Supporting Online Material for

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

Gad Getz,* Holger Höfling, Jill P. Mesirov, Todd R. Golub,
Matthew L. Meyerson, Robert Tibshirani, Eric S. Lander

*To whom correspondence should be addressed. E-mail: gadgetz@broad.mit.edu

Published 14 September 2007, *Science* **317**, 1500b (2007)
DOI: 10.1126/science.1138764

This PDF file includes:

Other Supporting Online Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/317/5844/1500b/DC1)

Supporting Online Material for Technical Comment on Sjöblom et al.

Gad Getz[*], Holger Höfling[*§], Jill P. Mesirov, Todd R. Golub,
Matthew L. Meyerson, Robert Tibshirani and Eric S. Lander

1 Background

The recent paper by Sjöblom et al. [1] identified somatic mutations in cancer and listed 122 and 69 candidate cancer genes (*CAN*-genes) in breast and colorectal cancers, respectively. Their experiment had two screens: a *discovery screen* in which 13,023 genes were sequenced in 11 samples, and a *validation screen* in which only the genes that had at least one somatic mutation in the discovery screen were sequenced in 24 additional samples (see [1] for further details).

Among the mutated genes are ones already implicated in breast and colon cancer. Table S1 shows the breakdown of the identified mutations according to tissue type, screen and known vs. unknown cancer genes.

Breast		Colon	
Known cancer genes	TP53	Known cancer genes	TP53, APC, KRAS, FBXW7, SMAD4
Discovery: known	Genes: 1	Discovery: known	Genes: 5
	Mutations: 10		Mutations: 26
Discovery: unknown	Genes: 671	Discovery: unknown	Genes: 514
	Mutations: 722		Mutations: 548
Discovery: ratio	722/671=1.08 mut/gene	Discovery: ratio	548/514=1.07 mut/gene
Validation: known	Genes: 1	Validation: known	Genes: 5
	Mutations: 8		Mutations: 47
Validation: unknown	Genes: 136	Validation: unknown	Genes: 100
	Mutations: 180		Mutations: 130
Validation: ratio	180/136=1.32 mut/gene	Validation: ratio	130/100=1.3 mut/gene
No. of *CAN*-genes	122	No. of *CAN*-genes	69
Mutation enrichment ratio	1.32/1.08=1.23	Mutation enrichment ratio	1.3/1.07=1.21

Table S1: Breakdown of mutations in breast and colon cancer as found by Sjöblom et al. (constructed from table S4 in Sjöblom et al.)

In order to identify cancer related genes, Sjöblom et al. applied a statistical model to distinguish between genes which harbor "driver" mutations that conferred a growth or maintenance advantage to the tumors from ones which carry "passenger" mutations due to random mutation events. The candidate cancer genes (*CAN*-genes) are ones that carry more mutations than one would expect by chance, after correcting for multiple hypothesis.

The statistical model applied by Sjöblom et al. (following Greenman et al. [2]) is based on background mutation rates for 7 different mutation categories defined by the mutated base and its surrounding (Table S2) . Importantly, these rates were estimated by multiplying an average background mutation rate of

[*] These authors contributed equally
[§] This material is based upon work supported under a Stanford Graduate Fellowship

$\mu=1.2x10^{-6}$ which was estimated using other studies ([3, 4] and reference S10 in [1]) with the empirical ratios among the different mutation rates seen in this study[1].

Mutation rates (x 10^{-6})	C, G in CpG	C in TpC, G in GpA	A	Remaining C	Remaining G	T	INS/DEL/DUP
Breast cancer	2.99	2.48	0.76	1.38	1.07	0.30	0.55
Colon cancer	7.73	0.96	0.56	0.95	0.85	0.51	0.55

Table S2: Estimated background mutation rates in breast and colon cancers. Rates are in units of mut/Mb.

For any gene, *g,* the authors calculated the probability to obtain the number of mutations observed across the 11 discovery and 24 validation samples. To this end, they first calculated for each category, *i*, the probability of observing x_{gi} mutations among the N_{gi} possible sites for this mutation category[2] in gene *g*. Assuming all mutations are generated according to the background mutation rates, f_i, this probability is given by,

$$b_{gi} = b(x_{gi}; N_{gi}, f_i) = \binom{N_{gi}}{x_{gi}} f_i^{x_{gi}} \left(1 - f_i\right)^{N_{gi} - x_{gi}} . \quad (1)$$

The probability for gene *g* was then calculated by taking the product of b_{gi} for the 7 mutation categories, i.e. $p_g = \prod_{i=1}^{7} b_{gi}$. In order to correct for multiple hypotheses, the authors applied the Benjamini-Hochberg False Discovery Rate procedure (BH-FDR)[5] to the resulting $\{p_g\}_{g=1}^{n}$ where n=13,023 (the total number of genes[3]). The FDR procedure yields a q-value for each gene, q_g, which is an upper bound on the expected fraction of false positives among all genes with equal or lower q-value. The authors defined a CaMP score for each gene as $-\log_{10}(q_g)$.

The 122 *CAN*-genes in breast cancer and 69 *CAN*-genes in colon cancer (listed in table 3 in [1]) were obtained by selecting those genes with CaMP scores greater than 1, which is equivalent to controlling the FDR at a rate of 0.1. At least 90% of these genes are said to have mutation rates greater than the background mutation rates indicating they harbor mutations which are selected for.

2 Two shortcomings

We identify two shortcomings that generate overly significant CaMP scores. Correcting these shortcomings decreases the CaMP scores and yields much shorter lists of candidate cancer genes. The issues are:

(i) The BH-FDR procedure is wrongly applied to point probabilities rather than p-values. Consequently, the CaMP scores are incorrect and overly significant.

(ii) The background mutation rates used in the model are too low for several reasons. Increasing the background mutation rates decreases the significance of each gene and, therefore, yields fewer *CAN*-genes. The rates are underestimated due to the following:

a. They do not fit the observed number of mutations (which are higher).

[1] The ratios among the mutation rates were estimated separately for breast and colon cancer and counted both mutations observed in the discovery and validation screens.

[2] In fact, N_{gi} represents the possible sites times the fraction of bases sequenced in gene g.

[3] The probability, p_g, for genes that did not have at least one mutation in the discovery screen and one mutation in the validation was set to 1.

b. Considering a variation in background mutation rates across genes results in an enrichment of genes with higher mutation rates among those that pass the screens. Hence higher mutation rates should be used when assessing the significance of these genes.

The following subsections elaborate on each of these issues.

2.1 P-value calculation

The probability calculated for each gene, p_g, is taken as the product of point probabilities for the 7 individual mutation types. In contrast to Greenman et al. [2] who used these point probabilities to calculate log-likelihood ratios, Sjöblom et al. use them as p-values in the BH-FDR multiple hypothesis correction procedure. This step is incorrect since these point probabilities are *not* p-values as they do not measure how likely it is to obtain a set of mutations which are at least as extreme as the observed ones by chance.

