Supporting Information for

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

Ziyue Gao[1,*], Priya Moorjani[2,3], Thomas A. Sasani[4], Brent S. Pedersen[4], Aaron R. Quinlan[4,5], Lynn B. Jorde[4], Guy Amster[6,†] and Molly Przeworski[6,7, †,*]

[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 Human Genetics, University of Utah School of Medicine.
[5] Department of Biomedical Informatics, University of Utah School of Medicine.
[6] Department of Biological Sciences, Columbia University.
[7] Department of Systems Biology, Columbia University.
† Contributed equally
* Correspondence to: ziyuegao@stanford.edu or mp3284@columbia.edu

Table of Contents

Supplementary Figures

Supplementary Notes

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:

$L^i = \mathrm{P}(Y_M{}^i = y_M{}^i, Y_P{}^i = y_P{}^i, Y_U{}^i = y_U{}^i \mid \beta_{0,M}, \beta_{0,P}, \beta_M, \beta_P, G_M{}^i, G_P{}^i, p^i)$

$= \mathrm{P}(y_M{}^i, y_P{}^i, y_U{}^i \mid X_M{}^i, X_P{}^i, p^i)\mathrm{P}(X_M{}^i \mid \beta_{0,M}, \beta_M, G_M{}^i)\mathrm{P}(X_P{}^i \mid \beta_{0,P}, \beta_P, G_P{}^i)$

$=\sum_{k=0}^{y_U{}^i} P\left((y_M{}^i, y_P{}^i, y_U{}^i \mid X_M{}^i = y_M{}^i + k,\ X_P{}^i = y_P{}^i + y_U{}^i - k, p^i\right) P\left(y_M{}^i + k \mid \beta_{0,M}, \beta_M, G_M{}^i\right) P\left(y_P{}^i + y_U{}^i - k \mid \beta_{0,P}, \beta_P, G_P{}^i\right)$

The first term of the addend can be expressed as:

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

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

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

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

The second term of the addend is:

$$P\left(y_M{}^i + k \middle| \beta_{0,M}, \beta_M, G_M{}^i\right) = Poisson\left(y_M{}^i + k \middle| \lambda_M{}^i\right) = \frac{(\lambda_M{}^i)^{y_M{}^i + k} e^{-\lambda_M{}^i}}{(y_M{}^i + k)!}$$

Similarly, the third term of the addend is:

$$P\left(y_P{}^i + y_U{}^i - k \middle| \beta_{0,P}, \beta_P, G_P{}^i\right) = Poisson\left(y_P{}^i + y_U{}^i - k \middle| \lambda_P{}^i\right) = \frac{(\lambda_P{}^i)^{(y_P{}^i + y_U{}^i - k)} e^{-\lambda_P{}^i}}{(y_P{}^i + y_U{}^i - k)!}$$

Therefore, the addend can be written together as:

$$\frac{p^{i(y_M{}^i + y_P{}^i)}(1-p^i)^{y_U{}^i}}{y_M{}^i!\, y_P{}^i!} \frac{1}{k!\,(y_U{}^i - k)!}(y_M{}^i + k)!\,(y_P{}^i + y_U{}^i - k)! \frac{(\lambda_M{}^i)^{y_M{}^i + k} e^{-\lambda_M{}^i}}{(y_M{}^i + k)!} \frac{(\lambda_P{}^i)^{y_P{}^i + y_U{}^i - k} e^{-\lambda_P{}^i}}{(y_P{}^i + y_U{}^i - k)!}$$

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

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

$$\frac{p^{i(y_M{}^i + y_P{}^i)}(1-p^i)^{y_U{}^i}(\lambda_M{}^i)^{y_M{}^i}(\lambda_P{}^i)^{y_P{}^i} e^{-(\lambda_M{}^i + \lambda_P{}^i)}}{y_M{}^i!\, y_P{}^i!} \times \frac{(\lambda_M{}^i)^k (\lambda_P{}^i)^{y_U{}^i - k}}{k!\,(y_U{}^i - k)!}$$

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

$$1 = \sum_{k=0}^{y_U{}^i} \frac{y_U{}^i}{k!\,(y_U{}^i - k)!}\left(\frac{\lambda_M{}^i}{\lambda_M{}^i + \lambda_P{}^i}\right)^k \left(\frac{\lambda_P{}^i}{\lambda_M{}^i + \lambda_P{}^i}\right)^{y_U{}^i - k}$$

We obtain:

$$\sum_{k=0}^{y_U{}^i} \frac{(\lambda_M{}^i)^k (\lambda_P{}^i)^{y_U{}^i - k}}{k!\,(y_U{}^i - k)!} = \frac{(\lambda_M{}^i + \lambda_P{}^i)^{y_U{}^i}}{y_U{}^i!}$$

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

$$L_i = \frac{{p^i}^{({y_M}^i + {y_P}^i)}(1-p^i)^{{y_U}^i}({\lambda_M}^i)^{{y_M}^i}({\lambda_P}^i)^{{y_P}^i}e^{-({\lambda_M}^i+{\lambda_P}^i)}}{{y_M}^i!\,{y_P}^i!}\frac{({\lambda_M}^i+{\lambda_P}^i)^{{y_U}^i}}{{y_U}^i!}$$

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 (1) and may experience fewer mitoses (2). 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 (3, 4). 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 (5).

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 (6–8). 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 in SI Appendix, 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 (5). 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).

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

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 excluded) based on 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.

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 S11). 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 GP slope is much higher in two generations families (1.47 vs 1.17), and the GM 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 GP and GM (Pearson's R=0.84) as the three-generation families, and found that 1.9% replicates produced estimates of GP slope lower than the estimate in three-generation families in Poisson regression of total DNM count (2.7% when subsampling with replacement), suggesting the sex and age dependencies of mutation rate differ unexpectedly between two-generation and three-generation families. We found similar statistically significant differences when considering C>A transversions only, despite their lower numbers (Table S12): the difference in GP slope is not significant (p=0.17), but the difference in GM slope is (p=0.015). Findings in three-generation families, however, support a maternal age effect on paternal mutations, suggesting that previous estimates of the paternal age effect may have been soaking up part of the maternal effect and should consequently be corrected downwards.

