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

Thomas C A Smith [Corresp., 1], **Antony M Carr** [2], **Adam C Eyre-Walker** [Corresp. 1]

[1] School of Life Sciences, University of Sussex, Brighton, East Sussex, United Kingdom

[2] Genome Damage and Stability Centre, University of Sussex, Brighton, East Sussex, United Kingdom

Corresponding Authors: Thomas C A Smith, Adam C Eyre-Walker
Email address: tom81_asp@hotmail.com, a.c.eyre-walker@sussex.ac.uk

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 between sites.

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

Thomas C. A. Smith[1]

Antony M. Carr[2]

Adam Eyre-Walker[1]

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

2. Genome Damage and Stability Centre, University of Sussex, Falmer, Sussex, BN1 9RQ, United Kingdom.

Corresponding Authors:

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

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

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

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

Abstract.

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 between sites.

48 **Introduction.**

49

50 There is currently huge interest in sequencing cancer genomes with a view to identifying the
51 mutations in somatic tissues that lead to cancer, the so called “driver” mutations. Driver
52 mutations are expected to cluster in particular genes or genomic regions, or to recur at particular
53 sites in the genome, because only a limited number of mutations can cause cancer. For example,
54 the driver mutations in the TERT1 promoter were identified because it had independently
55 occurred in multiple cancers (Huang et al., 2013). However, there are two other processes that
56 can potentially lead to the repeated occurrence of an apparent somatic mutation at a site. First, it
57 is known that the mutation rate varies across the genome at a number of different scales in both
58 the germ-line and soma (Hodgkinson & Eyre-Walker, 2011; Hodgkinson, Chen & Eyre-Walker,
59 2012; Michaelson et al., 2012; Francioli et al., 2015). Sites with recurrent SNVs could simply be
60 a consequence of sites with high rates of mutations. And second there is the potential for
61 sequencing error. Although, the average rate of sequencing error is thought to be quite low it is
62 evident that some types of sites, such as those in runs of nucleotides, are difficult to sequence
63 accurately. Furthermore, since the genome contains many similar sequences it can often be
64 difficult to map sequencing reads successfully (Treangen & Salzberg, 2013).

65

66 In the germ-line the density of point mutations varies at a number of different scales
67 (Hodgkinson & Eyre-Walker, 2011). At the mega-base scale the mutation varies by about 2-fold,
68 and ~50% of this variance can be explained by correlations with factors such as replication time,
69 recombination rate and distance from telomeres (as reviewed in (Hodgkinson & Eyre-Walker
70 2011)). However the greatest variance, reportedly up to ~30-fold, has been found at the single

nucleotide level (Hodgkinson, Chen & Eyre-Walker, 2012; Kong et al., 2012; Michaelson et al., 2012), whereby the nucleotide context, that is the identity of the bases immediately 5' and 3' of the mutated base, are highly influential on the rate of mutation (Gojobori, Li & Graur, 1982; Bulmer, 1986; Cooper & Krawczak, 1990; Nachman & Crowell, 2000; Hwang & Green, 2004). The most well known example is that of CpG hyper-mutation (Bird, 1980), which is thought to account ~20% of all mutations in the human genome (Fryxell & Moon, 2005). However there is also variation at the single nucleotide level that cannot be ascribed to the effects of neighbouring nucleotides; this has been termed cryptic variation in the mutation rate and is thought to account for at least as much variation in the mutation rate as does simple context (Hodgkinson, Ladoukakis & Eyre-Walker, 2009; Eyre-Walker & Eyre-Walker, 2014, Johnson & Hellman, 2011, Smith et al., 2016).

The somatic mutation rate is estimated to be at least an order of magnitude greater than that of the germ line (Lynch, 2010). It has been shown to vary between cancers (Lawrence et al. 2013) and different cancer types are known to vary in their relative contributions of different mutations to their overall mutational compositions (Alexandrov et al., 2013). For a review see (Martinocorena & Campbell, 2015). The aforementioned correlates of variation that are found in the germ line are also apparent in the soma (Hodgkinson, Chen & Eyre-Walker, 2012; Schuster-Bockler & Lehner, 2012; Lawrence et al., 2013; Liu, De & Michor, 2013), for example replication time correlates strongly with single nucleotide variant (SNV) density at the 1Mb base scale and can vary by up to 3-fold along the genome (Hodgkinson & Eyre-Walker, 2011; Woo & Li, 2012). However, as yet there has been no attempt to quantify the level of cryptic variation in the mutation rate at the single nucleotide level in the somatic genome. This is an important

property to understand; for example a site which experiences a recurrence of SNVs across many cancer genomes would be of interest as a potential driver of cancer (Lawrence et al., 2013), however, this site might simply be cryptically hypermutable (Hodgkinson, Ladoukakis & Eyre-Walker, 2009; Eyre-Walker & Eyre-Walker, 2014; Smith et al., 2016). Here we examine the distribution of recurrent SNVs taken from 507 whole genome sequences made publicly available by Alexandrov et al. (2013) to investigate the level of cryptic variation in the mutation rate for somatic tissues. We show that there is a large excess of sites that have been hit by recurrent SNVs. Since the density of these is greater in the non-coding, than the coding fraction of the genome, we conclude that most of them are unlikely to be drivers. We therefore investigate whether they are due to mutational heterogeneity or sequencing errors. In particular we investigate whether there might be cryptic variation in the mutation rate in cancer genomes. Unfortunately, the available evidence suggests that most sites with recurrent SNVs are likely to be due to sequencing error or errors in post-sequencing processing.

