Striking differences in patterns of germline mutation between mice and humans

Sarah J. Lindsay[1], Raheleh Rahbari[1], Joanna Kaplanis[1], Thomas Keane[1], Matthew E. Hurles[1]

Affiliations: [1]Wellcome 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[1,2], resulting in estimates of an average of ~1.2×10^{-8} mutation per basepair (bp) per generation, considerably lower than estimated from earlier evolutionary comparisons[3]. Previous estimates of murine generational germline mutation rates are also conflicting, with estimates from WGS [4,5] suggesting an average mutation rate of $3.5\text{-}5.4 \times 10^{-9}$, compatible with estimates based on phenotypic markers of $4\text{-}8 \times 10^{-9}$ [6], but not with higher estimates from transgenic loci of 37×10^{-9} [7]. A lower germline mutation rate in mice has been attributed to more efficient purifying selection in mice compared to humans.[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 [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 [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[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 ~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[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[12,13,14]. Mutational variability between germline stages has also been implicated in recent work in humans[9] and drosophila[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.

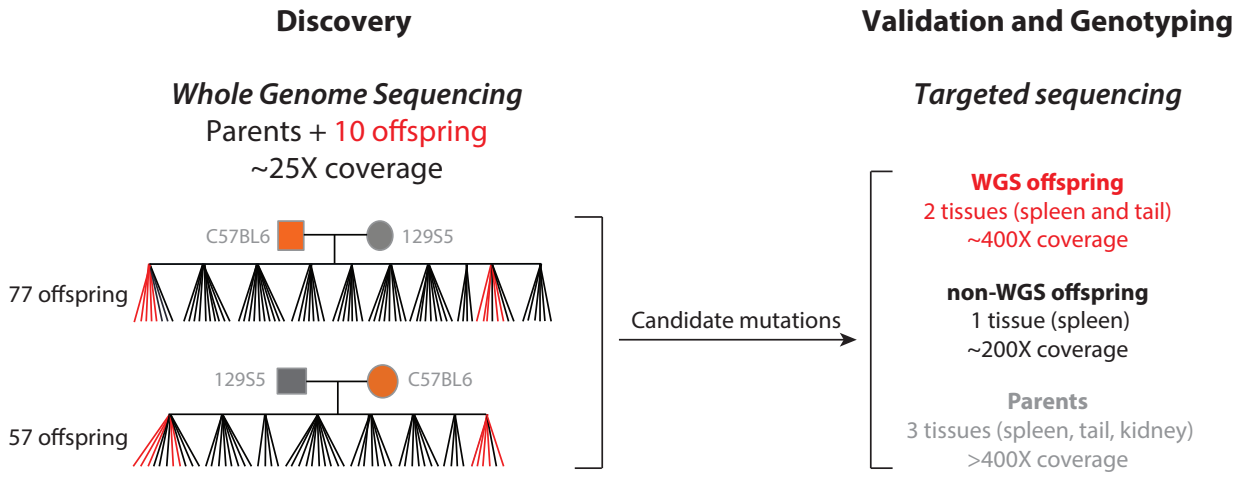


Figure 1: Mouse pedigree sequencing and genotyping strategy. Reciprocal crosses were repeated mated over their fertile lifespan. Three tissues (spleen, kidney and tail), were collected from the offspring at weaning, and the parents at the end of the experiment. Five pups (shown in red) from the time-matched earliest and latest litters were subject to WGS to ~25X in DNA extracted from spleen. Candidate de novo mutations were called, and then validated to high depth ~600X in the WGS offspring in spleen, and 300X in both other tissues, and to ~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).

84

85 **Germline mutation rates in mice**

86

87 We validated 402 unique DNMs across the two pedigrees, with a range of
88 14-36 DNMs per offspring (Supplementary Table 1).
89 Eight DNMs impacted on likely protein function with one nonsense and
90 seven missense DNMs, however, none of these were in genes known to have a
91 dominant phenotype in mice, or are associated with somatic driver mutations,
92 and so are assumed to be representative of underlying mutational processes
93 (Supplementary Table 2).
94 We determined that 2.6-fold more DNMs were of paternal (N=72) than
95 maternal (N=28) origin, similar to previous studies[4,5]. It is striking that mice and
96 humans have similar paternal biases in mutations (2.6:1 and 3.6:1
97 respectively[2,9,10]), despite the fact that the ratio of genome replications in the
98 paternal and maternal germlines are much more similar in mice (~2.5:1) than in
99 humans (~13:1)[11] (Figure 2A).

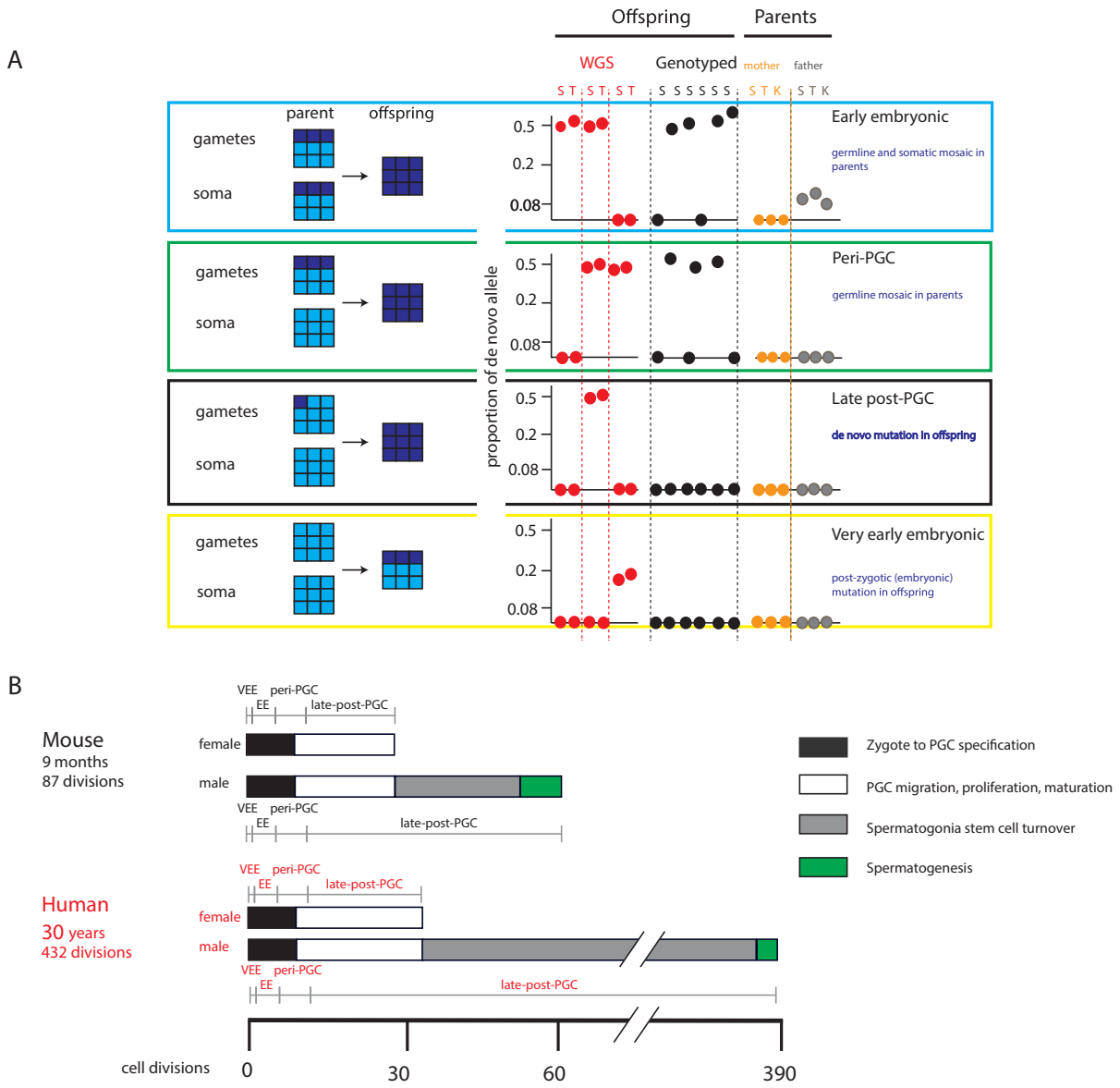


Figure 2 Temporal strata of observed mutations. A. Schema showing on the left, new mutations occurring in one of four temporal strata defined in the germline (above). On the right, the graphs show how the mutation that occurs at this stage manifests itself in very high depth sequencing data. **B.** Schematic showing the number of cell divisions occurring in the average mouse and human generation[11]. 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 4.7×10^{-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 67×10^{-10} per base per year, 16 times higher than the human mutation rate of 4×10^{-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 (5.7×10^{-11} compared to 2.8×10^{-11}).(Table 1).

