SOFTWARE Open Access

GRACOMICS: software for graphical comparison of multiple results with omics data

Minseok Seo[1†], Joon Yoon[1†] and Taesung Park[1,2*]

Abstract

Background: Analysis of large-scale omics data has become more and more challenging due to high dimensionality. More complex analysis methods and tools are required to handle such data. While many methods already exist, those methods often produce different results. To help users obtain more appropriate results (i.e. candidate genes), we propose a tool, GRACOMICS that compares numerous analysis results visually in a more systematic way; this enables the users to easily interpret the results more comfortably.

Results: GRACOMICS has the ability to visualize multiple analysis results interactively. We developed GRACOMICS to provide instantaneous results (plots and tables), corresponding to user-defined threshold values, since there are yet no other up-to-date omics data visualization tools that provide such features. In our analysis, we successfully employed two types of omics data: transcriptomic data (microarray and RNA-seq data) and genomic data (SNP chip and NGS data).

Conclusions: GRACOMICS is a graphical user interface (GUI)-based program written in Java for cross-platform computing environments, and can be applied to compare analysis results for any type of large-scale omics data. This tool can be useful for biologists to identify genes commonly found by intersected statistical methods, for further experimental validation.

Keywords: GUI, Microarray, NGS, Omics, RNA-seq, SNP, Visualization

Background

Over the last decade, success in microarray data studies has led to an expansion of large-scale omics data analyses and their data types. Vast amounts of data, in various forms, are produced for a common goal: to find genetic variants related to a phenotype of interest (e.g., disease status, etc.). In unison with technological advances, many statistical tools were developed for separate types of omics data analyses. In our study, we will illustrate the application of our tool for different omics data types.

Many microarrays studies aim to detect “gene expression signatures” specific to various human diseases by comparing expression levels between two distinct groups. The main idea is to identify overexpressed and underexpressed genes, as compared to a control group, and label them as deleterious or protective, respectively. The success of this approach in human cancer, and other diseases [1], promoted the development of many statistical methods. However, unifying the analysis results from disjointed methods cannot keep up with the explosive rate of publications concerning the specific phenotype of interest. Thus, annotation and replication studies are required in this current era. Many databases, such as the National Center for Biotechnology Information (NCBI), have been used to infer biological information from omics data and make note of novel findings that were detected as previously reported “markers.”

The popularity of another type of array-based study, focusing on single nucleotide polymorphism (SNP) association studies, has steadily increased. In fact, SNP analysis has been crucial in uncovering the genetic correlations of genomic variants with quantitative traits, complex diseases, and drug responses [2]. One well-known data source, the Wellcome Trust Case Control Consortium (WTCCC) database, which handles 14,000 cases of seven common diseases and 3,000 shared controls, has led to many influential publications. While various analysis methods have been published, and public databases such as dbSNP [3]

* Correspondence: tspark@stats.snu.ac.kr
†Equal contributors
[1]Interdiciplinary of Bioinformatics Program, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, Korea
[2]Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, Korea

© 2015 Seo et al.; licensee BioMed Central. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

and HapMap [4] are available, utilizing them well is another issue.

Following the footsteps of array-based approaches, an era of high-throughput sequencing began, and this technology has been applied to RNA-seq and whole exome and genome sequencing. RNA-seq has properties that are different from microarrays, for example, a high dynamic range and low background expression levels. To address these properties, several statistical methods using Poisson or negative binomial distributions have been proposed [5-7]. In the case of exome and genome sequencing, issues with missing heritability have led researchers to study more than just common variants, and various methods have now been proposed to handle rare variants [8-10].

As for visualization tools, there are only a few programs available for comparison. Multi Experiment Viewer (MeV) [11] is one of the most popular tools included in the TM4 suite, which is used to analyze microarray data. Although it supports several statistical methods of microarray data analysis, MeV provides only multiple outputs in treeview. Similar to MeV, PLINK [12] is a widely used genome association analysis toolset, but does not provide graphical interactive comparison of results.

Here, we focused on exploring the inconsistent results that can be produced from method-specific assumptions and parameters. Taking an extra step to check, understand, and interpret the different results can be challenging for scientists without computational proficiency. We aimed to ease such problems by proposing a visual comparison tool in a user-friendly environment. In addition to its accessibility, GRACOMICS can reflect a change in results according to an immediate alteration of significance levels. Such characteristics are valuable, and likely essential for effective, interactive, and integrative comparison of multiple results. Therefore, the proposed tool, GRACOMICS, provides a novel approach to visually compare several test results through graphical user interface (GUI) components.

In addition to its interactive GUI, our tool provides three distinctive layouts for comparison, including pairwise plots, summary tables, and a "heatmap-like" summary table highlighting pivotal markers, commonly detected by different methods. Two of the modules, the Pairwise Comprehensive Scatter Plots Module (Pair-CSP) and the Pairwise Detailed Scatter Plot Module (Pair-DSP), compare and contrast a pair of methods at the same time, while the third, the Multiple Results Comparison Module (Multi-RC), can handle all the employed methods (more than two) at once. Note that the user can define the top N significant markers (from input files) that will be used in the modules, for more interactive and efficient comparison. Furthermore, simple web-annotation functionality adds to the benefits, in terms of biological interpretation.

Implementation

Microarray dataset and statistical methods

For microarray studies, statistical tests were performed to detect differentially expressed genes (DEGs) between two groups: cases and controls. A pre-processing step is necessary for statistical analysis of the raw expression profiles, including background correction, global or local normalization, log-transformation, etc. Such processing steps may alter the results and should be performed only after fully understanding the platform and target probes of the analysis. We employed a microarray dataset, GSE27567 [13], from the Gene Expression Omnibus (GEO) database, consisting of 45,101 Affymetrix probes from 93 individual mice. To detect the DEGs from the microarray data, we perform two group comparison tests between tumor-bearing mice and non-transgenic controls. We employed statistical tests such as *t*-test, significant analysis of microarray (SAM) [14], permutation, and Wilcoxon rank-sum test.

