BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 23 2006, pages 2950–2951
doi:10.1093/bioinformatics/btl433

Gene expression

COPA—cancer outlier profile analysis

James W. MacDonald[1,*] and Debashis Ghosh[2]

[1]University of Michigan Cancer Center, Ann Arbor, MI, USA and [2]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

Received on April 3, 2006; revised and accepted on August 3, 2006
Advance Access publication August 7, 2006
Associate Editor: John Quackenbush

ABSTRACT

Summary: Chromosomal translocations are common in cancer, and in some cases may be causal in the progression of the disease. Using microarrays, in which the expression of thousands of genes are simultaneously measured, could potentially allow one to detect recurrent translocations for a particular cancer type. Standard statistical tests, such as the *t*-test are not suited for detecting these translocations, but a simple test based on robust centering and scaling of the data to help detect outlier samples, followed by a search for pairs of samples with mutually exclusive outliers, may be used to find genes involved in recurrent translocations. We have implemented this method, termed Cancer Outlier Profile Analysis (COPA) in an R package (that we call the copa package), and show its applicability on a publicly available dataset.

Availability: http://www.bioconductor.org

Contact: jmacdon@med.umich.edu

Chromosomal aberrations are a hallmark of cancer, and recurrent translocations have been shown to be causal in the progression of the disease, particularly in haematological disorders (Rowley, 2001). Functional translocations generally fall into one of two categories; those that result in a fusion protein that may have a novel effect on the cell and those in which the promoter region from one gene is translocated to the intact coding region of an oncogene, thus up-regulating the expression of the oncogene. Although there are fewer examples of recurrent translocations in common carcinomas, it is unclear whether this is because recurrent translocations are less common in these cancer types, or whether they have yet to be discovered. Given the importance of certain translocations in haematological disorders, it is worthwhile to determine whether recurrent translocations occur in common carcinomas. Microarray data are ideally suited to screen for candidate translocations, owing to the simultaneous measurement of thousands of genes.

Tomlins *et al*. (2005) proposed a method they call Cancer Outlier Profile Analysis (COPA) for detecting translocations of the second type using microarray data. They have implemented this procedure as part of the Oncomine database (www.oncomine.org). COPA on Oncomine is simple to use. However, the web-based implementation is not extensible, so one is only able to do what analyses are available. For instance, it is not possible at this time to assess significance of a particular outlier. One is also limited to analyzing those data that have been uploaded to the Oncomine database.

We have implemented the COPA procedure in an R package that is part of the BioConductor project (Gentleman *et al*., 2004). The copa package is Open Source, written in R, and is easily extensible. In addition, we have implemented a permutation-based method to assess significance using *p*-values, as well as false discovery rate (FDR) (Storey, 2002).

The idea behind COPA is very simple; it is well known that genetic translocations occur in cancer cells, and that these translocations can result in the up-regulation of oncogenes that may affect the progression of the cancer. This happens when the 5′ activation domain of a constitutively expressed (or up-regulated) gene is fused to the 3′ portion of a given oncogene, thus increasing the expression of the oncogene. This translocation can happen between the activating gene and multiple oncogenes, as was shown by Tomlins *et al*. (2005), as well as others (Fonseca *et al*., 2004).

Since a given translocation is only likely to occur once per sample, if there were multiple partners for a given activating gene, we would expect to see certain cancer samples with a high expression of say, gene A, whereas other cancer samples might have high expression of gene B, but these samples would be mutually exclusive. In addition, we would expect that the normal samples would not have high expression for either gene A or B. We can use this idea to both pre-filter genes as well as to find interesting genes that may be involved in translocations.

Common methods for detecting differences between tumor and normal samples will not work for finding these genes (e.g. *t*-tests); we need to find those genes where only a subset of the samples have high expression. To do this, we center and scale the data (on a row-wise basis) using the median and median average difference (MAD). Here we are assuming that the columns of our data matrix are samples and rows are genes. We can then select a common value (default is 5) as a cutoff for 'outlier' status and apply this to all genes. We then simply look for pairs of genes that have a large number of mutually exclusive outlier (cancer) samples, but few or no normal outliers. The candidate gene pairs will be ranked based on the sum of outlier samples for each pair. Note that this is a slightly different ranking system than used on Oncomine. For example, in Table 1 of the Tomlins paper the genes are ranked based on the 75th, 90th or 95th percentile of the centered and scaled expression values.

Since there may be several gene pairs with the same number of outliers we need to add an additional criterion to rank the ties. Because we are looking at pairs rather than individual genes, we take the difference between the 75th percentile of the tumor and normal samples, and then compute the sum of these differences for each gene pair. This value quantifies how different the outlier pairs are from their corresponding normals.

