nature genetics

Parent-of-origin-specific signatures of *de novo* mutations

Jakob M Goldmann[1,10], Wendy S W Wong[2,10], Michele Pinelli[3], Terry Farrah[4], Dale Bodian[2], Anna B Stittrich[4], Gustavo Glusman[4], Lisenka E L M Vissers[5], Alexander Hoischen[5], Jared C Roach[4], Joseph G Vockley[2,6], Joris A Veltman[5,7], Benjamin D Solomon[2,8,9], Christian Gilissen[5,11] & John E Niederhuber[2,9,11]

***De novo* mutations (DNMs) originating in gametogenesis are an important source of genetic variation. We use a data set of 7,216 autosomal DNMs with resolved parent of origin from whole-genome sequencing of 816 parent–offspring trios to investigate differences between maternally and paternally derived DNMs and study the underlying mutational mechanisms. Our results show that the number of DNMs in offspring increases not only with paternal age, but also with maternal age, and that some genome regions show enrichment for maternally derived DNMs. We identify parent-of-origin-specific mutation signatures that become more pronounced with increased parental age, pointing to different mutational mechanisms in spermatogenesis and oogenesis. Moreover, we find DNMs that are spatially clustered to have a unique mutational signature with no significant differences between parental alleles, suggesting a different mutational mechanism. Our findings provide insights into the molecular mechanisms that underlie mutagenesis and are relevant to disease and evolution in humans[1].**

Studies of *de novo* mutations (DNMs) in humans have estimated the mutation rate of single-nucleotide variants to be approximately 1×10^{-8} mutations per generation, giving rise to 45–60 DNMs per genome[2–5]. The susceptibility to DNMs varies by several orders of magnitude along the genome and may be influenced by factors such as nucleotide content, replication timing, distance to recombination hotspots, nucleosome occupancy, transcription, and chromatin 'openness'[4,6]. Several mechanisms of DNA mutation are known, most predominantly involving DNA replication[7]. The latter mechanism also explains the 3.9:1 ratio of DNMs on the paternal allele to the maternal allele, as there are many more germline cell divisions in spermatogenesis than in oogenesis[2]. We hypothesize that the different underlying biology of male and female gametogenesis results in differences in mutational signatures between paternally and maternally transmitted DNMs. These signatures will provide insight into the mechanisms underlying *de novo* mutations in human germline cells.

Studies to date have lacked sufficient sample size to determine the parental allele for large numbers of DNMs so as to compare DNMs of paternal and maternal origin. In this study, whole-genome sequencing (WGS) was performed on 832 offspring–parent trios, with an average of 60× coverage, by Complete Genomics Inc. (**Table 1**, **Supplementary Tables 1–4**; see Online Methods for a description of the cohort)[8]. After removing an outlier and one twin from each of the monozygotic twin pairs, *de novo* mutations were identified for the autosomes of 816 trios. A random forest classifier was used to remove potential false positives from the initial set of putative DNMs, resulting in 36,441 DNMs, or an average of 45 DNMs per individual (Online Methods and **Supplementary Tables 5–8**). Quality assessment of these results based on monozygotic twin concordance and Sanger validations of a subset of DNMs indicated high specificity (**Supplementary Tables 9–11**, Online Methods). Overall, the nucleotide substitution frequencies for DNMs were dominated by C–T and T–C changes, giving rise to a transition/transversion ratio (Ts/Tv) of 2.23. Haplotype assembly of all mutations successfully phased 19.8% of all DNMs, resulting in a set of 7,216 phased DNMs (Online Methods, **Supplementary Tables 12–15**). Assessing the parental origin of DNMs, we found that 5,640 DNMs were on the paternal allele and 1,576 on the maternal allele, giving rise to the expected median paternal/maternal ratio of 3.6:1 (**Supplementary Fig. 1**)[2,3].

Multiple studies have shown that the numbers of DNMs in offspring are positively correlated with increasing paternal age at the time of conception[2,3]. Using our phased DNMs, we were able to confirm

Table 1 Cohort description

Birth constellation	No. births	No. children	No. sequenced samples
Singletons	731	731	2,193
Dizygotic twins	35	70	140
Monozygotic twins	14	28	56
Triplet	1	3	5
Total	781	832	2,394

The cohort consists of 731 trios, 49 quartets, and one quintet, resulting in a total of 832 children.

[1]Department of Human Genetics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, the Netherlands. [2]Inova Translational Medicine Institute (ITMI), Inova Health Systems, Falls Church, Virginia, USA. [3]Telethon Institute of Genetics and Medicine (TIGEM), Naples, Italy. [4]Institute for Systems Biology, Seattle, Washington, USA. [5]Department of Human Genetics, Donders Centre for Neuroscience, Radboud University Medical Center, Nijmegen, the Netherlands. [6]Department of Pediatrics, Virginia Commonwealth University School of Medicine, Richmond, Virginia, USA. [7]Department of Clinical Genetics, GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre, Maastricht, the Netherlands. [8]Department of Pediatrics, Inova Children's Hospital, Inova Health System, Falls Church, Virginia, USA. [9]Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [10]These authors contributed equally to this work. [11]These authors jointly supervised this work. Correspondence should be addressed to C.G. (Christian.gilissen@radboudumc.nl) or J.E.N. (John.Niederhuber@inova.org).