SNP dataset and statistical methods

In genome-wide association (GWA) studies, researchers focus on the positions of genetic variants that are significantly related to the phenotype of interest. There is no gold standard for pre-processing such data, but a few guidelines exist. Many steps, such as normalization and bias removal are included in data pre-processing, and the analysis results are very dependent on those steps. In our analysis, we used a bipolar disorder data in the WTCCC database, which includes 354,019 SNPs from 4,806 individuals (1,868 bipolar disorder patients and 2,938 normal controls). As a first step, we conducted a quality control process based on specific criteria [15]. For the association test between genotype and phenotype, using SNP data, we used statistical methods such as chi-square test, Fisher's exact test, logistic regression with covariate adjusting, and logistic regression without covariate adjusting. These association tests were implemented using the PLINK tool.

RNA-seq dataset and statistical methods

We employed results from RNA-seq, another type of transcriptome measuring platform. Recently, its advantages over microarray platforms have been described by many comparative reports [16]. Thus, a more elaborated estimation became possible by RNA-seq, in short. However, RNA-seq gene expression is measured in counts (i.e., number of strands synthesized), and therefore direct application of RNA-seq methods to microarray analysis is impossible. Instead, RNA-seq analysis methods are developed by applying statistical methodologies based on analyzing serial analysis of gene expression (SAGE) platform data, a traditional approach for measuring gene expression in counts. Here, we employed RNA-seq data from a previous study [17] using edgeR, DESeq, and NBPSeq methods. The RNA-seq data from a MicroArray Quality Control Project (MAQC) had 7 replicates and one pooled sample each from two types of samples, Ambion's (Austin, TX, USA) human brain reference RNA, and Stratagene's (Santa Clara, CA, USA) human universal reference RNA. After filtering out the NA values; 10,473 genes remained, with three DE-analysis methods.

NGS dataset and statistical methods

Shortcomings of common variants in explaining the whole heritability of diseases has led to the study of rare variants. Unlike common variants, rare variant analyses, based on single genetic associations, often shows large false-negative results, unless the sample or effect sizes are very large. Hence, collapsed genotype scores for a set of rare variants are suggested for an analysis scheme. For our input, we employed the results from rare variant association tests such as C-alpha, burden test, and SKAT-O. These association tests were implemented using the FARVAT tool [18]. For illustrative purposes, we used the simulation dataset of FARVAT consisting of 100 SNPs and 16 genes which was enlarged to have 10,000 SNPs and 2,000 genes, using the same settings.

Implementation of GRACOMICS

GRACOMICS is a java-based stand-alone program using a GUI platform. It was developed under Java because statistical analysis tools are generally developed by diverse codes such as R, SAS, etc. Java programs are renowned for their compatibility with various computing environments, are supported by all operating systems, and can easily be executed by other programs written in different computer languages. GRACOMICS can read tabular types of tab-separated values (TSV) files containing p-values for each method in columns and genetic markers in rows. Also, using simple mouse clicks, rather than command lines as input, helps bridge the gap between biology-based researchers and computer science-based researchers. Our plan was to design and implement a user-friendly program any researcher could use in any environment. The proposed tool, GRACOMICS, has the following three interactive modules with distinct features:

Pairwise Comprehensive Scatter Plots Module (Pair-CSP)

Pair-CSP provides a scatter plot of pairwise comparisons between statistical method inputs simultaneously (Figures 1 and 2). Pair-CSP automatically generates these pairwise scatterplots using the p-values from the input file(s), letting the user interpret the similarities between the test results through correlation plots and correlation coefficients at a glance. When the significance level is manipulated, the pairwise scatterplots change accordingly, to display markers over the threshold only. There are two reasons behind this feature: one is to reduce computational time for drawing multitudinous points, and the other is to show only what the researcher wants to see, i.e., the meaningful results.

Pairwise Detailed Scatter Plot Module (Pair-DSP)

Pair-DSP is an interactive plot to compare the results between two methods on a more detailed level than Pair-CSP (Figures 3 and 4). This module is linked to Pair-CSP, enabling the user to directly access Pair-DSP from Pair-CSP for extended summarization of the chosen biomarkers. The summary organizes meaningful results via a Venn diagram, a table, and a marker list. For the known marker's function, simple annotation of a single biomarker is offered via the NCBI database. Its simple annotation function automatically provides a link to the NCBI web page corresponding to its marker type, for convenience. In addition, for pathway analysis of microarray data, GRACOMICS connects to the web-based DAVID database. As a result, researchers can summarize their list of significant results, and then check the biological functions of the chosen markers.

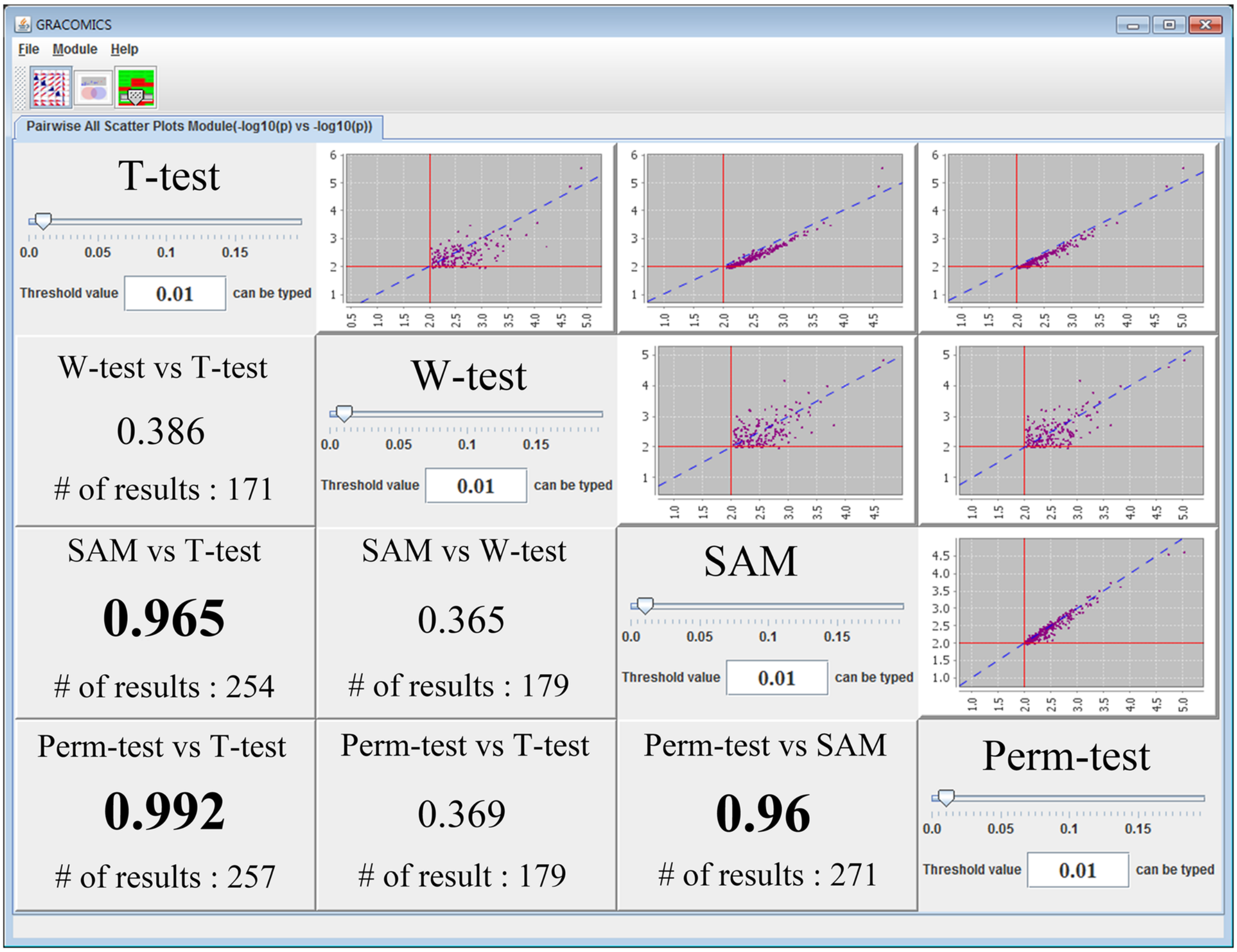
Multiple Results Comparison Module (Multi-RC)

