Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W441–W444***
DOI: 10.1093/nar/gkh397

KARMA: a web server application for comparing and annotating heterogeneous microarray platforms

Kei-Hoi Cheung[1,2,*], Janet Hager[3,4], Deyun Pan[1], Ranjana Srivastava[6], Shrikant Mane[3,4], Yuli Li[1], Perry Miller[1,5] and Kenneth R. Williams[3,4]

[1]Center for Medical Informatics, Department of Anesthesiology, [2]Department of Genetics, [3]W. M. Keck Biotechnology Resource Laboratory, [4]Department of Molecular Biophysics and Biochemistry, [5]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT, USA and [6]Celera, Rockville, MD, USA

Received February 14, 2004; Revised and Accepted March 24, 2004

ABSTRACT

We have developed a universal web server application (KARMA) that allows comparison and annotation of user-defined pairs of microarray platforms based on diverse types of genome annotation data (across different species) collected from multiple sources. The application is an effective tool for diverse microarray platforms, including arrays that are provided by (i) the Keck Microarray Resource at Yale, (ii) commercially available Affymetrix GeneChips® and spotted arrays and (iii) custom arrays made by individual academics. The tool provides a web interface that allows users to input pairs of test files that represent diverse array platforms for either single or multiple species. The program dynamically identifies analogous DNA fragments spotted or synthesized on multiple microarray platforms based on the following types of information: (i) NCBI-Unigene identifiers, if the platforms being compared are within the same species or (ii) NCBI-Homologene data, if they are cross-species. The single-species comparison is implemented based on set operations: intersection, union and difference. Other forms of retrievable annotation data, including LocusLink, SwissProt and Gene Ontology (GO), are collected from multiple remote sites and stored in an integrated fashion using an Oracle database. The KARMA database, which is updated periodically, is available on line at the following URL: http://ymd.med.yale.edu/karma/cgi-bin/karma.pl.

INTRODUCTION

As microarray technology has become increasingly used by academic researchers worldwide, the plethora of both commercial and academic arrays available to the research community has led to a need for a 'universal' tool for annotation and comparison of array platform content. This need arises both at the inception stage of microarray experimental design and subsequently, after array choice and experimental implementation, for annotation and mining of final gene expression data subsets. Additionally, annotation of gene lists is an integral part of preparing platform descriptions for web-posting of published data via NCBI's Gene Expression Omnibus (GEO) (1).

With advances in DNA sequencing technologies, the genomes of many organisms (including the human genome) have been sequenced completely. Many experimental and computational approaches have been used to extract functional information encoded in these complete sequences. As a result, a large amount and a wide variety of genome annotation data have been generated. Such annotation data are available through numerous (web-accessible) databases, including GenBank (2), LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/), Unigene (http://www.ncbi.nlm.nih.gov/), Gene Ontology (GO) (3) and SwissProt (4).

Genome-wide data analysis (e.g. microarray gene expression analysis) typically requires integration of diverse types of annotation data. Integrating these data and keeping them up to date is a significant informatics effort since these datasets evolve rapidly and are available in heterogeneous formats (including text, XML and database formats). Efforts including SOURCE (5), DRAGON (6), and Unchip (http://unchip.org:8080/bio/unchip) have been underway to provide users with timely access to such integrated data. In general, they provide a web interface for users to supply a list of gene identifiers (e.g. GenBank accession numbers), and the interface will return annotation data for each gene. We have created an integrated gene annotation database (KARMA—Keck ARray Manager and Annotator) and an associated web interface which allow users to compare and annotate their own array platforms (or simple lists of gene identifiers) as well

*To whom correspondence should be addressed. Tel: +1 203 737 5783; Fax: +1 203 737 5708; Email: kei.cheung@yale.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