Table1: Germline mutation rates per generation, per year and per cell division in humans and mice.

	Human	**Mouse**
Mutations per genome per generation	~63	~25
Mutation rate per genome per generation	**1.2×10^{-8} (0.8x10-8-1.3x10-8)**	0.5×10^{-8} (0.3×10^{-8} - 0.7×10^{-8})
Mutation rate per year	4×10^{-10} (2.8×10^{-10} -4.5×10^{-10})	**67×10^{-10} (40×10^{-10} - 91×10^{-10})**
Mutation rate per cell division	2.8×10^{-11} (1.9×10^{-11} - 3.1×10^{-11})	**5.7×10^{-11} (3.5×10^{-11} - 7.9×10^{-11})**

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.

132 **Timing of germline mutations in mice and humans**

133

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

144

145 We used the pattern of mutation sharing among offspring and parental
146 tissues to classify DNMs into four different temporal strata of the germline (Figure
147 2B). We refer to these four strata as very early embryonic (VEE), early embryonic
148 (EE), peri-primordial germ cell specification (peri-PGC) and late post-primordial
149 germ cell specification (late post-PGC).

150

151 VEE mutations were observed in 25-50% of cells reproducibly in different
152 offspring tissues, likely due to having arisen in one of the first two post-zygotic
153 cell divisions contributing to the developing embryo. EE mutations are observed
154 as DNMs present in parental somatic tissues in a low proportion of cells (2-20%),
155 compatible with them arising during later embyronic cell divisions, prior to PGC
156 specification. Peri-PGC mutations are shared among siblings, but are not
157 detectable in parental somatic tissues (<1.6% of cells), compatible with them
158 arising around the time of PGC specification and the split between germline and
159 soma. After specification, PGCs proliferate rapidly, generating thousands of germ
160 cell progenitors in both sexes [17,18,19]. Only mutations that occur prior to this
161 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[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[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).

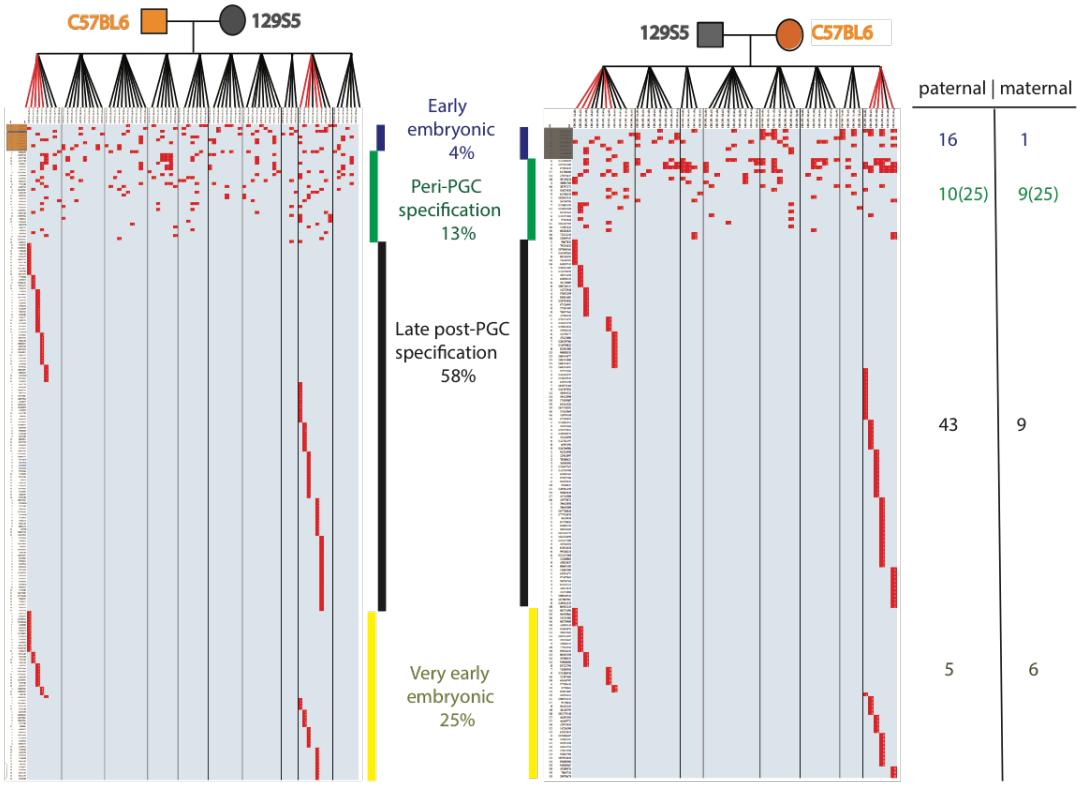


Figure 3: Validated mutations in two pedigrees. Offspring and their litters they belong to are shown vertically on the plot. Validated DNMs are shown horizontally. Sites that are present in an offspring are shown in red, while sites that are absent are shown in light blue. The sites are ordered by temporal 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.

201 We observed seventeen EE DNMs in mice (4% of DNMs), present at low
202 levels in all three parental somatic tissues (1.6-19%) (Figure 3, Supplementary
203 Table 1), representing a very similar proportion of all DNMs to that observed in
204 human pedigrees[10]. All but one EE mutations were observed in multiple offspring,
205 confirming germline mosaicism. We observed a striking parental sex bias for this
206 class of mutations in mice (16 paternal, 1 maternal, p=0.001) but not humans (9
207 paternal, 16 maternal, p=0.83). It is remarkable to observe such a biological
208 difference between the sexes prior to the specification of PGCs. We considered
209 and discounted a wide variety of possible technical artefacts that might explain
210 this apparent parental sex bias in mice (Methods). We propose two possible
211 biological explanations for this extreme paternal bias in EE mutations: (i) an
212 elevated paternal mutation rate per cell division or (ii) a later paternal split
213 between soma and germline (i.e. more shared cell divisions). Further work is
214 required to distinguish between these two scenarios, although the observation of
215 early sex dimorphism in pre-implantation murine and bovine embryos[20, 21] may
216 well be relevant.
217
218 We identified 54 peri-PGC DNMs shared among two or more offspring but not
219 present at detectable levels (>1.6% of cells) in parental somatic tissues (Figure
220 3). We did not observe any preferential sharing of these DNMs within litters as
221 opposed to between litters (Figure 3), as might be expected if only a subset of
222 spermatogonial stem cells (SSCs) were productive at any one time. Unlike EE
223 mutations, peri-PGC mutations arose approximately equally in the paternal and
224 maternal germlines (direct phasing: 10 paternal, 9 maternal; inferred parental
225 origin using co-occurence: 25 paternal, 25 maternal). The numbers of peri-PGC
226 DNMs are not comparable between mouse and human pedigrees, due to the
227 disparity in numbers of offspring per pedigree and therefore the power to observe
228 shared DNMs.
229

230 Taken together, these results show that for some mice, 40-50% of de
231 novo mutations observed in the offspring are derived from early stages of
232 embryonic development in the parents, which accords with estimates of germline
233 mosaicism from phenotypic studies[9].
234
235 **Mutation spectra in mice and humans**
236
237 Comparing low-resolution (6-class) mutational spectra of DNMs in mice
238 and a catalogue of compiled DNMs in humans[9] reveals a significant increase in
239 T>A (p=0.00032, Chi-squared test), and a significant decrease in T>C
240 (p=0.00002, Chi-squared test) in mice compared to humans (Figure 4A(i)), which
241 is supported by data from other mouse pedigrees[4]. However, we observed no
242 significant differences in the mutation spectra between maternally and paternally
243 derived DNMs in mice (p= 0.2426, Chi-squared test, Supplementary Figure 3).
244
245

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.

