1 **Frequency of mosaicism points towards mutation-prone early cleavage cell**
2 **divisions in cattle.**

3 *Chad Harland[1#], Carole Charlier[1#], Latifa Karim[2], Nadine Cambisano[2], Manon*
4 *Deckers[2], Myriam Mni[1], Erik Mullaart[3], Wouter Coppieters[2], Michel Georges[1].*

5 [1] Unit of Animal Genomics, GIGA-R & Faculty of Veterinary Medicine, University
6 of Liège, Belgium. [2] GIGA Genomics Platform, University of Liège, Belgium. [3] CRV,
7 Arnhem, The Netherlands.

8 [#] *Contributed equally to this work*

9 Correspondence: michel.georges@ulg.ac.be

10

11 **It has recently become possible to directly estimate the germ-line de novo**
12 **mutation (*dnm*) rate by sequencing the whole genome of father-mother-**
13 **offspring trios, and this has been conducted in human[1-5], chimpanzee[6],**
14 **mice[7], birds[8] and fish[9]. In these studies *dnm*'s are typically defined as**
15 **variants that are heterozygous in the offspring while being absent in both**
16 **parents. They are assumed to have occurred in the germ-line of one of the**
17 **parents and to have been transmitted to the offspring via the sperm cell or**
18 **oocyte. This definition assumes that detectable mosaïcism in the parent in**
19 **which the mutation occurred is negligible. However, instances of**
20 **detectable mosaïcism or premeiotic clusters are well documented in**
21 **humans and other organisms, including ruminants[10-12]. We herein take**
22 **advantage of cattle pedigrees to show that as much as ~30% to ~50% of**
23 ***dnm*'s present in a gamete may occur during the early cleavage cell**

24 **divisions in males and females, respectively, resulting in frequent**
25 **detectable mosaïcism and a high rate of sharing of multiple *dnm*'s between**
26 **siblings. This should be taken into account to accurately estimate the**
27 **mutation rate in cattle and other species.**

28 To study the process of *dnm*'s in the cattle germ-line, we sequenced the whole
29 genome of 54 animals from four pedigrees. Grand-parents, parents and offspring
30 (referred to as probands) were sequenced at average 26-fold depth (min = 21),
31 and grand-offspring at average 21-fold depth (min = 10). The source of DNA was
32 venous blood for females and sperm for males. The genome of one male proband
33 (Pr 1) was sequenced both from semen (26-fold depth) and blood DNA (37-fold
34 depth) (Figure 1A).

35 Using the standard definition, we identified 190 candidate *dnm*'s as variants that
36 were (i) detected in a proband, (ii) absent in both parents (and grand-parents
37 when available), (iii) transmitted to at least one grand-offspring, and (iv) not
38 previously reported in unrelated individuals from the 1,000 Bulls project[13]
39 (Suppl. Figure 1&2 and Suppl. Table 1). For confirmation, we developed
40 amplicons spanning 113 candidate *dnm*'s and sequenced them at average depth
41 of ~2,187 in the 54 animals plus 55 relatives (Figure 1A). This confirmed the
42 genuine nature of 110/113 variants, demonstrating the excellent specificity of
43 our bioinformatics pipeline. The three remaining ones were also detected in one
44 of the parents (although not in the grand-parents) in the confirmation, and
45 momentarily ignored.

46 We first examined what proportion of *dnm*'s detected in a proband might
47 actually have occurred during its development rather than being inherited via

48 the sperm or oocyte. An unambiguous distinction between the two types of
49 *dnm*'s is their degree of linkage with either the paternal or maternal haplotype
50 upon transmission to the next generation (i.e. the grand-offspring in Figure 1A).
51 *Dnm*'s that have occurred in the germ-line of the sire will show *perfect* linkage
52 with the proband's paternal haplotype in the grand-offspring (i.e. always
53 transmitted with the paternal haplotype, never transmitted with the maternal
54 haplotype), while *dnm*'s that have occurred in the germ-line of the dam will show
55 *perfect* linkage with the proband's maternal haplotype in the grand-offspring. On
56 the contrary, *dnm*'s that have occurred during the development of the proband
57 will be in *complete* (but imperfect) linkage with either the paternal or maternal
58 haplotype (i.e. sometimes transmitted with the maternal haplotype, never
59 transmitted with the paternal haplotype, or sometimes transmitted with the
60 paternal haplotype, never transmitted with the maternal haplotype)(Suppl.
61 Figure 1). Across the four pedigrees, 124 variants were in perfect linkage with
62 the paternal haplotype, 32 in perfect linkage with the maternal haplotype, 10 in
63 complete (but imperfect) linkage with the paternal haplotype and 21 in complete
64 (but imperfect) linkage with the maternal haplotype (Figure 1B). If the 10+21
65 *dnm*'s indeed occurred during the development of the proband rather then being
66 inherited from the sire or dam, the *dnm* dosage (defined as the proportion of
67 reads spanning the *dnm* site that carry the mutant allele) is expected to be < 50%
68 in the proband but equal to 50% in the grand-offspring inheriting the *dnm*. The
69 mean dosage was 0.26 in the proband, and 0.52 in the grand-offspring, and this
70 difference was highly significant ($p < 10^{-6}$). The corresponding means were 0.48
71 and 0.49 ($p = 0.40$) for the 124+32 mutations showing perfect linkage with
72 either the paternal or maternal haplotype (Figure 1C).

73 We conclude that in cattle ~17% of *dnm*'s detected in an animal using standard
74 procedures are not inherited from the sire or dam but correspond to premeiotic
75 clusters generated during the development of the individual. This is a lower
76 bound, as *dnm* will only be detected and recognized as having occurred in the
77 proband if (i) the *dnm* dosage is sufficiently high for the proband to be called
78 heterozygote, (ii) the *dnm* is transmitted to at least one grand-offspring, and (iii)
79 complete (but imperfect) linkage is demonstrated in the grand-offspring (Suppl.
80 Figure 3). We will refer to this type of *dnm*'s as Proband-Mosaic (PM), while the
81 others will be referred to as Sire-Non-Mosaic (SNM) (meaning that the sire is not
82 mosaic for a mutation transmitted via his sperm), or as Dam-Non-Mosaic (DNM)
83 (meaning that the dam is not mosaic for a mutation transmitted via her oocyte).
84 The proportion of PM (but not SNM and DNM) mutations differed significantly
85 between probands ($p = 0.004$); neither differed significantly between sexes ($p >$
86 0.30). Of particular interest, the three PM mutations of proband 1 were
87 detected in both sperm and blood DNA (Suppl. Table 1), indicating that they
88 occurred early in development (see hereafter).

