Bioinformatics, 35(22), 2019, 4776–4778
doi: 10.1093/bioinformatics/btz486
Advance Access Publication Date: 14 June 2019
Applications Note

OXFORD

Genome analysis

Testing clonal relatedness of two tumors from the same patient based on their mutational profiles: update of the *Clonality* R package

Audrey Mauguen , Venkatraman E. Seshan, Colin B. Begg and Irina Ostrovnaya*

Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY 10017, USA

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on March 4, 2019; revised on June 4, 2019; editorial decision on June 5, 2019; accepted on June 10, 2019

Abstract

Summary: The *Clonality* R package is a practical tool to assess the clonal relatedness of two tumors from the same patient. We have previously presented its functionality for testing tumors using loss of heterozygosity data or copy number arrays. Since then somatic mutation data have been more widely available through next generation sequencing and we have developed new methodology for comparing the tumors' mutational profiles. We thus extended the package to include these two new methods for comparing tumors as well as the mutational frequency estimation from external data required for their implementation. The first method is a likelihood ratio test that is readily available on a patient by patient basis. The second method employs a random-effects model to estimate both the population and individual probabilities of clonal relatedness from a group of patients with pairs of tumors. The package is available on Bioconductor.

Availability and implementation: Bioconductor (http://bioconductor.org/packages/release/bioc/html/Clonality.html).

Contact: ostrovni@mskcc.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

When a patient presents with two tumors that are separated anatomically clonality testing can help to diagnose whether the tumors arose independently, or one tumor is a metastasis of the other. Several years ago the *Clonality* R package (Ostrovnaya *et al.*, 2011) implemented methods for testing whether two tumors from the same patient are clonally related. These involved the comparison of the loss of heterozygosity at a set of candidate genetic loci and a method that involved genome-wide copy number arrays to search for identical copy number gains or losses. Since the initial publication of the *Clonality* package, it has become commonplace in clinical practice for patients' tumors to be evaluated for the presence of somatic mutations in major cancer genes, or from exome sequencing. Evaluation of clonal relatedness of tumors in this context is based on the presence or absence of identical mutations in the two tumors and requires knowledge of the probabilities of occurrence of each of the observed somatic mutations. Our software update includes a statistical test for clonal relatedness based on somatic mutations (Ostrovnaya *et al.*, 2015) along with material to assist users in calculating these marginal mutation probabilities. In order to estimate these probabilities, we utilize mutation frequencies estimated from external datasets, i.e. The Cancer Genome Atlas (TCGA) (Ellrott *et al.*, 2018). The R package is now updated with a function that allows easy estimation of mutational frequencies. Moreover, a feature of the clonality test is that it only provides quantitative evidence against the hypothesis that the tumors are independent. Specifically, if no shared mutations are observed the test will always give a *P*-value of 1. In recognition of this limitation we developed a method that analyses data from a database of patients in the same clinical setting using random-effects methodology, in order to estimate the individual probabilities of clonal relatedness for each patient in the study, and by extension the overall relative frequency of clonality in the clinical setting under study (Mauguen *et al.*, 2018). Of note, a key parameter that needs to be estimated in each of these methods is the clonality signal ξ. This parameter represents the relative time during the evolution of the tumors in which mutations were accumulating in the original, clonal cell. This parameter, which ranges in value from 0 to 1, will tend to be large in tumors where the seeding of the metastasis occurred late in the evolutionary process. Conversely, it is 0 in tumors that are independent. This note provides instructions on how to use the software. More extensive details of the underlying methodology are provided in the Supplementary Material.

2 Available functionality

We present first an example of data added to the package to which the two additional methods can be applied. We then present the functions to use and how to interpret the outputs.

2.1 The *lcis* dataset

These data come from a study investigating whether in situ lesions are precursors of invasive breast cancers by looking at clonal relatedness in patients having lobular carcinoma *in situ* (LCIS) and either invasive lobular carcinoma, ductal carcinoma *in situ* or invasive ductal carcinoma (Begg *et al.*, 2016). Exome sequencing data included are from 34 samples from 17 patients. The dataset is in the format of a mutation annotation format (MAF) file. Each row in the dataset represents a mutation that was observed in the specific tumor. The columns include patient ID, Tumor_Sample_Barcode (sample ID), Hugo_Symbol (gene), mutation Chromosome and Start_Position and reference and alternative alleles. The function *create.mutation.matrix()* has to be used to reformat the dataset for further analysis. By default, it creates a binary matrix with rows being mutations and columns being samples.

