BMC Bioinformatics

Software

Open Access

HDBStat!: A platform-independent software suite for statistical analysis of high dimensional biology data

Prinal Trivedi[1], Jode W Edwards[1,3], Jelai Wang[1], Gary L Gadbury[1,2], Vinodh Srinivasasainagendra[1], Stanislav O Zakharkin[1], Kyoungmi Kim[1], Tapan Mehta[1], Jacob PL Brand[1,4], Amit Patki[1], Grier P Page[1] and David B Allison*[1]

Address: [1]Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294, USA, [2]Department of Mathematics and Statistics, University of Missouri-Rolla, Rolla, MO 65409, USA, [3]Department of Agronomy, Iowa State University, Ames, IA 50011, USA and [4]Pennington Biomedical Research Center, 6400 Perkins Rd., Baton Rouge, LA 70808, USA

Email: Prinal Trivedi - PatelHP@uab.edu; Jode W Edwards - Jode@iastate.edu; Jelai Wang - JelaiW@uab.edu; Gary L Gadbury - GadburyG@umr.edu; Vinodh Srinivasasainagendra - Vinodh@uab.edu; Stanislav O Zakharkin - Stas@uab.edu; Kyoungmi Kim - Kyoungmi@uab.edu; Tapan Mehta - Tapan@uab.edu; Jacob PL Brand - BrandJP@pbrc.edu; Amit Patki - APatki@uab.edu; Grier P Page - GPage@uab.edu; David B Allison* - DAllison@uab.edu

* Corresponding author

Published: 06 April 2005

BMC Bioinformatics 2005, **6**:86 doi:10.1186/1471-2105-6-86

Received: 29 November 2004
Accepted: 06 April 2005

This article is available from: http://www.biomedcentral.com/1471-2105/6/86

Abstract

Background: Many efforts in microarray data analysis are focused on providing tools and methods for the qualitative analysis of microarray data. HDBStat! (High-Dimensional Biology-Statistics) is a software package designed for analysis of high dimensional biology data such as microarray data. It was initially developed for the analysis of microarray gene expression data, but it can also be used for some applications in proteomics and other aspects of genomics. HDBStat! provides statisticians and biologists a flexible and easy-to-use interface to analyze complex microarray data using a variety of methods for data preprocessing, quality control analysis and hypothesis testing.

Results: Results generated from data preprocessing methods, quality control analysis and hypothesis testing methods are output in the form of Excel CSV tables, graphs and an Html report summarizing data analysis.

Conclusion: HDBStat! is a platform-independent software that is freely available to academic institutions and non-profit organizations. It can be downloaded from our website http://www.soph.uab.edu/ssg_content.asp?id=1164.

Background

One of the most critical tasks in the field of biology is identifying how and which genes interact with each other under different conditions. Until a few years ago, researchers were only able to accomplish this task for a limited number of genes because the traditional methods in molecular biology allowed them to assess only one gene at a time. The advent of microarray technology has provided investigators the opportunity to simultaneously assess the expression levels of thousands of genes. Microarrays also generate a large amount of data in short period of time. Extracting statistically valid and biologically

