Annotation and Cross-Indexing of Array Elements on Multiple Platforms

William B. Mattes

Investigative Toxicology, Pfizer Inc, Kalamazoo, Michigan, USA

On the surface, transcript profiling using microarrays seems to offer a way of looking at the global response of the cell to perturbation, with a focus on changes in gene expression. The difficulty, however, is that the response of a particular gene is actually measured on the array by an element that is a short, defined nucleic acid sequence. Sequences that map back to the same genetic locus may actually be given different names and descriptions when they are deposited in public sequence databases; when such sequences are used in microarray construction, elements that monitor the same genetic locus may have different names and descriptions. The algorithm described here uses a hierarchical approach to assign a single best annotation to the elements in a given microarray in such a fashion that elements from one microarray platform may be cross-indexed with those of another. The algorithm relies on the nucleic acid accession number for a given array element, and uses that to retrieve annotation from the most recent versions of LocusLink and UniGene. Both database resources are searched, with a priority being given to annotation derived from the curated LocusLink database. In lieu of annotation found in these databases, the default GenBank annotation is used. As a final outcome, a cross-chip identifier is generated that may be used to cross-index array elements. The program is available as a practical extraction and report language (Perl) script that can run under any Perl interpreter. *Key words:* annotation, cross-platform, indexing, LocusLink, microarray, UniGene. *Environ Health Perspect* 112:506–510 (2004). doi:10.1289/txg.6698 available via *http://dx.doi.org/* [Online 15 January 2004]

On the surface, microarrays and other genomic technologies offer the toxicologist a look at the transcript levels for hundreds to thousands of genes. However, although toxicologists and cell biologists think in terms of genes and pathways, these technologies actually measure nucleic acid sequences. Thus, the challenge is to clearly associate a given nucleic acid sequence with the most current and consistent information on the gene of which it is part. This association is complicated by the fact that the same sequence can be submitted to public databases from several sources that may assign it different names and descriptions. For example, the gene, N-*myc* downstream regulated (*Ndrg1*) (LocusID 10397; http://www.ncbi.nih.gov/LocusLink/) was originally cloned and submitted by three laboratories as different sequences with different names: *RTP* (accession no. D87953; http://www.ncbi.nih.gov/GenBank), a homocysteine-respondent gene in vascular endothelial cells (Kokame et al. 1996); *DRG1* (GenBank accession no. X92845), a gene upregulated during colon epithelial cell differentiation (Van et al. 1997); and *CAP43* (GenBank accession no. AF004162), a gene specifically induced by Ni^{2+} compounds (Zhou et al. 1998). All three sequences are identical and represent the same gene. Microarrays are built using individual sequences or clones that are annotated in this fashion, and thus identifying microarray elements (i.e., spots) on a single array or on different arrays that represent a certain gene can be a frustrating exercise.

Our approach to annotate microarray elements makes use of two public databases: UniGene (http://www.ncbi.nih.gov/UniGene/; Wheeler et al. 2000) and LocusLink (http://www.ncbi.nih.gov/LocusLink/; Pruitt and Maglott 2001). Whereas UniGene is an experimental system for grouping GenBank sequences (http://www.ncbi.nih.gov/GenBank/) into gene-oriented clusters, LocusLink is a database of curated sequence and descriptive information about genetic loci. Together these resources allow us to map a given microarray element to a certain gene, using UniGene and the GenBank accession number of the element, and to annotate that gene using LocusLink information. Furthermore, the process for doing so is automated with a computer script that can be run on a regular basis to make use of current database information. Although our approach appears to be similar to that taken by the DRAGON database (http://pevsnerlab.kennedykrieger.org/dragon.htm; Bouton and Pevsner 2000) and the DAVID software (http://apps1.niaid.nih.gov/david/upload.asp) (Dennis et al. 2003), ours seeks to create a single best annotation for a sequence and, based upon this hierarchical process, to generate a cross-chip ID. Although there are caveats to this approach, the results show that it generally allows for intra- and interplatform identification of microarray elements representing a single gene. This approach has been applied to comparing results generated in the multilaboratory genomics research program coordinated by the International Life Sciences Institute (ILSI) Health and Environmental Sciences Institute (HESI) Committee on the Application of Genomics to Mechanism-Based Risk Assessment.

