Timing, rates and spectra of human germline mutation

Raheleh Rahbari[1,8], Arthur Wuster[1–3,8], Sarah J Lindsay[1], Robert J Hardwick[1], Ludmil B Alexandrov[1], Saeed Al Turki[1], Anna Dominiczak[4], Andrew Morris[5], David Porteous[6], Blair Smith[5], Michael R Stratton[1], UK10K Consortium[7] & Matthew E Hurles[1]

Germline mutations are a driving force behind genome evolution and genetic disease. We investigated genome-wide mutation rates and spectra in multi-sibling families. The mutation rate increased with paternal age in all families, but the number of additional mutations per year differed by more than twofold between families. Meta-analysis of 6,570 mutations showed that germline methylation influences mutation rates. In contrast to somatic mutations, we found remarkable consistency in germline mutation spectra between the sexes and at different paternal ages. In parental germ line, 3.8% of mutations were mosaic, resulting in 1.3% of mutations being shared by 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.

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]. On the basis of whole-genome sequencing studies of family trios, the average generational mutation rate for single-base substitutions in humans has been estimated to be ~1–1.5 × 10^{-8} (refs. 5–9).

In 1947, J.B.S. Haldane noted that the mutation rate of the hemophilia-associated gene is significantly higher in men than in women[10]. Recent genome sequencing studies have confirmed Haldane's observation that the male germ line is more mutagenic[5–8,11]. On average, each additional year in 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 germ line[14]. Whereas 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 germ line has experienced 160 genome replications in a 20-year-old male, with the number rising to 610 genome replications in a 40-year-old male[15].

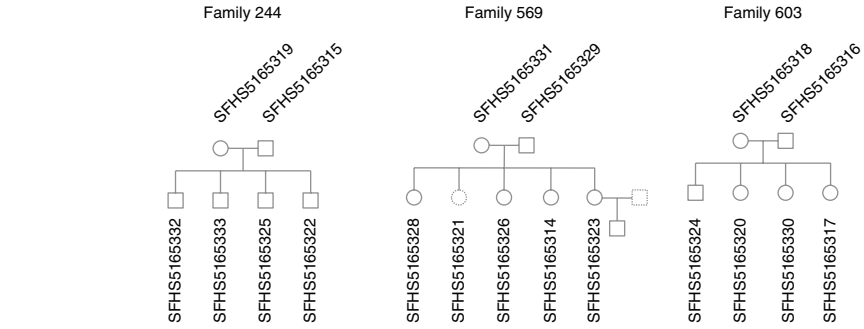
Mutation rate depends on local nucleotide context. Moreover, studies of somatic mutations in cancer have shown that observed mutation spectra can be decomposed into different 'mutational signatures' that reflect particular cellular contexts of exogenous and endogenous mutagen exposure and the efficiency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

The germ line comprises a lineage of different cellular contexts, from the zygote to the gamete[17] (**Supplementary Fig. 1**). Postzygotic mutations can potentially lead to germline mosaicism. Observing apparent DNMs shared by siblings—predominantly in studies of dominant disorders—has provided direct evidence for germline mosaicism[18]. Although recent studies have determined the average germline mutation rate and estimated the average effect of paternal age, a deeper understanding of germline mutational rates and spectra and the underlying mutational processes remains elusive. For example, it is not known whether mutation spectra differ between paternal and maternal germ lines, whether mutation rates and spectra vary significantly between families, or whether different stages of the cellular lineage from the zygote to the gamete differ in their mutation rates and spectra.

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

[1]Wellcome Trust Sanger Institute, Hinxton, UK. [2]Department of Human Genetics, Genentech, Inc., South San Francisco, California, USA. [3]Department of Bioinformatics and Computational Biology, Genentech, Inc., South San Francisco, California, USA. [4]Institute of Cardiovascular and Medical Sciences, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK. [5]Medical Research Institute, University of Dundee, Dundee, UK. [6]Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [7]A list of members and affiliations appears in the **Supplementary Note**. [8]These authors contributed equally to this work. Correspondence should be addressed to M.E.H. (meh@sanger.ac.uk).

Received 4 June; accepted 20 November; published online 14 December 2015; doi:10.1038/ng.3469



Age (years)	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 the sequenced families. Identifiers and relationship between individuals are shown for the three families in this study. Individuals who were sequenced are represented by circles or squares with solid outlines; other individuals are represented by circles or squares with dotted outlines. The ages of the mother and father at conception of each child and phasing information are summarized in the table. SFHS5165321 was only used for the part of the analysis related to mosaicism.

RESULTS

Family-specific paternal age effects

We sequenced the genomes of three multi-sibling families (**Fig. 1**). We discovered and validated 768 DNMs across the three families, with an average of 64 DNMs per child (range of 43–84; **Supplementary Table 1**). When taking into account genomic regions inaccessible to our analyses (Online Methods), the average number of mutations per individual increased to 76.9. This adjusted number of mutations is equivalent to an average mutation rate of 1.28×10^{-8} (95% confidence interval (CI) = $1.13–1.43 \times 10^{-8}$) mutations per nucleotide per generation 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 (**Fig. 1**). Our data confirm the effect of paternal age. Taking all families together, the number of DNMs increased with father's age by 2.87 mutations per year (95% confidence interval = 2.11–3.64). In all three families, there was a 12- to 13-year age 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 was considered separately (**Fig. 2**). The paternal age effect for families 244, 603 and 569 was 1.46 (95% CI = 1.15–1.78), 3.27 (95% CI = 2.07–4.47) and 3.65 (95% CI = 1.52–5.77) mutations per year, respectively. Overall, a model that takes both paternal age and family into account performed 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).

