Timing, rates and spectra of human germline mutation

Raheleh Rahbari[#1], **Arthur Wuster**[#1,5], **Sarah J. Lindsay**[1], **Robert J. Hardwick**[1], **Ludmil B. Alexandrov**[1], **Saeed Al Turki**[1], **Anna Dominiczak**[2], **Andrew Morris**[3], **David Porteous**[4], **Blair Smith**[3], **Michael R. Stratton**[1], **UK10K Consortium**[1,$], and **Matthew E. Hurles**[1]

[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom

[2]Institute of Cardiovascular and Medical Sciences, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom

[3]Medical Research Institute, University of Dundee, Dundee, United Kingdom

[4]Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, Scotland, United Kingdom

[5]Department of Human Genetics and Department of Bioinformatics and Computational Biology, Genentech Inc, 1 DNA Way, CA 94080 South San Francisco, USA

[#] These authors contributed equally to this work.

Abstract

Germline mutations are a driving force behind genome evolution and genetic disease. We investigated genome-wide mutation rates and spectra in multi-sibling families. Mutation rate increased with paternal age in all families, but the number of additional mutations per year differed more than two-fold between families. Meta-analysis of 6,570 mutations showed that germline methylation influences mutation rates. In contrast to somatic mutations, we found remarkable consistency of germline mutation spectra between the sexes and at different paternal

ages. 3.8% of mutations were mosaic in the parental germline, resulting in 1.3% of mutations being shared between siblings. The number of these shared mutations varied significantly between families. Our data suggest that the mutation rate per cell division is higher during both early embryogenesis and differentiation of primordial germ cells, but is reduced substantially during post-pubertal spermatogenesis. These findings have important consequences for the recurrence risks of disorders caused by *de novo* mutations.

Introduction

Mutations have manifold consequences, from driving evolution to causing disease. DNA damage can have exogenous causes such as ionizing radiation and mutagenic chemicals or endogenous causes such as oxidative respiration and errors in DNA replication[1,2]. Both endogenous and exogenous damage are restored by DNA repair pathways, which are highly conserved in mammals[2]. However, damage repair pathways are not perfect and *de novo* mutations (DNMs) occur in every generation.

Knowledge of the rates and mechanisms by which germline mutations arise has diverse applications, from empowering the discovery of the genetic causes of rare disorders[3], to dating critical periods in human evolution[4]. Based on whole-genome sequencing studies of trios the average generational mutation rate of single base substitutions in humans has been estimated[5-9] to be ~1-1.5×10^{-8}.

In 1947, Haldane noted that the mutation rate of the hemophilia gene is significantly higher in men than in women[10]. Recent genome sequencing studies confirmed Haldane's observation that the male germline is more mutagenic[5-8,11]. On average, each additional year in the father's age at conception results in ~2 additional DNMs in the child[6]. Correspondingly, the risk of dominant genetic disorders in the child increases with increasing paternal age[12,13]. The most likely cause of the paternal age effect is the increasing number of cell divisions in the male germline[14]. While oocytes are produced early in a woman's life and have a fixed number of genome replications, spermatogenic stem cells undergo continuous genome replication throughout a man's life. It has been estimated that the male germline experiences 160 genome replications in a 20 year old male, rising to 610 genome replications in a 40 year old male[15].

Mutation rate depends on local nucleotide context. Moreover, studies of somatic mutations in cancer have shown that the observed mutation spectra can be decomposed into different 'mutational signatures' that reflect particular cellular contexts of exogenous and endogenous mutagen exposure and the efficiency of different DNA repair pathways[16].

The germline comprises a lineage of different cellular contexts, from the zygote to the gamete[17] (Supplementary Figure 1). Post-zygotic mutations can potentially lead to germline mosaicism. Observing apparent DNMs shared between siblings – predominantly in studies of dominant disorders – has provided direct evidence for germline mosaicism[18]. While recent studies have determined the average germline mutation rate and estimated the average paternal age effect, a deeper understanding of germline mutational rates, spectra and their underlying mutational processes remains elusive. For example, it is not known whether

mutation spectra differ between paternal and maternal germlines, nor whether mutation rates and spectra vary significantly between families, or whether different stages of the cellular lineage between zygote and gamete differ in their mutation rates and spectra.

Here, we investigated human germline mutations within and between multi-sibling families. This allowed us to compare mutation rates and spectra between families and to detect instances of post-zygotic mosaicism. We also investigated mutational processes and spectra more broadly by combining our data with previously published datasets.

Results

Family-specific paternal age effects

We sequenced the genomes of three multi-sibling families (Figure 1). We discovered and validated 768 DNMs across the three families, with an average of 64 per child (range 43–84, Supplementary Table 1). When taking into account genomic regions inaccessible to our analyses (Methods), the average number of mutations per individual increases to 76.9. This adjusted number of mutations is equivalent to an average mutation rate of 1.28×10^{-8} (95% confidence interval $1.13–1.43 \times 10^{-8}$) at a mean paternal age of 29.8 years. In the following analyses, we used the adjusted number of mutations.

We determined the parental origin of 399 DNMs, 311 of which (78%) were of paternal origin (Figure 1). Our data confirm the paternal age effect. Taking all families together, the number of DNMs increases with the fathers' age by 2.87 per year (95% confidence interval 2.11–3.64). In all three families, there is a 12–13 year gap between the youngest and oldest siblings, which enabled us to estimate the parental age effect for each family separately. The correlation between paternal age and the number of DNMs in the child was even stronger when each family is considered separately (Figure 2). The parental age effect for family 244, 603, and 569 is 1.46 (95% confidence interval 1.15–1.78), 3.27 (CI 2.07–4.47), and 3.65 (CI 1.52–5.77) mutations per year, respectively. Overall, a model that takes both paternal age and family into account performs significantly better in predicting the number of mutations in the offspring than a model that only considers paternal age (p = 0.020, Analysis of Variance).

Germline mosaicism in parents

Mutations that occur during early development can lead to mosaicism in germline and/or somatic tissues. Germline mosaic mutations in parents could be passed on to more than one child. We used two orthogonal approaches to identify potential parental germline mosaic DNMs in our multi-sibling family sequencing data, by deeply sequencing every validated DNM in every individual in all three pedigrees to a mean depth of 567X per individual (Methods).

First, we identified 10 validated DNMs that are shared between at least two siblings in the same family, which are clearly not constitutively heterozygous in either parent (alternate allele fraction <10%). Based on this, the probability of any germline mutation being shared between two siblings is 1.3% (Supplementary Table 2).

Second, by identifying sites with a significant excess of alternative (ALT) reads in the DNA from a single parent (Methods), we distinguished sites among the validated DNMs that were potentially mosaic at low levels in parental blood (Figure 3B, Table 1, Supplementary Figure 2). This approach identifies germline mutations mosaic in at least one parental somatic tissue, and thus most likely occurred during early embryonic development of the parent, prior to the separation and proliferation of the germline and the soma, and consequently are mosaic in both tissues. We attempted further experimental validation of the candidate mosaic sites using orthogonal amplification and sequencing technologies (**Methods**). Taking these independent experiments together, we identified 25 DNMs with excess parental ALTs, ranging from 0.6% to 10% of the reads, with a median of 3%. We modeled our statistical power to detect parental somatic mosaicism (Figure 3A) and conclude that we have ~80% power to detect a mosaic variant present in 1% of parental blood cells and ~90% power to detect a variant in 2% of parental blood cells.

Six of the ten DNMs shared among siblings also exhibited parental somatic mosaicism, which is a significant enrichment (p=4.6e-7, Fishers Exact Test). Four DNMs were shared among siblings without excessive ALTs in parental blood. Hence they either occurred after the separation of germline and soma, or had parental somatic mosaicism below detectable levels. In total, 29 of validated DNMs have evidence of parental germline mosaicism (Table 1). Correcting for our incomplete power to detect mosaic mutations (Figure 3A), suggests that 4.2% of germline mutations may be may be mosaic in >1% of parental blood cells (**Methods**).

64% (16/25) of the parental mosaic DNMs were maternal in origin. This is compatible with a 1:1 ratio of paternal and maternal somatic mosaicism but represents a significantly different ratio of parental origins compared to paternal bias observed in all 768 DNMs (p=7.7e-6, binomial test). There is not likely due to differential sequencing-coverage between mothers and fathers (Supplementary Figure 3).

Germline mutational spectra

We compiled a catalogue of 6,570 high confidence DNMs from 109 trios based on six different sources, including the families we sequenced for this project (Supplementary Table 3). All DNMs were called from whole-genome sequencing data. For 10% of the mutations, data on parental origin were available.

We used this catalogue to evaluate evidence for distinct germline mutational processes. Low resolution mutational spectra, which we define as the relative frequency of the six possible point mutations confirm the expected preponderance of transitions over transversions (Figure 4A). There was no significant difference between the spectra of maternal and paternal mutations (p = 0.19, Chi-squared test; Figure 4B). Even though there is a significant difference in the magnitude of the paternal age effect between the three families, there is no significant difference between the mutational spectra of the three families (p = 0.925, Chi-squared test, nor between the spectra of DNMs of children born to young and old fathers (p = 0.83, Chi-squared test; Figure 4C).

As an independent assessment of potential differences in maternal and paternal mutation spectra, we contrasted variants identified on chrX and chrY in a genome-wide sequencing dataset based on 2,453 individuals from the UK10K project. All variation on chrY arose in the male germline, whereas variation on chrX is generated in both the maternal and paternal germline. We observed that only rare variants faithfully recapitulate the mutation spectra observed in de novo mutations[19], as the ratio of C:G>T:A and T:A>C:G transitions decreases dramatically with increasing derived allele frequency, most likely because of biased gene conversion[20] (Supplementary Figure 4). We did not observe any statistically significant difference (p = 0.10, Chi-squared test) in chrX and chrY mutation spectra (number of variants = 3,217) after accounting for base composition differences between the chromosomes (**Methods,** Supplementary Figure 5). This confirms our observation above that despite the differences in mutation rates, numbers of genome divisions and cellular contexts, the mutation spectra in the maternal and paternal germline are very similar.