Materials and Methods

Algorithm rationale. Most developers of microarrays, either private or commercial (e.g., Affymetrix, Inc., Santa Clara, CA) will provide for each array element (i.e., probe) a GenBank accession number indicating the sequence or clone that the element represents or is derived from. On the other hand, the descriptive information for such GenBank entries or the locus that they are associated with may change as new information is deposited in the public databases, especially UniGene and LocusLink. Furthermore, UniGene and LocusLink can serve as sequence "Rosetta stones" where *a*) UniGene serves to collate accession numbers, *b*) UniGene integrates with LocusLink, *c*) LocusLink serves as a curated annotation database with canonical gene names and curated gene information, and *d*) LocusLink integrates with other information such as OMIM (Online Mendelian Inheritance in Man). To represent the best information for a particular microarray element, a cross-chip ID (XChipID) can be created based upon UniGene and LocusLink information, as described below.

This article is part of the mini-monograph "Application of Genomics to Mechanism-Based Risk Assessment."

Address correspondence to W.B. Mattes, Gene Logic Inc., 610 Professional Dr., Gaithersburg, MD 20879 USA. Telephone: (240) 364-6238. Fax: (240) 364-6262. E-mail: wmattes@genelogic.com

The author thanks the many colleagues who offered support and advice. The input of B. Pennie (Pfizer Inc), P. Lord (Johnson & Johnson Pharmaceutical Research Division), R. Paules [National Institute of Environmental Health Sciences (NIEHS)], and D. Robinson (Pfizer Inc) from the International Life Sciences Institute Health and Environmental Sciences Institute Committee on the Application of Genomics to Mechanism-Based Risk Assessment was critical to the initiation and continuation of this effort. J. Fostel (NIEHS), I. Reardon (Pfizer Inc), C. Storer (Pfizer), and M. Lawton (Pfizer Inc) offered especially helpful comments over the course of this project on the algorithm and Perl programming in general. The author also thanks C. Bradfield (McArdle Laboratory for Cancer Research, University of Wisconsin) for a careful review of this article. Finally, the author is indebted to S. Pettit (ILSI HESI) for her constant support and suggestions.

The authors declare they have no competing financial interests.

Received 25 August 2003; accepted 12 January 2004.

Algorithm and logic flow. The logic flow of annotation is illustrated in Figure 1. Essentially, the program searches the UniGene database for the accession number in question. If the accession number is referenced in UniGene, the next step is to seek information in LocusLink, using the UniGene Cluster ID. If the accession number is not referenced in UniGene, then the LocusLink database is checked for the accession number. (Some accession numbers are referenced in LocusLink but not in UniGene.) If the accession number is not referenced in either the UniGene or LocusLink databases, then the annotation in GenBank associated with that accession number is used. As noted, a XChipID is constructed on the basis of the best ID available, a LocusID being preferred to a UniGene ID, and if neither is found, a GenBank accession number. The prefix to the XChipID indicates the origin of the identifier (LL., LocusLink; Rn., rat; UniGene; Ac., GenBank).

Input files and software programs. The files obtained from the National Center for Biotechnology Information (NCBI) are listed in Table 1, along with the key value and cross-indexed values obtained from each file. Data reported here made use of *Rattus norvegicus* UniGene Build no. 117 and LocusLink data current to 27 May 2003. Scripts (i.e., program code) were written in Perl, version 5.6.1, a programming language developed in 1988 by Larry Wall as Open Source software (http://www.perl.org). Perl scripts are text-based programs run by an interpreter program, which has been developed for almost every operating system (e.g., Mac, PC, UNIX). Five Perl scripts were developed: UgXRef.pl to extract data from the Rn.data UniGene file; UgDupe.pl to examine UniGene data for duplicate entries; ChipXAnno.pl to collate data from the LocusLink, UniGene, and microarray definition files and carry out the annotation; ChipDupe2.pl to examine the microarray annotation file for multiple entries based on the XChipID; ChipCompare8.pl to compare two different microarray annotation files for overlapping entries; and XChipData.pl to merge data sets from two different microarray platforms. The outputs of all programs are simple text files, most of which are tab delimited, that can be imported into analysis programs such as Microsoft's Excel and Access (Microsoft Corp., Bellevue, WA) and Spotfire DecisionSite. All these programs have been run in a disk operating system (DOS) command line window using ActivePerl (binary build 629; http://www.activestate.com), although after conversion of the end-of-line sequence they run under UNIX. UgXRef.pl processes UniGene files and as such is memory intensive: for large UniGene files (e.g., for mouse and human), these scripts must be run on either DOS or UNIX systems with > 1 GB RAM. The scripts are small and are available from the web site for the HESI Committee on the Application of Genomics to Mechanism-Based Risk Assessment (http://hesi.ilsi.org/publications/index.cfm?pubentityid=120).