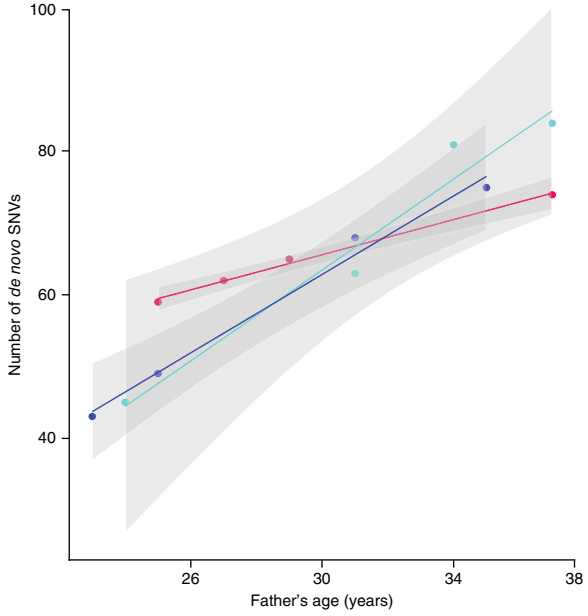


Figure 2 Paternal age versus number of *de novo* mutations. The number of DNMs has been corrected to take into account genomic regions inaccessible to our methods. Red, family 244; light blue, family 569; dark blue, family 603. Gray areas denote the region covered by the 95% confidence interval of the 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.

Germline mosaicism in parents

Mutations that occur during early development can lead to mosaicism in the germ line 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, deeply sequencing every validated DNM in every individual in all three pedigrees to a mean depth of 567× per individual (Online Methods).

First, we identified ten validated DNMs shared by at least two siblings in the same family that are clearly not constitutively heterozygous in either parent (alternative allele fraction <10%). On the basis of these criteria, the probability of any germline mutation being shared by two siblings is 1.3% (**Supplementary Table 2**).

Second, by identifying sites with a significant excess of reads for the alternative allele in DNA from a single parent (Online Methods), we distinguished sites among the validated DNMs that were potentially mosaic at low levels in parental blood (**Fig. 3b**, **Table 1** and **Supplementary Fig. 2**). This approach identifies germline mutations mosaic in at least one parental somatic tissue that thus most likely occurred during early embryonic development of the parent, before the separation and proliferation of the germ line and soma, and consequently are mosaic in both tissues. We attempted further experimental validation of the candidate mosaic sites using orthogonal amplification and sequencing technologies (Online Methods). Taking these independent experiments together, we identified 25 DNMs with excess parental alternative reads, constituting from 0.6% to 10% of the total reads for the site, with a median of 3%. We modeled our statistical power to detect parental somatic mosaicism (**Fig. 3a**) and conclude

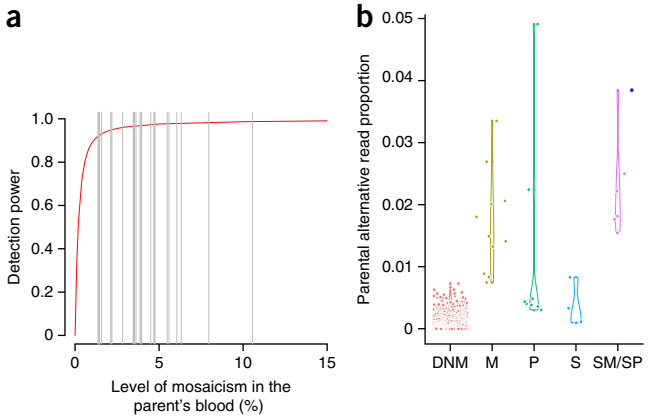


Figure 3 Detection of mutations mosaic in parents. (**a**) Simulation of detection power for a range of mosaicism levels in blood from parents using MiSeq depth of coverage for all DNMs ($n = 768$). For a mean validation coverage (MiSeq platform) of 567× in parents, we have >0.94 power to detect mosaicism at a rate of 2% and higher in blood from parents. (**b**) Comparison of parental alternative read ratios between DNMs and germline mosaic sites. M, mosaic sites with a significant excess of alternative reads in the mother's blood; P, mosaic sites with a significant excess of alternative reads in the father's blood; S, sites that are shared by the siblings but for which an excess of alternative reads could not be detected in blood from either parent; SM and SP, mosaic sites shared by the siblings for which a significant excess of alternative reads was detected in the mother's blood (SM; pink dots) or father's blood (SP; dark blue dot).

that we have ~80% power to detect a mosaic variant present in 1% of parental blood cells and ~90% power to detect a variant present in 2% of parental blood cells.

Six of the ten DNMs shared by siblings also exhibited parental somatic mosaicism, which is a significant enrichment in comparison to mutations observed in a single sibling ($P = 4.6 \times 10^{-7}$, Fisher's exact test). Four DNMs were shared by siblings without excess alternative reads in parental blood. Hence, these mutations either occurred after the separation of the germ line and soma or correspond to parental somatic mosaicism below detectable levels. In total, 29 of the validated DNMs had evidence of parental germline mosaicism (**Table 1**). Correcting for our incomplete power to detect mosaic mutations (**Fig. 3a**) suggests that 4.2% of germline mutations may be mosaic in >1% of parental blood cells (Online Methods).

Of the parental mosaic DNMs, 64% (16/25) 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 origin than the paternal bias observed for all 768 DNMs ($P = 7.7 \times 10^{-6}$, binomial test). This ratio is not likely to be due