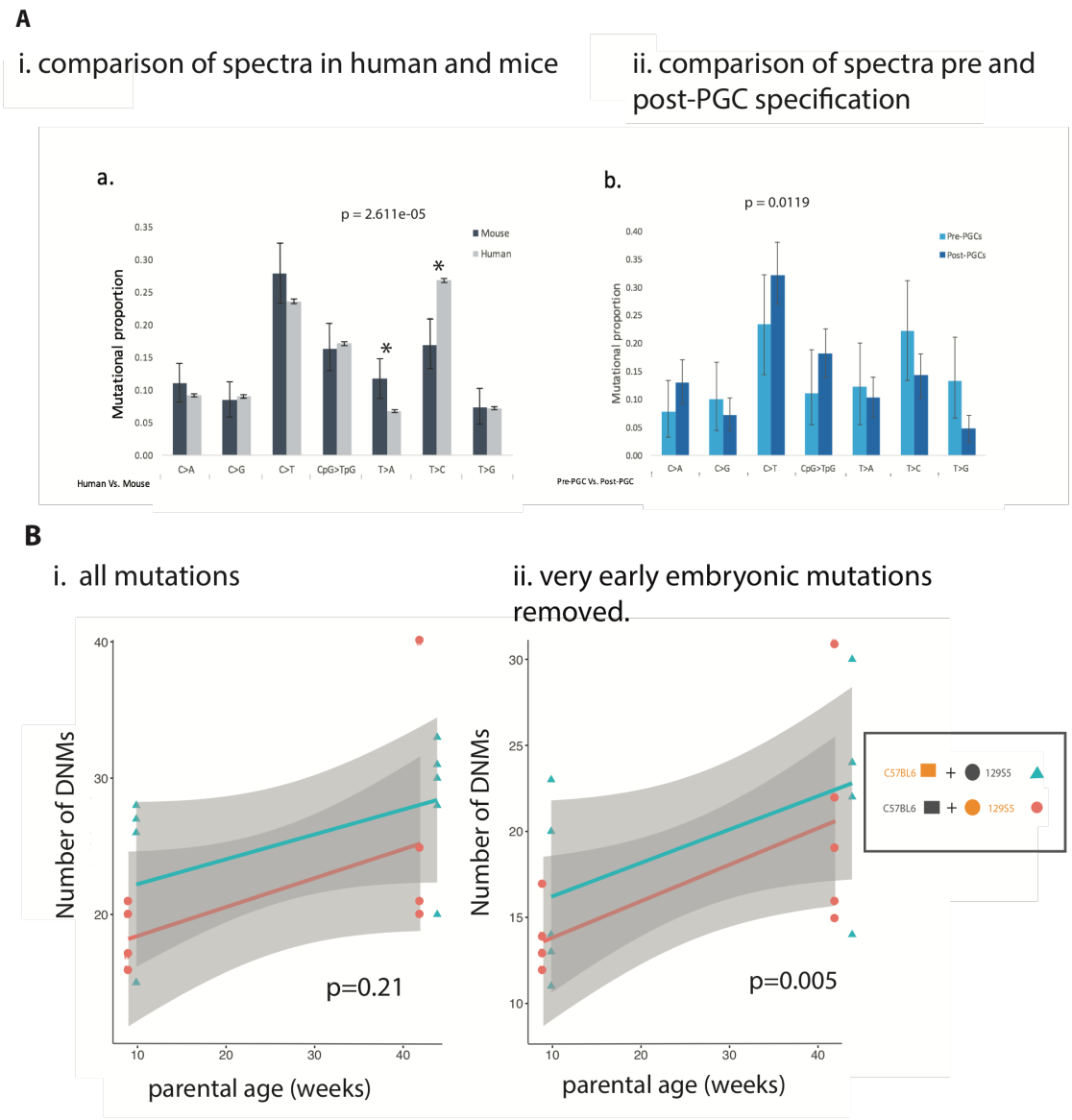


Figure 4: Plot showing the effect of parental age on the number of DNMs observed in each individual before (a) and after (b) the removal of very early embryonic mutations occurring in the offspring. (c) Comparison of mutational spectra in mice and humans using catalogue of compiled DNMs in humans as in Rahbari R[9]. (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).

264

265 **Parental age effect**

266

267 We observed an average increase of 6 DNMs over the 33 weeks between
268 earliest and latest mouse litters, which is 4.6 times greater than we would expect
269 in humans in the same time period [2,8,9,10]. This increase is greater than the 1.9-
270 fold increased rate of turnover of SSCs in mice compared to humans, suggesting
271 an increased mutation rate per SSC division in mice[11]. However, unlike in
272 humans, in mice parental age is not a significant predictor of the total number of
273 DNMs per offspring, either within each pedigree individually p=0.11 and 0.13) or
274 across both combined (p=0.21) (Figure 4B(i), Supplementary Table 1). This is
275 due in part to the lower number of mutations resulting in lower power to detect a
276 parental age effect. However, VEE mutations represent a large proportion of all
277 DNMs in mice, and yet we might expect only pre-zygotic mutations to be
278 influenced by parental age. Accordingly, we found that parental age was a
279 significant predictor of the total number of pre-zygotic DNMs across both
280 pedigrees (p=0.005)(Figure 4B(ii). As in humans, the parental age effect in mice
281 appears to be predominantly paternally driven, as pre-zygotic mutations exhibit
282 the greatest paternal bias (4.7:1 compared to 2.6:1 overall) and the ratio of
283 paternal mutations to maternal mutations is higher in offspring in later litters
284 compared to earlier litters.

285

286 **Comparing stage-specific mutation rates in mice and humans**

287

288 We calculated and compared mutation rates per cell division at different
289 phases of the germline in both mice and humans (Figure 5), by integrating
290 information on the known cellular demography of the germline in mice and
291 humans[11], the strength of the paternal age effects, and the numbers of mutations
292 arising in each temporal strata from our pedigree studies.

293

294 We observed that mutation rates per cell division are highest in the first
295 cell division of embryonic development than at any other germline stage, in both
296 humans (8X higher than average) and mice (9X higher than average). This
297 observation is supported by previous murine studies in which mosaic mutations
298 causing visible phenotypes were strongly enriched for mutations present in 50%
299 of cells [14].
300
301 The mutation rate per cell division during SSC turnover (post-puberty) is
302 considerably lower in humans than in mice (Figure 5). Moreover, in mice the
303 mutation rate per SSC division is only two-fold lower than during pre-pubertal
304 divisions, whereas in humans the concomitant reduction in mutation rate is ten-
305 fold. This discordance likely explains the marked difference in humans between
306 average germline mutation rates per cell division in males and females (Figure
307 5), whereas in mice the average mutation rates in the maternal and paternal
308 germline are much more similar. It is likely that the disproportionate contribution
309 of SSC divisions to the human germline (due to the lag between puberty and
310 average age at conception) has led to stronger selection pressures to reduce the
311 mutation rate per cell division in SSCs in humans than in mice.

