10-11-2007

STATISTICAL METHODS FOR THE ANALYSIS OF CANCER GENOME SEQUENCING DATA

Giovanni Parmigiani
The Sydney Kimmel Comprehensive Cancer Center, Johns Hopkins University & Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, gp@jimmy.harvard.edu

J. Lin
The Ludwig Center and the Howard Hughes Medical Institute, The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins

Simina Boca
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

T. Sjoblom
The Ludwig Centre and the Howard Hughes Medical Institute, The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins

K.W. Kinzler
The Ludwig Centre and the Howard Hughes Medical Institute, The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins

See next page for additional authors

Authors
Giovanni Parmigiani, J. Lin, Simina Boca, T. Sjoblom, K.W. Kinzler, V.E. Velculescu, and B. Vogelstein

Statistical methods for the analysis of cancer genome sequencing data

Parmigiani G, Lin J, Boca S, Sjöblom T, Kinzler KW, Velculescu VE, Vogelstein B

The Ludwig Center and the Howard Hughes Medical Institute, The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins.

Abstract

The purpose of cancer genome sequencing studies is to determine the nature and types of alterations present in a typical cancer and to discover genes mutated at high frequencies. In this article we discuss statistical methods for the analysis of data generated in these studies. We place special emphasis on a two-stage study design introduced by Sjöblom *et al.*[1]. In this context, we describe statistical methods for constructing scores that can be used to prioritize candidate genes for further investigation and to assess the statistical significance of the candidates thus identified.

1 Introduction

The systematic investigation of the genomes of human cancers has recently become possible with improvements in sequencing and bioinformatic technologies. Sjöblom *et al.* [1] determined the sequence of a set of 13,023 genes termed the consensus coding sequences (CCDS) in colorectal and breast cancers and provided a catalogue of somatic mutations. In this context, a somatic mutation is a tumor-specific mutation not present in the germline of the patient whose tumor contained it. More recently, Greenman *et al.* [2] investigated somatic mutations in the coding exons of 518 protein kinase genes in a large and diverse set of human cancers. Other projects are under way or planned, including the Cancer Genome Atlas Project (`http://cancergenome.nih.gov/`), whose goal is to identify all the genes mutated in at least 5% of human cancers, across cancer types.

Statistical analysis of the data generated in these studies poses new challenges that are worthy of careful consideration. Greenman *et al.* have provided an in-depth analysis of data generated by one-stage studies [3]. In this article, we present analytic approaches applicable to two-stage designs. To make optimal use of sequencing resources, Sjöblom *et al.* introduced such a two-stage design, with the stages termed ”Discovery” and ”Validation”. The Discovery Stage consists of a catalogue of mutations in all genes considered —for example all genes in the CCDS database. This permitted selection of the subset of genes that harbored at least one somatic mutation, termed ”Discovered”. This subset was further investigated in a Validation Stage which catalogued somatic mutations in discovered genes in an independent set of tumor samples. Genes that were mutated in at least one tumor in the Validation set were termed ”Validated”.

The somatic mutations found in cancers are either ”drivers” or ”passengers” [2]. Driver mutations are causally involved in the neoplastic process and are positively selected for during tumorigenesis. Passenger mutations provide no positive or negative selective advantage to the tumor but are retained by chance during repeated rounds of cell division and clonal expansion. The overarching goal of the statistical analysis is to identify genes that are most likely to contain driver mutations on the basis of their mutation type and frequency. This is done by quantifying the evidence that the mutations in a gene reflect underlying mutation rates that are higher than the passenger rates. From a statistical standpoint, the following three considerations are critical for this analysis:

Two-Stage Design. The first consideration arises from the two-stage approach. Only genes that harbor at least one mutation in the Discovery Stage and at least one mutation in the Validation Stage are considered contenders for further analysis. This screening condition needs to be taken into account when assessing significance using p-values or other methods that rely on the sampling distribution of the data, as it removes a large number of possible experimental outcomes.

Coverage. The second consideration is that the number of nucleotides successfully sequenced is generally smaller than the size of the gene times the number of tumors analyzed. For example, certain exons may be technically challenging to sequence. It is appropriate to apply stringent quality criteria to sequencing data, which lead to the exclusion of nucleotides whose sequence could not be identified with certainty. We refer to the portion of a gene that is successfully investigated as its ”coverage”. Nucleotides excluded, or not covered, are not eligible for false positive mutations and should therefore not be included in statistical evaluations. Moreover, when applying corrections for multiple testing, coverage should be a consideration for all the genes, including those for which no mutations were found.

Context. The third consideration is that the precise base and neighboring bases of the mutations that were observed, termed the ”mutation context”, are important. The priority given to a gene in further studies and the statistical significance of a gene should depend on how mutations are distributed across contexts.

Sjöblom *et al.* addressed these issues by computing a score, called the Cancer Mutation Prevalence (CaMP) score, that provides a ranking of the Validated genes for selecting promising candidates. The CaMP score also provides a heuristic approximation of the false discovery rate (FDR) associated with lists of genes that were both validated and had a CaMP score that exceeded a given threshold. More recently, Parmigiani *et al.* [4], further developed this methodology using an Empirical Bayes approach similar to that originally proposed in Efron *et al.* [5]. This more fully addressed the three challenges described above.

In this article we provide a general notation and and foundation for statistical analysis of cancer genomes, show how the Empirical Bayes method for false discovery rates can be applied to this context and discuss alternative possibilities for implementation, compare Empirical Bayes estimates of false discovery rates to frequentist alternatives in simulated experiments and in experimental data on the CCDS genes, and present additional aspects of the data analysis that are specific to the RefSeq study [6]. This article serves as a detailed statistical account of the techniques used in [4] and [6], and implemented, respectively, in versions 1.0 and 2.0 of the CancerMutationAnalysis R package.

2 Methods

2.1 Sampling distribution of mutations counts.

While cancer genome sequencing projects produce a wealth of information, for the purposes of this discussion we will focus on the somatic mutation counts, broken down by gene and context, and considered separately for the Discovery and Validation Stages. Table 1 summarizes the notation we will be using. Our discussion refers to a single tumor type, say colorectal cancer. When data on multiple types is collected, this analysis can simply be repeated.

Statistically, an important component of the analysis is the inference on whether each gene is mutated at a rate that is increased compared to the passenger mutation rate. This can be formalized by setting the null hypothesis, for gene g, to be that all the context-specific mutation rates are the same as the corresponding context-specific passenger rates.