Table 1 Germline mosaic single-nucleotide variants

Coordinates (chr.:position)	Mutation (ref>alt)	MiSeq mother (ref/alt)	MiSeq father (ref/alt)	Variant-specific error	Adjusted *P* value for excess alternative reads		Mosaicism (%)	PacBio validation[a]	Mosaic status[b]	Family ID[c]	Haplotype[d]
					Maternal	Paternal					
2:186,300,610	C:T>T:C	149/0	153/1	0.0008	1.00	1.00	NA	NA	S	603	NA
2:193,157,646	T:G>G:T	426/1	443/0	0.0009	1.00	1.00	NA	NA	S	603	Paternal
9:2,959,572	G:T>T:G	378/3	699/6	0.0017	1.00	9.84×10^{-1}	NA	NA	S	569	NA
X:110,276,581	C:T>T:C	607/1	427/0	0.0017	1.00	1.00	NA	NA	S	569	NA
5:109,729,461	C:A>A:C	38/0	37/3	0.0015	1.00	2.32×10^{-2}	7.50	Y	SP	569	NA
1:230,857,935	G:C>C:G	252/13	366/1	0.0034	9.16×10^{-9}	1.00	4.91	U	SM	569	NA
4:131,248,301	T:G>G:T	403/14	487/0	0.0004	2.06×10^{-20}	1.00	3.36	F	SM	569	Maternal
8:10,261,976	G:T>T:G	292/17	371/0	0.0026	1.43×10^{-14}	1.00	5.50	Y	SM	569	Maternal
8:92,146,874	C:G>C:G	546/22	792/2	0.0005	1.52×10^{-31}	1.00	3.87	Y	SM	569	NA
16:60,784,060	A:G>G:A	278/12	371/0	0.0007	6.32×10^{-15}	1.00	4.14	Y	SM	569	NA
1:47,735,584	A:G>G:A	574/19	329/0	0.0007	1.94×10^{-22}	1.00	3.20	Y	M	244	Maternal
2:170,651,456	C:A>A:C	391/24	388/3	0.0028	1.07×10^{-20}	1.00	5.78	Y	M	603	Maternal
2:170,651,804	A:G>G:A	966/51	986/3	0.0013	3.55×10^{-59}	1.00	5.01	Y	M	603	Maternal
2:191,908,075	A:G>G:A	445/13	599/1	0.0010	2.38×10^{-12}	1.00	2.84	Y	M	569	Maternal
2:213,698,262	C:T>T:C	888/27	917/10	0.0020	1.35×10^{-19}	1.71×10^{-2}	2.95	Y	M	603	Maternal
2:225,499,135	G:A>A:G	751/11	715/0	0.0003	2.77×10^{-12}	1.00	1.44	Y	M	603	Maternal
3:98,029,130	G:A>A:G	981/28	934/1	0.0015	6.39×10^{-23}	1.00	2.78	Y	M	603	Maternal
8:113,112,993	T:C>C:T	463/10	727/0	0.0004	9.81×10^{-12}	1.00	2.11	Y	M	569	NA
13:75,697,455	C:T>T:C	1,004/34	1,011/3	0.0009	1.64×10^{-37}	1.00	3.28	Y	M	603	Paternal
16:65,940,897	A:G>G:A	1,140/33	1,168/0	0.0013	4.22×10^{-30}	1.00	2.81	Y	M	603	Maternal
X:17,047,163	G:A>A:G	216/3	119/3	0.0005	1.65×10^{-1}	1.00	1.37	Y	M	603	NA
2:32,093,200	C:G>G:C	1,033/1	1,060/7	0.0002	1.00	4.98×10^{-6}	0.66	Y	P	603	NA
2:37,841,931	A:T>T:A	721/0	596/68	0.0003	1.00	2.29×10^{-139}	10.24	F	P	603	NA
3:133,108,055	A:G>G:A	860/0	869/7	0.0010	1.00	2.67×10^{-2}	0.80	U	P	603	Paternal
4:86,375,051	C:T>T:C	1,004/1	938/6	0.0004	1.00	1.87×10^{-3}	0.64	U	P	603	NA
5:146,765,532	C:T>T:C	1,011/4	1041/6	0.0002	9.09×10^{-2}	2.17×10^{-4}	0.57	U	P	603	NA
9:126,471,014	T:G>G:T	920/2	903/6	0.0003	1.00	1.16×10^{-4}	0.66	U	P	603	Paternal
12:8,090,871	G:T>T:G	1,138/5	1,083/46	0.0004	1.27×10^{-1}	4.90×10^{-70}	4.07	Y	P	603	Paternal
14:89,561,953	C:G>G:C	682/0	632/4	0.0002	1.00	4.83×10^{-3}	0.63	U	P	603	Paternal

Chr., chromosome; ref, reference allele; alt, alternative allele.
[a]NA, not applicable, F, sites not validated; U, uncertain sites; Y, sites validated (Online Methods). [b]S, site shared by the siblings but for which the alternative allele could not be detected in blood from either parent; SM, mosaic site shared by the siblings for which excess alternative reads were detected in the mother's blood; SP, mosaic site shared by the siblings for which excess alternative reads were detected in the father's blood; M, mosaic site with excess alternative reads in the mother's blood but not shared by the siblings; P, mosaic site with excess alternative reads in the father's blood but not shared by the siblings. [c]Family ID indicates the family origin of each of the mosaic sites: family 244, 603 or 569. [d]Sites for which the parental origin is known through the experimental analysis (Online Methods).

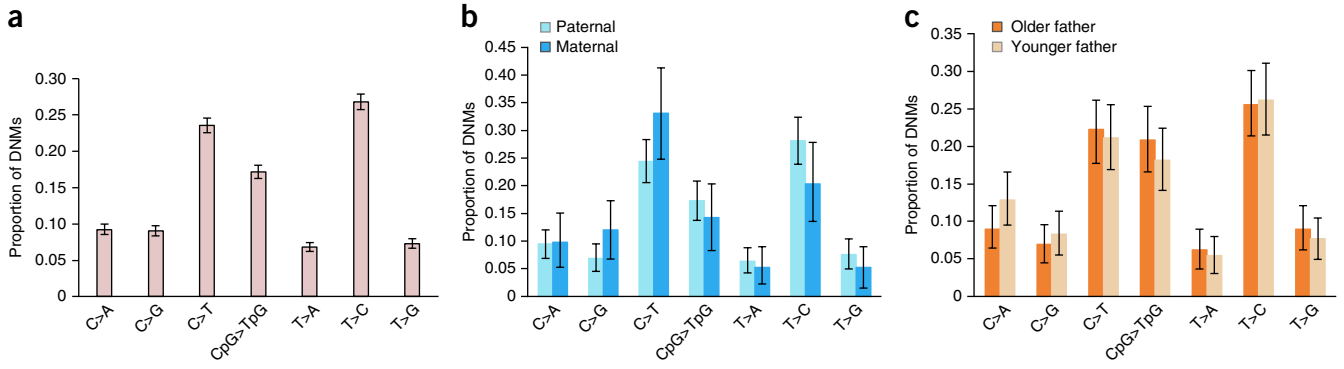


Figure 4 Mutational spectra. (**a**) Frequency of all mutation types in the catalog of 6,570 high-confidence DNMs. (**b**) Difference in the frequency of maternal and paternal mutations for the subset of DNMs with phasing information ($n = 556$). (**c**) Difference in the frequency of mutations of children from fathers younger and older than 30 years ($n = 680$). Error bars, 95% confidence intervals.

to differential sequencing coverage for mothers and fathers (**Supplementary Fig. 3**).

