Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W627–W632***
doi:10.1093/nar/gki443

VAMPIRE microarray suite: a web-based platform for the interpretation of gene expression data

Albert Hsiao[1,2], Trey Ideker[1], Jerrold M. Olefsky[3] and Shankar Subramaniam[1,4,*]

[1]Department of Bioengineering, [2]Medical Scientist Training Program, [3]Department of Medicine and [4]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA 92093, USA

Received February 11, 2005; Revised March 9, 2005; Accepted March 30, 2005

ABSTRACT

Microarrays are invaluable high-throughput tools used to snapshot the gene expression profiles of cells and tissues. Among the most basic and fundamental questions asked of microarray data is whether individual genes are significantly activated or repressed by a particular stimulus. We have previously presented two Bayesian statistical methods for this level of analysis, collectively known as variance-modeled posterior inference with regional exponentials (VAMPIRE). These methods each require a sophisticated modeling step followed by integration of a posterior probability density. We present here a publicly available, web-based platform that allows users to easily load data, associate related samples and identify differentially expressed features using the VAMPIRE statistical framework. In addition, this suite of tools seamlessly integrates a novel gene annotation tool, known as GOby, which identifies statistically overrepresented gene groups. Unlike other tools in this genre, GOby can localize enrichment while respecting the hierarchical structure of annotation systems like Gene Ontology (GO). By identifying statistically significant enrichment of GO terms, Kyoto Encyclopedia of Genes and Genomes pathways, and TRANSFAC transcription factor binding sites, users can gain substantial insight into the physiological significance of sets of differentially expressed genes. The VAMPIRE microarray suite can be accessed at http://genome.ucsd.edu/microarray.

INTRODUCTION

Gene expression microarrays are commonly used to study the transcriptional responses of cells and tissues. Most studies involve comparisons of an experimental treatment with a corresponding control, often with only 2–3 replicates within each treatment group. These experiments are commonly performed on either one-channel or two-channel microarray platforms, which simultaneously measure gene expression in one or two RNA samples, respectively. As biologists devise more sophisticated experimental designs, existing statistical tools for analyzing these data become increasingly unwieldy. For example, in large data sets of >50 arrays, users may wish to begin analysis before all microarrays have been completely processed. As users define more statistical tests to perform on this data, accounting for all of the tests and changes to the underlying data becomes extremely challenging. Furthermore, because of the high-throughput nature of these experiments, and because of their non-uniform error structure, the resulting data can be difficult to interpret. To address these issues, we have devised a tightly integrated, web-based suite of microarray analysis tools based on a robust Bayesian approach known as variance-modeled posterior inference with regional exponentials (VAMPIRE) (1).

The microarray analysis platform we present here provides an integrated interface for data management, statistical analysis and interpretation of gene expression data. To simplify the analysis of large data sets, this interface allows users to quickly load data and characterize experimental designs within the data set. Users can associate related samples and combine related groups. The variance structure of these groups can then be modeled to identify the coefficients of expression-dependent (A) and expression-independent variance (B). The analysis suite subsequently uses these models, stored in the user's account, and applies them to identify microarray features that are differentially expressed between treatment groups. Once this analysis is complete, the corresponding gene lists must still be interpreted and related to biological function. We have therefore integrated a novel statistical tool known as GOby, which can identify overrepresentation of previously defined functional categories. This is performed while respecting the hierarchical nature of annotation systems like Gene Ontology (GO) (2). Together, these data management and analytical tools provide users with a powerful new approach to microarray gene expression analysis.

*To whom correspondence should be addressed. Tel: +1 858 8220986; Fax: +1 858 8223752; Email: shankar@sdsc.edu

IMPLEMENTATION

The web application is implemented as a package of Java 1.5 servlets. In order to separate interface and the underlying content and logic, we applied a commonly used three-tier scheme. At the foundation of the VAMPIRE analysis suite is a structured query language (SQL)-based relational database. The current implementation uses a combination of MySQL and Oracle databases. Java interfaces have been appropriately designed to isolate SQL-specific code, which allows us to quickly re-deploy the system on other database platforms. Each servlet gathers data through these SQL interfaces and constructs its own intermediate extensible markup language (XML) document. XML is subsequently transformed via extensible stylesheet language transformation (XSLT) into HTML. The end-user is ultimately presented with the HTML interface. By implementing VAMPIRE in this fashion, each component can be re-designed or re-implemented without disturbing the remainder of the site.

