BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

DECODE: an integrated differential co-expression and differential expression analysis of gene expression data

Thomas WH Lui, Nancy BY Tsui*, Lawrence WC Chan, Cesar SC Wong, Parco MF Siu and Benjamin YM Yung*

Abstract

Background: Both differential expression (DE) and differential co-expression (DC) analyses are appreciated as useful tools in understanding gene regulation related to complex diseases. The performance of integrating DE and DC, however, remains unexplored.

Results: In this study, we proposed a novel analytical approach called DECODE (*D*iff*e*rential *Co*-expression and *D*ifferential *E*xpression) to integrate DC and DE analyses of gene expression data. DECODE allows one to study the combined features of DC and DE of each transcript between two conditions. By incorporating information of the dependency between DC and DE variables, two optimal thresholds for defining substantial change in expression and co-expression are systematically defined for each gene based on chi-square maximization. By using these thresholds, genes can be categorized into four groups with either high or low DC and DE characteristics. In this study, DECODE was applied to a large breast cancer microarray data set consisted of two thousand tumor samples. By identifying genes with high DE and high DC, we demonstrated that DECODE could improve the detection of some functional gene sets such as those related to immune system, metastasis, lipid and glucose metabolism. Further investigation on the identified genes and the associated functional pathways would provide an additional level of understanding of complex disease mechanism.

Conclusions: By complementing the recent DC and the traditional DE analyses, DECODE is a valuable methodology for investigating biological functions of genes exhibiting disease-associated DE and DC combined characteristics, which may not be easily revealed through DC or DE approach alone.
DECODE is available at the Comprehensive R Archive Network (CRAN): http://cran.r-project.org/web/packages/decode/index.html.

Background

The identification of complex gene connections and interactions that contribute to the function of living cells is one of the main challenges in functional genomics and system biology. Gene expression profiles provide rich functional information for the study of gene interrelationships. An early key approach in analyzing gene expression data was based on differential expression (DE). DE analysis has been widely used in many gene expression studies, in which the main task is to identify genes that showed different expression levels across different conditions [1-3]. The motivation is that the differentially expressed genes may have roles in the given phenotypes or conditions, and hence the studying of these genes may reveal the underlying biological mechanisms. In particular, DE analysis is a widely adopted approach that has been successfully applied in cancer research [4-6]. The analysis is useful in prioritising genes that may be dysregulated in cancer. It is popularly used in some challenging problems such as in identifying cancer-specific biomarkers for distinguishing patients and normal subjects, and in identifying potential candidate genes that response to drug treatment and environmental toxins, which will provide illuminative insight on better diagnosis and treatment of diseases at molecular level [4,5,7,8].

* Correspondence: nancytsui.cuhk@gmail.com; ben.yung@polyu.edu.hk
Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

However, DE analysis considers each gene individually and their potential interactions are ignored. Biomolecules such as genes, RNAs and proteins do not act alone; they coordinate as functional modules in biological processes and signalling pathways. They also physically aggregate to form nano-machineries such as ribosomes, chaperone and spliceosome to carry out specific functions in the cells [9]. Genes participate in same biological process tend to have similar expression pattern as demonstrated by numerous genome-wide expression studies [10-15]. Furthermore, evidence from previous studies showed that activating a metabolic pathway by small increasing expressions of many genes can be more substantial than a significant over-expression of an individual gene [16,17]. To address the gene independence model in DE analysis, approaches based on gene co-expression, gene sets, and gene clustering have been emerged. They were utilized to explore patterns of RNA expression, and hence intrinsic gene interactions [10-12,18-25].

Extending the gene co-expression concept, the analysis of differential co-expression (DC) has gained much attention in recent years [26-29]. It aims to gain insights into altered regulatory mechanisms between classes, such as disease and healthy controls, by studying their difference in gene co-expression patterns. The analysis is based on the rationale that co-regulated genes tend to share similar expression patterns. As complement to DE analysis, DC analysis is useful in identifying disease genes that may not show significant changes in expressional levels. One possible biological explanation is that given a disease gene, mutations in its coding region or post-translational modifications such as methylation, ubiquitination, and glycosylation, can impair its interactions with other gene counterparts without alternating expression level [26,30].

Evidence from previous studies showed that both DE and DC analyses are useful in identifying functionally important genes. From an informatics perspective, we questioned if relationship exists between these two types of information. Conceptually, if the two approaches extract independent information, we can simply deploy them separately and obtain distinct pieces of information (i.e. two statistically independent gene lists). On the other hand, if they extract dependent information, from a biological perspective, we seek for biological reasons such as cellular functions correspond to such dependency. Furthermore, we evaluated whether combining DE and DC criteria would improve the selection of functional relevant genes. The integrated DE and DC information may provide new opportunities for dissecting complex disease mechanism.

The benefit of integrating DC and DE approaches has been demonstrated by the study of Hudson *et al.* that compared two groups of cattle with or without a known mutation on the transcriptional regulator, namely the *myostatin* [31]. While no significant difference in *myostatin* mRNA levels was found between the two groups, *myostatin* was ranked the most important among 920 transcriptional regulators according to a scoring function that incorporates DC, DE, and expression level. After detailed examination of the scoring system of Hudson *et al.*, we concerned that the differential co-expression term was squared in the score in which the reason was unclear. Moreover, the DE genes were selected using a rather conservative statistical criterion, such that only 85 out of 11,057 genes were identified to be significant.

When integrating DE and DC approaches, one challenging problem is to define appropriate thresholds for selecting high DE genes and high DC gene pairs. Applying over-stringent thresholds may filter out many useful genes and gene pairs; whereas over-relaxing thresholds may lead to high false positivity. This problem is more apparent in DC analysis. Consider an expression data of *m* genes, the number of unique gene pairs is $m(m-1)/2$. Such huge number of gene pairs makes most multiple testing procedures powerless [32]. As a result, DC gene-pair selection methods were usually based on ad hoc criteria, such as by considering the highest *n%* of gene pairs [27] or by using pre-defined constant thresholds [33].

