International Journal of Knowledge Discovery in Bioinformatics
Volume 8 • Issue 1 • January-June 2018

GEView (Gene Expression View) Tool for Intuitive and High Accessible Visualization of Expression Data for Non-Programmer Biologists

Libi Hertzberg, Shalvata Mental Health Center, Affiliated to the Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel & Weizmann Institute of Science, Rehovot, Israel

Assif Yitzhaky, Weizmann Institute of Science, Rehovot, Israel

Metsada Pasmanik-Chor, Bioinformatics Unit, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel

ABSTRACT

This article describes how the last decade has been characterized by the production of huge amounts of different types of biological data. Following that, a flood of bioinformatics tools have been published. However, many of these tools are commercial, or require computational skills. In addition, not all tools provide intuitive and highly accessible visualization of the results. The authors have developed GEView (Gene Expression View), which is a free, user-friendly tool harboring several existing algorithms and statistical methods for the analysis of high-throughput gene, microRNA or protein expression data. It can be used to perform basic analysis such as quality control, outlier detection, batch correction and differential expression analysis, through a single intuitive graphical user interface. GEView is unique in its simplicity and highly accessible visualization it provides. Together with its basic and intuitive functionality it allows Bio-Medical scientists with no computational skills to independently analyze and visualize high-throughput data produced in their own labs.

KEYWORDS

ANOVA, Batch Correction, Bio-Med, Box Plots, High-Throughput, Next Generation Sequencing, PCA, Proteomics

INTRODUCTION

Advanced high-throughput technologies have been developed and extensively used in the last decades (microarrays since the mid-1980s (Schena, Shalon, Davis, & Brown, 1995) and Next Generation Sequencing (NGS) (Goodwin, McPherson, & McCombie, 2016) since the mid-2000s). Mass spectrometry technologies were originally invented almost 100 years ago and further developed during the 1990s (Glish & Vachet, 2003). All these technologies and others are being routinely used by Bio-medical (Bio-Med) scientists for production of high-throughput data. However, currently, it is very difficult for Bio-Med researchers with no computational skills to analyze their own data. Various free, online and user-friendly tools have been developed and are routinely used for gene

DOI: 10.4018/IJKDB.2018010107

expression and proteomics data analysis. For example, Expander (Ulitsky et al., 2010), Chipster (Kallio et al., 2011), SAM (Tusher, Tibshirani, & Chu, 2001), Limma (Ritchie et al., 2015), DESeq (Love, Huber, & Anders, 2014), Morpheus (https://software.broadinstitute.org/morpheus/), MeV (Howe, Sinha, Schlauch, & Quackenbush, 2011) for gene expression analysis, and MaxQuant (Cox & Mann, 2008) for proteomics data analysis. However, these tools usually perform a large range of tasks, which are sometimes not straightforward and may necessitate some computational skills. In addition, they usually lack a single-gene trivial visualization (such as box plots), which is a basic request of lab researchers in order to study the expression pattern of a specific gene of interest. While the Morpheus tool (https://software.broadinstitute.org/morpheus/) does provide box plots for single genes, it has the limitation of not providing differential expression statistical analysis of more than two groups of samples. The result is that many Bio-Med researchers are reluctant to use such tools and often outsource for the analysis of their data. Outsourcing creates a gap in time and expenditure of financial resources, in addition to the fact that it is performed by a computational expert, who doesn't necessarily have a biological view of results. Thus, there is a need for a more basic and intuitive tool, with functionalities such as quality control measures, simple statistics, differential expression analysis and intuitive visualization of the results, including single-gene visualizations.

GEView is a free, easily downloadable tool suitable for Windows and Linux platforms, written in MATLAB (https://www.mathworks.com). It provides intuitive quality control analysis of the experiment, outlier exclusion and easy identification of differential expression of gene, microRNA or protein data. The great advantage of GEView is its usage simplicity and the output Excel table which provides immediate visualization of the analyzed data on a single-gene basis. It serves all available expression platforms, given that expression values for each gene/microRNA/protein are provided in a tab-delimited file as input, after normalization. GEView is freely available to non-commercial users and can be downloaded from http://www.weizmann.ac.il/complex/compphys/software/geview/. It includes a convenient "help" menu and an example test data (Teuffel et al., 2004).

We believe GEView may be highly useful for Bio-Med researchers who are not computational experts. Such researches, who were previously reluctant to use computational tools and were dependent on external bioinformatics services, have now an opportunity to analyze their own data. In addition, the readily available single-gene graphical representation enables laboratory scientists to simply return to the results table each time they become interested in a new gene or protein and explore its expression pattern in their data. Thus, by improving the access to basic expression analysis, GEView improves the ability of laboratory scientists to produce significant results from their various experiments.

