1 **TITLE**
2 Selfish mutations dysregulating RAS-MAPK signaling are pervasive in aged human testes
3
4 **Running title:** Selfish *de novo* mutations in human testes
5
6 **Keywords:**
7 clonal expansion; germline mutation; somatic mutation; gametes; selfish spermatogonial
8 selection; mosaicism; paternal age effect; spermatogonial stem cells.
9
10 **Authors:**
11

12 **Geoffrey J. Maher[1,2¶], Hannah K. Ralph[1,2¶], Zhihao Ding[1,2*¶], Nils Koelling[1,2], Hana**
13 **Mlcochova[1,2], Eleni Giannoulatou[1,2^], Pawan Dhami[3$], Dirk S. Paul[3#], Stefan H. Stricker[3%],**
14 **Stephan Beck[3], Gilean McVean[4], Andrew OM Wilkie[1,2], Anne Goriely[1,2]**
15 [1]Clinical Genetics Group, MRC-Weatherall Institute of Molecular Medicine, University of
16 Oxford, Oxford OX3 9DS, UK; [2]Nuffield Division of Clinical Laboratory Sciences, Radcliffe
17 Department of Medicine, University of Oxford, Oxford OX3 9DS, UK; [3]Medical Genomics,
18 UCL Cancer Institute, University College London, London WC1E 6BT, UK; [4]Big Data Institute,
19 Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK.
20 ¶equal contribution
21
22 Current addresses:
23 * Genomics plc, King Charles House, Park End Street, Oxford OX1 1JD, UK. Zhihao Ding is an
24 employee of Genomics plc. His involvement in the conduct of this research was solely in his
25 former capacity as a Statistical Geneticist at the University of Oxford.
26 ^ Victor Chang Cardiac Research Institute, University of New South Wales, Sydney, Australia
27 $ Genomics and Genome Engineering core facility, Research department of Oncology, UCL
28 Cancer Institute, University College London, London WC1E 6BT, UK
29 # Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care,
30 University of Cambridge, Strangeways Research Laboratory, Cambridge CB1 8RN, UK
31 % MCN Junior Research Group, Munich Center for Neurosciences, Ludwig-Maximilian-
32 Universität, BioMedical Center, Grosshaderner Strasse 9, Planegg-Martinsried, 82152,
33 Germany
34
35 correspondence: anne.goriely@imm.ox.ac.uk
36

37   **ABSTRACT**

38   Mosaic mutations present in the germline have important implications for reproductive risk

39   and disease transmission. We previously demonstrated a phenomenon occurring in the male

40   germline, whereby specific mutations arising spontaneously in stem cells (spermatogonia)

41   lead to clonal expansion, resulting in elevated mutation levels in sperm over time. This

42   process, termed *selfish spermatogonial selection,* explains the high spontaneous birth

43   prevalence and strong paternal age-effect of disorders such as achondroplasia, Apert, Noonan

44   and Costello syndromes, with direct experimental evidence currently available for specific

45   positions of six genes (*FGFR2*, *FGFR3*, *RET*, *PTPN11*, *HRAS* and *KRAS*). We present a discovery

46   screen to identify novel mutations and genes showing evidence of positive selection in the

47   male germline, by performing massively parallel simplex PCR using RainDance technology to

48   interrogate mutational hotspots in 67 genes (51.5 kb in total) in 276 biopsies of testes from 5

49   men (median age: 83 years). Following ultra-deep sequencing (~16,000x), development of a

50   low-frequency variant prioritization strategy and targeted validation, we identified 61 distinct

51   variants present at frequencies as low as 0.06%, including 54 variants not previously directly

52   associated with selfish selection. The majority (80%) of variants identified have previously

53   been implicated in developmental disorders and/or oncogenesis and include mutations in six

54   newly associated genes (*BRAF, CBL, MAP2K1*, *MAP2K2*, *RAF1* and *SOS1*), all of which encode

55   components of RAS-MAPK pathway and activate signaling. Our findings extend the link

56   between mutations dysregulating the RAS-MAPK pathway and selfish selection, and show

57   that the ageing male germline is a repository for such deleterious mutations.

58 **INTRODUCTION**

59 The timing, location and functional effects of spontaneous mutations determine the

60 distribution and phenotypes of mutant cells within the body: this can have a variety of impacts

61 on the health of an individual, and potentially, their offspring. Spontaneous mutations

62 occurring during early post-zygotic development lead to widespread tissue mosaicism that,

63 depending on context, may be phenotypically undetectable or cause so-called 'somatic'

64 disorders (Campbell et al. 2015). Such early post-zygotic mosaicism occurs commonly, with

65 up to 22% of apparently *de novo* point mutations (DNMs) detectable in a child's blood sample

66 likely to have occurred after fertilization (Acuna-Hidalgo et al. 2015; Krupp et al. 2017). A

67 corollary is that a further ~4-10% of DNMs and ~4% of copy-number variants (CNVs) present

68 in a child can be detected at low-level in one of the parent's somatic (usually blood or saliva)

69 samples, and are therefore in fact inherited; as these would have occurred early during

70 parental post-zygotic development (before the separation of the somatic and gonadal

71 lineages), they are associated with a significant risk of recurrence (Campbell et al. 2014;

72 Acuna-Hidalgo et al. 2015; Rahbari et al. 2016; Krupp et al. 2017). By contrast, spontaneous

73 mutations occurring postnatally contribute to tissue-specific, low-level mosaicism, formation

74 of benign tumors, or cancer, depending on the functional consequence(s) of the acquired

75 mutation(s), the clonal dynamics of the tissue involved and the state of the niche (Klein et al.

76 2010a; Vermeulen et al. 2013; Holstege et al. 2014; Swanton 2015). This latter phenomenon

77 has been documented in apparently healthy somatic tissues that display stem cell

78 replacement (e.g. skin, colon, small intestine and blood), where low levels (~1-10%) of clonal

79 mutations are prevalent and their incidence and frequency increase with age (Hafner et al.

80 2010; Laurie et al. 2012; Genovese et al. 2014; Jaiswal et al. 2014; Martincorena et al. 2015;

81   McKerrell et al. 2015; Acuna-Hidalgo et al. 2017; Coombs et al. 2017; Martincorena et al.

82   2017; Zink et al. 2017).

83

84   Analogous to the postnatal occurrence of somatic mutations, we previously demonstrated a

85   similar phenomenon, termed selfish spermatogonial selection, that occurs in the testes of

86   adult men as they age. However, because the testis contains germ cells that, upon

87   fertilization, will carry the genetic information across generations, this process has important

88   reproductive implications, being associated with an increased prevalence of pathogenic

89   DNMs in the next-generation. Despite the relatively low average human germline point

90   mutation rate of ~1.2 x 10^{-8} per nucleotide per generation (Kong et al. 2012; Goldmann et al.

91   2016; Jonsson et al. 2017), specific 'selfish' DNMs in *FGFR2*, *FGFR3*, *HRAS*, *PTPN11* and *RET*

92   are observed up to 1000-fold more frequently in offspring (Goriely and Wilkie 2012). These

93   pathogenic mutations, which cause developmental disorders that show an extreme paternal

94   bias in origin and an epidemiological paternal age effect (collectively referred to as PAE

95   disorders; for example achondroplasia, Apert, Costello and Noonan syndromes, multiple

96   endocrine neoplasia type 2a/b), are identical (or allelic) to oncogenic driver mutations in

97   tumors (Goriely and Wilkie 2012). We proposed that although the mutational events arise at

98   low background rates in male germ cells, selfish mutations confer a selective advantage to

99   spermatogonia leading to their clonal expansion, which results in increased apparent

100   mutation levels in sperm over time (Goriely and Wilkie 2012; Maher et al. 2014).

101

102   Three methods have previously been used to detect selfish mutations in the male germline,

103   each of which has been limited in their ability to evaluate the process at scale: (1)

104   quantification in sperm, (2) quantification in testis biopsies and (3) direct identification in

105   seminiferous tubules. Detecting selfish mutations in sperm, in which individual mutations are

106   present at levels ranging from 10^{-3} to $<10^{-6}$, requires ultra-sensitive techniques that have

107   limited quantitative analysis to small regions of 1-6 nucleotides across five locations in *FGFR2*

108   (x2) (Goriely et al. 2003; Goriely et al. 2005; Yoon et al. 2009), *FGFR3* (x2) (Tiemann-Boege et

109   al. 2002; Goriely et al. 2009) and *HRAS* (Giannoulatou et al. 2013) (Supplementary Table 1).

110   To circumvent the technical challenges caused by mutational dilution within an entire

111   ejaculate, mutations may alternatively be identified following systematic dissection and

112   sequencing of DNA extracted from discrete testicular biopsies. The germ cells (from diploid

113   spermatogonia to haploid spermatozoa) are located in long (up to ~80 cm) highly convoluted

114   and tightly packed seminiferous tubules, comprising ~300-500 per testis (Glass 2005). As

115   clonally expanding mutant spermatogonia are physically restricted to the tubules in which

116   they arise, their geographical distribution within the testis is confined to specific regions: the

117   existence of such localized foci has been demonstrated for selfish mutations in four genes

118   (*FGFR2*, *FGFR3*, *PTPN11*, *RET*) (Qin et al. 2007; Choi et al. 2008; Dakouane Giudicelli et al.

119   2008; Choi et al. 2012; Shinde et al. 2013; Yoon et al. 2013; Eboreime et al. 2016). Finally,

120   mutant clones have been directly visualized in sections of formalin-fixed paraffin embedded

121   (FFPE) normal human testes using immunohistochemical approaches to reveal abnormal

122   expression of spermatogonial antigens (Lim et al. 2012; Maher et al. 2016a). Microdissection

123   of tubules exhibiting enhanced antigen staining and subsequent whole genome amplification

124   facilitated screening of over 100 genes, identifying 9 new selfish mutations, including one in

125   a novel gene (*KRAS*) (Supplementary Table 1). However this approach is limited both by the

126   need to source fixed testis samples with good tissue morphology and DNA preservation, and

127   by the high threshold required for successful immunohistochemical detection (Maher et al.

128   2016a; Maher et al. 2016b).

129

130   Owing to the limitations outlined above, experimental evidence of clonal expansion has so far

131   been restricted to activating mutations at 16 codons in only six genes (Supplementary Table

132   1), all encoding members of the receptor tyrosine kinase (RTK)-RAS-MAPK signaling pathway.

133   Here, we hypothesized that other variants dysregulating the RAS-MAPK pathway, and/or

134   other pathways controlling spermatogonial stem cell homeostasis, may be under positive

135   selection in the male germline (Goriely and Wilkie 2012; Goriely et al. 2013). To reduce the

136   required assay sensitivity compared with bulk semen analysis, and hence substantially widen

137   the extent of the genomic target that could feasibly be analyzed in a single experiment, we

138   exploited approach (2) above. By combining systematic dissection of 276 testicular biopsies

139   from 5 individuals with massively parallel simplex PCR and ultra-deep sequencing (~16,000x)

140   of mutational hotspots in 67 genes, we present the most comprehensive survey of mutations

141   clonally enriched in the human testis to date. We describe the identification of 61 distinct

142   variants across 15 genes with variant allele frequencies (VAF) as low as 0.06%, including 51

143   mutations and 6 novel genes with strong support for association with the process of selfish

144   spermatogonial selection.

