

The following resources related to this article are available online at www.sciencemag.org (this information is current as of December 24, 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/317/5844/1500d>

Supporting Online Material can be found at:
<http://www.sciencemag.org/cgi/content/full/317/5844/1500d/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/317/5844/1500d#related-content>

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>

Response to Comments on “The Consensus Coding Sequences of Human Breast and Colorectal Cancers”

Giovanni Parmigiani,^{1*} Jimmy Lin,¹ Simina M. Boca,¹ Tobias Sjöblom,¹ Siân Jones,¹ Laura D. Wood,¹ D. Williams Parsons,¹ Thomas Barber,^{1,2} Phillip Buckhaults,³ Sanford D. Markowitz,⁴ Ben Ho Park,⁵ Kurtis E. Bachman,⁶ Nickolas Papadopoulos,¹ Bert Vogelstein,^{1*} Kenneth W. Kinzler,^{1*} Victor E. Velculescu^{1*}

Forrest and Cavet, Getz *et al.*, and Rubin and Green describe a variety of statistical methods to analyze the mutational data published in Sjöblom *et al.* However, their conclusions are inaccurate because they are based on analyses that do not fully take into account the experimental design and other critical features of our study. When these factors are incorporated, their methods provide estimates similar to those we reported and support the conclusion that a large number of genes are mutated at rates greater than the passenger mutation rate.

We appreciate the considerable effort put forth by Forrest and Cavet (1), Getz *et al.* (2), and Rubin and Green (3) to analyze and expand upon the statistical techniques used in our study (4). Their Technical Comments raise three important questions: (i) Are the candidate cancer genes (*CAN* genes) identified in Sjöblom *et al.* mutated at rates higher than the experimentally determined passenger rate? (ii) How does the passenger mutation rate affect the estimates of *CAN* genes? and (iii) How does variation in passenger mutation rates among genes affect the analysis? We address each of these questions in order below.

Before discussing the statistical concerns relating to the issue of whether the *CAN* genes we identified are mutated at a higher rate than the experimentally determined passenger rate, a simple “reality check” can be used to obtain an intuitive answer to this question. The number of mutated genes that one would expect to identify as a result of any given passenger mutation rate can be assessed by *in silico* simulations using assumptions that are common to all approaches.

¹Ludwig Center for Cancer Genetics and Therapeutics, Howard Hughes Medical Institute and Departments of Biostatistics and Pathology, Johns Hopkins Medical Institutions, Baltimore, MD 21231, USA. ²Lilly Research Laboratories, Eli Lilly and Company, Indianapolis, IN 46285, USA. ³Department of Pathology and Microbiology, Center for Colon Cancer Research, and The South Carolina Cancer Center, Division of Basic Research, University of South Carolina, School of Medicine, Columbia, SC 29229, USA. ⁴Department of Medicine and Ireland Cancer Center, Case Western Reserve University and University Hospitals of Cleveland, and Howard Hughes Medical Institute, Cleveland, OH 44106, USA. ⁵Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore, MD 21231, USA. ⁶Departments of Radiation Oncology and Biochemistry and Molecular Biology, Marlene and Stewart Greenebaum Cancer Center, University of Maryland School of Medicine, Baltimore, MD 21201, USA.

*To whom correspondence should be addressed. E-mail: gp@jhu.edu (G.P.); vogelbe@jhmi.edu (B.V.); kinzke@jhmi.edu (K.W.K.); velculescu@jhmi.edu (V.E.V.)

These simulations show that only 17 genes in colon and 16 genes in breast are expected to be mutated in the two sequential screens employed in our study. In the Sjöblom *et al.* experiment (4), however, 105 and 137 genes were found to be mutated in both screens. These numbers are consistent with those we estimated to be candidate cancer genes and suggest that the vast majority of these are mutated at frequencies significantly higher than the passenger rate employed in Sjöblom *et al.*

Based on this check, the question then becomes why the Forrest and Cavet, Getz *et al.*,

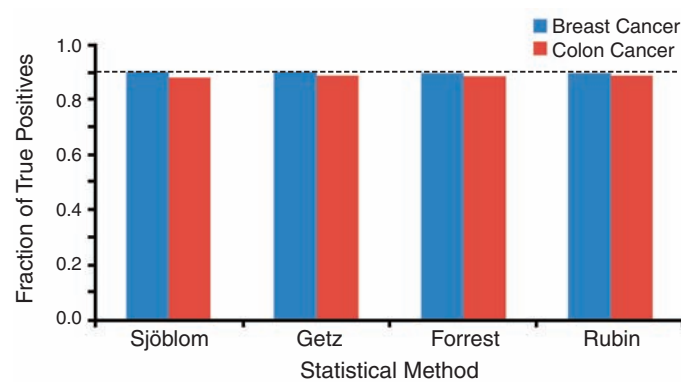


Fig. 1. Estimated proportion of *CAN* genes expected to be mutated above the passenger rate according to various statistical models. See (5) for details.

and Rubin and Green statistical models so greatly underestimated the number of genes mutated at rates greater than this passenger rate. The answer is that there were problems with their evaluations that led them to draw incorrect conclusions. These problems are described in detail in the Supporting Online Material (5) and include the following: (i) failure to fully account for the two-stage design of the experiment (discovery screen followed by validation screen); (ii) failure to fully

account for the number of nucleotides successfully sequenced, including those from genes for which no mutations were found; and (iii) failure to include the precise base and neighboring bases of the mutations that were observed. Each of these problems results in an artificial inflation of the *P* values and a consequent decrease in the number of *CAN* genes (5). When the otherwise reasonable models proposed by the three groups incorporate these critical parameters, they yield numbers of *CAN* genes almost identical to those we initially proposed (Fig. 1). We are pleased that the interaction between our groups has facilitated more precise statistical modeling of the experimental results and that, when implemented properly, all models support our original estimates.