IMPLEMENTATION

In this paper we present GEView, a useful tool for various types of expression analysis (genes, microRNAs and proteins). Our aim is to enable the Bio-Med researchers with no computer-science skills, to simply analyze and visualize high-throughput data produced in their wet experiments (see workflow in Figure 1), using an intuitive user interface (see Figure 2).

GEView performs a basic quality control step of Principal Component Analysis (PCA) (Jolliffe & Cadima, 2016). The researcher may correlate PCA output with experimental design and sample replicates and examine whether all biological replicates cluster together (otherwise, unknown variables may have affected the data). It may also indicate biological outliers among the samples. In addition, various experimental batches resulting from experimental design and/or other technical constraints can be easily corrected using GEView, implementing the ComBat batch correction algorithm (Johnson, Li, & Rabinovic, 2007; Leek, Johnson, Parker, Jaffe, & Storey, 2012). Moreover, GEView uses Analysis Of Variance (ANOVA) (Bewick, Cheek, & Ball, 2004) (http://www.mathworks.com/help/stats/anova1.html) to identify genes (/microRNAs /proteins) which are differentially expressed between the various experimental conditions. The single output Excel file provides immediate visualization of all genes' differential expression in the various experimental conditions (box plots) together with

Figure 1. GEView flow chart

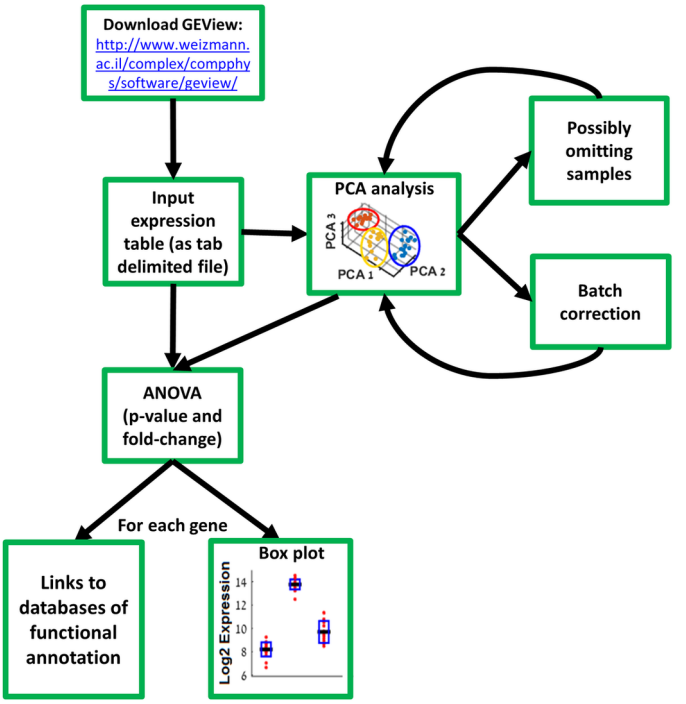
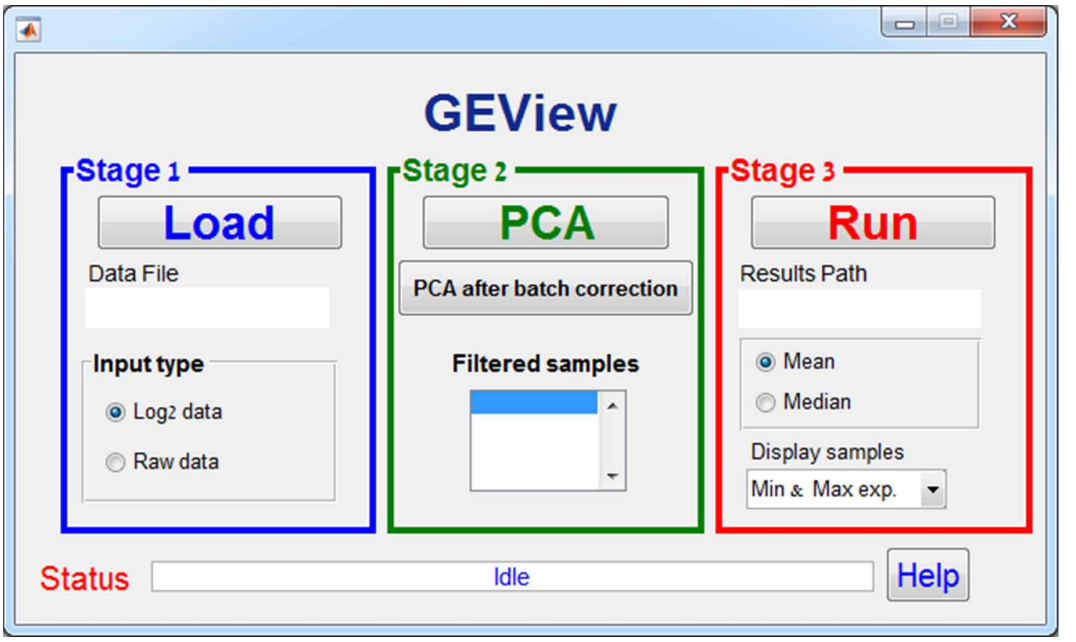