Multi-RC provides simultaneous comparison of numerous test results (Figures 5 and 6). Researchers can choose an interesting subset of methods and set their significance levels separately. A tabular output with rows as significant markers and columns as statistical methods, is provided (with p-values in each cell). Each cell is color-coded red or green, representing significant or not, respectively. Also, variation of color intensities are used to represent the degree of significance, with more significant markers colored more intensely. In addition, Multi-RC summarizes commonly significant results and provide links to their annotation. As an extra option (with a checkbox) for meta-studies, we implemented Fisher's method in combining p-values to provide overall importance in version 1.1.

Results

Application of GRACOMICS to real microarray data

In Figure 1, the plots provided by Pair-CSP compare the test results of *t*-test, Wilcoxon rank-sum test, SAM, and permutation test, displaying the top 1,500 markers by their average p-values (the user can designate the number or percentage of markers to be displayed). Pair-CSP reveals a close relationship between each pair of methods; most correlation coefficients are over 0.9, except for those with the Wilcoxon rank-sum test. Although both Wilcoxon rank-sum and permutation tests are nonparametric tests, the Wilcoxon rank-sum test uses only rank information, while the permutation test uses the variance



T-test — Threshold value 0.01

W-test vs T-test 0.386 # of results : 171

W-test — Threshold value 0.01

SAM vs T-test **0.965** # of results : 254

SAM vs W-test 0.365 # of results : 179

SAM — Threshold value 0.01

Perm-test vs T-test **0.992** # of results : 257

Perm-test vs T-test 0.369 # of result : 179

Perm-test vs SAM **0.96** # of results : 271

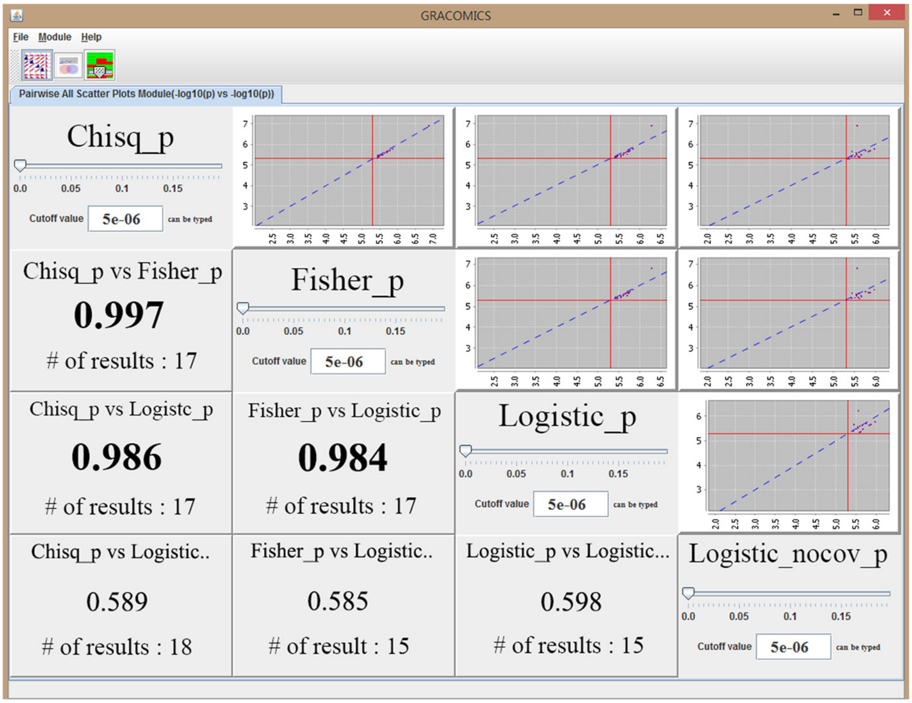
Perm-test — Threshold value 0.01

Figure 1 Pair-CSP plot with GSE27567 data. Four test results were compared, and all pairwise scatterplots and their correlation coefficients are given in the Pair-CSP module.

information that arises when defining *t*-test statistics. Thus, they provide different results.

