Striking differences in patterns of germline mutation between mice and humans

Sarah J. Lindsay¹, Raheleh Rahbari¹, Joanna Kaplanis¹, Thomas Keane¹, Matthew E. Hurles¹

Affiliations: ¹Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK
Summary

Little is known about differences in germline mutation processes between extant mammals. We analysed genome sequences of mouse and human pedigrees to investigate mutational differences between these species. We found that while the generational mutation rate in mice is 40% of that in humans, the annual mutation rate is 16 times higher, and the mutation rate per cell division is two-fold higher. We classified mutations into four temporal strata reflecting the timing of the mutation within the lineage from zygote to gamete. The earliest embryonic cell divisions are the most mutagenic in both species, but these earliest mutations account for a much higher proportion of all mutations in mice (~25%) than in humans (~5%). We observed a strong sex bias in the number of mutations arising in subsequent cell divisions in the early embryo in mice, but not in humans. Finally, we reconstructed partial genealogies of murine parental gametes that suggest markedly unequal contributions from founding primordial germ cells.
Introduction

Several studies have used whole genome sequencing (WGS) to estimate average germline mutation rates for single nucleotide substitutions in human pedigrees\textsuperscript{1,2}, resulting in estimates of an average of $\sim 1.2 \times 10^{-8}$ mutation per basepair (bp) per generation, considerably lower than estimated from earlier evolutionary comparisons\textsuperscript{3}. Previous estimates of murine generational germline mutation rates are also conflicting, with estimates from WGS\textsuperscript{4,5} suggesting an average mutation rate of 3.5-5.4 $\times 10^{-9}$, compatible with estimates based on phenotypic markers of 4-8 $\times 10^{-9}$\textsuperscript{6}, but not with higher estimates from transgenic loci of $37 \times 10^{-9}$\textsuperscript{7}. A lower germline mutation rate in mice has been attributed to more efficient purifying selection in mice compared to humans\textsuperscript{6,7}.

Most germline mutations in humans (75-80\%) are paternal in origin, and increasing paternal age is the major factor determining variation in numbers of mutations per offspring in humans\textsuperscript{2,8,9} with an average increase of 1-2 paternal de novo mutations (DNMs) per year. Recently a more modest effect of maternal age has been reported, equating to an additional 0.24-0.5 DNMs per year\textsuperscript{10}. However, parental age effects, and other factors that influence variation in germline mutation rate, have not been well characterized in other species. The paternal age effect has been attributed to the high number of ongoing cell divisions, and concomitant genome replications, in the male germline. However, as the ratio of the number of paternal and maternal germline cell divisions in humans considerably exceeds the ratio of paternal and maternal-derived mutations\textsuperscript{11}, it appears not all germline cell divisions are equally mutable.

Germline mutations can arise at any stage of the cellular lineage from zygote to gamete. Mutations that arise in the first $\sim 10$ cell divisions prior to the specification of primordial germ cells (PGCs) can be shared with somatic lineages. In humans, at least 4\% of de novo germline mutations are mosaic in parental somatic tissues\textsuperscript{9}. Mutations that arise just after PGC specification should
lead to germline mosaicism, although the typically small numbers of human offspring per family limit the detection of germline mosaicism, and thus our understanding of mutation processes post-PGC specification. Studies of phenotypic markers of germline mutation in mice have suggested variability in mutation rates and spectra at different stages of the germline\textsuperscript{12,13,14}. Mutational variability between germline stages has also been implicated in recent work in humans\textsuperscript{9} and drosophila\textsuperscript{15}.

To characterise mutation rates, timing and spectra in the murine germline, and compare with previously published human data, we analysed patterns of de novo mutation sharing among offspring and parental tissues in two large mouse pedigrees (Figure 1), using a combination of WGS and deep targeted sequencing.

**Figure 1: Mouse pedigree sequencing and genotyping strategy.** Reciprocal crosses were repeated mated over their fertile lifespan. Three tissues (spleen, kidney and tail), were collected from the offspring at weaning, and the parents at the end of the experiment. Five pups (shown in red) from the time-matched earliest and latest litters were subject to WGS to \textasciitilde 25X in DNA extracted from spleen. Candidate de novo mutations were called, and then validated to high depth \textasciitilde 600X in the WGS offspring in spleen, and 300X in both other tissues, and to \textasciitilde 200X in DNA extracted from spleen in all other individuals (including those from the reciprocal pedigree. Candidate sites were sequenced to extremely high depth in all three tissues of all four parents (400-800X).
Germline mutation rates in mice

We validated 402 unique DNMs across the two pedigrees, with a range of 14-36 DNMs per offspring (Supplementary Table 1).

Eight DNMs impacted on likely protein function with one nonsense and seven missense DNMs, however, none of these were in genes known to have a dominant phenotype in mice, or are associated with somatic driver mutations, and so are assumed to be representative of underlying mutational processes (Supplementary Table 2).

We determined that 2.6-fold more DNMs were of paternal (N=72) than maternal (N=28) origin, similar to previous studies\(^4,5\). It is striking that mice and humans have similar paternal biases in mutations (2.6:1 and 3.6:1 respectively\(^2,9,10\)), despite the fact that the ratio of genome replications in the paternal and maternal germlines are much more similar in mice (~2.5:1) than in humans (~13:1)\(^11\) (Figure 2A).
Figure 2 Temporal strata of observed mutations. A. Schema showing on the left, new mutations occurring in one of four temporal strata defined in the germline (above). On the right, the graphs show how the mutation that occurs at this stage manifests itself in very high depth sequencing data. B. Schematic showing the number of cell divisions occurring in the average mouse and human generation. The coloured bands show the order, ratio, and approximate timing of cell divisions that occur in the germline, as defined by the temporal stages in Figure 2B.
Accounting for our sensitivity to detect DNMs, we extrapolated the average generational mutation rate in mice to be $4.7 \times 10^{-9}$ per bp; similar to that observed in previous WGS studies\(^4,5\), and approximately 40% of that estimated in humans\(^2,9\). Assuming generation times of 30 years in humans, and 9 months in mice\(^7\), we estimated the annual mutation rate in mice to be $67 \times 10^{-10}$ per base per year, 16 times higher than the human mutation rate of $4 \times 10^{-10}$. Furthermore, using the known number of germline cell divisions in human and mice\(^11\), we calculated the average mutation rate per bp per cell division to be twice as high in mice as in humans ($5.7 \times 10^{-11}$ compared to $2.8 \times 10^{-11}$).(Table 1).

