BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 12 2011, pages 1698–1699
doi:10.1093/bioinformatics/btr267

Genome analysis

Advance Access publication May 5, 2011

Clonality: an R package for testing clonal relatedness of two tumors from the same patient based on their genomic profiles

Irina Ostrovnaya[1,*], Venkatraman E. Seshan[1], Adam B. Olshen[2] and Colin B. Begg[1]

[1]Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY 10065 and [2]Department of Epidemiology and Biostatistics and Helen Diller Family Comprehensive Cancer Center, University of California—San Francisco, San Francisco, CA 94158, USA

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: If a cancer patient develops multiple tumors, it is sometimes impossible to determine whether these tumors are independent or clonal based solely on pathological characteristics. Investigators have studied how to improve this diagnostic challenge by comparing the presence of loss of heterozygosity (LOH) at selected genetic locations of tumor samples, or by comparing genomewide copy number array profiles. We have previously developed statistical methodology to compare such genomic profiles for an evidence of clonality. We assembled the software for these tests in a new R package called '*Clonality*'. For LOH profiles, the package contains significance tests. The analysis of copy number profiles includes a likelihood ratio statistic and reference distribution, as well as an option to produce various plots that summarize the results.

Availability: Bioconductor (http://bioconductor.org/packages/release/bioc/html/Clonality.html) and http://www.mskcc.org/mskcc/html/13287.cfm.

Contact: ostrovni@mskcc.org

Received on December 3, 2010; revised on April 4, 2011; accepted on April 19, 2011

1 INTRODUCTION

Multiple cancerous lesions in the same patient can be either clonal or independent. Clonal tumors, for example, a primary tumor and its metastasis, originate from the same 'clonal' cell. A second primary tumor develops completely independently from the original primary. This distinction can have important clinical implications. For example, a lung nodule in a survivor with a previous head/neck cancer is potentially curable by surgery, if it is a primary lung cancer, but not if it is a metastasis from the original primary. In such cases, the genomic profiles of tumors can provide insight into their relationship, since clonal tumors necessarily possess at least some identical somatic mutations. Researchers have utilized two types of genetic data for evaluating clonality. The most common approach is to evaluate heterozygous markers at a set of candidate genetic loci for loss of heterozygosity (LOH). This strategy has promise as a clinical tool, because of its simplicity, and because it can be accomplished easily with minimal quantities of DNA. More recently, investigators have compared genomewide copy number arrays (CNAs) (e.g. CGH arrays) to search for identical copy number gains or losses. This is a more elaborate and expensive strategy, less likely to see clinical application in the near future, but potentially valuable for its greatly increased resolution compared with the use of candidate markers. We have previously published statistical methodology for formal statistical testing in both settings. For LOH data, we have developed a concordant mutations (CM) test (Begg *et al.*, 2007) and a likelihood ratio (LR) test (Ostrovnaya *et al.*, 2008). For CGH data, we proposed a LR statistic and an algorithm for calculating its reference distribution under the 'independence' hypothesis (Ostrovnaya *et al.*, 2010a). We have created the R/Bioconductor (Gentleman *et al.*, 2004; R Development Core Team, 2010) package called '*Clonality*' that implements these methods. This enables users to summarize genetic profiles of tumors and to perform tests for clonal relatedness.

*To whom correspondence should be addressed.

2 AVAILABLE FUNCTIONALITY

We describe the seven principal functions that are included in the package and discuss how to interpret the output. All tests are applied to pairs of tumors from the same patient. If multiple tumors per patient are provided, all pairwise comparisons are performed.

2.1 Testing clonality using LOH data

The function 'LOHclonality()' combines both tests we have developed for LOH data at the candidate loci (usually 10–30 markers). The main input is a matrix of LOH calls, where each marker has to be represented by one of three user-specified symbols denoting no LOH, LOH in allele 1 or LOH in allele 2. Markers that are not informative (e.g. homozygous) in a particular tumor should be given an NA instead of a call. The methodology assumes that the markers are independent (ideally from different chromosome arms) and that uninformative markers are missing at random. The user can choose to perform the CM test, the LR test or both. The test statistic for the CM test is the number of concordant losses (i.e. LOH on the same allele). This is printed in the output along with counts of markers with an LOH in one tumor only, with no LOH in both tumors and so on. The *P*-value is produced using a theoretical reference distribution that relies on assumptions that alleles 1 and 2 are equally likely to be lost, and that each marker has the same probability of an LOH. This test is valid when the deviations from these assumptions are not substantial (Begg *et al.*, 2007). The LR test does not make these assumptions, but it requires knowledge of the LOH frequencies for each marker. If they are not provided, they can be estimated from the original dataset or another group of reference patients with the same disease. The reference distribution for the LR test is obtained by simulating tumors with given LOH frequencies. Unlike the CM test, concordant LOH at an infrequently

occurring marker can provide stronger evidence for clonality than at a commonly occurring marker. The LR test, while free of the assumptions of the CM test, is preferred only when the true LOH frequencies in the specific patient population of interest are known or can be estimated with high confidence.

