Large, three-generation CEPH families reveal post-zygotic mosaicism and variability in germline mutation accumulation

Thomas A. Sasani[1], Brent S. Pedersen[1], Ziyue Gao[4], Lisa Baird[1], Molly Przeworski[5,6], Lynn B. Jorde[1,3*], Aaron R. Quinlan[1,2,3*]

1 Department of Human Genetics, University of Utah. Salt Lake City, UT
2 Department of Biomedical Informatics, University of Utah. Salt Lake City, UT
3 USTAR Center for Genetic Discovery, University of Utah. Salt Lake City, UT
4 Howard Hughes Medical Institute & Department of Genetics, Stanford University, Stanford, CA
5 Department of Biological Sciences, Columbia University, New York City, NY
6 Department of Systems Biology, Columbia University, New York City, NY
* to whom correspondence should be addressed

Keywords: De novo mutation, germline mutation, germline mosaicism, postzygotic mutation, gonosomal mutation, human pedigrees

Abstract

The number of de novo mutations (DNMs) found in an offspring's genome is known to increase with both paternal and maternal age. But does the rate of mutation accumulation in parental gametes differ across families? To answer this question, we analyzed DNMs in 33 large, three-generation families collected in Utah by the Centre d'Etude du Polymorphisme Humain (CEPH) consortium. We observed significant variability in parental age effects on DNM counts across families, ranging from 0.24 to 3.33 additional DNMs per year. Using up to 14 grandchildren in these families, we find that 3% of DNMs originated following primordial germ cell specification (PGCS) in a parent, and differ from non-mosaic germline DNMs in their mutational spectra. We also identify a median of 3 gonosomal mutations per sample in the F1 generation, which, along with post-PGCS DNMs, occur at equivalent frequencies on the paternal and maternal haplotypes. These results demonstrate that the rate of germline mutation accumulation varies among families with similar ancestry, and confirm that parental mosaicism is a substantial source of de novo mutations in children.

Data and code availability. Code used for statistical analysis and figure generation has been deposited on GitHub as a collection of annotated Jupyter Notebooks: https://github.com/quinlan-lab/ceph-dnm-manuscript. Data files containing germline de novo mutations, as well as the gonosomal and post-primordial germ cell specification (PGCS) mosaic mutations, are included with these Notebooks. To mitigate compatibility/version issues, we have also made all notebooks available in a Binder environment, accessible at the above GitHub repository.

Introduction

In a 1996 lecture at the National Academy of Sciences, James Crow noted that "without mutation, evolution would be impossible." [1] His remark highlights the importance of understanding the rate at which germline mutations occur, the mechanisms that generate them, and the effects of gamete-of-origin and parental age. Not surprisingly, continued investigation into the germline mutation rate has helped to illuminate the timing and complexity of human evolution and demography, as well as the key role of spontaneous mutation in human disease [2–7].

Some of the first careful investigations of human mutation rates can be attributed to J.B.S. Haldane and others, who cleverly leveraged an understanding of mutation-selection balance to estimate rates of mutation at individual disease-associated loci [8,9]. Over half of a century later, phylogenetic analyses inferred mutation rates from the observed sequence divergence between humans and related primate species at a small number of loci [10–12]. In the last decade, whole genome sequencing of pedigrees has enabled direct estimates of the human germline mutation rate by identifying mutations present in offspring yet absent from their parents (de novo mutations, DNMs) [2,13–18]. Numerous studies have employed this approach to analyze the mutation rate in cohorts of small, nuclear families, producing estimates nearly two-fold lower than those from phylogenetic comparison [2,11,14–17,19].

These studies have demonstrated that the number of DNMs increases with both maternal and paternal ages; such age effects can likely be attributed to a number of factors, including the increased mitotic divisions in sperm cells following puberty, an accumulation of damage-associated mutation, and substantial epigenetic reprogramming undergone by germ cells [14–16,20–22]. There is also evidence that the mutational spectra of de novo mutations differ in the male and female germlines [14,15,18,21]. Furthermore, a recent study of three two-generation pedigrees, each with 4 or 5 children, indicated that paternal age effects may differ across families [20]. However, two-generation families with few offspring provide limited power to quantify parental age effects on mutation rates, and restrict the ability to assign a gamete-of-origin to ~20-30% of DNMs [14,15,20].

Here, we investigate germline mutation among families with large numbers of offspring spanning many years of parental age. We describe de novo mutation dynamics across multiple conceptions using blood-derived DNA samples from large, three-generation families from Utah, which were collected as part of the Centre d'Etude du Polymorphisme Humain (CEPH) consortium [23]. The CEPH/Utah families have played a central role in our understanding of human genetic variation [24,25] by guiding the construction of reference linkage maps for the Human Genome Project [26], defining haplotypes in the International HapMap Project [27], and characterizing genome-wide variation in the 1000 Genomes Project [25].

The CEPH/Utah pedigrees are uniquely powerful for the study of germline mutation dynamics in that they have considerably more (min = 4, max = 16, median = 8) offspring than those used in many prior estimates of the human mutation rate (**Supplementary File 1**). Multiple offspring, whose birth dates span up to 27 years, motivated our investigation of parental age effects on DNM counts within families, and allowed us to ask whether these effects differed across families. The structure of all CEPH/Utah pedigrees **(Supplementary File 1)** also enables the use of haplotype sharing through three generations to determine the parental haplotype of origin for nearly all DNMs in the F1 generation. Using this large dataset of "phased" DNMs, we can investigate the effects of gamete-of-origin on human germline mutation in greater detail.

Finally, if a DNM occurs in the early cell divisions following zygote fertilization (considered gonosomal), or during the proliferation of the primordial germ cells, it may be mosaic in the germline of that individual. This mosaicism can then present as recurrent DNMs in the children of that parent. As DNMs are an important source of genetic disease [6,7,28–31], it is critical to understand the rates of mosaic DNM transmission in families. The structures of the CEPH/Utah pedigrees enable the identification of these recurrent DNMs, and can allow for the differentiation of mutations arising as post-zygotic gonosomal variants or as mosaic in the germline of the F1 generation.

Results

Identifying high-confidence DNMs using transmission to a third generation

We sequenced the genomes of 603 individuals from 33, three-generation CEPH/Utah pedigrees to a genome-wide median depth of ~30X (**Supplementary Figure 1, Supplementary File 1**), and removed 10 samples from further analysis following quality control using peddy [32]. After standard quality filtering we identified a total of 5,064 de novo mutations in 70 F1 individuals, each of which was transmitted to at least one offspring in the F2 generation (**Figure 1a, Supplementary File 2**). Approximately 92% (4,638/5,064) of F1 DNMs were single nucleotide variants (SNVs), and the remainder were small (<= 10 bp) insertion/deletion variants. The eight parents of four F1 samples were re-sequenced to a median depth of ~60X (**Supplementary Fig. 1d**), allowing us to estimate a false positive rate of 4.6% for our de novo mutation detection strategy. Taking all F1 samples together, we calculated median germline mutation rates of 1.20 x 10^{-8} and 1.12 x 10^{-9} per base pair per generation for SNVs and indels, respectively, which corroborate prior estimates based on family genome sequencing [14,16,20,33]. Extrapolating to a diploid genome size of ~6.4 Gbp, we therefore estimate an average number of 76.5 de novo SNVs and 7.2 de novo indels per genome, at average paternal and maternal ages of 29.1 and 26.0 years, respectively.

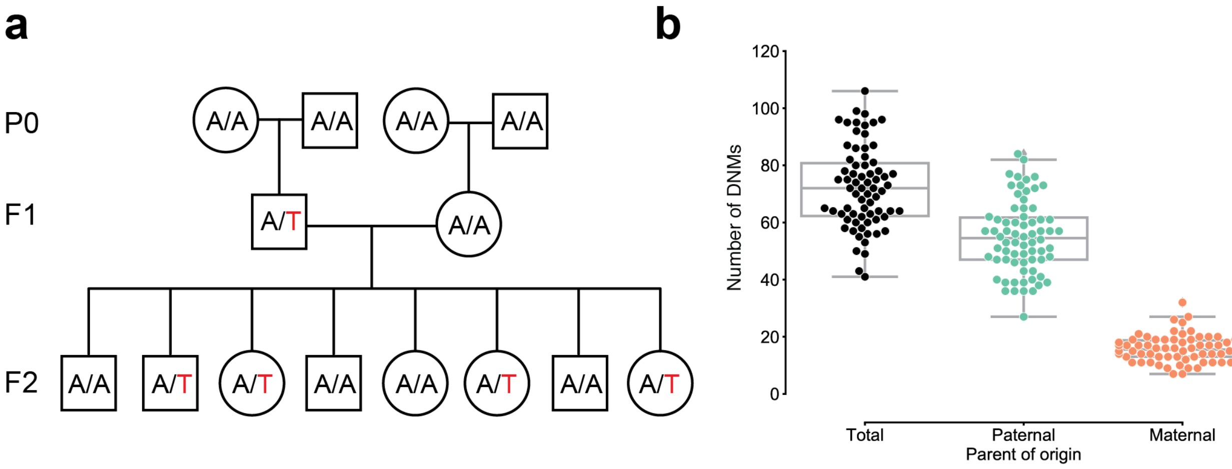

Figure 1. Estimating the rate of germline mutation using multigenerational CEPH/Utah pedigrees. (**a**) The CEPH/Utah dataset comprises 33 multi-generation families, each with a P0, F1, and F2 generation. After identifying candidate de novo mutations present in the F1 generation (e.g., the de novo "T" mutation shown in the F1 father), it is possible to assess their validity both by their absence in the parental (P0) generation and by transmission to one or more offspring in the F2 generation. (**b**) Total numbers of DNMs (both SNVs and indels) identified across F1 CEPH/Utah individuals and stratified by parent-of-origin.

