

1 **The mutational landscape of normal human endometrial epithelium**

2

3 **Authors**

4 Luiza Moore(1), Daniel Leongamornlert(1), Tim H. H. Coorens(1), Mathijs A. Sanders(1,2),
5 Peter Ellis(1), Kevin Dawson(1), Francesco Maura(1), Jyoti Nangalia(1), Patrick S.
6 Tarpey(1), Simon F. Brunner(1), Henry Lee-Six(1), Raheleh Rahbari(1), Sarah Moody(1),
7 Yvette Hooks(1), Krishnaa Mahbubani(3), Mercedes Jimenez-Linan (4), Jan J. Brosens(5),
8 Christine A. Iacobuzio-Donahue(6), Inigo Martincorena(1), Kourosh Saeb-Parsy(2), Peter J.
9 Campbell(1), Michael R. Stratton(1)

10

11 **Author information**

12 (1) Cancer, Ageing and Somatic Mutation (CASM), Wellcome Sanger Institute, Hinxton,
13 CB10 1SA, UK

14 (2) Department of Hematology, Erasmus University Medical Center, 3015 CN Rotterdam,
15 The Netherlands

16 (3) Department of Surgery, University of Cambridge, and Cambridge NIHR Biomedical
17 Research Centre, Cambridge Biomedical Campus, Cambridge, CB2 0QQ

18 (4) Department of Pathology, Addenbrooke's Hospital, Cambridge, CB2 0QQ

19 (5) Division of Biomedical Sciences, Warwick Medical School, University of Warwick,
20 Coventry CV2 2DX, UK

21 (6) Department of Pathology, Memorial Sloan Kettering Cancer Center, 1275 York Avenue,
22 Box 20, New York, NY 10065

23

24

25 **Abstract**

26 All normal somatic cells are thought to acquire mutations. However, characterisation of the
27 patterns and consequences of somatic mutation in normal tissues is limited. Uterine
28 endometrium is a dynamic tissue that undergoes cyclical shedding and reconstitution and is
29 lined by a gland-forming epithelium. Whole genome sequencing of normal endometrial
30 glands showed that most are clonal cell populations derived from a recent common ancestor
31 with mutation burdens differing from other normal cell types and manyfold lower than
32 endometrial cancers. Mutational signatures found ubiquitously account for most mutations.
33 Many, in some women potentially all, endometrial glands are colonised by cell clones
34 carrying driver mutations in cancer genes, often with multiple drivers. Total and driver
35 mutation burdens increase with age but are also influenced by other factors including body
36 mass index and parity. Clones with drivers often originate during early decades of life. The
37 somatic mutational landscapes of normal cells differ between cell types and are revealing
38 the procession of neoplastic change leading to cancer.

39

40 **Introduction**

41 Acquisition of mutations is a ubiquitous and essential feature of the cells of living organisms.
42 Although there has been comprehensive characterisation of the somatic mutation landscape
43 of human cancer¹⁻³, understanding of patterns of somatic mutation in normal cells is limited.
44 In large part this has been due to the challenge of detecting somatic mutations in normal
45 tissues and several strategies have recently been developed to address this including
46 sequencing of *in vitro* derived clonal cell populations from normal tissues⁴⁻⁸, sequencing
47 small biopsies containing limited numbers of microscopic clones^{9,10}, sequencing
48 microscopically distinguishable structural elements which are clonal units^{11,12}, highly error

49 corrected sequencing^{13,14} and sequencing single cells^{15,16}. Together, these have begun to
50 reveal differing mutation burdens between different cell types, their patterns of acquisition
51 over time and the signatures of the mutational processes generating them. They have also
52 shown that, in normal tissues, clones of normal cells with “driver” mutations in cancer genes
53 are present. In the glandular epithelium of the colon these are relatively uncommon¹² but, in
54 the squamous epithelia of the skin⁹ and oesophagus¹⁰ and other tissues, such as the
55 blood¹⁷⁻²¹, clones carrying drivers can constitute substantial proportions of normal cells
56 present after middle age.

57
58 The factors determining differences in mutation landscape between normal cell types are
59 incompletely understood. However, they plausibly include the intrinsic structural and
60 physiological features of each tissue. Endometrium is a uniquely dynamic tissue composed
61 of a stromal cell layer invaginated by a contiguous glandular epithelial sheet covering the
62 luminal surface. It adopts multiple different physiological states during life including
63 premenarche, menstrual cycling, pregnancy, and postmenopause. During reproductive
64 years it undergoes cyclical breakdown, shedding, repair and remodelling in response to
65 oscillating levels of oestrogen and progesterone which entail iterative restoration of the
66 contiguity of the interrupted glandular epithelial sheet that is effected by stem cells within
67 basal glands retained after menstruation²²⁻²⁵.

68
69 Characterisation of the mutational landscapes of normal tissues is beginning to provide
70 comprehensive understanding of the succession of intermediate neoplastic stages between
71 normal cells and cancers originating from them. There are two major histological classes of
72 endometrial carcinoma^{26,27}. Type I, endometrioid carcinoma, is commoner with the main
73 known risk factor being extent of oestrogen exposure, influenced by early menarche, late
74 menopause and body mass index (BMI)^{27,28}. Type II, including serous and clear cell
75 carcinomas, occurs in older women with smoking, age and elevated BMI as recognised risk
76 factors²⁹. Commonly mutated cancer genes include *PTEN*, *TP53*, *PIK3CA*, *KRAS*, *ARID1A*,
77 *FBXW7* and *PIK3R1*³⁰ and subsets of endometrial cancer carry large numbers of base
78 substitution and/or small insertion and deletion (indel) mutations due to defective DNA
79 mismatch repair, polymerase epsilon/delta mutations, or large numbers of copy number
80 changes and genome rearrangements^{26,31}.

81
82 Recent studies using exome and targeted sequencing have revealed the presence of driver
83 mutations in known cancer genes in a high proportion of endometrial glands in
84 endometriosis^{11,32,33}, and also in eutopic normal endometrial epithelium¹¹. Here, by whole
85 genome sequencing, we have further characterised the mutational landscape of normal
86 endometrial epithelium, explored how it is influenced by age, BMI and parity, estimated the
87 age of driver mutations and the relationship of clonal evolution to glandular architecture.

88

89 **Results**

90 **Samples and sequencing**

91 Using laser capture microdissection (LCM) 215 histologically normal endometrial glands
92 were isolated from 18 women aged 19 to 81 years. The samples were from biopsies taken
93 for infertility assessments (6), hysterectomies for benign non-endometrial pathologies (2),
94 residual tissues from transplant organ donors (6) and autopsies after death from non-
95 gynaecological causes (4). DNA from each gland was whole genome sequenced using a
96 library-making protocol modified to accommodate small amounts of input DNA¹². The mean

97 sequencing coverage was 28-fold and only samples with >15-fold coverage were included in
98 subsequent analyses (n=182, Supplementary Table 1, Supplementary Results 1). Somatic
99 mutations in each gland were determined by comparison with whole genome sequences
100 from pieces of uterus, cervix or Fallopian tube from the same individuals. From each of 18
101 glands two separate samples were obtained and subjected to independent DNA extraction,
102 library preparation and whole genome sequencing. Using these biological “near-replicates”
103 the mean sensitivity of somatic mutation variant calling was estimated at >86% (range 0.70 –
104 0.95%) (Methods).

105

106 **Clonality of endometrial glands**

107 To assess whether endometrial glands are clonal populations derived from single recent
108 ancestor cells the variant allele fractions (VAFs) of somatic mutations were examined. Most
109 somatic mutations are heterozygous. Heterozygous mutations present in all cells of a
110 population derived from a single ancestor will have VAFs of 0.5 whereas somatic mutations
111 in cell populations derived from multiple ancestors will have lower VAFs or be undetectable
112 by standard mutation calling approaches. 90% (163/182) of microdissected endometrial
113 glands showed distributions of base substitution VAFs with peaks between 0.3 and 0.5
114 indicating that each consists predominantly of a cell population descended from a single
115 epithelial progenitor stem cell with contamination by other cells potentially including
116 endometrial stromal cells, inflammatory cells and epithelial cells from other glands (Fig. 1,
117 Supplementary Results 1). Similar VAF distributions were observed for small insertions and
118 deletions (indels). Subsequent analyses (see below) have demonstrated that many
119 endometrial glands carry “driver” mutations in known cancer genes. However, endometrial
120 glands exhibited clonality irrespective of the presence of driver mutations with, for example,
121 somatic mutations in all 10 glands from a 19-year-old individual (PD37506) having a median
122 VAF >0.3 but no driver mutations identified (Extended Data Fig. 1a, b). Thus, colonisation of
123 endometrial glands by descendants of single endometrial epithelial stem cells is not
124 contingent on growth selective advantage provided by driver mutations and may occur by a
125 process analogous to genetic drift, as proposed for other tissues^{34,35}.

126

127 **Mutation burdens**

128 The somatic mutation burdens in normal endometrial glands from the 18 individuals ranged
129 from 225 to 2890 base substitutions (mean 1324) and 3 to 243 indels (mean 85) (Fig. 2a, b).
130 In large part this variation was attributable to the ages of the individuals with a linear
131 accumulation of ~28 base substitutions per gland per year during adult life (linear mixed-
132 effect model, SE = 3.1, P = 1.061e-07) (Supplementary Results 2). However, the
133 possibilities of lower mutation rates premenarche and postmenopause cannot be excluded.
134 The potential influences of BMI, a known risk factor for endometrial cancer, and the
135 presence of driver mutations on mutation burden were also examined. An additional 20
136 substitutions were acquired with each unit of BMI (SE = 8, P = 2.330e-02). Therefore, the
137 association between elevated BMI and increased endometrial cancer risk may, at least
138 partially, be mediated by this additional mutation burden induced by BMI in normal
139 endometrial epithelial stem cells. Positive driver mutation status conferred an addition of
140 ~177 substitutions (SE = 45.7, P = 1.632e-04). The basis of this correlation is unclear. It is
141 conceivable that an elevated total mutation load increases the chances of including, by
142 chance, a driver. It is also plausible, however, that drivers engender biological changes, for
143 example elevated cell division rates, that result in higher overall mutation loads. There was
144 no obvious correlation between parity and mutation burden.

145

146 In addition to endometrial glands, nearby normal endocervical glands were microdissected
147 from one individual (PD37506). There was a ~2-fold lower somatic mutation burden in
148 endocervical than endometrial glands (Extended Data Fig. 3). This may reflect the absence,
149 in endocervical glands, of the cyclical process of loss and regeneration that occurs in
150 endometrial glands.

151

152 **Mutational signatures**

153 To explore the underlying processes of somatic mutagenesis operative in normal
154 endometrial epithelial cells mutational signatures were analysed. Three previously described
155 single base substitution (SBS) mutational signatures were identified in all endometrial glands
156 (Extended Data Fig. 2): SBS1, predominantly characterised by NCG>NTG mutations and
157 likely due to spontaneous deamination of 5-methylcytosine; SBS5, a relatively featureless,
158 'flat' signature of uncertain cause; SBS18, predominantly characterised by C>A substitutions
159 and possibly due to reactive oxygen species³⁶. Overall, the mean signature exposures per
160 gland were 0.22 for SBS1, 0.59 for SBS5 and 0.17 for SBS18; interestingly, glands from one
161 donor with a history of recurrent missed miscarriage (RMM) showed much higher mean
162 SBS18 exposure (0.35) compared to the rest of the cohort. There were approximately 2.7-
163 fold more SBS5 than SBS1 mutations (SD 0.4171666). A positive linear correlation with age
164 for the mutation burden attributable to all three signatures was observed (Fig. 2d, e, f). To
165 ascertain the periods during which different mutational processes operate, phylogenetic
166 trees of endometrial glands were constructed for each individual using somatic mutations
167 (Figs. 3, 4). These revealed that the mutational processes underlying the three signatures
168 are active throughout life. With respect to small indels, composite mutational spectra for
169 each donor were generated and were similar across ages; however, due to the relative
170 sparsity of indels in normal endometrial glands, formal signature extraction was not
171 performed (Extended Data Figure 3).

172

173 Somatic copy number changes and structural variants (genome rearrangements) were found
174 in only 27 out of 182 (15%) normal endometrial glands (Fig. 2c, Supplementary Results 3).
175 These included copy number neutral loss of heterozygosity (cnN LOH) in six glands, whole
176 chromosome copy number increases in one and structural variants in eighteen (12 large
177 deletions, six tandem duplications and nine translocations). The majority of glands showed a
178 single change. However, one of two glands carrying a *TP53* mutation (see below) exhibited
179 nine structural variants, indicating that genomic instability caused by defective DNA
180 maintenance occurs in normal cells.

181

182 **Driver mutations**

183 To identify genes under positive selection a statistical method based on the
184 observed:expected ratios of non-synonymous:synonymous mutations was used³⁰. Eleven
185 genes showed evidence of positive selection in the 182 normal endometrial glands; *PIK3CA*,
186 *PIK3R1*, *ARHGAP35*, *FBXW7*, *ZFHX3*, *FOXA2*, *ERBB2*, *CHD4*, *KRAS*, *SPOP* and *ERBB3*
187 (Supplementary Results 4). All were present in a set of 369 genes previously shown to be
188 under positive selection in human cancer³⁰. In addition, four different truncating mutations
189 (and no other mutations) were observed in the progesterone receptor gene (PGR). Although
190 these did not attain standard significance levels the biological role that progesterone plays in
191 normal endometrium as an antagonist of oestrogen driven proliferation raises the possibility
192 that these inactivating mutations confer growth advantage. To comprehensively identify

193 drivers in the 182 endometrial glands, mutations with the characteristics of drivers in each of
194 the 369 genes were sought (Methods).

195

196 163 driver mutations were found in normal endometrial glands from 17/18 women
197 (Supplementary Results 5). The youngest carrier was a 24 year old (PD40535) with a *KRAS*
198 G12D mutation in 1/7 glands sampled. 58% (105/182) of endometrial glands carried at least
199 one driver mutation, 19% (35/182) carried at least two and 3% (5/182) carried at least four
200 drivers. Remarkably, in four women, aged 34 (19 glands), 44 (11 glands), 60 (14 glands)
201 and 81 (5 glands), all glands analysed carried driver mutations suggesting that the whole
202 endometrium had been colonised by microneoplastic clones (Figs 3, 4). The fraction of
203 endometrial glands carrying a driver (Fig. 2g), the average number of drivers per gland (Fig.
204 2h) and the number of different drivers in each individual (corrected for number of glands
205 sampled) (Fig. 2i) all positively correlated with age of the individual. However, there were
206 sufficient outliers from this age correlation to suggest that other factors influence colonisation
207 of the endometrium by driver carrying clones. Using a generalised linear mixed effect model,
208 we found that age has a positive association with the accumulation of driver mutations
209 (coefficient = 0.0336, SE = 0.0131), while parity has a negative association (coefficient = -
210 0.330, SE = 0.117) (Supplementary Results 6 and 7).

211

212 Driver mutations in both recessive (tumour suppressor genes) and dominant cancer genes
213 were found. *PIK3CA* was the most frequently mutated cancer gene, with at least one
214 missense mutation in 61% (11/18) of women and five different mutations found in two
215 women (Fig. 3 and 4, Extended Data Fig. 4). Most truncating driver mutations in recessive
216 cancer genes (including in *ZFHX3*, *ARGHAP35* and *FOXA2* which showed evidence of
217 selection in normal endometrial glands, see above) were heterozygous without evidence of a
218 mutation inactivating the second, wild type allele. Therefore, haploinsufficiency of these
219 genes appears sufficient to confer growth advantage in normal cells. Nevertheless, further
220 inactivating mutations, including copy number neutral LOH of the wild type allele and
221 truncating mutations, in the same genes in other glands indicate that additional advantage is
222 conferred by complete abolition of their activity (notably for *ZFHX3* in the 60 year old, Figure
223 3 and Exended Data Fig. 5). Driver mutations were found in genes encoding growth factor
224 receptors (*ERBB2*, *ERBB3*, *FGFR2*), components of signal transduction pathways (*HRAS*,
225 *KRAS*, *BRAF*, *PIK3CA*, *PIK3R1*, *ARHGAP35*, *RRAS2*, *NF1*, *PP2R1A*, *PTEN*), pathways
226 mediating steroid hormone responses (*ZFHX3*, *FOXA2*, *ARHGAP35*), pathways mediating
227 WNT signalling (*FBXW7*) and proteins involved in chromatin function (*KMT2D*, *ARID5B*).
228 Many different combinations of mutated cancer genes were found in individual glands.

