Comments on "Significance of candidate cancer genes as assessed by the CaMP score" by Parmigiani et al.

Holger Höfling*† Gad Getz‡ Robert Tibshirani§

June 26, 2007

1 Introduction

Identifying genes that are involved in the development of cancer has been a very important research goal. One approach tries to identify genes in tumors that have an increased mutation rate. [Sjöblom et al., 2006] sequenced 13,023 CCDS genes in breast and colorectal cancer tumors. CCDS genes are protein encoding genes and represent the most highly curated gene set currently available. The data collection phase consisted of two main parts in which genes that were deemed not to have an increased mutation rate were eliminated. In particular, these two parts were:

Discovery screen: All genes were sequenced in 11 breast and 11 colorectal cancer tumors. Initially 816,986 mutations were identified. In order to find true somatic mutations (i.e. present in the tumor but not present in the germline of the patient), a complex set of filtering steps was used and all but 1,307, mutations in 1,149 genes were discarded. The next step of the data collection was only performed on those genes that contained at least one of these mutations.

Validation screen: Genes with mutations in the Discovery screen were sequenced in additional 24 breast and 24 colorectal cancer tumors. Through a similar system as before, 133,693 initially identified mutations were filtered down to 365 in 236 genes. Only genes with at least one mutation in the Discovery as well as the Validation screen were used in the subsequent statistical data analysis. These genes were called "validated".

*Holger Höfling is supported by an Albion Walter Hewlett Stanford Graduate Fellowship.

†Dept. of Statistics, Stanford University, Stanford, CA, 94305, USA

‡Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02142, USA

§Dept. of Statistics, Stanford University, Stanford, CA, 94305, USA

Among the validated genes, those that have a significantly increased mutation rate have to be identified. Sjöblom et al proposed to use the Benjamini-Hochberg procedure to deal with multiple hypothesis testing and control the False Discovery Rate (FDR). In order to do this, they defined the CaMP score (for a more detailed description see appendix A). A validated gene is determined to be significant at an FDR level of 0.1 if its CaMP score is > 1. Using this score, 122 genes in breast and 69 genes in colon cancer were identified as significant. In [Getz et al., 2007], an error and other problems with this approach were identified and corrected. The corrected analysis yields only 2 significant genes in breast and 28 genes in colorectal cancer.

The main goal of [Parmigiani et al., 2006] is to refute the results of [Getz et al., 2007] and establish that the results of [Sjöblom et al., 2006] were in fact conservative. They claim that when using the Sjöblom background mutation rate, the FDR is 1.1% for breast and 1.9% for colon cancers. With the mutation rate of [Getz et al., 2007], which uses the Discovery screen for estimation (DS-rate), these values change to 15.9% for breast and 9.1% for colon cancer.

In the following sections, the approach of [Parmigiani et al., 2006] will be described, a new error and other issues identified and its implications on the results discussed. In section 5, a corrected method will be presented which gives much higher estimates for the FDR. A comparison of the results can be seen in Figure 6.

2 Analysis of CaMP score in [Parmigiani et al., 2006]

[Parmigiani et al., 2006] use an Empirical Bayes-like setting for their analysis. 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 $f(\cdot)$, which is a mixture of two underlying distributions $f_0(\cdot)$ and $f_1(\cdot)$, Specifically

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

where f_0 is the distribution of z_g when the null hypothesis is true, f_1 when the null is false and π is the proportion of true nulls. An estimate for the FDR when rejecting genes if $z_g \geq z$ can then be obtained as

$$FDR(z) = \pi \bar{F}_0(z) / \bar{F}(z)$$

with threshold z and $\bar{F}(z) = P(z_g \geq z)$.

In particular, [Parmigiani et al., 2006] used the CaMP score for gene g as statistic z_g. $\bar{F}$ was estimated from the Sjöblom data. In order to get $\bar{F}_0$, they used a simulation study. Assuming that the null hypothesis is true for **all** genes (i.e. mutations occur at the background mutation rate), mutations on all 13,023 genes, 7 categories and with separate Discovery and Validation screens were simulated using (a) the Sjöblom background rate and (b) the DS background

rate. Over 1000 simulations, the average number of genes with CaMP scores exceeding z was recorded. Assuming $\pi = 1$, the FDR was estimated as

$$FDR(z) = \frac{\text{average proportion of genes with CaMP score} \geq \text{z in simulations}}{\text{proportion of genes with CaMP score} \geq \text{z in Sjöblom data}}. \quad (1)$$

As already stated above, for a threshold $z = 1$ and using the Sjöblom rate, the FDR was 1.1% for breast and 1.9% for colon cancers. With the DS background rate, the FDR increased to 15.9% for breast and 9.1% for colon cancer.