In order to compare the Wilcoxon rank-sum test to other tests more systematically, we used Pair-DSP focusing on the *t*-test and the Wilcoxon rank-sum test. As shown in Figure 3, Pair-DSP displays a pairwise plot of the two methods using p-values, and summarizes the number of genes commonly identified by the two methods. Unlike the pairwise plot of Pair-CSP, the pairwise plot of Pair-DSP shows far more detailed information. For example, a red color represents the significant genes identified by *t*-test only, a blue color signifies those identified by Wilcoxon rank-sum test only, and purple color indicates those identified by both tests. The gene name, in tool tip form, of a point is provided when the cursor is put directly over the single point. The summary table, at the top right, shows a decrease in the number of significant genes commonly identified by the two methods goes from 1,049 to 12, as the cut-off value decreases from 5% to 0.1%. Pair-DSP also provides a Venn diagram displaying the numbers of genes identified commonly and separately by the two methods. Pair-DSP shows that 171 genes remained significant by both t-tests and Wilcoxon rank-sum tests at the 1% significance level. 86 genes were significant by *t*-test only and 141 genes by Wilcoxon rank-sum test only, at the same significance level. The bottom right table shows the list of genes identified by the two methods.

To investigate the functions of the identified genes, simple annotation is provided via the NCBI database. This simple annotation function automatically opens a link to the NCBI web page corresponding to the gene of interest, for convenience. In addition, for a pathway analysis annotation database, GRACOMICS provides connection to the web-based DAVID database [19]. For example, clicking the gene *Cyyr1*, followed by a right click shows a popup window with two menus of “Link to NCBI annotation database” and “Link to DAVID



Chisq_p
Cutoff value 5e-06

Chisq_p vs Fisher_p
0.997
# of results : 17

Fisher_p
Cutoff value 5e-06

Chisq_p vs Logistc_p
0.986
# of results : 17

Fisher_p vs Logistic_p
0.984
# of results : 17

Logistic_p
Cutoff value 5e-06

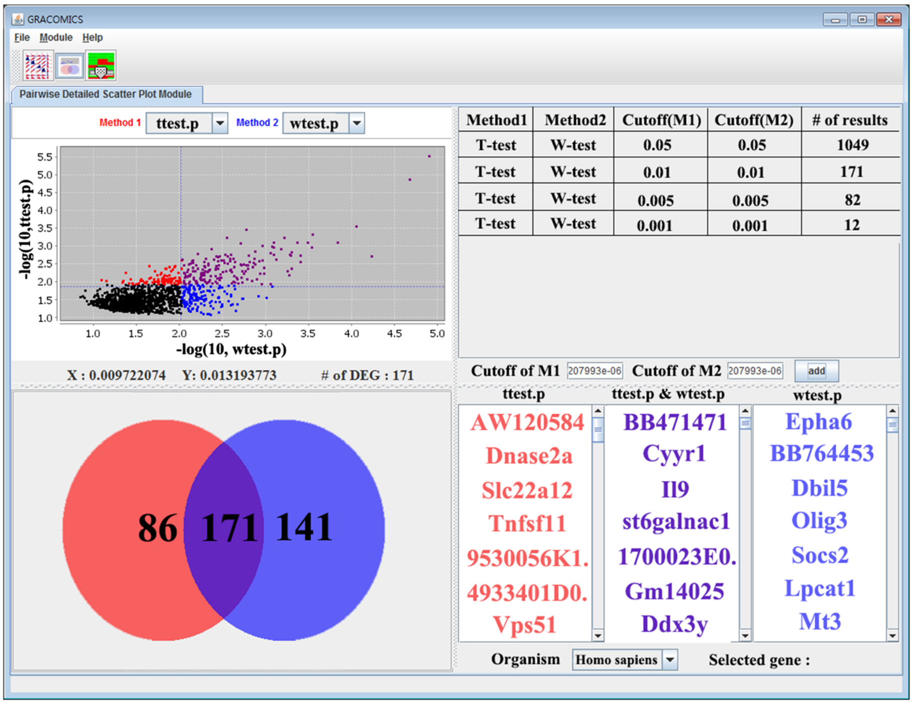
Chisq_p vs Logistic..
0.589
# of results : 18

Fisher_p vs Logistic..
0.585
# of result : 15

Logistic_p vs Logistic...
0.598
# of results : 15

Logistic_nocov_p
Cutoff value 5e-06

Figure 2 Pair-CSP plot with WTCCC SNP data. Four tests results were compared, and all pairwise scatterplots and their correlation coefficients are given in Pair-CSP.



Method 1 ttest.p Method 2 wtest.p

X : 0.009722074 Y: 0.013193773 # of DEG : 171

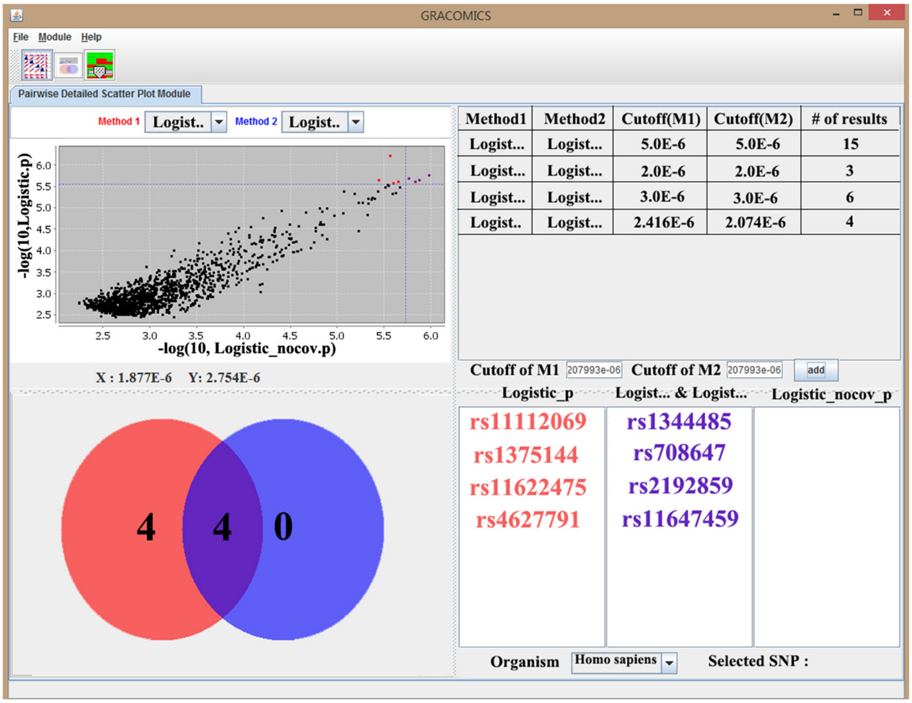
Method1	Method2	Cutoff(M1)	Cutoff(M2)	# of results
T-test	W-test	0.05	0.05	1049
T-test	W-test	0.01	0.01	171
T-test	W-test	0.005	0.005	82
T-test	W-test	0.001	0.001	12

