Are sites with multiple single nucleotide variants in cancer genomes a consequence of drivers, hypermutable sites or sequencing errors?

Thomas C A Smith, Antony M Carr, Adam C Eyre-Walker

Across indepedent cancer genomes it has been observed that some sites have been recurrently hit by single nucleotide variants (SNVs). Such recurrently hit sites might be either i) drivers of cancer that are postively selected during oncogenesis, ii) due to mutation rate variation, or iii) due to sequencing and assembly errors. We have investigated the cause of recurrently hit sites in a dataset of >3 million SNVs from 507 complete cancer genome sequences. We find evidence that many sites have been hit significantly more often than one would expect by chance, even taking into account the effect of the adjacent nucleotides on the rate of mutation. We find that the density of these recurrently hit sites is higher in non-coding than coding DNA and hence conclude that most of them are unlikely to be drivers. We also find that most of them are found in parts of the genome that are not uniquely mappable and hence are likly to be due to mapping errors. In support of the error hypothesis, we find that recurently hit sites are not randomly distributed across sequences from different laboratories. We fit a model to the data in which the rate of mutation is constant across sites but the rate of error varies. This model suggests that ~4% of all SNVs are error in this dataset, but that the rate of error varies by thousands-of-fold.

1   **Are sites with multiple single nucleotide variants in cancer genomes a**

2   **consequence of drivers, hypermutable sites or sequencing errors?**

3

4

5   Thomas C. A. Smith[1]

6   Antony M. Carr[2]

7   Adam Eyre-Walker[1]

8

9   1. School of Life Sciences, University of Sussex, Falmer, Sussex, BN1 9RQ, United

10   Kingdom.

11

12   2. Genome Damage and Stability Centre, University of Sussex, Falmer, Sussex, BN1

13   9RQ, United Kingdom.

14

15

16   Corresponding Authors:

17   Thomas C A Smith[1] and Adam Eyre-Walker[1].

18   1. School of Life Sciences, University of Sussex, Falmer, Sussex, BN1 9RQ, United

19   Kingdom.

20

21   Email address: t.c.a.smith@sussex.ac.uk

22   Email address: a.c.eyre-walker@sussex.ac.uk

23

24

25

1

26 **Abstract.**

27

28   Across indepedent cancer genomes it has been observed that some sites have been

29   recurrently hit by single nucleotide variants (SNVs). Such recurrently hit sites might

30   be either i) drivers of cancer that are postively selected during oncogenesis, ii) due to

31   mutation rate variation, or iii) due to sequencing and assembly errors. We have

32   investigated the cause of recurrently hit sites in a dataset of >3 million SNVs from

33   507 complete cancer genome sequences. We find evidence that many sites have been

34   hit significantly more often than one would expect by chance, even taking into

35   account the effect of the adjacent nucleotides on the rate of mutation. We find that the

36   density of these recurrently hit sites is higher in non-coding than coding DNA and

37   hence conclude that most of them are unlikely to be drivers. We also find that most of

38   them are found in parts of the genome that are not uniquely mappable and hence are

39   likly to be due to mapping errors. In support of the error hypothesis, we find that

40   recurently hit sites are not randomly distributed across sequences from different

41   laboratories. We fit a model to the data in which the rate of mutation is constant across

42   sites but the rate of error varies. This model suggests that ~4% of all SNVs are error

43   in this dataset, but that the rate of error varies by thousands-of-fold.

44

45

46

47

48

49

2

50 **Introduction.**

51

52 There is currently huge interest in sequencing cancer genomes with a view to

53 identifying the mutations in somatic tissues that lead to cancer, the so called "driver"

54 mutations. Driver mutations are expected to cluster in particular genes or genomic

55 regions, or to recur at particular sites in the genome, because only a limited number of

56 mutations can cause cancer. For example, the driver mutations in the TERT1 promoter

57 were identified because it had independently occurred in multiple cancers (Huang et

58 al., 2013). However, there are two other processes that can potentially lead to the

59 repeated occurrence of an apparent somatic mutation at a site. First, it is known that

60 the mutation rate varies across the genome at a number of different scales in both the

61 germ-line and soma (Hodgkinson & Eyre-Walker, 2011; Hodgkinson, Chen & Eyre-

62 Walker, 2012; Michaelson et al., 2012; Francioli et al., 2015). Sites with recurrent

63 SNVs could simply be a consequence of sites with high rates of mutations. And

64 second there is the potential for sequencing error. Although, the average rate of

65 sequencing error is thought to be quite low it is evident that some types of sites, such

66 as those in runs of nucleotides, are difficult to sequence accurately. Furthermore, since

67 the genome contains many similar sequences it can often be difficult to map

68 sequencing reads successfully (Treangen & Salzberg, 2013).

69

70 In the germ-line the density of point mutations varies at a number of different scales

71 (Hodgkinson & Eyre-Walker, 2011). At the mega-base scale the mutation varies by

72 about 2-fold, and ~50% of this variance can be explained by correlations with factors

73 such as replication time, recombination rate and distance from telomeres (as reviewed

3

74   in (Hodgkinson & Eyre-Walker 2011)). However the greatest variance, reportedly up

75   to ~30-fold, has been found at the single nucleotide level (Hodgkinson, Chen & Eyre-

76   Walker, 2012; Kong et al., 2012; Michaelson et al., 2012), whereby the nucleotide

77   context, that is the identity of the bases immediately 5' and 3' of the mutated base, are

78   highly influential on the rate of mutation (Gojobori, Li & Graur, 1982; Bulmer, 1986;

79   Cooper & Krawczak, 1990; Nachman & Crowell, 2000; Hwang & Green, 2004). The

80   most well known example is that of CpG hyper-mutation (Bird, 1980), which is

81   thought to account ~20% of all mutations in the human genome (Fryxell & Moon,

82   2005). However there is also variation at the single nucleotide level that cannot be

83   ascribed to the effects of neighbouring nucleotides; this has been termed cryptic

84   variation in the mutation rate and is thought to account for at least as much variation

85   in the mutation rate as does simple context (Hodgkinson, Ladoukakis & Eyre-Walker,

86   2009; Eyre-Walker & Eyre-Walker, 2014).