3 A quick example of the main error

As the main error of the method used in [Parmigiani et al., 2006] is quite subtle, we want to show it in a small example before getting into more detail in the next section. In simulation 1, assume that there are $G = 1000$ genes, and the statistics p_g are distributed as

$$p_g \sim \prod_{i=1}^{7} U_i \quad \forall g$$

where $U_i \sim U(0,1)$ i.i.d. Let q_g be the rank of p_g and define

$$CaMP_g = -\log_{10}(Gp_g/q_g).$$

In simulation 2, define everything as before, but this time only assume that 990 out of 1000 genes follow the null distribution. For the other 10 genes set $p_g = 10^{-10}$ (i.e. highly significant). A histogram of all p_g-statistics and the CaMP scores of only the true nulls that exceed 4.5 can be seen in Figure 1. It is evident that the 10 alternative hypotheses also lead to higher CaMP scores for the the null hypotheses. When all null hypotheses are true, in 1000 simulations on average .45 CaMP scores of null genes exceeded 5.5. When 10 genes are forced to be highly significant, this number increases to 2.11 (again only for null genes). The highly significant CaMP scores were not plotted so that they don't obscure the effect on the null gene CaMP scores. All CaMP scores smaller than 4.5 were not plotted so that the increase for large scores becomes visible in the histogram. The cutoff 5.5 was chosen arbitrarily.

From this example, it can be seen that the number of true null hypotheses that cross the threshold also depends on the number of genes that follow the alternative. Ignoring this dependence by assuming that all null hypotheses are true leads to a serious underestimate of the FDR in [Parmigiani et al., 2006]. In the example above, for a cutoff of 5.5, the FDR estimate would have been $.45/12.11 = 0.037$ instead of the better estimate $2.11/12.11 = .174$.