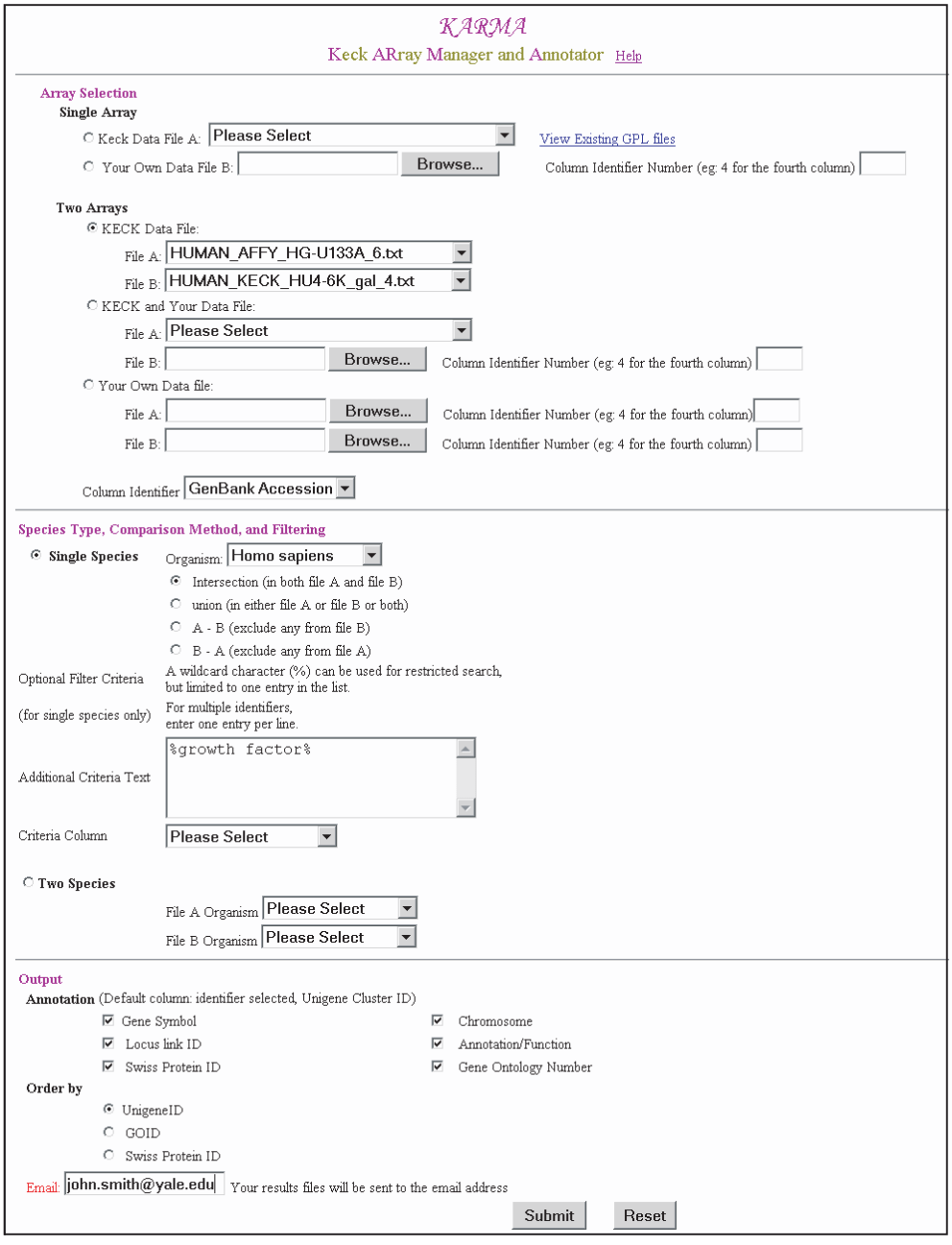

Figure 1. The KARMA web interface through which datasets can be selected or uploaded for comparison and annotation.

as commonly used platforms made available through the Keck Microarray and Affymetrix Resources at Yale (http://keck.med.yale.edu/dna_arrays.htm).

