D898–D901 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 25 October 2008*
doi:10.1093/nar/gkn786

Implementation of GenePattern within the Stanford Microarray Database

Jeremy Hubble[1], Janos Demeter[2], Heng Jin[2], Maria Mao[1], Michael Nitzberg[2], T. B. K. Reddy[2], Farrell Wymore[2], Zachariah K. Zachariah[2], Gavin Sherlock[1] and Catherine A. Ball[2,*]

[1]Departments of Genetics, Stanford University School of Medicine, CA 94305, USA and [2]Departments of Biochemistry, Stanford University School of Medicine, Stanford, CA 94305, USA

Received September 15, 2008; Accepted October 9, 2008

ABSTRACT

Hundreds of researchers across the world use the Stanford Microarray Database (SMD; http://smd.stanford.edu/) to store, annotate, view, analyze and share microarray data. In addition to providing registered users at Stanford access to their own data, SMD also provides access to public data, and tools with which to analyze those data, to any public user anywhere in the world. Previously, the addition of new microarray data analysis tools to SMD has been limited by available engineering resources, and in addition, the existing suite of tools did not provide a simple way to design, execute and share analysis pipelines, or to document such pipelines for the purposes of publication. To address this, we have incorporated the GenePattern software package directly into SMD, providing access to many new analysis tools, as well as a plug-in architecture that allows users to directly integrate and share additional tools through SMD. In this article, we describe our implementation of the GenePattern microarray analysis software package into the SMD code base. This extension is available with the SMD source code that is fully and freely available to others under an Open Source license, enabling other groups to create a local installation of SMD with an enriched data analysis capability.

INTRODUCTION

The Stanford Microarray Database [SMD; http://smd.stanford.edu/, (1)] provides a web-based research environment that supports the research of almost 1500 active users in more than 400 laboratories around the world. Users of the Stanford installation have used SMD to study the biology of over 60 organisms, including humans and various model organisms, and have entered data generated from more than 70 000 microarrays. Data stored in SMD have led to the publication of over 400 research papers and all raw data related to these publications (including results from about 9000 *Homo sapiens* and more than 5000 arrays from *Mycobacterium tuberculosis*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and *Mus musculus* combined) are made freely available via the SMD website. SMD stores data from gene expression as well as array CGH and ChIP–chip experiments, and supports multiple microarray platforms (spotted cDNA or oligonucleotide arrays, Affymetrix, Agilent, Combimatrix and Nimblegen arrays). SMD provides extensive biological annotation for genes and sequences for all supported organisms on each of the platforms, as well as annotations from the Gene Ontology where they are available. SMD also provides researchers with experiment annotation tools to make their data fully compliant to the Minimal Information About a Microarray Experiment standards (2) and has tools to write MAGE-ML files (3) and a data pipeline that can communicate published data directly to the international data repositories at ArrayExpress (4) and Gene Expression Omnibus (GEO) (5). The public data can be selected, viewed, downloaded and analyzed by the public using most of the data analysis and quality assessment tools that are available to registered SMD users.

ANALYSIS TOOLS WITHIN SMD

SMD has long provided a limited set of data analysis tools to its users, allowing them to directly analyze their data within SMD, and we recently added a repository feature that allows users to share the results of their analyses. However, the implementation of these various tools within SMD has certain shortcomings for both users

*To whom correspondence should be addressed. Tel: +1 650 724 3028; Fax: +1 650 724 3701; Email: ball@genome.stanford.edu and those maintaining the software. First, SMD users often want to experiment with different data analysis approaches and thus frequently request implementation of an analysis tool that is of limited use to other researchers. Due to limited resources, we had to treat some requests as a lower priority, which can be frustrating to users and developers both. Second, completing a data analysis requires several steps with many variables that need to be specified, and users will frequently want to re-execute these steps, varying the values of the parameters. This can be time-consuming to the user and also makes it difficult for the user to remember exactly what they did during a particular analysis or to replicate a complex analysis pipeline designed by someone with more expertise. Communicating, sharing and re-executing a data analysis was thus challenging for SMD users. Third, introduction of new data analysis methods into the existing SMD software architecture was time-consuming, required expert understanding of the algorithm being implemented and required a novel user interface for every algorithm. Fourth, since the analysis tools had been tightly integrated with the SMD data retrieval pipeline, careful engineering and testing was required to ensure that introduction of a new algorithm did not break the software pipeline. While mostly effective, the SMD architecture for data analysis was difficult to maintain, time-consuming to extend and was unable to keep up with the rapidly expanding number of microarray data analysis techniques, and the needs of our users. We thus found ourselves unable to quickly or easily implement and deploy new microarray analysis techniques.

