Nucleic Acids Research, 2011, Vol. 39, Database issue ***D1011–D1015***
doi:10.1093/nar/gkq1259

The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes

Matthew N. McCall[1], Karan Uppal[2], Harris A. Jaffee[1], Michael J. Zilliox[3,*] and Rafael A. Irizarry[1,*]

[1]Department of Biostatistics, Johns Hopkins University Bloomberg School of Public Health, 615 N. Wolfe Street, Baltimore, MD 21205, USA, [2]Bimcore and [3]Department of Microbiology and Immunology, Emory University School of Medicine, 1510 Clifton Road NE, Atlanta, GA 30322, USA

Received November 18, 2010; Accepted November 19, 2010

ABSTRACT

Various databases have harnessed the wealth of publicly available microarray data to address biological questions ranging from across-tissue differential expression to homologous gene expression. Despite their practical value, these databases rely on relative measures of expression and are unable to address the most fundamental question—which genes are expressed in a given cell type. The Gene Expression Barcode is the first database to provide reliable absolute measures of expression for most annotated genes for 131 human and 89 mouse tissue types, including diseased tissue. This is made possible by a novel algorithm that leverages information from the GEO and ArrayExpress public repositories to build statistical models that permit converting data from a single microarray into expressed/unexpressed calls for each gene. For selected platforms, users may upload data and obtain results in a matter of seconds. The raw data, curated annotation, and code used to create our resource are also available at http://rafalab.jhsph.edu/barcode.

INTRODUCTION

The completion of the human genome led to an unprecedented number of biological discoveries including the sequence and location of genes. However, knowledge of our genome provides little information about what distinguishes cell types. In contrast, knowledge of our transcriptome, the genes expressed in our cells, provides insight into what distinguishes cell types. Here we report on a draft of the human and murine transcriptomes based on public microarray data. Studies based on microarray data (1,2) typically report results in terms of relative expression, for example, which genes are differentially expressed in one condition compared to others. Minimal attention has been devoted to determining which genes are expressed in a specific tissue or cell type. Databases such as EBI's Human Gene Expression Map (3), TiGER (4), BODYMAP (5), BioGPS (6) and TiSGeD (7) provide comprehensive microarray data from thousands of samples but none focus on absolute measures of expression. Furthermore, these databases are useful for gene-by-gene analysis, but have limited functionality for researchers interested in a global view of transcription for specific tissue types.

Knowing which genes are expressed in which tissues will increase our fundamental understanding of cellular processes and provide a starting point for research targeting drug discovery and individualized medicine. The difficulty in obtaining absolute measures of expression stems from the fact that feature characteristics, such as probe sequence, can cloud the relationship between observed intensities and actual expression levels (8). This 'probe effect' is large, yet consistent between hybridizations on the same platform, which implies that it can be cancelled out by relative measures of expression (9). However, the probe effect makes it impossible to interpret differences in intensities between genes. This implies that one cannot infer presence of a transcript from the size of reported microarray values (Figure 1A).

To estimate transcriptomes we extended the original gene expression barcode methodology (8), which provides reliable absolute measures of expression but for a limited number of genes: 2519 for human and 5031 for mouse. This algorithm requires genes to show clear separation between low and high expression measurements to classify groups into silenced and expressed. However,

*To whom correspondence should be addressed. Tel: +1 410 614 5157; Fax: +1 410 955 0958; Email: rafa@jhu.edu
Correspondence may also be addressed to Michael J. Zilliox. Tel: +1 404 712 2596; Fax: +1 404 727 3722; Email: mzillio@emory.edu

most genes do not exhibit this property. To create a comprehensive estimate of cell-type transcriptomes, we extended the original barcode methodology to provide expression calls for all genes on the array. We achieved this by (i) developing a series of negative control experiments; (ii) leveraging information from 13 824, 18 656 and 9652 publically available chips from the Affymetrix Human Genome U133A (HGU133a), U133 Plus 2.0 (HGU133plus2) and Mouse Genome 430 2.0 (Mouse4302) platforms, respectively; and (iii) developing a novel application of the probability of expression (POE) model (10). This resulting approach provided a new version of the barcode algorithm that provided standardized values that for the first time permitted comparison across all genes (Figure 1B). These standardized values can then be converted to silenced and expressed calls by deciding on a single threshold. We refer to this binary version of the expression values as a 'barcode'. For details on the new statistical approach and comparisons to existing detection algorithms, see the Supplementary Data. The Supplementary Data also describes how the binary barcode version of the data provides protection against the ubiquitous batch effect (11); a fact described by Zilliox and Irizarry (8) and confirmed by the Microarray Quality Control (MAQC) II project (12). Below we include examples of practical uses of our database.

