



ARROGANT: an application to manipulate large gene collections

Amit V. Kulkarni¹, Noelle Sevilir Williams⁷, Yun Lian^{5, 6}, Jonathan D. Wren², David Mittelman^{5, 6}, Alexander Pertsemlidis^{4, 5} and Harold R. Garner^{3, 4, 5, 6,*}

¹Program in Biomedical Engineering, ²Program in Genetics and Development, Southwestern Graduate School of Biomedical Science, ³Department of Biochemistry, ⁴Department of Internal Medicine, ⁵McDermott Center for Human Growth and Development, ⁶Center for Biomedical Inventions and ⁷Simmons Comprehensive Cancer Center, The University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390, USA

Received on February 2, 2002; revised on April 10, 2002; accepted on April 20, 2002

ABSTRACT

Summary: ARROGANT (ARRay OrGANizing Tool) is a software tool developed to facilitate the identification, annotation and comparison of large collections of genes or clones. The objective is to enable users to compile gene/clone collections from different databases, allowing them to design experiments and analyze the collections as well as associated experimental data efficiently. ARROGANT can relate different sequence identifiers to their common reference sequence using the UniGene database, allowing for the comparison of data from two different microarray experiments. ARROGANT has been successfully used to analyze microarray expression data for colon cancer, to compile genes potentially related to cardiac diseases for subsequent resequencing (to identify single nucleotide polymorphisms, SNPs), to design a new comprehensive human cDNA microarray for cancer, to combine and compare expression data generated by different microarrays and to provide annotation for genes on custom and Affymetrix chips.

Availability: ARROGANT is freely available for use on the web at <http://lethargy.swmed.edu/>

Contact: harold.garner@utsouthwestern.edu

INTRODUCTION

With the 'completion' of the Human Genome Project, researchers are now applying the resulting sequence, mapping and annotation data towards the understanding of diseases which are due to altered regulation or sequence of genes belonging to one or more pathways, such as cancer and other multigenic diseases. With the recent revolution in gene expression and genotyping technologies, it is now

possible to conduct experiments with thousands of genes simultaneously. This increase in the data production rate and the development of methods to assess relationships among genes in a single experiment mean that we must be able to simultaneously study groups of genes as well as single genes. However, since the (functional) annotation necessary to analyze the results of such parallel experiments is in many cases unavailable, incomplete, or distributed across numerous databases, our ability to see the big picture is limited. Software tools to design these experiments and then analyze the results based on integrating data from various sources are therefore very important to our ability to extract knowledge from these enormous amounts of data. For example, clustering and visualization techniques for analyzing microarray results have proven invaluable in identifying patterns in expression data (Dysvik and Jonassen, 2001; Sturn *et al.*, 2002). Some prominent packages include GeneSpring, Cluster Treeview (Eisen *et al.*, 1998) and Spotfire (<http://www.spotfire.com>).

Of the approximately 35 000 human genes, many have more than one splice variant, leading to more than 100 000 proteins (Lander *et al.*, 2001). With the current technology, it is not practical to include all genes and potential variants on a single microarray experiment—indeed most research groups work with custom application-specific arrays or large, but incomplete sets of genes/clones. The selection of an optimal subset of genes is therefore important in the design of microarray experiments. Thus, there is a need for a tool to facilitate the identification, analysis, and comparison of gene or clone collections.

There have been several efforts to develop integrated analysis engines for data generated from microarray experiments. All of these efforts depend on the databases and

*To whom correspondence should be addressed.

tables produced and maintained by a number of groups, especially the National Center for Biotechnology Information. The NCBI website (<http://www.ncbi.nlm.nih.gov>) provides search engines for the various databases it curates, including GenBank (Benson *et al.*, 2000), UniGene (Schuler, 1997), and LocusLink (Pruitt and Maglott, 2001). While it is possible to use these search engines to generate a list of potential gene candidates for study, individual keyword searches are required for each database or dataset. The results must then be combined and more importantly, the redundancies must be eliminated. So, while it is possible to assemble and design large gene collections manually, it is inconvenient. While LocusLink can be used to annotate a gene collection in batch retrieval mode, the annotation cannot be accessed online and the researcher cannot include additional information, for example, experimental data, to be associated with such a collection.