To investigate the contribution to germline mutation of 30 previously identified and validated mutational signatures operative in somatic lineages leading to cancer[16], we characterised higher-resolution mutational spectra. For this, we calculated the relative frequency of mutations at the 96 triplets defined by the mutated base and its flanking base on either side (Figure 5A). The spectrum observed for germline mutations clearly recapitulates the known higher mutability of CpG dinucleotides.

We evaluated if any combination of the 30 previously identified signatures[16] is sufficient to explain the observed pattern of germline mutations (Figure 5B). Two of the mutational signatures, previously termed Signatures 1 (25% of DNMs) and 5 (75% of DNMs), explain the majority of the observed mutational pattern (Pearson correlation = 0.98; Figure 5C). Including any additional mutational signatures did not significantly improve this correlation. Signature 1 is characterised by C:G>T:A mutations at CpG dinucleotides, while Signature 5 is predominately characterised by T:A>C:G mutations (Supplementary Figure 6). These signatures are responsible for the generation of the majority of spontaneous pre-neoplastic somatic mutations[16], indicating that the mutational processes underlying these signatures in somatic cells are also operative in the germline.

Methylated CpG sites spontaneously deaminate, leading to TpG sites and increasing the number of C:G>T:A mutations[21]. To test whether methylation status in the germline has a detectable impact on mutations, we obtained cell-line methylation data for three cell types that had been generated by reduced representation bisulfite sequencing as part of the ENCODE project[22]. In the testis cell-line, 25.3% of CpG sites had more than 50% of reads methylated (Supplementary Table 4). 13 of those sites overlap with DNMs from our catalogue, of which 12 have more than 50% of reads methylated. This means that in the testis cell-line, methylated CpG sites are significantly more likely to mutate than unmethylated ones (p = 1.71×10^{-8}, Binomial test). All of the 12 DNMs that were methylated in the testis are CpG>TpG mutations (Supplementary Table 5). For B-lymphocyte and embryonic stem cell-lines, the association between methylation status and mutation is less significant (p = 0.04 and p = 2.39×10^{-6}, respectively).

Discussion

We sequenced the genomes of three multi-sibling families, identified candidate DNMs and validated 768 of them by targeted re-sequencing. Both the average genome-wide mutation rate of 1.28×10^{-8}, and the ratio of paternal to maternal mutations (3.5) are slightly higher than but compatible with previous estimates[6]. On average, the number of mutations in the child increased approximately linearly by 2.9 additional mutations with each additional year in the parents' age. The magnitude of this effect differed by a factor of >2-fold between families. While our observations corroborate a previous study[6] that proposed that the major factor influencing the number of mutations in a child rate is paternal age, our multi-sibling study design allows detection of more subtle differences between families. Given that the increase in mutations with parental age is driven by paternal mutations, we suggest that this observation could result from variation between males either in the rate of turnover of spermatogenic stem cells, or in the mutation rate per cell division. A recent review noted that the strength of the paternal age effect differs between studies[23]. Whilst this could be due to study design or analysis choices, our results highlight a more interesting possibility, namely that, due to the families in each study, the paternal age effect actually differed between the studies, most of which had a limited sample size.

We observed no difference in mutation spectra between the maternal and paternal germlines or between young and old fathers. The lack of large differences in mutation spectra between the sexes is perhaps counter-intuitive, given the different cellular contexts in the maternal and paternal germline, including the marked difference in cell divisions and thus the increased potential for replication-associated mutations in the paternal germline. Larger catalogues of paternal and maternal mutations will be required to identify any subtler differences in germline mutation spectra.

We have shown that a combination of two previously identified mutational signatures operative in somatic cell lineages are sufficient to explain the observed mutational spectrum of germline mutations. These two mutational signatures were originally extracted from somatic mutations derived from diverse cancer genomes and thus likely reflect mutation processes operative across somatic tissues[16]. This high concordance between the germline and the soma suggests that the mutation processes underlying these two signatures are associated with maintenance and replication of DNA in all cells. The generality of these two signatures, and their underlying mutation processes, across diverse cellular contexts, likely explains our observation of an absence of appreciable age- or sex-dependent variation in mutation spectrum. Nonetheless, despite this genome-wide concordance across different cellular lineages, our observation of increased mutation rate at sites known to be methylated in a testis-derived cell-line revealed that DNA methylation, and perhaps other cell-type specific factors, has a finer-grained role to play in influencing the precise location of mutations in specific cell-types.

With regard to the timing of mutations in the cellular lineage of the germline, we have shown that at least 3.8% of DNMs are mosaic in at least 1% of parental blood cells. This estimate represents a lower bound on the true proportion of DNMs that are mosaic in parental somatic tissues, as we only sampled a single somatic tissue and cannot exclude the

possibility of very low level (<1%) mosaicism in that tissue. This proportion is compatible with a recent estimate for parental somatic mosaicism of copy number variants[24]. We infer that DNMs that are mosaic in parental soma must have arisen early on during embryonic development of the parent (first 8-12 cell divisions [25,26]), prior to the specification of primordial germ cells (PGCs) and the concomitant separation of the germline from the soma. Whereas all DNMs showed a 3.5:1 ratio of paternal to maternal mutations, these early mutations were compatible with a 1:1 ratio of paternal and maternal origins, as might be expected given the origin of these mutations prior to sexual differentiation of the embryo.

We note that our observations are not compatible with monophyletic origins of blood and germline, but that each tissue must be founded by multiple cells with polyphyletic ancestry. A logical consequence is that some mesoderm founder cells are more closely related to primordial germ cells within the cellular genealogy of the early embryo than they are to other mesoderm founder cells, and vice versa.

One limitation of our study is not having complete ascertainment of all pre-PGC mutations. Mutations that arose in very early post-zygotic divisions may well be present at such high frequencies within parental tissues that our analytical workflow for identifying candidate de novo mutations fails them on the basis that such sites are much more likely to be inherited variants with a biased sampling of alleles. Moreover, pre-PGC mutations that arose in later cell divisions, only just before PGC specification, may be mosaic in parental somatic tissues at such low levels that our deep resequencing was unable to identify them. Nonetheless, the 20-fold difference in levels of somatic mosaicism that we could detect suggests that we were able to detect pre-PGC mutations across at least 4 rounds of early embryonic cell division ($2^4 < 20$).

Using the data we have generated on the paternal age effect and the prevalence of parental somatic mosaicism, we can interrogate the mutagenicity of different phases of gametogenesis. By assigning mutations to early embryonic cell divisions prior to PGC specification, we can estimate a credible range for the mutation rate in early cell divisions in parental germlines. Based on the sharing of pre-PGC mutations between gametes from the same parent, we can define a maximal and minimal number of pre-PGC cell divisions within which the observed pre-PGC mutations must have occurred, and from these estimate an upper and lower bound on the mutation rate per cell division. Our data suggest that the pre-PGC mutation rate per cell division is in a range of ~0.2 to 0.6 (for a haploid genome) in both parental germlines). The paternal age effect that we observed implies that a lower mutation rate per cell division of a range of ~0.09 to ~0.17 (~2-4 paternal mutations per year derived from 23 cell divisions) operates during post-pubertal spermatogenesis. By contrast, oogenesis appears to be significantly more mutagenic than post-pubertal spermatogenesis with a mutation rate per cell division of ~0.5 to ~0.7 (~10-14 maternal mutations arise during ~20 post-PGC cell divisions[27]). In the paternal germline, we also need to consider an intermediate phase of cell division, during the proliferation and differentiation of PGCs to form pre-spermatogonia during prenatal development. This phase of spermatogenesis is contemporaneous with oogenesis in females. By extrapolating the paternal age effect we can estimate the total number of paternal mutations at puberty (averaging across pedigrees and assuming no maternal age effect) to be ~19, and by subtracting the number of pre-PGC

mutations (~2-6, from ~10 divisions), we can estimate the number of paternal mutations that arose during this intermediate phase to be ~13-17. It has been estimated that there are ~24 cell divisions during this phase[27] giving a range of mutation rate per cell division of ~0.5-0.7, very similar to that observed during maternal PGC proliferation and differentiation to oogonia.

From these observations we derive a tentative model of germline mutation rate during gametogenesis (Figure 6), with two phases of oogenesis and three phases of spermatogenesis, wherein the mutation rate per cell division is higher during early embryogenesis and during PGC proliferation and differentiation during later embryogenesis, and reduces ~3-fold during post-pubertal spermatogenesis. This model is consistent with prior inferences that the average mutation rate per cell division must be higher in the female germline given the relative number of cell divisions and the ratio of paternal and maternal mutations, and this could be due to a lower error rate per cell division after puberty in males [23]. It has previously been suggested that the earliest embryonic divisions exhibit elevated mutagenicity with respect to structural variation[28]. Our data suggest that for SNVs, the main step change in mutation rate per cell division may be between embryonic and post-pubertal phases of gametogenesis in males, and a similar observation has been reported in mice spermatogenesis[29]. If the model that we have proposed above proves to be correct, then it suggests that evolutionary selection may have acted to lower the mutation rate per cell division during post-pubertal spermatogenesis, perhaps achieving a selective balance between producing sufficient numbers of sperm to maintain fertility, while minimizing the deleterious mutation rate.

