j\right) \sim Bin\left(M_1, p_j \cdot r_j/M_j\right)$, yielding

$$E\left(X_i \cdot X_j\right) = \frac{M_j}{M_1} \cdot E_{\vec{r}\sim f_j}\left(\frac{M_1^2 p_i p_j r_i r_j}{M_i M_j}\right) = \frac{M_1 p_i p_j W_{i,j}}{M_i}. \quad \text{(A2.25)}$$

Substituting the expressions from Eqs. A2.23 and A2.25 into Eq. A2.19 we find that

$$E(X) = 1 \text{ and } E(X^2) = 1 + M_1 \sum_i \frac{p_i^2 \cdot W_{i,i} + 2\sum_{j>i} p_i p_j W_{i,j}}{M_i} - \sum_i \frac{p_i^2 \cdot W_{i,i}}{M_i}. \quad \text{(A2.26)}$$

Assuming that the total population size is sufficiently large for the ratios M_i/M_j and terms $W_{i,j}$ to be approximated as fixed, and for the higher order term $\sum_i \frac{p_i^2 \cdot W_{i,i}}{M_i}$ to be negligible, we find that

$$E(X) = 1 \text{ and } E(X^2) \cong 1 + \frac{M_1}{M} W, \tag{A2.27}$$

and therefore the total reproductive variance is

$$V = E(X^2) - E^2(X) = \frac{M_1}{M} W, \tag{A2.28}$$

which is Eq. 2.11 of chapter 2. These assumptions correspond to the standard practice of neglecting higher order terms in $1/N$ in models with non-overlapping generations. From Eqs. A2.17 and A2.28 we find that the effective population size is

$$N_e = (G \cdot M_1)/V, \tag{A2.29}$$

which is the same form as in the case without age-structure (Eq. A2.18), and the general form presented in Eq. 2.10 of chapter 2.

A2.1.4 Age-structure alone

Felsenstein used a different approach to solve the haploid model without endogenous reproductive variance, relying on the definition of the effective population size as the inbreeding effective number (2). To see that his results agree with ours (as well as with those of Sagitov and Jagers (3)), consider the case without endogenous reproductive variance, where Eq. A2.17 reduces to

$$N_e = MG = \frac{G}{\sum_i \frac{p_i^2 + 2\sum_{j>i} p_i p_j}{M_i}} = \frac{G}{\sum_i \frac{p_i}{M_i}(q_i + q_{i+1})}, \tag{A2.30}$$

where $q_i = \sum_{j \geq i} p_i$. Noting that $p_i(q_i + q_{i+1}) = (q_i - q_{i+1})(q_i + q_{i+1}) = q_i^2 - q_{i+1}^2$, we find that

$$N_e = \frac{G}{\sum_i \frac{p_i}{M_i}(q_i + q_{i+1})} = \frac{GM_1}{\sum_i \frac{M_1}{M_i}(q_i^2 - q_{i+1}^2)} = \frac{GM_1}{1 + \sum_i q_{i+1}^2 (\frac{M_1}{M_{i+1}} - \frac{M_1}{M_i})}, \tag{A2.31}$$

where Felsenstein's functional form (p. 585 in (2)) is on the rightmost side.

To better understand the effect of age-structure on the effective population size, consider a simple example in which there is no endogenous reproductive variance, and no age-dependence in

reproductive success. In other words, the only difference among individuals' numbers of offspring arise from the stochasticity of mortality and reproduction. In this case, the probability of having a parent of age a is proportional to the size of the age class, i.e., $p_a = M_a/N$ where $N = \sum_a M_a$ is the census size. Following our derivations, the effective population size (Eq. 2.3 in chapter 2) then reduces to $N_e = \frac{G}{(2G-1)} N$, and if the generation time $G \gg 1$ then $N_e \approx \frac{1}{2} N$. In other words, the age structure reduces the effective population size to half of the census size

A2.2. Diploid Model

A2.2.1 Overview

Here we rigorously define and solve the diploid model with two sexes and endogenous reproductive variance, and derive formulas for the effective population sizes of X and autosomes. While the diploid model is more elaborate, the model and results follow along the same lines as we described for the haploid model. In Section A2.2.2 we detail the assumptions of the diploid model and introduce the notation required for the derivations that follow. In Section A2.2.3 we solve for the joint stationary distribution of the age and relative reproductive success of autosomal and X-linked alleles. We build on the joint stationary distribution to solving for the stationary per-generation coalescence rates and corresponding effective population sizes on X and autosomes. Since some of the explicit equations we derive are not presented in the chapter 2, we briefly review them here.

Notably, to extend the haploid formula for the effective population size, $N_e = MG/W$ (Eqs. 2.7 and A2.17), to the diploid case, we require explicit expressions for the effective age-class size M, generation time G, and W, corresponding to the X and autosomes. First, we define these measures

for each sex in the same way that we did in the haploid model (i.e., as in Eqs. A2.15 and A2.16). We then define G and W for X and autosomes, as simple weighted averages over their values in males and females:

$$G_X = \frac{2}{3}G_F + \frac{1}{3}G_M \text{ and } G_A = \frac{1}{2}(G_M + G_F) \tag{A2.32}$$

(which is Eq. 2.12 in chapter 2), and

$$W_X = \frac{2}{3}W_F + \frac{1}{3}W_M \text{ and } W_A = \frac{1}{2}(W_M + W_F), \tag{A2.33}$$

where the weights reflect the relative number of generations that X and autosomal linked loci spend in males and females (see Table A2.2 for notation). The effective age class sizes on X and autosomes are defined as weighted harmonic averages. In the case without endogenous reproductive variance, they are defined as

$$\frac{1}{M_X} = \frac{1/3}{M_M} + \frac{2/3}{M_F} \text{ and } \frac{1}{M_A} = \frac{1/2}{M_M} + \frac{1/2}{M_F}. \tag{A2.34}$$

To account for sex-specific endogenous reproductive variances, the weights further account for the endogenous reproductive variances effect on the relative probability of coalescence in males and females,

$$\frac{1}{M_X} = \frac{1/3(W_M/W_X)}{M_M} + \frac{2/3(W_F/W_X)}{M_F} \text{ and } \frac{1}{M_A} = \frac{1/2(W_M/W_X)}{M_M} + \frac{1/2(W_F/W_X)}{M_F}. \tag{A2.35}$$

Using these definitions, the effective population size for the X and autosomes take the form

$$N_e^A = \frac{2G_A M_A}{W_A} \text{ and } N_e^A = \frac{2G_X M_X}{W_X}, \tag{A2.36}$$

where the factor 2, which is absent in the haploid case (Eqs. 2.7 and A2.17), accounts for the effective number of age classes in the population (i.e., G classes in the haploid population, but $2G$ classes in the case with two sexes). To translate these effective sizes into coalescence rates, we also account for ploidy, yielding per generation rates of $1/2N_e^A = W_A/4G_A M_A$ on autosomes and $1/(3/2)N_e^X = W_X/3G_X M_X$ on the X. Based on Eq. A2.36, the mutation rate on X and autosomes,

and the standard forms for polymorphism levels, we obtain the following expression for the ratio of X to autosome polymorphism levels:

$$\frac{E(\pi_X)}{E(\pi_A)} = \frac{3}{4} \cdot \frac{f\left(\frac{\mu_M}{\mu_F}\right) \cdot f\left(\frac{G_M}{G_F}\right)}{f\left(\frac{W_M/W_F}{M_M/M_F}\right)} \tag{A2.37}$$

(note that this differs from Eq. 2.18 of chapter 2).