Figure 2. GEView user interface



ANOVA P-values and fold-change values, through a click on each gene-link in the output file. The expression pattern of each single gene can be easily and immediately viewed by clicking on the link in the output file, which opens a pre-computed box plot (for illustration, see an example of a results table in Table 1; the links in the column "Box Plots" open a box plot for each of the genes). In order to complete data examination and analysis, links are also provided for GeneCards (Safran et al., 2010), UCSC (Kent et al., 2002) and NCBI (https://www.ncbi.nlm.nih.gov/gene) databases, for each specific entry (for illustration, see the links in columns 3-5 of Table 1).

RESULTS

Test Data: The "Zurich" Dataset

In order to demonstrate GEView functionality, we use a gene expression dataset of leukemia blood samples that were processed at the department of Oncology, University Children's Hospital, Zurich, Switzerland (Teuffel et al., 2004) and measured by Affymetrix HG-U133A microarrays. We utilize as an example three sample groups that represent three leukemia subtypes, each with different genomic characteristics: E2A-PBX1 (E2A; 5 samples), High hyperdiploid (HD; 4 samples) and TEL-AML1 (TEL; 8 samples). E2A-PBX1 is characterized by a translocation that fuses two genes, E2A and PBX1, High hyperdiploid is characterized by gain of chromosomes (overall 51-67 chromosomes) and TEL-AML1 is characterized by a translocation that fuses two genes, TEL and AML1.

Data Input File and Details for Obtaining It

A tab-delimited text file with normalized expression values for each entry (resulting from microarray, NGS or proteomics) is required as input by GEView. Data can be provided as raw expression data or log 2 transformed, according to the user`s preference. The format is fully described in the "help" document; an example input table is given in Additional File 1 (http://www.weizmann.ac.il/complex/compphys/software/geview/0.4/Additional_File1_Zurich_HD_TEL_E2A_data_labeled.txt).

Various tools are available for free download in order to obtain the normalized expression values. EXPANDER tool (Ulitsky et al., 2010) (http://acgt.cs.tau.ac.il/expander/) or the Expression console tool (for Affymetrix microarray data; https://www.affymetrix.com/support/technical/byproduct.affx?product=expressionconsole) can be used for microarray data, Chipster (Kallio et al., 2011) (http://chipster.csc.fi/) may be used for NGS data, MaxQuant (Cox & Mann, 2008) may be used for proteomics data and R software (https://www.r-project.org/), which requires programing skills, can be used for all types of data. Commercial tools such as Partek Genomics suite (http://www.partek.com/) and MATLAB (https://www.mathworks.com) may also be used.

Load Data

Use the "Load" button (blue; see Figure 2) in order to browse through the input data file in the computer directory. After choosing the data file, the status bar will indicate when its loading is completed.

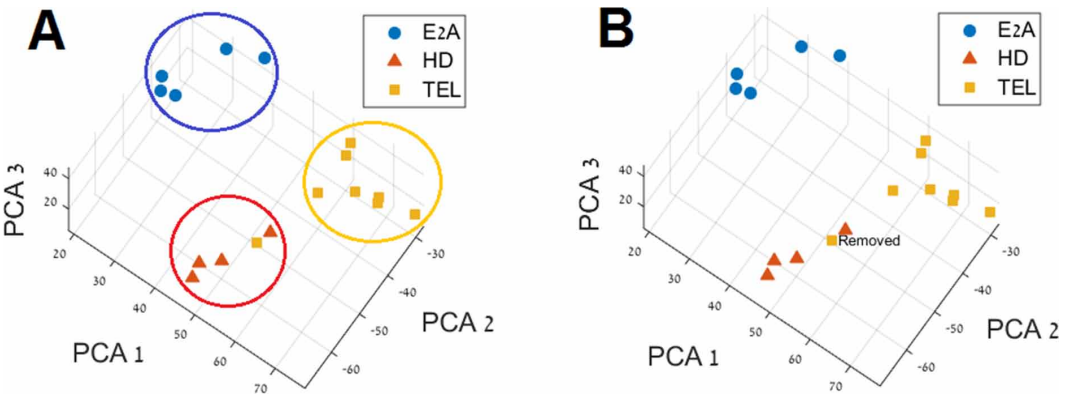
Principal Component Analysis (PCA) and Batch Effect Correction