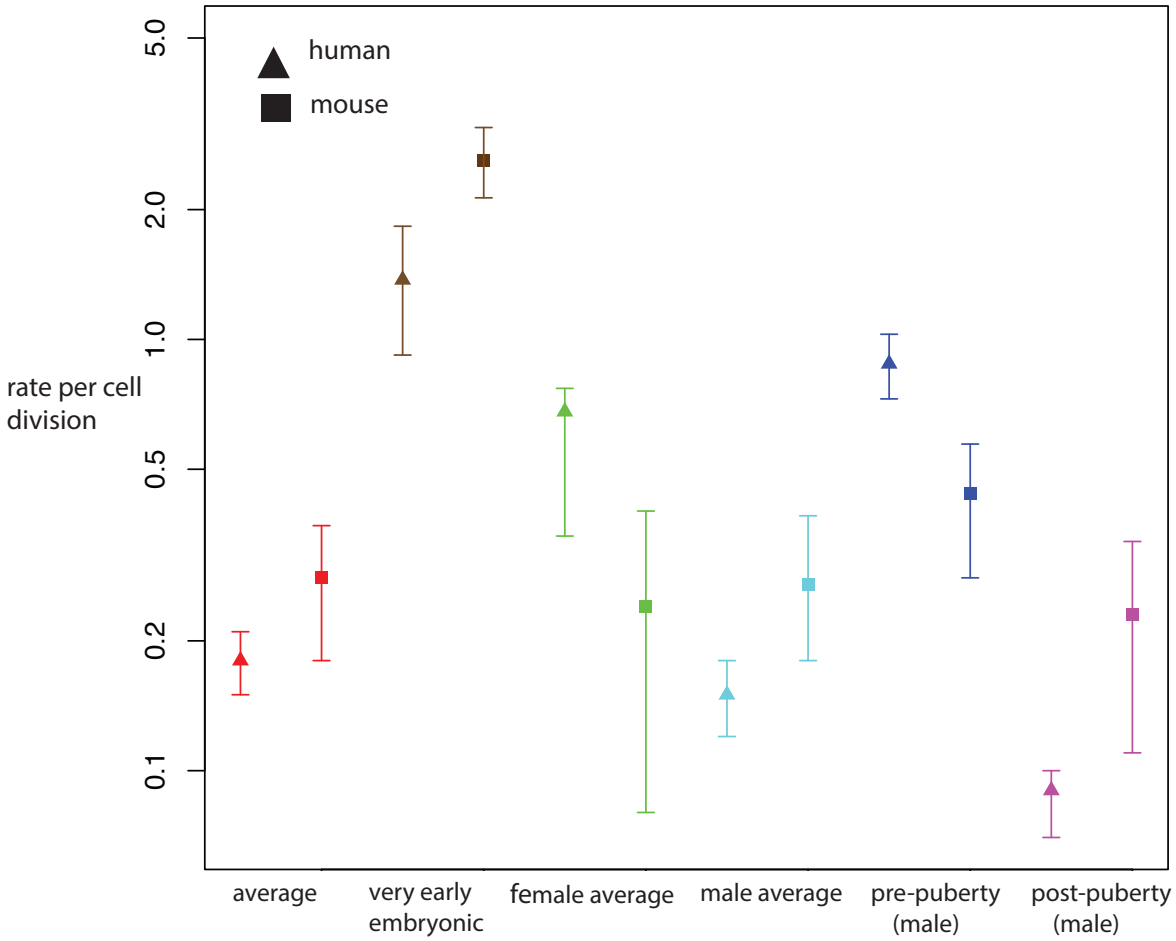


Figure 5: Estimation of mutation rates per cell division; species average in red, very early embryonic in brown, female average in green, male average in blue, and male pre and post puberty in dark blue and pink respectively. A description of how these were calculated can be found in the methods section.

Reconstruction of mouse geneaologies

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

325 by chance. Conversely, we would expect two shared mutations arising on
326 different cellular lineages in the same parent to be observed in mutually exclusive
327 sets of offspring. Finally, two shared mutations arising in different parents would
328 be expected to observed in the same offspring at random. Therefore, we
329 reconstructed four cellular genealogies, one for each parent, using an iterative
330 procedure to cluster shared mutations into lineages based on their correlation
331 across offspring, constrained by parental origin (see Methods).
332
333 Using this iterative clustering procedure, we assigned 67/71 shared
334 mutations to a specific parent, and defined partial cellular genealogies for each
335 parent (Figure 6). Each parental genealogy is characterised by 2-4 lineages
336 defined by early embryonic and peri-PGC mutations, and a residue of offspring
337 without shared mutations (representing 13-55% of all offspring). These primary
338 lineages are distributed randomly with respect to litter timing, suggesting that
339 their relative representation among gametes is stable over time and primarily
340 reflects processes operating prior to PGC specification and/or during the early
341 stages of PGC proliferation. We noted markedly unequal contributions from
342 different lineages, with individual lineages defined by early embryonic or peri-
343 PGC mutations accounting for 2-54% of offspring from a breeding pair. It has
344 been estimated that 6 cell lineages are set aside during mouse development
345 which later go on to specify 40-42 PGCs[17,18,22]. In principle, over-represented
346 lineages could have arisen from having begat multiple PGC founders, or from
347 relative fecundity during early PGC proliferation. The correlation between levels
348 of somatic mosaicism and germline mosaicism suggests that the former can be a
349 contributing factor, whereas the observation that the most over-represented
350 lineage (M2) is only defined by peri-PGC mutations, and the presence of major
351 sublineages defined by later peri-PGC mutations, suggests that lineage birth-
352 death during early PGC proliferation can also play a major role. These results
353 indicate that specified PGCs do not contribute equally to the final pool of
354 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.

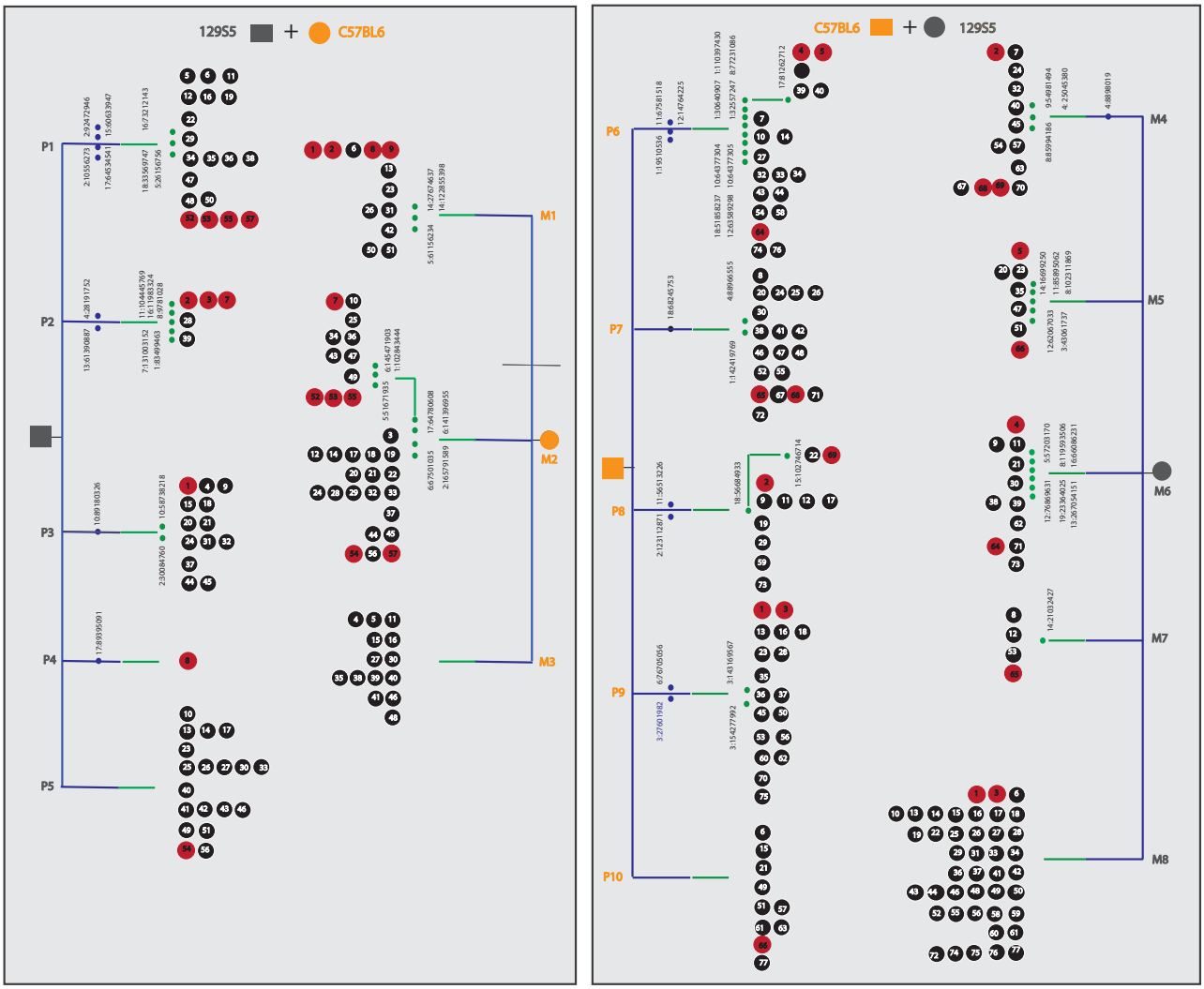


Figure 6: Lineage reconstructions showing reconstruction of putative maternal and paternal cell lineages using early embryonic and peri-PGC mutations. Individual offspring are numbered and coloured by litter.