In order to apply the BH-FDR procedure, one has to calculate a p-value for each gene. One can calculate p-values for variety of statistical tests (see Appendices A, B and C). The closest test to Sjöblom et al.'s methodology is to calculate, for each gene, the probability to obtain p_g or less under the null hypothesis. Only then can we apply the BH-FDR procedure to obtain q-values and CaMP scores. We also applied alternative statistical tests – a more standard likelihood-ratio test, a test that is uniformly most powerful against uniformly increased mutation rates in all categories as well as a plug-in estimator for the FDR using the s_g statistic (see Appendices A, B and C).

An additional minor issue is that the BH-FDR procedure as described in Sjöblom et al. is missing a monotonization step. After sorting the p-values and multiplying each p-value by the total number of hypothesis tested and dividing by its rank, as performed by Sjöblom et al., one needs to make sure that a gene with a lower p-value is not assigned a higher q-value. This is done by setting $q_{(t)} = \min(q_{(t)}, q_{(t+1)})$ where $q_{(n+1)}=1$ stepping down from the largest value, $t=n,n-1,\ldots,1$. This procedure can only decrease the q-values making genes more significant but in practice it has little affect on the top CaMP scores.

We reanalyzed the data in Sjöblom et al. using p-values, instead of point probabilities, and calculated corrected CaMP scores. Table S4 contains the corrected CaMP scores for a range of background mutations rates. The results for the original mutation rates are in the 4th column, titled "CaMP (1.0)" (the other columns are referred to in the next subsection). The table includes all the genes that are significant in at least one of the columns. Genes that pass significance, i.e. with CaMP score greater than 1, are marked with a red background. As a result of this reanalysis, the list of candidate cancer genes is reduced from 122 to 6 in breast cancer and from 69 to 28 in colon cancer. Moreover, many of the genes that pass significance are not highly significant with CaMP scores less than 1.3 (equivalent to q-value of 0.05). The results of the other tests can be found in Table S7.

2.2 Underestimated mutation rates

The calculation of probabilities and p-values is based on the background mutation rates, $\{f_i\}$. It is important to test the robustness of the results, i.e. the *CAN*-gene lists, with respect to changes in these mutation rates. Naturally, we are more concerned with underestimation of the mutation rates as it can lead to additional false positives.

To test the effect of having underestimated mutation rates we calculated the corrected CaMP scores using a range of mutation rates obtained by multiplying the given rates (Table S2) by a scaling factor

between 1 and 2 (in steps of 0.1). From Table S4, one can see that the lists of candidate genes are very sensitive to changes in mutation rates – using rates which are only 2 fold higher (last column) reduces the lists from 6 to 1 in breast and from 28 to 6 in colon. These remaining genes are known to be associated with the studied cancer types (the last of the 6 significant genes in colon, EPHA3, is known to be mutated in sporadic tumors). If these mutation rates were the case, then no new cancer genes would have passed significance in this study. That is not to say that there are no new cancer genes in the list but only that the number of samples used in this study are not sufficient to conclude that any new gene is significantly associated with these cancer types. Obviously, if additional data, such as synonymous somatic mutation rates, would yield lower estimates than the ones reported in Sjöblom et al., then the number of candidate cancer genes will increase.

The question that remains is whether the background mutation rates are underestimated by a factor of 2. In the next Subsections we consider different mechanisms that can lead to underestimated mutation rates.

2.2.1 Inaccurate background mutation rates

As described in Section 1, the relative ratios among the different mutation rates are determined from the observed data whereas the absolute scale is estimated based on external studies. This made us suspect that the mutation rates might be biased or inaccurate. There are several reasons for these external studies not to be reliable sources for background mutation rates: (i) Only a small subset of genes were studied. (ii) Lower number of nucleotides were sequenced (iii) Different samples were analyzed and (iv) Different screening procedure may have been applied (99.84% of initially identified mutations were filtered out in the discovery screen and 99.72% in the validation screen in Sjöblom et al.; small procedural changes may have a big effect). In fact, as elaborated below, if one estimates the mutation rates solely based on the observed mutations in the discovery screen, the mutation rates are higher by a factor of 1.9 in breast cancer and 1.43 in colon cancer. Consequently, the CaMP scores decrease and the lists of candidate genes are shortened such that they include only one gene in breast cancer and 11 genes in colorectal cancer (see Table S7 for results of other tests).

Table S5 lists statistics regarding the number of mutations seen in the discovery and validation screens in the experiment and in computer simulations. It is important to note that we excluded mutations that were found in genes known to be mutated in the specific cancer types (listed in Table S1). We simulated a 1000 experiments for the two cancer types, each as described in Sjöblom et al., including a discovery screen of 11 samples and a validation screen of an additional 24 samples. All the mutations were randomly generated according to the original background mutation rates (Table S2). For each experiment we calculated the same statistics as for the experimental data, such as the number of mutations seen in each screen, the ratio between the numbers of mutations and mutated genes and the number of *CAN*-genes that would have been listed by the method described in Sjöblom et al. We also report the top 10^{th} percentile of the number of *CAN*-genes in order to get a feel whether the observed number of *CAN*-genes could have been observed by chance. The results of this set of simulations appear under Simulation I in Table S5. Clearly, the numbers of mutations seen in the discovery screens in Simulation I are lower compared to the experimental ones – 380 vs. 722 in breast cancer and 384 vs. 548 in colon cancer. This indicates that the background mutation rates are too low.

In Simulation II, we increased the mutations rates to match the observed number of mutations in the discovery screen by multiplying the original rates by the ratio between number of mutations in the discovery screen in the experiment (after excluding mutations in known cancer genes) and in Simulation I (see Table S5). The scaling factors that were used are 1.9 and 1.43 for breast and colon cancer,

respectively. In order to address the possibility that additional true cancer genes exist which inflate our estimated background rates, we used a different method which is based only on genes which were not mutated in the validation screen (see Appendix D). The new scaling factors were very close to the ones above hence we continued using the initial values.

To test the effect of using these scaling factors on the candidate genes, one can read off Table S4 the lists in the appropriate columns which indicate that only one gene in breast cancer (TP53) and 11 in colon cancer (rounding 1.43 down to 1.4) are statistically significant. As expected, the results of Simulation II match the experimental ones in terms of number of mutations in the discovery screen but still have fewer mutations in the validation screen and also in other statistics such as mutations per gene and enrichment ratios. Next we consider the effect of having variable mutation rates across genes.

