BMC Bioinformatics

Software

CHESS (CgHExpreSS): A comprehensive analysis tool for the analysis of genomic alterations and their effects on the expression profile of the genome

Mikyung Lee and Yangseok Kim*

Address: Department of Physiology, College of Oriental Medicine, KyungHee University, #1 Hoegi-Dong, Dongdaemun-gu, Seoul 130-701, South Korea

Email: Mikyung Lee - mklee77@khu.ac.kr; Yangseok Kim* - yskim1158@khu.ac.kr

* Corresponding author

This article is available from: http://www.biomedcentral.com/1471-2105/10/424

Abstract

Background: Genomic alterations frequently occur in many cancer patients and play important mechanistic roles in the pathogenesis of cancer. Furthermore, they can modify the expression level of genes due to altered copy number in the corresponding region of the chromosome. An accumulating body of evidence supports the possibility that strong genome-wide correlation exists between DNA content and gene expression. Therefore, more comprehensive analysis is needed to quantify the relationship between genomic alteration and gene expression. A well-designed bioinformatics tool is essential to perform this kind of integrative analysis. A few programs have already been introduced for integrative analysis. However, there are many limitations in their performance of comprehensive integrated analysis using published software because of limitations in implemented algorithms and visualization modules.

Results: To address this issue, we have implemented the Java-based program CHESS to allow integrative analysis of two experimental data sets: genomic alteration and genome-wide expression profile. CHESS is composed of a genomic alteration analysis module and an integrative analysis module. The genomic alteration analysis module detects genomic alteration by applying a threshold based method or SW-ARRAY algorithm and investigates whether the detected alteration is phenotype specific or not. On the other hand, the integrative analysis module measures the genomic alteration's influence on gene expression. It is divided into two separate parts. The first part calculates overall correlation between comparative genomic hybridization ratio and gene expression level by applying following three statistical methods: simple linear regression, Spearman rank correlation and Pearson's correlation. In the second part, CHESS detects the genes that are differentially expressed according to the genomic alteration pattern with three alternative statistical approaches: Student's t-test, Fisher's exact test and Chi square test. By successive operations of two modules, users can clarify how gene expression levels are affected by the phenotype specific genomic alterations. As CHESS was developed in both Java application and web environments, it can be run on a web browser or a local machine. It also supports all experimental platforms if a properly formatted text file is provided to include the chromosomal position of probes and their gene identifiers.

Conclusions: CHESS is a user-friendly tool for investigating disease specific genomic alterations and quantitative relationships between those genomic alterations and genome-wide gene expression profiling.

Background

It is well-known that genomic alterations frequently occur in many cancer patients and play important mechanistic roles in pathogenesis of cancer. Recently, in order to identify genomic alteration regions, a comparative genomic hybridization (CGH) technology has been extensively applied to various types of cancer cases. CGH is a molecular cytogenetic method that detects gain or loss of genomic DNA content of an individual, which is accomplished by measuring the ratio between the intensity of test DNA and that of reference DNA [1]. As the technology of array-based CGH has advanced, the resulting unprecedented detailed examination of chromosomal regions has led to efforts to discover genomic alterations as the genetic markers for various diseases [2,3]. Those efforts have served to emphasize the fact that genomic alterations play important roles in a particular disease. Furthermore, genomic alterations can modify the expression level of genes due to changed copy number in the relevant chromosomal regions. Recent studies have been concerned with verifying the existence of a strong genome-wide correlation between DNA content and gene expression [4,5]. These important CGH-based biological discoveries have spurred the widespread use of the technique, which has prompted the need for a genomic alteration analysis tool.

Accordingly, statistical algorithms and programs such as GEAR, aCGHViewer, and CAPweb have been developed to analyze and visualize CGH data. However, these programs have concentrated on analyzing genomic alteration and do not deal with an integrative analysis of genomic alteration and gene expression [6-8]. A few non commercial programs have been introduced for integrative analysis such as ACE-it, ISACGH, VAMP, and SIGMA² [9-12]. While ACE-it and ISACGH are focused on the analysis of genomic CGH and gene expression, VAMP and SIGMA² are implemented to deal with various genomic molecular profiles. Those programs are well implemented according to their analysis goals. However, they have limited algorithms and visualization modules, it is not easy to perform comprehensive analysis of integrated analysis of genomic CGH and gene expression. Table 1 compares the features of the introduced programs, the number in parenthesis means the number of offered statistical methods. Additionally, some programs need the non familiar package like a R. Therefore, we have developed an integrative analysis tool named CHESS to allow comprehensive analysis of genomic CGH data and gene expression data within an easy-to-use single consistent interface using a variety of statistical methods. CHESS is composed of two successive modules: a genomic alteration analysis module and an integrative analysis module of genomic alteration and gene expression. The complete successive operations of these two modules provide two kinds of biologically meaningful information: phenotype or class specific genomic alteration regions, and a list of genes whose expressions are affected by those genomic alteration regions. CHESS supports all experimental platforms if a properly formatted text file is provided to include chromosomal position and their gene symbol of probes. We presently provide information concerning CHESS. A detailed description of the entire procedure and file format is provided in the online manual; http://bios tone.khu.ac.kr/CHESS/.