Methods.

Genome and data filtering.

The human genome (hg19/GRCh37) was masked to remove simple sequence repeats (SSR) as defined by Tandem Repeat Finder (Benson, 1999). The remaining regions were separated into three genomic fractions, consisting of 1,346,629,686 bp of non-coding transposable element DNA (TE), defined as LINEs, SINEs, LTRs and DNA transposons as identified by repeat masker (Smit et al. 1996), 1,322,985,768 bp of non-coding non-transposable element DNA

(NTE), and 119,806,141 bp of exonic non-transposable element DNA (EX) defined by Ensemble (Flicek et al., 2011). From the supplementary data of Alexandrov et al. (2013) we collated 3,382,737 single nucleotide variants (SNV), classified as "somatic-for-signature-analysis" (see (Alexandrov et al., 2013) for SNV filtering methods). These can be downloaded from ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl/. These came from 507 whole genome sequenced cancers and represent 10 different cancer types and were reduced to 3,299,881 SNVs when excluding SNVs in SSRs; 1,666,759 in TE and 1,535,069 in NTE and 98053 in EX.

Testing for mutation rate heterogeneity.

We were interested in whether some sites have more SNVs than expected by chance. Since the mutation rate is affected by the identity of the neighbouring nucleotides we need to control for those effects. To do this we separated each SNV into one of 64 categories based upon the triplet to which it was the central base. This was reduced to 32 triplets when accounting for base complementarity with the pyrimidine (C/T) taken as the central base. If the total number of triplets of type i (e.g. C**T**C in the non-TE fraction) is l_i and the number SNVs at that triplet is m_i then the expected number of sites hit x times can be calculated using a Poisson distribution:

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

where $\mu_i = m_i/l_i$ is the mean number of SNVs per site, The expected number of sites with x SNVs across all triplets was calculated by summing the values of $P_i(x)$. Whether the observed distribution deviated from the expected was tested using a chisquare test.

Model fitting

As well as testing whether there was significant hetereogeneity we were also interested in quantifying the level of variation. We fit two basic models. In the first we allowed the density of SNVs to follow a gamma distribution. Let the expected density of SNVs at a site be $\mu\alpha$ where μ is the mean density of SNVs for a particular triplet and α is the deviation from this mean which is gamma distributed, parameterised such that the gamma has a mean of one. Under this model the expected number of sites with x SNVs is

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

In a second model we imagine that the production of SNVs depends upon two processes, one of which is constant across sites, and one which varies across sites with the rate drawn from a gamma distribution. Let the proportion of SNVs due to the first process be ε. Under this model the expected number of sites with x SNVs is

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

Given the expected number of sites, the likelihood of observing $\hat{P}(x)$ sites with x SNVs is itself Poisson distributed

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

These likelihoods can be multiplied across triplets to obtain the overall likelihood. We estimated the maximum likelihood values of the model parameters using the Maximize function of Mathematica which implements the Nelder-Mead algorithm (Nelder et al., 1965).

.

Privacy analysis

To investigate whether the SNVs at some sites tended to be produced by a particular research group we took all sites with 3 or more SNVs from the same cancer type and then performed Fishers exact test on a 2 x 30 matrix using the the R stats package, version 3.2.4 (R Core Team, 2016).

Mappability.

Each nucleotide in genome was assigned a mappability score for uniqueness, as determined by the Mappability track (Derrien et al., 2012) downloaded from the UCSC table browser at http://genome.ucsc.edu/ (Karolchik et al., 2004). This feature assigns a value of 1 to unique *k*-mer sequences in the genome, 0.5 to those that occur twice, 0.33 to those that occur thrice etc. This is computed for every base in the human genome with the value being assigned to the first position of the k-mer. We used *k*-mers of 100 and 20 bases.

Results.

The distribution of recurrent SNVs.

If there is no variation in the density of single nucleotide variants (SNVs) then we should find them to be distributed randomly across the genome. To investigate whether this was the case we

187 calculated the expected number of sites with 1,2,3…etc SNVs, taking into account the fact that
188 some triplets have higher mutation rates than others. We found that there are some sites that have
189 7 SNVs whereas we expect very few sites to have more than 3 SNVs – the difference is highly
190 significant using the Chi-square goodness of fit test ($p < 0.0001$) for both the whole genome
191 (Total) and when separating the genome into non-coding transposable elements (TE), non-coding
192 non-transposable elements and (NTE) and exons (EX) (Table 1). We refer to sites with 3 or
193 more SNVs as excess sites. In total we observed 1187 excess sites (Table 1) with the density of
194 excess sites in TE being 3.9 and 3.4 fold greater than in NTE and EX respectively. The
195 probability of this level of SNV recurrence by chance alone is so low (Chi-squared goodness of
196 fit test, $p > 0.0001$) that these excess sites must either be (i) drivers, (ii) the result of mutation
197 rate heterogeneity across the genome or, (iii) the consequence of next generation sequencing
198 (NGS) pipeline errors.