In Section A2.2.4, we recast the results for the effective population size (Eq. A2.36) and polymorphism ratio (Eq. A2.37) in terms of male and female reproductive variances. First, we show that the reproductive variances in males and females, V_M and V_F, are given by

$$V_s = \frac{M_1}{\gamma_s}\frac{W_s}{M_s} - \frac{1-\gamma_s}{\gamma_s^2}, \tag{A2.38}$$

where the index *s* corresponds to *M* or *F*, and γ_M and γ_F are the proportions of males and females among newborns, respectively. Thus, this equation does not assume a sex ratio of 1. Rewriting Eq. A2.36 in terms of male and female reproductive variances we find that

$$N_e^X = \frac{4G_X M_1}{\frac{2}{3}\gamma_M V_M + \frac{4}{3}\gamma_F V_F + \frac{2}{3}\frac{\gamma_F}{\gamma_M} + \frac{4}{3}\frac{\gamma_M}{\gamma_F}} \text{ and } N_e^A = \frac{4G_A M_1}{\gamma_M V_M + \gamma_F V_F + \frac{\gamma_F}{\gamma_M} + \frac{\gamma_M}{\gamma_F}}, \tag{A2.39}$$

where M_1 is the number of newborns of both sexes per year, and that

$$\frac{E(\pi_X)}{E(\pi_A)} = \frac{3}{4} \cdot \frac{f(\mu_M/\mu_F) \cdot f(G_M/G_F)}{f\left(\frac{\gamma_F/\gamma_M + \gamma_M V_M}{\gamma_M/\gamma_F + \gamma_F V_F}\right)}. \tag{A2.40}$$

These equations reduce to Eqs. 2.15 and 2.18 in chapter 2 when the sex ratio at birth equals 1.

In Section A2.2.5, we derive the expressions for the effective population sizes in terms of the allelic reproductive variances for X and autosomes, V_X^* and V_A^* (Eq. 2.13 in chapter 2):

$$N_e^A = \frac{G_A \cdot M_1}{V_A^*} \text{ and } N_e^X = \frac{G_X \cdot M_1}{V_X^*}, \tag{A2.41}$$

where the expression for the X only hold when the sex ratio at birth equals 1 (see below).

Notation	Definition	Remarks
$\boldsymbol{p_{s,a}}$	Probability that a parent of sex s is of age a	$\Sigma_a p_{F,a} = \Sigma_a p_{M,a} = 1$
$\boldsymbol{q_{s,a}}$	Probability that a parent of sex s is of age $\geq a$	$q_{s,a} = \Sigma_{i \geq a} p_{s,i}$
$\boldsymbol{G_M, G_F}$	Male and female generation times	$G_s = \Sigma_a a \cdot p_{s,a}$
$\boldsymbol{G_X, G_A}$	Generation times for X and autosomes	See Eq. A2.32
$\boldsymbol{M_{s,a}}$	Number of individuals of sex s and age a	$M_{s,a+1} \leq M_{s,a}$
$\boldsymbol{M_1}$	Number of newborns of both sexes per-year	
$\boldsymbol{\gamma_M, \gamma_F}$	Proportions of males and females among newborns	$\gamma_s = M_{s,1}/M_1$
$\boldsymbol{\vec{r}}$	Relative reproductive success	
$\boldsymbol{f_{s,a}(\vec{r})}$	The proportion of individuals with relative reproductive success $\vec{r}$ among individuals of sex s and age a	
$\boldsymbol{g_{s,a}(\vec{r})}$	Given a newborn that descended from a parent of sex s and age a, $g_{s,a}(\vec{r})$ is the probability that the parent has relative reproductive success $\vec{r}$	$g_{s,a}(\vec{r}) = r_a f_{s,a}(\vec{r})$
$\boldsymbol{\epsilon^X(s,a,\vec{r})}$, $\boldsymbol{\epsilon^A(s,a,\vec{r})}$	Joint stationary probability of sex s, age a and relative reproductive success $\vec{r}$ for the X and autosomes	See Eqs. A2.49 and A2.50
$\boldsymbol{\epsilon^X_{s,a}, \epsilon^A_{s,a}}$	Marginal stationary distribution of age a for the X and autosomes	
$\boldsymbol{M_M, M_F}$	Effective male and female age-class sizes	See Eq. A2.61
$\boldsymbol{M_X, M_A}$	Effective X and autosome linked age-class sizes	See Eq. A2.65
$\boldsymbol{W_{s,i,j}}$	Expectation of $r_i \cdot r_j$ among individuals of sex conditional of surviving to age $a \geq j$	Defined for $j \geq i$
$\boldsymbol{W_M, W_F}$	Weighted averages of the $W_{M,i,j}$ and the $W_{F,i,j}$, respectively	See Eq. A2.60
$\boldsymbol{W_X, W_A}$	Weighted averages of W_M and W_F for X and autosome linked loci	See Eq. A2.33
$\boldsymbol{X_s, X_{s,a}}$	$X_{s,a}$ and X_s are Random variables describing the numbers of offspring an individual has at age a or throughout life, respectively	
$\boldsymbol{V_M, V_F}$	Male and female reproductive variances (i.e., $V_s = V(X_s)$)	See Eq. A2.83
$\boldsymbol{S_{s,a}}$	The event of a newborn of sex s surviving to age $\geq a$	
$\boldsymbol{f(x)}$	$f(x) \equiv (2x+4)/(3x+3)$	
$\boldsymbol{\mu_M, \mu_F}$	Male and female expected mutation rates per generation	See Section A2.3
$\boldsymbol{\mu_X, \mu_A}$	Expected mutation rates per generation on X and autosomes; $\mu_X = \frac{1}{3}\mu_M + \frac{2}{3}\mu_F$ and $\mu_A = \frac{1}{2}\mu_M + \frac{1}{2}\mu_F$	
$\boldsymbol{X^X_m, X^A_m}$	The number of newborns carrying a random X or autosome linked allele m	
$\boldsymbol{V^*_X, V^*_A}$	Reproductive variances of X and autosome linked alleles, respectively (i.e., $V^*_X \equiv V(X^X_m)$ and $V^*_A \equiv V(X^A_m)$)	See Eqs A2.89 and A2.93

Table A2.2: Notation for the diploid model with two sexes, with parameters of the model in red.

A2.2.2 Assumptions and notation

We consider a panmictic, diploid population of constant size, with two sexes, and sex- and age-dependent mortalities, fecundities and reproductive variances. We measure age in years, and assume that the number of individuals of sex *s* and age *a*, $M_{s,a}$, is constant. Specifically, the sizes of the newborn age classes, $M_{M,a}$ and $M_{F,a}$, may take any integer values, meaning that at this point we make do not assume that the sex-ratios at birth equals 1. More generally, the size of classes can vary between sexes, but for each sex they decrease with age, i. e. , $M_{s,a+1} \leq M_{s,a}$, reflecting sex- and age-specific mortalities. We further assume that age classes are partitioned according to individuals' age-dependent reproductive success. Namely, individuals are randomly assigned a vector $\vec{r}$ at birth, reflecting their expected relative reproductive success at each age (see below). We then assume that the number of individuals in the population of sex *s*, age *a* and relative reproductive success $\vec{r}$, is constant and equal to $M_{s,a} \cdot f_{s,a}(\vec{r})$, where $f_{s,a}$ is the probability mass function of $\vec{r}$ among individuals of sex *s* and age *a*. Individuals with the same value of $\vec{r}$ are chosen to survive to the next age class at random, i.e., there are no differences in viability, but $M_{s,a} \cdot f_{s,a}(\vec{r}) \geq M_{s,a+1} \cdot f_{s,a+1}(\vec{r})$ due to mortality, where rates of mortality can depend on the value of $\vec{r}$.