367 **Conclusions**

368

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

380

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

391

392 Our finding that generational mutation rates in mice are lower than in
393 humans while per division mutation rates are higher, raises an apparent paradox:
394 if purifying selection in mice is more efficient at reducing generational mutation
395 rates, why does the murine cellular machinery have lower fidelity per genome
396 replication? The answer likely lies in the expectation that the selection coefficient

397 of an allele that alters the absolute fidelity of genome replication will depend
398 critically on the number of genome replications per generation. Thus, given the
399 much greater number of genome replications in a human generation, an allele
400 that alters the fidelity of genome replication by a given amount will have a
401 considerably higher selection coefficient in humans than in mice. The reduction in
402 mutation rate in SSC divisions compared to previous cell divisions was far more
403 pronounced in humans than in mice. This is presumably as a result of stronger
404 selective pressures in humans due to the much greater contribution of this class
405 of genome replication to the overall number of genome replications in the
406 germline.
407
408 Much of the existing literature comparing germline mutation processes
409 between species focuses on the dependence of these processes on 'life history'
410 traits[25,26]. We contend that these 'life history' traits are imperfect proxies for the
411 true molecular and cellular basis of this variation between species, which relates
412 to the number of different classes of cell division within the germline, and the
413 mutation rates and spectra accompanying each temporal strata of the germline.
414 Broader application of the kinds of analyses performed here will catalyse the
415 transition from a demographic understanding of germline mutation towards a truly
416 molecular comprehension.
417

418 **Online Methods**

419

420 **Mice.**

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

435

436 **DNA Sample Preparation and QC.**

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

444

445 **Sequencing and variant calling.**

446 DNA extracted from the spleen of parents and offspring was sequenced using
447 standard protocols and Illumina HiSeq technologies. The resultant sequence

448 data was aligned to mouse reference GRCm38. The total mapped coverage after
449 duplicate removal had a mean of 25X and range 22-35X for CBGP, and 29X and
450 22X-40X for GPCB. Variants were called using bcftools and samtools and
451 standard settings[27].
452
453 **De novo mutation calling.**
454 De novo mutations were called on the variants supplied by bcftools by using
455 *DeNovoGear* version 0.5 using standard settings[28]. *DeNovoGear* called between
456 7711 and 11069 (mean 9736) short indels and SNVs in CB trios, and between
457 8578 and 12835 (mean 10916) candidates in GP trios respectively. Calls from
458 the X chromosome were discarded as SNVs and indels showed a strain/sex
459 specific inflation, for which it was not possible to correct for.
460
461 **Filtering of candidate de novo mutations.**
462 Candidate de novo mutations were filtered to exclude sites highly enriched for
463 false positives (simple sequence repeats (2% of sites on average), segmental
464 duplications (0.5% of sites on average), although these sites are not exclusive of
465 each other. In addition, strain-specific mapping artefacts (low quality areas
466 leading to clustered/low quality SNV/indel candidates were filtered by removing
467 sites that had a high alternative allele ratio (>0.2) in any pup in the reciprocal
468 (unrelated litter), or parent of reciprocal (unrelated) litter (>0.04). Assuming a
469 Poisson distribution for sequencing depth, sites with a depth greater than the
470 0.0001 quantile were removed due to the likelihood of mapping errors or low
471 complexity repeats introducing false positives (generally 13% of candidate sites).
472 Candidate sites where the de novo mutation was present in either parent in
473 greater than 5% of reads and where there were known SNPs in the parental
474 strain were also removed on the grounds that they were likely to be inherited (on
475 average, 79% of sites). Once these filters were applied, 272, 380, 225, 260, 205,
476 324, 166, 286, 284, 375 and 211, 174, 180, 346, 135, 101, 160, 143, 191, 300
477 candidate de novo mutations remained for CBGP and GPCB offspring

478 respectively.

479

480 **Experimental validation of de novo mutations.**

481 A total of 4460 unique sites across all 20 offspring were put forward for validation
482 by Agilent Sure Select Target Enrichment. Twenty-one sites were lost during
483 liftover conversion, leading to 4439 sites put forward for bait design. Bait design
484 included 2X tiling, moderate repeat masking, maximum boosting, across 100bp,
485 of sequence flanking the site of interest (extending to 200bp where baits could
486 not be designed on the initial attempt. Of these 4439 sites, 3253 sites were
487 successfully designed for with high coverage (>50% coverage), 222 with medium
488 coverage (>25% coverage), and 421 with low coverage (<25% coverage). 564
489 sites failed bait design, however, our previous analyses have showed that sites
490 that fail bait design are enriched for false positives. Initially, the target enrichment
491 set was run (2 lanes of 75bp PE Hiseq) on DNA extracted from the spleen of the
492 20 offspring subject to WGS and their parents, leading to an average of 300X
493 across each site. A subsequent run (5 lanes of 75bp PE Hiseq) was carried out
494 with tissues from the parents' kidney, tail and spleen, the WGS-sequenced
495 offspring spleen and tail, and the spleen from all the additional offspring from the
496 breeding pairs, leading to an average of 400-800X coverage for each site in
497 parental tissues, and an average of 200X coverage in offspring tissues. The
498 resultant sequence data were merged by individual and annotated with read
499 counts at the candidate site using an in-house python script. An in house R script
500 (http://www.Rproject.org) was then used to allocate a likelihood to each
501 candidate variant being a true de novo mutation, an inherited variant or a false
502 positive call, based on the allele counts of the parents and offspring at that locus.
503 A proportion of the SNV candidates (all sites put forward for validation for one
504 individual) as well as all of the indel candidates were reviewed manually using
505 Integrative Genomics Viewer (IGV)[29].

506

507 **Functional Annotation of variants**

508 Functional annotation of DNMs was carried out using ANNOVAR[30].

509

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

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

521

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

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

533

534 **Per generation mutation rate estimation.**

535 We calculated a mutation rate for autosomal SNVs in each individual as follows:
536 first, we calculated the proportion of the genome not covered in our analysis
537 because of the depth of the whole-genome sequencing: Bedtools[31] was used to

538 calculate the proportion of the genome not considered in our analysis due to low-
539 or high-sequence depths for each individual (mean 5.6%). We then calculated
540 the proportion of sites that were removed by our whole-genome filters (simple
541 sequence repeats and segmental duplications) after the depth filters were
542 applied (average 2.1%). Last, we used the posterior probability supplied by
543 *DeNovoGear* (>0.9) to calculate what proportion of sites that were not validated
544 (failed validation or removed by to filters), were likely to be true de novo
545 mutations. For human/mouse comparisons, generation times were assumed to
546 be 30 years and 9 months respectively. According to Drost[11], this would result in
547 ~432 cell divisons in the human germline, and ~87 cell divisions in the mouse
548 (paternal and maternal combined).
549
550 **Identification of very early embryonic mutations in offspring.**
551 We aggregated the alternate allele counts and total depths between tissues, after
552 testing that the allele ratios were concordant across tissues (Fishers Exact test).
553 Very early embryonic mutations (defined as occurring after in the individual after
554 fertilization, and therefore private to that offspring), was classified as follows :
555 A likelihood-based test was then carried out on the combined counts to test the
556 hypothesis that the alternate allele count was suggestive of a constitutive
557 (binomial p=0.5) or a VEE origin (binomial p=0.25), where a site with log
558 likelihood difference of >5 was designated VEE, <-5 was designated constitutive,
559 or unassigned if it falls between those values. Due to lower coverage, for 10% of
560 mutations in human pedigrees, and 4% in mouse pedigrees, we were unable to
561 confidently infer whether the mutations were constitutive or very early embryonic.
562
563 In addition, haplotype occupancy (HO) was ascertained where possible; the
564 nearest heterozygous variant to the de novo mutation should phase consistently
565 100% of the time for a zygotic (constitutive) mutation, whereas for a very early
566 embryonic mutation, the de novo allele mutation only be seen on a proportion of
567 haplotypes defined by the nearest variant. (Supplementary Figure 3). The HO for