Germline mutational spectra

We compiled a catalog 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 catalog to evaluate evidence for distinct germline mutational processes. Low-resolution mutational spectra, which we define as the relative frequencies of the six possible point mutations, confirmed the expected preponderance of transitions over transversions (**Fig. 4a**). There was no significant difference between the spectra of maternal and paternal mutations ($P = 0.19$, χ^2 test; **Fig. 4b**). Even though there was a significant difference in the magnitude of the paternal age effect between the three families, there was no significant difference between their mutational spectra ($P = 0.925$, χ^2 test) nor between the spectra of DNMs for children born to younger and older fathers ($P = 0.83$, χ^2 test; **Fig. 4c**).

As an independent assessment of potential differences in maternal and paternal mutation spectra, we contrasted variants identified on the X and Y chromosomes in a genome-wide sequencing data set based on 2,453 individuals from the UK10K project. All variation on the Y chromosome arose in the male germ line, whereas variation on the X chromosome was generated in both the maternal and paternal germ lines. We observed that only rare variants faithfully recapitulated the mutation spectra observed for DNMs[19], as the ratio of C:G>T:A and T:A>C:G transitions decreased dramatically with increasing derived allele frequency, most likely because of biased gene conversion[20] (**Supplementary Fig. 4**). We did not observe any statistically significant difference ($P = 0.10$, χ^2 test) in the X-chromosome and Y-chromosome mutation spectra (number of variants = 3,217) after accounting for differences in base composition between the chromosomes (Online Methods and **Supplementary Fig. 5**). This confirms our observation above that, despite differences in mutation rates, numbers of genome divisions and cellular contexts, the mutation spectra in the maternal and paternal germ lines 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 characterized higher-resolution mutational spectra. For this analysis, we calculated the relative frequency of mutations at the 96 triplets defined by the mutated base and the bases flanking it on each side (**Fig. 5a**). The spectrum observed for germline mutations clearly recapitulated the known higher mutability of CpG dinucleotides.

We evaluated whether any combination of the 30 previously identified signatures[16] was sufficient to explain the observed pattern of

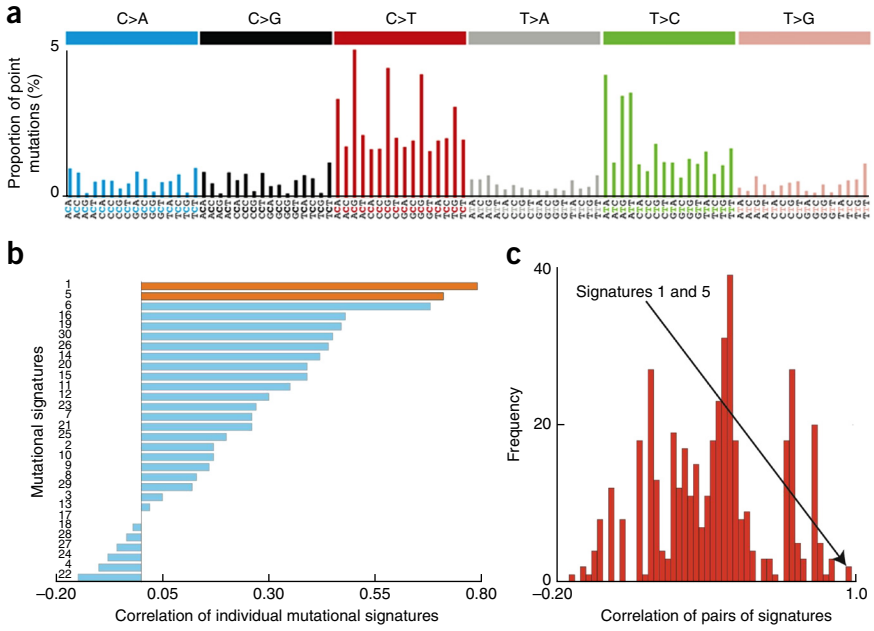


Figure 5 Mutational spectrum and signatures. (**a**) High-resolution mutational spectrum of DNMs. Each of the six possible point mutations is subdivided into 16 subclasses on the basis of the 3′ and 5′ nucleotides flanking the mutation. We note that C:G>T:A and T:A>C:G transitions are more common. Within these categories, CpG site mutations are particularly frequent. (**b**) Correlation of mutational signatures with observed mutations in the mutational catalog. Correlation is shown for each of the 30 signatures, with signatures 1 and 5 highlighted in orange. (**c**) Combination of all possible pairs of signatures; the combination of signatures 1 and 5 is indicated with an arrow.

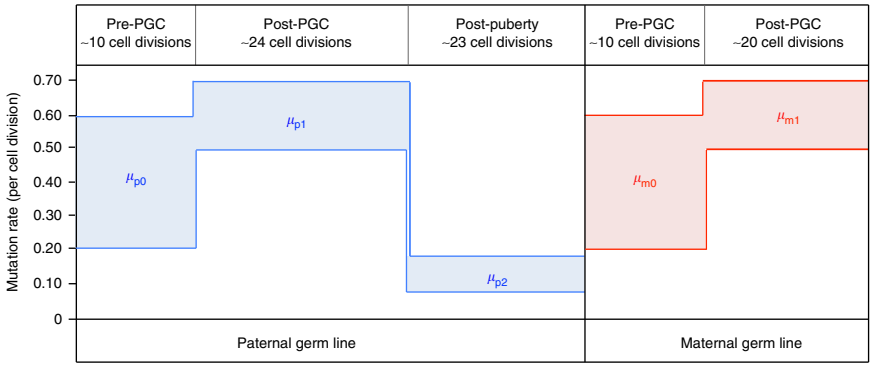


Figure 6 Mutation rate model during gametogenesis. Comparison of mutation rates between spermatogenesis (blue box) and oogenesis (red box). μ_p and μ_m are the mutation rates in paternal and maternal genomes, respectively, and the mutation rate for each stage of gametogenesis is denoted by number. Gametogenesis 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 germ lines) with ~0.2–0.6 mutations per haploid genome per cell division; this rate is similar in both maternal and paternal gametogenesis; stage 2: post-PGC specification; in the maternal germ line there are ~20 cell divisions and in the paternal germ line there are ~24 cell divisions post-PGC specification and up to puberty; the 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 germ line) 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.

germline mutations (**Fig. 5b**). Two of the mutational signatures, previously termed signature 1 (25% of DNMs) and signature 5 (75% of DNMs), explained the majority of the observed mutational patterns (Pearson correlation = 0.98; **Fig. 5c**). Including any additional mutational signatures did not significantly improve this correlation. Signature 1 is characterized by C:G>T:A mutations at CpG dinucleotides, whereas signature 5 is predominately characterized by T:A>C:G mutations (**Supplementary Fig. 6**). These signatures are responsible for generation of the majority of spontaneous preneoplastic somatic mutations[16], indicating that the mutational processes underlying these signatures in somatic cells are also operative in the germ line.