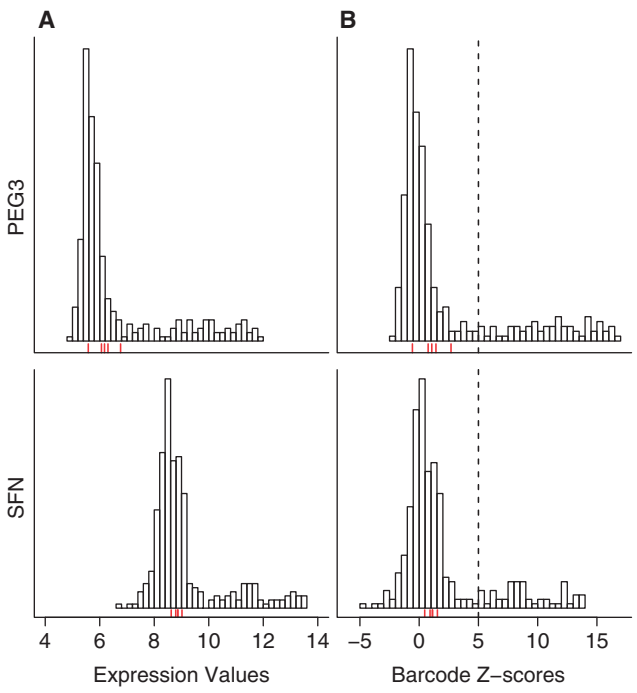

Figure 1. Histograms of reported expression measurements and barcode standardized values. (**A**) Reported expression values for two genes. Values from all samples in the barcode database are shown. The red tick marks on the *x*-axis represent values from yeast samples expected not to hybridize. Both genes each have a single mode with a long right tail. We assume values near the mode correspond to the gene being silenced and values well above the mode correspond to the gene being expressed. However, these two genes clearly have different modes. If we were to use the background distribution of the first gene (PEG3) to estimate whether the second gene (SFN) is expressed, SFN would appear to be expressed in nearly every tissue. (**B**) Values standardized with the barcode approach. Notice that the mode of each distribution is now approximately zero and the yeast samples are clustered near zero. The dash lines represent a possible threshold to convert the barcode standardized measurements into a gene expression barcode.

Estimated transcriptomes

The barcode database provides absolute calls of expression for 13 824, 18 656 and 9652 samples from the HGU133a, HGU133plus2 and Mouse4302 platforms, respectively, obtained from GEO and Array Express. We manually curated sample annotation associated with these arrays and required at least five biological replicates for each cell type. This resulted in 78, 131 and 89 different normal and cancer cell types for HGU133a, HGU133plus2 and Mouse4302, respectively. In each cell type, for each gene we computed the proportion of samples for which that gene was called expressed. These values defined our estimated transcriptome for each cell type. Tables containing these values for each cell type represented in each of the three platforms are available here: http://rafalab.jhsph.edu/barcode/index.php?page=transcriptome.

We verified the biological validity of our transcriptomes by grouping the genes expressed in $CD4^+$ T cells, cerebellum, liver and skeletal muscle by gene ontology. Functional annotation clustering using DAVID (13) showed that the most enriched biological groups were those expected for a given tissue (Table 1). For example, gene groups involved in synaptic transmission were found in the cerebellum, while groups involved in muscle contraction were specific to skeletal muscle. We provide the analysis for these four cell types only as an example and expect researchers with the appropriate expertise to take advantage of our resource and perform more detailed analysis of the specific genes found in each tissue.

Gene expression distributions

The consistency of barcode expression calls within cell-type replicates was remarkable. To illustrate this we focused on normal tissues from HGU133a, in which 90% of the proportions were exactly 0 or 1. Because 51% of the non-consistent calls were due to just 12% of the genes, we quantified consistency with a gene-specific entropy measure. We speculate that these genes are associated with poor performing probe sets. Another possibility is that they are universally varying genes such as those involved in the circadian rhythm. Because the transcriptome values of these genes should be interpreted with care, we flag them in our database. This classification can be useful to users that are interested in using barcode data in prediction applications such as clinical diagnostics.

