Supplementary Information for
Overlooked roles of DNA damage and maternal age in generating human germline mutations

Ziyue Gao[1,+], Priya Moorjani[2,3,4], Guy Amster[4,*] and Molly Przeworski[4,5,*,+]

[1] Howard Hughes Medical Institute & Department of Genetics, Stanford University
[2] Department of Molecular and Cell Biology, University of California, Berkeley
[3] Center for Computational Biology, University of California, Berkeley
[4] Department of Biological Sciences, Columbia University
[5] Department of Systems Biology, Columbia University
[*] Contributed equally
[+] To whom correspondence should be addressed: ziyuegao@stanford.edu and mp3284@columbia.edu

Table of Contents

Materials and Methods

Processing of *de novo* mutation (DNM) data
Estimation of sex-specific mutation parameters with a model-based approach
Derivation of the likelihood function for estimating sex-specific mutation parameters
Confidence intervals of male-to-female mutation ratio at given parental ages
Test for alternative models for parental age effects
Alternative hypotheses for maternal age effect on maternal mutation rate
Detection and estimation of maternal age effect on paternal mutation rate
Differences in mutation properties of trios with or without a third generation in Jónsson et al (2017)

Supplementary Tables

Table S1 Estimated parental age effects based on the maximum likelihood model and comparison to estimates from Jónsson et al (2017)
Table S2 Summary statistics of the two DNM data sets
Table S3 Comparison of models with linear and exponential parental age effects (in accompanying excel spreadsheets)
Table S4 Comparison of linear models fitted to trios with different maternal ages (in accompanying excel spreadsheets)
Table S5 Comparison of models with linear and exponential parental age effects fitted to all trios and trios with maternal age below 40 (in accompanying excel spreadsheets)
Table S6 Co-occurrence of C>G and indels on the same chromosome
Table S7 Estimates of maternal age effect on paternal mutations by different methods
Table S8 Estimating the probability of a spurious maternal age effect on paternal mutations
Table S9 Estimating the probability of a stronger maternal age effect on paternal C>A mutations than paternal age effect
Table S10 Differences in the age and sex dependencies of DNMs between the trios with or without a third generation

Supplementary Figures

Figure S1. Fraction of paternal mutations among phased mutations as a function of paternal age
Figure S2. Replication of the stable fraction of paternal mutations with paternal age in an independent dataset

Figure S3. Fraction of paternal mutations among phased mutations for different ratios of paternal age (G_P) to maternal age (G_M)
Figure S4. Comparison of parental age effects and predicted male-to-female mutation ratio at given ages estimated from two DNM datasets
Figure S5. Estimated sex-specific mutation rates and male-to-female mutation ratio as a function of parental ages, by mutation type
Figure S6. Distribution of distances to the closest deletion of ≥5bp for C>G mutations
Figure S7. Effect of maternal age on the number of paternal DNMs by deviation analysis

Supplementary Methods

Processing of *de novo* mutation data

For each *de novo* mutation (DNM), we obtained parental ages at conception of the child (proband) and the position, allele and parent-of-origin information from the Supplementary Material of the publication for one dataset (Jónsson, Sulem, Kehr, et al. 2017) and by personal communication with the authors for the replication dataset (Goldmann et al. 2016; Wong et al. 2016). We considered a mutation as "phased" if the parental haplotype on which it arose was determined by either informative flanking variant in the read or from transmission to a third generation (Jónsson et al. 2017, Goldmann et al. 2016). See Table S2 for a comparison of summary statistics of these datasets.

For both datasets, we removed indels and mutations on X chromosome (no Y-linked DNMs were reported), which resulted in 98,858 and 35,793 point mutations (or single nucleotide substitutions) for Jónsson et al (2017) and Goldmann et al (2016), respectively. Each of these mutations was then assigned into one of six mutation types (T>A, T>C, T>G, C>A, C>G, and C>T) based on the original allele present in homozygous state in both parents and the derived allele that is carried by the child in heterozygous state. Complementary combinations (such as C>T and G>A) were combined such that the original allele is always a pyrimidine (C or T). Moreover, each DNM was annotated to be in CpG or nonCpG context based on its two immediate flanking bases extracted from human reference genome. For analyses of C>T mutations at CpG sites, we excluded CpG transitions present in CpG islands (annotations downloaded from UCSC browser: CpG Islands track), because these sites are thought to be hypomethylated and thus behave differently in terms of mutation rate compared to CpG sites outside CGI (Moorjani et al. 2016). C>T mutations at CpG sites in CGIs were included in analysis of "all point mutations".

Estimation of sex-specific mutation parameters with a model-based approach

Similar to Jónsson et al (2017), we modeled the expected number of mutations from a parent as a linear function of her (or his) age at conception of the child, and assumed that the observed maternal (paternal) mutation count follows a Poisson distribution with this expectation. One difference from the Jónsson et al (2017) model is in how we account for the incomplete parental origin information for the unphased DNMs. Unlike Jónsson et al., we explicitly modeled the phasing process as a binomial sampling of DNMs with a proband-specific phasing rate parameter, assuming that the phasing probabilities of all mutations in the same individual are identical and independent (as seems sensible). This approach enabled us to fully leverage information from both phased and unphased mutations jointly.

Specifically, the increase in DNMs with sex-specific parental ages is modeled as the following:
$X_M^i \sim \text{Poisson}(\alpha_M + \beta_M G_M^i)$,
$X_P^i \sim \text{Poisson}(\alpha_P + \beta_P G_P^i)$,
where index i indicates the proband; X_M^i and X_P^i are the total numbers of maternal and paternal mutations; G_M^i and G_P^i are ages of the mother and the father at conception, respectively; α_M, β_M, α_P, and β_P are the mutation parameters that characterize the sex-specific parental age effects and are shared across all probands (note that α_M and α_P are the extrapolated intercepts at age zero, which are not necessarily non-negative). We assumed linear effects for both sexes in the initial model, but relaxed this assumption by testing for exponential effects for either or both sexes later (see "Test for alternative models for parental age effects" below; Table S3).

Because of incomplete phasing, X_M^i and X_P^i are not directly observed. Thus, we modeled the observed mutation counts as:
$Y_M^i \sim \text{Binomial}(X_M^i, p^i)$,
$Y_P^i \sim \text{Binomial}(X_P^i, p^i)$,
$Y_U^i = (X_M^i - Y_M^i) + (X_P^i - Y_P^i)$,

where p^i is the phasing rate in proband i and $Y_M{}^i$, $Y_P{}^i$ and $Y_U{}^i$ represent the numbers of phased maternal, phased paternal and unphased mutations, respectively. $Y_M{}^i$, $Y_P{}^i$ and $Y_U{}^i$ are defined as random variables, and we denote the observed values of these with lower case notations $y_M{}^i$, $y_P{}^i$ and $y_U{}^i$.

With the parameterization above, the likelihood of the observed data for proband i can be written as:

$$L^i = \mathrm{P}(Y_M{}^i = y_M{}^i,\ Y_P{}^i = y_P{}^i,\ Y_U{}^i = y_U{}^i \mid \alpha_M, \alpha_P, \beta_M, \beta_P, G_M{}^i, G_P{}^i, p^i)$$

$$= \mathrm{P}(y_M{}^i, y_P{}^i, y_U{}^i \mid X_M{}^i, X_P{}^i, p^i)\mathrm{P}(X_M{}^i \mid \alpha_M, \beta_M, G_M{}^i)\mathrm{P}(X_P{}^i \mid \alpha_P, \beta_P, G_P{}^i)$$

$$= \sum\nolimits_{x_M^i, x_P^i} P\left(y_M{}^i, y_P{}^i, y_U{}^i \middle| \ X_M{}^i = x_M{}^i,\ X_P{}^i = x_P{}^i, p^i\right) P(\ X_M{}^i = x_M{}^i | \alpha_M, \beta_M, G_M{}^i) P(\ X_P{}^i = x_P{}^i | \alpha_P, \beta_P, G_P{}^i)$$

$$= \sum\nolimits_{k=0}^{y_U{}^i} [P\left(\left(y_M{}^i, y_P{}^i, y_U{}^i \middle| \ X_M{}^i = y_M{}^i + k,\ X_P{}^i = y_P{}^i + y_U{}^i - k, p^i\right)\right)] P\left(y_M{}^i + k \middle| \alpha_M, \beta_M, G_M{}^i\right) P\left(y_P{}^i + y_U{}^i - k \middle| \alpha_P, \beta_P, G_P{}^i\right).$$

We note that the likelihood function of Jónsson et al. (2017) does not include the first term, which is the probability of the observed data given possible partitionings of the unphased mutations into paternal and maternal origins (assuming the same phasing rates of maternal and paternal mutations). As an illustration, the set of observations ($y_M{}^i$, $y_P{}^i$, $y_U{}^i$) = (10, 30, 80) is more probable under ($x_M{}^i$, $x_P{}^i$)=(30, 90), where one third of DNMs were phased for both parental origins, than under ($x_M{}^i$, $x_P{}^i$)=(80, 40), where 75% paternal DNMs were phased but only 12.5% of maternal DNMs.