The multiplicity of microarray platforms—manufactured by commercial vendors (e.g. Affymetrix GeneChips®, Agilent, MWG and ClonTech) or spotted by individual laboratories at different universities or research institutes using commercially available material (e.g. Operon, Compugen, MWG and Research Genetics) or academically generated targets— necessitates comparison and annotation of these platforms. By doing so, researchers are able to choose the microarray platform that best matches their needs. Resources including DRAGON, SOURCE and Unchip allow the user to submit a set of the identifiers (e.g. GenBank accession numbers) of the DNA elements laid down on chips/arrays and then return the latest annotation describing those elements. However, to perform comparison between chips/arrays based on such annotation, the user has to develop their own individual method. ARROGANT (7) allows Unigene-based comparison of two gene collections corresponding to two different microarray platforms, but it does not support cross-species comparison. RESOURCERER (8) allows comparison of different platforms involving a single or multiple species, but it can only be applied to a set of commonly used array platforms and not to a unique user-uploaded platform. In other words, it does not support comparison of custom array platforms. To address these issues, KARMA was designed to allow comparisons to be made across diverse microarray platforms

Count	HUMAN_AFFY_HG-U133A_6.txt	HUMAN_KECK_HU4-6K_gal_4.txt	clusterid	Gene_symbol	chromosome	Locuslink	Annotation/Function	Swissid	GOids	biological_process	cellular_component	molecular_function
1	NM_000596	AA233079	Hs.102122	IGFBP1	7	3484	insulin-like growth factor binding protein 1	P08833 P47876 P21743	GO:0001558 GO:0005576 GO:0005615 GO:0019838 GO:0007165 GO:0005520	GO:0001558~regulation of cell growth; GO:0007165~signal transduction;	GO:0005576~extracellular; GO:0005615~extracellular space;	GO:0019838~growth factor binding; GO:0005520~insulin-like growth factor binding;
2	BC000125 NM_000660	R36467	Hs.1103	TGFB1	19	7040	transforming growth factor, beta 1 (Camurati-Engelmann disease)	P04202	GO:0000074 GO:0005160 GO:0008283 GO:0008083 GO:0016049	GO:0000074~regulation of cell cycle; GO:0008283~cell proliferation; GO:0016049~cell growth;		GO:0005160~transforming growth factor beta receptor binding; GO:0008083~growth factor activity;
3	NM_000358	H94586 R31321	Hs.118787	TGFBI	5	7045	transforming growth factor, beta-induced, 68kD					
4	M58051 NM_000142	AA417654	Hs.1420	FGFR3	4	2261	fibroblast growth factor receptor 3 (achondroplasia, thanatophoric dwarfism)	P22607 Q07407	GO:0000165 GO:0001501 GO:0016740 GO:0016021 GO:0008543 GO:0007259 GO:0007048 GO:0006468 GO:0005887 GO:0005524 GO:0005007 GO:0004872 GO:0004713 GO:0004672	GO:0000165~MAPKKK cascade; GO:0001501~skeletal development; GO:0008543~fibroblast growth factor receptor signaling pathway; GO:0007259~JAK-STAT cascade; GO:0007048~oncogenesis; GO:0006468~protein amino acid phosphorylation;	GO:0016021~integral to membrane; GO:0005887~integral to plasma membrane;	GO:0016740~transferase activity; GO:0005524~ATP binding; GO:0005007~fibroblast growth factor receptor activity; GO:0004872~receptor activity; GO:0004713~protein-tyrosine kinase activity; GO:0004672~protein kinase activity;
5	NM_000597	H79047	Hs.162	IGFBP2	2	3485	insulin-like growth factor binding protein 2 (36kD)	P47877	GO:0001558 GO:0005520 GO:0005576 GO:0019838	GO:0001558~regulation of cell growth;	GO:0005576~extracellular;	GO:0005520~insulin-like growth factor binding; GO:0019838~growth factor binding;

Figure 2. A portion of the corresponding results presented as an HTML table.

