Exploratory Visual Analysis of Pharmacogenomic Results

D.M. Reif, S.M. Dudek, C.M. Shaffer, J. Wang, and J.H. Moore

Pacific Symposium on Biocomputing 10:296-307(2005)

EXPLORATORY VISUAL ANALYSIS OF PHARMACOGENOMIC RESULTS

DAVID M. REIF[1,2], SCOTT M. DUDEK[2], CHRISTIAN M. SHAFFER[2], JANEY WANG[2], JASON H. MOORE [1,†]

[1]*Computational Genetics Laboratory, Department of Genetics, Dartmouth Medical School, Lebanon, NH 03756*
[2]*Center for Human Genetics Research, Vanderbilt University Medical School, Nashville, TN 37232*
{reif, dudek, shaffer, wang}@chgr.mc.vanderbilt.edu, jason.h.moore@dartmouth.edu

Comprehensive analysis of expansive pharmacogenomic datasets is a daunting challenge. A thorough exploration of experimental results requires both statistical and annotative information. Therefore, appropriate analysis tools must bring a readily-accessible, flexible combination of statistics and biological annotation to the user's desktop. We present the Exploratory Visual Analysis (EVA) software and database as such a tool and demonstrate its utility in replicating the findings of an earlier pharmacogenomic study as well as elucidating novel biologically plausible hypotheses. EVA brings all of the often disparate pieces of analysis together in an infinitely flexible visual display that is amenable to any type of statistical result and biological question. Here, we describe the motivations for developing EVA, detail the database and custom graphical user interface (GUI), provide an example of its application to a publicly available pharmacogenomic dataset, and discuss the broad utility of the EVA tool for the pharmacogenomics community.

1. Introduction

1.1. *Visualizing results*

Recent years have seen an explosion in the sheer volume of data generated by modern experimental methods. Analysis of pharmacological results can be a maze of complex spreadsheets and arbitrary statistical significance thresholds. Visualization is a proven solution to this challenge of scale. In his work on visualizing quantitative information, Tufte states that "the most effective way to describe, explore, and summarize a set of numbers—even a very large set—is to look at pictures of those numbers. Furthermore, of all methods for analyzing

† Address for correspondence:
706 Rubin Bldg, HB 7937
One Medical Center Dr.
Dartmouth Hitchcock Medical Center
Lebanon, NH 03756, USA
jason.h.moore@dartmouth.edu.

and communicating statistical information, well-designed data graphics are usually the simplest and at the same time the most powerful" [1]. In what would otherwise be a sea of numbers, analysis tools must organize and display results so that readily distinguishable patterns emerge.

1.2. *Other analysis tools currently available*

Other analysis tools have been developed in recent years, such as DAVID [2], FatiGO [3], GoMiner [4], GoSurfer [5], GOTree Machine [6], and Onto-Express [7]. Each of these tools fills a specific niche in the community of genomic analysis methods. However, there are significant limitations that must be addressed.

While the tree-like representations common to many of the aforementioned tools are suited to the hierarchical structure of the Gene Ontology, such a display cannot present the volume of simultaneous information necessary to see otherwise hidden patterns emerge. Many current applications are designed for a single type of input (*e.g.* only GenBank Accession IDs), rather than compatibility across multiple platforms. The sluggishness of many of the available web-based applications can make exploratory analysis frustratingly slow. Many of the current applications take raw data as input, which leaves the user dependent on the statistical tests built into the package, rather than having the flexibility to perform any sort of analysis. This is fundamentally different from a results-based tool that gives the user freedom to implement any type of analysis. Another major concern with exploratory tools is the lack of mechanisms to replicate findings. These gaps in the functionality of other tools currently available leave a critical void in genomic analysis.

1.3. *The motivation for developing EVA*

The Exploratory Visual Analysis software (EVA) was developed to address the limitations of other approaches to analysis of genomic results. Combining flexibility, speed, and visualization of both statistical and annotative information into a single package, EVA fulfills a crucial role in comprehensive pharmacogenomic analysis.