Sex and age dependent fertility and reproductive success are described backwards in time, in terms of the probability of an individual being chosen as a parent. Every newborn has a mother and a father, which are chosen independently. The probability that the parent of sex *s* is of a given age is described by a discrete distribution $A_s = (p_{s,a})_{a=1}^{\infty}$, where the expectations $G_M = E(A_M)$ and $G_F = E(A_F)$ are the generation times for males and females, respectively. The average probability per individual of age *a* is therefore $p_{s,a}/M_{s,a}$, which can be viewed as the fertility associated with

that age and sex. The probability of being born to a specific parent of age a and relative reproductive success $\vec{r}$ is $p_{s,a} \cdot \frac{r_a}{M_{s,a}}$, where r_a is the a-th component of $\vec{r}$. The value of r_a thus reflects an individual's expected (rather than actual) relative reproductive success.

Similar to the haploid case (cf. Section A2.1.1), our assumptions imply several requirements on the form of the probability mass functions $f_{s,a}$. First, requiring that the probability of a parent of sex s being of age a is $p_{s,a}$, implies that for any sex s and age a, $E_{f_{s,a}}(r_a) = 1$. Second, requiring that $M_{s,a} \cdot f_{s,a}(\vec{r}) \geq M_{s,a+1} \cdot f_{s,a+1}(\vec{r})$ implies that $f_{s,a}(\vec{r})/f_{s,a+1}(\vec{r}) \geq M_{s,a+1}/M_{s,a}$. Third, requiring that for any sex s and age a $M_{s,a} \cdot f_{s,a}(\vec{r})$ is an integer, implies that the probability mass functions $f_{s,a}$ are discrete and can only take values $i/M_{s,a}$ for $i = 0, 1, \ldots, M_{s,a}$. While the latter requirement may appear to be highly restrictive, if we fix the ratios $M_{s,a}/M_{s',a}$ and assume that the total population size is sufficiently large, we can relax this requirement and assume any continuous or discrete distributions $f_{s,a}$ that satisfy the first two requirements (by the same reasoning we applied to the haploid case in Section A2.1.2).

A2.2.3 Solution backwards in time

Here, we extend the derivations of Pollak (5) to account for endogenous reproductive variance. Tracing an allele backward in time, the sex s_t, age a_t and relative reproductive success $\vec{r}_t$ of the individual I_t carrying the allele t years in the past defines a Markov chain, $(s_t, a_t, \vec{r}_t)$. To define the transition probabilities of the chain, we distinguish between two cases. First, if the allele is not carried by a newborn, i.e., if $a_t > 1$, then at time $t+1$ the individual carrying it was one year younger, and its sex s and relative reproductive success $\vec{r}$ remain unchanged, i.e., $(s_{t+1}, a_{t+1}, \vec{r}_{t+1}) = (s_t, a_t - 1, \vec{r}_t)$ with probability one. Second, if the allele is carried by a

newborn, i.e., if $a_t = 1$, then the sex of the parent, s_{t+1}, is equally likely to be male or female if the allele is autosomal or if it is X-linked and the newborn was a female, but if it is X-linked and the newborn was a male then the sex of the parent will be female with probability 1. Conditional on the parent's sex, s_{t+1}, its age $a_{t+1} = a$ with probability $p_{s_{t+1},a}$. The probability mass function of $\vec{r}_{t+1}$ conditional on (s_{t+1}, a_{t+1}), follows from Bayes' theorem, further conditioning on the fact that the parent, $I_{t+1} = I$, necessarily reproduced successfully

$$P(\vec{r}_I = \vec{r} | I_{t+1} = I, a_{t+1} = a)$$

$$= \frac{P(I_{t+1}=I|\vec{r}_{t+1}=\vec{r},a_{t+1}=a)\cdot P(\vec{r}_I=\vec{r}|\, a_{t+1}=a)}{P(I_{t+1}=I)} = \frac{(r_a/M_{s,a})\cdot f_{s,a}(\vec{r})}{\sum_{\vec{k}}(r_a/M_{s,a})\cdot f_{s,a}(\vec{k})} = r_a \cdot f_{s,a}(\vec{r}) \qquad \text{(A2.42)}$$

We denote this probability by $g_{s,a}(\vec{r}) \equiv r_a \cdot f_{s,a}(\vec{r})$, and conclude that when $a_t = 1$, s_{t+1} is distributed as we described above and $P\big((a_{t+1}, \vec{r}_{t+1}) = (a, \vec{r})\big|s_{t+1}\big) = p_{s,a} \cdot g_{s,a}(\vec{r})$.

$g_{s,a}$ is a proper probability mass function since $\sum_{\vec{r}} g_{s,a}(\vec{r}) = \sum_{\vec{r}} r_a \cdot f_{s,a}(\vec{r}) = 1$. Moreover, the parent's expected value of r_a is $E_{\vec{r}\sim g_{s,a}}(r_a) = E_{\vec{r}\sim f_{s,a}}(r_a^2) = 1 + V_{\vec{r}\sim f_{s,a}}(r_a) \geq 1$. The latter inequality makes intuitive sense, as it implies that the allele is more likely to be descended from an individual that has higher than average relative reproductive success in its age class.

We rely on the transition probabilities to derive and solve recursions for the stationary probabilities, $\epsilon_A(s, a, \vec{r})$ and $\epsilon_X(s, a, \vec{r})$, of sex, *s*, age, *a*, and relative reproductive successes, $\vec{r}$, of autosome and X linked alleles, respectively. For autosomal alleles

$$\epsilon^A(s, a, \vec{r}) = \epsilon^A(s, a+1, \vec{r}) + \Big(\sum_{t,\vec{k}} \epsilon^A\big(t, 1, \vec{k}\big)\Big) \cdot \frac{1}{2} p_{s,a} \cdot g_{s,a}(\vec{r}), \qquad \text{(A2.43)}$$

where the first term corresponds to aging by one year and the second term corresponds to parenting a newborn. For X linked alleles

$$\epsilon^X(s,a,\vec{r}) = \epsilon^X(s,a+1,\vec{r}) + \left(\frac{1}{2}\sum_{\vec{k}}\epsilon^X(F,1,\vec{k}) + \mathbb{I}_{s=F}\sum_{\vec{k}}\epsilon^X(M,1,\vec{k})\right)\cdot p_{s,a}\cdot g_{s,a}(\vec{r}),$$

(A2.44)

where $\mathbb{I}$ denotes an indicator function, and, similar to the autosomal case, the first term corresponds to aging by one year and the second term corresponds to parenting a newborn.