229

230 Driver mutations were placed on the phylogenetic trees of somatic mutations constructed for
231 each individual and, by assuming a constant somatic mutation rate during life, the time of
232 occurrence of a subset was estimated (Methods). Some driver mutations occurred early in
233 life. These included a *KRAS* G12D mutation in three glands from a 35 year old and a
234 *PIK3CA* mutation in two glands from a 34 year old, which are both likely to have arisen
235 during the first decade. A pair of drivers in *ZFHX3* and *PIK3CA*, co-occurring in six glands
236 from a 60 year old, was also acquired during the first decade indicating that driver
237 associated clonal evolution begins early in life. There was evidence, however, for continuing
238 acquisition and clonal expansion of driver mutations into the third and fourth decades and
239 further accumulation beyond this period is not excluded.

240

241 **Phylogeography of mutations within the endometrium**

242 Phylogenetically closely related glands were often in close physical proximity within the
243 endometrium (Fig. 3). In phylogenetic clusters for which the mutation catalogues were
244 almost identical, this may simply reflect multiple sampling of a single tortuous gland weaving
245 in and out of the plane of section, rather than distinct glands with their own stem cell
246 populations (e.g. glands C5 and E5, Figs. 3a, c). For other phylogenetic clusters, the
247 different branches within the clade have diverged substantially, sometimes acquiring
248 different driver mutations, and therefore are likely derived from different stem cell
249 populations. In such instances phylogenetically related glands can range over distances of
250 multiple millimetres suggesting that their clonal evolution has entailed capture and
251 colonisation of extensive zones of endometrium (e.g. glands C1, A2, B1, H2, A3, B3, Figs.
252 3b, d). Conversely, many glands in close physical proximity are phylogenetically distant (e.g.
253 glands E1 and G2, Figs 3a, c), indicating that the cell populations have remained isolated
254 from each other.

255

256 **Normal endometrium compared to other cells**

257 Endometrial cells exhibit lower mutation rates than normal skin epidermal⁹, colorectal^{4,12},
258 small intestinal^{4,12} and liver cells⁴, similar burdens to oesophageal cells¹⁰ and higher rates
259 than skeletal muscle cells⁷ (Extended Data Fig. 7). Of the mutational signatures found in
260 endometrial cells, SBS1 and SBS5 are found in all other cell types³⁷. However, the SBS1
261 mutation rate is higher in colorectal and small intestinal epithelial cells whereas the SBS5
262 mutation rate is higher in liver cells⁴. SBS18 has also been found ubiquitously in colonic
263 crypts¹².

264

265 The prevalence of driver mutations is substantially different in different normal cell types. In
266 colon, like the endometrium a tissue with glandular architecture, only ~1% crypts (glands) in
267 60 year old individuals carry a driver mutation¹² compared to the much higher fractions (up
268 to 100%) in the endometrial glands of 60 year old women. The biological basis of this
269 difference is unclear but is unlikely to be the difference in total mutation burden, which is
270 lower in the endometrium than the colon.

271

272 Endometrial cancers exhibit higher mutation loads than normal endometrial cells, for base
273 substitutions (~5-fold, medians of 1346 and 7330 substitutions observed in normal
274 endometrium and endometrial cancer respectively (Mann-Whitney test, $P = 7.629e-06$) (Fig.
275 5a)) and indels (Fig. 5b) and these differences also pertain to normal endometrial cells with
276 driver mutations. In most endometrial cancers these differences are attributable to higher
277 mutation burdens of the ubiquitous base substitution and indel mutational signatures. In
278 addition, however, the very high mutation loads of the subsets of endometrial cancers with
279 DNA mismatch repair deficiency and polymerase epsilon/delta mutations were not seen in
280 normal endometrial cells. Differences between endometrial cancers and normal cells were
281 even more marked for structural variants and copy number changes (median number zero in
282 normal endometrial cells and ~23 in endometrial cancers³⁸) and this again pertained to
283 normal endometrial cells with drivers.

284

285 There were also differences in the repertoire of cancer genes in which driver mutations were
286 found (Fig 5 d,e,f, Supplementary Results 4 and 8). Notably, mutations in *PTEN*, *CTCF*,
287 *CTNNB1* and *ARID1A* in endometrioid and in *TP53* in serous endometrial cancer accounted
288 for higher proportions of driver mutations than in normal endometrial cells. It is possible that

289 *PTEN*, *ARID1A*, *TP53* and *CTCF* require biallelic mutation to confer growth advantage and
290 this may account for their lower prevalence in normal cells. However, heterozygous
291 mutations in *PTEN* and *TP53* were found, albeit rarely and restricted to the two oldest
292 individuals studied (69 and 81-year old), and this explanation would not account for the
293 relative deficit of *CTNNB1* mutations. Overall, the results suggest that driver mutations in
294 some cancer genes may be relatively effective at enabling stem cell colonisation of normal
295 tissues but confer limited risk of conversion to invasive cancers. Conversely, others may
296 require biallelic mutation and/or confer limited advantage in colonising normal tissues but are
297 relatively effective at conversion to malignancy.

298

299 **Discussion**

300 This study of normal endometrial epithelium, together with recent studies of other normal cell
301 types^{4,5,9-12,17,18}, is revealing the landscape of somatic mutations in normal human cells. The
302 landscape is characterised by different somatic mutation rates in different cell types that, for
303 the most part, are generated by a limited repertoire of ubiquitous mutational processes
304 generating base substitutions, small indels, genome rearrangements and whole
305 chromosome copy number changes. These processes exhibit more or less constant
306 mutation rates during the course of a lifetime resulting in essentially linear accumulation of
307 mutations with age. However, the influences of BMI and the presence of driver mutations on
308 mutation burden in endometrial epithelium indicate that additional factors can modulate their
309 mutation rates. The reasons for the different mutation rates of ubiquitous signatures in
310 different tissues are unclear. For SBS1, which is likely due to deamination of 5-
311 methylcytosine, the differences may be related to the number of mitoses a cell has
312 experienced. Additional mutational signatures which are present only in some cells, only in
313 some cell types and/or are intermittent also operate in normal cells, supplementing the
314 mutation load contributed by ubiquitous signatures. The latter include exposures such as
315 ultraviolet light in skin⁹, APOBEC mutagenesis in occasional colon crypts and other
316 signatures of unknown cause in normal colon epithelium¹².

317

318 A small subset of mutations generated by these mutational processes have the properties of
319 driver mutations. The total somatic mutation rate is lower in endometrial than colonic
320 epithelial stem cells and thus the rate of generation of driver mutations is also likely to be
321 lower. However, numerous cell clones with different driver mutations, some carrying multiple
322 drivers, colonise much of, and in some cases potentially all of, the normal endometrial
323 epithelium in most women. This is in marked contrast to the colon where just 1% of normal
324 crypts in middle-aged individuals carry a driver¹². This dramatic difference may be due to
325 intrinsic differences in physiology between endometrium and colon. In the endometrium, the
326 cyclical process of tissue breakdown, shedding and remodelling iteratively opens up
327 denuded terrains for pioneering clones of endometrial epithelial cells with drivers to
328 preferentially colonise compared to wild type cells. By contrast, in the colon the selective
329 advantage of a clone with a driver is usually confined to the small siloed population of a
330 single crypt, with only occasional opportunities for further expansion. Thus, the endometrium
331 in some respects resembles more the squamous epithelia of skin and oesophagus in which
332 cell clones derived from basal cells directly compete against each other for occupancy of the
333 squamous sheet and in which substantial proportions of such sheets become colonised over
334 a lifetime by normal cell clones carrying driver mutations^{9,39}. Although this rampant
335 colonisation by driver clones in endometrium progresses with age, it is already well
336 advanced in some young women, and parity apparently has an inhibitory effect on it,

337 indicating that multiple factors influence its progression. More extensive studies of the
338 mutational landscape in normal endometrium are required to better assess how pregnancy,
339 the premenarcheal and postmenopausal states, hormonal contraceptive use and hormone
340 replacement therapies influence it and also the potential impact it has on pregnancy and
341 fertility.

342

343 The burdens of all mutation classes are lower in normal endometrial cells, including those
344 with drivers, than in endometrial cancers. However, these differences are most marked for
345 structural variants/copy number changes and for the extreme base substitution/indel
346 hypermutator phenotypes due to DNA mismatch repair deficiency and polymerase
347 delta/epsilon mutations which were not found in normal endometrium. The results therefore
348 suggest that in endometrial epithelium, and in other tissues thus far studied including colon,
349 oesophagus and skin, normal mutation rates are sufficient to generate large numbers of
350 microneoplastic clones with driver mutations behaving as normal cells, but that acquisition of
351 an elevated mutation rate and burden is associated with further evolution to invasive cancer.
352 Given that the endometrial epithelium is extensively colonised by clones of normal cells with
353 driver mutations in middle-aged and older women and that the lifetime risk of endometrial
354 cancer is only 3%⁴⁰, this conversion from microneoplasm to symptomatic malignancy
355 appears to be extremely rare. Driver mutations in normal endometrium often appear to arise
356 and initiate clonal expansion early in life. It is therefore plausible that some neoplastic clones
357 ultimately manifesting as cancer were initiated during childhood, although the fraction to
358 which this might apply is unclear.

359

360 This study has added endometrial epithelial cells to the set of normal cell types in which the
361 landscape of somatic mutations has been characterised. However, most normal tissues
362 have not been investigated in this way. The outcomes of the current studies showing
363 differences in mutation burdens, mutational signatures and prevalence of driver mutations
364 mandates a systematic characterisation of the somatic mutation landscape in all normal
365 human cell types.

366

367

368

369 **ACKNOWLEDGEMENTS**

370 This work was supported by the Wellcome Trust. LM is a recipient of a CRUK Clinical PhD
371 fellowship (C20/A20917) and Pathological Society of Great Britain and Ireland Trainee Small
372 Grant (Grant Reference No 1175). SFB was supported by the Swiss National Science
373 Foundation (P2SKP3-171753 and P400PB-180790). MAS is supported by a Rubicon
374 fellowship from NWO (019.153LW.038).

375

376 We thank Laura O'Neil, Calli Latimer and Paul Scott for technical support; Feran Nadeu and
377 Jingwei Wang for their advice on mutational signature extraction; Thomas J Mitchell, Nicola
378 Roberts and Andrew R.J. Lawson for their assistance with data analysis. We are also
379 grateful to the Cambridge Biorepository for Translational Medicine for the provision of
380 samples from deceased transplant organ donors.

381

382

383

384 **AUTHOR CONTRIBUTIONS**

385 MRS and LM designed the study and wrote the manuscript with contributions from all
386 authors. KSP, CAID, JJB, KM, MJL and LM obtained samples. PE and LM devised the
387 protocol for laser-capture microscopy, DNA extraction and sequencing of endometrial
388 glands. LM prepared sections, reviewed histology, micro-dissected and lysed endometrial
389 glands. YH assisted with tissue processing and section preparation. LM performed data
390 curation and analysis with the help from DL, THHC, MAS, KD, JN, PST, SFB, HLS, and RR.
391 THHC reconstructed phylogenetic trees. MAS devised filters for substitutions and structural
392 variants. DL, FM and SM assisted with signature analyses. IM assisted with statistical and
393 dnds analyses. PJC oversaw statistical analyses and performed analysis of structural
394 variants. MRS supervised the study.

395

396

397

398 REFERENCES

399

- 400 1 Alexandrov, L. B. Signatures of mutational processes in human cancer. *Nature* **500**,
401 415–421, doi:10.1038/nature12477 (2013).
- 402 2 Alexandrov, L. The Repertoire of Mutational Signatures in Human Cancer. *Environ
403 Mol Mutagen* **59**, 25-25 (2018).
- 404 3 Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**,
405 719-724, doi:10.1038/nature07943 (2009).
- 406 4 Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells
407 during life. *Nature* **538**, 260-264, doi:10.1038/nature19768 (2016).
- 408 5 Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic
409 mutations. *Nature* **561**, 473-478, doi:10.1038/s41586-018-0497-0 (2018).
- 410 6 Bae, T. et al. Different mutational rates and mechanisms in human cells at
411 pregastrulation and neurogenesis. *Science* **359**, 550-+, doi:10.1126/science.aan8690
412 (2018).
- 413 7 Franco, I. et al. Somatic mutagenesis in satellite cells associates with human skeletal
414 muscle aging. *Nat Commun* **9**, 800, doi:10.1038/s41467-018-03244-6 (2018).
- 415 8 Osorio, F. G. et al. Somatic Mutations Reveal Lineage Relationships and Age-
416 Related Mutagenesis in Human Hematopoiesis. *Cell Rep* **25**, 2308-2316 e2304,
417 doi:10.1016/j.celrep.2018.11.014 (2018).
- 418 9 Martincorena, I. et al. Tumor evolution. High burden and pervasive positive selection
419 of somatic mutations in normal human skin. *Science* **348**, 880-886,
420 doi:10.1126/science.aaa6806 (2015).
- 421 10 Martincorena, I. Somatic mutant clones colonize the human esophagus with age.
422 *Science* **362**, 911-917, doi:10.1126/science.aau3879 (2018).
- 423 11 Suda, K. et al. Clonal Expansion and Diversification of Cancer-Associated Mutations
424 in Endometriosis and Normal Endometrium. *Cell Rep* **24**, 1777-1789,
425 doi:10.1016/j.celrep.2018.07.037 (2018).
- 426 12 Lee-Six, H., et al. *The landscape of somatic mutation in normal colorectal
427 epithelial cells.*, <bioRxiv 416800> (2018).
- 428 13 Schmitt, M. W. et al. Detection of ultra-rare mutations by next-generation
429 sequencing. *P Natl Acad Sci USA* **109**, 14508-14513, doi:10.1073/pnas.1208715109
430 (2012).

431 14 Hoang, M. L. *et al.* Genome-wide quantification of rare somatic mutations in normal
432 human tissues using massively parallel sequencing. *Proc Natl Acad Sci U S A* **113**,
433 9846-9851, doi:10.1073/pnas.1607794113 (2016).

434 15 Michael A. Lodato, R. E. R., Craig L. Bohrson, Michael E. Coulter. Aging and
435 neurodegeneration are associated with increased mutations in single human
436 neurons. *Science* **359**, 555-559, doi:10.1126/science.aoa4426 (2018).

437 16 Lodato, M. A. *et al.* Somatic mutation in single human neurons tracks developmental
438 and transcriptional history. *Science* **350**, 94-98, doi:10.1126/science.aab1785 (2015).

439 17 Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse
440 outcomes. *N Engl J Med* **371**, 2488-2498, doi:10.1056/NEJMoa1408617 (2014).

441 18 Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood
442 DNA sequence. *N Engl J Med* **371**, 2477-2487, doi:10.1056/NEJMoa1409405
443 (2014).

444 19 Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion
445 and malignancies. *Nat Med* **20**, 1472-1478, doi:10.1038/nm.3733 (2014).

446 20 McKerrell, T. *et al.* Leukemia-associated somatic mutations drive distinct patterns of
447 age-related clonal hemopoiesis. *Cell Rep* **10**, 1239-1245,
448 doi:10.1016/j.celrep.2015.02.005 (2015).

449 21 Zink, F. *et al.* Clonal hematopoiesis, with and without candidate driver mutations, is
450 common in the elderly. *Blood* **130**, 742-752, doi:10.1182/blood-2017-02-769869
451 (2017).

452 22 Cousins, F. L., O, D. F. & Gargett, C. E. Endometrial stem/progenitor cells and their
453 role in the pathogenesis of endometriosis. *Best Practice & Research Clinical
454 Obstetrics & Gynaecology* **50**, 27-38, doi:10.1016/j.bpobgyn.2018.01.011 (2018).

455 23 Gargett, C. E., Schwab, K. E. & Deane, J. A. Endometrial stem/progenitor cells: the
456 first 10 years. *Hum Reprod Update* **22**, 137-163, doi:10.1093/humupd/dmv051
457 (2016).

458 24 Kaitu'u-Lino, T. J., Ye, L. & Gargett, C. E. Reepithelialization of the uterine surface
459 arises from endometrial glands: evidence from a functional mouse model of
460 breakdown and repair. *Endocrinology* **151**, 3386-3395, doi:10.1210/en.2009-1334
461 (2010).

462 25 Tempest, N., Maclean, A. & Hapangama, D. K. Endometrial Stem Cell Markers:
463 Current Concepts and Unresolved Questions. *Int J Mol Sci* **19**,
464 doi:10.3390/ijms19103240 (2018).

465 26 Le Gallo, M. & Bell, D. W. The emerging genomic landscape of endometrial cancer.
466 *Clin Chem* **60**, 98-110, doi:10.1373/clinchem.2013.205740 (2014).

