1-19-2007

TRAB: TESTING WHETHER MUTATION FREQUENCIES ARE ABOVE AN UNKNOWN BACKGROUND

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

Sining Chen
The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University

Victor E. Velculescu
The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University

TRAB: Testing Whether Mutation Frequencies Are Above an Unknown Background

Giovanni Parmigiani
Sining Chen
Victor E. Velculescu
The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University
January 18, 2007

Abstract

Summary: To rigorously determine whether a gene or a population of genes have alterations that are involved in carcinogenesis requires comparison of the prevalence of identified changes to the background mutation frequency present in tumor DNA. To facilitate this task, we develop a testing approach and the associated R library, called TRAB, that evaluates whether the frequency of somatic mutation is higher than an unknown, but estimable, background. We test the null hypothesis that the frequency belongs to background population of frequencies against the alternative hypothesis that the frequency is higher. Background mutation frequencies are themselves allowed to be variable. TRAB computes the *a posteriori* probability and the Bayes factor for the hypothesis using a hierarchical Bayesian approach.

Software Availability: `http://astor.som.jhmi.edu/~gp/trab/`

Contact: `gp@jhu.edu`

Introduction

A major avenue of study in cancer research is the identification of somatic mutations in key oncogenes and suppressor genes. However, to rigorously determine whether a gene or a population of genes have alterations that are pathogenically important requires comparison of the prevalence of identified changes to the background mutation frequency present in tumor DNA. Background or passenger mutations accumulate in tumor DNA from replication errors through repeated rounds of normal somatic cell division in tumor precursor cells, as well as through multiple waves of selection and clonal expansion that occurs throughout tumorigenesis. To facilitate research on the interpretation of potentially pathogenic mutations, we developed a testing approach and an R library, (Ihaka and Gentleman, 1996) called TRAB, that evaluates whether the frequency of somatic mutation is higher than an unknown, but estimable, background. The background rates themselves are allowed to vary across genes. We thus test the null hypothesis (H_0) that the frequency belongs to background population of frequencies against the alternative hypothesis (H_1) that the frequency is higher. The background population of frequencies is estimated based on the analysis described by Wang et al., 2002 and is currently built-in in the function. The TRAB library computes the *a posteriori* probability of the alternative hypothesis and the Bayes factor, using a hierarchical Poisson model that accounts for uncertainty in estimated input quantities as well as biological heterogeneity of background prevalence. This procedure is utilized in Wang et al., 2002, and summarized there in one line of text. This article provides the full details of the model, as well as a discussion of the functionality of the software. This methodology is likely to be of wide applicablility as large sequencing project are coming to the fore (Sjöblom et al., 2006).

Methods and Algorithm

Experimental data. Consider a candidate gene g^*, observed to have n_{g^*} mutations over a total of l_{g^*} Mb sequenced. This information needs to be compared to a background distribution. We develop a general testing approach and illustrate it using

the background distribution described in Wang et al., 2002, which includes the length $l_1, l_2, \ldots, l_G$ of the sections sequenced and the numbers $n_1, n_2, \ldots, n_G$ of mutations found for $G = 475$ genes. These genes are known not to be involved in carcinogenesis.

Statistical model for background frequencies. Our procedure acknowledges that there may be heterogeneity of somatic mutation frequencies within the background gene set by using a two-stage Poisson-Gamma model: the Poisson stage describes the randomness in the occurrence of mutations within a gene, and the Gamma stage describes the variation of the mutation frequency across genes (Schervish, 1995).

In the first stage of the model, each gene $g, g = 1, 2, \ldots, G$ is assumed to have its own Poisson rate λ_g of mutations per Mb. Gene counts are assumed to be independent conditional on the frequencies $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_G)$. The sampling distribution of the background group is then

$$p(n_1, \ldots, n_G | \boldsymbol{\lambda}, \boldsymbol{l}) = \prod_{g=1}^{G} \frac{1}{n_g!} (l_g \lambda_g)^{n_g} e^{-l_g \lambda_g}. \tag{1}$$

We can condition on the gene lengths because they are not likely to be related with increased frequencies of somatic mutations in cancer.