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 germ line 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 Encyclopedia of DNA Elements (ENCODE) Project[22]. In the testis cell line, 25.3% of CpG sites had more than 50% of reads methylated (**Supplementary Table 4**). Thirteen of these sites overlapped with DNMs from our catalog, of which 12 had 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 ($P = 1.71 \times 10^{-8}$, binomial test). All 12 of the DNMs that were methylated in the testis cell line were CpG>TpG mutations (**Supplementary Table 5**). For B-lymphocyte and embryonic stem cell lines, the association between methylation status and mutation was less significant ($P = 0.04$ and 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 resequencing. Both the average genome-wide mutation rate of 1.28×10^{-8} mutations per nucleotide per generation 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 mutations with each additional year in the parents' ages. The magnitude of this effect differed by a factor of greater than two between families. Although our observations corroborate a previous study[6] that proposed that the major factor influencing the number of mutations in a child is paternal age, our multi-sibling study design allows detection of more subtle differences between families. Given that the increase in the number of mutations with parental age is driven by paternal mutations, we suggest that this observation could result from variation among 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]. Although this variation could be due to study design or analysis choices, our results highlight a more interesting possibility, namely that the paternal age effect actually differed between the studies because of the families included, with most of the studies having a limited sample size.

We observed no difference in the mutation spectra for the maternal and paternal germ lines or with younger and older fathers. The lack of large differences in mutation spectra between the sexes is perhaps counterintuitive given the different cellular contexts in the maternal and paternal germ lines, including the marked difference in the number of cell divisions and thus the increased potential for replication-associated mutations in the paternal germ line. Larger catalogs 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 is 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 germ line 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 demonstrates that DNA methylation and perhaps other cell type–specific factors have a finer-grained role in influencing the precise location of mutations in specific cell types.

With regard to the timing of mutations in the cellular lineage of the germ line, 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 mosaicism at very low levels (<1%)

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 (within the first 8–12 cell divisions[25,26]), before the specification of primordial germ cells (PGCs) and the concomitant separation of the germ line 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 origin, as might be expected given the occurrence of these mutations before sexual differentiation of the embryo.

We note that our observations seem incompatible with monophyletic origins for the blood and germ line; instead, each tissue is likely to be founded by multiple cells with polyphyletic ancestry. A logical consequence is that some mesoderm founder cells are more closely related to PGCs 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 postzygotic divisions may well be present at such high frequencies within parental tissues that our analytical workflow for identifying candidate DNMs fails to identify 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 the levels of somatic mosaicism that we could detect suggests that we were able to detect pre-PGC mutations across at least four 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 before PGC specification, we can estimate a credible range for the mutation rate in early cell divisions in parental germ lines. On the basis of sharing of pre-PGC mutations by gametes from the same parent, we can define a maximum and minimum number of pre-PGC cell divisions within which the observed pre-PGC mutations must have occurred, and from these estimates, 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 germ lines. The paternal age effect that we observed implies that a lower mutation rate per cell division, ranging from ~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 considerably more mutagenic than post-pubertal spermatogenesis, with a mutation rate per cell division of ~0.5 to ~0.7 (with ~10–14 maternal mutations arising during ~20 post-PGC cell divisions[27]). In the paternal germ line, we also need to consider an intermediate phase of cell division, during the proliferation and differentiation of PGCs to form prespermatogonia 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 arise during this intermediate phase to be ~13–17. It has been estimated that there are ~24 cell divisions during this phase[27], giving a mutation rate range 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 (**Fig. 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 is reduced by ~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 germ line 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 single-nucleotide variants (SNVs), the main step change in mutation rate per cell division may be between the embryonic and post-pubertal phases of gametogenesis in males, and a similar observation has been reported in mouse 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 and minimizing the deleterious mutation rate.

It is important to note that the estimated ranges for the mutation rates per cell division presented above represent a combination of mutations that arise during genome replication and any spontaneous mutations occurring between cell divisions. The time interval between cell divisions differs markedly throughout the different phases of gametogenesis, and these mutation rate estimates therefore 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 germ line; indeed, we observed that the six parental somatic mosaic DNMs that were present in more than one child had significantly higher levels of somatic mosaicism, on average, than the other parental somatic mosaic DNMs that were not present in more than one child ($P = 0.009$, 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 children.

We identified four DNMs that were shared by siblings and thus are highly likely to be mosaic in the parental germ line, 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 germ line.

Previous studies of the germline mosaicism of sequence variants have been largely limited to case studies of sibling recurrence of pathogenic DNMs[30–34]. Our estimate of 1.3% for 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 germ line[35]. We note that these recurrent DNMs shared by siblings were not randomly distributed across families but were significantly ($P < 0.01$) enriched in one pedigree. This suggests that there may also be significant variation across families in patterns of germline mosaicism of DNMs.

These results on germline mosaicism have implications for genetic counseling on recurrence risks for families with children with genetic disorders caused by DNMs[17]. Although 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. Whereas only 1.3% of all DNMs were observed recurrently among siblings, this proportion increased to 24% for DNMs that were mosaic in >1% of parental blood cells and 50% for DNMs that were 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 a 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 are always limited by the number of offspring available. Deep sequencing of single gametes from different individuals[36] should enable us to characterize and compare their mutation rates and spectra at much higher resolution. This will also mitigate any biases associated with the selection that might occur 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 germ line would extend the currently limited understanding of the selective pressures operative throughout the genealogy of the germ line.

URLs. UK10K, http://www.uk10k.org/; Signatures of Mutational Processes in Human Cancer, http://cancer.sanger.ac.uk/cosmic/signatures; Generation Scotland, http://www.generationscotland.org/; UCSC Lift Genome Annotations, https://genome.ucsc.edu/cgi-bin/hgLiftOver; Avon Longitudinal Study of Parents and Children (ALSPAC), http://www.bristol.ac.uk/alspac/; TwinsUK, http://www.twinsuk.ac.uk/; Ensembl, Comparative Genomics, http://www.ensembl.org/info/genome/compara/index.html; UCSC ENCODE composite track, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibMethylRrbs/.