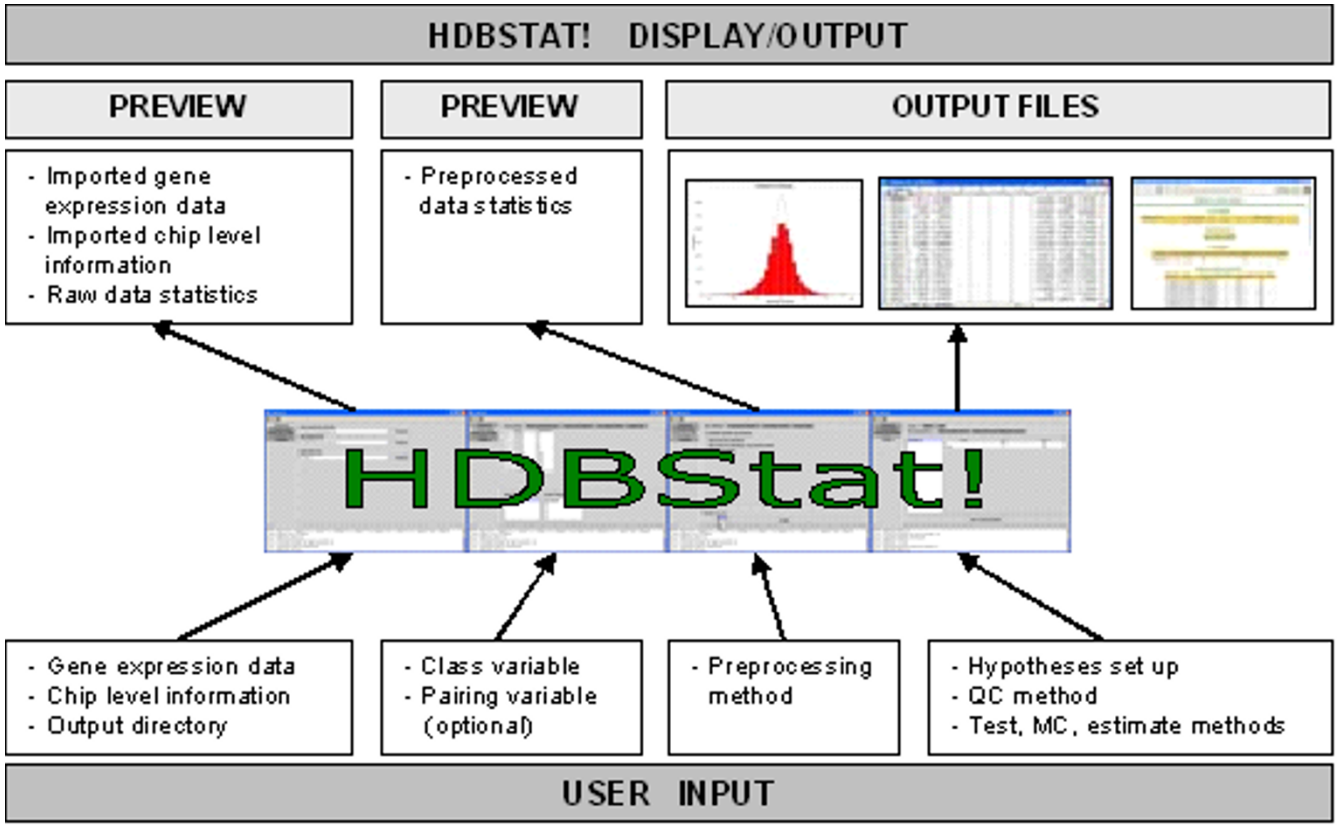


Figure 1
Data analysis in HDBStat! is divided into four steps – data import, data preprocessing, quality control and hypotheses testing. At each step, user input is required and in return, the results are displayed in the interface and/or output to a file.

relevant information from such massive data sets is a major challenge. HDBStat! is a user-friendly and platform-independent software designed for the statistical analysis of microarray data using well-validated methods for quality control of experiments and the identification of differentially expressed genes.

Implementation

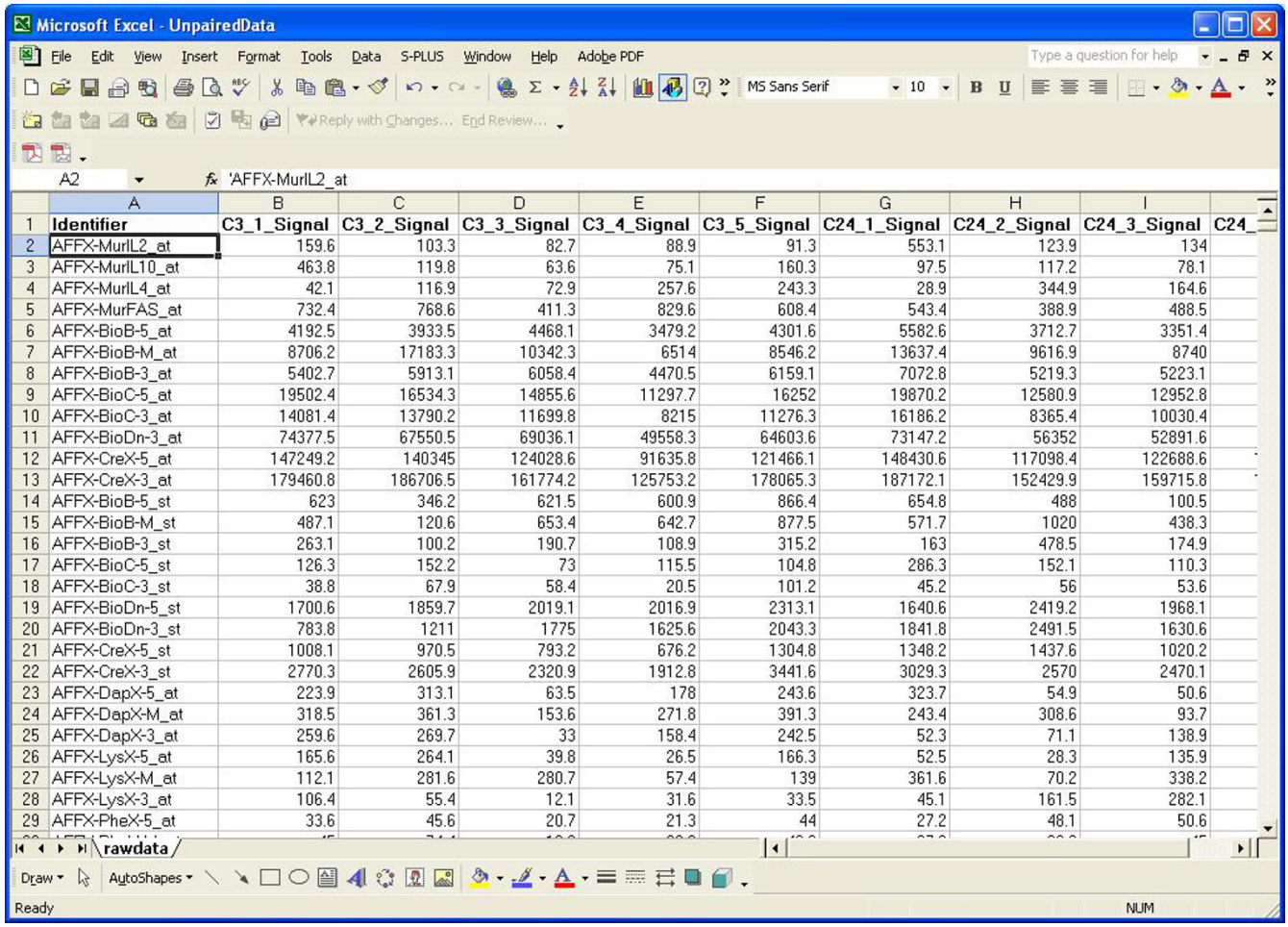
Data analysis in HDBStat! is divided into four steps – data import, data processing, quality control and hypotheses testing (Figure 1).

Data import

Data is imported into HDBStat! using two files, a gene expression data file (Figure 2) and chip level information file (Figure 3), both of which must be Microsoft Excel '97 or more recent format (.xls), or Comma Separated Values (.csv) files [see Additional file 1 and Additional file 2]. The gene expression data file contains the output from the chip image processing software, such as MAS 5.0, Bioconductor, or GenePix. The chip level file contains experimental variables such as treatment, time, experiment, and if appropriate, pairing variables for the chips. Upon import some descriptive statistics are automatically generated about the raw data such as Pearson's correlations between chips, mean, standard deviation, minimum and maximum values of gene expression levels for each chip and displayed in graphical and tabular formats.