89 If detectable mosaicism for *dnm*'s is common in the individual in whom they
90 occurred, requiring their absence in the DNA of the parents (as typically done)
91 will cause genuine *dnm*'s to be eliminated. We took advantage of the grand-
92 parents available in three pedigrees to recover such events as variants that were
93 (i) absent in the grand-parents, (ii) detected in either sire or dam with a dosage
94 significantly < 50% (Suppl. Table 1), (iii) transmitted to the proband with a
95 dosage of ~50%, (iv) transmitted to at least one grand-offspring with a dosage of
96 ~50%, and (v) not previously reported in unrelated individuals[13]. We will refer

97 to these types of mutations as Sire-Mosaic (SM) and Dam-Mosaic (DM),
98 respectively (meaning that the sire/dam is detectably mosaic for a *dnm*
99 transmitted via the sperm/oocyte)(Suppl. Figure 1). We detected 61 such
100 candidate events, including the 3/113 variants mentioned above (Suppl. Table 1
101 and Suppl. Figure 2). We developed amplicons for 34, and sequenced (average
102 1,498-fold depth) all 54 individuals plus 55 relatives (including ≥ 5 half-sibs of
103 the probands) (Figure 1A). We took advantage of whole genome sequence
104 information that became available for 27 half-sibs, to trace the inheritance of the
105 remaining 24 candidate variants. The ensuing data indicated that 11/61
106 candidates were genuine *dnm*'s but occurred in the germ-line of one of the
107 grand-parents rather than one of the parents (dosage ~50% in the sire or dam,
108 and perfect linkage in the half-sibs). The SM/DM status was unambiguously
109 demonstrated for 40 (dosage < 50% in the sire or dam in the confirmation,
110 transmission to half-sibs, and complete (but imperfect) linkage) and strongly
111 supported for the remaining 10 (dosage < 50% in the sire/dam in the
112 confirmation or complete (but imperfect) linkage yet without transmission)
113 (Suppl. Figure 2 and Figure 1B). Further supporting the genuine nature of the
114 SM/DM mutations, the dosage was 0.12 on average in the corresponding parent,
115 while being 0.51 in descendants ($p < 10^{-6}$) (Figure 1C).

116 For *dnm*'s that were detectably mosaic in sperm (SM and PM in male probands),
117 allelic dosage was significantly correlated with rate of transmission ($p = 0.025$)
118 and strength of linkage ($p = 0.0002$). These correlations were not significant for
119 *dnm*'s that were detectably mosaic in blood (DM and PM in female probands).
120 This suggests that the degree of mosaicism in the soma is a poor indicator of the

121 degree of mosaicism in the germ line (Figure 1D). Accordingly, the rate of
122 transmission of the three PM mutations of proband 1 to its 53 offspring was
123 better predicted by their dosage in sperm than in blood (Suppl. Figure 4).

124 Considering SNM/SM and DNM/DM mutations jointly, we conclude that on
125 average a sire is detectably mosaic (in sperm) for 29% of *dnm*'s present in a
126 sperm cell, while a dam is detectably mosaic (in blood) for 51% of *dnm*'s present
127 in an oocyte. These are lower bounds as we only considered *dnm*'s for which the
128 dosage was significantly < 0.5 in the parent (condition (ii) above). These figures
129 are possibly consistent with recent reports in the mouse (~25%)[7], but
130 considerably larger than current estimates in human (~5%)[14]. They are
131 certainly larger than expected assuming that the mutation rate per cell division
132 is uniform throughout development, and it suggests that the mutation rate is
133 higher for early cell divisions (Suppl. Figure 5). Moreover, when analyzing the
134 transmission patterns of SM and DM mutations to the half-sibs of the proband (in
135 whom the *dnm*'s were detected), we were struck by the fact that (i) >60% of half-
136 sibs share at least one *dnm* with the proband, while <50% are expected (p =
137 0.05), and (ii) half-sibs sharing multiple *dnm*'s with the proband appeared
138 surprisingly common (Suppl. Table 2). These findings also indicate that a
139 substantial proportion of *dnm*'s must occur early in development and be present
140 in the precursor cells common to the soma and germ line (Suppl. Figure 5).

141 In mammals, after fertilization, cleavage, and segregation of (i) the inner cell
142 mass from the trophoblast, (ii) the epiblast from the hypoblast, (iii) the
143 embryonic epiblast from the amniotic ectoderm, a small number of epiblast-
144 derived cells located in the wall of the yolk sac in the vicinity of the allantois are

145 induced to become primordial germ cells (PGCs). These migrate to the primitive
146 gonad where they expand and produce >1 million gametogonia. Oogonia initiate
147 meiosis prior to birth in females. Spermatogonia will resume mitotic divisions at
148 puberty allowing (i) the maintenance of a pool of stem cell like spermatogonia,
149 and (ii) sustained spermiogenesis involving ~3 additional mitotic divisions
150 followed by meiosis (Suppl. Figure 6). We simulated the process of de novo
151 mutagenesis in the male and female germ cell lineages assuming (i) uniform pre-
152 and post-natal mutation rates per cell division, and (ii) 40 PGCs sampled at
153 random from the embryonic epiblast-derived cells[15]. Pre- and post-natal
154 mutation rates were adjusted to match the observed number of mutations per
155 gamete (34 in sperm, 14 in oocytes). Under these conditions, we virtually never
156 observed the level of mosaicism, nor the sharing between sibs characterizing the
157 real data (Figure 2). We (i) increased the relative mutation rate during the early
158 cell divisions (keeping the mutation rate per gamete constant)(10 and 20-fold
159 increase during the first 4, 7, 11, 15 and 18 cell divisions; Suppl. Figure 6), (ii)
160 reduced the number of induced PGCs (4, 10, or 40), and (iii) varied the
161 relatedness between PGCs (i.e. sampled randomly amongst all embryonic
162 epiblast-derived cells or from a sub-sector)(Suppl. Figure 6). Increasing the
163 mutation rate during the very first cell divisions matched the real data much
164 better (Figure 2). To quantitatively evaluate model fitting we used (i) the
165 proportion of PM, SM and DM mutations with corresponding rate of mosaicism
166 in sperm and soma, and (ii) the proportion of sibs sharing 0, 1, 2, ... *dnm*'s with a
167 proband, to compute the likelihood of the data under different scenarios (see
168 M&M). A 20-fold increased mutation rate during the first four cell divisions,
169 combined with 4 related PGCs fitted the data best (Table 1 and Suppl. Table 2).

170 The data were ≥10^{16} times less likely under models assuming a uniform mutation
171 rate throughout development, and ≥10^{5} times less likely assuming an increased
172 mutation rate passed the 7th cell division (after segregation of inner cell mass
173 and throphoblast)(Table 1 and Suppl. Table 2).