In this study, we have developed a novel DECODE (***D***iffer***e***ntial ***Co***-expression and ***D***ifferential ***E***xpression) analytical approach that coherently integrates DC and DE aspects. In particular, DECODE aims to improve the identification of functional gene sets or pathways that may be missed out by DC or DE criterion alone. We systematically defined DC and DE thresholds based on the dependency pattern between DE and DC variables. The functional relevance of the identified genes was also evaluated.

Methods

DECODE consists of four steps: (1) calculating differential expression (DE), (2) calculating differential co-expression (DC), (3) selecting thresholds to define high or low values of DC and DE variables based on chi-square maximization, and statistically evaluating partitions divided by the thresholds, (4) comparing functional relevance of genes categorized into the partitions of high DC, high DE, or both. Figure 1 illustrates the overview of the analytical framework. Details are described in the following sections.

Differential Expression (DE)

Consider a gene expression data set with *m* genes from samples of two states (or classes): one state consists of case (e.g. disease) group x_D, while the other consists of control (e.g. normal or healthy) group x_N. We used

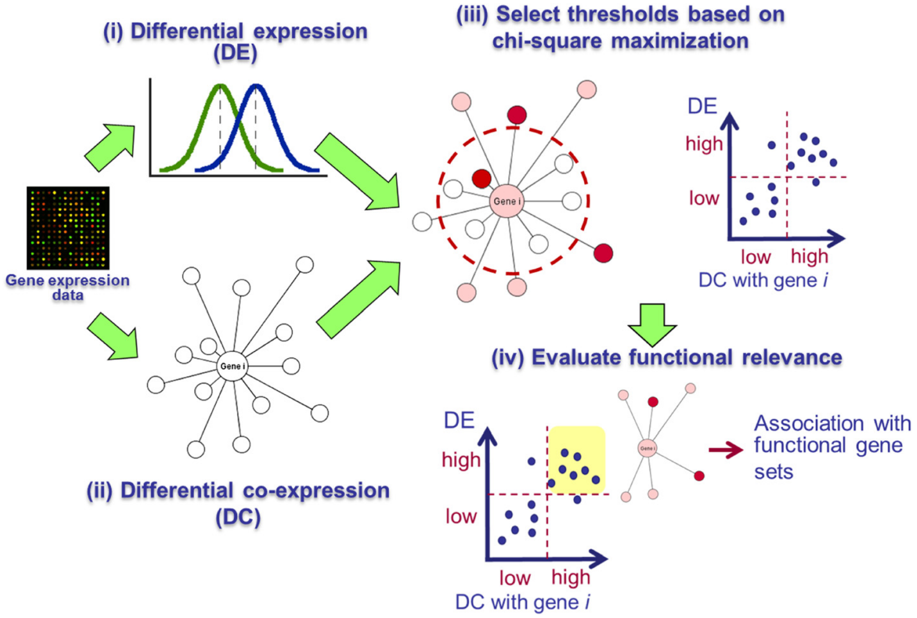
(i) Differential expression (DE)
Gene expression data
(ii) Differential co-expression (DC)
Gene i
(iii) Select thresholds based on chi-square maximization
DE
high
low
low high
DC with gene i
(iv) Evaluate functional relevance
DE
high
low
low high
DC with gene i
Association with functional gene sets

Figure 1 Overview of DECODE. (i) Calculating DE for every gene. (ii) Consider every individual gene i in turn, calculating DC between gene i with every other genes. Genes are represented by nodes. Higher DC between a gene and gene i is illustrated using longer edge. (iii) Selecting optimal thresholds to define high/low DE and high/low DC based on chi-square maximization. Genes with higher DE are illustrated by shading with deeper red colour. (iv) Evaluate functional relevance of selected gene partitions based on functional gene sets.

absolute *t*-value in *t* statistics to quantify the degree of differential expression of each gene. The *t*-value measures the difference of expression levels, in units of standard deviations, between the two states. A positive *t*-value (disease vs. normal) of a gene indicates an up-regulation in disease state; whereas a negative value indicates a down-regulation. A higher absolute *t*-value indicates a larger DE difference. The absolute *t*-value $|t_i|$ for a given gene *i*, where $i\in\{1,\ldots,m\}$, is defined as:

$$|t_i| = \frac{|\bar{x_D} - \bar{x_N}|}{\sqrt{\frac{{s_D}^2}{n_D} + \frac{{s_N}^2}{n_N}}} \tag{1}$$

where $\bar{x_D}$ and $\bar{x_N}$ are mean expression levels in disease and normal states, n_D and n_N are sample sizes of disease and normal states, and s_D and s_N are standard deviations of expression levels in disease and normal states.

Our current DECODE algorithm has been designed to handle gene expression profiles of large sample size because we have utilized ordinary *t*-statistic to measure DE. In the future, DECODE can be readily modified for expressional analysis of small dataset by incorporating the moderate *t*-statistic [34].

Differential Co-expression (DC)

We have adopted a widely used differential co-expression measure, *Z* [30,32,35-38]. The *Z* measure quantifies the correlation difference between expression levels of two genes in disease and normal samples. Consider any two genes *i* and *j* in the expression data, let r_{ij}^N and r_{ij}^D be the Pearson correlation coefficient calculated separately over the samples in normal and disease state, respectively. The measure for differential co-expression, Z_{ij}, between X_i and X_j is defined as:

$$Z_{ij} = \frac{\left|z_{ij}^N - z_{ij}^D\right|}{\sqrt{\frac{1}{n_N-3} + \frac{1}{n_D-3}}} \tag{2}$$

where n_N and n_D are sample sizes in the normal and disease states, z_{ij}^N and z_{ij}^D are the Fisher-transforms of the correlations for r_{ij}^N and r_{ij}^D, respectively, they are defined as:

$$z_{ij}^N = \frac{1}{2} ln \left| \frac{1 + r_{ij}^N}{1 - r_{ij}^N} \right| \tag{3.1}$$

$$z_{ij}^D = \frac{1}{2} ln \left| \frac{1 + r_{ij}^D}{1 - r_{ij}^D} \right| \tag{3.2}$$

After the transformation, z_{ij}^N and z_{ij}^D are both approximately normally distributed [39,40].

