Published in final edited form as:
Nat Methods. 2007 November ; 4(11): 911–913. doi:10.1038/nmeth1102.

A Gene Expression Barcode for Microarray Data

Michael J. Zilliox[1,3] and **Rafael A. Irizarry**[2,*]

[1]W. Harry Feinstone Department of Molecular Microbiology and Immunology, Baltimore, MD 21205

[2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205

Abstract

The ability to measure genome-wide expression holds great promise for characterizing cells and distinguishing diseased from normal tissues. Thus far, microarray technology has only been useful for measuring relative expression between two or more samples, which has handicapped its ability to classify tissue types. This paper presents the first method that can successfully predict tissue type based on data from a single hybridization. A preliminary web-tool is available at http://rafalab.jhsph.edu/barcode/

The high throughput analysis of cells and tissues is revolutionizing biological research. The ability of microarrays to measure thousands of RNA transcripts at one time allows for the characterization of cells and tissues in greater depth than was previously possible, but has not yet led to big advances in diagnosis or treatment. The main reason for this is that feature characteristics, such as probe sequence, can cloud the relationship between observed intensity and actual expression. Although this *probe effect* is large, it is also very consistent across different hybridizations, which implies that relative measures of expression are substantially more useful than absolute ones[1, 2]. To understand this, consider that when comparing intensities from different hybridizations for the same gene, the probe effect is very similar and cancels out. On the other hand, when comparing intensities for two genes from the same hybridization, the different probe effects can alter the observed differences. For this reason the overwhelming majority of results based on microarray data rely on measures of relative expression: genes are reported to be differentially expressed rather than expressed or unexpressed.

Approaches for thresholding noisy data have been successfully used in many applications including microarray studies[3, 4]. We used this as motivation to develop the first method that can accurately demarcate expressed from unexpressed genes and therefore defines a unique *gene expression barcode* for each tissue type. To do this we took advantage of the vast amount of publicly available datasets. These data were also used to assess the algorithm. With clinical data, we find near perfect predictability of normal from diseased tissue for three cancer studies and one Alzheimer's disease study. The barcode method also discovers new tumor subsets in previously published breast cancer studies that can be used for the prognosis of tumor recurrence and survival time.

[*]Corresponding author: Dr. Rafael A. Irizarry, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Baltimore, MD 21205. Telephone (410) 614-5157. ririzarr@jhsph.edu.
[3]Current address: Emory Vaccine Center and Department of Microbiology and Immunology, Emory University School of Medicine, 1510 Clifton Road, Atlanta, GA 30322.

For any given gene and microarray platform, we wanted to know what intensity relates to no expression. A simple way to determine this intensity would be to hybridize tissues for which we know the gene is not expressed and look at the distribution of the observed intensities. If a new sample is provided and we want to know if this gene is expressed, we would simply compare the observed intensity to the previously formed distribution. For a single lab, creating this training dataset is logistically impossible for two reasons: 1) we do not know what genes are expressed in which tissues and 2) it would require various hybridizations for each gene. Fortunately, a preliminary version of such a dataset already exists for some platforms/organisms. Raw data were obtained for more than a hundred tissues from the public repositories and pre-processed with the same algorithm[1, 5–7]. Now, for each gene the intensity distribution is determined. Because it is expected that any given gene will only be expressed in some tissues, multiple modes should be observed. It is assumed that the lowest intensity mode is due to lack of expression. Genes that are expected to be expressed are coded with ones and the unexpressed genes are coded with zeros. This information is referred to as the *gene expression barcode* (see the Supplemental Methods for more details on the statistical algorithm). Barcodes were created for 118 human tissues and 44 mouse tissues (Supplemental Tables 1 and 2). A dendrogram and heatmap displaying the human and mouse barcodes and related summary statistics can be seen in Supplemental Figure 1.

We compared the barcode to the present/absent/marginal calls from the Affymetrix Microarray Suite 5.0 (MAS 5.0). With MAS 5.0, only 10% of the 22215 genes represented in the human array achieve the same call in all samples within the same tissue. This number increases to 48% using the barcode approach (Supplemental Figures 2A–B). Similar results were obtained with mouse data (Supplemental Figures 2CD). To assess sensitivity we used results from an extensive study that reported proteins present in various mouse tissues[8]. We mapped these proteins and found that the barcode was more sensitive at declaring genes, approximated by proteins found in the tissues, present (Supplemental Figure 2E).