It is important to note that the estimated ranges for the mutation rate per cell division presented above represent a combination of mutations that arise during genome replication and any spontaneous mutations between cell divisions. The time interval between cell divisions differs markedly throughout the different phases of gametogenesis, and so these mutation rate estimates do not necessarily reflect the mutagenicity of genome replication in isolation.

We infer that germline DNMs that are mosaic in parental soma will also be mosaic in the germline, indeed we observed that the six parental somatic mosaic DNMs that were shared among siblings had significantly higher levels of somatic mosaicism, on average, than the other parental somatic mosaic DNMs that were not shared between siblings (p=0.009, Mann-Whitney test). This suggests that the extent of somatic mosaicism correlates with the extent of germline mosaicism, and hence the probability that a DNM will be observed recurrently among siblings.

We identified four DNMs that were shared among siblings, and thus are highly likely to be mosaic in the parental germline, although we observed no evidence for accompanying somatic mosaicism in parental blood. We infer that these mutations may have arisen in early cell divisions post-PGC specification and thus mosaicism is restricted to the germline.

Previous studies of germline mosaicism of sequence variants have been largely limited to case studies of sibling recurrence of pathogenic DNMs[30-34]. Our 1.3% estimate of the

average recurrence probability is compatible with those empirical studies, but they are not compatible with recent lower estimates of recurrence risks derived from theoretical modeling of the cellular genealogy of the germline[35]. We note that these recurrent DNMs between siblings were not randomly distributed between families, but were significantly enriched in one pedigree. This suggests that there may also be significant variation between families in patterns of germline mosaicism of DNMs.

These results on germline mosaicism have implications for the genetic counselling of recurrence risks for families with children with genetic disorders caused by DNMs [17]. While the currently used recurrence risk of ~1% is supported by our findings, our data suggest that this represents an average across DNMs with very different recurrence risks. While only 1.3% of all DNMs were observed recurrently among siblings, this increases to 24% for DNMs that were mosaic in >1% of parental blood cells and 50% for DNMs mosaic in >6% of parental blood cells. Our data suggest that deep sequencing of parental blood for pathogenic DNMs seen in children should enable meaningful stratification of families into a substantial majority with <1% recurrence risks, and a small minority with recurrence risks that could be at least an order of magnitude higher. Considerably more data will be required to enable more precise quantitative estimates of recurrence risks given an observed extent of parental somatic mosaicism.

Our data also show that in the absence of deep sequencing of parental somatic tissue(s), knowing the parental origin of DNM alters the recurrence risk, with maternal mutations likely having a ~3 to 4-fold higher recurrence risk, on average, than paternal mutations. As noted previously[24], the higher probability of germline mosaicism for maternally-derived DNMs results in a higher recurrence risk, on average, for DNMs causing X-linked recessive disorders than for autosomal dominant disorders.

Pedigree based analyses always are limited by the number of offspring available. Deep sequencing of single gametes from different individuals[36] should enable us to characterise and compare their mutation rates and spectra at much higher resolution. This will also mitigate any biases associated with the selection inherent during conception and fetal development, although it would still be prone to biases caused by mutations that confer enhanced proliferation on progenitors of gametes[12]. Moreover, sequencing progenitors of gametes from different stages of the germline would illuminate our current limited understanding of the selective pressures operative throughout the genealogy of the germline.

Online methods

We conducted a study of genome-wide germline mutations by sequencing the genomes of three healthy families who participated in the Scottish Family Health Study (SFHS). Informed consent was obtained from all participants. The families were selected based on genomic DNA quality, the number of children, and the age gap between the oldest and youngest sibling.

De novo mutation discovery

For each of the three families, the two parents and children were sequenced to 24.7×
coverage on average. In one of the families (Family 569), a child of one of the proband was
also sequenced. We used the DeNovoGear software[1] to identify 49,893 candidate DNMs in
the children. We identified likely false positives as those sites that overlapped low
complexity regions[2], which we defined as segmental duplications or simple repeats. Further,
we removed sites that had more than 5% of reads supporting the alternative allele in either of
the parents. To avoid regions with a large number of misaligned reads, we also removed sites
whose depth was in the top 0.01% quantile in terms of read depth. For this, we assumed read
depth to be Poisson-distributed, with the lambda parameter of the Poisson distribution being
equal to the mean read depth of the genome. Taken together, these filters resulted in 4,881
candidate sites.

For validation, we designed Agilent SureSelect probes around the sites that passed filtering
and resequenced the resulting pulldown library using Illumina to 139× coverage on average
(range: 88-191). We designed baits to cover a 200bp window around the candidate sites. The
bait design succeeded for 4,141 sites. To analyse the validation data, we classified each
putative DNM into one of three categories: Germline DNM, inherited variant or false
positive and evaluated the likelihood of the data under each model. The three models are
defined below. In addition, 37 of the DNMs were removed following manual inspection in
the IGV genome browser.

Model 1: Germline DNM—We defined the likelihood of the data under the DNM model
as:

$$LL.DNM = Pois\left(m_m, m_T * e\right) + Pois\left(d_m, d_T * e\right) + Bin\left(c_m, c_T * e, 0.5\right)$$

mm, dm, and *cm* are the number of reads supporting the mutant allele (mostly the alternative
allele) in the mother, the father and the child, respectively. *mT, dT,* and *cT* are the total
number of reads in the mother, the father and the child, respectively. *e* is the sequencing
error rate.

Model 2: Inherited variant—The likelihood that the variant is inherited is defined as:

$$LL.I = \max\left(LL.IFM, LL.IFD, LL.IFMD\right)$$

LL.IFM, LL.IFD, and LL.IFMD refer to the likelihood that the variant is maternally
inherited, paternally inherited, or inherited from both parents:

$$LL.IFM = Bin\left(m_m, m_T, 0.5\right) + Pois\left(d_m, d_T * e\right) + Bin\left(c_m, c_T, 0.5\right)$$
$$LL.IFD = Pois\left(m_m, m_T * e\right) + Bin\left(d_m, d_T, 0.5\right) + Bin\left(c_m, c_T, 0.5\right)$$
$$LL.IFMD = Bin\left(m_m, m_T, 0.5\right) + Bin\left(d_m, d_T, 0.5\right) + Bin\left(c_m, c_T, 0.5\right)$$

Model 3: False positive—

$$LL.FP = Pois\left(m_m, m_T * e\right) + Pois\left(d_m, d_T * e\right) + Pois\left(c_m, c_T * e\right)$$

Correction of the mutation rate

The correction accounts for the part of the genome that we could not interrogate because of insufficient depth in low complexity regions, filtering procedures to exclude false positives, and failed validation. To take into account the different karyotypes of the male and female genomes, the precise form of the correction depend on the sex of the proband:

$$Girls: (1- \quad noCvg) * (1 - filtered) * (1 - noVal * ppAdjust) * 2 * valDNM/genomeLength$$
$$Boys: (1- \quad noCvg) * (1 - filtered) * (1 - noVal * ppAdjust) * 2$$
$$* valDNM + valDNMX/genomeLength$$

noCvg is the proportion of the genome that is either N or not covered at 7x or more, *filtered* is the proportion that is a segmental duplication or a simple repeat (but not N or low coverage), *noVal* is the proportion that passed filtering but for which validation was not possible (mainly due to failed primer design), *ppAdjust* is the proportion of non-validatable calls that are likely to be true positives based on their posterior probability as calculated by DeNovoGear, *valDNM* is the number of validated DNMs, *valDNMX* is the number of validated DNMs on the X chromosome, and *genomeLength* is the length of the human reference genome build 37 without the Y chromosome, unmapped regions, and mitochondrial DNA. This correction assumes that the mutation rate is similar in the inaccessible regions of the genome. On average, 83.1% of the genome was accessible, ranging from 82.1% to 84.3% in different genomes.

Identification of DNMs mosaic in parents

We used two analytical methods to identify potential parental mosaic DNMs in our multi-sibling family sequencing data: DNMs shared among siblings and DNMs with excess alternative reads in DNA from one parent.

Method 1: Identification by recurrence in siblings—Only validated and therefore high confidence DNMs, were used for this analysis. Validation ensured that the DNMs were not constitutively heterozygous in either parent. This method involved the identification of DNMs that were present in more than one offspring from the same family.

Method 2: Identification by excess of alternative reads in a parent—Potential parental germline mosaic events were further investigated in the 768 validated DNMs by identifying instances of a significant excess of reads supporting the alternative allele in one of the parents. To improve our power to detect candidate germline mosaic sites, we performed an additional Miseq run of the custom pull down library we previously used for validation, which resulted in an average coverage of 500X for validated DNMs (n=768). The site-specific error rate for each DNM was estimated by dividing the total number of reads supporting the alternative allele by the total number of reads in all non-related individuals, from the two families, in which the DNM was not discovered. Hence the probability that the

observed number of parental ALT alleles resulted from sequencing error was calculated as follows:

$$\begin{aligned} p_{Maternal} &= Bin\left(m_m, m_{alt+ref}, e\right) \\ p_{Paternal} &= Bin\left(f_{alt}, f_{alt+ref}, e\right) \end{aligned}$$

Where m and f are the number of reads in the mother and father respectively, alt and ref are the alternate and reference alleles respectively, and e is the site-specific error rate. Both maternal and paternal p values for each DNM were adjusted for multiple testing using the Bonferroni correction. Sites that were significant at an adjusted $p < 0.05$ were considered as mosaic. In total, 24 mosaic sites were validated using this method. Six of these were also discovered by the sibling recurrence method described above.

Estimation of recurrence risk—The probability of an apparent DNM being shared between more than one sibling in the same family was calculated as number of instances of a mutation being shared between two siblings divided by the number of pairwise comparisons between two siblings, in all three families (Supplementary Table 2).