Cutoff of M1 207993e-06 Cutoff of M2 207993e-06

ttest.p	ttest.p & wtest.p	wtest.p
AW120584	BB471471	Epha6
Dnase2a	Cyyr1	BB764453
Slc22a12	Il9	Dbil5
Tnfsf11	st6galnac1	Olig3
9530056K1.	1700023E0.	Socs2
4933401D0.	Gm14025	Lpcat1
Vps51	Ddx3y	Mt3

Organism Homo sapiens Selected gene :

Figure 3 Pair-DSP plot with GSE27567 data. Wilcoxon rank sum tests and *t*-tests were chosen for detailed investigation, Venn diagram and the summary tables are key features of Pair-DSP.



X : 1.877E-6 Y: 2.754E-6

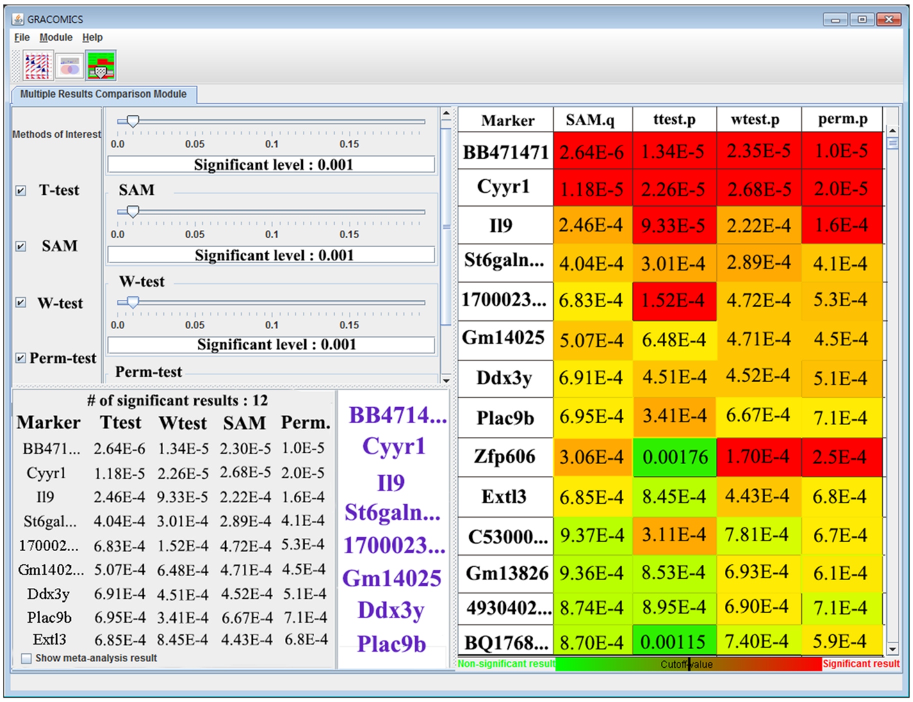
Method1	Method2	Cutoff(M1)	Cutoff(M2)	# of results
Logist...	Logist...	5.0E-6	5.0E-6	15
Logist...	Logist...	2.0E-6	2.0E-6	3
Logist...	Logist...	3.0E-6	3.0E-6	6
Logist..	Logist...	2.416E-6	2.074E-6	4

Cutoff of M1 207993e-06 Cutoff of M2 207993e-06

Logistic_p	Logist... & Logist...	Logistic_nocov_p
rs11112069	rs1344485	
rs1375144	rs708647	
rs11622475	rs2192859	
rs4627791	rs11647459	

Organism: Homo sapiens Selected SNP :

Figure 4 Pair-DSP plot with WTCCC SNP data. Two logistic models, one with and the other without covariates, has been chosen for detailed investigation, Venn diagrams and the summary tables are key features of Pair-DSP.



☑ T-test Significant level : 0.001

☑ SAM Significant level : 0.001

☑ W-test Significant level : 0.001

☑ Perm-test

# of significant results : 12

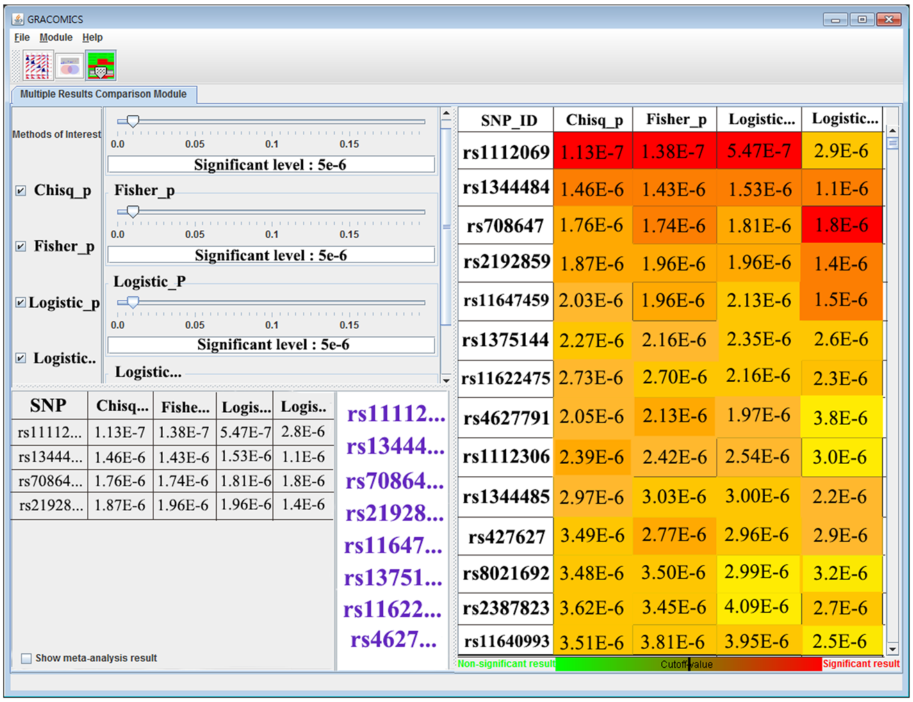
Marker	Ttest	Wtest	SAM	Perm.
BB471...	2.64E-6	1.34E-5	2.30E-5	1.0E-5
Cyyr1	1.18E-5	2.26E-5	2.68E-5	2.0E-5
Il9	2.46E-4	9.33E-5	2.22E-4	1.6E-4
St6gal...	4.04E-4	3.01E-4	2.89E-4	4.1E-4
170002...	6.83E-4	1.52E-4	4.72E-4	5.3E-4
Gm1402...	5.07E-4	6.48E-4	4.71E-4	4.5E-4
Ddx3y	6.91E-4	4.51E-4	4.52E-4	5.1E-4
Plac9b	6.95E-4	3.41E-4	6.67E-4	7.1E-4
Extl3	6.85E-4	8.45E-4	4.43E-4	6.8E-4