METHODS

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

Accession codes. Whole-genome sequencing data are accessible via the European Genome-phenome Archive (EGA) under accession EGAD00001001214.

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

ACKNOWLEDGMENTS

We thank D. Conrad and A. Ramu for their responsive development of the DeNovoGear software and A. Campbell and S. Kerr for their support in identifying relevant families. This research was funded by the Wellcome Trust (grant WT098051). Generation Scotland has received core funding from the Chief Scientist Office of the Scottish Government Health Directorates CZD/16/6 and the Scottish Funding Council HR03006. This study makes use of data generated by the UK10K Consortium, derived from samples from ALSPAC and TwinsUK. A full list of the investigators who contributed to the generation of the data is available from http://www.uk10k.org. Funding for UK10K was provided by the Wellcome Trust under award WT091310. Data can be accessed at the European Genome-phenome Archive (EGA) under accessions EGAS00001000108 and EGAS00001000090.

AUTHOR CONTRIBUTIONS

R.R., A.W. and M.E.H. developed analytical methods and/or analyzed sequencing data. R.R. performed mutation rate estimation, family comparison, analysis of germline mosaicism and validation. A.W. performed meta-analysis of the DNMs for mutational spectrum and methylation status. S.J.L. and R.J.H. contributed toward phasing and the detection and validation of DNMs. L.B.A. performed mutational signature analysis. S.A.T. contributed to whole-genome data analysis. A.D., A.M., D.P. and B.S. provided blood samples for the Scottish Family Health Study. M.R.S. advised on mutational processes. The UK10K Consortium contributed sequences for meta-data analysis. R.R., A.W. and M.E.H. wrote the manuscript. M.E.H. supervised the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Lindahl, T. & Wood, R.D. Quality control by DNA repair. *Science* **286**, 1897–1905 (1999).
2. Hoeijmakers, J.H. Genome maintenance mechanisms for preventing cancer. *Nature* **411**, 366–374 (2001).
3. MacArthur, D.G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).
4. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.* **13**, 745–753 (2012).
5. Michaelson, J.J. *et al.* Whole-genome sequencing in autism identifies hot spots for *de novo* germline mutation. *Cell* **151**, 1431–1442 (2012).
6. Kong, A. *et al.* Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
7. Conrad, D.F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* **43**, 712–714 (2011).
8. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
9. Campbell, C.D. & Eichler, E.E. Properties and rates of germline mutations in humans. *Trends Genet.* **29**, 575–584 (2013).
10. Haldane, J.B. The mutation rate of the gene for haemophilia, and its segregation ratios in males and females. *Ann. Eugen.* **13**, 262–271 (1947).
11. Venn, O. *et al.* Nonhuman genetics. Strong male bias drives germline mutation in chimpanzees. *Science* **344**, 1272–1275 (2014).
12. Momand, J.R., Xu, G. & Walter, C.A. The paternal age effect: a multifaceted phenomenon. *Biol. Reprod.* **88**, 108 (2013).
13. Goriely, A. & Wilkie, A.O. Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. *Am. J. Hum. Genet.* **90**, 175–200 (2012).
14. Crow, J.F. The origins, patterns and implications of human spontaneous mutation. *Nat. Rev. Genet.* **1**, 40–47 (2000).
15. Wilson Sayres, M.A. & Makova, K.D. Genome analyses substantiate male mutation bias in many species. *BioEssays* **33**, 938–945 (2011).
16. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
17. Lupski, J.R. Genetics. Genome mosaicism—one human, multiple genomes. *Science* **341**, 358–359 (2013).
18. Biesecker, L.G. & Spinner, N.B. A genomic view of mosaicism and human disease. *Nat. Rev. Genet.* **14**, 307–320 (2013).
19. Schaibley, V.M. *et al.* The influence of genomic context on mutation patterns in the human genome inferred from rare variants. *Genome Res.* **23**, 1974–1984 (2013).
20. Duret, L. & Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genomics Hum. Genet.* **10**, 285–311 (2009).
21. Cooper, D.N. & Krawczak, M. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. *Hum. Genet.* **83**, 181–188 (1989).
22. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
23. Ségurel, L., Wyman, M.J. & Przeworski, M. Determinants of mutation rate variation in the human germline. *Annu. Rev. Genomics Hum. Genet.* **15**, 47–70 (2014).
24. 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).
25. O'Rahilly, R., Müller, F. & Streeter, G.L. *Developmental Stages in Human Embryos: Including a Revision of Streeter's "Horizons" and a Survey of the Carnegie Collection* (Carnegie Institution of Washington, 1987).
26. Coticchio, G., Albertini, D.F. & De Santis, L. *Oogenesis* (Springer Verlag, 2013).
27. Drost, J.B. & Lee, W.R. Biological basis of germline mutation: comparisons of spontaneous germline mutation rates among *Drosophila*, mouse, and human. *Environ. Mol. Mutagen.* **25** (suppl. 26), 48–64 (1995).
28. Voet, T., Vanneste, E. & Vermeesch, J.R. The human cleavage stage embryo is a cradle of chromosomal rearrangements. *Cytogenet. Genome Res.* **133**, 160–168 (2011).

29. Walter, C.A., Intano, G.W., McCarrey, J.R., McMahan, C.A. & Walter, R.B. Mutation frequency declines during spermatogenesis in young mice but increases in old mice. *Proc. Natl. Acad. Sci. USA* **95**, 10015–10019 (1998).
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.* **20**, 15–23 (2014).
31. Anazi, S., Al-Sabban, E. & Alkuraya, F.S. Gonadal mosaicism as a rare cause of autosomal recessive inheritance. *Clin. Genet.* **85**, 278–281 (2014).
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.* **164A**, 2356–2359 (2014).
33. Tajir, M. *et al.* Germline mosaicism in Rubinstein-Taybi syndrome. *Gene* **518**, 476–478 (2013).
34. Bachetti, T. *et al.* Recurrence of CCHS associated *PHOX2B* poly-alanine expansion mutation due to maternal mosaicism. *Pediatr. Pulmonol.* **49**, E45–E47 (2014).
35. 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).
36. Wang, J., Fan, H.C., Behr, B. & Quake, S.R. Genome-wide single-cell analysis of recombination activity and *de novo* mutation rates in human sperm. *Cell* **150**, 402–412 (2012).

