Characterization of biochemical properties and biological activity of isolated antibodies during purification is a critical step that helps focus the proteomics process on the identification of monoclonal antibodies with desired functional properties. We demonstrate this principle in the second proof of concept, namely the isolation of HCMV-neutralizing human monoclonal antibodies from a naturally infected donor. To accomplish this task, we first screened for donors' plasma with potent neutralizing activity *in vitro* (**Supplementary Fig. 9a**). Using plasma from one such donor and a purification strategy restricted to the use of AD4 domain of gB from HCMV[13], we isolated monoclonal antibodies to HCMV with high affinities (up to 278 pM) and potent neutralization activity (IC_{50} values as low as 0.04 μg ml^{-1}). In future experiments, one could envision discovering additional neutralizing monoclonal antibodies from the same donor by using additional components of gB[13] or other HCMV glycoproteins[18] for the affinity purification step. In this way, one could reconstitute a combined pool of potent neutralizing, fully human monoclonal antibodies. Considering that pooled HCMV hyperimmune globulin preparations are still the only available antibody-based HCMV-specific therapy, a neutralizing mixture made by recombinant human monoclonal antibodies would provide an improved clinical tool for passive immunotherapy against HCMV.

Manipulating purification conditions upfront facilitates the isolation and identification of antigen-specific human monoclonal antibodies with various biophysical or biochemical characteristics such as acid resistance, high heat tolerance, specific association and/or dissociation rates, ability to compete with a specific ligand, binding to a protein domain or a combination of any of these properties. We tracked antigen-specific binding activity to monitor enrichment of the desired polyclonal fraction (**Fig. 1a,c**). Such enrichment before mass spectrometry analysis enhances the probability of reconstituting functional heavy and light chain matches by combinatorial pairing, and we speculate that some of the matches may correspond to cognate pairs.

Finally, functionally validated antibody sequences identified using this approach can be used as a guide for further mining of additional clonally related antibody chains from the next-generation sequencing database generated from the same donor[19] (**Supplementary Fig. 10**). These additional antibody chains could have been missed owing to their very low affinity or very low abundance in serum, or because they were encoded by memory B cells, that did not contribute to the serological response at the time the blood sample was drawn. In addition, we cannot rule out that other specific antibodies enriched through purification could not be identified because they were expressed only by plasma cells in the bone marrow or other lymphoid organs and thus their sequences may be absent in the cDNA sequence databases from circulating B cells.

Whether the goal is to identify a broad antibody pool against a whole protein antigen or a more restricted set of neutralizing antibodies against a smaller domain, these results demonstrate that our proteomics approach is applicable in humans and thus may be useful to address questions in humoral immunity and facilitate the development of human antibody therapeutics.

Note: Supplementary information is available at http://www.nature.com/doifinder/10.1038/nbt.2406.

ACKNOWLEDGMENTS

We thank J. Fisher for biotinylation of antigen; D. Moore-Lai, T. Manganaro, T. Palazzola and K. Riley for antibody expression and purification; J. Knott and J. MacNeill for peptide synthesis; C. Manning and M. Nelson for high-content analysis; A. Funicella for plasmid purification, C. Reeves for DNA sequencing of expression constructs; K. Lee and A. Moritz for insightful discussion on mass spectrometry; and S. Martin and E. Savinelli for coordinating donor blood collection. We thank M. Mach for allowing us to adopt the HCMV gB structural image in our manuscript. We thank R. Matthews, T. Singleton and D. Comb for designing graphics. We thank M. Comb, T. Sulahian, K. Huynh, P. Hornbeck, C. Hoffman and L. Morrison for insightful comments and discussion on the manuscript. Finally, we are very grateful to all the volunteers who donated blood, without whom this project would not have been feasible.

AUTHOR CONTRIBUTIONS

S.S., S.A.B., W.C.C. and R.D.P developed the methodology, designed experiments, analyzed data and wrote the manuscript. S.A.B. did bioinformatic analyses. S.S., S.A.B., L.P., J.G.B., R.K.R., X.Z., J.S.W. and S.M.S. did experiments.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available at http://www.nature.com/doifinder/10.1038/nbt.2406.

Shuji Sato[1,2], Sean A Beausoleil[1,2], Lana Popova[1], Jason G Beaudet[1], Ravi K Ramenani[1], Xiaowu Zhang[1], James S Wieler[1], Sandra M Schieferl[1], Wan Cheung Cheung[1] & Roberto D Polakiewicz[1]

[1]Cell Signaling Technology, Danvers, Massachusetts, USA. [2]These authors contributed equally to this work. Correspondence should be addressed to W.C.C. (gcheung@cellsignal.com) or R.D.P. (rpolakiewicz@cellsignal.com).