Received 1 December 2015; accepted 26 May 2016; published online 20 June 2016; doi:10.1038/ng.3597

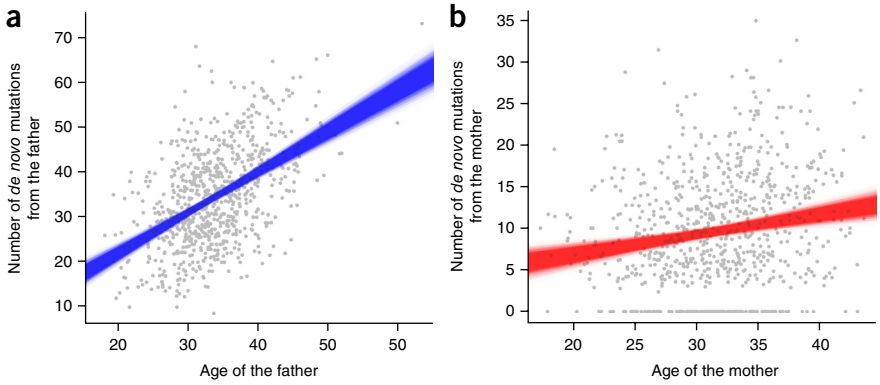

Figure 1 Correlation of paternal and maternal age with the number of DNMs with resolved parent of origin. (**a**,**b**) Simple linear regression plots on normalized number of phased DNMs versus the respective paternal (**a**) and maternal age (**b**), with 10,000 bootstrap resampling. The underlying data can be found in **Supplementary Table 14**. The number of phased DNMs was normalized by the proportion of phased variant for each proband. Where the number of normalized DNMs of a particular parental origin equals 0, this indicates that there are no DNMs in the proband that could be confidently assigned to the specified parent of origin. Regression plots for the observed number of phased DNMs are in **Supplementary Figure 2**, where a similar trend can be observed.

this correlation, and we found an increase of 0.91 paternally transmitted DNMs (95% confidence interval (CI) by bootstrap sampling: 0.81–1.02) per year (**Fig. 1a**, **Supplementary Tables 16** and **17**, **Supplementary Fig. 2**). Interestingly, our data also showed a smaller maternal age effect, consisting of 0.24 maternally transmitted DNMs (95% CI by bootstrap sampling: 0.15–0.34) per year (**Fig. 1b**, **Supplementary Table 18**). The result is consistent with a previous study on maternal age effect that used a subset of this cohort (693 singleton trios). In this previous study, a different algorithm was used to call *de novo* variants, and the maternal age effect was assessed by regressing both parents' ages on the total number of DNMs[9]. The finding of a maternal age effect is consistent with the speculation that spontaneous mutations accumulate over time in the female germline[10,11].

To identify local genomic factors that influence DNM susceptibility in male and female germline cells, we divided the human autosomes into 1-megabase (Mb) windows and examined the linear correlations between several genomic features and mutation rates that have previously been related to germline mutation rates[3,6,12] (**Fig. 2a**, **Supplementary Table 19**, **Supplementary Fig. 3**). For mutation rates in each of the age–gender groups, we performed multiple robust regressions using the subset of features selected by optimizing for residual variance (Online Methods).

Mutation rates in both older fathers and older mothers are strongly positively correlated with DNA methylation and negatively correlated with histone H3 Lys36 trimethylation (H3K36me3), indicating that the higher mutation rates are correlated with depletion of transcription. Interestingly, H3K9me3 was highly correlated with all but mutation rate within young mothers, and it was previously shown to account for more than 40% of the somatic mutation variation in cancer in the human genome[12].

We investigated how paternally and maternally derived DNMs were distributed across the genome. A multivariate Poisson hidden Markov model selected four hidden states of mutational patterns in young and old parents based on the Akaike information criterion

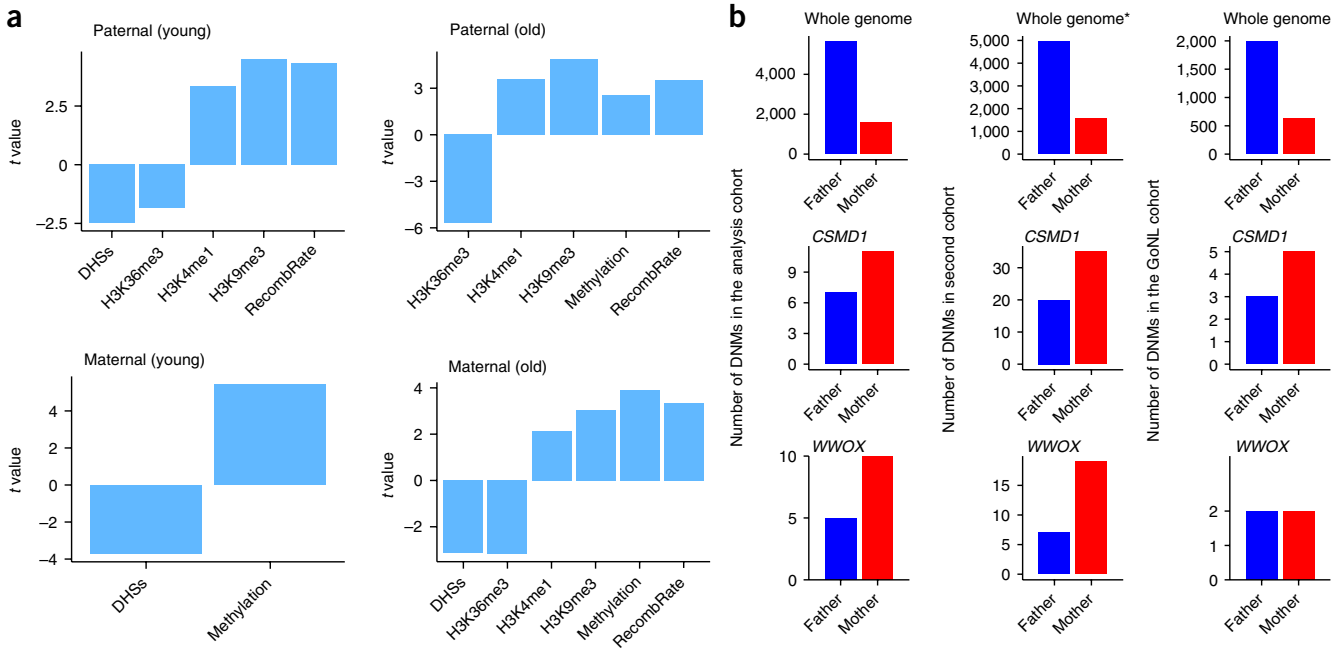

Figure 2 Regions enriched for maternally and paternally derived DNMs. (**a**) *t*-statistic of the features selected to be included in the multiple regressions of each of the age–gender category for DNM mutation rates. All features included have asymptotic approximate *P* value < 0.05. The mutation rates and values for each feature in each 1-Mb window are shown in **Supplementary Table 22**. (**b**) Number of DNMs on paternal and maternal allele in the whole genome and the genes *CSMD1* and *WWOX* in the analysis cohort (sequenced by CGI platform, based on all autosomes), in the second or validation cohort (independent samples sequenced by Illumina platform; * signifies data based on eight chromosomes only: chromosomes 2, 3, 4, 7, 8, 12, 14, 16), and in the GoNL cohort[15]. The number of counts is shown in **Supplementary Table 23**.