Parent-of-origin and parental age effects on de novo mutation in F1 children

We determined the parent-of-origin for a median of 98.3% of de novo variants per F1 individual (range: 92-100%) by leveraging haplotype sharing across all three generations in a family[14,16], as well as read tracing of DNMs to informative sites in the parents (**Fig. 1b, Supplementary Fig. 2**). The ratio of paternal to maternal DNMs was 3.47:1, and 77.6% of all F1 DNMs were paternal in origin. We then measured the relationship between the number of phased DNMs in each F1 child and the ages of the child's parents at conception (**Fig. 2a**).

After fitting Poisson regressions, we observed a significant paternal age effect of 1.57 (95% CI: 1.24-1.91, p < 2e-16) additional DNMs per year, and a significant maternal age effect of 0.48 (95% CI: 0.29-0.66, p = 5.4e-7) DNMs per year (**Fig. 2a**). These confirm prior estimates of the paternal and maternal age effects on de novo mutation accumulation, and further suggest that both older mothers and fathers contribute to increased DNM counts in children (**Supplementary Fig. 3**) [14–17,20,34].

We next compared the paternal and maternal fractions of phased autosomal F1 DNMs in eight mutational classes (**Fig. 2b**). In maternal mutations, there was an enrichment of C>T transitions in a non-CpG context (unadjusted p = 8.94e-6, Chi-squared test of independence), as well as a significant enrichment of small indels (unadjusted p = 5.57e-3, Chi-squared test of independence). We also observed an enrichment of T>G transversions in paternal mutations (unadjusted p = 6.45e-3, Chi-squared test of independence). Maternal and paternal enrichments of C>T and T>G, respectively, have been reported in recent studies of de novo mutation spectra, though the mechanisms underlying these observations are currently unclear [14,15]. We additionally stratified F1 children by the ages of their parents at conception and found no significant differences in the mutational spectra of children born to older or younger parents (**Supplementary Fig. 4**).

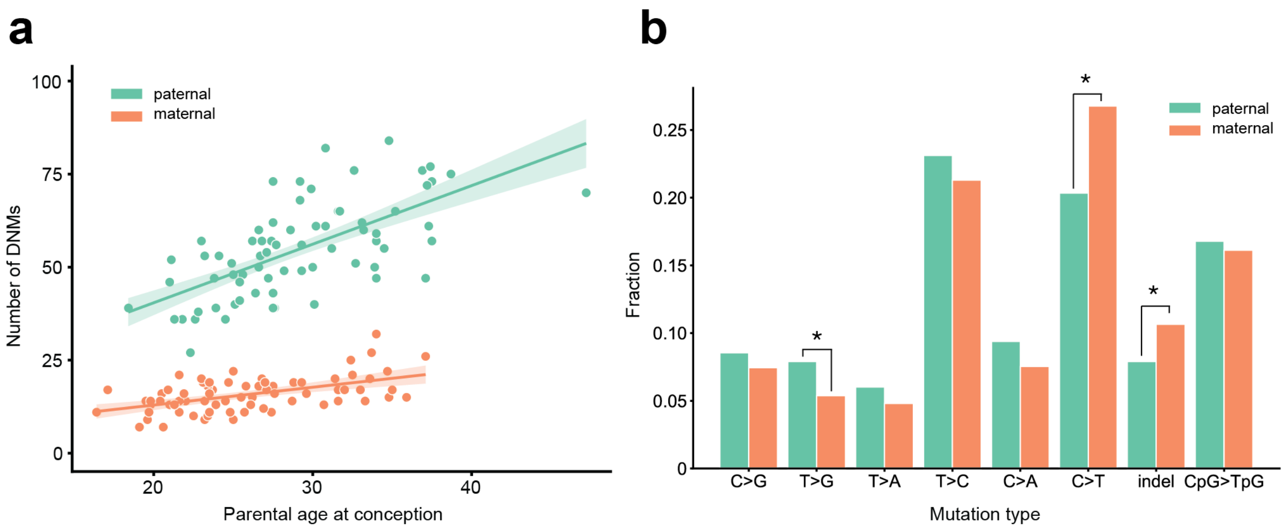

Figure 2. Effects of parental age and sex on autosomal DNM counts and mutation types in the F1 generation.

(**a**) Numbers of phased paternal and maternal de novo variants as a function of parental age at conception. Poisson regressions (with 95% confidence intervals) were fit for mothers and fathers separately using an identity link. (**b**) Mutation spectra in autosomal DNMs phased to the paternal (n=3,771) and maternal (n=1,061) haplotypes. Asterisks indicate significant differences between paternal and maternal fractions at a false-discovery rate of 0.05 (Benjamini-Hochberg procedure), using a Chi-squared test of independence. Unadjusted p-values for each comparison are: C>G: 0.281, T>G: 6.45e-3, T>A: 0.154, T>C: 0.226, C>A: 7.21e-2, C>T: 8.94e-6, indel: 5.57e-3, CpG>TpG: 0.638.

Evidence for inter-family variability of parental age effects on offspring DNM counts

A recent study of three two-generation pedigrees with multiple offspring suggested that the effect of paternal age on DNM counts in children may differ between families [20]. Given the large numbers of offspring in the CEPH/Utah pedigrees, we were motivated to perform a longitudinal investigation of parental age effects on mutation counts within individual families. To measure these effects in the CEPH dataset, we first generated a high-quality set of de novo variants in the F2 generation, excluding recurrent (mosaic) DNMs shared by multiple F2 siblings and "missed heterozygotes" in the F1 generation (1.2% of heterozygous variants). The latter represent apparent DNMs in the F2 generation that were, in fact, likely inherited from an F1 parent who was incorrectly genotyped as being homozygous for the reference allele (**Methods**). In total, we detected 25,346 de novo SNVs and small indels in 350 individuals in the F2 generation (**Supplementary File 3)**. Of these, we were able to confidently determine a parental origin for 5,651 (median of 22% per F2; range of 9-39%) using read tracing, and assign 4,454 (78.8%) of these to the paternal haplotype. Given the comparatively low phasing rate in the F2 generation, we focused our age effect analysis on the relationship between paternal age and the total number of autosomal DNMs in each individual, regardless of parent-of-origin. Taking all F2 individuals into account, we estimate the slope of the paternal age effect to be 1.74 DNMs per year (95% CI: 1.61-1.88, p < 2e-16). Within a given family, maternal and paternal ages are perfectly correlated; therefore, the paternal effect approximates the combined age effects of both parents.

When inspecting each family separately, we observed a wide range of paternal age effects among the CEPH/Utah families (**Fig. 3**). To test whether these observed effects varied significantly between families, we fit a Poisson regression that incorporated the effects of paternal age, family membership, and an interaction between paternal age and family membership, across all F2 individuals in CEPH/Utah pedigrees. As a small number of the CEPH/Utah pedigrees comprise multiple three-generation families (**Supplementary File 1**), we assigned each unique set of F1 parents and their F2 children a distinct ID, resulting in a total of 40 families (**Supplementary Fig. 5**). Overall, the effect of paternal age on offspring DNM counts varied widely across pedigrees, from only 0.24 (95% CI: -1.0-1.49) to nearly 3.33

(95% CI: 2.32-4.34) additional DNMs per year. A goodness-of-fit test supported the use of a "family-aware" regression model when compared to a model that ignores family membership, even after accounting for variable sequencing coverage across F2 samples (ANOVA: p = 3.88e-10). Inter-family differences involve both the initial number of mutations (i.e., the intercept of each family's regression) and the magnitude of the paternal age effect (i.e., the slope of each regression), as we found that family membership, as well as the interaction between paternal age and family membership, significantly improved the fit of the linear model (**Supplementary Table 1**). Finally, when compared to a multiple regression that includes the effects of both paternal and maternal age, a model that takes family membership into account remained a significantly better fit (ANOVA: p = 1.23e-5). The high degree of correlation between paternal and maternal ages makes it difficult to tease out the individual contributions of each parent to the observed inter-family differences. Nonetheless, these results suggest the existence of significant variability in parental age effects across CEPH/Utah families, which could involve both genetic and environmental factors that differ among families.