Data preprocessing

Optionally, a normalization and/or transformation method(s) can be applied prior to the primary statistical analyses. Normalization is a procedure intended to remove variability among chips that is unrelated to treatment conditions of interest. HDBStat! offers Chip Mean normalization, which divides each observation by the chip mean, and Quantile-Quantile normalization, which ranks each observation on the chip based on expression value and then converts to the value of a deviation that would be expected from the standard normal distribution based on the observation rank. Quantile-quantile normalization results in data from each chip with a mean



	A	B	C	D	E	F	G	H	I
1	Identifier	C3_1_Signal	C3_2_Signal	C3_3_Signal	C3_4_Signal	C3_5_Signal	C24_1_Signal	C24_2_Signal	C24_3_Signal
2	AFFX-MurIL2_at	159.6	103.3	82.7	88.9	91.3	553.1	123.9	134
3	AFFX-MurIL10_at	463.8	119.8	63.6	75.1	160.3	97.5	117.2	78.1
4	AFFX-MurIL4_at	42.1	116.9	72.9	257.6	243.3	28.9	344.9	164.6
5	AFFX-MurFAS_at	732.4	768.6	411.3	829.6	608.4	543.4	388.9	488.5
6	AFFX-BioB-5_at	4192.5	3933.5	4468.1	3479.2	4301.6	5582.6	3712.7	3351.4
7	AFFX-BioB-M_at	8706.2	17183.3	10342.3	6514	8546.2	13637.4	9616.9	8740
8	AFFX-BioB-3_at	5402.7	5913.1	6058.4	4470.5	6159.1	7072.8	5219.3	5223.1
9	AFFX-BioC-5_at	19502.4	16534.3	14855.6	11297.7	16252	19870.2	12580.9	12952.8
10	AFFX-BioC-3_at	14081.4	13790.2	11699.8	8215	11276.3	16186.2	8365.4	10030.4
11	AFFX-BioDn-3_at	74377.5	67550.5	69036.1	49558.3	64603.6	73147.2	56352	52891.6
12	AFFX-CreX-5_at	147249.2	140345	124028.6	91635.8	121466.1	148430.6	117098.4	122688.6
13	AFFX-CreX-3_at	179460.8	186706.5	161774.2	125753.2	178065.3	187172.1	152429.9	159715.8
14	AFFX-BioB-5_st	623	346.2	621.5	600.9	866.4	654.8	488	100.5
15	AFFX-BioB-M_st	487.1	120.6	653.4	642.7	877.5	571.7	1020	438.3
16	AFFX-BioB-3_st	263.1	100.2	190.7	108.9	315.2	163	478.5	174.9
17	AFFX-BioC-5_st	126.3	152.2	73	115.5	104.8	286.3	152.1	110.3
18	AFFX-BioC-3_st	38.8	67.9	58.4	20.5	101.2	45.2	56	53.6
19	AFFX-BioDn-5_st	1700.6	1859.7	2019.1	2016.9	2313.1	1640.6	2419.2	1968.1
20	AFFX-BioDn-3_st	783.8	1211	1775	1625.6	2043.3	1841.8	2491.5	1630.6
21	AFFX-CreX-5_st	1008.1	970.5	793.2	676.2	1304.8	1348.2	1437.6	1020.2
22	AFFX-CreX-3_st	2770.3	2605.9	2320.9	1912.8	3441.6	3029.3	2570	2470.1
23	AFFX-DapX-5_at	223.9	313.1	63.5	178	243.6	323.7	54.9	50.6
24	AFFX-DapX-M_at	318.5	361.3	153.6	271.8	391.3	243.4	308.6	93.7
25	AFFX-DapX-3_at	259.6	269.7	33	158.4	242.5	52.3	71.1	138.9
26	AFFX-LysX-5_at	165.6	264.1	39.8	26.5	166.3	52.5	28.3	135.9
27	AFFX-LysX-M_at	112.1	281.6	280.7	57.4	139	361.6	70.2	338.2
28	AFFX-LysX-3_at	106.4	55.4	12.1	31.6	33.5	45.1	161.5	282.1
29	AFFX-PheX-5_at	33.6	45.6	20.7	21.3	44	27.2	48.1	50.6

Figure 2
Screenshot of gene expression data file in Excel format

of zero and standard deviation of 1.0. Transformation is a process of applying a mathematical function to every observation in a data set in order to better satisfy assumptions of certain statistical models used for analysis. HDBStat! offers three different scales of logarithmic transformation, base-2, base-e, or base-10. Combinations of normalizations and transformations may be selected.

Quality control

HDBStat! provides a unique quality control procedure based upon Deleted Residuals (DR). Deleted residuals have traditionally been used in the statistical analysis of data when the number of observations in a group are small or may be influenced by outliers, as in the case in microarrays. In HDBStat!, the deleted residuals for each gene on each chip is calculated by taking the observed value of a gene on a chip subtracting the mean for the gene across all other chips in that group divided by the standard deviation of the mean for the gene across all the other chips in that group. The Probability Density Function (PDF) for the deleted residuals for a gene will follow a Student's t-distribution with n-2 degrees of freedom where n is the number of chips in the treatment group. If we assume that the genes across a chip are independent identically distributed (IID) the distribution of the deleted residuals should approximate a Student's t-distribution with n-2 degrees of freedom. The difference of the observed data from the expected t-distribution is graphically illustrated (Figure 4) and the significance of the difference is tested using a Kolmogorov-Smirnov test. If a chip is significantly different from the t-distribution it may be an indication that the particular chip is an outlier compared with the other chips in the group. Further, the