2.2.2 Variable mutation rates across genes

It is biologically plausible that somatic background mutation rates vary across genes. In fact, recent studies describe variation in germline mutation rates both in human [6-9] and mouse [10]. The coefficient of variation of these rates is estimated to be on the order of 0.2-0.3 (see Table S3). These estimates are based on the correlation between interspecies divergence and intraspecies diversity. In brief, denote by $\pi(x)$ the local rate of intraspecies differences and by $\delta(x)$ the local rate of interspecies differences. The coefficient of variation of the germline mutation rate, $CV_x[\mu]$, can be estimated by,

$$CV_x[\mu]=\sqrt{\frac{Var_x[\mu]}{E_x[\mu]^2}}=\sqrt{\frac{Cov_x[\delta,\pi]}{E_x[\delta]E_x[\pi]}} \qquad (2)$$

Table 3 describes the estimated values for $E_x[\pi]$, $E_x[\delta]$ and $CV_x[\mu]$ using different bin sizes. We do not have enough data to directly estimate the variation of somatic mutation rates in human cancers. However, collecting statistics regarding silent and non-coding somatic mutations across the genome will help get direct measurement of this variability.

Bin size	$E_x[\pi]$	$E_x[\delta]$	$Cov_x[\delta,\pi]$	$Var_x[\mu]/E_x[\mu]^2$	$CV_x[\mu]$
0.5 kb	0.0010	0.0136	1.3×10^{-6}	0.092	0.30
1 kb	0.0010	0.0133	9.1×10^{-7}	0.068	0.26
10 kb	0.0010	0.0131	6.6×10^{-7}	0.050	0.22
100 kb	0.0010	0.0131	5.5×10^{-7}	0.041	0.20

Table S3: Estimated coefficient of variation of mutation rates using human and chimpanzee data [6].

Any variation in somatic mutation rates (i.e. $CV[\mu]>0$) has the effect that the genes that are selected by the discovery and validation screens will be biased toward those with somewhat higher mutation rates. The distribution of mutation rates of genes that pass the screens will thus be shifted towards higher values, even if none are related to cancer. To correct for this, one should use higher *effective* background mutation rates when calculating the p-values and CaMP scores. In principle, one could try to estimate a specific mutation rate for each gene. However, this would require sequencing the gene in a huge number of samples. In practice, it would reasonably suffice to know the CV and then to estimate corresponding effective mutation rates by simulation. These effective rates can then be used for calculating the p-values and CaMP scores.

(a) Breast

Original Rank	Gene	Size	CaMP (1.0)	CaMP (1.1)	CaMP (1.2)	CaMP (1.3)	CaMP (1.4)	CaMP (1.5)	CaMP (1.6)	CaMP (1.7)	CaMP (1.8)	CaMP (1.9)	CaMP (2.0)
# of CAN-genes			6	5	3	3	3	3	3	1	1	1	1
1	TP53	1254	>6	>6	>6	>6	>6	>6	>6	>6	>6	>6	>6
2	FLJ13479	1619	1.80	1.66	1.51	1.39	1.30	1.19	1.10	1.00	0.89	0.80	0.71
3	C14orf155	1813	1.80	1.66	1.51	1.39	1.30	1.19	1.10	1.00	0.89	0.80	0.71
4	FLNB	7575	1.23	1.02	0.83	0.69	0.59	0.46	0.40	0.32	0.25	0.18	0.11
5	ATP8B1	3612	1.23	1.05	0.90	0.76	0.64	0.52	0.41	0.32	0.25	0.18	0.11
6	MYH1	5876	0.97	0.84	0.73	0.63	0.53	0.46	0.40	0.32	0.25	0.18	0.11
7	SPTAN1	7517	0.92	0.76	0.64	0.55	0.47	0.36	0.27	0.19	0.11	0.05	0.00
8	DBN1	1506	1.03	0.91	0.80	0.69	0.59	0.46	0.40	0.32	0.25	0.18	0.11
9	CUBN	9237	0.87	0.76	0.62	0.51	0.43	0.35	0.25	0.18	0.11	0.04	-0.03
10	GRIN2D	2369	0.97	0.84	0.73	0.63	0.53	0.46	0.40	0.32	0.25	0.18	0.11
11	TECTA	6031	0.87	0.76	0.64	0.54	0.46	0.35	0.26	0.19	0.11	0.05	0.00
12	KIAA1632	3925	0.92	0.76	0.66	0.55	0.47	0.37	0.28	0.21	0.17	0.12	0.04
13	SIX4	1950	0.97	0.84	0.73	0.63	0.53	0.46	0.40	0.32	0.25	0.18	0.11
14	KIAA0934	4297	0.87	0.76	0.64	0.55	0.47	0.36	0.27	0.19	0.12	0.05	0.00
15	LRRFIP1	1976	0.97	0.84	0.73	0.63	0.53	0.46	0.40	0.32	0.25	0.18	0.11
16	GLI1	2048	0.92	0.76	0.66	0.55	0.47	0.37	0.28	0.21	0.17	0.12	0.05
17	OTOF	4606	0.84	0.71	0.60	0.51	0.43	0.35	0.26	0.19	0.11	0.05	0.00
18	CDH20	2453	0.87	0.76	0.64	0.55	0.47	0.36	0.27	0.20	0.13	0.06	0.00
19	GSN	2504	0.87	0.76	0.64	0.55	0.47	0.36	0.27	0.19	0.12	0.06	0.00