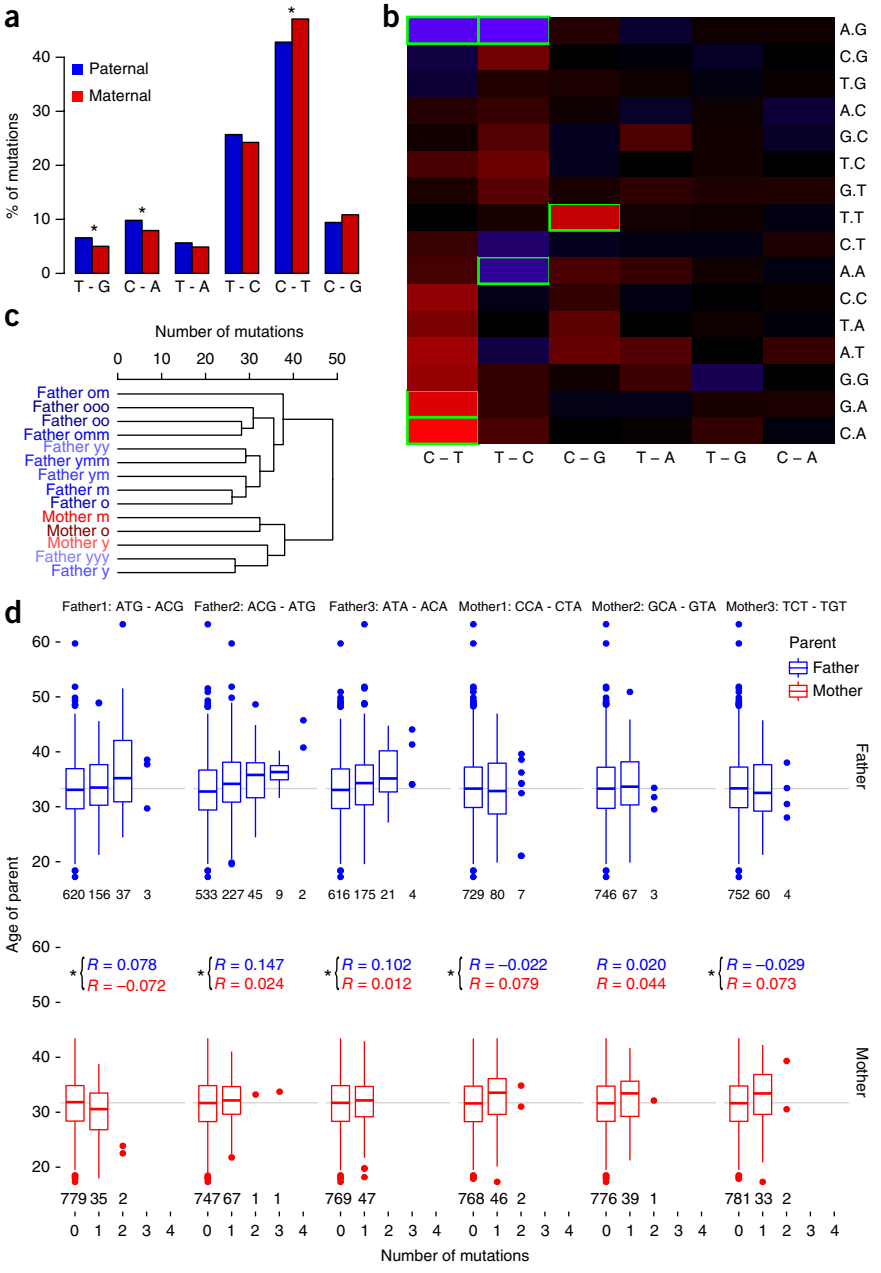

Figure 3 Differences in paternal and maternal mutation profiles and correlation with parental age at conception. (**a**) Nucleotide substitution ratios in maternal and paternal sets. Asterisks indicate substitutions whose ratios differ significantly between the two ($P < 0.05$ after Bonferroni correction). (**b**) Heatmap of parental mutation profile differences split by nucleotide substitutions (columns) and nucleotide context (that is, the adjacent nucleotides; rows); blue and red indicate overrepresentation in father or mothers, respectively. Green boxes highlight the mutation categories that differ more than 1% of mutation load with a bootstrapping P value <0.05. (**c**) Hierarchical clustering of paternal and maternal mutations, sorted by age of the parent and grouped into groups of approximately 500 mutations (Online Methods and **Supplementary Table 26**). Maternal signatures are more closely related to those of young fathers than those of young fathers are to those of old fathers ($P = 0.113$). Axis indicates number of mutations; the age categories are denoted as y = young, m = moderate and o = old, sorted from youngest to oldest as yyy < yy < y < ym < ymm < m < omm < om < o < oo < ooo. (**d**) Different coefficients of correlation for age of parent and number of mutations between the parents. Boxplots of the mutation categories highlighted in **b** (box, interquartile range; line, median; whiskers, extreme values <1.5 × interquartile ranges from box borders). Spearman correlation coefficients were Fisher-transformed to assess the significances of the difference. Asterisks mark categories that differ with $P < 0.05$ after bootstrapping. Numbers above the x axis indicate the number of offspring with the respective number of mutations.