2.2 Estimating mutational frequencies

The mutational frequencies required for performing these analyses are obtained for each possible locus separately and should be specific to the cancer type of the tumors. To assist in obtaining these frequencies we have included the data object *freqdata* with frequencies of all mutations in 33 cancer types profiled by TCGA. The function *get.mutation.frequencies()* extracts these frequencies based on two inputs: mutation IDs in the format {chromosome genomic.location reference.allele alternate.allele}, and TCGA subtype abbreviations. There is also an option for specifying a custom external mutation file from independent patients for the frequency estimation which can be used independently or added to TCGA data. An external dataset could be used if the cancer type was not profiled by TCGA, or if one wishes to use a specific subtype of interest (e.g. Luminal A breast cancer), or if the cohort assembled for clonality testing is substantial and can improve the accuracy of mutation frequency estimation. Note that the probability of an observed mutation can never be zero. To estimate the marginal probability of a mutation that was not observed in TCGA (or the source one elects to use) we have adopted the convention of estimating the probability of such a mutation as $1/(N+1)$, where N is the number of cases from which the estimate is based. The function *get.mutation.frequencies()* returns the vector of frequencies that can be used as input into the subsequent functions.

2.3 Testing for clonal relatedness using mutation data

The function *SNVtest()* tests whether the available mutation information of a tumor pair is evidence in favor of clonality using the method described in Ostrovnaya *et al.* (2015). It calculates the tail probability of the observed data under the null hypothesis that the two tumors are independent. The test is conditional on the mutations observed in either of the tumors. The arguments in the function are the two binary vectors of observed mutations in tumors 1 and 2, that can be obtained as the output of *create.mutation.matrix()*. The function also needs the vector *pfreq* of probabilities of occurrence for each mutation, which can be obtained using *get.mutation.frequencies()*. The null distribution is generated from simulations in which the same set of mutations is observed but matches are generated randomly assuming independence.

The function's output summarizes the numbers of mutations observed in each tumor, and the number of mutations shared by the two tumors. The likelihood ratio statistic is presented along with a *P*-value calculated using the null distribution generated using the specified mutational frequencies *pfreq*. A low *P*-value is evidence that the two tumors are clonally related. Finally, the estimate of the clonality signal ξ is presented. A value close to 1 indicates very similar mutational profiles, whereas low values indicate that most mutations occur in only one tumor.

Of note, our methods permit distinct testing approaches if both mutational and copy number profiles are available, as in Begg *et al.* (2016). In this study of LCIS, the two methods were broadly in agreement, with a few discordant cases. Clearly agreement is reassuring, but there are yet no formal methods for resolving discrepancies of this nature. A heuristic approach could be to conclude overall using Bonferroni correction, but the best approach is still to be determined.

2.4 Calculating individual diagnostic probabilities

The function *mutation.rem()* estimates the random-effects model parameters. It requires a collection of pairs of tumors from a suitable dataset. The first argument is a matrix of mutations formatted using *create.mutation.matrix()* with option *rem=TRUE*. As the estimation methods involves integration over the density of the clonality signal ξ, it is possible to control the range of values used in the integral calculation to obtain more precise estimates, at the price of more computational time. The function makes use of an Expectation-Maximization (EM) algorithm to estimate the parameters (see Supplementary Material), unlike in Mauguen *et al.* (2018), as subsequent research showed using the EM algorithm overall improved the precision of the estimates. Finally, the function allows the user to pick initial parameters values, e.g. if prior knowledge or assumptions on the population are available, the convergence criteria used and the maximum number of iterations, although the user can rely on the proposed default values.

The output presents the estimate of π, the proportion of clonal cases in the population of interest, as well as estimated values for μ and σ, the two parameters of the clonality signal density, assumed to be lognormally distributed. Criteria related to the likelihood maximization are presented: the value of the maximized likelihood, the convergence status and the number of iterations used to obtain convergence. In case of non-convergence, the user can increase the number of iterations or modify the initial values of the parameters.

Once the random-effects model parameters have been estimated, the probability of clonal relatedness for each case can be obtained in two ways. For the tumor pairs included in the analysis, adding the option *proba=TRUE* when calling *mutation.rem()* will compute the individual probabilities. To obtain the probability of clonality for a new pair, the function *mutation.proba()* can be called. Both methods will give, for each pair, a value between 0 and 1 representing the probability that the two tumors of the pair are clonal, based on their set of both shared and private mutations, and based on the a priori probability of clonality in this population π.

Funding

This work was supported by the National Cancer Institute [awards CA124504, CA163251 and CA008748].

Conflict of Interest: none declared.

References

Begg,C.B. *et al*. (2016) Clonal relationships between lobular carcinoma in situ and other breast malignancies. *Breast Cancer Res*., **18**, 66.

Ellrott,K. *et al*. (2018) Cancer Genome Atlas Research Network, scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst*., **6**, 271–281.e7.

Mauguen,A. *et al*. (2018) Estimating the probability of clonal relatedness of pairs of tumors in cancer patients. *Biometrics*, **74**, 321–330.

Ostrovnaya,I. *et al*. (2015) Using somatic mutation data to test tumors for clonal relatedness. *Ann. Appl. Stat*., **9**, 1533–1548.

Ostrovnaya,I. *et al*. (2011) Clonality: an R package for testing clonal relatedness of two tumors from the same patient based on their genomic profiles. *Bioinformatics*, **27**, 1698–1699.