Implementation

We chose Java as the programming language because it is publicly available and ensures cross-platform compatibility. Moreover, CHESS was developed in both Java application and webstart environments; it can be run on a web browser or a local machine. If CHESS is operated through a web browser, any uploaded data is not transmitted anywhere because all the analysis are performed locally using webstart. CHESS can deal with high-density arrays on commonly used desktop computers. For example, we were able to load 30 Agilent 244 k arrays in 3 min on a computer with 3 GB of memory and a 2.4 Ghz processor.

Results
Analysis flow
CHESS is composed of two primary modules: a genomic alteration analysis module and an integrative analysis module. The first module is responsible for the detection of genomic alteration regions and investigates whether the detected regions are phenotype specific or not. Genes located in the altered regions are automatically listed and

Table 1: CHESS's main features

| Features	ACE-it	ISACGH	VAMP	SIGMA²	CGH Analytics	CHESS
Built-in GA detection		3	3	3	3	3
Frequent GA region definition		3	3	3	3	3
Case/control study				3		3
Enrichment analysis for GO and KEGG		3				3
Integrative analysis of genomic CGH and gene expression	3(1)	3(2)	3(2)	3(3)	3(2)	3(6)
Free for academic	3	3	3	3	3	3
Web accessible		3	3			3

The number in parenthesis means the number of offered statistical methods

biological information is given by an implemented annotation module. The integrative analysis module includes a combined analysis of genomic alteration and gene expression. For user's rapid understanding of the complete analyzed results, CHESS provides a resulting figure on a whole chromosomal scale. CHESS also provides the analyzed results on a single chromosomal scale for further detailed analysis.

Data Requirements

CHESS was developed to support all experimental platforms if only a properly formatted text file is given including general information such as probe identification, chromosomal location and normalized log transformed ratio values for each probe. CHESS handles two sets of experimental data: signal ratio values from a CGH experiment and those from a gene expression experiment. Firstly, the CGH ratio file acquired from the CGH experi-

ment should define header information in the first four lines; total number of samples, total number of probes, sample names and their clinical information. The test/reference ratio values should be written from the fourth line, in which the first five fields record probe identification, probe name or alias, chromosome number, start position and stop position with tab separations. The rest of the columns are considered as the ratio values for each sample. When the user successfully loads the CGH ratio file and finishes detecting the genomic alteration regions, a GA (Genomic Alteration) file is automatically generated that has the exactly the same format as the CGH ratio file. However, the GA file contains discrete values concerning genomic alteration in the form of gain (+1), loss (-1) and no genomic alteration (0), instead of the continuous signal values of the CGH ratio file. To support the output from other detection algorithms as aCGH [13] and DNAcopy [14], CHESS allows direct loading of the GA file. Sec-
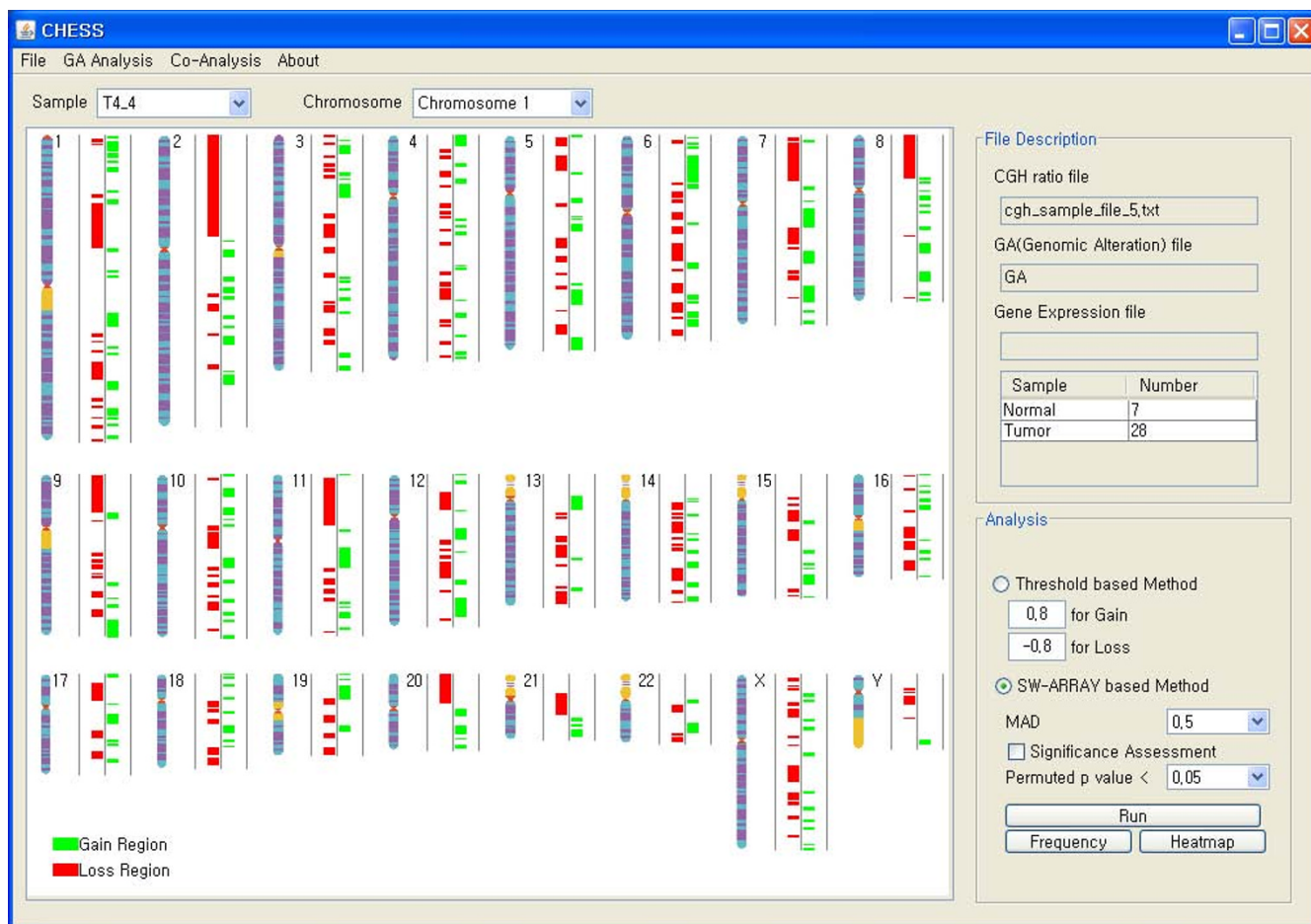
Figure 1
Screenshot of detection of genomic alteration region. After detection of genomic alterations, CHESS displays graphically detected altered regions for each sample on the whole chromosomal mode, in which green and red indicate an amplified or deleted region, respectively. Single chromosomal view can be shown by selecting the chromosome of interest in a dropdown list.