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

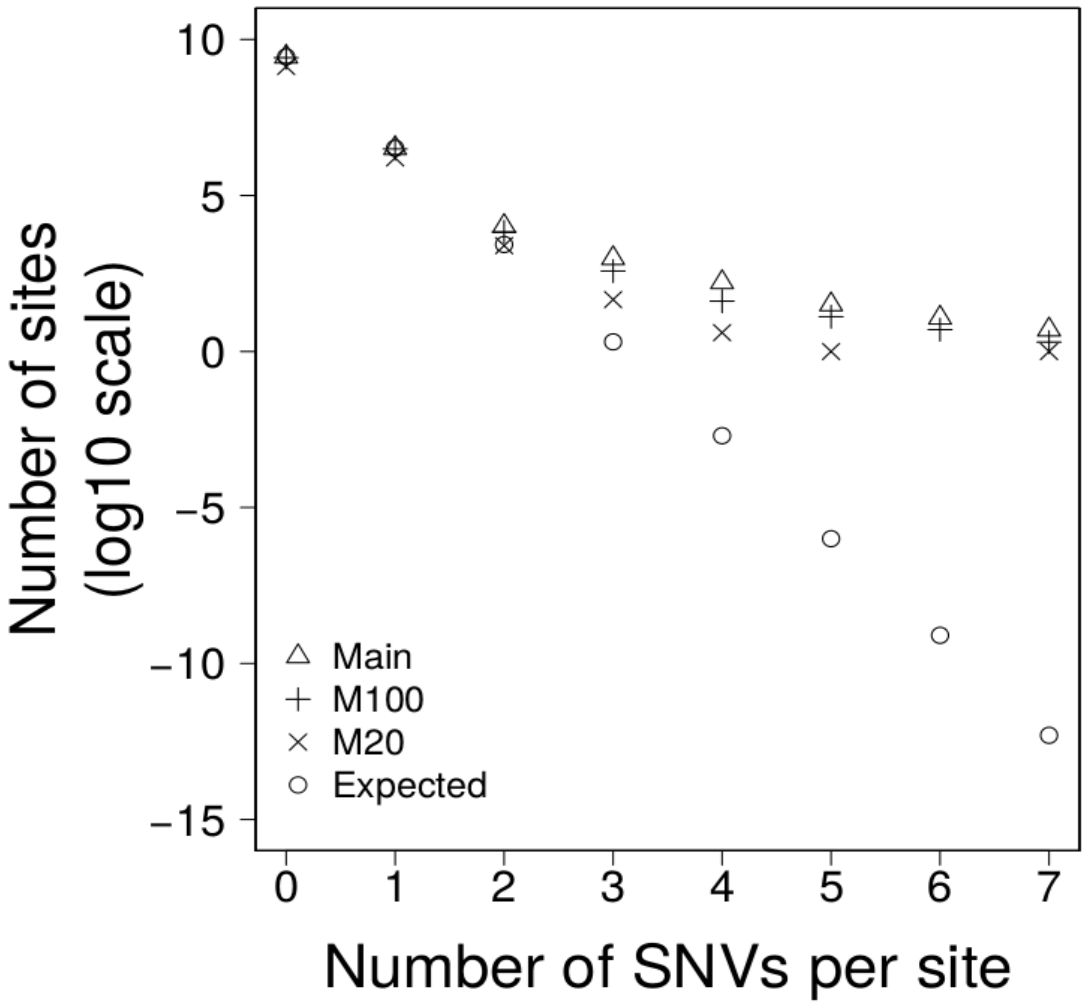
Site Type	0 hits	1 hit	2 hits	3 hits	4 hits	5 hits	6 hits	7 hits
Non-Exon TE obs (TE)	3.89E+8	4.81E+5	741	9	0	0	0	0
Non-Exon TE exp (TE)	3.89E+8	4.81E+5	417	0.34	3E-4	2E-7	2E-10	1E-13
Non-Exon Non-TE obs (NTE)	8.92E+8	1.06E+6	1621	31	4	1	0	1
Non-Exon Non-TE exp (NTE)	8.92E+8	1.06E+6	868	0.65	5E-4	3E-7	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.

It seems unlikely that the majority of the excess sites are due to drivers since the density of excess sites is higher in the TE and NTE parts of the genome than in EX (Table 1A). Furthermore, to date only one intergenic driver of cancer – an activating C>T mutation in the *TERT* promoter (Huang et al. 2013) at chr5:1,295,228 – has been confirmed, and although this is included in the excess sites with 7 SNVs, the remaining 1186 excess sites are unlikely to be under such selection. It therefore seems likely that the excess sites are either due to mutation rate variation or problems with sequencing.

Excess sites are enriched in non-unique sequences.

The human genome contains many duplicated sequences particularly within transposable elements, and these pose challenges for accurate alignment of the short ~100bp reads produced from NGS (Zhuang et al., 2014). If the excess sites were the result of NGS mapping errors then we might expect them to occur in regions of the genome that were hard to align. Using the mappability scores (Derrien et al., 2012) we excluded all bases that were not uniquely mappable at 100bp; this should give an overall indication of how easy it is to map reads to the region. This only reduced the interrogable genome by 6%, but the number of excess sites was reduced by 64% (Table 1B), demonstrating that a large proportion of the excess sites were in duplicated sequences and therefore likely originate from mapping errors. However, even with this large reduction in excess sites we still observed many excess sites far greater than chance expectation (Chi-squared goodness of fit test, $p < 0.0001$) (Table 1B & Figure 1).