Validation of DNMs mosaic in parents

We carried out further independent validation of 40 candidate parental mosaic DNMs (**method 1 and 2,** Supplementary table) using PacBio amplicon sequencing. These 40 candidate mosaic DNMs were selected as follows: 10 DNMs that were shared between siblings (for six of these shared DNMs we had previously identified a significant parental excess of alt alleles, as described above), and 30 candidate mosaic sites that had an excess of alt alleles in a parent's blood, with a nominal p value < 0.05 . Note this set of 30 candidates was based on nominal significance rather than Bonferroni corrected significance and so represents a less stringent set of candidate mosaic DNMs.

Primers were designed using Primer 3[3] in order to generate amplicons with an average length of 250 bp, with the candidate mosaic site in the middle of the amplicon. For each candidate mosaic site, amplicons were prepared for the mosaic children and their parents, including a unique 11 bp sequence in the forward primer to act as a barcode for each individual. The amplicons were prepared using a standard PCR protocol. Two of the candidate mosaic sites (chr2: 37841931, and chr4: 131248301) failed to amplify and therefore were not included in this validation experiment.

In total 114 amplicons were successfully prepared for the remaining 38 sites. Amplicons were pooled in equimolar amounts and prepared for circular consensus sequencing with shared libraries on PacBio SMRT cell.

Following PacBio sequencing, the filtered subreads and ROI (reads of insert) were generated using SMRTAnalysis (provided by Pacific Biosciences, Menlo Park CA). The resulting fastq files were demultiplexed based on the 11bp unique barcodes for each individual and mapped to the human reference genome GRCh37 (hg19). Average sequence coverage from the PacBio data was 158X across the 114 amplicons. Lastly, variants were called from the resulting BAM alignments using samtools[4] mpileup, version1.1. Each of the candidate

parental mosaic DNMs were only further analysed if we observed ~50% ref/alt in the child, and hence their parental alt/ref were counted. We categorised sites with this criteria into: 1-Validated, comprising sites where we observed alternative alleles in the relevant parent, 2-Uncertain, comprising sites where we had <90% power to detect the alternative alleles in the parents (PacBio detection power was calculated using the mosaicism level from the Miseq data), and finally 3-Not validated, comprising sites where we had>90% power to detect the alternative allele in the mosaic parents but failed to detect them.

We classified 29 of the 40 candidate sites set as parentally mosaic. Four mosaic DNMs were shared between the siblings in the same family but we could not observe alternative alleles in either parent in either validation dataset (Miseq and PacBio). 16 sites were validated as mosaics where the mosaic parent was confirmed on both platforms (all of these sites had significant p-value for their Miseq data after the Bonferroni correction). One additional site with significant nominal p-value but not significant adjusted p-value for the Miseq data was confirmed mosaic on PacBio. Two sites were confirmed as mosaic based on significant adjusted p-value from the Miseq only, as they failed the PacBio experiment. Six sites were confirmed mosaic based on Miseq only (with significant adjusted p-values), as their mosaicism level was below detection power on PacBio. In the remaining 11 sites, despite having significant nominal p-values, the adjusted Miseq p-value was not significant and the PacBio data was inconclusive (Table 1 and Supplementary table).

In summary, we attempted further experimental validation of 40 candidate mosaic sites by conducting deep amplicon sequencing (158X mean coverage per individual) in the child, mother, and father's blood using the PacBio platform. This validation experiment confirmed the presence of ALT reads in parental blood-derived DNA at 100% of DNMs (N=9) where the PacBio data had >90% power to detect the level of mosaicism observed in the MiSeq data. Furthermore, we observed 100% concordance (N=14) between the parental-origin determined by significant excess of ALT reads in maternal or paternal blood and that determined by phasing the DNM onto a parental haplotype.

Correction for mosaic power detection

In order to estimate the number of mosaic sites that we failed to detect due to power limitations, we ran 1000 simulations across our 768 validated *de novo* mutations with their given coverage (from Miseq sequencing) for a range of mosaicism levels. We calculated the number of sites with >2% mosaicism that we failed to identify. For this we defined two bins for the mosaic level (2%-4%, >4.0%). The average undetectable mosaic sites were calculated as a product of number of mosaic sites and average detection power for each bin. Hence, the number of germline mosaics after power adjustment is ~4% (31/768) of validated *de novo* mutations.

Parent of origin

To study the effect of parental age and sex on germline mutations we determined the parental origin for the validated germline DNM using three approaches.

Firstly, we used DeNovoGear's readpair algorithm1 to obtain parental phasing information. In short, this algorithm determines the parent of origin if haplotype informative sites are

present in phase with the mutation in the child and in the parents. The informative sites are those that are phased with respect to the mutation in the child because they are located on the same read pair. Furthermore, the genotype of the site must be informative in the parents. Using this method, we identified an informative haplotype for 198 mutations.

Secondly, a child of one of the probands (SFHS5165328 in Family 569) was also sequenced. For this proband, the parent of origin was determined using informative variants in a 20 kb window around the DNM. If the paternal haplotype was transmitted to the proband and the child also carried the DNM, then the mutation was classified as being of paternal origin. Similarly, if the child carrying the paternal haplotype did not have the DNM, then the mutation was classified as being of maternal origin. The same logic was applied when the child inherited the maternal haplotype of the proband. Using this method, we identified an informative haplotype for 30 mutations.

Thirdly, we experimentally ascertained the parental haplotype on which the DNM arose. Genomic DNA from the child was diluted to single molecule concentration and then re-amplified across 48 wells using Repli-G Midi Kit from Qiagen. The resultant amplified DNA, along with undiluted genomic DNA from the child and the parents, was then Sequenom genotyped at the putative DNM of interest, along with the nearest haplotype informative SNP (heterozygous in the child, and heterozygous in one of the parents). If genotyping assays were heterozygous in child, homozygous in parents at the putative DNM in the unamplified DNA, and homozygous in the single molecule amplified wells, then the raw genotype data from the 48 amplified single molecules was analyzed in two ways. Firstly, haplotype inference was obtained from examining peak height correlations between the genotype calls for the putative DNM and adjacent informative SNP, and clustering of calls was observed using an in-house script. Secondly, genotype calls (or peak heights pertaining to genotype calls) from the same well were counted for each locus and the haplotype derived from a likelihood ratio test as detailed by Konfortov *et al.*[5]

Mutational catalogue

We generated a catalogue of human DNMs based on previously published high confidence mutations obtained by whole genome sequencing (Supplementary Table 3). Only single nucleotide DNMs were included. Where necessary, we used the LiftOver tool to convert coordinates from NCBI build 36 to build 37.

Mutational spectra and signatures

Mutational spectra were derived directly from the reference and alternative (or ancestral and derived) allele at each variant site. The resulting spectra are composed of the relative frequencies of the six distinguishable point mutations (C:G>T:A, T:A>C:G, C:G>A:T, C:G>G:C, T:A>A:T, T:A>G:T). Significance of the differences between mutational spectra was assessed by comparing the number of the six mutation types in the two spectra by means of a Chi-squared test (df = 5).

Mutational signatures were detected by refitting of previously identified consensus signatures of mutational processes[6]. All possible combinations of at least seven mutational signatures were evaluated by minimizing the constrained linear function:

$$\min_{Exposures_i \geq 0} \| \overrightarrow{DeNovoMutations} - \sum_{i=1}^{N} \left(\overrightarrow{Signature_i} * Exposure_i \right) \|$$

Here, $\overrightarrow{DeNovoMutations}$ and $\overrightarrow{Signature_i}$ represent vectors with 96 components corresponding to the six types of single nucleotide variants and their immediate sequencing context and $Exposure_i$ is a nonnegative scalar reflecting the number of mutations contributed by this signature. N reflects the number of signatures being re-fitted and all possible combinations of consensus mutational signatures for N between 1 and 7 were examined, resulting in 2,804,011 solutions. Model selection framework based on Akaike information criterion was applied to these solutions to select the optimal decomposition of mutational signatures.

Diversity and divergence data

Diversity data was based on 2,453 individuals who were whole genome sequenced to 6-8× depth as part of the ALSPAC and Twins UK cohorts within the UK10K project. Ancestral alleles were defined by a maximum parsimony approach as those that appeared in most of five ape species (human, chimp, gorilla, orang-utan, macaque)[7]. Processing of great ape reference genome data is described below. Single nucleotide variants were determined to be equivalent to one of the six mutation types according to the identity of the ancestral and derived alleles.

Using an approach that was identical to that taken by others[8], variant sites that were likely to be under selection because they were located in exonic regions or because they were 2kb upstream or downstream of genes were filtered and excluded from the dataset. To avoid biases created by misalignment of sequencing reads, we also excluded sites that overlapped simple sequence repeats or segmental duplications. Where DNMs were compared to variants, they were subjected to the same filters.

Divergence data was based on multispecies alignments of the chimpanzee, gorilla, orang-utan, macaque and human reference genomes, as provided by Ensembl Compara. Sites likely to be under selection were removed in the same way as described for the diversity dataset. Sites that were different in humans compared to the other great ape species were defined as substitutions.

Sex chromosomes

We included only the rarest 5% of variants into this analysis, as the spectrum of those variants most resembles that of DNMs, as we show elsewhere in this study. From the resulting variants, we obtained raw mutational spectra for each chromosome, as well as mutational spectra corrected for chromosomal nucleotide composition. The correction for chromosomal nucleotide composition was done by counting the number of each of the four nucleotides in the interrogated regions of each chromosome. For each variant, we determined the ancestral and the derived allele. For each variant type, we then divided the

number of variants by the number of nucleotides that matched the ancestral allele (Supplementary Figure 5).