1. Cheung, W.C. *et al. Nat. Biotechnol.* **30**, 447–452 (2012).
2. Jin, A. *et al. Nat. Med.* **15**, 1088–1092 (2009).
3. Meijer, P.J. *et al. J. Mol. Biol.* **358**, 764–772 (2006).
4. Schmaljohn, C., Cui, Y., Kerby, S., Pennock, D. & Spik, K. *Virology* **258**, 189–200 (1999).
5. Wrammert, J. *et al. Nature* **453**, 667–671 (2008).
6. Purtha, W.E., Tedder, T.F., Johnson, S., Bhattacharya, D. & Diamond, M.S. *J. Exp. Med.* **208**, 2599–2606 (2011).
7. Zuckerman, J.N. & Zuckerman, A.J. *J. Infect.* **41**, 130–136 (2000).
8. Hilleman, M.R. *Infection* **15**, 3–7 (1987).
9. von Pawel-Rammingen, U., Johansson, B.P. & Bjorck, L. *EMBO J.* **21**, 1607–1615 (2002).
10. Tajiri, K. *et al. Antiviral Res.* **87**, 40–49 (2010).
11. Staras, S.A. *et al. Clin. Infect. Dis.* **43**, 1143–1151 (2006).
12. Marshall, G.S., Rabalais, G.P., Stout, G.G. & Waldeyer, S.L. *J. Infect. Dis.* **165**, 381–384 (1992).
13. Potzsch, S. *et al. PLoS Pathog.* **7**, e1002172 (2011).
14. Abai, A.M., Smith, L.R. & Wloch, M.K. *J. Immunol. Methods* **322**, 82–93 (2007).
15. Macagno, A. *et al. J. Virol.* **84**, 1005–1013 (2010).
16. Barrette, R.W., Urbonas, J. & Silbart, L.K. *Clin. Vaccine Immunol.* **13**, 802–805 (2006).
17. Scheid, J.F. *et al. Nature* **458**, 636–640 (2009).
18. Ryckman, B.J. *et al. J. Virol.* **82**, 60–70 (2008).
19. Wu, X. *et al. Science* **333**, 1593–1602 (2011).

Analyzing the association of SCNA boundaries with replication timing

To the Editor:

A paper by De and Michor[1] published in last December's issue claimed that DNA replication timing and long-range DNA interactions can predict mutational landscapes of cancer genomes. We would like to draw readers' attention to a statistical weakness in their analysis, which in our opinion negates one of the main claims of the article.

The paper, which presents integrative analysis of a database of somatic copy-number alterations (SCNAs)[2], DNA replication timing[3] and three-dimensional conformation data[4] (Hi-C), argues that the data provide evidence that the formation of SCNAs is governed by a DNA replication–driven model, in which alterations occur opportunistically when two active replication forks at the boundaries of the nascent SCNA are in close three-dimensional proximity. The empirical evidence presented has two essential components: first, an increase in the density

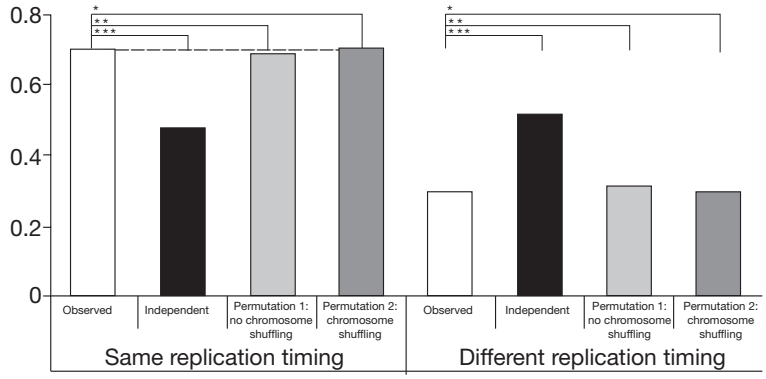

Figure 1 Only the overly simplistic 'independent' background model, whereby SCNA ends are chosen independently, neglecting long-range replication time correlation (black bars), yields an estimate for ends to reside in the same timing domain, which differs from the observed (white bars) fraction in a biologically meaningful way. Utilizing random permutation background models, the differences are diminishingly small or disappear entirely (dark- and light-gray bars; see text). The thin horizontal lines were introduced to aid the comparison. *P = 0.1351; q = 0.99. **P = 0.0006; q < 1 × 10–5. ***P < 1 × 10–5; q < 1 × 10–5

of Hi-C reads close to SCNA boundaries demonstrates spatial proximity of SCNA boundaries; and second, the association with the DNA replication process is supported by observing that the boundaries of the SCNA preferentially reside in the same replication timing domain.