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Mouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutations per genome per generation</td>
<td>~63</td>
<td>~25</td>
</tr>
<tr>
<td>Mutation rate per genome per generation</td>
<td>$1.2 \times 10^{-8}$ ($0.8 \times 10^{-8}$-$1.3 \times 10^{-8}$)</td>
<td>$0.5 \times 10^{-8}$ ($0.3 \times 10^{-8}$-$0.7 \times 10^{-8}$)</td>
</tr>
<tr>
<td>Mutation rate per year</td>
<td>$4 \times 10^{-10}$ ($2.8 \times 10^{-10}$-$4.5 \times 10^{-10}$)</td>
<td>$67 \times 10^{-10}$ ($40 \times 10^{-10}$-$91 \times 10^{-10}$)</td>
</tr>
<tr>
<td>Mutation rate per cell division</td>
<td>$2.8 \times 10^{-11}$ ($1.9 \times 10^{-11}$-$3.1 \times 10^{-11}$)</td>
<td>$5.7 \times 10^{-11}$ ($3.5 \times 10^{-11}$-$7.9 \times 10^{-11}$)</td>
</tr>
</tbody>
</table>

These figures are in broad agreement with the hypothesis that there is a negative correlation between generational mutation rate and effective population size\(^7\), but show that due to the greater number of germline cell divisions occurring per year in mice compared to humans, the mutation rates per cell division for mice and humans are closer than previously thought\(^6,7\). The 16-fold difference in annual mutation rate between extant mouse and human is substantially greater than the approximately two-fold greater accumulation of mutations on the mouse lineage since the split from the human-mouse common ancestor ~75 million years ago\(^16\). This is presumably due to much more similar annual germline mutation rates operating over much of this evolutionary time.
Timing of germline mutations in mice and humans

We deeply sequenced all validated DNMs in three tissues from the parents (mean coverage of 400-800X per tissue), two tissues from the WGS offspring (mean coverage of 400X) and a single tissue from all other offspring (mean coverage of 200X). We observed that 17/402 unique DNMs were also detected in parental somatic tissues. In addition, 70/402 DNMs were shared among 2-19 siblings, and on the same parental haplotype (where it could be determined), strongly implying a single ancestral mutation rather than recurrent mutation. The probability of two siblings sharing a DNM is three-fold higher in mice than in humans, suggesting that a higher proportion of DNMs in mice derive from early mutations in the parental germline.

We used the pattern of mutation sharing among offspring and parental tissues to classify DNMs into four different temporal strata of the germline (Figure 2B). We refer to these four strata as very early embryonic (VEE), early embryonic (EE), peri-primordial germ cell specification (peri-PGC) and late post-primordial germ cell specification (late post-PGC).

VEE mutations were observed in 25-50% of cells reproducibly in different offspring tissues, likely due to having arisen in one of the first two post-zygotic cell divisions contributing to the developing embryo. EE mutations are observed as DNMs present in parental somatic tissues in a low proportion of cells (2-20%), compatible with them arising during later embryonic cell divisions, prior to PGC specification. Peri-PGC mutations are shared among siblings, but are not detectable in parental somatic tissues (<1.6% of cells), compatible with them arising around the time of PGC specification and the split between germline and soma. After specification, PGCs proliferate rapidly, generating thousands of germ cell progenitors in both sexes. Only mutations that occur prior to this proliferation are likely to be observed in multiple siblings in our pedigrees. This
assertion is supported by studies of phenotypic markers of mutation that have shown that to induce mutant phenotypes shared among offspring, spermatogonial stem cells have to be highly depleted, almost to compete extinction\textsuperscript{13,14}. Finally, late post-PGC mutations are only observed in a single offspring, but in 100\% of cells. These encompass mutations arising during cell divisions from PGC proliferation onwards. In addition to the mouse pedigree data, we reanalyzed our previously published data on three human multi-sibling pedigrees\textsuperscript{9} to classify DNMs consistently between mouse and human.

In mice, we observed that ~25\% of all DNMs (104/402) (32\% of those private to a single offspring) were VEE mutations (Figure 3). We observed a much lower proportion, 4.3\% (33/768) in humans, despite having similar detection power. The number of VEE mutations per offspring in mice varied strikingly (0-58\% of all DNMs), much greater than expected under a Poisson distribution (p=0.002), and contributed significantly to the variance in the overall number of DNMs per individual, but not in humans (1-10\% of all DNMs). (Supplementary Table 1).

VEE mutations in mice arose at similar rates in both sexes, and approximately equally on paternal and maternal haplotypes (Figure 3). The distribution of allele proportions for the observed VEE mutations is consistent with the vast majority of these events occurring in the first cleavage cell division that contributes to the embryo (Supplementary Figures 2 and 3).
Figure 3: Validated mutations in two pedigrees. Offspring and their litters they belong to are shown vertically on the plot. Validated DNMs are shown horizontally. Sites that are present in an offspring are shown in red, while sites that are absent are shown in light blue. The sites are ordered by temporal time points; early embryonic sites (the site to the left of the DNM is shaded according to which parent it arises from), then peri-PGC sites, followed by late-PGC mutations and very early embryonic mutations which we observe in the offspring. The ratio of paternal/maternal haplotype on which the mutation arose is shown on the left, and both read pair phased and lineage inferred phasing (in brackets) is shown for peri-PGC sites. The ratio of sites observed in male:female offspring for very early embryonic mutations.
We observed seventeen EE DNMs in mice (4% of DNMs), present at low levels in all three parental somatic tissues (1.6-19%) (Figure 3, Supplementary Table 1), representing a very similar proportion of all DNMs to that observed in human pedigrees. All but one EE mutations were observed in multiple offspring, confirming germline mosaicism. We observed a striking parental sex bias for this class of mutations in mice (16 paternal, 1 maternal, p=0.001) but not humans (9 paternal, 16 maternal, p=0.83). It is remarkable to observe such a biological difference between the sexes prior to the specification of PGCs. We considered and discounted a wide variety of possible technical artefacts that might explain this apparent parental sex bias in mice (Methods). We propose two possible biological explanations for this extreme paternal bias in EE mutations: (i) an elevated paternal mutation rate per cell division or (ii) a later paternal split between soma and germline (i.e. more shared cell divisions). Further work is required to distinguish between these two scenarios, although the observation of early sex dimorphism in pre-implantation murine and bovine embryos may well be relevant.