Methylation data

We downloaded ENCODE methylation data from the UCSC server for three cell-lines: BC_Testis_N30 (testes of a 41-year old Asian donor), GM12878 (B-lymphocytes from a European Caucasian donor), and H1-hESC (embryonic stem cells). The methylation data had been obtained by reduced representation bisulfite sequencing (RRBS). For each cell-line, two replicates were available. We only included sites that were represented in both replicates. There were 1,151,596 such sites in BC_Testis_N30, 1,048,775 in GM12878, and 1,118,911 in H1-hESC. For each cell-line, we identified sites with more than 50% of reads were methylated in both replicates combined. We also identified sites that were present in our DNM catalogue. We computed binomial p values as as $Bin(q, n, p)$, where q is the number of methylated DNMs, n the total number of DNMs for which methylation data is available, and p the proportion of sites that are methylated in the dataset.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

References

1. Lindahl T, Wood RD. Quality control by DNA repair. Science. 1999; 286:1897–905. [PubMed: 10583946]

2. Hoeijmakers JH. Genome maintenance mechanisms for preventing cancer. Nature. 2001; 411:366–74. [PubMed: 11357144]

3. MacArthur DG, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014; 508:469–76. [PubMed: 24759409]

4. Scally A, Durbin R. Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet. 2012; 13:745–53. [PubMed: 22965354]

5. Michaelson JJ, et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell. 2012; 151:1431–42. [PubMed: 23260136]

6. Kong A, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 2012; 488:471–5. [PubMed: 22914163]

7. Conrad DF, et al. Variation in genome-wide mutation rates within and between human families. Nat Genet. 2011; 43:712–4. [PubMed: 21666693]

8. Roach JC, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science. 2010; 328:636–9. [PubMed: 20220176]

9. Campbell CD, Eichler EE. Properties and rates of germline mutations in humans. Trends Genet. 2013; 29:575–84. [PubMed: 23684843]

10. Haldane JB. The mutation rate of the gene for haemophilia, and its segregation ratios in males and females. Ann Eugen. 1947; 13:262–71. [PubMed: 20249869]

11. Venn O, et al. Nonhuman genetics. Strong male bias drives germline mutation in chimpanzees. Science. 2014; 344:1272–5. [PubMed: 24926018]

12. Momand JR, Xu G, Walter CA. The paternal age effect: a multifaceted phenomenon. Biol Reprod. 2013; 88:108. [PubMed: 23515674]

13. Goriely A, Wilkie AO. Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. Am J Hum Genet. 2012; 90:175–200. [PubMed: 22325359]

14. Crow JF. The origins, patterns and implications of human spontaneous mutation. Nat Rev Genet. 2000; 1:40–7. [PubMed: 11262873]

15. Wilson Sayres MA, Makova KD. Genome analyses substantiate male mutation bias in many species. Bioessays. 2011; 33:938–45. [PubMed: 22006834]

16. Alexandrov LB, et al. Signatures of mutational processes in human cancer. Nature. 2013; 500:415–21. [PubMed: 23945592]

17. Lupski JR. Genetics. Genome mosaicism--one human, multiple genomes. Science. 2013; 341:358–9. [PubMed: 23888031]

18. Biesecker LG, Spinner NB. A genomic view of mosaicism and human disease. Nat Rev Genet. 2013; 14:307–20. [PubMed: 23594909]

19. Schaibley VM, et al. The influence of genomic context on mutation patterns in the human genome inferred from rare variants. Genome Res. 2013; 23:1974–84. [PubMed: 23990608]

20. Duret L, Galtier N. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. Annual Review of Genomics and Human Genetics. 2009; 10:285–311.

21. Cooper DN, Krawczak M. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum Genet. 1989; 83:181–8. [PubMed: 2777259]

22. Bernstein BE, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]

23. Segurel L, Wyman MJ, Przeworski M. Determinants of mutation rate variation in the human germline. Annu Rev Genomics Hum Genet. 2014; 15:47–70. [PubMed: 25000986]

24. Campbell IM, et al. Parental somatic mosaicism is underrecognized and influences recurrence risk of genomic disorders. Am J Hum Genet. 2014; 95:173–82. [PubMed: 25087610]

25. O'Rahilly, R.; Müller, F.; Streeter, GL. Developmental stages in human embryos : including a revision of Streeter's "Horizons" and a survey of the Carnegie collection. Carnegie Institution of Washington; Washington, D.C.: 1987. p. 3061 leaf of plates

26. Coticchio, G.; Albertini, DF.; De Santis, L. Oogenesis. Vol. xii. Springer Verlag; London ; New York: 2013. p. 364

27. Drost JB, Lee WR. Biological Basis of Germline Mutation - Comparisons of Spontaneous Germline Mutation-Rates among Drosophila, Mouse, and Human. Environmental and Molecular Mutagenesis. 1995; 25:48–64. [PubMed: 7789362]

28. Voet T, Vanneste E, Vermeesch JR. The Human Cleavage Stage Embryo Is a Cradle of Chromosomal Rearrangements. Cytogenetic and Genome Research. 2011; 133:160–168. [PubMed: 21311182]

29. Walter CA, Intano GW, McCarrey JR, McMahan CA, Walter RB. Mutation frequency declines during spermatogenesis in young mice but increases in old mice. Proceedings of the National Academy of Sciences of the United States of America. 1998; 95:10015–10019. [PubMed: 9707592]

30. Liu G, et al. Maternal germline mosaicism of kinesin family member 21A (KIF21A) mutation causes complex phenotypes in a Chinese family with congenital fibrosis of the extraocular muscles. Mol Vis. 2014; 20:15–23. [PubMed: 24426772]

31. Anazi S, Al-Sabban E, Alkuraya FS. Gonadal mosaicism as a rare cause of autosomal recessive inheritance. Clin Genet. 2014; 85:278–81. [PubMed: 23551117]

32. Dhamija R, et al. Novel de novo heterozygous FGFR1 mutation in two siblings with Hartsfield syndrome: A case of gonadal mosaicism. Am J Med Genet A. 2014

33. Tajir M, et al. Germline mosaicism in Rubinstein-Taybi syndrome. Gene. 2013; 518:476–8. [PubMed: 23352794]

34. Bachetti T, et al. Recurrence of CCHS associated PHOX2B poly-alanine expansion mutation due to maternal mosaicism. Pediatr Pulmonol. 2014; 49:E45–7. [PubMed: 23460545]

35. Campbell IM, et al. Parent of origin, mosaicism, and recurrence risk: probabilistic modeling explains the broken symmetry of transmission genetics. Am J Hum Genet. 2014; 95:345–59. [PubMed: 25242496]

36. Wang J, Fan HC, Behr B, Quake SR. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. Cell. 2012; 150:402–12. [PubMed: 22817899]

OnlineMethods-References

1. Ramu A, et al. DeNovoGear: de novo indel and point mutation discovery and phasing. Nat Methods. 2013; 10:985–7. [PubMed: 23975140]

2. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014

3. Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. Bioinformatics. 2007; 23:1289–1291. [PubMed: 17379693]

4. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–9. [PubMed: 19505943]

5. Konfortov BA, Bankier AT, Dear PH. An efficient method for multi-locus molecular haplotyping. Nucleic Acids Res. 2007; 35:e6. [PubMed: 17158153]

6. Alexandrov LB, et al. Signatures of mutational processes in human cancer. Nature. 2013; 500:415–21. [PubMed: 23945592]

7. Schaibley VM, et al. The influence of genomic context on mutation patterns in the human genome inferred from rare variants. Genome Res. 2013; 23:1974–84. [PubMed: 23990608]

8. Wilson Sayres MA, Venditti C, Pagel M, Makova KD. Do variations in substitution rates and male mutation bias correlate with life-history traits? A study of 32 mammalian genomes. Evolution. 2011; 65:2800–15. [PubMed: 21967423]
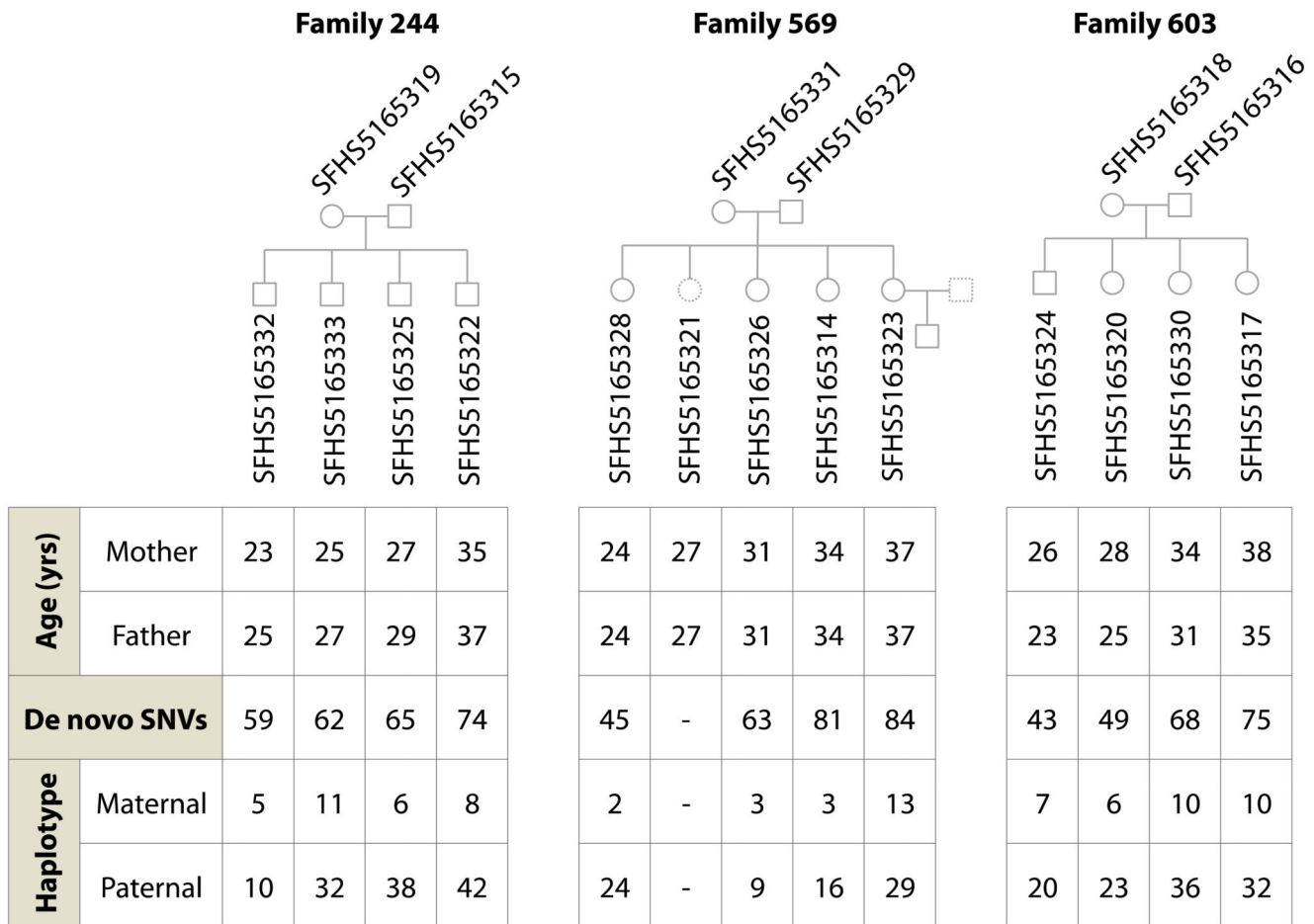
Family 244