In order to evaluate sample quality and homogeneity, PCA (Jolliffe & Cadima, 2016) may be plotted (using the "PCA" green button, see Figure 2), implemented by MATLAB software (http://www.mathworks.com/help/stats/pca.html). The PCA calculation is based on the 1,000 highest variance genes (if there exists less than 1000 genes, then all genes are used). In Figure 3A the PCA of the "Zurich" test data is plotted. It can be seen that one of the TEL (yellow square) samples is clustered within the HD group (red triangles), and not with its own group. Using GEView, the researcher can exclude this sample from the analysis where, for instance, there is evidence that it was mislabeled, by right clicking on it (see Figure 3B). It should be noted that such a decision should be taken carefully, as it has the potential to bias the results.

Table 1. Output Excel table

Probe Set ID	Gene Symbol	GeneCards Link	UCSC Link	NCBI Link	ANOVA P-Value	Q-Value (Corrected)	Box Plots	E2A/HD Fold-Change	p-Value	E2A/TEL Fold-Change	p-Value	HD/TEL Fold-Change	p-Value
212148_at	PBX1	GeneCards	UCSC	NCBI	2.30E-15	3.80E-11	Figure	20.3	1.90E-09	21	1.90E-09	1.03	0.9
212151_at	PBX1	GeneCards	UCSC	NCBI	3.50E-13	2.90E-09	Figure	25.2	1.90E-09	20.6	1.90E-09	0.817	0.2
200953_s_at	CCND2	GeneCards	UCSC	NCBI	2.10E-07	0.00022	Figure	0.0787	1.70E-07	0.223	1.40E-05	2.83	0.00089
201005_at	CD9	GeneCards	UCSC	NCBI	1.30E-06	0.00055	Figure	0.874	0.92	10.7	6.00E-06	12.3	6.90E-06
205253_at	PBX1	GeneCards	UCSC	NCBI	3.70E-06	0.0011	Figure	10.8	2.60E-05	10.8	6.10E-06	1	1
204849_at	TCFL5	GeneCards	UCSC	NCBI	2.20E-05	0.0029	Figure	0.639	0.34	0.152	2.70E-05	0.238	0.00069
200951_s_at	CCND2	GeneCards	UCSC	NCBI	3.80E-05	0.004	Figure	0.372	3.10E-05	0.744	0.083	2	0.00049
221773_at	ELK3	GeneCards	UCSC	NCBI	0.0011	0.024	Figure	0.195	0.0016	0.279	0.0039	1.43	0.56
200952_s_at	CCND2	GeneCards	UCSC	NCBI	0.13	0.27	Figure	0.789	0.18	0.812	0.18	1.03	0.97
206127_at	ELK3	GeneCards	UCSC	NCBI	0.28	0.43	Figure	0.988	0.99	0.905	0.32	0.916	0.45

Figure 3. PCA quality control visualization: A) PCA is plotted for the "Zurich" dataset. Each of the experimental conditions is shaped and colored differently: E2A in blue circles, HD in red triangles and TEL in yellow squares. The colored large circles were added in order to emphasize the clustering into 3 biological groups. B) PCA is plotted for the "Zurich" dataset in the same manner as in A), after removing the outlier TEL sample (marked by the text "Removed").



If various batches are seen by PCA or are known in advance, normalization according to experimental batches may be applied. "PCA after batch correction" button may be used to re-view the new PCA after the batch correction.