The second stage, or genomic distribution, describes the variation of λ_g's using a gamma distribution with parameters α_0 and β_0, defined as

$$p(\lambda_g | \alpha_0, \beta_0) = \frac{\beta^{\alpha_0}}{\Gamma(\alpha_0)} \lambda_g^{\alpha_0 - 1} e^{-\lambda_g \beta_0}, \qquad g = 1, \ldots, G \tag{2}$$

We work with the mean/coefficient-of-variation reparameterization given by letting $\mu_0 = \alpha_0/\beta_0$, which is the mean frequency in the background population, and $\nu_0 = \sqrt{\alpha_0}$, which is the coefficient of variation of the frequencies in the background population. Data provide information about μ_0, but are typically insufficient to distinguish among large values of ν_0 because of the high prevalence of genes with no mutations. We use half normal prior distributions on both parameters, that is $\mu_0 \sim N^+(0, \sigma_\mu)$ and $\nu_0 \sim N^+(0, \sigma_\nu)$. In the absence of more specific information, we use default values of $\sigma_\mu = 100$ and $\sigma_\nu = 10$, leading to widely dispersed priors. Higher values of σ_ν lead to similar conclusions but can produce numerical instabilities.

The sampling distribution for n_g given μ_0 and ν_0 is a negative binomial, obtained

by analytically integrating out λ_g. Application of Bayes rule lead to the *a posteriori* distribution $p(\mu_0, \nu_0 | n_1, \ldots, n_G)$.

Hypotheses. We compare two hypotheses: the null hypothesis H_0 states that the mutation frequency λ_{g^*} for the candidate gene belongs to the background group; the alternative hypothesis H_1 that λ_{g^*} belongs to a group of unknown average frequency μ_1 and unknown coefficient of variation ν_1. *A priori* uncertainty about μ_1 and ν_1 is assumed to be $\mu_1 \sim N^+(0, \sigma_\mu)$ and $\nu_1 \sim N^+(0, \sigma_\nu)$, the same as for the background group, with the additional constraint that the mean mutation frequency in the alternative population is greater than that in the background population, that is, $\mu_1 > \mu_0$.

Bayes factor. The Bayes factor (Kass and Raftery, 1995) in favor of H_0 is

$$B = \frac{P(n_{g^*} | H_0, l_{g^*})}{P(n_{g^*} | H_1, l_{g^*})}. \tag{3}$$

This is determined in two steps:

Step 1. We get

$$p(\lambda^* | H_0, l_{g^*}) = \int_0^\infty \int_0^\infty p(\lambda_g | \mu_0, \nu_0) p(\mu_0, \nu_0 | n_1, \ldots, n_G) d\mu_0 \, d\nu_0 \tag{4}$$

$$p(\lambda^* | H_1, l_{g^*}) = \int_0^\infty \int_0^\infty \int_{\mu_0}^\infty p(\lambda_g | \mu_1, \nu_1) p(\mu_1, \nu_1) p(\mu_1 | n_1, \ldots, n_G) d\mu_1 \, d\nu_1 \, d\mu_0 \tag{5}$$

In our implementation these are computed off-line by numerical integration, using the function `trab.setup()`. Modifications of the type of *a priori* distributions and background frequency data require changes to the source code of `trab.setup()`.