(**Supplementary Note** and **Supplementary Table 20**). Notably, the two 1-Mb windows that exhibit the highest maternal mutation rates are spanned by one large gene each, *CSMD1* and *WWOX* (**Supplementary Tables 21–23**). *WWOX* is at a well-known fragile site in the genome[13] and is known to be involved in human gonad development[14]. We confirmed the same pattern of enrichment of maternal mutations in these two genic regions in an external control cohort of 656 trios with WGS by the Illumina platform, as well as in data from the GoNL project[15] (Online Methods, **Fig. 2b**).

Cancer genome sequencing studies have identified mutational signatures thought to reflect distinct underlying mutational mechanisms[16]. Applying this line of reasoning to DNMs, we hypothesized that DNMs of different parental origin show discriminative mutational signatures. Indeed, we observe significant differences in nucleotide substitution patterns between paternal and maternal DNMs. Paternally derived DNMs contain a higher frequency of T–G and C–A substitutions than maternal DNMs (Bonferroni-corrected bootstrapping test $P = 1.49 \times 10^{-2}$ and $P = 1.91 \times 10^{-2}$, respectively), whereas maternally derived DNMs contain more C–T mutations (Bonferroni-corrected $P = 2.7 \times 10^{-3}$, **Fig. 3a**, **Supplementary Table 24**). Following these initial observations, we investigated whether paternal and maternal DNMs give rise to different mutation signatures and compared nucleotide substitutions within the context of the adjacent nucleotides. Indeed, within the set of 7,216 phased mutations, we found significant differences between the nucleotide contexts of paternally and maternally derived DNMs (χ^2 test $P = 2.9 \times 10^{-8}$; **Fig. 3b**, **Supplementary Table 25**, **Supplementary Fig. 4**). More specifically, paternally derived DNMs are enriched in transitions in A[.]G contexts, especially ACG>ATG and ATG>ACG (Bonferroni-corrected $P = 1.3 \times 10^{-2}$ and $P = 1 \times 10^{-3}$, respectively). Additionally, we observed overrepresentation of ATA>ACA mutations (Bonferroni-corrected $P = 4.28 \times 10^{-2}$) for DNMs of paternal origin. Among maternally derived DNMs, CCA>CTA, GCA>GTA and TCT>TGT mutations were significantly overrepresented (Bonferroni-corrected $P = 4 \times 10^{-4}$, $P = 5 \times 10^{-4}$, $P = 1 \times 10^{-3}$, respectively). Interestingly, these differences between signatures of paternally and maternally derived DNMs became more pronounced with increasing age of the parents at conception. Unsupervised hierarchical clustering of trinucleotide DNMs binned by age and parent of origin almost perfectly separates paternally and maternally derived DNMs (**Fig. 3c**,

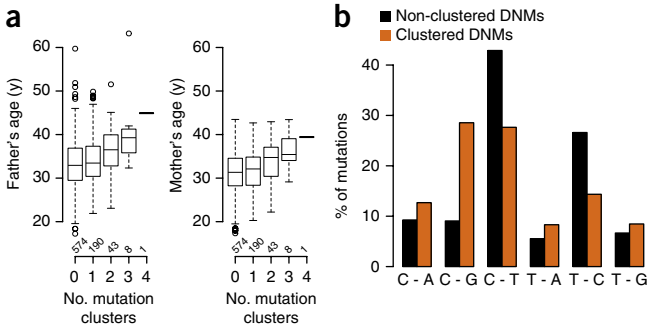

Figure 4 Mutation profiles of clustered DNMs. (**a**) Boxplots showing that the number of mutation clusters per individual rises with parents' age. The *y* axis shows boxplots of the father's and mother's age (box, interquartile range; line, median; whiskers, extreme values <1.5 × interquartile ranges from box borders). The *x* axis shows the number of mutation clusters that were identified. Numbers above the *x* axis indicate the number of trios per category. (**b**) Nucleotide substitution profile of clustered and nonclustered mutations. The *y* axis shows the percentage of mutations for each substitution; the *x* axis shows the six possible nucleotide substitutions. Mutations for which there was no evidence of clustering are in black and those with evidence for clustering in brown.

Supplementary Table 26). The only exceptions are DNMs from the youngest fathers, whose signatures are apparently more similar to maternally derived DNMs than to those from older fathers. Additionally, the observed parent-of-origin-specific mutations correlate with the age at conception of the respective parent (**Fig. 3d**).

The differences that we observe between signatures of paternally derived DNMs from older fathers and those of younger fathers and mothers may hint at distinct mutational processes, some of which become more significant with increasing paternal age. Our overall DNM spectrum closely resembles that of Rahbari *et al.*[17] (Pearson's $R = 0.98$) (**Supplementary Fig. 5**), who found that the spectrum could be decomposed into the two cancer signatures 1 and 5 (ref. 18). However, the difference we observed between maternal and paternal spectra did not show similarity to any of the known signatures (Pearson's $R < 0.35$, **Supplementary Fig. 6**). Of note, we find an enrichment of maternal DNMs with motifs of APOBEC-mediated mutagenesis (χ^2 test for enrichment $P = 0.029$), which are known to result from aberrant DNA double-strand break repair[16]. The efficiency of oocyte double-strand break repair is known to decrease in aging women[16], which might result in a higher susceptibility to APOBEC-mediated mutations. Overall, these results suggest that aging may trigger mutagenic processes in male sperm that do not occur in female oocytes.

