BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 21 2005, pages 3990–3992
doi:10.1093/bioinformatics/bti667

Gene expression

A method for detection of differential gene expression in the presence of inter-individual variability in response

David M. Rocke[1,*], Zelanna Goldberg[2], Chad Schweitert[2] and Alison Santana[2]
[1]Division of Biostatistics and [2]Department of Radiation Oncology, University of California, Davis, CA, USA

Received on August 19, 2005; revised on September 5, 2005; accepted on September 6, 2005
Advance Access publication September 13, 2005

ABSTRACT

Motivation: Many stimuli to biological systems result in transcriptional responses that vary across the individual organism either in type or in timing. This creates substantial difficulties in detecting these responses. This is especially the case when the data for any one individual are limited and when the number of genes, probes or probe sets is large.

Results: We have developed a procedure that allows for sensitive detection of transcriptional responses that differ between individuals in type or in timing. This consists of four steps: one is to identify a group of genes, probes or probe sets that detect genes that belong to a molecular class or to a common pathway. The second is to conduct a statistical test of the hypothesis that the gene is differentially expressed for each individual and for each gene in the set. The third is to examine the collection of these statistics to see if there is a detectable signal in the aggregate of them. The final step is to assess the significance of this by resampling to avoid correlational bias.

Availability: Software in the form of *R* code to perform the required test is available from the first author or from his website http://www.idav.ucdavis.edu/~dmrocke/software; however the procedures are also easily performed using any standard statistical software.

Contact: dmrocke@ucdavis.edu

Supplementary information: Datasets used in this paper may be obtained from the first author's website http://www.idav.ucdavis.edu/~dmrocke/software

1 INTRODUCTION

One of the major problems in analysis of highly parallel biological assay data, such as gene expression arrays, is detecting differential expression in the face of the huge multiplicity of tests. This is especially the case when the number of arrays in each test is small, because P-values sufficiently small for significance require quite extreme t- or F-statistics.

For human studies especially, this can be exacerbated by individual variation in response, a significant issue in drug response, for example. In the present study, we use exposure to low-dose ionizing radiation as a model for this phenomenon, since inter-individual variability in response was one of the major hypotheses of the study from which the data used here were derived.

The method described here is designed for cases in which one of two possible types of inter-individual variability exists. One possibility is that a response exists in most or all individuals in a particular molecular class (e.g. MAP Kinases) but may be detected by different probes/probe sets in different individuals as a result of polymorphisms in the response transcript or because of physiological differences. A second possibility concerns pathways that may be initiated by the radiation exposure or other stimulus. If samples are taken at a fixed time, and if the timing of the cascade is different in different individuals, then the response may be exhibited in different molecules in the pathway for different individuals.

The common feature here is that the response is diffused across different probes/probe sets for different individuals, such that detection may be difficult.

2 METHODS

We begin with a description of the study that motivated the development of this method. The biological activity of low doses of ionizing radiation (LDIR) is fundamentally undefined in the human when exposure is under normal physiologic conditions of intact 3D architecture, vasculature and cell–cell contacts (between epithelial cells, and between epithelial and stromal cells). We therefore undertook these studies to evaluate transcriptomic responses to a single exposure of LDIR in the normal skin of men undergoing therapeutic radiation for the treatment of their prostate cancers.

Patients were treated in a standard fashion for their localized prostate cancer and the areas of their abdominal wall skin which would receive 1, 10 or 100 cGy (tumor dose was the standard of 200 cGy; the skin doses were areas of scattered or non-targeted radiation exposure) were marked at the time of the patient's first radiation treatment. Prior to any radiation therapy, patients had a baseline control biopsy (zero dose). After 3 h of the completion of the first radiation treatment, the skin biopsies were obtained (local anesthetic, sterile conditions) and immediately placed into RNAlater (Ambion, catalog #7021) for preservation of the material and inactivation of the tissue RNAses. This was performed in a HIPPA compliant fashion within an IRB approved clinical protocol. There were eight patients on this arm of the study, which forms the basis for the data analysis described.

The biopsied tissue was then mechanically disrupted on a FastPrep Bead Beater (Qbiogene #6001-120) in lysis buffer (1 ml guanidine thiocyanate and the preloaded Lysing Matrix D) and the RNA extracted using a modified Trireagent procedure based on the methods of Chomczynski and Sacchi (1987). RNA was quantified, qualitated, amplified and primer labeled employing standard protocols for use on the Affymetrix GeneChip platform.

