Published in final edited form as:
Nat Methods. 2007 November ; 4(11): 879. doi:10.1038/nmeth1107-879.

AILUN: Re-annotating Gene Expression Data Automatically

Rong Chen[1], **Li Li**[2], and **Atul J Butte**[1,2,3]

[1] Stanford Medical Informatics, Department of Medicine, Stanford University School of Medicine, 251 Campus Drive, Stanford, CA 94305 USA

[2]Department of Pediatrics, Stanford University School of Medicine, 251 Campus Drive, Stanford, CA 94305 USA

[3]Lucile Packard Children's Hospital, 725 Welch Road, Palo Alto, CA 94304 USA

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 lies with the mapping of probes in GEO data sets to established gene identifiers, which can change as annotations for the underlying sequences change[2]. Therefore, microarray results need to be re-evaluated with the latest probe annotations. There have been several previous efforts to re-annotate microarray probe identifiers [3,4] but only for a few platforms and species.

We built a fully automated system, AILUN, to re-annotate all types of microarrays in GEO periodically by relating every probe ID to Entrez Gene IDs. 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 ID with Entrez Gene IDs. (Supplementary Methods **and** Supplementary Fig. 1 on line).

UGIT contained 75 million (M) gene identifiers of 90 types for 3585 species. AILUN successfully re-annotated 66% gene expression platforms, enabling reuse of 77% samples across 79 species. The platform annotation coverage was 5 times larger than GEO (Table 1) and 94% identical for those probes annotated by AILUN and GEO. To validate, 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 tied NetAffx at 97% precision and 97% recall, and outperformed GEO with 98% precision and 86% recall (Supplementary Table 1-3 and Supplementary Discussion on line).

The server (http://ailun.stanford.edu) offers four functions to help users re-annotate 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.

Corresponding Author: Atul J Butte, MD, PhD Email: E-mail: abutte@stanford.edu.

COMPETING INTERESTS STATEMENTS
The authors declare no competing financial interests.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

Supported by Lucile Packard Foundation for Children's Health, 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 Alex Skrenchuk and Annie Chiang from Stanford University for computer support and manuscript review, respectively.

REFERENCES

1. Barrett T, et al. Nucleic Acids Res 2007;35:D760–765. [PubMed: 17099226]
2. Perez-Iratxeta C, Andrade MA. BMC Bioinformatics 2005;6:183. [PubMed: 16033654]
3. Dai M, et al. Nucleic Acids Res 2005;33:e175. [PubMed: 16284200]
4. Tsai J, et al. Genome Biol 2001;2SOFTWARE0002
5. Liu G, et al. Nucleic Acids Res 2003;31:82–86. [PubMed: 12519953]

Table 1

Performance comparison. AILUN and GEO are compared on the number of re-annotated array platforms and the number of samples enabled for reuse.

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	1075	6	1075
Total (including other species)	2232	155,472	1469	119,358	294	71,531	266	70,424