SFHS5165319 SFHS5165315

SFHS5165332 SFHS5165333 SFHS5165325 SFHS5165322

Family 569

SFHS5165331 SFHS5165329

SFHS5165328 SFHS5165321 SFHS5165326 SFHS5165314 SFHS5165323

Family 603

SFHS5165318 SFHS5165316

SFHS5165324 SFHS5165320 SFHS5165330 SFHS5165317

Age (yrs)	Mother	23	25	27	35		24	27	31	34	37		26	28	34	38
	Father	25	27	29	37		24	27	31	34	37		23	25	31	35
De novo SNVs		59	62	65	74		45	-	63	81	84		43	49	68	75
Haplotype	Maternal	5	11	6	8		2	-	3	3	13		7	6	10	10
	Paternal	10	32	38	42		24	-	9	16	29		20	23	36	32

Figure 1. Pedigrees of sequenced families

Identifiers and relationship between the individuals in the three families in this study. Individuals that were sequenced are symbolised by full circles and squares, other individuals by dotted circles and squares. Age of mother and father at the conception of each child and phasing information are summarised in the table. SFHS5165321 was only used for the part of the analysis related to mosaicism.
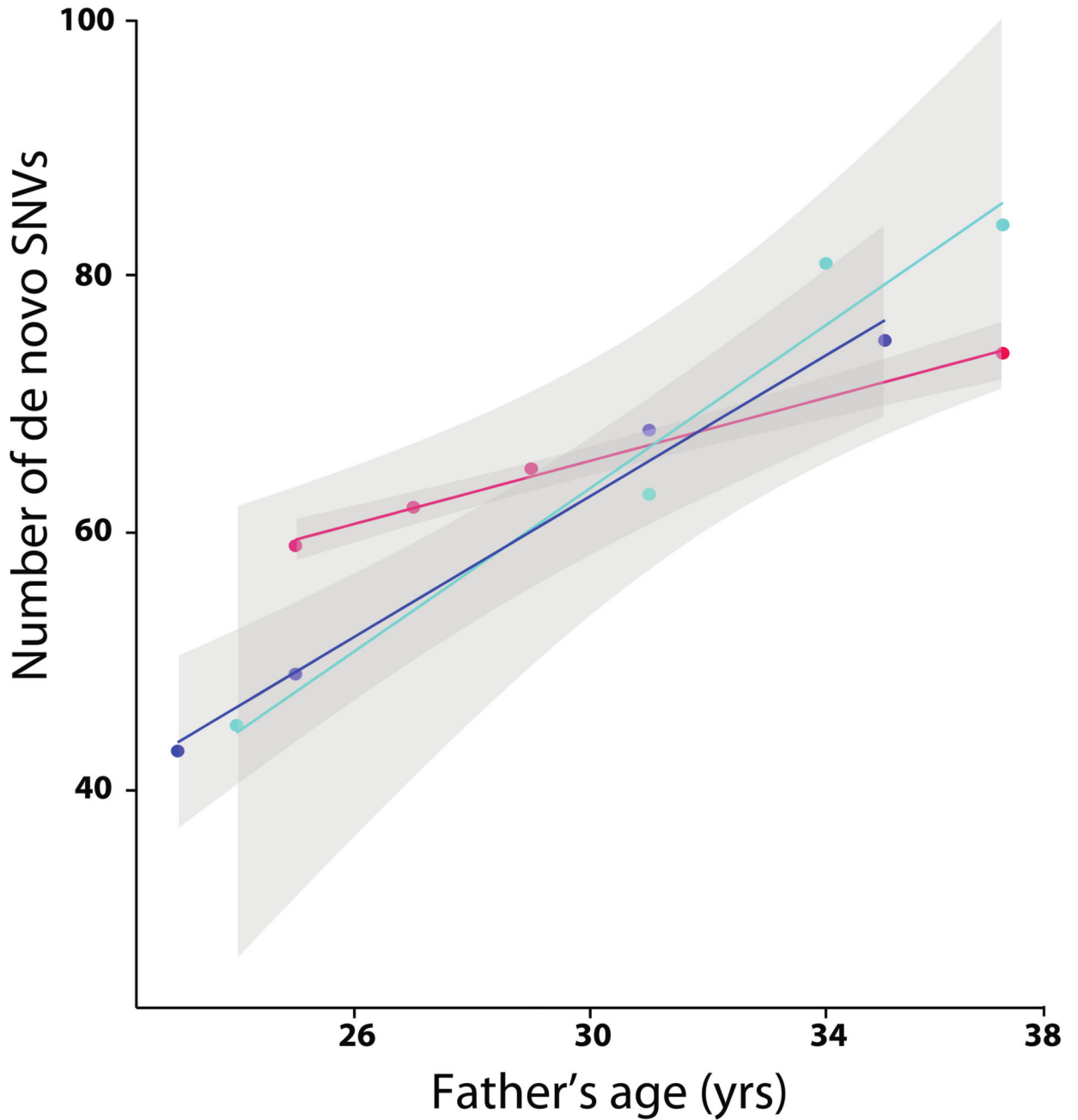
Figure 2. Paternal age vs. number of *de novo* mutations
The number of DNMs has been corrected to take into account genomic regions inaccessible to our methods. Red: Family 244. Yellow: Family 569. Blue: Family 603. Gray areas denote the regions covered by the 95% confidence interval of the intercept and slope of the linear regression line for each separate family. We note that the confidence intervals for families 244 and 603 do not overlap for younger fathers.
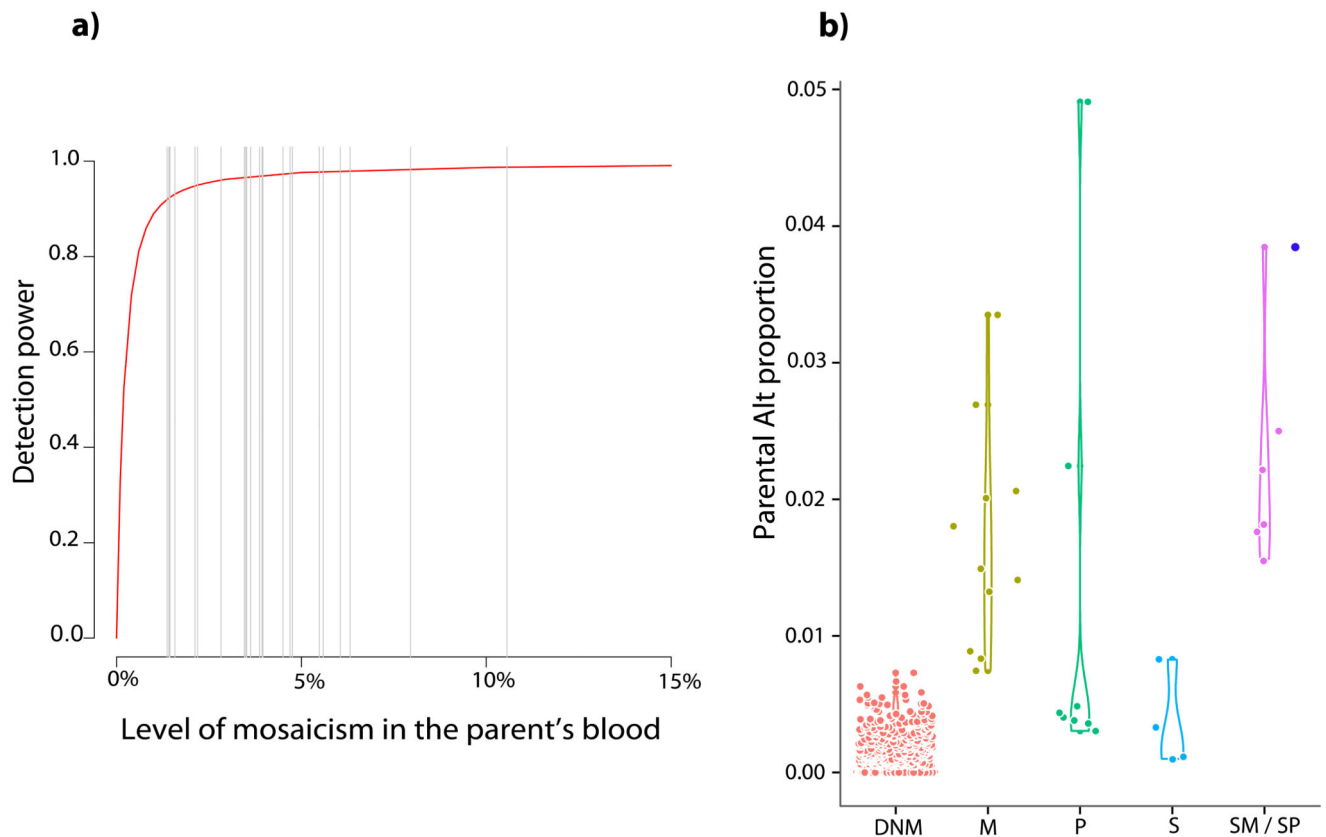
a)

b)