The data for this analysis consist of 32 Affymetrix HGU133 Plus 2.0 GeneChips, four from each of eight patients, at doses of 1, 10 and 100 cGy, as well as a pre-exposure control at 0 cGy. The HGU133 Plus 2.0 contains 1 354 896 probes divided into 54 675 probe sets, which have been summarized using the GLA expression index (Zhou and Rocke, 2005), though the results are very similar if RMA is used instead (Irizarry *et al*., 2003). This array platform demonstrates in stark form the advantages and disadvantages of whole genome assays, as this is certainly one of the most highly parallel transcriptomics assays available. The positive feature of this

[*]To whom correspondence should be addressed.

Table 1. Results of a standard analysis and the ToTS method for selected gene groups and pathways

Group or pathway	Probe sets	Significant probe sets	*P*-values for dose by *t*-test and Wilcoxon test	Empirical *P*-values for dose by *t*-test and Wilcoxon test	*P*-values for modified log dose by *t*-test and Wilcoxon test	Empirical *P*-values for modified log dose by *t*-test and Wilcoxon test
Zinc-finger protein	799	0	2.68×10^{-5} 7.78×10^{-5}	0.015 0.000	4.17×10^{-6} 2.20×10^{-16}	0.000 0.000
MAPK	131	1	0.0020 0.0025	0.012 0.002	0.0143 0.0165	0.013 0.002
Akt/PI3 kinase	99	0	0.0014 0.0002	0.003 0.000	0.1333 0.0006	0.141 0.000
Stress apoptosis	144	0	0.9343 0.0739	0.944 0.040	0.0359 0.0029	0.047 0.007

For each gene group or pathway, we identified the probe sets that measure transcripts in that category, the number of which is given in column 2. Column 3 shows how many of these are individually significant at a 5% FDR rate for either regression on dose or on modified log dose. In each cell with *P*-values, the upper *P*-value is for the *t*-test of all the *t*-statistics and the lower *P*-value is for the Wilcoxon test of all the *t*-statistics. The empirical *P*-values are based on resampling with 1000 random sets probe sets containing the same number of probe sets as the indicated gene group or pathway.

is its comprehensiveness, including multiple exons from most of the known genes in the human genome. The disadvantage is that with such a large number of probe sets, there will be many apparently significant changes that occur strictly by chance, and its control requires a very high bar for declaring differential expression to be significant. For example, suppose we conduct a regression analysis of expression on dose for each patient separately. This results in $(54\,675) \times (8) = 437\,400$ *t*-tests of the significance of the slope, which is a test for dose–response relationship, each of which has 2df for error. By pure chance, we will easily get *t*-scores of >1000, even if the data are completely random and show no dose–response pattern for any patient for any transcript (the *t*-score required to be significant at a Bonferroni-corrected 5% level is 2956). Thus, it is not surprising that none of the 437 400 regressions is significant at the 5% level after adjusting for false discovery rate (Benjamini and Hochberg, 1995; Reiner *et al*., 2003). The same is true if one uses modified log dose (in which the positive exposures are coded as 0, 1 and 2, and the zero exposure is coded as -1, as if it were 0.1 cGy). In part this is because of the very small residual degrees of freedom (2), which makes the test rather insensitive.

It is, of course, much easier to detect dose–response patterns that are consistent across patients, but radiation response is an area in which substantial inter-individual variability may occur (though evidence one way or the other under *in vivo* conditions in humans is essentially absent). One possibility is to fit a statistical model that allows for a patient by dose interaction, specifically

$$y_{ijk} = \bar{\mu}_i + \alpha_{ij} + \bar{\beta}_i x + \beta_{ij} x_k + \varepsilon_{ijk},$$

where the index i refers to probe sets, the index j to patients, and the index k to doses. This is likely to be more sensitive than the individual patient regressions since the residual degrees of freedom are 14 instead of 2. In the 54 765 probe sets, and using the methods of Rocke (2004), we found no cases in which the overall dose–response coefficient $\bar{\beta}_i$ is significant at a 5% FDR level, and 11 where the FDR-adjusted significance level is better than 10%.