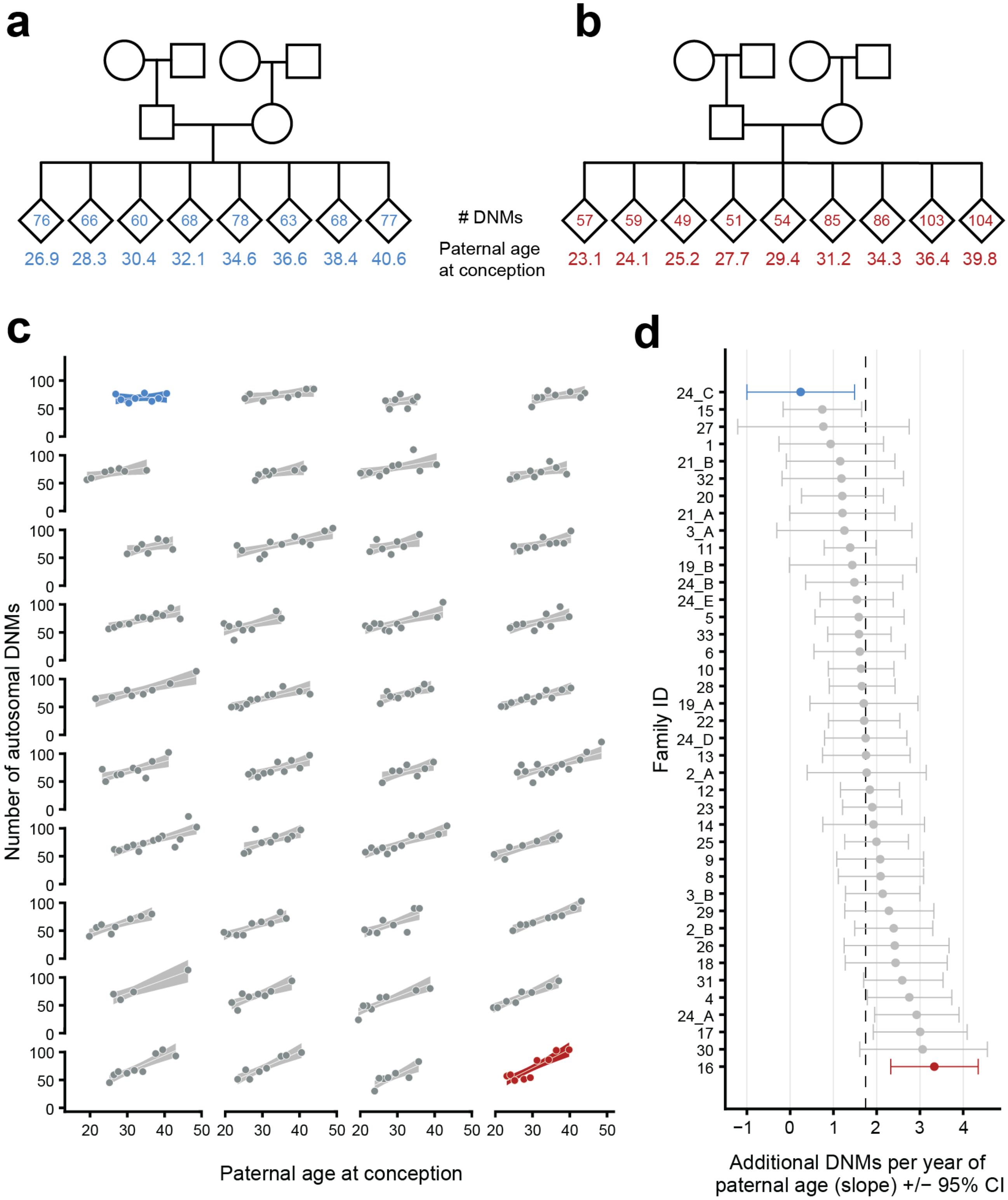

Figure 3: Parental age effects on autosomal germline mutation counts vary significantly among CEPH/Utah families

Illustrations of pedigrees exhibiting the smallest (family 24_C, panel **a**) and largest (family 16, panel **b**) paternal age effects on F2 DNMs demonstrate the extremes of inter-family variability. Given the perfect correlation of paternal and maternal ages in each family, paternal age models the overall contribution from both parents. Diamonds are used to anonymize the sex of each F2 individual. F2 individuals are arranged by birth order from left to right. The number of autosomal DNMs observed in each F2 is shown within each F2 diamond, and the age of the father at the F2's conception is shown below the diamond. The coloring for these two families is used to identify them in panels **c** and **d**. (**c**) The total number of autosomal DNMs is plotted versus paternal age at conception for F2 individuals from all CEPH/Utah families. Regression lines and 95% confidence bands indicate the predicted number of DNMs as a function of paternal age using a Poisson regression. Families are sorted in order of increasing slope, and families with the least and greatest paternal age effects are highlighted in blue and red, respectively. (**d**) A Poisson regression (predicting autosomal DNMs as a function of paternal age) was fit to each family separately; the slope of each family's regression is plotted, as well as the 95% confidence interval of the regression coefficient estimate. The same two families are highlighted as in **(a)**. A dashed black line indicates the overall paternal age effect (estimated using all F2 samples). Families are ordered from top to bottom in order of increasing slope, as in **(a)**.

Identifying gonadal, post-primordial germ cell specification (PGCS) mosaicism in the F1 generation

Generally, studies of de novo mutation focus on variants that arise in a single parental gamete. However, if a de novo variant arises during or after primordial germ cell specification (PGCS), that variant may be present in multiple resulting gametes and absent from somatic cells [20,28–30,35–37]. These post-PGCS variants can therefore be present in more than one offspring as apparent de novo mutations. Given the large number of F2 siblings in each CEPH/Utah family, we had substantially higher power to detect post-PGCS mosaicism in the F1 generation than in prior studies. In each family, we searched for post-PGCS mosaic variants by identifying high-confidence DNMs that were shared by 2 or more F2 individuals, and were absent from the blood DNA of any parents or grandparents within the family (**Fig. 4a**). In total, we identified 721 single-nucleotide post-PGCS mutations at a total of 303 unique sites, which were subsequently corroborated through visual inspection using the Integrative Genomics Viewer (IGV) (**Supplementary File 4**) [38]. Of the phased post-PGCS mutations, 114/249 (45.9%) occurred on a paternal haplotype, consistent with the expectation that these mutations occurred early following PGCS in the F1 generation, and were independent of the parental (P0) origin of the haplotype. Thus, approximately 3.0% (721/23,681) of all single-nucleotide DNMs observed in the F2 generation likely arose following PGCS in a parent's germline, confirming that these variants comprise a non-negligible fraction of all de novo germline mutations.

The mutation spectrum for non-shared germline de novo variants was significantly different than the spectrum for shared, post-PGCS mosaic variants (**Fig. 4b**). Specifically, we found enrichments of CpG>TpG and C>A mutations in post-PGCS variants when compared to all non-shared F2 germline de novo variants (**Fig. 4b**). An enrichment of CpG>TpG mutations in post-PGCS DNMs, which was also seen in a recent report on mutations shared between siblings [37], is particularly intriguing, as many C>T transitions in a CG dinucleotide context are thought to occur due to spontaneous deamination of methylated cytosine [39]. Indeed, DNA methylation patterns are highly dynamic during gametogenesis; evidence in mouse demonstrates that the early primordial germ cells are highly methylated, but experience a global loss of methylation during expansion and migration to the genital ridge, followed by a re-establishment of epigenetic marks (at different time points in males and females) [40,41].

We also tabulated the number of each F2 individual's DNMs that was shared with one or more of their siblings. As reported in the recent analysis of post-PGCS mosaicism [37], we observed that the number of post-PGCS DNMs does not increase with paternal age ($p = 0.77$, **Fig. 4c, Methods**). Thus, a gamete sampled from a younger father is more likely to possess a DNM that will recur in a future child, as early-occurring, potentially mosaic mutations comprise a larger proportion of all DNMs present among the sperm population (**Fig. 4d**). Conversely, a gamete sampled from an older father is more likely to possess a non-mosaic DNM, as the vast majority of DNMs in that father's gametes will have arisen later in life in individual spermatogonial stem cells (**Fig. 4d**) [37,42]. Consistent with this expectation, we observed a significant age-related decrease in the proportion of post-PGCS mosaic DNMs ($p = 2.3e-5$, **Fig. 4e**). Although families with large numbers of siblings are expected to offer greater power to detect shared, post-PGCS DNMs, we verified that the mosaic fraction is not significantly associated with the number of siblings in a family (**Methods**).