Estimate of the power to detect a maternal age effect on the paternal mutation rate by simulation

We simulated paternal mutation counts for various paternal (y-axis) and maternal (x-axis) age effect sizes on paternal mutation rate, assuming that the mutation count is Poisson distributed, and quantified the fraction of simulations with a significant maternal age effect in 1,000 replicates. We assumed that the extrapolated intercept at G_P=0 and G_M=0 is six mutations in the assayable regions of a diploid genome (estimated to be 5.56 by deCODE and 6.05 by our analysis), i.e., $X_P \sim \text{Poisson}(6 + \beta_P G_P + \beta_{Mp} G_M)$. We performed simulations under different combinations of parental effect sizes: $\beta_P \in \{1.2, 1.3, 1.4, 1.5, 1.6\} \times \beta_{Mp} \in \{0, 0.1, 0.2, 0.3, 0.4\}$ with a sample size of 100, 200, 500, 1000 or 2000 trios. The criteria for a significant maternal age effect are (1) that the fit of a model with both parental ages is improved compared to a model with paternal age only by at least Δ AIC=-2.4; and (2) that the estimated effect of maternal age is positive, and the p-value is smaller than 0.05 in the model with both parental ages. We assumed no error in DNM calling, when error rates are in fact non-zero and post-zygotic mutations in particular are more likely to be missed. For the case of incomplete phasing, we took a phasing rate of 0.3, which is the value typically obtained for whole-genome trio data based on informative heterozygous sites in reads, and assumed identical and independent phasing probabilities across mutations and trios, regardless of parental origin. Based on the thinning property of Poisson distribution, the number of phased paternal mutations also follows a Poisson distribution with the product of the mutation rate and the phasing rate as the rate parameter. However, the actual phasing rate is likely to be variable across trios (depending on

trio-specific sequencing coverage, and other factors), which will introduce additional variation in the counts of phased mutations and further reduce the power. Given the overly optimistic assumptions about DNM calling and phasing in the simulations, the true power to detect a maternal age effect on paternal mutations is likely lower. As expected, the power to detect a maternal age effect on paternal mutations increases with the simulated effect size, phasing rate and the sample size. See main texts for a brief description of the results.

Detection of a maternal age effect on the rate of paternal C>A mutations

Although C>A mutations only constitute 8% of all DNMs, for this mutation type, we found a significant effect of the maternal age (p=0.02) 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 replicates, 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 (see SI Appendix, Table S10). 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 among paternal mutations than maternal ones (constituting 8.3% of paternal DNMs vs 6.2% of maternal ones)(9), potentially reflecting DNA oxidative stress in spermatogenesis and lack of a complete base excision repair pathway in spermatozoa (10, 11). For C>A mutations in the 199 probands with >95% DNMs phased, we did not observe enrichment in the GCA or TCT trinucleotide context reported in Harland *et al*.(4), possibly due to lack of power.

Supplementary Tables

Table S1 Test for an effect of parental age on the fraction of paternal mutations based on generalized linear models. In all these regressions, the response variable is specified as a two-column integer matrix: the first column is the count of paternal mutations and the second the count of maternal mutations (see in "Test for an effect of parental age on the male mutation bias" section in Methods for the rationale of the regression).

Regression of the counts of all single-base substitution DNMs on **paternal age** for 719 trios with $0.9<G_P/G_M<1.1$ in Jonsson et al (2017)(9)

Model	Link	Intercept	SE of intercept	Slope of G_P	SE of slope	*p*-value of slope	Predicted alpha at age 20	Predicted alpha at age 40
Binomial	Logit	1.172576	0.092631	0.003198	0.003017	0.289	3.44	3.67
Binomial	Identity	0.764669	0.015880	0.0005267	0.000515	0.306	3.45	3.67
Quasibinomial	Logit	1.172576	0.099969	0.003198	0.003256	0.326	3.44	3.67
Quasibinomial	Identity	0.764669	0.017139	0.0005267	0.000556	0.344	3.45	3.67

Regression of the counts of all single-base substitution DNMs on **average parental age** for 719 trios with $0.9<G_P/G_M<1.1$ in Jonsson et al (2017)(9)

Model	Link	Intercept	SE of intercept	Slope of G_P	SE of slope	*p*-value of slope	Predicted alpha at age 20	Predicted alpha at age 40
Binomial	Logit	1.220205	0.093721	0.001634	0.003086	0.597	3.50	3.62
Binomial	Identity	0.772623	0.016064	0.0002659	0.000528	0.614	3.50	3.61
Quasibinomial	Logit	1.220205	0.101207	0.001634	0.003332	0.624	3.50	3.62
Quasibinomial	Identity	0.772623	0.017347	0.0002659	0.000570	0.641	3.50	3.61

Regression of the counts of all single-base substitution DNMs on **paternal age** for 486 trios with $0.9<G_P/G_M<1.1$ in Goldmann et al (2016)(12)

Model	Link	Intercept	SE of intercept	Slope of G_P	SE of slope	*p*-value of slope	Predicted alpha at age 20	Predicted alpha at age 40
Binomial	Logit	1.335809	0.263760	-0.002255	0.007881	0.775	3.64	3.48
Binomial	Identity	0.792269	0.045239	-0.000394	0.001354	0.771	3.64	3.47
Quasibinomial	Logit	1.335809	0.269583	- 0.00226	0.008055	0.78	3.64	3.48
Quasibinomial	Identity	0.792269	0.046237	-0.000394	0.001384	0.776	3.64	3.47

Table S2 Estimated parental age effects based on the maximum likelihood model and comparison to estimates from Jónsson et al (2017)(9). 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 S3 Summary statistics of the two DNM data sets

	Inova	deCODE
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
Mean paternal age	33.65	32.02
Mean 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)(13), which included 693 of the 816 trios in Goldmann et al. (2016)(12).

Table S4 Comparison of models with linear and exponential parental age effects. We took ΔAIC=-6 as the threshold for evidence of a significant better fit (approximately 20-fold more probable).