There are other tools for working with expression data such as DRAGON (Bouton and Pevsner, 2000) (<http://pevsnerlab.kennedykrieger.org/dragon.htm>). However, DRAGON has limitations: it searches only one database at a time, has a limited ability to manipulate results, and is unable to incorporate additional information from the user. Furthermore, DRAGON does not include some widely used databases such as GenBank (useful to study/identify non-transcribed as well as transcribed regions) and LocusLink. Other microarray data mining tools include the yeast Microarray Global Viewer (yMGV), which only includes published expression data from yeast microarrays (Marc *et al.*, 2001). There are a number of other microarray and gene analysis utilities, including SOURCE (Stanford Online Universal Resources for Clones & ESTs, <http://genome-www5.stanford.edu/cgi-bin/SMD/source/sourceSearch>), unCHIP (<http://unchip.org:8080/bio/unchip>), RESOURCERER (Tsai *et al.*, 2001) and Allgenes (<http://www.allgenes.org>). Each of these resources has unique capabilities, but is limited to specific functions which are packaged together in ARROGANT. ARROGANT provides in a single implementation wide annotation from multiple databases; flexibility in how it designs, compares, annotates and merges its collections and the ability to integrate experimental data with that annotation.

Finally, the interpretation of data can be more efficient and thorough in an environment where important biomedical information is collected, visualized and manipulated for all genes simultaneously, and facilitated by searching and sorting capabilities.

ARROGANT, as detailed below, was developed to address all of these needs. ARROGANT is a general-purpose tool for designing and analyzing experiments involving many genes/clones such as those from expression microarrays or DNA resequencing efforts for variant single nucleotide polymorphisms (SNP) discovery.

IMPLEMENTATION

Computational tools

ARROGANT is distributed over three separate components: a web server, a database server, and a compute server. The servers access a shared 3-terabyte RAID-5 storage system containing copies of several databases which are updated monthly in a semi-automated manner. The web server is used to receive input and display output results to the researcher. The database server hosts several Microsoft SQL Server 7.0 databases. Various compute-intensive programs, such as BLAST (Altschul *et al.*, 1990) for sequence comparison, and PRIMO (Li *et al.*, 1997) for primer design, are implemented on a four-processor Hewlett-Packard L2000.

Program organization

ARROGANT is a database driven tool. ARROGANT utilizes NCBI databases (GenBank, UniGene, LocusLink, HomoloGene), KEGG (Kanehisa and Goto, 2000), Genome Ontology Consortium data (Ashburner *et al.*, 2000), Research Genetics clone data and other custom databases (Rep-X (Wren and Foracs, 2000)). The local implementation of various databases and tools makes ARROGANT independent of server-based network connections and significantly improves its performance and reliability. ARROGANT has three modes of operation as shown in Figure 1 design mode, analysis mode and merge mode.

Design mode

In the design mode, ARROGANT can be used to compile a large collection of genes from different databases with a single keyword-based query. Each retrieved candidate gene/clone is hyperlinked to its annotation and flagged for inclusion in or exclusion from the collection. ARROGANT allows researchers to add GenBank accession numbers, LocusLink IDs or UniGene IDs to be included in the final collection. ARROGANT assists in the design of cDNA microarrays by creating FASTA files of input sequences from lists of commercially available clones and/or by designing primers for PCR product or cDNA clone microarrays. Primer design parameters can be set by the researcher. For oligonucleotide microarrays, FASTA files are created and probes are designed using the PRIMO oligonucleotide design code (Li *et al.*, 1997) developed in our laboratory. The output files are sent to the user by email.