ondly, the gene expression file is also a tab delimited text file to record the level of mRNA expression. It can have two kinds of data formats according to its experimental method, single channel array and dual channel array. The single channel array hybridizes two samples on the separate arrays and makes two separate files that record intensity values for test and reference samples. In the other hand, the dual channel array hybridizes two samples on the same array and creates one ratio file that records ratio values of test/reference intensity values. CHESS handles these two kinds of data formats and they have same file format. The first two lines include information on data dimension of used sample number and probe number. The third line lists the sample names separated by tabs. The rest of the lines record ratio values or intensity values for entire probe in which the first three columns list probe name, the corresponding gene symbol, chromosome number and the subsequent columns record actual expression values. Finally, the gene mapping file is needed to match probes used in CGH to a corresponding gene for biological interpretation. It has a very simple format, in which the gene symbol is followed by CGH probe identification with tab separations.
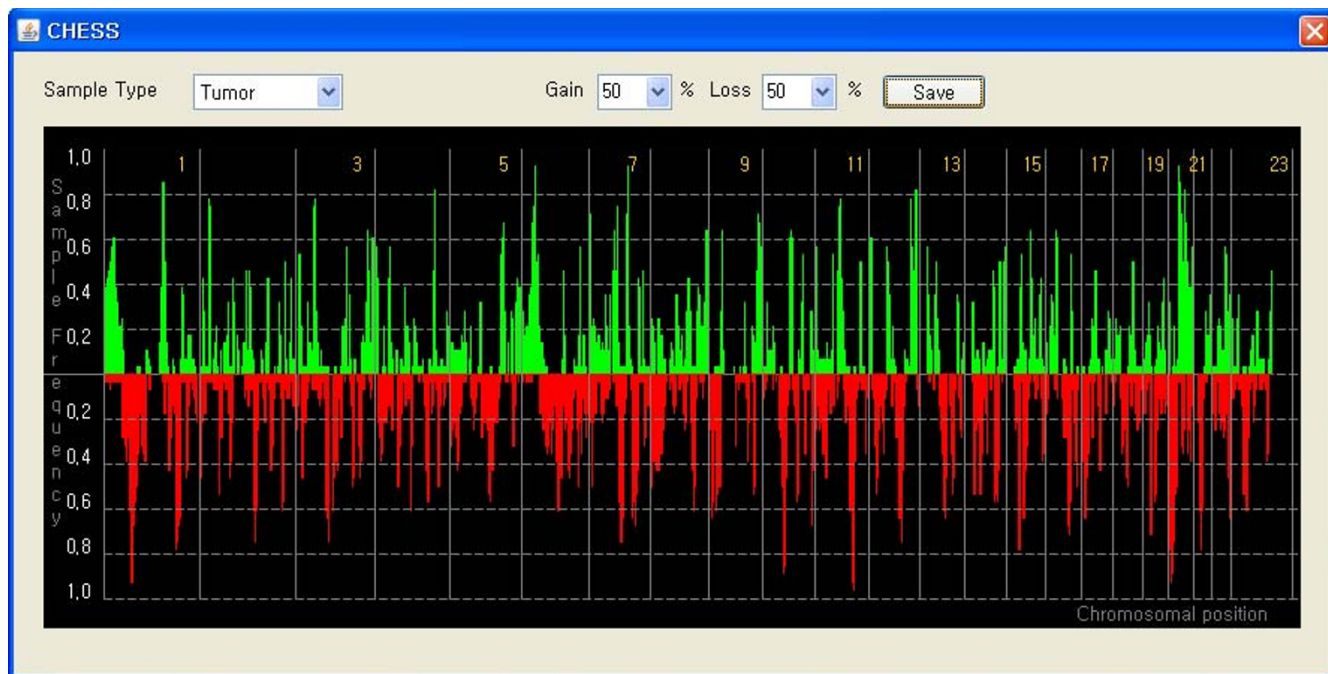
Definition of genomic alteration

