

The following resources related to this article are available online at www.sciencemag.org (this information is current as of November 15, 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/317/5844/1500c>

Supporting Online Material can be found at:

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

This article **cites 6 articles**, 2 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/317/5844/1500c#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”

Alan F. Rubin¹ and Phil Green^{1,2*}

Sjöblom *et al.* (Research Article, 13 October 2006, p. 268) reported many new genes with an apparent significant excess of mutations in breast and colorectal cancer. Reanalysis of their data with more appropriate statistical methods and background mutation rate assumptions reveals that few if any of these genes have significantly elevated mutation rates.

Sjöblom *et al.* (1) found a total of 189 genes with an apparent statistically significant excess of mutations in breast and colorectal tumors and concluded that the array of mutations selected in cancer growth is much larger than previously suspected. However, in calculating false discovery rates (FDRs), the estimated fraction of false positives among the genes reported at a given threshold, they incorrectly substituted probabilities of the specific observed mutation patterns for each gene in place of *P* values [which are required by FDR theory (2)]. This results in substantially underestimating the FDRs and overstating the significance of many genes. Correcting this, while retaining Sjöblom *et al.*'s assumed background mutation rate of 1.2 per megabase of tumor DNA, sharply reduces the number of genes that are significant at a 10% FDR (Fig 1).

However, we also question the assumed background mutation rate in (1). It is based on two much smaller studies, one of which (3) analyzed only 3.2 Mb of sequence data in colorectal cancer samples and obtained a very broad rate confidence interval of 0.22 to 2.5/Mb and the other (4) 518 kinase genes in 25 breast cancer samples, which (after eliminating two outlier samples) yielded a rate of 1.0/Mb. Because 367 of the kinase genes were also sequenced in Sjöblom *et al.*'s discovery screen, we could test consistency between studies by computing the discovery screen rates in these genes. We find a rate of 2.4/Mb in the breast cancer discovery screen, significantly higher ($P < 0.001$) than in (4). This discrepancy may reflect protocol or tumor sample differences between studies but in any case indicates that background rates should be determined separately for each study.

Valid background rates for the present study could be determined from mutation data at synonymous or other unselected sites in the same samples. In the absence of such data, the best available choice appears to be the discovery screen rate. Using this, we find that only a handful of

genes (*TP53* in breast cancer, and *APC*, *KRAS*, *TP53*, *SMAD4*, and *FBXW7* in colorectal cancer), all of them previously known to be mutated at high frequency in these cancers, are significant at an FDR of 10% (Fig. 1). Because some discovery screen mutations do reflect selection, using the discovery screen rate is conservative (i.e., will tend to underestimate the number of significant genes). However, the above conclusion is reasonably robust to the specific rate assumption: A smaller value of 2.0/Mb, which is well within the confi-

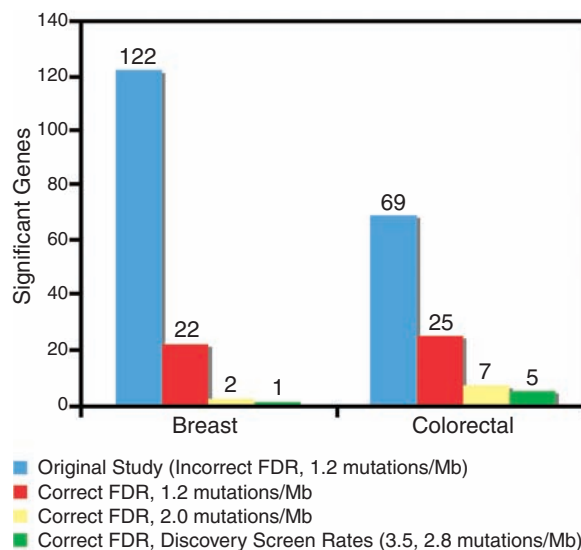


Fig. 1. Number of genes significant at an FDR of 10%, as originally found in (1) and upon reanalysis (8). Significant genes using the discovery screen rates (green bars) are *TP53* (breast cancer) and *APC*, *KRAS*, *TP53*, *SMAD4*, and *FBXW7* (colorectal cancer).

dence interval from (3), yields only three additional genes (Fig. 1). If the true background rate is substantially lower than this, the number of significant genes would of course go up, but there is no basis at present for presuming that a much lower rate is applicable. Tables S1 and S2 provide revised *P* values and FDR-based scores for each gene, for each of the background mutation rates shown in Fig. 1.

After eliminating the significant genes noted above, the mutation rate in the validation screen of (1) is still significantly higher than the overall

discovery screen rate (5.2/Mb versus 3.5/Mb for breast cancers; 4.5/Mb versus 2.8/Mb for colorectal cancers). The higher validation screen rate could reflect weak positive selection for cancer growth mutations, distributed over many genes. However, we believe a more likely possibility is that there is gene-to-gene heterogeneity in the strength of purifying selection and/or in the underlying (neutral) mutation rate, such that genes with intrinsically higher background rates are more likely to have been identified in the discovery screen, and their rates dominate the validation screen estimate. The existence of both kinds of heterogeneity is well established for germline mutations (5, 6).

Thus, the data of Sjöblom *et al.* do not appear to implicate any additional cancer genes beyond the handful known from previous studies. We cannot rule out the possibility that there are additional genes with subtle effects, but if they exist, establishing their importance will require a substantially larger set of samples, and/or alternative experimental approaches. We view the Sjöblom *et al.* study as a prelude to the Cancer Genome Atlas project, which seeks to identify causal mutations for a large number of cancers. Our analyses highlight the crucial importance of determining valid background rates for that project.

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. T. L. Wang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 3076 (2002).
4. P. Stephens *et al.*, *Nat. Genet.* **37**, 590 (2005).
5. Mouse Genome Sequencing Consortium, *Nature* **420**, 520 (2002).
6. W. H. Li, *Molecular Evolution* (Sinauer Assoc., Sunderland, MA, 1997).
7. F. Antequera, *Cell. Mol. Life Sci.* **60**, 1647 (2003).
8. We estimated relative rates for different substitution types as described (2), but with the following differences: We used only discovery screen data, we estimated distinct rates for C's and for G's within CpG and TpC / GpA dinucleotides, and we estimated distinct rates for CpGs inside and outside CpG islands (7). We then used these rates to estimate for each gene, based on its sequence composition, the expected total number, λ , of mutations in the discovery and validation screens combined. The *P* value for a gene having $n > 0$ observed mutations (in both screens combined) is then given by $\sum_{i=n}^{\infty} q_i r_i$ where $q_i = \binom{\lambda}{i} e^{-\lambda}$ is the Poisson probability, and $r_i \approx \left(1 - \left(\frac{\lambda}{35}\right)^i\right)$ is the probability that at least one of the *i* mutations occurs during the discovery screen. We assumed that sequencing failures were distributed proportionately across all genes.
9. This work was supported by an NIH training grant (A.F.R.) and by the Howard Hughes Medical Institute.

Supporting Online Material

www.sciencemag.org/cgi/content/full/317/5844/1500c/DC1
 Tables S1 and S2
 18 December 2006; accepted 2 July 2007
 10.1126/science.1138956

¹Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA. ²Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA.

*To whom correspondence should be addressed. E-mail: phg@u.washington.edu