568 mouse and human DNM sites was plotted against the alternate allele proportion;
569 this showed that, where HO could be determined, sites with a low alternate allele
570 ratio were enriched for sites with low HO, whereas shared sites that are
571 constitutive by definition only have high HO.
572
573 **Reconstruction and testing of parental lineages.**
574 Parental lineages were reconstructed using the distribution of mutations shared
575 between offspring, using the following expectations: Shared mutations that are
576 observed in the same offspring significantly more frequently than expected by
577 chance are likely to belong to the same parental lineage. Conversely, mutations
578 that are never observed together are likely to come from the same parent, but a
579 different lineage. Mutations that are shared in a random manner could come from
580 the same lineage in the same parent, or a lineage from the other parent.
581
582 In the first step, a pairwise test was carried out for each shared mutation, which
583 calculated the binomial probability of ***n*** pups sharing ***m*** mutations where the
584 frequencies of the mutations were ***p*** and ***q*** in the offspring. Then, the pair of
585 sites with the lowest resultant p-value were merged into a single pseudo site
586 containing all the offspring who have either site from the initial pair, as long as
587 the parental origin of the two mutations was not discordant. The pairwise test
588 was then repeated, followed by another merge of sites, either until a given p-
589 value threshold is reached, or the pseudo sites cannot be merged any further.
590 Given a p-value threshold of 0.05, all sites had completely collapsed into the
591 given clusters. All but four of the seventy shared mutations could be assigned to
592 either paternal or maternal lineages, the remaining mutations represent lineages
593 defined by a only single shared mutation.
594
595 The accuracy of the lineage reconstructions were tested using two simulations.
596 Firstly, for each pedigree, shared mutations were randomly re-assigned into the
597 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% conference 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:

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

626 where *scaling factor* scales the number of discovered mutations to the genome
627 wide corrected number of mutations, and where the 95% confidence intervals
628 were derived from the assumed Poisson distribution of numbers of mutations.
629 The putative numbers of paternal mutations per generation were then divided by
630 the estimated number of cell divisions per generation (62 in mice, 401 in
631 humans)[8].
632 The maternal per-generation average was calculated as above, using 25 and 31
633 cell divisions per generation (mouse and human, respectively)[11].
634
635
636 *Very Early Embryonic Mutations*
637
638 Very early embryonic mutations occur in the first cell divisions that contribute to
639 the embryo (rather than to extra-embryonic tissues). Assuming the founding cells
640 in the inner cell mass (ICM) of the blastocyst divide symmetrically, these
641 mutations occur in one or two consecutive cell divisions in the first two cells to
642 eventually comprise the embryonic tissues. We can only observe these in the
643 offspring; recovery of very early embryonic mutations that occur in the parents
644 will have been filtered as putative inherited variants. In addition, we can only
645 capture two symmetrical cell divisions at most; once the frequency of cells
646 carrying the alternate allele below falls 25% it is unlikely to be recovered during
647 de novo calling when WGS coverage is ~25X. We identified this class of mutation
648 arising in offspring using several different methods (Methods). As we are
649 estimating the rate from the offspring, we use the sex of the offspring rather than
650 haplotypes from the parents to define relative contributions by sex.
651 With 25X coverage for the WGS discovery phase, the vast majority of the VEE
652 mutations we detect will be from a single cell division. Modelling shows that our
653 mutation calling pipeline had very low power to detect VEE mutations in
654 subsequent cell divisions. In addition, the distribution of the alternate allele
655 proportion for VEE mutations is centred symmetrically around 0.25 as would be

656 expected for mutations arising in the first cleavage cell division contributing to the
657 embryo. These results suggest that the majority of VEE mutations we detected
658 arose in a single cell division (Supplementary Figure 3).
659
660 To estimate the VEE mutation rate per cell division we took the total number of
661 mutations that we determined to be VEE (104 in mice, 33 in humans), and
662 calculated the 95% Poisson confidence interval around this count. We then
663 divided this number by 2 (to obtain a haploid rate), and then by the total number
664 of offspring (20 for mouse, 12 for human).
665
666 The power to identify this class of mutation is based on WGS sequencing depth,
667 and the power to correctly discriminate it from a constitutive mutation is based on
668 validation sequencing depth. At ~100X sequencing coverage, we have 97%
669 power to correctly infer this class of mutation, and we have similar power to
670 detect this class of mutations in humans and mice.
671
672 *Pre-puberty in the male germline*
673
674 The total number of mutations occurring pre-puberty in the male germline were
675 defined as follows :
676
677 $N = mutations_{mean} - (age_{mean} - age_{puberty}) \times annual\ mutations_{mean}$
678
679 95% Poisson confidence intervals were derived from the mean number of
680 mutations per year.
681
682 *Post-puberty in the male germline*
683
684 As parentally-aged induced mutations accrue in an approximately linear manner,
685 the post-puberty mutation rate in males was calculated on the number of

686 mutations accrued in the mouse and human paternal germline in a single year.
687 The average number of mutations in mice increased by 6 over a 33 week
688 timespan, leading to an extrapolated annual increase of 9.45 mutations. The
689 largest human study to date suggests an increase of 2.01 mutations per year[2].
690 The annual number of mutations was divided by the annual number of cell
691 divisions occurring in that organism (42 for mice, 23 for humans[8]). Confidence
692 intervals were derived from the uncertainly of the slope of the linear models of
693 effect of age on number of mutations (estimates for human obtained from Kong
694 et al[2]).
695
696 **Analysis of mutation spectra**
697
698 Mutational spectra were derived directly from the reference and alternative (or
699 ancestral and derived) allele at each variant site. The resulting spectra are
700 composed of the relative frequencies of the six distinguishable point mutations
701 (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
702 differences between mutational spectra was assessed by comparing the number
703 of the six mutation types in the two spectra by means of a Chi-squared test (df =
704 5).
705
706 **Estimation of recurrence risk of DNMs in offspring**
707 The probability of an apparent DNM being present in more than one sibling in the
708 same family was calculated as the number of instances of a mutation being
709 shared by two siblings divided by the number of pairwise comparisons between
710 two siblings in both pedigrees
711
712 **Possibility of technical artefacts.**
713
714 We considered and discounted a wide variety of possible technical artefacts that
715 might explain the apparent parental sex bias we observe in early embryonic

716 mutations in mice. Firstly, sequencing depth, and thus power to detect somatic
717 mosaicism, was equal between maternal and paternal tissues, and the identity of
718 the WGS samples were checked using strain and gender specific SNPs.
719 Secondly, where parental origin could be independently determined by
720 haplotyping with nearby informative sites (N=6), the parental origin was
721 confirmed, thus excluding sample swaps. Thirdly, parental mosaicism was
722 supported by very low read counts in the WGS data in the parents at 6 of the
723 mosaic sites (2 and 3 sites from both fathers, and one from the mother). Fourth,
724 the same aliquot of DNA was used for WGS and validation of mutations in
725 parental spleen, lowering the possibility of sample swaps. Lastly, in all cases,
726 parental mosaicism was independently supported by sequencing data from two
727 additional tissues.
728
729
730
731

732 **Supplementary**:

733

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

737

738 **Supplementary Table 2:** DNMs with potentially functional consequences as
739 given by ANNOVAR[40] are listed.