Novel strategy in selecting optimal DE and DC thresholds based on chi-square (χ^2) maximization

With DE and DC measures defined, we investigated the relationship between DE and DC for every gene in the expression data in turn. Given *m* genes in the expression data, there are *m* pairs of relationships between DE and DC for consideration. Specifically, consider an individual gene *i* in the data, we explored the relationship between DE of every gene and DC between gene *i* and every other genes. Figure 2 illustrated some possible relationships using scatterplots.

Next, for each gene *i*, we questioned whether genes with higher DC to gene *i* tends to (or tends not to) have higher DE. To address this, we identify two thresholds for gene *i*. One is used for defining high or low DE; another is used for defining high or low DC. We selected these two thresholds for each gene based on chi-square (χ^2) maximization. In general, a Pearson's chi-square test is used to evaluate the dependency between two variables. For our purpose, the chi-square test is also used for selecting two optimal thresholds, one for each variable, such that the strongest statistical dependencies between the DE and DC variables can be observed. Defining a variable into three or more categories or comparing chi-square measure with other discretizing measures such as entropy based measure [41] is out of the scope of current study.

The threshold selection algorithm based on chi-square maximization is described as follows. Given *m* genes in the expression data, for each gene *i*, we seeked for a pair of optimal thresholds, z_i^* and t_i^* for the DC and DE variables respectively. The pair of optimal threshold is selected from a set of threshold candidates, $\{(z_{ij}, t_j)\}$ where $j = \{1,\ldots,m\}$. Consider each pair of threshold candidates in turn, every gene *k* where $k = \{1,\ldots,m\}$ can be categorized into one of following four partitions as illustrated in Figure 3 including (1) low DC and low DE (or LDC_LDE), denoted as S_{LDC_LDE}, (2) high DC and low DE (HDC_LDE), S_{HDC_LDE}, (3) low DC and high DE (LDC_HDE), S_{LDC_HDE}, (4) high DC and high DE (HDC_HDE), S_{HDC_HDE}. They can be formally defined as:

$$S_{LDC\_LDE} = \{(z_{ik}, t_k), \text{where } z_{ik} < z_{ij} \text{ and } t_k < t_j\} \tag{4.1}$$

$$S_{LDC\_HDE} = \{(z_{ik}, t_k), \text{ where } z_{ik} < z_{ij} \text{ and } t_k \geq t_j\} \tag{4.2}$$

$$S_{HDC\_LDE} = \{(z_{ik}, t_k), \text{ where } z_{ik} \geq z_{ij} \text{ and } t_k < t_j\} \tag{4.3}$$

$$S_{HDC\_HDE} = \{(z_{ik}, t_k), \text{ where } z_{ik} \geq z_{ij} \text{ and } t_k \geq t_j\} \tag{4.4}$$

Based on these four partitions, a two by two contingency table (Table 1) can be constructed in which the number of observed genes in each partition can be counted. The observed frequency for each partition can formally be defined as:

$$obs_{A\_B} = |S_{A_B}| \tag{5}$$

where $A = \{low\ DC, high\ DC\}$, $B = \{low\ DE, high\ DE\}$.

Given the contingency table, the chi-square value, χ_k^2, for gene *k* can be computed as follows:

$$\chi_k^2 = \sum_{A=\{low\ DC,\ \ high\ DC\}} \sum_{B=\{low\ DE,\ \ high\ DE\}} \left(\frac{obs_{A\_B} - exp_{A\_B}}{exp_{A\_B}}\right)^2 \tag{6}$$

where obs_{A_B} and exp_{A_B} are the observed and expected frequency respectively. Assume the two DE and DC variables are independent, the expected frequency can be calculated using the marginal totals of the contingency table (Table 1). They can be computed as follows:

$$exp_{A\_B} = \frac{obs_A obs_B}{m} \tag{7}$$

The pair of threshold candidate, z_i^* and t_i^*, that gives maximum chi-square value is then selected as the

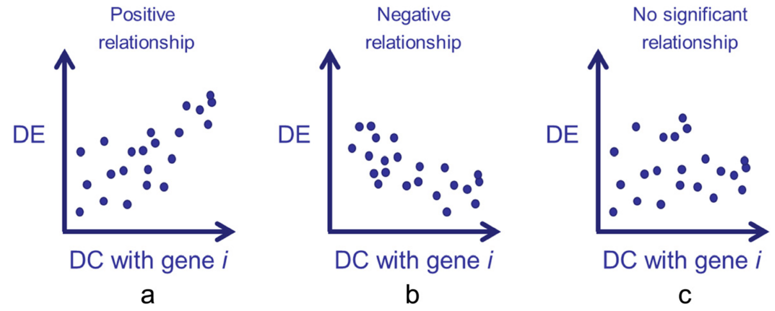

Figure 2 Some possible relationships between differential expression (DE) and the differential co-expression (DC) with gene i. Each point represents a gene. **(a)** Positive relationship **(b)** Negative relationship **(c)** No significant relationship.