174 When accounting for genome coverage, the estimated number of *dnm*'s per
175 gamete (SNM+SM, DNM+DM) averaged 46.6 for sperm cells and 18.1 for oocytes
176 (male/female ratio of 2.6), corresponding to an average mutation rate of
177 ~1.2×10^{-8} per base pair per gamete. Including an estimate (from the simulations)
178 of the number of missed SM (~3.3) / DM (~1.1) and misclassified PM mutations
179 (~2.9 to ~8 depending on the proband), yields an average mutation rate of
180 ~1.17×10^{-8} per base pair per gamete and a male/female ratio of 2.4. The
181 standard approach of ascertaining *dnm*'s (i.e. erroneously considering PM
182 mutations, ignoring SM and DM mutations) would have yielded a mutation rate
183 of 0.9×10^{-8} per bp per gamete, with a 2.5-fold higher mutation rate in bulls than
184 in cows.

185 Two hundred twenty of the 237 identified *dnm*'s were nucleotide substitutions,
186 17 small insertion-deletions. The non-mosaic classes of mutations (SNM and
187 DNM) were ~30-fold enriched in CpG>TpG transitions as expected. This
188 signature was also present but less pronounced for mosaic mutations (PM, SM
189 and DM). Mosaic mutations were ~2.6-fold enriched in C>A and/or G>T
190 transversions, largely due to GpCpA>GpApA and TpCpT>TpApT substitutions
191 (Figure 3). This was unlikely to be an artifact for reasons spelled out in Suppl.
192 Figure 7. It is noteworthy that this is exactly the same mutational signature as
193 the one recently reported for human embryonic somatic mutations[16]. There

194 was no obvious difference between the profile of *dnm*'s in the male and female
195 germ line (data not shown). In general, *dnm*'s appeared uniformly scattered
196 across the genome (Suppl. Figure 8).

197 The enrichment of C>A/G>T transversions in the mosaic mutations caused the
198 overall Ti/Tv ratio to be 1.33, well below expectations. This was likely due to
199 sampling variation (meaning that Ti/Tv ratios might differ between families and
200 that we by chance sampled families at the low end), as the Ti/Tv ratio was 1.99
201 for 2,530 candidate *dnm*'s detected with the same bioinformatics pipeline in a
202 follow-up study of 113 probands (excluding the ones analyzed in this work), i.e.
203 closer to expectations and the 2.2 Ti/Tv ratio of SNPs segregating in the
204 Holstein-Friesian dairy cattle population (MAF ≤ 0.01; rare allele considered to
205 be the derived allele). However, the spectrum of the 2,530 *dnm*'s remained
206 significantly different from the SNP spectrum, with an excess of C>A/G>T
207 transversions in the mosaic class of mutations, an excess of C>T/G>A transitions
208 in both mosaic and non-mosaic mutations, and a paucity of T>C/A>G transitions
209 in both mosaic and non-mosaic mutations (Suppl. Figure 7). This could point
210 towards recent alterations of the mutational profile in domestic cattle. It is
211 worth noting in this regard that most analyzed animals were bred using artificial
212 insemination and/or in vitro embryo production. It seems unlikely that artificial
213 insemination with frozen semen could explain the observed familial clustering of
214 specific *dnm*'s. However, it is conceivable that in vitro maturation, fertilization
215 and culture of oocytes and embryos affect the *dnm* rate, possibly by perturbing
216 DNA replication. It is important to determine whether this is the case, especially
217 as the same methods are increasingly used in human reproduction.

218 *Dnm*'s occurring during the development of an individual, should a priori affect
219 the maternal and paternal chromosome with equal probability. When
220 considering the PM, SM and DM jointly, 48 mosaic mutations occurred on the
221 maternal chromosome versus 31 on the paternal chromosome ($p = 0.11$). This
222 trend suggests that the maternal and paternal chromosomes might be
223 epigenetically distinct during early development and that this may affect their
224 mutability.

225 Our work points towards the fact that direct estimates of mutation rates from
226 sequencing families may have to be revisited, taken PM, SM and DM status into
227 account, to obtain more accurate estimates of the mutation rate per gamete and
228 per generation. This may affect both the overall mutation rate as well as its
229 male/female ratio. However, our analyses suggest that the effect is likely to be
230 modest and would, for instance, be insufficient to explain the present 2-fold
231 discrepancy between direct and indirect estimates in human studies[17,18]. We
232 confirmed by simulation that the rate of mosaicism does not significantly affect
233 the rate of nucleotide substitution per generation or average fixation time[19]
234 (Suppl. Figure 9).

235 Our work calls for a careful reevaluation of the importance of mosaicism for
236 *dnm*'s in humans. If more common than presently appreciated, the recurrence
237 risk of *dnm*-dependent disorders in sibs may be higher than generally assumed
238 [11,17]. Moreover, a non-negligible proportion of true *dnm*'s may have been ignored
239 (because they were detected at low dosage in the parents) in *dnm*-dependent
240 searches for genes underlying inherited disorders hence reducing the potential
241 power of such studies.

242

243 **Acknowledgements**

244 This work was funded by the DAMONA Advanced ERC project to Michel Georges.
245 Carole Charlier is Senior Research Associate from the Fonds de la Recherche
246 Scientifique - FNRS (F.R.S.-FNRS). Chad Harland has been funded in part by
247 Livestock Improvement Corporation (New Zealand). We are grateful to Erik
248 Mullaart and CRV (Arnhem the Netherlands) for providing us with the sperm and
249 blood samples. We used the supercomputing facilities of the Consortium des
250 Equipements de Calcul Intensif en Fédération Wallonie Bruxelles (CECI) funded
251 by the F.R.S-FNRS.

252

253 **Authors contributions**

254 MG, CH, CC: designed the experiments. EM: provided samples. LK, NC, MD, WC:
255 performed the sequencing. CH, MG, CC: analyzed data. MG, CH, CC: wrote the
256 paper.

257

258 **Data availability**

259 All sequence data will be made freely available in public databases.

260

261 **References**

262 1. Roach JC et al. Analysis of Genetic Inheritance in a Family Quartet by Whole-
263 Genome Sequencing. *Science* **328**: 636–639 (2010).

264 2. Conrad DF et al. Variation in genome-wide mutation rates within and between
265 human families. *Nat Genet* **43**: 712–714 (2011.).

266 3. Kong A et al. Rate of de novo mutations and the importance of father/'s age to
267 disease risk. *Nature* **488**: 471–475 (2012).

268 4. Campbell CD et al. Estimating the human mutation rate using autozygosity in a
269 founder population. *Nat Genet* **44**: 1277–1281 (2012).

270 5. Michaelson JJ et al. Whole-Genome Sequencing in Autism Identifies Hot Spots
271 for De Novo Germline Mutation. *Cell* **151**: 1431–1442 (2012).