87

88   The somatic mutation rate is estimated to be at least an order of magnitude greater

89   than that of the germ line (Lynch, 2010). It has been shown to vary between cancers

90   (Lawrence et al. 2013) and different cancer types are known to vary in their relative

91   contributions of different mutations to their overall mutational compositions

92   (Alexandrov et al., 2013). For a review see (Martinocorena & Campbell, 2015). The

93   aforementioned correlates of variation that are found in the germ line are also

94   apparent in the soma (Hodgkinson, Chen & Eyre-Walker, 2012; Schuster-Bockler &

95   Lehner, 2012; Lawrence et al., 2013; Liu, De & Michor, 2013), for example

96   replication time correlates strongly with single nucleotide variant (SNV) density at the

97   1Mb base scale and can vary by up to 3-fold along the genome (Hodgkinson & Eyre-

4

98   Walker, 2011; Woo & Li, 2012). However, as yet there has been no attempt to

99   quantify the level of cryptic variation in the mutation rate at the single nucleotide

100   level in the somatic genome. This is an important property to understand; for example

101   a site which experiences a recurrence of SNVs across many cancer genomes would be

102   of interest as a potential driver of cancer (Lawrence et al., 2013), however, this site

103   might simply be cryptically hypermutable (Hodgkinson, Ladoukakis & Eyre-Walker,

104   2009; Eyre-Walker & Eyre-Walker, 2014; Smith et al., 2016). Here we examine the

105   distribution of recurrent SNVs taken from 507 whole genome sequences made

106   publicly available by Alexandrov et al. (2013) to investigate the level of cryptic

107   variation in the mutation rate for somatic tissues. We show that there is a large excess

108   of sites that have been hit by recurrent SNVs. Since the density of these is greater in

109   the non-coding, than the coding fraction of the genome, we conclude that most of

110   them are unlikely to be drivers. We therefore investigate whether they are due to

111   mutational heterogeneity or sequencing errors. In particular we investigate whether

112   there might be cryptic variation in the mutation rate in cancer genomes.

113   Unfortunately, the available evidence suggests that most sites with recurrent SNVs are

114   likely to be due to sequencing error or errors in post-sequencing processing.

115

116

117   **Methods.**

118

119   *Genome and data filtering.*

120   The human genome (hg19/GRCh37) was masked to remove simple sequence repeats

121   (SSR) as defined by Tandem Repeat Finder (Benson, 1999). The remaining regions

5

122 were separated into three genomic fractions, consisting of 1,346,629,686 bp of non-

123 coding transposable element DNA (TE), defined as LINEs, SINEs, LTRs and DNA

124 transposons as identified by repeat masker (Smit et al. 1996), 1,322,985,768 bp of

125 non-coding non-transposable element DNA (NTE), and 119,806,141 bp of exonic

126 non-transposable element DNA (EX) defined by Ensemble (Flicek et al., 2011). From

127 the supplementary data of Alexandrov et al. (2013) we collated 3,382,737 single

128 nucleotide variants (SNV), classified as "somatic-for-signature-analysis" (see

129 (Alexandrov et al., 2013) for SNV filtering methods). These can be downloaded from

130 ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl/. These came from 507 whole

131 genome sequenced cancers and represent 10 different cancer types and were reduced

132 to 3,299,881 SNVs when excluding SNVs in SSRs; 1,666,759 in TE and 1,535,069 in

133 NTE and 98053 in EX.

134

135 *Testing for mutation rate heterogeneity.*

136 We were interested in whether some sites have more SNVs than expected by chance.

137 Since the mutation rate is affected by the identity of the neighbouring nucleotides we

138 need to control for those effects. To do this we separated each SNV into one of 64

139 categories based upon the triplet to which it was the central base. This was reduced to

140 32 triplets when accounting for base complementarity with the pyrimidine (C/T) taken

141 as the central base. If the total number of triplets of type i (e.g. C**T**C in the non-TE

142 fraction) is l_i and the number SNVs at that triplet is m_i then the expected number of

143 sites hit x times can be calculated using a Poisson distribution:

144

145
$$P_i(x) = l_i \frac{e^{-\mu_i} \mu_i^x}{x!} \qquad (1)$$

6

146   where $\mu_i = m_i/l_i$ is the mean number of SNVs per site, The expected number of sites

147   with x SNVs across all triplets was calculated by summing the values of $P_i(x)$.

148   Whether the observed distribution deviated from the expected was tested using a

149   chisquare test.

150

151   *Model fitting*

152   As well as testing whether there was significant hereogeneity we were also

153   interested in quantifying the level of variation. We fit two basic models. In the first we

154   allowed the density of SNVs to follow a gamma distribution. Let the expected density

155   of SNVs at a site be $\mu\alpha$ where μ is the mean density of SNVs for a particular triplet

156   and α is the deviation from this mean which is gamma distributed, parameterised such

157   that the gamma has a mean of one. Under this model the expected number of sites

158   with x SNVs is

159

160     $$P(x) = l \int_0^\infty \frac{e^{-\mu\alpha}(\mu\alpha)^x}{x!} D(\alpha)\,d\alpha \qquad (2)$$

161

162

163   In a second model we imagine that the production of SNVs depends upon two

164   processes, one of which is constant across sites, and one which varies across sites with

165   the rate drawn from a gamma distribution. Let the proportion of SNVs due to the first

166   process be ε. Under this model the expected number of sites with x SNVs is

167

168     $$P(x) = l \int_0^\infty \frac{e^{-\mu(\varepsilon+(1-\varepsilon)\alpha)}(\mu(\varepsilon+(1-\varepsilon)\alpha))^x}{x!} D(\alpha)\,d\alpha \qquad (3)$$

7

169   Given the expected number of sites, the likelihood of observing $\hat{P}(x)$ sites with x

170   SNVs is itself Poisson distributed

171

172    $$L(x) = \frac{e^{-P(x)}P(x)^{\hat{P}(x)}}{\hat{P}(x)!} \qquad (4)$$

173

174   These likelihoods can be multiplied across triplets to obtain the overall likelihood. We

175   estimated the maximum likelihood values of the model parameters using the

176   Maximize function of Mathematica which implements the Nelder-Mead algorithm

177   (Nelder et al., 1965).