Mutation type	Model				Estimates									
	Paternal effect	Maternal effect	Log likelihood	AIC	β_P	β_M	$\beta_{0,P}$	$\beta_{0,M}$	b_P	b_M	a_P	a_M	c_P	c_M
all point mutations	linear	linear	-12174.3	24356.6	1.41	0.39	5.5	2.0						
	linear	Exp.	-12158.4	24326.7	1.41		5.6			0.072		7.8		0.45
	Exp.	linear	-12174.6	24359.2		0.38		2.3	0.0051		-227		5.5	
	Exp.	Exp.	-12159.7	24331.4					0.0046	0.069	-252	7.6	5.6	0.54
C>A	linear	linear	-5334.4	10676.8	0.11	0.023	0.42	0.18						
	linear	Exp.	-5331.1	10672.3	0.11		0.42			0.14		0.67		-4.4
	Exp.	linear	-5335.6	10681.2		0.023		0.17	0.0075		-10.8		2.6	
	Exp.	Exp.	-5332.9	10677.8					0.0081	0.13	-9.5	0.65	2.4	-4.2
C>G	linear	linear	-5910.8	11829.6	0.14	0.073	0.52	-0.92						
	linear	Exp.	-5900.6	11811.3	0.14		0.56			0.10		0.35		-1.9
	Exp.	linear	-5906.5	11823.0		0.074		-0.94	0.022		-0.82		1.3	
	Exp.	Exp.	-5898.8	11809.6					0.011	0.090	-7.7	0.30	2.3	-1.7
C>T at nonCpG	linear	linear	-8495.5	16998.9	0.29	0.095	2.8	1.1						
	linear	Exp.	-8491.5	16993.0	0.29		2.8			0.078		2.6		-1.2
	Exp.	linear	-8496.0	17002.0		0.096		1.1	0.0087		-20.0		3.3	
	Exp.	Exp.	-8492.7	16997.5					0.0087	0.070	-19.7	2.4	3.3	-0.87
C>T at CpG	linear	linear	-7500.2	15008.4	0.25	0.038	0.75	1.2						
	linear	Exp.	-7499.8	15009.6	0.25		0.75			0.049		1.6		-1.1
	Exp.	linear	-7499.8	15009.7		0.039		1.2	0.0092		-17.7		3.1	
	Exp.	Exp.	-7499.7	15011.5					0.0096	0.055	-16.5	1.6	3.0	-1.3
T>A	linear	linear	-5171.5	10351.0	0.085	0.030	0.73	0.032						
	linear	Exp	-5171.5	10352.9	0.085		0.74			0.017		-0.87		0.29

	Exp.	linear	-5172.0	10353.9		0.029		0.043	0.0039		-18.2		3.0	
	Exp.	Exp.	-5171.8	10355.6					0.0025	0.0019	-30.6	-15.2	3.5	2.8
T>C	linear	linear	-8577.8	17163.7	0.40	0.11	0.75	0.36						
	linear	Exp.	-8577.8	17165.6	0.40		0.84			0.0029		-37.4		3.7
	Exp.	linear	-8579.0	17168.0		0.11		0.31	0.010		-23.3		3.4	
	Exp.	Exp.	-8577.9	17167.9					0.010	0.014	-24.3	-4.6	3.4	1.9
T>G	linear	linear	-5098.9	10205.8	0.13	0.024	-0.45	0.044						
	linear	Exp.	-5102.2	10214.5	0.13		-0.57			0.0011		-30.2		3.4
	Exp.	linear	-5099.3	10208.6		0.024		0.044	0.0066		-16.1		2.9	
	Exp.	Exp.	-5102.0	10216.1					0.0054	0.0011	-21.0	-30.2	3.1	3.4

Table S5 Comparison of linear models fitted to trios with different maternal ages.

		deCODE dataset				
		sample size	β_P	β_M	$\beta_{0,P}$	$\beta_{0,M}$
all point mutations	All trios	1548	1.41	0.39	5.50	2.04
	Trios with Gm<=40	1476	1.41	0.35	5.56	3.24
	Trios with Gm<=27	774	1.40	0.24	5.81	5.59
	Trios with Gm>27	774	1.41	0.56	5.52	-3.74
point mutations excluding C>G	All trios	1548	1.27	0.32	4.97	3.03
	Trios with Gm<=40	1476	1.28	0.28	4.94	3.93
	Trios with Gm<=27	774	1.26	0.20	5.23	5.80
	Trios with Gm>27	774	1.26	0.45	5.62	-1.35
point mutations excluding C>G and CpG>TpG	All trios	1548	1.02	0.28	4.23	1.81
	Trios with Gm<=40	1476	1.02	0.25	4.41	2.61
	Trios with Gm<=27	774	1.03	0.18	3.93	4.18
	Trios with Gm>27	774	1.01	0.40	4.89	-2.33

Table S6 Comparison of models with linear and exponential maternal age effects fitted to all trios and trios with G_m below 40.

All trios

Mutation type	Model					Estimates						
	Paternal effect	Maternal effect	Log likelihood	AIC	BIC	β_P	β_M	$\beta_{0,P}$	$\beta_{0,M}$	b_M	a_M	c_M
all point mutations	linear	linear	-12174.3	24356.6	24378.0	1.41	0.39	5.50	2.0			
	linear	exponential	-12158.4	24326.7	24353.5	1.41		5.58		0.072	7.8	0.45
point mutations excluding C>G	linear	linear	-11859.2	23726.4	23747.8	1.27	0.319	4.97	3.03			
	linear	exponential	-11849.6	23709.2	23736.0	1.27		5.01		0.067	7.42	0.4
point mutations excluding C>G and CpG>TpG	linear	linear	-11289.6	22587.1	22613.8	1.02	0.28	4.23	1.81			
	linear	exponential	-11279.6	22569.3	22596.0	1.02		4.30		0.070	5.83	0.2

Trios with G_m<=40

Mutation type	Model					Estimates						
	Paternal effect	Maternal effect	Log likelihood	AIC	BIC	β_P	β_M	$\beta_{0,P}$	$\beta_{0,M}$	b_M	a_M	c_M
all point mutations	linear	linear	-11544.1	23096.1	23117.5	1.41	0.35	5.56	3.2			
	linear	exponential	-11541.3	23092.5	23119.2	1.41		5.63		0.037	3.2	1.70
point mutations excluding C>G	linear	linear	-11251.0	22510.0	22531.4	1.28	0.282	4.94	3.93			
	linear	exponential	-11249.7	22509.3	22536.0	1.27		4.97		0.021	-1.57	2.3
point mutations excluding C>G and CpG>TpG	linear	linear	-10695.3	21398.7	21425.4	1.02	0.25	4.41	2.61			
	linear	exponential	-10693.6	21397.3	21424.0	1.02		4.44		0.033	1.88	1.5