From its inception, EVA was intended to be flexible across a wide range of research goals—allowing a truly exploratory analysis. The software can take as input any kind of statistical result(s) for any number of experiments. The user is thus free to use any statistic of choice or to define a custom statistic, rather than be limited by those implemented in the software. EVA's graphical results display can be organized into nested groupings for any combination of six biological categories: Gene Ontology (GO) [8], Biopath [9], Domain [10], Map Location [11], Chromosome [11], and Phenotype [12]. The statistical

significance of particular subsets of categories or particular genes within a category can be assessed through permutation testing. To complement the statistical analysis, EVA links to multiple annotation sources via Locus Link [11]. This aspect affords immediate evaluation of the pharmacological relevancy of candidate genes or groups of genes. To ensure that the user can replicate findings, EVA incorporates a printable command log feature.

The speed of the EVA tool enhances its interactive flexibility. Because all of the data and links are loaded into memory upon opening the software, the user can switch seamlessly between annotative groupings, statistics, significance levels, and display modes. The permutation testing and reporting features provide an immediate complement to visual exploration.

Visualization is the aspect that binds the various components of EVA into a single coherent exploratory tool. Color choices, sliding significance scales, category display thresholds, and other display parameters are modifiable in real-time according to user preferences and/or analysis goals. EVA can be adjusted to display highly significant genes, marginally significant genes, or both, with cut-offs decided by the user. Most importantly, the graphical nature of EVA facilitates interpretation of multitudes of information simultaneously. This is because the human eye is acutely trained to identify pictorial patterns, while digesting tables of lifeless numbers is a taxing exercise. EVA translates numbers into pictures and *vice-versa*. Thus, graphical discoveries can be verified statistically, and statistical significance can be verified graphically. Famous examples such as Anscombe's Quartet [13] validate the notion that both types of information are essential for confident analysis. The ability of graphics to rationally condense vast amounts of information is essential for evaluating biological systems in which the concerted action of numerous contributing factors is the final determinant of phenotype.

Taken together, the diverse abilities of EVA allow the kind of comprehensive analysis necessary to answer complex pharmacological questions. Because pharmacological phenotypes are the product of myriad interacting factors, the annotative groupings and immediate expert-knowledge links provided by EVA are essential to understanding biological questions of a systemic nature [14]. The synergistic pieces of EVA coalesce into an analysis tool wherein the visual, numerical, and annotative components are mutually complementary. EVA takes analysis beyond spreadsheets of flat statistical results and into the realm of integrated analysis.

1.4. *Application of EVA to a pharmacogenomic dataset*

To demonstrate the power of EVA, we apply the software to the leukemia drug response dataset published by Cheok *et. al* [15]. The study measured the

expression of 9,600 genes in leukemia cells using oligonucleotide microarrays before and after *in vivo* treatment with methotrexate and mercaptopurine given alone or in combination. The gene expression changes induced by these two widely-used antimetabolites are important for understanding the pharmacological action—including side-effects—of common chemotherapy agents, as well as for identifying potential treatment targets. The original authors' conclusions were based upon a number of statistical approaches.

Using only the results of three simple statistical tests as input, we demonstrate the ability of EVA to not only replicate but also extend the findings of the original authors, and to draw new biologically plausible conclusions. By grounding the entire exploratory analysis in biological relevancy, EVA renders moot the usual analytical step of constructing a pharmacologically relevant explanation from a collection of statistical significance values after the fact. Additionally, because EVA displays information graphically, we can see unifying patterns in the results, rather than a list of disjointed significance levels. This lends greater confidence to our conclusions, as correlated patterns are more robust than thousands of measurements taken individually [16]. The exploratory capabilities of EVA allow pharmacological discovery with or without *a priori* hypotheses—giving the user increased power to elucidate novel mechanisms or target pathways. Details of our analysis are given in Section 2.3, and the results are described in Section 3.

2. Methods

2.1. *Details of the EVA database*

The EVA database was designed with versatility, speed, and user-friendliness in mind. The user downloads the client, which provides a portal to the EVA web server via the custom GUI. The web server accesses the EVA database (stored on an Oracle server). This architecture provides ease of security, distributability, and expandability. Changes made by the database administrator pass seamlessly to the user through the EVA web service. Thus, once the user has downloaded the client, updates or expansions of the EVA modules are transparent from the user's perspective. For speed considerations, everything is stored in memory upon loading a particular experiment so that performance is not limited by demand on the web server. This aspect negates the query lag time typical of other analysis packages.