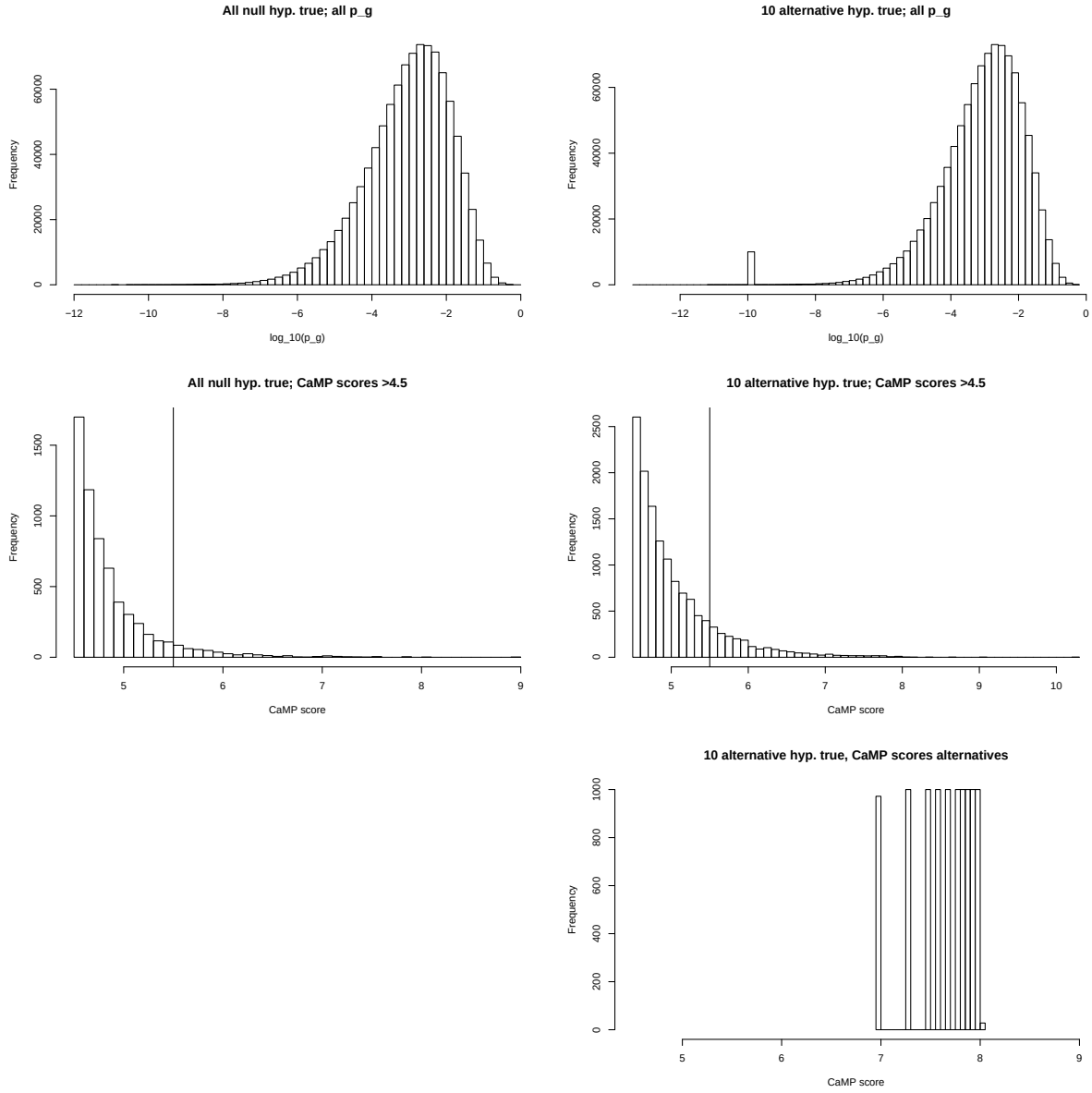

Figure 1: In the top left and right panels, histograms of all p_g statistics over 1000 simulation runs are plotted when all nulls are true and for the case when 10 follow the alternative, respectively. The p_g statistics corresponding to the alternative hyp. can be seen in the top right panel at value -10. In the middle left and right panels, histograms of CaMP scores > 4.5 (alternatives excluded), on the left when all null hyp. are true and on the right when 10 follow the alternative hyp. The average # of scores > 5.5 in the middle left panel is 450 (in 1000 simulations, so .45 on average) and 2160 (2.11 on average) in the middle right panel. The bottom right panel contains the CaMP scores of the highly significant genes which were plotted separately so that they don't obscure the effect on the CaMP scores of the null genes.

4 The error and other issues in [Parmigiani et al., 2006]

As shown in the previous section, the main error of [Parmigiani et al., 2006] is the use of the CaMP score. This section will go into more detail on this error. Apart from this, two other issues regarding estimation of the background mutation rate as well as control of the FDR by the CaMP score will be discussed. In particular, the topics of the following three subsections are:

1. The functional dependence of the CaMP score on all p_g statistics leads to an understatement of the FDR.
2. The background rate estimates in [Getz et al., 2007] are not too high. Excluding all validated genes leads to almost identical values.
3. The CaMP score does not control the FDR at the specified level.

4.1 The dependence structure of the CaMP scores on the p_g statistics leads to an understatement of the FDR

The key to estimating the FDR is formula (1), which is called a "plug-in" estimate in [Storey and Tibshirani, 2003] and it is equivalent to what [Parmigiani et al., 2006] called an Empirical Bayes estimate (see [Efron and Tibshirani, 2002]). Subsequently, we will refer to it as the "plug-in" estimator. In order to estimate the FDR using equation (1), the statistics are assumed to be independent. The most serious issue with the method of [Parmigiani et al., 2006] is the special functional dependence of the CaMP scores on the p_g statistics (for a definition see Appendix A). This dependence will lower the estimate for the FDR. The CaMP score is defined as

$$CaMP_g = -\log_{10}(N \cdot p_g/q_g).$$

The statistics p_g are independent between genes, but the ranks q_g of course depend on all statistics p_g of all genes simultaneously. The CaMP score being greater than a threshold c can be restated in terms of the independent statistics p_g as

$$CaMP_g \geq c \quad \Longleftrightarrow p_g \leq 10^{-c}\frac{q_g}{G}.$$

So, in terms of the p_g statistics, we do not have a single threshold but a set of thresholds the statistics have to cross. The largest statistic has to cross $10^{-c}/G$, the second largest $10^{-c} \cdot 2/G$ and the n-th largest statistic $10^{-c} \cdot n/G$. When the null hypothesis is true for all genes, the largest p_g of the null distribution has to cross the smallest threshold and so on. This situation however changes when for some genes the alternative hypothesis holds. These

will have low p_g statistics, crossing the lowest thresholds, leaving larger thresholds for the smallest p_g of the true null genes. So, when some of the null hypotheses do not hold, it is easier for the null genes to be significant and the number of falsely rejected null hypotheses will increase with the number of genes distributed according to the alternative. Therefore, the assumption that all null hypotheses are true, as in [Parmigiani et al., 2006], leads to a too low estimate for the number of falsely rejected null hypotheses and thus also for the FDR if in fact some null hypotheses do not hold.