In face of these shortcomings, we wanted to improve the code development and maintenance process and increase the speed with which we could adopt and implement new data analysis tools and techniques. In addition to considering building a new data analysis software platform from the ground up, we also considered adopting an existing microarray data analysis architecture. While using software written in-house would give us complete control over the project, we would also have complete responsibility for the maintenance of the software as well as implementing new analysis techniques. In contrast, using software created by others would require us to cede some control over the user interface, data formats used and overall software behavior, development and bug fixes. However, using software created by others might allow us to provide our users with significant utility more quickly and would also allow us to benefit from economies of scale—our users would be able to benefit from analysis techniques added by others. After careful consideration, we elected to integrate the stand-alone GenePattern microarray data analysis package. GenePattern has several attractive traits: users can chain together analysis steps into pipelines that can be carefully documented, re-executed and shared, externally developed tools are easy to incorporate, and most importantly, data from a database such as SMD can be easily entered into the GenePattern environment. Features of GenePattern have been fully described elsewhere (6). In this article, we describe the implementation of GenePattern as part of the SMD code base.

RESULTS

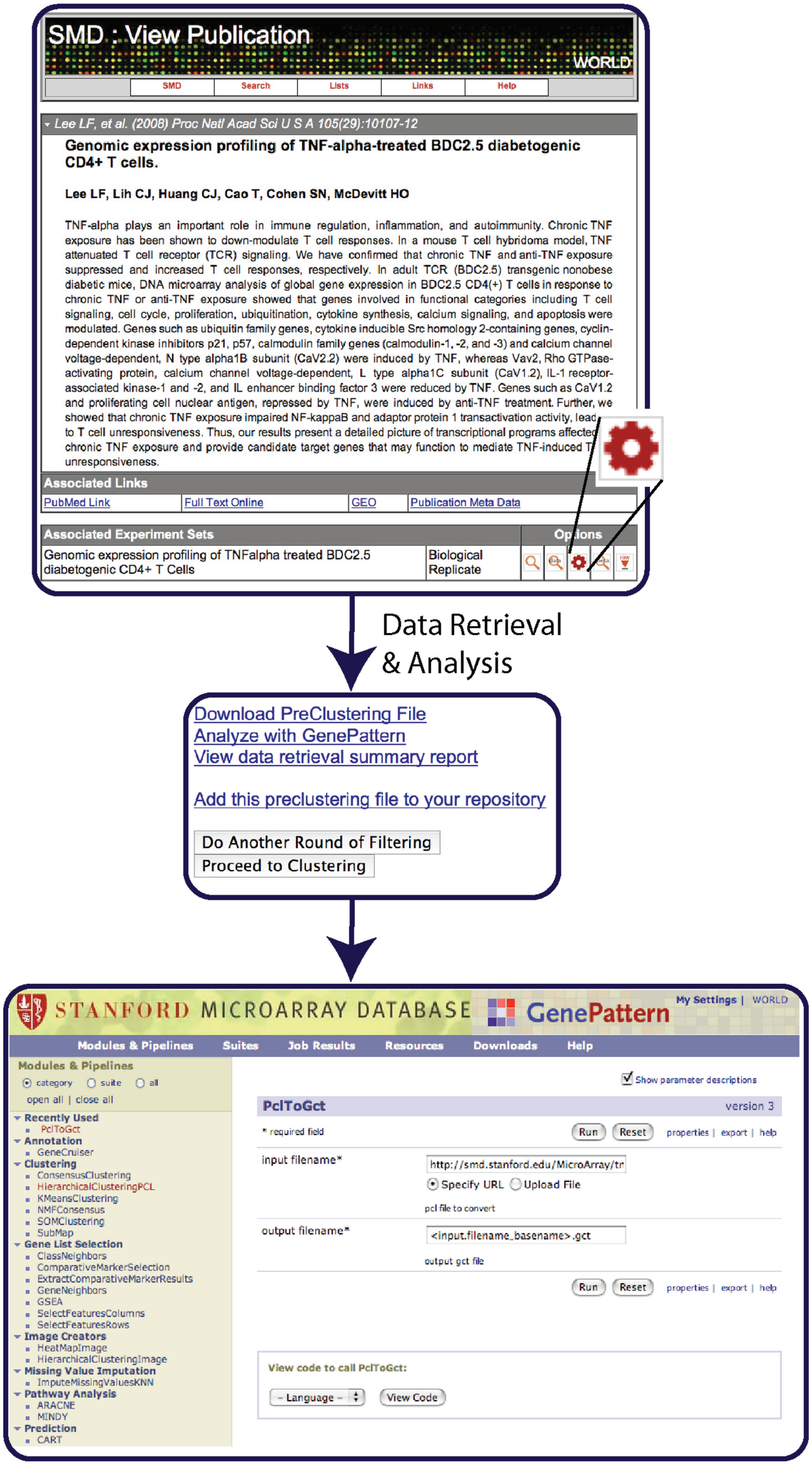
Deployment of GenePattern within SMD

Mesirov and co-workers maintain GenePattern both as a web application and as a downloadable software package that can be locally installed, either as a desktop application, or on a server. We elected to install a customized version of the GenePattern server on the same server as SMD, and to integrate our current data retrieval pipeline and analysis tools with GenePattern. The integration of GenePattern with SMD allows us to maintain a common environment where data and analyses can be associated and where users can use the same login for both tools. A custom implementation also allows us to provide needed tools for transforming data from formats used by the SMD into formats that can be used by some GenePattern tools.

Installing GenePattern in the SMD software environment posed some specific engineering challenges. First, the two systems were developed to run under different operating systems. SMD currently runs under Sun's Solaris operating system, while GenePattern runs on a Linux platform. Though the two operating systems are similar, they have enough differences that there were many challenges in the installation of GenePattern. In addition, the standard GenePattern installation requires supporting certain software (such as the R programming language and Tomcat webserver) to be located in specific places and to be used solely by GenePattern. In our environment, many tools were shared by a complex set of other software, and many different solutions were required. In some cases, we could simply install a different version of a tool in the location expected by GenePattern, or point GenePattern to the correct path. In other instances, much more involved modifications were needed. For example, GenePattern originally utilized a local HSQL database, which proved to be extremely unreliable in our Solaris environment. We were able to achieve a reliable level of performance by instead configuring GenePattern to run using our local Oracle database.