The first step is to derive the sampling distribution under the null hypothesis. For a specific gene and given mutation rates, we can derive the probability of observing mutation count vectors X^d_{mg} and X^v_{mg} under a product binomial model, based on the assumptions that mutations occur independently of each other both within and across contexts. This requires that the occurrence of a mutation does not change the total number of nucleotides available for other mutations, that is, that the nucleotides' contexts are mutually exclusive. In practice, in the Sjöblom *et al.* project, one

Nucleotides	M	number of mutation contexts.
	N_{gm}	number of nucleotides of context m in gene g.
Samples	K^d	number of samples in the Discovery Stage.
	K^v	number of samples in the Validation Stage.
	K	total.
Coverage	T_{gm}^d	number of nucleotides available for a mutation of type m in gene g in the Discovery Stage.
	T_{gm}^v	number of nucleotides available for a mutation of type m in gene g in the Validation Stage.
	T_{gm}	type-specific total $T_{gm}^d + T_{gm}^v$.
Mutation counts	X_{gm}^d	number of mutations of type m detected in gene g in the Discovery Stage.
	X_{gm}^v	number of mutations of type m detected in gene g in the Validation Stage.
	X_{gm}	type-specific total $X_{gm}^d + X_{gm}^v$.
	X_g	the vector $(X_{g1}, \dots, X_{gM})$.
Uppercase X's represent random variables; lowercase x's represent experimentally observed results		
Mutation rates	θ_m^d	probability of a passenger mutation in a nucleotide of context m in the Discovery Stage.
	θ_m^v	probability of a passenger mutation in a nucleotide of context m in the Validation Stage.
	θ_m	probability of a passenger mutation in a nucleotide of context m, when common to both Stages.
	θ_{gm}^d	probability of a mutation in a nucleotide of context m in gene g in the Discovery Stage.
	θ_{gm}^v	probability of a mutation in a nucleotide of context m in gene g in the Validation Stage.
	θ_{gm}	probability of a mutation in a nucleotide of context m gene g in, when common to both Stages.

Table 1: Summary of notation for the data produced by the study for gene g.

of the mutation types is indels (insertion-deletion mutations), which is applicable to all nucleotides, while others are nonsynonymous point mutations, which are catalogued in six mutually exclusive contexts. Each nucleotide is therefore susceptible to two types of mutations. However, given the large number of available nucleotides compared to the mutation rates, the assumption that the occurrence of a mutation does not change the total number of nucleotides available for other mutations is likely to have negligible effects. A refinement of this approach could consider a product of multinomial distributions.

Consider one gene at the time. Because the experiment's stopping rule is to continue sampling for gene g only if at least one mutation is found in the discovery phase, the sampling distribution takes two different forms depending on whether sequencing stops at the discovery or continues to validation. Define $b(x|N, \theta)$ the Binomial density for x successes in N independent trials each with success probability θ. If we stop at discovery then $\max_m(x_g^d) = 0$ and the sampling distribution is

$$\Pr\left(\max_m(X_g^d) = 0|T_{g1}^d, \ldots, T_{gM}^d, \theta_{g1}, \ldots, \theta_{gM}\right) = \prod_{m=1}^{M} b(0|T_{gm}^d, \theta_{gm}) \tag{1}$$

If we continue to validation then $\max_m(x_g^d) > 0$ and the sampling distribution is

$$\Pr(X_g^d = x_g^d, X_g^v = x_g^v|T_{g1}^d, \ldots, T_{gM}^d, T_{g1}^v, \ldots, T_{gM}^v, \theta_1, \ldots, \theta_M) = \\ \prod_{m=1}^{M} b(x_{gm}^d|T_{gm}^d, \theta_m) b(x_{gm}^v|T_{gm}^v, \theta_m) \tag{2}$$

Under the null hypothesis, mutations in all the genes occur at a common passenger rate θ_m for mutation type m. Call $\mathcal{G}$ the set of genes that are discovered, that is such that $\max_m(x_g^d) > 0$. The overall sampling distribution under the is

$$\Pr(\text{Data}|\text{Coverage}, \text{Passenger Rates}) = \\ \prod_{g \in \mathcal{G}^c} \prod_{m=1}^{M} b(0|T_{gm}^d, \theta_m) \prod_{g \in \mathcal{G}} \prod_{m=1}^{M} b(x_{gm}^d|T_{gm}^d, \theta_m) b(x_{gm}^v|T_{gm}^v, \theta_m) \tag{3}$$

In some cases it may be important to consider separate mutation rates for the Discovery and Validation Stages. For example, the Discovery Stage may employ cell lines while the Validation Stage may employ primary tumor samples. These two types of genomes may harbor different passenger mutation rates as a result of disparate selection conditions and different numbers of generations and bottlenecks through which the cells have traversed. When that is the case, using again the notation of Table 1, the sampling distribution becomes

$$\Pr(\text{Data}|\text{Coverage}, \text{Passenger Rates}) = \\ \prod_{g \in \mathcal{G}^c} \prod_{m=1}^{M} b(0|T_{gm}^d, \theta_m^d) \prod_{g \in \mathcal{G}} \prod_{m=1}^{M} b(x_{gm}^d|T_{gm}^d, \theta_m^d) b(x_{gm}^v|T_{gm}^v, \theta_m^v) \tag{4}$$

2.2 The CaMP score

In Sjöblom *et al.* we introduced the cancer mutation prevalence (CaMP) score, to provide a ranking of the Validated genes and select promising candidates. We define it formally here, based on the following algorithm. First, we compute

$$p_g = \prod_{m=1}^{M} b(x_{gm}|T_{gm}, \theta_m) \tag{5}$$

if the passenger rates are common to Discovery and Validation, and

$$p_g = \prod_{m=1}^{M} b(x_{gm}^d|T_{gm}^d, \theta_m^d) b(x_{gm}^v|T_{gm}^v, \theta_m^v). \tag{6}$$

if the passenger rates are allowed to differ in Discovery and Validation. Note that expression (6) does not collapse to expression (5) when $\theta^d_m = \theta^v_m$. More generally, the scale of the two versions of p_g is different and cannot be compared across the two analyses.