The VAMPIRE algorithm itself is computationally intensive. On an Intel Xeon 3.0 GHz processor, typical execution time for variance modeling ranges from 5 to 10 min depending on the number of features on the microarray, and whether unpaired or paired analysis is desired. Significance testing and GOby analysis usually complete within 2–3 min. Because of these computational costs, we have devised a Java remote-method invocation (RMI)-based distributed processing solution (Figure 1). Each analysis job constructed by the web-application is placed into a priority queue. An RMI host program provides access to each job as it arrives at the top of the queue to any number of RMI clients. The RMI hosts and clients may be executed on remote machines distant from the web-server. This design provides scalability to the VAMPIRE site, as additional machines may be added to handle the increased computational load. Since all of these transactions are stored in the VAMPIRE database, no data is lost when individual hosts or processing nodes fail. Jobs may be restarted by other RMI nodes or provided through alternative RMI hosts as long as the database itself is accessible.

USER INTERFACE

The web-based interface of the microarray analysis suite can be divided into three major sections: (i) data management, (ii) statistical analysis and (iii) interpretation. Throughout all three divisions is a consistent user-interface with multiple tools for exporting data (Figure 2). To ease integration with third-party tools, results may be downloaded as tab-delimited text files or as XML documents.

Data management

Management of microarray data sets and their corresponding analyses becomes increasingly complex as the numbers of treatment groups grow. Users must manage not only individual samples, but also all subsequent analyses. This is a particular problem for users with large data sets, who wish to begin interpreting gene expression data before all of samples have been completely processed. To accommodate these issues, we have created a data management system that can be quickly used to load, annotate and associate microarray measurements. Summary measures of gene expression, such as those obtained from Affymetrix MAS 5.0 or Agilent processed signal