467 27 Morice, P., Leary, A., Creutzberg, C., Abu-Rustum, N. & Darai, E. Endometrial
468 cancer. *Lancet* **387**, 1094-1108, doi:10.1016/S0140-6736(15)00130-0 (2016).

469 28 Onstad, M. A., Schmandt, R. E. & Lu, K. H. Addressing the Role of Obesity in
470 Endometrial Cancer Risk, Prevention, and Treatment. *J Clin Oncol* **34**, 4225-4230,
471 doi:10.1200/JCO.2016.69.4638 (2016).

472 29 Setiawan, V. W. *et al.* Type I and II Endometrial Cancers: Have They Different Risk
473 Factors? *Journal of Clinical Oncology* **31**, 2607-+, doi:10.1200/Jco.2012.48.2596
474 (2013).

475 30 Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic
476 Tissues. *Cell* **171**, 1029-1041 e1021, doi:10.1016/j.cell.2017.09.042 (2017).

477 31 Getz, G. Integrated genomic characterization of endometrial carcinoma (vol 497, pg
478 67, 2013). *Nature* **500**, doi:10.1038/nature12325 (2013).

479 32 Anglesio, M. S. *et al.* Cancer-Associated Mutations in Endometriosis without Cancer.
480 *N Engl J Med* **376**, 1835-1848, doi:10.1056/NEJMoa1614814 (2017).

481 33 Lac, V. *et al.* Iatrogenic endometriosis harbors somatic cancer-driver mutations. *Hum*
482 *Reprod*, doi:10.1093/humrep/dey332 (2018).

483 34 Lopez-Garcia, C., Klein, A. M., Simons, B. D. & Winton, D. J. Intestinal stem cell
484 replacement follows a pattern of neutral drift. *Science* **330**, 822-825,
485 doi:10.1126/science.1196236 (2010).

486 35 Snippert, H. J. *et al.* Intestinal crypt homeostasis results from neutral competition
487 between symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134-144,
488 doi:10.1016/j.cell.2010.09.016 (2010).

489 36 Rouhani, F. J. *et al.* Mutational History of a Human Cell Lineage from Somatic to
490 Induced Pluripotent Stem Cells. *PLoS Genet* **12**, e1005932,
491 doi:10.1371/journal.pgen.1005932 (2016).

492 37 Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat*
493 *Genet* **47**, 1402-1407, doi:10.1038/ng.3441 (2015).

494 38 Zhang, Y. Y., L., Kucherlapati, M. A Pan-Cancer Compendium of Genes Deregulated
495 by Somatic Genomic Rearrangement across More Than 1,400 Cases. *Cell Rep.* **24**,
496 515-527, doi:10.1016/j.celrep.2018.06.025 (2018).

497 39 Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with
498 age. *Science* **362**, 911-917, doi:10.1126/science.aau3879 (2018).

499 40 CRUK <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/uterine-cancer/risk-factors#heading-Zero>
500 Accessed on 19/12/2018

501

502

503

504 **FIGURE LEGENDS**

505

506 **Figure 1. Clonality of normal endometrial glands.** Individual normal endometrial glands
507 were laser-capture microdissected and whole genome sequenced. Majority (90%) of the
508 sampled endometrial glands were clonal, i.e. shared most recent common ancestor, with a
509 median variant allele frequency (VAF) between 0.3 and 0.5 for all substitutions.

510

511 **Figure 2. Mutation burden in normal endometrial glands.** (a) Substitutions accumulate in
512 the endometrium in a relatively linear fashion at an estimated rate of ~28 substitutions per
513 year (mixed-effect model, $P = 1.061\text{e-}07$). A positive correlation between age and
514 accumulation of indels (b), copy number and structural variants (c) and mutations attributed
515 to mutational signature SBS1 (d), SBS5 (e) and SBS18(f) was also observed. The fraction of
516 glands with driver mutations (g), mean number of driver mutations (h) and number of unique
517 (different) driver mutations (i) all show positive correlation with age.

518

519

520

521 **Figure 3. Histology images and phylogenetic trees of normal endometrial glands for**
522 **two selected individuals with an entirely neoplastic endometrium: 34-year-old (a,b)**
523 **and 60-year-old (c,d).** (a,b) Haematoxylin and eosin (H&E) images of individual endometrial
524 glands were taken after laser-capture microdissection (20x magnification). (c,d)
525 Phylogenetic trees were reconstructed using single base substitutions; the length of each
526 branch is proportional to the number of variants; a stacked barplot of attributed SBS

527 mutational signatures that contributed to each branch is then superimposed onto every
528 branch; signature extraction was not performed on branches with less than 100
529 substitutions. The ordering of signatures within each branch is for visualization purposes
530 only as it is not possible to time different signatures within individual branches. Endometrial
531 glands that shared more than 100 variants were considered to belong to the same clade
532 (indicated by the colour of the sample ID label). Labels for glands that did not belong to any
533 clades, are coloured white. The histology images are annotated accordingly. Single base
534 substitution (SBS) signatures are colour-coded (SBS1, SBS5 and SBS18); a small
535 proportion of substitutions across branches were not attributed to reference signatures
536 ('Unattributed').

537

538 **Figure 4. Phylogenetic trees of endometrial glands for all other donors.** Phylogenetic
539 trees for the other fifteen donors were reconstructed also using single base substitutions
540 with branch length proportional to the number of variants; the stacked bar plots represent
541 attributed SBS mutational signatures that contributed to each branch. Signature extraction
542 was not performed on branches with less than 100 substitutions. The ordering of signatures
543 within each branch is for visualization purposes only as it is not possible to time different
544 signatures within individual branches.

545

546 **Figure 5. Comparison between normal endometrial epithelium and endometrial
547 cancer. (a,b)** Normal endometrial glands show lower total mutation burden in comparison to
548 endometrial cancer. **(d,e)** Genes that are under significant positive selection ($dn/ds > 1$) in
549 normal endometrial epithelium and endometrial cancer. RHT, restricted hypothesis testing of
550 known cancer genes. *ERBB2* and *ERBB3* are under selection in normal endometrial
551 epithelium, but are not in endometrial cancer. **(f)** Identified driver mutations and their
552 distribution in normal endometrial glands and the two major types of endometrial cancer
553 (endometroid and serous carcinomas).

554

555

556 EXTENDED FIGURE LEGENDS

557

558 **Extended Data Figure 1. Clonality of endometrial glands and driver mutations.** **(a)** The
559 majority of sampled normal endometrial glands were clonal with a median variant allele
560 frequency (VAF) of 0.3 or above. The observed monoclonality of the glands was
561 independent of the driver status (Mann-Whitney two-sided test, $P = 0.1999$). **(b)** All glands
562 from the 19-year-old donor (PD37506) were clonal with a median VAF ≥ 0.3 , but there were
563 no detectable driver mutations.

564

565 **Extended Data Figure 2. Single Base Substitution (SBS) mutational signatures in
566 normal endometrial glands.** **(a)** Final catalogue of single base substitutions were used to
567 re-construct phylogenetic trees for 17 donors (due to the low number of high depth samples,
568 genomes from donor PD38812 were not included in this analysis). SBS signatures were
569 extracted on a per branch basis using a Hierarchical Dirichlet Process (HDP) with a set of 19
570 reference signatures that were identified in endometrial cancer ('priors') by the Mutational
571 Signatures working group of the Pan Cancer Analysis of Whole Genomes (PCAWG). **(b)**
572 HDP extracted components included the following: 'priors'/reference SBS signatures ($P1 =$
573 SBS1 and $P5 =$ SBS5); 'new' components that did not match any of the provided 19
574 reference signatures/priors ($N1, N2$ and $N3$) and 'Component 0' (Comp 0). **(c)** As $P1$ and $P5$

575 showed high cosine similarity (>0.95) to SBS1 and SBS5 signatures respectively (**f,g**), no
576 further deconvolution of these components was required. All other extracted components
577 were compared to the full set of 60 reference signatures. If a component had a cosine
578 similarity of >0.95 to any of the reference signatures (N2 = SBS18, **h**), no further
579 deconvolution was required (**d**). If a component did not show high cosine similarity to any of
580 the reference signatures, deconvolution was performed using a ‘deconvolution’ catalogue
581 comprising all of the extracted signatures (SBS1, SBS5 and SBS18). (**e**) Final exposures
582 were derived and signatures re-attributed to the individual branches.

583

584 **Extended Data Figure 3. Composite mutational spectra of small insertions and**
585 **deletions (indels) for each donor.** Indels were classified and composite mutational spectra
586 for each individual were generated; due to the relative sparsity of indels detected, no formal
587 signature extraction was performed.

588

589 **Extended Data Figure 4. Comparison between normal endometrial and endocervical**
590 **glands.** (**a**) An overview histology image of an ~2cm³ tissue biopsy sample from a 19-year-
591 old donor (PD37506). The image shows normal endometrial and adjacent endocervical
592 glands, which were subsequently micro-dissected. (**b**) Endometrial and endocervical glands
593 with a similar median variant allele frequency (VAF) of substitutions were compared. (**c**)
594 There was a ~2-fold difference in the mutation burden between the two types of glands.

595

596 **Extended Data Figure 5. Oncoplot of all driver mutations and their distribution across**
597 **individual endometrial gland samples and donors.** Each cell represents an individual
598 endometrial gland sample and is colour-coded to represent the total number of detected
599 driver mutations (0-3). *PIK3CA* was the most frequently mutated gene with at least one
600 mutation detected in 61% (11/18) of women. In some glands, these co-occurred with
601 mutations in *ZFHX3*, *ARHGAP35*, *FGFR2*, *FOXA2* and other genes that are also selected
602 for in endometrial cancer.

603

604 **Extended Data Figure 6. An example of copy-number neutral loss of heterozygosity**
605 **(cnn-LOH) in a normal endometrial gland.** (**a**) biallelic truncating mutation is seen in
606 *ZFHX3* (p.R715*) with every read carrying the variant. (**b**) an associated cnn-LOH is
607 observed on chromosome 16.

608

609 **Extended Data Figure 7. Comparison of mutation rates between endometrial**
610 **epithelium and other cell types.** The barplot shows a comparison of estimated mutation
611 rates (substitutions) for normal endometrial epithelial and other cell types from previously
612 published studies (liver, colon and small intestine⁴, oesophagus³⁹ and skeletal muscle⁷).

613

614

615

616

617

618

619

620

621

622

623

624

625

626 **SUPPLEMENTARY METHODS**

627

628 **Sample collection**

629 Anonymized snap-frozen endometrial tissue samples were obtained from five different
630 cohorts.

631

632 Cohort 1: Samples from individuals PD37605, PD37601, PD37607, PD37613, PD37594 and
633 PD37595 (age 29 to 46) were collected from women undergoing hysteroscopy examination
634 as part of infertility assessment at the Tommy's National Early Miscarriage Centre,
635 University Hospitals Coventry and Warwickshire NHS Trust. Informed consent was obtained
636 and biopsies collected and stored at the Arden Tissue Bank, University Hospitals Coventry
637 and Warwickshire NHS Trust in line with the protocols approved by the NRES Committee
638 South Central Southampton B (REC reference 12/SC/0526, 19/04/2013).

639

640 Cohort 2: Samples from individuals PD40535, PD39444, PD39953, PD39952, PD39954 and
641 PD40107 (age 24 to 69) were collected from residual tissues from transplant organ donors
642 with an informed consent obtained from donor's family (REC reference: 15/EE/0152 NRES
643 Committee East of England – Cambridge South).

644

645 Cohort 3: Individuals PD36804 and PD36805 (age 47 and 49), underwent total abdominal
646 hysterectomy for benign non-endometrial pathologies and biopsies were collected, snap
647 frozen and stored at the Human Research Tissue Bank, Cambridge University Hospitals
648 NHS Foundation Trust in line with the protocols approved by the NRES Committee East of
649 England (REC reference 11/EE/0011, 11/03/2011).

650

651 Cohorts 4 and 5: Samples from individuals PD37506, PD38812, PD37507 and PD40659
652 (age 19 to 81) were obtained at autopsy following death from non-gynaecological causes.
653 The use of this material was approved by the London, Surrey Research Ethics Committee
654 (REC reference 17/LO/1801, 26/10/2017) and East of Scotland Research Ethics Service
655 (REC reference: 17/ES/0102, 27/07/2017).

656

657 All endometrial biopsies underwent formal pathology review, which confirmed benign
658 histology.

659

660 **Laser-capture microdissection of endometrial glands**

661 Frozen and paraffin sections were used for laser-capture microdissection (LCM). For frozen
662 sections, endometrial tissue was embedded in optimal cutting temperature (OCT)
663 compound. 14 to 20-micron thick sections were generated at -20°C to -23°C, mounted on to
664 poly-ethylene naphthalate (PEN)-membrane slides (Leica), fixed with 70% ethanol, washed
665 twice with phosphate-buffered saline (PBS), and stained with Gill's haematoxylin and eosin
666 for 20 and 10 seconds respectively.

667

668 For paraffin sections, frozen endometrial tissue was first thawed at 4°C for 10-15 minutes,
669 then fixed in 70% ethanol and embedded in paraffin using standard histological tissue
670 processing. 8 to 10-micron thick sections were subsequently cut, mounted on to PEN-

671 membrane slides, and stained by sequential immersion in the following: xylene (two minutes,
672 twice), ethanol (100%, 1 minute, twice), deionised water (1 minute, once), Gill's
673 haematoxylin (10-20 seconds), tap water (20 seconds, twice), eosin (10 seconds, once), tap
674 water (10-20 seconds, once), ethanol (70%, 20 seconds, twice) and xylene or neo-clear
675 xylene substitute (10-20 seconds, twice).

676

677 Using laser-capture microscope (Leica LMD7), individual endometrial glands were first
678 visualised, then dissected (power 7, aperture 1, pulse 119 and speed 5) and collected into
679 separate wells in a 96-well plate. Overview pre- and post-dissection images were taken. In
680 addition, 200-500- μm^2 sections of either myometrium, endometrial stroma or Fallopian tube
681 epithelium were also obtained.

682

683 **Cell lysis, DNA extraction and whole genome sequencing of endometrial glands**

684 20 μl of an in-house lysis buffer containing 30 mM Tris-HCl pH 8.0 (Sigma Aldrich), 0.5%
685 Tween-20 (Sigma Aldrich), 0.5% NP-40/IGEPAL CA-630 (Sigma Aldrich) and 1.25 $\mu\text{g/ml}$
686 Proteinase K (Qiagen) was added to each well, vortexed (30 seconds) and spun down at
687 18°C (one minute at 1500 rpm). Samples were subsequently incubated in a thermal cycler
688 for 60 minutes at 50°C and 30 minutes at 75°C prior to storage at -80°C.