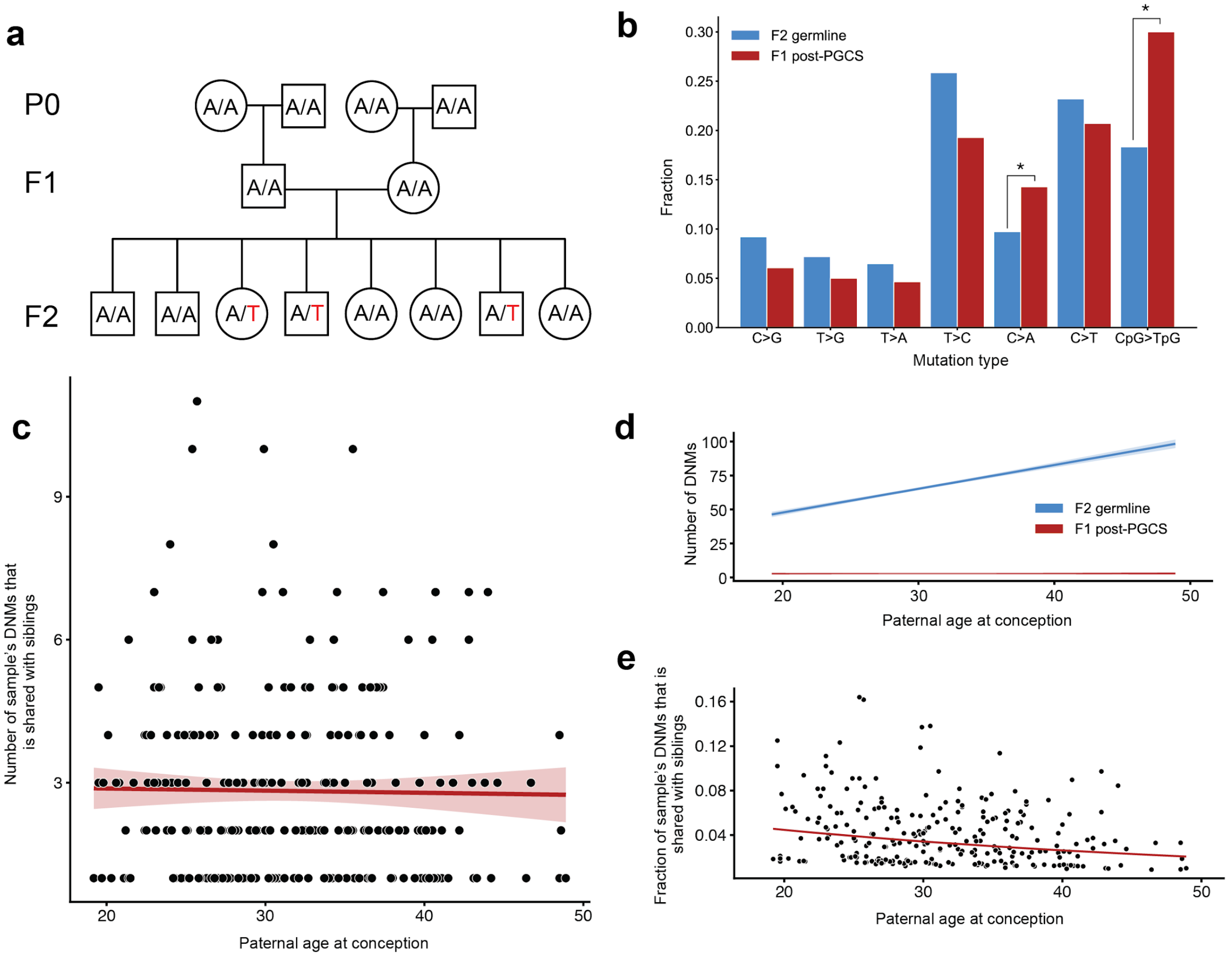

Figure 4. Identification of post-PGCS mosaicism in the F1 generation

(**a**) Mosaic variants occurring after primordial germ cell specification (PGCS) were defined as DNMs present in multiple F2 siblings, and absent from progenitors in the family. (**b**) Comparison of mutation spectra in F1 post-PGCS variants (red) and F2 germline de novo variants (non-shared) (blue). Asterisks indicate significant differences at a false-discovery rate of 0.05 (Benjamini-Hochberg procedure), using a Chi-squared test of independence. Unadjusted p-values for each comparison are: C>G: 6.54e-2, T>G: 0.154, T>A: 0.316, T>C: 3.02e-2, C>A: 8.62e-3, C>T: 0.298, CpG>TpG: 2.10e-6. (**c**) For each F2 individual, we calculated the number of their DNMs that was shared with at least one F2 sibling, and plotted this number against the F2 individual's paternal age at conception. The red line shows a Poisson regression predicting the mosaic number as a function of paternal age at conception. (**d**) We fit a Poisson regression predicting the total number of germline single-nucleotide DNMs in the F2 individuals as a function of paternal age at conception, and plotted the regression line (with 95% CI) in blue. In red, we plotted the line of best fit (with 95% CI) produced by the regression detailed in **(c)**. (**e**) For each F2 individual, we divided the number of their DNMs that occurred post-PGCS in a parent (i.e., that were shared with a sibling) by their total number of DNMs (germline + post-PGCS), and plotted this fraction of post-PGCS DNMs against their paternal age at conception.

Identifying gonosomal mosaicism in the F1 generation

We further distinguished post-PGCS mosaicism from mutations that occurred before primordial germ cell specification, but likely following the fertilization of F1 zygotes. De novo mutations that occur prior to PGCS can be present in both blood and germ cells; we therefore sought to characterize these "gonosomal" variants that likely occurred early during the early post-zygotic development of F1 individuals [20,28,29,37,42,43]. We assumed that gonosomal mutations would be genotyped as heterozygous in an F1 individual, but be observed in a small fraction (less than 20%) of that individual's sequenced reads (**Fig. 5a**). Additionally, if these variants occurred early in development, and were present in both the blood and germ cells, we could validate them by identifying F2 individuals that inherited the variants with a balanced number of reads supporting the reference and alternate alleles (**Fig. 5a**).