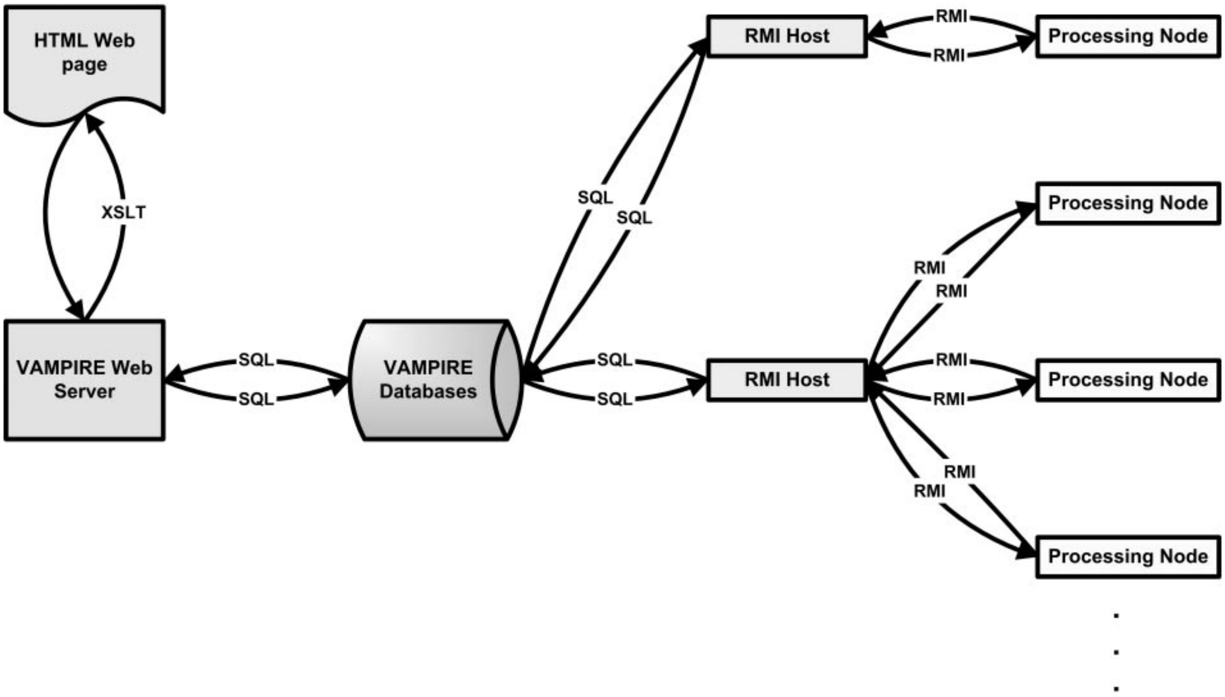
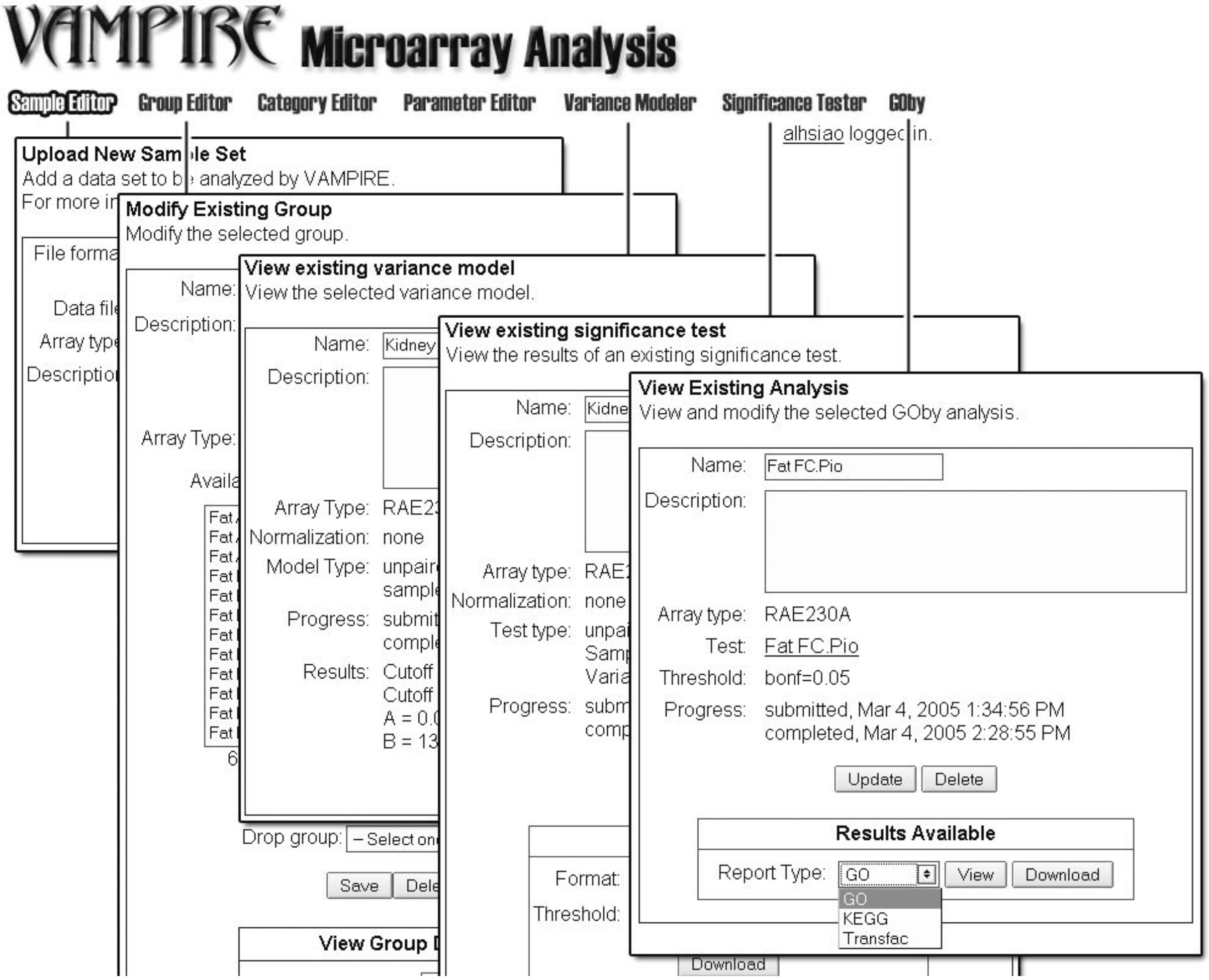


