LETTER

doi:10.1038/nature24018

Parental influence on human germline *de novo* mutations in 1,548 trios from Iceland

Hákon Jónsson[1], Patrick Sulem[1], Birte Kehr[1], Snaedis Kristmundsdottir[1], Florian Zink[1], Eirikur Hjartarson[1], Marteinn T. Hardarson[1], Kristjan E. Hjorleifsson[1], Hannes P. Eggertsson[1], Sigurjon Axel Gudjonsson[1], Lucas D. Ward[1], Gudny A. Arnadottir[1], Einar A. Helgason[1], Hannes Helgason[1], Arnaldur Gylfason[1], Adalbjorg Jonasdottir[1], Aslaug Jonasdottir[1], Thorunn Rafnar[1], Mike Frigge[1], Simon N. Stacey[1], Olafur Th. Magnusson[1], Unnur Thorsteinsdottir[1,2], Gisli Masson[1], Augustine Kong[1,3], Bjarni V. Halldorsson[1,4], Agnar Helgason[1,5], Daniel F. Gudbjartsson[1,3] & Kari Stefansson[1,2]

The characterization of mutational processes that generate sequence diversity in the human genome is of paramount importance both to medical genetics[1,2] and to evolutionary studies[3]. To understand how the age and sex of transmitting parents affect *de novo* mutations, here we sequence 1,548 Icelanders, their parents, and, for a subset of 225, at least one child, to 35× genome-wide coverage. We find 108,778 *de novo* mutations, both single nucleotide polymorphisms and indels, and determine the parent of origin of 42,961. The number of *de novo* mutations from mothers increases by 0.37 per year of age (95% CI 0.32–0.43), a quarter of the 1.51 per year from fathers (95% CI 1.45–1.57). The number of clustered mutations increases faster with the mother's age than with the father's, and the genomic span of maternal *de novo* mutation clusters is greater than that of paternal ones. The types of *de novo* mutation from mothers change substantially with age, with a 0.26% (95% CI 0.19–0.33%) decrease in cytosine–phosphate–guanine to thymine–phosphate–guanine (CpG>TpG) *de novo* mutations and a 0.33% (95% CI 0.28–0.38%) increase in C>G *de novo* mutations per year, respectively. Remarkably, these age-related changes are not distributed uniformly across the genome. A striking example is a 20 megabase region on chromosome 8p, with a maternal C>G mutation rate that is up to 50-fold greater than the rest of the genome. The age-related accumulation of maternal non-crossover gene conversions also mostly occurs within these regions. Increased sequence diversity and linkage disequilibrium of C>G variants within regions affected by excess maternal mutations indicate that the underlying mutational process has persisted in humans for thousands of years. Moreover, the regional excess of C>G variation in humans is largely shared by chimpanzees, less by gorillas, and is almost absent from orangutans. This demonstrates that sequence diversity in humans results from evolving interactions between age, sex, mutation type, and genomic location.

In a previous study, we found that the number of *de novo* mutations (DNMs) transmitted by fathers increases with age, at a rate of ~2 per year, with no significant effect of the mother's age[4]. Recently, studies have shown a maternal age effect[5–7] of 0.24 DNM per year (ref. 6). The greater impact of the father's age is consistent with repeated mitosis of spermatogonia (~23 per year (ref. 8)), whereas ova do not divide postnatally. Moreover, mothers transmit relatively more C>T, and fewer T>G and C>A, DNMs than fathers[6]. Nucleotide type[2,9], sequence context[2,9], replication timing[10], functional constraints[9,11], apolipoprotein B messenger RNA-editing enzyme catalytic (APOBEC) polypeptide activity[12], and epigenetics[13,14] have also been reported to affect the mutational landscape. DNM clusters in the human germline are characterized by an excess of C>G mutations[15,16], are often of maternal origin[17], and show strand concordance[18]. Despite many advances, our knowledge on how sex differences in germ cell development and maintenance affect their mutability is limited. To assess differences in the rate and class of DNMs transmitted by mothers and fathers, we analysed whole-genome sequencing (WGS) data from 14,688 Icelanders with an average of 35× coverage (Data Descriptor[19]). This set contained 1,548 trios, used to identify 108,778 high-quality DNMs (101,377 single nucleotide polymorphisms (SNPs); Methods and Fig. 1), resulting in an average of 70.3 DNMs per proband.

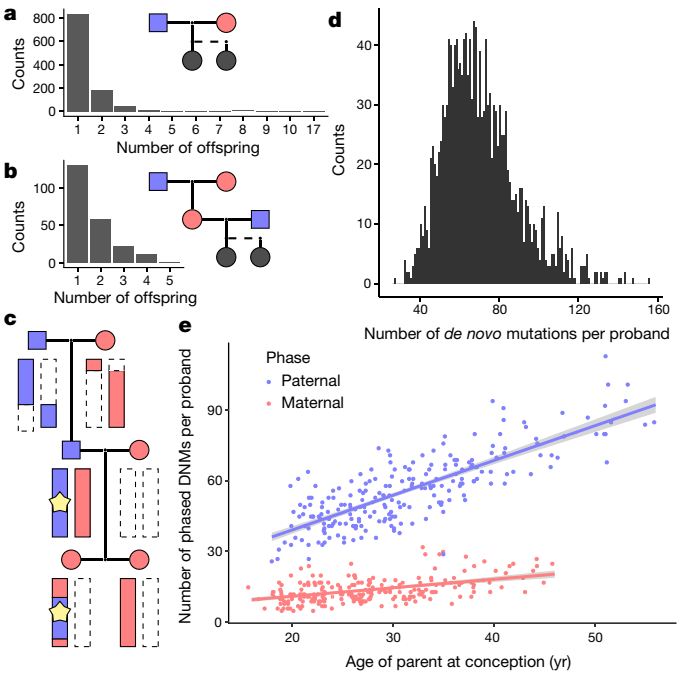


Figure 1 | Family relationships and phasing of DNMs in three-generation families. a, **b**, Number of offspring per parent pair and three-generation family proband. **c**, Schematic view of the three-generation phasing approach. The DNM (star), along with the paternal chromosome (blue segment), is transmitted from the proband to an offspring. We modelled DNM transmission in three-generation families and used the resulting prediction to define high-quality DNMs (Methods). **d**, Number of DNMs per proband. **e**, Phased DNMs as a function of the parent's age at conception (restricting to 225 three-generation probands). The lines are from a Poisson regression. Grey area, 95% CI. Dither was added to the ages in **e**.

[1]deCODE genetics/Amgen Inc., 101 Reykjavik, Iceland. [2]Faculty of Medicine, School of Health Sciences, University of Iceland, 101 Reykjavik, Iceland. [3]School of Engineering and Natural Sciences, University of Iceland, 101 Reykjavik, Iceland. [4]School of Science and Engineering, Reykjavik University, 101 Reykjavik, Iceland. [5]Department of Anthropology, University of Iceland, 101 Reykjavik, Iceland.

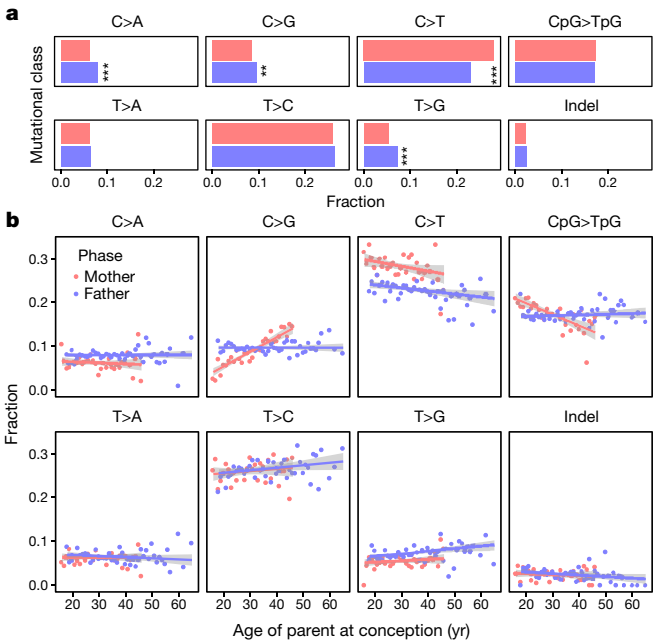


Figure 2 | Mutational spectra as a function of parents' ages at conception. a, Relative frequency of mutational classes by gender. The *P* values from Fisher's tests are 7.8×10^{-23} (C>T***), 3.8×10^{-10} (T>G***), 2.3×10^{-8} (C>A***), 2.7×10^{-3} (C>G**), 4.0×10^{-1} (T>A), 4.8×10^{-1} (T>C), 5.3×10^{-1} (Indel), and 5.4×10^{-1} (CpG>TpG). **b**, Mutation spectra as a function of the parent's age at conception. The absolute counts are depicted in Extended Data Fig. 2. **c**, Absolute age effect for each mutational class. The line is from a linear regression through the origin using the numbers of paternal DNMs for the mutational classes as weights. For this figure, 42,961 phased DNMs from 1,548 trios were used. Grey areas and crossbars, 95% CI.

The DNM call quality was also assessed using 91 monozygotic twins of probands (Methods). Of 6,034 DNMs observed in these probands, 97.1% were found in their twins. Sanger sequencing was used to validate 38 discordant calls in monozygotic twins, of which 57.9% were confirmed to be present only in the proband, and therefore postzygotic, with the rest deemed genotyping errors.

After determining the parental origin of 15,746 DNMs in the 225 three-generation families using haplotype sharing (Fig. 1c and Methods), 80.4% were found to be of paternal origin (Extended Data Fig. 1). Figure 1e shows a strong relationship between the number of paternal DNMs and the father's age at conception (1.47 per year, 95% CI 1.34–1.59) and a weaker impact of the mother's age on the number of maternal DNMs (0.37 per year, 95% CI 0.30–0.45).

The parental origin of all DNMs was also assessed by read pair tracing nearby phased variants, resulting in 42,961 DNMs phased by at least one method. This augmentation yielded similar age-effect estimates: 1.51 per year (95% CI 1.45–1.57) and 0.37 (95% CI 0.32–0.43) for fathers and mothers, respectively (Supplementary Table 6).

In line with a previous report[6], the classification of DNMs by mutation class revealed that the relative frequency of C>T DNMs was greater in maternal than paternal transmissions (odds = 1.31; $P = 7.8 \times 10^{-23}$), while T>G (odds = 0.73; $P = 3.8 \times 10^{-10}$) and C>A substitutions (odds = 0.76; $P = 2.3 \times 10^{-8}$) were relatively rarer (Fig. 2a). We extended the analysis to incorporate adjacent bases while accounting for mutational class (Methods). The class ATT>AGT was enriched in mothers (odds = 2.1; $P = 1.3 \times 10^{-5}$; 96 tests). However, we did not confirm a parental sex-bias for 16 other classes previously reported as significant[6]. This discrepancy is explained by a lack of multiple testing correction in the previous study (Supplementary Information).

The parental age effect was significant for each mutational class ($P < 0.05/16$; Fig. 2c), with the greatest effects observed for T>C mutations: 0.39 (95% CI 0.36–0.42) paternal and 0.12 (95% CI 0.09–0.15) maternal mutations per year. All but two mutational classes (Supplementary Table 10) were consistent with a proportional relationship, where the ratio of the paternal and maternal age effect was 3.1. The exceptions were CpG>TpG mutations ($P = 1.7 \times 10^{-2}$), where the paternal age effect (0.24; 95% CI 0.22–0.26) was sixfold that of the maternal one (0.04; 95% CI 0.02–0.07), and C>G mutations ($P = 9.3 \times 10^{-7}$), where maternal (0.09; 95% CI 0.07–0.10) and paternal (0.12; 95% CI 0.11–0.14) age effects were similar.

The mutational spectrum was affected more by maternal than paternal age (Fig. 2b, Extended Data Fig. 2 and Supplementary Table 11). The fraction of maternal C>G mutations increased (0.33% per year; 95% CI 0.28–0.38%), whereas CpG>TpG mutations decreased (0.26% per year; 95% CI 0.19–0.33%).

An assessment of the genomic distribution of phased DNMs (2-megabase (Mb) windows, Methods) revealed substantial regional enrichment of maternal DNMs (Extended Data Figs 3–5). Strikingly, the density of maternal DNMs on the first 20 Mb of chromosome 8p was 4.5 times greater than the genome-wide average (Fig. 3a, b). The maternal DNM fraction in this region (47.5%) was much greater than the genome-wide average (20.6%), primarily because of a 12.8-fold enrichment of C>G DNMs (Fig. 3c). The enrichment of maternal C>G DNMs was also found at several other chromosomal regions such as 2p16–22, 7p, 9p, 16p, and 16q, explaining a report of maternal DNM enrichment in two genes[6]. Our results demonstrate that this enrichment is primarily driven by C>G DNMs and is associated with sub-chromosomal regions rather than specific genes.

The region-specific enrichment of maternal DNMs should lead to greater SNP density, particularly of C>G SNPs. Accordingly, we observed a similar pattern of C>G SNP excess of rare and common SNPs both in the Icelandic WGS and in the 1000 Genomes (http://www.internationalgenome.org/data/) data (Fig. 3d and Extended Data Fig. 6). Thus, these regions have been subject to a high C>G mutation rate in humans both recently and in the distant past.