(b) Colon

Original Rank	Gene	Size	CaMP (1.0)	CaMP (1.1)	CaMP (1.2)	CaMP (1.3)	CaMP (1.4)	CaMP (1.5)	CaMP (1.6)	CaMP (1.7)	CaMP (1.8)	CaMP (1.9)	CaMP (2.0)
# of CAN-genes			28	26	23	16	11	10	10	8	7	6	6
1	APC	6562	>6	>6	>6	>6	>6	>6	>6	>6	>6	>6	>6
2	KRAS	621	>6	>6	>6	>6	>6	>6	>6	>6	>6	>6	>6
3	TP53	1209	>6	>6	>6	>6	>6	>6	>6	>6	>6	>6	>6
4	FBXW7	1863	3.33	3.16	2.99	2.84	2.70	2.57	2.45	2.34	2.23	2.13	2.03
5	SMAD4	1565	2.90	2.73	2.58	2.45	2.32	2.20	2.09	1.98	1.89	1.79	1.70
6	EPHA3	2771	2.47	2.30	2.14	2.00	1.86	1.74	1.62	1.51	1.40	1.30	1.21
7	MLL3	12500	1.50	1.34	1.18	1.01	0.87	0.75	0.64	0.54	0.41	0.34	0.25
8	PKHD1	11206	1.41	1.23	1.08	0.93	0.84	0.73	0.63	0.53	0.41	0.34	0.26
9	EPHB6	2758	1.77	1.61	1.47	1.32	1.20	1.10	1.01	0.92	0.81	0.72	0.63
10	GUCY1A2	1628	1.92	1.78	1.65	1.53	1.42	1.32	1.22	1.13	1.05	0.97	0.90
11	TBX22	1159	1.77	1.61	1.49	1.39	1.29	1.20	1.12	1.04	0.96	0.89	0.83
12	ADAMTSL3	5099	1.47	1.34	1.19	1.06	0.93	0.84	0.73	0.63	0.56	0.48	0.41
13	SMAD2	1251	1.62	1.49	1.38	1.28	1.18	1.10	1.01	0.92	0.82	0.75	0.68
14	OBSCN	18582	0.84	0.59	0.39	0.21	0.04	-0.07	-0.17	-0.28	-0.39	-0.48	-0.58
15	ABCA1	6903	1.16	1.02	0.88	0.76	0.64	0.55	0.46	0.36	0.28	0.21	0.13
16	TGFBR2	1730	1.47	1.34	1.19	1.06	0.96	0.86	0.78	0.69	0.62	0.55	0.48
17	TCF7L2	1669	1.47	1.34	1.19	1.06	0.96	0.86	0.78	0.69	0.62	0.55	0.48
18	ADAMTS18	1963	1.35	1.22	1.09	0.98	0.87	0.77	0.69	0.63	0.56	0.48	0.41
19	C10orf137	3277	1.22	1.14	1.01	0.92	0.82	0.72	0.63	0.54	0.45	0.37	0.30
20	GNAS	1123	1.41	1.26	1.15	1.02	0.93	0.84	0.73	0.65	0.58	0.51	0.44
21	HIST1H1B	661	1.47	1.34	1.20	1.11	1.02	0.95	0.88	0.81	0.74	0.68	0.62
22	RUNX1T1	1904	1.16	1.03	0.92	0.82	0.73	0.64	0.56	0.50	0.41	0.34	0.26
23	MMP2	2056	1.09	0.96	0.86	0.76	0.65	0.56	0.49	0.43	0.37	0.30	0.23
24	SYNE1	24397	0.27	0.12	-0.03	-0.14	-0.25	-0.36	-0.46	-0.54	-0.62	-0.70	-0.78
25	RET	2918	0.99	0.87	0.76	0.67	0.59	0.50	0.42	0.34	0.28	0.21	0.13
26	SEC8L1	3004	0.94	0.83	0.72	0.62	0.54	0.46	0.38	0.31	0.24	0.18	0.11
27	P2RY14	663	1.22	1.14	1.02	0.93	0.84	0.75	0.69	0.63	0.56	0.49	0.44
28	LGR6	2808	0.93	0.81	0.70	0.62	0.53	0.45	0.38	0.30	0.23	0.16	0.10
29	TTLL3	1076	1.22	1.14	1.02	0.93	0.84	0.75	0.66	0.59	0.53	0.47	0.41
30	PTPRD	4924	0.84	0.71	0.59	0.48	0.38	0.28	0.20	0.11	0.04	-0.04	-0.10
31	SDBCAG84	1254	1.22	1.14	1.02	0.93	0.84	0.75	0.66	0.59	0.52	0.46	0.40
32	MCP	1110	1.19	1.10	1.01	0.92	0.84	0.75	0.66	0.59	0.53	0.47	0.41
33	ADAM29	1470	1.15	1.03	0.93	0.85	0.77	0.70	0.63	0.54	0.45	0.39	0.34
34	LOC157697	1166	0.98	0.87	0.79	0.71	0.64	0.56	0.49	0.43	0.37	0.32	0.26
35	EVL	1345	0.94	0.83	0.75	0.67	0.59	0.52	0.46	0.38	0.33	0.27	0.22
36	ZNF442	1764	1.10	0.93	0.88	0.79	0.71	0.64	0.56	0.50	0.37	0.34	0.26

Table S4: Corrected CaMP scores using background mutation rates which are scaled by a factor denoted in parentheses in the top row. Red background indicates statistical significant genes (CaMP >1). The second row shows the total number of significant genes.

Experiment			
Breast		Colon	
Discovery: unknown	Genes: 671	Discovery: unknown	Genes: 514
	Mutations: 722		Mutations: 548
Discovery: ratio	722/671=1.08 mut/gene	Discovery: ratio	548/514=1.07 mut/gene
Validation: unknown	Genes: 136	Validation: unknown	Genes: 100
	Mutations: 180		Mutations: 130
Validation: ratio	180/136=1.32 mut/gene	Validation: ratio	130/100=1.3 mut/gene
Mutation enrichment ratio	1.32/1.08=1.23	Mutation enrichment ratio	1.3/1.07=1.21
No. of *CAN*-genes	122	No. of *CAN*-genes	69
Simulation I			
Scaling factor:	**1.0**	**Scaling factor:**	**1.0**
Discovery: unknown	Genes: 371±18	Discovery: unknown	Genes: 375±19
	Mutations: 380±19		Mutations: 384±20
Discovery: ratio	1.025±0.008 mut/gene	Discovery: ratio	1.025±0.008 mut/gene
Validation: unknown	Genes: 36±6	Validation: unknown	Genes: 37±6
	Mutations: 40±7		Mutations: 40±7
Validation: ratio	1.1±0.06 mut/gene	Validation: ratio	1.1±0.05 mut/gene
Mutation enrichment ratio	1.07±0.06	Mutation enrichment ratio	1.07±0.05
No. of *CAN*-genes	2.0±2.4 (in 10% ≥ 5)	No. of *CAN*-genes	1.8±2.3 (in 10% ≥ 5)
Simulation II			
Scaling factor:	**722/380=1.9**	**Scaling factor:**	**548/384=1.43**
Discovery: unknown	Genes: 688±25	Discovery: unknown	Genes: 528±23
	Mutations: 720±26		Mutations: 546±24
Discovery: ratio	1.048±0.008 mut/gene	Discovery: ratio	1.035±0.008 mut/gene
Validation: unknown	Genes: 117±10	Validation: unknown	Genes: 70±8
	Mutations: 137±13		Mutations: 80±10
Validation: ratio	1.18±0.05 mut/gene	Validation: ratio	1.13±0.05 mut/gene
Mutation enrichment ratio	1.12±0.04	Mutation enrichment ratio	1.10±0.05
No. of *CAN*-genes	63±14 (in 10% ≥ 81)	No. of *CAN*-genes	14±8 (in 10% ≥ 25)
Simulation III			
Scaling factor:	**1.9**	**Scaling factor:**	**1.43**
Log-normal σ:	**0.58**	**Log-normal σ:**	**0.78**
Effective scaling factor:	3.31±0.17	Effective scaling factor:	3.78±0.29
Discovery: unknown	Genes: 677r±24	Discovery: unknown	Genes: 515±22
	Mutations: 720±27		Mutations: 546±24
Discovery: ratio	1.064±0.01 mut/gene	Discovery: ratio	1.06±0.012 mut/gene
Validation: unknown	Genes: 139±12	Validation: unknown	Genes: 99±10
	Mutations: 180±17		Mutations: 130±14
Validation: ratio	1.29±0.06 mut/gene	Validation: ratio	1.31±0.07 mut/gene
Mutation enrichment ratio	1.21±0.05	Mutation enrichment ratio	1.24±0.07
No. of *CAN*-genes	101±14 (in 10% ≥ 118)	No. of *CAN*-genes	56±11 (in 10% ≥ 71)

Table S5: Number of mutations and number of mutated genes found among genes that were not previously implicated in breast and colon cancer. The top section summarizes the experimental results and Simulation I,II and III summarize results from different simulation sets. Each simulation set uses different parameters, marked with bold font, and represents a 1000 simulation runs. The simulation results are reported as average and standard deviation over the 1000 simulation runs.