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, and the study was approved by the National Health Service (NHS) East of Scotland Research Ethics Service REC 1 (reference 15/ES/0040). The families were selected on the basis of genomic DNA quality, number of children and age gap between the oldest and youngest siblings.

***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 probands was also sequenced. We used the DeNovoGear software[37] to identify 49,893 candidate DNMs in the children. We identified likely false positives as those sites that overlapped low-complexity regions[38], 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 λ parameter of the Poisson distribution 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 sequencing to 139× coverage on average (range of 88–191×). We designed baits to cover a 200-bp window around each candidate site. The bait design succeeded for 4,141 sites. To analyze 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 after manual inspection in the Integrative Genomics Viewer (IGV) genome browser.

Model 1: germline DNM. We defined the likelihood of the data under the DNM model as

$$\mathrm{LL.DNM} = \mathrm{Pois}(m_\mathrm{m}, m_\mathrm{T} \times e) + \mathrm{Pois}(d_\mathrm{m}, d_\mathrm{T} \times e) + \mathrm{Bin}(c_\mathrm{m}, c_\mathrm{T} \times e, 0.5)$$

where m_m, d_m and c_m are the number of reads supporting the mutant allele (mostly the alternative allele) in the mother, father and child, respectively. m_T, d_T and c_T are the total number of reads in the mother, father and child, respectively, and e is the sequencing error rate.

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

$$\mathrm{LL.I} = \max(\mathrm{LL.IFM}, \mathrm{LL.IFD}, \mathrm{LL.IFMD})$$

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

$$\mathrm{LL.IFM} = \mathrm{Bin}(m_\mathrm{m}, m_\mathrm{T}, 0.5) + \mathrm{Pois}(d_\mathrm{m}, d_\mathrm{T} \times e) + \mathrm{Bin}(c_\mathrm{m}, c_\mathrm{T}, 0.5)$$

$$\mathrm{LL.IFD} = \mathrm{Pois}(m_\mathrm{m}, m_\mathrm{T} \times e) + \mathrm{Bin}(d_\mathrm{m}, d_\mathrm{T}, 0.5) + \mathrm{Bin}(c_\mathrm{m}, c_\mathrm{T}, 0.5)$$

$$\mathrm{LL.IFMD} = \mathrm{Bin}(m_\mathrm{m}, m_\mathrm{T}, 0.5) + \mathrm{Bin}(d_\mathrm{m}, d_\mathrm{T}, 0.5) + \mathrm{Bin}(c_\mathrm{m}, c_\mathrm{T}, 0.5)$$

Model 3: false positive. Model 3 is written as

$$\mathrm{LL.FP} = \mathrm{Pois}(m_\mathrm{m}, m_\mathrm{T} \times e) + \mathrm{Pois}(d_\mathrm{m}, d_\mathrm{T} \times e) + \mathrm{Pois}(c_\mathrm{m}, c_\mathrm{T} \times e)$$

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 depends on the sex of the proband

$$\mathrm{Girls}: (1 - \mathrm{noCvg}) \times (1 - \mathrm{filtered}) \times (1 - \mathrm{noVal} \times \mathrm{ppAdjust}) \times 2 \times \mathrm{valDNM/genome\,length}$$

$$\mathrm{Boys}: (1 - \mathrm{noCvg}) \times (1 - \mathrm{filtered}) \times (1 - \mathrm{noVal} \times \mathrm{ppAdjust}) \times 2 \times \mathrm{valDNM} + \mathrm{valDNMX/genome\,length}$$

where noCvg is the proportion of the genome that is either N's or not covered by 7× or more, filtered is the proportion of the genome that is a segmental duplication or a simple repeat (but not N's or low-coverage sequence), noVal is the proportion of the genome that passed filtering but for which validation was not possible (mainly because of failed primer design), ppAdjust is the proportion of calls that could not be validated that are likely to be true positives on the basis of their posterior probabilities as calculated by DeNovoGear, valDNM is the number of validated DNMs, valDNMX is the number of validated DNMs on the X chromosome, and genome length is the length of 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, identifying DNMs shared by 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 for 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 pulldown library we previously used for validation, which resulted in an average coverage of 500× 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 alternative allele reads resulted from sequencing error was calculated as follows

$$p_\mathrm{maternal} = \mathrm{Bin}(m_\mathrm{m}, m_\mathrm{alt+ref}, e)$$

$$p_\mathrm{paternal} = \mathrm{Bin}(f_\mathrm{alt}, f_\mathrm{alt+ref}, e)$$

where m and f are the number of reads in the mother and father, respectively, alt and ref are the alternative 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 Bonferroni correction. Sites that were significant at adjusted $P < 0.05$ were considered to be 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 present in more than one sibling in the same family was calculated as the number of instances of a mutation being shared by 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 (**Supplementary Data Set 1**) using PacBio amplicon sequencing. These 40 candidate mosaic DNMs were selected as follows: ten DNMs that were shared by siblings (for six of these shared DNMs, we had previously identified a significant parental excess of reads for alternative alleles, as described above) and 30 candidate mosaic sites that had an excess of reads for alternative alleles in a parent's blood, with nominal P value < 0.05. Note that the 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 Primer3 (ref. 39) 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

doi:10.1038/ng.3469

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 (chr. 2: 37,841,931 and chr. 4: 131,248,301) 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 cells.

After PacBio sequencing, filtered subreads and ROI (reads of insert) were generated using SMRTAnalysis (provided by Pacific Biosciences). The resulting fastq files were demultiplexed on the basis of the 11-bp unique barcodes for each individual and mapped to the human reference genome GRCh37 (hg19). The average sequence coverage from the PacBio data was 158× across the 114 amplicons. Lastly, variants were called from the resulting BAM alignments using SAMtools[40] mpileup, version 1.1. Each of the candidate parental mosaic DNMs was only further analyzed if we observed ~50% reference/alternative reads in the child, and then the parental alternative/reference reads were counted. We categorized sites meeting this criterion into the following classes: (i) validated, comprising sites where we observed alternative alleles in the relevant parent; (ii) 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 (iii) not validated, comprising sites where we had >90% power to detect the alternative alleles in the mosaic parents but did not detect them.