223 Figure 1. The number of site with 0-7 SNVs per sites for: **Main** = all data, **M100** = sites that are uniquely mappable
224 at 100 base-pairs, **M20** = sites that are uniquely mappable at base-pairs and, **Expected** is the expected number of
225 SNVs per site drawn from a poisson distribution using all data.

226

227 The SNVs in this data were all called from >100bp reads. If the excess sites were errors of read
228 mapping, they should not be affected by the uniqueness of shorter sequences (i.e. there is no
229 reason why 100bp sequences that map uniquely to the genome should be mis-mapped if it
230 contains a non-unique 20bp sequence), however if the SNVs were the product of a biological
231 process that was more prevalent in non-unique or repetitive sequences, then we might expect to

see a reduction of excess sites when we exclude all bases that do not map uniquely at 20bp. When we excluded all bases that were not unique at 20bp we found that the interrogable genome was reduced by 52% and the excess sites were reduced by 96% (Table 1C & Figure 1). It is worth noting that, due to their proliferative nature throughout the genome, this reduction disproportionately affects TEs where the interrogable genome is reduced by 71% and the excess sites by >99%. This would suggest that the excess sites existing in sequences that were unique at 100bp but not unique at 20bp may represent some biological process and not error. Furthermore, th*e TE*RT promoter, whose recurrence is the result of positive selection, and is therefore the only excess site that that we can confidently say is not a product of error, remains in this most conservative of these analyses. Despite this large reduction in excess sites, significant heterogeneity still remains; the probability of observing the 52 excess sites in the part of the genome uniquely mappable at 20 bases is still extremely low (Chi-squared goodness of fit test, $p < 0.0001$).

One other potential problem with mapping reads to non-unique sequences occurs when a segmental duplication has been collapsed in the assembly of the reference genome; i.e. reads from two different locations are mapped to the same locus in the reference. Differences between the duplications will appear as SNVs. If this was the case we would expect to see an increase in escess site read coverage of ~2-fold or greater. To investigate whether this could be a problem in our data we compared the read coverage for excess sites and non-excess sites, which nevertheless had an SNV, in the one set of cancer genomes for which we had this information - the liver cancers sequenced by the RIKEN group. However, we found that the median read coverage for the excess sites (n=15) was actually lower than for non-excess sites (n=224602) (28 and 33 reads respectively; Mann-Whitney U test, $p = 0.043$)

256

257 *Privacy of mutations.*

258 To further investigate the origin of excess sites we exploited the fact that some types of cancer
259 were sequenced by different laboratories using different technologies and NGS pipelines. If the
260 SNVs at excess sites found in a particular cancer are due to hypermutable sites then we would
261 expect them to be randomly distributed across research groups (i.e. all research groups should
262 identify the same hypermutable sites). If however the SNVs at excess sites are due to error then
263 we might expect them to be heterogeneously distributed across research groups (i.e. the calling
264 of recurrent false positive SNVs should be systematic of individual research group NGS
265 pipelines). The liver cancers, which were all virus associated hepatocellular carcinomas, , were
266 sequenced by two different groups; 66 from the RIKEN group using the Illumina Genome
267 Analyser (https://dcc.icgc.org/projects/LIRI-JP) and 22 from the National Cancer Centre in Japan
268 using the IIlumina HiSeq platform (https://dcc.icgc.org/projects/LINC-JP). We found that the
269 excess SNVs were heterogeneously distributed amongst research groups (Fisher's exact test, P =
270 $4x10^{-6}$) suggesting that the 30 excess sites from liver cancers were predominantly errors
271 (Supplementary Table 1).

272

273 *Parameter estimation*

274 To gauge how much variation there is in the density of SNVs across the genome we fit two
275 models to the data using maximum likelihood. In model 1 we allowed the density of SNVs to
276 vary between sites according to a gamma distribution, estimating the shape parameter, and hence
277 the amount of variation there was between sites. We fitted two versions of this model. In the first
278 version, 1a, we constrained the model such that the mean SNV density, shape parameter, and
279 hence the level of variation, was the same for all triplets. In the second version, 1b, we allowed