272 6. Venn O et al. Strong male bias drives germline mutation in chimpanzees.
273 *Science* **344**:1272-1275 (2014).

274 7. Lindsay SJ et al. Striking differences in patterns of germline mutation between
275 mice and humans. doi: https://doi.org/10.1101/082297

276 8. Smeds L, Qvarnström A, Ellegren H Direct estimate of the rate of germline
277 mutation in a bird. *Genome Res.* **26**: 1211-1218 (2016).

278 9. Feng C et al. Moderate nucleotide diversity in the Atlantic herring is associated
279 with a low mutation rate. *eLife*, in press (2017).

280 10. Woodruff RC & Thompson JN. Have premeiotic clusters of mutation been
281 overlooked in evolutionary theory? *J. Evol. Biol.* **5:**457-464 (1992).

282 11. Campbell IM et al. Somatic mosaicism: implications for disease and
283 transmission genetics. *Trends Genet.* **31**:382-392 (2015).

284 12. Smit M et al. Mosaicism of Solid Gold supports the causality of a noncoding A-
285 to-G transition in the determinism of the callipyge phenotype. *Genetics* **163**:453-
286 456 (2003).

287 13. Daetwyler HD et al. Whole-genome sequencing of 234 bulls facilitates
288 mapping of monogenic and complex traits in cattle. *Nat Genet* **46**:858-865
289 (2014).

290 14. Rahbari R et al. Timing, rates and spectra of human germline mutation.
291 *Nature Genetics* **48:** 126-133 (2016).

292 15. Ohinata et al. Blimp1 is a critical determinant of the germ cell lineage in mice.
293 *Nature* **437**:207-213 (2005).

294 16. Ju YS et al. Somatic mutations reveal asymetric cellular dynamics in the early
295 human embryo. *Nature* **543**: 714-718 (2017).

296 17. Scally A. Mutation rates and the evolution of germline structure. *Philos Trans*
297 *R Soc Lond B Biol Sci* **371:** 20150137 (2016).

298 18. Segurel L, Wyman MJ, Przeworski M. Determinants of mutation rate variation
299 in the human germline. *Annu Rev Genomics Hum Genet* **15**:47-70 (2014).

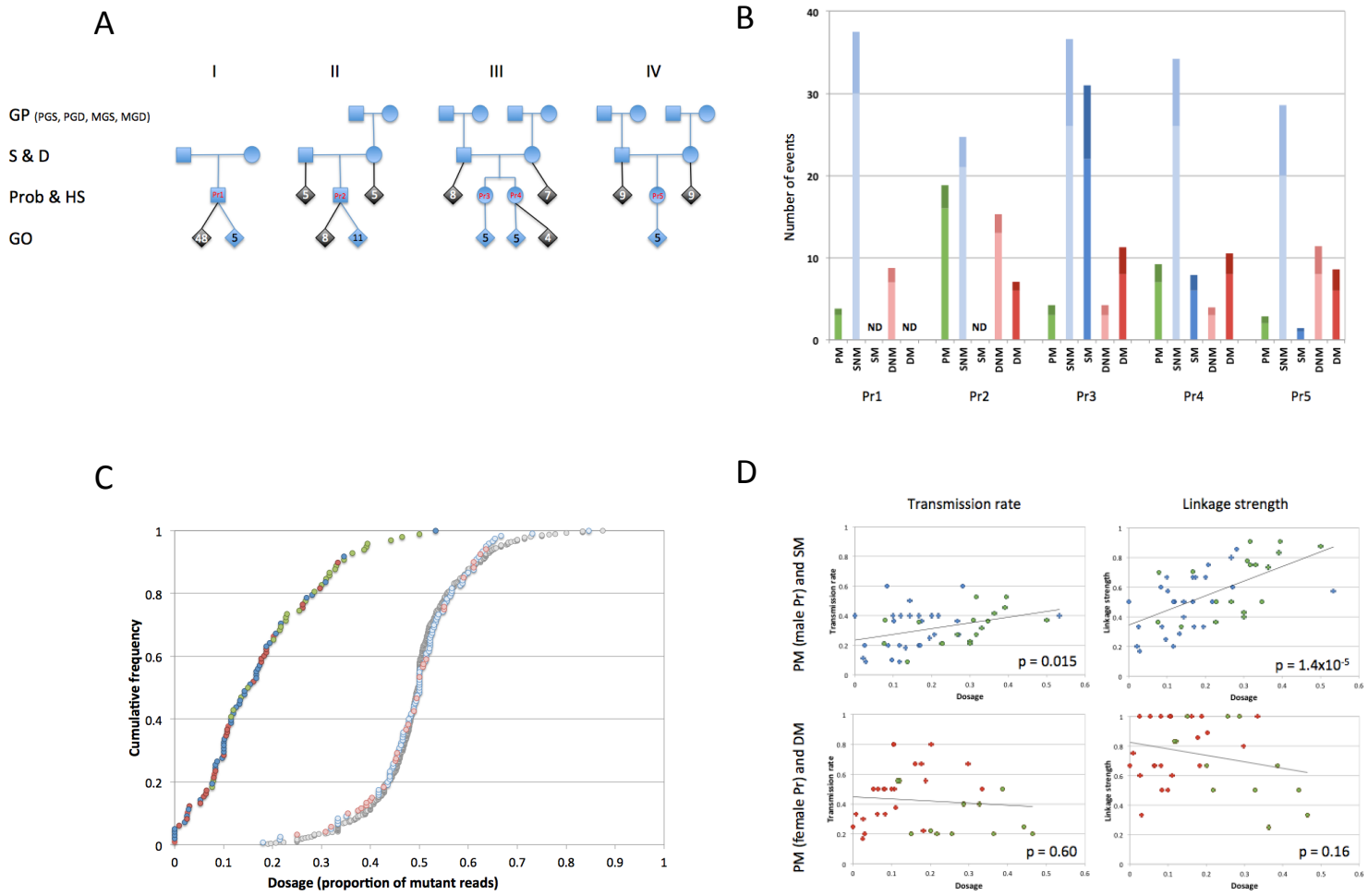
300 19. Woodruff RC & Thomson JN. The fundamental theorem of neutral evolution:
301 rates of substitution and mutations should factor in premeiotic clusters. *Genetics*
302 **125:** 333-339 (2005).