We classified 29 of the set of 40 candidate sites as parentally mosaic. Four mosaic DNMs were shared by siblings from the same family, but we could not observe alternative alleles in either parent in either validation data set (MiSeq or PacBio). Sixteen sites were validated as mosaic, with the mosaic parent confirmed on both platforms (all of these sites had a significant *P* value for the MiSeq data after Bonferroni correction). One additional site with a nominally significant *P* value but that was not significant when considering the adjusted *P* value for the MiSeq data was confirmed to be mosaic in PacBio data. Two sites were confirmed to be mosaic on the basis of significant adjusted *P* values from the MiSeq data only, as they failed the PacBio experiment. Six sites were confirmed to be mosaic on the basis of MiSeq data only (with significant adjusted *P* values), as their mosaicism level was below detection power in PacBio analysis. For the remaining 11 sites, despite their having nominally significant *P* values, the adjusted MiSeq *P* values were not significant and the PacBio data were inconclusive (**Table 1** and **Supplementary Data Set 1**).

In summary, we attempted further experimental validation of 40 candidate mosaic sites by conducting deep amplicon sequencing (158× mean coverage per individual) in blood from the child, mother and father using the PacBio platform. This validation experiment confirmed the presence of reads for the alternative allele in parental blood-derived DNA at 100% of the 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 a significant excess of reads for the alternative allele in maternal or paternal blood and that determined by phasing the DNM onto a parental haplotype.

Correction for mosaic power detection. To estimate the number of mosaic sites that we failed to detect because of power limitations, we ran 1,000 simulations across our 768 validated DNMs 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 calculation, we defined two bins for the mosaic level (2–4% and >4.0%). The average number of undetectable mosaic sites was calculated as a product of the number of mosaic sites and the average detection power for each bin. Hence, the proportion of germline mosaic sites after power adjustment is ~4% (31/768) of the validated DNMs.

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

First, 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. 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.

Second, 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 each 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.

Third, we experimentally ascertained the parental haplotype on which the DNM arose. Genomic DNA from the child was diluted to single-molecule concentration and then reamplified across 48 wells using the 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 the child and homozygous in the parents at the putative DNM in the unamplified DNA and were homozygous in the wells with single-molecule amplification, then the raw genotype data from the 48 amplified single molecules were analyzed in two ways. First, haplotype inference was obtained from examining peak height correlations between the genotype calls for the putative DNM and the adjacent informative SNP, and the clustering of calls was observed using an in-house script. Second, genotype calls (or peak heights pertaining to genotype calls) from the same well were counted for each locus, and the haplotype was derived from a likelihood-ratio test as detailed by Konfortov *et al.*[41].

Mutational catalog. We generated a catalog of human DNMs on the basis of 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) alleles 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 and T:A>G:T). The significance of the differences between mutational spectra was assessed by comparing the number of mutations for the six mutation types in the two spectra by means of a χ^2 test (5 degrees of freedom).

Mutational signatures were detected by refitting previously identified consensus signatures of mutational processes[16]. All possible combinations of at least seven mutational signatures were evaluated by minimizing the constrained linear function

$$\min_{\text{Exposures}_i \geq 0} \left\| \overrightarrow{\text{DNMs}} - \sum_{i=1}^{N} \left(\overrightarrow{\text{Signature}_i} \times \text{Exposure}_i \right) \right\|$$

Here $\overrightarrow{\text{DNMs}}$ and $\overrightarrow{\text{Signature}_i}$ represent vectors with 96 components corresponding to the six types of SNVs and their immediate sequencing context and Exposure_i is a non-negative scalar reflecting the number of mutations contributed by this signature. N reflects the number of signatures being refitted, and all possible combinations of consensus mutational signatures for N values between 1 and 7 were examined, resulting in 2,804,011 solutions. A model selection framework based on the Akaike information criterion was applied to these solutions to select the optimal decomposition of mutational signatures.

Diversity and divergence data. Diversity data were based on 2,453 individuals who were whole-genome sequenced to 6–8× depth as part of the ALSPAC and TwinsUK cohorts within the UK10K project. Ancestral alleles were defined by a maximum-parsimony approach as those that appeared in the majority

of five ape species (human, chimpanzee, gorilla, orangutan and macaque)[19]. Processing of great ape reference genome data is described below. SNVs were determined to be equivalent to one of the six mutation types on the basis of the identity of the ancestral and derived alleles.

Using an approach that was identical to that taken by others[42], variant sites that were likely to be under selection because they were located in exonic regions or because they were 2 kb upstream or downstream of genes were filtered out and excluded from the data set. 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 were based on multispecies alignments of the chimpanzee, gorilla, orangutan, 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 data set. Sites that were different in humans in comparison to the other great ape species were defined as substitutions.

Sex chromosomes. We included only the rarest 5% of variants in this analysis, as the mutational spectrum of those variants most resembled 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 performed 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 derived alleles. For each variant type, we then divided the number of variants by the number of nucleotides that matched the ancestral allele (**Supplementary Fig. 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. For each cell line, two replicates were available. We only included sites that were represented in both replicates. There were 1,151,596 of these sites in BC_Testis_N30, 1,048,775 of these sites in GM12878 and 1,118,911 of these sites in H1-hESC. For each cell line, we identified sites for which more than 50% of the reads were methylated in both replicates combined. We also identified sites that were present in our DNM catalog. We computed binomial P values as $\mathrm{Bin}(q, n, p)$, where q is the number of methylated DNMs, n is the total number of DNMs for which methylation data were available and p is the proportion of sites that were methylated in the data set.

37. Ramu, A. *et al.* DeNovoGear: *de novo* indel and point mutation discovery and phasing. *Nat. Methods* **10**, 985–987 (2013).
38. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).
39. Koressaar, T. & Remm, M. Enhancements and modifications of primer design program Primer3. *Bioinformatics* **23**, 1289–1291 (2007).
40. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
41. Konfortov, B.A., Bankier, A.T. & Dear, P.H. An efficient method for multi-locus molecular haplotyping. *Nucleic Acids Res.* **35**, e6 (2007).
42. Wilson Sayres, M.A., Venditti, C., Pagel, M. & Makova, K.D. Do variations in substitution rates and male mutation bias correlate with life-history traits? A study of 32 mammalian genomes. *Evolution* **65**, 2800–2815 (2011).

doi:10.1038/ng.3469