Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Two-group comparison of gene signatures: failure of conventional statistical methods and validation of a novel algorithm

Haseeb Ahmad Khan
King Saud University, Riyadh, Saudi Arabia

Method Article

Keywords: gene signatures, two-group comparison, nonparametric tests, microarray data, software, Excel add-in

Posted Date: May 22nd, 2009

DOI: https://doi.org/10.1038/nprot.2009.106

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Introduction

The gene expression profiling could aid the physicians to better understand the cellular morphology, resistance to chemotherapy and overall clinical outcome of disease \[1,2]. Such individualized treatment may significantly increase survival due to the optimization of treatment procedure according to clinical pathogenesis. Ein-Dor et al \[3] have pointed out that the gene sorted for the same clinical types of patients but different groups differed widely and possessed only few genes in common. An explanation to this lack of overlap between predictive signatures from different studies with the same goal may be due to the presence of more predictive genes than required to design an accurate predictor \[4]. However, the microarray technique itself has been shown to be highly reproducible within and across two high volume laboratories \[5]. Numerous statistical procedures including t-test \[6,7], analysis of variance \[8], Pearson correlation \[9], Wilcoxon signed-rank test \[10,11] and Mann Whitney U test \[12,13] have been used for comparison of microarray data. However, the validity of various conventional statistical methods for two-group comparison of gene signatures was never evaluated using carefully selected data sets. A novel algorithm with software support is presented herein for more realistic and comprehensive interpretation of gene signatures. **Computational method and theory of CalcHEPI** The formula used for computation of HEPI score is HEPI=Σ \[\(Ni\(0→t) Sj\(0→1)/Nt]x100. Where Ni is the number of genes with Score Sj. The subscript ‘i’ may vary between 0 and total number of genes in the signature and ‘j’ may vary between 0 \(minimum score) and 1 \(maximum score). Nt is the total number of genes in the signature. First, all the ratios of expression data are categorized according to a logical scale to get the respective Ni and Sj values. The percent contributions of each set of genes \(genes with same expression score) are computed and then summed up to get HEPI score. The fold-change strategy used in HEPI scores is robust, accurate and reproducible. Although the concept of fold-change has been described in microarray experiments it has never been utilized for collective interpretation of gene signatures. Technically, the ratio of the color intensity of each spot \(probe) measures the relative expression of the corresponding gene under two different experimental conditions. In general, a gene is said to be differentially expressed if the ratio in absolute value of the expression levels between the control and treated group exceeds certain thresholds. The most acceptable expression ratios for up- and down-regulated genes have been suggested as >1.5 and <0.5 respectively \[11,17,18]. While adopting the same cut-off values, additional sub-grading has been proposed in this protocol. HEPI scores are simple to interpret, easy to compare and prominent for visual cross checking. **Software design** CalcHEPI software has been developed in Microsoft Excel platform due to Excel’s flexibility, universal availability, and macro-based automation. Moreover, the spreadsheet layout of Excel is perfectly suitable for storing and analyzing microarray data as well as developing microarray analysis software. The data selection is controlled by input box to allow the users to select the paired expression values from any place of the worksheet \(Fig. 1). The software then utilizes Excel’s worksheet formula function together with a macro subroutine to compute HEPI scores \(Fig. 2). The percent contribution of norm-regulated \(green), down-regulated \(blue) and up-regulated \(red) genes is also shown as a color-coded bar. The output of the

software provides a comprehensive understanding of the results in terms of both qualitative \(up- or down-regulation) and quantitative \(gradation in fold-change) analysis of gene signature with the quick review indicator bar. The clarity and integrity of report format are quite helpful for any cross evaluation. HEPI scores are valid for any size of array signature as they are calculated according to percent \(and not number) of differentially expressed genes on a 10 point scale \(5 for up regulation and 5 for down regulation). **Software validation** The functional accuracy and reliability of software have been validated using the simulated and real gene signatures data for two-group comparisons. Six pairs of expression data were specifically designed to represent various degrees of similarity/differences \(details not shown). Among them, the two groups in pair 4 are not significantly different whereas the groups in pair 6 possess maximum difference. All these 6 pairs were subjected to nonparametric comparisons with Mann-Whitney U test, Kolmogorov-Smirnov test, Kruskal-Wallis test, Wilcoxon signed-rank test, Sign test, Friedman test and Kendall W test using SPSS \(Version 10). The real expression data of published signatures including ovarian carcinoma \[14], ulcerative colitis \[15], leukemia \[16] and adenocarcinoma \[6] were also analyzed by the above nonparametric tests as well as CalcHEPI. The characteristics of these real signatures have been summarized in our earlier report \[10].

Equipment

A personal computer with Microsoft Excel program.