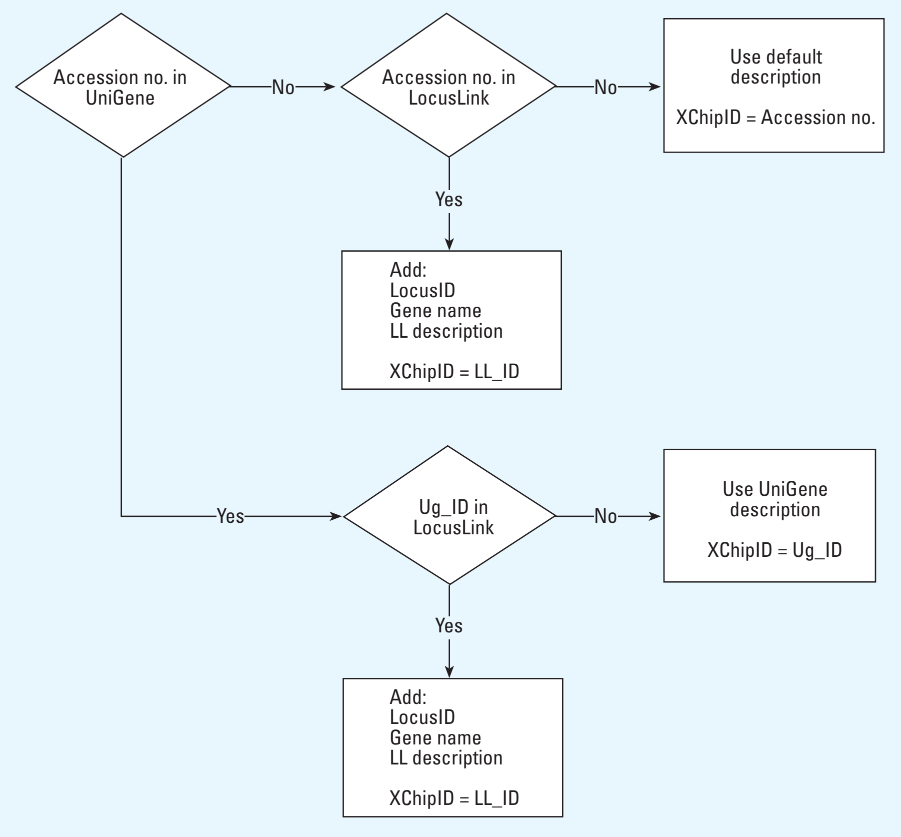

Figure 1. Algorithm flow in XChipAnno.pl.

Table 1. Input NCBI files for annotation.

File name	Source	Key value	Indexed values
loc2acc	LocusLink[a]	GenBank accession no.	LocusID
loc2UG	LocusLink	UniGene ID	LocusID
ll.out	LocusLink	LocusID	Gene symbol, LocusLink description
Rn.data	UniGene[b]	Used to create Acc2Ug_Rn.prn, Ug2Tit_Rn.prn	
Acc2Ug_Rn.prn	Rn.data	GenBank accession no.	UniGene ID
Ug2Tit_Rn.prn	Rn.data	UniGene ID	UniGene description

[a]ftp://ftp.ncbi.nlm.nih.gov/refseq/LocusLink/. [b]ftp//ftp.ncbi.nlm.nih.gov/repository/UniGene/.