Analysis mode

The function of the analysis mode is to allow the researcher to annotate a gene/clone collection and associate it with experimental data. All or selected portions of the data are displayed in a searchable and sortable tabular form to enable the researcher to organize it and make

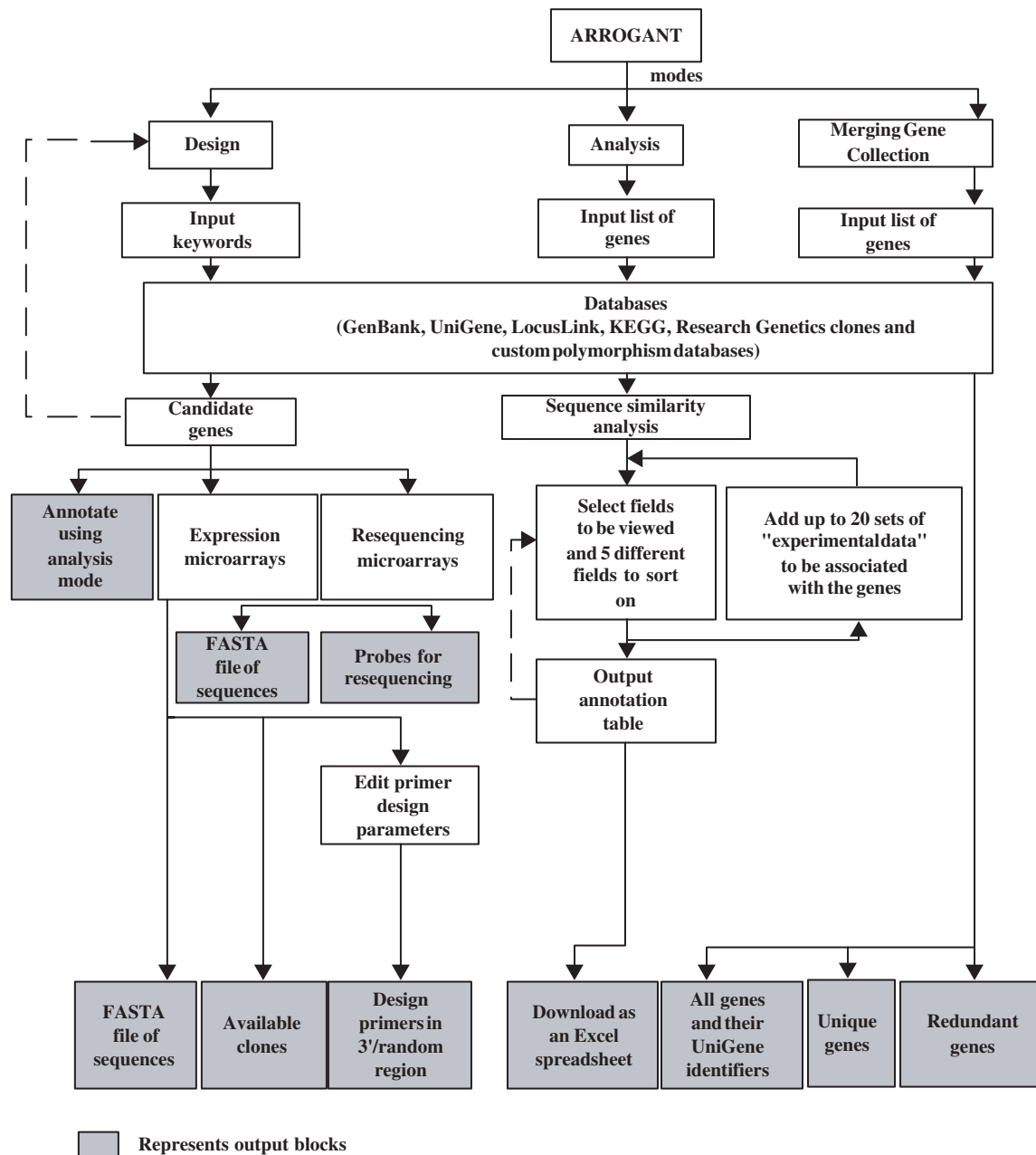


Fig. 1. Flow chart of events of ARROGANT in all the three modes of operation. The user interface, input, viewing and interacting with the results, is through the world wide web using a standard browser. Many of the results files, including text and hypertext files, are emailed to the user, and other results are downloadable.