Figure 1. Design of the VAMPIRE processing pipeline. At the core of the application is an SQL-based relational database. The web-based user-interface generates intermediate XML documents from this database, and transforms these documents into the HTML interface via XSLT. A distributed processing framework based on Java/RMI disperses computational load on to remote processing servers.



Upload New Sam le Set
Add a data set to b e analyzed by VAMPIRE.
For more in
File forma
Data fil
Array typ
Descriptio

Modify Existing Group
Modify the selected group.
Name:
Description:
Array Type:
Availa
Drop group: – Select on
View Group D

View existing variance model
View the selected variance model.
Name: Kidney
Description:
Array Type: RAE23
Normalization: none
Model Type: unpair sample
Progress: submit comple
Results: Cutoff Cutoff A = 0.0 B = 13

View existing significance test
View the results of an existing significance test.
Name: Kidne
Description:
Array type: RAE2
Normalization: none
Test type: unpai Samp Varia
Progress: subm comp
Format:
Threshold:

View Existing Analysis
View and modify the selected GOby analysis.
Name: Fat FC.Pio
Description:
Array type: RAE230A
Test: Fat FC.Pio
Threshold: bonf=0.05
Progress: submitted, Mar 4, 2005 1:34:56 PM
completed, Mar 4, 2005 2:28:55 PM

Results Available
Report Type: GO

Figure 2. Screenshots of the VAMPIRE web interface. The intuitive design allows users to easily manage data sets, perform VAMPIRE statistical analysis, and interpret gene expression results with GOby.

intensities, can be imported into the server as tab-delimited text files, a file format that is easily accessible to most biologists.

Once data have been loaded into the web application, the user may associate related microarray samples. Characterizing these relationships is crucial, as they help to describe the analyses that will be later performed. For example, in a two-channel tutorial provided on the web site, users create sample groups for (i) replicates of LPS-treated macrophages, (ii) replicates of control-treated macrophages and (iii) paired samples obtained from the same chips. These groups can be further combined to create 'categories' of related groups. Once these relationships are recorded, they can be used to compare gene expression across different treatment conditions. Since further changes to each sample group are recorded by the analysis suite, VAMPIRE can subsequently inform users when analyses need to be re-executed.

Statistical analysis

Statistical analysis by VAMPIRE requires two distinct steps: (i) modeling of the error structure of sample groups and (ii) significance testing with a priori-defined significance thresholds. This approach to microarray analysis is considerably different from the approaches taken by other analytical methods (3–6). Normalization methods are commonly applied to average out the error structure prior to performing statistical analysis. Statistical tests are then left to address the significance of expression differences found in the remaining data. In contrast, VAMPIRE studies the underlying error structure, without perturbing it, and uses this knowledge to distinguish signal from noise. The cutoff for statistical significance is then defined by the significance threshold and by the magnitude of the variance model coefficients. Because of the additional variance modeling step however, this kind of analysis can be quite challenging without a robust accounting system. In the web application that we present here, users can easily keep track of all variance models and the data sets for which they can be applied. We have initially incorporated two variants of VAMPIRE—classical unpaired analysis (1) and paired analysis (A. Hsiao and S. Subramaniam, manuscript submitted); both of which use variance models to detect significant changes in gene expression.