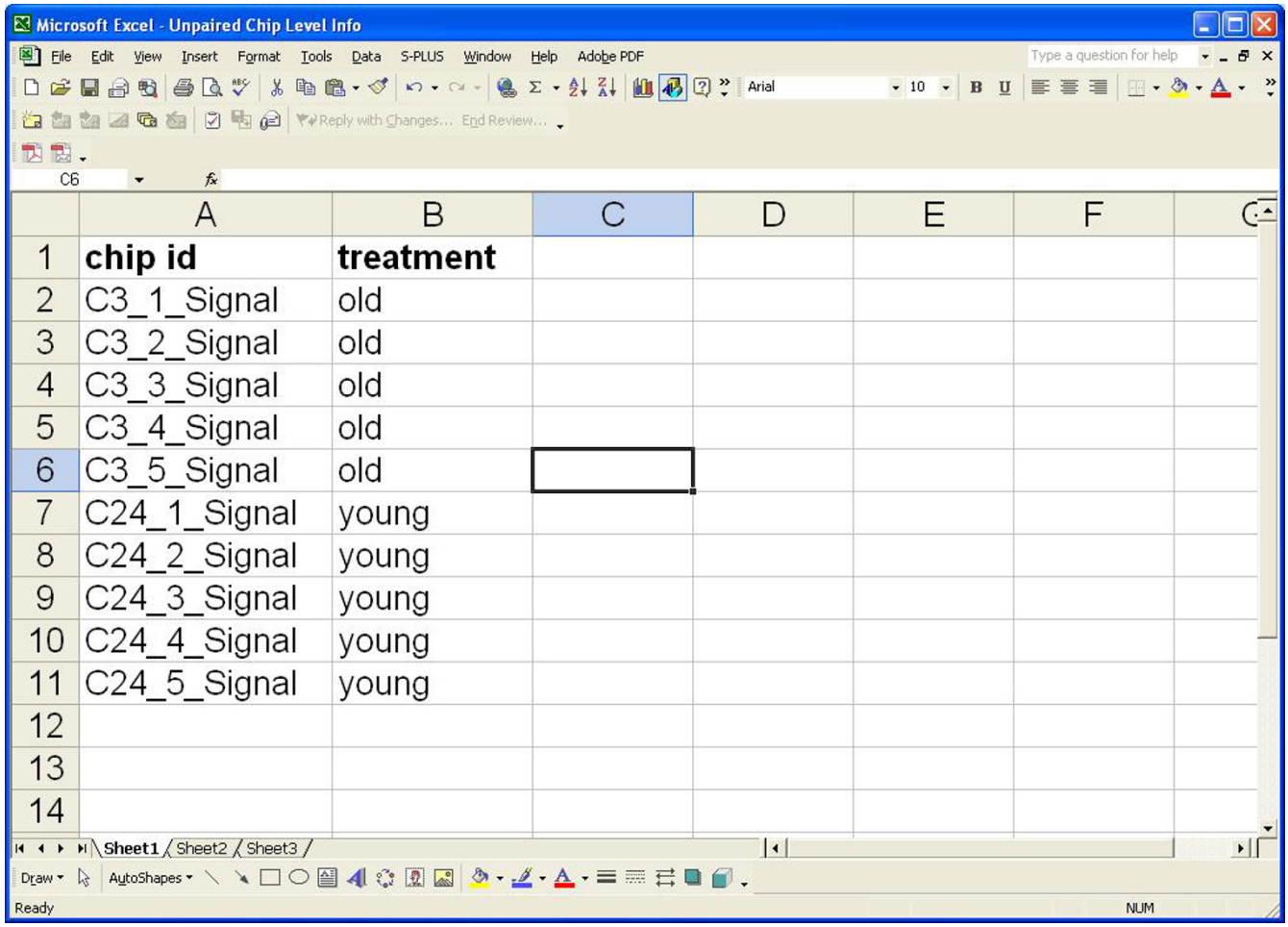


Figure 3
Screenshot of chip level information file in Excel format

user has the opportunity to remove chip(s) from the analysis and re-analyze the data.

Hypotheses testing

Currently, HDBStat! performs a series of pair wise comparison tests. Based on the information provided by user in chip level information file, a combination of all possible hypotheses is displayed in the user interface. User must select at least one hypothesis in order to perform two group comparisons.

HDBStat! includes parametric and non-parametric methods for estimating the significance of changes in gene expression between groups. Student's t-test, for which the user can choose an equal-variance t-test, which uses a pooled variance across treatments, or Welch's t-test, which assumes unequal variances between the two treatment groups [11]. Another method based on Chebyshev's inequality, Chebby Checker is extremely robust against departures from normality and equality of variance between treatment groups, but it also has very low power [2]. The Chebby Checker is useful for identifying genes that are almost certainly differentially expressed without considering any statistical assumptions. In addition a bootstrap resampling method [6,8] is implemented. One can either conduct an exact bootstrap (all possible permutations) or a random (used specified number of permutations) bootstrap. The bootstrap procedures implement both pivots and smoothes in order to calculate the significance more accurately. As exact bootstrap is more accurate than random bootstrap, it is preferred for computationally feasible cases, but once the n per groups exceeds 6 it is difficult to implement.