689

690 All samples in this study were processed using our recently developed low-input enzymatic
691 fragmentation-based library preparation method¹². Briefly, each 20 μl LCM lysate was mixed
692 with 50 μl Ampure XP beads (Beckman Coulter) and 50 μl TE buffer (Ambion; 10 mM Tris-
693 HCl, 1 mM EDTA) at room temperature. Following a 5 minute binding reaction and magnetic
694 bead separation, genomic DNA was washed twice with 75% ethanol. Beads were
695 resuspended in 26 μl TE buffer and the bead/genomic DNA slurry was processed
696 immediately for DNA library construction. Each sample (26 μl) was mixed with 7 μl of 5X
697 Ultra II FS buffer, 2 μl of Ultra II FS enzyme (New England BioLabs) and incubated on a
698 thermal cycler for 12 minutes at 37°C then 30 minutes at 65°C. Following DNA
699 fragmentation and A-tailing, each sample was incubated for 20 minutes at 20°C with a
700 mixture of 30 μl ligation mix and 1 μl ligation enhancer (New England BioLabs), 0.9 μl
701 nuclease-free water (Ambion) and 0.1 μl duplexed adapters (100 μM ; 5'-
702 AACTCTTCCCTACACGACGCTCTCCGATC*T-3', 5'-phos-
703 GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'). Adapter-ligated libraries were
704 purified using Ampure XP beads by addition of 65 μl Ampure XP solution (Beckman Coulter)
705 and 65 μl TE buffer (Ambion). Following elution and bead separation, DNA libraries (21.5 μl)
706 were amplified by PCR by addition of 25 μl KAPA HiFi HotStart ReadyMix (KAPA
707 Biosystems), 1 μl PE1.0 primer (100 μM ; 5'-
708 AATGATAACGGCGACCACCGAGATCTACACTCTTCCCTACACGACGCTCTCCGATC*T-
709 3') and 2.5 μl iPCR-Tag (40 μM ; 5'-
710 CAAGCAGAAAGACGGCATCGAGATXGAGATCGGTCTCGGCATTCCTGCTGAACCGCTC
711 TTCCGATC-3') where 'X' represents one of 96 unique 8-base indexes. The samples were
712 then mixed and thermal cycled as follows: 98 °C for 5 minutes, then 12 cycles of 98 °C for
713 30 s, 65°C for 30 s, 72 °C for 1 minute and finally 72 °C for 5 minutes. Amplified libraries
714 were purified using a 0.7:1 volumetric ratio of Ampure Beads (Beckman Coulter) to PCR
715 product and eluted into 25 μl of nuclease-free water (Ambion). DNA libraries were adjusted
716 to 2.4 nM and sequenced on the HiSeq X platform (illumina) according to the manufacturer's

717 instructions with the exception that we used iPCRtagseq (5'-
718 AAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTC-3') to read the library index.
719

720 All LCM samples were subjected to whole genome sequencing of 15-40.3x, using 150 base
721 pair clipped reads sequenced on HiSeq X platform (Illumina). For selected donors
722 (PD36804, PD36805 and PD37506), bulk samples (fragments of uterus or cervix) were also
723 whole genome sequenced.
724

725 **Variant calling**

726 **Substitutions**

727 Sequencing data were aligned to the reference human genome (NCBI build 37) using
728 Burrow-Wheeler Aligner (BWA-MEM)⁴¹. Duplicates were marked and removed and mapping
729 quality thresholds were set at 30. Single base somatic substitutions were called using
730 Cancer Variants through Expectation Maximization (CaVEMan) algorithm (major copy
731 number 5, minor copy number 2)⁴². To exclude germline variants, matched normal samples
732 (cervix, myometrium, Fallopian tube or endometrial stroma) were used to run the algorithm.
733

734 A set of previously described post-processing filters were subsequently applied:
735

736 (1) to remove common single nucleotide polymorphisms, variants were filtered against a
737 panel of 75 unmatched normal samples⁴²;

738 (2) to remove mapping artefacts associated with BWA-MEM, median alignment score of
739 reads supporting a mutation should be greater than or equal to 140 (ASMD>=140)
740 and fewer than half of the reads should be clipped (CLPM=0)¹²;

742 (3) to remove artefacts that are specific to the library preparation for LCM samples, two
743 additional filters were used. A fragment-based filter, which is designed to remove
744 overlapping reads resulting from relatively shorter insert sizes allowed in this
745 protocol that can lead to double counting of variants, and a cruciform filter, which
746 removes erroneous variants that can be introduced due to the incorrect processing
747 of cruciform DNA. For each variant, the standard deviation (SD) and median
748 absolute deviation (MAD) of the variant position within the read was calculated
749 separately for positive and negative strands reads. If a variant was supported by a
750 low number of reads for one strand, the filtering was based on the statistics
751 calculated from the reads derived from the other strand and it was required that
752 either: (a) ≤ 90% of supporting reads report the variant within the first 15% of the
753 read as determined from the alignment start, or (b) that the MASD >0 and SD>4.
754 Where both strands were supported by sufficient reads, it was required for both
755 strands separately to either: (a) ≤90% of supporting reads report the variant within
756 the first 15% of the read as determined from the alignment start, (b) that the MAD>2
757 and SD>2, or (c) that at least one strand has fulfilled the criteria MAD>1 and SD>10.
759

760 **Validation experiments and sensitivity**

761 To validate somatic variants, for selected donors, pairs of biological 'near-replicates' were
762 obtained. For these experiments, we collected two samples from the same endometrial
763 gland which was identified on two or more consecutive levels using z-stacking approach;
764 each sample was processed separately with an independent DNA extraction, library
765 preparation and whole genome sequencing. As these samples were obtained from the same
766 glands, they would represent derivatives of the same clone and therefore the same
767 sensitivity would be assumed in both samples in each pair. The maximum likelihood
768 estimate for sensitivity (s) was then calculated as follows:

769

770

$$S = \frac{2n_2}{n_1 + 2n_2}$$

771

772 where n_1 is the number of variants called only in one of the two LCM samples and n_2 is the
773 number of variants called in both LCM samples in each pair. Using this approach, the mean
774 sensitivity of somatic mutation variant calling was estimated at >86% (range 0.70-0.95%).

775

776 **Indels**

777 Insertions and deletions were called using cgpPindel^{43,44}. To remove germline variants the
778 algorithm was run with the same matched normal samples that were used for calling
779 substitutions. Post-processing filters were applied as previously described⁴². In addition, a
780 'Qual' filter (the sum of the mapping qualities of the supporting reads) of at least 300 and a
781 depth cut-off of at least 15 reads were used.

782

783 **Copy number variants and structural variants**

784 Allele-specific copy number profiles were reconstructed for the endometrial gland samples
785 by ASCAT^{45,46} using matched samples as described above, ploidy of 2 and contamination
786 with other cell types of 10%. Only samples with a minimum coverage of 15X and above
787 were used. All putative copy number changes were visually inspected for copy number
788 profiles on Jbrowse⁴⁷.

789

790 Structural variants (SVs) in endometrial glands were called using matched samples (as
791 described above) with the Breakpoints Via Assembly (BRASS) algorithm and further
792 annotated by GRASS (<https://github.com/cancerit/BRASS>). Potential SVs are detected for
793 the sample of interest and read-pairs clusters supporting the SV are used for breakpoint
794 sequence de novo assembly. Absence of supporting evidence in the matched control
795 indicates that the SV was acquired in the sample of interest. The isolation of minute amounts
796 of DNA for sequencing in combination with the LCM enzymatic fragmentation-based library
797 preparation procedure introduces additional artefacts and additional post-processing filtering
798 was performed in two phase:

799

800 **Further annotation of SVs with statistics that detect LCM specific artefacts**

801 All SVs detected by BRASS were further annotated by AnnotateBRASS. Each SV is defined
802 by two breakpoints and their genomic coordinates.

803

804 (A) The following statistics were determined for each breakpoint separately:

805

1. The total number of reads supporting the SV.
2. The total number of unique reads supporting the SV, based on alignment position and read orientation.
3. The standard deviation of the alignment positions of reads supporting the SV.
4. The number of chromosomes, based on read-pairs not supporting the SV, to which one read mapped while the mate-read aligned to the SV breakpoint.
5. The number of reads supporting the SV that had an alternative alignment (XA-tag).
6. The number of reads supporting the SV that had an alternative alignment score (XS-tag) similar to the current alignment score.
7. The percentage of read-pairs not supporting the SV with a discordant inferred insert size (default: $\geq 1000\text{bp}$).

817

818 (B) A wider search for read-pairs supporting the SV is initiated and the following statistics
819 were calculated for each breakpoint separately:

820

1. The total number of reads supporting the SV.
2. The total number of unique reads supporting the SV, based on alignment position and read orientation.
3. The standard deviation of the alignment positions of reads supporting the SV.
4. The number of reads supporting the SV that had an alternative alignment.
5. The number of reads supporting the SV that had an alternative alignment score similar to the current alignment score.

828

829 (C) Reads spanning the SV breakpoints are often clipped. Clipped sequences of sufficient
830 length can be aligned to other positions on the genome (i.e., supplementary alignment) and
831 it is expected that these align to the proximity of the other SV breakpoint. Based on the
832 clipping positions and supplementary alignments the following was determined for each SV:

833

1. Whether the clipped sequences of read-pairs spanning a SV breakpoint align in the proximity of the other SV breakpoint.
2. Whether the clipping within read-pairs supporting the SV occurred at roughly the same genomic position (default: all clipping positions occurred within 10bp of each other).

839

840 (D) BRASS uses a single matched control and a panel of normals (PoN, bulk WGS) to
841 determine whether a SV is somatic. SVs observed in the sample of interest but not in the
842 matched control or PoN are considered somatic. However, due to the difference in library
843 preparation and the variance of spatial genomic coverage observed it is not always possible
844 to accurately assess the validity of the SV. Two different approaches were implemented to
845 determine whether the SV is somatic:

846

847 1. A wider search in the matched control sample was performed to search for read-
848 pairs that could support the SV. The SV was still considered detected in case the
849 discovered read-pairs were insufficient for breakpoint sequence de novo assembly.
850 2. Additional controls can be defined in case multiple samples have been isolated for
851 the same individual. Samples from the same individual with little genetic relationship,
852 as determined from the SNVs and indels, can be used as controls to determine
853 whether the detected SV is germline or a recurrent artifact.

854

855 **Post-hoc filtering of SVs based on a combination of the above statistics.**

856 SVs were further filtered based on the described statistics. The optimal set of statistics and
857 their most practical thresholds depends on the achieved coverage and stringency of filtering
858 desired. At default the following criteria were used for detecting somatic SVs:

859 1. For each breakpoint there must be ≥ 4 unique reads supporting the SV (**A.2**).
860 2. The alignment position standard deviation must be > 0 (**A.3**).
861 3. At each breakpoint there are read-pairs not supporting the SV that map to < 5 other
862 chromosomes (**A.4**).
863 4. The total number of chromosomes mapped to by read-pairs not supporting the SV for
864 both breakpoints should be < 7 (**A.4**).
865 5. The percentage of reads supporting the SV with alternative alignments or alternative
866 alignments with similar alignment scores should be $\leq 50\%$ for both SV breakpoints
867 separately (**A.5-A.6**).
868 6. The percentage of discordant read-pairs not supporting the SV should be $\leq 7.5\%$ of
869 total read-pairs for both SV breakpoints separately (**A.7**).
870 7. For the wider search of SV-supporting read-pairs the same thresholds apply as
871 under criteria 1-6 (**B.1-B.5**).
872 8. There are no read-pairs in the matched control that support the SV (**C.1**).
873 9. The SV is not detected in any of the other control samples, or there were ≤ 2
874 samples carrying the same SV and the proportion of control samples carrying the SV
875 was $< 1/3$ of the defined control set (**C.2**).
876 10. It was not allowed for read-pairs supporting the SV to have widely divergent clipping
877 positions in terms of genomic location for both SV breakpoints separately (**D.2**).
878

879

880 **Detection of driver mutations**

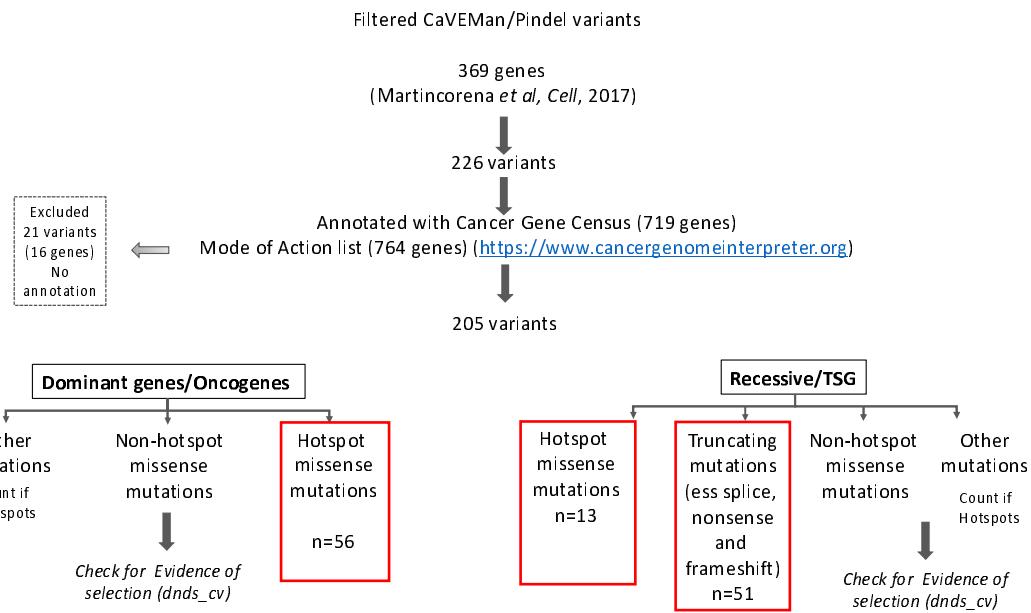
881 Analysis of driver variants in the normal endometrial glands was performed in two parts.
882 First, filtered CaVEMan and Pindel variants were intersected against a previously published
883 list of 369 genes that are under selection in human cancers³⁰. All non-synonymous
884 mutations were annotated to indicate mode of action using a Cancer Gene Census (719
885 genes) and a catalogue of 764 genes (<https://www.cancergenomeinterpreter.org>).
886 Truncating variants (nonsense, frameshift and essential splice), which resided in
887 recessive/tumour-suppressor genes (TSG) were declared likely drivers. Missense mutations
888 in recessive/TSG and dominant/oncogenes were triaged against a database of validated
889 hotspot mutations (http://www.ncbiportal.org/mutation_mapper). All mutations that were
890 shown to be known mutational hotspots or 'likely oncogenic' were declared drivers. In

891 addition, identified activating mutations in mutational hotspots in genes *RRAS2* and *SOS1*,
892 involving the RAS/MAPK pathway were declared as likely drivers.
893

894 Second, to identify genes that are under positive selection in normal endometrium we used
895 the dN/dS³⁰ method that is based on the observed:expected ratios of non-
896 synonymous:synonymous mutations. The analysis was carried out for the whole genome
897 ($q < 0.05$ and $q < 0.01$) and for 369 known cancer genes³⁰ (RHT, restricted hypothesis testing,
898 $q < 0.05$). Eleven genes were found to be under positive selection in normal endometrial
899 glands. The output of this analysis was also used to assess whether missense mutations in
900 genes that are under positive selection in normal and/or malignant endometrium (*PIK3CA*,
901 *ERBB2*, *ERBB3*, *FBXW7* and *CHD4*) but are not known mutational hotspots, are likely to be
902 drivers. If q -value was < 0.05 , we used the following calculation to assess the likelihood of a
903 variant being a driver:
904

$$f = \frac{\omega - 1}{\omega}$$

905
906
907 If f was ≥ 0.95 , then all missense mutations in that gene were declared likely drivers.
908



909
910 To compare patterns of selection in normal endometrial epithelium and cancer, we
911 performed dNdS analysis on publicly available calls from the The Cancer Genome Atlas,
912 TCGA.
913

914 **Phylogenetic tree reconstruction**

915 Phylogenies for endometrial glands were reconstructed for seventeen donors. Due to the
916 low number of available samples, donor PD38812 was not included in this analysis. We first
917 generated trees using substitutions called by CaVEMan; matched normal samples were
918 used to exclude germline variants and post-processing filters were applied as above. Final
919 variants were recalled in all samples from each donor using an in-house re-genotyping
920 algorithm (cgpVAF). Variants with a VAF>0.3 were noted to be present ('1'), VAF<0.1 absent

921 ('0') and between 0.1 and 0.3 as ambiguous ('?'). This approach excludes private sub-clonal
922 variants from the tree building. The tree was reconstructed using a maximum parsimony
923 approach⁴⁸ and branch support was calculated using 1000 bootstrap replicates. Nodes with
924 a confidence lower than 50 were collapsed into polytomies and branch lengths of the
925 collapsed tree were determined by the number of assigned substitutions.
926

927 The constructed phylogenies were validated using indels called by Pindel and filtered as
928 above. The same approach was applied for the final indel matrices. Although the lower
929 number of indels resulted in more polytomous tree, the overall tree topologies were
930 reconcilable with those generated using substitutions.
931

932 Cancer driver mutations, copy number and structural variants were annotated manually in
933 the trees.
934

935 **Mutational signature analysis**

936 Mutational signature extraction was performed using mutations assigned to every branch of
937 the reconstructed phylogenetic trees and each branch was treated as an individual sample.
938 Such approach allows characterisation and differentiation of specific mutational processes
939 that were operative at various times in individual glands. Substitutions were first categorised
940 into 96 classes following the method used by the Mutational Signature working group of the
941 Pan Cancer Analysis of Whole Genomes (PCAWG)². Single base substitution (SBS)
942 signatures were then extracted using the HDP package
943 (<https://github.com/nicolaroberts/hdp>) that utilises hierarchical Bayesian Dirichlet process.
944 Code and the input mutations are available at <https://github.com/LuizaMoore/Endometrium>.