□ Show meta-analysis result

BB4714...
Cyyr1
Il9
St6galn...
1700023...
Gm14025
Ddx3y
Plac9b

Marker	SAM.q	ttest.p	wtest.p	perm.p
BB471471	2.64E-6	1.34E-5	2.35E-5	1.0E-5
Cyyr1	1.18E-5	2.26E-5	2.68E-5	2.0E-5
Il9	2.46E-4	9.33E-5	2.22E-4	1.6E-4
St6galn...	4.04E-4	3.01E-4	2.89E-4	4.1E-4
1700023...	6.83E-4	1.52E-4	4.72E-4	5.3E-4
Gm14025	5.07E-4	6.48E-4	4.71E-4	4.5E-4
Ddx3y	6.91E-4	4.51E-4	4.52E-4	5.1E-4
Plac9b	6.95E-4	3.41E-4	6.67E-4	7.1E-4
Zfp606	3.06E-4	0.00176	1.70E-4	2.5E-4
Extl3	6.85E-4	8.45E-4	4.43E-4	6.8E-4
C53000...	9.37E-4	3.11E-4	7.81E-4	6.7E-4
Gm13826	9.36E-4	8.53E-4	6.93E-4	6.1E-4
4930402...	8.74E-4	8.95E-4	6.90E-4	7.1E-4
BQ1768...	8.70E-4	0.00115	7.40E-4	5.9E-4

Figure 5 Multi-RC plot with GSE27567 data. The Multi-RC module provides an overall summary in a heatmap-like tabular format which highlights markers with the lowest average p-values. The user can then choose which methods to investigate by using the checkboxes in the top-left panel.



☑ Chisq_p — Significant level : 5e-6

☑ Fisher_p — Significant level : 5e-6

☑ Logistic_p — Significant level : 5e-6

☑ Logistic..

SNP	Chisq...	Fishe...	Logis...	Logis..
rs11112...	1.13E-7	1.38E-7	5.47E-7	2.8E-6
rs13444...	1.46E-6	1.43E-6	1.53E-6	1.1E-6
rs70864...	1.76E-6	1.74E-6	1.81E-6	1.8E-6
rs21928...	1.87E-6	1.96E-6	1.96E-6	1.4E-6

rs11112...
rs13444...
rs70864...
rs21928...
rs11647...
rs13751...
rs11622...
rs4627...

☐ Show meta-analysis result

SNP_ID	Chisq_p	Fisher_p	Logistic...	Logistic...
rs1112069	1.13E-7	1.38E-7	5.47E-7	2.9E-6
rs1344484	1.46E-6	1.43E-6	1.53E-6	1.1E-6
rs708647	1.76E-6	1.74E-6	1.81E-6	1.8E-6
rs2192859	1.87E-6	1.96E-6	1.96E-6	1.4E-6
rs11647459	2.03E-6	1.96E-6	2.13E-6	1.5E-6
rs1375144	2.27E-6	2.16E-6	2.35E-6	2.6E-6
rs11622475	2.73E-6	2.70E-6	2.16E-6	2.3E-6
rs4627791	2.05E-6	2.13E-6	1.97E-6	3.8E-6
rs1112306	2.39E-6	2.42E-6	2.54E-6	3.0E-6
rs1344485	2.97E-6	3.03E-6	3.00E-6	2.2E-6
rs427627	3.49E-6	2.77E-6	2.96E-6	2.9E-6
rs8021692	3.48E-6	3.50E-6	2.99E-6	3.2E-6
rs2387823	3.62E-6	3.45E-6	4.09E-6	2.7E-6
rs11640993	3.51E-6	3.81E-6	3.95E-6	2.5E-6

Figure 6 Multi-RC plot with WTCCC SNP data. The Multi-RC module provides an overall summary in a heatmap-like tabular format which highlights markers with the lowest average p-values. Note rs1112069 is colored in red by 3 of the 4 tests, as discussed in the manuscript.

annotation database". From the NCBI database, researchers can investigate known gene functions, and related papers in PubMed, for each gene. We observed that *Cyyr1* [20] and *Il9* [21] are genes reported in PubMed. Next, when using DAVID to analyze the functional annotation of the 171 commonly identified genes from t-tests and Wilcoxon rank-sum tests, we observed the gene list to be enriched in the GO term "cell cycle arrest," with a p-value of 4.1e-3. As a result, researchers can summarize their list of significant results, and then check the biological functions and related publications of the chosen markers.

The Multi-RC module allows simultaneous comparison of two or more results, as shown in Figure 5. We selected four methods: *t*-test, SAM, Wilcoxon rank-sum test, and permutation test, with a cut-off value of 0.1%. In this setting, we observed 12 common significant genes between all the methods. The genes *BB471471*, *Cyyr1*, *Il9*, and *St6galnac1* [22] were consistent candidates from all four methods. However, while *BB471471* was at the top of the list, no reports were found of its association with tumours or any other diseases. Therefore, we suggest the *BB471471* is a worthy candidate to examine further for its possible association with tumours. By analyzing this real microarray data analysis with GRACOMICS, we identified several commonly significant DEGs from comparisons from each method, to obtain the most reliable candidate DEGs.