Batch Correction Example

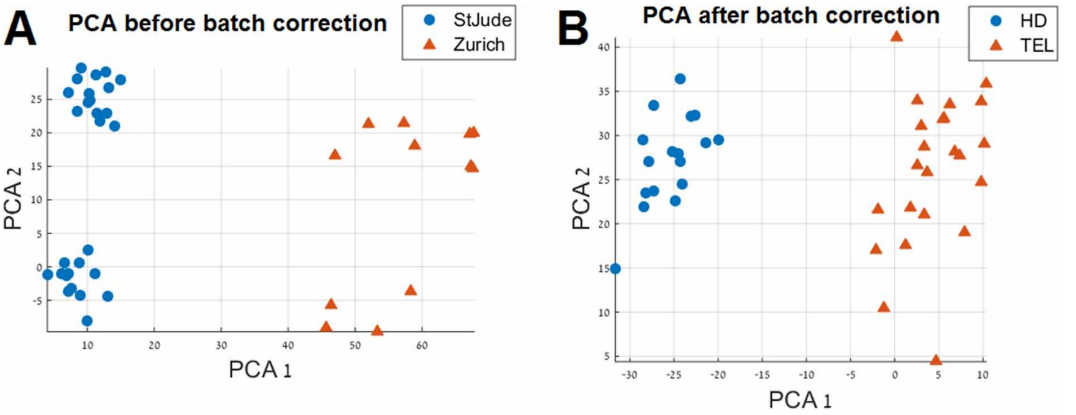
When combining datasets from various origins (for example, in case datasets were processed in different dates or laboratories), the samples are often clustered according to the technical origin instead of the experimental conditions. GEView implements the ComBat batch correction algorithm (Johnson et al., 2007; Leek et al., 2012) to enable correction of such technical artifacts. In the following example, we added to our test data that were processed at the department of Oncology, University Children's Hospital Zurich, Switzerland (Teuffel et al., 2004) (labeled "Zurich") an additional dataset, that was processed at St Jude Children's Research Hospital, Memphis, TN (Ross et al., 2003) (labeled "StJude"). These datasets share two of the sample groups, TEL and HD. However, it can be seen in Figure 4A, where the samples are shaped and colored according to data origin (batches: "Zurich" – red triangles, "StJude" – blue circles), that the samples are grouped according to the batches; Zurich to the right and StJude to the left. After applying the batch correction algorithm (using the "PCA after batch correction" button, see Figure 2), it can be seen in Figure 4B that the batch effect disappeared, and the samples are grouped solely according to the experimental conditions (HD and TEL).

Run Statistical Analysis

By pressing the "RUN" red button (see Figure 2), ANOVA is applied for each probe set (gene/microRNA/protein), testing whether the means of its expression significantly differ between each of the different conditions (see http://www.mathworks.com/help/stats/anova1.html for more details). P-values and Q-values (corrected P-values for multiple comparisons (Benjamini & Hochberg, 1995) are being calculated for each probe set. Running the statistical analysis of a typical dataset containing about 20,000 entries may take up to a few hours, using a computer with basic requirements such as 8GB RAM and Intel Core i7 processor of 3.40 GHz. Running progress can be viewed by the status line at the bottom of the user interface (see Figure 2).

This stage generates a single Excel file which summarizes in each line the results for a specific probe set. Table 1 represents the resulting Excel table for 10 probe sets (see Additional File 2 for the complete ANOVA table for ~ 17,000 probe sets http://www.weizmann.ac.il/complex/compphys/software/geview/0.4/Additional_File2_results_Zurich_HD_TEL_E2A_data_labeled.xlsx).

Figure 4. Batch correction example: A) PCA is plotted for the union of HD and TEL samples from both "Zurich" and "StJude" datasets, before the batch correction is applied. The samples are shaped and colored by batch (red triangles for "Zurich" and blue circles for "StJude"). B) PCA is plotted for the union of HD and TEL samples from both "Zurich" and "StJude" datasets after the batch correction was applied. The samples are shaped and colored by experimental conditions (blue circles for HD, red triangles for TEL).