(including both standard and custom-designed arrays) and also between different species using homologene information obtained from NCBI (http://www.ncbi.nlm.nih.gov/HomoloGene/). Homologous genes (based on sequence homology) do not necessarily have homologous function. The arrays to be compared may include both old and new versions (e.g. different versions of Affymetrix GeneChips), and the comparison is done based on the same version of annotation and homologene information collected by our system.

IMPLEMENTATION

The annotation data and homologene data are stored in an Oracle database. We adopted the source code (a set of Perl programs) distributed by Stanford's SOURCE, which integrates information from multiple data sources including SwissProt, GO and NCBI's databases including LocusLink and Unigene. The scripts load the integrated data into the Oracle database. Unigene Cluster IDs were used for performing comparison among platforms involving a single species. We have extended the GO scripts to incorporate annotation data for more species including Arabidopsis. In addition, we have written a Perl program to fetch the homologene dataset from the NCBI site and put the data in KARMA. Such data enable identification of homologous sequences spotted on different array platforms involving multiple organisms. Our web interface was implemented in Perl/CGI, interfacing with Apache web server running on a Linux PC.

APPLICATIONS AND ILLUSTRATION

Currently the KARMA tool is being used for the following types of applications.

1. To annotate individual whole arrays based on GenBank accession numbers. The flexibility of the database allows any spreadsheet text file to be uploaded from a local browser for annotation by users creating their own custom arrays or using commercial platforms.
2. To compare array content in pairwise fashion to aid in the platform decision-making process during experimental design.
3. To query array content for genes of interest based on GenBank accession number.
4. To annotate subsets of gene expression microarray data based on GenBank accession numbers.
5. To sort annotation results based on GO ID, GenBank accession numbers, SwissProt, etc.
6. To create expanded annotated datasets for GEO Platform submission and provide links to existing GEO Platform files.

Figure 1 shows a web page that allows a user to select Keck Microarray Resource array platforms or upload their own data files for annotation and comparison. In this case, the user wants to make a comparison between the Affymetrix human GeneChip (HU133A) and an oligo human spotted array. Currently, our program accepts only GenBank accession numbers as the spot IDs, but additional types of ID will be added. When users upload their own files, they need to enter the position of the column (column number) that contains the GenBank accession numbers. As described previously, KARMA allows both single- and multiple-species comparison. When two datasets (set A and set B) are compared for a single species, four set comparison operations are available, namely, $A \cap B$ (intersection or common elements between the two sets), $A \cup B$ (union or merging the non-redundant elements from the sets) and two set difference operations: $A - B$ (elements that are in A but not in B) and $B - A$ (elements that are in B but not in A). For datasets involving different species, there is only one type of comparison, namely, identification of homologous sequences (this is similar to the notion of intersection). Instead of comparing the whole datasets, our system allows comparison of subsets based on user-defined restriction conditions. Currently, we allow users to restrict the comparison by GenBank accession numbers, gene symbols, names or Unigene Cluster IDs. As shown in Figure 1, the comparison is restricted by gene names that contain the phrase 'growth factor'. In addition to comparison, KARMA provides up-to-date annotation including gene symbols, functional descriptions, Locus Link ID and Gene Ontology IDs, which can be selected for display in the output. An option is available to sort the output by Unigene ID, GO ID or SwissProt ID. Currently, the program

is designed to run in batch mode. In other words, the user does not need to wait for the program to finish. When the processing is done, the URL links to the output report files in HTML, Excel and text formats are sent to the email address entered by the user. Figure 2 shows a portion of the HTML-formatted comparison/annotation report corresponding to the input parameters as shown in Figure 1. As shown in Figure 2, the GO terms (as well as the GO IDs) associated with each entry are separated into three categories: biological process, cellular component and molecular function. This makes the output more informative for further analysis. In addition, the advantage of the HTML format is that it allows hypertext links for accessing more detailed information available at other websites (e.g. clicking on a GO ID will return the corresponding GO functional description). The Excel output format also allows the hypertext links to be preserved. If the user saves the results in text format, they can be re-uploaded to the server for comparing with another array (either Keck, commercial or custom). This will allow the user to compare more than two arrays.