CHESS defines genomic alteration regions with two distinct ways: user-specified threshold and SW-ARRAY-based algorithm. The threshold based method defines individ-

ual probe as gain or loss if the ratio value is higher or lower than user specified threshold, respectively. CHESS also has the widely used SW-ARRAY algorithm that is known to be a robust and reliable method for detection of genomic alterations [15]. It determines locally high-scoring segments whose score cannot be increased by shrinking or expanding the segment boundaries by applying Smith-Waterman dynamic-programming algorithm. After detecting genomic alterations, a GA file is automatically generated to record information regarding the alterations in the form of gain (+1), loss (-1) and no genomic alteration (0). After detection of genomic alterations is completed, CHESS shows these altered genomic regions in the form of a color bar, in which green and red indicate the regions on each chromosome that are amplified and deleted, respectively (Figure 1). A detailed view on the single chromosomal scale can be generated by selecting the chromosome of interest in a drop-down list. Sample frequency plot and heatmap drawn in a separate panel intuitively assists users in observing the overall genomic alteration pattern across whole samples intuitively (Figure 2).

Identification of phenotype specific genomic alteration regions

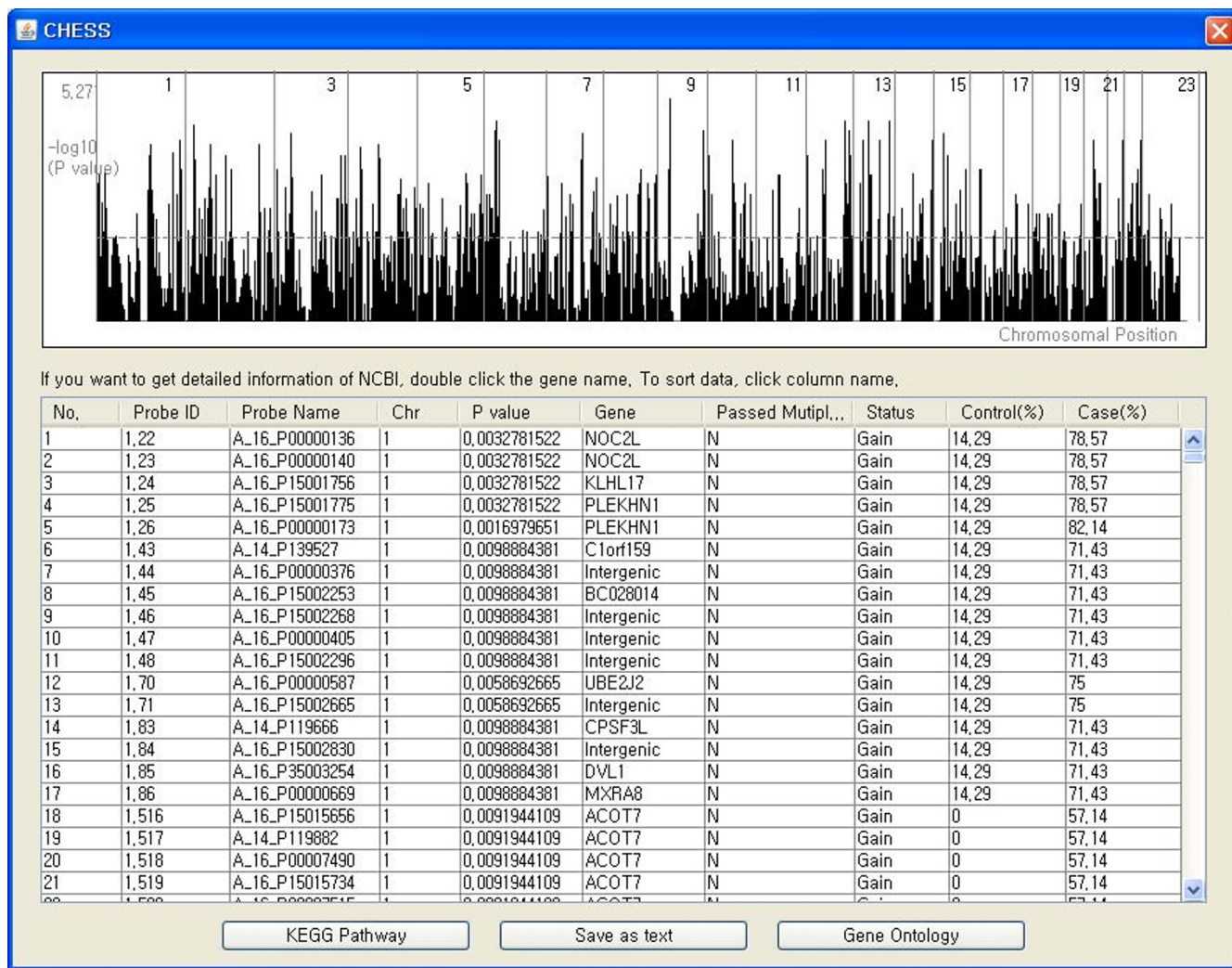CHESS identifies the phenotype or class specific genomic alteration regions based on the given clinical information. For this association study, continuous raw signal ratio val-
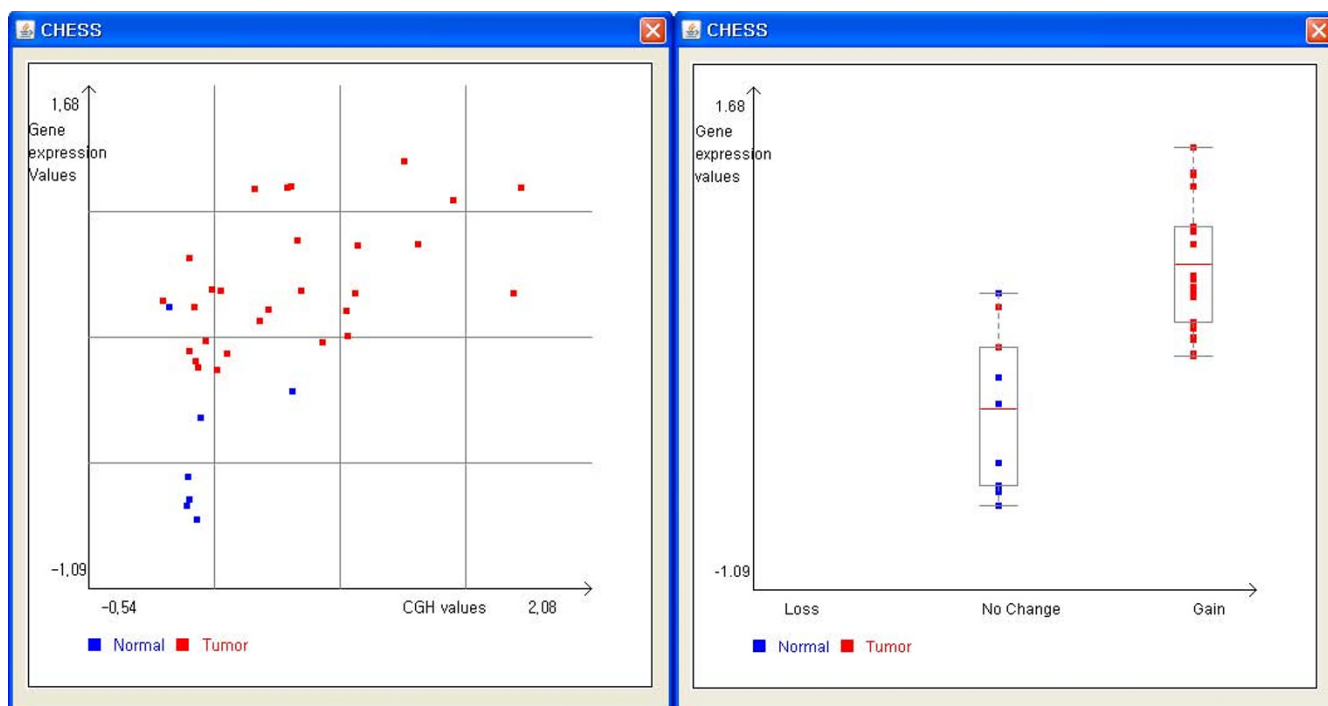