For example, if CV[μ]=0.4, the simulation results in Figure S1a show that the average increase in background mutation rates is 2.5-fold in breast and 1.9-fold in colon (relative to Sjöblom et al.'s mutation rates). Having 1.9-fold increased rates in colon cancer reduces the list of candidate genes to 6 genes, as presented in column "CaMP (1.9)" in Table S4b. Using the number of observed mutation in the validation screen, we were able to estimate the coefficient of variation in background mutation rates[6]. These levels of variation lead to effective background mutation rates that are more than 3-fold

higher than the original ones and yield lists of significant genes which include only genes already known to be mutated in cancer.

In more detail, to model the effect of variable mutation rates across genes we randomly generated gene-specific mutation rates by taking the ones used in Simulation II (after adjusting for the observed number of mutations in the discovery screen) and multiplying them by random gene-specific scaling factors drawn from a log-normal distribution with log mean of $-\sigma^2/2$ and log standard deviation σ. This distribution has a mean value of one to ensure the average mutation rate across all the genes is kept fixed[4]. Figure S1a shows the mean scaling factor across the genes that passed the screens as a function of σ (as measured in 1000 simulation runs). The scaling factors in Figure S1a represent the ratio to the original mutation rates and include the adjustments performed in Simulation II, which are reflected by the values for $\sigma=0$.

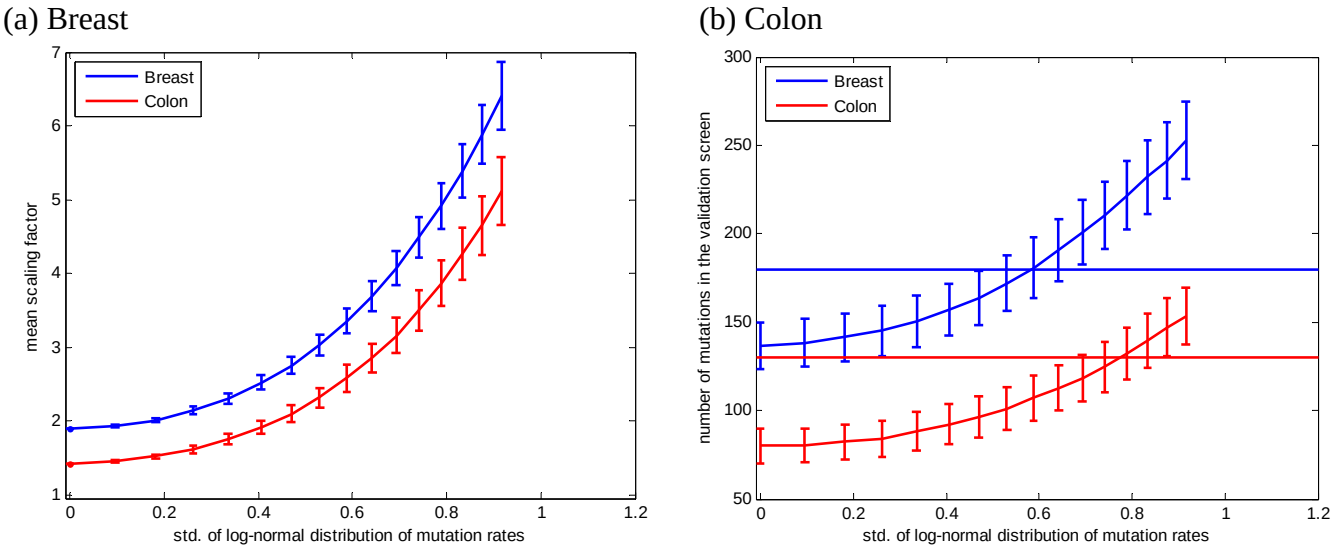


Figure S1: (a) Mean scaling factor across the genes that passed both screens as a function of σ, the standard deviation of the log-normal distribution of gene-specific scaling factors. The scaling factors are with respect to the original mutation rates. (b) Number of mutations in the validation screen as a function of σ. Blue lines are for breast cancer and red lines are for colon cancer. The horizontal line indicates the observed values.

Besides increasing the mean mutation rates, changing σ also affects the number of mutations observed in the validation screen (Figure S1b). This enables us to estimate σ by matching to the simulated and observed number of mutations in the validation screen[5]. The estimated values are 0.58 and 0.78 for breast and colon cancer, respectively. Simulation III, in Table S5, shows the resulting statistics when using these values. In this setting, the numbers of mutations match the observed numbers both in the discovery screen and the validation screen. It is striking, however, that the other statistics are also shifted such that they fit the observed ones. Particularly, the numbers of *CAN*-genes observed in Sjöblom et al. are not significantly higher than the simulated ones. This lends additional support for this model. Finally, the effect of enriching for genes with higher mutation rates gives rise to effective scaling factors of 3.31 and 3.78 in breast and colon cancer, respectively. Calculating the corrected CaMP scores using these scaling factors yields even shorter candidate lists which include TP53 in breast cancer and APC, KRAS and TP53 in colon cancer – all of which are already known to be mutated in the corresponding

[4] We actually divide by the gene-specific factors by their average to ensure the mean factor is precisely 1.
[5] We exclude mutations seen in genes known to be mutated in these cancer types.

cancer type. Of note, with these estimated levels of variation in background mutation rates, 90% of the genes have mutation rates within 2.6-fold and 3.6-fold from the average in breast and colon cancer, respectively. Combining Figure S1a and Table S4 shows that even lower levels of variation (σ=0.4) can increase the effective scaling factors to a level where only a few genes meet significance, all of which are known to be mutated in cancer. This level of variation is on the same scale as the reported coefficient of variation[6] in germline mutation rates.