The likelihood function for proband i can be simplified as (see "Derivation of the likelihood function for estimating sex-specific mutation parameters" section later):

$$L_i = \frac{{p^i}^{(y_M{}^i + y_P{}^i)} (1 - p^i)^{y_U{}^i}}{y_U{}^i!\, y_M{}^i!\, y_P{}^i!} \cdot \frac{\left(\alpha_M + \beta_M G_M{}^i\right)^{y_M{}^i} \left(\alpha_P + \beta_P G_P{}^i\right)^{y_P{}^i} \left(\alpha_M + \beta_M G_M{}^i + \alpha_P + \beta_P G_P{}^i\right)^{y_U{}^i}}{e^{(\alpha_M + \beta_M G_M{}^i + \alpha_P + \beta_P G_P{}^i)}}.$$

Note that the first term contains the phasing rate (p^i) but is independent of the mutation parameters, whereas the second term is dependent on the mutation parameters but independent of p^i. Therefore, the maximum likelihood estimator (MLE) of p^i and those of the mutation parameters can be identified by maximizing the first and second terms separately.

The log joint likelihood of all observed data under a set of mutation parameter values can be expressed as:

$$LL = log \prod_{i=1}^{N} L_i = \sum_{i=1}^{N} \log(L_i)$$

$$= C + \sum\nolimits_{i=1}^{N} [y_M{}^i \log\left(\alpha_M + \beta_M G_M{}^i\right) + y_P{}^i \log\left(\alpha_P + \beta_P G_P{}^i\right) + y_U{}^i \log\left(\alpha_M + \beta_M G_M{}^i + \alpha_P + \beta_P G_P{}^i\right) - (\alpha_M + \beta_M G_M{}^i + \alpha_P + \beta_P G_P{}^i)],$$

where C is a constant that is independent of the mutation parameters of interest.

We implemented this log likelihood function in R and found the MLEs of the mutation parameters by using function mle2 in the package "bbmle" (with the optimization method "BFGS"; Bolker, R Development Core Team 2017). To avoid being trapped in local maxima, we tested a grid of initial values for the slopes (β_P and β_M). We performed the estimation for all point mutations altogether as well as for each mutation type separately.

Derivation of the likelihood function for estimating sex-specific mutation parameters

Following the parameter setup and assumptions specified in the Methods section, the likelihood of the observed data for proband i is:

$$L^i = \mathrm{P}(Y_M{}^i = y_M{}^i,\ Y_P{}^i = y_P{}^i,\ Y_U{}^i = y_U{}^i \mid \alpha_M, \alpha_P, \beta_M, \beta_P, G_M{}^i, G_P{}^i, p^i)$$

$$= \mathrm{P}(y_M{}^i, y_P{}^i, y_U{}^i \mid X_M{}^i, X_P{}^i, p^i)\mathrm{P}(X_M{}^i \mid \alpha_M, \beta_M, G_M{}^i)\mathrm{P}(X_P{}^i \mid \alpha_P, \beta_P, G_P{}^i)$$

$$=\sum_{k=0}^{y_U{}^i} P\left(\left(y_M{}^i, y_P{}^i, y_U{}^i \middle| X_M{}^i = y_M{}^i + k,\ X_P{}^i = y_P{}^i + y_U{}^i - k, p^i\right) P\left(y_M{}^i + k \middle| \alpha_M, \beta_M, G_M{}^i\right) P\left(y_P{}^i + y_U{}^i - k \middle| \alpha_P, \beta_P, G_P{}^i\right)\right.$$

The first term of the addend can be expressed as:

$$P\left(\left(y_M{}^i, y_P{}^i, y_U{}^i \middle| X_M{}^i = y_M{}^i + k,\ X_P{}^i = y_P{}^i + y_U{}^i - k, p^i\right)\right.$$

$$= P\left(y_M{}^i \middle| X_M{}^i = y_M{}^i + k, p^i\right) P\left(y_P{}^i \middle| X_P{}^i = y_P{}^i + y_U{}^i - k, p^i\right)$$

$$= Bin\left(y_M{}^i \middle| y_M{}^i + k, p^i\right) Bin\left(y_P{}^i \middle| y_P{}^i + y_U{}^i - k, p^i\right)$$

$$= \frac{\left(y_M{}^i + k\right)!}{y_M{}^i!\,k!}(p^i)^{y_M{}^i}(1-p^i)^k \frac{\left(y_P{}^i + y_U{}^i - k\right)!}{y_P{}^i!\,(y_U{}^i - k)!}(p^i)^{y_P{}^i}(1-p^i)^{(y_U{}^i - k)}$$

$$= \frac{p^{i(y_M{}^i + y_P{}^i)}(1-p^i)^{y_U{}^i}}{y_M{}^i!\,y_P{}^i!} \frac{1}{k!\,(y_U{}^i - k)!}\left(y_M{}^i + k\right)!\left(y_P{}^i + y_U{}^i - k\right)!$$

The second term of the addend is:

$$P\left(y_M{}^i + k \middle| \alpha_M, \beta_M, G_M{}^i\right) = Poisson\left(y_M{}^i + k \middle| \alpha_M + \beta_M G_M{}^i\right) = \frac{(\alpha_M + \beta_M G_M{}^i)^{y_M{}^i + k} e^{-(\alpha_M + \beta_M G_M{}^i)}}{(y_M{}^i + k)!}$$

Similarly, the third term of the addend is:

$$P\left(y_P{}^i + y_U{}^i - k \middle| \alpha_P, \beta_P, G_P{}^i\right) = Poisson\left(y_P{}^i + y_U{}^i - k \middle| \alpha_P + \beta_P G_P{}^i\right)$$

$$= \frac{(\alpha_P + \beta_P G_P{}^i)^{y_P{}^i + y_U{}^i - k} e^{-(\alpha_P + \beta_P G_P{}^i)}}{(y_P{}^i + y_U{}^i - k)!}$$

Therefore, the addend can be written together as:

$$\frac{p^{i(y_M{}^i + y_P{}^i)}(1-p^i)^{y_U{}^i}}{y_M{}^i!\,y_P{}^i!} \frac{1}{k!\,(y_U{}^i - k)!}\left(y_M{}^i + k\right)!\left(y_P{}^i + y_U{}^i - k\right)! \frac{(\alpha_M + \beta_M G_M{}^i)^{y_M{}^i + k} e^{-(\alpha_M + \beta_M G_M{}^i)}}{(y_M{}^i + k)!} \frac{(\alpha_P + \beta_P G_P{}^i)^{y_P{}^i + y_U{}^i - k} e^{-(\alpha_P + \beta_P G_P{}^i)}}{(y_P{}^i + y_U{}^i - k)!}$$

$$= \frac{p^{i(y_M{}^i + y_P{}^i)}(1-p^i)^{y_U{}^i}}{y_M{}^i!y_P{}^i!} \frac{1}{k!(y_U{}^i - k)!}(\alpha_M + \beta_M G_M{}^i)^{y_M{}^i + k}(\alpha_P + \beta_P G_P{}^i)^{y_P{}^i + y_U{}^i - k} e^{-(\alpha_M + \beta_M G_M{}^i + \alpha_P + \beta_P G_P{}^i)}$$

If we re-organize it to isolate terms independent of *k*, the addend becomes:

$$\frac{p^{i(y_M{}^i + y_P{}^i)}(1-p^i)^{y_U{}^i}(\alpha_M + \beta_M G_M{}^i)^{y_M{}^i}(\alpha_P + \beta_P G_P{}^i)^{y_P{}^i} e^{-(\alpha_M + \beta_M G_M{}^i + \alpha_P + \beta_P G_P{}^i)}}{y_M{}^i!\,y_P{}^i!}$$

$$\times \frac{(\alpha_M + \beta_M G_M{}^i)^k (\alpha_P + \beta_P G_P{}^i)^{y_U{}^i - k}}{k!\,(y_U{}^i - k)!}$$

Given that the second term of the above resembles the binomial point mass function, and that:

$$1 = \sum_{k=0}^{y_U{}^i} \frac{y_U{}^i}{k!\,(y_U{}^i - k)!}\left(\frac{\alpha_M + \beta_M G_M{}^i}{\alpha_M + \beta_M G_M{}^i + \alpha_P + \beta_P G_P{}^i}\right)^k \left(\frac{\alpha_P + \beta_P G_P{}^i}{\alpha_M + \beta_M G_M{}^i + \alpha_P + \beta_P G_P{}^i}\right)^{y_U{}^i - k}$$

We obtain:

$$\sum_{k=0}^{y_U{}^i} \frac{(\alpha_M + \beta_M G_M{}^i)^k (\alpha_P + \beta_P G_P{}^i)^{y_U{}^i - k}}{k!\,(y_U{}^i - k)!} = \frac{(\alpha_M + \beta_M G_M{}^i + \alpha_P + \beta_P G_P{}^i)^{y_U{}^i}}{y_U{}^i!}$$

Therefore, the likelihood for proband *i* can be simplified to:

$$L_i = \frac{p^{i(y_M{}^i + y_P{}^i)} (1 - p^i)^{y_U{}^i} (\alpha_M + \beta_M G_M{}^i)^{y_M{}^i} (\alpha_P + \beta_P G_P{}^i)^{y_P{}^i} e^{-(\alpha_M + \beta_M G_M{}^i + \alpha_P + \beta_P G_P{}^i)}}{y_M{}^i!\, y_P{}^i!} \frac{(\alpha_M + \beta_M G_M{}^i + \alpha_P + \beta_P G_P{}^i)^{y_U}}{y_U{}^i!}$$

Confidence intervals of male-to-female mutation ratio at given parental ages

To account for uncertainties in the DNM parameter estimates, we used a bootstrap approach, randomly re-sampling the probands with replacement 500 times, keeping the same total number of probands in each run. For each replicate, we obtained the MLEs of the DNM parameters as described above, predicted the numbers of paternal and maternal mutations at given ages, and calculated the male-to-female mutation ratio; the actual average ages at conception in the Icelandic dataset are 28.2 and 32.0 for mothers and fathers, respectively). Thus, each bootstrap provides one point estimate for each of the quantities of interest, and the approximate distribution for each quantity can be obtained by aggregating results from the 100 replicates.