the mean SNV density and shape parameter to vary between triplets. The second of these models fits the data significantly better than the first according to a likelihood ratio test suggesting that the level of variation differs between triplets (Table 2). However, a goodness of fit test, comparing the number of sites predicted to have 1, 2, 3…etc SNVs per site to the observed data, suggests the model fits the data poorly. We therefore fit a second pair of models in which we allowed the rate of SNVs to be due to two processes. The first process, is constant across sites whereas the second process is variable and drawn from a gamma distribution. There are two parameters in the model, the proportion of SNVs at a site produced by the first process and the level of variation in the second process. This model might represent a situation where the rate of mutation is constant across sites but the rate of sequencing error is variable. As with the first model we fit two versions of this model; in Model 2a we constrained the model such that the parameters of the two processes were the same for all triplets. In Model 2b they were allowed to vary between triplets. Both models 2a and 2b fit the data significantly better than models 1a and 1b, and of this second pair of models, model 2b, which allows the parameters to vary between triplets fits the data significantly better than model 2a, in which the parameters are shared across triplets (Table 2). The best fitting model is therefore one in which we have two processes contributing to the production of SNVs, one that is constant across sites, although it differs between triplets, and one which is variable across sites. Although, we can formally reject this model using a goodness-of-fit test (Chi-square $p < 0.0001$), because we have so much data, it is clear that the model fits the data fairly well (Figure 2). Under this model we estimate that approximately 4.1%, 2.8% and 4.3% of SNVs are due to the process that varies across sites in the TE and NTE, and EX sequences respectively. However, the variation in the density between sites due to the variable process is extremely large. The median shape parameters are 0.0013,

0.0011 and 0.00075 for the TE and NTE, and EX sequences respectively. Under a gamma distribution with a shape parameter of 0.0004 we would expect more than 99% of sites to have no SNVs generated by this variable process, but some sites to have a density of SNVs that is 30,000-fold above the average rate.

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*

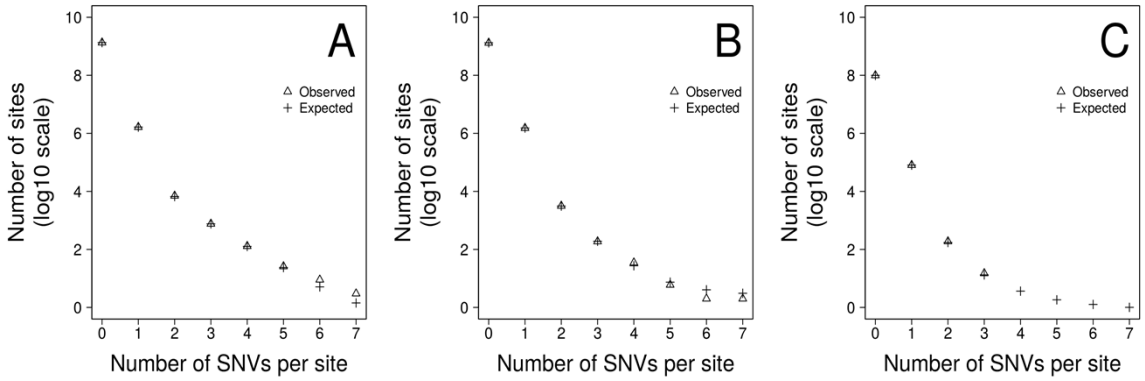
Non-Exon Non-TE (NTE)

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*

Table 2. The fit of 4 models to the observed distribution of recurrent SNVs in the three different genomic fractions A) TE, B) NTE and C) EX. **N** = number of parameters. *Italics* indicate the best fit as determined by a liklihood ratio test.



315

316 Figure 2. The fit of the observed recurrent SNV distribution to expected distribution under the favoured model, 2b,
317 for A) TE, B) NTE and C) EX genomic fractions.

318

319

320

321 **Discussion.**

322

323 Through our analysis of ~3 million SNVs from whole cancer genomes we have shown that there
324 are many sites at which there is a significant excess of SNVs. The majority of these are unlikely
325 to be drivers because the density of sites with an excess of SNVs is greater in the non-coding part
326 of the genome than in the exons. It therefore seems likely that the majority of the excess sites are
327 either due to hypermutation or problems with sequencing or the processing of the sequences.
328 Several lines of evidence point to sequencing problems being the chief culprit. First, many of the
329 excess sites disappear when regions of the genome with low mappability are removed. Second,
330 SNVs at a particular excess site tend to be found within the sequences from a particular
331 laboratory; for example, site 85,091,895 on chromosome 5 has 5 SNVs in liver cancers, but all of
332 these are found in the sequences from RIKEN not the sequences from the NCC. It is possbile that

this could be caused by biological differences between the cohorts, either environmentally induced or endogenous genetic variation, such as that seen between European and African populations and the differing frequency of 5'-TCC-3' > 5'-TTC-3' mutation (Harris, 2015). However the level of site and cohort specific, but cryptic, variation required would be huge and we have very little evidence to support such a hypothesis. Third, the level of variation in the density of SNVs is much greater than has been observed or suggested for variation in the mutation rate (Hodgkinson & Eyre-Walker, 2011; Kong et al., 2012; Michaelson et al., 2012) though see a recent analysis of de novo germ-line mutations which suggests there could be extreme mutational heterogeneity (Smith et al., 2016); some sites are estimated to have rates of SNV production that are tens of thousands of times faster than the genomic average.