Figure 2
Sample frequency display. After detection of genomic alterations, sample frequency plot is drawn in a separate panel to help users observe the genomic alteration pattern across whole samples in a straightforward manner.

ues or discrete values of genomic alteration status are needed as essential information. When continuous signal ratio values are read, the one-sided Student's t-test can be applied to investigate the null hypothesis that the ratio values of chosen control and case types have the same distribution in their CGH ratios. If the GA file is generated and the information about genomic alteration status is available, users can carry out Fisher's exact test or Chi-square test. When sample size is relatively small, Fisher's exact test is recommended because it calculates exactly the significance rather than assuming approximated distribution like a Chi-square test. After the association study between case and control is completed, the probes determined to be significant in distinguishing between case and control types are summarized with their p values and

corresponding gene symbol in a separate panel (Figure 3). To compensate for multiple test of association study, the adjusted result by Bonferroni correction will be also presented. When the association study is finished, CHESS marks the genomic regions including the significant probes as a green/red bar on the whole chromosomal scale. A single chromosomal panel plots the log transformed p value for entire probes according to chromosomal locations. Finally for biological interpretation, CHESS gives the significance value calculated from hypergeometric distribution to represent enrichment for KEGG pathways and Gene Ontology in which phenotype specific genomic regions are involved.

If you want to get detailed information of NCBI, double click the gene name, To sort data, click column name,

No,	Probe ID	Probe Name	Chr	P value	Gene	Passed Mutipl...	Status	Control(%)	Case(%)
1	1,22	A_16_P00000136	1	0,0032781522	NOC2L	N	Gain	14,29	78,57