When users submit a request to model the variance structure of a group of samples, a new 'processing job' is immediately submitted into a processing queue. Individual jobs require an average of 5–10 min to compute on an Intel Xeon 3.06 GHz processor. In the meantime, users may continue to use the remainder of the site, without waiting for each job to complete. An estimate of the date and time of completion is prominently displayed. Similarly, users may request that specific statistical tests be performed. Since these tests rely on variance model results, they will not be executed until their dependent models have been completed. As data can be continually added to the analysis platform, outdated models and tests are automatically flagged by the system to allow users to re-execute analyses with updated data. This particular feature facilitates 'on-the-fly' analysis. Users can monitor the results as they collect data, which may help them to decide which analyses require additional replication.

Interpretation

Differentially regulated features obtained from any statistical test must be interpreted biologically. We have developed a novel tool, known as GOby, to initiate biological interpretation, independent of whether VAMPIRE itself was used to derive the feature lists. This database-driven application curates annotation data from several sources: National Center for Biotechnology Information (NCBI), GO, Kyoto Encyclopedia of Genes and Genomes (7) (KEGG), TRANSFAC (8), Biocarta and Superarray. In addition, it can be readily updated with additional user-defined annotation lists.

GOby primarily uses its annotation database to identify overrepresented annotation groups. It does so by comparing a 'selected' list to a 'background'. In our experience, using the comprehensive feature list for each microarray as the background gives quite meaningful results. In a manner similar to other recently published tools (9–11), GOby uses exact probabilities to compute enrichment likelihoods, and displays the enrichment likelihood as a *P*-value. GOby reports as its *P*-value, the probability of finding no more than *s* features annotated with a given term among *k* 'selected' features:

$$P(\text{annotated} \leqslant s) = \sum_{i=0}^{s} P(\text{annotated} = i)$$

$$P(\text{annotated} = i) = \binom{k}{i} \prod_{j=0}^{i-1} \frac{b-j}{N-j} \cdot \prod_{j=0}^{k-i-1} \frac{(N-b)-j}{(N-i)-j} = \frac{\binom{b}{i}\binom{N-b}{k-i}}{\binom{N}{k}}$$