As C>G DNMs are more common in clusters than other DNMs[6,15], we investigated whether sex and age of parents and genomic location affect the rates of DNM clusters (Methods). Parent-of-origin information was available for 869 of 1,859 DNM clusters, of which 558 were paternal (64.2%). The number of DNM clusters increased faster ($P = 0.0062$) with the mother's age (0.032 per year, 95% CI 0.027–0.037) than with the father's age (0.019 per year, 95% CI 0.014–0.025).

Using the Icelandic WGS data to define C>G enriched regions as the genomic decile with the greatest density of C>G SNPs (Methods), we found that 20.6% and 11.4% of maternal and paternal DNMs occurred there, respectively. Importantly, 33.1% of the maternal age effect was attributable to the C>G enriched regions, in contrast to 8.8% of the paternal one (Supplementary Table 13). Strikingly, 56.5% of maternal, but only 13.0% of paternal, DNM clusters occurred within C>G enriched regions (Fig. 3e and Extended Data Fig. 7). Furthermore, the maternal clusters were longer than paternal ones (median 3,660 versus 261 bases; $P = 6.9 \times 10^{-12}$).

No APOBEC motif enrichment was detected for clustered DNMs after correcting for their mutational spectrum (Supplementary

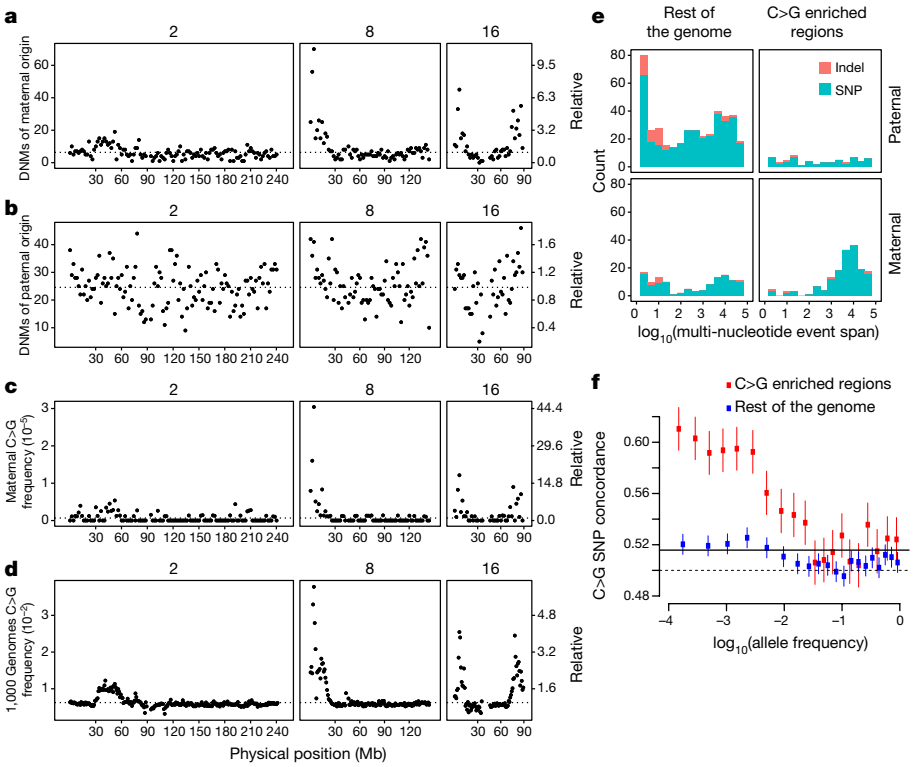


Figure 3 | Genomic distribution and clustering of DNMs. a, **b**, Number of maternal and paternal DNMs (2-Mb windows). **c**, Maternal C>G DNM frequency (2-Mb windows). **d**, 1000 Genomes C>G variant frequency (1-Mb windows). The chromosomes chr2, chr8, and chr16 were depicted as they have regions with proximal (chr2) and extreme enrichment of maternal DNMs (chr8 and chr16). The dotted line is the genome-wide average. Extended Data Figs 3–6 are genome-wide versions of **a**, **b**, and **d**. **e**, DNM cluster span. Extended Data Fig. 7 is a detailed version of **e**. **f**, Strand concordance among rare C>G SNPs. Points and vertical lines, means and 95% CI (normal approximation).

Information), indicating that the clusters are not formed by APOBEC activity. An alternative explanation is hyper-mutability of single-strand intermediates during repair of double-strand breaks[20], which would result in strand-coordinated DNM clusters. Supporting this hypothesis, we observed 86% concordance (95% CI 76.1–93.5%) between pairs of maternal C>G within clusters and 75% concordance (95% CI 59.7–87.5%) between paternal ones. The influx of strand-coordinated C>G clusters has affected sequence diversity in Icelanders, in whom we found greater concordance among rare than common C>G SNPs (allele frequency <1% and ≥1%; Extended Data Figs 7 and 8) and substantially greater concordance for rare C>G SNPs (59.0%) within the C>G enriched regions than outside (52.0%; Fig. 3f).

Recently, we reported a marked increase in the rate of non-crossover gene conversions (NCOGCs) with the mother's age, but no paternal age effect[21]. Intriguingly, we found their rate to be 2.5-fold greater within the C>G enriched regions than outside them ($P = 3.5 \times 10^{-12}$; Methods and Supplementary Table 16). The impact of the mother's age on the NCOGC rate was also greater within the C>G enriched regions than outside (2.87×10^{-6} and 2.37×10^{-7} per base pair per year; $P = 7.0 \times 10^{-8}$). These results suggest a common mechanism for the maternal-age-related C>G DNMs and NCOGCs.

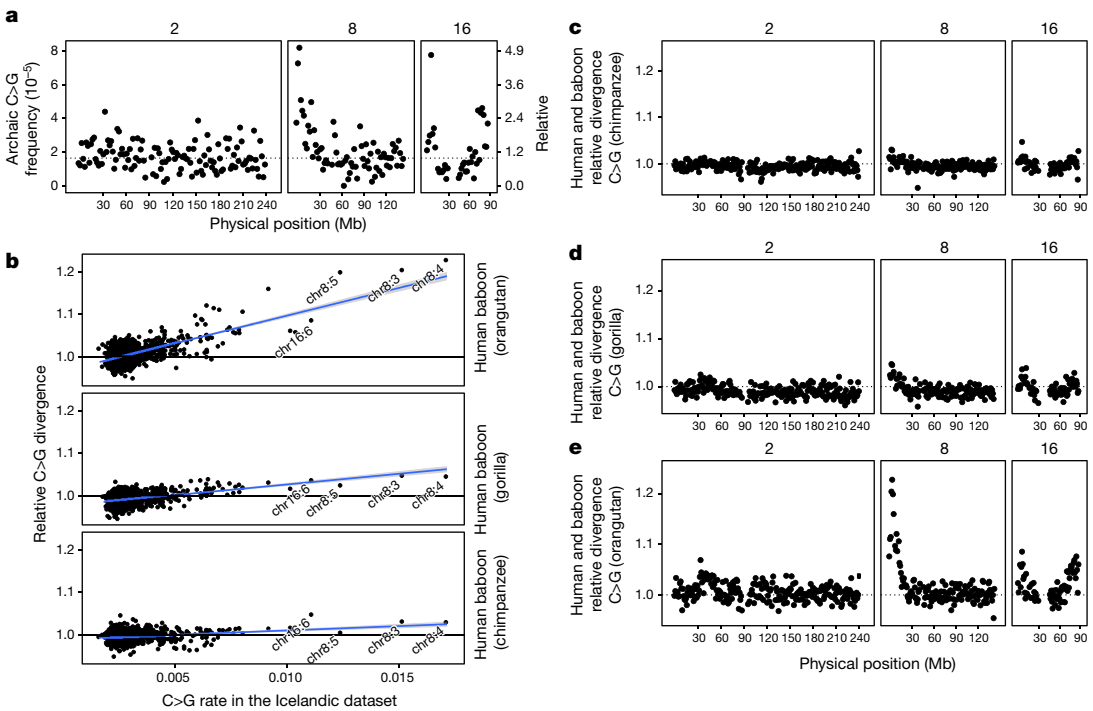


Figure 4 | Local C>G enrichment in an evolutionary context. **a**, Frequency of archaic C>G mutations (2-Mb windows). **b**, C>G normalized divergence ratios against C>G rate in the Icelandic dataset in 1-Mb windows. Grey area, 95% CI. **c**, C>G divergence between baboon and human, normalized by the C>G divergence between baboon and chimpanzee ($d_{b,h}/d_{b,c}$, 1-Mb windows). **d**, **e**, Same as **c** except the gorilla ($d_{b,h}/d_{b,g}$) and the orangutan ($d_{b,h}/d_{b,o}$) are used instead of the chimpanzee. **a**, **c**–**e**, Results for chromosomes 2, 8, and 16. The genome-wide versions of **c**–**e** are in Extended Data Fig. 9.

Since common variants in the Icelandic and 1000 Genomes datasets exhibit the regional C>G DNM enrichment, the mechanism underlying the regional distribution of C>G mutations is presumably common to all humans. Similar patterns were also observed in archaic hominins (Methods and Fig. 4a).

To dig deeper into evolutionary history, we extended the analysis of C>G mutation enrichment to sequence divergence between primate species. We devised divergence ratios (Methods), which deviate from 1 when orangutans, gorillas, or chimpanzees exhibit a fraction of C>G substitutions relative to baboons that differs from the human and baboon sequence divergence. Regression of the divergence ratios against C>G fractions revealed lower slope estimates for chimpanzees (2.05, 95% CI 1.62–2.48) and gorillas (5.04, 95% CI 4.53–5.55) than for orangutans (13.0, 95% CI 12.4–13.7) (Fig. 4b–e and Extended Data Figs 9 and 10). The considerable differences between orangutans and African apes indicate that the regional pattern of maternally transmitted C>G DNMs observed in contemporary Icelanders arose, or at least substantially increased, on the lineage leading to African apes.

One plausible explanation for the drastic age-related sex differences in transmitted DNMs is the relative lack of mitosis in ageing oocytes compared with spermatogonia, which may enrich for damage-induced DNMs. The maternal-age-related increase of both C>G DNMs and NCOGCs in particular genomic regions must be due to some distinctive property of ageing oocytes. One such property is the long-term structural stress acting on chromosomes in chiasmata, which deteriorate with age owing to depletion of cohesin from oocytes[22]. Our results indicate that double-strand breaks occur at a greater rate in C>G enriched regions, perhaps because of structural fragility. Another possibility is that oocyte viability is less impaired by double-strand breaks in these regions, although this would not be sufficient to create the approximately 50-fold enrichment of C>G DNMs on chromosome 8p (Supplementary Information).

Mutation rates are key parameters for calibrating the timescale of sequence divergence. We estimate the mutation rate as 1.29×10^{-8} per base pair per generation and 4.27×10^{-10} per base pair per year (Methods). Our findings have a direct bearing on the disparity that has emerged between mutation rates estimated directly from pedigrees ($\sim 4 \times 10^{-10}$ per base pair per year) and phylogenetic rates ($\sim 10^{-9}$ per base pair per year)[3], as they indicate that the molecular clock is affected by life-history traits in a sex-specific manner[23–25] and varies by genomic region within and across species. This allows us to predict the long-term consequences of a shift in generation times (Methods)[24]. Thus, a 10 year increase in the average age of fathers would increase the mutation rate by 4.7% per year. The same change for mothers would decrease the mutation rate by 9.6%, because extra mutations attributable to older mothers are offset by fewer generations.

It has been argued that CpG>TpG mutations are less affected by mitosis and thus more clocklike than other classes[26]. However, we show that they accumulate at a high rate in the paternal germline, perhaps because the repair of deaminated cytosines takes longer than the average interval between mitoses of spermatogonia[25]. An assumption about a strict molecular clock requires an absence of a generation effect, which is not compatible with our data (Supplementary Information).

One implication of the hyper-mutable regions of maternal origin is that the genomic distribution of variants could be used to make inferences about long-term sex differences in the age of parents in populations or species. Another is that caution is warranted in omitting these attributes in the use and interpretation of estimates of phylogenetic mutation rates.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 March; accepted 20 August 2017.
Published online 20 September 2017.