Our database also provides a global view of transcriptomes, allowing a user to examine global characteristics of gene expression. For example, in our estimate of the human transcriptome, most genes were primarily off, and a small proportion primarily on, across cell types: 76% of genes were off in at least 80% of tissues and 2% of genes were on in at least 80% of tissues. Among the remaining 22% of genes, 55% were high entropy genes.

Table 1. Functional annotation clustering using DAVID

$CD4^+$ T cells		Cerebellum		Liver		Skeletal muscle	
GO Term	ES	GO Term	ES	GO Term	ES	GO Term	ES
RNA metabolic process	12.4	Synaptic transmission	9.7	Cellular ketone metabolic process	26.2	Muscle contraction	15.8
Cellular macro-molecule cata-bolic process	8	Transport	9.5	Monocarboxylic acid metabolic process	16	Muscle organ development	9.1
Cellular protein metabolic process	7.6	Neurogenesis	7.4	Organic acid catabolic process	15.7	Striated muscle tissue development	7.1
Apoptosis	7.2	Nervous system development	7.3	Steroid metabolic process	11.4	Energy derivation by oxidation of organic compounds	5.9
Lymphocyte activation	6.2	Cytoskeleton organization	6	Wound healing	10.7	Anatomical structure development	4.5

The transcriptomes of four tissues were clustered using DAVID. The gene ontology (GO) term with the lowest *P*-value is shown to represent each cluster. ES, enrichment score.

Catalog and webtools

In our database, we used the estimated transcriptomes to create catalogs, available from http://rafalab.jhsph.edu/barcode/index.php?page=catalog, with gene-specific information. Specifically, for each gene we reported the cell types in which it was expressed, similar to other databases, and the estimated entropy for that gene. Our database can also be used to determine the genes expressed in a given cell type (http://rafalab.jhsph.edu/barcode/index.php?page=tissuegene) and to compare the genes expressed in two or more cell types (http://rafalab.jhsph.edu/barcode/index.php?page=tissuecomp). A unique aspect of our database is that it can be used to identify genes in a particular tissue that have not been well studied. For example, we found that there were 1298 expressed genes found in $CD4^+$ T cells (excluding genes found in 'many' tissues). We then narrowed the list down to Affymetrix grade E and R genes, which correspond to expressed sequence tags (ESTs) and have little or no functional annotation. We also excluded high entropy genes. Of the 130 genes that fit these criteria, 12 were only found in $CD4^+$ T cells and not any other tissue in the database, demonstrating a query unique to this catalog.

Comparison to other tools

While our database differs fundamentally from exiting web-tools in that it can produce barcodes from a single microarray and provides absolute measure of expression, three other databases can be used to create a list of genes that are specifically up-regulated for a given cell type; namely EBI's Human Gene Expression Map, TiGeR and BODYMAP. We assessed the performance of our resource relative to others using sec-gen RNA sequencing counts as a gold standard. Our database provided a substantial improvement in sensitivity over existing databases without sacrificing specificity. In fact our specificity was equal or superior in all comparisons (Table 2). More details are provided in the Supplementary Data.

Table 2. Comparison to other tools

Method	Tissue	Expressed	FP, %
Barcode	Kidney	761	13
TiGER	Kidney	320	13
EBI	Kidney	245	14
Barcode	Liver	695	21
TiGER	Liver	295	41
Bodymap	Liver	36	25

For the competing methods we determined genes that were up-regulated in kidney and liver as compared to other tissues. For the barcode we simply obtained genes called expressed. We then compared to sec-gen data to determine the false postive rate. The barcode finds the greatest number of expressed genes in both tissues (Column 3) while maintaining the lowest false positive rate (Column 4). Note that the EBI tool does not provide information for liver, and the Bodymap does not provide it for kidney.

Orthologous genes

Here we present a powerful illustration of the added value our resource provides the research community. There have been conflicting reports on the similarity of gene expression profiles between human and mouse orthologous genes. Some have gone so far as to claim that 'any human tissue is more similar to any other human tissue examined than to its corresponding mouse tissue' (14). Data produced with two different platforms, human and mouse, designed with different probes, supported this conclusion. This led others to claim that platform-specific technical variability is the cause of this apparent dissimilarity (15,16). Our resource easily clarifies this controversy. We selected a number of normal tissues represented in both our human and mouse databases. Using reported gene expression measurements to cluster human and mouse tissues, we observed a strong species-effect: species clustered together (Figure 2A). However, using barcode data from our database, the species-effect was removed and we observed perfect clustering among the tissues (Figure 2B). Our result supported the more intuitive biological conclusion that, for example, a