We identified 54 peri-PGC DNMs shared among two or more offspring but not present at detectable levels (>1.6% of cells) in parental somatic tissues (Figure 3). We did not observe any preferential sharing of these DNMs within litters as opposed to between litters (Figure 3), as might be expected if only a subset of spermatogonial stem cells (SSCs) were productive at any one time. Unlike EE mutations, peri-PGC mutations arose approximately equally in the paternal and maternal germlines (direct phasing: 10 paternal, 9 maternal; inferred parental origin using co-occurrence: 25 paternal, 25 maternal). The numbers of peri-PGC DNMs are not comparable between mouse and human pedigrees, due to the disparity in numbers of offspring per pedigree and therefore the power to observe shared DNMs.
Taken together, these results show that for some mice, 40-50% of de novo mutations observed in the offspring are derived from early stages of embryonic development in the parents, which accords with estimates of germline mosaicism from phenotypic studies\textsuperscript{9}.

**Mutation spectra in mice and humans**

Comparing low-resolution (6-class) mutational spectra of DNMs in mice and a catalogue of compiled DNMs in humans\textsuperscript{9} reveals a significant increase in T>A (p=0.00032, Chi-squared test), and a significant decrease in T>C (p=0.00002, Chi-squared test) in mice compared to humans (Figure 4A(i)), which is supported by data from other mouse pedigrees\textsuperscript{4}. However, we observed no significant differences in the mutation spectra between maternally and paternally derived DNMs in mice (p= 0.2426, Chi-squared test, Supplementary Figure 3).
In addition, we observed significant differences (p= 0.01, Chi-squared test) in the mutation spectra in mice before and after primordial germ cell specification (Figure 4A(ii)), primarily characterized by T>G mutations, highlighting differences in mutation processes between embryonic development and later gametogenesis.

With fewer pre-PGC mutations in humans, we are underpowered to detect a similar temporal difference in mutation spectra.
Figure 4: Plot showing the effect of parental age on the number of DNMs observed in each individual before (a) and after (b) the removal of very early embryonic mutations occurring in the offspring. (c) Comparison of mutational spectra in mice and humans using catalogue of compiled DNMs in humans as in Rahbari R². (d) Comparison of mutational spectra in mice, where very early embryonic and early embryonic mutations (Pre-PGCs) are compared against peri-PGC and late post-PGC mutations (Post-PGCs).
Parental age effect

We observed an average increase of 6 DNMs over the 33 weeks between earliest and latest mouse litters, which is 4.6 times greater than we would expect in humans in the same time period. This increase is greater than the 1.9-fold increased rate of turnover of SSCs in mice compared to humans, suggesting an increased mutation rate per SSC division in mice. However, unlike in humans, in mice parental age is not a significant predictor of the total number of DNMs per offspring, either within each pedigree individually (p=0.11 and 0.13) or across both combined (p=0.21) (Figure 4B(i), Supplementary Table 1). This is due in part to the lower number of mutations resulting in lower power to detect a parental age effect. However, VEE mutations represent a large proportion of all DNMs in mice, and yet we might expect only pre-zygotic mutations to be influenced by parental age. Accordingly, we found that parental age was a significant predictor of the total number of pre-zygotic DNMs across both pedigrees (p=0.005)(Figure 4B(ii). As in humans, the parental age effect in mice appears to be predominantly paternally driven, as pre-zygotic mutations exhibit the greatest paternal bias (4.7:1 compared to 2.6:1 overall) and the ratio of paternal mutations to maternal mutations is higher in offspring in later litters compared to earlier litters.

Comparing stage-specific mutation rates in mice and humans

We calculated and compared mutation rates per cell division at different phases of the germline in both mice and humans (Figure 5), by integrating information on the known cellular demography of the germline in mice and humans, the strength of the paternal age effects, and the numbers of mutations arising in each temporal strata from our pedigree studies.
We observed that mutation rates per cell division are highest in the first cell division of embryonic development than at any other germline stage, in both humans (8X higher than average) and mice (9X higher than average). This observation is supported by previous murine studies in which mosaic mutations causing visible phenotypes were strongly enriched for mutations present in 50% of cells. The mutation rate per cell division during SSC turnover (post-puberty) is considerably lower in humans than in mice (Figure 5). Moreover, in mice the mutation rate per SSC division is only two-fold lower than during pre-pubertal divisions, whereas in humans the concomitant reduction in mutation rate is ten-fold. This discordance likely explains the marked difference in humans between average germline mutation rates per cell division in males and females (Figure 5), whereas in mice the average mutation rates in the maternal and paternal germline are much more similar. It is likely that the disproportionate contribution of SSC divisions to the human germline (due to the lag between puberty and average age at conception) has led to stronger selection pressures to reduce the mutation rate per cell division in SSCs in humans than in mice.
Figure 5: Estimation of mutation rates per cell division; species average in red, very early embryonic in brown, female average in green, male average in blue, and male pre and post puberty in dark blue and pink respectively. A description of how these were calculated can be found in the methods section.

Reconstruction of mouse genealogies

Mutations shared among offspring are markers of the underlying cellular lineages from which parental gametes were derived. Although meiotic generation of haploid genomes can uncouple mutations present in the same ancestral diploid genome, we would expect two shared mutations arising on the same cellular lineage to be observed in the same offspring more often than expected.
by chance. Conversely, we would expect two shared mutations arising on
different cellular lineages in the same parent to be observed in mutually exclusive
sets of offspring. Finally, two shared mutations arising in different parents would
be expected to observed in the same offspring at random. Therefore, we
reconstructed four cellular genealogies, one for each parent, using an iterative
procedure to cluster shared mutations into lineages based on their correlation
across offspring, constrained by parental origin (see Methods).