biologically relevant observations. The input consists of a tab-delimited list of GenBank accession numbers, LocusLink IDs or UniGene IDs followed by up to 50 numeric fields, which correspond to experimental data, such as expression ratios from a microarray experiment, or the number of SNPs found in a resequencing experiment. Five levels of sort (and numerical summaries of

sort statistics) are provided to help the researcher group information of interest by various criteria (see Figure 2, panels a and b). The analysis mode of ARROGANT can help answer questions such as, 'Which of the genes in the collection are located on chromosome 3, are up-regulated by a factor of 3, have potentially polymorphic repeats, and also have homologs in mouse which could be

Please select the fields to be viewed by clicking its check box. To select all analysis fields press [Select All](#). Each field is a hot link to help. For example, to know more about Gene Name press [Gene Name](#).

<input type="checkbox"/> Accession	<input type="checkbox"/> IMAGE identifier	<input type="checkbox"/> Alias protein
<input type="checkbox"/> Experimental data	<input type="checkbox"/> Sequence Similarity	<input type="checkbox"/> Phenotype
<input type="checkbox"/> Unique identifier	<input type="checkbox"/> Tissues	<input type="checkbox"/> Phenotype ID/ OMIM ID
<input type="checkbox"/> Title	<input type="checkbox"/> GDB	<input type="checkbox"/> Chromosome
<input type="checkbox"/> PubMed	<input type="checkbox"/> Gene name	<input type="checkbox"/> Map location
<input type="checkbox"/> Related Proteins	<input type="checkbox"/> Gene Function	<input type="checkbox"/> Map link
<input type="checkbox"/> Related Sequences	<input type="checkbox"/> Synonyms	<input type="checkbox"/> Map type
<input type="checkbox"/> Taxonomy	<input type="checkbox"/> Pathway	<input type="checkbox"/> STS- marker name
<input type="checkbox"/> Repeats	<input type="checkbox"/> SNP analysis	<input type="checkbox"/> STS- chromosome
<input type="checkbox"/> Unigene ID	<input type="checkbox"/> STS- ID	<input type="checkbox"/> Homology
<input type="checkbox"/> Locus ID	<input type="checkbox"/> RefSeq	<input type="checkbox"/> Related GO#
<input type="checkbox"/> Research Genetics Clone	<input type="checkbox"/> Alias symbol	<input type="checkbox"/> Molecular Function
<input type="checkbox"/> Biological Process	<input type="checkbox"/> Cellular Component	<input type="checkbox"/> Summary Function

Sort Preferences: Please choose the 5 levels to sort and sub-sort the data. No Sort

1. 2. 3.
 4. 5.

If you are done selecting the fields, please view the output annotated table by pressing [Show Collection](#)

Analysis With User Supplied Data

Include sets of numbers (which could represent expression ratios for microarrays, purity of repeats for polymorphism studies etc.) to be associated with the annotated gene collection

[Add Experimental Data](#)

(a)

Results Pages: [Search](#)