Microarray definition files listing each microarray element and its associated accession number and description were obtained from individual vendors through the ILSI consortium.

The Blast2 program (http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html) was used to investigate the similarity and identity of various sequences at the protein level.

Results

Array annotation. The algorithm replaces frequently minimal sequence descriptions with biologically meaningful annotation. Thus, elements originally annotated as ESTs (Expressed Sequence Tags) are identified as corresponding to *Gstm2* and *Lgals1* (Table 2). It is important that in doing so the algorithm identifies multiple elements, including ESTs, that query the same locus. Examples given in Table 2 include cytochrome P450 1b1 (*Cyp1b1*), phosphodiesterase 4B (*Pde4b*), *Cyp4a10*, and endothelin receptor (*Ednrb*). Conversely, the algorithm can highlight elements incorrectly annotated. Thus, U39571, an element described as phosphatidylinositol

4-kinase (*Pik4ca*), was not annotated by the algorithm as *Pik4ca*. In fact, BLAST analysis (http://www.ncbi.nlm.nih.gov/BLAST/) showed that U39571 does not share significant sequence homology with the other *Pik4ca* sequences.

Occasionally microarray elements are incorrectly annotated and grouped. Although both accession number X81395 and accession number U10697 were annotated as carboxylesterase 1 (*Ces1*) (presumably because of DNA sequence homology), the amino acid sequences are divergent enough to suggest that these are indeed two different proteins (data not shown). However, as UniGene clusters and LocusLink information are updated, incorrect groupings can be resolved. Thus, when UniGene and LocusLink information from September 2001 was used, X14552 (alpha-2µ globulin, type 1) and M83298 (phosphatase 2A 55-kD regulatory subunit alpha) were annotated as caldesmon (LocusID 25687), based on short sequence overlaps. Using February 2002 UniGene and LocusLink data the sequences identified by these GenBank accession numbers were distinguished from caldesmon (data not shown). As with any system using these resources, the annotation is only as current as the UniGene and LocusLink files used for input.

The XChipID represents the best information available identifier for a given sequence element and as such offers a means to *a*) group elements that actually represent the same gene and *b*) estimate the number of unique genes queried by the microarray. Thus, the 8,740 elements on the Affymetrix RGd_U34a array are estimated to query a total of 6,385 unique genes (Table 3). Of course, the actual sequence queried by each element is different, and as such, these sequences may have different hybridization characteristics and give rise to quantitatively different signals.

Identification of homologous targets across array platforms. Using the XchipID, one can determine genes queried in common by two different microarray platforms and compare results at a relatively simplistic level.

Cross-array comparisons of the Affymetrix RG_U34a, the NIEHS 7K array (National Institute of Environmental Health Sciences, Research Triangle Park, NC), and the Clontech Atlas Tox2 arrays (Clontech, Palo Alto, CA, USA) indicate overlaps, as well as a substantial number of genes uniquely queried by each array (Figure 2).

Table 3. Summary of annotation results for Affymetrix Rat RG_U34a genome chip.[a]

No. of probe sets	8,740
LocusLink annotated	62.6%
UniGene-only annotated	22.4%
Unique	73.0%
Ambiguous ESTs	31.2%

[a]Summary output from ChipXAnnol.pl and ChipCompare8.pl. Control probesets were not included in the analysis.

Table 2. Summary of annotation results for Affymetrix Rat RG_U34a genome chip.[a]