945 SBS signature analysis was performed in 3 steps: extraction, deconvolution and re-
946 attribution. First, extraction was performed with conditioning on the set of mutational
947 signatures that have been previously reported in endometrial cancer by PCAWG: SBS1,
948 SBS2, SBS3, SBS5, SBS6, SBS7a, SBS7b, SBS10a, SBS10b, SBS13, SBS14, SBS15,
949 SBS20, SBS21, SBS26, SBS28, SBS30, SBS44 and SBS54 (reference). Such an approach
950 not only allows discovery of new mutational signatures, but also simultaneous matching to
951 the provided reference signatures. The extraction was run with 50,000 burn-in iterations
952 (parameter 'burnin'), with a spacing of 500 iterations (parameter 'space') and 250
953 samples/iterations were collected (parameter 'n'). After each Gibbs sampling iteration, 3
954 iterations of concentration parameter sampling were performed (parameter 'cpiter') and
955 components were extracted (Cos merge 0.9, sample 2).
956

957 Two reference signatures were extracted: SBS1 (P1, cosine similarity 0.997) and SBS5 (P5,
958 cosine similarity 0.983), which were added to the 'deconvolution' catalogue. Other extracted
959 components (New 1 = N1, New 2 = N2, New 3 = N3 and Component 0 = Comp 0) that did
960 not fit the provided set of 19 reference signatures were examined for similarity to the full set
961 of 60 reference SBS signatures (PCAWG)². Component N2 showed high cosine similarity to
962 SBS18 (0.968), therefore did not require further deconvolution and was added to the
963 'deconvolution' catalogue. No other HDP components showed cosine similarity of >0.95 to
964 any of the reference signatures and therefore required further deconvolution. The final
965 'deconvolution' catalogue comprising the three extracted SBS signatures (SBS1, SBS5 and
966 SBS18) was then used to decipher all other components. Final SBS signature exposures
967 were re-calculated and signatures re-attributed to individual samples.

968
969 Indels were classified using PCAWG method² and composite mutational spectra were
970 generated for each donor. However, given the relatively low numbers of indels, no formal
971 signature extraction was performed.
972

973 **Calculations of mutation burden and estimation of mutation rate**
974

975 To account for the non-independent sampling per patient we used mixed effects models. We
976 tested features with a known effect on mutation burden or endometrial cancer risk; age,
977 Read depth & VAF, BMI and Parity. All statistical analyses were performed in R and are
978 summarised in the Supplementary Results.
979

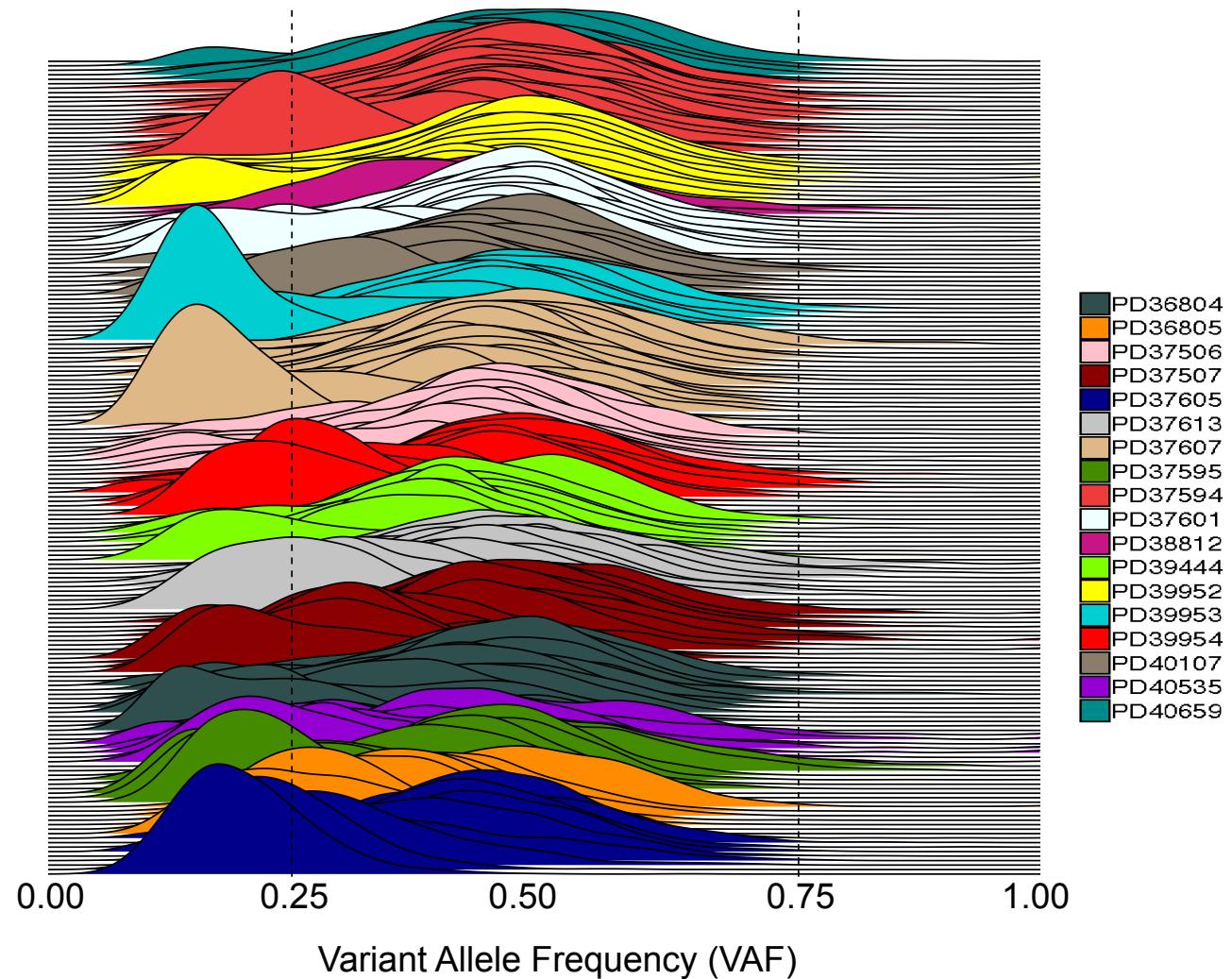
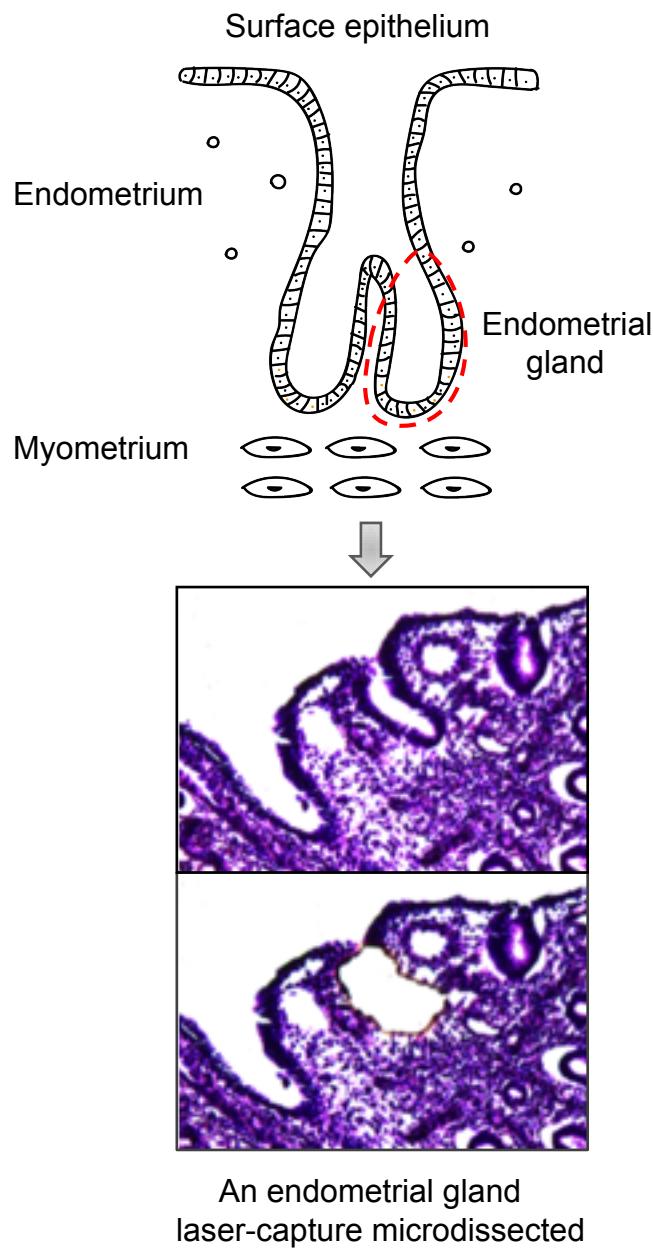
980
981 **Data availability**
982 Whole genome sequencing data are deposited in the European Genome 715 Phenome
983 Archive (EGA) with accession number 716 EGAS00001002471.
984

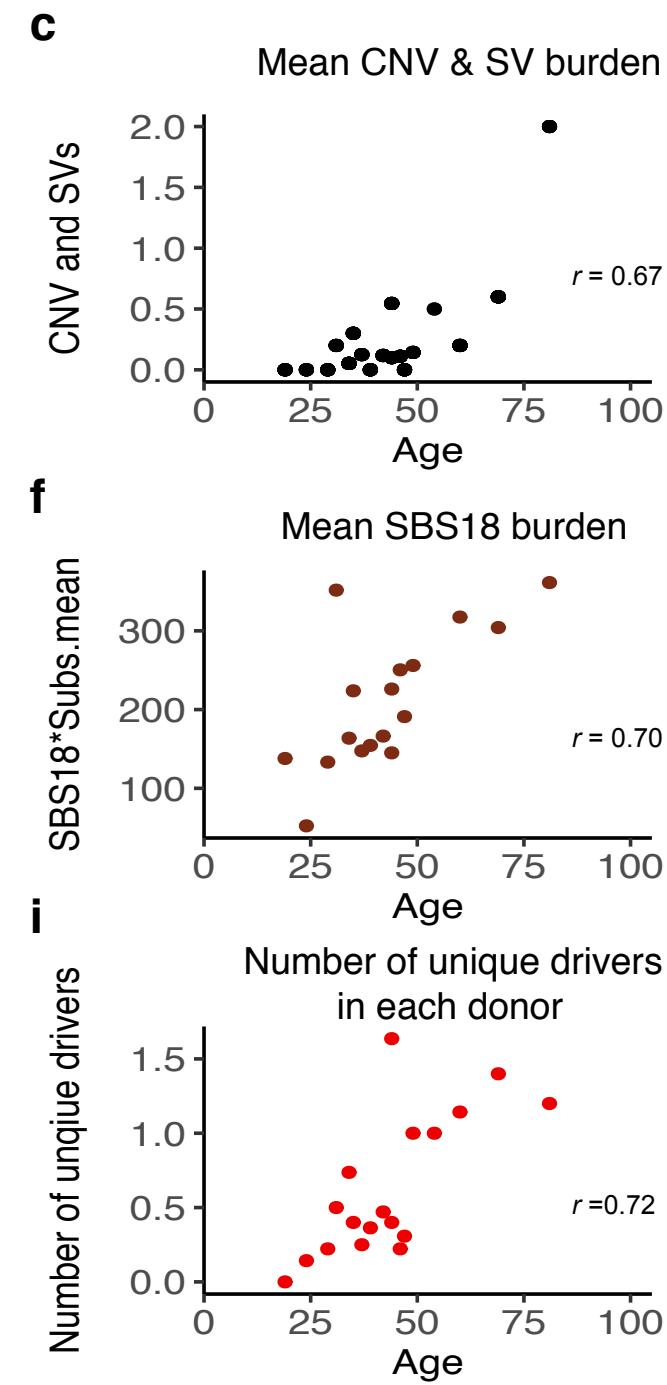
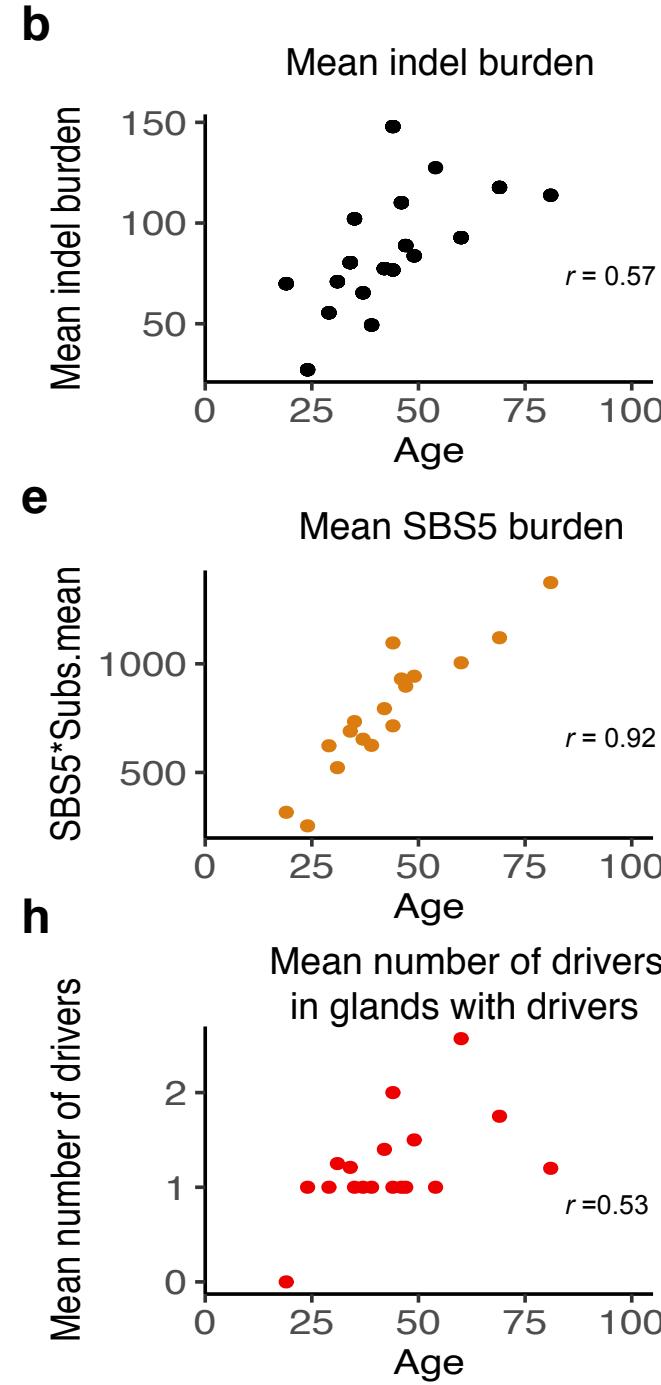
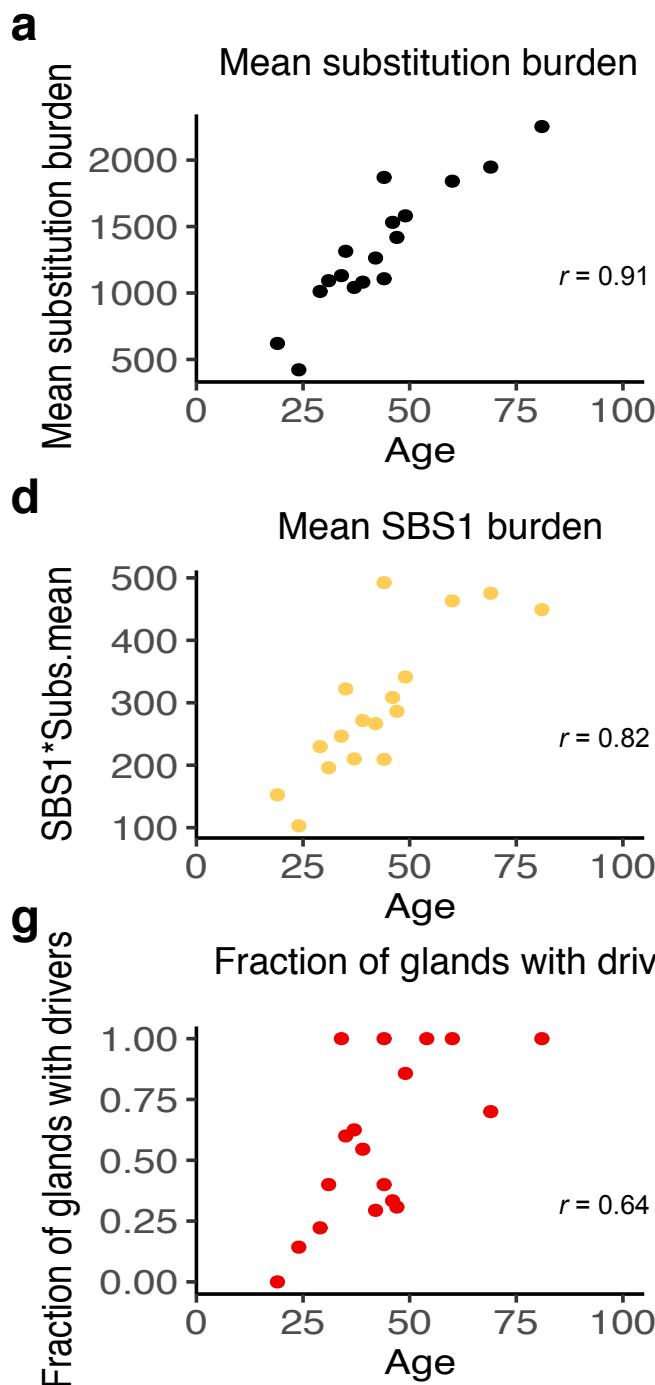
985
986 **Code availability**
987 Code for statistical analyses on total substitution and driver mutation burdens is included in
988 the supplementary material. Code for mutational signature extraction is deposited on GitHub
989 at <https://github.com/LuizaMoore/Endometrium>. All other code is available on request.
990