The patient effect was significant at the 5% FDR-adjusted level for 13 514 of the 54 675 probe sets, and significant at the 10% FDR-adjusted level for 18 605 of the probe sets. This shows how important the individual variation is in gene expression. The patient by dose interaction was significant at the 5% FDR-adjusted level for 22 probe sets and at the 10% FDR-adjusted level for 60 probe sets. We also find 6 of the 11 probe sets significant for the dose effect to be also significant for the patient by dose interaction.

These results are somewhat disappointing in that the radiation doses used should have caused important changes in expression. Clearly, much of the problem is caused by the low-error degrees of freedom for each individual regression for each patient (df $= 2$), and the very large number of such tests. This means that multiplicity adjustments require extremely small *P*-values to be significant.

2.1 Gene groups and pathways

The first tactic in approaching the problem of low degrees of freedom is to reduce the number of tests performed by using biological information. This may not always be necessary, and indeed it would be valuable if this were not needed as a broad screen would promote discovery of unknown molecular responses. However, it seems necessary in this case in order to have any real chance of detecting a response, given the very large *t*-score required to show significance with a residual df of 2, 8 patients and 54 675 tests.

We did this in two ways. We identified a number of gene groups that, from the literature, would be likely to respond to LDIR. We will use two for illustration: Zinc-finger proteins (representing 799 probe sets) and MAP Kinases (representing 131 probe sets). A complete analysis of these data may be found in Goldberg *et al*. (2005).

We also identified pathways that are expected to be involved in LDIR response, also from a literature review, of which we use for illustration the Akt PI3 kinase pathway (99 probe sets) and the stress apoptosis pathway (144 probe sets).

This reduction in the number of probe sets examined is helpful, in increasing the power of the analysis, but as shown in Table 1, few or no probe sets have an FDR-adjusted *P*-value <0.05. This motivated the development of a potentially more sensitive method of detecting differential expression in gene groups in pathways, when individual heterogeneity and small sample size may otherwise result in very low power.

2.2 The test of test–statistics (ToTS) method

Consider the data on the 799 zinc-finger probe sets and 8 patients. We have $799 \times 8 = 6392$ datasets of four points each. For each of dose or modified log dose as a regressor we can conduct a test of the hypothesis of no dose–response by the usual *t*-statistic, which here has 2 degrees of freedom. We hypothesize that there are effects in at least some members of this group for perhaps most individuals, but it may be reflected in different transcripts in different individuals. Thus, if there is this kind of diffuse, highly variable upregulation, we would expect that the *t*-scores would be biased in a positive direction. The histogram in Figure 1 shows that this is in fact the case: that there is an apparent positive bias in the *t*-scores. The upward shift in the *t*-scores is small, as might be expected from the hypothesized diffuse, variable upregulation, but it may be of interest to verify that it is not an accident or an artifact. We propose testing for this positive bias in two ways: by computing the one-sample *t*-statistic for the hypothesis that the mean value of the collection of test statistics is zero and the one-sample Wilcoxon

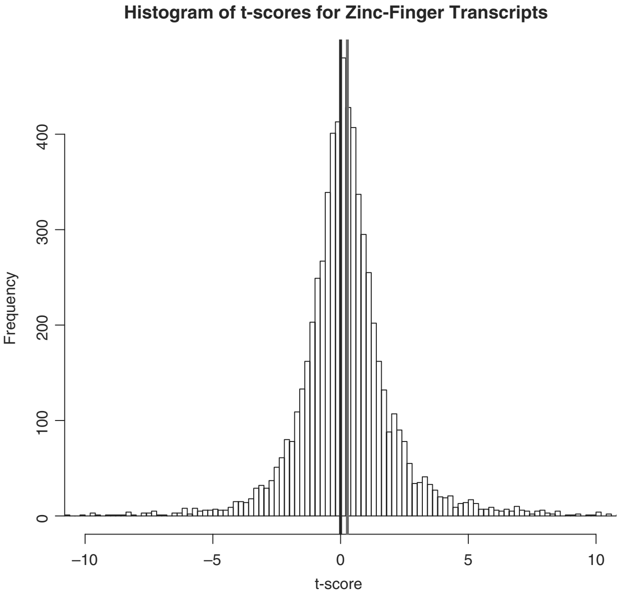

Fig. 1. Histogram of 6392 *t*-scores from the 799 zinc-finger probe sets and eight patients for regression of response on modified log dose. The left-hand (black) vertical line is at 0, and the right-hand (red) vertical line at the mean of the *t*-scores, showing an apparent slight positive bias.

statistic for the hypothesis that the median value of the collection of test statistics is zero. It would also be possible to filter the probes or probe sets first, e.g. by choosing only highly expressed ones, though we favor using all the data.