1. Veltman, J. A. & Brunner, H. G. *De novo* mutations in human genetic disease. *Nat. Rev. Genet.* **13,** 565–575 (2012).
2. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
3. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.* **13,** 745–753 (2012).
4. Kong, A. *et al.* Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488,** 471–475 (2012).
5. Wong, W. S. W. *et al.* New observations on maternal age effect on germline *de novo* mutations. *Nat. Commun.* **7,** 10486 (2016).
6. Goldmann, J. M. *et al.* Parent-of-origin-specific signatures of *de novo* mutations. *Nat. Genet.* **48,** 935–939 (2016).
7. Rebolledo-Jaramillo, B. *et al.* Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA. *Proc. Natl Acad. Sci. USA* **111,** 15474–15479 (2014).
8. Qin, J. *et al.* The molecular anatomy of spontaneous germline mutations in human testes. *PLoS Biol.* **5,** 1912–1922 (2007).
9. Aggarwala, V. & Voight, B. F. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat. Genet.* **48,** 349–355 (2016).
10. Koren, A. *et al.* Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet.* **91,** 1033–1040 (2012).
11. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47,** 435–444 (2015).
12. Chan, K. & Gordenin, D. A. Clusters of multiple mutations: incidence and molecular mechanisms. *Annu. Rev. Genet.* **49,** 243–267 (2015).
13. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518,** 360–364 (2015).
14. Rahbari, R. *et al.* Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48,** 126–133 (2016).
15. Francioli, L. C. *et al.* Genome-wide patterns and properties of *de novo* mutations in humans. *Nat. Genet.* **47,** 822–826 (2015).
16. Besenbacher, S. *et al.* Multi-nucleotide *de novo* mutations in humans. *PLoS Genet.* **12,** e1006315 (2016).
17. Yuen, R. K. C. *et al.* Genome-wide characteristics of *de novo* mutations in autism. *npj Genomic Med.* **1,** 16027 (2016).
18. Seplyarskiy, V. B., Andrianova, M. A. & Bazykin, G. A. APOBEC3A/B-induced mutagenesis is responsible for 20% of heritable mutations in the TpCpW context. *Genome Res.* **27,** 175–184 (2017).
19. Jónsson, H. *et al.* Whole genome characterization of sequence diversity of 15,220 Icelanders. *Sci. Data* **4,** 170115 (2017).
20. Roberts, S. A. & Gordenin, D. A. Hypermutation in human cancer genomes: footprints and mechanisms. *Nat. Rev. Cancer* **14,** 786–800 (2014).
21. Halldorsson, B. V. *et al.* The rate of meiotic gene conversion varies by sex and age. *Nat. Genet.* **48,** 1377–1384 (2016).
22. Herbert, M., Kalleas, D., Cooney, D., Lamb, M. & Lister, L. Meiosis and maternal aging: insights from aneuploid oocytes and trisomy births. *Cold Spring Harb. Perspect. Biol.* **7,** a017970 (2015).
23. Ségurel, L., Wyman, M. J. & Przeworski, M. Determinants of mutation rate variation in the human germline. *Annu. Rev. Genomics Hum. Genet.* **15,** 47–70 (2014).
24. Amster, G. & Sella, G. Life history effects on the molecular clock of autosomes and sex chromosomes. *Proc. Natl Acad. Sci. USA* **113,** 1588–1593 (2016).
25. Gao, Z., Wyman, M. J., Sella, G. & Przeworski, M. Interpreting the dependence of mutation rates on age and time. *PLoS Biol*. **14,** e1002355 (2016).
26. Moorjani, P., Amorim, C. E., Arndt, P. F. & Przeworski, M. Variation in the molecular clock of primates. *Proc. Natl Acad. Sci. USA* **113,** 10607–10612 (2016).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank all the participants in this study. This study was performed in collaboration with Illumina.

Author Contributions H.J., F.Z., E.H., M.T.H., K.E.H., E.A.H., and D.F.G. analysed the data. H.J., B.K., S.K., F.Z., E.H., M.T.H., K.E.H., H.P.E., E.A.H., A.G., and D.F.G. created methods for analysing the data. Ad.J., As.J., and O.Th.M. performed the experiments. S.A.G., L.D.W., G.A.A., H.H., T.R., and M.F. collected the samples and information. H.J., P.S., U.T., G.M., A.K., B.V.H., A.H., D.F.G., and K.S. designed the study. H.J., P.S., B.V.H., A.H., D.F.G., and K.S. wrote the manuscript with input from S.N.S., U.T., G.M., and A.K.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to D.F.G. (daniel.gudbjartsson@decode.is) or K.S. (kari.stefansson@decode.is).

Reviewer Information *Nature* thanks S. Sunyaev and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

The dataset described here represents an addition of 12,803 whole-genome sequences to our previous dataset of 2,636 Icelanders, resulting in 15,220 samples after filtering (Supplementary Table 1). The description of the data acquisition and processing is described in greater detail in the accompanying Data Descriptor[19].

Data access. Access to these data are controlled; the data access consists of lists of variants without genotypes and DNMs with enumerated proband identifiers. In the latter dataset, all of the information necessary to reproduce the analysis presented in this manuscript is provided. We deposited the panel of identified variants in the European Variant Archive (accession number PRJEB15197). In addition to the variant panel, we submitted a list of high-quality DNMs along with determination of parent of origin to the European Variant Archive (accession number PRJEB21300; Supplementary Table 4). We made both panels publicly available.

Ethics statement. The National Bioethics Committee and the Icelandic Data Protection Authority approved this study. Blood or buccal samples were taken from individuals participating in various studies, after receiving informed consent from them or their guardians.

Code availability. Code availability is described in the accompanying Data Descriptor[19].

Preparation of samples for WGS. Three different sample preparation kits from Illumina were used: TruSeq DNA (method A), TruSeq Nano (method B), and TruSeq PCR-Free (method C). Samples were prepared for sequencing according to the manufacturer's instructions (Illumina). In short, either 50 ng (method B) or 1 μg (methods A and C) of genomic DNA, isolated from either frozen blood samples or buccal swabs, were fragmented to a mean target size of 300–400 base pairs (bp) using a Covaris E220 instrument. End repair, generating blunt-ended fragments, was performed followed by size selection using different ratios of AMPure XP magnetic purification beads. 3′-Adenylation and ligation of indexed sequencing adaptors containing a T nucleotide overhang was performed, followed either by AMPure purification (method C) alone or purification followed by PCR enrichment (ten cycles) using appropriate primers (methods A and B). The quality and concentration of all sequencing libraries were assessed using either an Agilent 2100 Bioanalyzer (12 samples) or a LabChip GX (96 samples) instrument from Perkin Elmer. Sequencing libraries were diluted and stored at −20 °C. Further quality control of sequencing libraries was done by multiplexing and pooling either 24 or 96 samples and sequencing each pool on an Illumina MiSeq instrument to assess optimal cluster densities, library insert size, duplication rates, and library diversities. All steps in the workflow were monitored using an in-house laboratory information management system with barcode tracking of all samples and reagents.

DNA WGS. Sequencing libraries were hybridized to the surface of paired-end flowcells using an Illumina cBot. Paired-end sequencing-by-synthesis was performed on Illumina sequencers, including $GAII_x$, HiSeq 2000/2500, or HiSeq X instruments, respectively. Read lengths depended on the instrument and/or sequencing kit being used and varied from 2 × 76 cycles to 2 × 150 cycles of incorporation and imaging. Real-time analysis involved conversion of image data to base-calling in real-time. The largest number of samples (approximately 12,000) was sequenced on HiSeq X instruments with read lengths of 2 × 150 cycles, using either v1 or v2 flowcells and sequencing chemistries, respectively (Supplementary Table 18).

Variant calling. We aligned the raw sequences against hg38 (excluding alternative contigs, but including decoy sequences) with BWA version 0.7.10 mem[27]. The sequences in the BAM files were realigned around indels with GenomeAnalysisTKLite/2.3.9 (ref. 28) using a public set of known indels plus a set of indels previously discovered in the Icelandic data. We marked PCR duplicates with Picard tools 1.117. Here, we applied a filtering step after merging the alignments and before the variant calling. We required the alignment to contain at least 45 matching bases (not necessarily consecutive) and we removed parts of reads extending from the template and into the adaptor. For the adaptor removal, we hard clipped bases from the forward read extending further than the rightmost alignment of the reverse read, and vice versa for the bases in the reverse read extending further than the leftmost alignment in the forward read.

Subsequently, we merged BAM files in segments of 50,000 bases using SAMtools (version 1.3)[29]. These merged BAM files were used as input to GATK-unified genotype caller[28], resulting in a pooled call of 15,220 individuals. This resulted in 39,020,168 autosomal variants passing the GATK-recommended filters[11], of which 31,079,378 were SNPs and 7,940,790 indels. We annotated these variants with VEP version 80 as described in ref. 11 with the ordering defined in Supplementary Table 19. The tabulation of variants in the annotation categories is reported in Supplementary Table 20. Of the 15,220 individuals used for the sequence genotyping, we restricted the *de novo* extraction to 14,688 individuals over 20× coverage (Supplementary Table 2).

Extraction of *de novo* candidates. Here we define allelic balance as the fraction of reads supporting the alternative allele out of the reads supporting the reference and alternative alleles. We used the genotypes from GATK to define possible carriers. We defined likely carriers of a *de novo* variant as those that met the following requirements: read depth over 12; allele balance between 0.25 and 0.75; and genotype likelihood difference greater than 20 between the highest and second-highest scoring genotype (GQ). We restricted our DNM analysis to the primary assembly of the autosomes and chromosome X (hg38).

We extracted *de novo* candidates from the variants satisfying the following criteria: the proband had to be an alternative allele carrier; for homozygous alternative allele carriers, we only considered candidates with ≤1 read supporting the reference allele; minimum depth in the parent of 12 reads for the autosomes and 6 reads for hemizygous chromosomes; maximum of 1 read supporting the alternative allele in the parent; maximum allelic balance for parent of 0.05; minimum depth of 12 reads for proband; minimum allelic balance for proband of 0.15; maximum of 10 possible and 3 likely carriers beyond the descendants of the parent pair; maximum of 10% average soft clipping per read covering the DNM candidate.

Furthermore, we removed probands from the DNM analysis with more than 10% average soft clipping per read, 1.5% average fraction of ambiguous bases (Ns) in the read alignment, and 300 DNM candidates.

Quality control using transmission in three-generation families. The *de novo* candidates from the previous section present in three-generation families were used to tune the *de novo* calling by using propagation of alleles to the next generation. We took discrepancy between haplotype sharing and segregation of *de novo* alleles as an indication that the *de novo* candidate was not present in the germline. More specifically, we dichotomized these *de novo* variants by whether they were consistent with haplotype sharing among the probands' chromosomes and the chromosomes inherited in the offspring from the proband (Fig. 1c). We restricted ourselves to cases where two or more offspring shared distinct haplotypes from the proband in three-generation families.

We imposed a biological constraint on the DNM candidates on the X chromosome: that is, male probands and offspring carriers had to be homozygous for the alternative allele, otherwise we considered it inconsistent. To avoid inconsistent calls due to low-quality genotypes in the offspring, we treated genotype calls of the offspring as missing if they did not meet the following requirements: at least two reads supporting an alternative allele; allelic balance greater than or equal to 0.1; and depth greater than or equal to ten reads.

This evaluation of the *de novo* candidates gave us a binary outcome Y, which we incorporated into the following generalized additive model:

$Y \sim s(\text{Psroband\_allelic\_balance}) + s(\text{oxoG\_FoxoG\_metric}) + s(\text{Trio\_NPOSS}) + \text{gatk\_filter}$ where $s()$ denotes a smooth term and the covariates are described in the following list: Proband_allelic_balance, allelic balance of the proband; oxoG_FoxoG_metric, 0 for non-C>A substitutions and for C>A substitutions it measures the strand-specific affinity of the C>A variant, as defined in ref. 30; Trio_NPOSS, the number of possible carriers of the DNM allele beyond the descendants of parent-family pair; gatk_filter, a binary covariate dichotomizing whether the variant passes variant filters recommended by GATK best practices.

Using the fitted model, we predicted the response for all *de novo* candidates, including those not present in three-generation family probands. We used a predicted probability of being transmitted of strictly greater than 0.8 to define high-quality *de novo* variants. In total, we identified 108,778 high-quality DNMs, consisting of 7,401 indels (2,071 insertions and 5,330 deletions) and 101,377 SNPs (69,907 transitions, 31,470 transversions; Supplementary Table 4).

Monozygotic twin discordance. For DNMs identified in probands with a sequenced and genotyped monozygotic twin (Supplementary Table 3), we checked whether the genotype of the monozygotic twin was concordant with the DNM. We restricted this to comparisons where the genotype was not missing from the monozygotic twin of the proband. We treated all genotype calls of the monozygotic twin as missing if the depth was less than 10 and heterozygous calls as missing if they did not meet the following requirements: at least two reads supporting an alternative allele and allelic balance ≥0.1. Using these requirements, we could verify the absence or presence of the genotype in the twin of the proband for 6,000 out of 6,034 DNMs.

We randomly selected 38 discordant DNM calls from monozygotic twin probands for Sanger validation (Supplementary Table 5), to confirm presence in the proband and absence from the monozygotic twin. We designed primers using Primer 3 software (http://www.broadinstitute.org). We performed PCR and cycle sequencing reactions in both directions on MJ Research PTC-225 thermal cyclers, using a Big Dye Terminator Cycle Sequencing Kit version 3.1 (Life Technologies) and AMPure XP and CleanSeq kits (Agencourt) for cleanup of the PCR products. We loaded sequencing products onto a 3730 XL Genetic Analyser (Applied Biosystems) and analysed them with Sequencher 5.0 software (GeneCodes).