Exploration of the data can be facilitated using graphical summaries. As an example, we use the prostate cancer data of

*To whom correspondence should be addressed.

Table 1. Gene pairs with a given number of outliers

Outliers	17	15	14	13	12	11	10	9	8
Gene Pairs	1	6	23	1	12	43	65	142	126

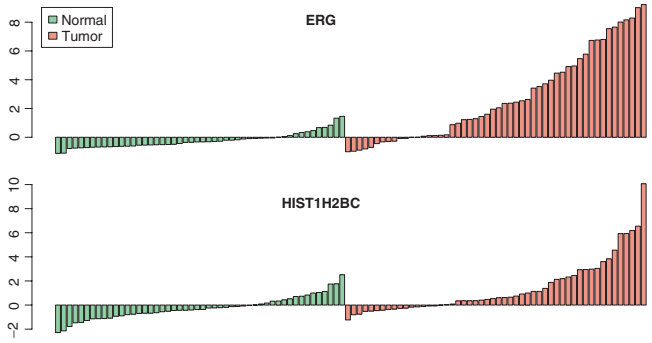

Fig. 1. Ordered barplot of the top gene pair from the Singh prostate cancer data. Salmon colored bars correspond to the tumor samples and green bars correspond to the normal adjacent samples. The ranking for *ERG* is in agreement with results from Tomlins *et al*. (2005). The other gene, *HIST1H2BC* encodes a histone protein.

Singh *et al*. (2002). This dataset consists of 52 prostate tumor samples and 50 normal adjacent samples, that were run on Affymetrix HG-U95av2 chips. We converted raw data to expression values using a robust multi-array average (RMA) (Irizarry *et al*., 2003) and then ran the copa procedure. We plotted the top ranked gene pair using both ordered bar plots (Fig. 1), and a scatter plot (Fig. 2). The top gene pair is *ERG*, which was also the top ranked gene in the Tomlins paper, and *HIST1H2BC*, which encodes a histone protein.

The number of gene pairs with a given number of outliers from Singh *et al*. (2002) is given in Table 1. There was one gene pair that had 17 mutually exclusive outliers, six gene pairs that had 15 mutually exclusive outliers, and 23 gene pairs that had 14 mutually exclusive outliers. We might want to test the significance of observing 30 gene pairs that had 14 or more mutually exclusive outliers in the shuffled datasets. The copa package has a function that will repeatedly permute the data and then count the number of gene pairs that had 14 or more mutually exclusive outliers. It will then calculate a *p*-value based on the number of permuted datasets that had 30 or more such gene pairs, divided by the number of

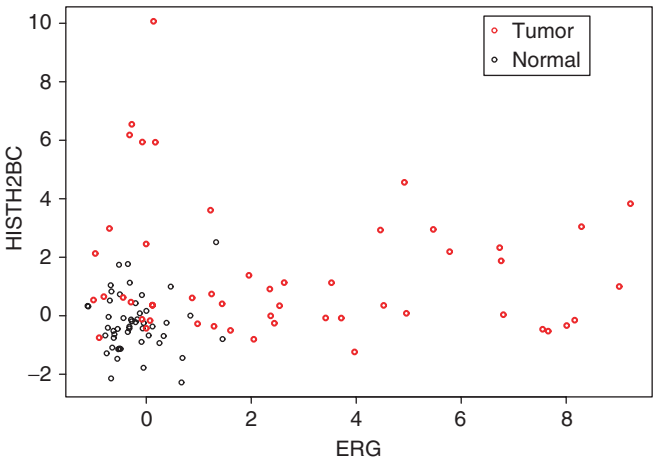

Fig. 2. Scatterplot of the top gene pair from the Singh prostate cancer dataset. Tumor samples are colored red and normal adjacent samples are black. This representation may be better for visualizing the number of outlier samples for each gene.

permutations. We can also calculate the FDR, which estimates the percentage of the 30 observed gene pairs that are likely to be false positives. This is calculated by dividing the mean number of gene pairs from the permutated data by the observed number of gene pairs (30 in our case).

We calculated a permuted *p*-value of <0.001, and an FDR of 0.04%, which implies that our observed result was not likely to arise by chance alone.

REFERENCES

Fonseca,R *et al*. (2004) Genetics and cytogenetics of multiple myeloma: a workshop report. *Cancer Res*., **64**, 1546–1558.

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

Irizarry,R. *et al*. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Rowley,J.D. (2001) Chromosome translocations: dangerous liaisons revisited. *Nat. Rev. Cancer*, **1**, 245–250.

Singh,D. *et al*. (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.

Storey,J. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. B*, **64**, 479–498.

Tomlins,S.A. *et al*. (2005) Recurrent fusion of *TMPRSS2* and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644–648.