The EVA database schema is available upon request. Briefly, the gene identifier chosen by the user (Affymetrix ID, GenBank ID, or user/custom-defined gene number) links particular genes to the various EVA modules. At present, these modules are Locus Link, Wormbase (*C. Elegans*), and Linkage.

2.2. *Details of the EVA GUI*

The GUI is the portal through which the user manipulates the components of the EVA package (see Figure 1). Presently, the interface is written in Visual Basic, with a platform-independent Java version in production. A brief overview of features is provided here, and a comprehensive, illustrated help menu is included with the software. Upon opening the software, the user supplies a username and secure password. Login grants the user access to interfaces for various administrative tasks, including creating, updating, deleting, or loading experiments and results. Defining a new experiment involves deciding upon a descriptive name, choosing the type of gene identifier (Affymetrix, Locus Link, Wormbase, or Linkage), selecting the statistical tests used, and uploading the text file of results. Once a new experiment has been defined or an existing experiment selected, all of the results and links for that experiment are loaded for viewing on the user's desktop.

EVA displays genes organized according to the selected biological category (Chromosome, Biopath, Domain, Map Location, GO, or Phenotype). Each time a relevant parameter is changed, EVA dynamically resizes all boxes displayed to provide the most efficient use of screen space possible. Additionally, the categories can be re-organized by a second biological group, which will re-partition the results into new group boxes based upon the intersection of the two categories. The differential color display is a function of the significance threshold selected for a particular statistical test. The threshold range is modified via a slider. In order to customize the display, the user can alter both the color palette and number of colors into which the chosen range is divided.

The volume of information displayed on the screen can be modified in a number of ways. The display threshold sets the minimum number of results a group box must contain for it to appear in the display panel. Only groups exceeding this threshold are shown. Additionally, the 'Filter' tool will restrict the categories displayed to those containing the search text.

The 'Find' tool will search the currently selected biological group for the text in question. Once found, the title bar of that results box will be highlighted in yellow, and the display panel will scroll to that box. Clicking 'Find Next' will zoom to subsequent boxes containing the search text.

Hovering over any individual gene square with the mouse brings up a text box containing summary annotation for that gene. Right-clicking accesses a wealth of biological annotation for the selected gene through Locus Link.

EVA includes command log and reporting features. The log can be cleared or saved to file at any time. The reporting feature generates reports listing all results from a single group, from all groups currently on display, or from all

groups in the current biological category. The 'Preferences' menu includes tabs for 'Log,' 'Reports,' and 'Web Service' options. The log options govern which types of information are stored in the log. The reports options determine the organization of the printable analysis reports and the type of file to which they are saved. The web service tab stores the EVA web service URL. Taken together, these features allow findings and/or the accompanying EVA parameters to be written to external files for further study. They also afford a mechanism for replicating findings, which is a vital quality in an exploratory analysis tool.

EVA uses a permutation testing strategy to assess the significance of the statistical results for a biological group. This feature complements visual inspection and provides statistical validation for the relative enrichment of particular groups. A permutation testing significance range can be selected by one of three methods: clicking on a particular gene in a particular group, selecting a range of significance colors for a particular group, or running a batch permutation on a selected significance range for all visible groups. The procedure is as follows:

1. Count the actual number of results within the selected significance range for a particular category.
2. Determine the total number of genes in that category.
3. Fill a box of the same size as that category with a randomly selected set of results.
4. Increment a total counter every time the number of randomly assigned results within the selected significance range is greater than or equal to the actual count from Step 1 above.
5. Perform 1,000 iterations of Steps 3 and 4.
6. Calculate the final p-value by dividing the total counter from Step 4 by 1,000.

This p-value represents the probability of obtaining the observed number of genes within the selected significance range in a given category by chance alone. In the current implementation, it is left to the user to account for multiple testing and the fact that certain annotation categories (*e.g.* GO) may have the same genes appear within several groups. The permutation testing result(s) can be saved in the log and written to an external file for resorting.