In order to solve these recursions, we first consider the marginal stationary distribution of age and sex, $\epsilon^A_{s,a} = \sum_{\vec{r}}\epsilon^A(s,a,\vec{r})$ for autosomes and $\epsilon^X_{s,a} = \sum_{\vec{r}}\epsilon^X(s,a,\vec{r})$. To this end, we sum the recursions over $\vec{r}$ to obtain recursions on the marginal distributions,

$$\epsilon^A_{s,a} = \epsilon^A_{s,a+1} + \left(\epsilon^A_{M,1} + \epsilon^A_{F,1}\right)\cdot\frac{1}{2}p_{s,a} \text{ and } \epsilon^X_{s,a} = \epsilon^X_{s,a+1} + \left(\frac{1}{2}\epsilon^X_{F,1} + \mathbb{I}_{s=F}\epsilon^X_{M,1}\right)\cdot p_{s,a},$$

(A2.45)

where we also require that $\sum_{s,a}\epsilon^A_{s,a} = \sum_{s,a}\epsilon^X_{s,a} = 1$. These recursions were solved by Pollack (5) for the case without endogenous reproductive variance, yielding

$$\epsilon^A_{s,a} = q_{s,a}/2G_A,\ \epsilon^X_{M,a} = q_{M,a}/3G_X \text{ and } \epsilon^X_{F,a} = 2q_{F,a}/3G_X, \quad \text{(A2.46)}$$

where $q_{s,j} \equiv \sum_{j\geq a} p_{s,j}$ is the probability that a parent of sex *s* is at least *j* years old. Substituting these expressions into Eqs. A2.43 and A2.44, the recursions simplify to

$$\epsilon^A(s,a,\vec{r}) = \epsilon^A(s,a+1,\vec{r}) + \frac{1}{2G_a}p_{s,a}\cdot g_{s,a}(\vec{r}) \quad \text{(A2.47)}$$

for autosomes and

$$\epsilon^X(s,a,\vec{r}) = \epsilon^X(s,a+1,\vec{r}) + \frac{1+\mathbb{I}_{s=F}}{3G_X}\cdot p_{s,a}\cdot g_{s,a}(\vec{r}) \quad \text{(A2.48)}$$

for the X, where we further require that $\sum_{s,a,\vec{r}}\epsilon^A(s,a,\vec{r}) = \sum_{s,a,\vec{r}}\epsilon^X(s,a,\vec{r}) = 1$. The solution to these recursions is

$$\epsilon^A(s,a,\vec{r}) = \frac{1}{2G_A}\epsilon(s,a,\vec{r}) \quad \text{(A2.49)}$$

for autosomes and

$$\epsilon^X(s,a,\vec{r}) = \frac{1+\mathbb{I}_{s=F}}{3G_X}\epsilon(s,a,\vec{r}) \quad \text{(A2.50)}$$

for the X, where $\epsilon(s,a,\vec{r}) \equiv \sum_{j\geq a} p_{s,j} \cdot g_{s,j}(\vec{r})$.

The marginal stationary probability mass function of $\vec{r}$ follows,

$$\epsilon^A_{\vec{r}} = \sum_{s,a} \epsilon^A(s,a,\vec{r}) = \sum_{s,j} \frac{j \cdot p_{s,j}}{2G_A} \cdot g_{s,j}(\vec{r}) \quad \text{(A2.51)}$$

for autosomes, and

$$\epsilon^X_{\vec{r}} = \sum_{s,a} \epsilon^X(s,a,\vec{r}) = \sum_{s,j} \frac{(1+\mathbb{I}_{s=F}) \cdot j \cdot p_{s,j}}{3G_X} \cdot g_{s,j}(\vec{r}) \quad \text{(A2.52)}$$

for the X. These are proper probability mass functions since they are weighted averages of the probability mass functions $g_{s,j}$, since $\sum_{s,j} \frac{j \cdot p_{s,j}}{2G_A} = \sum_{s,j} \frac{(1+\mathbb{I}_{s=F}) \cdot j \cdot p_{s,j}}{3G_X} = 1$.

Similar to the haploid case, we rely on the stationary distribution to derive the probability of coalescence of two alleles. Consider the autosomal case first. For coalescence to occur at time *t* in the past, one of the alleles (A) would descent from the other (B) or both would descent from the same parental allele at that time. Specifically, if allele B is in an individual of sex *s*, age *a* and relative reproductive success $\vec{r}$ at time *t* (with probability $\epsilon^A(s,a,\vec{r})$), then allele A must be in a newborn at time *t*-1 (with probability $\epsilon_{M,1} + \epsilon_{F,1}$), having descended from the same individual carrying allele B (with probability $\frac{1}{2} p_{s,a} \cdot \frac{r_a}{M_a}$) and from allele B specifically (with probability $1/2$). Summing over the individual's possible sexes, ages and reproductive success vectors, we obtain the probability

$$\sum_{s,a,\vec{r}} \epsilon^A(s,a,\vec{r}) \cdot \left(\epsilon^A_{M,1} + \epsilon^A_{F,1}\right) \cdot \frac{1}{2} p_{s,a} \cdot \frac{r_a}{2M_{s,a}} = \frac{1}{8(G_A)^2} \sum_{s,a} \frac{\sum_{j\geq a} p_{s,a} p_{s,j} \sum_{\vec{r}} r_a \cdot g_{s,j}(\vec{r})}{M_{s,a}}$$

$$= \frac{1}{8(G_A)^2} \sum_{s,a} \frac{\sum_{j\geq a} p_{s,a} p_{s,j} W_{s,a,j}}{M_{s,a}}, \quad \text{(A2.53)}$$

where, for $j \geq i$,

$$W_{s,i,j} \equiv E_{\vec{r}\sim f_{s,j}}\left(r_i \cdot r_j\right) = E_{\vec{r}\sim g_{s,j}}(r_i) = \sum_{\vec{r}} r_i \cdot g_{s,j}(\vec{r}) \tag{A2.54}$$

is the expectation of $\left(r_i \cdot r_j\right)$ over individuals of sex s and age j. Further allowing for either allele or both to be the newborn (using the inclusion-exclusion principal to subtract the probability

$$\left(\epsilon^A_{M,1} + \epsilon^A_{F,1}\right)^2 \sum_{s,a,\vec{r}} \left(\frac{1}{2} p_{s,a}\right)^2 \cdot \frac{r_a g_{s,a}(\vec{r})}{2M_{s,a}} = \frac{1}{8(G_A)^2} \sum_{s,a} \frac{p_{s,a} p_{s,a} W_{s,a,a}}{M_{s,a}} \tag{A2.55}$$

that both alleles were in a newborn prior to coalescence), the autosomal stationary coalescence rate per year is

$$\frac{1}{8(G_A)^2} \sum_{s,a} \frac{p^2_{s,a} W_{s,a,a} + 2\sum_{j>a} p_{s,a} p_{s,j} W_{s,a,j}}{M_{s,a}}. \tag{A2.56}$$

The per generation coalescence rate (in terms of the autosomal generation time G_A) and corresponding effective population size are therefore

$$\frac{1}{2N^A_e} = \frac{1}{8 \cdot G_A} \sum_{s,a} \frac{p^2_{s,a} W_{s,a,a} + 2\sum_{j>a} p_{s,a} p_{s,j} W_{s,a,j}}{M_{s,a}}. \tag{A2.57}$$