Using this iterative clustering procedure, we assigned 67/71 shared
mutations to a specific parent, and defined partial cellular genealogies for each
parent (Figure 6). Each parental genealogy is characterised by 2-4 lineages
defined by early embryonic and peri-PGC mutations, and a residue of offspring
without shared mutations (representing 13-55% of all offspring). These primary
lineages are distributed randomly with respect to litter timing, suggesting that
their relative representation among gametes is stable over time and primarily
reflects processes operating prior to PGC specification and/or during the early
stages of PGC proliferation. We noted markedly unequal contributions from
different lineages, with individual lineages defined by early embryonic or peri-
PGC mutations accounting for 2-54% of offspring from a breeding pair. It has
been estimated that 6 cell lineages are set aside during mouse development
which later go on to specify 40-42 PGCs\textsuperscript{17,18,22}. In principle, over-represented
lineages could have arisen from having begat multiple PGC founders, or from
relative fecundity during early PGC proliferation. The correlation between levels
of somatic mosaicism and germline mosaicism suggests that the former can be a
contributing factor, whereas the observation that the most over-represented
lineage (M2) is only defined by peri-PGC mutations, and the presence of major
sublineages defined by later peri-PGC mutations, suggests that lineage birth-
death during early PGC proliferation can also play a major role. These results
indicate that specified PGCs do not contribute equally to the final pool of
gametes, although further work is required to determine the relative contribution
of selective and stochastic factors to the disproportionate representation of cellular lineages among gametes.

Figure 6: Lineage reconstructions showing reconstruction of putative maternal and paternal cell lineages using early embryonic and peri-PGC mutations. Individual offspring are numbered and coloured by litter.
Conclusions

We have characterized DNMs in two mouse pedigrees assigning the mutations to different time points within embryonic development and gametogenesis, and compared to similar data in humans. Some of the differences we observed between mouse and humans can be attributed to the differences in cellular genealogies of the germline (e.g. the greater number of SSC divisions in humans), however, others cannot, and must result from biological differences within the same stage of embryogenesis or gametogenesis. For example, the likely cause of the striking paternal bias of EE mutations in mice, which is not observed in humans, is unknown, but perhaps relates to poorly understood, but fundamental, sex differences in how cell lineages are specified in early embryonic development in mice\textsuperscript{23,24}.

One notable similarity between mouse and human germlines was the hypermutability of the first post-zygotic cell division contributing to the developing embryo, although the relative contribution of VEE mutations to the mutation rate per generation was much higher in mice. The strikingly high variance in numbers of VEE mutations between mouse offspring suggests that this stage is much more mutagenic for some zygotes than others. In addition, reconstructing partial genealogies for the mouse germline has revealed highly unequal contributions of different founding lineages to the ultimate pool of gametes. These observations motivate a deeper understanding of the demography of primordial germ cell lineages.

Our finding that generational mutation rates in mice are lower than in humans while per division mutation rates are higher, raises an apparent paradox: if purifying selection in mice is more efficient at reducing generational mutation rates, why does the murine cellular machinery have lower fidelity per genome replication? The answer likely lies in the expectation that the selection coefficient
of an allele that alters the absolute fidelity of genome replication will depend critically on the number of genome replications per generation. Thus, given the much greater number of genome replications in a human generation, an allele that alters the fidelity of genome replication by a given amount will have a considerably higher selection coefficient in humans than in mice. The reduction in mutation rate in SSC divisions compared to previous cell divisions was far more pronounced in humans than in mice. This is presumably as a result of stronger selective pressures in humans due to the much greater contribution of this class of genome replication to the overall number of genome replications in the germline.

Much of the existing literature comparing germline mutation processes between species focuses on the dependence of these processes on ‘life history’ traits\textsuperscript{25,26}. We contend that these ‘life history’ traits are imperfect proxies for the true molecular and cellular basis of this variation between species, which relates to the number of different classes of cell division within the germline, and the mutation rates and spectra accompanying each temporal strata of the germline. Broader application of the kinds of analyses performed here will catalyse the transition from a demographic understanding of germline mutation towards a truly molecular comprehension.
Online Methods

Mice.

Ten male and female mice from each strain (CB57BL/6 and 129S5) were obtained from sib-sib inbred lines previously established at the Wellcome Trust Sanger Institute. Twenty breeding pairs were established (Ten CB57BL/6 ♂ x ten 129S5 ♀ (GPCB), and ten reciprocal crosses (CBGP)). Breeding pairs were introduced at regular intervals over a period of several months, if a pregnancy resulted, the pups were left to wean and then culled at 3-4 weeks of age. Tissue samples of spleen, kidney and tail were taken from pups, and from the parents either when one of them died or became ill, or when no pregnancies resulted after matings over a period of three months. At the onset of the experiment, the ages of the GPCB breeding pairs were 9.9 weeks (male), and 7.8 weeks (female), and the CBGP pairs were 8.1 weeks (father) and 9.8 weeks (mother). Strain specific SNPs were identified in the WGS data to verify the identity of the parents was correctly assigned. To prevent sample swaps, the litters were stored apart and extractions carried out separately for each litter and each pedigree.

DNA Sample Preparation and QC.

Tissues were stored at -80°C immediately after harvest. DNA was prepared DNA using Qiagen DNeasy tissue prep kits in litter specific/parent specific batches to minimize possible sample swaps. Where possible, single DNA aliquots from the same tissues were used for multiple studies; for example, the DNA from the same tube was used for WGS and validation sequencing. After WGS was carried out, parental samples were genotyped referenced against strain and sex specific SNVs.

Sequencing and variant calling.

DNA extracted from the spleen of parents and offspring was sequenced using standard protocols and Illumina HiSeq technologies. The resultant sequence
data was aligned to mouse reference GRCm38. The total mapped coverage after
duplicate removal had a mean of 25X and range 22-35X for CBGP, and 29X and
22X-40X for GPCB. Variants were called using bcftools and samtools and
standard settings\textsuperscript{27}.