Procedure

Installation of CalcHEPI Add-in 1. Open the Excel program and insert the program CD in drive \(if the CalcHEPI software is in a CD). NOTE: Alternatively the software file \(115 KB) can be obtained as an e-mail attachment from the author. 2. In the ‘Tool’ menu of Excel workbook, click on ‘Add-Ins’ and then click on ‘Browse’. Locate your appropriate drive and double click on ‘HEPI’. 3. The message “Copy HEPI to …” will appear, click ‘Yes’; the appearance of ‘HEPI’ on the menu bar indicates the proper installation of the Add-in. 4. Remove the software CD \(if applicable). 5. For un-installation, follow all the above steps except selecting ‘HEPI Uninstall’ instead of ‘HEPI’; the removal of ‘HEPI’ from the menu bar will indicate the proper un-installation of the Add-in. **Running the program** 1. For computing HEPI score, enter the gene expression data of different groups in separate columns. IMPORTANT: Prior to activating the ‘Input’ window, ensure that the data to be compared reside in adjacent columns. CAUTION: Since most of the gene signatures are composed of tens to a few hundred specific genes the upper limit of the software has been fixed at 1000 genes to prevent unnecessary system-busy situation due to mistake in data selection. The software will alarm the user if more than 1000 rows are selected in the worksheet. 2. Once the data entry has been completed, click the ‘HEPI’ button in the menu bar to popup the ‘Input’ window \(Fig. 1). 3. Select the range of data \(numbers only) without including the header row \(if any) as shown in Fig. 1. 4. Now clicking the ‘OK’ button executes the software and a comprehensive report is displayed \(Fig. 2).

Timing

Once the installation of Add-in has been done, the computation of HEPI score is performed instantly after selection of the desired gene expression data.

Critical Steps

1. Please ensure that the data to be compared reside in adjacent columns. 2. The group against which the other group is to be compared should be on the left column. 3. The data selection for analysis should only be from TWO ADJACENT columns. 4. Only numerical data should be selected; any header row with text must not be selected as shown in Fig. 1. 5. The software accepts only up to 1000 rows data \ (exceeding this limit will be indicated by a message). However, the users who want to use the software for more number of genes can contact the author to get a modified version with more capacity.

Troubleshooting

Installation of Add-in for Excel 2007 The instructions for Add-in installation given above are valid for Excel 2003. For installing the Add-in in Excel 2007, follow the following steps: 1. Click on the Office Button, the big round button in the top left of the Excel window. 2. Click the Excel Options button at the bottom of this menu to open the dialog box. 3. Click the Add-Ins item in the list along the left edge of the dialog to see the Add-Ins panel. Make sure the Manage dropdown at the bottom shows Excel Add-Ins 4. Press the Go button to show the Add-Ins dialog. 5. Click Browse and use the Browse dialog to locate the add-in file. **Troubleshooting** So far, I have not faced any problem with the installation or running this Add-in. However, the users are requested to contact the author in case they encounter any problem associated with this software.

Anticipated Results

The anticipated results format is shown in Fig. 2. For users' information, the results of software validation using simulated and real signatures clearly demonstrated the incompatibility of conventional statistical methods for comparing gene expression data. Paradoxical outcomes were observed while comparing 6 simulated gene signatures using 7 nonparametric tests \(Table 1). Five tests including Mann-Whitney U test, Kruskal-Wallis test, Sign test, Friedman test and Kendall W test resulted same but logically unrealistic P values for all these signature pairs. Surprisingly, these tests showed P=1 for a gene signature with maximum difference \(Pair 6, HEPI = 100) and P = 0.001 for a signature with a slight difference \(Pair 5, HEPI = 4) \(Table 1). The remaining two tests including Kolmogorov-Smirnov test and Wilcoxon signed-rank test also failed to effectively handle these comparisons. For instance, Kolmogorov-Smirnov test resulted P = 0.001 both for similar groups \(Pair 4, HEPI = 0) as well as the groups with maximum difference \(Pair 6, HEPI = 100). Wilcoxon signed-rank test showed ambiguous results for Pairs 3 and 5 \ (Table 1). Statistical inconsistency also prevailed while comparing real signatures \(Table 1) affixing a question mark on the reliability of nonparametric tests for two-group comparison of gene signatures. Thus the conventional statistical methods may not be able to handle the peculiar microarray expression data, particularly for two-group comparison of gene signatures. More accurate and unified statistical

methods and/or coding systems are therefore needed to ensure routine and uniform clinical application of gene signatures. CalcHEPI is one such effort that may serve as a convenient and robust tool for two-group comparison of gene signatures.