Step 2. For each user–supplied n_{g^*} and l_{g^*}, we evaluate

$$P(n_{g^*} | H_0, l_{g^*}) = \int_0^\infty p(n_{g^*} | \lambda^*, l_{g^*}) p(\lambda^* | H_0, l_{g^*}) d\lambda^* \tag{6}$$

$$P(n_{g^*} | H_1, l_{g^*}) = \int_0^\infty p(n_{g^*} | \lambda^*, l_{g^*}) p(\lambda^* | H_1, l_{g^*}) d\lambda^* \tag{7}$$

For a given *a priori* probability π that the candidate gene has a higher mutation frequency, the *a posteriori* probability π^* is computed as

$$\pi^* = \frac{\pi}{\pi + (1 - \pi) B} \tag{8}$$

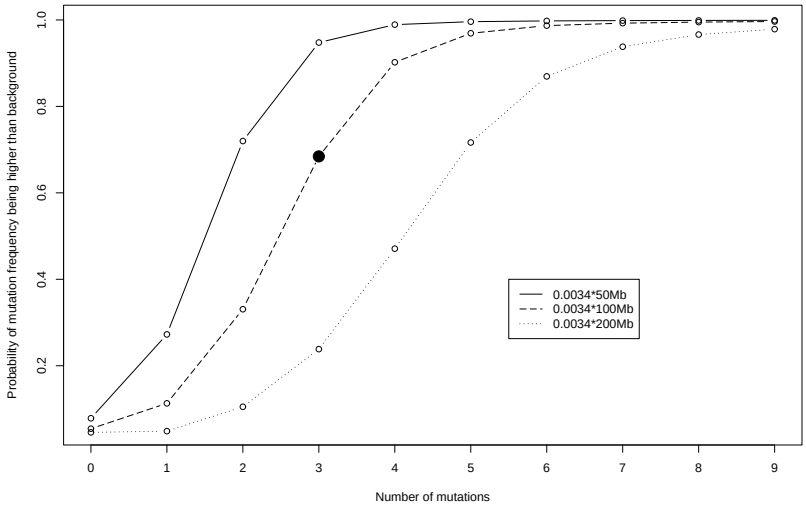

Figure 1: Posterior probability of the mutation frequency being higher than the background frequency versus total number of mutations. Each line corresponds to a specific length of DNA sequenced. For example, the solid circle represents the posterior probability of the frequency being higher than background if three mutations are observed on a total length of 0.0034Mb $\times$ 100. It can be calculated by `trab(3, 0.0034*10)`.

Illustration. Generally, posterior probabilities of H_1 increase with the number of mutations and decrease with the size of the region sequenced, all else being equal. See Figure 1 for an illustration. For very extreme frequencies, especially those produced by extremely small regions, this behavior may not hold. The reason is that the alternative population is a proper distribution of frequencies and exorbitant mutation frequencies become unlikely even under the alternative. These situations are unlikely to occur in practice.

Multiple candidate genes. When multiple candidate genes are input, we assume that the frequencies are conditionally independent given hyperparameters σ_μ and σ_ν, and compute the probability π for each. We then determine the most likely assignment of genes to the background or elevated frequency, and compute its joint probability under independence of the genes. Unlike in a clustering algorithm, correlations of genes are not considered: each gene is tested against a separate alternative population.

Implementation

TRAB is distributed as a library under the open source environment R (Ihaka and Gentleman, 1996), which has important applications in computational biology (Gentleman et al., 2004).

The main function that tests the hypothesis H_0 is `trab()`. It takes the following inputs:

`Mb`**: Gene length(s)** The number(s) of base-pairs of the candidate gene(s) sequenced. It should be a vector of length G if a total of G genes are considered.

`Nmutations`**: Number(s) of mutations** The number(s) of mutations found in the candidate gene(s). If there are G candidate genes, each number corresponds to one gene.

`priorHo`**: Prior probability of the H_0 hypothesis** The prior probability that the mutation frequency of each candidate gene belongs to the same population of frequencies as the background. The default is 0.5 for each gene, representing even odds that the frequency is above the background. When multiple genes are tested, each entry in this vector should be set at the likely value of the the precentage of genes in the group sequenced that are expected to be above background.

`marginal`**: Marginal distribution of the background frequencies** A list that contains the numerical marginal probability density function of the background frequencies. The default frequency can be loaded from the data object "lambda" that is built into the package. It is calculated using the function `trab.setup()` from background data reported in Wang et al., 2002. See the description of function `trab.setup()`.