303

304

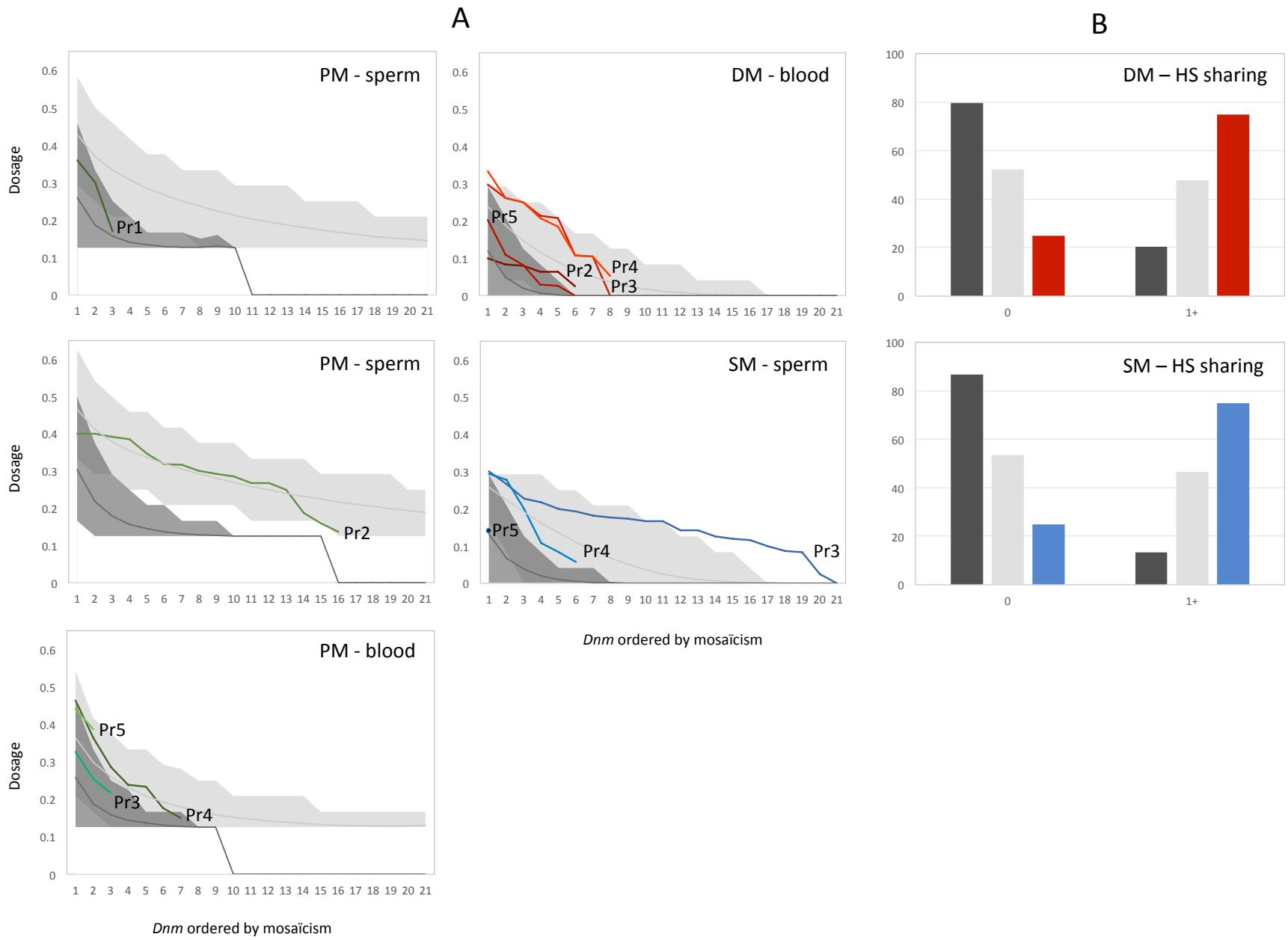
305

306


A
I
II
III
IV
GP (PGS, PGD, MGS, MGD)
S & D
Prob & HS
GO
B
Number of events
ND
PM
SNM
SM
DNM
DM
Pr1
Pr2
Pr3
Pr4
Pr5
C
Cumulative frequency
Dosage (proportion of mutant reads)
D
Transmission rate
Linkage strength
PM (male Pr) and SM
PM (female Pr) and DM
p = 0.015
p = 1.4x10-5
p = 0.60
p = 0.16
Dosage

325 **Figure 1: (A) Four pedigrees (I, II, III, IV) used for the detection of *dnm*'s.** GP: grand-parents (PGS: paternal grand-sires, PGD,
326 paternal grand-dams, MGS: maternal grand-sires, MGD: maternal grand-dams), S: sires, D: dams, Pr: probands, HS: half-sibs (of the
327 proband), GO: grand-offspring. The five probands are labeled in red. Animals in blue were genome-sequenced at average depth of 23
328 and used for the detection of *dnm*'s. Animals in gray were used for confirmation by whole genome (average sequence depth of 20) or
329 targeted sequencing (see Supplemental Methods). DNA was extracted from venous blood for females, and semen from males, except for
330 Proband 1 for which both semen and blood DNA were analyzed. **(B) Numbers and types of *dnm*'s detected in the five probands (Pr1,**
331 **Pr2, Pr3, Pr4), Pr5).** Green: PM = proband mosaic, Light blue: SNM = sire non mosaic, Dark blue: SM = sire mosaic, Light red: DNM =
332 dam non mosaic, Dark red: DM = dam mosaic. For each bar, the lower light section corresponds to the actual number of detected *dnm*'s,
333 the upper darker section to an extrapolation to the whole genome based on the estimated coverage. ND = not done (because the
334 corresponding grand-parents were not sequenced). **(C) Cumulative frequency distribution of *dnm* dosage estimated as the**
335 **proportion of reads carrying the mutation.** Green circles: PM mutations in the probands. Light blue circles: SNM mutations in the
336 probands. Light red circles: DNM mutations in the probands. Dark blue circles: SM mutations in the sires. Dark red circles: DM
337 mutations in the dams. Grey circles: corresponding PM, SNM, DNM, SM and DM mutations in the grand-offspring. The three SM and one
338 DM variant with dosage of 0, were shared between the proband and at least one half-sib yet not detectable in the semen or blood of the
339 corresponding parent. **(D)** Relationship between the *dnm* dosage (fraction of mutant reads) and the rate of transmission to offspring
340 (left) and strength of linkage (right) for mutations that are detectably mosaic in the sperm of a male parent (upper), or in the blood of a
341 female parent (lower). Green circles: PM mutations, blue circles: SM mutations, red circles: DM mutations. The corresponding
342 correlations were significant in males ($p = 0.015$ and 1.4×10^{-5}) but not in females ($p = 0.60$ and 0.16).