Record	Accession	Cluster ID	PubMed	Gene Name	Molecular Function
1	NM_009512	Mm.10984	9671728	Slc27a5, solute carrier family 27 (fatty acid transporter), member 5, Slc27a5, FATP5, Vlacsr	-
2	NM_004960	Hs.99969	8510758	FUS, fusion, derived from t (12;16) malignant liposarcoma, FUS, TLS	RNA binding
3	NM_000236	Hs.9994	10660332	LIPC, lipase, hepatic, LIPC, HL	-
4	NM_000023	Hs.99931	8069911	SGCA, sarcoglycan, alpha (50kD dystrophin-associated glycoprotein), SGCA, A2, ADL, DAG2, LGMD2D, SCARM1	-
5	NM_001807	Hs.99918	1988041	CEL, carboxyl ester lipase (bile salt-stimulated lipase), CEL, BAL, BSSL	-
6	NM_000684	Hs.99913	10212248	ADRB1, adrenergic, beta-1-, receptor, ADRB1, B1AR, ADRB1R	beta adrenergic receptor;beta ₁ -adrenergic receptor
7	NM_005036	Hs.998	10860941	PPARA, peroxisome proliferative activated receptor, alpha, PPARA, PPAR, HPPAR, NR1C1	receptor;DNA binding;peroxisome receptor;transcription factor
8	NM_005194	Hs.99029	10821850	CEBPB, CCAAT, enhancer binding protein (C, EBP), beta, CEBPB, LAP, CRP2, TCF5, NFIL6, IL6DBP	DNA binding;transcription factor;transcription activating factor
9	NM_005544	Hs.96063	8513971	IRS1, insulin receptor substrate 1, IRS1, HIRS-1	signal transduction;transmembrane receptor protein tyrosine kinase docking protein

(b)

Fig. 2. (a) After collection annotation in the analysis mode, the user selects which data to view and can also choose to be prompted for local numerical data entry. The researcher can select the fields to be displayed and sorted on the output table. Each of the fields is hot linked to help, a description of the source of the data and how it was prepared for display. (b) A portion of the output table for a given collection. Some fields are hyperlinked to the primary data source for additional detail. The search function enables a text search on specified columns of data to quickly find items of interest within the displayed data set

used for knockout experiments?’ ARROGANT also has a search function to identify keywords appearing within the generated annotation in its tables.

Merge mode

The merge mode functions to integrate data from different gene collections, eliminate redundancies and identify similarities and differences in member content of the collections. The comparison is done by first converting the input accession numbers, LocusLink IDs or UniGene IDs to their corresponding UniGene (Schuler, 1997) identifiers. Microarrays produced by different groups or companies contain their own unique clones with different accession numbers that represent fragments of the same gene. The clones associations can be resolved via a common UniGene cluster ID. Expression data from different microarrays can then be compared using the unique UniGene cluster identifiers. Once the different identifiers have been mapped to their common UniGene IDs, the list of UniGene identifiers is then sorted and duplicates are eliminated. The UniGene identifiers are converted back to accession numbers that represent the longest sequence for the UniGene cluster, providing only one accession number per UniGene cluster. The merge function works within the analysis mode to allow researchers to gather experimental data from different microarray designs and display the results together.

Results and Discussion

ARROGANT has been used in a number of different applications, from microarray design and analysis to candidate gene selection. Some of those applications are detailed below.

ARROGANT was used to develop a new human microarray for cancer studies

There was a critical need to develop a ‘comprehensive’ human cancer microarray, an array that contains every commercially available clone whose gene or expressed sequence tag (EST) has been implicated in cancer. ARROGANT’s ability to look for candidate sequences based on keywords, integrate clone IDs from different sources and eliminate duplicate clones greatly simplified the task of adding genes to the existing University of Texas Southwestern Medical Center (UTSW) human 10 000 clone microarray (<http://microarray.swmed.edu>). For ‘comprehensive’ coverage, UTSW cancer researchers desired the inclusion of the collection of 13 969 genes implicated in cancer by the Cancer Genome Anatomy Program (CGAP) (<http://cgap.nci.nih.gov>) and approximately 800 genes compiled by UTSW investigators. The design mode of ARROGANT suggested an additional 237 genes based on keywords (e.g. cancer, metastasis, etc.) compiled by faculty researchers. The merge mode was used to com-

bine these three collections. The combined collection was then compared with the existing members of our current human cDNA microarray, which determined that 9315 genes (clones) needed to be added to our existing clone collection. The analysis mode found commercially available Research Genetics, Inc clones for 7593 of these genes. The resulting array is in wide use, and details on the array can be found on our core facility web site (<http://microarray.swmed.edu>). Expansion of the array demonstrated the utility of ARROGANT in developing a new gene collection using all three modes of operation.