Test for alternative models for parental age effects

In addition to the linear model described in the above, we also considered models with exponential parental age effects post-puberty for either or both sexes. Specifically, we modeled the exponential parental age effect as follows:

$X_M{}^i \sim$ Poisson(a_M + Exp[$b_M(G_M{}^i\text{-}P)+c_M$]);
$X_P{}^i \sim$ Poisson(a_P + Exp[$b_P(G_P{}^i\text{-}P)+c_P$]),
where P=13 is the age of onset of puberty assumed for both sexes. We note that results are not sensitive to the choice of the value of P. Under this formulation, models with different values of P are mathematically equivalent to models with the same b_P (or b_M) value but different c_P (or c_M) values. Indeed, we confirmed the MLEs for b_P and b_M are the same for different P values (even for P=0).

We obtained the MLEs and corresponding log likelihoods of all four models for all point mutations combined and for each mutation type separately, and used the Akaike information criterion (AIC) to compare the relative fits of different models (a smaller AIC indicates a better fit of the model). We took ΔAIC<-6 as the threshold for evidence for a significantly better fit (approximately 20-fold more probable). The models with exponential paternal age effect provide worse fits (ΔAIC>0) for all mutation types.

For all DNMs combined, models with exponential effects of maternal age or both parental ages provide significantly better fits but are not significantly different from each other (Table S3). As verification, we split the 1,548 trios into two groups with maternal age at conception over and under 27 years (the median maternal ages in the dataset), respectively, and fitted both with linear parental age effects. The estimate of the maternal age effect is greater for older mothers than for younger mothers (0.56 vs 0.24, 95% CI: [0.45,0.66] vs [0.12,0.38]), whereas the estimates of paternal age effect are similar for the two groups (1.41 vs 1.40, 95% CI: [1.31, 1.51] vs [1.29, 1.53]; Table S3). The improved fit by an exponential maternal age was no longer significant when excluding the 72 trios with maternal age above 40 (Table S5). We therefore considered the linear model fitted to trios with maternal ages under 40 years for analyses for all point mutations combined throughout the manuscript. The predictions based on linear models fitted to all trios and trios with Gm under 40 show similar results: for instance, the estimated male-to-female mutation ratio at puberty (*P*=13) is 3.1 vs 3.3 (95% CI [2.8, 3.5] vs [3.0, 3.7]).

Among all mutation types considered, C>G transversions are the only type for which the model with exponential maternal age effect provides a significantly better fit by the criterion of ΔAIC<-6 (Table S3). Therefore, in all analyses for C>G transversions (e.g., calculation of α), we used the estimates from the model with an exponential maternal age effect (and linear paternal age effect) fitted to all 1548 trios. For all DNMs combined, the model with an exponential maternal age effect also provides a significantly better fit than the linear model. Interestingly, considering all trios, even after C>G transversions (or C>G transversions and CpG transitions) are excluded, an exponential maternal age effect still provides a significantly better fit for other point mutations combined (ΔAIC<-9; Table S5), suggesting that the signal is not driven by C>G mutations alone. Again, this effect is no longer discernable when trios with maternal age above 40 are excluded (Table S4).

Alternative hypotheses for maternal age effect on maternal mutation rate

The accumulation of DNA lesions and damage-induced mutations in aging oocytes is not the only logical explanation for a maternal age effect, as there are two (non-mutually exclusive) hypotheses that could allow for a maternal age effect due to replication-driven mutations.

Under hypothesis 1, all or most female germline DNMs arise from replication errors in mothers and therefore predate the formation of the primary oocytes, but there exists some mechanism by which oocytes with fewer replicative point mutations tend to be ovulated in earlier menstrual cycles. While this hypothesis is conjecture, evidence from mouse suggests that oogonia that enter meiosis earlier are ovulated earlier (Polani and Crolla 1991) and may experience fewer mitoses (Fulton et al. 2005). Given the roughly two-fold difference in maternal mutation rate between ages 17 and 40, this scenario would require oocytes of a 40-year-old mother to have experienced about two times the number of cell divisions of a 17-year-old mother—potentially more, depending on how mutagenic the first few cell divisions are compared to subsequent cell divisions (Lindsay et al. 2016; Harland et al. 2017). In this scenario, depending on unknown specifics of germ cell lineage relationships, older oocytes may not only accumulate more point mutations, but also share more mutations with other older oocytes. Thus, it is unclear if this hypothesis is consistent with the observation that the offspring of older mothers share a *smaller* fraction of maternal *DNM*s with their siblings (Jónsson, Sulem, Arnadottir, et al. 2017).

Under hypothesis 2, mutations increase with maternal age because proteins or mRNA transcripts in the oocytes deteriorate with maternal age (or the oocyte or sperm accumulates mutagens with parental ages), such that the first few divisions after fertilization generate more post-zygotic mutations in older mothers. This scenario is plausible, as a human zygote relies on the protein/transcript reservoir of the oocyte until the 4-cell or 8-cell stage (Braude, Bolton, and Moore 1988; Dobson et al. 2004; Zhang et al. 2009). It predicts that the number of DNMs on the paternal chromosomes should also increase with maternal age. We detected such an effect in the 202 trios with almost all DNMs phased (see details in "Detection and estimation of a maternal age effect on paternal mutation rate" section below, and main text). This finding does not distinguish between replication-driven and damage-induced mutations, however, as it can also arise from the deterioration of maternal repair proteins responsible for correcting DNA lesions during the embryonic cleavage stage, i.e., from damage-induced mutations (see main text). This hypothesis further predicts that offspring of older mothers should share a smaller fraction of maternal DNMs, since a larger fraction will have arisen post fertilization and hence be child specific (Figure 4A,B), as observed (Jónsson, Sulem, Arnadottir, et al. 2017).

Importantly, however, neither hypothesis 2 nor hypothesis 1 alone explains why the male-to-female mutation ratio is already high at puberty and remains stable with parental age beyond puberty (Fig 1, Fig 2B) or why paternal mutations increase roughly proportionally to paternal age (Fig 2A) for mutations other than C>G and CpG>TpG. Instead, at least two additional and very specific conditions would have to be met, involving balancing acts of the per cell division mutation rates and the numbers of cell divisions in multiple developmental stages (as well as the strength of maternal age effect on the paternal genome in the case hypothesis 2). In contrast, both the stable male-to-female mutation ratio and parental age effects can be explained if most mutations are induced by DNA damage and male and female germlines have distinct but roughly constant damage rates (per unit of time) throughout life. Thus, taken together, our observations suggest a role for hypothesis 2—a maternal age effect on early embryonic development—and a role for damage induced mutations in both sexes (see main text).

Detection and estimation of maternal age effect on paternal mutation rate

For analyses in this section, we focused on the 199 probands in which almost all DNMs were phased (>95% DNM phased). We first did a Poisson regression (with an identity link) of the count of paternal point mutations on both parental ages and found a significant effect of the maternal age (p= 0.035) and a slight but non-significant improvement in the fit compared to a model with paternal age only (ΔAIC=-2.4; approximately 3.3-fold more probable); p-values and AIC obtained by glm() function in R (with option "family = poisson(link = "identity")"). In contrast, regressing the maternal mutation count on both parental ages does not provide any improvement in the fit compared to a model with maternal age alone (ΔAIC=0.2). See Table S7 for estimated effect sizes by Poisson regression.

Motivated by this finding, we re-estimated the mutation parameters by maximum likelihood under models including a maternal age effect on paternal mutations (i.e., "maternal-on-paternal effect") of the same size (model 1) or a different size (model 2) than the maternal age effect on maternal mutations. Both models provide slight but insignificant improvements in fit compared to a model without a maternal age effect on paternal mutations (model 0) (Table S7), and the model with the same maternal effect on both maternal and paternal mutations gives the best fit based on AIC (ΔAIC =-3.7; MLE of maternal age effect is 0.34 mutations per year; Table S7).