991
992
993 **REFERENCES FOR SUPPLEMENTARY METHODS**
994

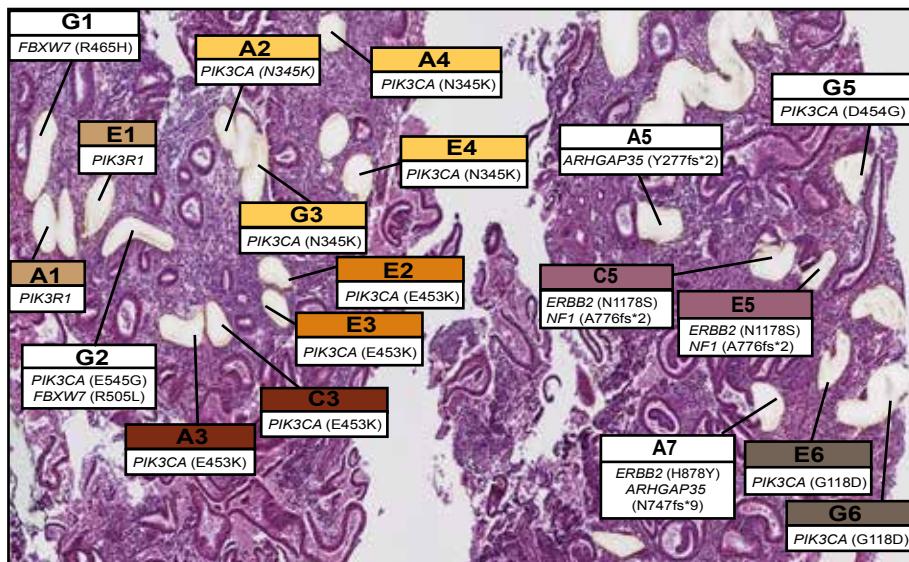
995 41 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
996 transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
997 42 Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers.
998 *Cell* **149**, 979-993, doi:10.1016/j.cell.2012.04.024 (2012).
999 43 Raine, K. M. *et al.* cgpPindel: Identifying Somatically Acquired Insertion and Deletion
1000 Events from Paired End Sequencing. *Curr Protoc Bioinformatics* **52**, 15 17 11-12,
1001 doi:10.1002/0471250953.bi1507s52 (2015).
1002 44 Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth
1003 approach to detect break points of large deletions and medium sized insertions from
1004 paired-end short reads. *Bioinformatics* **25**, 2865-2871,
1005 doi:10.1093/bioinformatics/btp394 (2009).
1006 45 Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci
1007 U S A* **107**, 16910-16915, doi:10.1073/pnas.1009843107 (2010).
1008 46 Raine, K. M. *et al.* ascatNgs: Identifying Somatically Acquired Copy-Number
1009 Alterations from Whole-Genome Sequencing Data. *Curr Protoc Bioinformatics* **56**, 15
1010 19 11-15 19 17, doi:10.1002/cpbi.17 (2016).
1011 47 Buels, R. *et al.* JBrowse: a dynamic web platform for genome visualization and
1012 analysis. *Genome Biol* **17**, 66, doi:10.1186/s13059-016-0924-1 (2016).

1013 48 Hoang, D. T. *et al.* MPBoot: fast phylogenetic maximum parsimony tree inference
1014 and bootstrap approximation. *BMC Evol Biol* **18**, 11, doi:10.1186/s12862-018-1131-3
1015 (2018).
1016

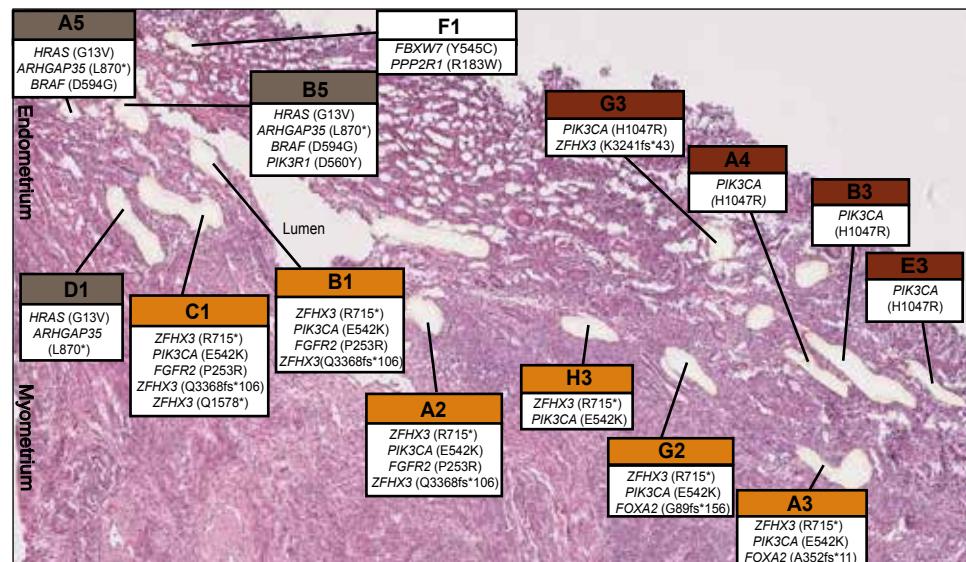




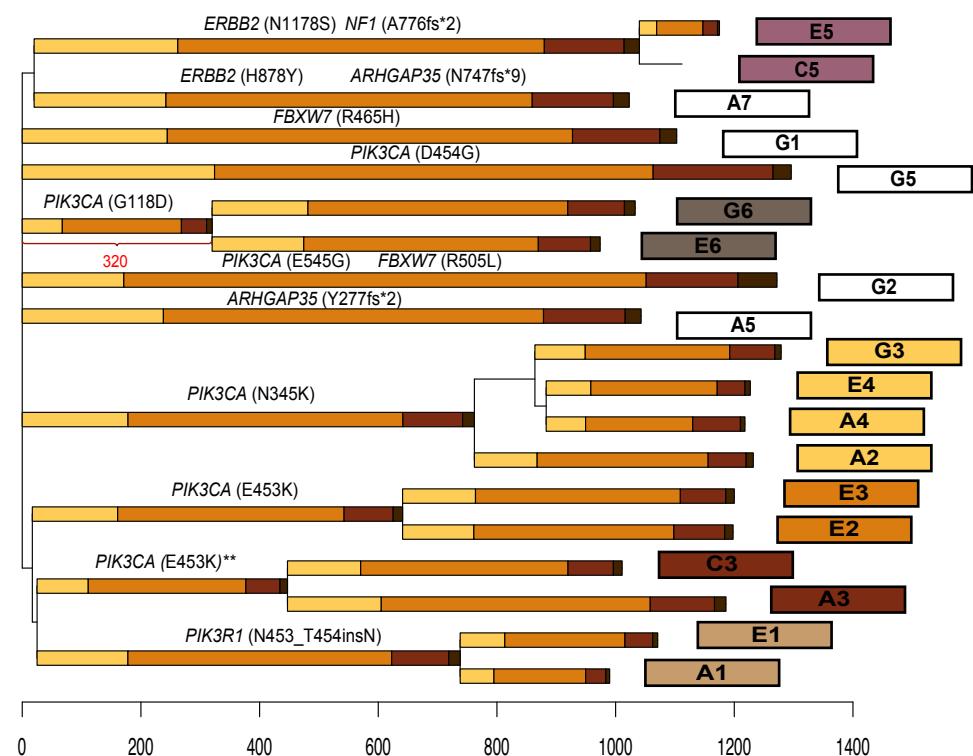
34-year-old donor (PD37607)



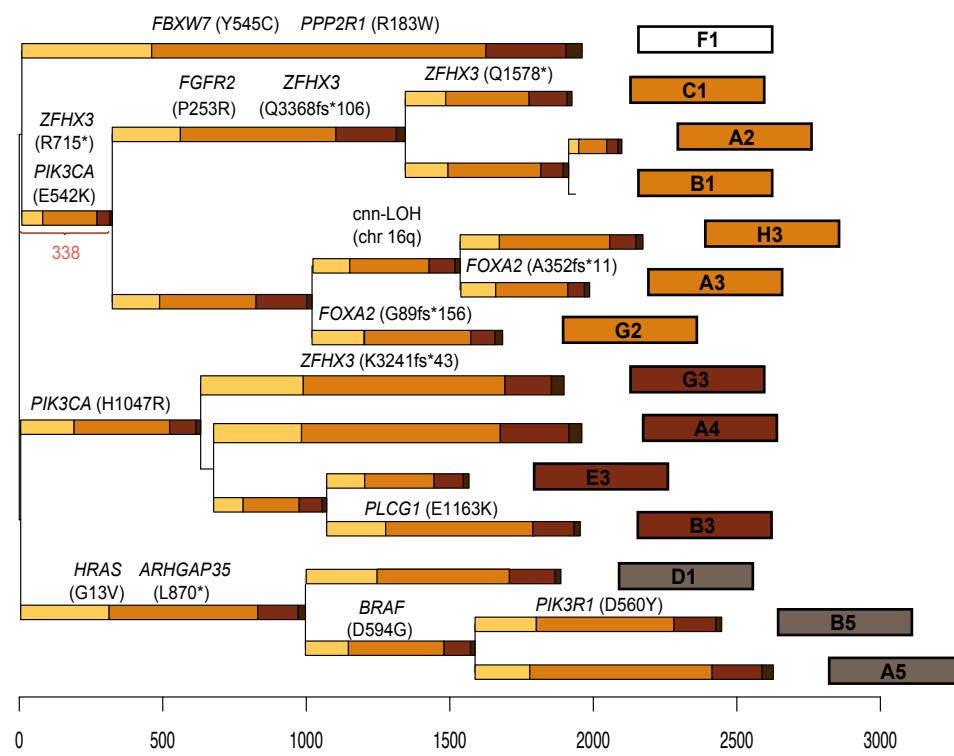
60-year-old donor (PD37507)



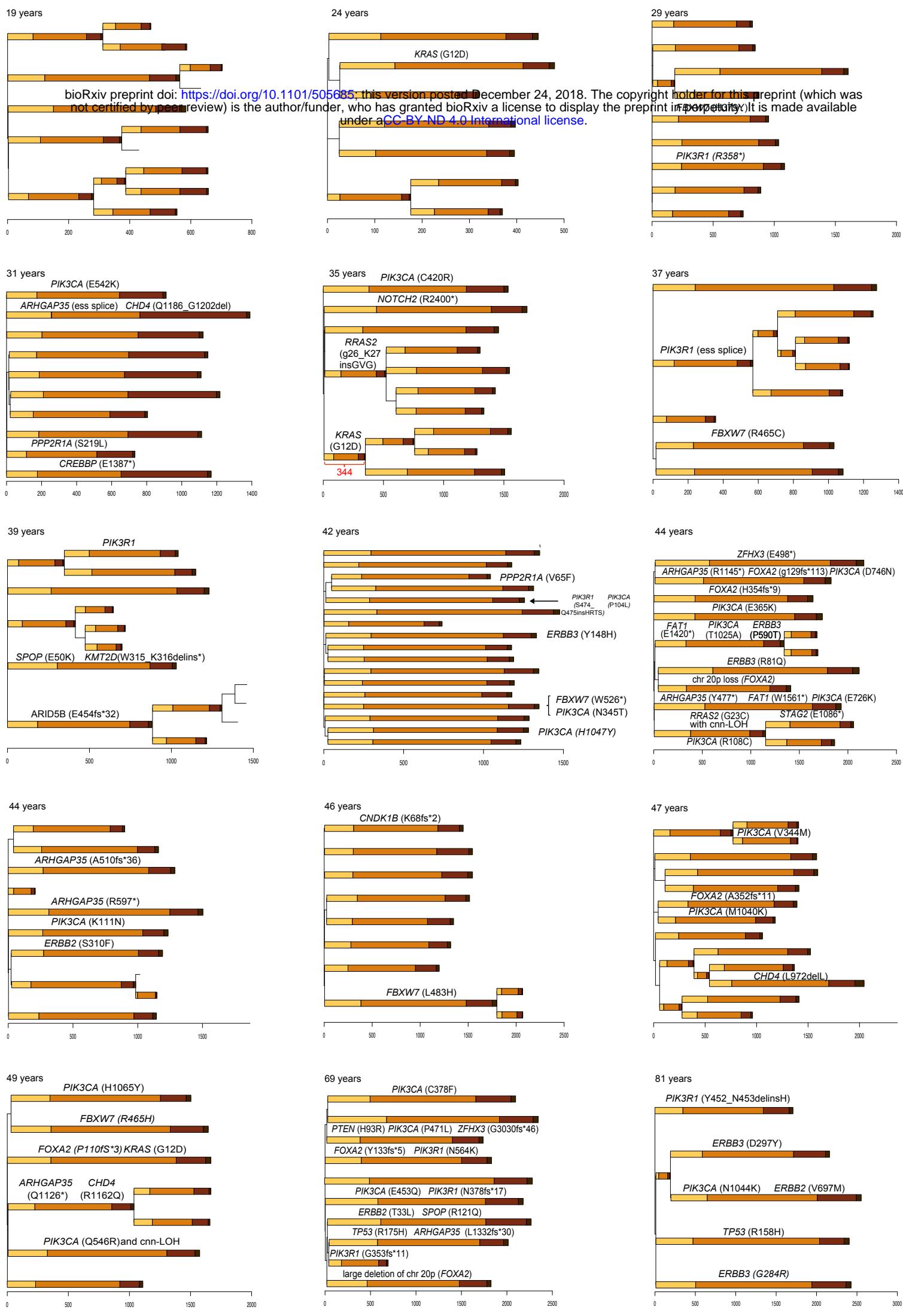
C

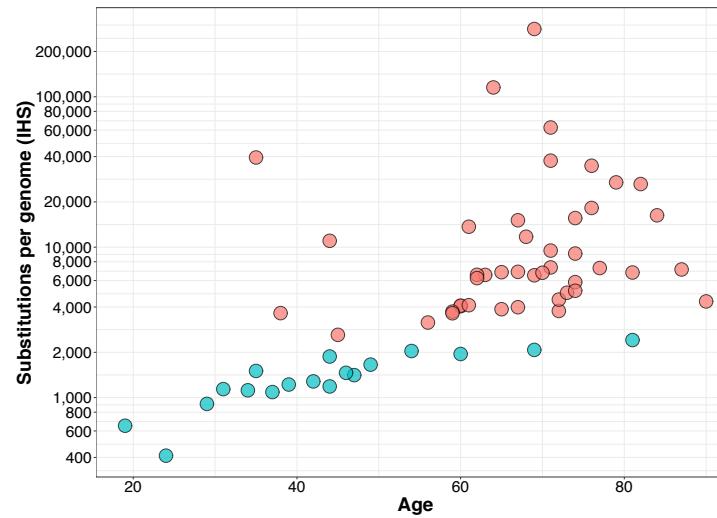
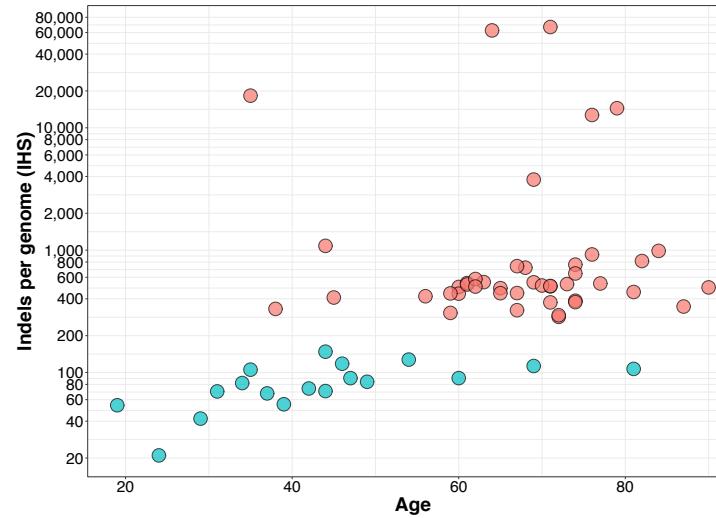