ARROGANT was used to study colon cancer by analyzing multiple microarray experiments

Previous to the creation of the ‘comprehensive’ cancer array, the UTSW human 10 000 clone microarray was used for colon cancer studies. Expression level data obtained from five experiments were analyzed using the ARROGANT analysis mode. mRNA samples were extracted from normal and cancerous tissue of four colon cancer patients and one patient having familial adenomatous polyposis (FAP), a precursor to colon cancer (Griffioen *et al.*, 1998). The final annotation table with its experimental data (http://lethargy.swmed.edu/hideandsort.asp?txt_array=60110) was initially sorted based on the expression level data. The threshold for flagging up-regulated or down-regulated genes was that their expression ratios were 2 or 2.5 times greater, respectively, in at least three of five colon cancer samples relative to their matched normal samples. The intent was to select a few genes whose regulation was greatly altered and then inspect other genes in the affected pathways with the help of other annotation. Any gene found to be up regulated in three of the four colon tumors, but unaffected in the polyp, was then potentially implicated in the final stages of tumor development. By these criteria, 28 up-regulated genes and 69 down-regulated genes appeared to be related to cancer. For example, AA680186 was down regulated in all five of the microarray datasets. The annotation provided by ARROGANT revealed that this gene was a chemotactic factor for lymphocytes, leading to the hypothesis that fewer lymphocytes are recruited to tumor sites allowing for growth of the tumor unchecked by the immune system.

With the help of annotation from ARROGANT, these preliminary studies confirmed the validity of this approach to identifying genes relevant to the growth of cancerous cells. These studies have now been expanded to 20 patients and genes up- or down-regulated by two-fold or more in one third of the patients were selected as indicated above. Expression levels of ten of these genes have been confirmed by quantitative real-time PCR. ARROGANT provided critical information that enabled fast and efficient analysis to select a subset of genes to be inspected further in follow-up experiments.

ARROGANT was used for identifying and annotating genes in polymorphism studies

ARROGANT was used in research being conducted as part of the NIH Program for Genomic Applications (PGA, <http://www.nhlbi.nih.gov/resources/pga/index.htm>) to identify and provide annotation for human genes putatively associated with cardiac disease and inflammation. In this project, 750 genes suspected to play a role in cardiac disease are being identified for subsequent genotype/phenotype association studies. The selection of candidate genes was done in several ways: as of January 1, 2002, 253 genes were chosen by faculty researchers in their areas of expertise; 100 human homologs of mouse genes were chosen from mouse microarray experiments on mouse models of cardiac disease; and additional genes were chosen using ARROGANT. The design mode of ARROGANT was used to identify 578 candidate genes. These candidate genes were then categorized based on associated keywords provided by expert researchers (e.g. cholesterol, inflammation, G-protein, etc.) and annotated using the analysis mode.

Overall, ARROGANT facilitated the selection process by providing sufficient relevant integrated biomedical information to make the selections. Of the 26 new genes identified using the keyword cholesterol and not included in other earlier lists, nine genes were selected for inclusion in the study. For example, the MEDLINE entry retrieved by ARROGANT for one of the nine genes chosen, NPC1, revealed that it was associated with the disease, Niemann-Pick type II, characterized by a defect in intracellular trafficking of sterols (Greer *et al.*, 1999), thus justifying its inclusion.

ARROGANT was used to annotate large gene collections

ARROGANT has been used to pre-compute the annotation for our local gene collections and microarrays and to annotate various Affymetrix (<http://www.affymetrix.com>) GeneChip gene collections. These pre-computed annotated collections are freely available at <http://lethargy.swmed.edu/precomputed.asp>. ARROGANT has been used by researchers to further annotate and analyze their Affymetrix data. While the suite of software tools from Affymetrix provides various graphical ways of analyzing GeneChip data, these tools do not provide an integrated view of gene annotations that some researchers desire.