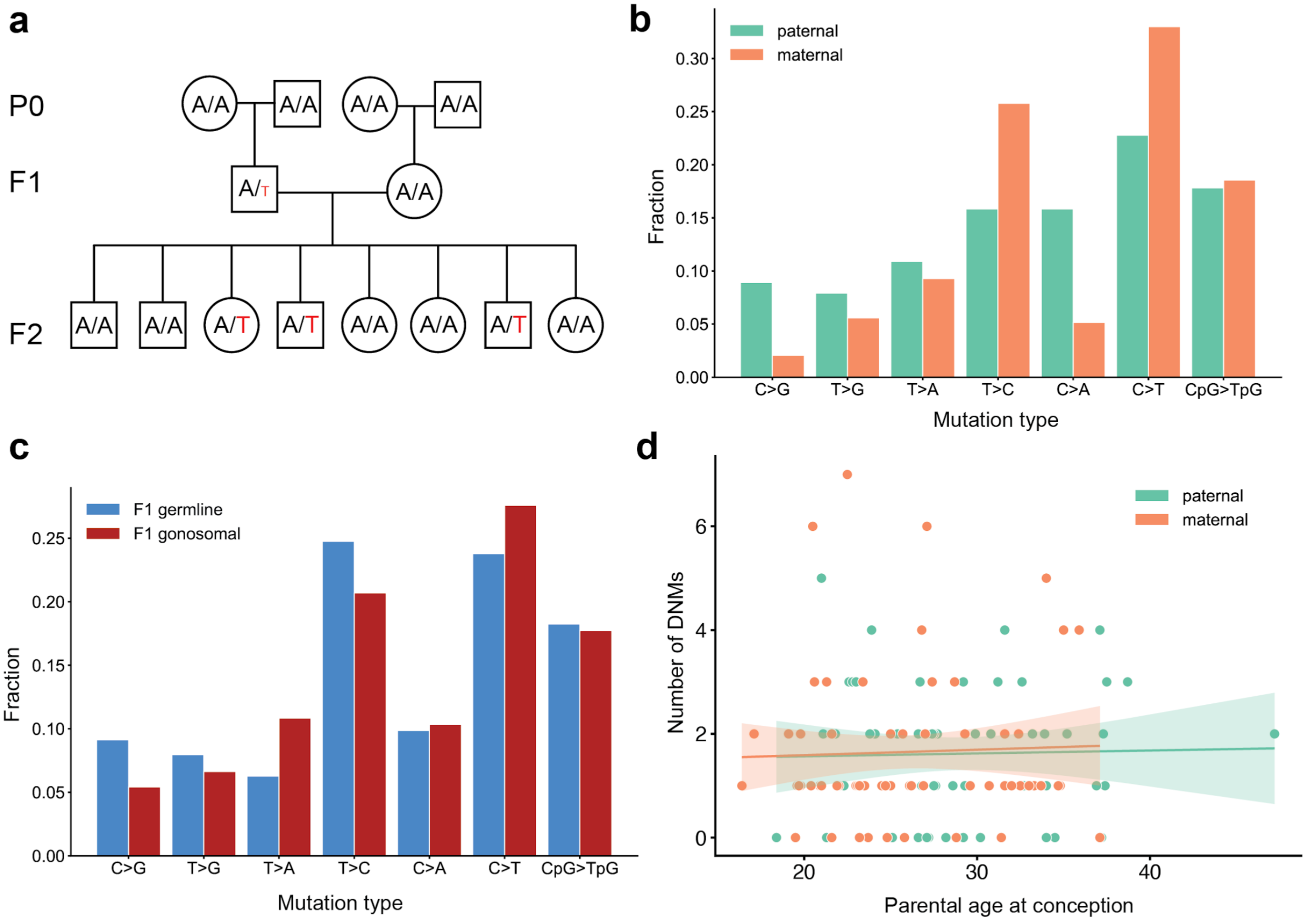

Figure 5. Identification of gonosomal mutations in the F1 generation

(a) Gonosomal post-zygotic variants were identified as DNMs present at low allele balance (i.e., the proportion of aligned sequences supporting the de novo allele; indicated by the smaller "T" allele in the F1 father) in an F1 individual, but inherited with approximately equal allele balance by one or more F2s. **(b)** Comparison of mutation spectra in paternal (n = 100) and maternal (n = 100) autosomal gonosomal variants. Unadjusted p-values for each comparison are: C>G: 6.27e-2, T>G: 0.766, T>A: 1.0, T>C: 0.161, C>A: 4.20e-2, C>T: 0.156, CpG>TpG: 1.0. **(c)** Comparison of mutation spectra in autosomal F1 germline DNMs (non-gonosomal) and putative gonosomal mutations in the F1 generation. Unadjusted p-values for each comparison are: C>G: 0.130, T>G: 0.457, T>A: 8.79e-3, T>C: 0.168, C>A: 0.809, C>T: 0.189, CpG>TpG: 0.967. **(d)** Numbers of phased gonosomal variants as a function of parental age at conception. Poisson regressions (with 95% confidence intervals) were fit for mothers and fathers separately using an identity link.

In total, we identified 207 putative autosomal gonosomal DNMs, which were also validated by visual inspection (**Supplementary File 5**). In contrast to germline F1 DNMs, gonosomal mutations appeared to be sex-balanced with respect to the parental haplotype on which they occurred; 50% (100/200) of all phased gonosomal DNMs occurred on a paternal haplotype, as compared to ~78% of germline F1 DNMs. Similarly, no significant enrichment of particular mutation types was observed on either parental haplotype at a false discovery rate of 0.05 (**Fig. 5b)**. We also found that the overall gonosomal mutation spectrum is similar to the spectrum for F1 germline de novo mutations, as seen in a previous analysis that identified putative germline and somatic DNMs based on allele balance levels, though T>A transversions may be enriched in gonosomal DNMs (unadjusted p = 8.79e-3)[43] (**Fig. 5c**). Unlike germline DNMs, there were no significant effects of parental age on gonosomal DNM counts (**Fig. 5d**). However, a recent study found tentative evidence for a maternal age effect on de novo mutations that arise in the early stages of zygote development [21]. Mutations generated during these first few cell divisions are expected to be present at relatively high mosaic levels in the F1 offspring, and would likely be classified as germline, rather than gonosomal, DNMs in this study.

We note that our analysis pipeline may erroneously classify some gonosomal DNMs. True germline de novo mutations may be incorrectly classified as gonosomal mutations if there is a biased under-sampling of the de novo allele in sequencing reads. Our count of gonosomal DNMs may also be an underestimate, since our requirement that the F1 be heterozygous, yet exhibit low alternate allele counts, precludes the detection of post-zygotic mosaic mutations at both high (> 20%) or very low frequency in each F1. Finally, blood cells represent only a fraction of the total somatic cell population, and we cannot rule out the possibility that apparently gonadal mosaicism may, in fact, be present in other somatic cells that were not sampled in this study [30].

Discussion

Using a cohort of large, multi-generational CEPH/Utah families, we have identified a high-confidence set of de novo mutations that are validated by transmission to the following generation. We determined the parent of origin for nearly all of these DNMs in the F1 generation and produced estimates of the maternal and paternal age effects on the number of DNMs in offspring. Then, by comparing parental age effects among pedigrees with large F2 generations whose birth dates span as many as 27 years, we find that families significantly differ with respect to these age effects. Finally, we identify gonosomal and post-PGCS de novo variants which appear to differ from single-gamete germline DNMs with respect to mutational spectra and magnitude of the sex bias.

Understanding family differences in both mutation rates and parental age effects could enable the identification of developmental, genetic, and environmental factors that impact this variability. The fact that there were detectable differences in DNM age effects between families is striking in light of the fact that the CEPH/Utah pedigrees comprise ostensibly healthy individuals, and that at the time of collection they resided within a relatively narrow geographic area [23,44]. We therefore suspect that our results understate the true extent of variability in mutation rates and age effects among families with diverse inherited risk for mutation accumulation, and who experience a wide range of exposures, diets, and other environmental factors. Supporting this hypothesis, a recent report identified substantial differences in mutation spectra across human ancestries, suggesting that genetic modifiers of the mutation rate may exist in humans, as well as possible differences in environmental exposures [45,46]. Candidate mutator alleles have also been described [47], though they are not significantly associated with elevated counts of de novo mutations in a population of Dutch individuals who share relatively homogenous genetic ancestry. One possible explanation (that we are unable to explore) for the range of de novo mutation counts in firstborn children across families (i.e., the intercepts of the regressions in **Fig. 3c**) is variability in the age at which parents enter puberty. In male F1 parents, for example, entering puberty at an older age would result in a smaller window of time between the start of spermatogenesis and the conception of his first child; compared to another male parent of the same age, his sperm will have accumulated fewer mutations by the time of conception. Indeed, prior studies have demonstrated heritable variation in puberty timing, further suggesting that inter-family differences might be driven at least in part by genetic factors that differ between pedigrees [48–50]. We note, however, that replication is unlikely to be the sole source of de novo germline mutations [21]. Additionally, we observe inter-family differences in both initial mutation counts and the magnitude of parental age effects, indicating that delayed puberty timing cannot fully explain the inter-family variability we observe across CEPH/Utah pedigrees.

Our observation of post-PGCS mosaicism has broad implications for the study of human disease and estimates of recurrence risks within families [28,30,31,37,51]. If a de novo mutation is found to underlie a genetic disorder in a child, it is critical to understand the risk of mutation recurrence in future offspring. We estimate that 3% of germline de novo mutations originated as a mosaic in the germ cells of a parent. This result corroborates recent reports [20,37] and demonstrates that a substantial fraction of all germline DNMs may be recurrent within a family. We also find that the mutation spectrum of post-PGCS DNMs is significantly different than the spectrum for single-gamete germline DNMs, raising the intriguing possibility that different mechanisms contribute to de novo mutation accumulation throughout the proliferation of primordial germ cells and later stages of gametogenesis. For instance, the substantial epigenetic reprogramming that occurs following primordial germ cell specification may predispose cells at particular developmental time points to certain classes of de novo mutations, such as C>T transitions at CpG dinucleotide sites.

Recurrent DNMs across siblings can also manifest as a consequence of gonosomal mosaicism in parents [30]. Although it can be difficult to distinguish gonosomal mosaicism from both germline de novo mutation and post-PGCS mosaicism, we have identified a set of putative gonosomal mosaic mutations that differs from germline DNMs. Namely, gonosomal mosaics are sex-balanced with respect to the parental haplotype and do not exhibit any detectable dependence on parental age at conception. Both of these observations are expected if gonosomal mutations spontaneously arise after zygote fertilization, rather than during the process of gametogenesis. Although we find that the mutation spectrum for putative gonosomal mutations is similar to that of the F1 germline de novo mutations, detecting true differences in mutation spectra is limited by the low numbers of post-zygotic mutations reported by both this and prior studies[43,52].

Taken together, these results underscore the power of large, multi-generational pedigrees for the study of germline mutation and yield greater insight into the mutation dynamics that exist among parental age and sex, as well as family of origin. Given that we studied only 33 large pedigrees, it is almost certain that the mutation rate variability we observe is an underestimate of the full range of variability worldwide. We therefore anticipate future studies of multi-generational pedigrees that will help to dissect the relative contributions of genetic background, developmental timing, and myriad environmental factors.