145

146   **RESULTS**

147   To perform a discovery screen and identify novel mutations and genes under selection in the

148   male germline, we systematically biopsied human testes following the experimental design

149   summarized in Supplementary Figure 1. A total of 276 small biopsies (~60–180 mm^3) from 5

150   men (age range 34-90 years, median 83 years) were screened by ultra-deep Illumina

151   sequencing (~16,000x post-filtering) of a panel of candidate loci (corresponding to 66.5 kb

152   genomic sequence across 500 amplicons, covering mutational hotspots in 71 genes; see

153   Methods for criteria used to include loci in screen), amplified using massively parallel simplex

154   PCR (RainDance Thunderstorm). To detect low level mosaicism (~0.1-3.0%), the background

155   at each genomic location was independently estimated for all 431 (of 500) amplicons (in 67

156   of 71 genes) that passed quality control (Supplementary Table 2). After normalization, a

157   statistical model was applied to call outlier non-consensus variants at each genomic position

158   (within each amplicon): a minimum threshold of 10 variant reads and median coverage of >

159   5,000x was implemented to reduce false positive calls. As a conservative prioritization

160   strategy, only variants with two or more independent calls were further studied, resulting in

161   a set of 374 variant calls located at 361 genomic locations (see Methods). Visualization and

162   manual curation of each of these calls identified 115 higher confidence candidate variants,

163   distributed at 105 genomic positions across 165 biopsies (Supplementary Figure 1 and

164   Supplementary Table 3).

165

166   As calling variants at low levels (<1%) is subject to PCR artefacts and sequencing errors

167   (Minoche et al. 2011; Hestand et al. 2016; Salk et al. 2018), we developed a tiered strategy

168   for further variant prioritization. We reasoned that variants called independently in

169   overlapping amplicons or in sample replicates (12 biopsies were amplified and sequenced in

170   duplicate) were least likely to be artefactual (Tier 1 variants, Table 1). 18 of the 40 Tier 1

171   variants (with VAF ranging from 0.10% to 2.63%) were re-screened by PCR or using single

172   molecule molecular inversion probes (smMIPs) and ultra-deep MiSeq sequencing (~30,000x).

173   Seventeen of the 18 (94%) variants were validated, suggesting the great majority of Tier 1

174   variants are true positive calls (Table 1, Supplementary Table 3). Amongst the Tier 1 variants

175   are five mutations previously associated experimentally with selfish selection: *FGFR2*

176   c.755C>G (p.Ser252Trp – Apert syndrome), c.758C>G (p.Pro253Arg – Apert syndrome) and

177   c.870G>T (p.Trp290Cys – Pfeiffer syndrome), *KRAS* c.182A>G (p.Gln61Arg – oncogenic) and

178   *PTPN11* c.215C>T (p.Ala72Val – oncogenic) (Table 1). This strong enrichment for canonical

179   examples of selfish mutations (Supplementary Table 1) provided initial validation of our

180   experimental approach and starting hypothesis.

Table 1

Tier	Variant number	Gene	Variant position (hg19) and predicted amino acid substitution$	VAF range (%)	Testis	Number of positive pieces	gnomAD exome frequency	COSMIC v82	Germline disorder
1	1	AKT3	chr1:243668575 G>C (p.Ser472Ser)	0.09	4	1	0.0008568	0 (1, 1)	-
1	2	APC	chr5:112175762 T>G (p.Phe1491Val)	0.47	4	1	0	0 (0, 1)	-
1	3	BRAF	chr7:140482928 G>C (p.Pro403Ala)	0.12	4	1	0.00001219	0 (0, 1)	-
1	4	BRAF	chr7:140481402 C>G (p.Gly469Ala)	0.16 - 0.32	1;4	1;3	4.062E-06	52 (52, 123)	-
1	5	BRAF	chr7:140449196 T>G (p.Gln628Pro)	0.19	1	1	0	-	-
1	6	BRAF	chr7:140439614 G>T (p.Gln709Lys)	0.26	4	1	0	-	-
1	7	FGFR2	chr10:123279677 G>C (p.Ser252Trp)*	0.14 - 1.55	1;2;4	2;7;14	4.086E-06	54 (54, 57)	Apert syndrome
1	8	FGFR2	chr10:123279674 G>C (p.Pro253Arg)*	0.06 - 0.56	1;4	4;4	0	9 (9, 11)	Apert syndrome
1	9	FGFR2	chr10:123279605 A>C (p.Phe276Cys)	0.32	4	1	0	1 (1, 1)	-
1	10	FGFR2	chr10:123279566 T>G (p.Gln289Pro)	0.12	4	1	0	-	Crouzon syndrome
1	11	FGFR2	chr10:123279562 C>A (p.Trp290Cys)*	0.18 - 0.69	1;4	1;2	0	0 (7, 7)	Pfeiffer syndrome
1	12	FGFR2	chr10:123279562 C>G (p.Trp290Cys)#	0.12 - 0.34	1;4	2;1	0	7 (7, 7)	Pfeiffer syndrome
1	13	FGFR2	chr10:123274803 G>C (p.Ser372Cys)	0.11 - 0.33	4	5	0	1 (1, 2)	Beare Stevenson
1	14	FGFR3	chr4:1805519 C>T (p.Ser344Phe)	0.21	4	1	4.065E-06	1 (1, 2)	-
1	15	FGFR3	chr4:1807371 C>A (p.Asn540Lys)	0.07 - 0.18	2;4	1;1	0	-	Hypochondroplasia
1	16	FGFR3	chr4:1807371 C>G (p.Asn540Lys)	0.07	2	1	0	-	Hypochondroplasia
1	17	FGFR3	chr4:1807488 G>A (p.Val553Met)	0.37	4	1	0.0001	0 (0,1)	-
1	18	KRAS	chr12:25398284 C>G (p.Gly12Ala)	0.12 - 0.37	1;2;4	1;1;1	0	2255 (2256, 33497)	-
1	19	KRAS	chr12:25398284 C>T (p.Gly12Asp)	0.12 - 1.82	4	6	4.094E-06	14126 (14128, 33497)	-
1	20	KRAS	chr12:25380282 G>C (p.Ala59Gly)	0.29	4	1	0	8 (8, 41)	-
1	21	KRAS	chr12:25380282 G>T (p.Ala59Glu)	0.14 - 0.50	4	2	0	6 (6, 41)	-
1	22	KRAS	chr12:25380276 T>C (p.Gln61Arg)*	0.62 - 2.63	4	9	0	115 (116, 601)	-
1	23	MAP2K1	chr15:66727455 G>C (p.Lys57Asn)	0.14	4	1	0	1 (14, 19)	-
1	24	PTPN11	chr12:112888197 T>G (p.Phe71Leu)	0.48	4	1	0	1 (22, 22)	Noonan syndrome
1	25	PTPN11	chr12:112888198 G>C (p.Ala72Pro)	0.28	2	1	0	0 (0, 137)	Noonan syndrome
1	26	PTPN11	chr12:112888199 C>A (p.Ala72Asp)	0.13 - 0.25	2	3	0	11 (11, 137)	-
1	27	PTPN11	chr12:112888199 C>G (p.Ala72Gly)	0.11	1	1	0	2 (2, 137)	Noonan syndrome
1	28	PTPN11	chr12:112888199 C>T (p.Ala72Val)*	0.88	2	1	0	72 (73, 137)	-
1	29	PTPN11	chr12:112888202 C>T (p.Thr73Ile)	0.39 - 0.59	2	2	0	19 (19, 19)	Noonan syndrome
1	30	PTPN11	chr12:112888210-112888211 GA>CT (p.Glu76Leu)	0.19	1	1	0	0 (0, 203)	-
1	31	PTPN11	chr12:112888211 A>C (p.Glu76Ala)	0.45 - 0.69	1;2	1;1	0	20 (20, 203)	-
1	32	PTPN11	chr12:112888211 A>T (p.Glu76Val)	0.24 - 0.93	2;4	1;1	0	11 (11, 203)	-
1	33	PTPN11	chr12:112924336 G>A (p.Val428Met)	0.54	4	1	4.063E-06	3 (3, 3)	-
1	34	PTPN11	chr12:112924336 G>T (p.Val428Leu)	0.52	4	1	0	0 (0, 3)	-
1	35	PTPN11	chr12:112926908 C>A (p.Gln510Lys)	0.13 - 0.30	4	2	0	3 (3, 21)	-
1	36	RET	chr10:43613906 G>C (p.Leu790Phe)	0.15	4	1	4.063E-06	0 (0, 1)	MEN2A
1	37	RET	chr10:43613906 G>T (p.Leu790Phe)	0.10 - 0.42	4	3	0.00002032	0 (0, 1)	MEN2A
1	38	RET	chr10:43615613 G>T (p.Asp898Tyr)	0.15	2	1	4.072E-06	0 (0, 1)	MEN2
1	39	SOS1	chr2:39250292 T>G (p.Gln426Pro)	0.37	4	1	0	0 (0, 1)	-
2	40	BRAF	chr7:140453155 C>G (p.Asp594His)	0.06 - 0.23	2	2	0	3 (3, 126)	-
2	41	BRAF	chr7:140453132 T>G (p.Lys601Asn)	0.22 - 0.42	4	2	0	7 (18, 129)	-
2	42	CBL	chr11:119148991 G>A (p.Cys404Tyr)	0.52 - 0.63	5	2	8.126E-06	15 (15, 19)	-
2	43	FGFR2	chr10:123276865 G>C (p.Ser351Cys)	0.09 - 0.26	4	4	0	0 (0, 1)	Pfeiffer syndrome
2	44	FGFR2	chr10:123276893 A>T (p.Cys342Ser)^	0.26 - 2.95	1	7	0	-	Crouzon syndrome
2	45	FGFR2	chr10:123258034 A>C (p.Asn549Lys)	0.14 - 0.34	4	2	0	10 (34, 44)	-
2	46	FGFR3	chr4:1808029 C>G (p.Arg669Gly)	0.14 - 0.24	1	2	4.105E-06	0 (0, 1)	-
2	47	LRP5	chr11:68115514 C>T (p.Ala97Ala)	0.53 - 1.20	4	4	0.0004877	-	-
2	48	MAP2K2	chr19:4110584 A>T (p.Cys125Ser)	0.14 - 0.21	2	2	0	1 (3, 5)	-
2	49	NF1	chr17:29554264 G>A (p.Met760Ile)	0.87 - 2.07	4	9	0	-	-
2	50	PTPN11	chr12:112888166 A>C (p.Asp61Ala)	0.26 - 1.02	2	2	0	1 (1, 121)	Noonan syndrome
2	51	PTPN11	chr12:112888166 A>G (p.Asp61Gly)	0.71 - 0.73	4	2	0	6 (6, 121)	Noonan syndrome
2	52	PTPN11	chr12:112891083 G>T (p.Glu139Asp)	0.15 - 0.56	4	2	0	1 (5, 6)	Noonan syndrome
2	53	PTPN11	chr12:112915455 T>G (p.Phe285Cys)	0.09 - 0.24	2	2	0	-	Noonan syndrome
2	54	PTPN11	chr12:112915523 A>G (Asn308Asp)*	0.68 - 0.69	4	2	1.219e-5	4 (4, 7)	Noonan syndrome
2	55	PTPN11	chr12:112926884-112926885TC>AA (p.Ser502Lys)	0.25	4	1	0	0 (0, 44)	-
2	56	PTPN11	chr12:112926890 A>G (p.Met504Val)	0.67 - 0.82	4	2	4.061E-06	1 (1, 1)	Noonan syndrome
2	57	RAF1	chr3:12645699 G>A (p.Ser257Leu)	0.44 - 0.56	4	2	0	14 (14, 15)	Noonan syndrome
3	58	BRAF	chr7:140453096 C>G (p.Leu613Phe)	0.10	4	1	0	-	-
3	59	MAP2K1	chr15:66727451 A>C (p.Gln56Pro)	0.07 - 0.11	4	3	0	8 (8, 8)	-
3	60	PTPN11	chr12:112888162 G>C (p.Gly60Arg)	0.09 - 0.17	1	2	0	6 (6, 55)	-
3	61	RAF1	chr3:12645688 G>C (p.Pro261Ala)	0.15	4	1	0	0 (0, 8)	Noonan syndrome

Table 1. List of 61 validated variants identified in this study.