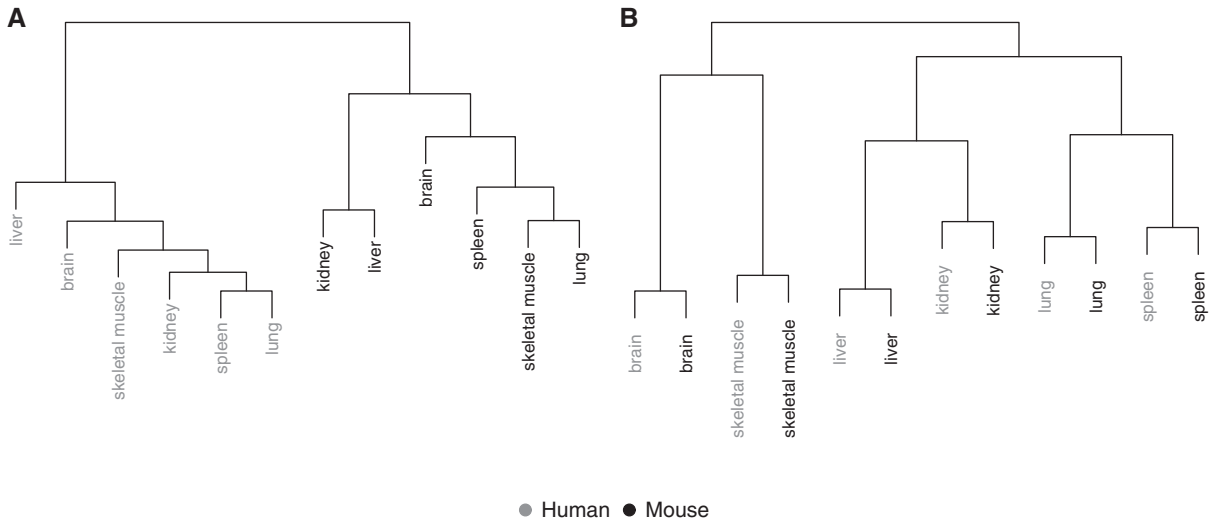

Figure 2. Hierarchical clustering of human and mouse tissue samples using orthologous genes. These are based on (**A**) average expression microarray measurements and (**B**) tissue specific transcriptomes based on averaged barcodes. The same genes were used in (A) and (B).

human kidney is more like a mouse kidney than a human brain. Because barcodes are designed to be robust to probe-effects, these results provide strong evidence that large species-specific differences are in fact driven by systematic biases. These biases are absent in our database, allowing users to compare transcriptomes between species.

Cell type predictions

Note that while most computational pipelines require data from multiple microarray experiments for normalization purposes, our resource is able to process data from a single array. A sample of unknown origin can be uploaded here: http://rafalab.jhsph.edu/barcode/index.php?page=sample_process, our resource then processes the sample, compares the result to each pre-computed cell type transcriptome, and reports back distances between the sample-specific barcode and tissue-type-specific transcriptome. The barcode methodology is robust to batch effects and can be applied to data from a single chip, which makes it particularly useful for across study cell type prediction (8). Because our database contains normal and cancer cell types, one can upload a tumor of unknown origin and determine the cancer cell type to which it is most similar. The Supplementary Data includes a promising example illustrating the potential of using our resource for cancer diagnosis.

CONCLUSIONS

The Gene Expression Barcode is the first resource to reliably call genes silenced or expressed using data from a single microarray. We harness this functionality to estimate the human and mouse transcriptomes and provide results via our database. Specifically, we can create a gene expression barcode for a single microarray that provides information about the expression states of all genes. We have combined thousands of gene expression barcodes to create vast catalogs of transcriptome information spanning hundreds of cell types and tens of thousands of genes. These catalogs are easily accessible via a series of webtools that allow an investigator to readily access gene and/or cell type specific information. Users can also calculate the transcriptomes for their own samples, which is useful for researchers looking at epigenetic markers compared to gene expression or those conducting genome wide association studies (GWAS) where gene expression is a phenotype. These resources will be updated and expanded to other platforms as more data becomes publicly available; we have created tools that automatically do this.