Acknowledgments

We thank all of the Utah individuals who participated in the CEPH consortium. We also thank Ray White, Jean-Marc Lalouel, and Mark Leppert, who were instrumental in the ascertainment of the CEPH/Utah pedigrees. We additionally thank Andrew Farrell for assistance with interpreting DNM calls. Finally, we thank Tim Formosa, Richard Cawthon, Amelia Wallace and many other members of the Quinlan and Jorde laboratories for insightful discussion related to the manuscript.

Author Contributions

L.B.J. and A.R.Q. designed the experiment and organized the study. T.A.S. led all research, methodology development, and data analysis. B.P.S., Z.G., L.B., M.P., A.R.Q., and L.B.J. contributed to methodologies used and the analyses conducted. T.A.S. and A.R.Q. wrote the manuscript.

Funding

A.R.Q. was supported by the US National Institutes of Health through grants from the National Human Genome Research Institute (R01HG006693 and R01HG009141), and the National Institute of General Medical Sciences (R01GM124355). L.B.J. was supported by the US National Institutes of Health through a grant from the National Institute of General Medical Sciences (R35GM118335). M.P. was supported by the US National Institutes of Health through a grant from the National Institute of General Medical Sciences (R01GM122975). T.A.S. was supported by a T32 award from the National Institute of General Medical Sciences (T32GM007464).

Methods

Genome Sequencing

Whole-genome DNA sequencing libraries were constructed with 500 ng of genomic DNA isolated from blood, utilizing the KAPA HTP Library Prep Kit (KAPA Biosystems, Boston, MA) on the SciClone NGS instrument (Perkin Elmer, Waltham, MA) targeting 350bp inserts. Post-fragmentation (Covaris, Woburn, MA) the genomic DNA was size selected with AMPure

XP beads using a 0.6x/0.8x ratio. The libraries were PCR amplified with KAPA HiFi for 4-6 cycles (KAPA Biosystems, Boston, MA). The final libraries were purified with two 0.7x AMPureXP bead cleanups. The concentration of each library was accurately determined through qPCR (KAPA Biosystems, Boston, MA). Twenty-four libraries were pooled and loaded across four lanes of a HiSeqX flow cell to ensure that the libraries within the pool were equally balanced. The final pool of balanced libraries was loaded over an additional 16 lanes of the Illumina HiSeqX (Illumina, San Diego, CA). 2x150 paired-end sequence data was generated. This efficient pooling scheme targeted ~30X coverage for each sample.

DNA Sequence Alignment

Sequence reads were aligned to the GRCh37 reference genome (including decoy sequences from the GATK resource bundle) using BWA-MEM v0.7.15 [53]. The aligned BAM files produced by BWA-MEM were de-duplicated with samblaster [54]. Realignment for regions containing potential short insertions and deletions and base quality score recalibration was performed using GATK v3.5.0 [55]. Alignment quality metrics were calculated by running samtools stats & flagstats [56] on aligned and polished BAM files.

Variant calling

Single-nucleotide and short insertion/deletion variant calling was performed with GATK v3.5.0 [55] to produce gVCF files for each sample. Sample gVCF files were then jointly genotyped to produce a multi-sample project level VCF file.

Sample quality control and filtering

We used peddy [32] to perform relatedness and sample sequencing quality checks on all CEPH/Utah samples. We discovered a total of 10 samples with excess levels of heterozygosity (ratio of heterozygous to homozygous alternate calls > 0.2). Many of these samples were also listed as being duplicates of other samples in the cohort, indicating possible sample contamination prior to sequencing. We therefore removed all 10 samples with a ratio of heterozygous to homozygous genotypes exceeding 0.2 from further analysis. In total, we were left with 593 P0, F1, and F2 samples with high-quality sequencing data.

Identifying DNM Candidates

We identified high-confidence de novo mutations in the F1 and F2 generations as follows. For each variant, we required that the child possessed a unique genotyped allele absent from both parents; when identifying de novo variants on the X chromosome, we required male offspring genotypes to be homozygous. We required the aligned sequence depth in the child and both parents to be >= 12 reads; allele balance (defined as the proportion of reads supporting the de novo allele) to be >= 0.2 in the child; Phred-scaled genotype quality (GQ) to be >= 20 in the child and both parents, and no reads supporting the de novo allele in either parent. We removed de novo variants within low-complexity regions [19,57], and any variants that were not listed as “PASS” variants by GATK HaplotypeCaller. Finally, we removed DNMs with likely DNM carriers in the cohort; we define carriers as samples that possess the DNM allele, other than the sample with the putative DNM and his/her children. We adapted a previously published strategy [14] to discriminate between “possible carriers” of the DNM allele (samples genotyped as possessing the de novo allele), and “likely carriers” (a subset of “possible carriers” with depth >= 12, allele balance >= 0.2, and Phred-scaled genotype quality >= 20). We removed putative DNMs for which there were any “likely carriers” of the allele in the cohort. We then separated the candidate F1 variants into true and false positives based on transmission to the F2 generation. For each candidate F1 variant, we assessed whether the DNM was inherited by at least one member of the F2 generation; to limit our identification of false positive transmission events, we required F2 individuals with inherited DNMs to have a depth >= 12 reads at the site and Phred-scaled genotype quality >= 20. We defined “transmitted” F1 variants as variants for which the median allele balance across transmissions was >= 0.3. One CEPH/Utah family (family ID 26) contains only 4 sequenced F2 grandchildren (**Supplementary File 1**); therefore, we did not include the two F1 individuals from this family in our analysis of F1 DNMs, as we lacked power to detect high-quality transmission events.

Because we were unable to validate DNMs in the F2 generation by transmission, we applied a more stringent set of quality filters to all F2 DNMs. We required the same filters as applied to all F1 DNMs, but additionally required that the allele balance in each DNM was >= 0.3. We further required that there were no possible carriers of the de novo allele in the rest of the cohort. Finally, for each DNM in the F2 generation, we assessed if any of the F2 individual’s grandparents were genotyped as possessing the DNM allele; if so, we removed that DNM from further analysis (see section entitled “Estimating a missed heterozygote rate”).

Determining the parent of origin for DNMs

To determine the parent of origin for each de novo variant in the F1 generation, we phased mutation alleles by transmission to a third generation, a technique which has been described previously [14–16,20] (**Supplementary Figure 2a**). We searched 200 kbp upstream and downstream of each F1 DNM for informative variants, defined as alleles present as a heterozygote in the F1, observed in only one of the two parents, and observed in at least one of the F2 individuals that inherited the DNM. For each of these informative variants, we asked if the informative variant was always transmitted with the DNM; if so, we could infer that the heterozygous variant was present on the same haplotype as the DNM (assuming recombination did not occur between the DNM and the flanking informative variants), and assign the P0 parent with the informative variant as the parent of origin (**Supplementary Figure 2a**). For each F1 DNM, we identified all transmission patterns (i.e., combinations of a P0, F1, and set of F2s that inherited both the informative variant and the DNM). We only assigned a confident parent-of-origin at sites where the most frequent transmission pattern occurred at >= 75% of all informative sites. Because we identified putative gonosomal DNMs by searching for de novo heterozygous variants in the F1 generation (albeit at lower allele balance than germline DNMs), we also phased these using haplotype sharing across three generations.

We additionally phased de novo variants in the F1 generation, as well as all DNMs in the F2 generation, using "read tracing" (also known as "read-backed phasing") [14,15]. Briefly, for each de novo variant, we first searched for nearby (within 1 read fragment length, 500 bp) variants present in the proband and one of the two parents. Thus, if the de novo variant was present on the same read as the inherited variant, we could infer haplotype sharing, and determine that the de novo event occurred on that parent's chromosome (**Supplementary Figure 2b**). Similarly, if the de novo variant was not present on the same read as the inherited variant, we could infer that the de novo event occurred on the other parent's chromosome.

We were also able to determine the parent-of-origin for many of the post-PGCS mosaic variants by leveraging haplotype sharing across three generations [37]. If all F2 individuals with a post-PGCS DNM shared a haplotype with a particular P0 grandparent, we assigned that P0 grandparent's child (i.e., one of the two F1 parents) as the parent of origin.

In the F1 generation, the read tracing and haplotyping sharing phasing strategies were highly concordant, and the parent-of-origin predictions agreed at 99.4% (1,057/1,063) of all DNMs for which both strategies could be applied.