ARROGANT was used to compare expression data from microarrays of different designs

Cluster analysis of gene expression data produced at the Stanford Microarray Center identified distinct types of diffuse large B-cell lymphoma. (Alizadeh *et al.*, 2000) (<http://llmmp.nih.gov/lymphoma/data/clones.txt>). As a demonstration, this dataset (representing 17 125 clones) was

merged with colon cancer studies using the UTSW human 10 000 clone microarray. The output consisted of three separate lists, one containing the genes common to the collections followed by genes unique to each collection. There were 4784 genes on the Stanford microarray that were also represented on the UTSW 10 000 clone microarray. There were 6684 genes unique to the UTSW microarray and 8728 genes unique to the Stanford microarray. A portion of the lists is shown in Table 1. As seen in Table 1(a), H89664 and H89517 are EST clones of an unknown genes, which are similar to amyloid-like protein 2 precursor. Amyloid proteins are thought to play a role in various cancers such as breast, colon, and prostate cancers (Makimattila *et al.*, 2001). As seen in Table 1(b), H94720 (poly ADP-ribose glycohydrolase) was found on the UTSW microarray but not on the Stanford microarray. This gene is known to be involved in the development of lymphoma (<http://www.ich.ucl.ac.uk/cmgs/apoptos.htm>) and would be a useful addition to the Stanford microarray (Park *et al.*, 2002). Similarly in Table 1(c), M77477 (human aldehyde dehydrogenase) is known to play a role in lung cancer and would be a useful addition in the UTSW human cancer microarray.

CONCLUSION

ARROGANT was successfully used to design a 'comprehensive' human cDNA microarray for cancer studies. The design mode was used to identify genes potentially involved in cardiac diseases and other similar applications. The analysis mode was used to annotate microarray experiments for colon cancer studies. ARROGANT has also been successfully used to annotate several gene collections including genes on Affymetrix and UTSW microarrays, and genes thought to be involved in cardiac disease as found using ARROGANT's design mode. The merge mode was successfully demonstrated by comparing data from a Stanford microarray experiment (Alizadeh *et al.*, 2000) with a UTSW microarray experimental data. The comparison suggested genes that could be added to both microarrays by finding clones unique to each collection. This feature enables the analysis of data from many types of microarray designs, solving a current difficulty in expression data sharing.

ACKNOWLEDGEMENTS

This work was funded by NIH/NCI (5R33CA8165603), NIH/NCI SPORE (50CA70907) and NIH/NHLBI PGA (5U01HL6688002) and the P.O'B. Montgomery distinguished chair. We wish to thank John Fondon III, Dr Helen Hobbs and Dr Jonathan Cohen for important feedback. We would like to thank Affymetrix, Inc for providing accession numbers for the sequences on their microarrays.

Table 1. A portion of the displayed output obtained from the comparison of the UTSW microarray clone list with the Stanford microarray clone list, by GenBank accession number (accsno). Note that experimental expression ratios are also displayed and properly associated with a given gene although the particular clone may be different in each array. (a) Clone members common to both. (b) Those unique to the UTSW microarray. (c) Those unique to the Stanford microarray

(a)

UTSW accsno	Stanford accsno	UniGene ID	Description	Ratio 1	Ratio 2
AA676804	AA179189	Hs.2227	CCAAT/enhancer binding protein (C/EBP), gamma	6.91	3.27
H89664	H89517	Hs.279518	Amyloid beta (A4) precursor-like protein 2	6.53	6.92
AA700005	R02721	Hs.19413	calcium-binding protein A12 (calgranulin C)	5.71	3.56
AA419177	N28006	Hs.184601	SLC7A5 Solute carrier family 7, member 5	3.21	4.24

(b)

UTSW accsno	Stanford accsno	UniGene ID	Description	Ratio 1	Ratio 2
AA115919	-	Hs.278589	General transcription factor II, i	29.75	-
AA007699	-	Hs.75790	PIGC Phosphatidylinositol glycan, class C	14.25	-
H94720	-	Hs.91390	PARG Poly (ADP-ribose) glycohydrolase	8.55	-
N24004	-	Hs.271353	MutY (E. coli) homolog	7.42	-