Figure 3. Detection of mutations mosaic in parents

(A) Simulation of detection power for ranges of mosaicism levels in the parents blood using the Miseq depth of coverage for all the *de novo* mutations (n=768). For mean validation coverage (Miseq platform) of 567X in the parents, we have >0.94 power to detect mosaicism of 2% and higher in the parents blood. (B) Comparison of parental Alt ratios between *de novo* mutations vs. germline mosaic sites. M is mosaic sites with significant excess of alt in the mother's blood, P is the sites with significant excess of alt in the father's blood, S is corresponding to the sites that are shared between the siblings but we could not detect any excess of alt allele in the either of the parents blood. SM/SP refer to the mosaic sites that are shared between the siblings and we have detected significant excess of alt in the mother's blood (SM presented in pink dots) or father's blood (SP shown in dark blue dot).
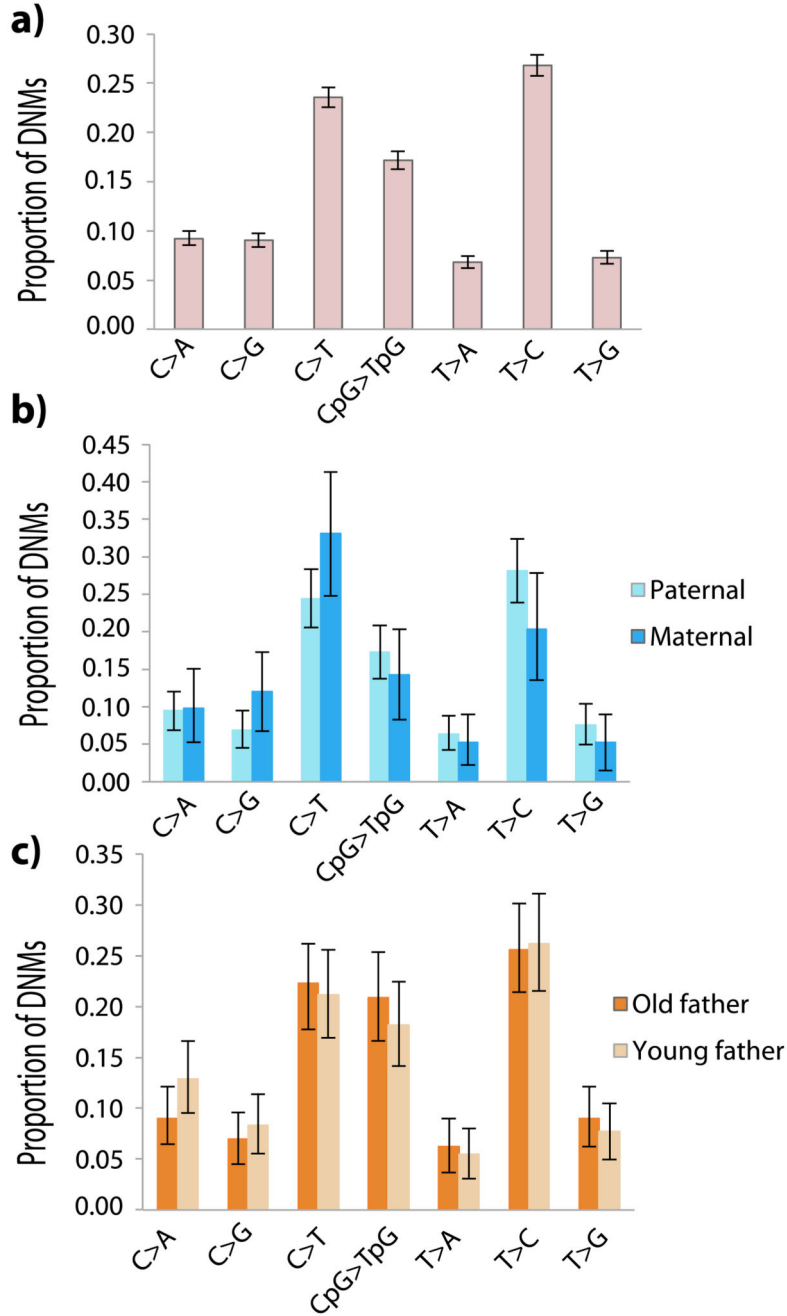
Figure 4. Mutational spectra
(A) Frequency of all mutation types in the catalogue of 6,570 high confidence DNMs (B) Difference in the frequency of maternal and paternal mutations for the subset of DNMs with phasing information (n=556) (C) Difference in the frequency of mutations of children fathers younger and older than 30 years (n=680). Error bars represent 95% confidence intervals.
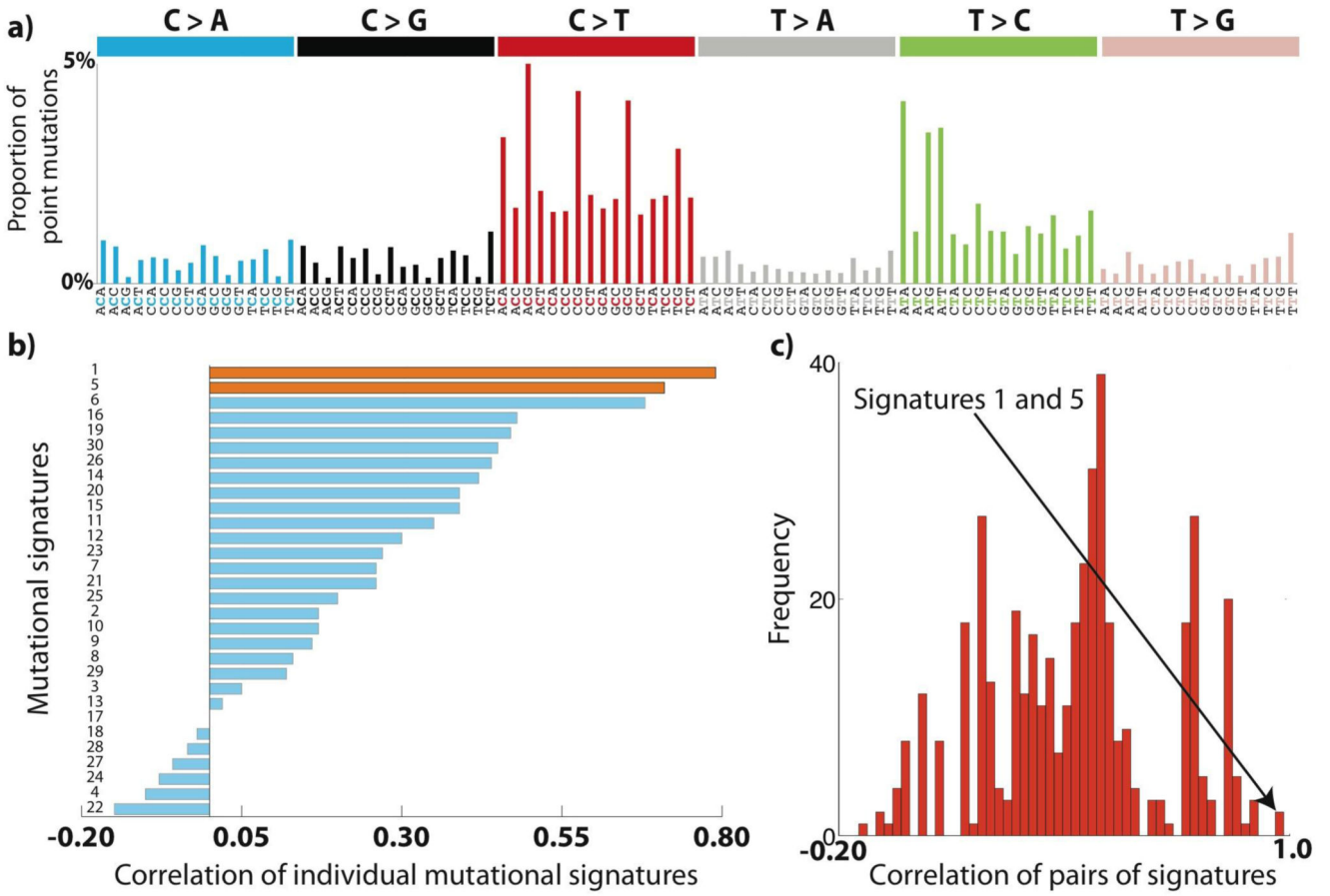
Figure 5. Mutational spectrum and signatures

(A) High resolution mutational spectrum of *de novo* mutations. Each of the six possible point mutations is subdivided into 16 subclasses based on the 3′ and 5′ nucleotide flanking the mutation. We note that C:G>T:A and T:A>C:G transitions are more common. Within those categories, CpG sites are particularly frequent (B) Correlation of mutational signatures with observed mutations in mutational catalogue, correlation of each of the 30 signatures, with signatures 1 and 5 highlighted in orange (C) Combination of all possible pairs of signatures, with the combination of signatures 1 and 5 shown with an arrow.