Only one line of evidence suggests that there might also be substantial variation in the mutation rate as well as variation in the error rate. When we eliminate sites that are not uniquely mappable at 20bp we find a great reduction in the number of excess sites relative to the case when we remove sites that are not uniquely mappable at 100bp, and yet the read length is greater than 100bp in the data that we have used. This might suggest that there are some repetitive sequences that are prone to a process of hyper-mutation. However, it might also be that mappability at 100bp is not a good guide to mappability during sequence processing. First, some level of mismatch must be allowed during the mapping of reads to the reference because there are single nucleotide variants segregating in the population and there are somatic mutations in cancer genomes. Second, the mappability score is assigned to the first nucleotide of the *k*-mer that can be mapped.. Third, although the read length was greater than 100bp, some shorter reads may have been used. Next generation sequencing involves a number of biological processes, such as

the polymerase chain reactions in the pre-sequencing creation of libraries and the polymerization of nucleotides during sequencing by synthesis, any one of which can result in technology-specific sequencing artefacts (Quail et al., 2008; Nazarian et al., 2010), In addition to the considerable post-sequencing processing, such as filtering and mapping, which can also generate errors (Harismendy & Frazer, 2009; Minoche, Dohm & Himmelbauer, 2011). Unfortunately it is not possible to say which of these factors is most important.

We have fit two models to the data in which the density of SNVs varies across sites. In the first we imagine that the variation is due to a single variable process and in the second we imagine it is due to two processes, one of which is constant across sites and one which is variable. We find that this second model fits the data much better than the first model, although it can be formally rejected by a goodness-of-fit test. In this second model we estimate the proportion of SNVs that are due to the two processes and the level of variation. We estimate that approximately 2.8-4.3% of SNVs are due to the second process and that this second process is highly variable between sites, such that a few sites have a density of SNVs that is ten of thousands higher than the average density. It is possible that the first process is mutation and the second is sequencing error, but we cannot rule out the possibility that the second process includes variation in the mutation rate as well. Studies of germ-line (Hodgkinson & Eyre-Walker, 2011; Michaelson et al., 2012) and somatic (Hodgkinson, Chen & Eyre-Walker, 2012; Woo & Li, 2012; Lawrence et al., 2013; Liu, De & Michor, 2013; Polak et al., 2015) mutations have indicated that the mutation rate varies between sites on a number of different scales. However, indications are that the

variation is probably fairly modest (Hodgkinson, Chen & Eyre-Walker, 2012; Michaelson et al., 2012).

A model including two processes fits the data well (figure 2). However, we can reject this model in a goodness-of-fit test, because we have a huge amount of data. Possible reasons for the less than perfect fit include large scale variation in the mutation rate (Hodgkinson & Eyre-Walker, 2011; Schuster-Bockler & Lehner, 2012; Makova & Hardison, 2015) and multi-nucleotide-mutations (MNMs) (Rosenfeld, Malhotra & Lencz, 2010; Schrider, Hourmozdi & Hahn, 2011; Harris & Nielsen, 2014); the latter represent ~2% of all human single nucleotide polymorphisms (SNPs).

In conclusion it seems likely that many sites in somatic tissues that have experienced recurrent SNVs are due to sequencing errors or artefacts of post-sequencing processing and there seems to be little evidence of cryptic variation in the somatic mutation rate. However, this not necessarily mean that such variation does not exist – it would be extremely difficult to detect it given the high level of site-specific sequencing error. As sequencing technology and processing pipelines improve in accuracy, we would expect similar future analyses to be able to confidently estimate the true underlying variation in the somatic mutation rate. Accompanied by the flow of data from projects such as the 100k genomes project, it should soon be possible to achieve per triplet mutation rate variation map for individual cancer types and not just pooled across multiple cancers.

References.

Alexandrov LB., Nik-Zainal S., Wedge DC., Aparicio S a JR., Behjati S., Biankin A V., Bignell GR., Bolli N., Borg A., Børresen-Dale A-L., Boyault S., Burkhardt B., Butler AP., Caldas C., Davies HR., Desmedt C., Eils R., Eyfjörd JE., Foekens J a., Greaves M., Hosoda F., Hutter B., Ilicic T., Imbeaud S., Imielinski M., Imielinsk M., Jäger N., Jones DTW., Jones D., Knappskog S., Kool M., Lakhani SR., López-Otín C., Martin S., Munshi NC., Nakamura H., Northcott P a., Pajic M., Papaemmanuil E., Paradiso A., Pearson J V., Puente XS., Raine K., Ramakrishna M., Richardson AL., Richter J., Rosenstiel P., Schlesner M., Schumacher TN., Span PN., Teague JW., Totoki Y., Tutt ANJ., Valdés-Mas R., van Buuren MM., van 't Veer L., Vincent-Salomon A., Waddell N., Yates LR., Zucman-Rossi J., Futreal PA., McDermott U., Lichter P., Meyerson M., Grimmond SM., Siebert R., Campo E., Shibata T., Pfister SM., Campbell PJ., Stratton MR. 2013. Signatures of mutational processes in human cancer. Nature 500:415–21. DOI: 10.1038/nature12477.

Benson G. 1999. Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Research 27:573–580. DOI: 10.1093/nar/27.2.573.

Bird AP. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Research 8:1499–1504. DOI: 10.1093/nar/8.7.1499.

Bulmer M. 1986. Neighboring base effects on substitution rates in pseudogenes. Molecular biology and evolution 3:322–329.

Cooper DN., Krawczak M. 1990. The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions. Human Genetics 85:55–74. DOI: 10.1007/BF00276326.

Derrien T., Estell?? J., Sola SM., Knowles DG., Raineri E., Guig?? R., Ribeca P. 2012. Fast computation and applications of genome mappability. PLoS ONE 7. DOI: 10.1371/journal.pone.0030377.

