AILUN: reannotating gene expression data automatically

To the editor: Gene Expression Omnibus (GEO)[1] is a public repository for gene expression data. While the amount of data in GEO has grown exponentially, the number of publications citing GEO has only grown linearly. The difficulty in data reuse is the mapping of probes in GEO datasets to established gene identifiers, which can change as annotations for the underlying sequences change[2]. Therefore, microarray results need to be reevaluated with the latest probe annotations. There have been several previous efforts to reannotate microarray probe identifiers[3,4], but only for a few platforms and species.

We built a fully automated system, Array Information Library Universal Navigator (AILUN), to reannotate all types of microarrays in GEO periodically by relating every probe identifier to Entrez Gene identifiers. First, we collected all gene identifiers from Entrez Gene and UniGene and built a universal gene identifier table (UGIT). We then matched each column of every GEO platform with UGIT to find the best matching column and type of external identifier, and annotated each probe identifier with Entrez Gene identifiers. (**Supplementary Methods** and **Supplementary Fig. 1** online).

UGIT contained 75 million gene identifiers of 90 types for 3,585 species. AILUN successfully reannotated 66% gene expression platforms, allowing reuse of 77% of samples across 79 species. The platform annotation coverage was 5 times greater than that in GEO (**Table 1**), and 94% identical for probes annotated by both AILUN and GEO. To validate the accuracy of annotation, we compared the annotations on Affymetrix U133A 2.0 across AILUN, GEO and NetAffx[5] using Brainarray[3] as the gold standard, which is based on probe-sequence matching. AILUN performed as well as NetAffx with 97% precision and 97% recall, and outperformed GEO with 98% precision and 86% recall (**Supplementary Tables 1–3** and **Supplementary Discussion** online).

The server (http://ailun.stanford.edu) offers four functions to help users reannotate platforms. 'Platform annotation' adds the latest annotations to any uploaded result file. 'Cross-species mapping' maps platform annotations to other species. 'Platform comparison' compares any two platforms to find corresponding probes mapping to the same gene. 'Gene search' finds deposited platforms and samples in GEO for any list of genes.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

Supported by Lucile Packard Foundation for Children's Health, US National Library of Medicine (K22 LM008261), National Institute of General Medical Sciences (R01 GM079719), Howard Hughes Medical Institute, and Pharmaceutical Research and Manufacturers of America Foundation. We thank A. Skrenchuk for computer support and A. Chiang for manuscript review.

Rong Chen[1], Li Li[2] & Atul J Butte[1–3]

[1]Stanford Medical Informatics, Department of Medicine, and [2]Department of Pediatrics, Stanford University School of Medicine, 251 Campus Drive, Stanford, California 94305, USA. [3]Lucile Packard Children's Hospital, 725 Welch Road, Palo Alto, California 94304, USA.

e-mail: abutte@stanford.edu

1. Barrett, T. *et al. Nucleic Acids Res.* **35**, D760–D765 (2007).
2. Perez-Iratxeta, C. & Andrade, M.A. BMC *Bioinformatics* **6**, 183 (2005).
3. Dai, M. *et al. Nucleic Acids Res.* **33**, e175 (2005).
4. Tsai, J. *et al. Genome Biol.* **2**, Software0002 (2001).
5. Liu, G. *et al. Nucleic Acids Res.* **31**, 82–86 (2003).

Table 1 | Performance comparison

Species	Total in GEO		Annotated by AILUN		Annotated by GEO		Annotated by AILUN and GEO	
	Platforms	Samples	Platforms	Samples	Platforms	Samples	Platforms	Samples
Human	813	80,543	602	61,132	144	40,885	140	40,624
Mouse	367	27,083	321	25,586	70	18,096	67	17,923
Rat	87	11,324	71	11,131	27	8,590	27	8,590
Yeast	204	8,069	80	2,851	5	873	1	841
Arabidopsis	68	5,833	43	5,154	9	303	9	303
Fruit fly	60	3,129	54	3,088	6	10,75	6	1,075
Total (including other species)	2,232	155,472	1,469	119,358	294	71,531	266	70,424

AILUN and GEO comparison based on the number of reannotated array platforms and the number of samples enabled for reuse.