Confirming previous observations[3,15], we found that a subset of DNMs are spatially clustered on the genome, with mutual proximities below 10 kb (χ^2 test for enrichment $P < 2 \times 10^{-16}$). This affects 662 DNMs in 304 clusters (1.8% of all DNMs). Interestingly, the number of such clusters per individual correlates with the age of both parents (**Fig. 4a**, **Supplementary Tables 27** and **28**; Kruskal–Wallis test $P = 1.37 \times 10^{-4}$ and $P = 2.19 \times 10^{-4}$ for mother and father, respectively). The nucleotide substitution profile of DNMs in clusters differs markedly from nonclustering DNMs (χ^2 test $P < 2 \times 10^{-16}$, **Fig. 4b**), as is most prominently demonstrated by a Ts/Tv of 0.72 for clustered DNMs compared to 2.23 for all nonclustered DNMs, which accords with previous observations[15]. We determined the parent of origin for 53 out of the 304 mutation clusters. In all but one phased cluster, the DNMs came from the same parent, supportive of a single mutational event as the cause (**Supplementary Table 15**). Interestingly, the DNM clusters did not show a paternal bias but were evenly divided between the maternal and paternal alleles (Fisher's exact test, $P = 0.74$; **Supplementary Table 29**). This may indicate that the underlying mechanism of DNM clusters may be the same for fathers and mothers. We did not observe significant differences in nucleotide substitution profiles between DNM clusters from paternal and maternal origin, which is consistent with this hypothesis, but could also be due to a lack of statistical power due to the low number of events. Because many mechanisms that have been proposed to cause clustered DNMs encompass action of endogenous proteins of APOBEC and activation-induced deaminase (AID)[19], we scanned the nucleotides surrounding the clustered and nonclustered DNMs for the specific APOBEC motif and AID motif. Clustered DNMs contained significantly more APOBEC-like mutations (χ^2 test, $P = 1.06 \times 10^{-6}$) and AID-like mutations (χ^2 test, $P = 0.017$), together accounting for about a quarter of all clustered DNMs (**Supplementary Table 30**).

Taken together, our results show that the difference in biology of male and female gametogenesis gives rise to distinct mutational signatures in offspring that diverge with increasing parental age.

URLs. Glu-genetics: https://code.google.com/p/glu-genetics.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. *De novo* mutation calls used in this manuscript are available in dbGaP under accession number phs001055.v1.p1.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank all the clinical, laboratory, information technology, and informatics staff for their support on this research project, especially R. Haridas and R. Smith for Sanger sequencing. We would like to thank D. Aguiar and S. Istrail for helpful discussions on their HapCompass software. We would also like to express our gratitude to the participating individuals and their families. The ITMI was supported by the Inova Health System, a nonprofit healthcare system in Northern Virginia. This work was partly financially supported by grants from the Netherlands Organization for Scientific Research (918-15-667 to J.A.V., 916-14-043 to C.G., and SH-271-13 to C.G. and J.A.V.), the European Research Council (ERC Starting grant DENOVO 281964 to J.A.V.), the German Academic Exchange Service DAAD (postdoctoral grant to A.B.S.), and the German Research Foundation DFG (Postdoc grant to A.B.S.).

AUTHOR CONTRIBUTIONS

J.A.V., C.G., and J.E.N. designed the study. J.M.G., W.S.W.W., and M.P. performed the data analyses. M.P., L.E.L.M.V., and A.H. provided and analyzed preliminary data and assisted in writing the final manuscript. J.G.V., B.D.S., and J.E.N. supervised the data collection, sequencing and writing of the manuscript. D.B., A.B.S., G.G., and J.C.R. assisted in data analyses and interpretation. T.F. assisted in data processing. J.M.G., W.S.W.W., and C.G. drafted the manuscript. All authors contributed to the final version of the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Veltman, J.A. & Brunner, H.G. *De novo* mutations in human genetic disease. *Nat. Rev. Genet.* **13**, 565–575 (2012).
2. Kong, A. *et al.* Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
3. Michaelson, J.J. *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**, 1431–1442 (2012).
4. Campbell, C.D. & Eichler, E.E. Properties and rates of germline mutations in humans. *Trends Genet.* **29**, 575–584 (2013).

5. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
6. Makova, K.D. & Hardison, R.C. The effects of chromatin organization on variation in mutation rates in the genome. *Nat. Rev. Genet.* **16**, 213–223 (2015).
7. Crow, J.F. The origins, patterns and implications of human spontaneous mutation. *Nat. Rev. Genet.* **1**, 40–47 (2000).
8. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
9. Wong, W.S. *et al.* New observations on maternal age effect on germline de novo mutations. *Nat. Commun.* **7**, 10486 (2016).
10. Forster, P. *et al.* Elevated germline mutation rate in teenage fathers. *Proc. R. Soc. Lond. B* **282**, 20142898 (2015).
11. 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).
12. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).
13. Smith, D.I., Zhu, Y., McAvoy, S. & Kuhn, R. Common fragile sites, extremely large genes, neural development and cancer. *Cancer Lett.* **232**, 48–57 (2006).
14. White, S. *et al.* A multi-exon deletion within WWOX is associated with a 46,XY disorder of sex development. *Eur. J. Hum. Genet.* **20**, 348–351 (2012).
15. Francioli, L.C. *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.* **47**, 822–826 (2015).
16. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
17. Rahbari, R. *et al.* Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48**, 126–133 (2016).
18. Titus, S. *et al.* Impairment of BRCA1-related DNA double-strand break repair leads to ovarian aging in mice and humans. *Sci. Transl. Med.* **5**, 172ra21 (2013).
19. Chan, K. & Gordenin, D.A. Clusters of multiple mutations: incidence and molecular mechanisms. *Annu. Rev. Genet.* **49**, 243–267 (2015).

ONLINE METHODS

Patient cohort. All participating families were enrolled in the Inova Translational Medicine Institute's (ITMI) IRB-approved research protocol entitled "Molecular Study of Preterm Birth"' (WIRB #20110624), with informed consent obtained for all participants. The eligibility criteria for families are listed in **Supplementary Table 31**. The clinical information was extracted from electronic medical records (EMR) and self-reported questionnaires. This study includes 832 newborns at the Inova Fairfax Hospital (**Table 1**). For all analyses performed in this manuscript, except for the monozygotic twin comparison and the coverage analysis, we randomly removed one twin from each of the monozygotic twin pairs; we also removed one outlier who has more than twice the mean number of DNMs in this cohort. The final analysis cohort consists of 816 trios. Of these 816 trios, 292 were born preterm (<37 weeks) (**Supplementary Table 1**). Babies in 51 families (76 probands) were conceived with assisted reproductive technology (**Supplementary Table 2**). The external cohort of 656 trios (630 singletons and 13 twin pairs) is a subset of the families enrolled in the First 1,000 Days of Study, conducted the Inova Translational Medicine Institute. The details on cohort demographics and sequencing information was previously described in ref. 20.