Enabling GenePattern to function properly using microarray data from SMD data was the next major challenge. First, we had to convert microarray data from SMD-specific format (in PreCLustering files or PCL format) to the GCT format that is used by GenePattern. This required us to create and implement a PCL to GCT converter within GenePattern. The most significant difference between the two files was the optional presence of either gene or array weights in the PCL files, which can be used to downweight genes or arrays when the data are hierarchically clustered.

We encountered several challenges implementing individual data analysis modules in GenePattern. Some were resolved by simply changing configuration options—for example, to get the Heat Map module to work, we needed to modify files so that no pathnames included hyphens. Other times we had to install different versions of some of our tools (such as R) in order to get consistently successful results. We did encounter insurmountable problems implementing GenePattern's hierarchical clustering module because it depends on the use of



Lee LF, et al. (2008) Proc Natl Acad Sci U S A 105(29):10107-12

Genomic expression profiling of TNF-alpha-treated BDC2.5 diabetogenic CD4+ T cells.

Lee LF, Lih CJ, Huang CJ, Cao T, Cohen SN, McDevitt HO

TNF-alpha plays an important role in immune regulation, inflammation, and autoimmunity. Chronic TNF exposure has been shown to down-modulate T cell responses. In a mouse T cell hybridoma model, TNF attenuated T cell receptor (TCR) signaling. We have confirmed that chronic TNF and anti-TNF exposure suppressed and increased T cell responses, respectively. In adult TCR (BDC2.5) transgenic nonobese diabetic mice, DNA microarray analysis of global gene expression in BDC2.5 CD4(+) T cells in response to chronic TNF or anti-TNF exposure showed that genes involved in functional categories including T cell signaling, cell cycle, proliferation, ubiquitination, cytokine synthesis, calcium signaling, and apoptosis were modulated. Genes such as ubiquitin family genes, cytokine inducible Src homology 2-containing genes, cyclin-dependent kinase inhibitors p21, p57, calmodulin family genes (calmodulin-1, -2, and -3) and calcium channel voltage-dependent, N type alpha1B subunit (CaV2.2) were induced by TNF, whereas Vav2, Rho GTPase-activating protein, calcium channel voltage-dependent, L type alpha1C subunit (CaV1.2), IL-1 receptor-associated kinase-1 and -2, and IL enhancer binding factor 3 were reduced by TNF. Genes such as CaV1.2 and proliferating cell nuclear antigen, repressed by TNF, were induced by anti-TNF treatment. Further, we showed that chronic TNF exposure impaired NF-kappaB and adaptor protein 1 transactivation activity, lead to T cell unresponsiveness. Thus, our results present a detailed picture of transcriptional programs affected chronic TNF exposure and provide candidate target genes that may function to mediate TNF-induced T unresponsiveness.