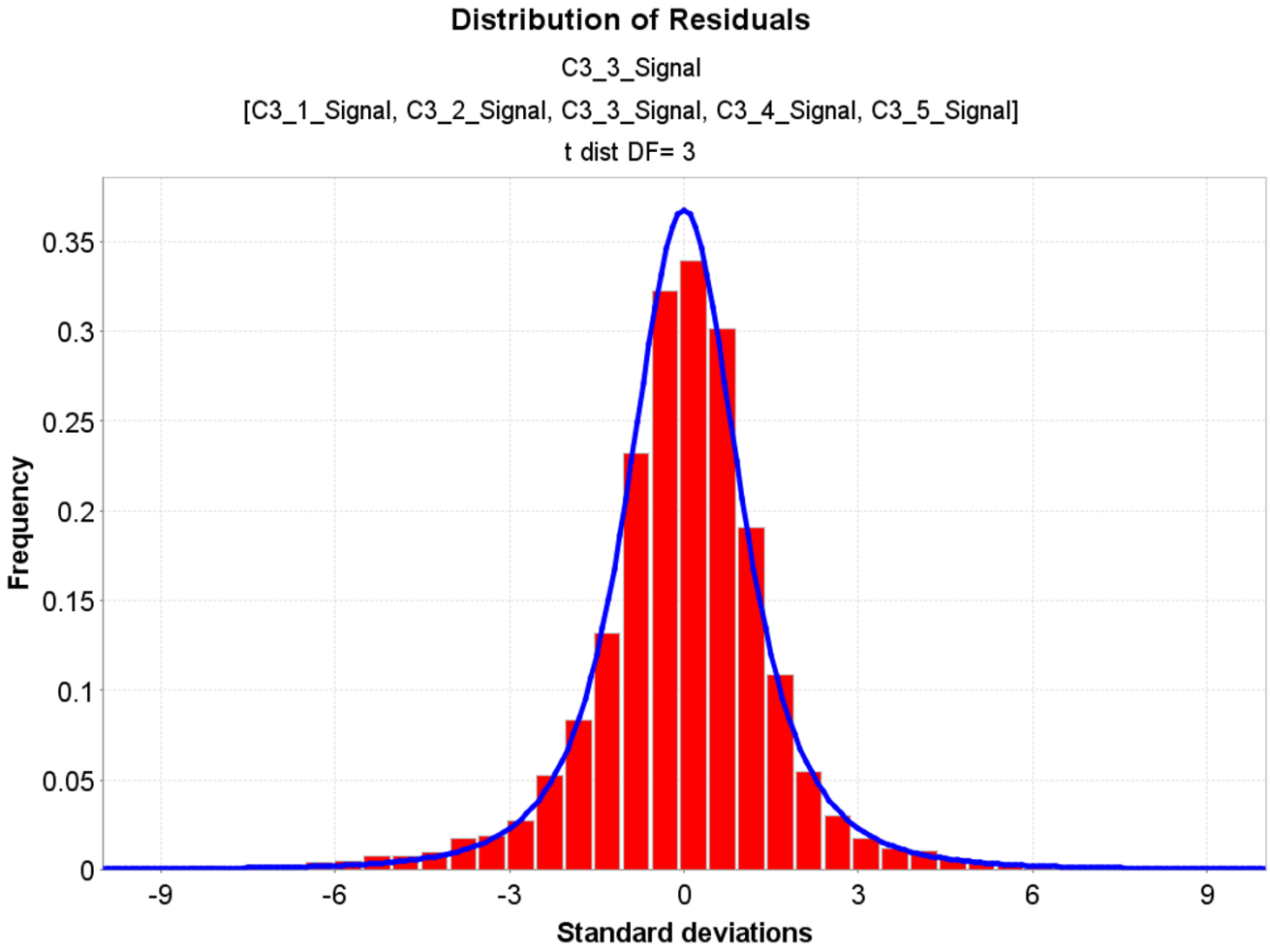


Figure 4
Deleted residuals graph

Because of the large number of simultaneous hypotheses tested in high dimensional biology experiments, an adjustment for multiple testing is appropriate in order to avoid falsely calling too many genes significant. In HDBStat! several multiplicity control methods are available to be applied to any hypothesis testing method. The available multiplicity control adjustments are Bonferroni [4], Sidak [10], two False Discovery Rate (FDR) estimation methods [3,5], and a method based on a mixture modeling of observed p-values [1], referred to as the "Mix-o-matic" (Figure 5) in the HDBStat! software. The Bonferroni and Sidak methods provide experiment-wise (or Family-wise) type I control. The FDR methods are designed to control the proportion of false positives among all genes declared differentially expressed. The mixture modeling method allows for the Bayesian estimation of the probability that each gene is a false positive or negative and this approach is also conveniently for projecting power estimates for future studies [9].

For the planning of future studies HDBStat! implements the method of Gadbury et al to extrapolate power from pilot data [9]. HDBStat! allows for the calculation of the expected discovery rate (EDR), posterior true positive, and posterior true negative rates for large and smaller samples sizes than were entered as pilot data. (Figure 6)

If an investigator is interested in empirically comparing the size of the observed differences in gene expression, an Empirical Bayes method is provided to provide shrinkage estimators of the true differences in gene expression [7]. In addition group means and fold changes in expression are