We note that our approach provides only an early draft of the transcriptomes. We can expect various improvements. For example, although we provide the original expression data, the default barcode data currently dichotomizes the data into two classes mainly to protect against batch effects. Future versions will further stratify the expressed strata. We also realize that much of the non-coding RNA expression is yet undefined, and hence much work on the transcriptome remains. We therefore plan to include second generation RNA sequencing data as soon as enough samples become publicly available to implement an adapted version of our algorithm. However, despite these limitations, we have provided various examples of powerful analyses permitted by the estimated transcriptomes. We expect many more applications to flourish and discoveries to be made as more experts become aware and use our tools.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Yongqiang Zhang for providing us with the yeast RNA and Tianwei Yu for help with the sample

parsing code. We thank the maintainers of GEO and ArrayExpress for making the data publicly available. We thank Thomas Louis for his feedback on the statistical aspects of the article. M.N.M. conceived the study, developed the software and drafted the article. R.A.I. conceived the study and drafted the article. M.J.Z. conceived the study, revised the article and provided the data. K.U. and H.A.J. developed software.

FUNDING

National Institutes of Health, partial (GM083084, RR021967 and UL1RR025005 to R.A.I.); National Cancer Institute (CA132480 to M.J.Z.); National Institutes of Health, partial (T32GM074906 to M.N.M.). Funding for open access charge: R01GM083084.

Conflict of interest statement. None declared.

REFERENCES

1. Parkinson,H., Kapushesky,M., Kolesnikov,N., Rustici,G., Shojatalab,M., Abeygunawardena,N., Berube,H., Dylag,M., Emam,I., Farne,A. *et al.* (2009) ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, **37**, D868–D872.
2. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
3. Lukk,M., Kapushesky,M., Nikkila,J., Parkinson,H., Goncalves,A., Huber,W., Ukkonen,E. and Brazma,A. A global map of human gene expression. *Nat. Biotechnol.*, **28**, 322–324.
4. Liu,X., Yu,X., Zack,D.J., Zhu,H. and Qian,J. (2008) TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, **9**, 271.
5. Ogasawara,O., Otsuji,M., Watanabe,K., Iizuka,T., Tamura,T., Hishiki,T., Kawamoto,S. and Okubo,K. (2006) BodyMap-Xs: anatomical breakdown of 17 million animal ESTs for cross-species comparison of gene expression. *Nucleic Acids Res.*, **34**, D628–D631.
6. Su,A.I., Cooke,M.P., Ching,K.A., Hakak,Y., Walker,J.R., Wiltshire,T., Orth,A.P., Vega,R.G., Sapinoso,L.M., Moqrich,A. *et al.* (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA*, **99**, 4465–4470.
7. Xiao,S.J., Zhang,C. and Ji,Z.L. (2010) TiSGeD: a Database for Tissue-Specific Genes. *Bioinformatics*.
8. Zilliox,M.J. and Irizarry,R.A. (2007) A gene expression bar code for microarray data. *Nat. Methods*, **4**, 911–913.
9. Irizarry,R.A., Warren,D., Spencer,F., Kim,I.F., Biswal,S., Frank,B.C., Gabrielson,E., Garcia,J.G., Geoghegan,J., Germino,G. *et al.* (2005) Multiple-laboratory comparison of microarray platforms. *Nat. Methods*, **2**, 345–350.
10. Parmigiani,G., Garrett,E.S., Anbazhaghan,R. and Gabrielson,E. (2002) A statistical framework for expression-based molecular classification in cancer. *J. Roy. Stat. Soc. B (Stat. Meth.)*, **64**, 20.
11. Leek,J.T., Scharpf,R.B., Bravo,H.C., Simcha,D., Langmead,B., Johnson,W.E., Geman,D., Baggerly,K. and Irizarry,R.A. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, **11**, 733–739.
12. Shi,L., Campbell,G., Jones,W.D., Campagne,F., Wen,Z., Walker,S.J., Su,Z., Chu,T.M., Goodsaid,F.M., Pusztai,L. *et al.* The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.*, **28**, 827–838.
13. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
14. Yanai,I., Graur,D. and Ophir,R. (2004) Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *OMICS*, **8**, 15–24.
15. Liao,B.Y. and Zhang,J. (2006) Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol. Biol. Evol.*, **23**, 530–540.
16. Xing,Y., Ouyang,Z., Kapur,K., Scott,M.P. and Wong,W.H. (2007) Assessing the conservation of mammalian gene expression using high-density exon arrays. *Mol. Biol. Evol.*, **24**, 1283–1285.