GenBank[a]		Updated NCBI-based annotation			
accession no.	Original Affymetrix description[b]	XChipID[c]	LocusID[d]	Gene symbols[d]	LocusLink description[d]
AI172064	EST218059 *Rattus norvegicus* cDNA, end/clone=RMUBU47	LL.56646	56646	*Lgals1*	Lectin, galactose 3´ binding, soluble 1
J02810	RATGSTYBX Rat prostate glutathione transferase mRNA, complete cds	LL.24423	24423	*Gstm1*	Glutathione *S*-transferase, mu 1
X04229	RNGSTYBR Rat mRNA for glutathione *S*-transferase (GST) Y(b) subunit	LL.24423			
H32189	EST107045 *Rattus norvegicus* cDNA 5´ end /clone=RPCBK23	LL.24423			
S65355	Nonselective-type endothelin receptor	LL.50672	50672	*Ednrb*	Endothelin receptor type B
X57764	Rat mRNA for ET-B endothelin receptor	LL.50672			
AA818970	UI-R-A0-as-g-05-0-UI.s1 *Rattus norvegicus* cDNA, 3´ end	LL.50672			
U09540	RNU09540 *Rattus norvegicus* Sprague-Dawley cytochrome P450 (CYP1B1) mRNA, complete cds	LL.25426	25426	*Cyp1b1*	Cytochrome P450 1b1
X83867	CYP1B1 *Rattus norvegicus* CYP1B1 mRNA for cytochrome P450	LL.25426			
AI176856	EST220459 *Rattus norvegius* cDNA, 3´ end /clone=ROVBX74	LL.25426			
M14972	Rat cytochrome P-450-LA-omega (lauric acid omega-hydroxylase) mRNA, complete cds	LL.50549	50549	*Cyp4a10*	Cytochrome P450, 4a10
AA924267	UI-R-A1-ds-g-03-0-UI.s1 *Rattus norvegicus* cDNA, 3´ end	LL.50549			
D83538	Rat mRNA for 230 kDa Phosphatidylinositol 4-kinase, complete cds	LL.64161	64161	*Pik4ca*	Phosphatidylinositol 4-kinase
U39572	RNU39572 *Rattus norvegicus* phosphatidylinositol 4-kinase mRNA, complete cds	LL.64161			
J04563	Rat cAMP phosphodiesterase mRNA, 3´ end	LL.24626	24626	*Pde4b*	Phosphodiesterase 4B, cAMP-specific [dunce (*Drosophila*)-homolog phosphodiesterase E4]
M25350	RATPHOCAMB Rat cAMP phosphodiesterase (PDE4) mRNA, partial cds	LL.24626			
X81395	*Rattus norvegicus* mRNA for pI 5.5 esterase (ES-3)	LL.29225	29225	*Ces1*	Carboxylesterase 1
U10697	*Rattus norvegicus* kidney microsomal carboxylesterase mRNA	LL.29225			

[a]http://www.ncbi.nlm.nih.gov/GenBank/. [b]Affymetrix descriptions are those provided with the original chip definition file (RG_U34.GIN). [c]Data represent selected output from XChipData.pl. [d]From LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/).

In fact, the three arrays query only 209 genes in common, and even the Clontech array queries a significant number of genes not queried by the other two arrays. On a case-by-case basis, the results for a given gene on one platform can be compared with those for the same gene on a different platform, using the XChipID (Thompson et al. 2004), taking into account that each platform may query the same gene more than once. It is critical to reiterate, however, that the quality and intensity of the signal from any given microarray element, querying a given gene, will depend on the sequence of that element, preparation of the target hybridization material, and technical aspects of the hybridization and signal processing. Furthermore, comparing platforms based on the XChipIDs depends on these platforms being annotated from the same input UniGene and LocusLink files. When these files are updated, the annotation process must be repeated for all platforms to be compared. Finally, comparing data from one array platform to another on a whole-array level is not a trivial effort, as the redundancy of genes queried on each platform creates what is called in database terminology a "many-to-many" relationship. XChipData.pl was designed to merge such data, and an example of the output from this program is given in Table 4.

Discussion

As microarrays are used more and more to investigate questions of biology and toxicology, a key technical issue becomes more and more problematic: that of associating the signals from each microarray sequence element with the known literature and biological context associated with that sequence. This issue is complicated because element descriptions are current only at the time of array construction and must be updated to reflect evolving information on the gene associated with the element. Such information can include an updated description, a standard gene/locus name (Wain et al. 1999; White et al. 1999), and gene ontology information (Ashburner et al. 2000). Several automated annotation systems have been described, including the DRAGON system (Bouton and Pevsner 2000), the DAVID software (Dennis et al. 2003), and the NetAffx resource specifically for Affymetrix arrays (http://www.affymetrix.com; Liu et al. 2003). Information from this latter resource can be automatically retrieved using the ChipInfo software (http://biosun1.harvard.edu/complab/chipinfo/; Zhong et al. 2003). The XChipAnno script described here differs in that it is designed to create a single best annotation and a XChipID. Although conceptually simple, the XChipID does group elements that, by annotation, should be querying the same gene, and in doing so allows for comparison of data across a microarray, between different versions of a microarray, and between different microarray platforms. This annotation can be carried out on a regular basis as public database information is updated. In addition, this annotation procedure requires only the GenBank accession number for a microarray element, not the actual sequence, and does not require extensive computer resources. The RESOURCERER database (http://pga.tigr.org/tigr-scripts/nhgi_scripts/resourcerer.pl; Tsai et al. 2001) carries out a similar annotation approach using the TIGR Gene Indices and extending this cross-indexing to across species. In contrast to XChipAnno, RESOURCERER focuses on a number of selected common microarray platforms and is accessible by a web interface.