We then carried out two types of analyses conditional on paternal age. First, we performed "pairwise analysis", in which we compared all pairs of trios with the same paternal age, G_P, but different maternal ages, G_M, (i.e., a "pairwise analysis"). In 619 such pairs, the child born to the older mother carries more paternal mutations than does the child with the younger mother in 319 cases, fewer in 280 cases, and the same number in 20 cases, and the difference in the number of paternal mutations is significantly associated with the difference in the maternal age (Kendall's rank test tau=0.09, p= 0.0015). Because some of the pairs include the same probands and are thus not independent, we did a permutation test by swapping the maternal ages within paternal age bin and calculating the adjusted z-score of Kendall's tau-b statistic. 220 out of 10,000 permutations had statistics equal to or greater than that observed with in real data (corresponding to an empirical one-tailed p-value of 0.022). To estimate the effect size of maternal age, we ran weighted linear regression of the difference in paternal counts on the difference in maternal ages for each pair of trios with the same paternal age (with an intercept of zero), with the weight of each data point specified as the inverse of the paternal age, which is approximately proportional to the variance in the observed difference in paternal mutation counts (Table S7). Because the mutation counts are integers and do not follow a normal distribution, the standard errors are inaccurate. We also did similar pairwise analysis for (1) the difference in maternal mutation counts against the difference in G_M, conditional on the same G_P, (2) the difference in paternal mutation counts against the difference in G_P, conditional on the same G_M, and (3) the difference in maternal mutation counts against the difference in G_P, conditional on the same G_M (Table S7); significance levels were evaluated with permutation tests, as described above.

To minimize the issue of non-independence among pairs of trios and to estimate the maternal age effect on paternal mutations, we also ran a "deviation analysis." For each proband, we found all probands with the same G_P and calculated the average number of paternal mutations and average maternal age for them.

Then we calculated the deviation of the number of paternal mutations for the proband of interest by subtracting the average from the number of his/her paternal mutations. We applied Kendall's rank test on the two deviation values across probands and found a significant association (p= 0.020); p-values were confirmed with a permutation test swapping maternal ages among trios with the same paternal age. We also estimated the maternal age effect by weighted linear regression as described above. We then did the reciprocal analysis to evaluate the effect of paternal age on maternal mutations and found no significant signal (Table S7).

One concern is that parental ages are assigned to integer bins in the Icelandic dataset, and there is potentially a subtle correlation between maternal and paternal ages even within a paternal age bin, in which case variation in paternal age counts caused by small G_P variation may be mistakenly ascribed to an effect of G_M. To address this concern, we simulated data of 202 trios with similar parental age structure but no maternal age effect on paternal DNMs, and asked how frequently analysis of simulated data based on binned parental ages would generate signals comparable to those observed in actual data. To mimic the distributions of maternal and paternal ages and the correlation between them in the actual dataset, we simulated an exact maternal age for each trio by adding a random variable that is uniform on (0,1) to the integer maternal age given in the dataset, and a corresponding exact paternal age taken from $2.70 + 1.076G_M + e$, where e follows Normal(0, 4.5) (parameters obtained by ordinary linear regression on the binned parental ages in the dataset). We then simulated the paternal DNM count as a Poisson random variable with expectation of either $1.51G_P$+6.05 (as estimated by Jónsson et al. 2017) or $1.41G_P$+5.56 (estimated by our maximum likelihood model) and ran Poisson regression or pairwise analysis on the mutation counts and integer parts of parental ages, as described above. The simulated data generated either a greater or equal maternal age effect on paternal mutations by Poisson regression or a more z-score of Kendall's tau-b statistic as significant or more significant in about 3.5% of 10,000 replicates and both in ~0.7% (Table S8).

Although C>A mutations only constitute 8% of all DNMs, for this mutation type, we found a significant effect of the maternal age (p= 0.020) and a slight improvement in the fit compared to a model with paternal age only (ΔAIC=-3.05; approximately 4.6-fold more probable) by Poisson regression (with identity link) of the number of paternal mutations. More surprisingly, the point estimate of the maternal age effect on paternal genome (0.095; se=0.041) is even stronger than that of the paternal age (0.057, se=0.033) and also stronger than the effect of maternal age on maternal genome (0.024, se=0.0094 by Poisson regression of maternal mutations on maternal age). To test the significance of this finding, we used simulations to examine whether the observations of C>A can happen by chance, conditional on the maternal age effect on paternal mutations on overall DNMs. We focused on 199 trios with >95% DNMs phased and simulated data with two schemes (1) randomly subsampling 8.3% paternal DNMs as C>A mutations for each trio, and (2) shuffling the mutation type labels across all paternal DNMs of the 199 trios. We then ran Poisson regression on the simulated paternal C>A mutation counts and found that in only 4.5% of the 20,000 replication, the model with maternal age would provide a better fit with ΔAIC<-3 and a greater point estimate of maternal age effect than paternal age effect (Table S9). These results suggest that paternal C>A mutations are more strongly affected by maternal age compared to other DNMs. In addition, the fraction of C>A DNMs is higher in paternal mutations than in maternal ones (constituting 8.3% paternal DNMs vs 6.2% maternal) (Jónsson, Sulem, Kehr, et al. 2017), potentially reflecting DNA oxidative stress in spermatogenesis and lack of a complete base excision repair pathway in spermatozoa (De Iuliis et al. 2009; Smith et al. 2013). We also noted that this mutation type is 15-20% under-represented than expected from rare variants (present in 3-9 copies) in 7,509 non-Finish Europeans in the gnomAD dataset, after accounting for GC content and effect of GC-biased gene conversion (Gao, Moorjani and Przeworski, unpublished), which is consistent with an under-detection of early embryonic mutations by standard trio approach.

Differences in mutation properties of trios with or without a third generation in Jónsson et al (2017)

Motivated by the significant maternal age effect on paternal mutations detected in 199 trios with >95% DNMs phased, we added such an effect in our maximum likelihood model for all 1548 trios but found no improvement in fit of the model (Table S10). We noted that all trios with >95% phasing rate were families with a third generation, whereas all trios without a third generation had low phasing rate (32% on average), based only on read tracing-based phasing. Applying the maximum likelihood model to all 225 trios with a third generation (phasing rate >80%) still provided support for a maternal-on-paternal effect (Table S10), suggesting mutation properties differ between trios with or without a third generation. In fact, DNMs were identified in different ways in three-generation and two-generation families: a large fraction of DNM candidates in three-generation families were directly validated or invalidated by transmission to the next generation, whereas DNMs in two-generation families were inferred from a candidate pool by a generalized additive model trained on the true positive and false positive calls in the three-generation families. Therefore, error rates in DNM calling are likely to be higher for trios without a third generation and may blur the subtle signals of maternal-on-paternal effect, especially in the face of large sampling variance, low phasing rate and high correlation between maternal and paternal ages. The false discovery rate of this study was estimated to be 3%, and the false negative rate at least 4% (Jónsson, Sulem, Kehr, et al. 2017; Jónsson, Sulem, Arnadottir, et al. 2017).

Consistent with this hypothesis, we observed substantial differences in the age and sex dependencies of DNMs between the two subsets of families (225 with a third generation and 1323 without) by maximum likelihood inference and Poisson regression of the total DNM count (Table S10). In principle, a Poisson regression of the total number of mutations on both parental ages should correctly assign a maternal-on-paternal effect, if there exists one, to maternal age. Yet, applying this method separately to two-generation and three-generation families, we found that the G_P slope is much higher in two generations families (1.47 vs 1.17), and G_M slope is much lower (0.32 vs 0.66). To assess the significance of these differences, we considered 1,000 random subsets of the two-generation families of the same size (225) and similar or higher correlation between G_P and G_M (Pearson's R=0.84) as the three-generation families, and found that 1.9% replicates produced estimates of G_P slope lower than the estimate in three-generation families in Poisson regression of total DNM count (2.7% when subsampling with replacement), suggesting the differences in sex and age dependencies of mutation rate is significant between two-gen and three-generation families. We found similar statistically significant differences considering C>A transversions only, despite the lower mutation counts (Table S11): while the difference in G_P slope is not significant (p=0.17), the difference in G_M slope is (p=0.015). Whether these differences between the two-generation and three-generation families are due to biological differences among individuals (Rahbari et al. 2016) or technical biases requires further investigation. Findings in three-generation families, however, strongly support a maternal age effect on paternal mutations, and suggest that previous estimates of the paternal age effect may have been soaking this effect, and should be corrected downwards.

Supplementary Tables

Table S1 Estimated parental age effects based on the maximum likelihood model and comparison to estimates from Jónsson et al (2017). One important distinction between the two models is that, to take into account the incomplete parental origin information, we explicitly modeled the phasing process as a binomial sampling of DNMs with a proband-specific phasing rate parameter, assuming that the phasing probabilities of all mutations in the same individual are identical and independent. This approach enabled us to fully leverage information of phased and unphased mutations (see in "Estimation of sex-specific mutation parameters with a model-based approach" section for more information).

