

The following resources related to this article are available online at www.sciencemag.org (this information is current as of November 8, 2009):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/317/5844/1500a>

Supporting Online Material can be found at:

<http://www.sciencemag.org/cgi/content/full/317/5844/1500a/DC1>

A list of selected additional articles on the Science Web sites **related to this article** can be found at:

<http://www.sciencemag.org/cgi/content/full/317/5844/1500a#related-content>

This article **cites 7 articles**, 2 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/317/5844/1500a#otherarticles>

This article appears in the following **subject collections**:

Medicine, Diseases

<http://www.sciencemag.org/cgi/collection/medicine>

Technical Comments

http://www.sciencemag.org/cgi/collection/tech_comment

Information about obtaining **reprints** of this article or about obtaining **permission to reproduce this article** in whole or in part can be found at:

<http://www.sciencemag.org/about/permissions.dtl>

Comment on “The Consensus Coding Sequences of Human Breast and Colorectal Cancers”

William F. Forrest¹ and Guy Cavet^{2*}

Sjöblom *et al.* (Research Articles, 13 October 2006, p. 268) used data from cancer genome resequencing to identify genes with elevated mutation rates. Their analysis used point probabilities when it should have used *P* values for the hypotheses they intended to test. Reimplementing their analysis method with exact *P* values results in far fewer genes with mutation rates that achieve statistical significance.

Sjöblom *et al.* (1) sequenced ~13,000 genes in human breast and colorectal cancers and identified a large number of genes that were mutated at significant frequency. We show that reimplementation of their analysis using more appropriate test statistics results in a much smaller number of significant genes.

The statistical analysis procedure in (1) included three steps. In the first step, the probability of the observed mutation profile for each gene was calculated. A mutation profile consists of the counts of mutations found in each of seven categories defined on the basis of mutation type and nucleotide context. For example, the mutation profile of *SPTAN1* in breast cancer is shown in Table 1. The binomial probability of the exact number of mutations in a category was calculated for each of the seven categories using estimated background mutation rates and the numbers of residues sequenced in each category. The overall profile probability was calculated for each gene as the product of the seven individual probabilities. In the second step, these profile probabilities were treated as if they were *P* values and adjusted for multiple hypothesis testing according to the procedure of Benjamini and Hochberg [although without a monotonization step (2)]. This procedure was used to calculate “*q*” scores [a term adopted from the Excel spreadsheet mentioned in (1)]. In the third step, a CaMP (cancer mutation prevalence) score was calculated for each gene as the negative log₁₀-transformed *q* score. The CaMP score reflects the false discovery rate (FDR). As stated in (1), “90% of the genes with CaMP scores of >1.0 are predicted to have mutation frequencies higher than the background mutation frequency.” Genes with CaMP scores ≥ 1 are considered to be candidate cancer genes (*CAN* genes). This procedure identified 122 *CAN* genes in breast cancer and 69 in colorectal cancer.

We argue that the CaMP scores calculated as above do not fulfill the stated intent of Sjöblom *et al.* in that they do not reflect “the probability that the number of mutations observed in a gene reflects a mutation frequency that is higher than that expected to be observed by chance given the background mutation rate.” Calculating the probability of the observed mutation profile for each gene fails to take into account the many other possible mutation profiles that would result in equal or greater numbers of mutations. For example, five mutations in *SPTAN1* could have occurred in 462 distinct profiles across the authors’ seven nucleotide categories. For the null hypothesis that the observed number of mutations in a gene is due to the background mutation process only, the appropriate *P* value is the sum of probabilities of all mutation profiles for equal or greater numbers of mutations. We calculated *P* values of this kind exactly by enumerating all possible mutation profiles for up to 10 total mutations per gene. For each gene, we calculated the probability of each profile, as in (1), found the sum of the probabilities of all profiles with fewer mutations than were observed, and subtracted from 1. We calculated *q* scores and CaMP scores as described in (1) (Fig. 1). The probabilities for *TP53* in breast cancer and for *TP53*, *APC*, and *KRAS* in colorectal cancer are so small that they make exact calculation of CaMP scores computationally difficult. For these genes, we report a lower bound on the CaMP scores of 9 (equivalent to a FDR of one in one billion). This procedure identifies only 2 *CAN* genes in breast cancer and 11 in colorectal cancer (see Supporting Online Material).

The choice of Sjöblom *et al.* to define the hypothesis of interest in terms of the total number of mutations in a gene compromises the sensitivity of the analysis. For a given number of mutations, some mutation profiles have much lower probabilities than others (for example, those with mutations at T residues, which have low estimated mutation rates). Greater sensitivity to detect *CAN* genes can be achieved by addressing the more general null hypothesis that

the observed mutation profile is due to the background mutation process only. *P* values can be calculated for this hypothesis by finding the sum of the probabilities of all mutation profiles that have probabilities less than or equal to the probability of the observed profile. We calculated *P* values of this kind exactly by finding the sum of the probabilities of all profiles with probabilities greater than that of the observed profile and subtracting from 1. We calculated *q* scores and CaMP scores as before (Fig. 1). Again, we report a lower bound of 9 for the CaMP scores of *TP53* in breast cancer and of *TP53*, *APC*, and *KRAS* in colorectal cancer. This procedure identifies 6 *CAN* genes in breast cancer and 28 in colorectal cancer (see SOM). The rank order of genes in each of our analyses is also different from that obtained by Sjöblom *et al.*

It is worth noting that these calculations employ background mutation rate estimates that can be made only approximately using studies published to date (3, 4). Lower background rates would result in more statistically significant genes, whereas higher background rates would result in fewer significant genes. The discovery phase data of Sjöblom *et al.* (which employed breast cancer cell lines) suggests higher rates, even if known cancer genes (5) are excluded and an additional 23% of the mutations are dismissed as drivers (6). However, other studies report consistent or lower rates in tumor samples (4, 6). It is also clear that mutation rates vary considerably between tumors (even those without microsatellite instability) (1, 3). This variation in mutation rates is not accounted for in the method of Sjöblom *et al.*, or in our analysis, and could further affect the number of *CAN* genes.

