Signatures of neutral tumor evolution are conserved despite the complexity of cancer genomics data

Marc J. Williams[1-3], Benjamin Werner[4], Timon Heide[4], Chris P. Barnes[2,5], Trevor A. Graham[1,*] & Andrea Sottoriva[4,*]

[1] Evolution and Cancer Laboratory, Barts Cancer Institute, Queen Marry University of London, London, UK.
[2] Department of Cell and Developmental Biology, University College London, London, UK.
[3] Centre for Mathematics and Physics in the Life Sciences and Experimental Biology (CoMPLEX), University College London, London, UK.
[4] Evolutionary Genomics & Modelling Lab, Centre for Evolution and Cancer, The Institute of Cancer Research, London, UK.
[5] Department of Genetics, Evolution and Environment, University College London, London, UK.

[*] Correspondence should be addressed to T.A.G. (t.graham@qmul.ac.uk) or A.S. (andrea.sottoriva@icr.ac.uk)

Balapariya and De question the applicability of the power-law neutral evolution model to adequately describe the pattern of subclonal somatic mutations in bulk cancer sequencing data.

The authors' letter focuses on the issues of the inherent noise in next generation sequencing data, whereby random sampling of alleles, PCR amplification during library preparation, limited depth sequencing, and subclonal copy number changes may cause considerable noise in variant allele frequency (VAF) measurement. The authors suggest that these errors lead to VAF measurements that, due to over-dispersion, follow a beta-binomial and not a binomial distribution. We thank Balapariya and De for the insightful comments and address their point in the response below.

The issue of VAF measurement accuracy is a very important point and something that concerned us in our original Williams, Werner *et al.* 2016[1] study. For this reason, we provided extensive error propagation analysis in our original manuscript to identify the inherent biases that affect VAF estimation (see Williams, Werner *et al.* 2016, Online Methods, equations 12-14). We aimed at starting from the analytical form of neutral evolution (Eq. 7 in the main manuscript) as the expected signal (S) and adding the different sources of noise (N), such as purity and allele sampling during library prep, to generate the expected pattern S+N reported by the data. Our results demonstrate that the signature of neutral evolution is detectable with current moderate sequencing depth ($\geq$75x, see Online Methods and Supplementary Figure 10), and we fully acknowledged that the signature of neutral evolution vs selection cannot be reasonably extracted (or rejected) from lower depth datasets.

In an effort to address Balapariya and De's critics, we have tested the ability of our model to recover neutral evolutionary dynamics in the presence of beta-binomially distributed noise and found no significant differences with respect to the binomial noise we used in our original manuscript, although with very high dispersion (ρ=0.1) a degree of difference is appreciable (Figure 1A). Moreover, we estimated the degree of over-dispersion in the data we analysed in Williams, Werner *et al.* 2016 by fitting a beta-binomial model to the clonal cluster using Markov chain Monte Carlo (MCMC) inference. In both the 100X whole-genome gastric cancer (Wang *et al.* 2014[2]) and whole-exome colon cancer (Sottoriva *et al.* 2015[3]) we estimated the dispersion parameter ρ to be less than 0.005 (Figure 1B and C respectively), notably 10X times lower than what postulated by Balapariya and De in their Figures 1C,D). Given that as $\rho \to 0$, the beta-binomial distribution converges to a binomial distribution, we argue that using a binomial distribution to model noise in sequencing data was appropriate in our original analysis.

Balapariya and De also suggest that, since copy number alterations impact VAF distributions, very

strict thresholds are necessary to make sure regions analysed with our method are truly diploid. This is an important point and we concord that the original threshold of absolute LRR≤0.5 may have been too lenient. To test the impact of this confounding factor, we have reanalysed the TCGA pan-cancer dataset using the new publicly accessible Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/) variant calls that were not available at the time of our original manuscript. We downloaded 8,455 TCGA tumours from different cancer types and selected cancers with high purity (≥70%) using variants called with Mutect2, as we did in our original manuscript. We then adjusted the VAF for the purity obtained from Aran et al. 2015[4] and identified diploid regions using absolute copy number calls from Forbes et al. 2017[5]. We then classified tumours as neutral/non-neutral using mutations in regions with copy number n=2. We employed the f=[0.12,0.24] integration range and requiring a minimum of 12 subclonal mutations, as in our original manuscript. Out of the 8,455 tumours, 724 satisfied all the conditions (see Supplementary Figure 1 for reasons of exclusions). This analysis confirmed our original findings and even improved the classification, with 40% of neutral cases identified in this cohort (290/724) with respect to the 31% in our original analysis (Table S1). Hence, these confounding factors are more likely to lead to incorrect rejection of the neutral model, rather than incorrect acceptance. In terms of copy number alterations, this makes sense because those would generate binomial clusters in the VAF distribution that would erroneously appear as subclonal clusters.