Mutation type	Method	β_P	β_M	α_P	α_M
All point mutations	MLE	1.41	0.39	5.50	2.04
	Jónsson et al (2017)	1.51	0.37	6.05	3.61
C>A	MLE	0.11	0.023	0.42	0.18
	Jónsson et al (2017)	0.10	0.040	0.73	0.25
C>G*	MLE	0.14	0.073	0.52	-0.92
	Jónsson et al (2017)	0.12	0.09	0.85	-0.80
C>T at nonCpG sites	MLE	0.29	0.095	2.76	1.12
	Jónsson et al (2017)	0.29	0.09	2.23	1.75
C>T at CpG sites	MLE (excluding sites in GC islands)	0.25	0.038	0.75	1.22
	Jónsson et al (2017) (including sites in GC islands)	0.24	0.04	0.71	1.71
T>A	MLE	0.085	0.030	0.73	0.032
	Jónsson et al (2017)	0.07	0.04	1.27	0.21
T>C	MLE	0.40	0.11	0.75	0.36
	Jónsson et al (2017)	0.39	0.12	0.41	0.83
T>G	MLE	0.13	0.024	-0.45	0.044
	Jónsson et al (2017)	0.12	0.03	0.14	0.21

* A model with exponential maternal age effect and linear paternal age effect was used for downstream analyses (see Table S3 for the parameter estimates).

Table S2 Summary statistics of the two DNM data sets

	Goldmann	Jónsson
Number of trios	816	1548
Average number of DNMs per proband	43.86	63.86
% DNMs phased (% by informative flanking variant in the read)	20.2% (20.2%)	41.5% (31.6%)
Estimated number of callable base pairs	1.62Gb[1]	2.68G
Average paternal age	33.65	32.02
Average maternal age	31.51	28.18
Correlation between parental ages (Pearson's R)	0.72	0.78
Estimated paternal age effect (slope) for all DNMs	0.92	1.41
Estimated maternal age effect (slope) for all DNMs	0.24	0.39
Ratio of paternal to maternal slope	3.78	3.58
%increase in DNM for one year increase in reproductive age in both sexes	2.64%	2.83%

[1] Estimated in the study of Wong et al. (2016), which included 693 of the 816 trios in Goldmann et al. (2016).

Table S3 Comparison of models with linear and exponential parental age effects (see accompanying excel spreadsheet). We took ΔAIC=-6 as the threshold for evidence of a significant better fit (approximately 20-fold more probable).

Table S4 Comparison of linear models fitted to trios with different maternal ages (see accompanying excel spreadsheets)

Table S5 Comparison of models with linear and exponential parental age effects fitted to all trios and trios with maternal age below 40 (see accompanying excel spreadsheets)

Table S6 Co-occurrence of *de novo* C>Gs and indels on the same chromosome. Shown in the table are the numbers of C>G and other point mutations that co-occur with an indel on the same chromosome in the same individual. Conditional on occurrence on the same chromosome and within 10Mb, C>Gs are also closer to deletions ≥5bp than are other mutation types. See Fig S6 for a comparison between C>G and other point mutations in the distance to the closest deletions of ≥5bp conditional on co-occurrence. Indels can arise from non-homologous end joining (NHEJ) or microhomology mediated end joining (MMEJ) repair of DSBs and polymerase slippage during replication, but the former mechanism is more likely to lead to deletions of intermediate size (Montgomery et al. 2013; Kloosterman et al. 2015), so the highly significant association of C>G DNMs with deletions greater than 4bp points to DSBs as the main source of both.

		Co-occurrence with a deletion≥5bp ($p < 2.2e-16$)		Co-occurrence with a deletion<5bp ($p = 0.702$)		Co-occurrence with an insertion≥5bp ($p = 0.168$)		Co-occurrence with an insertion<5bp ($p = 0.0554$)	
	Total number	Number	Prob.	Number	Prob.	Number	Prob.	Number	Prob.
C>G	9467	321	0.0339	1441	0.152	91	0.00961	656	0.0693
Non-C>G point mutations	89391	2529	0.0283	13745	0.154	733	0.00820	5734	0.0641

P-values were calculated based on Chi-square test for independence between (C>G vs. not) and (co-occurrence with an indel vs. not).

Table S7 Estimates of maternal age effect on paternal mutations by different methods
Poisson regression (with identity link)

Response variable	Explanatory variable (s)	Slope of G_P	SE of slope of G_P	p-value of slope of G_P	Slope of G_M	SE of slope of G_M	p-value of slope of G_M	AIC
Paternal mutation count	G_p, G_M	1.16	0.12	< 2e-16	0.30	0.14	0.035	1419.5
Paternal mutation count	G_P	1.37	0.062	< 2e-16	--	--	--	1422.0
Maternal mutation	G_P, G_M	-0.075	0.055	0.17	0.42	0.069	1.13e-09	1141.2

count								
Maternal mutation count	G_M	--	--	--	0.34	0.038	< 2e-16	1141.0

Maximum likelihood approach

Model	β_P	β_M	α_P	α_M	β_{Mp} *	Log likelihood	AIC
Model 0: Paternal mutation count ~ Pois($\alpha_P + \beta_P G_P$) Maternal mutation count ~ Pois($\alpha_M + \beta_M G_M$)	1.39	0.34	8.33	3.18	--	-1435.4	2878.8
Model 1: Paternal mutation count ~ Pois($\alpha_P+\beta_P G_P+\boldsymbol{\beta_M G_M}$) Maternal mutation count ~ Pois($\alpha_M + \beta_M G_M$)	1.16	0.34	6.24	3.29	--	-1433.5	2875.1
Model 2: Paternal mutation count ~ Pois($\alpha_P+\beta_P G_P+\boldsymbol{\beta_{Mp} G_M}$) Maternal mutation count ~ Pois($\alpha_M + \beta_M G_M$)	1.20	0.34	6.60	3.20	0.29	-1433.5	2876.9

* β_{Mp} represents the effect of maternal age on paternal DNMs, when it is different than that on maternal mutations

Pairwise analysis (weighted linear regression, intercept forced to zero)

Condition	Response variable	Explanatory variable (s)	Weight	Slope	SE of slope$^+$	P-value of slope$^+$	One-tailed p-value by permutation
Same G_P	Difference in **paternal** mutation counts	G_M	$1/G_P$	0.37	0.093	6.52e-05	0.022
Same G_P	Difference in **maternal** mutation counts	G_M	$1/(G_M1+G_M2)$*	0.38	0.047	1.69e-15	0.0033
Same G_M	Difference in **paternal** mutation counts	G_P	$1/(G_P1+G_P2)$*	1.01	0.069	<2e-16	<1e-4
Same G_M	Difference in **maternal** mutation counts	G_P	$1/G_M$	-0.019	0.033	0.57	0.31

* G_M1 and G_M2 are the maternal ages of the two probands in the pair with the same paternal age; similar for G_P1 and G_P2.

$^+$ Note that the standard errors and p-values may be unreliable due to violation of the linear regression assumptions.

Deviation analysis (weighted linear regression, intercept forced to zero)

Condition	Response variable	Explanatory variable (s)	Weight	Slope	SE of slope$^+$	p-value of slope$^+$	P-value by permutation

Same G_P	Deviation in **paternal** mutation count	Deviation in G_M	$1/G_P$	0.38	0.17	0.025	0.018
Same G_M	Difference in **maternal** mutation counts	Deviation in G_P	$1/G_M$	-0.10	0.065	0.13	0.088

[+] Note that the standard errors and p-values may be unreliable due to violation of the linear regression assumptions.

Table S8 Estimating the probability of a spurious maternal age effect on paternal mutations

Paternal mutations*	Parental ages for analysis	Number of replicates	Poisson regression: ΔAIC<-2.4 & maternal slope>0.3	Pairwise analysis: z-score of tau-b>3.1	Both
Poisson(1.51G_P'+6.05)	Integer part of simulated ages	10000	208	220	67
Poisson(1.41G_P'+5.56)	Integer part of simulated ages	10000	183	169	47

* G_P' is the exact paternal age used in simulations

Table S9 Estimating the probability of a stronger maternal age effect on paternal C>A mutations than paternal age effect

Simulation scheme	Number of replicates	ΔAIC<-3	ΔAIC<-3 & G_M slope > G_P slope	Rate
Subsample 8.3% paternal DNMs as C>A for each trio	20,000	926	900	4.50%
Shuffle of mutation type labels across all paternal DNMs	20,000	935	897	4.49%

Table S10 Differences in the age and sex dependencies of DNMs between the trios with or without a third generation

Poisson regression (with identity link) of total DNM count on G_M and G_P

	# Trios	G_P slope (SE)	P-value of G_P	G_M slope (SE)	P-value of G_M	Intercept (SE)	Ratio of point estimates of two slopes
Decode all	1548	1.44 (0.040)	< 2e-16	0.35 (0.052)	1.53e-11	7.89 (0.88)	4.11
225 3-gen trios	225	1.17 (0.12)	< 2e-16	0.66 (0.15)	1.53e-05	9.98 (2.30)	1.78
1323 2-gen trios	1323	1.47 (0.042)	< 2e-16	0.32 (0.055)	5.16e-09	7.45 (0.96)	4.53

Maximum likelihood inference

	Model	β_P	β_M	α_P	α_M	β_{Mp}	Log likelihood	AIC
All 1548 trios	**No maternal-on-paternal effect**	**1.41**	**0.39**	**5.50**	**2.04**	**--**	**-12174.3**	**24356.6**
	With maternal-on-paternal effect	1.41	0.39	5.50	2.04	0.00	-12174.3	24358.6

225 3-gen trios	No maternal-on-paternal effect	1.40	0.33	7.95	3.54	--	-1634.1	3276.1
	With maternal-on-paternal effect	**1.22**	**0.33**	**6.37**	**3.55**	**0.26**	**-1632.2**	**3274.5**
1323 2-gen trios	**No maternal-on-paternal effect**	**1.41**	**0.42**	**5.08**	**1.63**	**--**	**-10524.4**	**21056.8**
	With maternal-on-paternal effect	1.41	0.41	5.08	1.76	0.00	-10524.4	21058.8

Bold font indicates the model with a better fit.

