Bioinformatics Advance Access published June 22, 2007

Gene Expression

Web-Based GeneChip Analysis System for Large-Scale Collaborative Projects

Manhong Dai, Pinglang Wang, Elvis Jakupovic, Stanley J. Watson and Fan Meng*

Molecular and Behavioral Neuroscience Institute and Department of Psychiatry, U of Michigan, Ann Arbor, MI48109,U.S.

Associate Editor: Dr. Joaquin Dopazo

ABSTRACT

Summary: The Web-Based GeneChip Analysis System (WGAS) is developed to overcome limitations in analysis setup efficiency, data and procedure sharing, as well as security issues in existing commercial and public domain solutions. It also incorporates unique functions and resources for more accurate and flexible GeneChip analysis.

Availability: WGAS is freely available at:
http://arrayanalysis.mbni.med.umich.edu/arrayanalysis.html.

Contact: mengf@umich.edu

1 INTRODUCTION

Affymetrix GeneChip for expression analysis is probably the most widely adopted platform for large-scale expression profiling projects. While many commercial and public domains solutions incorporate a plethora of functions for data analysis and exploration, none of them are designed to deal with tasks involving hundreds of samples from different sources with complex properties. Sharing proprietary data/results, analysis methods and computing power across different research groups is also a challenge.

We developed a Web-based GeneChip Analysis System (WGAS) that provides centralized data analysis and management for large collaborative projects. Our system is based on the BioConductor platform, thus it enables easy integration of third-party analysis procedures. Sample descriptions, GeneChip probe level data files (CEL files) and parameters used in each analysis are tracked by an Oracle database. Data analysis is performed at a central LINUX cluster thus each research group can conduct large-scale analysis efficiently without the need to setup and manage computer clusters locally. We also created a distributed security network that allows encrypted data transfer across different geographic locations and within intranets.

2 FUNCTION DESCRIPTION

Data upload

WGAS supports both automatic and manual data uploading to a centralized Oracle database. The automatic uploading method utilizes a script to periodically scan pre-defined data folders at different geographical locations. Once new GeneChip files are detected, it will extract the chip type, sample barcode information etc. recorded in specific fields of the GeneChip experiment information file (EXP file) during GeneChip assays. If the information in the EXP file is consistent with the pre-defined descriptive data file name and sample information, the script will automatically upload the corresponding CEL file into our database. A notification e-mail will be automatically sent to the data creator if any missing field or inconsistency is found. To safeguard the confidentiality of clinical information, anonymized sample description data created by clinical researchers are manually loaded into our database.

The manual GeneChip data uploading method enables users to analyze their proprietary CEL files in their own private account without sharing data with external users. Setting up a private account involves three main steps 1) designate the account manager and define other account members allowed to access data 2) establish CEL property description data format, i.e., the name and data type of each property used to describe a CEL file in an private account. Most of the properties are for describing samples used to generate the corresponding CEL files such as treatment/control. 3) download a python package from our website and installed it on a local computer for remote CEL files and CEL file property data uploading. Detailed instructions about account signup, the python package download and installation, as well as data upload are at: (http://arrayanalysis.mbni.med.umich.edu/MBNIUM.html#MBNIUM). We also have a flash tutorial on the WGAS homepage showing how to perform CEL files and CEL file property uploading.

Since re-analysis and comparison with public domain GeneChip data is very useful in large-scale analysis, WGAS keeps a nightly updated version of GeneChip CEL files deposited in the NCBI Gene Expression Omnibus (GEO) database. There are over 33,000 CEL files from over 600 experiments in WGAS already. Consequently, researchers can easily re-analyze GEO CEL files using unique analysis methods and newer probe annotations in WGAS.

Probe level analysis

All popular and stable R implementation of probe-level analysis functions, including MAS5.0 (Affymetrix, 2005), dCHIP (Li and Wong, 2001), RMA (Irizarry, et al., 2003), GCRMA (Wu, et al., 2003), affyPLM (Bolstad, et al., 2004), as well as the threestep (Bolstad, 2007) and the expresso (Gautier, et al., 2005) functions are supported in WGAS. Most of these functions incorporate the normalization and background subtraction procedures for GeneChip analysis. The use of the BioConductor platform allows the easy addition of newer analysis function, too.

WGAS natively supports the use of custom GeneChip library files (Chip Definition Files or CDFs) through a straightforward dropdown menu. We believe such updated probe set definitions are critical for the accurate interpretation of GeneChip results (Dai, et al., 2005). Recently there are also independent systematic analyses demonstrating that our custom CDF can provide higher precision, accuracy as well as reduce false positive rate (Lu and Zhang, 2006; Sandberg and Larsson, 2007). While we also provide direct custom CDF download, the ability of selecting various CDFs

*To whom correspondence should be addressed.

through a graphic user interface greatly facilitates the analysis of GeneChip data using significantly better probe set definitions.