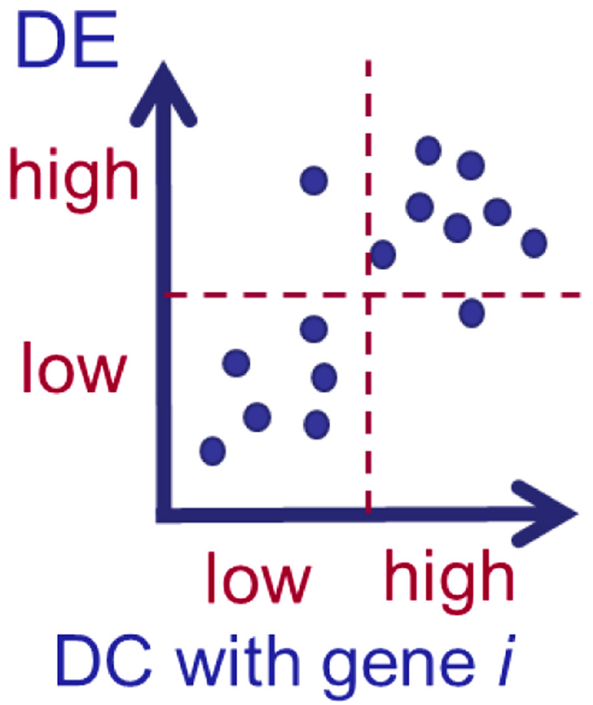

Figure 3 Genes are divided into four partitions based on the optimal DC and DE thresholds. The four divided partitions are (1) low DC and low DE (or LDC_LDE), (2) high DC and low DE (HDC_LDE), (3) low DC and high DE (LDC_HDE), and (4) high DC and high DE (HDC_HDE).

optimal threshold pair for gene *i*. For each gene *i* in the expression data, we perform the same procedure above and obtain their optimal threshold pairs. The chi-square maximization threshold selection procedure can be summarized as follows:

(1) For every gene *i*
- (1.1) For every pair of threshold candidates
  - (1.1.1) Based on current threshold candidate, all genes can be divided into 4 partitions including
    - Low DC and low DE
    - Low DC and high DE
    - High DC and low DE
    - High DC and high DE
  - (1.1.2) From the four partitions, construct a 2 × 2 contingency table to count their observed frequencies.
  - (1.1.3) Compute the chi-square value based on the contingency table.
- (1.2) Select the threshold candidate pair with maximized chi-square value as the pair of optimal thresholds for gene *i*.

Table 1 2×2 contingency table for DE and DC

	Low DC (LDC)	High DC (HDC)	Marginal total for DE
High DE (HDE)	obs_{HDE_LDC}	obs_{HDE_HDC}	obs_{HDE}
Low DE (LDE)	obs_{LDE_LDC}	obs_{LDE_HDC}	obs_{LDE}
Marginal total for DC	obs_{LDC}	obs_{HDC}	

We further evaluated the statistical significance for the association between DC and DE for every gene *i*. For every chi-square value generated in the above procedure, a corresponding *p*-value can also be obtained based on the chi-square distribution. The *p*-values have to be adjusted for multiple testing. First, for every gene *i*, since the chi-square tests are performed for *m* possible threshold candidates, there are *m* tests in total. Here, the *p*-values are adjusted using Bonferroni corrections [42]. Next, since a maximum chi-square value is used for selecting the optimal thresholds for every gene *i*, there are *m* maximum chi-square values in total for comparisons. We further corrected the adjusted *p*-values using a less stringent Benjamini and Hochberg's method [43]. In later section, we evaluated the false positive control of these adjustments using simulated data.

The chi-square test only examines the significance of the association between DC and DE variables. However, to further evaluate whether the association between high DC and high DE is significant, adjusted residual can be used [44]. If the observed number of genes (formula 5) found in high DC and high DE partition is higher than the expected frequency (formula 7), the association between high DC and high DE is regarded as positive. Conversely, if observed frequency is less than expected, the association is regarded as negative.

When the gene partitions are identified based on the optimal thresholds, they provide a flexible framework to study genes with different DC and DE characteristics. For instance, to understand the functional roles of the selected genes in a partition, gene set analysis can be performed. Furthermore, not only these gene partitions can be examined individually, studying on the combinations of partitions is also possible. Figure 4 illustrates some possible combinations. For example, by combining high DC and high DE partition (HDC_HDE) with high DC and low DE partition (HDC_LDE), the resulting partition is high DC (HDC), which can be regarded as a partition selected by using a single or individual high DC criterion.

Evaluating functional relevance

For a gene partition identified with specific DC and DE characteristics in previous step, we further examined the functional relevance of the partition genes using known functional gene sets. Pre-defined gene sets from Gene Ontology (GO) sets [45,46], Reactome pathways [47] and KEGG pathways [48,49] were used in the current

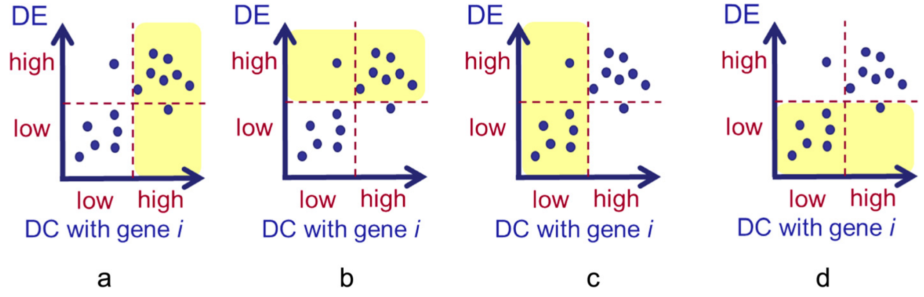

Figure 4 Some possible partitions by combining individual partitions. **(a)** High DC (HDC) only. **(b)** High DE (HDE) only. **(c)** Low DC (LDC) only. **(d)** Low DE (LDE) only.

analysis. Two-tailed Fisher's exact test [50] based on the hyper-geometric distribution is conducted to determine whether a set of partition genes is significantly over-represented in a functional gene set.

To simplify the analysis and interpretation, for each gene partition, only the best associated functional gene set was considered instead of all significant gene sets. Given a partition, the best associated gene set can be defined as the most significant gene set associated to the partition with lowest *p*-value computed from the Fisher's exact test. The *p*-values have to be adjusted for multiple testing. First, suppose the number of pre-defined functional gene sets is *k*, there are *k* tests between a gene partition and the gene sets. The *p*-values are then adjusted using Bonferroni corrections [42]. Next, consider *m'* selected gene partitions, there are *m'* best associated gene sets. The *p*-values are further adjusted using a less stringent Benjamini and Hochberg's method [43].