\textbf{De novo mutation calling.}

De novo mutations were called on the variants supplied by bcftools by using
\textit{DeNovoGear} version 0.5 using standard settings\textsuperscript{28}. \textit{DeNovoGear} called between
7711 and 11069 (mean 9736) short indels and SNVs in CB trios, and between
8578 and 12835 (mean 10916) candidates in GP trios respectively. Calls from
the X chromosome were discarded as SNVs and indels showed a strain/sex
specific inflation, for which it was not possible to correct for.

\textbf{Filtering of candidate de novo mutations.}

Candidate de novo mutations were filtered to exclude sites highly enriched for
false positives (simple sequence repeats (2\% of sites on average), segmental
duplications (0.5\% of sites on average), although these sites are not exclusive of
each other. In addition, strain-specific mapping artefacts (low quality areas
leading to clustered/low quality SNV/indel candidates were filtered by removing
sites that had a high alternative allele ratio (>0.2) in any pup in the reciprocal
(unrelated litter), or parent of reciprocal (unrelated) litter (>0.04). Assuming a
Poisson distribution for sequencing depth, sites with a depth greater than the
0.0001 quantile were removed due to the likelihood of mapping errors or low
complexity repeats introducing false positives (generally 13\% of candidate sites).
Candidate sites where the de novo mutation was present in either parent in
greater than 5\% of reads and where there were known SNPs in the parental
strain were also removed on the grounds that they were likely to be inherited (on
average, 79\% of sites). Once these filters were applied, 272, 380, 225, 260, 205,
324, 166, 286, 284, 375 and 211, 174, 180, 346, 135, 101, 160, 143, 191, 300
candidate de novo mutations remained for CBGP and GPCB offspring

\textsuperscript{27} CC-BY 4.0 International license
\textsuperscript{28} certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under
\textsuperscript{a CC-BY 4.0 International license.}
respectively.

**Experimental validation of de novo mutations.**

A total of 4460 unique sites across all 20 offspring were put forward for validation by Agilent Sure Select Target Enrichment. Twenty-one sites were lost during liftover conversion, leading to 4439 sites put forward for bait design. Bait design included 2X tiling, moderate repeat masking, maximum boosting, across 100bp, of sequence flanking the site of interest (extending to 200bp where baits could not be designed on the initial attempt. Of these 4439 sites, 3253 sites were successfully designed for with high coverage (>50% coverage), 222 with medium coverage (>25% coverage), and 421 with low coverage (<25% coverage). 564 sites failed bait design, however, our previous analyses have showed that sites that fail bait design are enriched for false positives. Initially, the target enrichment set was run (2 lanes of 75bp PE Hiseq) on DNA extracted from the spleen of the 20 offspring subject to WGS and their parents, leading to an average of 300X across each site. A subsequent run (5 lanes of 75bp PE Hiseq) was carried out with tissues from the parents’ kidney, tail and spleen, the WGS-sequenced offspring spleen and tail, and the spleen from all the additional offspring from the breeding pairs, leading to an average of 400-800X coverage for each site in parental tissues, and an average of 200X coverage in offspring tissues. The resultant sequence data were merged by individual and annotated with read counts at the candidate site using an in-house python script. An in house R script (http://www.Rproject.org) was then used to allocate a likelihood to each candidate variant being a true de novo mutation, an inherited variant or a false positive call, based on the allele counts of the parents and offspring at that locus. A proportion of the SNV candidates (all sites put forward for validation for one individual) as well as all of the indel candidates were reviewed manually using Integrative Genomics Viewer (IGV)²⁹.

**Functional Annotation of variants**
Functional annotation of DNMs was carried out using ANNOVAR\textsuperscript{30}.

**Identification and power to detect parental mosaics.**

In order to identify DNMs that could be mosaic in one of the parents, the site specific error was calculated for each site (\% of reads that map to non-reference allele in unrelated individuals from the reciprocal pedigree). This error was then used to calculate the binomial probability of observing \( n \) non-reference reads at the mutated site in each tissue in each individual. The probabilities were corrected for multiple testing, using both FDR and Bonferroni correction (yielding the same results), using a threshold of \( p<0.05 \) to identify candidate sites, which were then viewed in IGV\textsuperscript{29}. In addition, the power to detect mosaicism at different levels (0.5\%, 1, and 1.5\% respectively), in each tissue in each parent was calculated using the sequence depth from the validation data.

**Haplotyping of de novo mutations in offspring.**

We used the read-pair algorithm supplied with the DeNovogear software to determine the parent of origin of our validated de novo mutations using the deep whole-genome sequence data. DeNovoGear uses information from flanking variants that are not shared between parents to calculate the haplotype on which the mutation arose. Using this technique, we were able to confidently assign the parental haplotype in 100 of 402 unique validated de novo mutations. We were also able to infer the parent of origin for 12 additional sites that were assigned as being mosaic in one of the parents. We were also able to infer the phase of 37 additional mutations that were shared between offspring and were assigned to a parental lineage.

**Per generation mutation rate estimation.**

We calculated a mutation rate for autosomal SNVs in each individual as follows: first, we calculated the proportion of the genome not covered in our analysis because of the depth of the whole-genome sequencing: Bedtools\textsuperscript{31} was used to
calculate the proportion of the genome not considered in our analysis due to low- or high-sequence depths for each individual (mean 5.6%). We then calculated the proportion of sites that were removed by our whole-genome filters (simple sequence repeats and segmental duplications) after the depth filters were applied (average 2.1%). Last, we used the posterior probability supplied by DeNovoGear (>0.9) to calculate what proportion of sites that were not validated (failed validation or removed by to filters), were likely to be true de novo mutations. For human/mouse comparisons, generation times were assumed to be 30 years and 9 months respectively. According to Drost\textsuperscript{11}, this would result in ~432 cell divisions in the human germline, and ~87 cell divisions in the mouse (paternal and maternal combined).