Table S7 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(14, 15), 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 = 0.00212)		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 #	#	Prob.	#	Prob.	#	Prob.	#	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 S8 Estimates of maternal age effect on paternal mutations obtained by different methods

Poisson regression (with identity link) on 199 trios with >95% phasing rate

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 count	G_P, G_M	-0.075	0.055	0.17	0.42	0.069	1.1e-09	1141.2
Maternal mutation count	G_M	--	--	--	0.34	0.038	< 2e-16	1141.0

Negative binomial regression (with identity link) on 199 trios with >95% phasing rate

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	1.15	0.14	< 2e-16	0.31	0.17	0.062	1411.8
Paternal mutation count	G_P	1.36	0.074	< 2e-16	--	--	--	1413.2
Maternal mutation count	G_P, G_M	-0.073	0.065	0.26	0.41	0.081	4.5e-07	1129.5
Maternal mutation count	G_M	--	--	--	0.34	0.045	1.6e-13	1128.7

Poisson regression (with identity link) on 130 trios with >98% phasing rate

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	1.00	0.14	2.7e-13	0.48	0.17	0.0037	912.4
Paternal mutation count	G_P	1.33	0.080	< 2e-16	--	--	--	918.8
Maternal mutation count	G_P, G_M	-0.083	0.064	0.19	0.43	0.079	4.2e-08	747.8
Maternal mutation count	G_M	--	--	--	0.35	0.048	1.5e-13	747.5

Negative binomial regression (with identity link) on 130 trios with >98% phasing rate

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	0.99	0.15	4.1e-11	0.49	0.18	0.0075	912.3

Paternal mutation count	G_P	1.32	0.091	< 2e-16	--	--	--	917.2
Maternal mutation count	G_P, G_M	-0.078	0.075	0.29	0.42	0.092	4.4e-06	742.4
Maternal mutation count	G_M	--	--	--	0.35	0.057	8.0e-10	741.5

Maximum likelihood approach

Model	β_P	β_M	α_P	α_M	β_{Mp}*	Log likely-hood	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	6.5e-05	0.022
Same G_P	Difference in **maternal** mutation counts	G_M	$1/(G_M1+ G_M2)$*	0.38	0.047	1.7e-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 these standard errors and *p*-values may be unreliable due to violations of the linear regression assumptions.

Table S9 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 S10 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	4.5%
Shuffle of mutation type labels across all paternal DNMs	20,000	935	897	4.5%

Table S11 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	1.439 (0.0398)	< 2e-16	0.350 (0.0519)	1.53e-11	7.893 (0.883)	4.11
225 3-gen trios	225	1.167 (0.124)	< 2e-16	0.657 (0.152)	1.53e-05	9.981 (2.296)	1.78
1323 2-gen trios	1323	1.467 (0.0421)	< 2e-16	0.324 (0.0554)	5.16e-09	7.447 (0.957)	4.53

Negative binomial 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	1.441 (0.0491)	< 2e-16	0.343 (0.0629)	5.08e-08	8.073 (1.056)	4.20
225 3-gen trios	225	1.159 (0.145)	< 2e-16	0.659 (0.177)	1.92e-04	10.149 (2.669)	1.76
1323 2-gen trios	1323	1.467 (0.0521)	< 2e-16	0.315 (0.0672)	2.78e-06	7.611 (1.144)	4.66

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 S12 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	0.104 (0.0110)	< 2e-16	0.0396 (0.0144)	0.00594	0.478 (0.244)	2.62
225 3-gen trios	225	0.0804 (0.0347)	0.0205	0.105 (0.0426)	0.0137	-0.287 (0.640)	0.77
1323 2-gen trios	1323	0.106 (0.0116)	< 2e-16	0.0311 (0.0153)	0.0421	0.600 (0.265)	3.39

Negative binomial 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	0.104 (0.0112)	< 2e-16	0.0393 (0.0146)	0.00721	0.48 (0.24)	2.62
225 3-gen trios	225	0.0805 (0.0351)	0.0219	0.105 (0.0431)	0.0150	-0.288 (0.648)	0.77
1323 2-gen trios	1323	0.106 (0.0118)	< 2e-16	0.0309 (0.0155)	0.047	0.602 (0.268)	3.43

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

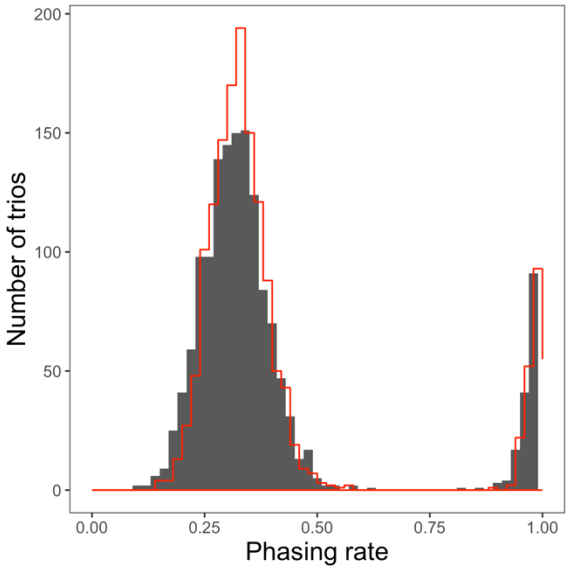

Figure S1. Fraction of phased mutations among detected DNMs for each trio in the deCODE dataset. The bimodal distribution reflects the drastically different phasing rates for probands with or without a third-generation. The red line represents the expected distribution under binomial sampling with a success rate of 0.977 or 0.318 per mutation, for trios with or without third generation data, respectively. These success rates are the actual fractions of phased mutations among all aggregated DNMs for these two sets of trios, respectively.