How does the passenger mutation rate affect the estimates of *CAN* genes? The passenger mutation rate is indeed important for determining the number of *CAN* genes. This is why we devoted so much time to measuring this rate in an unbiased fashion in colorectal cancers. The statement in the SOM for (2) that “only a small subset of genes were studied” to determine these rates is incorrect. In fact, in a separate study, we analyzed more than 250 Mb of noncoding DNA sequence in colorectal cancers, representing the largest such sequencing effort ever performed. In Sjöblom *et al.* (4), we assumed a passenger mutation rate of 1.2 mutations per Mb, twice the experimentally determined passenger rate, which we felt was conservative. In contrast, Rubin and Green and Getz *et al.* use internal data from Sjöblom *et al.* to estimate passenger mutation rates. Determining passenger rates from these data requires prior

assumptions about the number of true cancer genes as well as passengers, and therefore incorporates circular reasoning. But even with higher presumed passenger rates, our conclusion that a large number of genes are mutated at rates greater than the passenger rate is confirmed as long as the factors noted above are incorporated into the estimates (see table S7). Furthermore, in a check similar to that described above but using internal mutation rates based on the arguments of the Rubin and Getz groups, we find that a minimum of 66 and 63 more genes are mutated in breast and colorectal cancers, respectively, than predicted by chance.

Finally, how does variation in passenger mutation rates among genes affect the analysis? Although their corrected statistical models yield results that agree with ours (Fig. 1), Forrest and Cavet, Getz *et al.*, and Rubin and Green highlight

a fundamental disagreement between our groups with regard to the type of knowledge that can be gained from sequencing studies. We believe that sequencing data can, in general, only point to candidate genes worthy of further study. In contrast, the other groups equate candidate genes with true “cancer genes.” This is particularly emphasized by the attempt of Getz *et al.* to factor in variations among genes’ intrinsic mutation rates in their statistical models. Sequence data can only identify genes that are mutated at unusually high rates; in general, such data cannot determine whether the higher rate is the result of higher intrinsic mutability or of positive selection during tumorigenesis. The most one can do with sequencing data is to prioritize genes on the basis of their mutation characteristics and frequency. Notably, the ranking of genes is very similar regardless of which statistical model is used: The rank correlations between CaMP scores and the scores described in the comments (1–3) are >0.85 .

This disagreement has profound implications for the future sequencing projects mentioned in

the comments. For example, the stated goal of the Cancer Genome Atlas Project (TCGA) is to find 94% of the cancer genes mutated in at least 5% of tumors of a given type and to unequivocally distinguish passengers from true cancer genes (6). If the background rates and variations thereof estimated by Getz *et al.* are valid, however, it will be impossible to reach this goal, even if 25,000 tumors are studied—greater by a factor of more than 100 than is currently envisioned (5). On the other hand, we believe that the rates proposed by Getz *et al.* are unreasonably high and that if mutations are properly interpreted only as valuable clues for future investigation, larger studies such as the TCGA will provide a wealth of useful and unique information.

The Sjöblom *et al.* study (4) shows that the unbiased examination of the sequences of genes in cancers is a very powerful way to identify potential cancer genes. The fact that our study found virtually all the cancer genes previously identified in breast and colorectal cancers has provided unequivocal evidence of the utility of this ap-

proach. However, such sequencing studies have limitations in that only lists of candidate genes, not bona fide cancer genes, can generally be inferred from the data. These limitations apply to all such studies, whether 25 or 25,000 cancers are evaluated.

References

1. W. F. Forrest, G. Cavet, *Science* **317**, 1500 (2007); www.sciencemag.org/cgi/content/full/317/5844/1500a.
2. G. Getz *et al.*, *Science* **317**, 1500 (2007); www.sciencemag.org/cgi/content/full/317/5844/1500b.
3. A. F. Rubin, P. Green, *Science* **317**, 1500 (2007); www.sciencemag.org/cgi/content/full/317/5844/1500c.
4. T. Sjöblom *et al.*, *Science* **314**, 268 (2006).
5. See supporting material on *Science* Online.
6. http://cancergenome.nih.gov/about/NCABReport_Feb05.pdf

Supporting Online Material

www.sciencemag.org/cgi/content/full/317/5844/1500d/DC1

SOM Text

Tables S1 to S7

Software codes and data

References

20 February 2007; accepted 22 August 2007

10.1126/science.1138773