Table S11 Differences in the age and sex dependencies of C>A DNMs between the trios with or without a third generation

Poisson regression (with identity link) of total DNM count on G_M and G_P

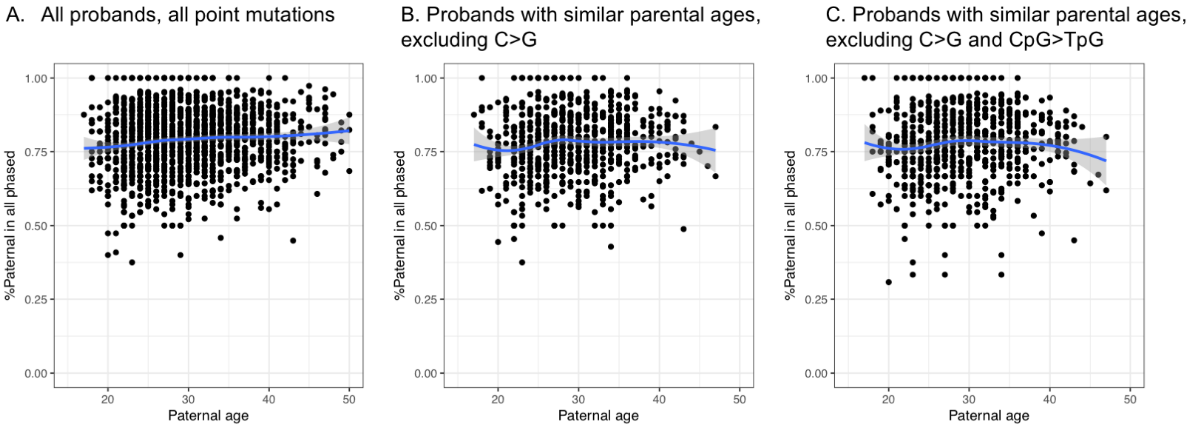
	# Trios	G_P slope (SE)	P-value of G_P	G_M slope (SE)	P-value of G_M	Intercept (SE)	Ratio of point estimates of two slopes
Decode all	1548	0.10 (0.011)	< 2e-16	0.039 (0.014)	0.0059	0.48 (0.24)	2.62
225 3-gen trios	225	0.080 (0.035)	0.021	0.10 (0.043)	0.014	-0.29 (0.64)	0.77
1323 2-gen trios	1323	0.11 (0.012)	< 2e-16	0.031 (0.015)	0.042	0.60 (0.26)	3.39

Maximum likelihood inference

	Model	β_P	β_M	α_P	α_M	β_{Mp}	Log likelihood	AIC
All 1548 trios	**No maternal-on-paternal effect**	**0.11**	**0.023**	**0.42**	**0.18**	**--**	**-5334.4**	**10676.8**
	With maternal-on-paternal effect	0.10	0.022	0.26	0.21	0.018	-5333.6	10677.1
225 3-gen trios	No maternal-on-paternal effect	0.14	0.023	-0.04	0.16	--	-764.5	1537.1
	With maternal-on-paternal effect	**0.083**	**0.023**	**-0.45**	**0.15**	**0.079**	**-762.7**	**1535.3**
1323 2-gen trios	**No maternal-on-paternal effect**	**0.11**	**0.024**	**0.48**	**0.18**	**--**	**-4565.9**	**9139.7**
	With maternal-on-paternal effect	0.11	0.023	0.39	0.21	0.009	-4565.7	9141.4

Bold font indicates the model with a better fit.

Supplementary Figures


A. All probands, all point mutations

B. Probands with similar parental ages, excluding C>G

C. Probands with similar parental ages, excluding C>G and CpG>TpG

Figure S1. Fraction of paternal mutations among phased mutations as a function of paternal age. Each point represents the data for one child (proband) in the Icelandic data set (Jónsson et al. 2017) with at least three phased mutations under the corresponding category. The blue line is the LOESS curve fitted to the scatterplot, with the shaded area representing the 95% confidence interval of the LOESS curve (calculated with the geom_smooth function in R). (A) For all probands (children) with at least three phased point mutations; (B) For probands with similar parental ages ($0.9<G_P/G_M<1.1$) and at least three non-C>G phased DNMs; (C) For probands with similar parental ages ($0.9< G_P/G_M<1.1$) and at least three DNMs that are not C>G or CpG>TpG.

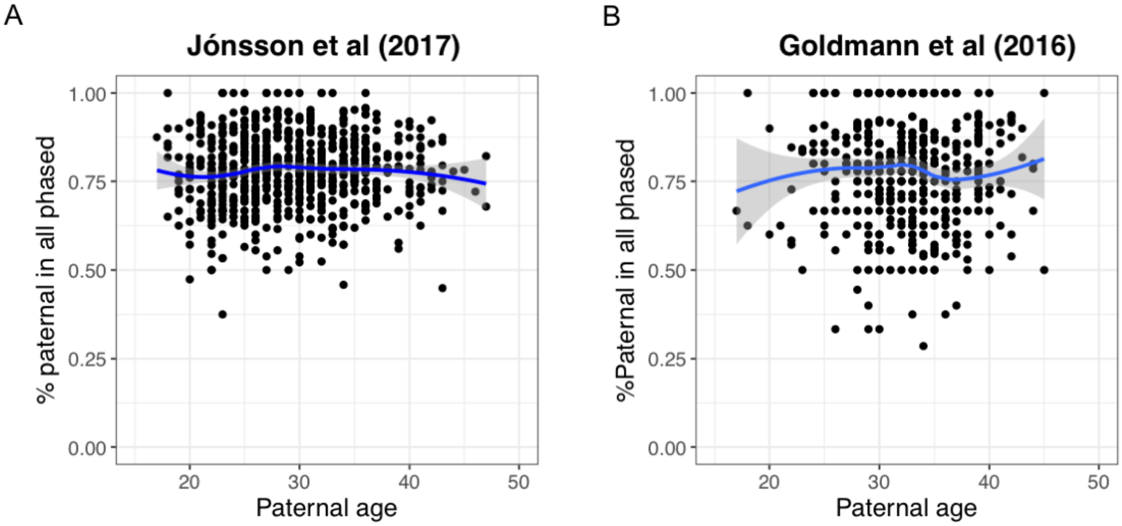


Figure S2. Replication of the stable fraction of paternal mutations with paternal age in an independent dataset. Each point represents the data for one child (proband) with at least three phased point mutations and similar parental ages (paternal-to-maternal age ratio between 0.9 to 1.1; 719 trios total). The blue line is the LOESS curve fitted to the scatterplot, with the shaded area representing the 95% confidence interval of the LOESS curve. (A) Same as Figure 1; (B) Similar plot for data from Goldmann et al. (2016), which includes a total of 35,793 DNMs (7,216 of which were phased). See SOM for details.

