BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 4 2014, pages 574–575
doi:10.1093/bioinformatics/btt720

Gene expression Advance Access publication December 13, 2013

ProbeSelect: selecting differentially expressed probes in transcriptional profile data

Raghavendra Hosur[1], Suzanne Szak[2], Alice Thai[2], Norm Allaire[2] and Jadwiga Bienkowska[1,*]

[1]Patient Stratification group and [2]Genetics and Genomics group, Biogen Idec, Cambridge, MA, USA

Associate Editor: Ziv Bar-Joseph

ABSTRACT

Summary: Transcriptional profiling still remains one of the most popular techniques for identifying relevant biomarkers in patient samples. However, heterogeneity in the population leads to poor statistical evidence for selection of most relevant biomarkers to pursue. In particular, human transcriptional differences can be subtle, making it difficult to tease out real differentially expressed biomarkers from the variability inherent in the population. To address this issue, we propose a simple statistical technique that identifies differentially expressed probes in heterogeneous populations as compared with controls.

Availability and implementation: The algorithm has been implemented in Java and available at www.sourceforge.net/projects/probeselect.

Contact: jbienkowska@gmail.com or jadwiga@csail.mit.edu

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

Received on June 27, 2013; revised on September 4, 2013; accepted on December 9, 2013

1 INTRODUCTION

High-throughput gene expression, protein or metabolite profiling of patient samples has become a standard tool of biomarker discovery for disease subtypes and progression. Irrespective of the particular technology platform, be it Affymetrix Chip, miRNA profiles, mass spec proteomics, metabolomics or so on, the challenge of focusing analyses on the most informative markers is similar. For heterogeneous diseases, the first challenging step analyzing multidimensional data is identification of markers that distinguish disease and control groups. Because of heterogeneity of the disease, grouping all disease samples for comparison against healthy controls may fail in identifying markers that could be specific to only a small percentage of disease samples. Once such candidate markers are identified, unsupervised clustering methods are applied to uncover the disease substructure in the biomarker space.

Consider a hypothetical case of a marker that is 2-fold different (1 on the log2 scale) in just 10% of disease samples with zero variability and with 90% of disease samples the same as control group. The standard deviation in disease groups for this marker will be 0.1 and the log2-fold difference between cases and controls only 1.1. Assuming zero variability in controls, such a marker would only have a z score of 1, thus not significant by either fold or P-value. Nevertheless, this marker clearly distinguishes a disease subtype from control and the rest of the population.

First approach to address this challenge has been proposed in the seminal paper by Perou *et al*. (2000) where breast cancer subtypes were identified using whole-genome expression profiling. In this article, probes with variability >10% in the cohort were selected for further unsupervised clustering, and this clustering successfully recapitulated major breast cancer molecular subtypes. This simple method has been applied in numerous subsequent publications and a variety of diseases and sample types. Another method frequently used is comparison of disease and control samples and identification of probes significantly different between disease and control cohorts, e.g. as implemented in LIMMA BioConductor package (Smyth, 2004). Often decisions determining significance thresholds and their stringency of probe selection are arbitrary.

In this article, we propose a simple statistical formalism for selection of the probes that distinguish heterogeneous disease samples from healthy controls. The standard clustering approach considers all probes with a high coefficient of variation (CV > 10% typically) and subsequently clusters both controls and cases to discover a structure that may separate controls from cases. In comparison, our method considers only the probes that are significantly different from healthy controls in a subgroup of patients. We apply this method to the whole-blood profiles of Secondary Progressive Multiple Sclerosis (SPMS) patients and uncover distinct subgroups in the patient population that, when treated as a uniform group, are indistinguishable from the control-group profiles. That is, a standard LIMMA approach to identification of genes differentially expressed between SPMS and control groups does not identify any significant differences. In addition, the method used by Perou *et al*. (2000) (>10% CV) selects 6013 probes, 328 of which overlap with our method's selection. However, a standard principal component analysis (PCA) or hierarchical clustering in the space of these probes does not separate out the controls from the cases (Supplementary Fig. S1).