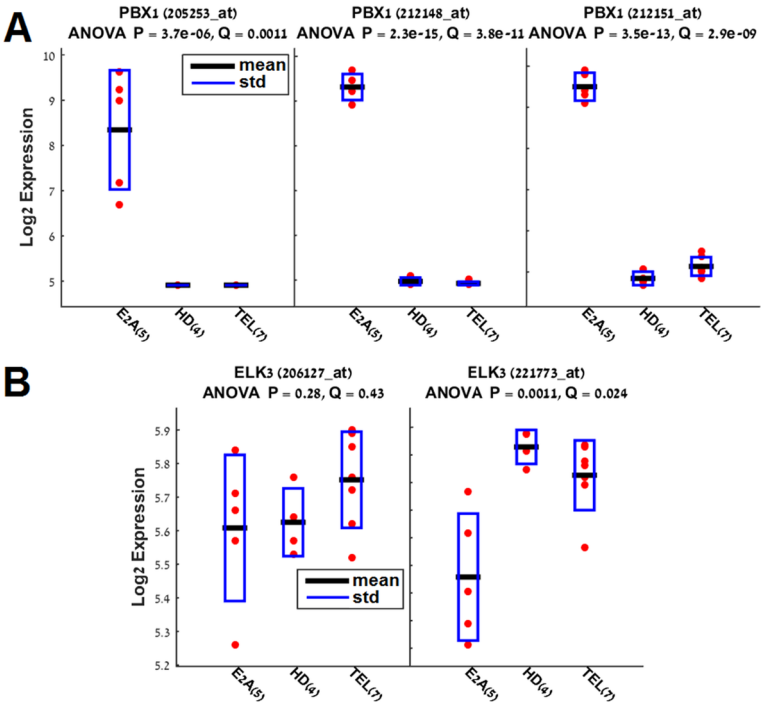
Exploring the Results

The output single Excel file provides immediate visualization of all genes' differential expression in the various experimental conditions (box plots) together with ANOVA P-values and fold-change values, through a click on each gene-link in the output file (Table 1 contains 10 rows, out of ~17,000, for illustration; the links on the "Box Plots" column can be clicked on). For example, in Figure 5, box plots of two genes are plotted, PBX1 and ELK3.

Each row represents a probe set ID, identified by a gene symbol (second column). Please note that 3 different probe set IDs for PBX1 gene and 2 for ELK3 gene are present. The Box plots column contains a link to a box plot of the mean expression values with standard errors and statistics resulting from ANOVA for each probe set. For each entry, links to GeneCards, UCSC and NCBI genome browser databases are provided. In addition, the same Excel file contains fold change values between each pair of experimental conditions. In case there are more than two experimental conditions, Tukey's test (Tukey JW, n.d.) for multiple comparisons is performed (http://www.mathworks.com/help/stats/multcompare.html), comparing between each pair, and the p-values are given.

As shown in Figure 5, probe sets belonging to the same gene are plotted together. This is done in order to estimate whether the different probe sets that belong to the same gene share a similar expression pattern (if they do, it significantly strengthens the validity of the pattern). It can be seen that PBX1 has 3 probe sets, all of which show a similar and significant expression pattern – high expression in E2A-PBX1 sample group and low in the rest (Figure 5A). E2A-PBX1 leukemia subtype is characterized by a fusion between the transcription factor E2A and PBX1 gene. While normally PBX1 is not expressed in lymphoid cells, the fusion with E2A causes its expression (Aspland, Bendall, & Murre, 2001). Thus, PBX1 is indeed expected to be expressed in E2A-PBX1 sample group and not expressed in the rest. The concordance between all 3 probe sets (Figure 5A) significantly increases the validation of this pattern. Another example of the gene ELK3, can be seen in Figure 5B. It has two probe sets, one showing a significant differential expression (Q= 0.024), while the other shows a different, non-significant pattern. Thus, the expression pattern of ELK3 is less validated than that of PBX1. When there is a need to decide on which gene to focus and perform further investigations and wet experiments, this information is crucial. Therefore, the simple grouping of the box plots of different probe sets of the same gene can improve the decision-making process.

Figure 5. Box plots of PBX1 and ELK3 genes: A) Box plots for the 3 probe set IDs present in the "Zurich" dataset for the gene PBX1 are plotted. In the title of each plot, the gene symbol and probe set ID (in parentheses) are presented. In addition, the ANOVA P and Q values (corrected p-value (Benjamini & Hochberg, 1995)) are provided. The Y-axis represents the Log2 expression value and the X-axis represents the 3 experimental conditions (see E2A, HD and TEL X-labels). The number of samples in each experimental condition is written in parentheses. Each sample expression value is marked by a red dot and the mean and standard deviation in each experimental condition are marked by a black line and a blue box, respectively. B) Box plots for the 2 probe set IDs present in the "Zurich" dataset for the gene ELK3 are plotted in the same manner as in A.



HELP DOCUMENT

A comprehensive Help document can be easily accessed through GEView (press the "Help" button in the right lower corner of the user interface, as plotted in Figure 2; the full Help document is also given as Additional File 3 http://www.weizmann.ac.il/complex/compphys/software/geview/0.4/Additional_File3_help.docx). The Help document contains:

- Installation instructions;
- Input format description for all data types;
- Description of the analysis flow, including a detailed description of the user interface functionality and the running parameters;
- Detailed description of each analysis type provided by GEView, including results of the "Zurich" test dataset.