Whole-genome sequencing (WGS). Whole blood samples were collected from all 2,394 subjects as previously described[9,21]. Genome sequences were assembled with Complete Genomics' Assembly Pipeline versions 2.0.0–2.0.3 using the NCBI build 37 human genome reference assembly[22]. All samples passed internal Complete Genomics quality control parameters. Coverage statistics were calculated using weight-sum sequence coverage depth. On average across all genomes, 70% of each genome and 80% of the coding regions had >40× coverage (**Supplementary Table 3**). WGS of the external cohort of 656 trios was performed by Illumina Services as described in Bodian *et al.* (2015)[20].

Quality control. For all samples we gathered the Complete Genomics summary files that are intended for quality control of the samples. We performed a principal-component analysis (PCA) on statistics from all samples (**Supplementary Fig. 7**). The major differentiating factor between samples was the number of identified single nucleotide variants. We found that the differences between samples for this statistic could be attributed to the ancestry admixture of individuals and the date of sequencing, which corresponded to the software version that was used for analysis. The latter component was, however, minor compared to the impact of admixture. We used 1000 genomes[23] phase 1 genotype calls in regions that qualify strict mask for high quality as the reference panel for admixture calculation. The samples were assigned admixture proportions of the 4 super populations (European, African, East Asian and Americas) using glu-genetics.

Callable genome fraction. Although WGS was performed, not all positions in the genome could be interrogated and this fraction varied considerably per trio. Therefore we calculated the callable fraction on a per trio basis by taking the union of all regions that were not called according to the Complete Genomics var-file based on all autosomes of the GRCh37 genome, excluding positions with "N" sequences (**Supplementary Table 4**). On average we found that across all chromosomes the minimal callable fraction was 88.94% and that on average 94.59% of the genome was called.

Identification of de novo mutations. The initial set of candidate DNMs was called as described previously[24]. Briefly, *de novo* variant calls for the autosomes were generated using the cgatools calldiff program (Complete Genomics). Calldiff compares two genomes and determines whether a variant is truly found in only the child's genome by gathering variants into superloci and performing a local refinement of variant calls, explicitly assuming a diploid genome. For this reason, only DNMs on autosomes were included in this study. The individual comparisons to each parent were merged and filtered for variants with "varQuality = PASS". Only high confidence *de novo* calls were extracted by selecting variants with the scores for both parents greater than or equal to 5. This resulted in a total of 55,049 DNMs and an average of 66 *de novo* calls per individual.

Filtering of DNMs. In order to obtain only the most reliable set of *de novo* variants, we developed a random forest classifier. An elaborate training set of more than 4,000 true positive and false positive single nucleotide DNM calls was established consisting of putative *de novo* mutations validated by three different methods (**Supplementary Table 5**). We randomly selected 90% of our data as a training set, and the remaining 10% as a validation set. Feature selection showed highest contribution to correct classification for 7 features[25] (**Supplementary Table 6**). The out of bag estimate of error rate on the training data was 3.55%. The error rate on the remaining 10% test data was 2.93%. All called *de novo* mutations were then intersected with the callable genome fraction as well as regions that occur >4 times within the genome based on Duke 35bp uniqueness values[26] (**Supplementary Table 7**).

To further validate the results of our classification, we compared *de novo* calls between 15 monozygotic twins, which should theoretically be fully concordant, and 35 dizygotic twins, which should be fully discordant. Comparison of 15 monozygotic twins showed a low error rate of less than 10% (**Supplementary Table 9**). As expected, concordance between 35 dizygotic twins was as low as 0.41%, of which some may be attributable to low-level mosaicism of one of the parents[27] (**Supplementary Table 10**). In total, our algorithm classified 36,441 variants as *de novo* mutations (66.56%), yielding an average of 44 (95% CI: 43.1, 44.6) mutations per individual (**Supplementary Table 8**). 35,793 of those variants are single nucleotide variants. For the further analyses, we randomly removed one twin out of all monozygotic twin pairs, resulting in a cohort of 816 trios.

While we attempted to filter out post-zygotic mutations by including the fraction of reads with the variant allele in our random forest classifier, there are inevitably a small number of post-zygotic mutations included in the final set. It was previously estimated that the proportion of post-zygotic mutations is around 6.5%[27]. Based on the discordance rate of the 15 monozygotic twins, the proportion of mosaicism in our final set is at most 10%. The true proportion is likely to be much lower, as allelic ratios for the filtered DNMs are closer to 0.5 compared to known heterozygous SNPs, while allelic ratios of the unfiltered DNMs have a wider range (**Supplementary Fig. 8**). Furthermore, post-zygotic mutations will affect both paternal and maternal chromosomes and thereby is unlikely to introduce any biases for our analyses.

Experimental validation. In order to access the accuracy of our DNM calls, we randomly selected subsets of DNMs from each variant type, namely, single nucleotide substitutions (SNVs), clustered single nucleotide substitutions (clustered SNVs), InDels (small insertions and deletions) and block substitutions (**Supplementary Table 11**) and Sanger sequenced the proband and both the parents. Sanger sequencing succeeded for 92 variants (43 SNVs, 16 Indels/Block Substitutions, and 33 clustered SNVs). Among these, the clustered SNVs had the highest validation rate of 93.9%, whereas SNVs and InDels/block substitutions achieved validation rates of 88.4% and 87.5% respectively.

Phasing of *de novo* mutations. In order to identify the parental origin of the DNM allele, we applied a haplotype assembly strategy: considering that the human autosomes are diploid and that a DNM affects only one of the two alleles, we attempted a reconstruction of the two distinct alleles. The reconstruction is based on the inherited variants close to the DNM. Some of the variants can only be inherited by one of the parents (informative SNPs). We applied the HapCompass algorithm[27] to assemble the haplotypes of the region 1 kb on both sides around the DNM. If an assembled allele carried an informative SNP, we could advise a parental origin to the DNM. Only single nucleotide substitution variants were phased.