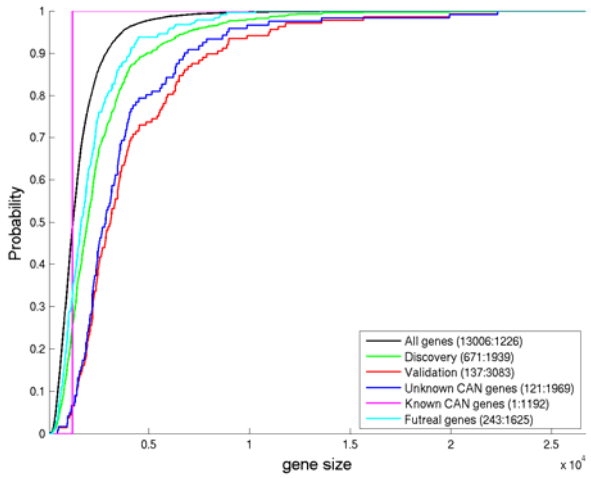
2.2.3 Variable mutation rates across samples

It is also reasonable to expect that background mutation rates vary across samples. This variation can contribute to larger error-bars on the p-values and can generate correlations among the p-values which need to be addressed when assessing statistical significance. We do not treat this type of variation in this Technical Comment.

3 Size distribution of the *CAN*-genes

In order to get another perspective on the *CAN*-genes, as listed by Sjöblom et al., and to confirm our belief that many of them may be false positives, we examined the size distribution of these genes. As one may expect, the screening process enriches for long genes as they are more likely to harbor random mutations. Figure S2 shows the cumulative distribution of gene sizes of the full set of genes, the genes that passed the discovery screen, those that passed both screens and the *CAN*-genes. Here, as before, we separated the known cancer genes from the other *CAN*-genes since their mutations are less likely to be due to random events. We also compare these distributions to that of genes known to be mutated in cancers taken from Table 1 in Futreal et al.[11]. One can clearly see the enrichment effect of the screening process. Despite the fact that the statistical model penalizes long genes, the *CAN*-genes are still enriched with longer genes compared to the ones described in Futreal et al.

(a) Breast



(b) Colon

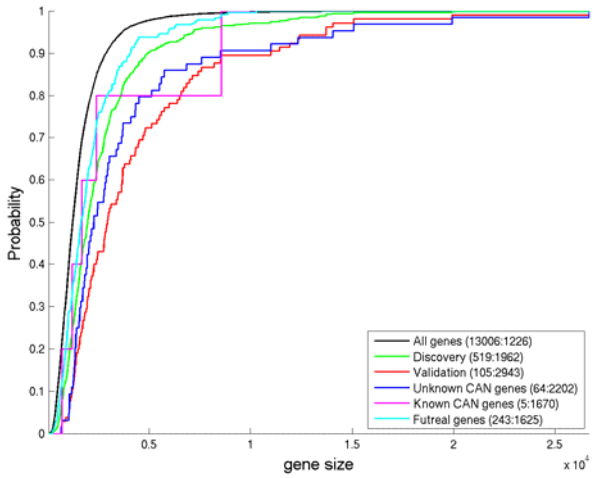


Figure S2: (a) Cumulative size distribution of genes in the breast cancer screens compared to all genes tested and known cancer genes. (b) Same for colon cancer. The numbers in the parentheses in the legend represent the population size and its median gene size.

[6] Note that the coefficient of variation and the standard deviation of a log-normal distribution are close for small values since $\mathrm{CV}=\sqrt{\exp[\sigma^2]-1} \approx \sigma + O(\sigma^2)$

Appendix A: A semi-exact calculation of the distribution of the product of probabilities, likelihood ratio test and UMP-test

In the model described by Sjöblom et al. [1] one calculates for each gene, p_g, the product of binomial probabilities for the seven mutation types. Our aim is to calculate the probability to obtain p_g or less by chance. It is convenient to work with a transformed test statistic, $s_g = -\ln\left(p_g\right) = -\ln\left[\prod_{i=1}^{7} b(x_{gi}; N_{gi}, f_i)\right] = \sum_{i=1}^{7} -\ln\left(b(x_{gi}; N_{gi}, f_i)\right) = \sum_{i=1}^{7} \beta_i^g$. Since under the null hypothesis, s_g is a sum of 7 independent random variables, its distribution is the convolution of the distributions of each of the individual random variables, β_i^g . The individual random variables are distributed according to,

$$\beta_i^g \sim \sum_{k=0}^{\infty} b\left(k; N_{gi}, f_i\right) \delta\left(x + \ln\left(b\left(k; N_{gi}, f_i\right)\right)\right), \qquad \text{(A1)}$$

where δ is the Dirac delta function. In practice, histograms are generated for each individual variable, according to (A1) and these are then convoluted to obtain the histogram describing the distribution of their sum. The p-value is the tail of the resulting histogram, from s_g to its end, Practically, one can calculate the p-value by taking one minus the left-tail, i.e. the probability to obtain a score smaller than the observed score. This semi-exact method approaches the exact value as the resolution of the histogram increases. We use histograms from 0 to 25 with bin size of 0.001. The advantage of using this semi-exact method is that it can scale up to handle many mutations when more samples are used.

Note that when the number of mutations is small one can calculate exactly the p-value by distributing mutations in the 7 different mutation types and summing the probabilities for all those cases in which the score is smaller than the observed score (which gives the left-tail). Then, the p-value is obtained by taking one minus this left-tail sum. The results of the semi-exact method and the exact one are nearly identical, giving the same significant genes.

In addition to the score used by Sjöblom et al., we explored other, more common, statistical tests in order to identify cancer genes. The likelihood-ratio test is the most widely used standard statistical test for this problem. In order to compute the test statistic, we need the maximum likelihood estimator (MLE) for the mutation rate in each category, which is $\hat{f}_{gi} = x_{gi} / N_{gi}$. Therefore, the log-likelihood-ratio test statistic is

$$LLRT_g = \sum_{i=1}^{7} \log\left(\binom{N_{gi}}{x_{gi}} \hat{f}_{gi}^{x_{gi}} (1-\hat{f}_{gi})^{N_{gi}-x_{gi}}\right) - \log\left(\binom{N_{gi}}{x_{gi}} f_i^{x_{gi}} (1-f_i)^{N_{gi}-x_{gi}}\right). \qquad \text{(A2)}$$

The distribution under the null hypothesis and the p-value can be calculated similar to the method above. This method yields 7 significant genes in breast cancer and 28 significant genes in colon cancer when using the background mutation rates as in Sjöblom et al.. When the mutation rates are adjusted to fit the observed data, the number of significant genes reduces to 1 and 13 in breast and colon cancer, respectively (see Table B-1).