2.3. *Details of the application to a pharmacogenomic dataset*

The pharmacogenomic dataset to which we applied EVA is described fully in [15]. Affymetrix oligonucleotide arrays measured the expression levels of

9,600 human genes in acute lymphoblastic leukemia cells isolated from patients both before and after treatment with common chemotherapy agents. Patients

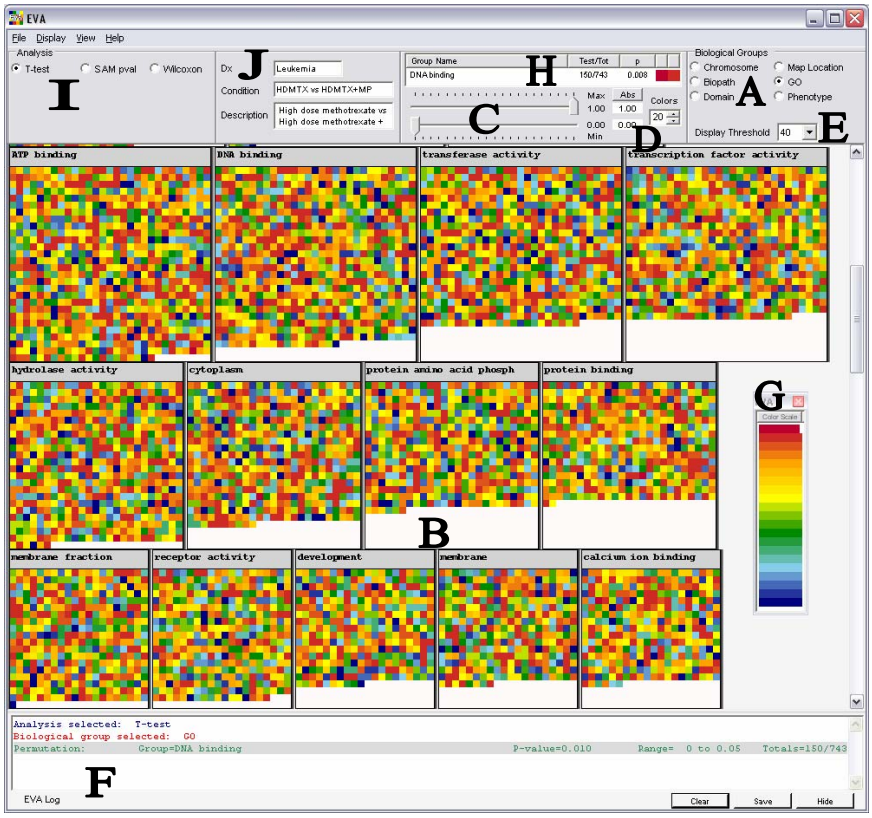

Figure 1. Details of the EVA graphical user interface. Each colored square represents a single gene. Visible features include: (A) – Biological group categories, (B) – Group boxes, (C) – Significance threshold color sliders, (D) – Number of color intervals displayed, (E) – Display threshold, (F) – Log, (G) – Significance range selector for permutation testing, (H) – Permutation testing results (also shown in log), (I) – Analysis types, (J) – Description of current experiment. In addition to options under the File, Display, View, and Help menus, right-clicking accesses a number of features not illustrated here.

were treated with either methotrexate (n = 22), mercaptopurine (n = 12), or a combination of both drugs (n = 10). This dataset provides the rare opportunity to measure the *in vivo* gene expression response of human patients to treatment with pharmacological agents.

The original authors' analysis involved a battery of statistical methods, including principal components analysis, hierarchical clustering, linear discriminant analysis, analysis of variance, support vector machines, and Fisher's exact test (for enrichment of selected GO groups). This approach,

while grounded in statistical rigor, leaves the researcher to bridge the usual gap between statistical results and biological interpretation. This is exactly the role EVA was designed to fulfill.