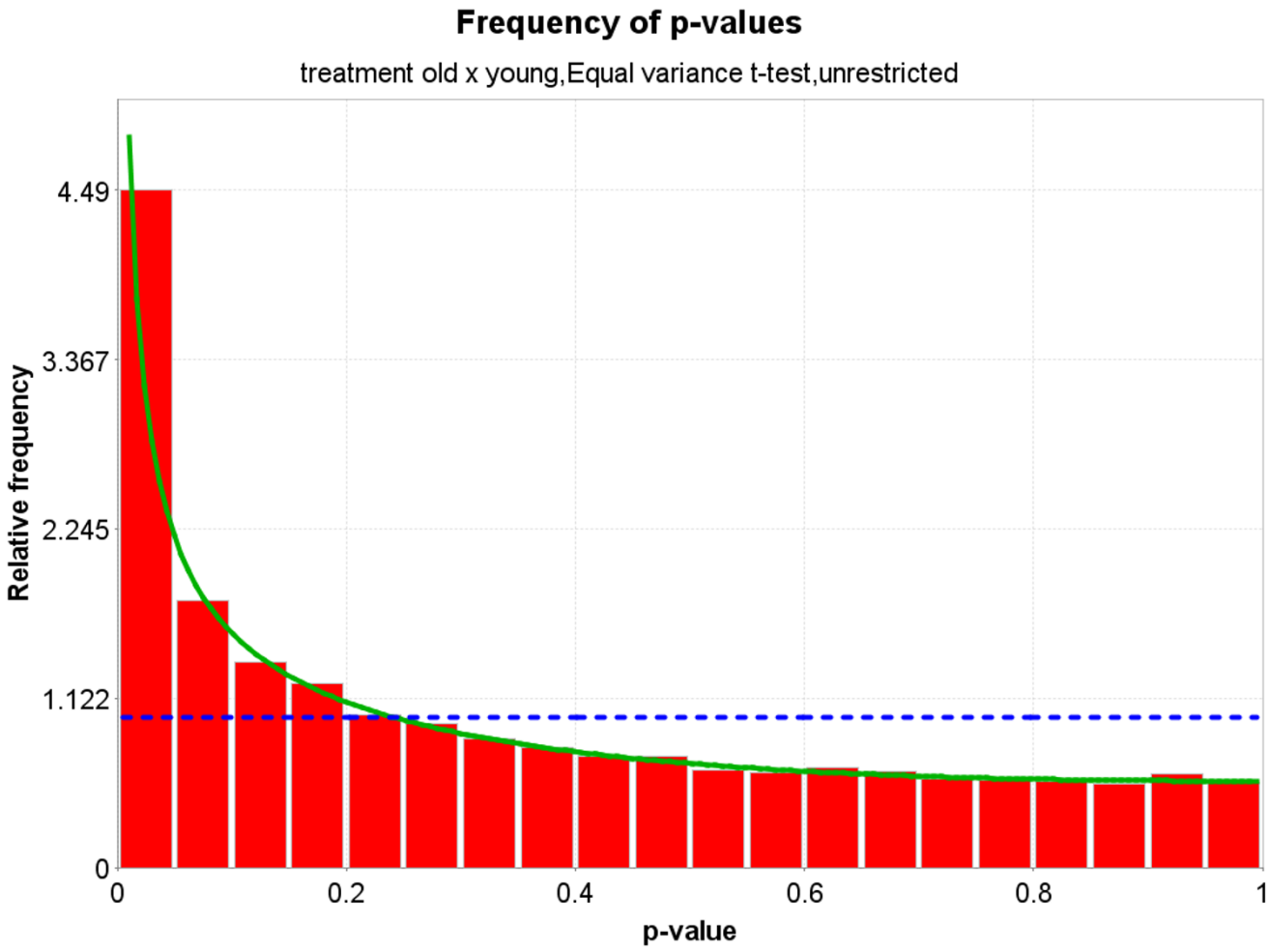


Figure 5
Mix-o-matic graph

calculated and output to results directory specified by the user.

Programming details

HDBStat! is implemented using the Java programming language using various licensed and open source libraries such as Visual Numerics JMSL, Jakarta POI, Velocity, and JFreeChart. Extensive software testing is performed using JUnit library.

Results

At the completion of calculating the deleted residuals and analyzing a hypothesis, results are output to a date/time-stamped directory into the user specified directory. Chip level statistics, preprocessed data, deleted residuals, standard outliers, various pair wise comparison tests (Table 1), mix-o-matic and power analysis results are output in the form of Excel CSV files. Graphs generated from chip level statistics, deleted residuals, mix-o-matic and power analysis results are output in .png format image files. HDBStat! also generates a HTML file that provides a summary of the analysis including the hypotheses tested, chips in each group. This mechanism of outputting results provides the user an opportunity to view quality control results and modify hypotheses, preprocessing methods, and/or chip selections before proceeding to the next step.

Discussion

The goal of HDBStat! is to help researchers analyze microarray data to extract valid inferences, estimates and interpretations via a flexible and user-friendly graphical interface. It allows the user to skip preprocessing and