178   .

179   *Privacy analysis*

180   To investigate whether the SNVs at some sites tended to be produced by a particular

181   research group we took all sites with 3 or more SNVs from the same cancer type and

182   then performed Fishers exact test on a 2 x 30 matrix using the the R stats package,

183   version 3.2.4 (R Core Team, 2016).

184

185   *Mappability.*

186   Each nucleotide in genome was assigned a mappability score, as determined by the

187   Mappability track (Derrien et al., 2012) downloaded from the UCSC table browser at

188   http://genome.ucsc.edu/ (Karolchik et al., 2004). This feature assigns a value of 1 to

189   unique k-mer sequences in the genome, 0.5 to those that occur twice, 0.33 to those

190   that occur thrice etc. This is computed for every base in the human genome with the

191   value being assigned to the first position of the k-mer. We used k-mers of 100 and 20

192   bases.

8

193   **Results**.

194

195   *The distribution of recurrent SNVs.*

196   If there is no variation in the density of single nucleotide variants (SNVs) then we

197   should find them to be distributed randomly across the genome. To investigate

198   whether this was the case we calculated the expected number of sites with 1,2,3…etc

199   SNVs, taking into account the fact that some triplets have higher mutation rates than

200   others. We found that there are some sites that have 7 SNVs whereas we expect very

201   few sites to have more than 3 SNVs – the difference is highly significant using the

202   Chi-square goodness of fit test ($p < 0.0001$) for both the whole genome (Total) and

203   when separating the genome into non-coding transposable elements (TE), non-coding

204   non-transposable elements and (NTE) and exons (EX) (Table 1). We refer to sites

205   with 3 or more SNVs as excess sites. In total we observed 1187 excess sites (Table 1)

206   with the density of excess sites in TE being 3.9 and 3.4 fold greater than in NTE and

207   EX respectively. The probability of this level of SNV recurrence is so low (Chi-

208   squared goodness of fit test, $p > 0.0001$) that these excess sites must either be (i)

209   drivers, (ii) the result of mutation rate heterogeneity across the genome or, (iii) the

210   consequence of next generation sequencing (NGS) pipeline errors.

9

A) – All Sites

Site Type	0 hits	1 hit	2 hits	3 hits	4 hits	5 hits	6 hits	7 hits
Non-Exon TE obs (TE)	1.34E+9	1.65E+6	7034	762	130	26	9	3
Non-Exon TE exp (TE)	1.34E+9	1.66E+6	1430	1.14	9E-4	7E-7	5E-10	4E-13
Non-Exon Non-TE obs (NTE)	1.32E+9	1.53E+6	3171	188	35	6	2	2
Non-Exon Non-TE exp (NTE)	1.32E+9	1.53E+6	1206	0.86	6E-4	4E-7	3E-10	2E-13
Exon obs (EX)	1.20E+8	9.75E+4	245	23	0	0	1	0
Exon exp (EX)	1.20E+8	9.79E+4	57	0.03	2E-5	7E-9	3E-12	1E-15
Total obs	1.44E+9	1.63E+6	10450	973	165	32	12	5
Total exp	1.44E+9	1.63E+6	2692	2.04	2E-3	1E-6	8E-10	5E-13

B) – Mappable 100

Site Type	0 hits	1 hit	2 hits	3 hits	4 hits	5 hits	6 hits	7 hits
Non-Exon TE obs (TE)	1.22E+9	1.52E+6	3927	266	25	11	5	1
Non-Exon TE exp (TE)	1.22E+9	1.52E+6	1322	1.07	9E-4	7E-7	5E-10	4E-13
Non-Exon Non-TE obs (NTE)	1.28E+9	1.50E+6	2698	97	16	2	0	1
Non-Exon Non-TE exp (NTE)	1.28E+9	1.50E+6	1201	0.88	6E-4	5E-7	3E-10	2E-13
Exon obs (EX)	1.12E+8	9.31E+4	185	16	0	0	0	0
Exon exp (EX)	1.12E+8	9.34E+4	55	0.03	2E-5	7E-9	3E-12	1E-15
Total obs	1.39E+9	1.59E+6	6810	379	41	13	5	2
Total exp	1.39E+9	1.60E+6	2578	2	2E-3	1E-6	8E-10	6E-13

C) – Mappable 20

Site Type	0 hits	1 hit	2 hits	3 hits	4 hits	5 hits	6 hits	7 hits
Non-Exon TE obs (TE)	3.89E+8	4.81E+5	741	9	0	0	0	0
Non-Exon TE exp (TE)	3.89E+8	4.81E+5	417	0.34	3E-4	2E-7	2E-10	1E-13
Non-Exon Non-TE obs (NTE)	8.92E+8	1.06E+6	1621	31	4	1	0	1
Non-Exon Non-TE exp (NTE)	8.92E+8	1.06E+6	868	0.65	5E-4	3E-7	2E-10	2E-13
Exon obs (EX)	7.47E+7	6.10E+4	103	6.00	0	0	0	0
Exon exp (EX)	7.47E+7	6.12E+4	36	0.02	9E-6	4E-9	2E-12	7E-16
Total obs	9.67E+8	1.12E+6	2465	46	4	1	0	1
Total exp	9.67E+8	1.12E+6	1321	1	8E-4	6E-7	4E-10	3E-13

Table 1. Observed and expected values for the distribution of SNVs for sites hit from 0-7 times. A) shows data for the whole interrogable human genome, excluding simple sequence repeats. B) shows data for all bases in the genome that are uniquely mappable at 100 base pairs. C) the same as B but for 20 base pairs. $P < 0.001$ for observing >7 sites with 3 SNVs in A),B) and C) if SNVs were randomly distributed throughout the genome.

10

216   It seems unlikely that the majority of the excess sites are due to drivers since the

217   density of excess sites is higher in the TE and NTE parts of the genome than in EX

218   (Table 1A). Furthermore, to date only one intergenic driver of cancer – an activating

219   C>T mutation in the *TERT* promoter (Huang et al. 2013) at chr5:1,295,228 – has been

220   confirmed, and although this is included in the excess sites with 7 SNVs, the

221   remaining 1186 excess sites are unlikely to be under such selection. It therefore seems

222   likely that the excess sites are either due to mutation rate variation or problems with

223   sequencing.