Figure 6. Mutation rate model during gametogenesis

Comparison of mutation rate between spermatogenesis (blue-box) vs. oogenesis (red-box). μ_p and μ_m are mutation rate in paternal and maternal genome in respective order and mutation rate per each stage of gametogenesis i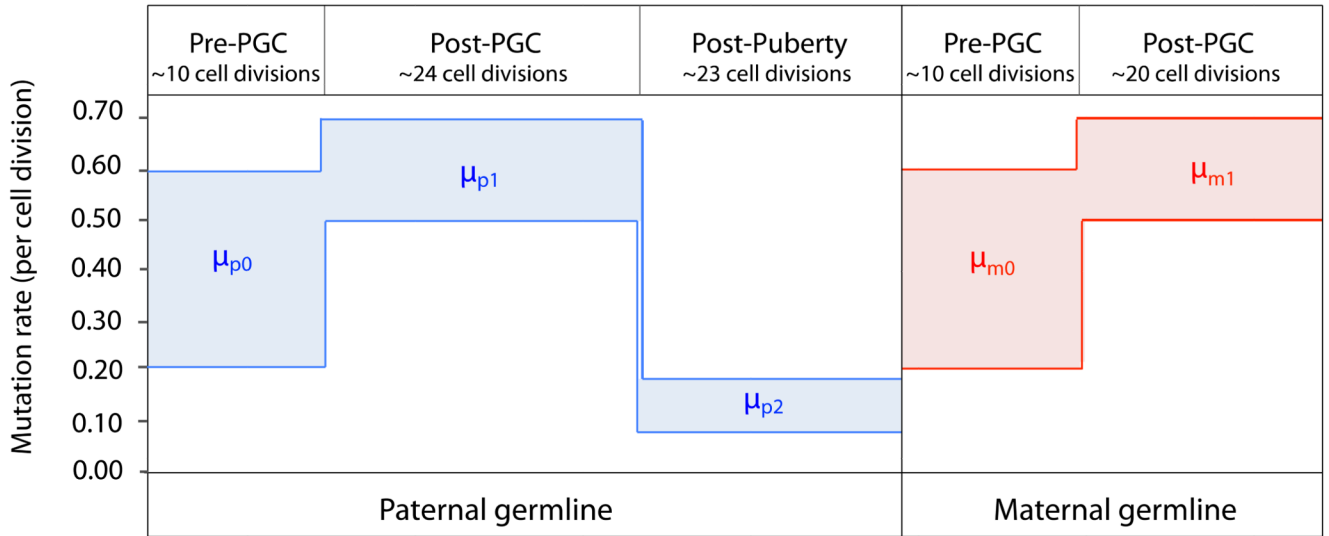s denoted by number. Gametogenesis is divided into three stages with different ranges of mutation rates. Stage 1: Pre-PGC specification (8-12 cell divisions in both maternal and paternal germline) ~0.2-0.6 mutations per haploid genome per cell division and this rate is the similar in both maternal and paternal gametogenesis, stage 2: post-PGC specification, in maternal germline there are ~20 cell divisions, in paternal germline there are ~24 cell divisions post-PGC up to puberty, mutation rate is similar at this stage in both sexes, (~0.5-0.7 mutations per haploid genome per cell division). Stage 3: post-puberty (only applicable to the paternal germline) sperm are continuously produced through the asymmetric division of self-renewing spermatogonial stem cells with ~23 cell divisions per year. The mutation rate falls to a range of ~0.09 to 0.17 mutations per haploid genome per cell division. This model is tentative and does not yet take all possible sources of uncertainty into account.

Table 1

Germline mosaic SNVs

PacBio validation; N/A: not applicable, U: indicates uncertain sites, Y: indicates sites that are validated (Methods), F: not validated, U: indicates uncertain sites, Y: indicates sites that are validated (Methods), ◆mosaic status: S refers to the site that is shared between the siblings but the alternative allele could not be detected in either of the parent's blood, SP is the mosaic site that is shared between the siblings and excess of alt allele were detected in the father's blood, SM: is the mosaic site which is shared between the sibling and the excess alt allele were detected in the mother's blood, SP is the mosaic site that is shared between the siblings and excess of alt allele were detected in the father's blood, M represent the mosaic site which has excess of alt in the mother's blood but not shared between the siblings, P is the mosaic site that excess of alt was detected in the father's blood and it is not shared between the siblings. Family Id indicates which of the three families the mosaic site is from: 244, 603, 569. Haplotype: refers to the sites that their parental origin is known through the experimental analysis (Methods).

Coordinates (Chr:Position)	Mutations (Ref>Alt)	Miseq Mother (Ref/Alt)	Miseq Father (Ref/Alt)	Variant Specific Error	Adjusted p-value for Excess Alt		Mosaicism (%)	*PacBio Validation	◆Mosaic Status	Family ID	Haplotype
					Maternal	Paternal					
2:186300610	C:T>T:C	149/0	153/1	0.0008	1.00E+00	1.00E+00	N/A	N/A	S	603	N/A
2:193157646	T:G>G:T	426/1	443/0	0.0009	1.00E+00	1.00E+00	N/A	N/A	S	603	Paternal
9:2959572	G:T>T:G	378/3	699/6	0.0017	1.00E+00	9.84E-01	N/A	N/A	S	569	N/A
X:110276581	C:T>T:C	607/1	427/0	0.0017	1.00E+00	1.00E+00	N/A	N/A	S	569	N/A
5:109729461	C:A>A:C	38/0	37/3	0.0015	1.00E+00	2.32E-02	7.50%	Y	SP	569	N/A
1:230857935	G:C>C:G	252/13	366/1	0.0034	9.16E-09	1.00E+00	4.91%	U	SM	569	N/A
4:131248301	T:G>G:T	403/14	487/0	0.0004	2.06E-20	1.00E+00	3.36%	F	SM	569	Maternal
8:10261976	G:T>T:G	292/17	371/0	0.0026	1.43E-14	1.00E+00	5.50%	Y	SM	569	Maternal
8:92146874	C:G>C:G	546/22	792/2	0.0005	1.52E-31	1.00E+00	3.87%	Y	SM	569	Maternal
16:60784060	A:G>G:A	278/12	371/0	0.0007	6.32E-15	1.00E+00	4.14%	Y	SM	569	N/A
1:47735584	A:G>G:A	574/19	329/0	0.0007	1.94E-22	1.00E+00	3.20%	Y	M	244	Maternal
2:170651456	C:A>A:C	391/24	388/3	0.0028	1.07E-20	1.00E+00	5.78%	Y	M	603	Maternal
2:170651804	A:G>G:A	966/51	986/3	0.0013	3.55E-59	1.00E+00	5.01%	Y	M	603	Maternal
2:191908075	A:G>G:A	445/13	599/1	0.0010	2.38E-12	1.00E+00	2.84%	Y	M	569	Maternal
2:213698262	C:T>T:C	888/27	917/10	0.0020	1.35E-19	1.71E-02	2.95%	Y	M	603	Maternal
2:225499135	G:A>A:G	751/11	715/0	0.0003	2.77E-12	1.00E+00	1.44%	Y	M	603	Maternal
3:98029130	G:A>A:G	981/28	934/1	0.0015	6.39E-23	1.00E+00	2.78%	Y	M	603	Maternal
8:113112993	T:C>C:T	463/10	727/0	0.0004	9.81E-12	1.00E+00	2.11%	Y	M	569	N/A
13:75697455	C:T>T:C	1004/34	1011/3	0.0009	1.64E-37	1.00E+00	3.28%	Y	M	603	Paternal
16:65940897	A:G>G:A	1140/33	1168/0	0.0013	4.22E-30	1.00E+00	2.81%	Y	M	603	Maternal
X:17047163	G:A>A:G	216/3	119/3	0.0005	1.65E-01	1.00E+00	1.37%	Y	M	603	N/A

Coordinations (Chr:Position)	Mutations (Ref>Alt)	Miseq Mother (Ref/Alt)	Miseq Father (Ref/Alt)	Variant Specific Error	Adjusted p-value for Excess Alt		Mosaicism (%)	*PacBio Validation	◆Mosaic Status	Family ID	Haplotype
					Maternal	Paternal					
2:32093200	C:G>C:G	1033/1	1060/7	0.0002	1.00E+00	4.98E-06	0.66%	Y	P	603	N/A
2:37841931	A:T>T:A	721/0	596/68	0.0003	1.00E+00	2.29E-139	10.24%	F	P	603	N/A
3:133108055	A:G>G:A	860/0	869/7	0.0010	1.00E+00	2.67E-02	0.80%	U	P	603	Paternal
4:86375051	C:T>T:C	1004/1	938/6	0.0004	1.00E+00	1.87E-03	0.64%	U	P	603	N/A
5:146765532	C:T>T:C	1011/4	1041/6	0.0002	9.09E-02	2.17E-04	0.57%	U	P	603	N/A
9:126471014	T:G>G:T	920/2	903/6	0.0003	1.00E+00	1.16E-04	0.66%	U	P	603	Paternal
12:8090871	G:T>T:G	1138/5	1083/46	0.0004	1.27E-01	4.90E-70	4.07%	Y	P	603	Paternal
14:89561953	C:G>C:G	682/0	632/4	0.0002	1.00E+00	4.83E-03	0.63%	U	P	603	Paternal