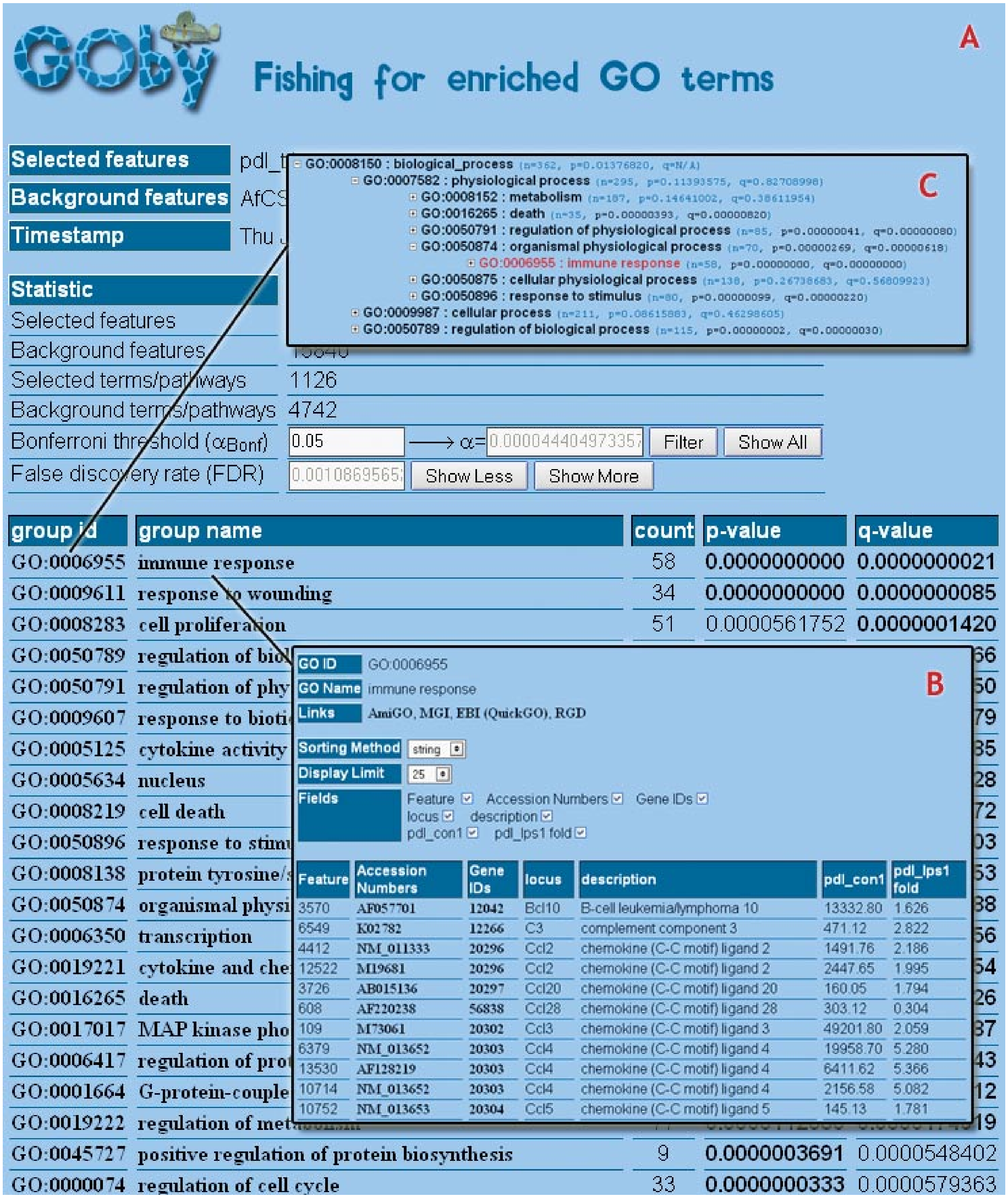
where *b* is the number of 'background' features annotated with the term; *s*, the number of 'selected' features annotated with the term; *N*, the total number of 'background' features; and *k*, the total number of 'selected' features.

Unlike similar tools however, GOby is also able to compute a 'conditional-enrichment likelihood', or *Q*-value, for each term in a hierarchical annotation system. This *Q*-value is based on the idea that truly meaningful enrichment in a hierarchical system like GO will occur at specific nodes in the annotation tree. The *Q*-value computes the enrichment likelihood for a particular term conditioned on the enrichment of its parent terms. In other words, instead of using the entire set of array features as the 'background', we use only the subset of features that are annotated with one of the parent terms. Unlike the *P*-value previously described, the *Q*-value prevents annotation terms from reaching significance simply because they lie in an area of the tree near other terms that are enriched. It can therefore narrow the user's focus by reporting only the optimal level of functional detail while excluding both more general and more specific terms (unless these terms fall into a second area of functional enrichment independent of the first). Since both methods have their own advantages, both are displayed in the results.

GOby provides two options for viewing the results of its statistical analysis. Results may be downloaded in an XML format, or they may be viewed directly from the VAMPIRE web site (Figure 3). Since each of the result pages is rendered for GOby via XSLT from the XML export file, all necessary information is easily accessible to third-party applications. Highly skilled users may wish to render their own views of GOby reports. Most users, however, will rely on automatically generated web pages. For each annotation system, GOby renders at least two types of web pages. First, a result table displays the annotation terms that are overrepresented given a specified significance cutoff. Significant *P*-values and *Q*-values, according to a user-selected significance threshold, are displayed in bold. Users may choose between Bonferroni-corrected thresholds and false-discovery rates to correct for multiple testing (A. Hsiao and S. Subramaniam, manuscript submitted). Second, each entry in the result table can be clicked to view a page that displays differentially expressed features that are annotated with the selected term. Because of careful integration with VAMPIRE statistical tests, fold-changes and *P*-values for each array feature can be viewed directly from these pages, eliminating the need to manually cross-reference additional lists. Each feature is linked by accession numbers and LocusLink IDs to corresponding web pages at NCBI. When available, each annotation term has also been linked to pages that display term-specific information. For example, clicking on the KEGG link from a KEGG report page will display the relevant KEGG pathway, while highlighting differentially expressed genes. In addition, we have included several Javascript-based functions to show/hide columns and to sort tables by column data. For hierarchical systems such as GO, GOby also renders a third view, the tree view. Here, *P*-values and *Q*-values are again visible to allow users to quickly address the importance of each node in annotation tree.