References

1. Buckhaults P. Gene expression determinants of clinical outcome. Curr Opin Oncol 2006; 18: 57-61. 2. Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R, Cottrill H, et al. Genomic signatures to guide the use of chemotherapeutics. Nat Med 2006; 12: 1294-300. 3. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci USA 2006; 103: 5923-8. 4. Roepman P, Kemmeren P, Wessels LF, Slootweg PJ, Holstege FC. Multiple robust signatures for detecting lymph node metastasis in head and neck cancer. Cancer Res 2006; 66: 2361-6. 5. Anderson K, Hess KR, Kapoor M, et al. Reproducibility of gene expression signature-based predictions in replicate experiments. Clin Cancer Res 2006; 12: 1721-7. 6. Notterman DA, Alon U, Sierk AJ, Levine AJ. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. Cancer Res 2001; 61: 3124-30. 7. Tanaka TS, Jaradat SA, Lim MK, Kargul GJ, Wang X, et al. Genome-wide expression profiling of mid-gestation placenta and embryo using a 15000 mouse developmental Cdna microarray. Proc Natl Acad Sci USA 2000; 97: 9127-32. 8. Bushel PR, Hamadeh HK, Bennett L, Green J, Ableson A, et al. Computational selection of distinct class- and subclass-specific gene expression signatures. J Biomed Inform 2002; 35: 160-70. 9. Bouras T, Southey MC, Chang AC, Reddel RR, Willhite D, et al. Stanniocalcin 2 is an estrogen-responsive gene coexpressed with the estrogen receptor in human breast cancer. Cancer Res 2002; 62: 1289-95. 10. Khan HA. ArrayVigil: a methodology for statistical comparison of gene signatures using segregated-one-tailed \ (SOT) Wilcoxon's signed-rank test. J Mol Biol 2005; 345: 645-9. 11. Khan HA. ArraySolver: an algorithm for colour-coded graphical display and Wilcoxon signed-rank statistics for comparing microarray gene expression data. Comp Func Genom 2004; 5: 39-47. 12. Kihara C, Tsunoda T, Tanaka T, Yamana H, Furukawa Y, et al. Prediction of sensitivity of esophageal tumors to adjuvant chemotherapy by cDNA microarray analysis of gene-expression profiles. Cancer Res 2001; 61: 6474-9. 13. Rus V, Atamas SP, Shustova V, Luzina IG, Selaru F, et al. Expression of cytokine- and chemokine-related genes in peripheral blood mononuclear cells from lupus patients by cDNA array. Clin Immunol 2002; 102: 283-90. 14. Wang K, Gan L, Jeffery E, Gayle M, Gown AM, et al. Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray. Gene 1999; 229, 101-8. 15. Dooly TP, Curto EV, Reddy SP, Davis RL, Lambert GW, et al. Regulation of gene expression in inflammatory bowel disease and correlation with IBD drugs screening by DNA microarrays. Infalmm Bowel Dis 2004; 10, 1-14. 16. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999; 286: 531-7. 17. Bull JH, Ellison G, Patel A, et al. Identification of potential diagnostic markers of prostate cancer and prostatic intraepithelial neoplasia using cDNA microarray. Br J Cancer 2001; 84: 1512-19. 18. Wang W, Marsh S, Cassidy J, McLeod HL. Pharmacogenomic dissection of resistance to thymidylate synthase inhibitors. Cancer Res 2001; 61: 5505-10.

Acknowledgements

The author is highly thankful to the research groups of Dr. Kai Wang \(Chiroscience R&D, Inc., Bothell, WA, USA); Dr. Thomas P. Dooley \(IntegriDerm Inc., Birmingham Alabama, USA); Dr. Todd R. Golub \(Massachusetts Institute of Technology, Cambridge, MA, USA) and Dr. Daniel A. Notterman \(Princeton University, Princeton, NJ, USA) for using their published data to validate CalcHEPI protocol.

Figures

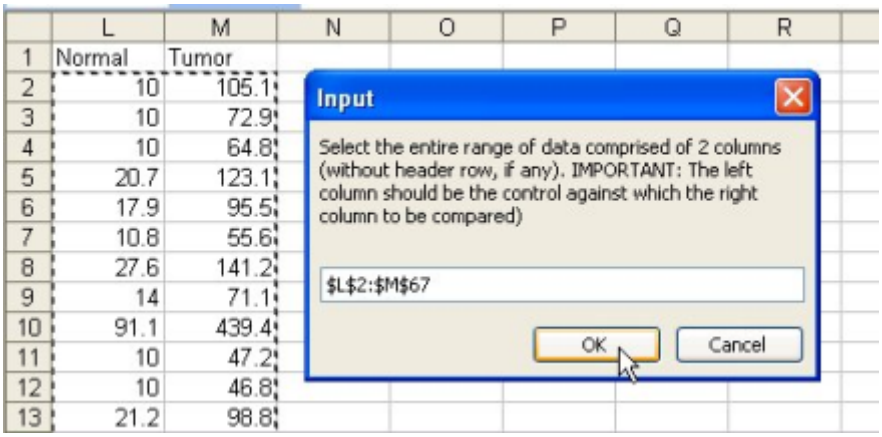

Figure 1

Data input box in the Excel worksheet. Excel worksheet displaying a portion of gene expression data of signature-D as well as the functioning of CalcHEPI software. The figure also shows the selection of paired gene expression values for all 66 genes (Cell L2 to Cell M67). Clicking the ‘OK’ button executes the program and the results are displayed (as shown in Fig. 2).