To facilitate the comparison of adjusted *p*-values from different partitions obtained in gene set analysis, a measure, referred to as functional information (*FI*), was used to quantify the significance of association between a gene's high DC and high DE partition S_{HDE_HDC} (formula 4.4) and a functional gene set *G*. It is defined as:

$$FI_{S_{HDE\_HDC},G} = -log_2(p) \tag{8}$$

where *p* is the adjusted *p*-value. When the significance of the association is high, *p* is small and in turn *FI* is high.

In this study, we questioned whether the functional information yielded from a partition selected based on the combining criteria of high DC and high DE is higher than that based on either of the individual high DC or high DE criterion alone. For a fair comparison, we considered the thresholds of individual criteria were the same as the optimal threshold pairs, z_i^* and t_i^*, obtained from the DC and DE criteria for each gene *i* as described in the previous section.

The gain of functional information by combining the high DC and high DE criteria over an individual criterion of DE for a given gene set *G* can be defined as:

$$\Delta'_G = FI_{S_{HDC\_HDE},G} - FI_{S_{HDE},G} \tag{9}$$

where $FI_{S_{HDE},G}$ is the functional information for the association between a high DE gene partition S_{HDE} and the function gene set *G*.

Similarly, the gain of functional information by combining the DC and DE criteria over an individual criterion of DC for a given gene set *G* can be defined as:

$$\Delta''_G = FI_{S_{HDC\_HDE},G} - FI_{S_{HDC},G} \tag{10}$$

where $FI_{S_{HDC},G}$ is the functional information for the association between a high DC gene partition S_{HDE} and the function gene set *G*.

To highlight the functional information gain by combining DC and DE criteria over individual DC or DE criteria for a given gene set *G*, the minimum of individual *FI* gains can be computed using formula 9 and formula 10, which is defined as:

$$\Delta^*_G = min\left(\Delta'_G, \Delta''_G\right) \tag{11}$$

The minimal *FI* gain is high only when both of the individual gains are high. It is low when any one of the individual gains is low. A negative gain means *FI* based on the combining criteria is lower than either one or both of the individual criteria.

Sample size estimation

The method uses three common statistical measures including *t*-statistic, differential co-expression measure based on *z*-transform of correlation coefficient, and chi-square statistics. For *t*-statistics, the sample size requirement depends on factors including alpha-level (α), power (1-β), and the anticipated effect size (Cohen's *d*) [51]. For example, consider $\alpha = 0.05$, 1-$\beta = 0.8$, and $d = 0.5$, the minimum sample size for a two-tailed *t*-test is 128. For differential co-expression measure, consider $\alpha = 0.05$, $1\text{-}\beta = 0.8$, the difference between two Fisher's z transforms is 0.5, the minimum sample size for a two-tailed t-test is 87 [52]. The chi-square test is used to categorized the genes into high/low DC and DE. In applying the test on a 2×2 table, the expected frequencies in every cells are required to be greater than 3 or 5. In these examples, the overall minimum sample size required would be 128 given the specification on the expected significant level and power.

Results and discussion

Simulation study

The proposed DECODE method provides a way to select thresholds for DC and DE variables for every gene based on chi-square maximization. Based on the maximum chi-square values, the significance of the dependencies between the DC and DE variables were evaluated. The p-values were adjusted for multiple testing as described in the method session. We performed simulation to test whether significantly high maximum chi-square values can be generated by chance even when DC and DE were independent. In addition, we evaluated whether the p-value adjustment provided good control on false positives rate.

DC and DE variables were simulated for different number of genes (m) including 10000, 15000, 20000, and 25000. For each of the m genes, we simulated m pairs of random t and Z values for the DC and DE variables respectively. The random t and Z values are simulated independently. Since DE measure, calculated based on t-statistics (formula 1), follows a t-distribution, the random t-values were generated based on t-distribution. The DC measure, calculated based on Fisher-transforms of the correlations (formula 2), are approximately normally distributed [40,53]. Here, the random Z values were generated based on normal distribution. All generated t and Z values were then converted to absolute values. Next, we performed chi-square maximization on these m pairs of DC and DE values.

The distribution of the maximum chi-square values for different m was shown in Figure 5. The average maximum chi-square values and maximum chi-square values at $\alpha = 0.05$ were shown in Table 2. Since DE and DC variables were simulated independently, any significant results were regarded as false positives. For example, consider $m = 10000$, the highest 500 chi-square values were false positives at $\alpha = 0.05$. When p-values of maximum chi-square values were not adjusted, all 10000 values were significant with confident level of 0.05 as their values greater than the corresponding tabulated value of 3.841.

To control these false positives, first adjustment was made for selecting maximum chi-square values from 10000 chi-square values using Bonferroni corrections. This resulted in 155 maximum chi-square values with adjusted p-value less than 0.05. In other words, there were 155 false positives. Second adjustment based on Benjamini and Hochberg's method was then made when comparing 10000 maximum chi-square values. This resulted in only 4 false positives. Results for other values of m were shown in Table 2. From the simulation, high maximum chi-square values could be observed because of multiple testing. We showed that the p-value adjustments could provide a stringent control on the false positive rate.

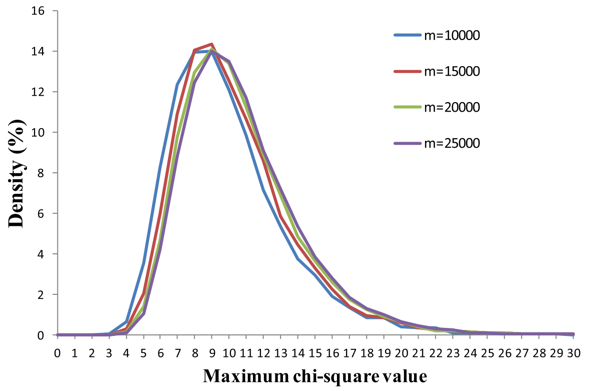

Figure 5 The distributions of the maximum chi-square values for different m number of genes.