d



■ SBS1 ■ SBS5 ■ SBS18 ■ Unattributed

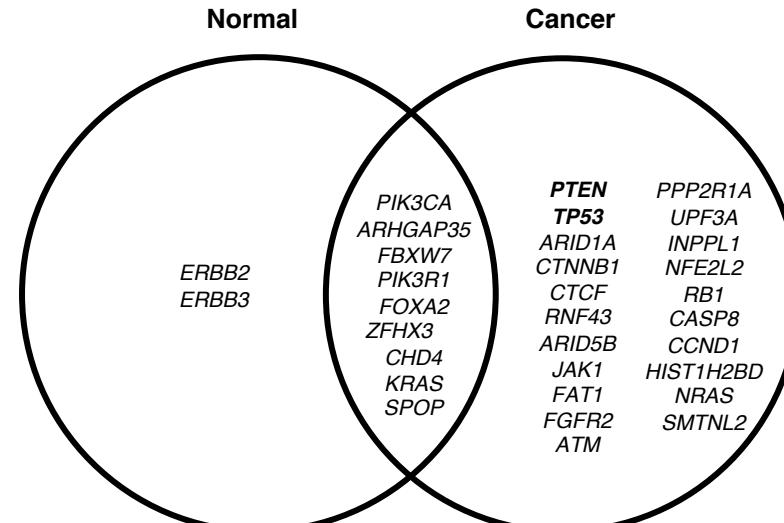
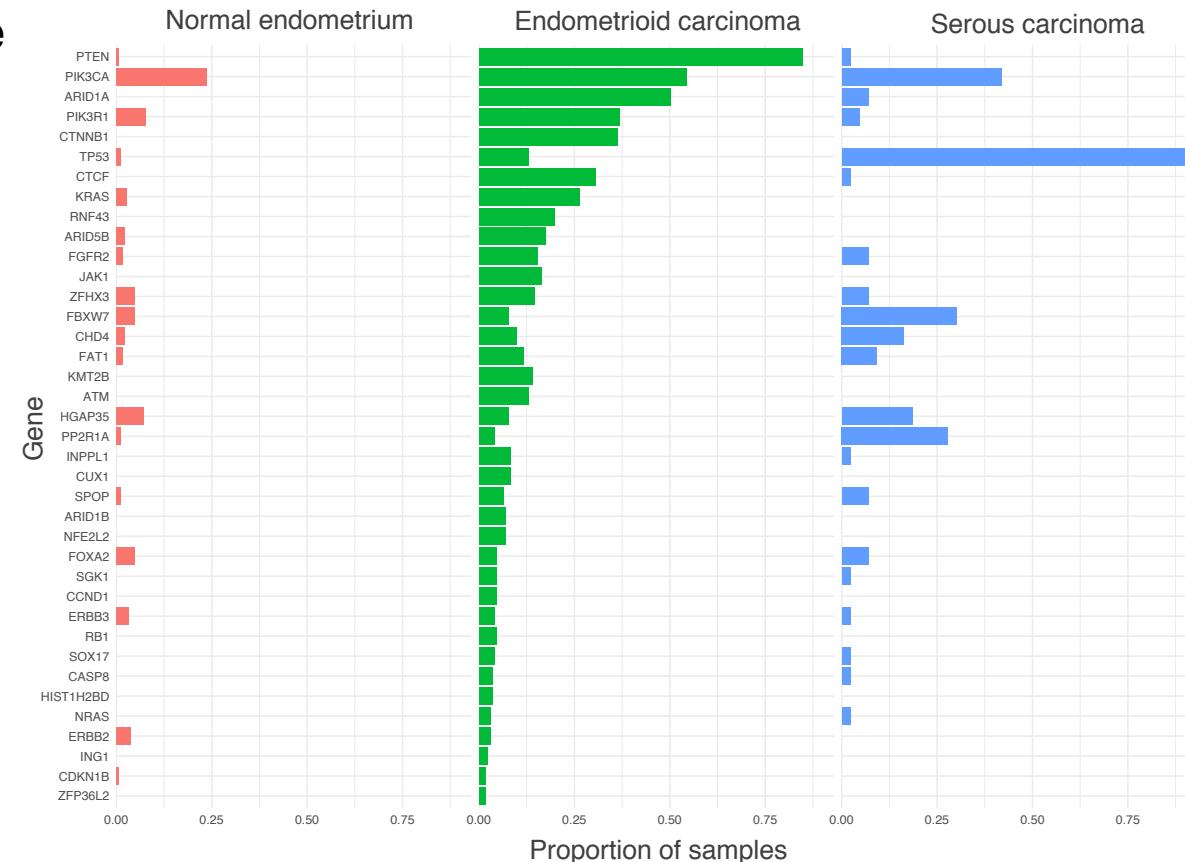


a**b****c**

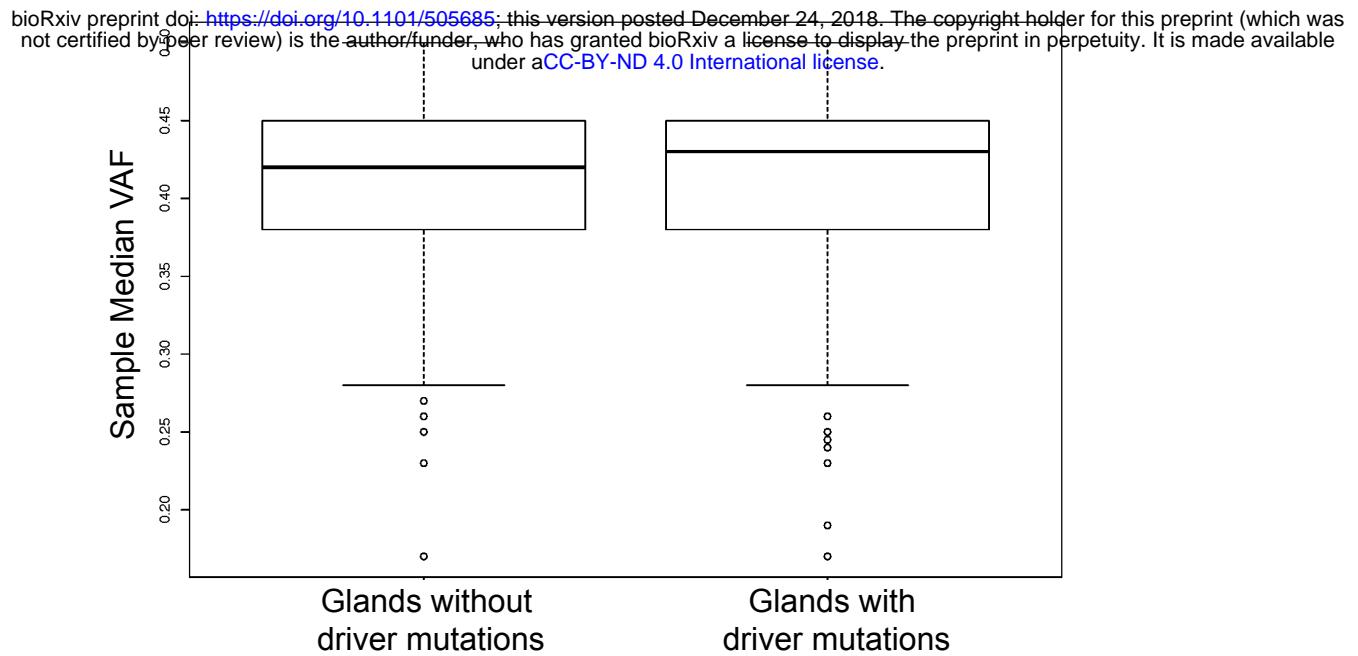
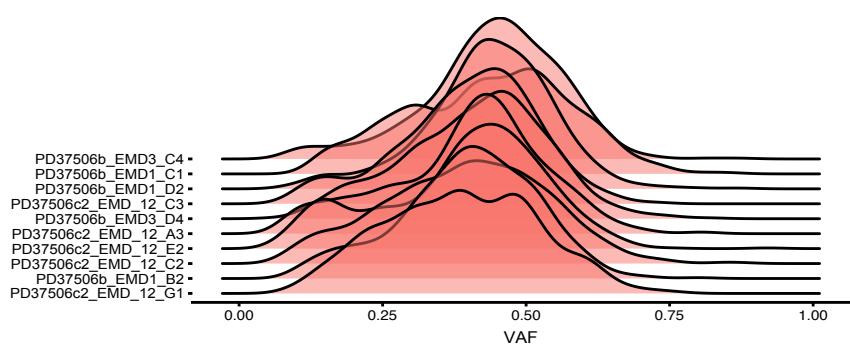
Cancer (n=218) **Normal (n=182)**

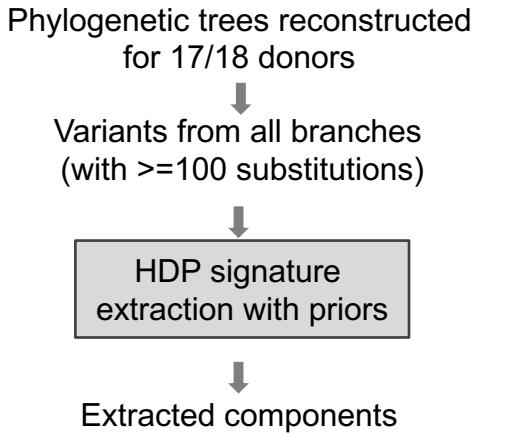
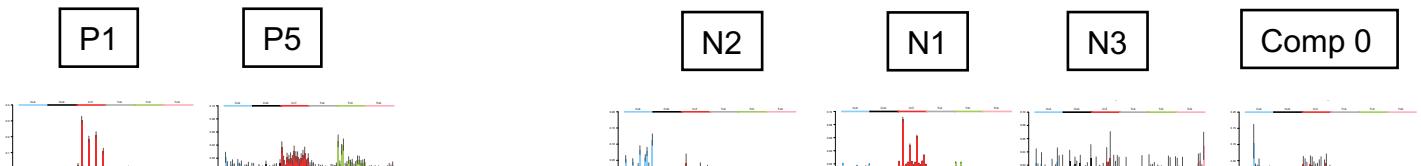
<i>PTEN</i>
<i>TP53</i>
<i>ARHGAP35</i>
<i>PIK3CA</i>
<i>CTNNB1</i>
<i>KRAS</i>
<i>CTCF</i>
<i>ARID1A</i>
<i>PIK3R1</i>
<i>FBXW7</i>
<i>ARHGAP35</i>
<i>ARID5B</i>
<i>ZFHX3</i>
<i>SPOP</i>
<i>FOXA2</i>
<i>PPP2R1A</i>
<i>FGFR2</i>
<i>RNF43</i>
<i>CHD4</i>
<i>NFE2L2</i>
<i>FAT1</i>
<i>INPPL1</i>
<i>ARID1B</i>
<i>SOX17</i>
<i>JAK1</i>
<i>KMT2B</i>
<i>HIST1H2BD</i>
<i>CCND1</i>
<i>ATM</i>
<i>ING1</i>
<i>CASP8</i>
<i>RB1</i>
<i>NRAS</i>
<i>ZFP36L2</i>
<i>CDKN1B</i>
<i>SGK1</i>
<i>CUX1</i>

- q<0.01
- q<0.05
- RHT (q<0.05)

d**e**

Proportion of samples

a**b**

a**b****c**

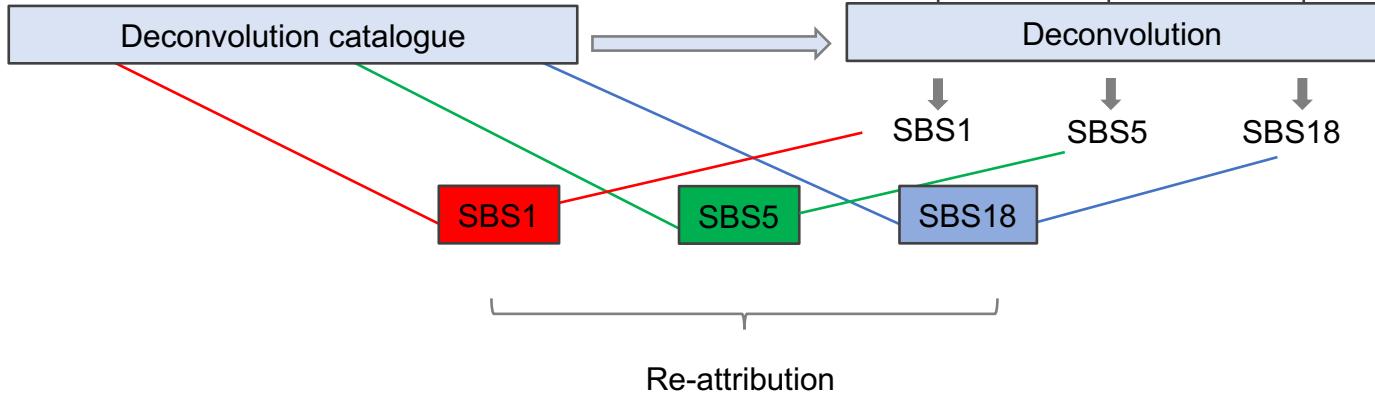
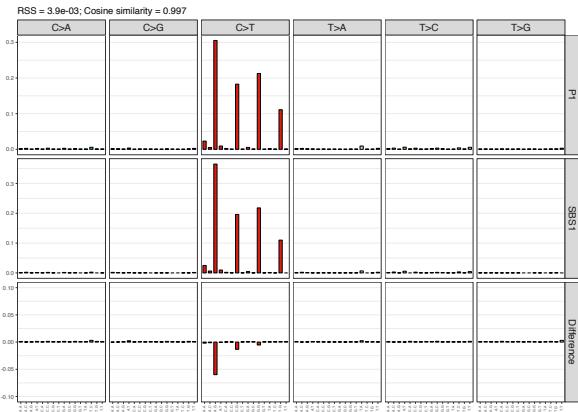
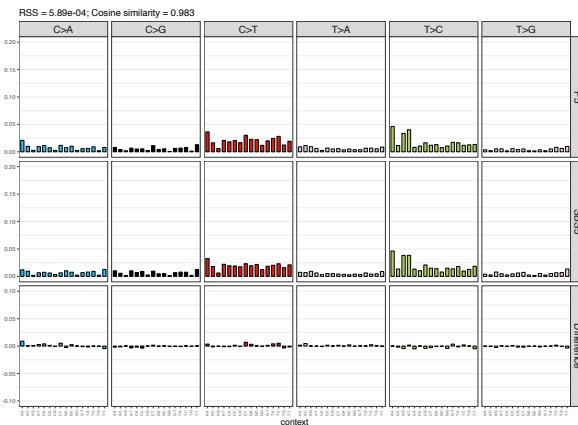
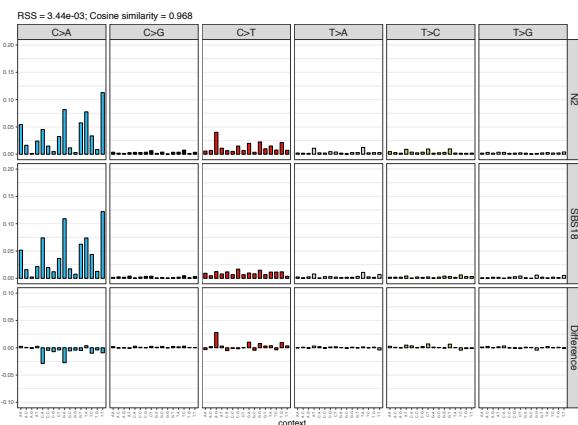
No deconvolution
required