Our analysis of the statistical methods used to establish the latter evidence concerning replication timing has identified a potential weakness introduced by a bias resulting from replication timing correlations within relatively large DNA domains. The null hypothesis used in the main body of the paper to model the random, 'trivial' expectation of finding both SCNA ends in the same replication timing domain essentially assumes that the two ends of SCNA are chosen independently from early and/or late replication regions. But independence is a rather strong assumption in this context. By their very nature, SCNA boundaries are comparatively close along the linear DNA space, thus increasing the likelihood that both ends of an SCNA will reside in the same replication time category. Although individual replication domains are indeed smaller than the median SCNA used in the study (4.5 Mb), much larger replication timing structures are common throughout the human genome, where long stretches of early- or late-replicating domains are interrupted only briefly by short bursts of a different timing. The decision of De and Michor to ignore these structures in the paper leads to a severe overestimation of how strongly replication timing is linked to the creation of SCNAs. In fact, in our own analysis, the association may not even be significant at all.

Based on that overly simplistic independent-ends model, **Figure 1** (left) indicates an almost 50% increase of same-timing ends of SCNAs over the random expectation, thus suggesting a strong association of replication timing with the events leading to the formation of SCNAs. In contrast, applying a permutation-based method, such as the one used by De and Michor in the Supplementary Module 10 accompanying their article[1], would provide a more realistic estimate. Even so, background frequencies were not reported for the permutation test. We therefore repeated the analysis; this reproduced the published results but also provided an estimate that same-timing boundaries were only 2% more frequent than expected by chance (**Fig. 1**, right). Even though our analysis confirmed the highly significant q value of $<10^{-5}$ for this enrichment reported in Supplementary Module 10 of the paper, this is driven mostly by the large number of SCNAs analyzed, and one may suspect that the association might disappear entirely when other, unknown biases were taken into account.

In fact, even small modifications of the permutation algorithm render the results nonsignificant, as is the case when the permutation test is carried out separately for each chromosome or when SCNAs are shuffled across chromosomes. A permutation test generates different results for individual chromosomes (**Table 1**). Out of 22 autosomal chromosomes, we found five with a significant preference for SCNA boundaries to reside in different timing domains. Three of these five chromosomes are acrocentric. Additionally, a permutation test that assigns SCNAs to chromosomes at random (**Fig. 1**) shows that the boundaries of an SCNA are not

Table 1 SCNA boundaries within the same replication timing domain

	No. of SCNAs	Observed frequency	Average permutation frequency	Effect strength[a]	Numerical permutation q value
1	21,146	0.712	0.680	0.047	8.33×10^{-7}
2	17,995	0.683	0.611	0.118	0
3	16,348	0.670	0.645	0.039	5.54×10^{-3}
4	13,136	0.795	0.723	0.100	0
5	14,954	0.703	0.647	0.086	0
6	14,698	0.712	0.666	0.069	1.67×10^{-6}
7	13,677	0.615	0.582	0.057	1.32×10^{-3}
8	17,050	0.707	0.648	0.092	0
9	15,724	0.806	0.791	0.019	4.15×10^{-2}
10	13,487	0.653	0.607	0.076	8.33×10^{-6}
11	14,235	0.705	0.653	0.079	0
12	14,799	0.567	0.669	–0.153	1.00
13	11,305	0.541	0.734	–0.263	1.00
14	11,123	0.704	0.627	0.124	0
15	9,842	0.657	0.703	–0.065	1.00
16	10,493	0.712	0.658	0.082	0
17	11,525	0.833	0.821	0.015	2.09×10^{-2}
18	7,914	0.720	0.690	0.043	1.54×10^{-2}
19	6,753	0.857	0.859	–0.002	6.00×10^{-1}
20	8,160	0.684	0.706	–0.031	9.89×10^{-1}
21	3,775	0.604	0.644	–0.062	9.84×10^{-1}
22	6,805	0.931	0.888	0.048	0