Since GeneChips include a noteworthy number of allele-specific probes and they may interfere with genetical genomic analysis, where samples are grouped according to individual genotypes (Sliwerska, et al., 2006), WGAS incorporates the ProbeFilter function developed in our group. The ProbeFilter function can be used to remove allele-specific probes meeting user's criteria (e.g., mismatch location on the probe and SNP heterozygosity) during probe-level data analysis. It can also remove predefined species-specific probes for better cross-species GeneChip applications.

Furthermore, since large-scale GeneChip analysis often need to merge legacy data generated by older generations GeneChips, we implemented a ConsensusCDF function in R to generate a consensus CDF on-the-fly when more than one chip type is used in an analysis. It can utilize probe sequence information to create consensus probe sets with at least 3 identical probes across all chip types, such as HG Focus, HG-133A and HG-U133P. As a result, the ConsensusCDF function can remove chip-specific systematic bias caused by probe set content difference across different generations of GeneChips.

Gene level analysis

WGAS includes standard gene (probe set) level statistical functions such as t-test, one-way ANNOVA and False Discovery Rate in the SAMR package (Tusher, et al., 2001).

To deal with the inevitable batch variations originated from different labeling reactions, scanner settings, chip type, etc. in large collaborative projects, we implemented the median centering procedure, which has been shown to be an effective method for reducing such systematic variations (Eckel, et al., 2005). Users can set up median centering correction for systematic data variations based on the relevant combinations of sample properties through WGAS web interface. Combining the median centering function with the accessibility of all GEO CEL files in WGAS greatly facilitates the meta-analysis of results from different laboratories.

In addition, we include the SPLICING function developed in our group for more accurate alternative splicing analysis with the new GeneChip exon arrays as well as extracting splicing signals from earlier generation of GeneChips. The SPLICING function uses the latest gene and exon definition from the ENSEMBL database. It calculates exon/gene signal ratio based on the RMA analysis of the corresponding exon and gene probe sets, followed by SAMR FDR ranking of such gene normalized exon signals.

Furthermore, since many researchers prefer to use commercial microarray packages for interactive gene-level analysis and presentation quality graphics, WGAS allows users to download the probe-level analysis results, quality control images and the gene set analysis results mentioned in the next section.

Gene-set level analysis

Gene-set level analysis provides the possibility of detecting subtle but coordinated expression changes among functionally related genes. WGAS currently supports two popular gene set-level analysis algorithms: the improved GSEA algorithm (Subramanian, et al., 2005) and the SigPathway method (Tian, et al., 2005). In addition, our own FunctionScore method, designed for measuring gene set activity in individual samples rather than sample groups, is accessible through WGAS. Our FunctionScore method enables the investigation of functional heterogeneity at gene-set level in large-scale analysis.

Since extensive functional annotation is critical for gene-set level analysis, we developed an internal function for the addition and updating of gene function annotations. We have currently integrated annotations from Gene Ontology, KEGG, BioCarta, GenMAPP, cytoband, microRNA target and neurotransmitter/neuropeptide-related function categories curated in-house. We are working with researchers in the National Center for Integrated Biomedical Information to add more function annotations related to transcription regulation and protein-protein interaction extracted from free text literature. To enable the use of extensive WGAS function annotations in standalone gene set analysis packages, the most current gene function annotations in WGAS can also be downloaded through an expandable function tree at http://arrayanalysis.mbni.med.umich.edu/genefunc/ .

3 WEB INTERFACE

The WGAS web interface guides users step-by-step through the analysis workflow for each data analysis job. In addition, we provide nine Flash tutorials for various data analysis tasks at the WGAS homepage. Every WGAS page title in the data analysis workflow is also linked to a Flash document describing the choices available in a page. Once a job is submitted, it will be processed or queued on our LINUX cluster, which can process analysis jobs involving several thousand HG-U133 Plus CEL files.

Selecting desired CEL files based on complex sample properties is the first bottleneck in large-scale analysis setup. In order to efficiently deal with about 200 clinical and sample property parameters in our internal projects, we organize them into more than a dozen property categories, where each category contains no more than 20 properties. Such a two-tiered sample property organization provides efficient access to every property and it is essential for analyzing complex sample properties in large projects. The default Boolean relationship setting for various selection parameters is "and". A user may also modify the CEL file selection string by using other type of relationships (e.g., "or", "not").