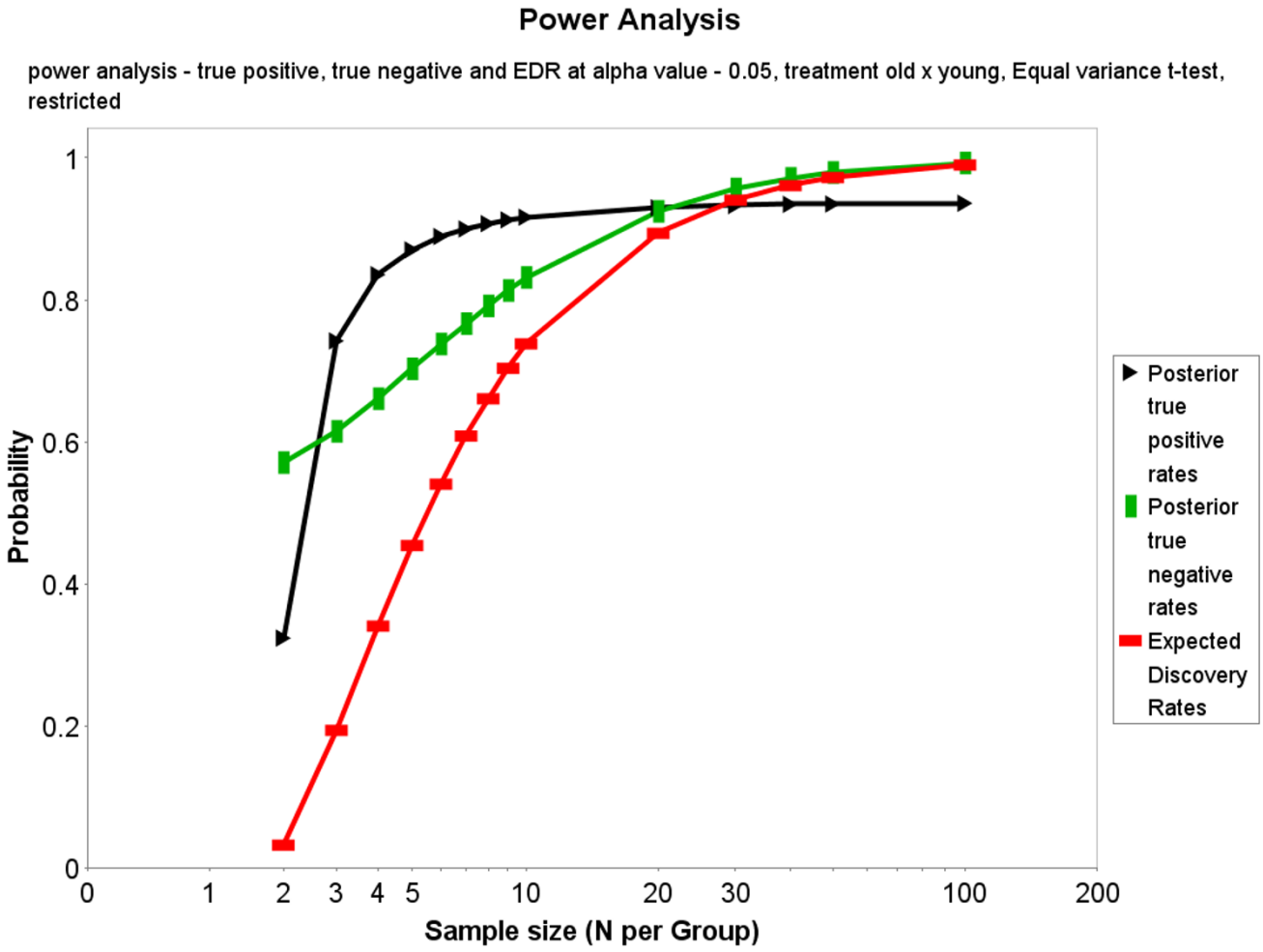


Figure 6
Power analysis graph

quality control methods by simply not selecting those methods. After previewing the preliminary results of raw data, preprocessed data or deleted residuals, user has flexibility to drop a chip by simply un-checking checkbox in the user interface. This feature allows the user to design any number of possible comparisons while the analysis is in progress.

To assist novice users with using HDBStat!, video clips demonstrating how to analyze paired and unpaired data, examples of how to set up input files for paired and unpaired data analyses, screen shots, and FAQ are available on our website. A detailed description of methods as well as additional explanations of the output files in this software is also available in a PDF format on our website.

Additional statistical methods and features are added on an ongoing basis. Support for data import from a text file and results output to a text file will be available for large data sets. In the current version, only single channel or common reference design microarray data can be analyzed using two group comparisons. In the near future, we will add the capability to analyze two channel data and support for ANOVA, and GLM.

There are many software programs available to analyze microarray data, each offering various features and functions. In Table 2, we have compared the features and functions of HDBStat! to SAM, BRB Array Tools and TM4.

Table 1: Hypothesis testing results

Probe	Mean (group1, Raw data)	Mean (group2, Raw data)	Fold change	t (Equal variance t-test)	p-value (Equal variance t-test)	PTP, unrestricted (Equal variance t-test)
AFFX-MurIL2_at	105.16	200.34	1.905097	-0.81193	0.440318	0.226179
AFFX-MurIL10_at	176.52	253.58	1.436551	-0.49777	0.632041	0.194216
AFFX-MurIL4_at	146.56	197.86	1.350027	-0.05477	0.957668	0.147303
AFFX-MurFAS_at	670.06	487.42	0.727427	2.264698	0.05333	0.32061
AFFX-BioB-5_at	4074.98	5165.36	1.267579	-1.10225	0.302405	0.251946
AFFX-BioB-M_at	10258.4	12376.04	1.20643	-0.85246	0.418745	0.230018
AFFX-BioB-3_at	5600.76	6784.9	1.211425	-0.82656	0.432446	0.227573
AFFX-BioC-5_at	15688.4	17406.88	1.109538	-0.23558	0.819676	0.16602
AFFX-BioC-3_at	11812.54	13288.74	1.124969	-0.2571	0.803593	0.168321
AFFX-BioDn-3_at	65025.2	68451.72	1.052695	0.467328	0.65273	0.190965
AFFX-CreX-5_at	124944.9	152682.8	1.222001	-0.39488	0.70325	0.183174
AFFX-CreX-3_at	166352	192790.6	1.158932	-0.26323	0.799024	0.168979
AFFX-BioB-5_st	611.6	677.8	1.108241	0.280987	0.785851	0.170886
AFFX-BioB-M_st	556.26	779.84	1.401934	-1.1822	0.27107	0.25832
AFFX-BioB-3_st	195.62	359.32	1.836827	-1.51174	0.169049	0.281499
AFFX-BioC-5_st	114.36	183.8	1.607205	-2.17372	0.061461	0.316596
AFFX-BioC-3_st	57.36	94.86	1.653766	-0.62822	0.547374	0.2079
AFFX-BioDn-5_st	1981.88	2327.5	1.17439	-0.99146	0.350498	0.242595
AFFX-BioDn-3_st	1487.74	2265.06	1.522484	-2.01951	0.078118	0.309434
AFFX-CreX-5_st	950.56	1221.04	1.284548	-1.42444	0.192137	0.275814
AFFX-CreX-3_st	2610.3	3429.5	1.313834	-1.21894	0.257585	0.261147
AFFX-DapX-5_at	204.42	267.12	1.306721	0.187522	0.85592	0.16092
AFFX-DapX-M_at	299.3	347.62	1.161443	0.119799	0.907597	0.153866
AFFX-DapX-3_at	192.64	95.82	0.497404	1.5611	0.157123	0.284581
AFFX-LysX-5_at	132.46	115.9	0.874981	0.445594	0.667702	0.188634
AFFX-LysX-M_at	174.16	204.14	1.172141	-0.09641	0.92557	0.151477
AFFX-LysX-3_at	47.8	167.94	3.513389	-1.2216	0.256632	0.261348
AFFX-PheX-5_at	33.04	55.5	1.679782	-2.04285	0.07534	0.310549