[a]Effect strength was calculated by dividing the difference between observed and average permutation frequencies by the average permutation frequency.

more likely than expected by chance to have the same replication timing. The results of these tests lead us to conclude that the association of SCNA boundaries with replication timing is, if present at all, not only rather weak but also not universal.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Sven Bilke & Yevgeniy Gindin

National Cancer Institute, Bethesda, Maryland, USA.
e-mail: bilkes@mail.nih.gov

1. De, S. & Michor, F. *Nat. Biotechnol.* **29**, 1103–1107 (2011).
2. Beroukhim, R. *et al. Nature* **463**, 899–905 (2010).
3. Hansen, R.S. *et al. Proc. Natl. Acad. Sci. USA* **107**, 139–144 (2010).
4. Lieberman-Aiden, E. *et al. Science* **326**, 289–293 (2009).

De and Michor respond:

Our paper[1] analyzed DNA replication timing patterns at the two boundaries of somatic copy-number alterations (SCNAs) in a large number of cancer samples and reported that, in general, these boundaries tend to be replicated at the same time during cell division. We found that the observed frequency of such events is significantly higher than the pattern expected by chance, based upon two independent statistical strategies: Fisher's test and a permutation analysis. Bilke and Ginden[2] now highlight the fact that although individual replication timing domains are typically short, the human genome contains many higher-order replication timing structures, wherein long stretches of early- or late-replicating domains are interrupted only briefly by short regions with different replication timing. They argue that ignoring such structures would lead to an overestimation of the association between replication timing and SCNA boundaries, and therefore they prefer permutation analyses to Fisher's test. Nevertheless, when Bilke and Ginden[2] repeat our analyses, they reproduce the published results and confirm the q value of $<1 \times 10^{-5}$ that we reported[1]. They caution, however, that the statistical significance of the permutation analysis could be driven by the large size of the data set and might decrease when unknown biases are taken into account (for example, permutation across chromosomes instead of within chromosomes).

Each statistical approach depends on its underlying assumptions. For instance, Fisher's test determines the significance of the difference between observed and expected values under the assumption that the two boundaries of SCNAs are chosen independently from early- versus late-replication timing zones. The permutation test, in contrast, assumes that the observations are exchangeable under the chosen null hypothesis. Therefore, in our original paper we decided to report the statistical significance of our findings using both statistical strategies[1].

We have now performed additional analyses to investigate the concern about higher-order replication timing domains. We first divided the SCNAs into four groups, depending on their lengths: <500 kb, 500 kb–1 Mb, 1–5 Mb and >5 Mb. We then performed the permutation analysis performed as described in Supplementary Module 10 of our original paper[1]. First, we calculated the number of instances (N_1) in which the two boundaries of SCNAs had the same replication timing, and the number of instances (N_2) in which the two boundaries of SCNAs had different replication timing in our data set. We then performed a permutation analysis in which we randomized the positions of the SCNAs, keeping their lengths unchanged, and counted the number of cases in which the simulated SCNAs had the same (N_{1sim}) or different (N_{2sim}) replication timing. We repeated this permutation analysis 10^6 times and calculated the proportion of cases (the q value) in which the quantity N_{1sim}/N_{2sim} was greater than N_1/N_2. The q value represents the probability of observing an enrichment for N_1 over N_2 by chance. The distribution of this statistic is provided in Figure 1. We found that in a vast majority of cases, $(N_1/N_2)/(N_{1sim}/N_{2sim})$ was significantly greater than 1, although the exact value was typically modest. We then repeated the analysis using a filtered set of SCNAs, as described in Supplementary Module of our original paper[1], after excluding complex SCNAs. These analyses confirmed that due to the higher-order replication structures brought up by Bilke and Gindin[2], the ratio of observed to expected values was indeed smaller when using the permutation test as compared to Fisher's test. Nevertheless, the enrichment remained significant in 3 out of 4 cases, and the observed/expected ratio was between 2% and 12% when all SCNAs from Beroukhim *et al.*[3] were taken into account (Fig. 1). Note that when analyzing a filtered list of SCNAs from which complex alterations, involving multiple DNA breakpoints, were excluded, the enrichment was found to be much higher—up to 30% in some cases and reaching significance ($P<1 \times 10^{-5}$) in all instances. We hypothesized that over large genomic distances, higher-order replication timing structures have limited effects, and therefore for large SCNAs, both the permutation test and Fisher's test should yield comparable results. Indeed, for the SCNAs of length >5 Mb, we found that Fisher's exact test and the permutation test produced very similar results (Fig. 2).