$ for amino acid numbering see transcript accession code details. * Variant previously associated with selfish selection. # Distinct DNA substitution but same amino acid substitution as previously described. ^ Same clone as previously described (Lim et al. 2012; Maher et al. 2016a). For variants covered by more than one amplicon or called in sample replicates, variant allele frequencies (VAF) are the averages of calls per piece. Variant frequencies in gnomAD exome dataset were accessed September 2017. Counts from the COSMIC database (v82) refer to identical DNA substitutions, identical amino acid substitutions and total substitutions at the specific amino acid, respectively.

191 Within the panel, the majority (88.7%) of callable (i.e. excluding primer sequences and

192 amplicons with low QC) regions were represented by a single amplicon and only 12 biopsies

193 were sequenced in duplicate (Supplementary Table 4): hence, we next investigated variants

194 that were called in single amplicons in two or more biopsies, at VAF of ≥0.2% in at least one

195 biopsy (Tier 2). Twenty-six Tier 2 variants were identified, 18 (69%) of which were validated

196 upon resequencing (Table 1, Supplementary Table 3). Notably, all (14/14) of the known

197 pathogenic variants were validated, but only four of the twelve variants without prior disease

198 association were true positives. In biopsy 4D25, *PTPN11* c.1504T>A (p.Ser502Thr - Noonan

199 syndrome) was called as a single nucleotide variant but on validation it was identified as a

200 double nucleotide substitution c.1504_1505delTCinsAA (p.Ser502Lys). Next, 29 variants with

201 a VAF of 0.1 - <0.2% called in a single amplicon in two or more biopsies (Tier 3) were identified.

202 Only 4 of the 22 (18%) resequenced Tier 3 variants were validated, suggesting that in this

203 lower frequency range, the majority of calls are artefactual (Table 1, Supplementary Table 3).

204 Owing to the low validation rate of variants with VAFs of 0.1 - <0.2%, none of the remaining

205 20 calls that exhibited VAF <0.1% (Tier 4) variants were re-screened for validation

206 (Supplementary Table 3).

207

208 Overall we identified 61 distinct variants that we classified as independently validated,

209 present in 15 of the 67 genes that passed quality control and were analyzed in the experiment.

210 Based on the identification of the same variant in testes sourced from different men, we

211 conclude that at least 72 independent mutational events (clones) could be distinguished

212 across the five testes (Table 1, Figure 1, Supplementary Figs 2-3). Two variants (*FGFR2*

213 c.755C>G (p.Ser252Trp) (#7) and *KRAS* c.35G>C (p.Gly12Ala) (#19)) occurred in three testes

214 and seven in two testes (Figure 1; Supplementary Fig 2). Strikingly, these variants are all either

215   recurrent mutations causative of congenital skeletal disorders, or known hotspots in cancer

216   (COSMIC) that may be associated with lethal or as yet undescribed congenital disorders (Table

217   1). Figure 2 details all validated variants for the two genes most highly represented in this list,

218   *FGFR2* and *PTPN11* (15 independent mutational events responsible for 10 distinct variants in

219   *FGFR2* (encoding nine pathogenic protein changes); and 22 independent mutational events

220   of 20 distinct variants in *PTPN11*). Their relative locations on the respective protein products

221   shows striking overlap with mutational hotspots previously associated with developmental

222   disorders and cancer. The corollary is that our observations of these mutations in testes are

223   likely to be relevant to the biological origins of the cognate diseases. Similar plots for 13 other

224   genes with validated variants are presented in Supplementary Figure 3.

225

226

227   **Figure 1. Distribution of validated variants in testis slices 1D, 2F, 4B, and 5J.** Testicular biopsy numbers are

228   located to the left of each testis slice. Some biopsies were further dissected into two pieces of which the

229   orientation is unknown – these are indicated with a diagonal dashed line (e.g. Tes2F 30a,b). Each variant has a

230   distinct number (as listed in Table 1) and is colored according to gene: *FGFR2* (purple), *FGFR3* (orange), *KRAS*

231   (black), *PTPN11* (blue), *RET* (pink), newly associated gene (red). The size of each circle is proportional to the

232   observed variant allele frequency (VAF) in each biopsy as indicated by black dots on the figure key. Identical

233   variants in different biopsies have been connected by lines that likely track the seminiferous trajectory across

234   the testis and therefore may represent a single 'clonal event'; note that the path of the clone has been arbitrarily

235   drawn and may not represent the true geography of a seminiferous tubule across the testis. Dark gray segments

236   represent biopsies that were not sequenced due to insufficient material quality/quantity (see Methods). Light

237   gray segments represent non-tubular regions of tissue. The age of the individual from whom the sample was

238   collected is indicated on the figure (See Supplementary Table 5 for further details on the testicular samples). The

239   remaining five slices of Tes4 are presented in Supplementary Figure 2. Tes3D is omitted as no variants were
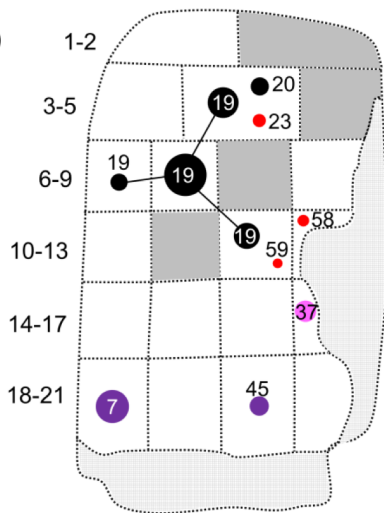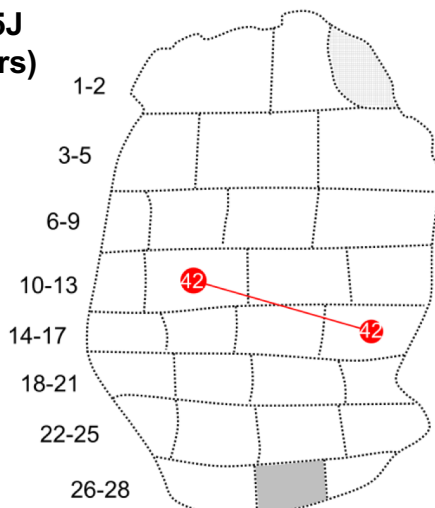
240   identified

241

Figure 1

242

243

12

244 **Figure 2. Spontaneous mutations in *FGFR2* (A) and *PTPN11* (encoding SHP2) (B) identified in testicular biopsies**

245 (A) (I) Ten validated variants positioned along the amino acid sequence of FGFR2 (*x*-axis, see panel V), ranging in

246 VAF from 0.06% to 2.95% (*y*-axis), identified in Tes1D, Tes2F and Tes4. Numbers correspond to those in Table 1;

247 two different variants (c.870G>C or T) predicted to cause the same p.Trp290Cys substitution (#11, #12) were

248 identified. (II) Relative location and length of amplicons used to sequence main hotspots of *FGFR2* are plotted

249 on the *x*-axis. Median coverage per amplicon is plotted on the *y*-axis. All amplicons had median coverage above

250 the cut-off (red dashed line) of 5,000x. (III) Number of reported constitutional variants encoding amino acid

251 substitutions in FGFR2 associated with developmental disorders (sqrt scale) (updated from (Wilkie 2005)). (IV)

252 Number of reported somatic amino acid substitutions in FGFR2 in cancer (COSMIC v82). (v) Protein domains of

253 FGFR2. Annotations and protein structure are based on transcript ID NM_000141 and Uniprot ID P21802

254 (v2017_01), respectively.

255 (B) (I) Twenty validated variants positioned along the amino acid sequence of SHP2 (*x*-axis, see panel (V), ranging

256 in VAF from 0.09% to 1.02% (*y*-axis), identified in Tes1D, Tes2F and Tes4. (II) Location and size of amplicons used

257 to sequence main hotspots of *PTPN11* are plotted on the *x*-axis. Median coverage per amplicon is plotted on the

258 *y*-axis. All amplicons except one had median coverage above the cut-off of 5,000x. (III) Number of reported

259 constitutional variants encoding amino acid substitutions in SHP2 associated with developmental disorders (sqrt

260 scale). (IV) Number of reported somatic amino acid substitutions in SHP2 in cancer (COSMIC v82). (V) Protein

261 domains of SHP2. Annotations and protein structure are based on transcript ID NM_002834 and Uniprot ID

262 Q06124 (v2017_01), respectively.

13
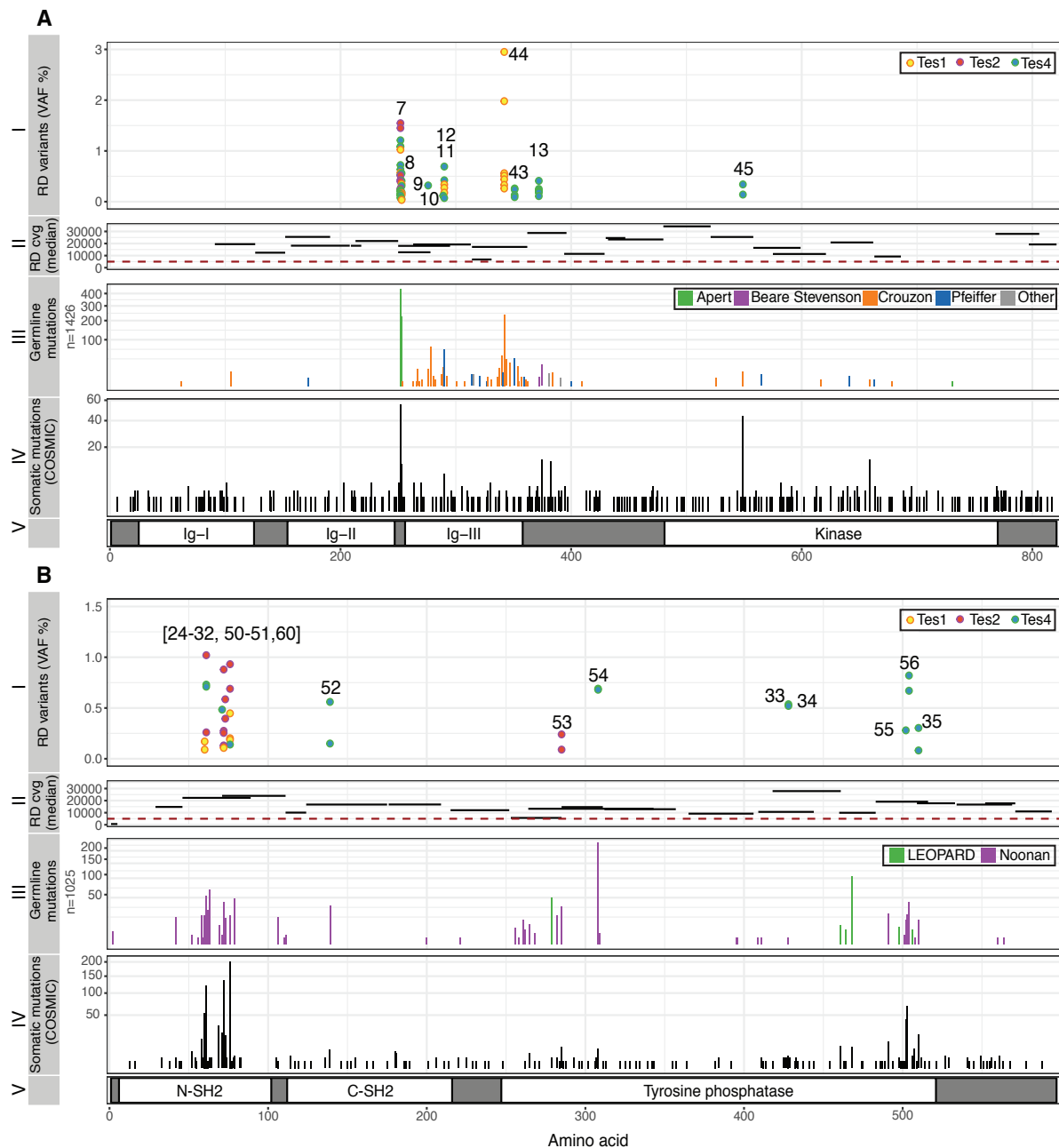
Figure 2

Next, using the geographical register of the multiple biopsies, the spatial distribution of each variant across the testicular biopsies was investigated (Figure 1, Supplementary Figure 2). For example, in 6 of 153 biopsies across three slices from Tes4 we identified a *KRAS* c.35G>A (p.Gly12Asp) mutation (#18). *KRAS* c.35G>A is one of the most frequently reported substitutions in cancer (>14,000 records in COSMIC v82) and post-zygotic *KRAS* c.35G>A mutations have been reported to cause arteriovenous malformations of the brain (Nikolaev