343


A
B
PM - sperm
Pr1
DM - blood
Pr5
Pr2
Pr4
Pr3
DM – HS sharing
PM - sperm
Pr2
SM - sperm
Pr5
Pr4
Pr3
SM – HS sharing
PM - blood
Pr5
Pr3
Pr4
Dosage
0
1+
Dnm ordered by mosaïcism
Dnm ordered by mosaïcism

361 **Figure 2: (A)** *Dnm*'s with detectable mosaicism in sperm DNA of male probands (PM – sperm) or sires (SM – sperm), or in blood DNA of
362 female probands (PM – blood) or dams (DM – blood), ranked by observed rate of mosaicism. Colored lines: real data. Pr1-5: proband 1-
363 5. Dark gray shaded area: 95% confidence interval obtained from simulations assuming uniform mutation rate per cell division and 40
364 unrelated PGCs. Light gray shaded area: 95% confidence interval obtained from simulations assuming 20-fold higher mutation rate
365 during the first 4 cell divisions, and 4 related PGCs (Table 1). **(B)** Distribution of the proportion of half-sibs (HS) of the probands that
366 share 0, or at least 1 (1+) of the *dnm*'s detected in the corresponding proband. Red bars: real observations for *dnm'*s transmitted by the
367 dam (DM+DNM). Blue bars: real observations for *dnm'*s transmitted by the sire (SM+SNM). Dark grey bars: expectation under the null
368 hypothesis of uniform prenatal mutation rate per cell division and 40 unrelated PGCs. Light grey bars: expectation under the best
369 alternative model assuming a 20x increased mutation rate during the first 4 cell division and 4 related PGCs (Table 1).

370

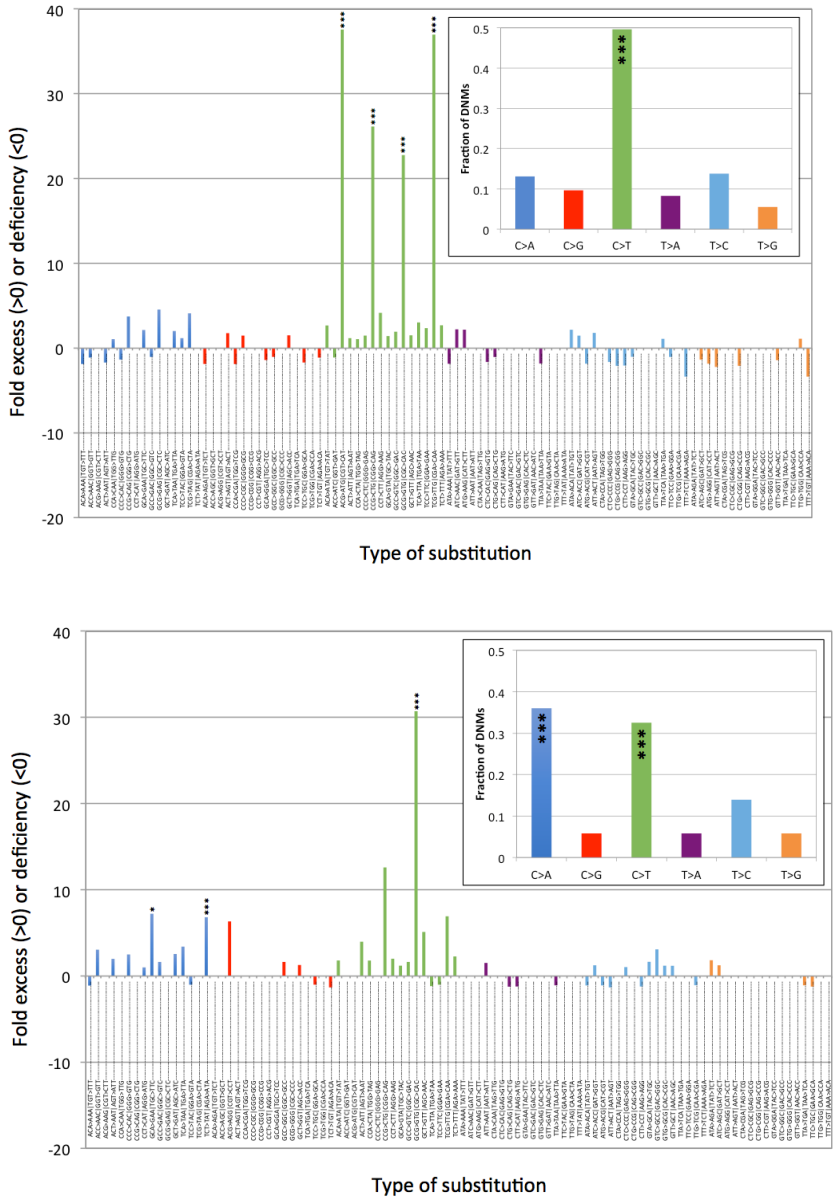


Figure 3: (A) SNM and DNM (i.e. *dnm*'s assumed to have occurred in the later stages of gametogenesis): fold excess or deficiency over expected for specific nucleotide substitutions accounting for trinucleotide context. Inset: Proportion of *dnm*'s corresponding to the six possible types of nucleotide substitutions. **(D)** Idem for PM, SM and DM (i.e. *dnm*'s assumed to have occurred in the early stages of gametogenesis). ***: p< 0.001, *: p< 0.05 (accounting for multiple testing by Sidak correction).

399 **Table 1: Relative likelihood of the observations under different models of**
400 **gametogenesis**

Fold increase mutation rate	During first x cell divisions	Number of PGCs	Related PGCs or not	Log(LR)
20x	**4**	**4**	**T**	0.00
10x	4	4	T	-0.35
20x	4	4	**F**	-0.40
10x	**7**	4	F	-1.57
20x	4	**10**	T	-1.74
20x	4	**40**	T	-3.45
10x	**11**	4	T	-5.09
20x	**15**	4	T	-7.03
10x	**18**	4	T	-9.46
1x	7	4	T	-16.17

401 The first four columns correspond to the parameters that were tested in the model: (i)
402 the fold increase of the mutation rate (1x, 10x, 20x), (ii) during the x first cell divisions
403 (4, 7, 11, 15, 18), (iii) the number of PGCs (4, 10, 40), and (iv) the ontogenetic
404 relatedness of the PGCs (F(alse) or T(rue)). Log(LR) corresponds to the logarithm
405 (base 10) of the likelihood of the data relative to the best model (first line).
406 Parameters are in bold when the corresponding model is the best given that
407 parameter value. We only show results for models that are the best given at least one
408 parameter value. Likelihoods of all models are given in Suppl. Table 3.