Out of the 38 discordant DNMs, 22 were validated, for 6 the Sanger genotypes did not match the WGS genotypes, and 10 failed in one or both of the monozygotic twins. For a conservative estimate of validation rate, we considered failed cases as non-validated resulting in a validation rate of 57.9%. If we restricted cases to successful Sanger genotyping, the validation rate was 78.6%.

Phasing of DNMs. We used two approaches to phase DNMs: haplotype sharing in three-generation families, and read pair tracing DNMs with phased variants.

In the former approach, we determined the parent of origin as in our previous analysis[4]. For example, if an offspring of the proband was a carrier of the DNM allele and had haplotype sharing to paternal chromosome of the proband, we assigned the mutation to the father. Meanwhile, if the offspring was not a DNM allele carrier, we would assign it to the maternal germline. We restricted the haplotype sharing analysis to segments of genetic length of at least 0.8 centimorgans and with 200 consecutive markers present in the long-range phased panel.

In the latter approach, we used a set of imputed and phased markers processed as described in ref. 31 to serve as a reference panel. We counted the number of read pairs traversing each biallelic combination between the DNM variant and the neighbouring imputed and phased variants. We also recorded the number of read pairs supporting three or more haplotypes and reads not consistent with any biallelic combination. We aggregated these numbers per parent pair and DNM site and phased DNMs by assigning the DNM allele to the parental chromosome with the read support. We only considered DNMs with read support for only one parent of origin. Finally, we aggregated the phases from both methods into a consensus phase. Of the 4,566 DNMs that were phased by both methods, we excluded 53 that were inconsistent.

Sex-specific estimation of age DNM accumulation. We estimated the sex-specific age effect by taking into account the unphased high-quality DNMs. More specifically for the three-generation probands, we were not able to phase DNMs falling outside the marker grid in our genotyping panels or sites where there was haplotype sharing between the probands' chromosomes. Similarly, not all DNMs were physically linked to a phased variant.

We fitted a Poisson regression model where we integrated over the latent state of the unphased DNMs. We define α_P and α_M as the paternal and maternal intercepts, and β_P and β_M as the sex-specific slopes. For each individual, we denote y_P and y_M as the sex-specific counts and A_P and A_M as the age of the parents at conception.

We modelled the sex-specific counts as Poisson random variables and we denote the Poisson likelihood function with f. If there are no ambiguities in the phase and if the age contributions are independent, then the likelihood function is

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}; y_P, y_M) = f(y_P; \alpha_P + A_P\beta_P) \times f(y_M; \alpha_M + A_M\beta_M)$$

However, as mentioned above, a fraction of the data are unphased. We define y_U as the observed number of DNMs with unknown phase and for convenience also define $y_T = y_P + y_M + y_U$, the total number of DNMs for the proband. We refer to the unobserved latent true number of paternal and maternal DNMs as y^*_P and y^*_M as, respectively. The likelihood is then the probability of observing the data under any combination of the latent phased counts that is compatible with the total number of DNMs for the proband and the observed number of phased paternal and maternal DNMs:

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}; y_P, y_M, y_U) = P(y^*_P \geq y_P, y^*_M \geq y_M, y^*_P + y^*_M = y_T; \boldsymbol{\alpha}, \boldsymbol{\beta}).$$

If the lack of phasing is independent of the parent of origin, then

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta}; y_P, y_M, y_U) &= \sum_{y^*_P = y_P}^{y_P + y_U} P(y^*_P \geq y_P, y^*_M \geq y_M, y^*_P + y^*_M = y_T | y^*_P; \boldsymbol{\alpha}, \boldsymbol{\beta}) \times P(y^*_P; \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &= \sum_{y^*_P = y_P}^{y_P + y_U} P(y^*_M \geq y_M, y^*_P + y^*_M = y_T | y^*_P; \boldsymbol{\alpha}, \boldsymbol{\beta}) \times P(y^*_P; \boldsymbol{\alpha}, \boldsymbol{\beta}). \end{aligned}$$

Note that $P(y^*_M \geq y_M | y^*_P + y^*_M = y_T, y^*_P; \boldsymbol{\alpha}, \boldsymbol{\beta}) = 1$ for $y_P \leq y^*_P \leq y_P + y_U$ and thus

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta}; y_P, y_M, y_U) &= \sum_{y^*_P = y_P}^{y_P + y_U} P(y^*_P + y^*_M = y_T | y^*_P; \boldsymbol{\alpha}, \boldsymbol{\beta}) \times P(y^*_P; \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &= \sum_{y^*_P = y_P}^{y_P + y_U} P(y^*_M = y_T - y^*_P | y^*_P; \boldsymbol{\alpha}, \boldsymbol{\beta}) \times P(y^*_P; \boldsymbol{\alpha}, \boldsymbol{\beta}). \end{aligned}$$

Note that the probability of whether the unphased mutation is of maternal or paternal origin is modelled through the age-effect parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$. In summary, we can address the lack of phasing by summing out the latent state:

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta}; y_P, y_M, y_U) &= \sum_{y^*_P = y_P}^{y_P + y_U} P(y^*_M = y_T - y^*_P | y^*_P; \boldsymbol{\alpha}, \boldsymbol{\beta}) \times P(y^*_P; \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &= \sum_{y^*_P = y_P}^{y_P + y_U} f(y^*_M = y_T - y^*_P \alpha_M + A_M\beta_M) \times f(y^*_P; \alpha_P + A_P\beta_P)). \end{aligned}$$

We implemented the model in R and found the maximum likelihood estimates using the nonlinear optimizing function nlm. We used a grid of starting points, where the intercepts (α_P and α_M) set to values 1, 5, and 10, and slopes (β_P and β_M) in 0, 0.5, 1, 1.5, and 2. For the sake of computational performance, we restricted the grid to plausible starting values; that is, where the sum of the slopes is 2 ($\beta_P + \beta_M = 2$). This resulted in 45 starting points and we used the maximum likelihood estimates corresponding to the highest likelihood value.

We considered different subsets of the data to observe whether the coefficients were robust to the phasing approach, as follows. Consensus: aggregation of the different phase approaches, excluding cases where the phase assignment differed. Physical: restricting to phase derived from read pair tracing. Physical HQ: subset of the physical set defined by requiring at least two read pairs supporting the phase. Three gen. pars.: restricting to sites phased by the three-generation approach. Three gen. pars. trans: subset of the Three gen. pars. set defined by observed transmission of the DNM allele to an offspring. Three gen. pars. conclusive: subset of the Three gen. pars. trans. set where the offspring carry different haplotypes from the proband.

The resulting coefficients are given in Supplementary Table 6. We also applied the regression method on the datasets resulting from stratifying the DNMs if they were present in C>G enriched regions (Supplementary Table 13). In addition, we ran a regression for each mutation class separately (Supplementary Table 9). For the stratified and the mutational class regressions, we used the following start values: $\alpha_P = 1$, $\alpha_M = 1$, $\beta_P = 1.5$, and $\beta_M = 0.5$. To assess significance of an age effect, we fitted a nested model with β_P or β_M set to 0, and evaluated the log likelihood difference between the full and nested model with a χ^2 approximation (Supplementary Tables 6 and 9). We also considered two other models with constrained ratios of the parental age effects: that is, β_P/β_M set to 1 and 3.12 (Supplementary Table 10).

Mutational classes. We used the consensus subset from the previous section for the analysis described here and in the following sections. Furthermore, we only considered the autosomes to avoid compositional bias due to the hemizygous X chromosome in male probands. Similar to refs 2, 14, we categorized the 12 different mutations into 6 classes, corresponding to the mutation and its complement. However, here we considered indels and dichotomized C>T substitutions depending on CpG context (C>T and CpG>TpG), resulting in a total of 8 classes.

We calculated the enrichment of the mutational classes between the phases. More specifically, we iterated over the mutational classes and dichotomized DNMs by whether they belonged to this class (foreground) or not (background), effectively creating a 2×2 table. This table was then used as input to the Fisher's exact test in R (fisher.test). We ordered the levels such that unconditional odds ratios were:

$$\mathrm{OR}_{\mathrm{unconditional}} = \frac{\#(\mathrm{background, father})\#(\mathrm{foreground, mother})}{\#(\mathrm{foreground, father})\#(\mathrm{background, mother})}.$$

The term '#(…)' denotes the number of DNMs in the category.

Mutation rate per generation. We estimated the number of reference bases that were accessible by short read sequencing by averaging the coverage in 10,000 base windows for a random subset of 100 probands and counting windows where the average coverage was above 12× and below 120×. We only considered the autosomal genome and only included reads with mapping quality greater than 20. This resulted in 268,289 windows or 2,682,890,000 base pairs (R) within our coverage range.

The mutation rate per base pair per generation was estimated by dividing the average number of DNMs (μ_a) by twice the R count.

$$\hat{\mu}_g = \frac{\hat{\mu}_a}{2 \times R}.$$

This resulted in a rate of 1.28×10^{-8} per base pair per generation. We estimated the average generation time (g) in our set by calculating the average parental age at conception for both parents, resulting in an estimate of $\hat{g} = 30.1$. Subsequently, we estimated the mutation rate per base per year by dividing the mutational per generation with the estimated generation time.

$$\hat{\mu}_y = \frac{\hat{\mu}_g}{\hat{g}}.$$

This resulted in a rate of 4.24×10^{-10} per base pair per year.

We next estimated the false positive rate in regions within our coverage range. We applied the DNM extraction procedure on simulated sequence data from eight trios. Instead of simulating ~70 DNMs per proband, with inherent limited precision per genomic window, we simulated on average 290,162 SNPs and 2,932 indels uniformly across the genome of each proband with mason_variator using the

default settings (version 2.0.0 (05113b7)). Subsequently, we simulated 334,054,203 read pairs with mason_simulator (version 2.0.0 (2f7f307)) with Illumina-like errors per family member for all trios. The parents were simulated as homozygous reference at all positions, while the proband was simulated as heterozygous at loci with simulated variation. The average, minimum, and maximum fragment sizes of the read pairs were selected to be 500, 301, and 1,500, respectively. The read length was selected to be 150, resulting in a targeted ~37 average genome-wide coverage. We called variants per trio with GATK-UG (version 2015.1-3.4.0.1-ga5ca3fc) and extracted DNMs as with the real data. We used the generalized additive model to define high-quality DNMs as with the real data, with two exceptions: that is, we set Trio_NPOSS and the oxoG_FoxoG_metric to zero for all DNM candidates. We defined false negative calls as those in which the simulated variant was not present among the simulated high-quality DNMs. For the SNPs, we required the alleles and position to match exactly, and we considered an indel to be discovered if there was a high-quality indel DNM within ten bases of the simulated indel. The last condition was to reduce undercalling of indels due to ambiguous representation of indels in VCF files.

We determined the negative rate to be 3.86%. We estimated a false positive rate of 3.0%, using monozygotic twin discordance for the regions satisfying the coverage range, resulting in a correction factor of 1.009. The mutation rates adjusted for false positive and negative rates were 1.29×10^{-8} per base pair per generation and 4.27×10^{-10} per base pair per year.

There are implicit assumptions behind these estimates, such as the following: the false negative rate is controlled by restricting the analysis to covered regions; the mutation rate is not different in regions that are inaccessible by short read sequencing; and that dividing the mutational rate per generation by the average generation time is a good surrogate for the year effect (see review[23] for a more comprehensive discussion).

The age effects for both parents allow us to address the last concern; that is, the sex-specific contributions of the germlines to the generation effect. We derived the sex-specific generation effects using our age effect and offset estimates from Supplementary Table 6 (the consensus dataset):

$$\hat{\mu}_{\mathrm{f},g} = 6.05 + 1.51 \times a_{\mathrm{f}}$$

$$\hat{\mu}_{\mathrm{m},g} = 3.61 + 0.37 \times a_m$$

with a_{f} and a_{m} representing the average age of fathers and mothers at conception. We incorporated these estimates into the modelling framework of ref. 24, which estimates long-term mutation rate per year and base for sex-specific generation effects.

$$\hat{\mu}_{a \text{ and } s,y} = \frac{\hat{\mu}_{\mathrm{m},g} + \hat{\mu}_{\mathrm{f},g}}{(a_{\mathrm{m}} + a_{\mathrm{f}}) \times R}.$$

We calculated mutation rate per base pair per year for several generation time combinations; the results are in Supplementary Table 17.

Comparison with sex-specific motifs from ref. 6. We determined whether the adjacent bases to DNMs were informative about the parent of origin. To allow comparison with the sex-specific motifs from ref. 6, we categorized all 12 possible substitutions between 4 bases into 6 mutational classes (C>A, C>G, C>T, T>A, T>C and T>G) along with the 4 possible reference bases adjacent to the substitution, resulting in 96 motifs (ACA>AAA, ACC>AAC, …. TTG>TGG, TTT>TGT). Furthermore, we restricted the analysis to SNPs as ref. 6 only considered SNPs. To determine whether a particular motif was informative about the parent of origin of a DNM beyond the mutational class, we used a logistic regression correcting for the mutational classes (the R implementation is in the Supplementary Information). More specifically, the parent of origin was used as a dependent variable, whereas mutational class and motif were used as additive covariates. We ran a separate logistic regression for each of the 96 motifs; the significance of the motif term was assessed with a likelihood ratio test. The estimates for the motif covariates are in Supplementary Table 7 and a description of the discrepancies between the studies is in the Supplementary Information.