272   et al. 2018) and linear nevus sebaceous syndrome (Wang et al. 2015), but it has never been

273   reported as a constitutional mutation. In slice 4B (slice B of Testis 4) (Figure 1 and

274   Supplementary Figure 3), this *KRAS* mutation was detected at VAF ranging from 0.26% to

275   1.82% in four adjacent biopsies, suggestive of an expansion of a mutational event tracking

276   along the length of a single seminiferous tubule. The same *KRAS* variant was also detected in

277   two neighboring biopsies from slices 4D and 4E, apparently at a distance from the larger clone

278   in slice 4B (Supplementary Figure 2); this smaller clone may represent a distinct mutational

279   event having occurred in an independent tubule, but the resolving power of the experiment

280   does not exclude the possibility that this is a large clonal event spreading along the length of

281   a single seminiferous tubule (that measure up to ~80 cm in humans).

282

283   Owing to the convoluted packing of the seminiferous tubules, individual testicular biopsies

284   contain segments of multiple individual tubules and in 43 biopsies more than one variant was

285   identified (Figure 1, Supplementary Figure 2 and Supplementary Table 3). Mutations with

286   similar distributions across multiple biopsies may represent clones either within the same

287   tubule, or in distinct intermingled tubules running alongside each other. For example, two

288   distinct mutations, *MAP2K2* c.373T>A (p.Cys125Ser) (oncogenic) and *PTPN11* c.215C>A

289   (p.Ala72Asp) (oncogenic)] are both found in the adjacent biopsies 2F11 and 2F16 (Figure 1),

290   with the latter mutation extending into the neighboring biopsy 2F21. In Tes4, four of the six

291   biopsies positive for the oncogenic *KRAS* c.182A>G (p.Gln61Arg) mutation (4E18, 4E25, 4F27,

292   4G1) were also positive for a synonymous variant in *LRP5* [c.291C>T (p.Ala97Ala); no prior

293   disease association] (Supplementary Figures 2 and 4).

294

295   In contrast to selfish mutations that arise in adult spermatogonia and are therefore restricted

296   to the seminiferous tubules in which they arise, 'classical' post-zygotic mosaic mutations

297   occurring in embryonic primordial germ cells, before the formation of the seminiferous

298   tubules, are expected to have a wider distribution in one or both testes. We found one

299   suggestive example of this, an *NF1* c.2280G>A (p.Met760Ile) variant, which exhibited a

300   pattern of occurrence in Tes4 distinct from all the other identified mutations. The variant was

301   originally called in nine biopsies at relatively high VAF (median 1.1%, range 0.9-2.1%)

302   (Supplementary Figure 2), and inspection of the mutation frequency in each sample

303   (Supplementary Figure 5) showed numerous other biopsies in Tes4 with elevated VAFs,

304   compatible with an earlier post-zygotic mosaic event. Unfortunately, no other tissue was

305   available from this individual to test whether the variant was restricted to a single testis

306   and/or to the germline tissue.

307

308

309   To explore the relationship between mutational events identified using RainDance technology

310   (which inherently involves destruction of the tissue structure of the testis) and the occurrence

311   of mutations in individual seminiferous tubules, we exploited the availability of adjacent FFPE

312   material for two of the testes. In Tes1D, our deep-screening strategy identified a *FGFR2*

313   c.1024T>A (p.Cys342Ser) variant at VAFs ranging from 0.26% to 2.95% in seven contiguous

314   biopsies, suggestive of a clonal event tracking a single seminiferous tubule across the testis

315   (Figures 1 and 2, variant #44). For this testis, we had previously studied the adjacent FFPE

316   tissue block (Tes1-1 described in (Lim et al. 2012; Maher et al. 2016a)) using

317   immunohistochemical staining for markers of selfish clones (enhanced MAGEA4 and pAKT

318   immunostaining), followed by laser capture microdissection and targeted resequencing.

319   Strikingly, we previously identified and validated the identical *FGFR2* variant, strongly

320   suggesting that this large mutant clone is present within a significant portion of a single

321   seminiferous tubule that tracks across adjacent testis slices (Maher et al. 2016a). To seek

322   further examples, we undertook a new analysis of putative mutant clones within Tes2E, a

323   FFPE tissue block adjacent to the Tes2F slice, to identify individual tubular cross-sections

324   exhibiting enhanced MAGEA4 immunostaining; laser capture microdissection of six distinct

325   groups of tubular cross-sections, followed by PCR and Illumina sequencing confirmed the

326   presence of the *FGFR2* c.755C>G (p.Ser252Trp – Apert syndrome) and *PTPN11* c.214G>C

327   (p.Ala72Pro – Noonan syndrome) mutations in distinct enhanced MAGEA4-tubules,

328   consistent with the geographic location of these specific variants identified by deep-

329   sequencing in the adjacent Tes2F slice (Figure 3). For the three other testes, FFPE blocks were

330   not available.

331

---

332   **Figure 3. Visualization of mutant tubules in Testis 2**

333   (A) A 5 μm thin section from Tes2E, a FFPE block of tissue adjacent to the testis slice 2F, immunostained with

334   anti-MAGEA4 antibody to label spermatogonia. Seminiferous tubules with enhanced MAGEA4

335   immunopositivity, suggestive of the presence of mutant clones are labelled with small red pins and boxed. Scale

336   bar = 5 mm. (B) High magnification view of cross-sections with MAGEA4-enhanced immunopositivity in two

337   localized areas are labelled with dotted lassoes representing the laser-microdissected regions. Scale bars = 100

338   μm. (C) Results from targeted resequencing of the microdissected seminiferous tubules labelled by dotted

339   lassoes in (B) viewed in IGV (Integrated Genome Viewer); spontaneous pathogenic *FGFR2* c.755C>G #7 (top) and

340   *PTPN11* c.214G>C #25 (bottom) variants were identified in DNA extracted from microdissected tubule cross-

341   sections, but not in DNA from the whole tissue section. Comparison of the MAGEA4 section (A) with adjacent

342   testis slice 2F from the Raindance screen (D) (the same image as in Figure 1 but showing only the targeted *FGFR2*

343   and *PTPN11* mutations), shows that both variants match to a mutation previously identified in the corresponding

344   position of testis slice 2F.
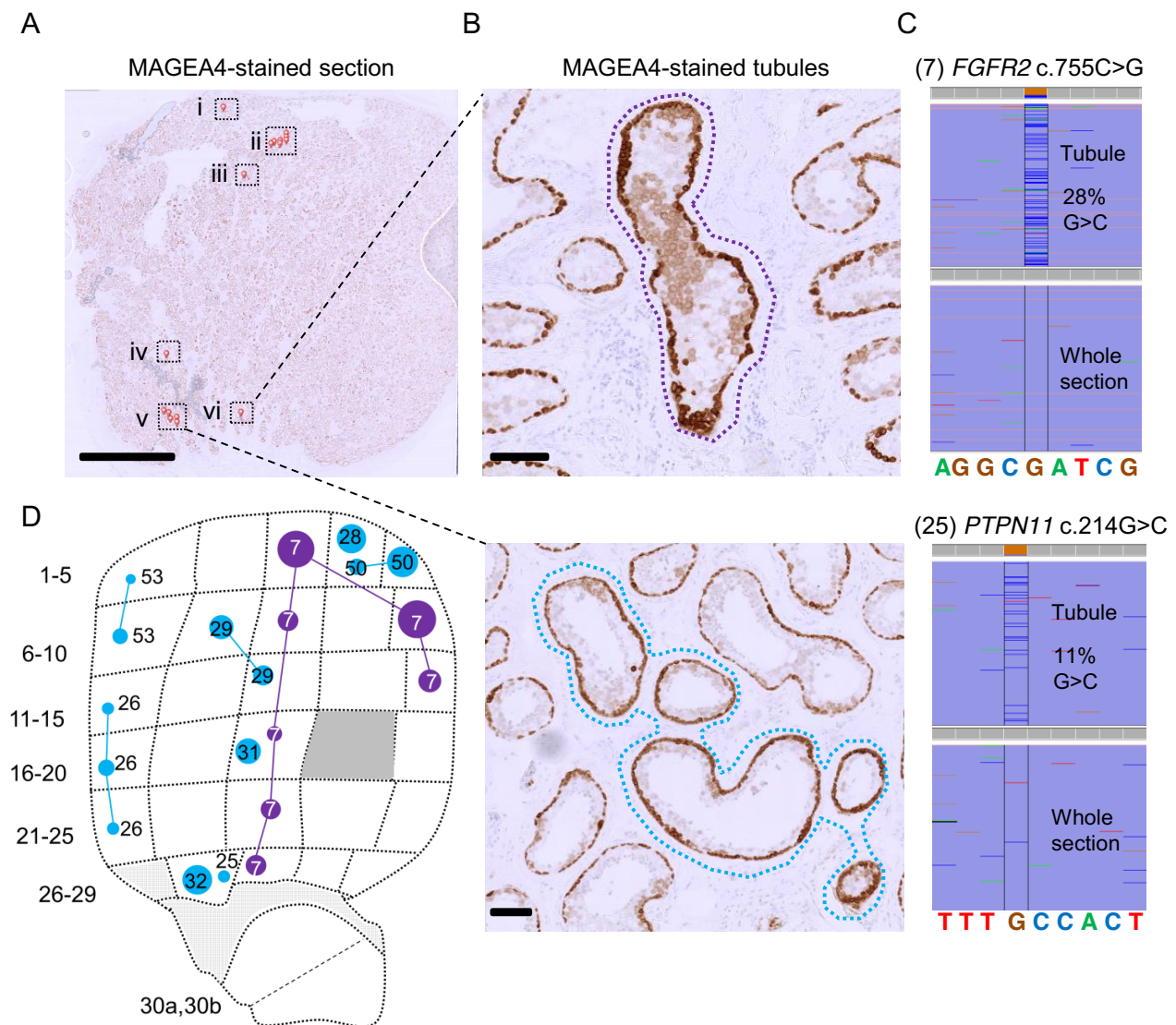
Figure 3

DISCUSSION

We present a new broad-scale approach to studying clonal *de novo* germline mutations directly in human adult testes, the tissue where the majority of DNMs originate. Utilizing massively parallel multiplex PCR and ultra-deep sequencing of 51.5 kb in 276 discrete human testicular biopsies followed by the implementation of a statistical prioritization calling

18

353   strategy, we identified 61 different variants in a total of 111 mutation-positive biopsies, 59 of

354   which encode non-synonymous substitutions (Table 1).