Eyre-Walker A., Eyre-Walker YC. 2014. How much of the variation in the mutation rate along the human genome can be explained? G3 4:1667–70. DOI: 10.1534/g3.114.012849.

Flicek P., Amode MR., Barrell D., Beal K., Brent S., Carvalho-Silva D., Clapham P., Coates G., Fairley S., Fitzgerald S. 2011. Ensembl 2012. Nucleic acids research:gkr991.

Francioli LC., Polak PP., Koren A., Menelaou A., Chun S., Renkens I., van Duijn CM., Swertz M., Wijmenga C., van Ommen G., Slagboom PE., Boomsma DI., Ye K., Guryev V., Arndt PF., Kloosterman WP., de Bakker PIW., Sunyaev SR. 2015. Genome-wide patterns and properties of de novo mutations in humans. Nature Genetics 47:822–826. DOI: 10.1038/ng.3292.

Fryxell KJ., Moon WJ. 2005. CpG mutation rates in the human genome are highly dependent on local GC content. Molecular Biology and Evolution 22:650–658. DOI: 10.1093/molbev/msi043.

Gojobori T., Li WH., Graur D. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. Journal of Molecular Evolution 18:360–369. DOI: 10.1007/BF01733904.Harismendy O., Frazer KA. 2009. Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology. BioTechniques 46:229–231. DOI: 10.2144/000113082.

445 Harris K. 2015. Evidence for recent, population-specific evolution of the human mutation rate.
446 *Proceedings of the National Academy of Sciences of the United States of America*
447 112:3439–3444. DOI: 10.1073/pnas.1418652112.

448 Harris K., Nielsen R. 2014. Error-prone polymerase activity causes multinucleotide mutations in
449 humans. *Genome Research* 24:1445–1454. DOI: 10.1101/gr.170696.113.

450 Hodgkinson A., Chen Y., Eyre-Walker A. 2012. The large-scale distribution of somatic
451 mutations in cancer genomes. Human Mutation 33:136–143. DOI: 10.1002/humu.21616.

452 Hodgkinson A., Eyre-Walker A. 2011. Variation in the mutation rate across mammalian
453 genomes. Nature Reviews Genetics 12:756–766. DOI: 10.1038/nrg3098.

454 Hodgkinson A., Ladoukakis E., Eyre-Walker A. 2009. Cryptic variation in the human mutation
455 rate. PLoS Biology 7:0226–0232. DOI: 10.1371/journal.pbio.1000027.

456 Huang FW., Hodis E., Xu MJ., Kryukov G V., Chin L., Garraway LA. 2013. Highly Recurrent
457 TERT Promoter Mutations in Human Melanoma. Science 339:957–959. DOI:
458 10.1126/science.1229259.

459 Hwang DG., Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals
460 varying neutral substitution patterns in mammalian evolution. Proceedings of the National
461 Academy of Sciences of the United States of America 101:13994–14001. DOI:
462 10.1073/pnas.0404142101.

463 Johnson PLF., Hellmann I. 2011. Mutation rate distribution inferred from coincident SNPs and
464 coincident substitutions. *Genome Biology and Evolution* 3:842–850. DOI:
465 10.1093/gbe/evr044.

466 Karolchik D., Hinrichs AS., Furey TS., Roskin KM., Sugnet CW., Haussler D., Kent WJ. 2004.
467 The UCSC Table Browser data retrieval tool. Nucleic acids research 32:D493–D496. DOI:
468 10.1093/nar/gkh103.

469 Kong A., Frigge ML., Masson G., Besenbacher S., Sulem P., Magnusson G., Gudjonsson S a.,
470 Sigurdsson A., Jonasdottir AA., Jonasdottir AA., Wong WSW., Sigurdsson G., Walters
471 GB., Steinberg S., Helgason H., Thorleifsson G., Gudbjartsson DF., Helgason A.,
472 Magnusson OT., Thorsteinsdottir U., Stefansson K. 2012. Rate of de novo mutations and
473 the importance of father's age to disease risk. Nature 488:471–475. DOI:
474 10.1038/nature11396.

475 Lawrence MS., Stojanov P., Polak P., Kryukov G V., Cibulskis K., Sivachenko A., Carter SL.,
476 Stewart C., Mermel CH., Roberts S a., Kiezun A., Hammerman PS., McKenna A., Drier Y.,
477 Zou L., Ramos AH., Pugh TJ., Stransky N., Helman E., Kim J., Sougnez C., Ambrogio L.,
478 Nickerson E., Shefler E., Cortés ML., Auclair D., Saksena G., Voet D., Noble M., DiCara
479 D., Lin P., Lichtenstein L., Heiman DI., Fennell T., Imielinski M., Hernandez B., Hodis E.,
480 Baca S., Dulak AM., Lohr J., Landau D-A., Wu CJ., Melendez-Zajgla J., Hidalgo-Miranda
481 A., Koren A., McCarroll S a., Mora J., Lee RS., Crompton B., Onofrio R., Parkin M.,
482 Winckler W., Ardlie K., Gabriel SB., Roberts CWM., Biegel J a., Stegmaier K., Bass AJ.,
483 Garraway L a., Meyerson M., Golub TR., Gordenin D a., Sunyaev S., Lander ES., Getz G.
484 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes.
485 Nature 499:214–8. DOI: 10.1038/nature12213.