(c)

UTSW accsno	Stanford accsno	UniGene ID	Description	Ratio 1	Ratio 2
-	M77477	Hs.575	Aldehyde dehydrogenase 3 family, member A1	-	44.88
-	AA102113	Hs.227730	ITGA6 Integrin, alpha 6	-	21.34
-	M19645	Hs.75410	Heat shock 70kD protein 5	-	17.69
-	T39516	Hs.75517	Laminin, beta 3	-	13.75

REFERENCES

- Alizadeh, A. Eisen, M. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.
- Altschul, S.F. and Warren, G. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
- Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
- Benson, D.A. Karsch-Mizrachi, I. *et al.* (2000) GenBank. *Nucleic Acids Res.*, **28**, 15–18.
- Bottner, M. Laaff, M. *et al.* (1999) Characterization of the rat, mouse, and human genes of growth/differentiation factor-15/macrophage inhibiting cytokine-1 (GDF-15/MIC-1). *Gene*, **237**, 105–111.
- Bouton, C.M. and Pevsner, J. (2000) DRAGON: Database Referencing of Array Genes Online. *Bioinformatics*, **16**, 1038–1039.
- Dysvik, B. and Jonassen, I. (2001) J-Express: Exploring Gene Expression Data using Java. *Bioinformatics*, **17**, 369–370.
- Eisen, M.B. Spellman, P.T. *et al.* (1998) Cluster Analysis and Display of Genome-Wide Expression Patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
- Lander, E.S. Linton, L.M. *et al.* (2001) Initial sequencing and analysis of the human genome. The genome international sequencing consortium. *Nature*, **409**, 860–921.
- Greer, W.L. Dobson, M.J. *et al.* (1999) Mutations in NPC1 highlight a conserved NPC1-specific cysteine-rich domain. *Am. J. Hum. Genet.*, **65**, 1252–1260.
- Griffioen, G. Bus, P.J. *et al.* (1998) Extracolonic manifestations of familial adenomatous polyposis: desmoid tumours, and upper gastrointestinal adenomas and carcinomas. *Scand. J. Gastroenterol Suppl.*, **225**, 85–91.
- Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
- Ko, C.Y. Rusin, L.C. *et al.* (2001) Long-term outcomes of the ileal pouch anal anastomosis: the association of bowel function and quality of life 5 years after surgery. *J. Surg. Res.*, **98**, 102–107.
- Le Crom, S. Devaux, F. *et al.* (2002) yMGV: helping biologists with yeast microarray data mining. *Nucleic Acids Res.*, **30**, 76–79.
- Li, P. Kupfer, K.C. *et al.* (1997) *Genomics*, **40**, 476–485.
- Makimattila, S. Hietaniemi, K. *et al.* (2001) In vivo glucose-stimulated amylin secretion is increased in nondiabetic patients with pancreatic cancer. *Metabolism Sep.*, **50**, 1036–1042.
- Marc, P. Devaux, F. *et al.* (2001) yMGV: a database for visualization and data mining of published genome-wide yeast expression data. *Nucleic Acids Res.*, **29**, E63–3.

- Park,K.S. Cho,S.Y. *et al.* (2002) Proteomic alterations of the variants of human aldehyde dehydrogenase isozymes correlate with hepatocellular carcinoma. *Int. J. Cancer*, **97**, 261–265.
- Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
- Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
- Sturn,A., Quackenbush,J. and Trajanoski,Z. (2002) Genesis: cluster analysis of microarray data. *Bioinformatics*, **18**, 207–208.
- Tsai,J., Sultana,R. *et al.* (2001) RESOURCERER: a database for annotating and linking microarray resources within and across species. *Genome Biol.*, **2**, software 0002.1–0002.4.
- Walker,N.J. (2001) Real-time and quantitative PCR: applications to mechanism-based toxicology. *J. Biochem. Mol. Toxicol.*, **15**, 121–127.
- Wren,J. and Foracs,E. (2000) Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. *Am. J. Hum. Genet.*, **67**, 345–356.