Mutational composition regression. We modelled the mutational composition of phased DNMs against the respective parental ages as a linear function. For each year of conception and parent of origin, we aggregated the number of DNMs per mutational class. Subsequently, we calculated the mutational class fraction in each aggregate. We regressed the fractions as a function of parental age and sex for each mutational class, resulting in eight regressions. The numbers of probands per year and parent-of-year aggregate were used as weights for the lm function in R. We restricted the analysis to aggregates with more than one proband. The regression results are in Supplementary Table 11.

Regional analysis of DNMs and variants. We aggregated the DNMs in 2-Mb non-overlapping windows by parent of origin and mutational class. In addition, we counted the number of C>G variants per non-overlapping 1-Mb window. Subsequently, we calculated the C>G DNM rate or C>G variant fraction per window by normalizing the C>G DNM/variant counts by the occurrences of C/G bases in the reference within the coverage range of 10^5–10^6 reads in the Icelandic dataset. We only considered windows where at least 50% of the reference bases were covered within the coverage range (10^5–10^6 reads). We defined the C>G enriched regions as the 10% of the genome windows with the greatest number of C>G SNPs normalized by the reference GC composition (Supplementary Table 12). The covered reference genome in C>G enriched regions had a slightly lower CG composition (39.6%) than the rest of the genome (40.9%).

The regional enrichments of maternal DNMs were not affected by the removal of DNMs in segmental duplications, repeats, or regions with abnormal read coverage or false negative rates (Extended Data Figs 4 and 5).

Segmental duplication, repeat, and coverage annotation. We annotated whether the DNM was in a segmental duplication, repeat, or an abnormally covered region. We fetched the genomicSuperDups (version 2014-10-14) and rmsk tables (version 2014-01-11) from the University of California, Santa Cruz (UCSC) genome browser that corresponded to the segmental duplication and repeat tracts, respectively. We used tracts from the 'Mutation rate per generation' section to estimate regions of abnormal coverage (<12× or >120×; 10,000-bp windows) and elevated false negative rates (>5%; 100,000-bp windows).

We intersected the DNM positions with the intervals from the segmental duplication, repeat, coverage, and false negative rate tracts. Subsequently, we aggregated the number of phased DNMs per non-overlapping 2-Mb window, excluding DNMs within segmental duplications, annotated repeats (Extended Data Fig. 5), abnormally covered regions, and regions with a high false negative rate (Extended Data Fig. 4). Finally, we restricted the phased DNM aggregation only to DNMs identified in probands from three-generation families (Extended Data Fig. 4).

Clustered DNMs. We defined a set of DNMs occurring in a proband on a single chromosome as clustering if the difference between each pair of adjacent DNMs in the cluster was less than $10^{14/3} = 46{,}415$. This threshold was chosen to avoid edge effects when plotting histograms of mutations on the log scale with a bin width of $\log_{10}(1/3)$. We note that the results are not sensitive to a choice of bin width between 10^4 and 10^5.

DNM strand concordance. We define a pair of clustered C>G DNMs as strand concordant if they both mutate a reference C to a G, or a reference G to a C. We estimated strand concordance among clustered C>G DNMs with

$$\frac{\text{number of concordant pairs of clustered C} > \text{G DNMs}}{\text{number of pairs of clustered C} > \text{G DNMs}}.$$

SNP strand concordance. We identified all pairs of C>G SNPs present in the Icelandic population that were within a 1-Mb distance of each other and in strong linkage disequilibrium ($r^2 > 0.9$). As we were counting pairs, we only used the first seven SNPs in each cluster of linked SNPs to avoid placing too much weight on very large clusters of linked SNPs. We defined a pair of C>G SNPs as concordant if the two C alleles mostly occurred on the same haplotype in the population. We then estimated strand concordance among linked C>G SNPs in the population with

$$\frac{\text{number of concordant pairs of linked C} > \text{G SNPs}}{\text{number of pairs of linked C} > \text{G SNPs}}.$$

NCOGC regional analysis. We analysed the regional NCOGC rate using a previously described set of gene conversions[21] with correction for the regional base composition of the reference genome. We estimated the NCOGC rate inside and outside the C>G enriched regions for both parents. We found significantly higher maternal NCOGC rate within C>G enriched regions than outside them (2.74×10^{-5} and 1.12×10^{-5} per base pair per generation, respectively; Supplementary Table 16). No such difference was found for the paternal NCOGC rate.

The enrichment of NCOGCs with C>G enriched regions and the impact of the mother's age on the NCOGC gene conversion rate both increased slightly when we adjusted only for the maternal recombination rate. After accounting for distance from telomeres, we did not observe an increased maternal recombination rate inside the C>G enriched regions.

1000 Genomes. We lifted the 1000 Genomes dataset (phase 3) to build hg38 with gortools and we restricted the analysis to variants that mapped unambiguously. We used bedtools[32] (version 2.25.0-76-g5e7c696z) map command to aggregate the 1000 Genomes variants per 1-Mb window, where we counted variants per mutational class in the same manner as the DNMs. We excluded windows with fewer than 5,000 variants in the 1000 Genomes dataset and/or fewer than 500,000 reference bases covered in the Icelandic dataset. For each window, we normalized the C>G counts by the occurrences of C/G bases in the reference within the coverage range of 10^5–10^6 reads in the Icelandic dataset. In addition to analysis

of all variants, we considered two subsets of the data: variants with minor allele frequency strictly less than 1% (rare) or greater than 1% (Extended Data Fig. 6).

Archaic hominins. Reference 33 created a list of polarized variants (using the chimpanzee as an outgroup) where the Altai Neanderthal and Denisova specimens were homozygous with same derived allele, whereas the ancestral allele was at high frequency (90%) or fixed in modern humans (see Supplementary Information 18 provided by ref. 33). We fetched this list of variants at http://cdna.eva.mpg.de/neandertal/altai/AltaiNeandertal/catalog/ArchaicSpecific/ArchaicDerived_SNC_bothgq30.all_combined_maxsco_ranked.tsv. We lifted the coordinates of the variants from build hg19 to hg38, with the liftOver tool (version 247). Similar to the 1000 Genomes processing, we aggregated the variants in 2-Mb windows using the bedtools map command and counted variants per mutational class. We excluded windows with fewer than 1,000,000 reference bases covered in the Icelandic dataset and normalized the C>G mutation counts with the C/G composition of the reference in the window.

C>G divergence between primate species. We used EPO alignments of eight primate species from Ensembl (v84) available at ftp://ftp.ensembl.org/pub/release-84/maf/ensembl-compara/multiple_alignments/epo_8_primate/. We processed alignment blocks by keeping only one sequence per species, which best matched the consensus of the alignment block. Subsequently, we oriented the strand of alignment blocks by the reference strand of the human genome: that is, the human sequence was always on the forward strand. The filtering and strand orientation was done with the programs mafDuplicateFilter and mafStrander, respectively (mafTools[34]; https://github.com/dentearl/mafTools). We only considered alignment blocks containing more than 1,000 bases of human sequence.

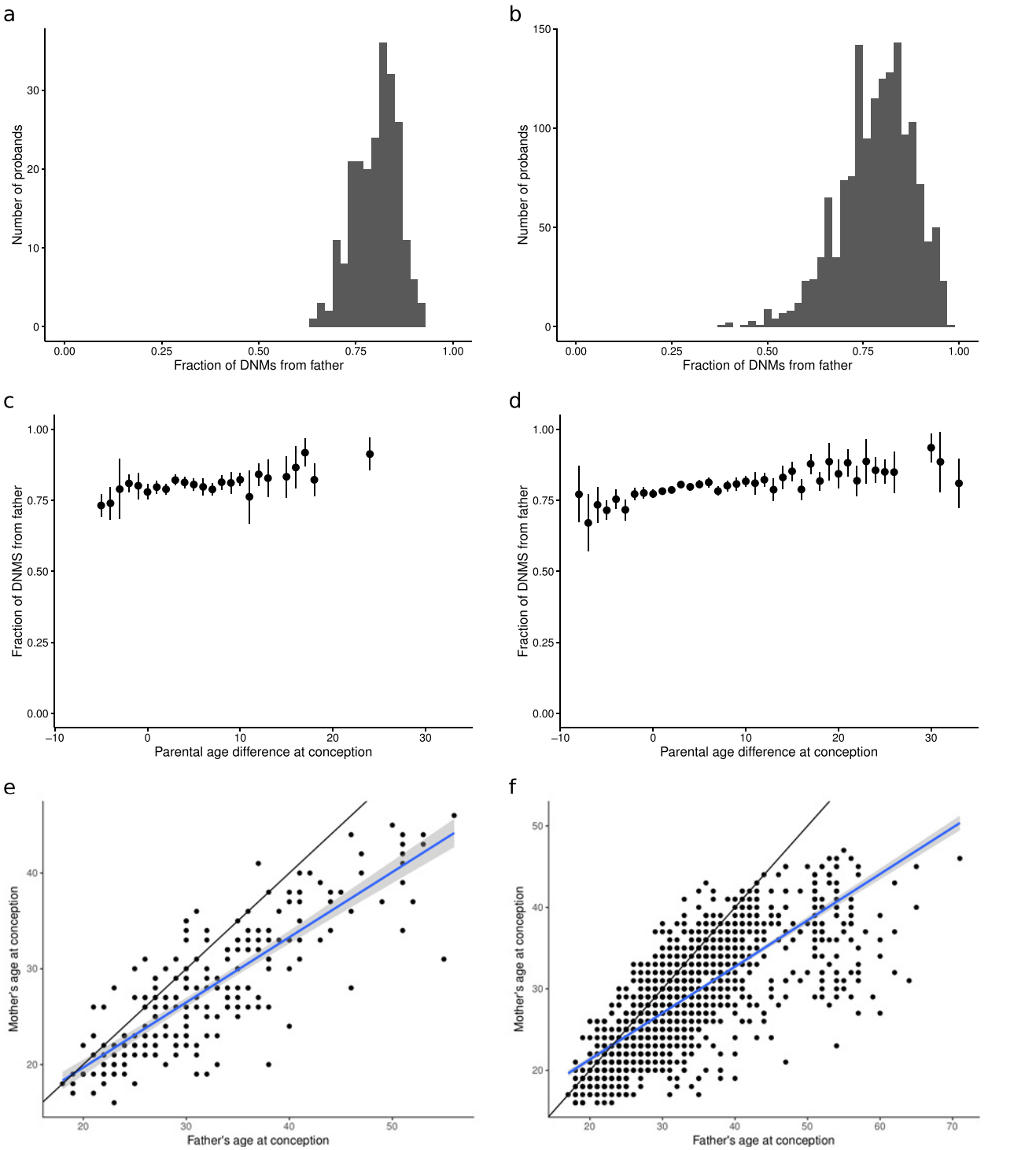
We parsed the alignments using the position on the human genome as a reference. More specifically, we iterated over each alignment position and recorded a sequence difference between two species only if the position was additionally matched to a position in the human genome and not to a gap. The alignment to a position in the human genome was required for placing the difference in the human genome as we then aggregated these differences in 1-Mb non-overlapping windows across the human genome. We excluded windows from pairwise-comparison analysis with alignment coverage of fewer than 500,000 bases of the human reference. We calculated the normalized C>G divergence fraction between pair species by dividing the occurrences of C>G differences by the number of all differences in the window.

We used the normalized C>G divergence for the reference genomes of four hominoid species and baboons ($d_{b,o}$, $d_{b,g}$, $d_{b,c}$, and $d_{b,h}$, for orangutans, gorillas, chimpanzees, and humans, respectively) to calculate the divergence ratios $d_{b,h}/d_{b,o}$, $d_{b,h}/d_{b,g}$, and $d_{b,h}/d_{b,c}$. These divergence ratios deviate from a value of 1 when orangutans, gorillas, or chimpanzees exhibit a fraction of C>G substitutions relative to baboons that differs from that observed in the sequence divergence between humans and baboons. We estimated the relationship between the divergence ratios and the C>G enrichment in humans by fitting a linear model with $d_{b,h}/d_{b,o}$, $d_{b,h}/d_{b,g}$, or $d_{b,h}/d_{b,c}$ as a dependent variable and the C>G variant fraction in Iceland as a covariate.