224

225   *Excess sites are enriched in non-unique sequences.*

226   The human genome contains many duplicated sequences particularly within

227   transposable elements, and these pose challenges for accurate alignment of the short

228   ~100bp reads produced from NGS (Zhuang et al., 2014). If the excess sites were the

229   result of NGS mapping errors then we might expect them to occur in regions of the

230   genome that were hard to align. Using the mappability scores (Derrien et al., 2012)

231   we excluded all bases that were not uniquely mappable at 100bp. This only reduced

232   the interrogable genome by 6%, but the number of excess sites was reduced by 64%

233   (Table 1B), demonstrating that a large proportion of the excess sites were in

234   duplicated sequences and therefore likely originate from mapping errors. However,

235   even with this large reduction in excess sites we still observed many excess sites far

236   greater than chance expectation (Chi-squared goodness of fit test, $p > 0.0001$) (Table

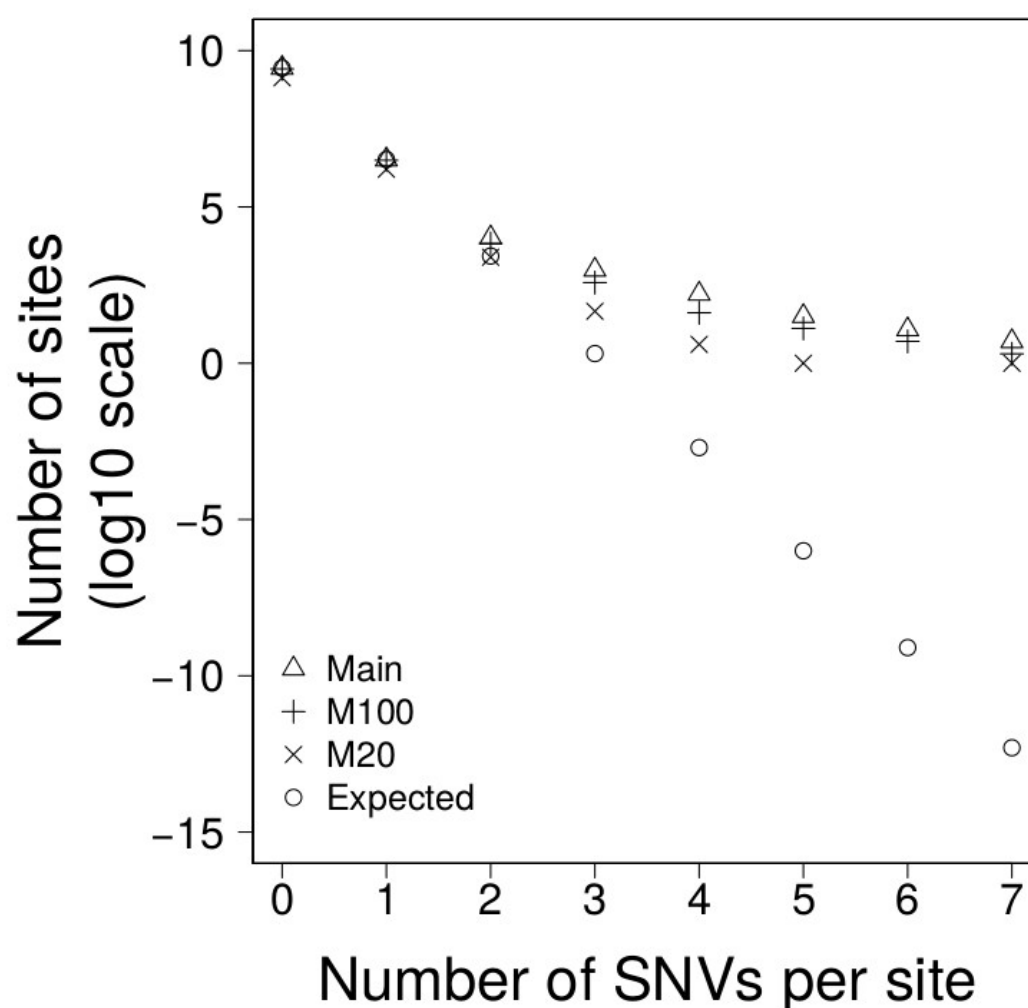237   1B & Figure 1).

11

238   Figure 1. The number of site with 0-7 SNVs per sites for: **Main** = all data, **M100** = sites that are

239   uniquely mappable at 100 base-pairs, **M20** = sites that are uniquely mappable at base-pairs and the

240   expected number drawn from a poisson distribution.

241

242   The SNVs in this data were all called from >100bp reads. If the excess sites were

243   errors of read mapping, they should not be affected by the uniqueness of shorter

244   sequences (i.e. there is no reason why 100bp sequences that map uniquely to the

245   genome should be mis-mapped if it contains a non-unique 20bp sequence), however if

246   the SNVs were the product of a biological process that was more prevalent in non-

12

247   unique or repetitive sequences, then we might expect to see a reduction of excess sites

248   when we exclude all bases that do not map uniquely at 20bp. When we excluded all

249   bases that were not unique at 20bp we found that the interrogable genome was

250   reduced by 52% and the excess sites were reduced by 96% (Table 1C & Figure 1). It

251   is worth noting that, due to their proliferative nature throughout the genome, this

252   reduction disproportionately affects TEs where the interrogable genome is reduced by

253   71% and the excess sites by >99%. This would suggest that the excess sites existing in

254   sequences that were unique at 100bp but not unique at 20bp likely represent some

255   biological process and not error. Furthermore, th*e TE*RT promoter, whose recurrence

256   is the result of positive selection, and is therefore the only excess site that that we can

257   confidently say is not a product of error, remains in this most conservative of

258   analyses. Despite this large reduction in excess sites, significant heterogeneity still

259   remains; the probability of observing the 52 excess sites in the part of the genome

260   uniquely mappable at 20 bases is still extremely low (Chi-squared goodness of fit test,

261    $p < 0.0001$).