409

410 **Methods**

411 ***Whole genome sequencing.*** DNA was extracted from sperm (for males) or
412 whole blood (females and one male) for the four families and their relatives
413 using standard procedures. Familial relationships were confirmed by
414 genotyping all DNAs with the 10K Illumina SNP chip. We constructed 550bp
415 insert size whole genome Illumina Nextera PCR free libraries following the
416 protocols recommended by the manufacturer. All samples where then sequenced
417 on Illumina HighSeq 2000 instruments, using the 2x100bp paired end protocol
418 by the GIGA Genomics platform (University of Liege). Data was mapped using
419 BWA mem (version 0.7.9a-r786)[20] to the BosTau6 reference genome. Alignments
420 were processed according to the GATK[21] best practices version 2 with PCR
421 duplicates marked, Indel realignment and Base Quality Score Recalibration using
422 known sites. GATK HaplotypeCaller (version 3.4) was used according to the N+1
423 workflow to generate variants from the alignments. Common variants were then
424 compared to a 10K Illumina SNP chip for each individual to confirm the identity
425 of the library.

426 ***Detection of de novo mutations.*** We developed a suite of scripts to identify
427 *dnm*'s from a vcf file produced by GATK and containing sequence information
428 about members of four-generation pedigrees such as the ones described in Fig. 1.
429 The first ("4_phaser_4_gen.pl") generates the linkage phase for the parents (sire
430 and dam), the proband, and the grand-offspring. Phasing is done based on high
431 quality variant positions and genotypes (f.i. QUAL score ≥ 50,000; PL scores ≥
432 100; sequence depth ≤ 2.5x the average sequence depth). The outcome is
433 knowledge of the grand-parental origin of the paternal and maternal

434 chromosomes of the proband including the identification of cross-over events, as
435 well as the grand-parental origin of the chromosomes transmitted by the
436 proband to the grand-offspring including the identification of cross-over events.
437 The second module ("5_de_novo_detector_4_gen.pl") identifies the candidate
438 *dnm*'s per se. It first identifies diallelic variant positions for which all grand-
439 parents, sire, dam and proband have a genotype and sequence coverage between
440 set limits (f.i. 10 and 60). The proportion of variants sites satisfying these depth
441 limits was used to estimate the proportion of the genome (of the total of the
442 2,670,422,299 base pairs in the bovine Bostau6 build) that was explored.
443 Candidate *dnm*'s were then identified as sites for which (i) the QUAL score was ≥
444 100, (ii) the proband had genotype 0/1 with corresponding PL-scores ≥ 40, (iii)
445 none of the grand-parents had reads with the alternate allele (AD = x,0), (iv)
446 either the sire or the dam had no reads with the alternate allele, and (v) reads
447 with the alternate allele were found in the grand-offspring. The same script also
448 determines the genotype frequencies (0/0, 0/1 and 1/1) at the corresponding
449 position in sequenced individuals outside of the pedigree that are not
450 descendants of the sire or the dam. The third module
451 ("6_germline_assigner_4_gen.pl") combines the output of the first and second
452 module to determine in which individual the *dnm* is most likely to have occurred
453 (one of the four grand-parents, sire or dam, or proband) and on which grand-
454 parental chromosome it occurred. All candidate *dnm*'s were manually curated
455 using the Integrated Genome Viewer (IGV)[22]. *Dnm*'s with mutant reads in either
456 sire or dam (even if called 0/0 by GATK) were relabeled as SM or DM, provided
457 that the *dnm* segregated in complete (but imperfect) linkage with the paternal or
458 maternal haplotype (respectively), in the half-sibs of the proband.

459 To test the corresponding pipeline, we identified 11,255 variants (of which
460 10,093 SNPs) for which Pr2 was heterozygous and which were not present in
461 unrelated individuals including from the 1,000 Bulls project. The corresponding
462 "genotype fields" of the parents and grand-offspring were modified in the vcf file
463 such that the genotype (GT) was set at 0/0 and the unfiltered allele depth (AD) of
464 the derived allele set at 0. The Ti/Tv ratio for the corresponding SNPs was
465 1.9922. The proportion of the genome explored was estimated at 76% as
466 described above. The pipeline detected 9,325 variants (83% sensitivity) of
467 which 8,409 SNPS (81% sensitivity). The Ti/Tv ratio amongst detected *dnm*'s
468 was 1.9989, indicating that the pipeline did not introduce a Ti/Tv bias.

469 ***Confirmation of candidate dnm's.*** PCR primers were then designed for each
470 candidate passing the quality check in IGV using BatchPrimer[23] targeting a
471 product size of 200-1000bp with at least one primer being present in unique
472 (non-repeat) sequence (as identified by repeatmasker). The resulting amplicons
473 were sequenced on an Illumina MiSeq instrument using the 2x250bp paired end
474 protocol. The sequenced amplicons were aligned to the BosTau6 reference
475 genome using BWA mem and candidate de novo mutations were checked in IGV
476 and variants were called using freebayes (v1.0.2-15-g357f175)[24].

477 ***Modeling gametogenesis.*** <u>(i) Data types:</u> To compare the adequacy of the
478 different gametogenesis models we computed the likelihood of three types of
479 data. The first is the degree of mosaicism in the parent across *dnm*'s detected in
480 a given gamete. Thus we may have detected *n* SM and *m* SNM *dnm*'s in a given
481 sperm cell. The *n* SM *dnm*'s have dosages in the paternal sperm DNA of x_1, x_2,
482 x_3, ... $x_n > 0$ while the *m* SNM *dnm*'s have a dosage of 0. We have three such lists

483 for Pr3, Pr4 and Pr5. Likewise, we may have detected *n* OM and *m* ONM *dnm*'s in
484 a given oocyte. The *n* OM *dnm*'s have dosages in the maternal blood DNA of x_1, x_2,
485 x_3, ... $x_n > 0$ while the *m* ONM *dnm*'s have a dosage of 0. We have four such lists
486 for Pr2, Pr3, Pr4 and Pr5.

487 The second data set consists in lists of PM *dnm*'s and their dosage (x_1, x_2, x_3, ... x_n
488 > 0) in sperm (Pr1 and Pr2) or blood DNA (Pr3, Pr4, Pr5).

489 The third data type consists in the number of *dnm*'s detected in a gamete
490 transmitted to a proband, and the numbers of those shared by the studied half-
491 sibs of the proband. Thus, we may have detected *n* SM and *m* SNM *dnm*'s in a
492 sperm cell (or oocyte) transmitted to a given proband, of which half-sib 1 will
493 share x_1, half-sib 2 x_2, ..., where $x_i \leq n$.