To assess the correctness, we compared the HapCompass results of six independent trios sequenced by CGI (not part of this cohort) to results from Long Fragment Read (LFR) Sequencing[28]. In total, 53 variants were successfully phased for which LFR data also provided phasing information. For 51 of these (96%) both technologies were concordant for the parent of origin (**Supplementary Table 12**). In our cohort, we obtained phasing information for 7,216 DNMs of the total 35,793 (20.16%). This percentage is comparable to the percentages of phased mutations in other peer-reviewed studies, although slightly lower due to the limited sequencing read size (**Supplementary Table 13b**). In total 5,640 variants were paternal in origin (78.16%) and 1,576 were maternal in origin (21.84%) (**Supplementary Table 13a**). We discerned no major differences between phased and unphased DNMs (**Supplementary Figs. 9** and **10**). Further, we compared the results of phasing in 65 trios in our cohort that were sequenced by both CGI and Illumina technologies. We used the DNMs detected using data generated by CGI, and phased using GATK

doi:10.1038/ng.3597

HaplotypeCaller, PhaseByTransmision, and ReadBackedPhasing with Illumina sequenced and assembled reads (**Supplementary Table 32**). Comparison of phasing results showed a 99.75% concordance.

Variant annotation. All 36,441 variants were annotated with SNPEff version 3.4e (ref. 29) and then loaded into the Gemini software Version 0.13.1 (ref. 30). We found 1.67% of variants affected the coding regions (**Supplementary Table 33**). We additionally annotated the variants with custom bed files on CpG positions and GC content based on 200 bp sliding window.

Simulation of DNMs. In order to compare our data to a random set of DNMs, we generated 1,000,000 positions across the callable human genome with the same base frequencies and substitution rate as we observed for the DNMs in the 816 trios. The variant allele was then created by mutating the reference base according to empirical distribution of the substitution rates observed from our filtered single nucleotide DNMs. All simulated variants were annotated by SNPEff-Gemini as described above (**Supplementary Table 34**).

Identification of influence factors on the number of DNMs. We sought to examine the possible factors that are correlated with number of DNMs in each proband. We first performed least-squares multiple linear regression with the total number of DNMs (SNVs only, $N = 35{,}793$) as the response variable, ages of the parents at conception, proband's ethnicity (European, Asian, Americas, African, others), CGI software pipeline version (batch effect), mode of conception (natural versus assisted) and gestation status at delivery (preterm versus full term) as the predictor variables. We then fitted the multiple regression model again with only the predictors that were significant at the 0.05 level, namely, ages of parents at conception and the mode of pregnancy (**Fig. 1a**, **Supplementary Table 16**). Both parents' ages are positively correlated with the total number of DNMs ($P < 10^{-16}$ and $P = 3 \times 10^{-3}$, paternal and maternal respectively).

Next, we examined the linear correlation between unambiguous parent-of-origin resolved number of DNMs and their respective parents' ages. Since only ~20% of the DNMs were phased, we estimated the true number of DNMs of each parental origin by dividing the number of DNMs phased by the proportion phased in each trio. We then fit simple least-squares linear regression models to study the normalized number of phased paternal DNMs with their respective parental ages (**Supplementary Tables 17** and **18**). Paternal age is significantly associated with number of paternal DNMs, with an estimated 0.91 DNMs per year increase in age ($P < 2 \times 10^{-16}$). Mother's age is also significantly associated with number of maternal DNMs, with an estimated 0.24 DNMs per year increase in age ($P = 4.49 \times 10^{-7}$).

Mutation signatures. DNMs were grouped by nucleotide substitution and, where applicable, nucleotide contexts. Statistical significance of group comparisons was assessed in two ways: First, to get an overall indication of whether the mutation distribution depends on the grouping variable, we applied Pearson's χ^2 test for independence. The contingency table that we applied the test to lists the numbers of mutations by group and by mutation type. Second, if there was a significant difference between the two groups, we used a bootstrapping approach to identify the individual mutation types that differed. For this, we re-sampled the grouping variable 10,000-fold and calculated the difference between the relative mutation frequencies of every mutation category. These bootstrapped relative mutation frequencies were then compared to the observed differences between the groups. *P* values give the relative frequencies of bootstrapped differences that were larger than the observed ones. To account for multiple testing, we applied Bonferroni-correction by dividing the obtained *P* values by the number of possible mutation categories (six in the case of substitutions, 96 in the case of substitutions and surrounding nucleotides).

Hierarchical clustering was performed using the hclust function from the "stats" package of the R statistical software[31], using complete linkage method and Euclidean distances. Mutations were grouped by age of the parent into groups of approximately 500 mutations. This resulted in 3 groups of maternal mutations, labeled "y" (young), "m"(middle), "o" (old), corresponding to mothers of younger, intermediate and older ages. For paternal mutations, this resulted in 11 groups, labeled "yyy", "yy", "y", "ym", "ymm", "m", "omm", "om", "o", "oo", "ooo" to indicate the ages of the fathers (in ascending order). The age ranges of each group are given in **Supplementary Table 26**. To assess the validity of the calculated clusters, we used the R package pvclust to calculate so-called AU *P* values[32]. For comparing paternal and maternal age correlation with the incidence of mutation categories, we transformed Spearman correlation coefficients by Fisher's *z* transformation to a normal distribution. From these, one-sided *P* values were calculated.

Identification of DNM clusters. For each DNM, the distance to its closest neighbor on the same chromosome of the same individual was calculated. All DNM with distances below 10 kb were considered as clustered (**Supplementary Table 27**). The nucleotide profiles were compared with a χ^2 test.