355

356   Several observations support the notion that the mutations identified are strongly enriched

357   for clonal events that are promoted by positive selection of mutant stem cells via the

358   phenomenon of selfish spermatogonial selection. Out of the 61 validated variants (Table 1),

359   43 are located in five (*FGFR2*, *FGFR3*, *KRAS*, *PTPN11*, *RET*) of the six genes associated with

360   strong prior experimental evidence for this process (Supplementary Table 1). As detailed in

361   Table 1 and illustrated in Figure 2 and Supplementary Figure 3, the vast majority of variants

362   identified across these five genes overlap with those observed in dominant congenital

363   disorders and/or cancer, strongly suggestive of a functional role via a gain-of-function

364   mechanism. The most commonly observed individual mutation was *FGFR2* c.755C>G

365   (p.Ser252Trp - Apert syndrome) detected in 23 biopsies. In this and other cases, the

366   identification of identical variants in multiple neighboring testis biopsies (Figure 1 and

367   Supplementary Figure 2) is supportive of clonal expansion along the length of the

368   seminiferous tubules, and in three cases this process could be directly validated at a cellular

369   level by visualizing the selfish expansion characterized by enhanced MAGEA4 staining in the

370   adjacent testis block (Figure 3 and (Maher et al. 2016a)). The largest number of mutations

371   was observed for *PTPN11* (encoding the SHP2 tyrosine phosphatase), in which we identified

372   20 different variants (across 33 biopsies) (Table 1 and Figure 2B). We observed 12 distinct

373   variants located within the N-SH2 domain of SHP2, a region of the protein known to repress

374   the catalytic phosphatase domain in its wild-type state (Neel et al. 2003), including each of

375   the possible nucleotide substitutions at *PTPN11* c.215C encoding three distinct amino acids

376   (p.Ala72Asp, p.Ala72Gly and p.Ala72Val) that have been associated with Noonan syndrome

377   or oncogenesis. This wide mutational spectrum is consistent with epidemiological data that

378   concur that *PTPN11*-associated Noonan syndrome mutations have a high spontaneous birth

379   prevalence (~1/10,000 births) (Goriely and Wilkie 2012). We also identified two dinucleotide

380   substitutions in *PTPN11*: both the c.226_227delGAinsCT (p.Glu76Leu (#30)) and the

381   c.1504_1505delTCinsAA (p.Ser502Lys (#55)) variants encode amino acid substitutions that,

382   owing to the nature of the genetic code, cannot arise from single-nucleotide changes. These

383   observations are reminiscent of other previously described selfish mutations encoded by

384   double and triple substitutions, which in some cases, were shown to result via a 'double-hit'

385   mechanism (Goriely et al. 2005; Goriely and Wilkie 2012; Giannoulatou et al. 2013). In

386   humans, the *de novo* tandem mutation rate is estimated to be ~0.3% of the single nucleotide

387   variant rate (Besenbacher et al. 2016); in this small set of 61 variants, we find a ~10-fold

388   enrichment over the background rate.

389

390

391   Given this strong support for positive clonal selection of pathogenic variants in previously

392   known selfish genes, the next question is whether the other 18 validated variants present in

393   novel candidate genes might also signal the presence of selfish selection. We first excluded

394   from consideration one variant, *NF1* c.2280G>A p.(Met760Ile) (variant #49), which presented

395   with a different pattern of occurrence characterized by an extended geographical distribution

396   across ~1/3 of the testis from individual Tes4, raising the possibility of an early post-zygotic

397   (as opposed to adult-onset) mutational event (Supplementary Figure 5). Although this NF1

398   variant exhibits a high CADD (24.6)/Polyphen score, has been reported in one case of lung

399   cancer (Redig et al. 2016) and is located within the cysteine-serine-rich domain (CSRD), a

400   region where several missense mutations associated with breast cancer and

401 neurofibromatosis have been identified (Koczkowska et al. 2018), its pathogenic status - and

402 potential for positive selection - remain uncertain.

403

404 Of the remaining 17 variants, all but three are accounted for by six genes (*BRAF*, *CBL*, *MAP2K1*,

405 *MAP2K2*, *RAF1* and *SOS1*) encoding members of the RAS-MAPK pathway, among which nine

406 variants have previously been reported in either congenital disorders or cancer (Table 1 and

407 Supplementary Figure 3). Moreover, for several variants (BRAF p.Gly469Ala, MAP2K1

408 p.Lys57Asn and p.Gln56Pro, MAP2K2 p.Cys125Ser, RAF1 p.Ser257Leu and p.Pro261Ala),

409 direct biochemical evidence of a dominant gain-of-function activity is available (Wan et al.

410 2004; Kobayashi et al. 2010; Van Allen et al. 2014; Arcila et al. 2015). In fact, only three

411 validated variants (#1,2,47), for which evidence of involvement in selfish selection is weak or

412 can be ruled out, were found in genes (*APC*, *AKT3*, *LRP5*) that function outside the RTK-RAS-

413 MAPK pathway (see Supplementary Note). Hence, although only 41.9% of the callable

414 sequence of our panel comprised RTK-RAS-MAPK candidate genes, 95% (57/60) of the

415 validated variants represented known or very likely pathogenic changes within members of

416 this signaling pathway (p value = 4.233e-13, Fisher's two tailed test), reinforcing the proposal

417 that activation of the RAS-MAPK pathway is the predominant mechanism underlying selfish

418 spermatogonial selection (Goriely et al. 2003; Goriely et al. 2009; Goriely and Wilkie 2012;

419 Maher et al. 2016a). Mutations in other core cellular pathways in human testes may either

420 not be associated with positive selection or may lead to milder clonal expansions that will

421 require more sensitive screening approaches to uncover. Although it can be difficult formally

422 to distinguish signals of selection from normal turnover/neutral drift dynamics whereby the

423 random loss of some clones is compensated by the expansion of others over time (Klein et al.

424 2010b; Simons 2016; Zink et al. 2017), the highly significant enrichment of functionally

425 significant (biochemically activating) mutations affecting a single signaling pathway argues

426 against a neutral process.

427

428 Among the variants we identified, we observed a high proportion of strongly oncogenic

429 mutations with 23 of the 35 non-synonymous variants reported in COSMIC (v82) having never

430 been described as constitutional mutations (Table 1). Strong gain-of-function mutations

431 would be more likely to promote efficient expansion of spermatogonial stem cells and result

432 in larger clones that are easier to detect. However, in order to be transmitted, the mutations

433 must be compatible with formation of functional sperm and with embryonic development.

434 We previously showed that tubules with spermatogonia harboring strongly oncogenic

435 variants are associated with reduced numbers of post-meiotic cells (Maher et al. 2016a). This

436 would represent a mechanism by which the testis 'filters' the transmission of pathogenic

437 mutations across generations, although proof of this concept would require development of

438 ultra-sensitive assays to screen large numbers of sperm samples. It is noteworthy that despite

439 the relative abundance of strongly oncogenic mutations in the adult male germline, testicular

440 tumors originating from adult spermatogonia (spermatocytic tumors) are extremely rare,

441 with an incidence of ~1 per million men and are mostly benign in nature (Ghazarian et al.

442 2015; Giannoulatou et al. 2017).

443

444 The age range of the testes analyzed in this study was highly skewed, with four being sampled

445 from older individuals (aged 71-90 years), and one (Tes5J) from a 34-year old man. While for

446 three of the four older individuals we identified multiple mutation-positive biopsies, Tes5J

447 from the younger man contained only two mutation-positive biopsies – likely representing a

448 single clonal event - carrying the oncogenic *CBL* c.1211G>A (p.Cys404Tyr) variant (at VAF 0.5-

22

449   0.6%), in keeping with the expectation that the prevalence and size of mutant clones increases

450   with time. It was however surprising that no variants were detected in Tes3D, given the

451   advanced age of the donor (87 years). Although it is possible that this individual may have had

452   a low propensity to accumulation of selfish mutations, a more likely explanation is that few

453   or no germ cells were present in this testis slice, either due to Sertoli-cell only syndrome or

454   due to age-related atrophy (Paniagua et al. 1987). Unfortunately, as no tissue had been

455   preserved for histological analysis, we were unable to determine the status of

456   spermatogenesis in this sample.

457

458   Our study has several technical limitations. The majority of variants identified were present

459   at VAFs <1%, close to the typical detection limits attributable to the error rates associated

460   with DNA damage (10^{-2}-10^{-4}) (Arbeithuber et al. 2016; Chen et al. 2017), PCR (10^{-4}-10^{-6})

461   (Hestand et al. 2016; Potapov and Ong 2017) and Illumina sequencing (~10^{-3}) (Minoche et al.

462   2011) (Salk et al. 2018). To account for such technical confounders, we employed a

463   conservative custom statistical approach to determine the background error rate at each

464   position and to prioritize variants (Supplementary Figure 1). Although we confirmed variants

465   with a frequency as low as 0.06% using this approach, the majority (81.8%) of the prioritized

466   variants called in single amplicon at VAFs of 0.1-0.2% (Tier 3) were false positives. In the

467   twelve samples amplified and sequenced in duplicate, only 7 of 15 variants were called in

468   both replicates (Supplementary Table 4). The best predictor of true positives was the

469   presence of a call in more than one amplicon (100% validation rate); for calls in single

470   amplicons the best predictor was the pathogenicity of the variant (17 of 18 (94.4%)

471   pathogenic variants vs. 5 of 30 (16.7%) without prior disease association validated). Broad-

472   scale approaches that target both DNA strands and use unique molecular indexes such as

473 Duplex sequencing (Kennedy et al. 2014) or smMIPs (Hiatt et al. 2013) (used here to validate

474 a subset of variants) represent valuable alternatives to direct PCR amplification in future

475 studies to reduce background errors (Salk et al. 2018). Overall 14% of the designed amplicons

476 did not pass quality control (due to insufficient coverage, mapping error…), which included

477 those targeting candidate PAE mutations such as eight mutational hotspots in *FGFR3,* six in

478 *PTPN11*, one in *RET* (p.Val804), and other key hotspots in *SKI* (Shprintzen-Goldberg

479 syndrome), *SETBP1* (Schinzel-Giedion syndrome) and *AKT1* (p.Glu17Lys – Proteus syndrome,

480 oncogenesis). Although considered to be the most frequently mutated nucleotide in the

481 germline with a birth prevalence of ~1:30,000 (Bellus et al. 1995), we did not detect the

482 *FGFR3* c.1138G>A or c.1138G>C achondroplasia-associated mutations due to exclusion of this

483 region because of insufficient coverage (<5,000x) (Supplementary Table 2; Supplementary Fig

484 3E).

485

486 In summary this work represents a new approach to studying DNMs directly in their tissue of

487 origin. By utilizing the clonal nature of mutations that leads to focal enrichment, we

488 circumvented the technical difficulties associated with calling DNMs in single sperm or the

489 poor DNA quality associated with immunopositive tubules from FFPE material. In a single

490 biopsy a whole population of *de novo* mutations can be assessed. Studying mutations within

491 the testis facilitates identification of mutations and pathways under positive selection in

492 spermatogonia but that may be incompatible with life, either by impairing gamete

493 differentiation and sperm production or by causing early embryonic lethality. Our approach

494 reveals the prevalence and geographical extent of clonal mutations in normal human testes,

495 suggesting that the ageing male germline is a repository for functionally significant, often

496 deleterious mutations. Based on an estimated total birth prevalence of DNMs causing

497   developmental disorders of 1 in 295 (DDD 2017), such PAE mutations may contribute 5-10%

498   of the total burden of pathological mutations, depending on paternal age. Investigating the

499   clonal nature of spontaneous testicular variants also provides insights into the regulation of

500   the poorly-studied human spermatogonial stem cell dynamics and how spontaneous

501   pathogenic mutations hijack homeostatic regulation in this tissue to increase their likelihood

502   of transmission to the next generation.