We repeated and expanded the simulations in [Parmigiani et al., 2006]. Assume that all except F randomly chosen genes have mutations according to the background mutation rate in the discovery and validation screens. The F other genes are assumed not to follow the null distribution and instead have as many mutations as possible, leading to certain significance. Taking $F = 0$ gives the same situation as in [Parmigiani et al., 2006]. Then the CaMP score is calculated, using a threshold of 1 for rejecting the null hypotheses. In Figure 2 and 3 the # of falsely rejected null hyp. depending on the # of alternative and the corresponding FDR are plotted, both for breast and colon cancer as well as for the background mutation rate in [Sjöblom et al., 2006] and the DS-rate based on the mutations in the discovery screen as in [Getz et al., 2007].

All these FDRs are a lot higher than the reported results in [Parmigiani et al., 2006]. It is important to note that this estimate of the FDR is only a lower bound for the FDR in the Sjöblom dataset as here it was assumed that all genes for which the alternative holds have an extremely high mutation rate, leading to mutations on every single nucleotide. In a situation where the alternative genes are less obvious, the FDR will tend to be larger. A more reliable way of estimating the FDR than this approach will be presented in section 5.

The underlying reason for the error is, that methods that simulate or permute to create a null distribution, make an important assumption. The simulated data contains only genes distributed according to the null distribution. The original data contains a mix of null and non-null genes. The assumption is that the presence of the non-null genes in the original data has no effect on the computation of the test statistic for the null genes. With this assumption, the null simulation provides a good estimate of the number of false positive genes in the original data.

This really means that the test statistic for each gene must be computed independently of the other genes. Otherwise, the FDR estimate will be biased, sometimes quite badly. In the present context, use of the point-probabilities p_g satisfies our assumption, but the CaMP score does not. The ranking operation means that every gene has an effect on the test statistic for every other gene. In Figure 1, bottom right panel, the presence of the non-null genes increases the scores for the null genes, as compared to that in the bottom left panel.

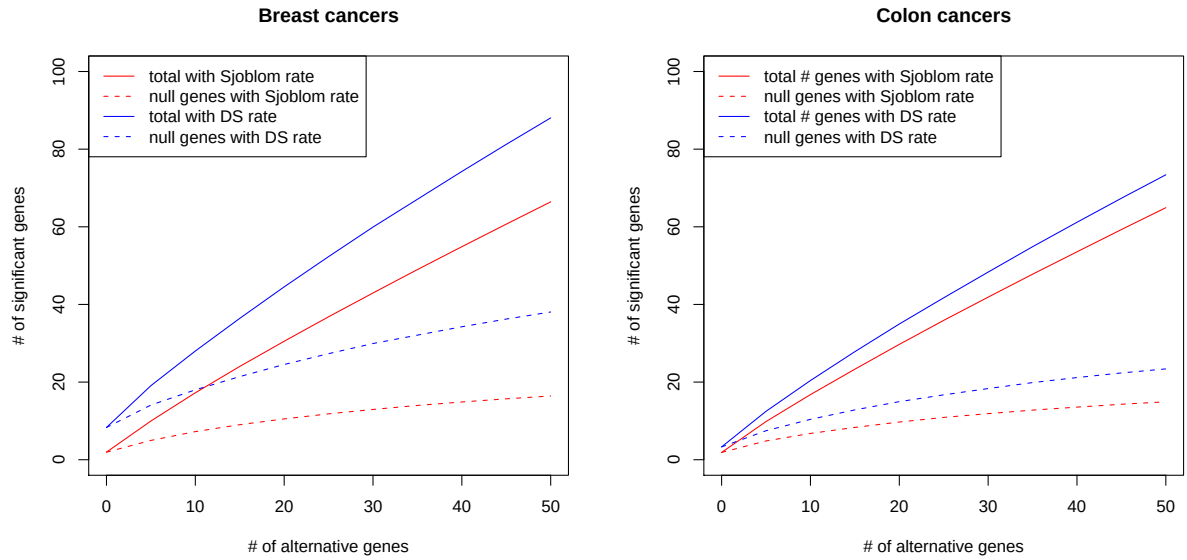

Figure 2: # of significant genes and # of significant null genes by # of alternatives; both for breast and colorectal cancer as well as the background rate from [Sjöblom et al., 2006] and the one estimated in [Getz et al., 2007].