CONCLUSIONS

The VAMPIRE microarray analysis suite provides a fundamentally different approach to gene expression analysis. While other tools independently exist for management of microarray data, analysis and interpretation, none are yet equipped with the statistical tools that we have described. The VAMPIRE suite gracefully handles incremental analysis of large data sets, applies some of the most rigorous statistical methods available for microarray analysis, and provides



Selected features	pdl_t
Background features	AfCS
Timestamp	Thu J

Statistic	
Selected features	
Background features	15840
Selected terms/pathways	1126
Background terms/pathways	4742
Bonferroni threshold (αBonf)	0.05 ⟶ α=0.000044404973357 Filter Show All
False discovery rate (FDR)	0.0010869565 Show Less Show More

group id	group name	count	p-value	q-value
GO:0006955	immune response	58	0.0000000000	0.0000000021
GO:0009611	response to wounding	34	0.0000000000	0.0000000085
GO:0008283	cell proliferation	51	0.0000561752	0.0000001420
GO:0050789	regulation of bio			66
GO:0050791	regulation of phy			50
GO:0009607	response to bioti			79
GO:0005125	cytokine activity			35
GO:0005634	nucleus			28
GO:0008219	cell death			72
GO:0050896	response to stimu			03
GO:0008138	protein tyrosine/s			53
GO:0050874	organismal physi			38
GO:0006350	transcription			56
GO:0019221	cytokine and che			54
GO:0016265	death			26
GO:0017017	MAP kinase pho			37
GO:0006417	regulation of prot			43
GO:0001664	G-protein-couple			12
GO:0019222	regulation of met			19
GO:0045727	positive regulation of protein biosynthesis	9	0.0000003691	0.0000548402
GO:0000074	regulation of cell cycle	33	0.0000000333	0.0000579363

```
⊟ GO:0008150 : biological_process (n=362, p=0.01376820, q=N/A)
        ⊟ GO:0007582 : physiological process (n=295, p=0.11393575, q=0.82708998)
                ⊞ GO:0008152 : metabolism (n=187, p=0.14641002, q=0.38611954)
                ⊞ GO:0016265 : death (n=35, p=0.00000393, q=0.00000820)
                ⊞ GO:0050791 : regulation of physiological process (n=85, p=0.00000041, q=0.00000080)
                ⊟ GO:0050874 : organismal physiological process (n=70, p=0.00000269, q=0.00000618)
                        ⊞ GO:0006955 : immune response (n=58, p=0.00000000, q=0.00000000)
                ⊞ GO:0050875 : cellular physiological process (n=138, p=0.26738683, q=0.56809923)
                ⊞ GO:0050896 : response to stimulus (n=80, p=0.00000099, q=0.00000220)
        ⊞ GO:0009987 : cellular process (n=211, p=0.08615883, q=0.46298605)
        ⊞ GO:0050789 : regulation of biological process (n=115, p=0.00000002, q=0.00000030)
```

GO ID	GO:0006955
GO Name	immune response
Links	AmiGO, MGI, EBI (QuickGO), RGD
Sorting Method	string
Display Limit	25
Fields	Feature ☑ Accession Numbers ☑ Gene IDs ☑ locus ☑ description ☑ pdl_con1 ☑ pdl_lps1 fold ☑