We then rank the p_g's and call q_g the resulting ranks. The CaMP score is

$$\begin{array}{lcll}
\text{CaMP}_g(x^d_{g1},\dots,x^d_{gM},x^v_{g1},\dots,x^v_{gM}) & = & -\infty & \text{if } \max_m\{x^d_{gm}\}=0 \\
\text{CaMP}_g(x^d_{g1},\dots,x^d_{gM},x^v_{g1},\dots,x^v_{gM}) & = & -\infty & \text{if } \max_m\{x^v_{gm}\}=0 \\
\text{CaMP}_g(x^d_{g1},\dots,x^d_{gM},x^v_{g1},\dots,x^v_{gM}) & = & -\log_{10}(Gp_g/q_g) & \text{otherwise}
\end{array} \qquad (7)$$

The top two rows correspond to genes that are eliminated at the Discovery and Validation Stages.

The goal of the CaMP score is to rank genes according to the strength of the evidence that they may be mutated at rates higher than the passenger rates. When passenger rates are common to both stages, Sjöblom *et al.*p_g can also be used as a heuristic approximation to the p-value to provide an estimate of the false discovery rate associated with lists of candidate genes drawn in the basis of CaMP. When this is done, a threshold of 1 on the CaMP score was considered sufficient by Sjöblom *et al.* to generate a list with estimated FDR of 10%. The p_g score is not a tail probability, as p-values should be. However, for very small rates, the approximation can be accurate, as discussed in Section 3.2.

2.3 False Discovery Rates

The false discovery rate (FDR) is defined as the estimated proportion of false discoveries among a set of reported discoveries [7], and has become popular for reporting uncertainty about the result of screening experiments whose goals are to identify lists of candidates for further biological evaluation. To address all three of the challenges described in the introduction, we suggest that the FDR can be optimally evaluated using an Empirical Bayes approach [5]. In such approaches, one computes a data summary, or score, that captures departure from the null hypothesis, such as a p-value, a likelihood ratio test (LRT) or other statistics. Examples specific to mutation analysis are given in Section 2.5. Then, via theoretical arguments or simulation, one determines the genomic distribution of this score under the assumption that no genes are real discoveries (the null distribution). For a given threshold, the counts of genes whose score is above a threshold in the observed and null distributions gives information on the FDR at that threshold. Such approaches do not necessarily require the calculation of p-values and have a variety of other theoretical and practical advantages [8, 9].

The notation used here follows that of [9]. Let $\delta_g \in \{0,1\}$ denote the indicator of whether gene g is a reported discovery (in the Sjöblom study whether gene g is a CAN gene), let $D = \sum_{g=1}^{G} \delta_g$ denote the total number of discoveries, and let $r_g \in \{0,1\}$ denote the unknown truth (in the Sjöblom study an indicator of whether gene g is truly mutated above the passenger rates in cancer). The false discovery rate is defined as $\text{FDR} = (\sum(1-r_g)\delta_g)/D$, the fraction of false discoveries, relative to the total number of discoveries. This ratio is unknown and needs to be estimated. Let z_g denote some univariate summary score for the g-th gene, such that larger values of z_g are stronger evidence against the gene's null hypothesis. Examples are the negative log of a p-value or the log of the LRT. Two-sided versions are also available, but are not considered here as the mutation analysis situation is naturally one-sided. Assume an i.i.d. sampling model for z_g, from a mixture distribution $f(\cdot)$ with terms f_0 and f_1 that describe the variation of the scores across genes within the subpopulations of truly mutated genes and not, respectively. Formally this distribution has the form

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

where π is the proportion of truly null genes. Using the unknown true status variables r_g introduced earlier, the mixture is equivalent to the two-stage model:

$$p(z_g \mid r_g = j) = f_j(z_g) \text{ and } \Pr(r_g = 0) = \pi \tag{8}$$

The distribution f is generally empirically observable, while f_0, f_1 and π need to be estimated.

It is common to estimate f_0 by permutation of class labels or by direct simulation from the null, if possible. In our analysis, the null distribution of a score can be obtained by simulating mutations at the passenger rates in a way that precisely replicates the two-stage experimental plan. Specifically, for the Discovery Stage, we consider each gene in turn and simulate the number of mutations of each type from a binomial distribution with success probability equal to the context-specific passenger rate. The number of available nucleotides in each context is the number of successfully sequenced nucleotides for that particular context and gene in the samples studied in the Discovery Stage. For all genes in which at least one mutation is generated in this simulation, the process is repeated, this time with the number of samples used in the Validation Stage. Potentially, a different passenger mutation rate may be used in the Validation Stage than was used in the Discovery Stage of the simulations. We finally apply to the simulated data any filters that were applied to the experimental data, for example excluding genes whose mutation rates are below a given threshold.

Estimation of π is more complex. Assuming $\pi = 1$ provides a practical and conservative alternative when the proportion of true discoveries is low. However, a better upper bound can be obtained as follows. We start by constructing histograms of the observed and simulated values of z_g for all genes, using bins of one unit in the log 10 scale. As an example, set $z_g = -\log_{10} p_g$, and consider the bin ranging from 0 to 1, which is composed mostly of genes with no mutations. Suppose that there are 1000 experimental genes and 1050 simulated genes in that bin. The 1000 genes include both passengers and non-passengers, while the 1050 should contain only passengers. Thus we can conclude that the number of passengers in the simulated set is too large and that π is at most 1000/1050. Because this argument can be applied to all bins, we can estimate π to be the reciprocal of the largest ratio between the simulated and observed bin counts. Estimates of π are stable over a wide range of bin sizes. This method is an adaptation of the approach proposed in Efron *et al.* [5]. In their approach, bin counts are modeled as a function of the scores using Poisson regression. In our case, a similar smoothing is achieved more simply by binning neighboring score values.

For given estimates of f_0 and π, the Empirical Bayesian false discovery rate [8] considers genes generated by the condition $\{z_g \geq z\}$ and is

$$\text{Fdr}(z) \equiv \pi \bar{F}_0(z)/\bar{F}(z)$$

where z is a threshold, and $\bar{F}(z) = \Pr\{z_g \geq z\}$. Even though this is technically not a posterior probability, it earned a Bayesian name because of the use of Bayes theorem to find the probability of false discovery given the observation that $z_g \geq z$, which can be shown to be equivalent to the defined Fdr statistic [8] above. The probability statement is in the context of the assumed mixture model, for assumed known f_0, f_1 and π. However, the Fdr statistic provides an excellent approximation for the predictive probability that a hypothetical additional gene randomly drawn from the same population, and exceeding the threshold, would be a real discovery $P(r_{G+1} = 1 \mid z_{G+1} \geq z, Y)$, if one were to use a full Bayesian model with flexible priors on f_1 and f_0 [10]. There is also a close connection between the empirical Bayes procedure proposed in [5] and the frequentist approach to control FDR using the Benjamini-Hochberg [11] procedure. Specifically, the empirical Bayes procedure can be shown to provide an upper bound to the frequentist FDR [8].