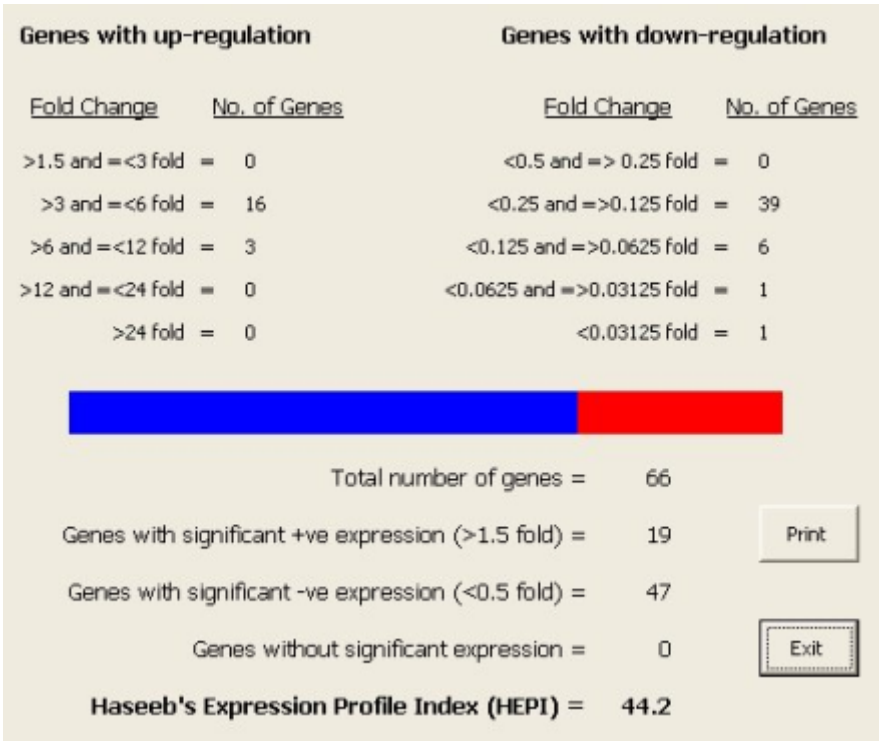

Figure 2

Results window of CalcHEPI software. A representative output of the results for ‘Signature-D’. Color bar represents percent contribution of norm-regulated (green, absent in this case), down-regulated (blue) and up-regulated (red) genes.

(A) Using simulated data

Statistical test	P value (2-tailed) or HEPI score					
	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5	Pair 6
Mann Whitney U test	1	1	1	1	0.001	1
Kolmogorov-Smirnov test	0.001	0.001	1	0.001	0.001	0.001
Kruskal Wallis test	1	1	1	1	0.001	1
Wilcoxon signed rank test	0.007	0.007	1	1	0.001	0.006
Sign test	1	1	1	1	0.001	1
Friedman test	1	1	1	1	0.001	1
Kendall W test	1	1	1	1	0.001	1
HEPI Score	60	45	50	0	4	100

(B) Using real gene signature data

Statistical test	P value (2-tailed) or HEPI score			
	Signature-A Ovarian carcinoma	*Signature-B* Ulcerative colitis	*Signature-C* Leukemia	*Signature-D* Adenocarcinoma
Mann Whitney U test	1.000	0.399	0.229	0.000
Kolmogorov-Smirnov test	0.001	0.010	0.249	0.000
Kruskal Wallis test	1.000	0.399	0.229	0.000
Wilcoxon signed rank test	0.021	0.067	0.182	0.000
Sign test	1.000	1.000	0.885	0.001
Friedman test	1.000	0.835	0.773	0.001
Kendall W test	1.000	0.835	0.773	0.001
HEPI Score	42	36.5	29.1	44.2

Figure 3

Table 1 Software validation for two-group comparisons of (A) simulated gene expression data and (B) real gene signatures using different statistical methods and CalcHEPI. Owing to the peculiarity of expression data, all the conventional statistical tests appear to be invalid for two-group comparisons. A huge disparity in P values can be seen while using different statistical tests. However, HEPI provides a realistic quantitative evaluation of differential gene expression with in-depth information about expression pattern (Fig. 2).