DISCUSSION

There are many tools that perform gene or protein expression analysis. We offer GEView, a tool that implements basic analyses and intuitive visualization of the results for the non-programmer biologist. The great advantage of GEView is its intuitive, accessible and comprehensive results table, given in an Excel file with links to graphics and databases. Specifically, it provides a single results table,

including pre-calculated ANOVA p-values and links to pre-computed figures of the expression pattern of each of the genes/microRNAs/proteins across the experimental conditions. This is a significant improvement to the commonly available results tables, which usually include only p-values and fold-change differences and don't provide visualization of expression patterns of single genes. The intuitive graphical representation, using box plots, enables the laboratory scientist to simply investigate many genes of interest. Gene expression patterns may easily be explored in the same results table, without the need for further external resources, through direct links to GeneCards, UCSC and NCBI databases for each gene (see Table 1).

CONCLUSION

GEView was developed for Bio-Med researchers who are not computational experts. We believe that researchers who were previously dependent on external services for the analysis of their own data, will have a new opportunity to analyze expression results with GEView. Upon input of expression data of various types, GEView provides PCA analysis, batch correction and analysis of differential expression, with immediate graphical view and functional annotation of the results, in a single simple Excel file. While all algorithms used are published, implemented by standard and validated MATLAB functions (in the case of ComBat algorithm, using the original R software code (Leek et al., 2012), wrapping them together in a simple and intuitive tool with immediate access to graphical representation on a single-gene basis will enable the Bio-Med researcher to:

1. Perform basic analyses without the need for external resources;
2. Have immediate access to the visualization and statistics of the data;
3. Compare the expression pattern validity of different gene probes for better wet experiments decision making;
4. Improve their ability to produce significant results out of their various experiments.

COMPETING INTERESTS

The authors declare that they have no competing interests.

FUNDING

This work was funded by a NARSAD Young Investigator Grant from the Brain & Behavior Foundation, the National Institute of Psychobiology in Israel (Grant number 132-14-15) and the Leir Charitable Foundation.

ACKNOWLEDGMENT

We deeply thank Professor Eytan Domany for his wise advice and for providing the conditions to develop GEView. We thank Dr. David Gurwitz for providing the first dataset for testing and his wise advices.

REFERENCES

Aspland, S. E., Bendall, H. H., & Murre, C. (2001). The role of E2A-PBX1 in leukemogenesis. *Oncogene*, *20*(40), 5708–5717. doi:10.1038/sj.onc.1204592 PMID:11607820

Bengtsson, H. (n.d.). R.matlab. Retrieved from https://github.com/HenrikBengtsson/R.matlab

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300.

Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 9: One-way analysis of variance. *Critical Care (London, England)*, *8*(2), 130–136. doi:10.1186/cc2836 PMID:15025774

Bornhauser, B. C., Bonapace, L., Lindholm, D., Martinez, R., Cario, G., Schrappe, M., & Bourquin, J.-P. et al. (2007). Low-dose arsenic trioxide sensitizes glucocorticoid-resistant acute lymphoblastic leukemia cells to dexamethasone via an Akt-dependent pathway. *Blood*, *110*(6), 2084–2091. doi:10.1182/blood-2006-12-060970 PMID:17537996

Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, *26*(12), 1367–1372. doi:10.1038/nbt.1511 PMID:19029910

Glish, G. L., & Vachet, R. W. (2003). The basics of mass spectrometry in the twenty-first century. *Nature Reviews. Drug Discovery*, *2*(2), 140–150. doi:10.1038/nrd1011 PMID:12563305

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews. Genetics*, *17*(6), 333–351. doi:10.1038/nrg.2016.49 PMID:27184599

Howe, E. A., Sinha, R., Schlauch, D., & Quackenbush, J. (2011). RNA-Seq analysis in MeV. *Bioinformatics (Oxford, England)*, *27*(22), 3209–3210. doi:10.1093/bioinformatics/btr490 PMID:21976420

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (Oxford, England)*, *8*(1), 118–127. doi:10.1093/biostatistics/kxj037 PMID:16632515

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *374*(2065), 20150202. doi:10.1098/rsta.2015.0202 PMID:26953178

Kallio, M. A., Tuimala, J. T., Hupponen, T., Klemelä, P., Gentile, M., Scheinin, I., & Korpelainen, E. I. et al. (2011). Chipster: User-friendly analysis software for microarray and other high-throughput data. *BMC Genomics*, *12*(1), 507. doi:10.1186/1471-2164-12-507 PMID:21999641