`verbose`**: Format of output** Logical: if TRUE `trab()` outputs both text and numerical results. If FALSE outputs only Bayes factors.

When there is only one candidate gene, the `trab()` function outputs two quantities (only the latter if `verbose` is FALSE): 1) the probability that the input mutation frequency is above the background; 2) the Bayes factor in favor of the frequency being same as the background.

When there is more than one candidate genes, the function `trab()` outputs: 1) the IDs of the group of genes whose posterior probabilities of the alternative hypothesis are greater than 0.5; 2) the probability of the genes in the above group have above-background mutation frequencies while the rest do not. In other words, when sequence data of multiple genes are available, `trab()` reports the most likely assignment of genes to background or elevated frequency, and the probability of that assignment.

If the user wishes to establish the background mutation frequency from a different tumor type based on experimental data, or change/update the current background distribution when new data are available, this can be readily incorporated using the function `trab.setup()`, which can produce a new object that contains the new numerical marginal probability density distribution using the new data. The inputs to `trab.setup` are:

`priorsd.mm` The prior standard deviation for the population mean of the of background mutation frequencies. The default value is 100.

`priorsd.cv` The prior standard deviation for the Coefficient of Variation. The default value is 10. Used in conjuction with a `priorsd.mm` of 100, it gives rise to a weak prior for the background mutation frequencies.

`Ngrid` The number of grid points for the numerical integration in Expressions 4 and 5.

`backgr.lengths, backgr.nmutations` Data used to estimate the population parameters of background frequencies. Units should be the same as the inputs `Mb, Nmutation`. The default values are taken from Wang et al., 2002 and built-in with the package as a list "trab.input" with two fields "lg" and "ng", originally made with the function `trab-makeinput()`. The user can specify alternative values by either 1) constructing a similar list and saving it as an external R object file (this can be accomplished by changing and re-running the function `trab.makeinput()`) or 2) entering the vectors directly.

`outfile` Name of the R data file containing the output.

The output of `trab.setup()` is an R data object that is intended to be used as

the input `marginal` in the main function `trab()`. It will be saved directly to the file specified by `output`, such that `trab.setup()` only needs to run once for each new background dataset.

The function `trab.makeinput()` prepares the inputs to `trab.setup()`: `backgr.lengths` and `backgr.nmutations`.

Discussion

The methodology and software we have developed should be immediately useful to researchers wishing to evaluate whether mutations in a gene or collection of genes occur at a prevalence that is higher than the background mutation frequency present in tumors. However, some caveats should be considered. First, the background population of rates is estimated using colorectal cancers. Usage in investigation of other tumor types is legitimate as long as those cancers behave similarly in terms of the accumulation of background mutations. This is likely to be the case for many solid cancers. Additionally, this test has been designed for analysis of cancers that do not have known abnormalities in DNA repair mechanisms (i.e. mismatch repair deficiencies). Finally, the test should not be used to determine if a single observed mutation is above the background mutation prevalence, as it is impossible to establish an accurate mutation frequency based on a single alteration. In such a case, further sequencing should be performed to identify additional mutations in the same gene, and then those results can be compared to the background frequency.

The TRAB library is publicly available at `http://astor.som.jhmi.edu/~gp/software/trab/`.

References

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5(10):R80

Ihaka R, Gentleman R (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 5:299–314

Kass RE, Raftery AE (1995). Bayes factors. *Journal of the American Statistical Association* 90:773–795

Schervish MJ (1995). *Theory of Statistics*. Springer-Verlag

Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JKV, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE (2006). The consensus coding sequences of human breast and colorectal cancers. *Science* 314(5797):268–274

Wang TL, Rago C, Silliman N, Ptak J, Markowitz S, Willson JKV, Parmigiani G, Kinzler KW, Vogelstein B, Velculescu VE (2002). Prevalence of somatic alterations in the colorectal cancer cell genome. *Proc Natl Acad Sci U S A* 99(5):3076–3080