\textbf{Identification of very early embryonic mutations in offspring.}

We aggregated the alternate allele counts and total depths between tissues, after testing that the allele ratios were concordant across tissues (Fishers Exact test). Very early embryonic mutations (defined as occurring after in the individual after fertilization, and therefore private to that offspring), was classified as follows: A likelihood-based test was then carried out on the combined counts to test the hypothesis that the alternate allele count was suggestive of a constitutive (binomial p=0.5) or a VEE origin (binomial p=0.25), where a site with log likelihood difference of >5 was designated VEE, <5 was designated constitutive, or unassigned if it falls between those values. Due to lower coverage, for 10% of mutations in human pedigrees, and 4% in mouse pedigrees, we were unable to confidently infer whether the mutations were constitutive or very early embryonic.

In addition, haplotype occupancy (HO) was ascertained where possible; the nearest heterozygous variant to the de novo mutation should phase consistently 100% of the time for a zygotic (constitutive) mutation, whereas for a very early embryonic mutation, the de novo allele mutation only be seen on a proportion of haplotypes defined by the nearest variant. (Supplementary Figure 3). The HO for
mouse and human DNM sites was plotted against the alternate allele proportion; this showed that, where HO could be determined, sites with a low alternate allele ratio were enriched for sites with low HO, whereas shared sites that are constitutive by definition only have high HO.

**Reconstruction and testing of parental lineages.**

Parental lineages were reconstructed using the distribution of mutations shared between offspring, using the following expectations: Shared mutations that are observed in the same offspring significantly more frequently than expected by chance are likely to belong to the same parental lineage. Conversely, mutations that are never observed together are likely to come from the same parent, but a different lineage. Mutations that are shared in a random manner could come from the same lineage in the same parent, or a lineage from the other parent.

In the first step, a pairwise test was carried out for each shared mutation, which calculated the binomial probability of \( n \) pups sharing \( m \) mutations where the frequencies of the mutations were \( p \) and \( q \) in the offspring. Then, the pair of sites with the lowest resultant p-value were merged into a single pseudo site containing all the offspring who have either site from the initial pair, as long as the parental origin of the two mutations was not discordant. The pairwise test was then repeated, followed by another merge of sites, either until a given p-value threshold is reached, or the pseudo sites cannot be merged any further.

Given a p-value threshold of 0.05, all sites had completely collapsed into the given clusters. All but four of the seventy shared mutations could be assigned to either paternal or maternal lineages, the remaining mutations represent lineages defined by a only single shared mutation.

The accuracy of the lineage reconstructions were tested using two simulations. Firstly, for each pedigree, shared mutations were randomly re-assigned into the lineages defined by the reconstruction above. They were then checked for
biological concordance - each individual can only belong to one paternal and one maternal lineage. This test was carried out 10,000 times for each pedigree, none of which were biologically concordant (ie at least one offspring would have more than one paternal or maternal lineage). Secondly, for each pedigree, mutations were randomly clustered into lineages containing differing numbers of mutations (from 2-10 mutant sites) and tested again for concordance as above, 10,000 times. In this way, 40000 simulations across both pedigrees showed no other possible concordant lineage structures. All phase and haplotype information was concordant between offspring.

**Estimation of mutation rates per cell division.**

Haploid rates were calculated as listed below:

*Average mutation rates*

Average mutation rates across species were calculated using the per-generation average number of mutations, corrected for genome wide coverage (see methods above), and the 95% confidence intervals were calculated assuming numbers of mutations fall in a Poisson distribution. The number of mutations were then divided by the sum of paternal and maternal cell divisions in a generation (87 and 432 respectively assuming a generation time of 9 months for mice, and 30 years for humans)\(^{11}\).

To calculate the paternal per-generation average, the total number of per-generation genome wide corrected mutations was used in the following formula:

$$\mu_{\text{paternal}} = k \times \frac{n_{\text{phased paternal}}}{n_{\text{phased}}} \times \frac{mutations_{\text{total}}}{n_{\text{offspring}}}$$

\(^{11}\)
where scaling factor scales the number of discovered mutations to the genome wide corrected number of mutations, and where the 95% confidence intervals were derived from the assumed Poisson distribution of numbers of mutations. The putative numbers of paternal mutations per generation were then divided by the estimated number of cell divisions per generation (62 in mice, 401 in humans).\(^8\). The maternal per-generation average was calculated as above, using 25 and 31 cell divisions per generation (mouse and human, respectively).\(^11\).

**Very Early Embryonic Mutations**

Very early embryonic mutations occur in the first cell divisions that contribute to the embryo (rather than to extra-embryonic tissues). Assuming the founding cells in the inner cell mass (ICM) of the blastocyst divide symmetrically, these mutations occur in one or two consecutive cell divisions in the first two cells to eventually comprise the embryonic tissues. We can only observe these in the offspring; recovery of very early embryonic mutations that occur in the parents will have been filtered as putative inherited variants. In addition, we can only capture two symmetrical cell divisions at most; once the frequency of cells carrying the alternate allele below falls 25% it is unlikely to be recovered during de novo calling when WGS coverage is ~25X. We identified this class of mutation arising in offspring using several different methods (Methods). As we are estimating the rate from the offspring, we use the sex of the offspring rather than haplotypes from the parents to define relative contributions by sex. With 25X coverage for the WGS discovery phase, the vast majority of the VEE mutations we detect will be from a single cell division. Modelling shows that our mutation calling pipeline had very low power to detect VEE mutations in subsequent cell divisions. In addition, the distribution of the alternate allele proportion for VEE mutations is centred symmetrically around 0.25 as would be
expected for mutations arising in the first cleavage cell division contributing to the embryo. These results suggest that the majority of VEE mutations we detected arose in a single cell division (Supplementary Figure 3).

To estimate the VEE mutation rate per cell division we took the total number of mutations that we determined to be VEE (104 in mice, 33 in humans), and calculated the 95% Poisson confidence interval around this count. We then divided this number by 2 (to obtain a haploid rate), and then by the total number of offspring (20 for mouse, 12 for human).