503

504

505   **METHODS**

506   **Testis samples**

507   Ethical approval was given for the use of human testicular tissue by the Oxfordshire Research

508   Ethics Committee A (C03.076: Receptor tyrosine kinases and germ cell development:

509   detection of mutations in normal testis, testicular tumors and sperm). Testes from five men

510   aged 34, 71, 83, 87 and 90 years were either commercially sourced or obtained locally from

511   research banks or post-mortems, with appropriate consent (Supplementary Table 5). Each

512   testis was cut into slices ~3-5 mm thick and either stored frozen at -80°C or formalin-fixed.

513   After thawing slices of frozen testis, extraneous tissue (epididymis or tunica albuginea) was

514   removed and slices were further dissected into 21-36 biopsies (Supplementary Table 5).

515   Biopsies were pulverized using a pestle and DNA extraction was performed using the Qiagen

516   DNeasy Blood & Tissue Kit. Samples with insufficient DNA quantity (determined using Qubit

517   fluorometer (Life Technologies)) or quality (determined using Nanodrop spectrophotometer

518   (Thermo Scientific)) were excluded, resulting in a total of 276 biopsies [Tes1D (34 biopsies),

519   Tes2F (30 biopsies), Tes3D (32 biopsies), Tes4B-4G (153 biopsies from 6 slices), Tes5J (27

520   biopsies)].

521

522   **RainDance library preparation and sequencing**

523   Primer pairs (tailed with common RainDance sequences (RD)) targeting 500 genomic regions

524   (20-169 bp [average 133 bp, median 143 bp]) in 71 genes (66.5 kb in total) were designed by

525   RainDance Technologies. The panel comprised mutational hotspots in the six established PAE

526   genes, genes encoding other RTKs and members of the RAS-MAPK signaling pathway, genes

527   in other pathways associated with spontaneous disorders that display narrow mutational

528   spectra suggestive of gain-of-function effects but lacking epidemiological data for paternal

529   age-effect, oncogenes commonly mutated in cancer, some of which are also associated with

530   germline disorders, and regions of 10 control genes. Details of all targeted regions and

531   primers used for amplification are provided in Supplementary Table 6. To maximize the

532   number of different molecules amplified, massively parallel simplex PCR was performed using

533   the RainDance Thunderstorm target enrichment system following the manufacturer's

534   instructions. Briefly, for each sample, 6 µg of genomic DNA (gDNA) was sheared to an average

535   size of 3,000 bp (using a Covaris blue AFA miniTUBE) and purified using a minElute column

536   (Qiagen). One microliter (out of 20 µl) was run on a gel to verify that the gDNA had been

537   sheared to the correct size range and the remaining gDNA was quantified using a Qubit

538   fluorometer (Life Technologies). The custom primer library, 1.75 µg of sheared gDNA and PCR

539   mix (Platinum Taq Polymerase High Fidelity reagents (Invitrogen), 2.5 mM $MgSO_4$, 0.35 µM

540   dNTPs, 0.6 M betaine, 7% dimethyl sulfoxide (DMSO), in 25 µl volume) were loaded onto a

541   ThunderStorm enrichment chip (48 samples at a time). Droplets containing up to 5 primer

542   pairs were merged with gDNA droplets to generate an average of 2 x 10^6 droplets per sample

543   (525,000 haploid genomes; average of 1 haploid genome per 3-4 droplets; ~1000

544   genomes/individual primer pair (Supplementary Figure 1). Following the merge, libraries

545 were PCR-amplified (94°C for 2 min; 54 cycles of 94°C, 54°C, 68°C for 30 s each; 68°C for 10

546 min) and the emulsion was broken down with 75-100 µl of Droplet Destabilizer (RainDance)

547 before being purified using AMPure beads (Agencourt). An aliquot of each sample was run on

548 a Bioanalyzer high sensitivity chip (Agilent) to verify the amplification profile and determine

549 the sample concentration. Sixteen different Illumina sequencing tailed libraries were

550 constructed using a set of barcoded (8 bp barcode (BC)) Illumina PE2-RD-rev adaptors, a

551 common PE1-RD-Fwd, 4 ng of merged amplicons and Phusion Hot Start Flex DNA Polymerase

552 (New England BioLabs) with 8% DMSO (98°C for 30 s, followed by 10 cycles of 98°C for 15 s,

553 56°C for 30 s, 72°C for 40 s, and a final extension at 72°C for 10 min). Following purification

554 (Qiagen MinElute), the relative concentration of the secondary tailing PCR samples was

555 estimated by Real-Time PCR using PE1 and PE2 primers. For each of the 16 libraries, 18

556 samples with BC1-18 were pooled in equimolar ratio and each final library was diluted to 10

557 nM. A total of 288 samples (264 singletons and 12 in duplicate) were amplified across 6

558 ThunderStorm enrichment chips (48 samples each) and subsequently ultra-deep sequenced

559 (~22,000x) on two flow cells (16 lanes; 18 samples per lane) of Illumina HiSeq 2000 (2 x 100

560 bp) using RD-Read1 and RD-Read2 custom sequencing primers generating 14-20 x 10^7 paired-

561 end reads per library.

562

563 **Sequence alignment and variant calling and prioritization**

564 Low quality reads with more than 20 bases below Q20, read pairs with one or two short (<50

565 bp) reads and reads pairs with unmatched or mismatched sequences between the forward

566 and reverse primer pairs expected for each amplicon were removed. Reads passing QC (on

567 average 86% of reads) were aligned to the human genome (hg19) using BWA-MEM version

568 0.7.10 (Li 2014) with default parameter settings. Primer sequences were included in the

569 alignment but ignored during variant quantification. The Python library Pysam was used to

570 fetch reads mapped to each amplicon and mapped bases (indicated as letter "M") were

571 identified from the CIGAR string. Pileup was then performed for each amplicon

572 independently. Nine amplicons that did not map to the targeted genomic regions were

573 excluded from subsequent analyses (Supplementary Table 2). Reads with more than 10 non-

574 reference bases were removed (<1% of coverage on average). For amplicons shorter than 200

575 bp, to avoid double-counting reads at positions where Read 1 and Read 2 overlapped, only

576 the base with the higher quality was considered.

577

578 Data exploration of the non-consensus variant counts within each amplicon across the

579 different samples revealed clear data structure with differences between flow cells,

580 sequencing lanes, coverage depths and base quality scores. To reduce false-positive calls,

581 primer sequences were trimmed and only variants supported by at least 10 reads were called.

582 To account for the technical confounders, the data were normalized (accounting for flow cell,

583 lane, and average base quality at each position) using a simple linear model

$$y_{i,s} = f_s + l_s + n_s + q_{i,s} + \epsilon_{i,s}$$

585 where $y_{i,s}$ is the nucleotide count for sample s at position i; f_s, l_s and n_s are the flow cell

586 identifier, the sequencing lane identifier, and individual identifier for sample s respectively;

587 and $q_{i,s}$ is the average base quality of sample s at position i. We used the glm package in R for

588 model inference (glm(y ~ f + l + n + q, family=gaussian())). Values of $\epsilon_{i,s}$ are the normalized

589 signals after accounting for the technical confounders and were used as the inputs for the

590 subsequent analyses. To further account for the effect of the sequencing lane structure, we

591 removed the median effect from each lane to reduce the background signal $g_{i,s} = \epsilon_{i,s} -$

592 median[$\epsilon_{i,s}$ for all s in same lane as s], before stabilizing the variance using the transformation

28

593    $g'_{i,s} = g_{i,s}/\text{IQR}[g_{i,.}]$, where $\text{IQR}[g_{i,.}]$ is the inter-quantile range at site i across all samples.

594   Following these normalization steps, variant calling was performed using a normal model to

595   test for an increase in non-consensus variant calls. Assuming that under the null hypothesis

596   the normalized variant quantification follows a normal distribution H_0: $P(g_i) = N(c\hat{\mu}_i, \hat{\sigma}_i^2)$ with

597   mean $\hat{\mu}_i$ and variance $\hat{\sigma}_i$, estimated using signals from all samples. We applied a one-sided z

598   test in R (pnorm(g, mean=mu0, sd=sigma2, lower.tail=FALSE). Non-consensus calls at each

599   genomic position across the 288 samples were tested independently in each amplicon that

600   passed QC. Variant prioritization was performed using a P-value cutoff of $-log_{10}P > 20$, which

601   resulted in a total of 19,625 genomic positions with at least one non-reference call.

602

603   As samples or amplicons with an excessive number of variants were more likely to represent

604   technical artefacts, these outliers were identified using a Chi Square (χ^2) test, where the

605   expected number of substitutions is defined as the median across all samples. Using a χ^2

606   threshold of $-log_{10}P > 3$, seven amplicons and 185 sample-mutation combinations were

607   removed from further analysis. Notably, the majority of these were C>A (=G>T) variant calls

608   (Supplementary Figure 4), which represent a known mutational signature associated with

609   oxidative stress that likely arose during sample preparation (Arbeithuber et al. 2016; Chen et

610   al. 2017). Further filtering was performed to remove potential sources of artefacts: calls

611   positioned 1 base from the amplification primer 3'-end were excluded; calls with a maximum

612   VAF of ≥3% were excluded to avoid calling SNPs and to eliminate gross alignment errors or

613   calling of non-consensus variants resulting from homologous genomic regions or pseudogene

614   amplification; positions with a median depth coverage below 5,000x across all samples were

615   excluded (this removed a 53 further amplicons (10.6%) from the analysis; Supplementary

616   Table 2). This resulted in a total of 5729 calls (5659 distinct variants) at 5421 positions, the

617   majority (90.2%) of which were made in a single amplicon and sample. As singleton calls were

618   more likely to represent PCR or sequencing artefacts, we further prioritized calls made in two

619   or more samples and/or present in overlapping amplicons. To exclude potential batch effects,

620   variants were excluded if all calls were made from a single library and the number of calls was

621   >3. This strategy identified 374 variants at 361 genomic positions. VAFs across all samples at

622   each of the 361 genomic positions were plotted and manually inspected for sequencing

623   library preparation or batch effects; raw sequencing reads from calls with suspected sequence

624   misalignment were visualized in Integrative Genomics Viewer (IGV) (Robinson et al. 2011).

625   Variant calls showing evidence of library-specific batch or sequence misalignment effects

626   were excluded from further analysis. Variants in *PTPN11* that matched bases at homologous

627   positions in one of its four pseudogenes were also excluded. The remaining 115 variants at

628   105 genomic positions were annotated with ANNOVAR version 2015Jun17 (Wang et al. 2010).

629

630   **Variant validation**

631   DNA from at least one putative-positive biopsy sample and at least 8 control samples

632   (unrelated blood gDNA and gDNA from other testicular biopsies) was screened by PCR or

633   single molecule molecular inversion probes (smMIPs) (primer and smMIP details in

634   Supplementary Table 6) and sequenced using Illumina MiSeq 300v2 (PCR) or 150v3 (smMIP)

635   kits (further details in Supplementary Methods). Demultiplexed reads were aligned to the

636   human genome (hg19) using BWA-MEM version 0.7.12 (Li 2014). Summary tables of the calls

637   across the aligned target region for PCR and smMIPs were generated using SAMtools mpileup

638   and a custom script (Amplimap – see Supplementary Methods), respectively. A base call was

639   only considered if its mapping quality was ≥Q20 and phred score ≥Q30. Validated variants

640   were annotated according to the following transcripts - *APC*: NM_001127510, *AKT3*:

641   NM_005465, *BRAF*: NM_004333, *CBL*: NM_005188, *FGFR2*: NM_000141, *FGFR3*:

642   NM_000142, *KRAS*: NM_033360, *LRP5:* NM_002335, *MAP2K1*: NM_002755, *MAP2K2*:

643   NM_030662, *NF1*: NM_001042492, *PTPN11*: NM_002834, *RAF1*: NM_002880, *RET*:

644   NM_020975, *SOS1*: NM_005633.

645

646   **Immunohistochemistry, microdissection and targeted mutation screen**

647   Where mutations had been identified in frozen sections for which an adjacent FFPE tissue

648   block was available, we attempted to visualize the corresponding mutant clone in sections of

649   the FFPE block. Immunohistochemical staining with anti-MAGEA4 antibody (clone 57B, gifted

650   by Prof. Giulio C. Spagnoli) to identify tubules with enhanced spermatogonial MAGEA4

651   staining, followed by laser capture microdissection and DNA extraction of adjacent FFPE

652   sections, was performed as described (Maher et al. 2016a). DNA was subsequently amplified

653   by PCR (40 cycles) using CS-tagged primers (Supplementary Table 6) and barcoded for Illumina

654   MiSeq 300v2 sequencing as described above (see also Supplementary Methods). DNA

655   samples extracted from the whole tissue section and from adjacent tubules with a normal

656   MAGEA4 staining appearance were used as controls. Reads were aligned to the human

657   genome (hg19) using BWA-MEM version 0.7.12 (Li 2014) and were visualized in IGV.