An additional technical difficulty arises because the Validation Stage was carried out only for a small and highly nonrandom subset of genes. However, as long as the score z_g is defined to take its minimum value, or $-\infty$, unless the Discovery and Validation Stages are passed, using the observed proportion of scores exceeding a finite threshold z still provides an unbiased estimate for $F(z)$.

2.4 Passenger Probabilities

While FDRs are statements about lists of genes, gene-specific assessments of uncertainty are also interesting. From a Bayesian standpoint, one would use the evidence more efficiently by conditioning on z_g rather than $\{z_g \geq z\}$, as can be proven using decision theoretic arguments [12]. In the EB approach, one makes gene-specific statements using the so-called local false discovery rate (fdr) [5], that is

$$\text{fdr}(z) \equiv \pi f_0(z)/f(z).$$

In terms of the mixture model, and conditioning on f_0, f_1 and π, the fdr statistic is the ”passenger probability” or the probability that a gene with score z is mutated at the passenger rates, $\text{fdr}(z) = Pr(r_g = 1 \mid z_g = z, Y, f_0, f_1, \pi)$. As before, one can argue that under a sufficiently flexible prior probability model on f_0, f_1, π, reasonable point estimates can be substituted for the unknown quantities, allowing us to interpret fdr’s as posterior probabilities, without reference to a specific prior model [10].

2.5 Scores for Empirical Bayes analysis

Both the FDR and passenger probability analyses rely on the selection of a score for ranking the genes. Several options are available. An intuitive approach is to choose $-\log_{10} p_g$, as this quantity captures in a natural way the information about context and coverage. A closely related score is the LRT [3, 13] for the null hypothesis that all the contexts are mutated at passenger rates, analyzed one gene at the time. For a gene such that we perform both discovery and validation, the LRT is the ratio p_g^*/p_g, where p_g^* is

$$p_g^* = \prod_{m=1}^{M} b(x_{gm}|T_{gm}, \hat{\theta}_{gm}) \tag{9}$$

and $\hat{\theta}_{gm}$ are the experimentally observed rates. Because mutations are rare, variation in the p_g^* terms across genes is small and the LRT is strongly associated with the p_g score, with rank correlations in excess of 99% in relevant experimental situations. The LRT can be generalized straightforwardly to the case when separate rates are used in Discovery and Validation.

An alternative set of scores can be constructed using frequentist p-values, obtained by evaluating the gene-specific probability of exceeding the observed value of a score of interest. If z_g is the score for gene g, then, keeping in mind that genes that are not validated are by definition not eligible to be declared significant, the frequentist p-value for the significance of the gene is

$$\begin{aligned}\text{p-value}_g = \Pr\{\max_m\{X^d_{g1}, \ldots, X^d_{gM}\} > 0 \text{ and } \max_m\{X^v_{g1}, \ldots, X^d_{gM}\} > 0 \text{ and} \\ Z_g(X^d_{g1}, \ldots, X^d_{gM}, X^v_{g1}, \ldots, X^d_{gM}) \geq z_g|T_{g1}, \ldots, T_{gM}, \theta_1, \ldots, \theta_m\}\end{aligned} \tag{10}$$

where Z_g is the random variable defined by mapping the experimental outcome on the scores' space. To express this as a more conventional tail probability one can equivalently define Z_g to take the value $-\infty$ when $\max_m\{X^d_{g1}, \ldots, X^d_{gM}\} = 0$ or $\max_m\{X^v_{g1}, \ldots, X^d_{gM}\} = 0$.

This p-value can be used directly in the Benjamini-Hochberg (BH) algorithm [11] or other procedures for controlling frequentist FDR. The use of frequentist p-values coupled with BH adjustment has been suggested as an alternative to the original analysis of Sjöblom *et al.* in three recent Technical Comments [14, 13, 15] to that paper.

Even when Z_g is a function only of the vector x_g the expression above differs from a binomial tail probability in that fewer outcomes are possible. For example, consider the case when there is only one type and two mutations are found: breaking things down into Discovery and Validation, a binomial tail probability would add up the probabilities of three cases: $(0, 2)$, $(1, 1)$, $(2, 0)$, while under the two-stage design with exclusion of the non-validated genes, only the $(1, 1)$ case is relevant for the p-value calculation.

More generally, consider a gene g for which we performed both stages of the two-stage design and found mutations in both. The mutation data is then such that $\max_m(X^d_{gm}) > 0$ and $\max_m(X^v_{gm}) > 0$. When passenger rates are constant across stages, the probability of observing $x = (x_1, ..., x_M)$ satisfying these constraints, for the p-value calculation, can be expressed as:

$$\Pr(X_g = x|T_{g1}, \ldots, T_{gM}, \theta_1, \ldots, \theta_M) =$$
$$\prod_{m=1}^{M} b(x_m|T_{gm}, \theta_m) - \prod_{m=1}^{M} b(0|T^d_{gm}, \theta_m) \prod_{m=1}^{M} b(x_m|T^v_{gm}, \theta_m) - \prod_{m=1}^{M} b(x_m|T^d_{gm}, \theta_m) \prod_{m=1}^{M} b(0|T^v_{gm}, \theta_m).$$

b

Table 2: Scores used for Empirical Bayes analyses.