Mutation rates along the genome. We divided the human genome (hg19) into nonoverlapping 1-Mb windows. We chose 1 Mb as the window size because it is commonly believed that this captures the natural variation in mutation rate in human genomes[33]. We calculated the number of callable base pairs in each of the window by intersecting the windows with the callable regions using bedtools[34], and discarding those windows with fewer than 50% callable bases. This yielded 2,659 1-Mb windows for analysis. For each window, we calculated the number of unique old and young fathers and mothers who passed on DNMs in the window, and we normalized it by the number of callable bases in the window. The old and young fathers are defined by whether their ages are greater than the median age. The same definition applies to the mothers. This provided multivariate mutation rates matrix consisting of 2,659 rows with 4 columns for old fathers, young fathers, old mothers and young mothers (**Supplementary Table 21**; **Supplementary Fig. 11**).

We obtained the sex-averaged male and female recombination rates from the UCSC table browser[35], the replication timing data from Table S2 from Koren *et al.*[36] and the 1-Mb GC content sliding windows from direct calculation from the hg19 genome using a custom script. The rest of the genomic features, including DNase-hypersensitive sites (DHSs) from various tissues, BisulfiteSeq from ovary and testis, and H3K36me3, H3K4me1, H3K9me3 and H3K27ac from adult ovary, were downloaded from the Roadmap Epigenomics Project[37]. We used bedtools map to calculate the mean value for each 1-Mb window for the values described above. We used the R package FWDselect 2.1.0 to perform feature selection using residual variance criteria with cross validation. Where a feature is available for both ovary and testis, we used the testis track for paternal mutations and the ovary track for maternal mutations. Only DHSs from fetal tissues were used in the multiple linear regressions due to the high correlation between DHSs tracks. We then performed multiple robust linear regressions with selected features for each of the mutation rate category.

To study the mutation rate along the genome, we developed a multivariate Poisson hidden Markov model (PHMM) using the normalized DNM counts in each category as response variables. The model was implemented with the R package depmixS4 Version 1.3-2 (ref. 38) (see **Supplementary Note** for more details). The number of hidden states is determined by fitting models with 2 to 6 states and comparing the Akaike information criterion (**Supplementary Table 20a**). The states are given names according to the parameter estimates. To further confirm that the regions enriched with maternal mutations are not due to random chance, we investigated the same segments in the genomes in a separate cohort of 656 trios sequenced by the Illumina platform. In addition, we compared to the cohorts of other published studies of DNM (**Supplementary Table 35**).

Identification of APOBEC-like and AID-like mutations. Both APOBEC and AID have nucleotide preferences in the mutations that they cause. The APOBEC motif was reported as TCW (mutated nucleotide underlined, W = A or T)[39], while the AID signature is reported as WRCY (Y = C or T)[40,41]. We scanned the nucleotides surrounding the DNMs for the presence of these motifs. χ^2 test for independence was used to assess differences.

Allele ratio comparisons. To assess the allele ratio distribution of the DNMs, we obtained the allele ratios of 100 heterozygous inherited SNPs per trio. The distribution of 83,200 SNP allele ratios, 55,049 unfiltered DNM allele ratios and 35,793 filtered single nucleotide DNM allele ratios are compared

in **Supplementary Figure 8a**,**b**. A Wilcoxon rank-sum test was used to assess the significance of the difference between groups.

Coverage comparisons. We compared the coverage of the DNM sites and 100 randomly selected heterozygous inherited SNP sites in the parents by examining the weighted sum sequence from the coverageRefScore files. The distribution of the coverage of 83,200 random sites in each parent is compared with that of 35,793 single nucleotide DNM sites in each parent (**Supplementary Fig. 8c**).

DNM spectrum comparison to previous studies. In order to compare the spectrum of DNMs in our cohort to the previous knowledge of DNMs, we collected the publicly available DNMs of earlier studies[2–5,17,42]. We chose the studies such that the resulting set matches the set analyzed by Rahbari *et al.*[17]. We obtained the spectra and compared them to our spectrum by calculating Pearson correlations (**Supplementary Fig. 5**).

20. Bodian, D.L. *et al.* Utility of whole-genome sequencing for detection of newborn screening disorders in a population cohort of 1,696 neonates. *Genet. Med.* 221–230 (2016).
21. Bodian, D.L. *et al.* Germline variation in cancer-susceptibility genes in a healthy, ancestrally diverse cohort: implications for individual genome sequencing. *PLoS One* **9**, e94554 (2014).
22. Carnevali, P. *et al.* Computational techniques for human genome resequencing using mated gapped reads. *J. Comput. Biol.* **19**, 279–292 (2012).
23. 1000 Genomes Project Consortium. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
24. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344–347 (2014).
25. Glusman, G., Caballero, J., Mauldin, D.E., Hood, L. & Roach, J.C. Kaviar: an accessible system for testing SNV novelty. *Bioinformatics* **27**, 3216–3217 (2011).
26. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).
27. Acuna-Hidalgo, R. *et al.* Post-zygotic point mutations are an underrecognized source of de novo genomic variation. *Am. J. Hum. Genet.* **97**, 67–74 (2015).
28. Peters, B.A. *et al.* Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487**, 190–195 (2012).
29. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
30. Paila, U., Chapman, B.A., Kirchner, R. & Quinlan, A.R. GEMINI: integrative exploration of genetic variation and genome annotations. *PLoS Comput. Biol.* **9**, e1003153 (2013).
31. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2008).
32. *pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling* (2015).
33. Hellmann, I. *et al.* Why do human diversity levels vary at a megabase scale? *Genome Res.* **15**, 1222–1231 (2005).
34. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
35. Rosenbloom, K.R. *et al.* The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.* **43**, D670–D681 (2015).
36. 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).
37. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
38. Visser, I.S.M. depmixS4: an R package for hidden Markov models. *J. Stat. Softw.* **36**, 1–21 (2010).
39. Roberts, S.A. *et al.* An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* **45**, 970–976 (2013).
40. Pettersen, H.S. *et al.* AID expression in B-cell lymphomas causes accumulation of genomic uracil and a distinct AID mutational signature. *DNA Repair (Amst.)* **25**, 60–71 (2015).
41. Qian, J. *et al.* B cell super-enhancers and regulatory clusters recruit AID tumorigenic activity. *Cell* **159**, 1524–1537 (2014).
42. Conrad, D.F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* **43**, 712–714 (2011).