Application of GRACOMICS to real SNP data

In Figure 2, the plots are provided by Pair-CSP, which compares the test results of chi-square test, Fisher's exact test, and logistic regression analyses. In the figure, two results from logistic regression analyses are provided: one is without covariates and the other is with the adjusting covariate effects of sex, age and the first two principal components. Although the significance of covariates can be easily tested, it is not always straightforward to determine which adjusting covariates to include in the model [23]. Here, we focused on the results from the two logistic models and demonstrate how efficiently GRACOMICS can be used to compare these two results, showing that the correlation between the two logistic regression models was 0.598.

For a further detailed comparison between these two results, Pair-DSP, in Figure 4, was conducted on these two logistic models. The summary table, at the top right, shows that the number of significant genes commonly identified by the two methods gradually decreases from 15 to 4, as the cut-off value decreases from 5.0e-6 to 2.4e-6. The Venn diagram illustrates that Pair-DSP successfully identified rs1344484 [24], rs708647, rs2192859 [25], rs11647459 and rs4627791 [26], in purple, as the most commonly detected SNPs. The four SNPs in red, rs11112069, rs1375144, rs11622475, and rs4627791, were detected by the with-covariates model only. We found rs11112069 as the top result (in average p-value), with low p-values in all four analyses. This SNP is within intron-2 of *CHST11*, a gene which has previously been reported as bipolar disorder-associated [27].

In the next module (Multi-RC; Figure 6), users can see the change in p-values for each marker, according to the method used or adjustments for covariates. Rs11112069 is displayed at the top of the list, and is marked in red (very significant) from 3 of the 4 tests, with a fairly low p-value for the fourth test also. To further analyze the top results, GRACOMICS can automatically distinguish marker types and links to dbSNP in the NCBI database for selected SNPs. From the annotation, researchers can attain detailed SNP information, such as location of the SNP, its mapped gene, clinical significance, etc. Unlike the microarray example, DAVID is not directly applicable to SNP data. However, we expect that other annotation databases will be added to future updates.

Application of GRACOMICS to real RNA-seq data

As shown in Additional file 1: Figure S1, Pair-CSP shows that all three analysis methods; edgeR, DESeq, and NBPSeq, yield very similar results. All of the correlation coefficients are over 0.86, and the highest was between edgeR and DESeq. In addition, the plots illustrate that edgeR generates lower P-values than the others, due to scattered points being skewed toward the y-axis (edgeR). Under the 1% significance level, approximately 7,000 genes were detected as DEGs by each method. In Additional file 2: Figure S2, Pair-DSP shows that more DEGs were identified by edgeR, as compared to DESeq. In the Venn diagram, significant genes that intersected ranged from 7087 to 1621, when decreasing the cut-off values from 0.01 to 1.0E-100. Finally, we can observe that most of genes are very significant in Multi-RC. As shown in Additional file 3: Figure S3, 6983 genes were detected by all the methods under a 1% significant level. Here, the gene symbol of RNA-seq data is its Ensemble ID, and these should be converted to official gene symbols for successful functional annotation. Although implementing the Ensemble annotation function on the web is possible, we did not include it in the current version of GRACOMICS, because accommodating several symbols in the program may lead to user confusion. Although we determined that only official gene symbols should be accommodated in the tool, later versions can be updated with such functions, if there are user demands.

Application of GRACOMICS to simulated NGS data

Using a simulated rare variant dataset, we successfully cataloged significant genes that were test-specific or marginal in all tests. The results are shown in Additional file 4: Figures S4, Additional file 5: Figure S5 and Additional file 6: Figure S6. In this analysis, gene names were masked as Genes 1 ~ 2000 and therefore could not be annotated to NCBI or DAVID. However, if a real dataset is used, the genes can be annotated in similar fashion as microarray and RNA-seq datasets. In accord with the above three applications, we could infer which methods showed higher correlation, in terms of p-values, from the Pair-CSP, followed by a detailed comparison of the number of significant genes detected in each method, and finally, by comparing the p-values in a tabular heatmap form. Here, we observed the highest correlation of 0.961 between the C-Alpha and SKAT-O methods, and these two methods shared 129 genes with a p-value threshold of < 0.05. The top-ranked genes all showed p-values ~ 0.001 using all the methods, and would be candidates of interest for end-users if this was a real data analysis.

Discussion

From the aforesaid illustration, we demonstrated the potential of GRACOMICS to successfully highlight biologically meaningful results from multiple methods. Traditional bioinformatics studies, and some recent works, show that simple comparison of results has been widely used for biological interpretation. For example, a transcriptome study concluded that in a situation where the most reliable list of markers is desirable, the best approach was to examine the intersection of genes identified by all tried methods, or by more conservative tests. Since checking the underlying assumptions of all methods is not easy, and even if the assumptions are met, each method may provide different results, which are hard to interpret. The easiest and most conventional method is to find commonly identified markers to trim down the candidate list, and carry on further analysis. While GRACOMICS cannot give conclusive evidence that the highlighted markers are significant, it can help the biologist narrow down the candidate list, based on the intersection of markers for efficiency for further validations, such as RT-PCR.

In addition to comparison of multiple results of the same datasets, GRACOMICS can be applicable to other types of studies. First, GRACOMICS can compare the results from different datasets, such as different tissues or organs. An RNA study compared differentially expressed test results from various tissues, such as liver, adipose tissue, muscle, and brain. GRACOMICS can effectively provide the list of common genes, as well as tissue-specific genes. Second, GRACOMICS can compare results from different platforms, such as microarray vs. RNA-seq [28]. Here, GRACOMICS can trim down the list of candidates significant to both platform results, for further biological validation. Finally, meta-analysis combining independent results from different studies can be analyzed by GRACOMICS; the p-values from each study can be efficiently compared to others and can be combined easily by Fisher's method. For meta-analysis, the compared results should be from independent datasets. However, when one single dataset was analyzed by multiple methods, the independent assumption is violated; the interpretation of this Fisher's combined p-value should be made with caution.

Conclusions

Comparative study of omics data analyses is unavoidable; however, many researchers skip the comparative step because it is a complicated process. GRACOMICS enables easy comparison of several methods for analyzing specific omics data platforms by any user. The four omics data employed are active areas of study in bioinformatics. We employed microarray & RNA-seq data at the transcriptomic level, and SNP and NGS data at the genomic level, to display the utility of GRACOMICS. So far, GRACOMICS can also employ proteomic analysis results, and will be extended to accommodate other types of annotations for proteomics data in a future study. In summary, we believe that this will be a highly valuable and straightforward tool for non-computational biologists, strongly assisting them in their interpretation of results from new cutting-edge technologies.