Method	Score*	Two Stages Included	Mutation Context Included	Label used in Figure 1
S	p_g	Not Appl.	Yes	Sjöblom
F	Forrest	No	No	Forrest
GF	Getz, Forrest	No	Yes	Getz
R	Rubin	Yes	No	Rubin
GB	Getz (Appendix B)	Yes	Yes	

**Forrest and Cavet described two p-values, one used in GF and one used in F; Getz also described two p-values, one in the main text of their Comment and one in their appendix B. Their conclusions in the main text were based on the GF p-values and did not incorporate those in Appendix B. Getz et al. also suggest using the Likelihood Ratio Test (LRT). As noted, the rank correlation of the LRT with p_g is so high that there is no need to consider it separately from p_g.*

Replacing the expression for the Binomial density in the expression above and rearranging terms one obtains:

$$\Pr(X_g = x | T_{g1}, \ldots, T_{gM}, \theta_1, \ldots, \theta_M) = \left[\prod_{m=1}^{M} \binom{T_{gm}}{x_m} - \prod_{m=1}^{M} \binom{T^v_{gm}}{x_m} - \prod_{m=1}^{M} \binom{T^d_{gm}}{x_m} \right] \prod_{m=1}^{M} \theta_m^{x_m} (1-\theta_m)^{T_{gm}-x_m} \quad (11)$$

When Binomial probabilities are small, Binomial distributions could also be accurately approximated using the Poisson distributions. This approach leads to the following expression

$$\Pr(X_g = x | T_{g1}, \ldots, T_{gM}, \theta_1, \ldots, \theta_M) \approx \left[\prod_{m=1}^{M} \frac{T_{gm}^{x_m}}{x_m!} - \prod_{m=1}^{M} \frac{(T^d_{gm})^{x_m}}{x_m!} - \prod_{m=1}^{M} \frac{(T^v_{gm})^{x_m}}{x_m!} \right] \prod_{m=1}^{M} \theta_m^{x_m} e^{-T_{gm}\theta_m}.$$

Here we considered four types of p-values, representative of the approaches suggested in the Technical Comments. The four p-value types differed by whether they accounted for the two-stage design, context, neither or both. The issue of coverage needs to be addressed by EB approaches in this case, because the p-value is 1 for all genes that are not discovered, irrespective of coverage. The "context-independent" approach is to set z_g to be the total number of mutations found in gene g irrespective of context, and compute the p-value as the probability of observing a total that is as large or larger. The "context-specific", by contrast, is to use the p_g as a one-dimensional summary of the vector of mutation counts and compute the p-value as the probability of a p_g as small or smaller. Approaches accounting for the two stages use expression (10) for determining the tail probability, while approaches not accounting for the two stages use a product of binomials. Accounting for the two-stages in the p-value calculations significantly increases the computational

Context	**Breast**			**Colon**		
	External	2.0	Internal	External	2.0	Internal
C or G in CpG	2.990	4.339	6.188	7.730	13.678	14.658
C or G in TpC or GpA	2.480	4.288	6.115	0.960	1.363	1.460
A	0.760	1.200	1.711	0.560	0.768	0.823
C not in CpG or TpC	1.380	1.680	2.397	0.950	1.464	1.569
G not in CpG or GpA	1.070	1.687	2.406	0.850	1.201	1.287
T	0.300	0.489	0.697	0.510	0.860	0.922
ins del	0.097	0.155	0.222	0.079	0.128	0.138

Table 3: Passenger rates scenarios (mutations per Mb)

burden. Abbreviations for these p-values are summarized in Table 2. For clarity, we use the names of the lead authors of the Technical Comments to identify the corresponding approaches.

3 Results

3.1 Passenger Rates

Determination of passenger rates is complex and controversial. See [4] for a discussion of some of the challenges. For the analyses presented here, we consider three sets of rates, summarized in Table 3. The External rates are based on the prior studies described in [4]. The rates for non-synonymous point mutations are determined so that the overall mutation rate, when applied to the CCDS genes, is set at 1.2 mutations per Mb. This is the rate used by Sjöblom *et al.* and is more then twice the experimentally determined passenger rate. The rate for insertions and deletion is set so that, at a rate of 1.2 mutations per Mb, the passengers' ratio of non-synonymous point mutations and indels is the same as that observed in the Discovery Stage. This results in a lower estimate than used in Sjöblom *et al.*, where it was assumed that the indels rates were much higher than the rates observed in the Discovery Stage. This assumption was unnecessarily conservative, as the Discovery Stage rate itself provided a theoretical upper bound.

The analysis assumed that the mutation spectrum observed in the study was no different from that of unselected passenger mutations and that both were a result of the same underlying processes and exposures to exogenous agents.

The Internal rates were calculated by eliminating all the Validated genes, and assuming that the remainder were harbor only passenger mutations. These rates constitute an upper bound to the true passenger rates as long as there are fewer passenger mutations in the validated gene set than there are non-passenger mutations in the set of genes that were discovered but not validated. This circumstance is highly plausible, though not estimable from this data without risking circularity. The rates labeled 2.0 are determined in the same way except that the factor 1.2 is replaced by 2.0, providing an intermediate scenario. This was the rate used by Rubin and Green.

	F	R	GF	GB	S
AUC	0.916	0.916	0.923	0.923	0.921
True Sensitivity: Frequentist	0.697	0.788	0.848	0.922	0.994
True Sensitivity: Empirical Bayes	0.978	0.979	0.980	0.982	0.978
True FDR: Frequentist	0.012	0.027	0.036	0.059	0.112
True FDR: Empirical Bayes	0.105	0.105	0.103	0.103	0.103

Table 4: Performance of alternative methodologies using a genome of 12521 passengers and 500 non-passengers

3.2 Simulation Results

In this section we consider simulation experiments in which a known number of non-passenger genes is included in the simulated genomes. Specifically, we consider genomes composed of 500 genes mutated at rates above the passenger rates, and 12521 genes mutated at the breast cancer passenger rates. Such cancer genomes are consistent with the results of Sjöblom *et al.* but have the advantage that, in the simulated case, we know exactly which genes are passengers and which are not. We simulated the mutations observed in each non-passenger by assigning it the rate empirically observed in one of the validated genes in breast cancer, randomly chosen.

For each method, we computed the true proportion of false discoveries in lists formed using a nominal FDR of 10%. A well calibrated method is one wherein the true proportion of false discoveries is also 10%. If the true proportion is lower, the method is unnecessarily conservative in setting the threshold. We perform this calculation with both the Benjamini-Hochberg correction (labeled Frequentist in the table) and with the EB correction. The EB values are within Monte Carlo error of the correct value of 10% and show no bias. The Frequentist values are biased towards lower values, i.e. they are too conservative. For example, the true FDR was actually only 1% in method F and 4% in method GF, whereas it should be 10% for both. The original Sjöblom *et al.* method, consisting of applying the Benjamini-Hochberg correction directly to p_g is .112, provides a slightly anticonservative estimate, as expected given the nature of the approximation.