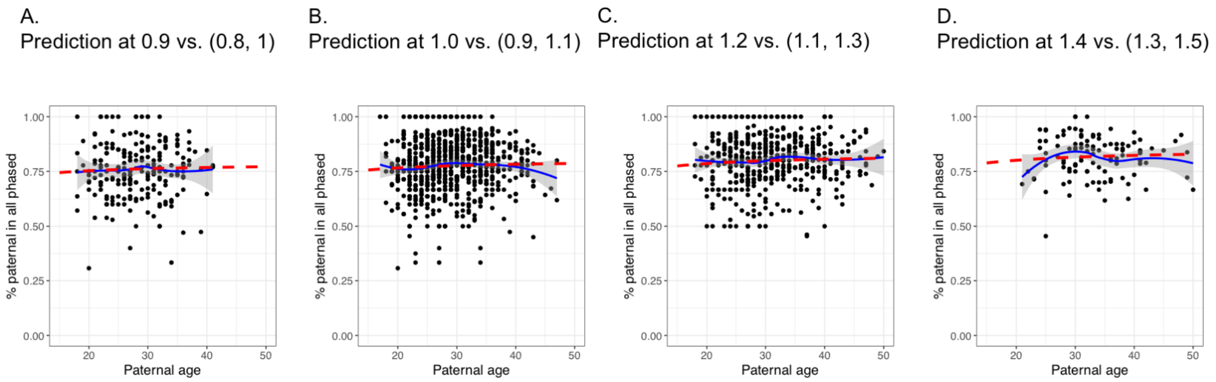


Figure S3. Fraction of paternal mutations among phased mutations for different ratios of paternal age (G_P) to maternal age (G_M). Each point represents the data for one child (proband) with at least three phased point mutations that are not C>G or CpG>TpG in the Icelandic data set (Jónsson et al. 2017). The blue line is the LOESS curve fitted to the scatterplot, with the shaded area representing the 95% confidence interval of the LOESS curve. The red dashed line is the prediction based on estimated parental age effects on mutation rate from our maximum likelihood model. (A) Data for probands with $0.8<G_P/G_M<1$ versus prediction for $G_P/G_M=0.9$; (B) Data for probands with $0.9<G_P/G_M<1.1$ versus prediction for $G_P/G_M=1$; (C) Data for probands with $1.1<G_P/G_M<1.3$ versus prediction for $G_P/G_M=1.2$; (D) Data for probands with $1.3<G_P/G_M<1.5$ versus prediction for $G_P/G_M=1.4$.

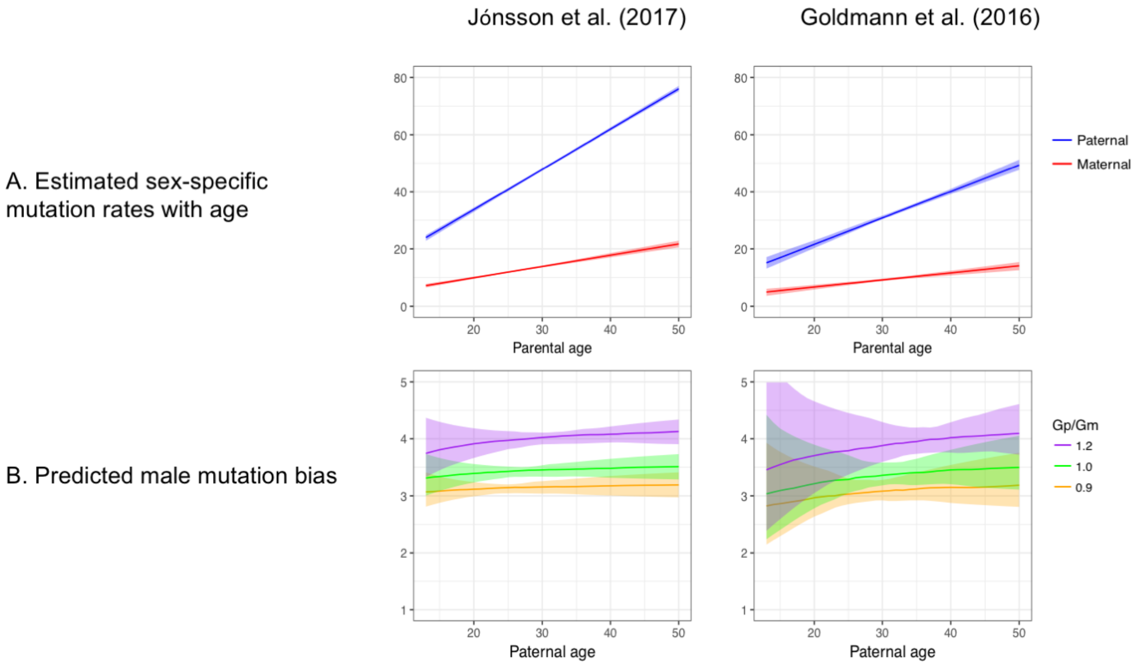


Figure S4. Comparison of parental age effects and predicted male-to-female mutation ratio at given ages estimated from two DNM datasets. The two data sets differ in their sample sizes (1548 vs 816 trios), average numbers of detected DNMs per proband (63.86 vs 43.86) and the fraction of DNMs that were phased (41.5% vs 20.2%), which lead to different absolute effects of parental ages on the count of DNMs (A). Despite all these differences, the male-to-female mutation ratio is inferred to be stable with paternal age for both data sets. (A) Estimated sex-specific mutation rates with paternal age; (B) Predicted male-to-female mutation ratio.

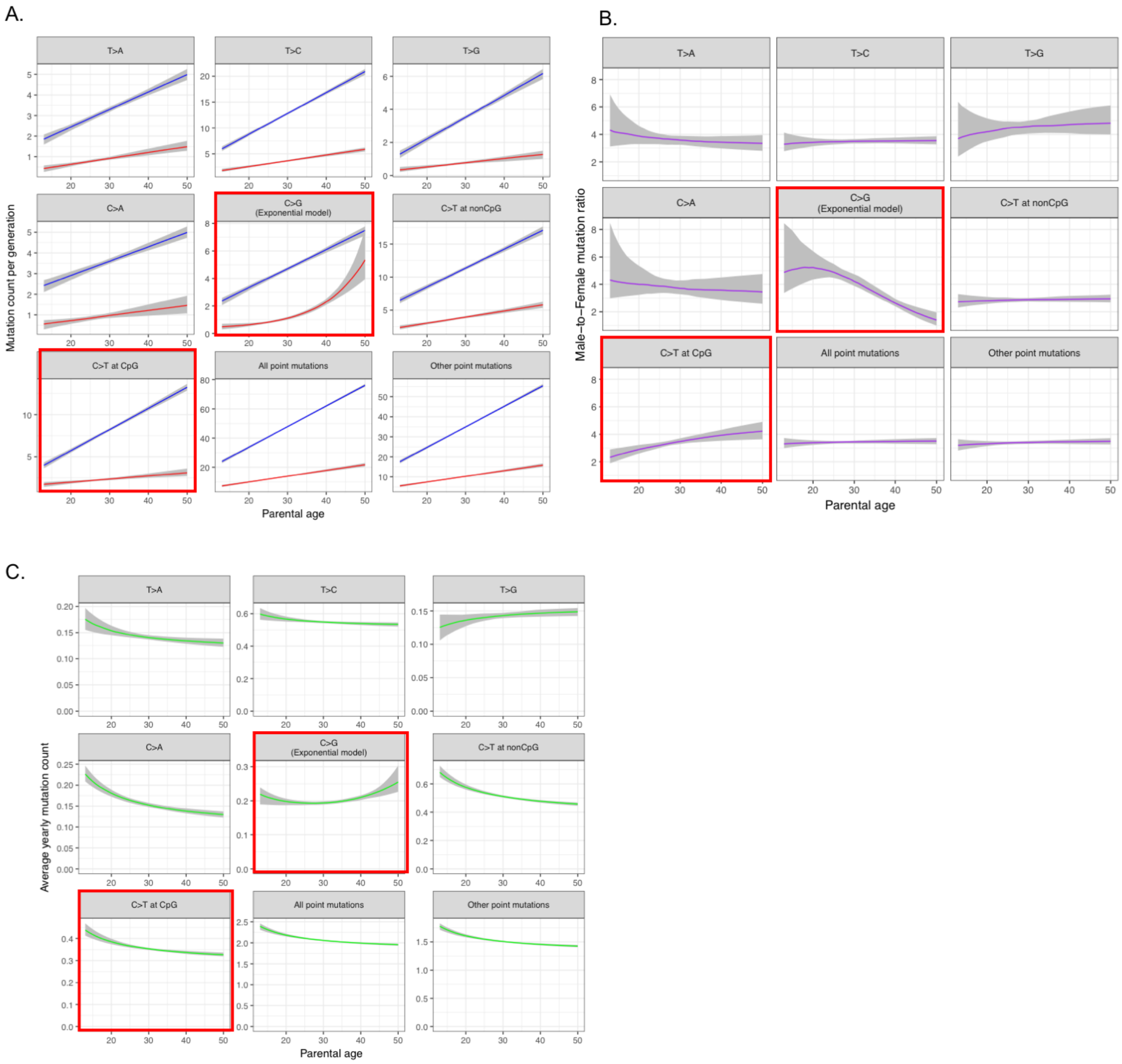


Figure S5. Estimated sex-specific mutation rates and male-to-female mutation ratio as a function of parental ages, by mutation type. Red boxes indicate the two mutation types highlighted in the main text. The extent of male mutational bias and average yearly mutation rate are estimated assuming the same paternal and maternal age. “Other point mutations” refers to point mutations excluding C>G and CpG>TpG mutations. (A) Estimated paternal and maternal mutation rates per generation; (B) Estimated male-to-female mutation ratio; (C) Estimated average yearly mutation rates.