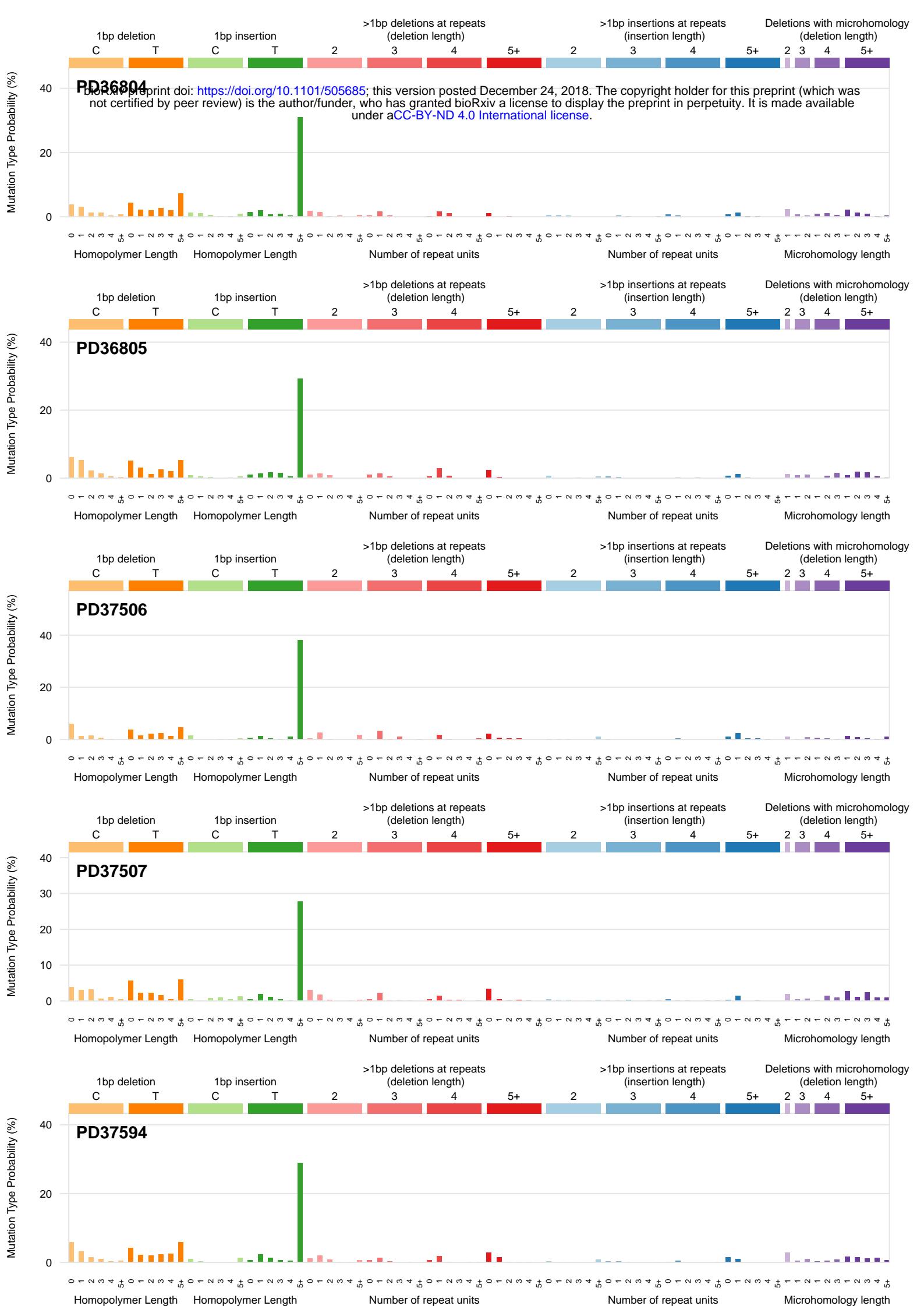
Cos Sim ≥ 0.95 to any
of the 60 PCAWG signatures?

SBS1 SBS5 SBS18

YES

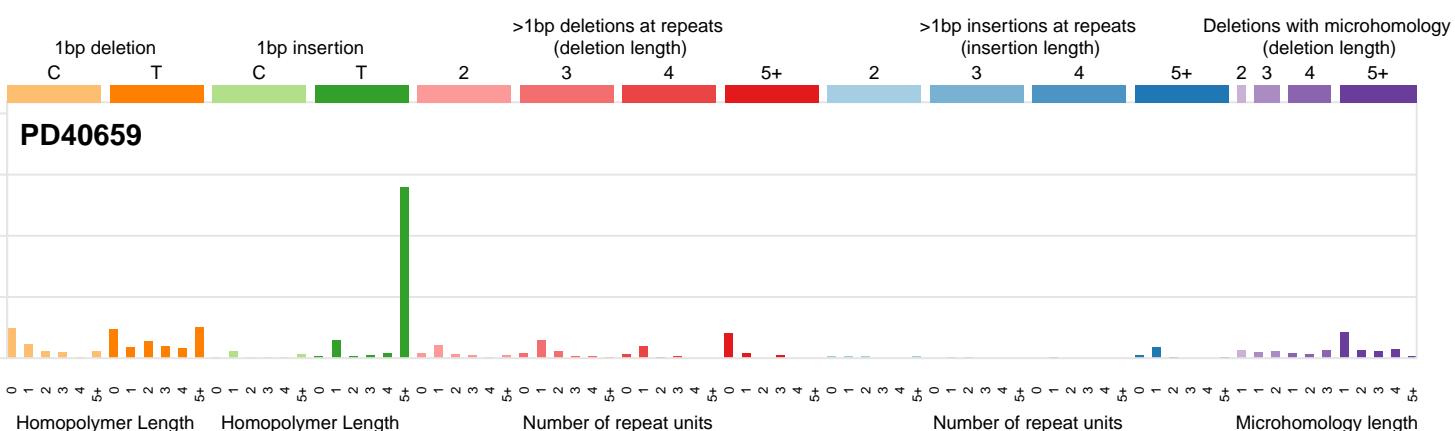
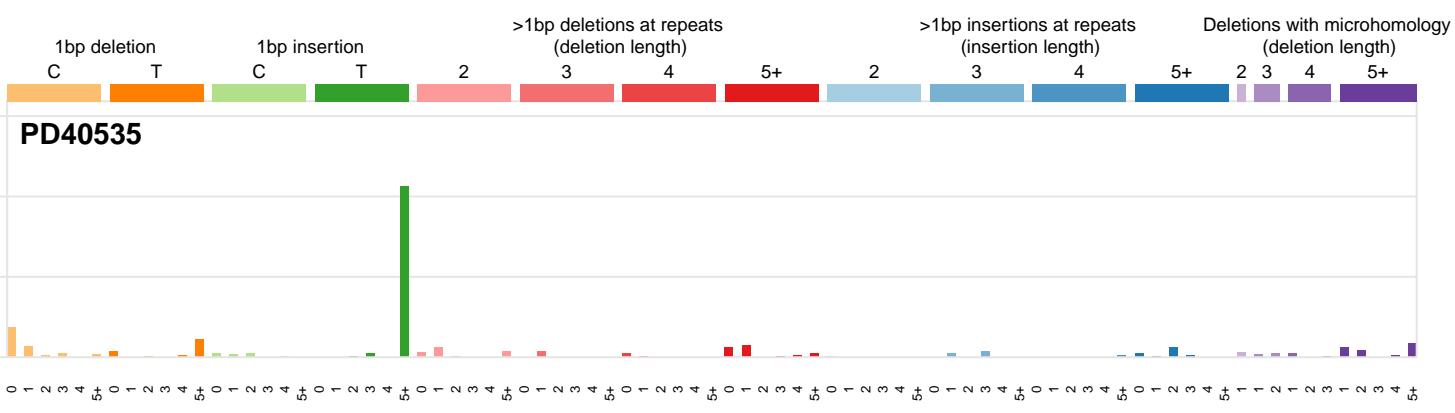
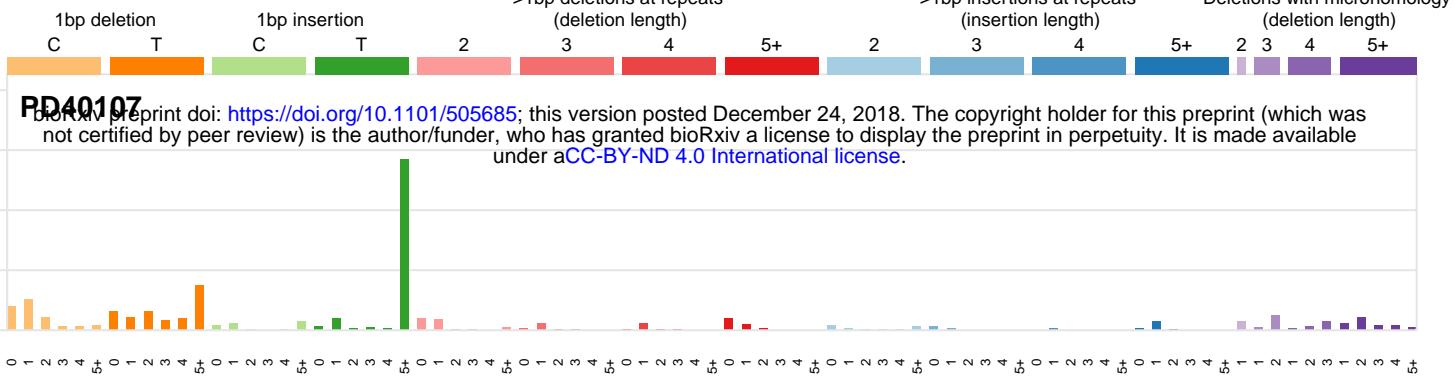
NO

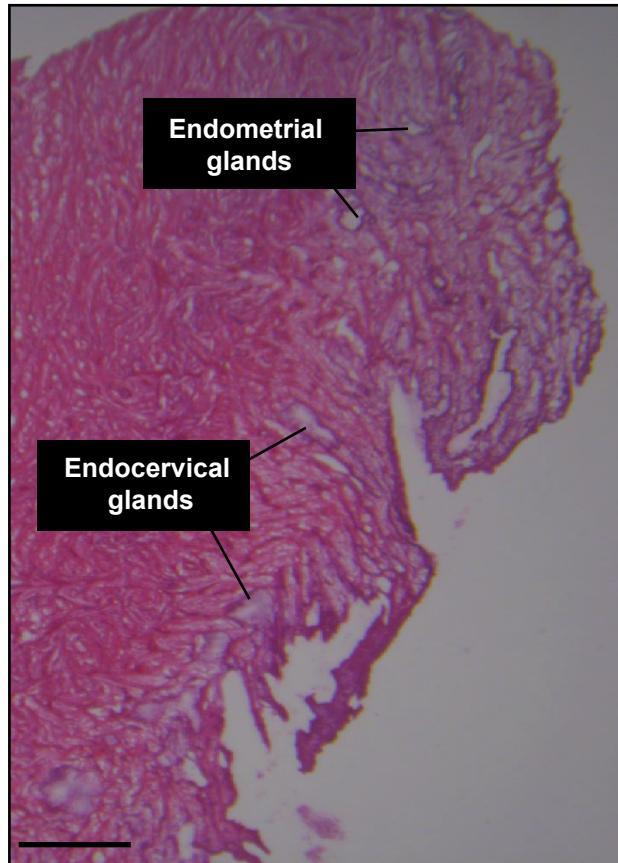
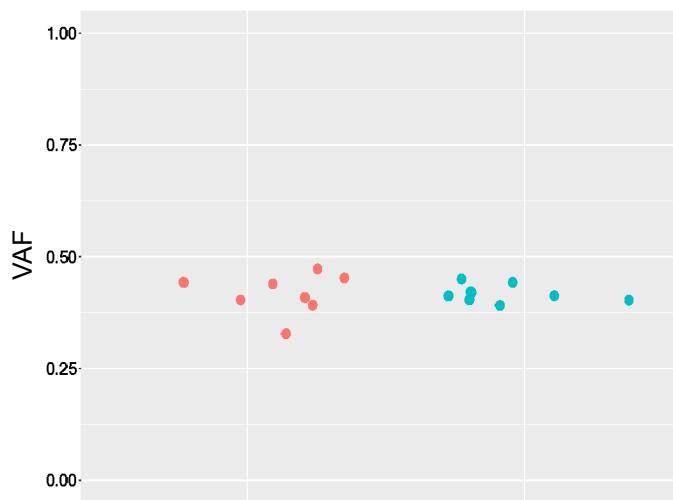
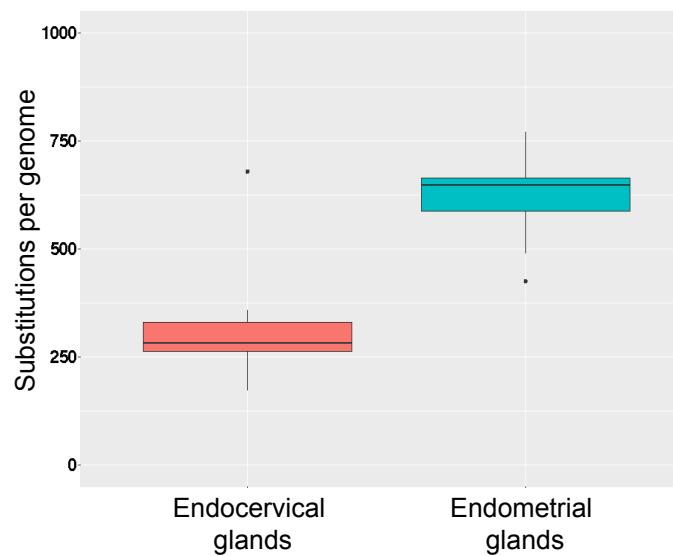
d**f****g****h**

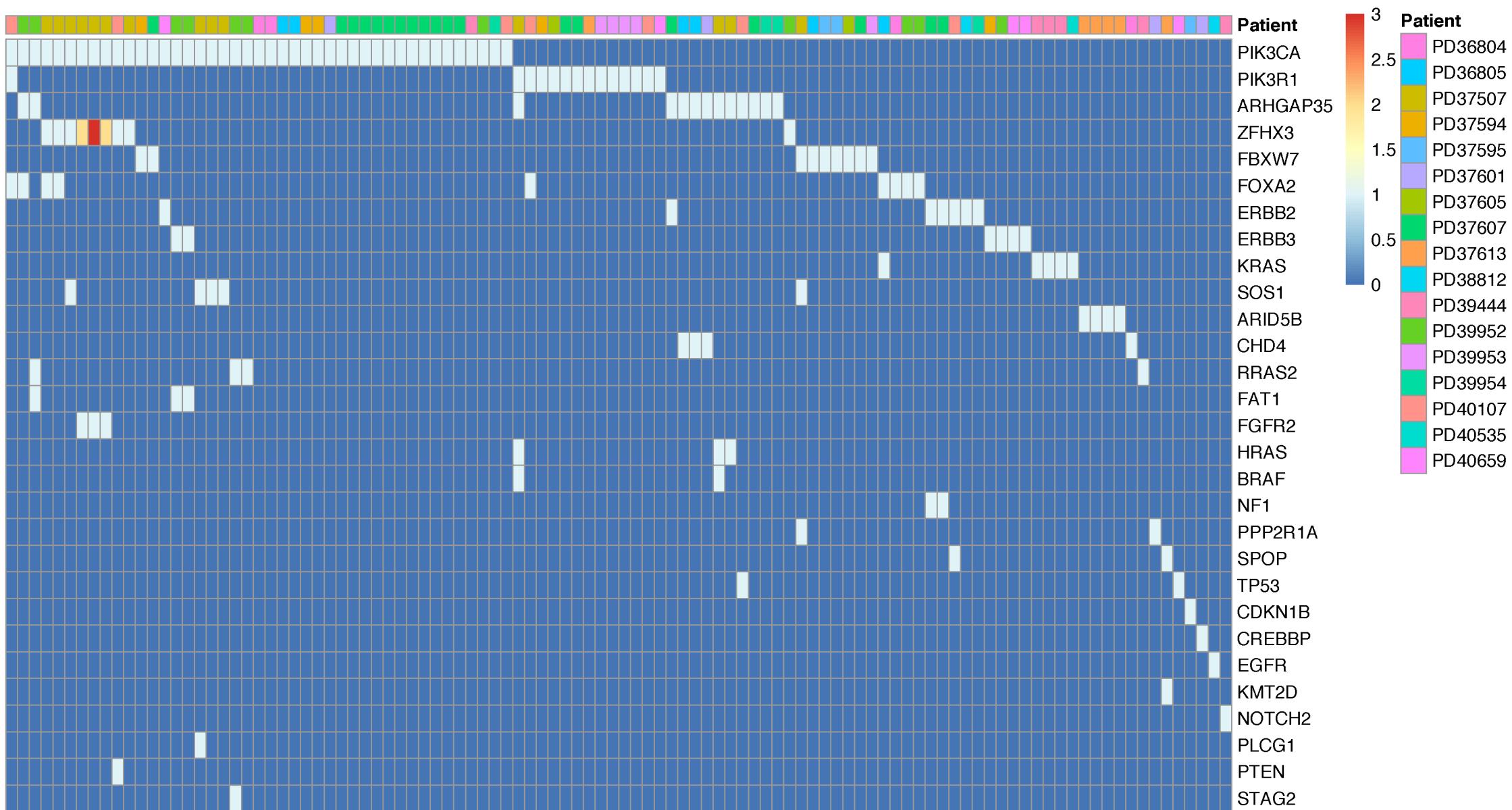








a**b****c**



a

b