262

263   *Privacy of mutations.*

264   To further investigate the origin of excess sites we exploited the fact that some types

265   of cancer were sequenced by different laboratories using different technologies and

266   NGS pipelines. If the SNVs at excess sites found in a particular cancer are due to

267   hypermutable sites then we would expect them to be randomly distributed across

268   research groups (i.e. all research groups should identify the same hypermutable sites).

269   If however the SNVs at excess sites are due to error then we might expect them to be

270   heterogeneously distributed across research groups (i.e. the calling of recurrent false

13

271 positive SNVs should be systematic of individual research group NGS pipelines). The

272 liver cancers, which were all virus associated hepatocellular carcinomas, , were

273 sequenced by two different groups; 66 from the RIKEN group using the Illumina

274 Genome Analyser (https://dcc.icgc.org/projects/LIRI-JP) and 22 from the National

275 Cancer Centre in Japan using the IIlumina HiSeq platform

276 (https://dcc.icgc.org/projects/LINC-JP). We found that the SNVs were

277 heterogeneously distributed amongst research groups (Fisher's exact test, $P = 4\text{x}10^{-6}$)

278 suggesting that the 30 excess sites from liver cancers were predominantly errors

279 (Supplementary Table 1).

280

281 *Parameter estimation*

282 To gauge how much variation there is in the density of SNVs across the genome we

283 fit two models to the data using maximum likelihood. In model 1 we allowed the

284 density of SNVs to vary between sites according to a gamma distribution, estimating

285 the shape parameter, and hence the amount of variation there was between sites. We

286 fitted two versions of this model. In the first version, 1a, we constrained the model

287 such that the mean SNV density, shape parameter, and hence the level of variation,

288 was the same for all triplets. In the second version, 1b, we allowed the mean SNV

289 density and shape parameter to vary between triplets. The second of these models fits

290 the data significantly better than the first according to a likelihood ratio test

291 suggesting that the level of variation differs between triplets (Table 2). However, a

292 goodness of fit test, comparing the number of sites predicted to have 1, 2, 3…etc

293 SNVs per site to the observed data, suggests the model fits the data poorly. We

294 therefore fit a second pair of models in which we allowed the rate of SNVs to be due

14

295 to two processes. The first process, is constant across sites whereas the second process

296 is variable and drawn from a gamma distribution. There are two parameters in the

297 model, the proportion of SNVs at a site produced by the first process and the level of

298 variation in the second process. This model might represent a situation where the rate

299 of mutation is constant across sites but the rate of sequencing error is variable. As

300 with the first model we fit two versions of this model; in Model 2a we constrained the

301 model such that the parameters of the two processes were the same for all triplets. In

302 Model 2b they were allowed to vary between triplets. Both models 2a and 2b fit the

303 data significantly better than models 1a and 1b, and of this second pair of models,

304 model 2b, which allows the parameters to vary between triplets fits the data

305 significantly better than model 2a, in which the parameters are shared across triplets

306 (Table 2). The best fitting model is therefore one in which we have two processes

307 contributing to the production of SNVs, one that is constant across sites, although it

308 differs between triplets, and one which is variable across sites. Although, we can

309 formally reject this model using a goodness-of-fit test (Chi-square $p < 0.0001$),

310 because we have so much data, it is clear that the model fits the data fairly well

311 (Figure 2). Under this model we estimate that approximately 4.1%, 2.8% and 4.3% of

312 SNVs are due to the process that varies across sites in the TE and NTE, and EX

313 sequences respectively. However, the variation in the density between sites due to the

314 variable process is extremely large. The median shape parameters are 0.0013, 0.0011

315 and 0.00075 for the TE and NTE, and EX sequences respectively. Under a gamma

316 distribution with a shape parameter of 0.0004 we would expect more than 99% of

317 sites to have no SNVs generated by this variable process, but some sites to have a

318 density of SNVs that is 30,000-fold above the average rate.

15

319

Non-Exon TE (TE)

Model	N	Log-likelihood	Shape	Median ε
1a	2	-269283	0.13	
1b	64	-2936	0.12	
2a	3	-266889	0.00021	0.044
2b	*96*	*-1302*	*0.0013*	*0.041*

Non-Exon Non-TE (NTE)

Model	N	Log-likelihood	Shape	Median ε
1a	2	-227728	0.31	
1b	64	-1207	0.37	
2a	3	-227026	0.0012	0.037
2b	*96*	*-566*	*0.0011*	*0.028*

Exon (EX)

Model	N	Log-likelihood	Shape	Median ε
1a	2	-13878	0.18	
1b	64	-270	0.22	
2a	3	-13842	0.00081	0.034
2b	*96*	*-240*	*0.00076*	*0.043*

320 Table 2. The fit of 4 models to the observed distribution of recurrent SNVs in the three different

321 genomic fractions A) TE, B) NTE and C) EX. **N** = number of parameters. *Italics* indicate the best fit as

322 determined by a liklihood ratio test.
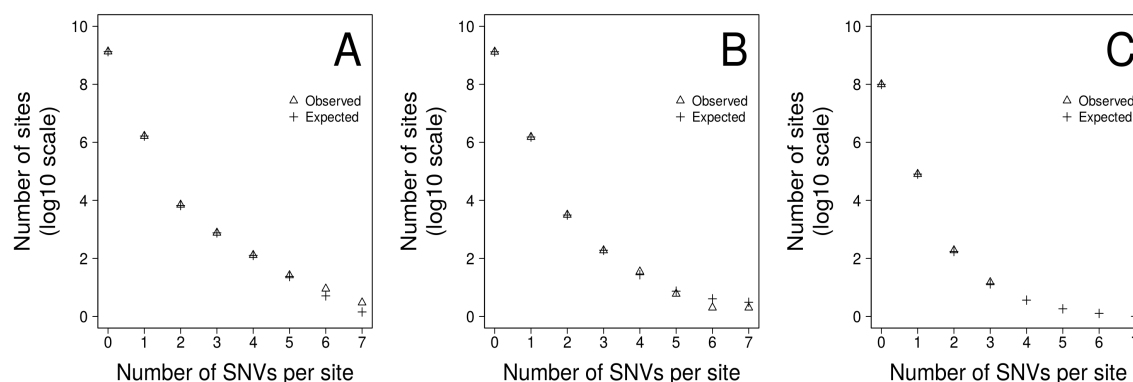
323

324

326 Figure 2. The fit of the observed recurrent SNV distribution to expected distribution under the favoured

327 model, 2b, for A) TE, B) NTE and C) EX genomic fractions.