486 Liu L., De S., Michor F. 2013. DNA replication timing and higher-order nuclear
487 organization determine single-nucleotide substitution patterns in cancer genomes. Nature
488 communications 4:1502. DOI: 10.1038/ncomms2502.

Lynch M. 2010. Evolution of the mutation rate. Trends in Genetics 26:345–352. DOI: 10.1016/j.tig.2010.05.003.

Makova KD., Hardison RC. 2015. The effects of chromatin organization on variation in mutation rates in the genome. *Nature Reviews Genetics* advance on:213–223. DOI: 10.1038/nrg3890.Martinocorena I., Campbell PJ. 2015. Somatic mutation in cancer and normal cells. Science 349:1483–1489. DOI: 10.1126/science.aab4082.

Michaelson JJ., Shi Y., Gujral M., Zheng H., Malhotra D., Jin X., Jian M., Liu G., Greer D., Bhandari A., Wu W., Corominas R., Peoples Á., Koren A., Gore A., Kang S., Lin GN., Estabillo J., Gadomski T., Singh B., Zhang K., Akshoomoff N., Corsello C., McCarroll S., Iakoucheva LM., Li Y., Wang J., Sebat J. 2012. Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation. Cell 151:1431–1442. DOI: http://dx.doi.org/10.1016/j.cell.2012.11.019.

Minoche AE., Dohm JC., Himmelbauer H. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. Genome Biology 12:R112. DOI: 10.1186/gb-2011-12-11-r112.

Nachman MW., Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. Genetics 156:297–304. DOI: papers2://publication/uuid/E46268CF-E7EF-4A7D-A85A-821D27B7F178.

Nazarian R., Shi H., Wang Q., Kong X., Koya RC., Lee H., Chen Z., Lee M-K., Attar N., Sazegar H., Chodon T., Nelson SF., McArthur G., Sosman JA., Ribas A., Lo RS. 2010. Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. Nature 468:973–7. DOI: 10.1038/nature09626.

511 Nelder JA., Mead R., Nelder BJA., Mead R. 1965. A Simplex Method for Function
512 Minimization. The Computer Journal 7:308–313. DOI: 10.1093/comjnl/7.4.308.

513 Polak P., Karlic R., Koren A., Thurman R., Sandstrom R., Lawrence MS., Reynolds A., Rynes
514 E., Vlahovicˇek K., Stamatoyannopoulos JA., Sunyaev SR. 2015. Cell-of-origin chromatin
515 organization shapes the mutational landscape of cancer. Nature 518:360–364. DOI:
516 10.1038/nature14221.

517 Quail MA., Kozarewa I., Smith F., Scally A., Stephens PJ., Durbin R., Swerdlow H., Turner DJ.
518 2008. A large genome center's improvements to the Illumina sequencing system. Nature
519 methods 5:1005–10. DOI: 10.1038/nmeth.1270.
520 Schuster-Bockler B., Lehner B. 2012. Chromatin organization is a major influence on
521 regional mutation rates in human cancer cells. Nature 488:504–507. DOI:
522 10.1038/nature11273.

523 Schuster-Bockler B., Lehner B. 2012. Chromatin organization is a major influence on regional
524 mutation rates in human cancer cells. *Nature* 488:504–507. DOI: 10.1038/nature11273.

525 Smith T., Ho G., Christodoulou J., Price EA., Onadim Z., Gauthier-Villars M., Dehainault C.,
526 Houdayer C., Parfait B., van Minkelen R., Lohman D., Eyre-Walker A. 2016. Extensive
527 Variation in the Mutation Rate Between and Within Human Genes Associated with
528 Mendelian Disease. Human mutation:n/a–n/a. DOI: 10.1002/humu.22967.

529 R Core Team. 2016. R: A Language and Environment for Statistical Computing.
530 Treangen TJ., Salzberg SL. 2013. Repetitive DNA and next-generation sequencing:
531 computational challenges and solutions. Nat Rev Genet. 13:36–46. DOI:
532 10.1038/nrg3117.Repetitive.

Rosenfeld JA., Malhotra AK., Lencz T. 2010. Novel multi-nucleotide polymorphisms in the human genome characterized by whole genome and exome sequencing. *Nucleic Acids Research* 38:6102–6111. DOI: 10.1093/nar/gkq408.

Schrider DR., Hourmozdi JN., Hahn MW. 2011. Pervasive multinucleotide mutational events in eukaryotes. *Current Biology* 21:1051–1054. DOI: 10.1016/j.cub.2011.05.013.

Woo YH., Li W-H. 2012. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. Nature Communications 3:1004. DOI: 10.1038/ncomms1982.

Zhuang J., Wang J., Theurkauf W., Weng Z. 2014. TEMP: A computational method for analyzing transposable element polymorphism in populations. Nucleic Acids Research 42:6826–6838. DOI: 10.1093/nar/gku323.

Supplementary table 1.

Excess SNVs from liver cancers split between the two labs of origin. RK indicates SNVs from the RIKEN lab and HX from the NCC. Significant heterogeneity of excess sites originating from different labs was tested using fishers exact test (see methods).

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