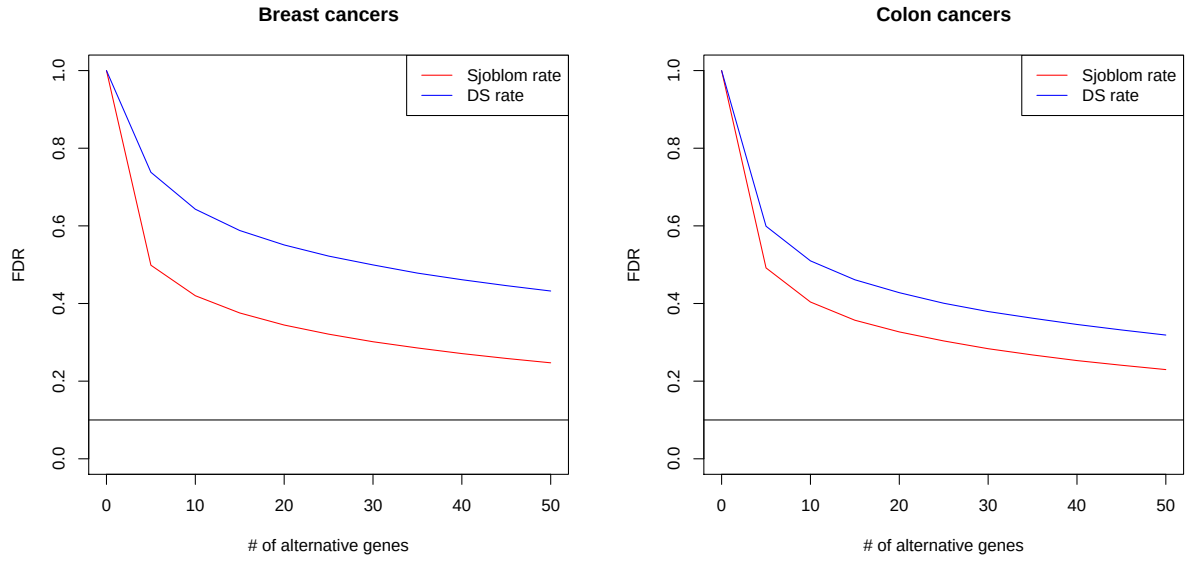

Figure 3: False Discovery rate by # of alternatives; both for breast and colorectal cancer as well as the background rate from [Sjöblom et al., 2006] and the one estimated in [Getz et al., 2007].

4.2 Background mutation rate estimation from data

As noted before in [Getz et al., 2007], the background mutation rate used has a big impact on the number of significant genes. There, the background rate for colon cancers was estimated to be 1.43 times as high as in [Sjöblom et al., 2006] and 1.9 times as high for breast cancers. [Parmigiani et al., 2006] states that these estimates are "a rather extreme perturbation of the background assumptions and is likely to be a significant overestimate of the background rate because the discovery phase included mutations present in known cancer genes".

In order to get an estimate of the background mutation rate that is less susceptible to this problem, exclude data that possibly has a high level of genes with increased mutation rates. In the Sjöblom dataset, relying only on mutations of genes that did not pass the validation screen achieves just that. This way all known cancer genes do not contaminate the background rate estimate. In addition, it is also unlikely that an important unknown cancer gene that has a mutation in the discovery screen fails the subsequent validation screen. This would only happen if in all 24 cancer samples in the validation screen, no mutation occurs. In a gene that is mutated in about 10% of all cancers, the probability of this event is about 8%. So, important unknown cancer genes that passed the discovery screen will most likely also pass the validation screen. Therefore, estimating the background rate only from genes that did not pass the screen is very reliable.

Fixing a background rate λ, the mean number of mutations in the discovery screen of genes that do not pass the validation screen (say $M(\lambda)$) is obtained by averaging over 1000 simulation runs. The estimate for the background mutation rate is then the rate λ for which $M(\lambda)$ is equal to the observed non-validated discovery screen mutations. For breast cancer, this is 1.92 times the Sjöblom rate and 1.38 times for colon cancer. Subsequently, these rates will be called non-validated genes mutation rates or NVG-rates. It is interesting to note that the NVG-rates are very close to what was already assumed in [Getz et al., 2007].

4.3 CaMP score does not control the FDR

It has been pointed out earlier in [Getz et al., 2007] that the CaMP score does not control the FDR. The simulations of [Parmigiani et al., 2006] show what the extent of this failure exactly is. In their simulations all genes were assumed to satisfy the null hypothesis, i.e.. that mutations occur according to the background mutation rate. When the background mutation rate is assumed to be as in [Sjöblom et al., 2006], an average of 1.3 genes are significant for colon and 1.4 genes in breast cancers. When the background rate is set to be the average mutation rate in the discovery phase, these values increase to 4.26 significant genes for colon and 7.46 for breast cancers.