Therefore, although a perfectly reasonable critique, the confounding factors pointed out by Balapariya and De do not invalidate our original conclusions.

Balapariya and De suggest an alternative explanation of 1/f tails, namely that there is a subclone with high over-dispersion (emulating 1/f neutral tails) exactly positioned around 0.15 VAF in >30% of all tumours (Balapariya and De's Fig 1C). We consider this alternative explanation highly unlikely. Although we agree with the authors that we cannot exclude that specific combinations of parameters from specific distributions may also fit the 1/f cumulative distribution, there are problems with the biological interpretation of low frequency beta-binomially distributed sub-clonal clusters as sole alternative explanation. Specifically, the number of mutations we detect in the 1/f tail is consistently very large, almost always larger than the number of mutations in the clonal cluster. This makes sense in terms of neutral evolution because these mutations accumulate naturally in different lineages at the same time. According to the authors' proposition, instead in 30% of the cases, there would be an extremely old sub-clone (containing a large amount of variants), possibly older than the time between the patient's birth and cancer initiation (as the subclonal cluster contains more mutations than the clonal cluster). Additionally, this very old sub-clone would be always partially hidden from sight because of its low frequency, showing only the right-hand tail of its distribution but never presenting itself in full (Balapariya and De's Figure 1C). We argue that this interpretation of the data is rather implausible, and more importantly, it is in conflict with the body of evidence from tumor phylogenetic trees that consistently show a long trunk and shorter diversification branches. Therefore, although we cannot definitively exclude this alternative explanation based on the observed power law distribution alone, we would argue that the alternative explanation is an artificial construction that immediately implies many implausible consequences. Furthermore, we note that in a recent study led by colleagues[6], when we leveraged on multi-region sequencing of colorectal cancer to analyze neutral tails in exome data, we did demonstrate that those tails describe mutations that are not in the same cells, hence that 1/f tails are not subclones (see Figure 4 in Sun *et al.* 2017[6]).

Balapariya and De suggest further alternative explanations for a 1/f tail in their Figures 1D-F. These simulated examples show how a specifically chosen mixture of 2 or 3 sub-clones produces a similar 1/f pattern in the cumulative distribution and suggest that such mixtures may confound our calling of neutrality. Although we cannot exclude that any specific mixture of subclones may by chance present a signature that is similar to the null model, it seems extremely unlikely that such a 'perfect mixture' of sub-clones exists in >30% of all tumours, especially because the frequency and number of mutations in these clones has to be within a very tight range (in the example shown by Balapariya and De and colleagues there are precisely 1000 mutations in each clone and the dispersions are also very specific). From a phenomenological perspective, a clonal mixture to appear neutral-like requires both the time of appearance of a new subclone (which sets the

number of mutations in the subclone) and its selective advantage (which defines the size of the sub-clone at the time the tumour is sampled)[6] to be 'just right' . This fine tuning is only exacerbated as more subclones are introduced.

We would like to stress that our analysis is built from first principles: we construct a dynamical model of tumour growth that incorporates the fundamental mechanisms of cell division and mutation, and we do not make assumptions on what clonal mixtures should look like. We argue that Balapariya and De's synthetic examples are not based on a model of tumour evolution. In particular, their sub-clones are isogenic, and mutation frequencies are normalised, hence we argue neutrality is a more parsimonious and much more likely explanation of the available data.

Balapariya and De also report that TCGA tumours classified as neutral with our method were classified as polyclonal (2-8 subclones) with other methods[7]. The authors however do not mention which specific cases they refer to, and do not present their VAF distribution for examination, hence we are unable address this point. We note however, that further work has shown that neutral tails can be mistaken for subclones[6], hence it is not surprising that previous clustering methods are biased towards over-clustering. We do acknowledge however that the 1/f analytical test is sensitive to the integration range, hence why we have been developing new methods that instead use the whole VAF distribution[8].

Overall, we maintain that if the data are adequately described by the null model of what we expect in the absence of selection, then by Occam's razor, we would need additional strong evidence for selection to incorporate additional complexity into our explanations of the available data. As we collect higher resolution data we may later be able to reject the null hypothesis.