328

16

329 **Discussion.**

330

331 Through our analysis of ~3 million SNVs from whole cancer genomes we have

332 shown that there are many sites at which there is a significant excess of SNVs. The

333 majority of these are unlikely to be drivers because the density of sites with an excess

334 of SNVs is greater in the non-coding part of the genome than in the exons. It therefore

335 seems likely that the majority of the excess sites are either due to hypermutation or

336 problems with sequencing or the processing of the sequences. Several lines of

337 evidence point to sequencing problems being the chief culprit. First, many of the

338 excess sites disappear when regions of the genome with low mappability are removed.

339 Second, SNVs at a particular excess site tend to be found within the sequences from a

340 particular laboratory; for example, site 85,091,895 on chromosome 5 has 5 SNVs in

341 liver cancers, but all of these are found in the sequences from RIKEN not the

342 sequences from the NCC. Third, the level of variation in the density of SNVs is much

343 greater than has been observed or suggested for variation in the mutation rate

344 (Hodgkinson & Eyre-Walker, 2011; Kong et al., 2012; Michaelson et al., 2012)

345 though see a recent analysis of de novo germ-line mutations which suggests there

346 could be extreme mutational heterogeneity (Smith et al., 2016); some sites are

347 estimated to have rates of SNV production that are tens of thousands of times faster

348 than the genomic average.

349

350 Only one line of evidence suggests that there might also be substantial variation in the

351 mutation rate as well as variation in the error rate. When we eliminate sites that are

352 not uniquely mappable at 20bp we find a great reduction in the number of excess sites

17

353 relative to the case when we remove sites that are not uniquely mappable at 100bp,

354 and yet the read length is greater than 100bp in the data that we have used. This might

355 suggest that there are some repetitive sequences that are prone to a process of hyper-

356 mutation. However, it might also be that mappability at 100bp is not a good guide to

357 mappability during sequence processing. First, some level of mismatch must be

358 allowed during the mapping of reads to the reference because there are single

359 nucleotide variants segregating in the population and there are somatic mutations in

360 cancer genomes. Second, the mappability score is assigned to the first nucleotide of

361 the k-mer that can be mapped; in reality what we really need is the average

362 mappability of all k-mers that overlap a site. Third, although the read length was

363 greater than 100bp, some shorter reads may have been used. Next generation

364 sequencing involves a number of biological processes, such as the polymerase chain

365 reactions in the pre-sequencing creation of libraries and the polymerization of

366 nucleotides during sequencing by synthesis, any one of which can result in

367 technology-specific sequencing artefacts (Quail et al., 2008; Nazarian et al., 2010), In

368 addition to the considerable post-sequencing processing, such as filtering and

369 mapping, which can also generate errors (Harismendy & Frazer, 2009; Minoche,

370 Dohm & Himmelbauer, 2011). Unfortunately it is not possible to say which of these

371 factors is most important.

372

373

374 We have fit two models to the data in which the density of SNVs varies across sites.

375 In the first we imagine that the variation is due to a single variable process and in the

376 second we imagine it is due to two processes, one of which is constant across sites

18

377   and one which is variable. We find that this second model fits the data much better

378   than the first model, although it can be formally rejected by a goodness-of-fit test. In

379   this second model we estimate the proportion of SNVs that are due to the two

380   processes and the level of variation. We estimate that approximately 2.8-4.3% of

381   SNVs are due to the second process and that this second process is highly variable

382   between sites, such that a few sites have a density of SNVs that is ten of thousands

383   higher than the average density. It is possible that the first process is mutation and the

384   second is sequencing error, but we cannot rule out the possibility that the second

385   process includes variation in the mutation rate as well. Studies of germ-line

386   (Hodgkinson & Eyre-Walker, 2011; Michaelson et al., 2012) and somatic

387   (Hodgkinson, Chen & Eyre-Walker, 2012; Woo & Li, 2012; Lawrence et al., 2013;

388   Liu, De & Michor, 2013; Polak et al., 2015) mutations have indicated that the

389   mutation rate varies between sites on a number of different scales. However,

390   indications are that the variation is probably fairly modest (Hodgkinson, Chen &

391   Eyre-Walker, 2012; Michaelson et al., 2012).

392

393   In conclusion it seems likely that many sites in somatic tissues that have experienced

394   recurrent SNVs are due to sequencing errors or artefacts of post-sequencing

395   processing and there seems to be little evidence of cryptic variation in the somatic

396   mutation rate. However, this not necessarily mean that such variation does not exist –

397   it would be extremely difficult to detect it given the high level of site-specific

398   sequencing error. As sequencing technology and processing pipelines improve in

399   accuracy, we would expect similar future analyses to be able to confidently estimate

400   the true underlying variation in the somatic mutation rate. Accompanied by the flow

19

401 of data from projects such as the 100k genomes project, it should soon be possible to

402 achieve per triplet mutation rate variation map for individual cancer types and not just

403 pooled across multiple cancers.

404

405

406 **References.**

407

408 Alexandrov LB., Nik-Zainal S., Wedge DC., Aparicio S a JR., Behjati S., Biankin A

409   V., Bignell GR., Bolli N., Borg A., Børresen-Dale A-L., Boyault S., Burkhardt

410   B., Butler AP., Caldas C., Davies HR., Desmedt C., Eils R., Eyfjörd JE., Foekens

411   J a., Greaves M., Hosoda F., Hutter B., Ilicic T., Imbeaud S., Imielinski M.,

412   Imielinsk M., Jäger N., Jones DTW., Jones D., Knappskog S., Kool M., Lakhani

413   SR., López-Otín C., Martin S., Munshi NC., Nakamura H., Northcott P a., Pajic

414   M., Papaemmanuil E., Paradiso A., Pearson J V., Puente XS., Raine K.,

415   Ramakrishna M., Richardson AL., Richter J., Rosenstiel P., Schlesner M.,

416   Schumacher TN., Span PN., Teague JW., Totoki Y., Tutt ANJ., Valdés-Mas R.,

417   van Buuren MM., van 't Veer L., Vincent-Salomon A., Waddell N., Yates LR.,

418   Zucman-Rossi J., Futreal PA., McDermott U., Lichter P., Meyerson M.,

419   Grimmond SM., Siebert R., Campo E., Shibata T., Pfister SM., Campbell PJ.,

420   Stratton MR. 2013. Signatures of mutational processes in human cancer. Nature

421   500:415–21. DOI: 10.1038/nature12477.