Table 4 also indicates the sensitivity of the various approaches when assessed through simulation. Sensitivity is defined as the proportion of validated non-passengers that are included in the lists drawn using a nominal 10% FDR. As the task at hand is to identify non-passengers among validated genes, this is the relevant sensitivity to consider. The results emphasize the high proportion of true discoveries that are missed by the frequentist methods as a result of the lack of calibration. One of the reasons for the poor performance (technically bias, or lack of calibration) of the frequentist approaches is that, when no mutations are found in a gene, p-values are equal to one, irrespective of gene size, context, and coverage. Therefore, any method that estimates the FDR based only on the mutated genes will end up ignoring critical information about the vast majority of the analyzed genes.

These conclusions remain true over a broad range of choices for the total number of non-passengers. The extent of the biases will depend on the specific distribution of rates for non-

	F	R	GF	GB	S
AUC	0.893	0.893	0.907	0.907	0.900
True Sensitivity: Frequentist	0.520	0.566	0.679	0.749	0.953
True Sensitivity: Empirical Bayes	0.738	0.735	0.779	0.776	0.762
True FDR: Frequentist	0.019	0.031	0.057	0.092	0.283
True FDR: Empirical Bayes	0.109	0.112	0.113	0.111	0.111

Table 5: Performance of alternative methodologies using a genome of 12521 passengers and 500 non-passengers assuming a higher passenger mutation rate

passengers.

The results of simulations including non-passengers and using the Internal or 2.0 rates instead of the External rates are shown in Table 5. The major conclusions indicated earlier remain true, as all the methods used to draw the conclusions in the Technical Comments retain a large bias. The frequentist method that employs two-stages and context specific p-values (method GB) is now less biased, while the method of Sjöblom *et al.*(method S) is more biased than before. With EB, the FDR obtained with both the GB and S methods are identical (11.1%), close to the intended FDR of 10%.

Finally, table S4 shows the discrimination performance of the four sets of p-values for distinguishing between drivers and passengers, as measured by the area under the ROC curve (AUC). The results indicate that the small differences in ranking between CaMP and other methods do not translate into a loss of discrimination: the CaMP score offers the same discrimination between passengers and non-passengers as do the most discriminating among the scores proposed in the Technical Comments. The AUC for scores that account for context is higher than that for methods that do not, though in the scenario examined here the AUC is high in both cases.

3.3 Experimental Results on CCDS Genes

We now move to a comparison of the analytic approaches presented so far on the data generated by Sjöblom *et al.* One of the most important aspects of all approaches is the ranking of the candidate genes that will be used to prioritize them for future studies. The ranking of genes by cancer mutation prevalence (CaMP) scores is similar to that provided by the other statistical methods used. The rank correlations between CaMP and all the methods proposed in the Technical Comments are greater than 0.85.

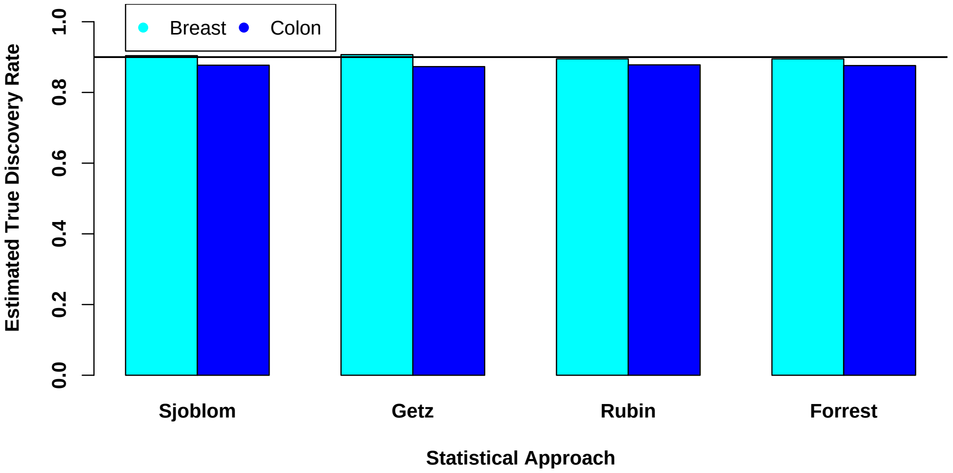
Next, in Table 6, we consider the results of a simulation that enumerates the genes that are mutated in the Discovery and Validation Stages when generating data from passenger rates. The results show that one would expect only about 17 genes to be mutated in both the Discovery and the Validation Stages for either breast or colorectal cancers, while 137 and 105 genes were actually found to be mutated in the Validation Stage of the experiment. The remaining genes (120 and 88) are therefore expected to be mutated at higher than passenger rates.

To compare the effect of the alternative passenger rates on the FDR, we used each in turn in

Table 6: Comparison of experimental results to the simulated results of a genome consisting entirely of passengers.

		Number of genes in Discovery	Number of genes in Validation	
Colon	Observed	519	105	
	Simulated	255	16	(s.d. 4)
Breast	Observed	673	137	
	Simulated	264	17	(s.d. 4)

Fig. 1: Proportion of CAN−genes expected to be mutated above background according to various statistical models. See SOM for details.



our EB analysis, which properly accounts for the three issues outlined in the introduction. The differences between the analyses are minimal, as shown in Figure 3.3. A reanalysis of the results in Figure 1 using Internal and 2.0 rates instead of External rates is presented in table 7. While it is true that increasing the passenger rates will increase the FDR of the CAN-gene list, the lists retain FDR's that support the conclusions of the Sjöblom *et al.* study, for the reasons discussed above. In particular, the conclusion that a large number of genes are mutated at rates greater than the passenger rate is confirmed: 61 to 95 of the CAN-genes identified in breast cancers and 41 to 49 of the CAN-genes identified in colorectal cancers are predicted to be mutated at frequencies above these higher passenger mutation rates based on table 7.

4 Additional Methodology used in Wood *et al.*

Wood *et al.* introduced several modifications for the analysis of the larger number of genes included in the RefSeq database. These included 1) a more detailed classification of mutations types; 2) a more accurate accounting of coverage; 3) different approaches to the estimation of passenger

Method	2.0		Internal	
	Breast	Colon	Breast	Colon
S	0.235	0.306	0.481	0.359
GF	0.232	0.315	0.474	0.376
F	0.250	0.306	0.506	0.360
R	0.247	0.306	0.506	0.360
GB	0.228	0.337	0.474	0.401

Table 7: Estimated False Discovery Rates for for the top 69 genes in colon and top 123 genes in breast assuming higher passenger mutation rates

rates; 4) the use of passenger probabilities rather then false discovery rates as the main measure of statistical uncertainty; 5) an additional phase of validation which required a different analytic approach.