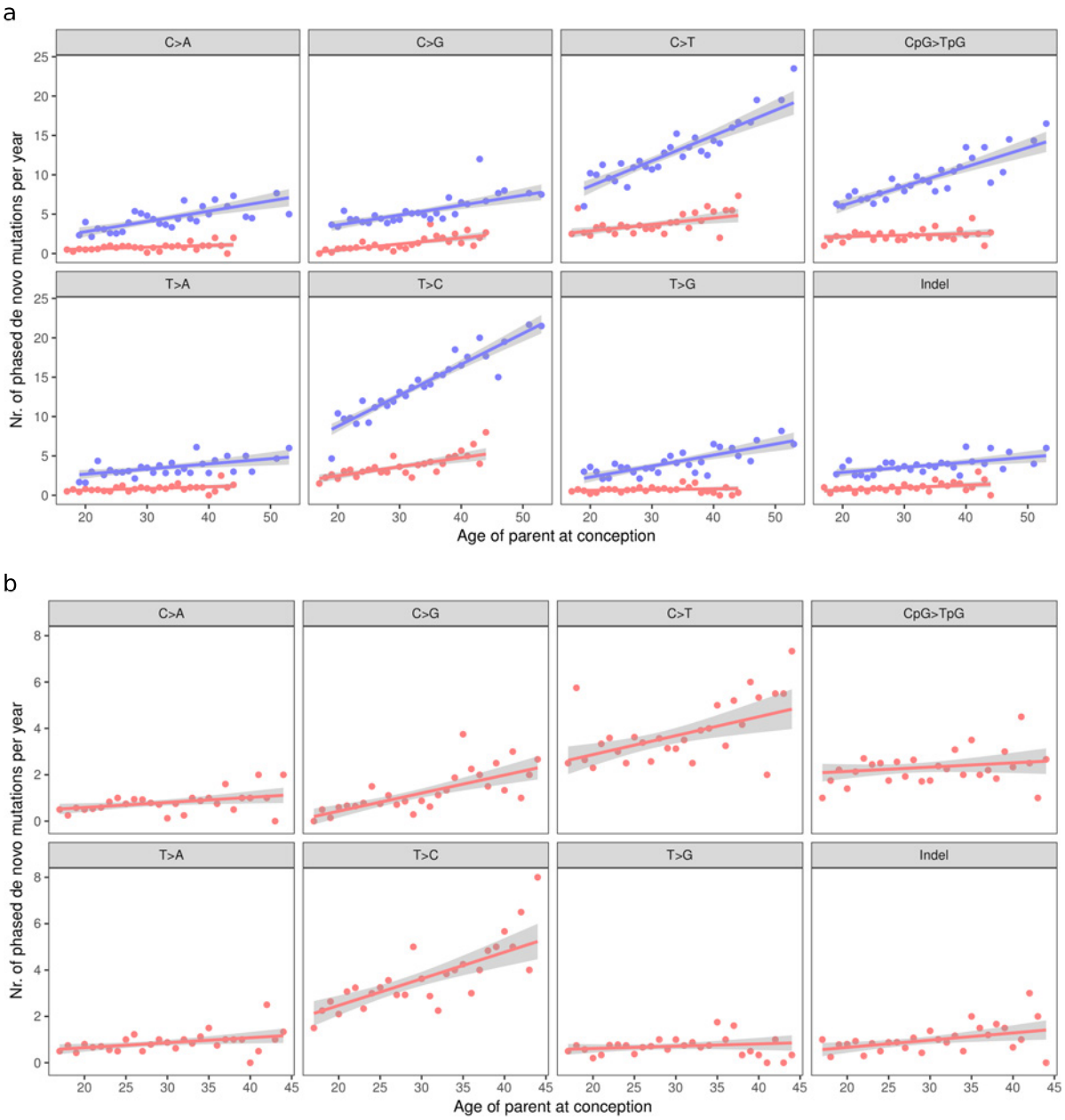
In addition to analysing the pairwise divergence, we considered the alignment of macaque, baboon, orangutan, gorilla, chimpanzee, and human jointly. More specifically, we identified sites where at least one of species of orangutan, gorilla, chimpanzee, and human carried a different nucleotide from the macaque and baboon. We restricted results to sites with only two types of nucleotide: where macaque and baboon had the same nucleotide (ancestral) and a subset of the orangutan, gorilla, chimpanzee, and human species had the same alternative nucleotide (derived) with respect to the macaque and baboon species. If there were gaps for any of the orangutan, gorilla, chimpanzee, and human species, and for macaque and baboon, we omitted the site.

We aggregated the sites per lineage in the phylogenetic tree in a parsimonious manner. For example, if human and chimpanzee shared a derived nucleotide, whereas the orangutan and gorilla had the ancestral nucleotide, we assigned the nucleotide difference to the ancestral lineage of chimpanzees and humans. We omitted sites that were not congruent with the species tree (macaque, (baboon, (orangutan, (gorilla, (chimpanzee, human))))). We normalized the lineage-specific C>G divergence similarly as in the pairwise case, by dividing the number of C>G differences by the total number of differences per lineage in 1-Mb windows. We modelled the dependency of C>G relative divergence on the C>G mutation rate in Icelanders in 1-Mb windows with a linear model for each lineage. For the lineage-specific linear fits, we omitted windows with fewer than 50 C>G events mapped to the lineage. We report the slope from the linear fit for each lineage in Extended Data Fig. 10a.

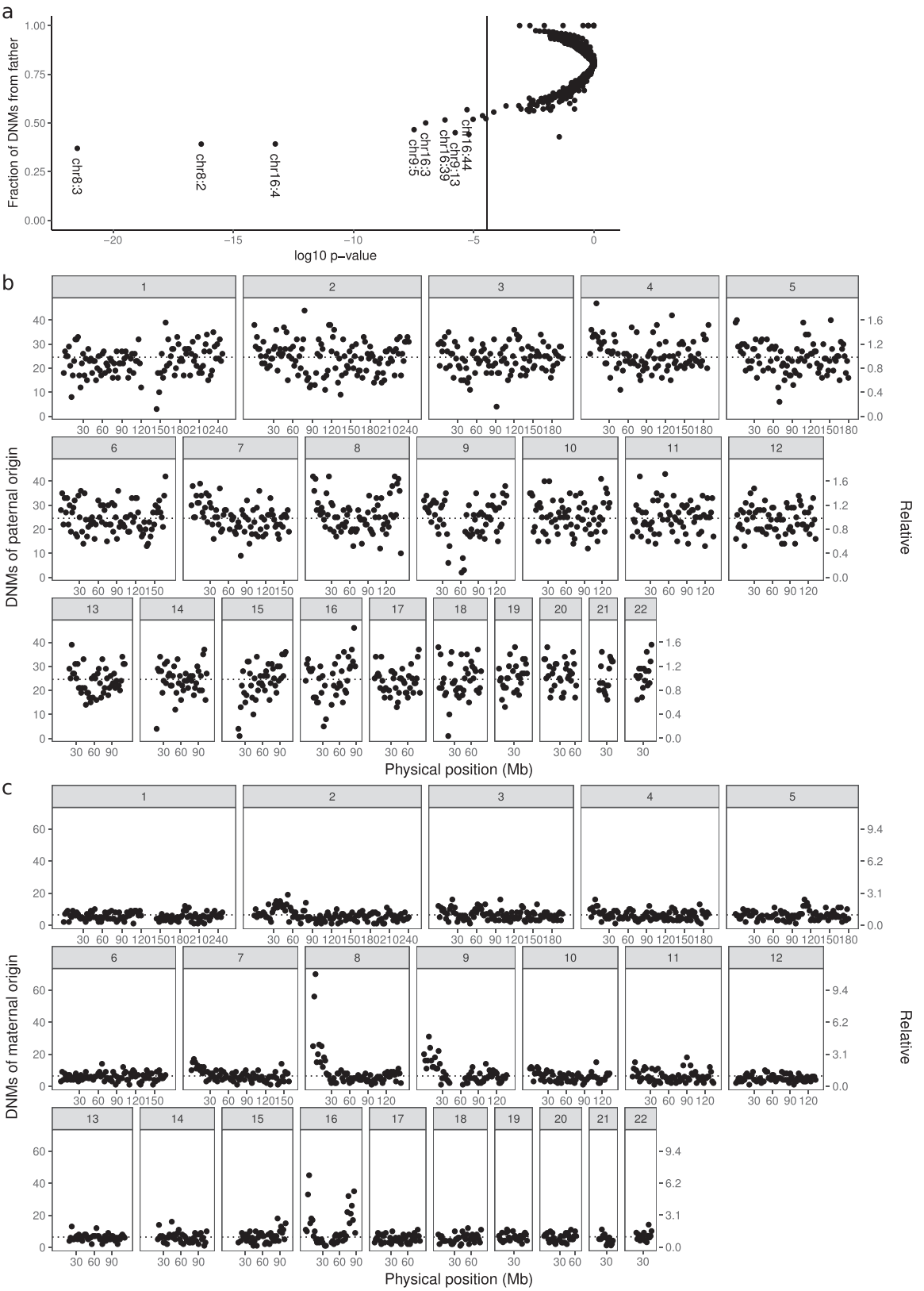
27. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
28. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
29. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
30. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41,** e67 (2013).
31. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462,** 868–874 (2009).
32. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
33. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505,** 43–49 (2014).
34. Earl, D. *et al.* Alignathon: a competitive assessment of whole-genome alignment methods. *Genome Res.* **24,** 2077–2089 (2014).



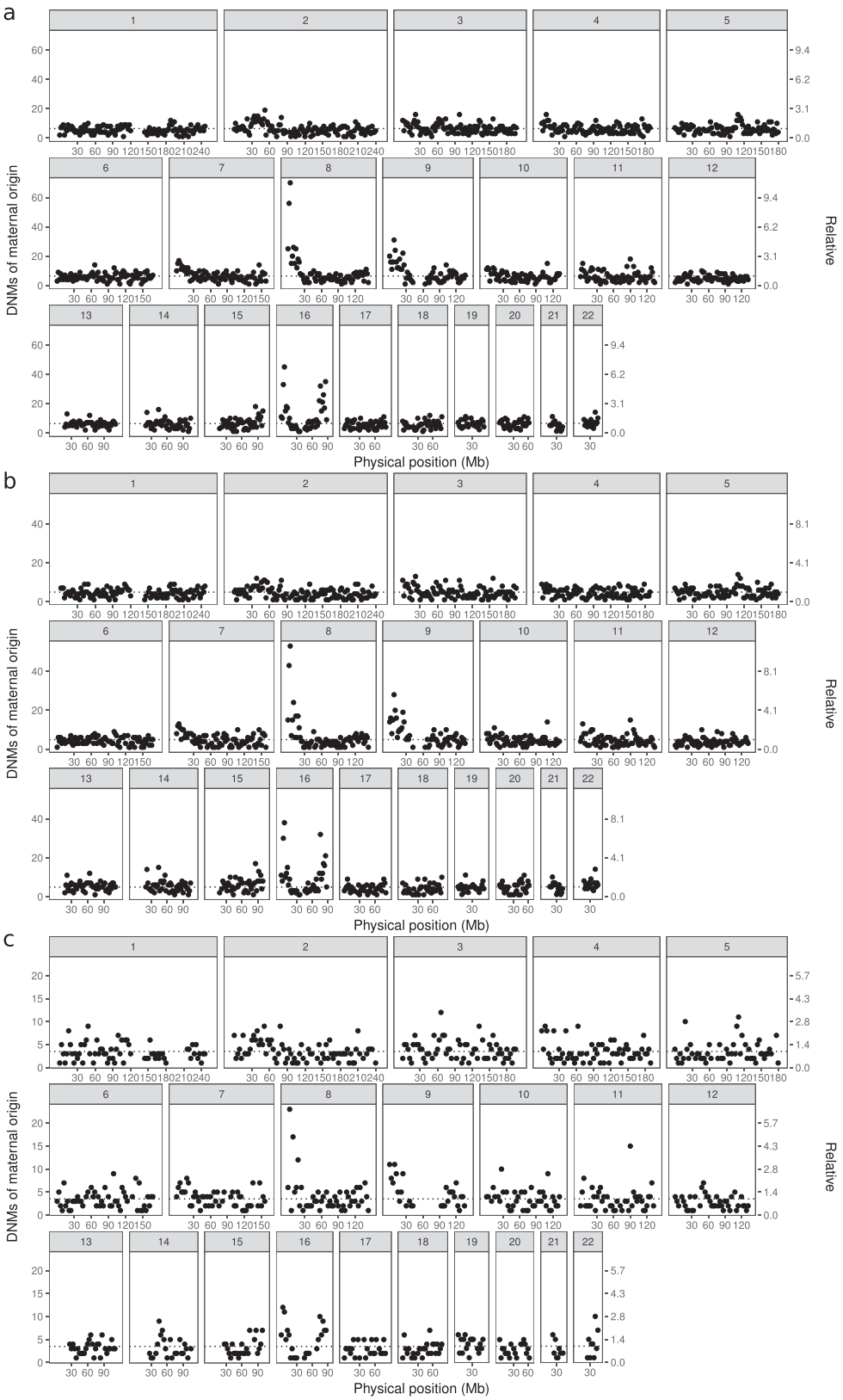
Extended Data Figure 1 | Fraction of DNMs of paternal origin. **a**, **b**, Fraction of DNMs from father per proband. **c**, **d**, Fraction of DNMs from father per proband against parental age difference at conception (father's age − mother's age). **e**, **f**, Paternal age at conception against maternal age at conception. In **b**, **d**, and **f**, all probands with phased DNMs were used; meanwhile, in **a**, **c**, and **e**, the analysis was restricted to 225 three-generation probands. The vertical bars in **c** and **d** represent 95% confidence intervals using a normal approximation.