422 Benson G. 1999. Tandem repeats finder: A program to analyze DNA sequences.

423   Nucleic Acids Research 27:573–580. DOI: 10.1093/nar/27.2.573.

20

424   Bird AP. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic

425   Acids Research 8:1499–1504. DOI: 10.1093/nar/8.7.1499.

426   Bulmer M. 1986. Neighboring base effects on substitution rates in pseudogenes.

427   Molecular biology and evolution 3:322–329.

428   Cooper DN., Krawczak M. 1990. The mutational spectrum of single base-pair

429   substitutions causing human genetic disease: patterns and predictions. Human

430   Genetics 85:55–74. DOI: 10.1007/BF00276326.

431   Derrien T., Estell?? J., Sola SM., Knowles DG., Raineri E., Guig?? R., Ribeca P.

432   2012. Fast computation and applications of genome mappability. PLoS ONE 7.

433   DOI: 10.1371/journal.pone.0030377.

434   Eyre-Walker A., Eyre-Walker YC. 2014. How much of the variation in the mutation

435   rate along the human genome can be explained? G3 4:1667–70. DOI:

436   10.1534/g3.114.012849.

437   Flicek P., Amode MR., Barrell D., Beal K., Brent S., Carvalho-Silva D., Clapham P.,

438   Coates G., Fairley S., Fitzgerald S. 2011. Ensembl 2012. Nucleic acids

439   research:gkr991.

440   Francioli LC., Polak PP., Koren A., Menelaou A., Chun S., Renkens I., van Duijn

441   CM., Swertz M., Wijmenga C., van Ommen G., Slagboom PE., Boomsma DI.,

442   Ye K., Guryev V., Arndt PF., Kloosterman WP., de Bakker PIW., Sunyaev SR.

443   2015. Genome-wide patterns and properties of de novo mutations in humans.

444   Nature Genetics 47:822–826. DOI: 10.1038/ng.3292.

21

445   Fryxell KJ., Moon WJ. 2005. CpG mutation rates in the human genome are highly

446   dependent on local GC content. Molecular Biology and Evolution 22:650–658.

447   DOI: 10.1093/molbev/msi043.

448   Gojobori T., Li WH., Graur D. 1982. Patterns of nucleotide substitution in

449   pseudogenes and functional genes. Journal of Molecular Evolution 18:360–369.

450   DOI: 10.1007/BF01733904.Harismendy O., Frazer KA. 2009. Method for im

451 proving sequence coverage uniformity of targeted genomic intervals amplified by LR-

452   PCR using Illumina GA sequencing-by-synthesis technology. BioTechniques

453   46:229–231. DOI: 10.2144/000113082.

454   Hodgkinson A., Chen Y., Eyre-Walker A. 2012. The large-scale distribution of

455   somatic mutations in cancer genomes. Human Mutation 33:136–143. DOI:

456   10.1002/humu.21616.

457   Hodgkinson A., Eyre-Walker A. 2011. Variation in the mutation rate across

458   mammalian genomes. Nature Reviews Genetics 12:756–766. DOI:

459   10.1038/nrg3098.

460   Hodgkinson A., Ladoukakis E., Eyre-Walker A. 2009. Cryptic variation in the human

461   mutation rate. PLoS Biology 7:0226–0232. DOI: 10.1371/journal.pbio.1000027.

462   Huang FW., Hodis E., Xu MJ., Kryukov G V., Chin L., Garraway LA. 2013. Highly

463   Recurrent TERT Promoter Mutations in Human Melanoma. Science 339:957–

464   959. DOI: 10.1126/science.1229259.

465   Hwang DG., Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis

466   reveals varying neutral substitution patterns in mammalian evolution.

22

467   Proceedings of the National Academy of Sciences of the United States of

468   America 101:13994–14001. DOI: 10.1073/pnas.0404142101.

469 Karolchik D., Hinrichs AS., Furey TS., Roskin KM., Sugnet CW., Haussler D., Kent

470   WJ. 2004. The UCSC Table Browser data retrieval tool. Nucleic acids research

471   32:D493–D496. DOI: 10.1093/nar/gkh103.

472 Kong A., Frigge ML., Masson G., Besenbacher S., Sulem P., Magnusson G.,

473   Gudjonsson S a., Sigurdsson A., Jonasdottir AA., Jonasdottir AA., Wong WSW.,

474   Sigurdsson G., Walters GB., Steinberg S., Helgason H., Thorleifsson G.,

475   Gudbjartsson DF., Helgason A., Magnusson OT., Thorsteinsdottir U., Stefansson

476   K. 2012. Rate of de novo mutations and the importance of father's age to disease

477   risk. Nature 488:471–475. DOI: 10.1038/nature11396.

478 Lawrence MS., Stojanov P., Polak P., Kryukov G V., Cibulskis K., Sivachenko A.,

479   Carter SL., Stewart C., Mermel CH., Roberts S a., Kiezun A., Hammerman PS.,

480   McKenna A., Drier Y., Zou L., Ramos AH., Pugh TJ., Stransky N., Helman E.,

481   Kim J., Sougnez C., Ambrogio L., Nickerson E., Shefler E., Cortés ML., Auclair

482   D., Saksena G., Voet D., Noble M., DiCara D., Lin P., Lichtenstein L., Heiman

483   DI., Fennell T., Imielinski M., Hernandez B., Hodis E., Baca S., Dulak AM.,

484   Lohr J., Landau D-A., Wu CJ., Melendez-Zajgla J., Hidalgo-Miranda A., Koren

485   A., McCarroll S a., Mora J., Lee RS., Crompton B., Onofrio R., Parkin M.,

486   Winckler W., Ardlie K., Gabriel SB., Roberts CWM., Biegel J a., Stegmaier K.,

487   Bass AJ., Garraway L a., Meyerson M., Golub TR., Gordenin D a., Sunyaev S.,

488   Lander ES., Getz G. 2013. Mutational heterogeneity in cancer and the search for

489   new cancer-associated genes. Nature 499:214–8. DOI: 10.1038/nature12213.