2	1,23	A_16_P00000140	1	0,0032781522	NOC2L	N	Gain	14,29	78,57
3	1,24	A_16_P15001756	1	0,0032781522	KLHL17	N	Gain	14,29	78,57
4	1,25	A_16_P15001775	1	0,0032781522	PLEKHN1	N	Gain	14,29	78,57
5	1,26	A_16_P00000173	1	0,0016979651	PLEKHN1	N	Gain	14,29	82,14
6	1,43	A_14_P139527	1	0,0098884381	C1orf159	N	Gain	14,29	71,43
7	1,44	A_16_P00000376	1	0,0098884381	Intergenic	N	Gain	14,29	71,43
8	1,45	A_16_P15002253	1	0,0098884381	BC028014	N	Gain	14,29	71,43
9	1,46	A_16_P15002268	1	0,0098884381	Intergenic	N	Gain	14,29	71,43
10	1,47	A_16_P00000405	1	0,0098884381	Intergenic	N	Gain	14,29	71,43
11	1,48	A_16_P15002296	1	0,0098884381	Intergenic	N	Gain	14,29	71,43
12	1,70	A_16_P00000587	1	0,0058692665	UBE2J2	N	Gain	14,29	75
13	1,71	A_16_P15002665	1	0,0058692665	Intergenic	N	Gain	14,29	75
14	1,83	A_14_P119666	1	0,0098884381	CPSF3L	N	Gain	14,29	71,43
15	1,84	A_16_P15002830	1	0,0098884381	Intergenic	N	Gain	14,29	71,43
16	1,85	A_16_P35003254	1	0,0098884381	DVL1	N	Gain	14,29	71,43
17	1,86	A_16_P00000669	1	0,0098884381	MXRA8	N	Gain	14,29	71,43
18	1,516	A_16_P15015656	1	0,0091944109	ACOT7	N	Gain	0	57,14
19	1,517	A_14_P119882	1	0,0091944109	ACOT7	N	Gain	0	57,14
20	1,518	A_16_P00007490	1	0,0091944109	ACOT7	N	Gain	0	57,14
21	1,519	A_16_P15015734	1	0,0091944109	ACOT7	N	Gain	0	57,14

KEGG Pathway   Save as text   Gene Ontology

Figure 3
Association study display. The panel on the top displays the log transformed p values as a function of chromosomal location. The table below summarizes the probes determined to be significant for discriminating between case and control types. Double clicking the interesting gene name leads user to the web page of NCBI's Entrez gene.

Figure 4
Scatter plot of CGH ratio and gene expression values. The scatter plot graphically expresses CGH ratios and gene expression values in horizontal and vertical axes, respectively. It is useful for investigating the influence of genomic alteration on gene expression level in an intuitive fashion.
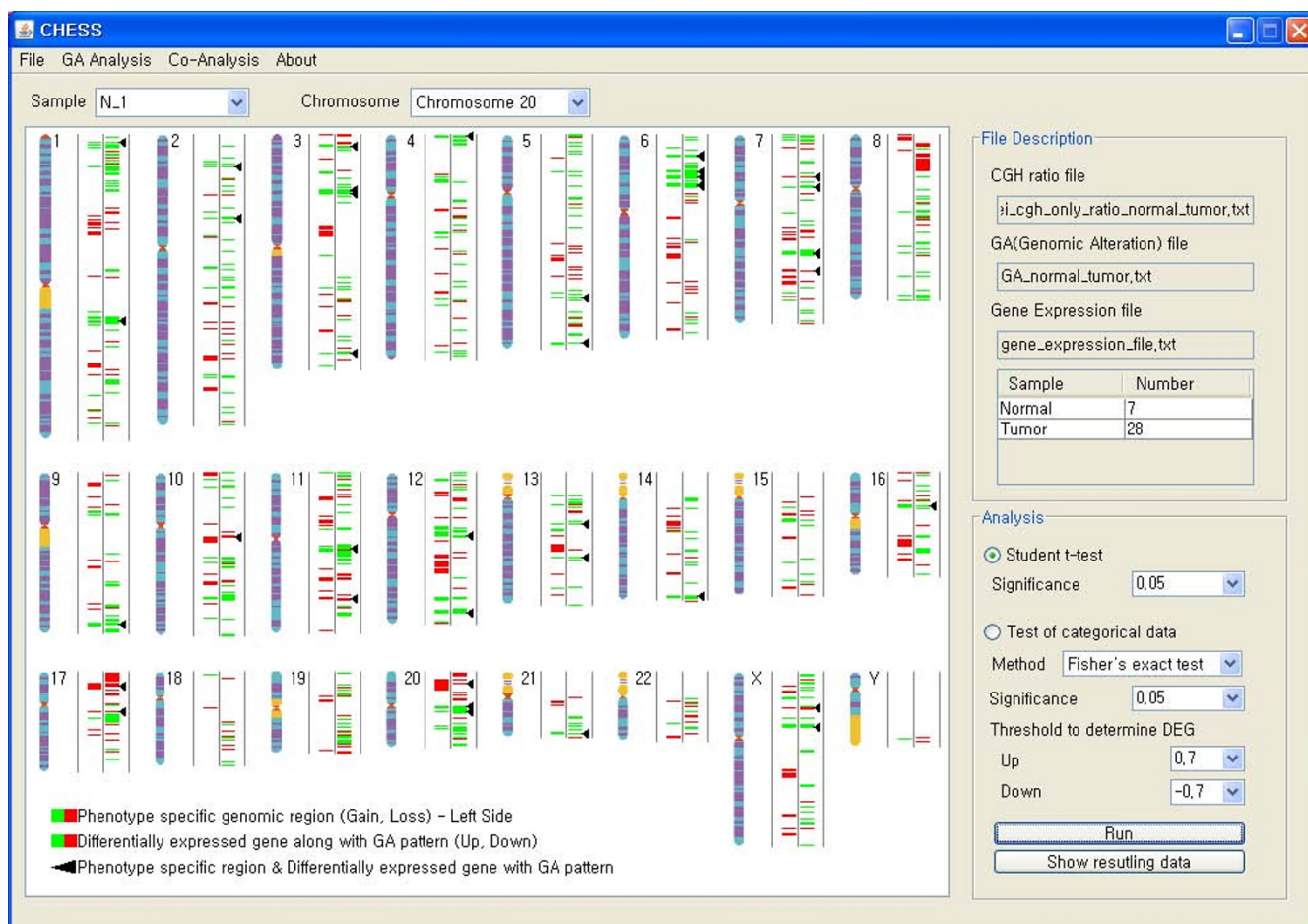
Integrative analysis of genomic alteration and gene expression data