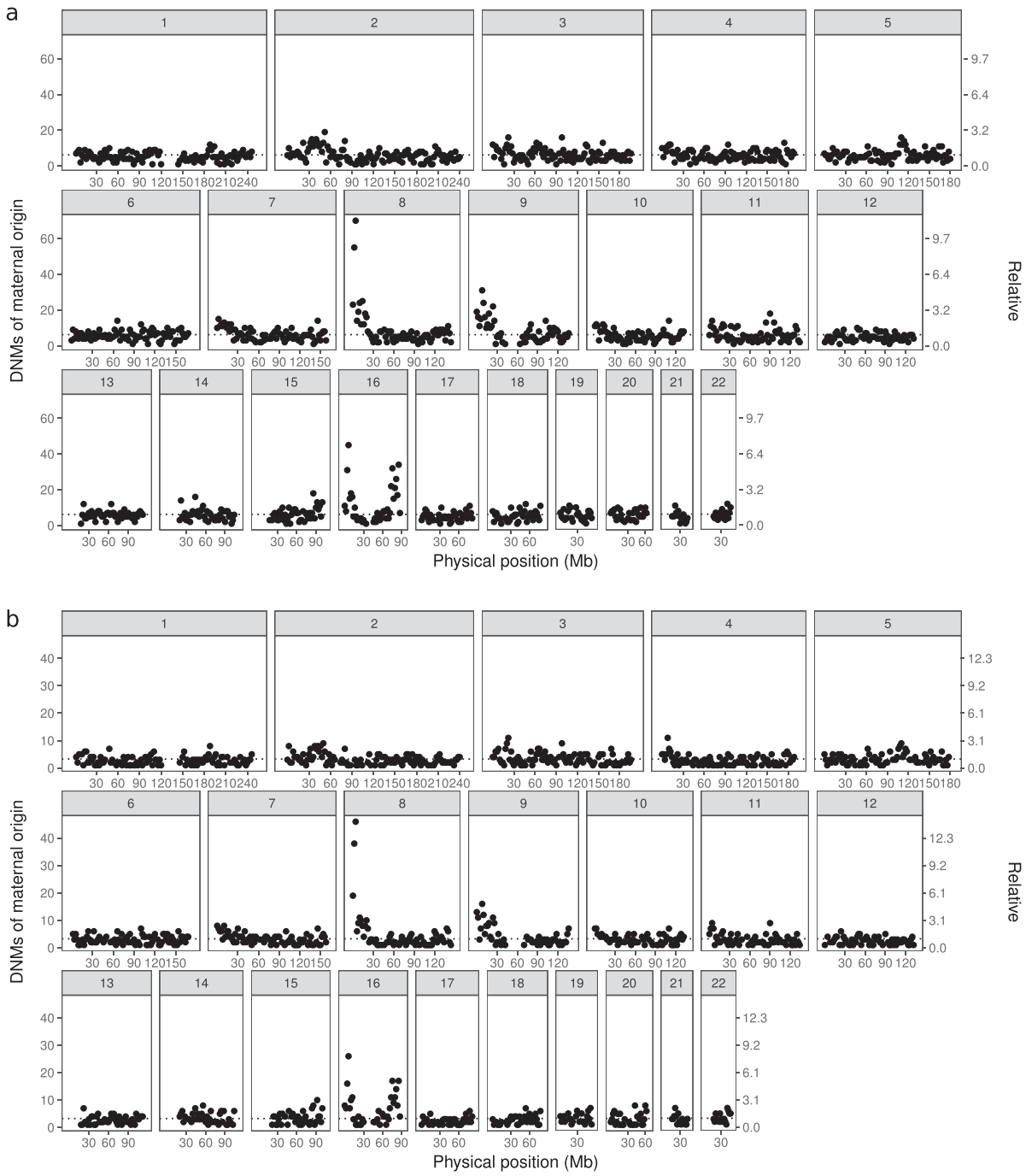
Extended Data Figure 2 | Absolute mutational spectra as a function of parents' ages at conception. a, Phased DNMs from both parents. **b**, DNMs from mothers. The mutations were aggregated per year, and parent age groups with only one proband were excluded. The numbers of probands per aggregate year were used as weights for linear regression. We restricted results to 225 three-generation probands for this figure. The *y* axes are different for **a** and **b**. The grey areas in **a** and **b** represent 95% confidence intervals using a normal approximation.



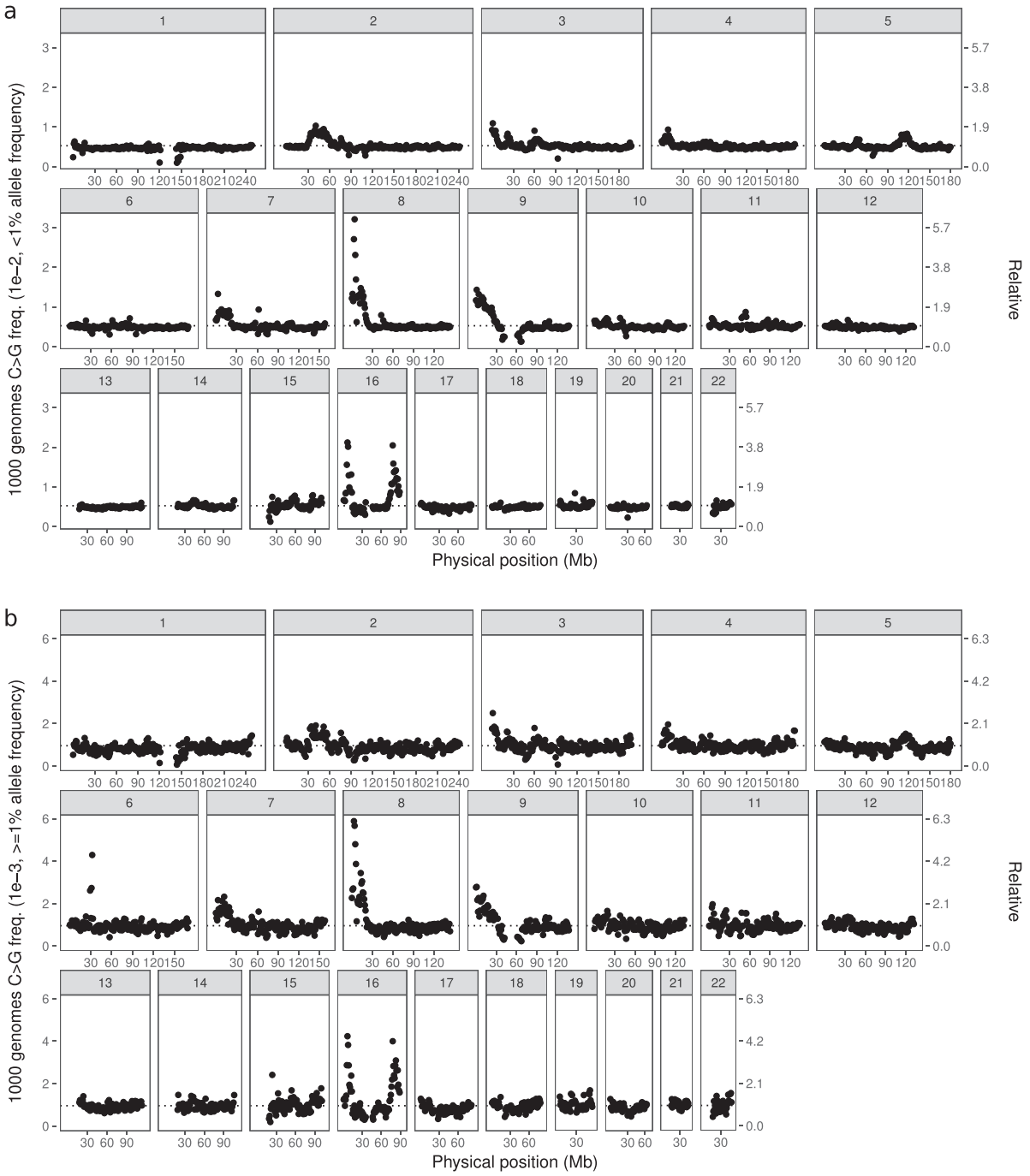
Extended Data Figure 3 | Localized enrichment of the DNM sex ratio. **a**, The maternal DNM enrichment in genome-wide context. We contrasted the local DNM sex ratio in each 2-Mb window against the genome-wide average, using a binomial test. The 2-Mb window number is depicted after the chromosome. The vertical line corresponds to $\log_{10}(0.05/1352)$. **b**, DNMs of paternal origin. **c**, DNMs of maternal origin. The dotted horizontal line is the genome-wide average. All phased DNMs (42,961) were used for this figure.



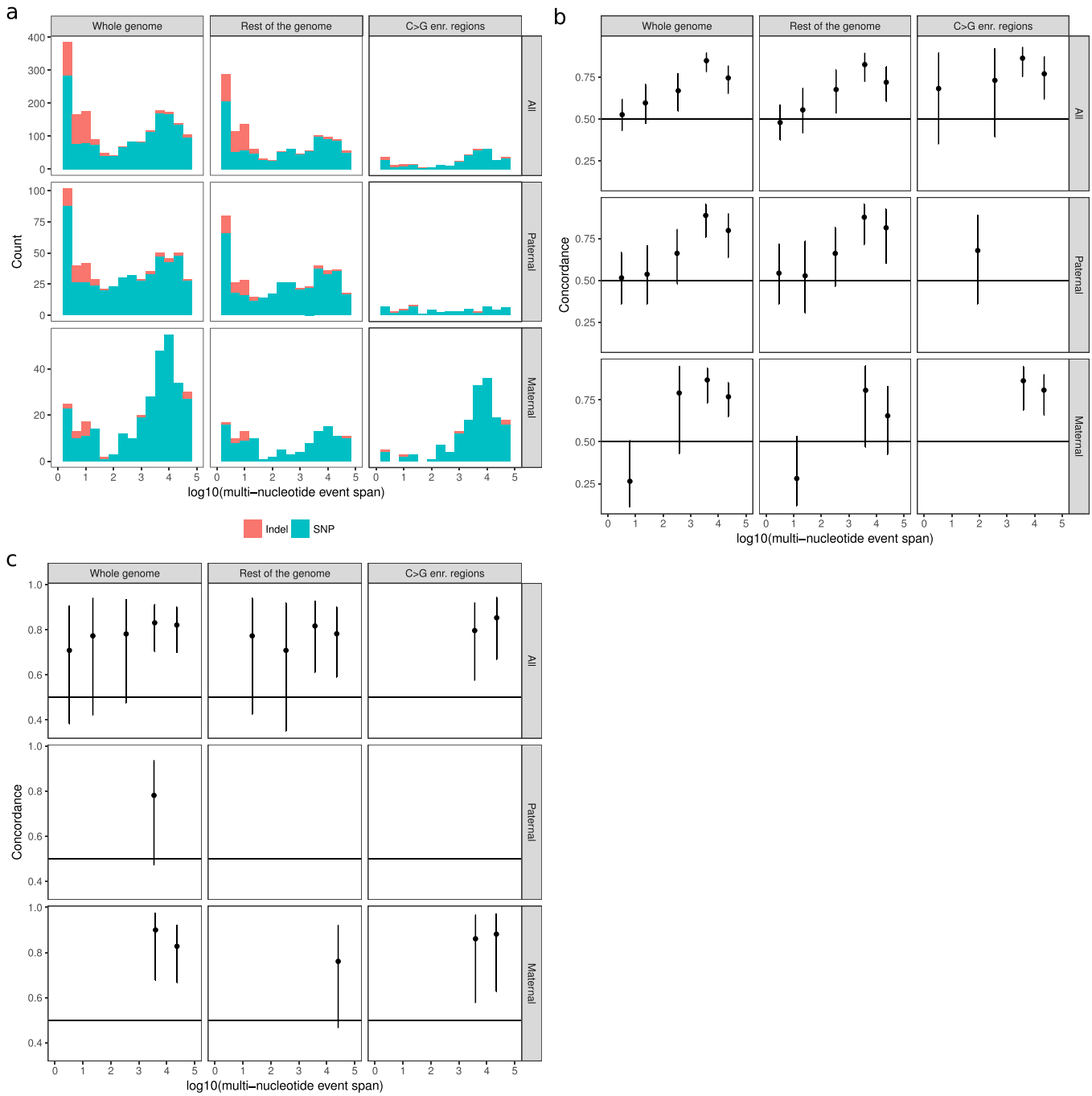
Extended Data Figure 4 | The number of maternal DNMs per window for various subsets. a, Excluding regions (10,000 bp) with average coverage of <12× or >120× (2-Mb windows). **b**, Excluding regions (100,000 bp) with false negative rate above 5% (2-Mb windows). **c**, Restricting to DNMs phased with the three-generation approach (15,746 DNMs, 3Mb windows).