We chose to demonstrate the power of EVA by feeding it the results of three basic statistical tests on the aforementioned dataset: the Student's t-test, the Wilcoxon Rank-Sum test, and the modified t-test implemented in the Significance Analysis of Microarrays (SAM) procedure. It would also have been possible to give EVA results from more complex statistical tests, such as those used in the original authors' analysis. Instead, we wanted to demonstrate the ability of EVA to translate widely-understood statistical results—of the type that do not require a thorough mastery of complex statistics—into plausible biological conclusions. For each of the three statistical tests, EVA was given the list of resultant p-values for all genes on the array. The p-values indicate the probability of obtaining the observed value of the test statistic for the given gene by chance alone. The statistical analyses were programmed in R version 1.8.1 [17]. The modified SAM t-test was implemented using the "siggenes" package, freely available from Bioconductor (www.bioconductor.org). Figure 2 depicts our statistical analyses and the treatment groups compared using EVA, which were: 1. methotrexate alone versus methotrexate plus mercaptopurine, 2. mercaptopurine alone versus methotrexate plus mercaptopurine, and 3. methotrexate alone versus mercaptopurine alone.

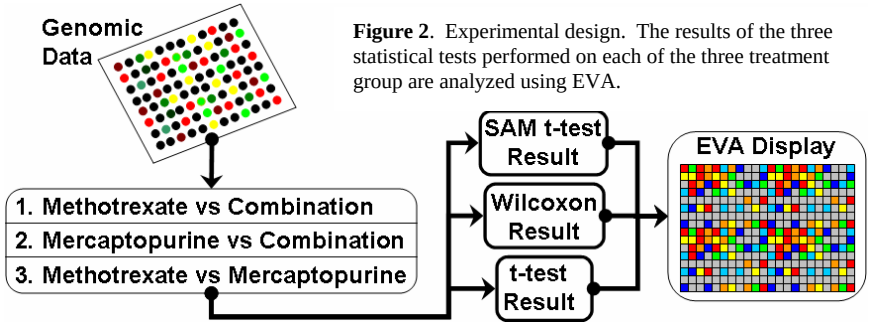

Figure 2. Experimental design. The results of the three statistical tests performed on each of the three treatment group are analyzed using EVA.

3. Results

3.1. *Replication of original authors' findings using EVA*

Using the results of the three basic statistical tests outlined above as input for EVA, we were able to replicate the major findings of the original authors. The original authors' main conclusions were that changes in gene expression are treatment-specific and that gene expression can illuminate differences in cellular response to drug combinations versus single agents. Through EVA's visualization of statistical results, these conclusions were immediately apparent.