Feature	Accession Numbers	Gene IDs	locus	description	pdl_con1	pdl_lps1 fold
3570	AF057701	12042	Bcl10	B-cell leukemia/lymphoma 10	13332.80	1.626
6549	K02782	12266	C3	complement component 3	471.12	2.822
4412	NM_011333	20296	Ccl2	chemokine (C-C motif) ligand 2	1491.76	2.186
12522	M19681	20296	Ccl2	chemokine (C-C motif) ligand 2	2447.65	1.995
3726	AB015136	20297	Ccl20	chemokine (C-C motif) ligand 20	160.05	1.794
608	AF220238	56838	Ccl28	chemokine (C-C motif) ligand 28	303.12	0.304
109	M73061	20302	Ccl3	chemokine (C-C motif) ligand 3	49201.80	2.059
6379	NM_013652	20303	Ccl4	chemokine (C-C motif) ligand 4	19958.70	5.280
13530	AF128219	20303	Ccl4	chemokine (C-C motif) ligand 4	6411.62	5.366
10714	NM_013652	20303	Ccl4	chemokine (C-C motif) ligand 4	2156.58	5.082
10752	NM_013653	20304	Ccl5	chemokine (C-C motif) ligand 5	145.13	1.781

Figure 3. GOby-rendered report pages. Three types of pages are automatically rendered by GOby for navigation of GOby results. The term table (**A**) displays annotation terms that are enriched among differentially expressed features. The term pages (**B**) show differentially expressed features that are annotated with each term. The tree term (**C**) displays the hierarchical structure of the annotation system with corresponding enrichment likelihoods.

powerful tools for interpretation of the results. We have also presented a novel tool, GOby, for interpreting of sets of differentially regulated genes. It can compute both 'global enrichment' of annotation terms, as well as 'conditional enrichment' at specific nodes of a hierarchical annotation system, like GO. This suite therefore represents a substantial advance in bringing the latest analytical algorithms into the hands of biologists.

ACKNOWLEDGEMENTS

We thank Chris Benner for several useful scripts, Yi-Xiong Zhou and Joshua Li for assisting in the Oracle-based implementation of the database backend, and Wendy Ching, Dorothy Sears and Nicholas Webster for their valuable comments. A.H. is a graduate student in the UCSD Medical Scientist Training Program and is supported by a Fellowship from the Whitaker Foundation and the UCSD MSTP Training Grant T35-GM07198. T.I. is supported by NIH grant NIGMS GM070743-01 and a David and Lucille Packard Fellowship Award. S.S. and J.M.O. acknowledge the NIH grants NIGMS K54 GM62114 and NIGMS K54 GM69338, and NIDDK RO1-DK33651, respectively, as well as a Life Science Informatics grant from the State of California, Pfizer La Jolla, and a grant from the Hillblom Foundation. Funding to pay the Open Access publication charges for this article was provided by a Hillblom Foundation Grant.

Conflict of interest statement. None declared.

REFERENCES

1. Hsiao,A., Worrall,D.S., Olefsky,J.M. and Subramaniam,S. (2004) Variance-modeled posterior inference of microarray data: detecting gene-expression changes in 3T3-L1 adipocytes. *Bioinformatics*, **20**, 3108–3127.
2. Consortium,T.G.O. (2001) Creating the Gene Ontology resource: design and implementation. *Genome Res*., **11**, 1425–1433.
3. Baldi,P. and Long,A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.
4. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
5. Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.
6. Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.
7. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.
8. Wingender,E., Chen,X., Fricke,E., Geffers,R., Hehl,R., Liebich,I., Krull,M., Matys,V., Michael,H., Ohnhauser,R. *et al*. (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res*., **29**, 281–283.
9. Beissbarth,T. and Speed,T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.
10. Zhang,B., Schmoyer,D., Kirov,S. and Snoddy,J. (2004) GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics*, **5**, 16.
11. Zhong,S., Tian,L., Li,C., Storch,K.F. and Wong,W.H. (2004) Comparative analysis of gene ontology space under the multiple hypothesis testing framework. *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB2004)*, 16–19 August, Stanford, CA.