As all genes follow the null hypothesis, rejecting any genes corresponds to a false discovery proportion of 1. Therefore, controlling the FDR when all genes follow the null hypothesis

is the same as controlling the familywise error rate at the same level. In order to control the FDR at a level of .1, only in 10% of all simulations any genes can be rejected and the average number of genes over all simulations would be 0.15. This number would also not vary with a varying background mutation rate. From the results in [Parmigiani et al., 2006], it can be seen how severely the CaMP score fails, rejecting at least 9 times as many genes in this situation as it should. This reinforces the point already made in [Getz et al., 2007], that the CaMP score should not be used to control the FDR.

5 An accurate estimate of the FDR

In order to get an accurate plug-in estimate of the FDR, we will use the NVG-rates together with a simulation of the null distribution of the $s_g = -\log_{10}(p_g)$ statistic, which are independent between genes. A gene will be deemed significant if $s_g > t$ for a threshold t. The FDR(t) will then be estimated by

$$FDR(t) = \pi \bar{F}_0(t)/\bar{F}(t).$$

As both for breast and colon cancer, only roughly 1% of all genes passed the validation screen, it is reasonable to assume that $\pi \approx 1$. $\bar{F}(t)$ will be estimated using the Sjöblom data and $\bar{F}_0(t)$ by simulation using the NVG-rates, similar to what was done in [Parmigiani et al., 2006]. For completeness, the rates used in [Sjöblom et al., 2006] will also be used. As the [Getz et al., 2007] rates are very similar to the NVG-rates, they will be omitted in the subsequent analysis.

The results can be seen in Figures 4 and 5. Based on a 10% False Discovery Rate, the number of significant genes using the NVG-rate are 1 ($t = 6.725$) for breast and 22 ($t = 4.75$) for colon cancer. When assuming the Sjöblom rates, these become 59 ($t = 3.95$) for breast and 44 ($t = 4.125$) for colon cancer.

A comparison of the FDR estimates of [Parmigiani et al., 2006] and the results above can be seen in Figure 6. It can clearly be seen that the rates above are much higher than those obtained from [Parmigiani et al., 2006] with the same background mutation rate, no matter how many genes are being rejected. It shows how much bias the dependence structure of the CaMP score causes. Its results are much too low and therefore completely unreliable.

6 Conclusion

The arguments and simulations above show clearly that the methods employed in [Parmigiani et al., 2006] are faulty. Thus both the original analysis in [Sjöblom et al., 2006] and the followup analysis in [Parmigiani et al., 2006] contain serious errors that invalidate

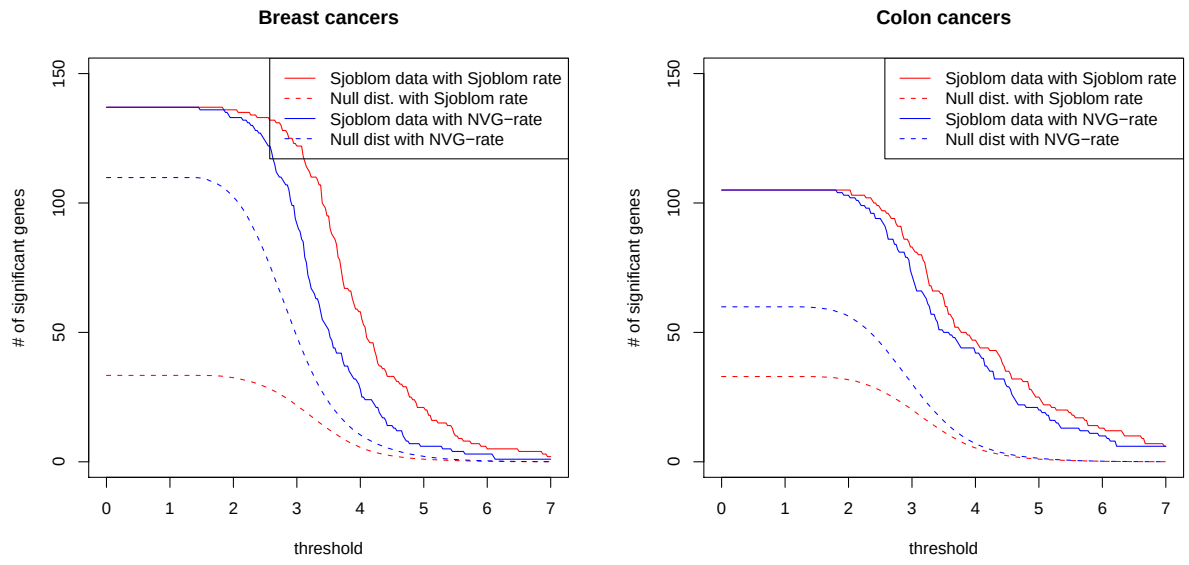

Figure 4: Number of significant genes of Sjöblom data by threshold; both for breast and colorectal cancer as well as the background rate from [Sjöblom et al., 2006] and the NVG-rate. Same for # of significant true null genes.