740

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

747

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

760

761 **Supplementary Figure 2**. Histograms of the proportion of alternate allele in
762 validated DNMs in high depth sequence data in humans (A) and mice (B). Sites
763 in red are constitutive and are have an alternate allele proportion centred around
764 50% of reads (100% of cells). Sites classified as very early embryonic are shown
765 in blue, are found in around 25% of reads (50% of cells). Red, blue and black
766 lines show the expected distribution of alternate allele proportions given a
767 binomial distribution of reads centred around constitutive, first division and
768 second division mutations, in our high depth sequence data.
769
770 **Supplementary Figure 3**
771 Low resolution mutation spectra in maternal and paternally derived DNMs in
772 mouse and human data. Error bars show the 95% confidence intervals.
773

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 :

1.Conrad, D F et al, Variation in genome-wide mutation rates within and between human families Nature Genetics 43:712-714 (2011)

2. Kong A, et al Rate of de novo mutations and the importance of father's age to disease risk Nature 488:471-475 (2012)

3.Scally A, and Durbin, R, Revising the human mutation rate:implications for understanding human evolution. Nature Reviews Genetics 13:745-753 (2012)

4. Adewoye, A G, et al The genome-wide effects of ionizing radiation on mutation induction in the mammalian germline. Nature Communications 6:6684 | DOI: 10.1038/ncomms7684 | (2015)

5. Uchimura A, et al Germline mutation rates and the long term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. Genome Research 25 1125-1134 (2015)

6. Drake, JW et al. Rates of Spontaneous Mutation Genetics 148: 1667-1686 (1998)

7. Lynch, M, Evolution of the Mutation Rate Trends in Genetics 26:345-352 (2010)

8. Goldmann, J M et al, Parent-of-origin-specific signatures of de novo mutations. Nature Genetics Published online 20 June (2016)

9.Rahbari R, et al, Timing, rates and spectra of human germline mutation. Nature Genetics 48:126-33 (2016)

10. Wong et al, Nature Communications 7 New observations on maternal age

811 effect on germline *de novo* mutations Nature Communications
812 l 7:10486 l DOI: 10.1038/ncomms1048 (2016)
813
814 11. Drost J and Lee W, Biological Basis of Germline Mutation ,Comparisons of
815 Spontaneous Germline Mutation Rates Among Drosophilia, Mouse and Human.
816 Environmental and Molecular Mutagenesis 25, Supplement 26:48-64 (1995)
817
818 12. Russell L, Effects of male germ-cell stage on the frequency, nature, and
819 spectrum of induced specific-locus mutations in the mouse Genetica 122: 25–
820 36, (2004).
821
822 13. Russell, B, Significance of the Perigametic Interval as a Major
823 Source of Spontaneous Mutations That Result in Mosaics Environmental and
824 Molecular Mutagenesis 34:16‹23 (1999)
825
826 14. Russell L, and Russell W, Spontaneous mutations recovered as mosaics in
827 the mouse specific-locus test Proc. Natl. Acad. Sci. USA Vol. 93, pp. 13072–
828 13077 (1996)
829
830 15. Gao J et al, Pattern of Mutation Rates in the Germline of
831 Drosophila melanogaster Males from a Large-Scale Mutation Screening
832 Experiment, Genes, Genomes, Genetics 4:1503-1514 (2014)
833
834 16. Mouse Genome Sequencing Consortium Initial sequencing and analysis of
835 the mouse genome. Nature 420:520-561 (2002)
836
837 17. Lawson, K A and Hage W J, Clonal analysis of the origin of primordial germ
838 cells in the mouse. Germline Development. Wiley, Chichester (Ciba Foundation
839 Symposium 182) 68-91(1994)
840

841 18. Saitou M and Yamaji , Primordial Germ Cells in Mice. Cold Spring Harb,
842 Perspect Biol 4:a008375 (2012)
843
844 19. Ehmcke J, Wistuba J, Schlatt S, Spermatogonial stem cells: questions,
845 models and perspectives Human Reproduction Update, Vol.12, No.3 pp. 275–
846 282, (2006)
847
848 20. Burgoyne P, S et al The genetic basis of XX-XY differences present before
849 gonadal sex differentiation in the mouse. Phil. Trans. R. Soc. Lond B. 350 253-
850 261 (1995)
851
852 21. Tan K, et al, IVF affects embryonic development in a sex-biased manner in
853 mice Reproduction 151 443–453 (2016)
854
855 22. De Felici, M. Origin, Migration, and Proliferation of Human Primordial
856 Germ Cells Oogenesis pp19-37 Springer Press (2012)
857
858 23 Bedzhov I, et al Developmental plasticity, cell fate specification and
859 morphogenesis in the early mouse embryo. Phil. Trans. R. Soc. B 369: 20130538
860 (2014)
861
862 24 Mihajlovic AI, Thamodaran V, Bruce AW, The first two cell fate decisions of
863 preimplantation mouse embryo development are not functionally independent.
864 Nature Scientific Reports 5:15034 (2016)
865
866 25 Amster, G, Sella, G Life history effects on the molecular clock of autosomes
867 and sex chromosomes. PNAS 113:6 1588-1593 (2016)
868

869 26 Gao Z, Wyman MJ, Sella G, Przeworski M Interpreting the Dependence of
870 Mutation Rates on Age and Time. PLoS Biol 14(1): e1002355.
871 doi:10.1371/journal.pbio.1002355 (2016)
872
873 27. Rooij, D and Li, H. *et al*. The Sequence alignment/map (SAM) format and
874 SAMtools. Bioinformatics 25, 2078–2079 (2009).
875
876 28. Ramu, A. *et al*. DeNovoGear: *de novo* indel and point mutation discovery and
877 phasing. Nat. Methods 10, 985–987 (2013)
878
879 29. James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell
880 Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. **Integrative Genomics**
881 **Viewer**. Nature Biotechnology 29, 24–26 (2011)
882
883 30. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of
884 genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38,
885 e164 (2010) .
886
887 31. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing
888 genomic features. Bioinformatics 26, 841–842 (2010) .
889
890
891
892
893
894
895
896
897
898

Supplementary Tables and Figures

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

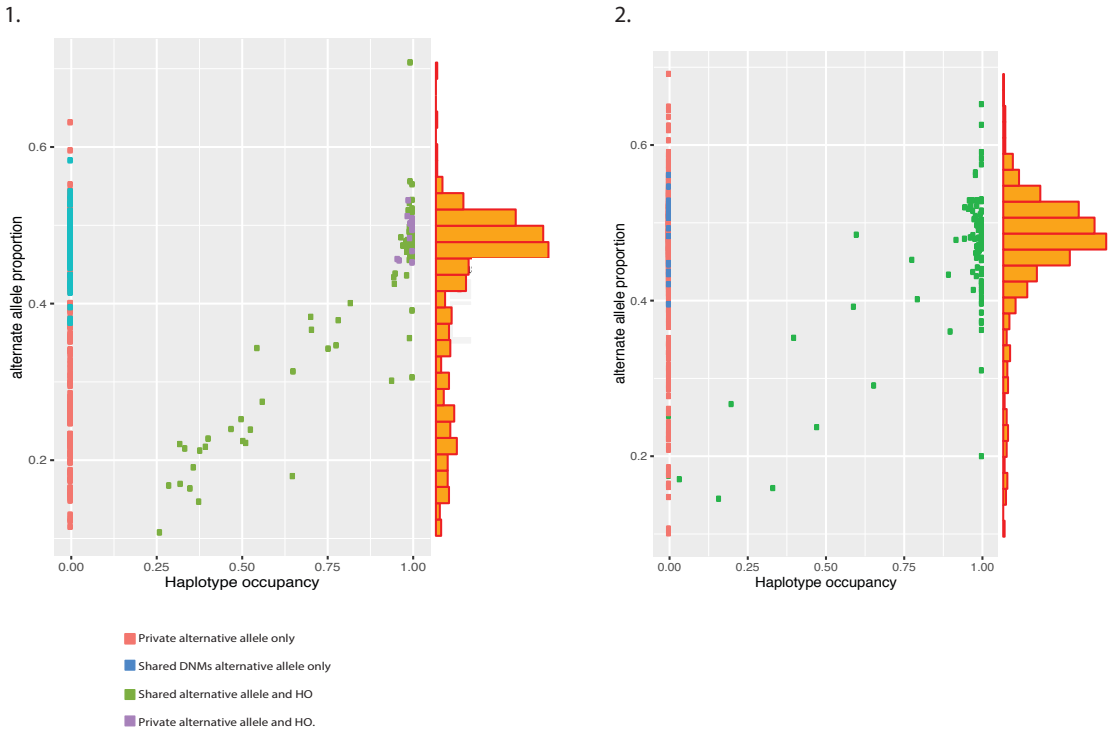
IID	early embryonic	peri_PGC specification	late post-PGC specification	very early embryonic	indels
CBGP8_1a	1	3	7	5	1
CBGP8_1b	0	5	6	7	0
CBGP8_1c	2	5	8	4	0
CBGP8_1g	1	6	4	5	0
CBGP8_1h	1	1	11	2	0
CBGP8_8a	1	3	14	1	1
CBGP8_8b	2	3	8	5	0
CBGP8_8c	0	2	13	5	1
CBGP8_8d	3	5	19	9	0
CBGP8_8f	4	5	11	3	0
GPCB2_1a	1	0	11	14	1
GPCB2_1b	1	4	5	4	0
GPCB2_1c	1	1	15	8	0
GPCB2_1d	1	9	11	3	0
GPCB2_1e	1	7	6	1	1
GPCB2_9a	2	8	14	4	2
GPCB2_9b	1	2	10	6	1
GPCB2_9c	0	5	16	7	0
GPCB2_9e	2	3	24	0	2
GPCB2_9f	2	3	15	11	1