658

659   **DATA ACCESS**

660   **Databases and online resources**

661   gnomAD: http://gnomad.broadinstitute.org/

662   COSMIC: http://cancer.sanger.ac.uk/cosmic/

663   ClinVar: https://www.ncbi.nlm.nih.gov/clinvar/

664   OMIM: http://www.omim.org/

665

Author contributions:

Experiments: GJM, HKR, AG; Technical support: HM, PD, DSP, SS, SB; Data analysis: GM, HKR, ZD, NK, EG, GMcV, AG; Manuscript writing: GJM, AOMW, AG; Conception, design and supervision: GMcV, AOMW, AG

Supplementary material

Supplementary Figure 1 – Schematic of experimental design.

Supplementary Figure 2 – Distribution of mutations in slices Tes4B-4G from individual 4. Testicular biopsy numbers are located outside and to the left of each testis slice. Each variant has a distinct number (as listed in Table 1) and is colored according to gene: *FGFR2* (purple), *FGFR3* (orange), *KRAS* (black), *PTPN11* (blue), *RET* (pink), newly associated gene (red), NF1 mosaic (yellow with red surround). The size of each circle is proportional to the mutation frequency. Lines connect biopsies in the same slice with identical mutations; in cases where more than two biopsies are positive, the path of the clone has been arbitrarily drawn. Solid grey regions represent biopsies that were not sequenced due to quality control issues. Gridded grey regions represent non-tubular regions of tissue.

Supplementary Figure 3 – Individual gene plots showing the location of spontaneous mutations identified in testicular biopsies for AKT3 (A), APC (B), BRAF (C), CBL (D), FGFR3 (E), KRAS (F), LRP5 (G), MAP2K1 (H), MAP2K2 (I), NF1 (J), RAF1 (K), RET (L), and SOS1 (M). (Panel I) Validated variants (with VAF on *y*-axis) positioned along the amino acid sequence of the relevant protein (*x*-axis, see Panel V). (Panel II) Location and size of amplicons used to sequence main hotspots of the relevant genes are plotted on the *x*-axis. Median coverage per amplicon is plotted on the *y*-axis. Line indicates coverage cut-off of 5,000x. (Panel III) Number of reported constitutional variants encoding amino acid substitutions associated with developmental disorders (sqrt scale). (Panel IV) Number of reported somatic amino acid substitutions in cancer (COSMIC v82). (Panel V) Protein domains. Annotations are based on the transcripts accessions listed in the methods.

Supplementary Figure 4 - Variant allele frequencies of *KRAS* c.182A>G (p.Gln61Arg) and *LRP5* c.291C>T (p.Ala97Ala) in all 288 samples.

Supplementary Figure 5 – Heatmap of *NF1* c.2280G>A and *KRAS* c.35G>A. Heatmap of G>A variants in *NF1* (called in 9 biopsies in Tes4 – surrounded by black lines) and *KRAS* (called in 6 biopsies in Tes4 – surrounded by black lines) reveals that there were a number of additional pieces with relatively high levels of the NF1 c.2280G>A variant that were not called. Heatmaps of the same variants in Tes1 and Tes2 demonstrate that the higher levels are specific to Tes4.

Supplementary Figure 6 – Mutation loadings per sample. Note that a number of samples show excessive C>A(G>T) mutations, which is typically associated with oxidative stress during the experimental procedure. Filtering of specific sample-mutation combinations and amplicons with excessive number of variants resulted in 6054 variant calls.

List of Supplementary Tables and other supplementary files:
Supplementary Table 1 – Literature review showing loci with evidence for selfish selection
Supplementary Table 2 – Coverage analysis of 500 amplicons
Supplementary Table 3 – Table of prioritized calls (Tiers 1, 2, 3, 4)
Supplementary Table 4 – Variant calls in replicate samples

736   **Supplementary Table 5 – Sample information**
737   **Supplementary Table 6 – Primers and smMIPs**
738   **Supplementary Note**
739   **Supplementary Methods**
740
741
742

References

Acuna-Hidalgo R, Bo T, Kwint MP, van de Vorst M, Pinelli M, Veltman JA, Hoischen A, Vissers LE, Gilissen C. 2015. Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation. *Am J Hum Genet* **97**: 67-74.

Acuna-Hidalgo R, Sengul H, Steehouwer M, van de Vorst M, Vermeulen SH, Kiemeney L, Veltman JA, Gilissen C, Hoischen A. 2017. Ultra-sensitive Sequencing Identifies High Prevalence of Clonal Hematopoiesis-Associated Mutations throughout Adult Life. *Am J Hum Genet* **101**: 50-64.

Arbeithuber B, Makova KD, Tiemann-Boege I. 2016. Artifactual mutations resulting from DNA lesions limit detection levels in ultrasensitive sequencing applications. *DNA Res* **23**: 547-559.

Arcila ME, Drilon A, Sylvester BE, Lovly CM, Borsu L, Reva B, Kris MG, Solit DB, Ladanyi M. 2015. MAP2K1 (MEK1) Mutations Define a Distinct Subset of Lung Adenocarcinoma Associated with Smoking. *Clin Cancer Res* **21**: 1935-1943.

Bellus GA, Hefferon TW, Ortiz de Luna RI, Hecht JT, Horton WA, Machado M, Kaitila I, McIntosh I, Francomano CA. 1995. Achondroplasia is defined by recurrent G380R mutations of FGFR3. *Am J Hum Genet* **56**: 368-373.

Besenbacher S, Sulem P, Helgason A, Helgason H, Kristjansson H, Jonasdottir A, Magnusson OT, Thorsteinsdottir U, Masson G, Kong A et al. 2016. Multi-nucleotide de novo Mutations in Humans. *PLoS Genet* **12**: e1006315.

Campbell IM, Shaw CA, Stankiewicz P, Lupski JR. 2015. Somatic mosaicism: implications for disease and transmission genetics. *Trends Genet* **31**: 382-392.

Campbell IM, Yuan B, Robberecht C, Pfundt R, Szafranski P, McEntagart ME, Nagamani SC, Erez A, Bartnik M, Wisniowiecka-Kowalnik B et al. 2014. Parental somatic mosaicism is underrecognized and influences recurrence risk of genomic disorders. *Am J Hum Genet* **95**: 173-182.

Chen L, Liu P, Evans TC, Jr., Ettwiller LM. 2017. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science* **355**: 752-756.

Choi SK, Yoon SR, Calabrese P, Arnheim N. 2008. A germ-line-selective advantage rather than an increased mutation rate can explain some unexpectedly common human disease mutations. *Proc Natl Acad Sci USA* **105**: 10143-10148.

Choi SK, Yoon SR, Calabrese P, Arnheim N. 2012. Positive selection for new disease mutations in the human germline: evidence from the heritable cancer syndrome multiple endocrine neoplasia type 2B. *PLoS Genet* **8**: e1002420.

Coombs CC, Zehir A, Devlin SM, Kishtagari A, Syed A, Jonsson P, Hyman DM, Solit DB, Robson ME, Baselga J et al. 2017. Therapy-Related Clonal Hematopoiesis in Patients with Non-hematologic Cancers Is Common and Associated with Adverse Clinical Outcomes. *Cell Stem Cell* **21**: 374-382 e374.

Dakouane Giudicelli M, Serazin V, Le Sciellour CR, Albert M, Selva J, Giudicelli Y. 2008. Increased achondroplasia mutation frequency with advanced age and evidence for G1138A mosaicism in human testis biopsies. *Fertil Steril* **89**: 1651-1656.

DDD. 2017. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**: 433-438.

Eboreime J, Choi SK, Yoon SR, Arnheim N, Calabrese P. 2016. Estimating Exceptionally Rare Germline and Somatic Mutation Frequencies via Next Generation Sequencing. *PLoS One* **11**: e0158340.

789 Genovese G, Kahler AK, Handsaker RE, Lindberg J, Rose SA, Bakhoum SF, Chambert K, Mick
790   E, Neale BM, Fromer M et al. 2014. Clonal hematopoiesis and blood-cancer risk
791   inferred from blood DNA sequence. *N Engl J Med* **371**: 2477-2487.
792 Ghazarian AA, Trabert B, Graubard BI, Schwartz SM, Altekruse SF, McGlynn KA. 2015.
793   Incidence of testicular germ cell tumors among US men by census region. *Cancer*
794   **121**: 4181-4189.
795 Giannoulatou E, Maher GJ, Ding Z, Gillis AJM, Dorssers LCJ, Hoischen A, Rajpert-De Meyts E,
796   McVean G, Wilkie AOM, Looijenga LHJ et al. 2017. Whole-genome sequencing of
797   spermatocytic tumors provides insights into the mutational processes operating in
798   the male germline. *PLoS One* **12**: e0178169.
799 Giannoulatou E, McVean G, Taylor IB, McGowan SJ, Maher GJ, Iqbal Z, Pfeifer SP, Turner I,
800   Burkitt Wright EM, Shorto J et al. 2013. Contributions of intrinsic mutation rate and
801   selfish selection to levels of de novo HRAS mutations in the paternal germline. *Proc*
802   *Natl Acad Sci USA* **110**: 20152-20157.
803 Glass J. 2005. Testes and epididymes. In *Gray's Anatomy: The anatomical basis of clinical*
804   *practice (39th edition)*, (ed. S Standring), pp. 1304–1310. Churchill Livingston,
805   Edinburgh, UK.
806 Goldmann JM, Wong WS, Pinelli M, Farrah T, Bodian D, Stittrich AB, Glusman G, Vissers LE,
807   Hoischen A, Roach JC et al. 2016. Parent-of-origin-specific signatures of de novo
808   mutations. *Nat Genet* **48**: 935-939.
809 Goriely A, Hansen RM, Taylor IB, Olesen IA, Jacobsen GK, McGowan SJ, Pfeifer SP, McVean
810   GA, Rajpert-De Meyts E, Wilkie AOM. 2009. Activating mutations in FGFR3 and HRAS
811   reveal a shared genetic origin for congenital disorders and testicular tumors. *Nat*
812   *Genet* **41**: 1247-1252.
813 Goriely A, McGrath JJ, Hultman CM, Wilkie AOM, Malaspina D. 2013. "Selfish
814   spermatogonial selection": a novel mechanism for the association between
815   advanced paternal age and neurodevelopmental disorders. *Am J Psychiatry* **170**: 599-
816   608.
817 Goriely A, McVean GA, van Pelt AM, O'Rourke AW, Wall SA, de Rooij DG, Wilkie AOM. 2005.
818   Gain-of-function amino acid substitutions drive positive selection of FGFR2
819   mutations in human spermatogonia. *Proc Natl Acad Sci USA* **102**: 6051-6056.
820 Goriely A, McVean GAT, Rojmyr M, Ingemarsson B, Wilkie AOM. 2003. Evidence for selective
821   advantage of pathogenic FGFR2 mutations in the male germ line. *Science* **301**: 643-
822   646.
823 Goriely A, Wilkie AOM. 2012. Paternal age effect mutations and selfish spermatogonial
824   selection: causes and consequences for human disease. *Am J Hum Genet* **90**: 175-
825   200.
826 Hafner C, Toll A, Fernandez-Casado A, Earl J, Marques M, Acquadro F, Mendez-Pertuz M,
827   Urioste M, Malats N, Burns JE et al. 2010. Multiple oncogenic mutations and clonal
828   relationship in spatially distinct benign human epidermal tumors. *Proc Natl Acad Sci*
829   *USA* **107**: 20780-20785.
830 Hestand MS, Van Houdt J, Cristofoli F, Vermeesch JR. 2016. Polymerase specific error rates
831   and profiles identified by single molecule sequencing. *Mutat Res* **784-785**: 39-45.
832 Hiatt JB, Pritchard CC, Salipante SJ, O'Roak BJ, Shendure J. 2013. Single molecule molecular
833   inversion probes for targeted, high-accuracy detection of low-frequency variation.
834   *Genome Res* **23**: 843-854.