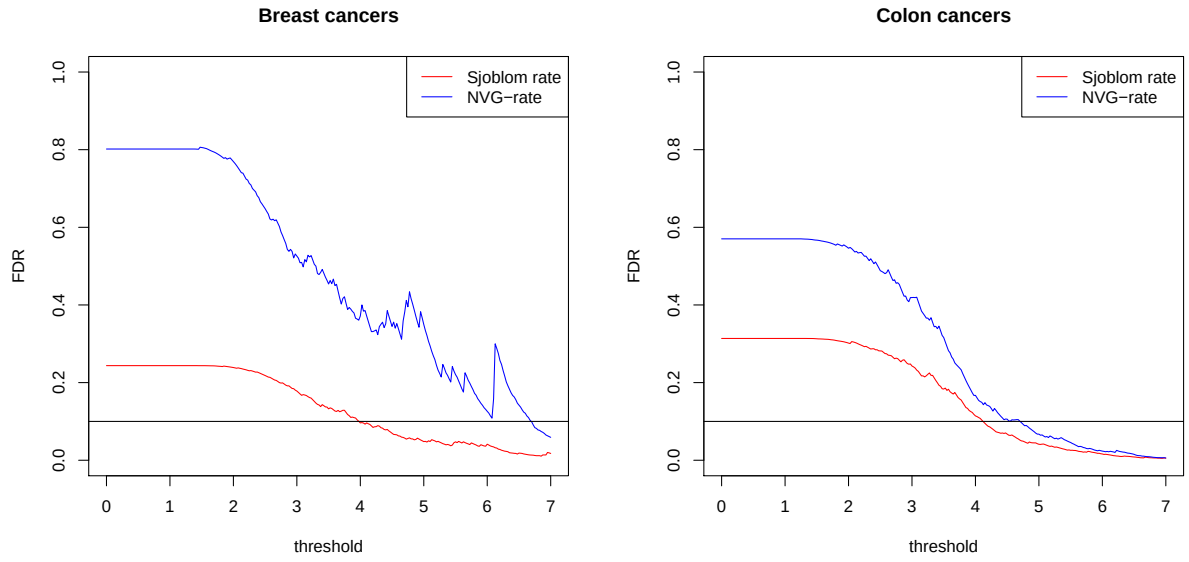

Figure 5: False Discovery rate by threshold; both for breast and colorectal cancer as well as the background rate from [Sjöblom et al., 2006] and the NVG-rate.