The power to identify this class of mutation is based on WGS sequencing depth, and the power to correctly discriminate it from a constitutive mutation is based on validation sequencing depth. At ~100X sequencing coverage, we have 97% power to correctly infer this class of mutation, and we have similar power to detect this class of mutations in humans and mice.

Pre-puberty in the male germline

The total number of mutations occurring pre-puberty in the male germline were defined as follows:

\[ N = \text{mutations}_{\text{mean}} - \left( \text{age}_{\text{mean}} - \text{age}_{\text{puberty}} \right) \times \text{annual mutations}_{\text{mean}} \]

95% Poisson confidence intervals were derived from the mean number of mutations per year.

Post-puberty in the male germline

As parentally-aged induced mutations accrue in an approximately linear manner, the post-puberty mutation rate in males was calculated on the number of
mutations accrued in the mouse and human paternal germline in a single year.

The average number of mutations in mice increased by 6 over a 33 week timespan, leading to an extrapolated annual increase of 9.45 mutations. The largest human study to date suggests an increase of 2.01 mutations per year\(^2\).

The annual number of mutations was divided by the annual number of cell divisions occurring in that organism (42 for mice, 23 for humans\(^6\)). Confidence intervals were derived from the uncertainly of the slope of the linear models of effect of age on number of mutations (estimates for human obtained from Kong et al\(^2\)).

**Analysis of mutation spectra**

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).

**Estimation of recurrence risk of DNMs in offspring**

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 both pedigrees

**Possibility of technical artefacts.**

We considered and discounted a wide variety of possible technical artefacts that might explain the apparent parental sex bias we observe in early embryonic
mutations in mice. Firstly, sequencing depth, and thus power to detect somatic
mosaicism, was equal between maternal and paternal tissues, and the identity of
the WGS samples were checked using strain and gender specific SNPs.
Secondly, where parental origin could be independently determined by
haplotyping with nearby informative sites (N=6), the parental origin was
confirmed, thus excluding sample swaps. Thirdly, parental mosaicism was
supported by very low read counts in the WGS data in the parents at 6 of the
mosaic sites (2 and 3 sites from both fathers, and one from the mother). Fourth,
the same aliquot of DNA was used for WGS and validation of mutations in
parental spleen, lowering the possibility of sample swaps. Lastly, in all cases,
parental mosaicism was independently supported by sequencing data from two
additional tissues.
Supplementary:

**Supplementary Table 1:** Table showing counts of DNMs in each category for each individual, offspring CBGP8_1a-h, GPCB2_1a-e and CBGP8_8a-f, GPCB2_9a-f derive from the earliest and latest litters respectively.

**Supplementary Table 2:** DNMs with potentially functional consequences as given by ANNOVAR are listed.

**Supplementary Table 3:** All DNMs are listed, with columns in order of chromosome, position, type, reference allele, alternative allele, which offspring they were called in (CBGP8_1a,CBGP8_1aT are sequences from the spleen and tail of the same individual), the number of individuals the site is shared with, whether the site is mosaic, called as VEE or Zygotic, which lineage it belongs to, and finally read-pair haplotyping results.

**Supplementary Figure 1**

Plots showing haplotype occupancy in heterozygous sites directly adjacent to de novo sites plotted against the alternate allele proportion at the validated site. The histogram shows the distribution of individuals along the y axis. It can be observed that the mouse DNM sites that are shared cluster around the 0.5 alternate allele proportion, and where ascertained, have a HO of ~1. Compared to the human data, the mouse DNMs have a greater skew towards low alternate allele proportion and a greater number of putative post-zygotic sites where HO and alternate allele proportion are both low. b) Haplotype occupancy (HO) defined as a DNM (in this case, A->G, which does not segregate fully with the variant on the haplotype on which it arose (in this case, on the paternal haplotype.)
Supplementary Figure 2. Histograms of the proportion of alternate allele in validated DNMs in high depth sequence data in humans (A) and mice (B). Sites in red are constitutive and are have an alternate allele proportion centred around 50% of reads (100% of cells). Sites classified as very early embryonic are shown in blue, are found in around 25% of reads (50% of cells). Red, blue and black lines show the expected distribution of alternate allele proportions given a binomial distribution of reads centred around constitutive, first division and second division mutations, in our high depth sequence data.

Supplementary Figure 3

Low resolution mutation spectra in maternal and paternally derived DNMs in mouse and human data. Error bars show the 95% confidence intervals.
Acknowledgements

We are very grateful for the expert assistance of James Bussell and the Sanger Institute Mouse Facility for mouse breeding. We thank Art Wuster, Saeed Al-Turki, Jeremy McRae, Ludmil Alexandrov, Aylwyn Scally, Kirstie Lawson, Ian Adams and Robin Lovell-Badge for thoughtful discussions and sharing of scripts. This work was supported by the Wellcome Trust [grant number WT098051].
References:


10. Wong et al, Nature Communications 7 New observations on maternal age
effect on germline *de novo* mutations Nature Communications

DOI: 10.1038/ncomms10486 (2016)


13. Russell, B, Significance of the Perigametic Interval as a Major Source of Spontaneous Mutations That Result in Mosaics Environmental and Molecular Mutagenesis 34:16-23 (1999)


24 Mihajlovic AI, Thamodaran V, Bruce AW, The first two cell fate decisions of preimplantation mouse embryo development are not functionally independent. Nature Scientific Reports 5:15034 (2016)


Supplementary Tables and Figures

Supplementary table 1: Mouse individuals and counts of each class of mutation.