A limitation of this approach, and any approach that groups accession numbers on the basis of UniGene clusters, is that any given build of UniGene may incorrectly cluster certain sequences. Sequence homology can cause closely related but nonidentical genes to cluster together and hence be given the same annotation by this approach. Thus, discordant results for microarray elements having the same annotation (i.e., XChipID) are best resolved by a rigorous BLAST comparison of element sequences with each other and with the target gene sequence. Although a BLAST comparison of each microarray element sequence with the entire sequence database is technically daunting, a simple comparison of such a sequence with a target sequence is quite simple using the LALIGN program (part of the FASTA package; ftp://ftp.virginia.edu/pub/fasta/) (Chao et al. 1992) and could be automated as a quality control check for the annotation of the entire microarray.

Another serious limitation in comparing different microarray platforms is encountered if one array uses sequences from several species, for example, a rat cDNA-based microarray that includes mouse sequences. Although these sequences may hybridize with a rat transcript, annotation by this method is not feasible, as individual species are clustered in UniGene separately. Such cross-species comparisons are desirable but may be best handled by large public database resources that link individual sequences with genomic information (Mattes et al. 2004).

Although any automated procedure to group and annotate DNA sequences is inherently flawed by the absence of human

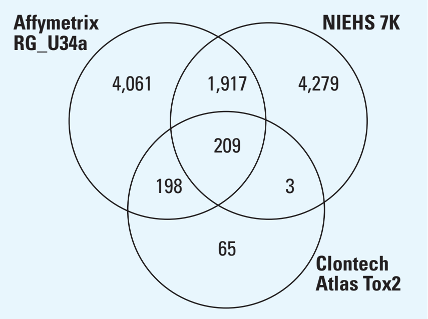

Figure 2. Overlaps of genes queried by different platforms, determined by ChipCompare.pl using the XChipID.

Table 4. Comparison of data from two platforms using the XchipID.[a]

XChipID	Affymetrix ID	Ratio (Affymetrix)	Change	GenBank[b] accession no.	Gene symbol[c]	NIEHS_ID	Ratio (NIEHS)
LL.83783	L19998_g_at	1.36	I	L19998	*Sult1a1*	AA874816	1.30
LL.24791	rc_AA891204_s_at	0.74	D	AA891204	*Sparc*	AA963036	0.81
LL.24791	rc_AA946313_s_at	0.76	D	AA946313	*Sparc*	AA963036	0.81
LL.24791	U75928UTR#1_s_at	0.63	D	U75928	*Sparc*	AA963036	0.81
LL.24791	U75929UTR#1_f_at	0.64	D	U75929	*Sparc*	AA963036	0.81
LL.24791	Y13714_at	0.78	D	Y13714	*Sparc*	AA963036	0.81
LL.171341	J03752_at	1.31	I	J03752	*Mgst1*	AA818422	1.42
LL.299331	rc_AA944397_at	1.77	I	AA944397	*Hsp86*	AA819777	1.80
LL.299331	rc_AI176546_at	1.86	I	AI176546	*Hsp86*	AA819777	1.80
LL.83687	AF093536_at	0.91	D	AF093536	*Defb1*	AA999116	0.85
LL.24854	M64733mRNA_s_at	2.50	I	M64733	*Clu*	AA818413	1.54
LL.113902	L46791_at	1.80	I	L46791	*Ces3*	AA955163	1.42
LL.113902	X65296cds_s_at	2.48	I	X65296	*Ces3*	AA955163	1.42
LL.29144	L18889_at	1.27	I	L18889	*Canx*	AA858850	1.33
LL.29144	rc_AA893328_at	1.98	I	AA893328	*Canx*	AA858850	1.33
LL.29144	rc_AI010725_at	1.41	I	AI010725	*Canx*	AA858850	1.33
LL.64202	D78308_at	1.22	I	D78308	*Calr*	AA859488	1.39
LL.64202	D78308_g_at	1.32	I	D78308	*Calr*	AA859488	1.39
LL.64202	X53363cds_s_at	2.04	I	X53363	***Calr***	AA859488	1.39