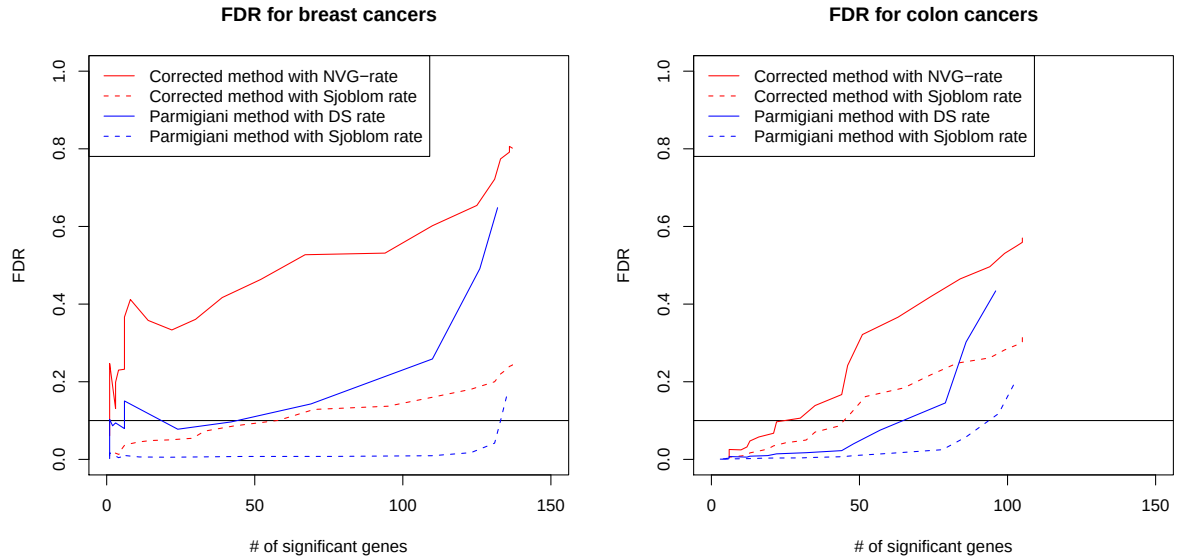

Figure 6: Comparison of the FDR estimates of [Parmigiani et al., 2006] and the corrected FDR estimates. The left panel shows for breast cancers in blue the FDR of Parmigiani using the Sjöblom rate as well as the DS-rate. In red, the corrected FDR can be seen, using the Sjöblom rates and the NVG-rates. The right panel shows the same information for colon cancer.

their conclusions. Most of the 189 genes (122 breast; 69 colon) that were claimed to be significant at an FDR level of 0.1 are in fact not significant. The genes that were found to be significant in this paper are consistent with the findings in [Getz et al., 2007], which corrected the statistical errors in [Sjöblom et al., 2006].

All in all, it can be seen that the CaMP score is not a good measure of statistical significance. It has no theoretical basis, does not control the FDR at the specified level and cannot be used in the plug-in estimator. Using the p_g-statistic (or LLRT-statistic) as proposed in [Getz et al., 2007] overcomes these problems. It has a solid theoretical foundation, can be used in the plug-in estimator as above and after transformation into p-values and application of the Benjamini-Hochberg procedure, it controls the False Discovery Rate.

APPENDIX

A Definition of the CaMP score

In order to explain their approach, a few definitions are needed (taken from [Parmigiani et al., 2006]):

G**:** Total number of genes sequenced ($G = 13,023$ in [Sjöblom et al., 2006]).

M**:** Number of mutation types ($M = 7$ in [Sjöblom et al., 2006]).

N_{gm}**:** Number of nucleotides for which a mutation of category m could occur in gene g.

T_{gm}**:** Total number of accurately sequenced nucleotides of type m in gene g.

X_{gm}**:** Total number of mutations of category m in gene g. May be broken down into X^d_{gm} and X^v_{gm} for discovery and validation screen.

θ_m**:** background probability of a mutation in a nucleotide of type m

The point probability of exactly having the observed number of mutations under the background mutation rate and disregarding separate Discovery and Validation screen is

$$p_g = \prod_{m=1}^{M} b(X_{gm}|T_{gm}, \theta_m)$$

where $b(x|n,p)$ is the binomial probability of x successes in n independent trials with success probability p. Then sort all p_g in increasing order and let q_g be the rank of p_g. The CaMP score is defined as

$$CaMP_g = \begin{cases} -\infty & \text{if } X^d_{g1} = \cdots = X^d_{gM} = 0 \text{ or } X^v_{g1} = \cdots = X^v_{gM} = 0 \\ -\log_{10}(Gp_g/q_g) & \text{otherwise} \end{cases}.$$

The top row specifies that genes that were not validated have a CaMP score of $-\infty$.

References

[Efron and Tibshirani, 2002] Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23:70–86.

[Getz et al., 2007] Getz, G., Höfling, H., Mesirov, J., Tibshirani, R., Golub, T., Meyerson, M., and Lander, E. (2007). Technical Comment on 'The consesus coding sequence of human breast and colorectal cancers' by Sjöblom et al. *Technical Report*.

[Parmigiani et al., 2006] Parmigiani, G., Lin, J., Sjöblom, T., Kinzler, K., Vogelstein, B., and Velculescu, V. (2006). Significance of candidate cancer genes as assessed by the CaMP score. *Johns Hopkins University, Dept. Of Biostatistics Working Papers*.

[Sjöblom et al., 2006] Sjöblom, T., Jones, S., Wood, L., Parsons, D., Lin, J., Barber, T., Mandelker, D., Leary, R., Ptak, J., Silliman, N., Szabo, S., Buckhaults, P., Farrell, C., Meeh, P., Markowitz, S., Willis, J., Dawson, D., Willson, J., Gazdar, A., Hartigan, J.,

Wu, L., Liu, C., Parmigiani, G., Park, B., Bachman, K., Papadopoulos, N., Vogelstein, B., Kinzler, K., and Velculescu, V. (2006). The consensus coding sequence of human breast and colorectal cancers. *Science*, pages 268–274.

[Storey and Tibshirani, 2003] Storey, J. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *PNAS*, pages 9440–9445.