Finally, the approach of Sjöblom *et al.* is intended to identify specific genes with anomalous patterns of mutation in cancer, but it does not estimate the total number of genes that can, when mutated, contribute to cancer. Furthermore, alternative methods may identify *CAN* genes with greater sensitivity, for example, by taking into account the distinct discovery and validation phases of the project. Simulation methods allowing empirical estimation of FDRs are an attractive proposal (7, 8), although CaMP scores cannot be used as a test statistic in these approaches because the scores of different genes

Table 1. Observed mutation profile for *SPTAN1* in breast cancer, giving the number of mutations observed in each category.

Category	No. of mutations observed
C or G (in CpG)	1
C or G (in TpC or GpA)	2
A	0
C (not in CpG or TpC)	1
G (not in CpG or GpA)	0
T	0
ins/del	1

¹Department of Biostatistics, Genentech, Inc., South San Francisco, CA 94080, USA. ²Department of Bioinformatics, Genentech, Inc., South San Francisco, CA 94080, USA.

*To whom correspondence should be addressed. E-mail: cavet.guy@gene.com

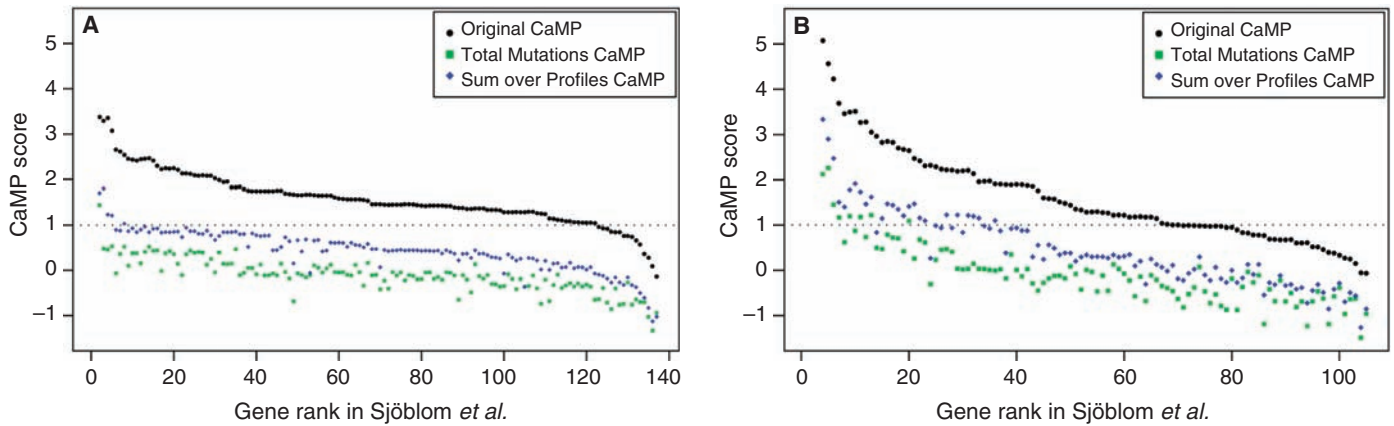


Fig. 1. CaMP scores from original and revised analyses. Black circles indicate CaMP scores calculated according to (1). Green squares indicate CaMP scores calculated using P values reflecting whether a gene has more mutations than expected by chance. Blue diamonds indicate CaMP scores calculated using P values reflecting whether a gene's mutational

profile is consistent with the background mutation rates. Dotted lines indicate the CaMP score threshold for CAN genes. **(A)** Results for breast cancer (results for TP53 are off scale and not shown). **(B)** Results for colorectal cancer (results for TP53, APC, and KRAS are off scale and not shown).

are not independent. Further advances in analysis methods and better knowledge of background mutation rates will allow more confident identification of new cancer genes.

References

1. T. Sjöblom *et al.*, *Science* **314**, 268 (2006).
2. Y. Benjamini, Y. Hochberg, *J. R. Stat. Soc. Ser. B* **57**, 289 (1995).
3. P. Stephens *et al.*, *Nat. Genet.* **37**, 590 (2005).
4. T. L. Wang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 3076 (2002).
5. P. A. Futreal *et al.*, *Nat. Rev. Cancer* **4**, 177 (2004).
6. C. Greenman *et al.*, *Nature* **446**, 153 (2007).
7. G. Parmigiani *et al.*, Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 126, www.bepress.com/jhubiostat/paper126 (December 2006).
8. B. Efron, R. Tibshirani, *Genet. Epidemiol.* **23**, 70 (2002).

Supporting Online Material

www.sciencemag.org/cgi/content/full/317/5844/1500a/DC1
Tables S1 and S2

29 November 2006; accepted 2 July 2007
10.1126/science.1138179