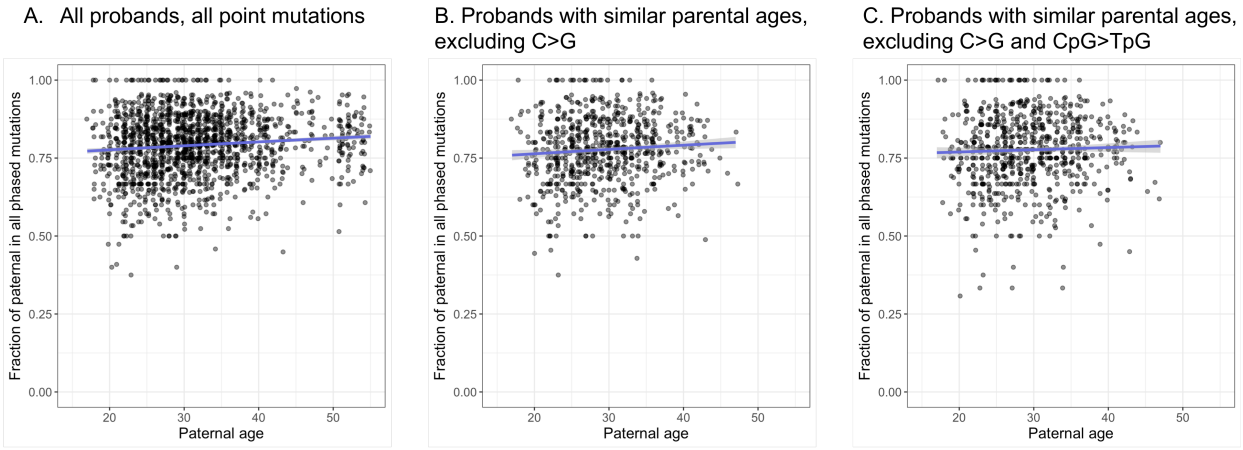

Figure S2. Fraction of paternal mutations among phased mutations as a function of paternal age. Each point represents the data for one child (proband) in the Icelandic data set(9) with at least three phased mutations under the corresponding category. The blue line is the predicted fraction of paternal mutations by binomial regression with logit link, with the shaded area representing the 95% confidence interval (calculated with the "predict" 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.

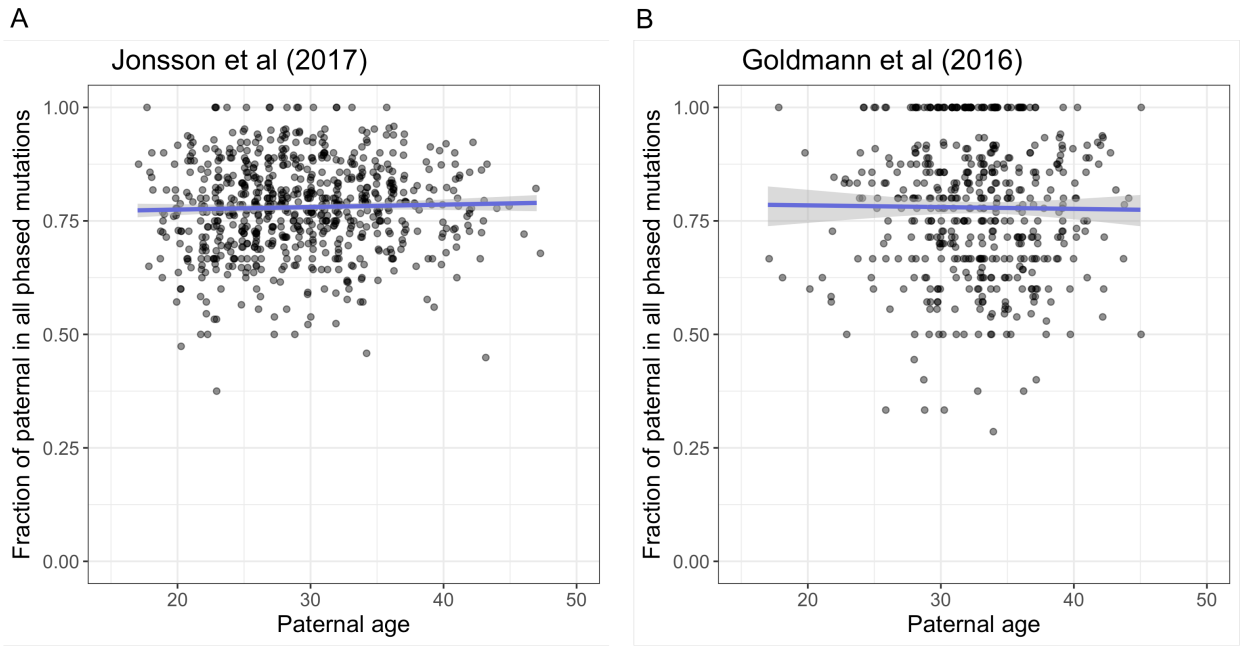

Figure S3. Replication of the stable fraction of paternal mutations with paternal age in an independent dataset. Each point represents the data for one child (proband) with at least three phased point mutations and similar parental ages (paternal-to-maternal age ratio between 0.9 to 1.1; 719 trios total). The blue line is the predicted fraction of paternal mutations by binomial regression with logit link, with the shaded area representing the 95% confidence interval (calculated with the "predict" function in R). (A) Same as Figure 1; (B) Similar plot for data from Goldmann et al. (2016)(12), which includes a total of 35,793 DNMs (7,216 of which were phased). See Materials and Methods for details.

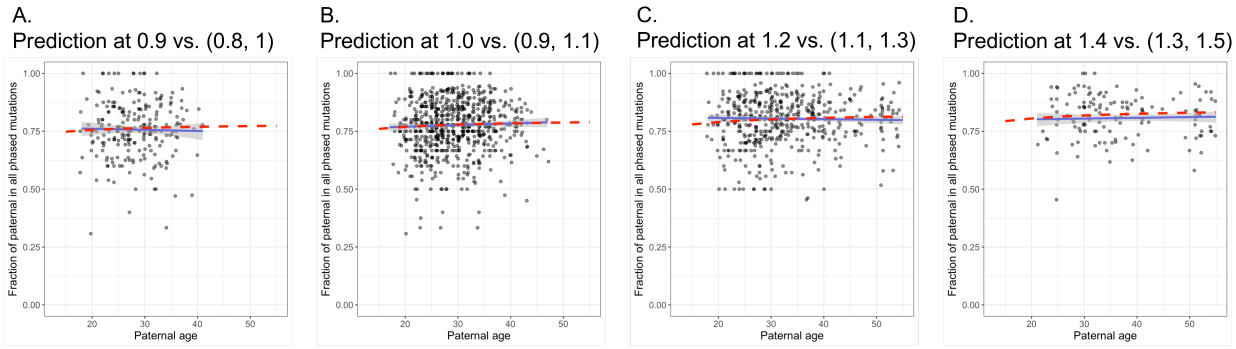

Figure S4. Fraction of paternal mutations among phased mutations for different ratios of paternal age (G_P) to maternal age (G_M). Each point represents the data for one child (proband) with at least three phased point mutations that are not C>G or CpG>TpG in the Icelandic data set(9). The blue line is the predicted fraction of paternal mutations by binomial regression with logit link, with the shaded area representing the 95% confidence interval (calculated with the "predict" function in R). 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 $G_P/G_M=1.4$.