4.1 Mutation Types, Coverage, and Passenger Rates.

The classification of mutation types in Wood *et al.* included 24 possible nonsynonymous substitutions, plus indels, giving $M = 25$. The types are listed in Tables 8 and 9. The calculation of coverage also differed between Sjöblom *et al.*and Wood *et al.* , not only because the number of mutation types was different, but also because of inclusion of a factor that takes into account gene-specific differences in potential nonsynonymous mutations. When considering base pair substitution mutations, Wood *et al.* considered only nucleotides-at-risk, that is those nucleotides that could result in a non-synonymous mutation when altered. The nucleotides-at-risk were determined for each gene, in a way broadly similar to that used by Greenman *et al.* [2]. For example, missense mutations at the third position of many codons would not result in a nonsynonymous mutation so were excluded from consideration. In both studies, all the nucleotides successfully sequenced are considered to be available for indels.

To determine passenger mutation rates, three independent approaches were used, termed ”External”, ”SNP” and ”NS/S” for nonsynonymous over synonymous. The External rates in colorectal cancers were determined using experimentally observed rates of mutations in noncoding sequences in variant detection oligonucleotide microarrays, as described in [6]. This led to an estimate of 0.55 nonsynonymous mutations/Mb, based on

$$0.55 = 1.2 \times 0.5 \times (1 - 0.16) + 0.6 \times 0.5 \times 0.16$$

Here 1.2 is the observed rate of mutations per Mb in non-coding diploid DNA, .5 is the fraction of non-coding diploid DNA, and .16 is the fraction of haploid tumors. The External rate for breast cancers was assumed to be 0.33 nonsynonymous mutations/Mb.

The SNP-based and NS/S-based approaches both estimate the passenger mutation rates from the synonymous mutations discovered in the study by first determining the expected nonsynonymous to synonymous mutation ratios. These were estimated in two ways. In the SNP-based case

Contexts	External	SNP: Dis	SNP: Val	NS/S: Dis	NS/S: Val
C to G in CpG	0.196	0.831	0.438	2.149	1.132
C to A in CpG	0.115	0.489	0.258	1.264	0.666
C to T in CpG	0.658	2.785	1.469	7.206	3.794
G to C in CpG	0.219	0.928	0.490	2.402	1.265
G to A in CpG	0.923	3.908	2.062	10.114	5.326
G to T in CpG	0.196	0.831	0.438	2.149	1.132
C to G in TpC	0.595	2.516	1.327	6.512	3.429
C to A in TpC	0.158	0.667	0.352	1.726	0.909
C to T in TpC	0.423	1.789	0.944	4.629	2.437
G to C in GpA	0.648	2.743	1.447	7.098	3.737
G to A in GpA	0.377	1.596	0.842	4.131	2.175
G to T in GpA	0.186	0.787	0.415	2.036	1.072
A to C	0.051	0.215	0.113	0.557	0.293
A to G	0.102	0.430	0.227	1.114	0.586
A to T	0.074	0.314	0.166	0.813	0.428
C to G not in CpG or TpC	0.098	0.415	0.219	1.073	0.565
C to A not in CpG or TpC	0.127	0.538	0.284	1.393	0.733
C to T not in CpG or TpC	0.119	0.503	0.265	1.301	0.685
G to C not in CpG or GpA	0.092	0.391	0.206	1.012	0.533
G to A not in CpG or GpA	0.191	0.810	0.427	2.095	1.103
G to T not in CpG or GpA	0.140	0.591	0.312	1.530	0.806
T to C	0.049	0.208	0.110	0.539	0.284
T to G	0.033	0.139	0.073	0.359	0.189
T to A	0.046	0.194	0.103	0.503	0.265
indels	0.030	0.120	0.060	0.320	0.170

Table 8: Passenger mutation rates used in the Wood *et al.* study to analyze mutations in breast tumors.

Contexts	External	SNP: Dis	SNP: Val	NS/S: Dis	NS/S: Val
C to G in CpG	0.109	0.198	0.289	0.468	0.680
C to A in CpG	0.196	0.357	0.520	0.843	1.225
C to T in CpG	4.515	8.202	11.954	19.393	28.168
G to C in CpG	0.131	0.238	0.346	0.562	0.816
G to A in CpG	5.212	9.470	13.802	22.391	32.523
G to T in CpG	0.327	0.594	0.866	1.405	2.041
C to G in TpC	0.244	0.443	0.645	1.047	1.520
C to A in TpC	0.149	0.271	0.394	0.640	0.929
C to T in TpC	0.284	0.516	0.753	1.221	1.773
G to C in GpA	0.110	0.201	0.292	0.474	0.689
G to A in GpA	0.401	0.729	1.063	1.725	2.505
G to T in GpA	0.321	0.584	0.850	1.380	2.004
A to C	0.083	0.151	0.220	0.357	0.518
A to G	0.119	0.217	0.316	0.513	0.745
A to T	0.096	0.174	0.254	0.413	0.599
C to G not in CpG or TpC	0.063	0.114	0.167	0.271	0.393
C to A not in CpG or TpC	0.158	0.286	0.417	0.677	0.983
C to T not in CpG or TpC	0.158	0.286	0.417	0.677	0.983
G to C not in CpG or GpA	0.081	0.148	0.215	0.349	0.507
G to A not in CpG or GpA	0.260	0.472	0.688	1.117	1.622
G to T not in CpG or GpA	0.211	0.384	0.559	0.907	1.318
T to C	0.096	0.175	0.254	0.413	0.600
T to G	0.105	0.191	0.279	0.453	0.658
T to A	0.062	0.113	0.164	0.266	0.387
indels	0.040	0.070	0.100	0.150	0.210

Table 9: Passenger mutation rates used in the Wood *et al.* study to analyze mutations in colorectal tumors.

we calculated this ratio based on coding SNPs identified in previous sequencing studies [16, 3]. The ratio of nonsynonymous (NS) to synonymous (S) mutations in these studies was 1.02. This ratio may be an underestimate of the true passenger mutation rate because the selection against NS mutations may be more stringent in the germline than during tumor development. In the NS/S-based approaches we therefore determined the NS/S ratio as follows. Context-specific mutation rates were used to determine the expected frequency of mutations that would create NS vs. S mutations. Each nucleotide of each codon was mutated in silico to determine whether a particular change would result in a NS or S change, thereby accounting for all possible changes to all bases of each codon. The fraction of changes resulting in NS and S alterations were adjusted to account for the type of base that was mutated, the base change that resulted from the mutation, the immediate 5' and 3' neighbors to the mutated base, and codon usage. Through analysis of all RefSeq genes, we determined that the expected NS/S ratios were 2.41 and 2.65 in colorectal and breast cancers, respectively. These estimates provide a theoretical upper bound to the true mutation rate because they do not take into account the fact that nonsynonymous mutations that retard cell growth will be selected against during tumorigenesis.