FUTURE DIRECTION

In the ever-expanding informatics domain faced by biological investigators, KARMA is a tool that filters, queries, compares and provides annotation in a universal manner. KARMA now forms an expandable backbone to which other database tools may be linked, such as transcription factor and sequence databases including AGRIS (http://arabidopsis.med.ohio-state.edu/) and TFSearch (http://siriusb.umdnj.edu:18080/EZRetrieve/multi_t.jsp). In the future, links from MAC—Microarray Convoluter (9), a tool for generating convoluted array format from array source plates—and from YMD (Yale Microarray Database) (10) query output could be implemented. Additionally GEO Platform file creation in soft format, available now at YMD, may be linked to KARMA to generate annotation information for GEO Platform descriptions. As the number and nature of databases and analytic tools expand, so will the interoperability of KARMA with these databases and tools.

ACKNOWLEDGEMENTS

This work was supported in part by NIH grants K25 HG02378 from the National Human Genome Research Institute, NIH 5 U24 DK58776 from the National Institute of Diabetes and Digestive and Kidney Diseases and Anna and Argall Hull Foundation to K.W. and J.H., T15 LM07056 from the National Library of Medicine, and NSF grant DBI-0135442.

REFERENCES

1. 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.
2. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res*., **31**, 23–27.
3. Ashburner,M., Ball,C., Blake,J., Botstein,D., Butler,H., Cherry,M., Davis,A., Dolinski,K., Dwight,S., Eppig,J. *et al*. (2000) Gene ontology: tool for the unification of biology. *Nature Genet*., **25**, 25–29.
4. Bairoch,A. and Boeckmann,B., (1994) The SWISS-PROT protein sequence data bank: current status. *Nucleic Acids Res*., **22**, 3578–3580.
5. Diehn,M., Sherlock,G., Binkley,G., Jin,H., Matese,J., Hernandez-Boussard,T., Rees,C., Cherry,J., Botstein,D., Brown,P. and Alizadeh,A. (2003) SOURCE: a unified genomic resource of functional annotations, ontologies, and gene. *Nucleic Acids Res*., **31**, 219–223.
6. Bouton,C. and Pevsner,J. (2000) DRAGON: database referencing of array genes online. *Bioinformatics*, **16**, 1038–1039.
7. Kulkarni,A., Williams,N., Lian,Y., Wren,J., Mittelman,D., Pertsemlidis,A. and Garner,H. (2002) ARROGANT: an application to manipulate large gene collections. *Bioinformatics*, **18**, 1410–1417.
8. Tsai,J., Sultana,R., Lee,Y., Pertea,G., Karamycheva,S., Antonescu,V., Cho,J., Parvizi,B., Cheung,F. and Quackenbush,J. (2001) RESOURCERER: a database for annotating and linking microarray resources within and across species. *Genome Biol*., **2**, 1–4.
9. Cheung,K.H., Hager,J., Nelson,K., White,K., Li,Y., Snyder,M., Williams,K. and Miller,P. (2002) A dynamic approach to mapping coordinates between microplates and microarrays. *J. Biomed. Inform*., **35** 306–312.
10. Cheung,K.H., White,K., Hager,J., Gerstein,M., Reinke,V., Nelson,K., Masiar,P., Srivastava,R., Li,Y., Li,J. *et al*. (2002) YMD: a microarray database for large-scale gene expression analysis. *Proceedings of the American Medical Informatics Association 2002 Symposium*, Hanley and Belfus, Inc., San Antonio, TX, pp. 140–144.