For the X, the stationary coalescence rate per year is

$$\begin{aligned} &2\left(\epsilon^X_{M,1} + \frac{1}{2}\epsilon^X_{F,a}\right) \sum_{\mathrm{a},\vec{r}} \epsilon^X(F, a, \vec{r}) \cdot p_{F,a} \cdot \frac{r_a}{2M_{F,a}} + 2 \cdot \frac{1}{2}\epsilon^X_{F,a} \sum_{\mathrm{a},\vec{r}} \epsilon^X(M, a, \vec{r}) \cdot p_{M,a} \cdot \frac{r_a}{M_{M,a}} \\ &- \left(\epsilon^X_{M,1} + \frac{1}{2}\epsilon^X_{F,1}\right)^2 \sum_{a,\vec{r}} p^2_{F,a} \cdot \frac{r_a g_{F,a}(\vec{r})}{2M_{F,a}} - \left(\frac{1}{2}\epsilon^X_{F,1}\right)^2 \sum_{a,\vec{r}} p^2_{M,a} \cdot \frac{r_a g_{M,a}(\vec{r})}{M_{M,a}} \\ &= \frac{1}{9(G_X)^2} \sum_s (1 + \mathbb{I}_{s=F}) \sum_a \frac{p^2_{s,a} W_{s,a,a} + 2\sum_{j>a} p_{s,a} p_{s,j} W_{s,a,j}}{M_{s,a}}, \end{aligned} \tag{A2.58}$$

and the corresponding per generation coalescence rate, which defines the effective population size for the X, N^X_e, is

$$\frac{1}{(3/2)N^X_e} = \frac{1}{9G_X} \sum_s (1 + \mathbb{I}_{s=F}) \sum_a \frac{p^2_{s,a} W_{s,a,a} + 2\sum_{j>a} p_{s,a} p_{s,j} W_{s,a,j}}{M_{s,a}} \tag{A2.59}$$

(defined in terms of the X-linked generation time G_X).

As outlined in Section A2.2.1, the effective population sizes, N_e^X and N_e^A, can be rewritten in terms of the effective age class sizes, to obtain expressions that are analogous to Eqs. 2.7 and A2.17 in the haploid case. To this end, the terms G, W and M in Eq. A2.17 need to be defined for the X and autosomes. First, we define these terms separately for males and females, by applying the haploid definitions. Specifically, we define

$$W_s = \sum_i p_{s,i}^2 W_{s,i,i} + 2\sum_{i<j} p_{s,i} p_{s,j} W_{s,i,j} \quad \text{(A2.60)}$$

as a weighted average of the $W_{s,i,j}$, and define

$$\frac{1}{M_s} = \sum_a \frac{w_{s,a}}{M_{s,a}} \quad \text{(A2.61)}$$

as a weighted harmonic average of the age classes sizes of sex *s*, with weights

$$w_{s,i} = (p_{s,i}^2 W_{s,i,i} + 2\sum_{j>i} p_{s,i} p_{s,j} W_{s,i,j})/W_s \quad \text{(A2.62)}$$

(note that $\sum_a w_{s,a} = 1$).

To extend the definitions of G, W and M to the X and autosomes, we define them as weighted averages over males and females. Specifically, *G* and *W* are defined as simple weighted averages,

$$G_A = \frac{1}{2}(G_M + G_F) \text{ and } G_X = \frac{2}{3}G_F + \frac{1}{3}G_M \quad \text{(A2.63)}$$

(this is Eq. 2.12 in chapter 2) and

$$W_A = \frac{1}{2}(W_M + W_F) \text{ and } W_X = \frac{2}{3}W_F + \frac{1}{3}W_M. \quad \text{(A2.64)}$$

The effective age class size *M* for X and autosomes is defined as a weighted harmonic average,

$$\frac{1}{\mathrm{M_A}} = \frac{1/2(\mathrm{W_M}/\mathrm{W_A})}{\mathrm{M_M}} + \frac{1/2(\mathrm{W_F}/\mathrm{W_A})}{\mathrm{M_F}} \text{ and } \frac{1}{\mathrm{M_X}} = \frac{1/3(\mathrm{W_M}/\mathrm{W_X})}{\mathrm{M_M}} + \frac{2/3(\mathrm{W_F}/\mathrm{W_X})}{\mathrm{M_F}}. \quad \text{(A2.65)}$$

Expressing Eqs. A2.57 and A2.59 in these terms, we find that

$$N_e^A = \frac{2M_A G_A}{W_A} \text{ and } N_e^X = \frac{2M_X G_X}{W_X}. \quad \text{(A2.66)}$$

The factor 2, which is absent in the haploid case (Eq. A2.17), reflects the effective number of age classes (i.e., G classes of size M in the haploid model, but $2G$ classes in the diploid model with two sexes).

Assuming the standard expressions for neutral heterozygosity, $E(\pi_A) = 4N_e^A\mu_A$ and $E(\pi_X) = 3N_e^X\mu_X$ (see Section A2.3), and rearranging the expressions in Eq. A2.66, we find that

$$\frac{E(\pi_X)}{E(\pi_A)} = \frac{3}{4} \cdot \frac{f(\mu_M/\mu_F)\cdot f(G_M/G_F)}{f\left(\frac{W_M/W_F}{M_M/M_F}\right)}. \qquad \text{(A2.67)}$$

When the mutation rate, age structure, and endogenous reproductive variance are identical in both sexes Eq. A2.67 reduces to the naïve neutral expectation of ¾. When these factors differ among sexes, Eq. A2.67 provides a simple expression for the effect of each factor.

A2.2.4 Reproductive variance

To recast our results for the effective population sizes in terms of total reproductive variances in males and females, V_M and V_F, we follow the same steps as described for the haploid case (Section A2.1.3). First, we consider the case with non-overlapping generations in a diploid population of constant size, with N_M males and N_F females. We denote the total population size by $N \equiv N_M + N_F$, the proportions of males and females by $\gamma_s \equiv \frac{N_s}{N}$, and the number of offspring of the i^{th} individual of sex s by k_i^s. To maintain a constant population size, we require that the number of offspring arising from parents of each sex equals N, and therefore the sex-specific expectations are $E(k_i^s) = \frac{1}{N_s}\sum_i k_i^s = \frac{N}{N_s}$. We denote the sex-specific variances by $V_s \equiv V(k_i^s)$.

We are interested in the probability that two distinct alleles descend from the same allele in the previous generation, as this probability equals $1/2N_e^A$ for autosomes and $1/(3/2)N_e^X$ for the X.

For autosomes, the probability that the two alleles descend from individuals of sex *s* is ¼, the probability that they descend from the same individual of that sex is $\sum_{i=1}^{N_s} \frac{k_i^s}{N} \cdot \frac{k_i^s-1}{N-1}$, and the probability that they descend from the same allele is 1/2 , therefore

$$\frac{1}{2N_e^A} = \frac{1}{8}\sum_s \sum_{i=1}^{N_s} \frac{k_i^s}{N} \cdot \frac{k_i^s-1}{N-1}. \quad \text{(A2.68)}$$

Substituting $\sum_{i=1}^{N_s} \frac{k_i^s}{N} \cdot \frac{k_i^s-1}{N-1} = \frac{\gamma_s}{N-1}\left(E\left(k_i^{s\,2}\right) - E(k_i^s)\right) = \frac{1}{N-1}(\gamma_s V_s + \frac{1}{\gamma_s} - 1)$ into Eq. A2.68, we find that

$$N_e^A = \frac{4(N-1)}{\gamma_M V_M + \gamma_F V_F + \frac{\gamma_F}{\gamma_M} + \frac{\gamma_M}{\gamma_F}} \cong \frac{4N}{\gamma_M V_M + \gamma_F V_F + \frac{\gamma_F}{\gamma_M} + \frac{\gamma_M}{\gamma_F}} \quad \text{(A2.69)}$$

(cf. (4)).