<table>
<thead>
<tr>
<th>IID</th>
<th>early embryonic</th>
<th>peri_PGC specification</th>
<th>late post-PGC specification</th>
<th>very early embryonic</th>
<th>indels</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBGP8_1a</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>CBGP8_1b</td>
<td>0</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>CBGP8_1c</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>CBGP8_1g</td>
<td>1</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>CBGP8_1h</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>CBGP8_8a</td>
<td>1</td>
<td>3</td>
<td>14</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CBGP8_8b</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>CBGP8_8c</td>
<td>0</td>
<td>2</td>
<td>13</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>CBGP8_8d</td>
<td>3</td>
<td>5</td>
<td>19</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>CBGP8_8f</td>
<td>4</td>
<td>5</td>
<td>11</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>GPCB2_1a</td>
<td>1</td>
<td>0</td>
<td>11</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>GPCB2_1b</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>GPCB2_1c</td>
<td>1</td>
<td>1</td>
<td>15</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>GPCB2_1d</td>
<td>1</td>
<td>9</td>
<td>11</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>GPCB2_1e</td>
<td>1</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GPCB2_9a</td>
<td>2</td>
<td>8</td>
<td>14</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>GPCB2_9b</td>
<td>1</td>
<td>2</td>
<td>10</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>GPCB2_9c</td>
<td>0</td>
<td>5</td>
<td>16</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>GPCB2_9e</td>
<td>2</td>
<td>3</td>
<td>24</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>GPCB2_9f</td>
<td>2</td>
<td>3</td>
<td>15</td>
<td>11</td>
<td>1</td>
</tr>
</tbody>
</table>
## Supplementary Table 2: Functional consequences of de novo mutations

<table>
<thead>
<tr>
<th>ID</th>
<th>chr</th>
<th>position</th>
<th>ref</th>
<th>alt</th>
<th>site</th>
<th>consequence</th>
<th>gene</th>
<th>mutation class</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBGP8_8b</td>
<td>2</td>
<td>10556273</td>
<td>G</td>
<td>A</td>
<td>exonic</td>
<td>synonymous SNV</td>
<td>Sfmbt2</td>
<td>paternal early embryonic</td>
</tr>
<tr>
<td>CBGP8_8f</td>
<td>2</td>
<td>28556769</td>
<td>A</td>
<td>T</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
<td>Cel</td>
<td>late post-PGC specification</td>
</tr>
<tr>
<td>CBGP8_1a</td>
<td>2</td>
<td>30084760</td>
<td>C</td>
<td>T</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
<td>Pkn3</td>
<td>paternal peri-PGC specification</td>
</tr>
<tr>
<td>CBGP8_8a</td>
<td>7</td>
<td>104975169</td>
<td>T</td>
<td>C</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
<td>Olfr671</td>
<td>late post-PGC specification</td>
</tr>
<tr>
<td>CBGP8_1g</td>
<td>9</td>
<td>123480538</td>
<td>C</td>
<td>T</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
<td>Limd1</td>
<td>late post-PGC specification</td>
</tr>
<tr>
<td>GPCB2_1d</td>
<td>11</td>
<td>50873775</td>
<td>G</td>
<td>A</td>
<td>exonic</td>
<td>stopgain</td>
<td>Zfp454</td>
<td>late post-PGC specification</td>
</tr>
<tr>
<td>CBGP8_1h</td>
<td>13</td>
<td>100154877</td>
<td>C</td>
<td>T</td>
<td>exonic</td>
<td>synonymous SNV</td>
<td>Naip2</td>
<td>late post-PGC specification</td>
</tr>
<tr>
<td>CBGP8_1h</td>
<td>13</td>
<td>100154880</td>
<td>G</td>
<td>A</td>
<td>exonic</td>
<td>synonymous SNV</td>
<td>Naip2</td>
<td>late post-PGC specification</td>
</tr>
<tr>
<td>CBGP8_1h</td>
<td>13</td>
<td>100154911</td>
<td>T</td>
<td>A</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
<td>Naip2</td>
<td>late post-PGC specification</td>
</tr>
<tr>
<td>CBGP8_1h</td>
<td>13</td>
<td>100154951</td>
<td>C</td>
<td>T</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
<td>Naip2</td>
<td>late post-PGC specification</td>
</tr>
<tr>
<td>CBGP8_1a</td>
<td>19</td>
<td>44935143</td>
<td>G</td>
<td>A</td>
<td>exonic</td>
<td>nonsynonymous SNV</td>
<td>Fam178a</td>
<td>late post-PGC specification</td>
</tr>
<tr>
<td>CBGP8_1c</td>
<td>8</td>
<td>83722794</td>
<td>G</td>
<td>A</td>
<td>splicing</td>
<td>NA</td>
<td>Ddx39</td>
<td>late post-PGC specification</td>
</tr>
</tbody>
</table>
Supplementary Figure 1

a) Alternate allele proportion and haplotype occupancy in SNVs in mouse (1, n=402) and human (2, n=768) offspring.

1. Haplotype occupancy

2. Haplotype occupancy

- Private alternative allele only
- Shared DHdI alternative allele only
- Shared alternative allele and HO
- Shared alternative allele and HO only
- Private alternative allele and HO

b) definition of haplotype occupancy
Supplementary Figure 1

Plots showing haplotype occupancy in heterozygous sites directly adjacent to de novo sites plotted against the alternate allele proportion at the validated site. The histogram shows the distribution of individuals along the y axis. It can be observed that the mouse DNM sites that are shared cluster around the 0.5 alternate allele proportion, and where ascertained, have a HO of ~1. Compared to the human data, the mouse DNMs have a greater skew towards low alternate allele proportion and a greater number of putative post-zygotic sites where HO and alternate allele proportion are both low. b) Haplotype occupancy (HO) defined as a DNM (in this case, A-G, which does not segregate fully with the variant on the haplotype on which it arose (in this case, on the paternal haplotype.)
**Supplementary Figure 2**

A

- constitutive sites
- VEE sites

expected distribution (constitutive)
expected distribution (first cell division)
expected distribution (second cell division)

B

- constitutive sites
- VEE sites

expected distribution (constitutive)
expected distribution (first cell division)
expected distribution (second cell division)

**Supplementary Figure 2.** Histograms of the proportion of alternate allele in validated DNMs in high depth sequence data in humans (A) and mice (B). Sites in red are constitutive and are have an alternate allele proportion centred around 50% of reads (100% of cells). Sites classified as very early embryonic are shown in blue, are found in around 25% of reads (50% of cells). Red, blue and black lines show the expected distribution of alternate allele proportions given a poisson distribution of reads centred around constitutive, first division and second division mutations.
Supplementary Figure 3

Low resolution mutation spectra in maternal and paternally derived DNMs in mouse and human data. Error bars show the 95% confidence intervals.