Calculating the rate of germline mutation

Given the filters we employed to identify high-confidence de novo mutations, we needed to calculate the fraction of the genome that was considered in our analysis. To this end, we used mosdepth [58] to calculate per-base genome coverage in all CEPH/Utah samples, excluding low-complexity regions [57] and reads with mapping quality < 20 (the minimum mapping quality threshold used by GATK HaplotypeCaller in this analysis). For each F1 and F2 child, we then calculated the fraction of all genomic positions that had at least 12 aligned sequence reads in the child's, mother's, and father's genome (excluding the X chromosome). In the F1 generation, the median number of callable autosomal base pairs per sample was 2,582,336,232. For each individual, we then divided their count of autosomal de novo mutations by the resulting number of base pairs, and divided the result by 2 to obtain a diploid human mutation rate per base pair per generation. The median F1 germline SNV mutation rate was calculated to be 1.239 x 10^{-8} per base pair per generation. We then adjusted this mutation rate based on our estimated false positive rate (FPR) and our estimated "missed heterozygote rate" (MHR; see section entitled "Estimating a missed heterozygote rate") as follows:

```
adj_mu = mu * (1 - FPR / 1 - MHR)

adj_mu = 1.239e-8 * (1 - 0.046 / 1 - 0.012)
```

Assessing age effect variability between families

Using the full call set of de novo variants in the F2 generation (excluding the recurrent, post-PGCS DNMs) we first fit a simple Poisson regression model that calculated the effect of paternal age on total autosomal DNM counts in the R statistical language (v3.5.1) as follows:

```
glm(autosomal_dnms ~ dad_age, family=poisson(link="identity"))
```

This model returned a highly significant effect of paternal age on total DNM counts (1.74 DNMs per year of paternal age, p < 2e-16), but was agnostic to the family from which each F2 individual was "sampled." Importantly, a number of F2 individuals in the CEPH/Utah cohort share grandparents, and may therefore be considered members of the same family, despite having unique F1 parents (**Supplementary Figure 5**). For all subsequent analysis, we defined a “family” as the unique group of two F1 parents and their F2 offspring (**Supplementary Figure 5)**. In the CEPH/Utah cohort, there are a total of 40 “families” meeting this definition.

To test for significant variability in paternal age effects between families, we fit the following model:

```
glm(autosomal_dnms ~ dad_age * family_id,
family=poisson(link=”identity”))
```

Which can also be written in an expanded form as:

```
glm(autosomal_dnms ~ dad_age + family_id + dad_age:family_id,
family=poisson(link=”identity”))
```

To assess the significance of each term in the fitted model, we performed an analysis of variance (ANOVA) as follows:

```
m = glm(autosomal_dnms ~ dad_age + family_id + dad_age:family_id,
family=poisson(link=”identity”))

anova(m, test=”Chisq”)
```

The results of this ANOVA are shown in **Supplementary Table 1**. In summary, this model contained the fixed effect of paternal age, as well as different regression intercepts within each “grouping factor” (i.e., family ID). Additionally, this model includes an interaction between paternal age and family ID, allowing for the effect of paternal age (i.e., the slope of the regression) to vary within each grouping factor.

To account for variable sequencing coverage across CEPH/Utah samples, we additionally calculated the callable autosomal fraction for all F2 individuals by summing the total number of nucleotides covered by >= 12 reads in the F2 and both of their F1 parents, excluding low-complexity regions and reads with mapping quality < 20 (see section entitled “Calculating the rate of germline mutation”).

Since we only consider the effect of paternal age on the mutation rate, we can model the mutation rate (`mu`) as:

```
mu = Bp * Ap + B0
```

Where `Bp` is the paternal age effect, `Ap` is the paternal age, and `B0` is an intercept term.

Therefore, the number of DNMs in a sample is assumed to follow a Poisson distribution, with the expected mean of the distribution defined as:

```
E(# DNMs) = mu * callable_fraction

E(# DNMs) = (Bp * Ap + B0) * callable_fraction

E(# DNMs) = (Bp * Ap * callable_fraction) + (callable_fraction * B0)
```

As our analysis only considers the effect of paternal age on total DNM counts, we can thus scale `Ap` (paternal age at conception) by the `callable_fraction`, generating a term called `dad_age_scaled`, and fit the following model, which takes each sample's callable fraction into account:

```
glm(autosomal_dnms ~ dad_age_scaled + autosomal_callable_fraction +
0, family=poisson(link=”identity”))
```

Then, we can determine whether inter-family differences remain significant by comparing the above null model to a model that takes family into account:

```
glm(autosomal_dnms ~ dad_age_scaled * family_id +
autosomal_callable_fraction + 0, family=poisson(link=”identity”))
```

After running an ANOVA to compare the two models, we find that the model incorporating family ID is a significantly better fit (ANOVA: p = 3.88e-10).

We previously identified significant effects of both maternal and paternal age on DNM counts (**Figure 2a**). Therefore, to account for the non-negligible effect of maternal age on DNM counts, we fit a final model that incorporated the effects of both maternal and paternal age, as well as family ID, on total DNM counts as follows:

```
glm(autosomal_dnms ~ dad_age + mom_age + family_id,
family=poisson(link="identity"))
```

We then performed an ANOVA on the model, and found that a model incorporating a family term is a significantly better fit than a model that includes the effects of paternal and maternal age alone (p = 1.23e-5)

Identifying post-PGCS mosaic mutations

To identify post-PGCS mosaic variants, we searched the previously generated callset of single-nucleotide DNMs in the F2 generation ("Identifying DNM candidates") for de novo single-nucleotide mutations that appeared in 2 or more F2 siblings. As a result, all filters applied to the germline F2 DNM callset were also applied to the post-PGCS mosaic variants. We validated all putative post-PGCS mosaic variants by visual inspection using the Integrative Genomics Viewer (IGV) [38]. In a small number of cases (32), we found evidence for the post-PGCS mosaic variant in one of the two F1 parents (**Supplementary File 3**). Reads supporting the post-PGCS mosaic variant were likely filtered from the joint-called CEPH/Utah VCF output following local re-assembly with GATK, though they are clearly present in the raw BAM alignment files. We removed these 32 variants, at which an F1 parent possessed 2 or more reads of support for the mosaic DNM allele in the aligned sequencing reads, and instead included these variants in our callset of "gonosomal" variants (described in section "Identifying gonosomal variants").

Assessing age effects on post-PGCS DNMs

To identify a paternal age effect on the number of post-PGCS DNMs transmitted to F2 children, we tabulated the number of each F2's DNMs that was shared with at least one of their siblings. We then fit a Poisson regression as follows, regressing the number of mosaic DNMs in each F2 against their father's age at conception:

```
glm(mosaic_number ~ dad_age, family=poisson(link=”identity”))
```

We did not find a significant effect of paternal age (p = 0.77).

Using the predicted paternal age effects on germline DNM counts and post-PGCS DNM counts, we determined that the fraction of post-PGCS DNMs should decrease non-linearly with paternal age (**Fig. 4e**). Therefore, to assess the effect of paternal age on the fraction of each F2's DNMs that occurred post-PGCS in a parent, we fit the following model:

```
lm(log(mosaic_fraction) ~ dad_age)
```

We found a significant effect of paternal age on the post-PGCS mosaic fraction (p = 2.3e-5).

As we may be more likely to identify shared, post-PGCS DNMs in families with larger numbers of F2 siblings, we additionally tested whether the fraction of post-PGCS DNMs in each child was dependent on the number of their siblings in the family by performing a correlation test as follows:

```
cor.test(mosaic_fraction, n_siblings)
```

We did not observe a significant correlation between an F2's number of siblings and the fraction of their DNMs that was shared with a sibling (p = 0.966).

Identifying gonosomal mutations

To identify variants that occurred early in post-zygotic development, we identified de novo single-nucleotide variants in the F1 generation using the same genotype quality and population-based filters as described previously (“Identifying DNM candidates”). However, we required that the allele balance in the F1 sample to be greater than 0 and less than 0.2, as we expected gonosomal DNMs to be present at low frequency in the F1 samples' sequenced reads. As with germline F1 DNMs, we assessed whether each gonosomal DNM was inherited by at least one member of the F2 generation; as for germline F1 DNMs, to limit our identification of false positive events, we required F2 individuals with inherited DNMs to have a depth >= 12 reads at the site, Phred-scaled genotype quality (GQ) >= 20, and for the median allele balance across transmissions to be >= 0.3. We validated all putative gonosomal variants by visual inspection using the Integrative Genomics Viewer (IGV) [38]. As described above (section "Identifying post-PGCS mosaic variants"), we also included 32 post-PGCS mosaic variants, that upon manual inspection, exhibited evidence in the F1 generation, in our gonosomal callset (**Supplementary File 3**).