We wholly agree with the authors' point that the limited depth of sequencing data restricts the resolvability of 'microscopic' evolutionary dynamics, and we discussed this in Williams, Werner *et al.* 2016. As Figure 5 in our original manuscript described, our analysis connects time with the allelic frequency spectrum so that only the earliest branches of the evolutionary process are observable via bulk sequencing at the current usually moderate depth. The deeper the sequencing, the more recent evolutionary events we can observe. Therefore, the majority of what we can observe with current bulk sequencing occurred early in tumour development, and microevolution, as also stated by the authors, cannot be observed with current standards. This however, does not invalidate our analysis, but rather informs on the level of accuracy at which we can understand tumour evolution with current sequencing data. We entirely agree that potentially variable 'microevolution' occurs in tumours at a scale that is currently undetectable in the data and we maintain that to study microevolution, multi-region sequencing[9] is likely the only solution. This is the reason why in follow-up studies we did extend our approach to multi-region profiling[6].

Importantly, the salient point that we would like people to take away from Williams, Werner *et al.* 2016 is that in many cancers neutral evolution – the most parsimonious description of the evolutionary process of cancer growth - is an entirely adequate description of the available sequencing data.

In conclusion, we thank Balapariya and De for providing constructive criticism that has led to important new analyses and improvement of our methods. All these new analyses however confirm that effectively neutral evolution does play an important role in cancer and, although selection remains the critical force that leads to adaptation, neutrality is fundamental to the understanding of how genomic alterations accumulate in a growing tumour within and between selected subpopulations.

Figure 1. Signatures of neutral evolution are robust to confounding factors in the data. Using our stochastic simulation of neutral tumor growth described in Williams, Werner *et al.* we generated 1000 in silico tumors and then subjected the same 1000 simulations to different sampling regimes to generate synthetic datasets. From these synthetic datasets we calculated R^2 values by fitting the neutral model. (A) As the degree of dispersion increases (increased ρ) the mean R^2 becomes lower. However, using our cutoff of 0.98 we would still correctly identify 86% of

samples as neutral even in the most extreme case of over-dispersion (ρ=0.1) compared to 92% using binomial sampling. (B,C) Using Markov chain Monte Carlo to fit a Beta-Binomial model to the clonal cluster on our data we demonstrate that the dispersion ρ<0.005 (10 times lower than what proposed by Balapariya and De) in both our whole genome sequencing gastric cancer cohort and whole exome sequencing colon cancer cohort. The Beta-Binomial distribution becomes a Binomial distribution for $\rho \to 0$. As we find low ρ in our cancer genomic data consistently, we argue that binomial sampling and beta-binomial sampling converge.

Author contributions

M.W. performed beta-binomial noise analysis. T.H. performed TCGA classification. All authors participated in the discussion and addressed the critiques. M.W., B.W., C.B., T.G. and A.S. wrote the manuscript.

Acknowledgments

A.S. is supported by The Chris Rokos Fellowship in Evolution and Cancer and by Cancer Research UK (A22909). T.A.G. is supported by Cancer Research UK (A19771). C.P.B. is supported by the Wellcome Trust (097319/Z/11/Z). B.W. is supported by the Geoffrey W. Lewis Post-Doctoral Training fellowship. A.S. and T.A.G. are jointly supported by the Wellcome Trust (202778/B/16/Z and 202778/Z/16/Z respectively). This work was also supported by Wellcome Trust funding to the Centre for Evolution and Cancer (105104/Z/14/Z).

Data Availability Statement

Gastric cancers from Wang et al. 2014 and colorectal cancers from Sottoriva et al. 2015 were analyzed with our framework in our original manuscript. The new TCGA pan-cancer analysis is based on publicly available GDC calls that can be downloaded from https://portal.gdc.cancer.gov/ (see Table S1).

Conflict of Interest

The authors declare no conflict of interest.

References

1. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nature Genetics* **48,** 238–244 (2016).
2. Wang, K. *et al.* Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nature Genetics* **46,** 573–582 (2014).
3. Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth. *Nature Genetics* **47,** 209–216 (2015).
4. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat Comms* **6,** 8971 (2015).
5. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45,** D777–D783 (2017).
6. Sun, R. *et al.* Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nature Genetics* **49,** 1015–1024 (2017).
7. Andor, N., Harness, J. V., Müller, S., Mewes, H. W. & Petritsch, C. EXPANDS: expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics* **30,** 50–60 (2014).
8. Williams, M. J. *et al.* Quantification of subclonal selection in cancer from bulk sequencing data. *Nature Genetics* (2018). In press.
9. Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501,** 338–345 (2013).