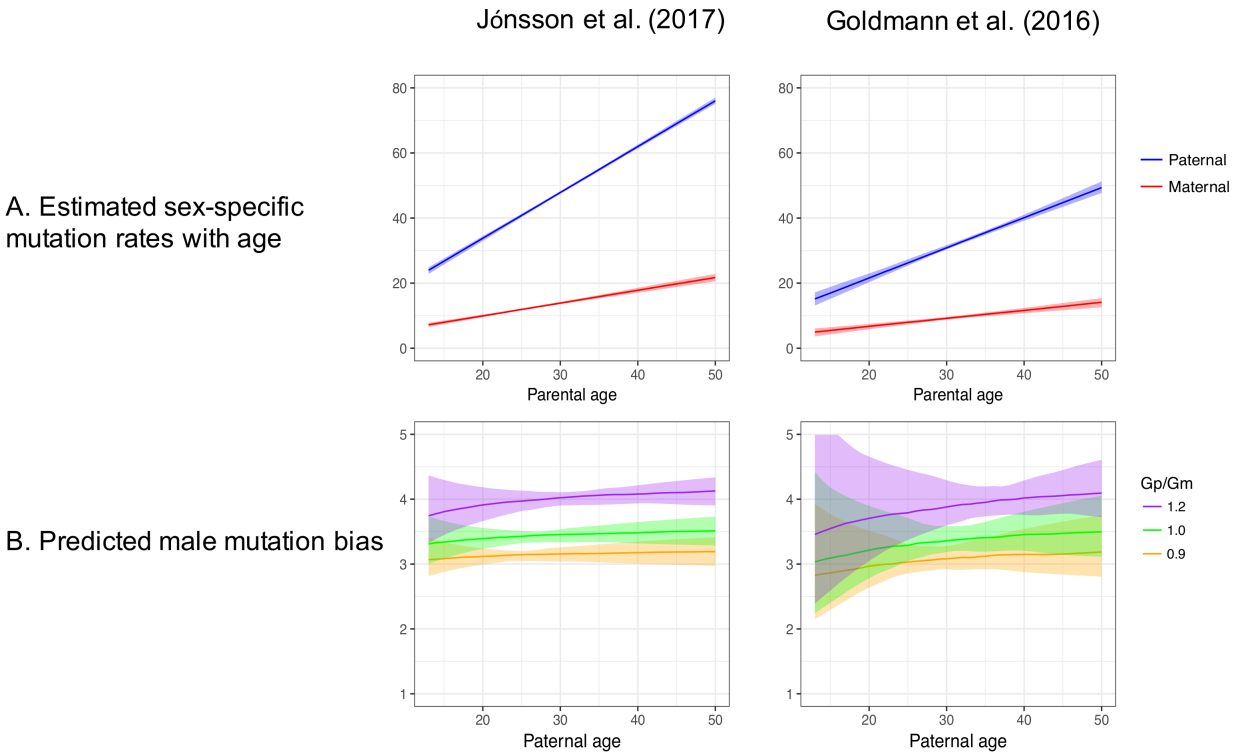

Figure S5. Comparison of parental age effects and predicted male-to-female mutation ratio at given ages estimated from two DNM datasets(9, 12, 13). The two data sets differ in their sample sizes (1548 vs 816 trios), average numbers of detected DNMs per proband (63.86 vs 43.86) and the fraction of DNMs 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.