Table 1: Hypothesis testing results *(Continued)*

AFFX-PheX-M_at	40.64	29.66	0.729823	1.223746	0.255865	0.261511
AFFX-PheX-3_at	497.66	217.12	0.436282	2.826877	0.022257	0.342932
AFFX-ThrX-5_at	92.06	184.74	2.006735	-0.26462	0.797994	0.169127
AFFX-ThrX-M_at	225.32	272.46	1.209214	-0.33299	0.747704	0.17649
AFFX-ThrX-3_at	114.54	95.08	0.830103	-0.17193	0.86776	0.15928
AFFX-TrpnX-5_at	73	96.68	1.324384	-0.82529	0.433126	0.227452
AFFX-TrpnX-M_at	30.12	37.74	1.252988	-0.31889	0.757976	0.174969

Table 2: Comparison of HDBStat! with other software packages. All these software packages are still in active development and new functions will undoubtedly be added over time.

	HDBStat!	SAM	BRB-Array Tools 3.2.2	TM4
Two color data handling	Common reference and balanced block designs	No	Yes	Yes
Database	No	No	No	Yes
Ratio Statistics	Yes	No	Yes	Yes
Normalization	Yes	Yes	Yes	Yes
Max number of arrays	No limit	255	249	No limit
Discriminate Analysis	No	No	Yes	Yes
ANOVA	Yes	No	Yes	Yes
Bootstrapping	Yes	Yes	Yes	Yes
Non-normal and heteroskedastic data handling	Yes	Yes	Via normalization	Via normalization
Non-parametric statistics	Yes	No	Yes	No
Cluster analysis	No	No	Yes	Yes
FDR (number)	8	1	2	1
Family Wise Error rate corrections	2	0	1	1
Quality Control	Yes	No	No	Yes

Table 2: Comparison of HDBStat! with other software packages. All these software packages are still in active development and new functions will undoubtedly be added over time. *(Continued)*

Power Analysis	Yes	No	No	No
Automatic report generation	Yes	No	Yes	No
Gene Class testing	No	No	Yes	No
Automatic Annotation	No	Link out	Yes	No
Platform	Single program implemented in Java & available via Java Web Start technology	Microsoft Excel Add-in	Microsoft Excel Add-in	4 separate programs, 3 of which implemented in java & 1 in C++

Availability and requirements

System requirements for an end-user are the Java Runtime Environment (JRE 1.4.2 or higher), at least 256 MB RAM and 25 MB hard disk space. Using Java Web Start technology, HDBStat! can be easily downloaded from our website at http://www.soph.uab.edu/ssg_content.asp?id=1164.

Authors' contributions

JWE, JPLB, KK, GPP wrote the implementation specification documents for various methods. GLG and DBA developed mix-o-matic method, developed prototype implementation in S-Plus and wrote the implementation specification document. GPP and JWE developed the deleted residuals approach. DBA and JWE developed prototype implementation of Empirical Bayes Estimates. JWE and PT developed a prototype implementation of statistical methods in SAS. JW, PT, VS and TM implemented and tested the java code. JWE, PT, JW, SOZ, GPP, VS, TM and AP tested the software. GPP and JWE directed the content of HDBStat! All authors read and approved the manuscript.

Acknowledgements

This work is supported by grants from the UAB HSF GEF, NSF 0217651 and NIH U54CA100949.

References

1. Allison DB, Gadbury GL, Moonseong H, Fernandez JR, Lee C, Prolla TA, Weindruch R: **A mixture model approach for the analysis of microarray gene expression data.** *Comp Statist & Data Anal* 2002, **39(1):**1-20.
2. Beasley TM, Page GP, Brand JPL, Gadbury GL, Mountz JD, Allison DB: **Chebyshev's inequality for non-parametric testing with small *N* and a in microarray research.** *J R Statist Soc C* 2004, **53:**95-108.
3. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Statist Soc B* 1995, **57:**289-300.
4. Bland JM, Altman DG: **Multiple significance tests: the Bonferroni method.** *BMJ* 1995, **310(6973):**170.
5. Benjamini Y, Yekutieli D: **The control of the false discovery rate in multiple testing under dependency.** *Ann Statist* 2001, **29(4):**1165-1188.
6. Davison AC, Hinkley DV: **Bootstrap methods and their application.** Cambridge University Press, United Kingdom; 1997.
7. Edwards JW, Page GP, Gadbury G, Heo M, Kayo T, Weindruch R, Allison DB: **Empirical Bayes estimation of gene-specific effects in micro-array research.** *Funct Integr Genomics* 2005, **5(1):**32-9.
8. Effron B, Tibshirani RJ: **An Introduction to the Bootstrap.** Chapmann and Hall New York; 1993.
9. Gadbury GL, Page GP, Edwards JW, Kayo T, Prolla TA, Weindruch R, Permana PA, Mountz J, Allison DB: **Power and Sample Size Estimation in High Dimensional Biology.** *Stat Meth Med Res* 2004, **13:**325-338.
10. Sidak Z: **Rectangular confidence regions for the means of the multivariate normal distributions.** *J Am Stat Assoc* 1967, **62:**626-633.
11. Welch BL: **The significance of the difference between two means when the population variances are unequal.** *Biometrika* 1938, **29:**350-362.