2.2 Testing clonality using CGH data

For the analysis of CNAs, the software requires the package DNAcopy (Olshen *et al.*, 2004; Venkatraman *et al.*, 2007). The input into the main function, 'clonality.analysis()', has to be a CNA object used in DNAcopy; it combines chromosome, genomic locations and each sample's log-ratios in columns. The chromosomes should be split into arms, since it increases the number of independent genomic units for the analysis. This can be done using function 'splitChromosomes()'. The clonality analysis consists of several steps:

- The data are segmented with the one step Circular Binary Segmentation (CBS) algorithm that selects at most one prominent copy number change per chromosome arm (another segmentation algorithm can potentially be used with a user-defined segmentation function instead of the built-in 'oneseg', as long as it detects at most one copy number change per chromosome arm. Details are given in the vignette).
- The chromosome arms are classified as gain/loss/normal based on the central or most outstanding segment. Number of Median Absolute Deviations (MADs) of the residuals, selected by the user, defines the gain/loss threshold. Users should investigate the plots of the segmentation to make sure that the large visible gains and losses pass the MAD criteria in most samples and adjust number of MADs ('nmad') parameter accordingly. Frequencies of gain or loss for each chromosome arm are needed for the calculation of the LRs. The frequencies are evaluated based on the original dataset, but there is also an option to estimate them based on a dataset of other patients with the same disease.
- A likelihood ratio (LR1) is calculated based on the concordance of the gain/loss/normal profiles. The final statistic, LR2, is the product of LR1 and LRs from each chromosome arm in which concordant gains or losses are observed in the two tumors. These reflect the odds that the chromosome arm specific mutation is clonal. LR2 quantifies the odds that the two tumors are clonal.
- The reference distribution (option in 'clonality.analysis') is calculated by comparing pairs of tumors from different patients, which are necessarily independent. This distribution is used to calculate *P*-values for the original tumor pairs. It is a preferred option to the hierarchial clustering method often used in practice, since it effectively uses the known paired structure of the data and for other technical reasons (Ostrovnaya *et al.*, 2010b). Note that calculating the reference distribution might take a substantial amount of time (hours) depending on the number of patients and the resolution of the assay. However, we recommend choosing this option unless the clonality signal is very strong and the noise level is small.

It is important to make sure that small gains and losses that could potentially be germline copy number variants (CNVs) are excluded from the arrays *prior to clonality analysis*. These appear as pronounced changes that are exactly the same in both tumors, and are thus mistaken to be strong evidence toward clonality by our algorithm. In order to exclude CNVs, the arrays should be either compared to their matching normal arrays, or to the database of genomic variants (http://projects.tcag.ca/variation/), or some other screening method can be applied (e.g. Ostrovnaya *et al.*, 2010c).

Since only one genomic change per chromosome arm is allowed, we suggest limiting the resolution of the array to roughly a total of 5000–15 000 markers. If the array has more markers, then mutually exclusive blocks of several adjacent markers can be averaged using the function 'ave.adj.probes()'. The number of markers averaged should be calculated based on the original resolution of the array. Averaging also helps reduce noise in the arrays, it removes waves in log-ratios that can result in false genomic changes, and it simplifies the copy number patterns. If the user is not confident that all CNVs are removed, then a lower resolution (e.g. 2000 markers) is recommended.

The help file for 'clonality.analysis()' contains a real data example—analysis of pairs of breast cancer lesions published by (Hwang *et al.,* 2004) and available from the authors' web site.

2.3 Visualization of CGH data

Two functions, 'chromosomePlots()' and 'genomewidePlots()', provide per chromosome and genomewide plots of the data and one step segmentation for pairs of tumors. These plots show where the clonality signal is coming from; they illustrate the concordance between general mutational patterns and the match between specific gains and losses. The overlaying histograms of the LRs and the reference distribution under the hypothesis of independence are produced by 'histogramPlot()'. The extent of overlap between the two histograms can help define the cut-offs for diagnosing patients: the upper right tail of the reference distribution might be interpreted as an equivocal area, while the LRs above or below the equivocal area receive unambiguous diagnoses.

ACKNOWLEDGEMENT

We are grateful to Kevin Eng for programming work conducted early in the development of this project.

Funding: National Cancer Institute award (CA124504).

Conflict of Interest: none declared.

REFERENCES

Begg,C.B. *et al.* (2007) Statistical tests for clonality, **63**, 522–530.

Gentleman,R. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics 2004. *Genome Biol.*, **5**, R80.

Hwang,E.S. *et al.* (2004) Clonality of lobular carcinoma in situ and synchronous invasive lobular cancer. *Cancer*, **100**, 2562–2572.

Olshen,A.B. *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Ostrovnaya,I. *et al.* (2008) Comparison of properties of tests for assessing tumor clonality. *Biometrics*, **68**, 1018–1022.

Ostrovnaya,I. *et al.* (2010a) A metastasis or a second independent cancer? Evaluating the clonal origin of tumors using array copy number data. *Stat. Med.*, **29**, 1608–1621.

Ostrovnaya,I. and Begg,C.B. (2010b) Testing clonal relatedness of tumors using array comparative genomic hybridization: a statistical challenge. *Clin. Cancer Res.*, **16**, 1358–1367.

Ostrovnaya,I. *et al.* (2010c) A classification model for distinguishing copy number variants from cancer-related alterations. *BMC Bioinformatics*, **11**, 297.

R Development Core Team (2010) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Venkatraman,E.S. and Olshen,A.B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, **23**, 657–663.