Integrative analysis module of CHESS can be divided into two separate parts. The first part calculates the overall correlation values between CGH ratio values and gene expression level. The second part determines genes whose expression patterns are significantly different according to the status of genomic alterations. For correlation analysis, CHESS provides three statistical methods: simple linear regression, Spearman rank correlation and Pearson's correlation. These calculate correlation values for all possible pairs of CGH probes and gene expression probes mapped to the same gene. While simple linear regression calculates the r^2 with real ratio values, Spearman rank analysis calculates the ρ (Spearman's rho) using their corresponding rank values instead of real ratios. When the correlation analysis is finished, the obtained correlation values can be saved in a new text file and are indicated in the form of colour bars on both the whole chromosomal and single chromosomal scales. The length and colour of the bar corresponds to the size and the direction of correlation, respectively. The scatter plot between CGH ratio values and gene expression values is also given to users for the intuitive understanding of the general correlation pattern (Figure 4). In addition to correlation analysis, CHESS determines genes whose expression patterns are signifi-

cantly affected by their genomic alteration status through three statistical methods: One sided Student's t-test, Fisher's exact test and Chi square test. Student's t-test examines the null hypothesis that the gene expression values have the same distribution regardless of genomic alteration pattern. Fisher's exact test and Chi square test investigate whether gene expression level is affected by genomic alteration after defining the status of gene expression (up, down and no change) under the user-specified threshold. These two tests are carried out twice to test the hypothesis that the amplified genomic region leads to a significant up-regulation of mRNA and the deleted region leads to a significant down-regulation of mRNA. After this analysis, the plot to express the gene expression values and genomic alteration status in vertical and horizontal axes, respectively, is presented in a separate panel (Figure 4).

Case study of CHESS using colorectal cancer data set

To illustrate the usability of CHESS, we have used our own colorectal cancer data set. The data set consists of genomic CGH data from Agilent 244 k CGH arrays and 17 k cDNA gene expression data for 7 normal and 28 colorectal cancer patient samples. To predict the genes whose expression levels are significantly affected by genomic alteration, we have performed two kinds of statistical methods,

Figure 5
Whole chromosomal panel after performing genomic alteration and integrative analyses. In this whole chromosomal panel, genomic regions identified as colorectal cancer specific regions by Fisher's exact test (p < 0.0001) are colored green (gain) and red (loss) in the left of bar. The red and green colors in the right of bar indicate genes whose expression patterns are significantly different according to the genomic alteration patterns (Student's t-test, p < 0.01). The black arrows represent genes whose expression levels are affected by colorectal cancer specific genomic alteration regions.

Spearman rank correlation and student t-test. As the result of analysis, 565 and 300 genes were identified from Spearman rank correlation ($\rho > 0.5$) and student t-test ($P < 0.01$) respectively. We have compared gene lists from each analysis, and found only 76 genes are common in both analyses. The different result from each analysis may be due to the characteristics of each method. The correlation based methods like a Spearman rank correlation and Pearson correlation coefficient are more sensitive than the categorical data analysis methods such as student t-test, Fisher's exact test and chi-square test. However, it cannot reflect biological phenomena that genomic alteration has a discrete nature. Therefore, each analysis method has its own strength and weakness, it is very important for users to attempt various statistical methods and to interpret the biological meaning in various angles. Among 76 common genes, BCL2L1 known to be the anti-apoptotic gene has

shown a clear correlation between genomic alteration and expression pattern (Figure. 4). Additionally, figure 5 shows the analysis result after conducting two successive modules (genomic alteration analysis, integrative analysis). The red/green in the left side of a bar represents the colorectal cancer specific genomic alteration regions (loss/gain). The red/green of the right side means the significantly differentially expressed genes according to the genomic alteration pattern. The black arrows represent genes whose expression levels are affected by colorectal cancer specific genomic alteration regions.