Assuming mutation rates increase uniformly over all categories, we can construct a uniformly most powerful (UMP) test. Let f_i be the mutation rate of category *i* under the null hypothesis and f_i^{alt} under the alternative. Then assume that

$\frac{f_i^{alt}}{1-f_i^{alt}} = \theta \frac{f_i}{1-f_i}$ for all *i*=1…7 for some θ (which is the same for all categories).

The joint distribution of x_{gi} for *i*=1…7 is an exponential family with parameter θ which has a monotone likelihood ratio $T(x) = \sum_{i=1}^{7} x_{gi}$. Therefore, a uniformly most powerful test for testing $\theta \le 1$ vs. $\theta > 1$ exists and rejects for large values of $T(x)$. Its distribution under the null hypothesis is easily calculated as the sum of independent binomial random variables. The number of significant genes obtained using this test statistic is 2 in breast cancer and 11 in colon. With the increased mutation rates, these numbers reduce to 1 in breast and 5 in colon cancer.

In Appendix B we suggest a correction to the scores and p-values that were described here, taking into account Sjöblom et al.'s experimental design.

Appendix B: Modeling the discovery and validation screens

Prior to submitting this Technical Comment we shared our analysis with Sjöblom et al. They suggested that the analysis should also take into account the two-staged nature of their design. In this appendix we model the two screens. The overall conclusion remains the same: after correcting the p-values, adjusting the mutation rates to match the observed rate in the discovery screen and taking into account the effect of variation in mutation rates, essentially no novel cancer gene meets significance.

In order to address the two-staged design, we propose a different score in which we consider separately the mutations in the two screens. The new score is defined as

$$s_g^{new} = \begin{cases} 0 & \text{if } \sum_i x_{gi}^D = 0 \text{ or } \sum_i x_{gi}^V = 0 \\ -\ln\left[\prod_{i=1}^{7} b(x_{gi}^D; N_{gi}^D, f_i) \prod_{i=1}^{7} b(x_{gi}^V; N_{gi}^V, f_i)\right] & \text{otherwise} \end{cases}$$

where the D and V superscripts designate the discovery and validation screens, respectively. Using the method described in Appendix A one can calculate the probability to observe a score greater than or equal to the observed one. The observed numbers of mutations in each of the screens were constructed from Sjöblom et al.'s Table S4.

As expected, this new method yields a larger number of significant genes: 34 in breast cancer and 38 in colon cancer. But, as with the original score, these gene lists present a strong sensitivity to slight increase in background mutation rates (see Table S7).

We can also adjust the likelihood-ratio test to account for the separate discovery and validation screen. The joint distribution of x_{gi}^D, x_{gi}^V for *i*=1…7 for $\sum_i x_{gi}^D > 0, \sum_i x_{gi}^V > 0$ is

$$p(x_{g1}^D,...,x_{g7}^D, x_{g1}^V,...,x_{g7}^V) = \prod_{i=1}^{7} \binom{N_{gi}^D}{x_{gi}^D} f_i^{x_{gi}^D} (1-f_i)^{N_{gi}^D - x_{gi}^D} \binom{N_{gi}^V}{x_{gi}^V} f_i^{x_{gi}^V} (1-f_i)^{N_{gi}^V - x_{gi}^V}.$$

The MLE is then $\hat{f}_{gi} = (x_{gi}^D + x_{gi}^V)/(N_{gi}^D + N_{gi}^V)$. For $\sum_i x_{gi}^D = 0$ or $\sum_i x_{gi}^V = 0$, the number of mutations is not being reported, which happens with probability

$$P(\text{not reported}) = P(\sum_i x_{gi}^D = 0) + P(\sum_i x_{gi}^D = 0) - P(\sum_i x_{gi}^D = 0, \sum_i x_{gi}^V = 0) =$$
$$= \prod_{i=1}^{7} (1-f_i)^{N_{gi}^D} + \prod_{i=1}^{7} (1-f_i)^{N_{gi}^V} - \prod_{i=1}^{7} (1-f_i)^{N_{gi}^D} \prod_{i=1}^{7} (1-f_i)^{N_{gi}^V}$$

and the MLE in this case is $\hat{f}_{gi} = 0$. The new test statistic is therefore

$$LLRT_g^{new} = \begin{cases} -\log\left(P(\text{not reported})\right) & \text{if } \sum_i x_{gi}^D = 0 \text{ or } \sum_i x_{gi}^V = 0 \\ \sum_{i=1}^{7} \log\left(\hat{f}_{gi}^{x_{gi}} (1-\hat{f}_{gi})^{N_{gi}-x_{gi}}\right) - \log\left(f_i^{x_{gi}} (1-f_i)^{N_{gi}-x_{gi}}\right) & \text{otherwise} \end{cases}$$

and its distribution under the null hypothesis and p-values can be calculated as explained above.
Table S7 summarizes the number of significant genes found by the various tests using the original mutation rates and increased rates that account for higher observed mutation rates and effect of variation in mutation rates.

Appendix C: A plug-in estimator approach for the FDR

Another method for estimating the FDR was proposed by Storey and Tibshirani [12]and Efron and Tibshirani [13] [14]. Let z_g be a univariate summary statistic for gene g. Assume that all z_g are i.i.d. random variables from a distribution $f()$, which is a mixture of two underlying distributions $f_0()$ and $f_1()$. Specifically

$$f(z_g) = \pi f_0(z_g) + (1-\pi) f_1(z_g)$$

where f_0 is the distribution of z_g when the null hypothesis is true, f_1 when the alternative is true and π is the proportion of true null hypotheses. An estimate for the FDR when rejecting genes with $z_g \geq z$ is then

$$FDR(z) = \pi \overline{F_0}(z) / \overline{F}(z)$$

where $\overline{F}(z) = P(z_g \geq z)$. An estimate for $\overline{F}(z)$ is obtained from the data by calculating the proportion of genes with $z_g \geq z$. In order to estimate $\overline{F_0}(z)$, simulations of the z_g-statistic under the null can be used. As π is very close to 1 (only a small percentage of the 13,203 genes are non-null), setting $\pi = 1$ is convenient and only slightly conservative. The estimate for the FDR is then

$$FDR(z) = \frac{\text{Proportion of genes with } z_g \geq z \text{ in simulations of the null distribution for all genes}}{\text{Proportion of genes with } z_g \geq z \text{ in observed data}}$$

Various choices are possible for z_g, for example s_g and $LLRT_g$. Here we used $z_g = s_g$ (for results see Table S7).