494 <u>(ii) Computing probabilities under various models of gametogenesis:</u> For data
495 type 1, we simulated the process of de novo mutation in the female and male cell
496 lineages described in Suppl. Figure 6. For the null hypothesis, the mutation rate
497 per cell division before birth was set at an average of 0.77 (Poisson distributed),
498 such that the number of *dnm*'s per oocyte averaged 14 (as observed). The
499 mutation rate per cell division after birth was set at an average of 0.3 (Poisson
500 distributed), such that the number of *dnm*'s per sperm cell averaged 34 (as
501 observed). For alternative hypotheses, the mutation rate for the early cleavage
502 cell divisions (4, 7, 11, 15 and 18 first cell divisions, corresponding to the
503 different development stages in Suppl. Figure 6) was increased 10- or 20-fold
504 when compared to the remaining prenatal cell divisions, for which the mutation
505 rate was concomitantly reduced such that the overall number of *dnm*'s per
506 oocyte remained unaffected (average of 14). We further tested 4, 10 and 40

507 induced PGCs, and unrelated or related induced PGCs as described in Suppl. Fig.
508 6. For all 90 possible scenarios, we determined by simulation what proportion of
509 *dnm*'s found in a sperm cell (respectively oocyte) were characterized by a dosage
510 in paternal sperm (respectively maternal soma) of 0-0.05, 0.05-0.10, ... These
511 proportions were then used as probabilities in computing the likelihood of the
512 data (i.e. a series *dnm*'s with corresponding rate of mosaicism in the parental
513 tissue) under the corresponding model. Given our experimental design, SM and
514 DM are only recognized as such, if (i) their dosage in the sire (SM) or dam (DM)
515 is significantly < 0.5, and (ii) they show complete (but imperfect) linkage in the
516 available half-sibs. We considered a fixed number of eight half-sibs in the
517 simulations. These conditions were included in the simulations. Thus a mutation
518 was only considered if it satisfied these two criteria. The dosage that was
519 considered was not the true dosage for that mutation, but the "realized" dosage
520 assuming a sequence depth of 24.

521 To compute the likelihood of the second type of data, we simulated
522 gametogenesis in exactly the same way as for data type 1. We then randomly
523 sampled *n* gametes, where *n* corresponds to the number of GO (hence 5 for Pr1,
524 Pr3-5, and 11 for Pr2). For all *dnm*'s in these *n* gametes, we then determined the
525 dosage in the germ-line (Pr1, Pr2) or soma (Pr3-5). For all 90 possible scenarios,
526 we determined by simulation what proportion of PM *dnm*'s detected in sperm of
527 blood DNA were characterized by a dosage of 0-0.05, 0.05-0.10, ... These
528 proportions were then used as probabilities in computing the likelihood of the
529 data (i.e. a series PM *dnm*'s with corresponding dosage in sperm or blood) under
530 the corresponding model. With the real data, PM mutations are only recognized

531 as such (i) if they are transmitted to at least one of the *n* offspring, (ii) if the
532 proband is called heterozygous for the corresponding *dnm* by GATK, and (iii) if
533 we demonstrate complete (but imperfect) linkage in the GO. Condition (i) is
534 achieved in the simulation by sampling *n* gametes at random. Condition (ii) and
535 (iii) were modeled in the simulations. We considered five and eleven GO to
536 match the real data. Thus a mutation was only considered if it satisfied these
537 two criteria. The dosage that was considered was not the true dosage for that
538 mutation, but the "realized" dosage assuming a sequence depth of 24.

539 For data type 3, we modified the simulations in order to exactly generate a
540 predetermined number *n* of *dnm*'s in a given "reference" gamete. Thus if in the
541 real data an oocyte was characterized by 11 *dnm*'s (f.i. Pr3 and Pr4), we would in
542 the simulations distribute 11 *dnm*'s across the (7+4+4+3+18) cell divisions
543 leading to a simulated reference oocyte and track their segregation (according to
544 their point of occurrence) across the entire germ line lineage. Under the null
545 hypothesis of uniform prenatal mutation rate, all 36 cell divisions would have
546 equal chance to be hit by anyone of the 11 mutations. Under the alternative
547 hypotheses, early cleavage cell divisions would have a 10- or 20-fold higher
548 chance than the remaining ones. Under the hypothesis of related PGCs the
549 segregation pattern of early mutations in the germ line lineage would be
550 concomitantly affected (see Suppl. Fig. 6). We would then samples gametes at
551 random from the same germ line tree and count the number of mutations shared
552 with the "reference" gamete. This would generate a frequency distribution of
553 gametes sharing 0, 1, 2, ... , *n* *dnm*'s with the reference gamete. The
554 corresponding frequencies were then used as probabilities in computing the

555 likelihood of the data (i.e. a series half-sibs sharing 0, 1, ... , *n dnm*'s with the
556 reference gamete transmitted to the proband) under the corresponding model.

557 Likelihoods of the data under the 90 tested models were then simply computed
558 as the product of the probabilities of all *dnm*'s (data type 1 and 2) and half-sibs
559 (data type 3) extracted from the simulations performed under the corresponding
560 model.

561 <u>(iii) Estimating the number of missed SM and DM and misclassified PM</u>
562 <u>mutations:</u> The simulations for dataset 1 allowed us to estimate the number of
563 SM and DM mutations missed either because the "realized" dosage was to high in
564 the parent, or because we could not demonstrate complete (but imperfect)
565 linkage in the half-sibs. Likewise the simulations for dataset 2 allowed us to
566 estimate the number of PM mutations that, although detected (realized dosage
567 sufficient to be called heterozygous by GATK), were misclassified as SNM or DNM
568 mutations because showing perfect linkage in the available GO. Under the best
569 biological model (20x increased mutation rate during the first 4 cell divisions, 4
570 related PGCs, see Table 1), these numbers were: (i) average loss of 3.3 SM
571 mutations, (ii) average loss of 1.1 lost DM mutations, (iii) average gain of 1.45
572 SNM and 1.45 DNM mutations for a male proband with 11 GO, (iv) average gain of
573 4 SNM and 4 DNM mutations for a male proband with 5 GO, and (v) average gain
574 of 1.8 SNM and 1.8 DNM mutations for a female proband with 5 GO.

575

576 20. Li H & Durbin R. Fast and accurate short read alignment with Burrows-
577 Wheeler transform. *Bioinformatics* **25:** 1754-1760 (2009).

578 21. McKenna A *et al.* The Genome Analysis Toolkit: a MapReduce framework for
579 analyzing next-generation DNA sequencing data. *Genome Res.* **20:** 1297-1303
580 (2010).

581 22.Robinson JT et al. Integrative Genomics Viewer. *Nature Biotechnology* **29**:24–
582 26 (2011).

583 23. You FM *et al.* BatchPrimer3: a high throughput web application for PCR and
584 sequencing primer design. *BMC Bioinformatics* **9:**253 (2008).

585 24. Garrison E, Marth G. Haplotype-based variant detection from short-read
586 sequencing. arXiv preprint arXiv:1207.3907 [q-bio.GN]. (2012) .

587

588