For the X chromosome, the probability that two alleles descend from individuals of sex *s* depends on γ_M and γ_F. However, as we go further backwards in time, this probability approaches 1/9 for both being male and 4/9 for both being female, regardless of γ_M and γ_F. The probability that both alleles descend from the same individual of that sex is $\sum_{i=1}^{N_s} \frac{k_i^s}{N} \cdot \frac{k_i^s-1}{N-1}$, and the probability that they descend from the same allele is ½ for females and 1 for males, and therefore

$$\frac{1}{(3/2)N_e^X} = \frac{1}{9}\sum_{i=1}^{N_M} \frac{k_i^M}{N} \cdot \frac{k_i^M-1}{N-1} + \frac{1}{2} \cdot \frac{4}{9}\sum_{i=1}^{N_F} \frac{k_i^F}{N} \cdot \frac{k_i^F-1}{N-1}, \quad \text{(A2.70)}$$

and thus

$$N_e^X = \frac{4(N-1)}{\frac{2}{3}\gamma_M V_M + \frac{4}{3}\gamma_F V_F + \frac{2}{3}\frac{\gamma_F}{\gamma_M} + \frac{4}{3}\frac{\gamma_M}{\gamma_F}} \cong \frac{4N}{\frac{2}{3}\gamma_M V_M + \frac{4}{3}\gamma_F V_F + \frac{2}{3}\frac{\gamma_F}{\gamma_M} + \frac{4}{3}\frac{\gamma_M}{\gamma_F}}. \quad \text{(A2.71)}$$

Assuming a sex ratio of 1 (i.e., $\gamma_M = \gamma_F = 1/2$), Eqs. A2.69 and A2.71 reduce to

$$N_e^A = \frac{4N}{2 + \frac{1}{2}V_M + \frac{1}{2}V_F} \text{ and } N_e^X = \frac{4N}{2 + \frac{1}{3}V_M + \frac{2}{3}V_F}. \quad \text{(A2.72)}$$

To extend Eq. A2.72 to the case with overlapping generations, we consider the first two moments of an individual's number of offspring, X_s, throughout its lifetime. First, we note that an

individual's number of offspring can be expressed as a sum over the number at each age, i.e., $X_s = \sum_a X_{s,a}$, where $X_{s,a}$ denotes the number of offspring at age a; and $X_{s,a} = 0$ if the individual does not survive to age a. In these terms, the first two moments are

$$E(X_s) = \sum_a E(X_{s,a}) \text{ and } E(X_s^2) = \sum_a E(X_{s,a}^2) + 2\sum_{j>i} E(X_{s,i} \cdot X_{s,j}). \quad \text{(A2.73)}$$

Denoting the event of surviving to age $\geq a$ by $S_{s,a}$, we note that

$$E\left(X_{s,a}^i\right) = Pr\left(S_{s,a}\right) \cdot E\left(X_{s,a}^i \middle| S_{s,a}\right) = \frac{M_{s,a}}{M_{s,1}} \cdot E\left(X_{s,a}^i \middle| S_{s,a}\right). \quad \text{(A2.74)}$$

The latter term, $E\left(X_{s,a}^i \middle| S_{s,a}\right)$, can be simplified further by conditioning on $\vec{r}$. Since the probability mass function of $\vec{r}$ conditional on $S_{s,a}$ is $f_{s,a}$,

$$E\left(X_{s,a}^i \middle| S_{s,a}\right) = E_{\vec{r}\sim f_{s,a}} E\left(X_{s,a}^i \middle| S_{s,a}, \vec{r}\right). \quad \text{(A2.75)}$$

Moreover, the distribution of $X_{s,a}$ conditional on $S_{s,a}$ and $\vec{r}$ is

$$\left(X_{s,a} \middle| \vec{r}, S_{s,a}\right) \sim Bin\left(M_1, p_{s,a} \cdot r_a / M_{s,a}\right), \quad \text{(A2.76)}$$

where $M_1 = M_{M,1} + M_{F,1}$ is the number of newborns of both sexes per-year, and therefore

$$E\left(X_{s,a} \middle| S_{s,a}\right) = E_{\vec{r}\sim f_{s,a}}\left(\frac{M_1 r_a p_{s,a}}{M_{s,a}}\right) = \frac{M_1 p_{s,a}}{M_{s,a}} \quad \text{(A2.77)}$$

and

$$E\left(X_{s,a}^2 \middle| S_{s,a}\right) = E_{\vec{r}\sim f_{s,a}}\left(M_1 \frac{r_a p_{s,a}}{M_{s,a}} + 2\binom{M_1}{2}\left(\frac{r_a p_{s,a}}{M_{s,a}}\right)^2\right) = \frac{M_1 p_{s,a}}{M_{s,a}} + 2\binom{M_1}{2}\left(\frac{p_{s,a}}{M_{s,a}}\right)^2 W_{s,a,a}.$$

Substituting these expression into Eq. A2.74, we find that

$$E(X_{s,a}) = \frac{p_{s,a}}{\gamma_s} \text{ and } E\left(X_{s,a}^2\right) = \frac{p_{s,a}}{\gamma_s} + \frac{M_1 - 1}{M_{s,a}} \frac{p_{s,a}^2}{\gamma_s} \cdot W_{s,a,a}, \quad \text{(A2.78)}$$

where γ_M and γ_F are the proportions of males and females at birth (i.e., $\gamma_s = M_{s,1}/M_1$). To calculate the remaining terms in Eq. A2.73, $E(X_{s,i} \cdot X_{s,j})$ for $j > i$, we note that conditioning on $S_{s,j}$, and on $\vec{r}|S_{s,j}$,

$$E\left(X_{s,i}\cdot X_{s,j}\right)=P\left(S_{s,j}\right)\cdot E\left(X_{s,i}\cdot X_{s,j}\middle|S_{s,j}\right)=\frac{M_{s,j}}{M_{s,1}}\cdot E_{\vec{r}\sim f_{s,j}}E\left(X_{s,i}\cdot X_{s,j}\middle|S_{s,j},\vec{r}\right). \quad \text{(A2.79)}$$

The latter term is easily calculated, since conditional on $S_{s,j}$ and $\vec{r}$, $X_{s,i}$ and $X_{s,j}$ are independent binomial variables: $\left(X_{s,i}\middle|\vec{r},S_{s,j}\right)\sim Bin\left(M_1,p_{s,i}\cdot r_i/M_{s,i}\right)$ and $\left(X_{s,j}\middle|\vec{r},S_{s,j}\right)\sim Bin\left(M_1,p_{s,j}\cdot r_j/M_{s,j}\right)$, and therefore

$$E\left(X_{s,i}\cdot X_{s,j}\right)=\frac{M_{s,j}}{M_{s,1}}\cdot E_{\vec{r}\sim f_{s,j}}\left(\frac{M_1^2p_{s,i}p_{s,j}r_ir_j}{M_{s,i}M_{s,j}}\right)=\frac{M_1p_{s,i}p_{s,j}W_{s,i,j}}{\gamma_sM_{s,i}}. \quad \text{(A2.80)}$$

Substituting the expressions from Eqs. A2.78 and A2.80 into Eq. A2.73 we obtain

$$E(X_s)=\frac{1}{\gamma_s}\text{ and }E(X_s^2)=\frac{1}{\gamma_s}+\frac{M_1}{\gamma_s}\sum_i\frac{p_{s,i}^2\cdot W_{s,i,i}+2\sum_{j>i}p_{s,i}p_{s,j}W_{s,i,j}}{M_{s,i}}-\sum_a\frac{p_{s,a}^2\cdot W_{s,a,a}}{\gamma_sM_{s,a}}. \quad \text{(A2.81)}$$