CaMP score cannot be used for z_g:
One statistic that does not fulfill the assumptions for statistic z_g is the *CaMP*-score. Using $CaMP_g$ in the procedure to estimate the FDR as described above can lead to serious underestimation of the true FDR. Let us briefly illustrate the problems with a quick example:

Let $u_g = \prod_{i=1}^{7} U_{ig}$ where U_{ig} are independent, uniformly distributed random variables. Let r_g be the rank of u_g and set $CaMP_g = -\log_{10}(G u_g / r_g)$ with $G=1000$ the number of genes. As an arbitrary threshold for the *CaMP*-score we pick 5.5. In Simulation A let all u_g be distributed as above. In Simulation B, pick 10 genes and set their u_g values to 10^{-10} (very significant genes that follow the alternative hypothesis).

	# alternatives > cutoff	# nulls > cutoff	FDR estimate	True FDR
All genes follow the null hypothesis (Sim. A)	0	.45	1	1
10 follow the alternative hypothesis (Sim. B)	10	2.11	.45/12.11=0.037	2.11/12.11=0.174

Table S6: Results of Simulations A and B. Simulation B shows the underestimation of the FDR in case true cancer genes exist.

As seen in the second row of the Table S6 above, the estimate for the FDR using the *CaMP* score severely underestimates the true FDR. The reason for this is the larger number of null genes exceeding the cutoff in a setting where genes that follow the alternative (true cancer genes) are present. In terms of the p_g statistic, the *CaMP* score is defined as $CaMP_g = -\log_{10}(Gp_g / r_g)$ with G the number of genes (13,023) and r_g the rank of p_g. Then, the *CaMP* score exceeds a cutoff c if

$$CaMP_g \geq c \quad \Leftrightarrow \quad p_g \leq 10^{-c} r_g / G .$$

If genes that follow the alternative are present and they are very significant, the rank of the most significant null gene will be larger. This, in turn, increases the threshold for p_g, letting more null genes become significant. However, the plug-in procedure described above does not account for this as $\overline{F_0}(z)$ is simulated assuming all genes follow the null hypothesis. For a more detailed treatment of this issue see [15].

	Breast				**Colon**			
Test statistic	**Sjöblom et al.'s mutation rates**	**Increased mutation rates to match discovery screen**	**Increased mutation rates to match discovery screen and model variation in mutation rates**	**Increased mutation rates to match discovery screen and model variation in mutation rates**	**Sjöblom et al.'s mutation rates**	**Increased mutation rates to match discovery screen**	**Increased mutation rates to match discovery screen and model variation in mutation rates**	**Increased mutation rates to match discovery screen and model variation in mutation rates**
Scaling factor	1	1.9	1.9	1.9	1	1.43	1.43	1.43
Log normal σ (~CV)	0	0	0.4	0.58	0	0	0.4	0.78
Effective scaling factor	**1**	**1.9**	**2.5**	**3.31**	**1**	**1.43**	**1.9**	**3.78**
s_g	6	1	1	1	28	11	6	3
$LLRT_g$	7	1	1	1	28	13	8	4
UMP-Test	2	1	1	1	11	5	5	3
s_g^{new}	34	1	1	1	38	18	6	3
$LLRT_g^{new}$	42	2	1	1	39	21	8	4
Plug-in estimator	58	1	1	0	44	21	12	3

Table S7: Number of *CAN*-genes in breast and colon cancer for different tests and mutation rates. The original score used by Sjöblom et al., s_g, is marked with a red background. The effective scaling factor appears in boldface. See supporting data tables in accompanying Excel file.

Appendix D: A second approach to estimating the background mutation rates

The estimates of the background mutation rates in Section 2.2.1 are based on the mutations observed in the discovery screen after removing known cancer genes. A possible criticism of this approach is that other yet unknown cancer genes may still be present and these inflate our estimation of background mutation rates.In order to minimize this possible contamination, we removed all genes from the discovery screen that had a mutation in the validation screen and counted the number of mutations observed in them. The remaining genes are less likely to be cancer related. Next, we simulated an experiment with discovery and validation screens using different values of background mutation rates and similarly counted the number of mutations seen in non-validated genes. The estimate for the background mutation rate is then the value for which the simulated number of mutations in non-validated genes matched the observed number of mutations in non-validated genes. This was done separately for breast and colon cancer. This resulting scaling factors for the Sjöblom rates were 1.92 for breast and 1.38 for colon cancer. As these were very close to the initial estimates of 1.9 and 1.43 (from Section 2.2.1), we continued using the initial values.

Acknowledgements

We wish to thank the authors of Sjöblom et al. for a fruitful dialog regarding the issues raised in this Technical Comment (see Appendix B). We also thank Steve Schaffner for sharing with us his analysis and results regarding variation in germline mutation rates.

References

S1. Sjöblom, T., et al., *The Consensus Coding Sequences of Human Breast and Colorectal Cancers.* Science, 2006.

S2. Greenman, C., et al., *Statistical analysis of pathogenicity of somatic mutations in cancer.* Genetics, 2006. **173**(4): p. 2187-98.

S3. Stephens, P., et al., *A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer.* Nat Genet, 2005. **37**(6): p. 590-2.

S4. Wang, T.L., et al., *Prevalence of somatic alterations in the colorectal cancer cell genome.* Proc Natl Acad Sci U S A, 2002. **99**(5): p. 3076-80.

S5. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* J Roy Stat Soc, Ser B, 1995. **57**: p. 289-300.

S6. Schaffner, S., *Regional variation in mutation rate*. 2006.

S7. Webster, M.T., et al., *Gene expression, synteny, and local similarity in human noncoding mutation rates.* Mol Biol Evol, 2004. **21**(10): p. 1820-30.

S8. Hellmann, I., et al., *A neutral explanation for the correlation of diversity with recombination rates in humans.* Am J Hum Genet, 2003. **72**(6): p. 1527-35.

S9. Hellmann, I., et al., *Why do human diversity levels vary at a megabase scale?* Genome Res. 10.1101/gr.3461105, 2005. **15**(9): p. 1222-1231.

S10. Gaffney, D.J. and P.D. Keightley, *The scale of mutational variation in the murid genome.* Genome Res, 2005. **15**(8): p. 1086-94.

S11. Futreal, P.A., et al., *A census of human cancer genes.* Nat Rev Cancer, 2004. **4**(3): p. 177-83.

S12. Storey, J.D. and R. Tibshirani, *Statistical significance for genomewide studies.* Proc Natl Acad Sci U S A, 2003. **100**(16): p. 9440-5.

S13. Efron, B., et al., *Empirical Bayes Analysis of a Microarray Experiment.* Journal of the American Statistical Association, 2001. **96**(456): p. 1151-1160.

S14. Efron, B. and R. Tibshirani, *Empirical bayes methods and false discovery rates for microarrays.* Genet Epidemiol, 2002. **23**(1): p. 70-86.

S15. Höfling, H., G. Getz, and R. Tibshirani, *Comments on "Significance of candidate cancer genes as assessed by the CaMP score".* 2007.http://www-stat.stanford.edu/~tibs/ftp/ParmigianiComment.pdf