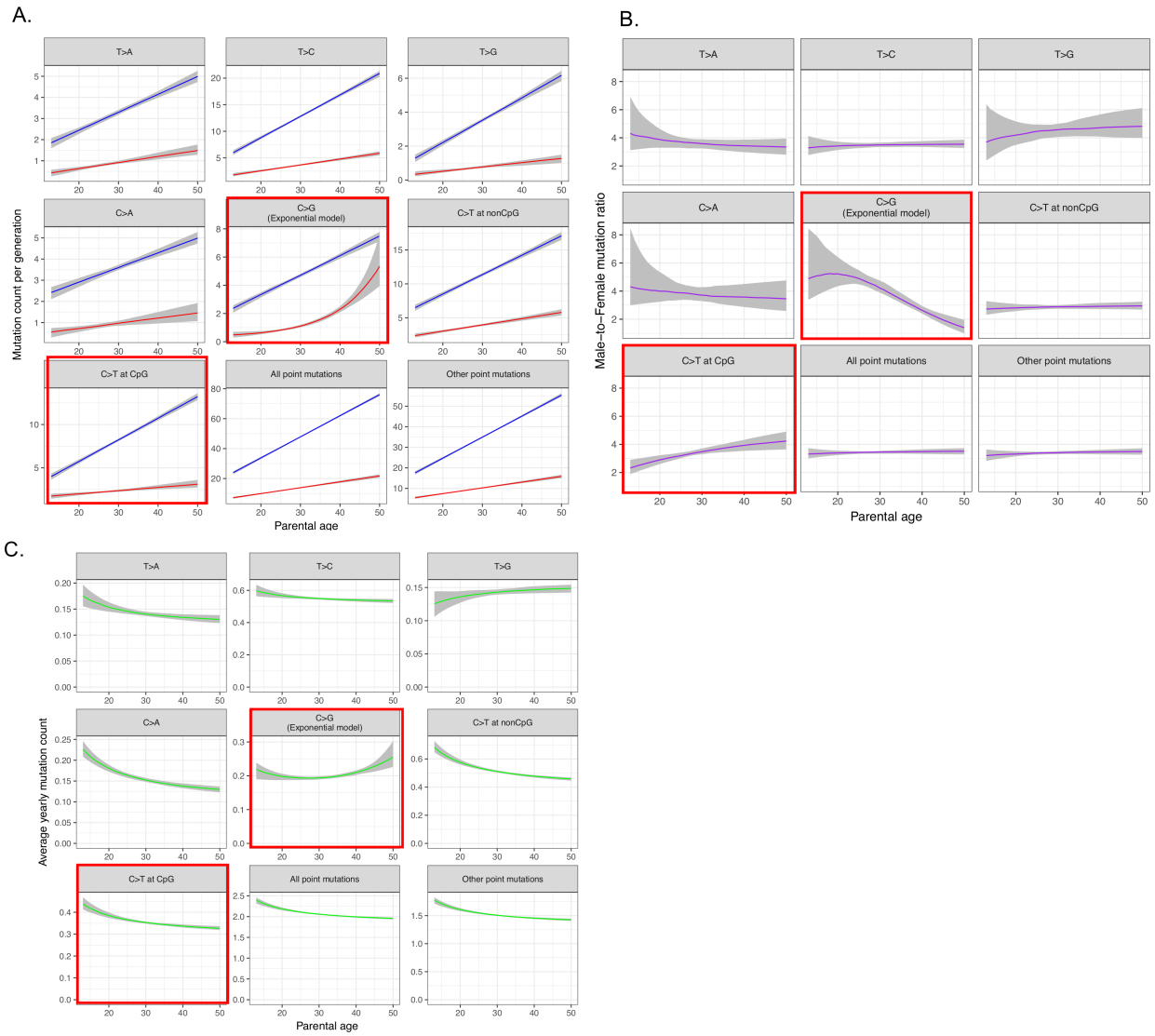

Figure S6. Estimated sex-specific mutation rates and male-to-female mutation ratio as a function of parental ages, by mutation type. Red boxes indicate the two mutation types highlighted in the main text. The extent of male mutational bias and average yearly mutation rate are estimated assuming the same 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.

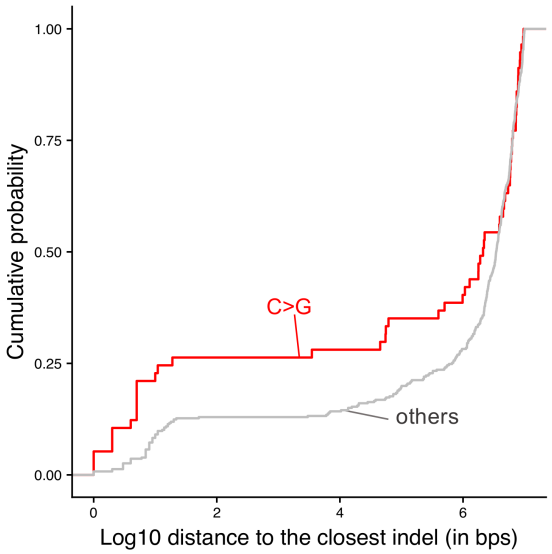

Figure S7. Distribution of distances to the closest deletion of ≥5bp for C>G mutations. Cumulative distribution of distance to the closest *de novo* deletion (≥ 5bp) for C>G transversions and for other point mutations, conditional on co-occurrence within 10Mb.

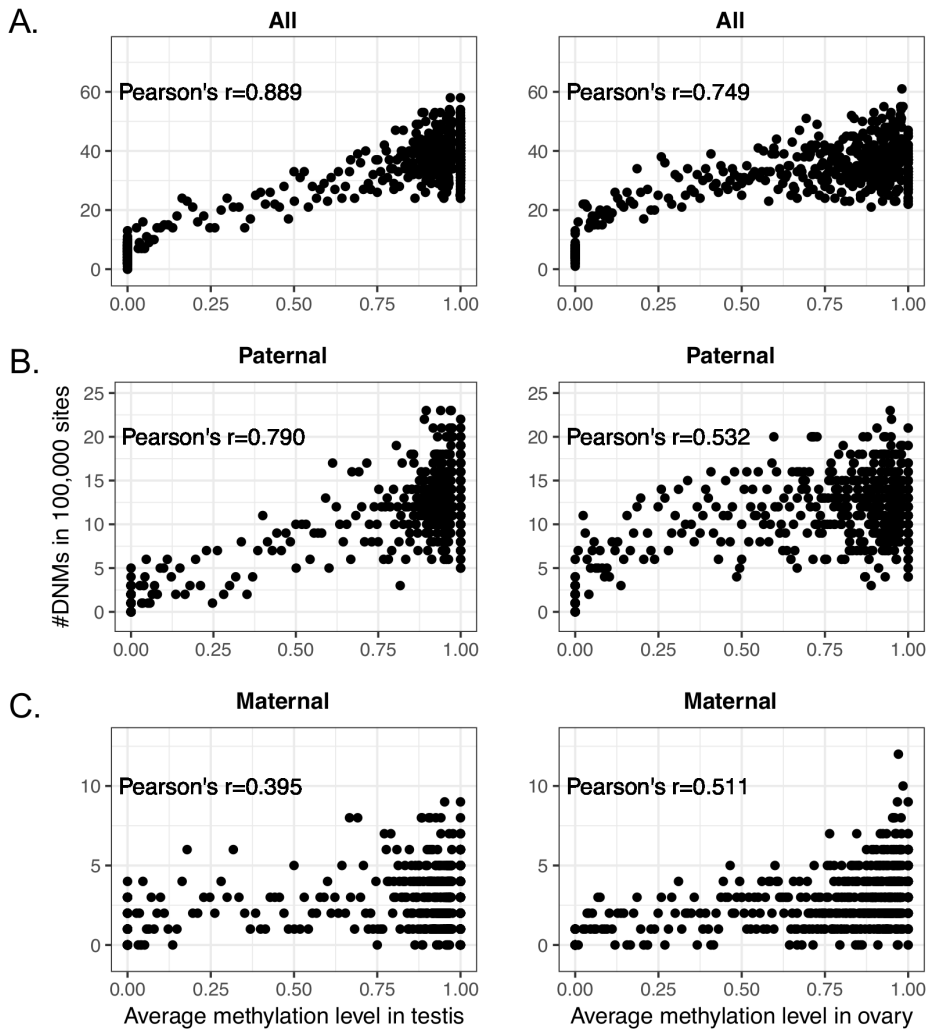

Figure S8. Correlation between the genomic distribution of CpG>TpG DNMs and methylation levels in testis and ovary. Genome-wide methylation level for each CpG site is measured by bisulfate sequencing of testis spermatozoa primary cells and ovary cells. CpGs are sorted based on methylation levels (in the two tissues separately) and grouped into bins of 100,000 sites each (see “Processing of ovary and testis methylation data at CpG sites” section for methods). The x-axis shows the average methylation level in each bin, and the y-axis is the total number of C->T DNMs in the 1548 Icelandic trios that occurred at the 100,000 sites. All correlations reported are highly significant with *p*-values below 2e-16. We note that the methylation profile of ovary cells may be a poor proxy for that of (primary) oocytes, so the correlation between CpG>TpG DNM rate and methylation is likely somewhat under-estimated.