835   Holstege H, Pfeiffer W, Sie D, Hulsman M, Nicholas TJ, Lee CC, Ross T, Lin J, Miller MA, Ylstra
836   B et al. 2014. Somatic mutations found in the healthy blood compartment of a 115-
837   yr-old woman demonstrate oligoclonal hematopoiesis. *Genome Res* **24**: 733-742.
838   Jaiswal S, Fontanillas P, Flannick J, Manning A, Grauman PV, Mar BG, Lindsley RC, Mermel
839   CH, Burtt N, Chavez A et al. 2014. Age-related clonal hematopoiesis associated with
840   adverse outcomes. *N Engl J Med* **371**: 2488-2498.
841   Jonsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT,
842   Hjorleifsson KE, Eggertsson HP, Gudjonsson SA et al. 2017. Parental influence on
843   human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**: 519-
844   522.
845   Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, Prindle MJ, Kuong KJ, Shen JC,
846   Risques RA et al. 2014. Detecting ultralow-frequency mutations by Duplex
847   Sequencing. *Nat Protoc* **9**: 2586-2606.
848   Klein AM, Brash DE, Jones PH, Simons BD. 2010a. Stochastic fate of p53-mutant epidermal
849   progenitor cells is tilted toward proliferation by UV B during preneoplasia. *Proc Natl
850   Acad Sci U S A* **107**: 270-275.
851   Klein AM, Nakagawa T, Ichikawa R, Yoshida S, Simons BD. 2010b. Mouse germ line stem cells
852   undergo rapid and stochastic turnover. *Cell Stem Cell* **7**: 214-224.
853   Kobayashi T, Aoki Y, Niihori T, Cave H, Verloes A, Okamoto N, Kawame H, Fujiwara I, Takada
854   F, Ohata T et al. 2010. Molecular and clinical analysis of RAF1 in Noonan syndrome
855   and related disorders: dephosphorylation of serine 259 as the essential mechanism
856   for mutant activation. *Hum Mutat* **31**: 284-294.
857   Koczkowska M, Chen Y, Callens T, Gomes A, Sharp A, Johnson S, Hsiao MC, Chen Z,
858   Balasubramanian M, Barnett CP et al. 2018. Genotype-Phenotype Correlation in NF1:
859   Evidence for a More Severe Phenotype Associated with Missense Mutations
860   Affecting NF1 Codons 844-848. *Am J Hum Genet* **102**: 69-87.
861   Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA,
862   Sigurdsson A, Jonasdottir A, Wong WS et al. 2012. Rate of de novo mutations and the
863   importance of father's age to disease risk. *Nature* **488**: 471-475.
864   Krupp DR, Barnard RA, Duffourd Y, Evans SA, Mulqueen RM, Bernier R, Riviere JB, Fombonne
865   E, O'Roak BJ. 2017. Exonic Mosaic Mutations Contribute Risk for Autism Spectrum
866   Disorder. *Am J Hum Genet* **101**: 369-390.
867   Laurie CC, Laurie CA, Rice K, Doheny KF, Zelnick LR, McHugh CP, Ling H, Hetrick KN, Pugh
868   EW, Amos C et al. 2012. Detectable clonal mosaicism from birth to old age and its
869   relationship to cancer. *Nat Genet* **44**: 642-650.
870   Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage
871   samples. *Bioinformatics* **30**: 2843-2851.
872   Lim J, Maher GJ, Turner GD, Dudka-Ruszkowska W, Taylor S, Rajpert-De Meyts E, Goriely A,
873   Wilkie AO. 2012. Selfish spermatogonial selection: evidence from an
874   immunohistochemical screen in testes of elderly men. *PLoS One* **7**: e42382.
875   Maher GJ, Goriely A, Wilkie AOM. 2014. Cellular evidence for selfish spermatogonial
876   selection in aged human testes. *Andrology* **2**: 304-314.
877   Maher GJ, McGowan SJ, Giannoulatou E, Verrill C, Goriely A, Wilkie AO. 2016a. Visualizing
878   the origins of selfish de novo mutations in individual seminiferous tubules of human
879   testes. *Proc Natl Acad Sci U S A* **113**: 2454-2459.
880   Maher GJ, Rajpert-De Meyts E, Goriely A, Wilkie AO. 2016b. Cellular correlates of selfish
881   spermatogonial selection. *Andrology* **4**: 550-553.

882 Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, Davies H, Stratton
883   MR, Campbell PJ. 2017. Universal Patterns of Selection in Cancer and Somatic
884   Tissues. *Cell* **171**: 1029-1041 e1021.
885 Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, Wedge DC, Fullam A,
886   Alexandrov LB, Tubio JM et al. 2015. Tumor evolution. High burden and pervasive
887   positive selection of somatic mutations in normal human skin. *Science* **348**: 880-886.
888 McKerrell T, Park N, Moreno T, Grove CS, Ponstingl H, Stephens J, Crawley C, Craig J, Scott
889   MA, Hodkinson C et al. 2015. Leukemia-associated somatic mutations drive distinct
890   patterns of age-related clonal hemopoiesis. *Cell Rep* **10**: 1239-1245.
891 Minoche AE, Dohm JC, Himmelbauer H. 2011. Evaluation of genomic high-throughput
892   sequencing data generated on Illumina HiSeq and genome analyzer systems.
893   *Genome Biol* **12**: R112.
894 Neel BG, Gu H, Pao L. 2003. The 'Shp'ing news: SH2 domain-containing tyrosine
895   phosphatases in cell signaling. *Trends Biochem Sci* **28**: 284-293.
896 Nikolaev SI, Vetiska S, Bonilla X, Boudreau E, Jauhiainen S, Rezai Jahromi B, Khyzha N,
897   DiStefano PV, Suutarinen S, Kiehl TR et al. 2018. Somatic Activating KRAS Mutations
898   in Arteriovenous Malformations of the Brain. *N Engl J Med* **378**: 250-261.
899 Paniagua R, Martin A, Nistal M, Amat P. 1987. Testicular involution in elderly men:
900   comparison of histologic quantitative studies with hormone patterns. *Fertil Steril* **47**:
901   671-679.
902 Potapov V, Ong JL. 2017. Examining Sources of Error in PCR by Single-Molecule Sequencing.
903   *PLoS One* **12**: e0169774.
904 Qin J, Calabrese P, Tiemann-Boege I, Shinde DN, Yoon SR, Gelfand D, Bauer K, Arnheim N.
905   2007. The molecular anatomy of spontaneous germline mutations in human testes.
906   *PLoS Biol* **5**: e224.
907 Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Turki SA, Dominiczak A, Morris
908   A, Porteous D, Smith B et al. 2016. Timing, rates and spectra of human germline
909   mutation. *Nat Genet* **48**: 126-133.
910 Redig AJ, Capelletti M, Dahlberg SE, Sholl LM, Mach S, Fontes C, Shi Y, Chalasani P, Janne PA.
911   2016. Clinical and Molecular Characteristics of NF1-Mutant Lung Cancer. *Clin Cancer*
912   *Res* **22**: 3148-3156.
913 Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP.
914   2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24-26.
915 Salk JJ, Schmitt MW, Loeb LA. 2018. Enhancing the accuracy of next-generation sequencing
916   for detecting rare and subclonal mutations. *Nat Rev Genet* **19**: 269-285.
917 Shinde DN, Elmer DP, Calabrese P, Boulanger J, Arnheim N, Tiemann-Boege I. 2013. New
918   evidence for positive selection helps explain the paternal age effect observed in
919   achondroplasia. *Hum Mol Genet* **22**: 4117-4126.
920 Simons BD. 2016. Deep sequencing as a probe of normal stem cell fate and preneoplasia in
921   human epidermis. *Proc Natl Acad Sci U S A* **113**: 128-133.
922 Swanton C. 2015. Cancer evolution constrained by mutation order. *N Engl J Med* **372**: 661-
923   663.
924 Tiemann-Boege I, Navidi W, Grewal R, Cohn D, Eskenazi B, Wyrobek AJ, Arnheim N. 2002.
925   The observed human sperm mutation frequency cannot explain the achondroplasia
926   paternal age effect. *Proc Natl Acad Sci U S A* **99**: 14952-14957.

927 Van Allen EM, Wagle N, Sucker A, Treacy DJ, Johannessen CM, Goetz EM, Place CS, Taylor-
928   Weiner A, Whittaker S, Kryukov GV et al. 2014. The genetic landscape of clinical
929   resistance to RAF inhibition in metastatic melanoma. *Cancer Discov* **4**: 94-109.
930 Vermeulen L, Morrissey E, van der Heijden M, Nicholson AM, Sottoriva A, Buczacki S, Kemp
931   R, Tavare S, Winton DJ. 2013. Defining stem cell dynamics in models of intestinal
932   tumor initiation. *Science* **342**: 995-998.
933 Wan PT, Garnett MJ, Roe SM, Lee S, Niculescu-Duvaz D, Good VM, Jones CM, Marshall CJ,
934   Springer CJ, Barford D et al. 2004. Mechanism of activation of the RAF-ERK signaling
935   pathway by oncogenic mutations of B-RAF. *Cell* **116**: 855-867.
936 Wang H, Qian Y, Wu B, Zhang P, Zhou W. 2015. KRAS G12D mosaic mutation in a Chinese
937   linear nevus sebaceous syndrome infant. *BMC Med Genet* **16**: 101.
938 Wang K, Li M, Hakonarson H. 2010. ANNOVAR: Functional annotation of genetic variants
939   from next-generation sequencing data. *Nucleic Acids Res* **38**.
940 Wilkie AO. 2005. Bad bones, absent smell, selfish testes: the pleiotropic consequences of
941   human FGF receptor mutations. *Cytokine Growth Factor Rev* **16**: 187-203.
942 Yoon SR, Choi SK, Eboreime J, Gelb BD, Calabrese P, Arnheim N. 2013. Age-Dependent
943   Germline Mosaicism of the Most Common Noonan Syndrome Mutation Shows the
944   Signature of Germline Selection. *Am J Hum Genet* **92**: 917-926.
945 Yoon SR, Qin J, Glaser RL, Jabs EW, Wexler NS, Sokol R, Arnheim N, Calabrese P. 2009. The
946   ups and downs of mutation frequencies during aging can account for the Apert
947   syndrome paternal age effect. *PLoS Genet* **5**: e1000558.
948 Zink F, Stacey SN, Norddahl GL, Frigge ML, Magnusson OT, Jonsdottir I, Thorgeirsson TE,
949   Sigurdsson A, Gudjonsson SA, Gudmundsson J et al. 2017. Clonal hematopoiesis,
950   with and without candidate driver mutations, is common in the elderly. *Blood* **130**:
951   742-752.
952