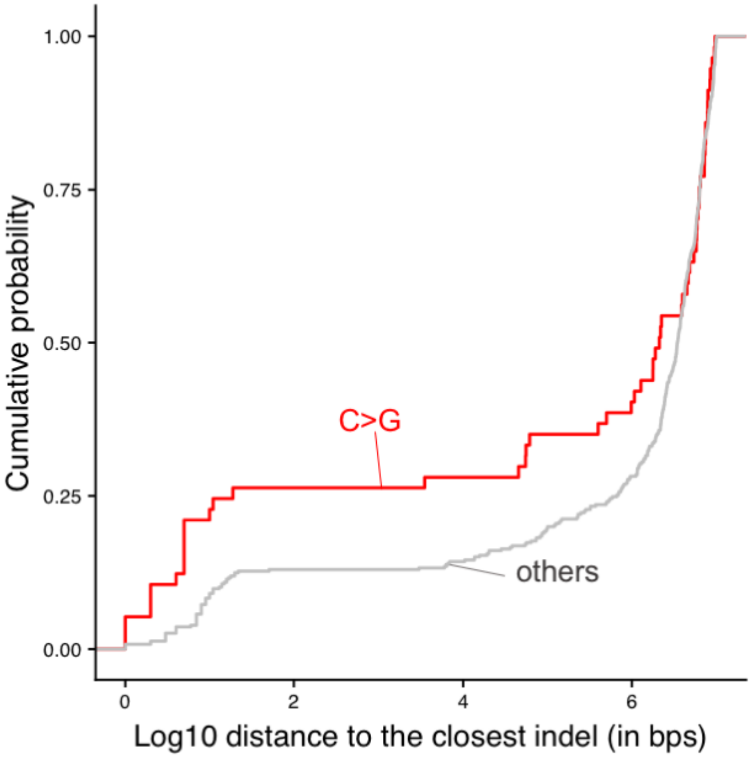


Figure S6. Distribution of distances to the closest deletion of ≥5bp for C>G mutations. Cumulative distribution of distance to the closest de novo deletion (≥ 5bp) for C>G transversions and for other point mutations, conditional on co-occurrence within 10Mb.

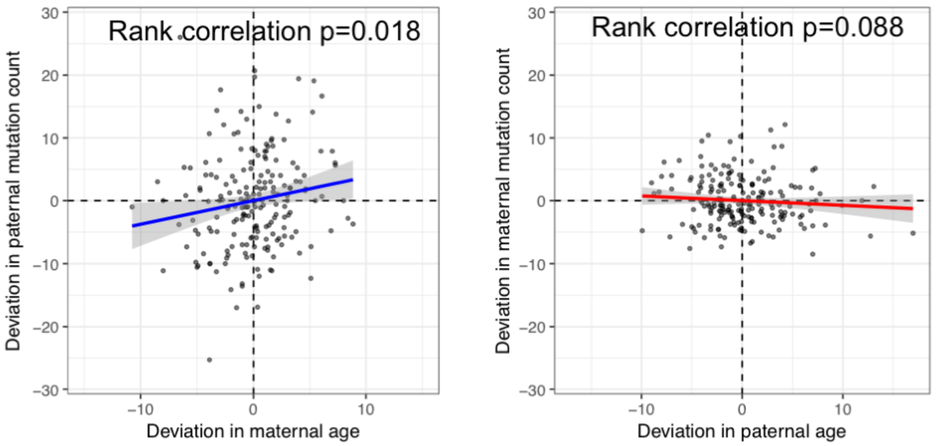


Figure S7. Effect of maternal age on the number of paternal DNMs by deviation analysis. Each point represents one proband (in which almost all mutations phased). The blue (or red) line is the weighted linear regression line (see SOM for details), with the shaded area representing the 95% confidence interval of the slope.

References

Braude, Peter, Virginia Bolton, and Stephen Moore. 1988. "Human Gene Expression First Occurs between the Four- and Eight-Cell Stages of Preimplantation Development." *Nature* 332(6163): 459–61.

Dobson, Anthony T, Rajiv Raja, Michael J Abeyta, Theresa Taylor, Shehua Shen, Christopher Haqq, and Renee A Reijo Pera. 2004. "The Unique Transcriptome through Day 3 of Human Preimplantation Development." *Human Molecular Genetics* 13 (14): 7–9. https://doi.org/10.1093/hmg/ddh157.

Fulton, N, S J Martins Silva, R A L Bayne, and R A Anderson. 2005. "Germ Cell Proliferation and Apoptosis in the Developing" 90 (May): 4664–70. https://doi.org/10.1210/jc.2005-0219.

Goldmann, Jakob M., Wendy S.W. Wong, Michele Pinelli, Terry Farrah, Dale Bodian, Anna B. Stittrich, Gustavo Glusman, et al. 2016. "Parent-of-Origin-Specific Signatures of de Novo Mutations." *Nature Genetics* 48 (8): 935–39. https://doi.org/10.1038/ng.3597.

Harland, Chad, Carole Charlier, Latifa Karim, Nadine Cambisano, Manon Deckers, Myriam Mni, Erik Mullaart, Wouter Coppieters, and Michel Georges. 2017. "Frequency of Mosaicism Points towards Mutation-Prone Early Cleavage Cell Divisions." *BioRxiv*.

Iuliis, Geoffry N. De, Laura K. Thomson, Lisa A. Mitchell, Jane M. Finnie, Adam J. Koppers, Andrew Hedges, Brett Nixon, and R. John Aitken. 2009. "DNA Damage in Human Spermatozoa Is Highly Correlated with the Efficiency of Chromatin Remodeling and the Formation of 8-Hydroxy-2′-Deoxyguanosine, a Marker of Oxidative Stress1." *Biology of Reproduction* 81 (3): 517–24. https://doi.org/10.1095/biolreprod.109.076836.

Jónsson, Hákon, Patrick Sulem, Gudny Arnadottir, Gunnar Pálsson, Eggertsson Hannes, Snaedis Kristmundsdottir, Florian Zink, Birte Kehr, and etc. 2017. "Recurrence of de Novo Mutations in Families." *BioRxiv*.

Jónsson, Hákon, Patrick Sulem, Birte Kehr, Snaedis Kristmundsdottir, Florian Zink, Eirikur Hjartarson, Marteinn T. Hardarson, et al. 2017. "Parental Influence on Human Germline de Novo Mutations in 1,548 Trios from Iceland." *Nature* 549 (7673). Nature Publishing Group: 519–22. https://doi.org/10.1038/nature24018.

Kloosterman, Wigard P, Laurent C Francioli, Fereydoun Hormozdiari, Tobias Marschall, Jayne Y Hehir-kwa, Abdel Abdellaoui, Eric-wubbo Lameijer, et al. 2015. "Characteristics of de Novo Structural Changes in the Human Genome," 792–801. https://doi.org/10.1101/gr.185041.114.19.

Lindsay, Sarah J., Raheleh Rahbari, Joanna Kaplanis, Thomas M. Keane, and Matthew E. Hurles. 2016. "Striking Differences in Patterns of Germline Mutation between Mice and Humans." *BioRxiv*.

Montgomery, Stephen B, David L Goode, Erika Kvikstad, Cornelis A Albers, Zhengdong D Zhang, Xinmeng Jasmine Mu, Guruprasad Ananda, et al. 2013. "The Origin , Evolution , and Functional Impact of Short Insertion – Deletion Variants Identified in 179 Human Genomes," 749–61. https://doi.org/10.1101/gr.148718.112.Freely.

Moorjani, Priya, Carlos Eduardo G. Amorim, Peter F. Arndt, and Molly Przeworski. 2016. "Variation in the Molecular Clock of Primates." *Proceedings of the National Academy of Sciences* 113 (38): 10607–12. https://doi.org/10.1073/pnas.1600374113.

Polani, Paul E, and John A Crolla. 1991. "A Test of the Production Line Hypothesis of Mammalian Oogenesis," 64–70.

Rahbari, Raheleh, Arthur Wuster, Sarah J. Lindsay, Robert J. Hardwick, Ludmil B. Alexandrov, Saeed Al Turki, Anna Dominiczak, et al. 2016. "Timing, Rates and Spectra of Human

Germline Mutation." *Nature Genetics* 48 (2). Nature Publishing Group: 126–33. https://doi.org/10.1038/ng.3469.

Smith, T. B., M. D. Dun, N. D. Smith, B. J. Curry, H. S. Connaughton, and R. J. Aitken. 2013. "The Presence of a Truncated Base Excision Repair Pathway in Human Spermatozoa That Is Mediated by OGG1." *Journal of Cell Science* 126 (6): 1488–97. https://doi.org/10.1242/jcs.121657.

Wong, Wendy S. W., Benjamin D. Solomon, Dale L. Bodian, Prachi Kothiyal, Greg Eley, Kathi C. Huddleston, Robin Baker, et al. 2016. "New Observations on Maternal Age Effect on Germline de Novo Mutations." *Nature Communications* 7 (May 2015). Nature Publishing Group: 10486. https://doi.org/10.1038/ncomms10486.

Zhang, Pu, Marco Zucchelli, Sara Bruce, Fredwell Hambiliki, Anneli Stavreus-evers, and Lev Levkov. 2009. "Transcriptome Profiling of Human Pre-Implantation Development" 4 (11). https://doi.org/10.1371/journal.pone.0007844.