In our analyses, we used a simple permutation model, which conserves SCNA length and chromosome information. Alternative permutation models, based on alternative biological hypotheses of the generation of SCNAs, are possible. For instance, we have now also permuted the SCNAs across the chromosomes, as suggested by Bilke and Gindin[2] (Fig. 3). The results were similar to those shown in Figure 1. Once again, analysis of the filtered

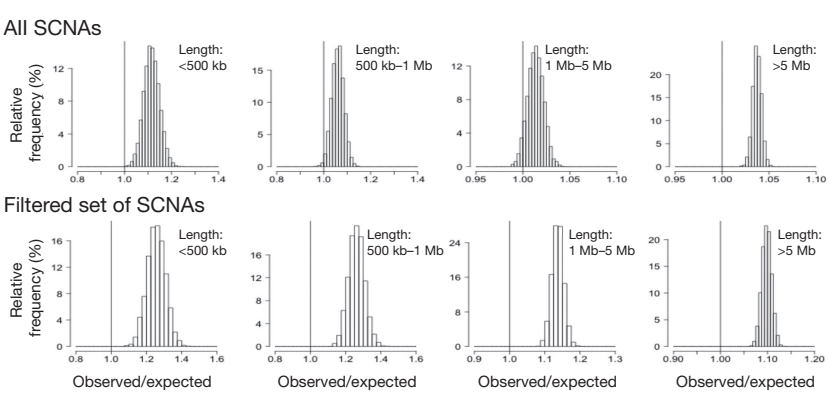

Figure 1 The distribution of observed/expected values obtained using a simple permutation approach for SCNAs grouped according to their size, with permutation performed within chromosomes. (**a**,**b**) We display the results using (**a**) all SCNAs from Beroukhim *et al.*[3] and (**b**) a filtered list of those SCNAs that excludes those arising via multiple DNA breaks as reported in Supplementary Module 9 of ref. 1.

set of SCNAs (Fig. 3) produced similar results, albeit with a stronger enrichment across all size categories. Therefore, it is possible that the complex SCNAs, which arise due to multiple genomic alteration events and for which determination of pairs of boundaries is challenging, contribute to the observed discrepancies. Moreover, the frequency of SCNAs, and their typical length, differ between chromosomes, and thus after randomization across chromosomes, the sets of observed and expected SCNAs for each chromosome can differ in their frequencies and length distributions, which can potentially bias the conclusions. For this reason, in our original paper, we refrained from permuting the SCNAs across the chromosomes, as suggested by Bilke and Gindin[2]. Before employing alternative permutation models, further work needs to be done to systematically evaluate different scenarios and their potential implications.

We found that, in general, the two boundaries of SCNAs more often than not had the same replication timing; this observation was invariant across different data sets and size thresholds of SCNAs. For instance, >95% of SCNAs with size <500 kb had the same replication timing at the two boundaries. This proportion was >85%, >70% and >50% for SCNAs of size <500 kb, 0.5–1 Mb, 1–5 Mb and >5 Mb, respectively. We then created a 3 × 3 matrix $N = \{n_{\text{early-early}}, n_{\text{early-mid}}, n_{\text{early-late}}, n_{\text{mid-early}}, n_{\text{mid-mid}}, n_{\text{mid-late}}, n_{\text{late-early}}, n_{\text{late-mid}}, n_{\text{late-late}}\}$ such that n_{ij} represents the number of SCNAs with the two end-points in the i^{th} and j^{th} replication timing category. We generated the matrix N for SCNAs with length <500 kb, 0.5–1 Mb, 1–5 Mb and >5 Mb, and in each case calculated Goodman and Kruskal's symmetric lambda (λ; http://faculty.vassar.edu/lowry/lambda.html). We found that the value of λ was usually high for small SCNAs but decreased as SCNA length increased (<500 kb, λ = 0.93; 0.5–1 Mb, λ = 0.76; 1–5 Mb, λ = 0.30; and >5 Mb, λ = 0.08). Thus, the two boundaries of large SCNAs usually do not reside within a single higher-order replication timing zone. We also calculated λ for the filtered set of SCNAs and found similar results (<500 kb, λ = 0.94; 0.5–1 Mb, λ = 0.78; 1–5 Mb, λ = 0.34; and >5 Mb, λ = 0.10). These results are consistent with our findings presented in Figures 1 and 2 and suggest that the bias driving the difference in the results obtained with Fisher's versus permutation tests arises because of small SCNAs, typically smaller than 1 Mb in length. Our observations that SCNA endpoints usually have similar replication timing offer a plausible mechanistic explanation for the generation of SCNAs in cancer genomes. Given that many SCNAs arise as a result of replicative stress, common replication timing at the two boundaries and their physical proximity in the three-dimensional organization of the nucleus offer a provocative mechanistic hypothesis for the generation of SCNAs. At this point, we would like to reiterate some of the caveats mentioned previously[1]. It would be important to validate our observations by analyzing DNA replication timing, copy number and long-range interaction data obtained from the same samples. Cancer cells often have an abnormal genome and epigenome, which could alter local DNA replication timing and long-range interactions, and thus further experiments are required to firmly establish our hypothesis. Currently available experimental evidence[4–6], however, is consistent with our findings.