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Res., 12*(6), 996–1006. doi:10.1101/gr.229102

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics (Oxford, England)*, *28*(6), 882–883. doi:10.1093/bioinformatics/bts034 PMID:22257669

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. doi:10.1186/s13059-014-0550-8 PMID:25516281

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47–e47. doi:10.1093/nar/gkv007 PMID:25605792

Ross, M. E., Zhou, X., Song, G., Shurtleff, S. A., Girtman, K., Williams, W. K., & Downing, J. R. et al. (2003). Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood*, *102*(8), 2951–2959. doi:10.1182/blood-2003-01-0338 PMID:12730115

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., & Lancet, D. et al. (2010). GeneCards Version 3: The human gene integrator. *Database : The Journal of Biological Databases and Curation*, *2010*(0), baq020. doi:10.1093/database/baq020 PMID:20689021

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray. *Science*, *270*(5235), 467–470. doi:10.1126/science.270.5235.467 PMID:7569999

Teuffel, O., Dettling, M., Cario, G., Stanulla, M., Schrappe, M., Bühlmann, P., & Schäfer, B. W. et al. (2004). Gene expression profiles and risk stratification in childhood acute lymphoblastic leukemia. *Haematologica*, *89*(7), 801–808. PMID:15257931

Tukey JW. (n.d.). Comparing individual means in the analysis of variance.

Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(9), 5116–5121. doi:10.1073/pnas.091062498 PMID:11309499

Ulitsky, I., Maron-Katz, A., Shavit, S., Sagir, D., Linhart, C., Elkon, R., & Shamir, R. et al. (2010). Expander: From expression microarrays to networks and functions. *Nature Protocols*, *5*(2), 303–322. doi:10.1038/nprot.2009.230 PMID:20134430

APPENDIX

Availability and Requirements

Project name: GEView
Project site: http://www.weizmann.ac.il/complex/compphys/software/geview/_
Programming Language: MATLAB (https://www.mathworks.com). In order to facilitate the installation step, GEView contains the following R packages: sva (Leek et al., 2012), R.matlab (Bengtsson, n.d.) and their subsidiaries.
Operating Systems: The compiled version can be launched on either Windows or Linux. Using the MATLAB source code, that can be downloaded from GEView website (http://www.weizmann.ac.il/complex/compphys/software/geview/), GEView can run on any operating system with MATLAB installed, including Mac OS.
Restrictions to use by non-academics: None
How to Cite GEView: The present paper should be cited if GEView is used in the preparation of a manuscript

Declarations

Ethics approval and consent to participate

As was stated before (see (Teuffel et al., 2004), (Bornhauser et al., 2007) for the "Zurich" dataset and (Ross et al., 2003) for the "StJude" dataset), patients who were enrolled in these studies or their guardians had given informed consent in accordance with the Declaration of Helsinki. Approval was obtained from each institutions' review board.

Consent for Publication

Not applicable

Availability of Data and Materials

The two datasets that were analyzed during the current study are available through download-links that appear in the following references: St Jude Children's Research Hospital, Memphis, TN (Ross et al., 2003) for the "StJude" dataset, and Departments of Oncology University Children's Hospital Zurich, Switzerland (Teuffel et al., 2004) for the "Zurich" dataset.

Libi Hertzberg, M.D./Ph.D. is the Deputy head, Psychiatric Emergency Department, Shalvata mental health center, Israel, a researcher at the Tel Aviv University, Israel and a scientific consultant at the Weizmann Institute of Science, Israel. She combines psychiatric clinical work with scientific research; her research interests include biological psychiatry and computational biology. Her background in mathematics and computer science allows her to study the biology of psychiatric disorders using computational methods. The resulting research is published in scientific Journals and presented in conferences.

Assif Yitzhaky has been working as a software engineer for 15 years at the Weizmann Institute of Science. He has been involved in various projects and publications in the fields of bioinformatics and computational biology. Programming languages: Matlab, R, C.

Metsada Pasmanik-Chor, Ph.D. is the head of Bioinformatics unit at Tel-Aviv University for the last 15 years. Her activities include Bioinformatics consultation in various topics related to genomics and high-throughput analysis. Consultation is provided to researchers in the University, hospitals and Bio-Tech companies. The resulting research is published in scientific Journals: https://en-lifesci.tau.ac.il/bioinformatics-unit/publications and presented in conferences. In addition to research, she conducts and teaches workshops in various Bioinformatics fields.