Estimating a "missed heterozygote rate" for DNM detection

Infrequently, variant calling methods such as GATK may incorrectly assign genotypes to samples at particular sites in the genome. When identifying de novo variants, we require that children possess genotyped alleles that are absent from either parent; thus, genotyping errors in parents could lead us to assign variants as being de novo, when in fact one or both parents possessed the variant and transmitted the allele. Given the multi-generational structure of our study cohort, we were able to estimate the rate at which our variant calling and filtering pipeline mis-genotyped an F1 parent as being homozygous for a reference allele. To estimate this "missed heterozygote" rate in our dataset, we looked for any cases in which one or more F2 individuals possessed a putative de novo variant (i.e. possessed an allele absent from both F1 parents). Then, we looked at the sample's grandparental (P0) genotypes for evidence of the same variant. If one or more grandparents was genotyped as having high-quality evidence for the de novo allele (depth >= 12 and Phred-scaled genotype quality >= 20), we inferred that the variant could have been "missed" in the F1 generation, despite being truly inherited. We estimate the missed heterozygote rate (MHR) to be 1.2%, by dividing the total number of F2 DNMs with grandparental support by the total number of F2 DNMs (305/25,651). In a small number of CEPH/Utah pedigrees, some members of the P0 (grandparental) generation were not sequenced (6 grandparents in 5 families, **Supplementary File 1**). As a result, in these families, we are underpowered to detect evidence of F2 DNM alleles in the P0 generation, and our MHR is likely a slight underestimate.

Estimating a false positive rate for de novo mutation detection

A total of 8 P0 grandparents were re-sequenced to a greater genome-wide median depth of 60X (**Supplementary Fig. 1D**). However, when variant calling and joint genotyping was performed on all 603 CEPH/Utah samples, the 30X data for these grandparents was used. Therefore, we sought to estimate a false positive rate for our de novo mutation detection strategy using the de novo mutation calls in the children of these 8 P0 individuals. For each of the children (F1) of these high-coverage P0 individuals, we looked for evidence of the F1 DNMs in the 60X alignments from their parents. Specifically, for each F1 DNM, we counted the number of reads supporting the DNM allele in each of the P0 parents, excluding reads with mapping quality < 20 (the minimum mapping quality imposed by GATK HaplotypeCaller in our analysis), and excluding bases with base qualities < 20 (the minimum base quality imposed by GATK HaplotypeCaller in our analysis). If we observed two or more reads supporting the F1 DNM in a P0 parent's 60X alignments, we considered the F1 DNM to be a false positive. Of the 216 de novo mutations called in the four F1 children of the high-coverage P0 parents, we find 10 mutations with at least two reads of supporting evidence in the 60X P0 alignments. Thus, we estimate our false positive rate for de novo mutation detection to be approximately 4.6% (10/216).

References

1. Crow, J. F. The high spontaneous mutation rate: is it a health risk? *Proc. Natl. Acad. Sci. U. S. A.* **94**, 8380–8386 (1997).
2. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.* **13**, 745–753 (2012).
3. Moorjani, P., Gao, Z. & Przeworski, M. Human Germline Mutation and the Erratic Evolutionary Clock. *PLoS Biol.* **14**, e2000744 (2016).
4. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
5. Yuen, R. K. C. *et al.* Genome-wide characteristics of de novo mutations in autism. *npj Genomic Medicine* **1**, 509 (2016).
6. Acuna-Hidalgo, R., Veltman, J. A. & Hoischen, A. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol.* **17**, 241 (2016).
7. Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nat. Rev. Genet.* **13**, 565–575 (2012).
8. Haldane, J. B. S. The rate of spontaneous mutation of a human gene. *J. Genet.* **31**, 317 (1935).
9. Nachman, M. W. Haldane and the first estimates of the human mutation rate. *J. Genet.* **87**, 317–317 (2008).
10. Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304 (2000).
11. Shendure, J. & Akey, J. M. The origins, determinants, and consequences of human mutations. *Science* **349**, 1478–1483 (2015).
12. Kondrashov, A. S. Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases. *Hum. Mutat.* **21**, 12–27 (2002).
13. 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).
14. Jónsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548 trios from

Iceland. *Nature* **549**, 519–522 (2017).

15. Goldmann, J. M. *et al.* Parent-of-origin-specific signatures of de novo mutations. *Nat. Genet.* **48**, 935–939 (2016).
16. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
17. Roach, J. C. *et al.* Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science* **328**, 636–639 (2010).
18. Francioli, L. C. *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.* 1–8 (2015).
19. Turner, T. N. *et al.* Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell* **171**, 710–722.e12 (2017).
20. Rahbari, R. *et al.* Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48**, 126–133 (2015).
21. Gao, Z., Moorjani, P., Amster, G. & Przeworski, M. Overlooked roles of DNA damage and maternal age in generating human germline mutations.
22. Crow, J. F. The origins, patterns and implications of human spontaneous mutation. *Nat. Rev. Genet.* **1**, 40–47 (2000).
23. Dausset, J. *et al.* Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics* **6**, 575–577 (1990).
24. Prescott, S. M., Lalouel, J. M. & Leppert, M. From linkage maps to quantitative trait loci: the history and science of the Utah genetic reference project. *Annu. Rev. Genomics Hum. Genet.* **9**, 347–358 (2008).
25. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
26. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
27. Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype

structure in the human genome. *Nat. Genet.* **29**, 229–232 (2001).

28. Campbell, I. M. *et al.* Parental somatic mosaicism is underrecognized and influences recurrence risk of genomic disorders. *Am. J. Hum. Genet.* **95**, 173–182 (2014).
29. Campbell, I. M., Shaw, C. A., Stankiewicz, P. & Lupski, J. R. Somatic mosaicism: implications for disease and transmission genetics. *Trends Genet.* **31**, 382–392 (2015).
30. Biesecker, L. G. & Spinner, N. B. A genomic view of mosaicism and human disease. *Nat. Rev. Genet.* **14**, 307–320 (2013).
31. Forsberg, L. A., Gisselsson, D. & Dumanski, J. P. Mosaicism in health and disease - clones picking up speed. *Nat. Rev. Genet.* **18**, 128–142 (2017).
32. Pedersen, B. S. & Quinlan, A. R. Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy. *Am. J. Hum. Genet.* **100**, 406–413 (2017).
33. Besenbacher, S. *et al.* Multi-nucleotide de novo Mutations in Humans. *PLoS Genet.* **12**, e1006315 (2016).
34. Wong, W. S. W. *et al.* New observations on maternal age effect on germline de novo mutations. *Nat. Commun.* **7**, 10486 (2016).
35. 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).
36. Tang, W. W. C., Kobayashi, T., Irie, N., Dietmann, S. & Surani, M. A. Specification and epigenetic programming of the human germ line. *Nat. Rev. Genet.* **17**, 585–600 (2016).
37. Jónsson, H. *et al.* Multiple transmissions of de novo mutations in families. *Nat. Genet.* (2018). doi:10.1038/s41588-018-0259-9
38. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
39. Fryxell, K. J. & Zuckerkandl, E. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol. Biol. Evol.* **17**, 1371–1383 (2000).
40. Seisenberger, S. *et al.* The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Mol. Cell* **48**, 849–862 (2012).

41. Reik, W., Dean, W. & Walter, J. Epigenetic reprogramming in mammalian development. *Science* **293**, 1089–1093 (2001).

42. Campbell, I. M. *et al.* Parent of origin, mosaicism, and recurrence risk: probabilistic modeling explains the broken symmetry of transmission genetics. *Am. J. Hum. Genet.* **95**, 345–359 (2014).

43. Besenbacher, S. *et al.* Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat. Commun.* **6**, 5969 (2015).

44. Malhotra, A., Cromer, K., Leppert, M. F. & Hasstedt, S. J. The power to detect genetic linkage for quantitative traits in the Utah CEPH pedigrees. *J. Hum. Genet.* **50**, 69 (2005).

45. Harris, K. & Pritchard, J. K. Rapid evolution of the human mutation spectrum. *Elife* **6**, 415 (2017).

46. Mathieson, I. & Reich, D. Differences in the rare variant spectrum among human populations. *PLoS Genet.* **13**, e1006581 (2017).

47. Seoighe, C. & Scally, A. Inference of Candidate Germline Mutator Loci in Humans from Genome-Wide Haplotype Data. *PLoS Genet.* **13**, e1006549 (2017).

48. Mostafavi, H. *et al.* Identifying genetic variants that affect viability in large cohorts. *PLoS Biol.* **15**, e2002458 (2017).

49. Pickrell, J. K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).

50. Day, F. R. *et al.* Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nat. Genet.* **49**, 834–841 (2017).

51. Krupp, D. R. *et al.* Exonic Mosaic Mutations Contribute Risk for Autism Spectrum Disorder. *Am. J. Hum. Genet.* **101**, 369–390 (2017).

52. Ju, Y. S. *et al.* Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**, 714–718 (2017).

53. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).

54. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).

55. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

56. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

57. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).

58. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).