2 METHODS

Our technique overcomes the limitation of poor group statistics (in the case of 'disease' samples) by treating each sample individually. Suppose we have 'N' cases and 'r' controls, each sample having 's' probes. We want to identify probes that are differentially expressed in cases, or subgroups thereof, versus controls. A challenge resulting from the inherent heterogeneity in the cases is that none of the probes show significant fold

[*]To whom correspondence should be addressed.

changes when comparing the cases and controls as groups. Our method works by first computing a z score for each probe in each sample:

$$z_{ik} = \frac{e_{ik} - \mu_{control}}{\sigma_{control}}$$

where 'i' indexes the probe ($i = 1 \ldots s$), 'k' indexes the sample ($k = 1 \ldots N$), 'e_{ik}' is the expression value of the i^{th} probe (typically log transformed units) in the k^{th} sample, 'μ' and 'σ' are the mean expression and standard deviation for the i^{th} probe in the control samples, respectively. Individual probes are selected if they pass a z-score cutoff, z_{cutoff}. The number of samples in which a probe passes this selection criterion is denoted as m_i. The z_{cutoff} can be set to a low value, e.g 1.5, to include large number of probes to begin with.

Once we have computed m_i for each probe, we use a binomial model to filter the probes. If 'q' is the selection probability for a sample $q = P(|z| \geq z_{cutoff})$, then the probability of selecting m_i samples under the binomial model is as follows:

$$p(q, m_i, N) = {}^{N}C_{m_i} q^{m_i} (1 - q)^{N - m_i}$$

The individual z-scores computed previously that pass the cutoffs are all treated equally (in q). We can then explicitly compute a P-value associated with selecting m_i samples under this model:

$$p(n \geq m_i | q, N) = \sum_{n = m_i}^{N} {}^{N}C_{m_i} q^{m_i} (1 - q)^{N - m_i}$$

Because there are 's' probes, we use a Bonferroni or False Discovery Rate correction to control for false-positives. The probes that pass a cutoff (0.05) after correction are the probes of interest. The algorithm essentially is the same for selection of only upregulated or downregulated probes. Only the calculation of q and m_i are changed to reflect the directionality. In this case, 'q' represents only one tail of the normal distribution, and only those samples that satisfy the direction as well as the cutoff are included in m_i.

3 RESULTS

We applied this method to select differentially expressed probes in 190 SPMS samples against 30 age- and gender-matched healthy controls. A standard differential analysis using the LIMMA package does not identify any probes as significant (at a fold-change cutoff of 1.5). On the other hand, using our approach, we select 1752 probes from the same dataset. Using simple PCA of the entire transcriptome we cannot discover any substructure in this patient population and there is no clear distinction from controls (Fig. 1A). However, when we apply PCA to our selected 1752 probes, two major subgroups can be easily identified, and one of these subgroups is clearly distinct from controls (Fig. 1B).

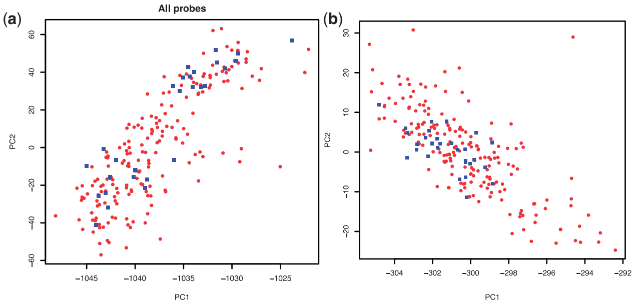

Fig. 1. Patients (red) and controls (blue) projected onto the top two principal components (PC). (**a**) There is no clear structure in the PC space of all probes. (**b**) Selected probes projected onto the top two PCs reveals the underlying structure (two subgroups in the cases). These subgroups do not represent gender-based stratification

ACKNOWLEDGEMENTS

We thank Accelerated Cure Project and patients for the blood samples used for profiling.

Conflict of Interest: none declared.

REFERENCES

Perou,C.M. *et al.* (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.

Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article 3.