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

Gad Getz,[1]*† Holger Höfling,[2]* Jill P. Mesirov,[1] Todd R. Golub,[1,3,4,5,6] Matthew Meyerson,[1,3] Robert Tibshirani,[2,7] Eric S. Lander[1,6,8]

Sjöblom *et al*. (Research Article, 13 October 2006, p. 268) reported nearly 200 novel cancer genes said to have a 90% probability of being involved in colon or breast cancer. However, their analysis raises two statistical concerns. When these concerns are addressed, few genes with significantly elevated mutation rates remain. Although the biological methodology in Sjöblom *et al*. is sound, more samples are needed to achieve sufficient power.

Sjöblom *et al*. (*1*) reported the first genome-wide effort to identify genes mutated in cancer. They also introduced a two-stage design in which they screened a large set of genes (13,023) for somatic mutations in a discovery set (11 breast and 11 colorectal cancers) and then screened only the small subset of genes that harbored at least one somatic mutation in a validation set (24 breast or colorectal tumors). They identified genes as candidate cancer genes (*CAN* genes) by applying a statistical model designed to assess the likelihood that the observed somatic nonsynonymous mutations would occur by chance. The approach employed the false discovery rate (FDR) approach of Benjamini and Hochberg (*2*) and used an assumed background mutation rate of $\mu = 1.2 \times 10^{-6}$.

The Sjöblom *et al*. analysis yielded rank-ordered lists of candidate genes with 122 and 69 genes in breast and colorectal cancers, respectively. These genes were said to have a 90% chance of being true cancer genes, that is, harboring mutations at a frequency significantly greater than expected by chance, based on the FDR approach (that is, FDR $\leq$ 10%). Reassuringly, 6 genes known to be mutated in these cancer types appear at the top of these lists (1 in breast and 5 in colon cancer). Extrapolating from these lists to the entire genome, the authors estimate that the total number of genes harboring important somatic mutations in breast and colon cancer, respectively, exceeds 189 and 107 genes, with the typical tumor carrying 14 and 20 mutations. These observations are of great interest because the number of genes is much higher than previously thought. However, this analysis raises two methodological concerns that, when addressed, eliminate the statistical evidence for almost all of the yet unknown candidate cancer genes.

First, the authors incorrectly apply the FDR formula. The formula requires the tail probabilities [Prob($X \geq T$)] as input, but Sjöblom *et al*. instead use the point probabilities [Prob($X = T$)]. Consequently, their probabilities are smaller than they should be and therefore falsely appear to be more significant. When *P* values rather than point probabilities are used, the number of candidate genes falls from 122 to 6 in breast cancer and from 69 to 28 in colorectal cancer.

Second, the analysis is highly sensitive to the background mutation rate μ used in the statistical model (see Supporting Online Material). Different tumors and cell lines may have different background mutation rates, and accurate estimation of μ requires large amounts of sequence data generated from the same tumor population. Sjöblom *et al*. estimated μ based on a different, smaller data set. However, an estimate based on their own data yields substantially higher mutation rates—by factors of about 1.9 and 1.4 in breast and colorectal cancers, respectively (estimated in two ways; see SOM). If these rates are inserted into the analysis, the number of candidate genes falls to only 1 for breast cancer and 11 for colorectal cancer. Only four of these genes were not previously reported as mutated in cancer.

We also note that the analysis assumes that μ is constant across the genome. It is well known that the germline mutation rate shows regional variation (see SOM), and similar variation could be estimated in cancer from silent mutations in adjacent sites. Such variation would bias the discovery screen to select genes with higher background mutation rates; therefore, an increased effective value of μ should be used in calculating significance. Allowing for plausible variation among genes (CV = 0.4) would increase the effective value of μ by a factor of more than 1.3. The candidate lists would be reduced to only known cancer genes.

We note that the authors have recently performed a simulation study (*3*) based on the empirical Bayes or plug-in approach of Efron *et al*. (*4*) as an alternative way to estimate the FDR of their gene lists. The results are said to indicate that the results of Sjöblom *et al*. are conservative, that is, that the true FDR is even lower than 10%. However, we have discovered that their simulation study contains a subtle but important statistical shortcoming (SOM). Specifically, their analysis uses a score (CaMP score) for each gene that is highly sensitive to the presence of true cancer genes in the data. Therefore, a simulation that assumes no true cancer genes cannot be used to estimate the FDR in settings in which even a single true cancer gene exists. Replacing their CaMP score with one that does not suffer from this functional dependency among genes yields much higher FDRs.

We emphasize that the mathematical shortcomings discussed above do not simply reflect different but reasonable approaches to the analysis but are fundamental statistical problems.

We also suggest other statistical tests to detect candidate cancer genes, some of which are more powerful than the one above (see SOM, Appendices A to D). Using our estimated background mutation rates, even these more powerful tests yield few candidate genes.

After correcting the statistical analysis and using a background mutation rate that better fits the data, one cannot conclude that the ~200 candidate genes reported in Sjöblom *et al*. have >90% probability of being cancer-related. The issue is simply one of statistical power: Much larger sample sizes are required to detect cancer genes. With smaller sample sizes, most candidate genes are expected to be false positives. Nevertheless, we strongly support the authors' experimental approach and urge its adoption in future large-scale cancer genome sequencing efforts. We suspect that there are indeed many more important cancer genes waiting to be discovered, some of which may well be on the lists of Sjöblom *et al*. In the end, statistical validation of a candidate gene will require study of large samples to show such properties as a high frequency of mutations and a high ratio of nonsynonymous to synonymous mutations.

[1]Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02142, USA. [2]Department of Statistics, Stanford University, Stanford, CA 94305, USA. [3]Dana-Farber Cancer Institute, Boston, MA 02115, USA. [4]Department of Medicine, Children's Hospital Boston, Boston, MA 02115, USA. [5]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA. [6]Harvard Medical School, Boston, MA 02115, USA. [7]Department of Health Research and Policy, Stanford University, Stanford, CA 94305, USA. [8]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: gadgetz@broad.mit.edu

References and Notes

1. T. Sjöblom *et al*., *Science* **314**, 268 (2006).
2. Y. Benjamini, Y. Hochberg, *J. Roy. Statist. Soc. Ser. B. Methodological* **57**, 289 (1995).
3. G. Parmigiani *et al*., Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 126, www.bepress.com/jhubiostat/paper126 (December 2006).
4. B. Efron *et al*., *J. Am. Stat. Assoc.* **96**, 1151 (2001).
5. This work was supported in part by the Broad Institute of Harvard and MIT and by grants from the National Cancer Institute, the National Human Genome Research Institute, the National Institute of General Medical Sciences, NSF, and Howard Hughes Medical Institute. H. H. was supported by a Stanford Graduate Fellowship.

Supporting Online Material

www.sciencemag.org/cgi/content/full/317/5844/1500b/DC1
SOM Text
Figs. S1 and S2
Tables S1 to S7
Data Tables
References

11 December 2006; accepted 22 August 2007
10.1126/science.1138764