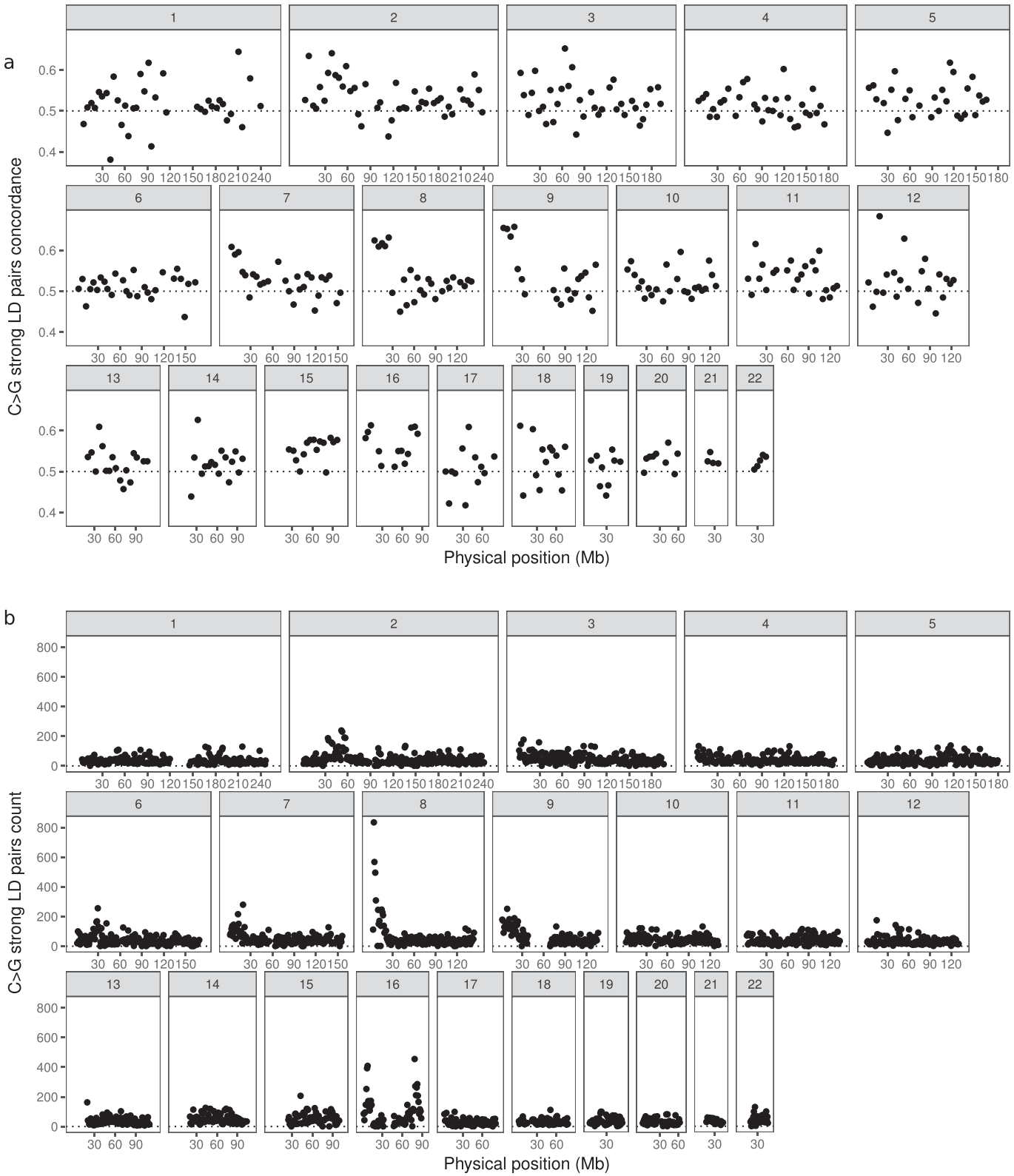
Extended Data Figure 5 | The number of maternal DNMs per 2-Mb window without DNMs in segmental duplications or repeat regions. a, Without DNMs in annotated segmental duplications. **b**, Without DNMs in annotated repeat regions.



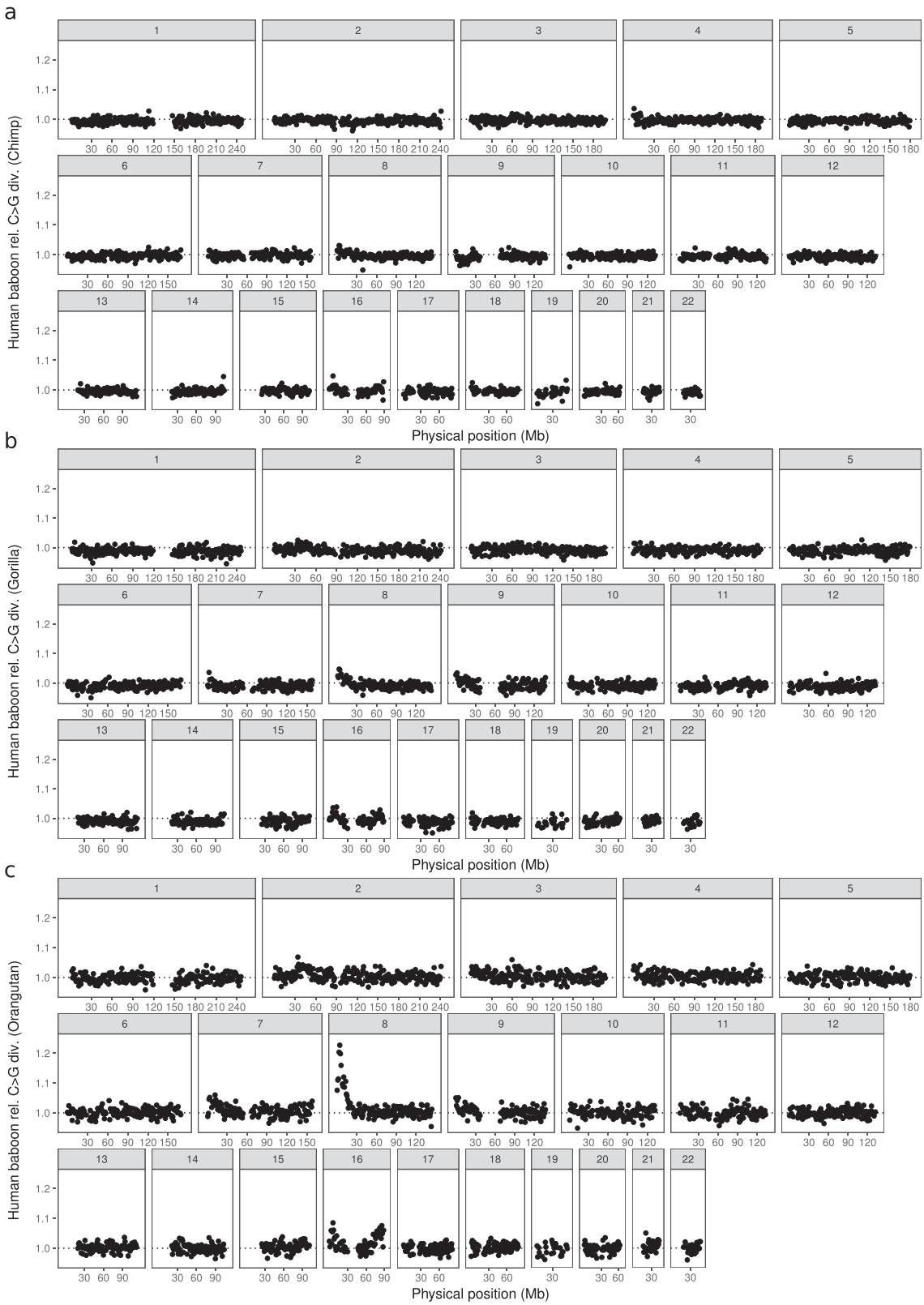
Extended Data Figure 6 | The frequency of C>G variants in 1-Mb windows for the 1000 Genomes dataset. a, Rare variants (<1%). **b**, Common variants (≥1%). The dotted horizontal line is the genome-wide average.



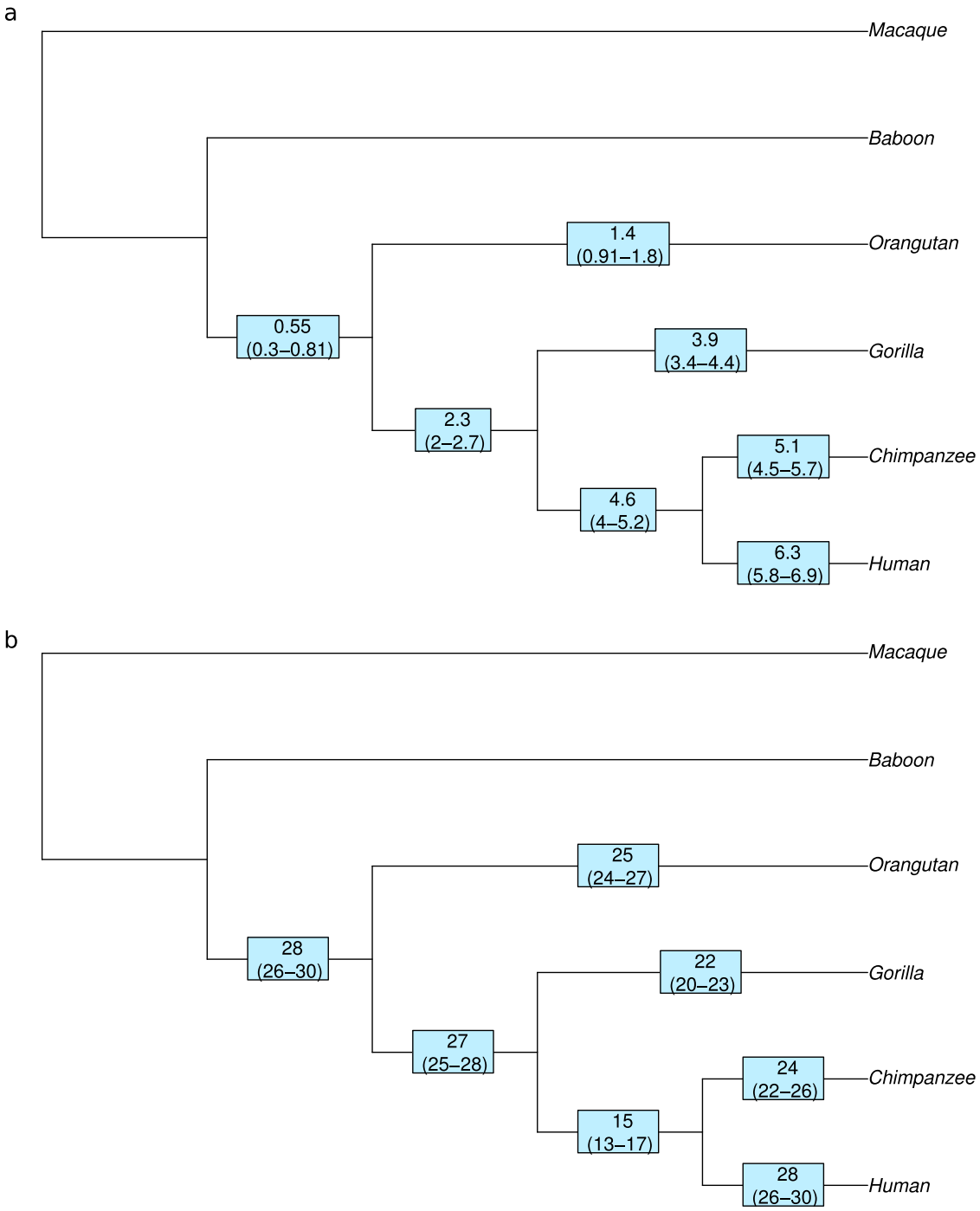
Extended Data Figure 7 | Strand concordance and cluster length. **a**, Span of multi-nucleotide events in base pairs. **b**, The concordance of C or G reference bases within multi-nucleotide events as a function of span length. **c**, Same as **b** except restricted to C>G DNMs. The vertical bars in **b** and **c** represent 95% confidence intervals using a normal approximation.



Extended Data Figure 8 | Strand concordance among C>G SNP pairs as a function of genomic position. a, The concordance ratio. **b**, The absolute concordance counts.



Extended Data Figure 9 | Local C>G enrichment in an evolutionary context for all of the autosomes. This figure corresponds to Fig. 4c–e for all of the autosomes. **a**, The C>G divergence between baboon and human, normalized by the C>G divergence between baboon and chimpanzee ($d_{b,h}/d_{b,c}$, 1-Mb windows). **b**, Same as **a** except the gorilla is used instead of the chimpanzee ($d_{b,h}/d_{b,g}$). **c**, Same as **b** except the orangutan is used instead of the chimpanzee ($d_{b,h}/d_{b,o}$).



Extended Data Figure 10 | The phylogenetic context of the dependence of C>G and T>A relative divergence on the rate in the Icelandic dataset. a, C>G relative divergence against C>G rate. **b**, T>A relative divergence against T>A rate. The dependency of the relative divergence (C>G or T>A) was modelled against the rate (C>G or T>A) in Icelandic dataset with a linear model. The coefficients of the slopes are reported in the blue rectangles with 95% confidence intervals below the estimates. The dependency of the C>G relative divergence on the C>G SNP patterns in humans reaches its minimum in the ancestral lineage of great apes.