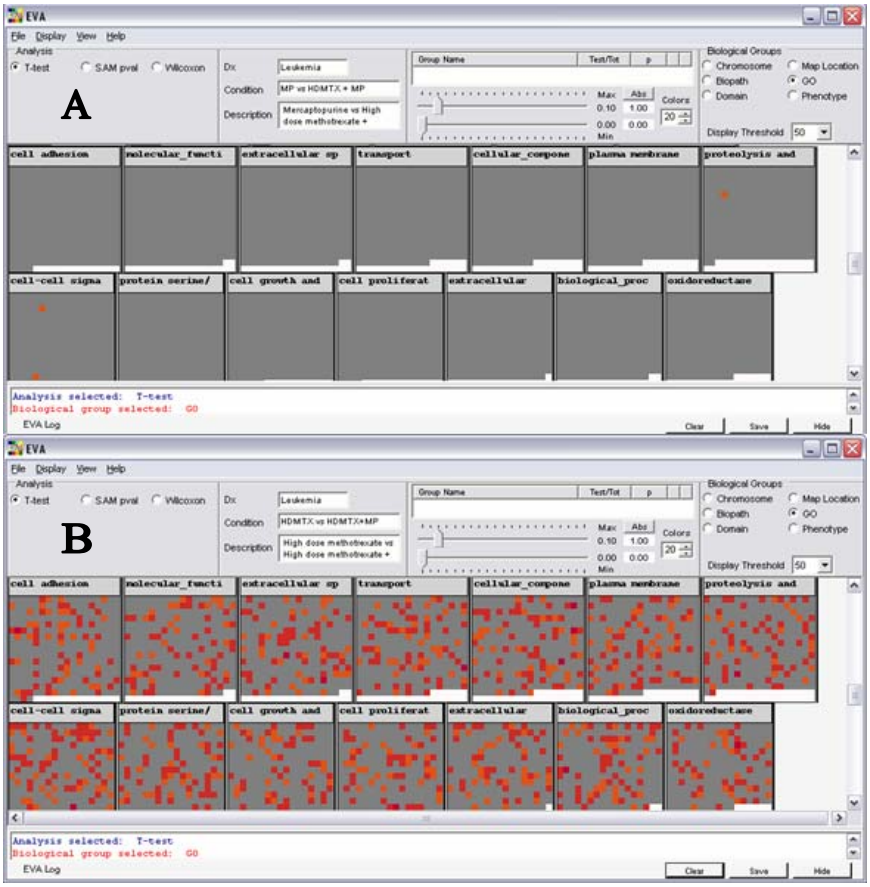

Figure 3. Visual inspection of statistical results for the treatment comparisons (A) – mercaptopurine versus combination and (B) – methotrexate versus combination. For each comparison, colored squares indicate p-values at or below the 0.05 (red) or 0.10 (orange) significance levels. The relative paucity of colored squares in (A) contrasted with the abundance in (B) indicates that there are fewer differentially expressed genes when statistically comparing mercaptopurine chemotherapy to combination chemotherapy.

The global patterns of gene expression were readily discernible between the three treatment groups, and the methotrexate plus mercaptopurine combination treatment exhibited an expression profile distinct from that of either agent given alone.

Demonstrating EVA's flexibility to incorporate any type of result into the analysis, we also took the list of genes highlighted by the original authors as input for the methotrexate alone versus combination chemotherapy comparison. Because EVA organized these genes into their annotation categories, we could instantly see where these individual genes fit into the broader biological picture.

EVA's reporting feature and links to the public annotation databases afforded one-stop evaluation of the biological relevancy of the genes in our list.

3.2. *Novel findings using EVA*

Drawing on exploratory visual, statistical, and annotative abilities of EVA enabled us to draw new biologically credible conclusions. Representative findings reached by starting at each of these three exploratory avenues are outlined below.

Upon visual inspection in EVA, it was immediately apparent that the gene expression pattern of the methotrexate alone treatment differs markedly from the combination chemotherapy, whereas the mercaptopurine alone treatment shows a gene expression pattern that was relatively closer to the combination chemotherapy (Figure 3). Corroborating visual evidence was provided by comparing the two treatments alone, where the global gene expression pattern resembled that of the methotrexate versus combination chemotherapy comparison. The next step called upon EVA's ability to back this conclusion statistically, and we found no genes significant at the 0.05 level and only a sparse few significant at 0.10 by any of the three statistical tests for the mercaptopurine versus combination comparison. This suggests that changes in gene expression induced by the combination chemotherapy are dominated by the action of mercaptopurine.

Starting with EVA's statistical capabilities, permutation testing for enrichment of biological categories with respect to the differences in gene expression patterns comparing methotrexate versus the combination treatment or mercaptopurine alone revealed a number of relevant biological findings. For example, there was a marked difference in expression of genes involved in cytoskeletal function. Statistically significant categories included the GO groups "epidermal differentiation" and "structural molecule activity," the Domain groups "kinesin like protein," "spectrin," "tubulin," and "gamma tubulin," and the Map Location 12q13, on which many of the genes in these categories are found. Keratins, a major component of hair follicles, are found throughout these annotative groups, which makes pharmacological sense because hair loss is one of the characteristic symptoms of methotrexate chemotherapy, though not mercaptopurine.

Incorporating expert-knowledge into the analysis, EVA's straightforward links to annotative information shed light on more interesting connections. For instance, because chemotherapy drugs affect the cell cycle, it is logical to look into that Biopath group. Sure enough, permutation testing showed the "cell cycle" group to be significantly enriched with respect to the number of genes differentially expressed at the 0.05 level of significance (Figure 4). Many of

these genes also appear in the Domain groups “DNA topoisomerase,” “G-protein beta subunit,” and “NERF transcription factor,” as well as the GO groups “DNA binding,” “DNA topological change,” “GTP binding,” “GTPase activator activity,” “kinase activity,” and “nucleotide excision repair,” all of which are significant or near-significant by permutation testing. Importantly, these relevant biological groups near the significance borderline would have been missed by a purely statistical analysis.