Assuming that the total population size is sufficiently large for the ratios $M_{s,i}/M_{t,j}$ and terms $W_{s,i,j}$ to be approximated as fixed, and for the higher order terms $\sum_a\frac{p_{s,a}^2\cdot W_{s,a,a}}{\gamma_sM_{s,a}}$ to be negligible, we find that

$$E(X_s)=\frac{1}{\gamma_s}\text{ and }E(X_s^2)\cong\frac{1}{\gamma_s}+\frac{M_1}{\gamma_s}\frac{W_s}{M_s}. \quad \text{(A2.82)}$$

The total reproductive variances of sex *s*, are therefore

$$V_s=E(X_s^2)-E^2(X_s)\cong\frac{M_1}{\gamma_s}\frac{W_s}{M_s}-\frac{1-\gamma_s}{\gamma_s^2}. \quad \text{(A2.83)}$$

From Eqs. A2.66 and A2.83, we obtain that

$$N_e^A=\frac{4G_AM_1}{\gamma_MV_M+\gamma_FV_F+\frac{\gamma_F}{\gamma_M}+\frac{\gamma_M}{\gamma_F}}\text{ and }N_e^X=\frac{4G_XM_1}{\frac{2}{3}\gamma_MV_M+\frac{4}{3}\gamma_FV_F+\frac{2}{3}\frac{\gamma_F}{\gamma_M}+\frac{4}{3}\frac{\gamma_M}{\gamma_F}}, \quad \text{(A2.84)}$$

where G_AM_1 and G_XM_1 are the total numbers of newborns per-generation, for autosomes and the X, respectively. Eq. A2.84 thus generalizes Eqs. A2.69 and A2.71 to the case with age-structure. Assuming that $E(\pi_A)=4N_e^A\mu_A$ and $E(\pi_X)=3N_e^X\mu_X$ (see Section A2.3), we find that

$$\frac{E(\pi_X)}{E(\pi_A)}=\frac{3}{4}\cdot\frac{f(\mu_M/\mu_F)\cdot f(G_M/G_F)}{f\left(\frac{\gamma_F/\gamma_M+\gamma_MV_M}{\gamma_M/\gamma_F+\gamma_FV_F}\right)}. \quad \text{(A2.85)}$$

Further assuming that the sex ratio at birth is 1 (i.e. that $\gamma_M = \gamma_F = {}^1/_2$), Eqs. A2.84 and A2.85 reduce to

$$N_e^A = \frac{4G_A M_1}{2+\frac{1}{2}V_M+\frac{1}{2}V_F},\ N_e^X = \frac{4G_X M_1}{2+\frac{1}{3}V_M+\frac{2}{3}V_F} \text{ and } \frac{E(\pi_X)}{E(\pi_A)} = \frac{3}{4}\cdot\frac{f(\mu_M/\mu_F)\cdot f(G_M/G_F)}{f\left(\frac{2+V_M}{2+V_F}\right)}, \qquad \text{(A2.86)}$$

which are the expressions presented in Eqs. 2.15 and 2.18 of chapter 2.

A2.2.5 Allelic reproductive variance

In chapter 2, we derived the effective population size for X and autosomes in terms of alleles rather than individuals, on the premise that we can then use the expressions obtained in the haploid model (Eqs. 2.10 and A2.29). Here we establish this premise, showing it to always be correct for autosomal alleles, whereas for the X it applies so long as the sex ratio at birth equals 1 (i.e., $\gamma_M = \gamma_F = 1/2$).

Consider an allele m carried by an individual I_m of sex s_m. We define the allele's (realized) reproductive success as the number of I_m's offspring who endogenous a copy of m, and denote it by X_m^A when m is autosomal and by X_m^X when it is X-linked. To obtain expressions for the effective population size in allelic terms, we calculate the first two moments of X_m^A and X_m^X, where m is chosen at random among alleles in newborns (in particular., m is carried by a male with probability $2M_{M,1}/(2M_{M,1}+2M_{F,1})$ for autosomes and $M_{M,1}/(2M_{M,1}+2M_{F,1})$ for the X).

First consider the case of an autosomal allele. To that end, we denote the total number of offspring of individual I_m by X_I. Since each offspring of I_m carries a copy of m with probability ½, the conditional distribution $X_m^A|X_I \sim Bin(X_I, \frac{1}{2})$. Based on the law of total variance, therefore

$$E(X_m^A) = \frac{1}{2}E(X_I) \text{ and } V(X_m^A) = \frac{1}{4}[E(X_I)+V(X_I)]. \qquad \text{(A2.87)}$$

Further conditioning on the sex of the individual carrying the allele, s_I, we note that $E(X_I|s_I) = 1/\gamma_{s_I}$ (Eq. A2.82) and $V(X_I|s_I) = V_{s_I}$, where the individual I_m is male with probability γ_M and female with probability γ_F. Applying the law of total variance again, we obtain

$$E(X_I) = 2 \text{ and } V(X_I) = \gamma_M V_M + \gamma_F V_F + \frac{(\gamma_M - \gamma_F)^2}{\gamma_M \gamma_F}. \quad \text{(A2.88)}$$

Substituting these expressions into Eq. A2.88, we find that $E(X_m^A) = 1$ and

$$V_A^* \equiv V(X_m^A) = \frac{1}{4}[\gamma_M V_M + \gamma_F V_F + \frac{\gamma_M}{\gamma_F} + \frac{\gamma_F}{\gamma_M}]. \quad \text{(A2.89)}$$

When the sex-ratio at birth is 1, and thus $\gamma_M = \gamma_F = 1/2$, Eq. A2.89 reduces to Eq. 2.14 for V_A^* in chapter 2. From Eqs. A2.84 and A2.89, we obtain

$$N_e^A = \frac{G_A \cdot M_1}{V_A^*} \quad \text{(A2.90)}$$

for any sex-ratio, which is Eq. 2.13 in chapter 2. While direct analogy with Eqs. 2.10 and A2.29 for the haploid case would result in an effective population size of $N_e = 2G_A \cdot M_1/V_A^*$, given that the effective population sizes are defined by requiring coalescence rates of $1/N_e$ in haploids and $1/(2 \cdot N_e^A)$ in diploids, Eq. A2.90 is, in fact, analogous to Eqs. 2.10 and A2.29.