Associated Links

PubMed Link	Full Text Online	GEO	Publication Meta Data

Associated Experiment Sets		Options
Genomic expression profiling of TNFalpha treated BDC2.5 diabetogenic CD4+ T Cells	Biological Replicate	

Download PreClustering File
Analyze with GenePattern
View data retrieval summary report

Add this preclustering file to your repository

Do Another Round of Filtering
Proceed to Clustering

Modules & Pipelines

● category ○ suite ○ all

open all | close all

- Recently Used
  - PclToGct
- Annotation
  - GeneCruiser
- Clustering
  - ConsensusClustering
  - HierarchicalClusteringPCL
  - KMeansClustering
  - NMFConsensus
  - SOMClustering
  - SubMap
- Gene List Selection
  - ClassNeighbors
  - ComparativeMarkerSelection
  - ExtractComparativeMarkerResults
  - GeneNeighbors
  - GSEA
  - SelectFeaturesColumns
  - SelectFeaturesRows
- Image Creators
  - HeatMapImage
  - HierarchicalClusteringImage
- Missing Value Imputation
  - ImputeMissingValuesKNN
- Pathway Analysis
  - ARACNE
  - MINDY
- Prediction
  - CART

☑ Show parameter descriptions

PclToGct version 3

* required field Run Reset properties | export | help

input filename* http://smd.stanford.edu/MicroArray/tr
● Specify URL ○ Upload File
pcl file to convert

output filename* <input.filename_basename>.gct
output gct file

Run Reset properties | export | help

View code to call PclToGct:
- Language - View Code

Figure 1. Accessing GenePattern microarray data analysis package from a publication record in SMD. Clicking on the red gear icon in a publication record will take the user to the SMD data retrieval pipeline. At any point after the data are retrieved, the user will be able to download the data files or click on the ‘Analyze with GenePattern’ option. GenePattern is launched with the data already loaded into the PCL to GCT conversion module.

Linux libraries that are not present on the Solaris operating system. To address this, we simply wrote and implemented a new GenePattern module that integrated XCluster, the existing SMD hierarchical clustering software, into GenePattern.

Accessing GenePattern from SMD

SMD users can access GenePattern by several methods. Registered users can launch GenePattern and load a given data set by clicking on the 'GP' icon next to a data set in their personal data repository. Since SMD-produced PCL files need to be converted to GCT files for further analysis, users are sent directly to the PCL to GCT conversion module, with the current data set pre-filled. Registered users can also click the 'Launch GenePattern' button to view all data in their repository within GenePattern, and 'drag and drop' data sets into different GenePattern modules. All users (public users or registered users) can also access GenePattern from every step in the SMD data retrieval pipeline where a PCL file is produced (Figure 1).

FUTURE WORK

Now that our users are able to use GenePattern within SMD, we look forward to contributing to the GenePattern community. We plan to convert all of our analysis tools to GenePattern modules that can be shared by other GenePattern users. We will also implement increased connectivity between GenePattern and SMD. This will include allowing direct invocation of GenePattern modules from within SMD, as well as sending GenePattern results and pipelines back to SMD. We aim to enable users to publish both their data sets and analysis pipelines within SMD/GenePattern, thus enabling their analysis to be easily repeated, and also easily changed and reanalyzed. We are working closely with our users to construct data analysis pipelines in GenePattern that can be specified by biostatisticians and re-used by others. Our experience implementing the GenePattern software package, while not without difficulties and challenges, has shown that incorporation of well-constructed third-party open source software can indeed be a pragmatic and effective move even in a software environment as complex as SMD.

ACKNOWLEDGEMENTS

We thank the users of SMD for their constant feedback that helps us to improve SMD, and Jill Mesirov's group (especially Ted Liefeld, Peter Carr and Michael Reich) for help with deploying the GenePattern instance within SMD.

FUNDING

National Institute of Health (R01 HG003469 to G.S. and C.A.B.); Bill and Melinda Gates Foundation (the incorporation of GenePattern within SMD). Funding for open access charge: National Institute of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Demeter,J., Beauheim,C., Gollub,J., Hernandez-Boussard,T., Jin,H., Maier,D., Matese,J.C., Nitzberg,M., Wymore,F., Zachariah,Z.K. *et al.* (2007) The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res.*, **35**, D766–D770.
2. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.
3. Spellman,P.T., Miller,M., Stewart,J., Troup,C., Sarkans,U., Chervitz,S., Bernhart,D., Sherlock,G., Ball,C., Lepage,M. *et al.* (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.*, **3**, RESEARCH0046.
4. Parkinson,H., Kapushesky,M., Shojatalab,M., Abeygunawardena,N., Coulson,R., Farne,A., Holloway,E., Kolesnykov,N., Lilja,P., Lukk,M. *et al.* (2007) ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.*, **35**, D747–D750.
5. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, D5–D12.
6. Kuehn,H., Liberzon,A., Reich,M. and Mesirov,J.P. (2008) Using GenePattern for gene expression analysis (Chapter 7, Unit 7.12). In Baxevanis,A.D. (ed.) *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc., USA