Abbreviations: D, decrease; I, increase.
[a]Data represent selected output from XChipData.pl. Both data sets were analyses of RNA pooled from kidneys of rats treated for 7 days with 80 mg/kg/day gentamycin (Kramer et al. 2004). [b](http://www.ncbi.nih.gov/GenBank/). [c]From LocusLink (http://www.ncbi.nih.gov/LocusLink/).

wisdom, such an automated approach is simply required to handle the vast amount of information contained within and generated by microarray technology. The approaches described in this article do help reduce the complexity and redundancy of microarray annotation in a straightforward fashion. The files required by this approach are readily available, and the output files generated may be directly used and manipulated with a variety of software packages such as Excel, Access, or Spotfire. Although microarray results are always best considered on a sequence-by-sequence basis, global annotation procedures can offer a way to provide an initial sift and analysis of the data with biological context.

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29.

Bouton CM, Pevsner J. 2000. DRAGON: Database referencing of array genes online. Bioinformatics 16:1038–1039.

Chao KM, Pearson WR, Miller W. 1992. Aligning two sequences within a specified diagonal band. Comput Appl Biosci 8:481–487.

Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4:P3.

Kokame K, Kato H, Miyata T. 1996. Homocysteine-respondent genes in vascular endothelial cells identified by differential display analysis. GRP78/BiP and novel genes. J Biol Chem 271:29695–29665.

Kramer JA, Pettit SD, Amin RP, Bertram TA, Car BD, Cunningham M, et al. 2004. Overview of the application of transcription profiling using selected nephrotoxicants for toxicology assessment. Environ Health Perspect 112:495–505.

Liu G, Loraine AE, Shigeta R, Cline M, Cheng J, Valmeekam V, et al. 2003. NetAffx: Affymetrix probesets and annotations. Nucleic Acids Res 31:82–86.

Mattes WB, Pettit SD, Sansone S-A,Bushel PR, Waters M, et al. 2004. Database development in toxicogenomics: issues and efforts. Environ Health Perspect 112:495–505.

Pruitt KD, Maglott DR. 2001. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 29:137–140.

Thompson KL, Afshari CA, Amin R, Bertram TA, Car B, Cunningham M, et al. 2004. Identification of platform-independent gene expression markers of cisplatin nephrotoxicity. Environ Health Perspect 112:488–494

Tsai J, Sulatan R, Lee Y, Pertea G, Karamycheva S, Antonescu V, et al. 2001. Resourcerer: a database for annotating and linking microarray resources within and across species. Genome Biol 2:1–4.

Van BN, Dinjens WN, Diesveld MP, Groen NA, Van der Made AC, Nozawa Y, et al. 1997. A novel gene which is up-regulated during colon epithelial cell differentiation and down-regulated in colorectal neoplasms. Lab Invest 77:85–92.

Wain H, White J, Povey S. 1999. The changing challenges of nomenclature. Cytogenet Cell Genet 86:162–164.

Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, et al. 2000. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 28:10–14.

White J, Wain H, Bruford E, Povey S. 1999. Promoting a standard nomenclature for genes and proteins. Nature 402:347.

Zhong S, Li C, Wong WH. 2003. ChipInfo: software for extracting gene annotation and gene ontology information for microarray analysis. Nucleic Acids Res 31:3483–3486.

Zhou D, Salnikow K, Costa M. 1998. *Cap43*, a novel gene specifically induced by Ni^{2+} compounds. Cancer Res 58:2182–2189.