DECODE analysis on breast cancer data

Design of experiment

We aimed to systematically determine whether the combining high DC and high DE (or HDC_HDE) criteria outperform individual criteria in selecting functional relevant genes. Specifically, after the best associated gene set was identified for each significant partition and the corresponding functional information was obtained, we evaluated whether the functional information based on the HDC_HDE criteria was higher than that based on individual HDC or HDE criteria.

Data sets

Breast cancer data of 25236 genes consisted of 1992 breast tumor samples and 144 normal samples was obtained from European genome-phenome archive [53]. From their study, the tumor data was pre-defined into two random subsets including a discovery set and a validation set. Here, we also randomly split the normal samples into 2 subsets, each with 72 samples. Consequently, we conducted DECODE analysis on two independent sets of tumor and normal samples including a *discovery vs. normal* set and a *validation vs. normal* set. The *validation vs. normal* set was used for evaluating the reproducibility of DECODE in detecting functional

Table 2 The false positive control using the p-value adjustments on simulated data

# of genes (*m*)	Average of maximum chi-square value	Observed max. chi-square at α = 0.05	# of false positive at α = 0.05	# of sig. genes after 1st adjustment	# of sig. genes after 2nd adjustment
10000	11.38	17.70	500	155	4
15000	11.79	17.97	750	215	7
20000	12.10	18.36	1000	231	13
25000	12.27	18.56	1250	280	11

gene sets. In addition, to evaluate whether the detection is an artifact, we also performed the same analysis using a *normal vs. normal* set in which both of the case and control groups are the two independent sets of 72 normal samples.

In evaluating the functional relevance of selected genes in the analysis, a total of 7114 functional gene sets were used, including 5895 sets from Gene Ontology (GO) sets (as of Jan 14, 2014) [45], 999 sets from Reactome pathways (release 37) [47] and KEGG pathways (as of July 1, 2011) [48].

Overview of the results

In analyzing the *discovery vs. normal* set of the breast cancer data, 17930 genes out of all genes in the breast cancer data have a significant and positive HDC and HDE association (adjusted p-value < 0.05). The best associated gene set was then identified for each gene partition of these positive associations. The number of unique best associated gene sets found was 99. For each unique best associated gene set, the mean minimum *FI* gains were calculated.

Comparing distribution of average functional information of HDC_HDE partitions in normal vs. normal set

Among the HDC_HDE partitions of 17930 genes selected from the *discovery vs. normal* set, we investigated the distribution of functional information of their best associated gene sets and compared it to those using individual HDC or HDE criteria. The distributions were shown in Figure 6a. From the figure, a noticeable observation is that when using the HDE criteria, a large group of 1609 partitions were obtained at a high functional information between 120 and 125. Despite of such large group, these partitions were only best associated to two functional gene set including "Cell Cycle (REACT_115566)" and "Cell Cycle, Mitotic (REACT_152)". In general, the functional information obtained using HDC criteria was lower than HDE or HDC_HDE criteria for these selected partitions.

To determine whether the functional information obtained was an artifact, we performed the same analysis to the *normal vs. normal* set. Out of all genes, only 6870 genes have a significant and positive HDC and HDE association (adjusted p-value < 0.05), compared to 17930 in the *discovery vs. normal* set. This difference was expected because the gene co-expressions were less differential when using the *normal vs. normal* set. Figure 6b showed the distribution of the function information of the best associated gene sets obtained for the selected 6870 partitions using different criteria. In comparing Figure 6a and b, the levels of functional information obtained were apparently lower when using the *normal vs. normal* set.

HDC_HDE vs. individual HDC or HDE criteria

Figure 7 showed the top 10 best associated gene sets with highest mean minimum FI gain for HDC_HDE partitions. More detail results were shown in Additional file 1: Table S1. The combined HDC_HDE criteria outperformed both of the individual criteria in six gene sets, as marked by both red and blue asterisks in Figure 7. An investigation of these gene sets provided useful insights on the mechanisms that are highly altered and highly activated (or inhibited) in breast cancer.

Among the identified gene sets, cellular response to type I interferon (GO:0071357) and regulation of T cell activation (GO:0050863) (Figure 7), are related to the immune response system. Type I interferons are key coordinators of the interactions between tumors and the immune system [54]. They regulate innate and adaptive immune responses such as the activation, migration, differentiation and survival of immune cells including macrophages, monocytes, NK cells, dendritic cells, B cells and T cells [55]. Furthermore, type I interferon response also plays an important role in preventing breast cancer spread to the bone [56].

Cell-cell junction (GO:0005911), regulation of cell adhesion (GO:0030155), and adherens junction (GO:0005912) (Figure 7) are closely related to metastasis in cancer. Metastasis is the process by which cancer spreads from the place of a primary tumor to distant locations in the body. The cell adhesion molecules play a crucial role in metastasis by promoting cell-cell interactions between tumor cells and the endothelium in distant tissues [57].

Lipid particle (GO:0005811), monocarboxylic acid metabolic process (GO:0032787), monosaccharide metabolic process (GO:0005996), and glucose metabolic process (GO:0006006) are related to lipid and glucose metabolism in breast cancer. In breast cancer, metabolisms including