Conclusions
We have implemented a Java-based program named CHESS for the comprehensive analysis of genomic alteration. Functionally, CHESS can be divided into two parts. The first function is responsible for detection of genomic

alteration region from the CGH data, and investigation of the relationship between detected alterations and the particular phenotype. The other function is the statistical analysis of the influence of genomic alteration on gene expression profiling. CHESS provides various optional statistical methods for these kinds of analysis, which enables users to choose the proper algorithm for their own data. Additionally, CHESS's detailed visualization module helps users understand massive data easily and intuitively. Finally, CHESS can be used as an essential tool for researchers who study genomic alteration as a molecular marker and characterize its underlying role on downstream mechanism(s) in the pathogenesis of a disease.

Availability and requirements

Project name: CgHExpreSS

Project homepage: http://biostone.khu.ac.kr/CHESS/

Operating systems: Windows and Linux

Programming language: Java

Other requirements: JRE 6 or higher (Java Runtime Environment)

License: free non-commercial research use license

Any restrictions to use by non-academics: none

Authors' contributions

ML designed and developed the software and wrote the manuscript. YK supervised the project and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

References

1. Lockwood WW, Chari R, Chi B, Lam WL: **Recent advances in array comparative genomic hybridization technologies and their applications in human genetics.** *Eur J Hum Genet* 2006, **14(2):**139-148.
2. van Beers EH, Nederlof PM: **Array-CGH and breast cancer.** *Breast Cancer Res* 2006, **8(3):**210.
3. Diep CB, Kleivi K, Ribeiro FR, Teixeira MR, Lindgjaerde OC, Lothe RA: **The order of genetic events associated with colorectal cancer progression inferred from meta-analysis of copy number changes.** *Genes Chromosomes Cancer* 2006, **45(1):**31-41.
4. Hyman E, Kauraniemi P, Hautaniemi S, Wolf M, Mousses S, Rozenblum E, Ringner M, Sauter G, Monni O, Elkahloun A, Kallioniemi O, Kallioniemi A: **Impact of DNA amplification on gene expression patterns in breast cancer.** *Cancer Res* 2002, **62(21):**6240-6245.
5. Tsafrir D, Bacolod M, Selvanayagam Z, Tsafrir I, Shia J, Zeng Z, Liu H, Krier C, Stengel RF, Barany F, Gerald WL, Paty PB, Domany E, Notterman DA: **Relationship of gene expression and chromosomal abnormalities in colorectal cancer.** *Cancer Res* 2006, **66(4):**2129-2137.
6. Kim TM, Jung YC, Rhyu MG, Jung MH, Chung YJ: **GEAR: genomic enrichment analysis of regional DNA copy number changes.** *Bioinformatics* 2008, **24(3):**420-421.
7. Shankar G, Rossi MR, McQuaid DE, Conroy JM, Gaile DG, Cowell JK, Nowak NJ, Liang P: **aCGHViewer: A Generic Visualization Tool For aCGH data.** *Cancer Inform* 2006, **2:**36-43.
8. Liva S, Hupe P, Neuvial P, Brito I, Viara E, La Rosa P, Barillot E: **CAPweb: a bioinformatics CGH array analysis platform.** *Nucleic Acids Res* 2006:W477-481.
9. van Wieringen WN, Belien JA, Vosse SJ, Achame EM, Ylstra B: **ACE-it: a tool for genome-wide integration of gene dosage and RNA expression data.** *Bioinformatics* 2006, **22(15):**1919-1920.
10. Conde L, Montaner D, Burguet-Castell J, Tarraga J, Medina I, Al-Shahrour F, Dopazo J: **ISACGH: a web-based environment for the analysis of Array CGH and gene expression which includes functional profiling.** *Nucleic Acids Res* 2007:W81-85.
11. La Rosa P, Viara E, Hupe P, Pierron G, Liva S, Neuvial P, Brito I, Lair S, Servant N, Robine N, Manie E, Brennetot C, Janoueix-Lerosey I, Raynal V, Gruel N, Rouveirol C, Stransky N, Stern MH, Delattre O, Aurias A, Radvanyi F, Barillot E: **VAMP: visualization and analysis of array-CGH, transcriptome and other molecular profiles.** *Bioinformatics* 2006, **22(17):**2066-2073.
12. Chari R, Coe BP, Wedseltoft C, Benetti M, Wilson IM, Vucic EA, MacAulay C, Ng RT, Lam WL: **SIGMA2: a system for the integrative genomic multi-dimensional analysis of cancer genomes, epigenomes, and transcriptomes.** *BMC Bioinformatics* 2008, **9:**422.
13. Fridlyand J, Snijders AM, Pinkel D, Albertson DG, Jain A: **Hidden Markov models approach to the analysis of array CGH data.** *J Multivariate Analysis* 2004, **90:**132-153.
14. Olshen AB, Venkatraman ES, Lucito R, Wigler M: **Circular binary segmentation for the analysis of array-based DNA copy number data.** *Biostatistics* 2004, **5(4):**557-572.
15. Price TS, Regan R, Mott R, Hedman A, Honey B, Daniels RJ, Smith L, Greenfield A, Tiganescu A, Buckle V, Ventress N, Ayyub H, Salhan A, Pedraza-Diaz S, Broxholme J, Ragoussis J, Higgs DR, Flint J, Knight SJ: **SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data.** *Nucleic Acids Res* 2005, **33(11):**3455-3464.