Of course, downregulation would cause a negative bias, so two-sided tests are needed. In other cases, the individual test statistics may be *F*-statistics instead of *t*-statistics, in which case we would analyze the logs of the test statistics instead.

2.3 Empirical *P*-values

A possible objection to this procedure is that all 6392 test statistics come from the same 32 arrays, so that the test statistics for a given patient, all of which come from the same 4 arrays, could be correlated, in spite of exercising care in normalization. As this is not unlikely, we propose not to use the standard *P*-values for the *t*-test or Wilcoxon test, but rather to use empirical *P*-values from a resampling-based method. For the zinc-finger case, we repeatedly sample 799 random probe sets and perform the test on the random set of probe sets, and then count the fraction of the time that the test statistics from the random probe sets exceeded that for the actual zinc-finger probe sets. In this case, we used 1000 trials, so that we can obtain sufficient accuracy.

2.4 Comparison with gene set enrichment analysis

Another possible approach to testing sets of genes is enrichment analysis. If one has a set of genes which are, for example, significant at the 5% level, and a set of genes corresponding to a biological category, one can test for enrichment of the gene set among the significant genes. The most highly developed version of this idea is gene set enrichment analysis (GSEA) (www.broad.mit.edu/gsea/) in which this test is in effect conducted at all significance cutoffs simultaneously, which equivalently is a Kolmogorov–Smirnov test of the difference in the distribution of the test statistics of genes in the biological gene set versus those not in the gene set. In this form, the ToTS method can be seen as a kind of generalization of the GSEA method, but one which uses the actual values of the test statistics, not just their ranks.

However, for the application of assessing diffuse expression in the presence of individual variation in response, GSEA in its present form is not suitable, since it requires a ranking of genes by some test statistic, and we have in this case eight test statistics for each gene, and there are multiple ways in which the genes could then be ranked.

3 RESULTS

The results of this analysis for the two gene groups and two pathways are shown in Table 1. The standard *P*-value calculations from the *t*-test and Wilcoxon test (obtained using the *R* functions t.test and wilcox.test) correspond quite well to the empirical *P*-values for modified log dose, for which the regressors are equally spaced. For the regression on dose, which is unbalanced at 0, 1, 10 and 100, with the largest value having high influence, the correspondence is poor in some cases, suggesting that in general the resampling-based method is a necessary precaution. Since the *t*-statistics on two degrees of freedom have long tails, the Wilcoxon test in this case often is more sensitive, and is likely to be preferred in such cases.

In several cases, the evidence for a diffuse upregulation is quite convincing using the ToTS method, whereas using standard methods there appears to be no effect of the treatment on gene expression. Thus this method seems to be a useful adjunct to more standard approaches when individual variation in response is likely, and when the data on each individual are limited.

ACKNOWLEDGEMENTS

The authors are grateful for useful comments from three referees that helped improve this paper from its initial draft. The research reported in this paper was supported by grants from National Institute of Environmental Health Sciences, National Institutes of Health (P43-ES04699), the Department of Energy (DE-FG02-01ER63237), the National Institute of Arthritis and Musculoskeletal and Skin Diseases (R01-AR044518-06) and the UC Davis Health System.

Conflict of Interest: none declared.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Chomczynski,P. and Sacchi,N. (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction. *Anal Biochem*., **162**, 156–159.

Goldberg *et al*. (2005) In vivo transcriptional response in humans to controlled, low--dose ionizing radiation exposure. In press

Irizarry,R.A. *et al*. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res. Methods Online*, **31**, e15.

Reiner,A. *et al*. (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368–375.

Rocke,D.M. (2004) Design and analysis of experiments with high throughput biological assay data. *Sem. Cell Dev. Biol*., **15**, 708–713.

Zhou,L. and Rocke,D.M. (2005) An expression index for Affymetrix GeneChips based on the generalized logarithm. *Bioinformatics*, **21**, 3983–3989.