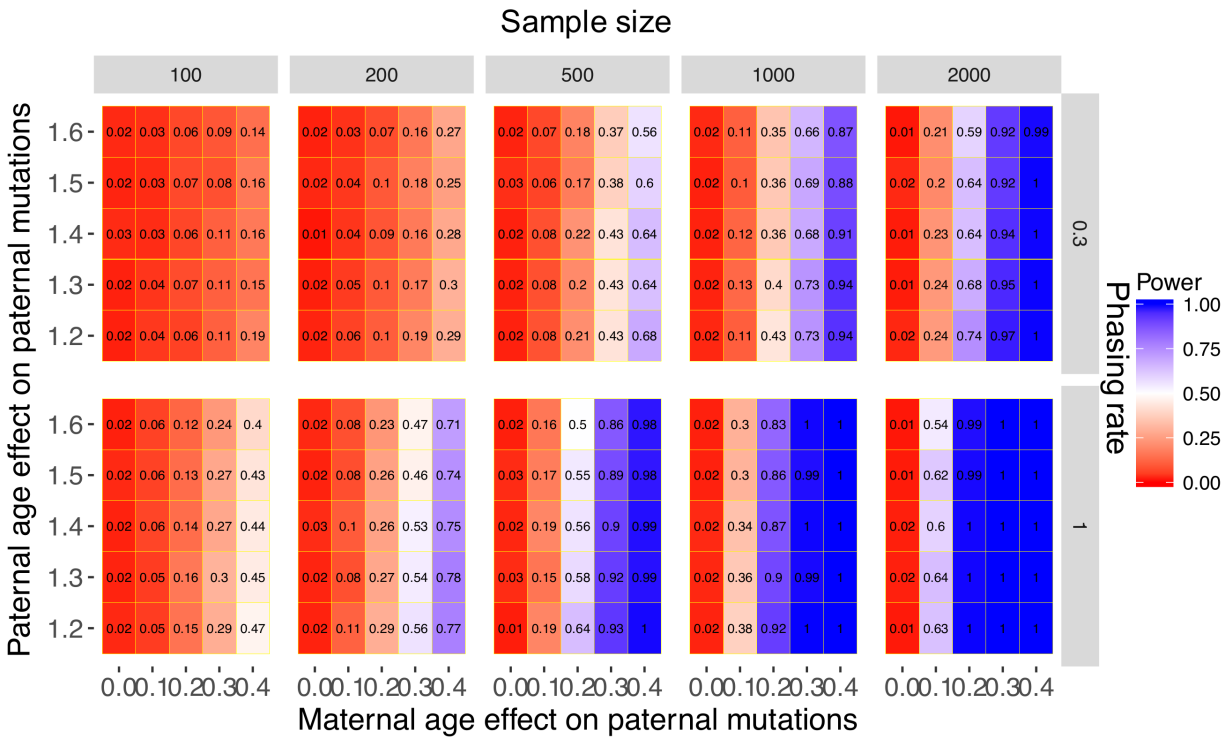

Figure S9. Power to detect a maternal age effect on paternal mutations estimated by simulation. In each panel, the x-axis and y-axis represent the presumptive maternal and paternal age effects (measured in increase in mutation counts per diploid genome per year) on the paternal mutation rate, respectively (see “Estimate of the power to detect a maternal age effect on the paternal mutation rate by simulations” section for more details for the simulations). The panels are ordered by sample size (the number of trios), in columns, and by the phasing rate, in rows. The color of each cell indicates the power of detecting a significant maternal age effect for this pair of parameter values, with white representing exactly 50% power, blue >50% and red <50% (the exact power is also provided in each cell). We assume no errors in DNM calling and phasing, as well as uniform phasing probability across mutations and families. Given these overly optimistic assumptions, power is likely over-estimated in our simulations, especially for the case with incomplete phasing.

References

1. Polani PE, Crolla JA (1991) A test of the production line hypothesis of mammalian oogenesis. 64–70.
2. Fulton N, Silva SJM, Bayne RAL, Anderson RA (2005) Germ Cell Proliferation and Apoptosis in the Developing. 90(May):4664–4670.
3. Lindsay SJ, Rahbari R, Kaplanis J, Keane TM, Hurles ME (2016) Striking differences in patterns of germline mutation between mice and humans. *bioRxiv*.
4. Harland C, et al. (2017) Frequency of mosaicism points towards mutation-prone early cleavage cell divisions. *bioRxiv*.
5. Jónsson H, et al. (2018) Multiple transmissions of de novo mutations in families. *Nat Genet* 50(December). doi:10.1038/s41588-018-0259-9.
6. Braude P, Bolton V, Moore S (1988) Human gene expression first occurs between the four- and eight-cell stages of preimplantation development. *Nature* 332(6163):459–61.
7. Dobson AT, et al. (2004) The unique transcriptome through day 3 of human preimplantation development. *Hum Mol Genet* 13(14):7–9.
8. Zhang P, et al. (2009) Transcriptome Profiling of Human Pre-Implantation Development. 4(11). doi:10.1371/journal.pone.0007844.
9. Jónsson H, et al. (2017) Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* 549(7673):519–522.
10. De Iuliis GN, et al. (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. *Biol Reprod* 81(3):517–524.
11. Smith TB, et al. (2013) The presence of a truncated base excision repair pathway in human spermatozoa that is mediated by OGG1. *J Cell Sci* 126(6):1488–1497.
12. Goldmann JM, et al. (2016) Parent-of-origin-specific signatures of de novo mutations. *Nat Genet* 48(8):935–939.
13. Wong WSW, et al. (2016) New observations on maternal age effect on germline de novo mutations. *Nat Commun* 7(May 2015):10486.
14. Montgomery SB, et al. (2013) The origin , evolution , and functional impact of short insertion – deletion variants identified in 179 human genomes. 749–761.
15. Kloosterman WP, et al. (2015) Characteristics of de novo structural changes in the human genome. *Genome Res* 25:792–801.