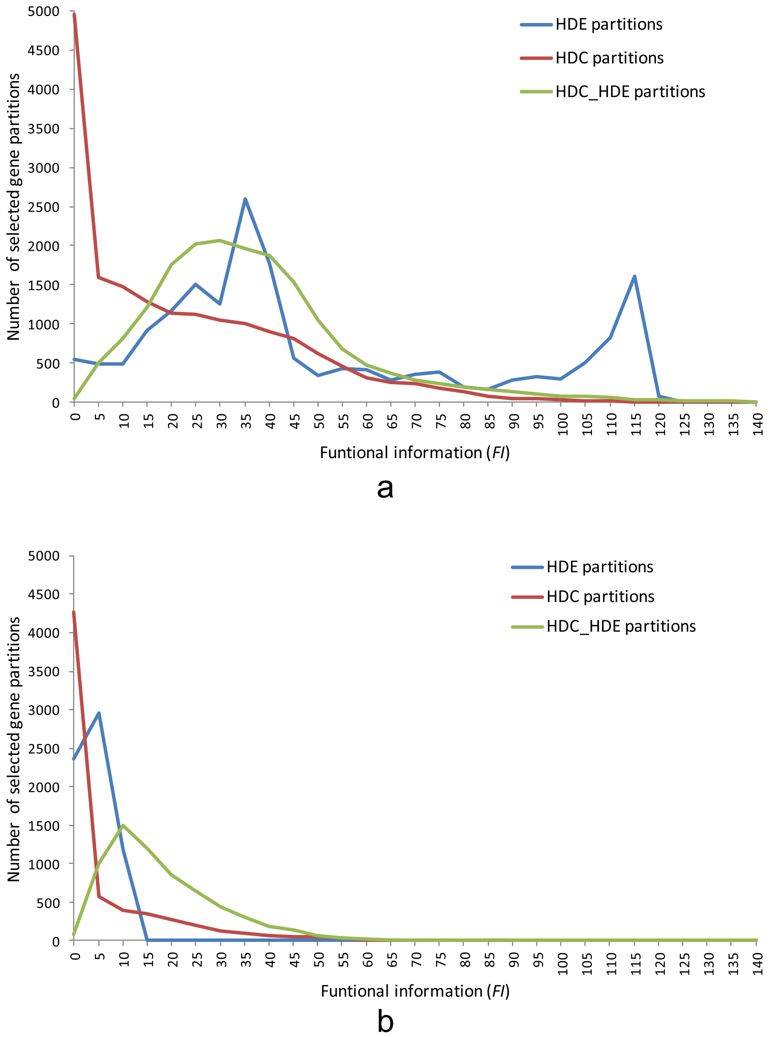

Figure 6 Distribution of function information for different gene partitions using **(a)** discovery vs. normal set, and **(b)** normal vs. normal set.

lipid and glucose metabolic processes are rewired [58,59], which happen as a result of mutations in cancer genes and alterations in cellular signalling [59]. A well-known metabolic rewiring in cancer is an increase of glucose uptake but a decrease in the proportion of glucose oxidized [60]. These rewired cancer metabolisms maintain the fitness of tumour cells for rapid proliferation and growth [59].

An increased understanding of these innate immune triggers, metastasis mechanisms, and cancer metabolisms can be important in developing new therapeutic strategies aimed at promoting immune responses against tumors, preventing metastasis, and targeting metabolisms in cancer cells. Remarkably, our proposed method was useful in detecting these functions that exhibit high DC and high DE characteristics in breast cancer. The detection on these functional gene sets based on combining criteria outperformed that based on individual high DC or high DE criteria alone.

Detecting association between Type I interferon and TRIM22

Next, to illustrate the DC and DE analysis in more detail, we selected the first ranked best association gene set for further exploration. As shown in Figure 7, the first ranked gene set was "cellular response to type I interferon". It was the best associated gene set of a total of 27 HDC_HDE partitions. Among these partitions, those of the gene *TRIM22* attained highest minimum *FI* gain of 20.5. Specifically, the gene set was associated to the HDC_HDE, HDC, and HDE partitions with the adjusted *p*-values of 2.73×10^{-18} ($FI = 58.3$ bits), 4.18×10^{-12} ($FI = 37.8$), and 1.85×10^{-2} ($FI = 5.8$) respectively. The average expression of *TRIM22* in disease state was significantly lower than that in normal state with FDR of 4.54×10^{-22}. It ranked 3166 among 18007 significant differential expressed genes (FDR <0.05). Figure 8 showed the scatterplots of DE and DC for *TRIM22*. The optimal thresholds were selected based on chi-square

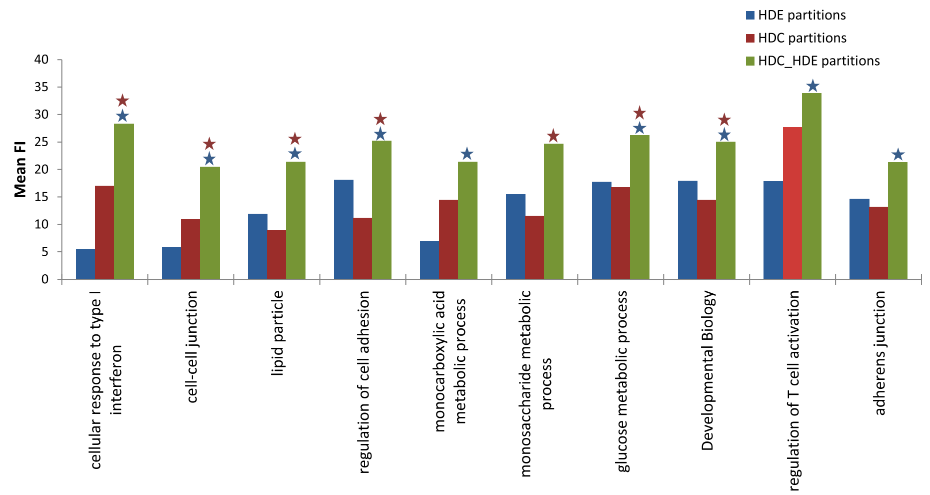

Figure 7 Top 10 best associated gene sets with highest mean minimum functional information (FI) gain, $\bar{\Delta}_G^*$, for HDC_HDE partitions in breast cancer data (discovery vs. normal set). The HDC_HDE partitions (in green) yield significantly higher mean FI than HDC partitions (in red) or HDE partitions (in blue) are marked by red or blue asterisks respectively. The combining HDC_HDE criteria outperformed both of the individual criteria in six gene sets (marked by both red and blue asterisks).