The utility of our algorithm was demonstrated by developing a classification scheme that assigns tissue types to unknown samples by comparing their barcode to predefined ones (using Euclidean distance). Various sample classification algorithms have been proposed for microarray data. Many of these were compared on the original expression estimates. Predictive Analysis of Microarrays (PAM) produced the best results (data not shown)[9]. We compared our approach to PAM using leave-one-out cross-validation (CV). Tissues with detailed annotation for which there were 3 or more samples were included. Table 1A shows the results, which include various clinical datasets[10–16]. The barcode outperformed PAM in all comparisons except two, where it performed as well. Because CV has a tendency to overestimate the performance of a classification algorithm, we assessed performance on six independent datasets not included in the CV process (Table 1B). Here only the barcode performed well, with similar accuracy to that seen with CV.

The fact that the barcode greatly outperformed PAM on the independent datasets is likely due to the lab/batch effect. Because studies usually target a particular tissue, a primary concern is that a strong lab effect will confound the ability to classify tissues from the ability to classify labs[2]. An example of the lab effect is shown in Figure 2A, where the correlations between samples from study E-AFMX-5 are high despite originating from a wide variety of tissues. The barcode approach can remove many of these effects because subtle changes in intensity values are not strong enough to make an absent gene appear present, or vice-versa. Notice that the barcode removes most of the correlations in the E-AFMX-5 study without removing the correlation between the brain tissues, both within the study and between studies (Figures 2B, 2C and 2D).

To assess the ability of the barcode algorithm to find undiscovered tissue subsets, we used data from three breast cancer studies that did not include normal breast tissue samples, but did include patient survival data[14–16]. The distance to all tissue barcodes was obtained and 499 of the 500 samples were classified as breast tumor (1 as bladder cancer). When we took out the breast tumor barcode 37 of these samples were close to a variety of normal tissues and the other 463 samples to a variety of cancer tissues. We then formed good and bad prognosis barcodes using these 37 and 463 samples, respectively. This new barcode was then used to re-classify the 500 samples. We iterated this procedure until the good and bad prognosis groups did not change. The final barcodes resulted in a powerful prognosis tool that outperformed the methods described in the original papers. The details can be found in the Supplemental Results.

We expect the barcode approach to classification and discovery presented in this paper to improve in various ways. First, the classification algorithm implemented on the barcode was based on a very simple detection method and distance calculation. Many aspects can be optimized for prediction purposes. We expect the machine learning community will help improve this already powerful algorithm. Second, as microarray technology improves so will the barcode performance. In particular, the emergence of better gene annotation and arrays that probe for individual exons are the most promising developments. Finally, we have only implemented the barcode for two widely used platforms: Affymetrix HGU133A human array and MOE430 mouse array. However, as soon as enough public data is available, the barcode will be defined for other platforms.

In conclusion, we would like to acknowledge the efforts from the Microarray Gene Expression Data (MGED) Society to promote the sharing of microarray data. The work presented here would not have been possible without the existing public repositories. In particular, the availability of raw data was key as the methods used to process raw data into gene level measurements also contribute to study-to-study variability[17, 18]. We hope this trend continues, as we believe it to be necessary for microarray technology to fulfill its promise to help diagnose and treat disease.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

REFERENCES

1. Irizarry, RA.; Gautier, L.; Cope, LM. The Analysis of Gene Expression Data: Methods and Software. Parmigiani, G.; Garrett, ES.; Irizarry, RA.; Zeger, SI., editors. New York: Springer-Verlag; 2003. p. 102-119.
2. Irizarry RA, et al. Multiple-laboratory comparison of microarray platforms. Nat. Methods. 2005; 2:345–350. [PubMed: 15846361]
3. Kim S, et al. Multivariate measurement of gene expression relationships. Genomics. 2000; 67:201–209. [PubMed: 10903845]
4. Pal R, Datta A, Fornace AJ Jr, Bittner ML, Dougherty ER. Boolean relationships among genes responsive to ionizing radiation in the NCI 60 ACDS. Bioinformatics. 2005; 21:1542–1549. [PubMed: 15598836]
5. Barrett T, et al. NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res. 2005; 33:D562–D566. [PubMed: 15608262]
6. Parkinson H, et al. ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. 2005; 33:D553–D555. [PubMed: 15608260]
7. Carter SL, Eklund AC, Mecham BH, Kohane IS, Szallasi Z. Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in

cancer-associated gene expression measurements. BMC Bioinformatics. 2005; 6:107. [PubMed: 15850491]