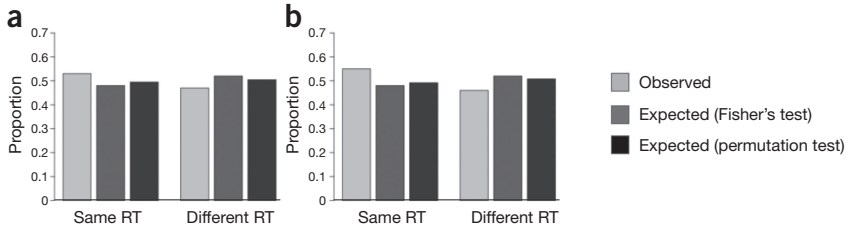

Figure 2 Comparison of the enrichment using Fisher's test and permutation test for SCNAs of length > 5 Mb. (**a**,**b**) We display the results using (**a**) all SCNAs from Beroukhim *et al.*[3] (**a**) and a filtered list of those SCNAs that excludes those arising via multiple DNA breaks as reported in Supplementary Module 9 of ref. 1 (**b**). The enrichment is statistically significant in both the permutation test and Fisher's test.

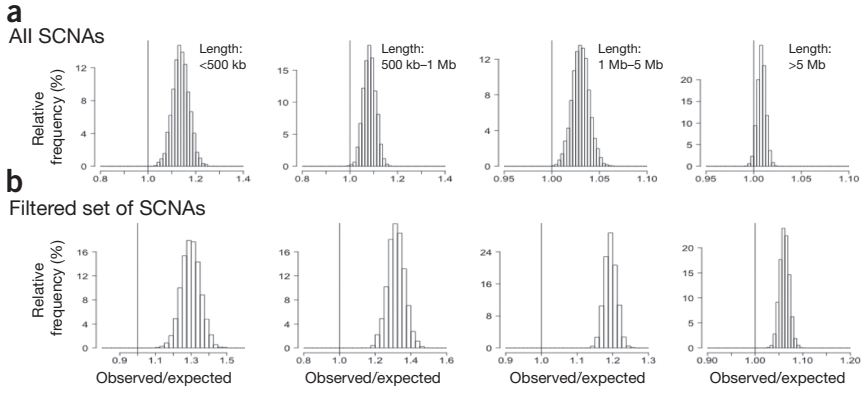

Figure 3 The distribution of observed/expected values obtained using a simple permutation approach for SCNAs grouped according to their size, with permutation performed across chromosomes. (**a**,**b**) We display the results using all SCNAs from Beroukhim *et al.*[3] (**a**) and a filtered list of those SCNAs which exclude those arising via multiple DNA breaks as reported in Supplementary Module 9 of ref. 1 (**b**).

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

***Subhajyoti De*[1,2] *& Franziska Michor*[1,2]**

[1]*Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA.* [2]*Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA.*
e-mail: michor@jimmy.harvard.edu

1. De, S. & Michor, F. *Nat. Biotechnol* **29**, 1103–1108 (2011).
2. Bilke, S. & Gindin, Y. *Nat. Biotechnol.* **30**, 1043–1045 (2012).
3. Beroukhim, R. *et al. Nature* **463**, 899–905 (2010).
4. Hastings, P.J., Lupski, J.R., Rosenberg, S.M. & Ira, G. *Nat. Rev. Genet.* **10**, 551–564 (2009).
5. Misteli, T. & Soutoglou, E. *Nat. Rev. Mol. Cell Biol.* **10**, 243–254 (2009).
6. Zhang, Y. *et al. Cell* **148**, 908–921 (2012).