Supplementary Table 2: Functional consequences of de novo mutations

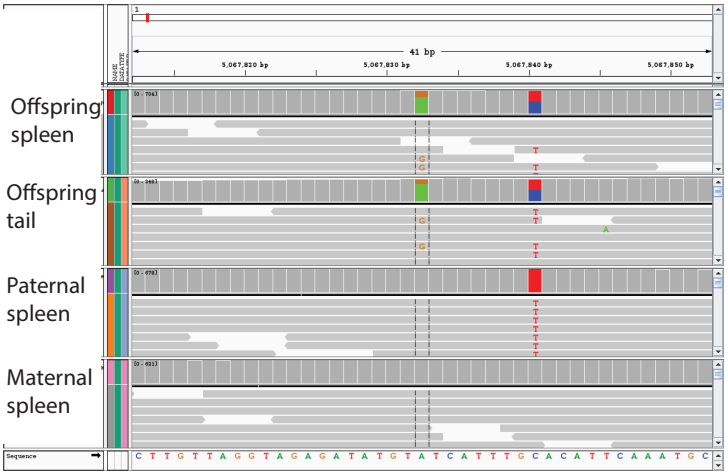
ID	chr	position	ref	alt	site	consequence	gene	mutation class
CBGP8_8b	2	10556273	G	A	exonic	synonymous SNV	Sfmbt2	paternal early embryonic
CBGP8_8f	2	28556769	A	T	exonic	nonsynonymous SNV	Cel	late post-PGC specification
CBGP8_1a	2	30084760	C	T	exonic	nonsynonymous SNV	Pkn3	paternal peri-PGC specification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

Supplementary Figure 1

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



b) definition of haplotype occpancy



921

922

923 **Supplementary Figure 1**

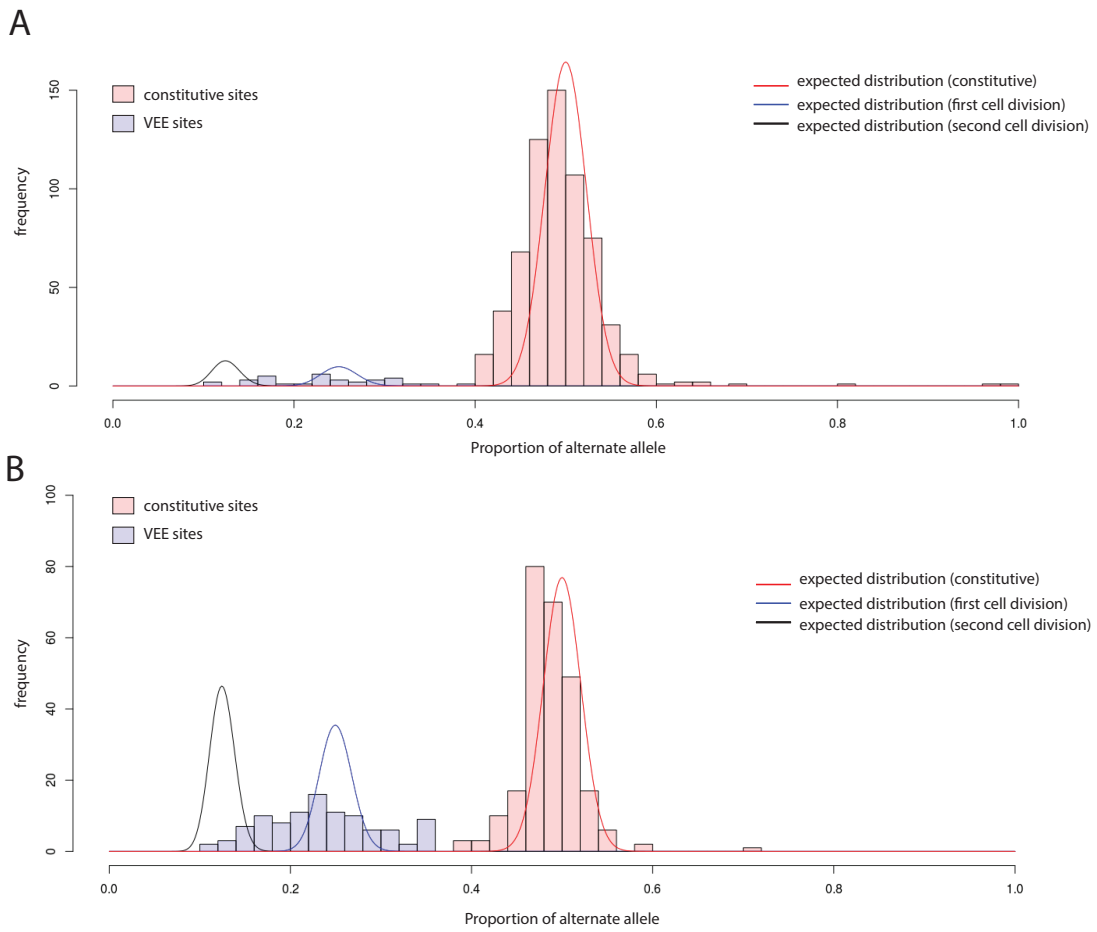
924 Plots showing haplotype occupancy in heterozygous sites directly adjacent to de novo sites
925 plotted against the alternate allele proportion at the validated site. The histogram shows the
926 distribution of individuals along the y axis. It can be observed that the mouse DNM sites that are
927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

933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952

953 **Supplementary Figure 2**

954

955



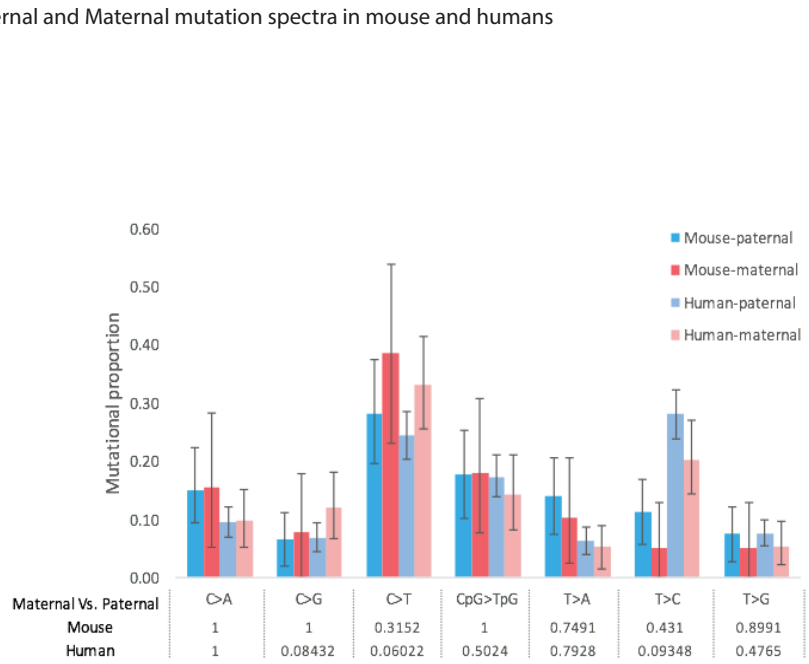
956

957

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

965

966 **Supplementary Figure 3**



967

968

969 **Supplementary Figure 3**

970 Low resolution mutation spectra in maternal and paternally derived DNMs in mouse and human

971 data. Error bars show the 95% confidence intervals.

972