Figure 4. Permutation testing for significant enrichment of the Biopath group ‘Cell cycle’. The Find tool was used to highlight the ‘Cell cycle’ group box. The permutation testing results are shown in the permutation testing results display and the log (see Figure 2). The probability of observing the given number of differentially expressed genes significant at the 0.05 level by chance alone is less than 0.001 for this group box.

4. Discussion

4.1. *Utility of EVA for pharmacogenomics*

As demonstrated by our results, EVA is adept at integrating multiple types of information to build cohesive biological conclusions supported by a variety of sources. This is vital in a field such as pharmacogenomics, where the cost of following false leads is prohibitively high. With EVA, the exploration of results can start down any of three avenues—visual, statistical, or annotative—to reflect the expertise or prior notions of the user. The various aspects of EVA are mutually complementary, and the flexibility, speed, and user-friendliness of the EVA interface allow users to move effortlessly between these three avenues.

EVA bridges the gap between raw statistical output and biological discovery, empowering the researcher to biologically validate statistical findings and to statistically test biological findings. The researcher can then plan the next experimental step by linking results to bodies of literature for particular genes.

4.2. *Future directions*

The development of EVA is an ongoing process. Future studies will incorporate results from machine learning and multivariate statistical methods. There are plans to integrate the tool with other publicly available data sources, including those for model organisms. Additionally, while EVA was developed for genomic applications, it can be naturally extended to genetic or proteomic analyses. A command line interface will allow programmable analyses and provide a mechanism to combine EVA with output from other analysis tools, such as sequence homology engines or alternative permutation testing strategies that address the issues of multiple testing and non-independence across tests. The new platform-independent Java version of the EVA GUI will be available at no cost to academic users. Contact the authors for distribution information.

Acknowledgments

This work was supported by National Institutes of Health grants LM07613, HL68744, and AI59694. This work was also supported by generous funds from the Robert J. Kleberg, Jr. and Helen C. Kleberg Foundation.

References

1. E. R. Tufte, *The Visual Display of Quantitative Information*, 2nd ed, (Graphics Press, Cheshire, CT, 2001).
2. G. Dennis, Jr. et al., *Genome Biol.* **4**, 3 (2003).
3. F. Al Shahrour, R. Diaz-Uriarte, and J. Dopazo, *Bioinf.* **20**, 578-580 (2004).
4. B. R. Zeeberg et al., *Genome Biol.* **4**, R28 (2003).
5. S. Zhong et al., *Applied Bioinf.* (*In press*), (2004).
6. B. Zhang, D. Schmoyer, S. Kirov, J. Snoddy, *BMC Bioinf.* **5**, 16 (2004).
7. S. Draghici et al., *Nucl. Acids Res.* **31**, 3775-3781 (2003).
8. M. A. Harris et al., *Nucl. Acids Res.* **32**, D258-D261 (2004).
9. M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, *Nucl. Acids Res.* **32**, D277-D280 (2004).
10. A. Marchler-Bauer et al., *Nucl. Acids Res.* **31**, 383-387 (2003).
11. K. D. Pruitt and D. R. Maglott, *Nucl. Acids Res.* **29**, 137-140 (2001).
12. McKusick-Nathans Institute for Genetic Medicine, *Online Mendelian Inheritance in Man*, National Center for Biotechnology Information, National Library of Medicine (2004).
13. F. J. Anscombe, *American Statistician* **27**, 17-21 (1973).
14. L. Hood, *Mech. Ageing Devel.* **124**, 9-16 (2003).
15. M. H. Cheok et al., *Nat. Genet.* **34**, 85-90 (2003).
16. P. O. Brown and D. Botstein, *Nat. Genet.* **21**, 33-37 (1999).
17. R. Ihaka and R. Gentleman, *J. Comp. Graph. Stat.* **5**, 3 (1996).