Assigning group/factor to each sample for median centering and group statistics is not only time-consuming but also error-prone in large-scale analysis. WGAS enables highly efficient and flexible multi-level factor assignment through simple dropdown menus.

WGAS has two functions for advanced users. The wizard function is for altering parameters used in analysis, such as the number of permutations or cutoff threshold in FDR analysis. The scripting window allows users to modify the WGAS-generated R-script for implementing user-specified functions before submitting the analysis job. Since every analysis job is easily retrievable in WGAS, complex custom analysis procedures can be shared among collaborators and used repetitively on different data sets without recoding.

Once an analysis job is finished on WGAS, the submitter will receive an e-mail message containing the web address for the result download. Existing analysis jobs can be retrieved and modified by users in the same account group for re-analysis. The actual R script used for a given analysis can also be downloaded thus a user may run or modify a WGAS generated R script on a local computer. WGAS includes an alert function that can notify a user about new CEL files satisfying criteria defined in an earlier analysis. These

functions greatly increase the efficiency of setting up complex large-scale GeneChip data analysis.

4 SECURITY

To ensure secured data transfer across different geographical locations as well as within intranet, we added a network security layer based on Checkpoint's Firewall/VPN and Microsoft's Active Directory (AD) technologies. All remote locations linked through VPN appear on the same physical network and we can publish and access all resources.

We utilize a RADIUS service using LDAP for authentication to the firewall. The authentication of VPN access can be easily managed from a central location.

In summary, we developed a flexible and powerful GeneChip analysis system that greatly facilitates large-scale collaborative research across different laboratories. WGAS also incorporates several unique functions developed in our group for more efficient and accurate GeneChip analysis. We are committed to keep functions and resources in WGAS updated as well as to improve our system based on users' feedbacks.

ACKNOWLEDGEMENTS

The authors are members of the Pritzker Neuropsychiatric Disorders Research Consortium, which is supported by the Pritzker Neuropsychiatric Disorders Research Fund L.L.C. This work is also partly supported by the National Center for Integrated Biomedical Informatics through NIH grant 1U54DA021519.

Conflict of Interest: none declared.

REFERENCES

Affymetrix (2005) GeneChip® Operating Software Manual, *http://www.affymetrix.com/support/technical/manuals.affx*.

Bolstad, B. (2007) affyPLM: the threestep function, *http://rss.acs.unt.edu/Rdoc/library/affyPLM/doc/ThreeStep.pdf*.

Bolstad, B.M., Collin, F., Simpson, K.M., Irizarry, R.A. and Speed, T.P. (2004) Experimental design and low-level analysis of microarray data, *Int Rev Neurobiol*, **60**, 25-58.

Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H., Watson, S.J. and Meng, F. (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data, *Nucleic Acids Res*, **33**, e175.

Eckel, J.E., Gennings, C., Therneau, T.M., Burgoon, L.D., Boverhof, D.R. and Zacharewski, T.R. (2005) Normalization of two-channel microarray experiments: a semiparametric approach, *Bioinformatics*, **21**, 1078-1083.

Gautier, L., Irizarry, R., Cope, L. and Bolstad, B. (2005) Description of affy, *http://www.bioconductor.org/repository/devel/vignette/affy.pdf*.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B. and Speed, T.P. (2003) Summaries of Affymetrix GeneChip probe level data, *Nucleic Acids Res*, **31**, e15.

Li, C. and Wong, W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection, *Proc Natl Acad Sci U S A*, **98**, 31-36.

Lu, X. and Zhang, X. (2006) The effect of GeneChip gene definitions on the microarray study of cancers, *Bioessays*, **28**, 739-746.

Sandberg, R. and Larsson, O. (2007) Improved precision and accuracy for microarrays using updated probe set definitions, *BMC Bioinformatics*, **8**, 48.

Sliwerska, E., Meng, F., Speed, T.P., Jones, E.G., Bunney, W.E., Akil, H., Watson, S.J. and Burmeister, M. (2006) SNPs on Chips: The Hidden Genetic Code in Expression Arrays, *Biol Psychiatry*.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proc Natl Acad Sci U S A*, **102**, 15545-15550.

Tian, L., Greenberg, S.A., Kong, S.W., Altschuler, J., Kohane, I.S. and Park, P.J. (2005) Discovering statistically significant pathways in expression profiling studies, *Proc Natl Acad Sci U S A*, **102**, 13544-13549.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response, *Proc Natl Acad Sci U S A*, **98**, 5116-5121.

Wu, Z., Irizarry, R., Gentleman, R., Murillo, F. and Spencer, F. (2003) A Model Based Background Adjustment for Oligonucleotide Expression Arrays, *Technical Report, John Hopkins University, Department of Biostatistics Working Papers, Baltimore, MD*.