490 Liu L., De S., Michor F. 2013. DNA replication timing and higher-order nuclear

23

491   organization determine single-nucleotide substitution patterns in cancer

492   genomes. Nature communications 4:1502. DOI: 10.1038/ncomms2502.

493 Lynch M. 2010. Evolution of the mutation rate. Trends in Genetics 26:345–352. DOI:

494   10.1016/j.tig.2010.05.003.

495 Martinocorena I., Campbell PJ. 2015. Somatic mutation in cancer and normal cells.

496   Science 349:1483–1489. DOI: 10.1126/science.aab4082.

497 Michaelson JJ., Shi Y., Gujral M., Zheng H., Malhotra D., Jin X., Jian M., Liu G.,

498   Greer D., Bhandari A., Wu W., Corominas R., Peoples Á., Koren A., Gore A.,

499   Kang S., Lin GN., Estabillo J., Gadomski T., Singh B., Zhang K., Akshoomoff

500   N., Corsello C., McCarroll S., Iakoucheva LM., Li Y., Wang J., Sebat J. 2012.

501   Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo

502   Germline Mutation. Cell 151:1431–1442. DOI:

503   http://dx.doi.org/10.1016/j.cell.2012.11.019.

504 Minoche AE., Dohm JC., Himmelbauer H. 2011. Evaluation of genomic high-

505   throughput sequencing data generated on Illumina HiSeq and Genome Analyzer

506   systems. Genome Biology 12:R112. DOI: 10.1186/gb-2011-12-11-r112.

507 Nachman MW., Crowell SL. 2000. Estimate of the mutation rate per nucleotide in

508   humans. Genetics 156:297–304. DOI: papers2://publication/uuid/E46268CF-

509   E7EF-4A7D-A85A-821D27B7F178.

510 Nazarian R., Shi H., Wang Q., Kong X., Koya RC., Lee H., Chen Z., Lee M-K., Attar

511   N., Sazegar H., Chodon T., Nelson SF., McArthur G., Sosman JA., Ribas A., Lo

512   RS. 2010. Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK

513   or N-RAS upregulation. Nature 468:973–7. DOI: 10.1038/nature09626.

24

514   Nelder JA., Mead R., Nelder BJA., Mead R. 1965. A Simplex Method for Function

515   Minimization. The Computer Journal 7:308–313. DOI: 10.1093/comjnl/7.4.308.

516   Polak P., Karlic R., Koren A., Thurman R., Sandstrom R., Lawrence MS., Reynolds

517   A., Rynes E., Vlahovicˇek K., Stamatoyannopoulos JA., Sunyaev SR. 2015. Cell-

518   of-origin chromatin organization shapes the mutational landscape of cancer.

519   Nature 518:360–364. DOI: 10.1038/nature14221.

520   Quail MA., Kozarewa I., Smith F., Scally A., Stephens PJ., Durbin R., Swerdlow H.,

521   Turner DJ. 2008. A large genome center's improvements to the Illumina

522   sequencing system. Nature methods 5:1005–10. DOI: 10.1038/nmeth.1270.

523   Schuster-Bockler B., Lehner B. 2012. Chromatin organization is a major

524   influence on regional mutation rates in human cancer cells. Nature 488:504–507.

525   DOI: 10.1038/nature11273.

526   Smith T., Ho G., Christodoulou J., Price EA., Onadim Z., Gauthier-Villars M.,

527   Dehainault C., Houdayer C., Parfait B., van Minkelen R., Lohman D., Eyre-

528   Walker A. 2016. Extensive Variation in the Mutation Rate Between and Within

529   Human Genes Associated with Mendelian Disease. Human mutation:n/a–n/a.

530   DOI: 10.1002/humu.22967.

531   R Core Team. 2016. R: A Language and Environment for Statistical Computing.

532   Treangen TJ., Salzberg SL. 2013. Repetitive DNA and next-generation

533   sequencing: computational challenges and solutions. Nat Rev Genet. 13:36–46.

534   DOI: 10.1038/nrg3117.Repetitive.

25

535   Woo YH., Li W-H. 2012. DNA replication timing and selection shape the landscape of

536   nucleotide variation in cancer genomes. Nature Communications 3:1004. DOI:

537   10.1038/ncomms1982.

538   Zhuang J., Wang J., Theurkauf W., Weng Z. 2014. TEMP: A computational method

539   for analyzing transposable element polymorphism in populations. Nucleic Acids

540   Research 42:6826–6838. DOI: 10.1093/nar/gku323.

541

542   **Supplementary table 1.**

543   Excess SNVs from liver cancers split between the two labs of origin. RK indicates SNVs from the

544   RIKEN lab and HX from the NCC. Significant heterogeneity of excess sites originating from different

545   labs was tested using fishers exact test (see methods).

546

547

548

549

550

551

552

553

554

555

556

557

558

559

560

561

locus	RK	HX	sum
chrX:56209339	6	0	6
chr10:96652829	6	0	6
chr10:96652827	6	0	6
chrX:56209340	5	0	5
chr5:85091859	5	0	5
chr5:1295228	0	5	5
chr9:121267366	4	0	4
chr8:119547627	4	0	4
chr19:22314552	1	2	3
chr14:95832895	1	2	3
chr9:16932821	2	1	3
chr7:27901228	2	1	3
chr4:162437670	2	1	3
chr3:164903710	2	1	3
chrY:4796240	3	0	3
chrX:84996701	3	0	3
chr7:11432162	3	0	3
chr7:11432157	3	0	3
chr3:174306603	3	0	3
chr2:49173787	3	0	3
chr2:139556678	3	0	3
chr19:8673262	3	0	3
chr1:190881448	3	0	3
chrX:79125571	0	3	3
chr6:78532352	0	3	3
chr5:97912191	0	3	3
chr4:190837614	0	3	3
chr19:44959650	0	3	3
chr15:73206445	0	3	3
chr14:74659965	0	3	3

26