8. Kislinger T, et al. Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. Cell. 2006; 125:173–186. [PubMed: 16615898]
9. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. U.S.A. 2002; 99:6567–6572. [PubMed: 12011421]
10. Blalock EM, et al. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proc. Natl. Acad. Sci. U.S.A. 2004; 101:2173–2178. [PubMed: 14769913]
11. Kimchi ET, et al. Progression of Barrett's metaplasia to adenocarcinoma is associated with the suppression of the transcriptional programs of epidermal differentiation. Cancer Res. 2005; 65:3146–3154. [PubMed: 15833844]
12. Dyrskjot L, et al. Gene expression in the urinary bladder: a common carcinoma in situ gene expression signature exists disregarding histopathological classification. Cancer Res. 2004; 64:4040–4048. [PubMed: 15173019]
13. Lenburg ME, et al. Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data. BMC Cancer. 2003; 3:31. [PubMed: 14641932]
14. Miller LD, et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc. Natl. Acad. Sci. U.S.A. 2005; 102:13550–13555. [PubMed: 16141321]
15. Pawitan Y, et al. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res. 2005; 7:R953–R964. [PubMed: 16280042]
16. Sotiriou C, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J. Natl. Cancer Inst. 2006; 98:262–272. [PubMed: 16478745]
17. Irizarry RA, Wu Z, Jaffee HA. Comparison of Affymetrix GeneChip expression measures. Bioinformatics. 2006; 22:789–794. [PubMed: 16410320]
18. Shi L, et al. Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential. BMC Bioinformatics. 2005; 6 Suppl 2:S12. [PubMed: 16026597]

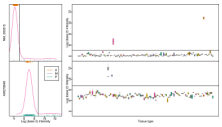

Figure 1.

Across sample gene expression estimates distributions. Data for two human genes are shown with the Genbank accession number on the y-axis. The left pane shows across sample distribution for all tissues. The vertical line is automatically drawn by the barcode method and distinguishes the intensity range associated with expressed and unexpressed genes. The orange, purple and green ticks denote the observed values with color denoting the call provided by the manufacturer. Absent calls are shown on the top axis while present and marginal calls are shown on the bottom axis. Notice that for the gene shown above, the calls appear consistent with the plot. However, these calls appear unable to distinguish expressed from unexpressed for the gene shown in the bottom row. The boxplots stratify these calls by tissue. The horizontal line denotes the expressed/unexpressed boundary. Notice that all samples of the same tissue are consistently present or consistently absent.

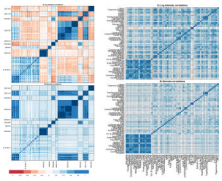

Figure 2.
Demonstration of the lab effect and its removal by the barcode algorithm. The correlation between every pair of samples from studies containing 10 or more arrays is shown in different shades of blue (negative correlation in red). For each gene the across-sample mean value was subtracted from the expression. The different studies are divided by black lines and labeled in the axes. In A) the blue high-correlation block on the bottom left demonstrates the lab effect. Most of the tissues in that study are completely different types and therefore the lab effect must be driving the high correlation. B) The same plot but using the barcode to define correlation. C) and D) The same plot zooming in to study E-AFMX-5. The lines now separate the different tissues.

Table 1

A) Percentage accuracy comparison: PAM versus the barcode approach in six data sets described in the text. In the normal tissues we assessed the ability to distinguish between tissues. For the Alzheimer's data we assessed the ability to: 1) distinguishing normals from all disease samples and 2) distinguishing normal samples from severe Alzheimer's samples. For the adenocarcinoma data we assessed the ability to: 1) distinguish normal, Barrett's esophagus (a precursor) and adenocarcinoma samples and 2) distinguish normal from precursor and cancer samples. For the bladder cancer data we assessed the ability to: 1) distinguish normal, muscle invasive transitional cell carcinoma (mTCC) and superficial transitional cell carcinoma (sTCC) samples and 2) distinguish normal and cancer samples. Finally, the annotation for the renal cell carcinoma data was simply normal and cancer thus that was the only comparison performed. B) Percentage accuracy comparison on independent datasets: PAM versus the barcode approach in six randomly selected data sets not included in the original database. The data described in Supplemental Table 1 was used to train the prediction algorithms.

A

Samples	Comparison Type	PAM (% correct)	Barcode (% correct)
Human Normal Tissues	Different tissues	95	98
Mouse Normal Tissues	Different tissues	91	96
	Normal versus disease	60	70
Alzheimer's disease	Normal versus severe disease	83	91
	Three different conditions	83	83
Adenocarcinoma	Normal versus cancer/precursor	91	91
	Three different conditions	73	83
Bladder Cancer	Normal versus cancer	90	96
Renal Cell Carcinoma	Normal versus cancer	94	100

B

GEO ID	Data type	PAM (% Correct)	Barcode (% Correct)
GSE5388	Cortex	100	100
GSE2395	Respiratory system epithelia	0	100
GSE2665	Lymph node/tonsil	35	95
GSE1561	Breast tumor	69	100
GSE2603	Breast tumor	77	90
GSE6344	Kidney: normal versus cancer	100	100