Next, consider the case of an X-linked allele. If the individual carrying the allele, I_m, is a male, then only his female offspring will inherit the allele, and thus, $X_m^X|(s_I = M, X_I) \sim Bin(X_I, \gamma_F)$. Since $E(X_I|s_I = M) = 1/\gamma_M$ (Eq. A2.82) and $V(X_I|s_I) = V_{s_I}$, the law of total variance implies that

$$E(X_m^X|s_I = M) = \gamma_F/\gamma_M \text{ and } V(X_m^X|s_I = M) = \gamma_F^2 V_M + \gamma_F. \quad \text{(A2.91)}$$

The case in which I_m is a female is similar to the autosomal case, and thus, $X_m^X|(s_I = F, X_I) \sim Bin(X_I, 1/2)$,

$$E(X_m^X|s_I = F) = \frac{1}{2\gamma_F} \text{ and } V(X_m^X|s_I = F) = \frac{1}{4}V_F + \frac{1}{4\gamma_F}. \quad \text{(A2.92)}$$

Given that there are $M_{M,1}$ X-linked alleles in newborn males and $2M_{F,1}$ in newborn females, the probability that an X-linked allele in a newborn is in a male is $\gamma_M/(1+\gamma_F)$ and the probability it is in a female is $2\gamma_F/(1+\gamma_F)$. Applying the law of total variance therefore implies that

$$E(X_m^X) = 1 \text{ and } V_X^* = Var(X_m^X) = \frac{\gamma_M\gamma_F^2}{1+\gamma_F}V_M + \frac{\gamma_F}{2(1+\gamma_F)}V_F + \frac{1+2\gamma_M\gamma_F}{2(1+\gamma_F)} + \frac{(1-2\gamma_F)^2}{2\gamma_M\gamma_F}. \quad \text{(A2.93)}$$

When the sex-ratio at birth is 1, and thus $\gamma_M = \gamma_F = 1/2$, Eq. A2.93 reduces to Eq. 2.14 for V_X^* in chapter 2. From Eqs. A2.84 and A2.93, we find that

$$N_e^X = \frac{G_X \cdot M_1}{V_X^*}, \quad \text{(A2.94)}$$

which is Eq. 2.13 in chapter 2, only holds when $\gamma_M = \gamma_F = 1/2$. Thus, the haploid result (Eqs. 2.10 and A2.29) applies to X-linked alleles only when the sex ratio at birth equals 1.

To gain some intuition as to why this result fails in the general case, consider the reproductive success of an X-linked allele in consecutive generations. As we have shown above, an allele's expected reproductive success is γ_F/γ_M in males and $1/(2\gamma_F)$ in females (averaged over the sexes the expectation is 1). Now consider the expected reproductive success in the next generation: if the allele was in a male in the previous generation it will necessarily be in a female, and the expected reproductive success of the offspring allele would be $1/(2\gamma_F)$; if the allele was in a female in the previous generation, the expected reproductive success is obtained by averaging over the sex of the offspring, and is $\frac{1}{2} + \gamma_F$. Thus, unless $\gamma_M = \gamma_F = 1/2$, the reproductive success of an X-linked allele will be negatively correlated between parents and offspring. Thus, the assumption of the haploid model that the reproductive success of individuals and their offspring are independent variables is clearly violated in this case.

A2.3. Mutational process

Here we describe the assumptions on the mutational model and derive formulas for the expected levels of heterozygosity. To incorporate what has recently been revealed about the dependencies of mutation rates on sex and age (e.g., (6-8)), we allow for mutation rate in the diploid model to depend on sex and age. Namely, we assume that the number of de novo mutations that a parent of sex s and age a bequeaths to its newborn is a random variable with expectation $\mu_{s,a}$ per base pair. While in chapter 2, we consider a specific model for $\mu_{s,a}$ motivated by pedigree studies in humans, our derivation here treat $(\mu_{s,a})_{a=1}^{\infty}$ as parameters and assumes no specific form. Since mutation rates vary with sex and age, the mutation rates per generation in males and females depend on the distributions of their breeding ages (i.e. A_M and A_F, which were defined in Section A2.2). We denote the expected mutation rate per generation in males by $\mu_M = E_{A_M}(\mu_{M,a}) = \sum_a p_{M,a} \cdot \mu_{M,a}$ and the expected rate in females by $\mu_F = E_{A_F}(\mu_{F,a})$. The average rates on the autosomes and the X are given by $\mu_A = \frac{1}{2}(\mu_M + \mu_F)$ and $\mu_X = \frac{2}{3}\mu_F + \frac{1}{3}\mu_M$. For the haploid model, we assume the expected number of mutations μ_a to be dependent of age and define the per generation rate as $\mu = E_A(\mu_a)$. In the special case in which the parameters $\mu_{s,a}$ (or the μ_a in the haploid case) depend linearly on age, these expectations will depend only on the expected generation times G_M and G_F, i.e., they are insensitive to higher moments of the distributions of breeding ages in males and females. As we show below, higher moments of the distributions of mutation rates per generation do not affect our results, which is how we avoid any further assumptions about these distributions. The standard expressions for heterozygosity (e.g., $E(\pi_A) = 4N_e^A\mu_A$) are usually derived assuming that the genealogical and mutational processes are independent (9). This assumption is violated in our case, because both the time to the most recent common ancestor and the number of accumulated mutations depend on the ages of the individuals along the lineage. To derive the

expected autosomal heterozygosity $E(\pi_A)$ under these conditions, we track alleles *A* and *B* backwards in time. Let X_i denote the number of mutations occurring on the lineage leading from allele *A* in the i^{th} generation and *T* denote the number of generations until the alleles coalesce. The number of mutations on the lineage leading to allele *A* is then $\sum_{i=1}^{T} X_i$. Although X_i and *T* are dependent variables, Wald's equation (10) implies that $E(\sum_{i=1}^{T} X_i) = E(T) \cdot E(X_i)$ (to see that Wald's equation holds, note that the indicator function $\mathbb{I}_{T \geq n}$ is independent of X_n, since the first depends on the sexes and ages in the first $n-1$ generations, and the second on the n^{th} generation). We have shown previously that $E(T) = 2N_e^A$. Since $E(X_i|s_i, a_i) = \mu_{s,a}$ (where s_i and a_i are the sex and age in the i^{th} generation), it follows that $E(X_i) = E(\mu_{s,a}) = \mu_A$. We conclude that the lineage leading to allele A has on average $E(\sum_{i=1}^{T} X_i) = 2N_e^A \cdot \mu_A$ mutations and therefore $E(\pi_A) = 4N_e^A \mu_A$. A similar argument shows that for the haploid model $E(\pi) = 2N_e\mu$.

The same argument cannot be readily applied to the X-chromosome, as the sexes s_i and s_{i+1} in consecutive generations along the lineage are dependent variables, leading to a dependence between X_{i+1} and s_i, in violation of the conditions for Wald's equation to hold. Instead, we define *T* as the number of females on the lineage until the coalescence occurs, and define X_i as the number of mutations between the i^{th} and $i+1$ females on the lineage. Under this definition, Wald's equation holds and $E(\pi_X) = 2E(\sum_{i=1}^{T} X_i) = 2E(X_i)E(T)$. It is easily shown that $E(X_i) = \frac{3}{2}\mu_X$ and $E(T) = N_e^X$, so that $E(\pi_X) = 3N_e^X \mu_X$.

A2.4. References

1. Hill WG (1972) Effective size of populations with overlapping generations. *Theor Popul Biol* 3(3):278-289.
2. Felsenstein J (1971) Inbreeding and variance effective numbers in populations with overlapping generations. *Genetics* 68(4):581-597.
3. Sagitov S & Jagers P (2005) The coalescent effective size of age-structured populations. *Annals of Applied Probability* 15(3):1778-1797.
4. Wright S (1939) Statistical genetics in relation to evolution, Vol. 802. *Hermann et Cie.: Paris*.
5. Pollak E (2011) Coalescent theory for age-structured random mating populations with two sexes. *Math Biosci* 233(2):126-134.
6. Kong A*, et al.* (2012) Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488(7412):471-475.
7. Segurel L, Wyman MJ, & Przeworski M (2014) Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet* 15:47-70.
8. Wong WS*, et al.* (2016) New observations on maternal age effect on germline de novo mutations. *Nat Commun* 7:10486.
9. Hudson RR (1990) Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology* 7(1):44.
10. Blackwell D (1946) On an Equation of Wald. *Annals of Mathematical Statistics* 17(1):84-87.