The products of these ratios and the observed synonymous mutation rates in each screen yielded the two different estimates of the passenger mutation rates, termed "SNP" and "NS/S" respectively. For example, the rate of synonymous mutations in the colorectal cancer Discovery Screen was 0.97 mutations/Mb. The SNP-based passenger rate was therefore estimated to be 0.99 NS mutations/Mb (=0.97 x 1.02) while the NS/S-based passenger rate was 2.35 NS mutations/Mb (=0.97 x 2.41). These were further broken down by context according to the relative frequency of observed mutation types. The complete set of rates is given in Tables 8 and 9.

4.2 Analysis of mutation prevalence study.

In addition to the Discovery and Validation stages, Wood *et al.* experimentally tested 40 CAN-genes in a separate cohort of 96 cancers. Because the process of selection of these 40 genes for further study could not be easily represented in terms of mutation counts, it was difficult to generate reference distributions such as the ones used to compute passenger probabilities for the Discovery and Validation Screens. We therefore developed an analytic method designed to be insensitive to the selection process. For this we used the Empirical Bayes approach of Section 2.3 to estimate of the probability π_g of gene g being a passenger, and used this as our prior probability. For each of the 40 genes in the mutation prevalence study, we then computed a Bayes Factor, based on the results of the mutation prevalence study alone, for the hypothesis that the gene was mutated at the passenger mutation rate. Computation of the Bayes Factor requires specification of a prior distribution of mutation rates that corresponds to the alternative hypothesis. To construct this distribution, we assumed that, for each of non-passenger gene, each of the non-passenger mutation rates θ'_{gm} followed Gamma distributions

$$\theta'_{gm} \sim \mathrm{Ga}(\alpha, \beta_m)$$

These are assumed to have the same shape parameter α and scale parameters β_m set so that the mean non-passenger rates are equal to the corresponding passenger mutation rates multiplied by a single scaling factor ϕ common to all contexts, that is

$$\beta_m = \frac{\alpha - 1}{\phi \theta_m}.$$

We next assumed that mutations have a Poisson distribution. This formulation was preferred to the binomial for computational convenience: using standard conjugate analysis calculations we can integrate out the θ_m's and operate with marginal distributions that depend on α and ϕ only. For the null case of passenger mutation rates we set $\phi = 1$ and set the shape parameter to a fixed value that allows for a very small amount of variation in the rates. In this way both the null and the alternative model have the same dimensionality.

The shape parameter α and the scaling factor ϕ were estimated empirically as follows. Drawing from the prior probabilities we randomly assigned each of the 140 CAN-genes to a true status of either passenger or non-passenger. We then estimated α and ϕ by maximizing the marginal likelihood. For each draw, we evaluated the Bayes' factor at the estimated α and ϕ and used Bayes' rule to combine the prior and Bayes Factor into the posterior probabilities. We then averaged the posterior probabilities over draws to determine the final estimate. Posterior estimates varied only slightly across draws. This method controlled for multiple testing via the prior distribution, which is the local false discovery rate from the Empirical Bayes analysis.

Acknowledgments

We thank Robert Elston, Gaddy Getz, Holger Hölfing, Sandy Markowitz, Peter Müller, Michael Newton, Terry Speed and Rob Tibshirani, for useful comments on earlier drafts of this analysis.

Software

A free, open-source, R package for replicating these analyses is available with the supplementary materials to [4] and at `http://astor.som.jhmi.edu/∼gp`.

References

[1] Sjöblom T; Jones S; Wood LD; et al. The consensus coding sequences of human breast and colorectal cancers. *Science*, 314(5797):268–274, Oct 2006.

[2] Greenman C; Stephens P; Smith R; et al. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–158, Mar 2007.

[3] Greenman C; Wooster R; Futreal PA; Stratton MR; Easton DF. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics*, 173(4):2187–2198, Aug 2006.

[4] Parmigiani G; Lin J; Sjöblom T; et al. Response to comments on ’The consensus coding sequences of breast and colorectal cancers’. *Science*, 317 (5844):1500d, 2007.

[5] Efron B; Tibshirani R; Storey JD; Tusher V. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

[6] Wood L; et al. The genomic landscapes of human breast and colorectal cancers. *Science*, 2007.

[7] Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73:751–754, 1986.

[8] Efron B; Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23(1):70–86, 2002.

[9] Müller P; Parmigiani G; Rice K. FDR and Bayesian multiple comparisons rules. In JM Bernardo; S Bayarri; JO Berger; A Dawid; D Heckerman; AFM Smith; M West, eds., *Bayesian Statistics 8*. Oxford University Press, 2007.

[10] Do KA; Müller P; Tang F. A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society Series C Applied Statistics*, 54(3):627–644, 2005.

[11] Benjamini Y; Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57:289–300, 1995.

[12] Müller P; Parmigiani G; Robert C; Rousseau J. Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association*, 99:990–1001, 2005.

[13] Getz G; Höfling H; Mesirov JP; Golub TR; Meyerson M; Tibshirani R; Lander ES. Comment on ”the consensus coding sequences of human breast and colorectal cancers”. *Science*, 317(5844):1500b, Sep 2007.

[14] Forrest WF; Cavet G. Comment on ”the consensus coding sequences of human breast and colorectal cancers”. *Science*, 317(5844):1500a, Sep 2007.

[15] Rubin AF; Green P. Comment on ”the consensus coding sequences of human breast and colorectal cancers”. *Science*, 317(5844):1500c, Sep 2007.

[16] Ng PC; Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*, 31(13):3812–3814, Jul 2003.