Availability and requirements

Project name: GRACOMICS
Project home page: http://bibs.snu.ac.kr/software/GRACOMICS
Operating system: Platform-independent
Programming language: Java
Other requirements: Java 1.7.0_45 or higher
License: LGPL 2.1

Additional files

Additional file 1: Figure S1. Pair-CSP plot with MAQC RNA-seq data. Three tests results have been compared, and all pairwise scatterplots and their correlation coefficients are given on Pair-CSP.

Additional file 2: Figure S2. Pair-DSP plot with MAQC RNA-seq data. EdgeR and DESeq were chosen for detailed investigation. Venn diagrams and the summary tables are key features of Pair-DSP.

Additional file 3: Figure S3. Multi-RC plot with MAQC RNA-seq data. The Multi-RC module provides an overall summary in a heatmap-like tabular format which highlights markers with the lowest average p-values.

Additional file 4: Figure S4. Pair-CSP plot with simulated NGS data. Three tests results were compared, and all pairwise scatterplots and their correlation coefficients are given on the Pair-CSP GUI.

Additional file 5: Figure S5. Pair-DSP plot with simulated NGS data. C-alpha and SKAT-O were chosen for detailed investigation. Venn diagrams and the summary tables are key features of Pair-DSP.

Additional file 6: Figure S6. The Multi-RC plot with simulated NGS data. The Multi-RC module provides an overall summary in a heatmap-like tabular format which highlights markers with the lowest average p-values.

Abbreviations

SNP: Single nucleotide polymorphism; GUI: Graphic user interface; NCBI: National center for biotechnology information; WTCCC: Wellcome trust case control consortium; Mev: Multi experiment viewer; Pair-CSP: Pairwise comprehensive scatter plots module; Pair-DSP: Pairwise detailed scatter plot module; Multi-RC: Multiple results comparison module; DEGs: Differentially expressed genes; GEO: Gene expression omnibus; SAM: Significant analysis of microarray; GWA: Genome-wide association; TSV: Tab separated values TSV; SAGE: Serial analysis of gene expression; MAQC: Microarray quaility control project.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MS wrote the software, carried out the analyses, and contributed to writing the manuscript. JY contributed to program testing, carried out the analyses, and wrote the manuscript. TP managed the whole project and contributed to the writing. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (No. 2012R1A3A2026438, 2008–0062618) and by the Bio & Medical Technology Development Program of the NRF grant (2013M3A9C4078158).

Received: 14 October 2014 Accepted: 9 March 2015
Published online: 01 April 2015

References

1. Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics. 2002;18(4):546–54.
2. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics. 2005;6(2):95–108.
3. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic acids research. 2001;29(1):308–11.
4. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, et al. The international HapMap project. Nature. 2003;426(6968):789–96.
5. Anders S, Huber W. Differential expression analysis for sequence count data. Genome biol. 2010;11(10):R106.
6. Di Y, Schafer DW, Cumbie JS, Chang JH. The NBP negative binomial model for assessing differential gene expression from RNA-Seq. Statistical Applications in Genetics and Molecular Biology. 2011;10(1):1–28.
7. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
8. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, et al. Testing for an unusual distribution of rare variants. PLoS genetics. 2011;7(3):e1001322.
9. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. The American Journal of Human Genetics. 2011;89(1):82–93.
10. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. The American Journal of Human Genetics. 2012;91(2):224–37.
11. Howe E, Holton K, Nair S, Schlauch D, Sinha R, Quackenbush J. Mev: multiexperiment viewer. In: Biomedical Informatics for Cancer Research. Springer US; 2010. p. 267–77.
12. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics. 2007;81(3):559–75.
13. LaBreche HG, Nevins JR, Huang E. Integrating factor analysis and a transgenic mouse model to reveal a peripheral blood predictor of breast tumors. BMC medical genomics. 2011;4(1):61.
14. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences. 2001;98(9):5116–21.

15. Oh S, Lee J, Kwon M-S, Weir B, Ha K, Park T. A novel method to identify high order gene-gene interactions in genome-wide association studies: Gene-based MDR. BMC bioinformatics. 2012;13 Suppl 9:S5.
16. Morozova O, Hirst M, Marra MA. Applications of new sequencing technologies for transcriptome analysis. Annual review of genomics and human genetics. 2009;10:135–51.
17. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC bioinformatics. 2010;11(1):94.
18. Choi S, Lee S, Nöthen MM, Lange C, Park T, Won S. FARVAT: a family-based rare variant association test. Bioinformatics. 2014;30(22):3197–205.
19. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols. 2008;4(1):44–57.
20. Vitale L, Frabetti F, Huntsman SA, Canaider S, Casadei R, Lenzi L, et al. Sequence. BMC cancer. 2007;7(1):66.
21. Nagato T, Kobayashi H, Kishibe K, Takahara M, Ogino T, Ishii H, et al. Expression of interleukin-9 in nasal natural killer/T-cell lymphoma cell lines and patients. Clinical cancer research. 2005;11(23):8250–7.
22. PATANI N, Jiang W, MOKBEL K. Prognostic utility of glycosyltransferase expression in breast cancer. Cancer Genomics-Proteomics. 2008;5(6):333–40.
23. Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics. 2002;18(11):1454–61.
24. Palo OM. Genetic background of bipolar disorder and related cognitive impairments. 2010.
25. Kwon M-S, Park M, Park T. IGENT: efficient entropy based algorithm for genome-wide gene-gene interaction analysis. BMC medical genomics. 2014;7 Suppl 1:S6.
26. Jiang Y, Zhang H. Propensity score-based nonparametric test revealing genetic variants underlying bipolar disorder. Genetic epidemiology. 2011;35(2):125–32.
27. Chen Y-H, Lu R-B, Hung H, Kuo P-H. Identifying Potential Regions of Copy Number Variation for Bipolar Disorder. Microarrays. 2014;3(1):52–71.
28. Nookaew I, Papini M, Pornputtpong N, Scalcinati G, Fagerberg L, Uhlén M, et al. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic acids research. 2012;40(20):10084–97.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central