maximization. The optimal thresholds for DC and DE were 2.263 and 5.654 respectively, which were represented using red dash lines in Figure 8. Figure 8a showed a heatmap for the chi-square values for each pair of threshold candidates. The optimal point was placed in the region of the high chi-square values. The high chi-square values were more spread horizontally along the DC than vertically along the DE dimension. It may implicate a narrower range of DE for detecting high DC and DE dependency in this case. With the optimal thresholds, genes were divided into four partitions including HDC_HDE (999 genes), HDC_LDE (1090 genes), LDC_HDE (8403 genes), and LDC_LDE (14744 genes). The number of genes in HDC_HDE (999) was significantly more than the expected number (778.3) with adjusted p-value of 7.49×10^{-21}. Genes of "cellular response to type I interferon" in these four partitions were highlighted using triangle as shown in Figure 8b. The scatterplot of differential expression (DE) and correlation between genes and *TRIM22* in breast cancer state and in normal state were shown in Figure 9a and b respectively. Most selected genes in HDC_HDE partition, colour in red, were more positively correlated with *TRIM22* in the breast cancer state (Figure 9b) in compare to the normal state (Figure 9a). Twenty-three out of twenty-seven (85.2%) selected genes in the HDC_HDE partition attained a higher expression in disease state whereas the remaining four genes attain a higher expression in normal state. A network between *TRIM22* and the genes of cellular response to type I interferon was shown in Figure 10.

TRIM22, tripartite motif-containing 22, previously known as Staf50 (stimulated transacting factor 50 kDa), is a member of the tripartite motif (TRIM) subfamily of RING finger proteins. *TRIM22* underwent self-ubiquitination *in vitro* and *in vivo*, suggesting its functional role as a RING finger E3 ligase [61]. Remarkably, the identified relationship between *TRIM22* and type I interferon was coherent to previous experimental finding [62-65]. *TRIM22* was reported to be inducible by type 1 IFN *in vitro* [62,65]. The association between *TRIM22* and type 1 IFN expression *in vivo* was recently identified in HIV studies [63,64]. *TRIM22* was suggested as an antiviral effector *in vitro* and *in vivo* [63,64]. The expression of *TRIM22* was found to be negatively correlated with plasma HIV viral load but positively correlated with CD4-cell counts in primary HIV-1 infection. Silencing of *TRIM22*, in the presence of IFN-α, could increase HIV infection and virus release. These evidences supported the immune pressure of *TRIM22* against HIV-1. Moreover, *TRIM22* is a p53 target gene and contribute to viral defence by restriction of viral replication [66]. Although the promoter region of *TRIM22* is not p53-responsive, a p53-responsive motif is located in intron 1 of *TRIM22*. The over-expression of *TRIM22* can moderate the clonogenic growth of leukemic U-937 cells suggests an anti-proliferative role of leukemic cells. Since *TRIM22* is inducible by both p53 and type I IFN, it may involve in the crosstalk of p53 related pathways and interferon pathways. In short, we demonstrated that the proposed method can generate hypothesis on the relationship between a gene and its associated functional gene sets with high DC and high DE characteristics, plausibly implicated some rewired biological functions in breast cancer for follow-up investigations.

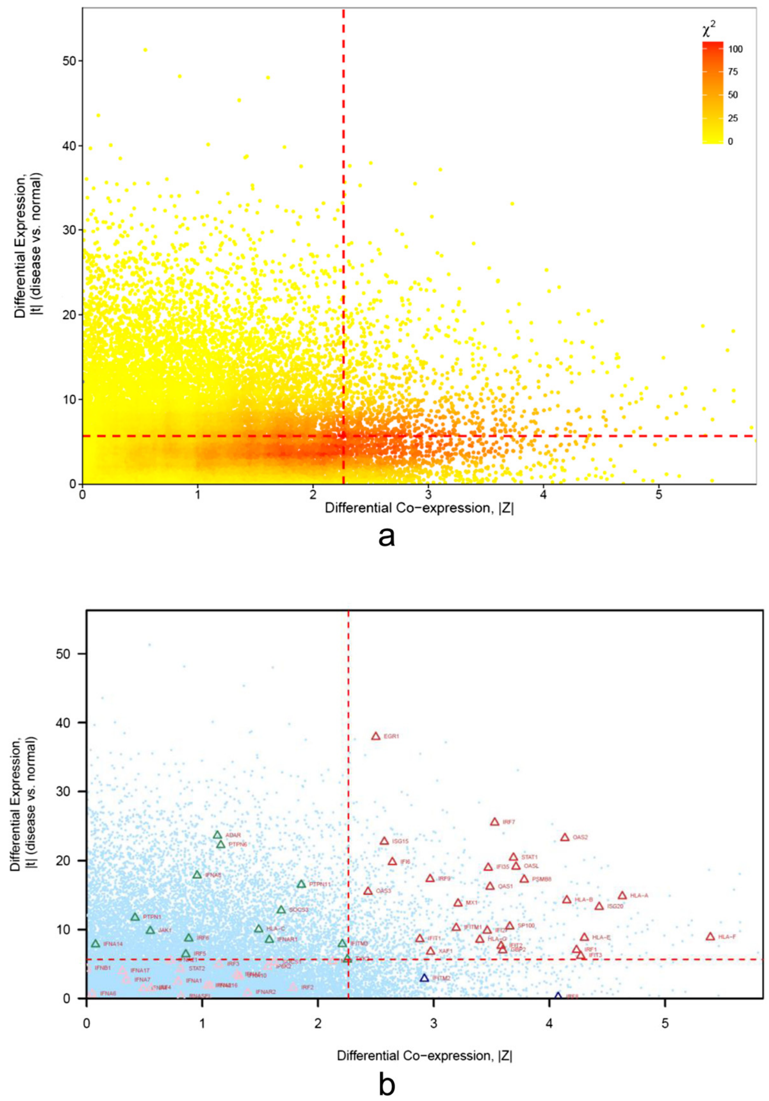

Figure 8 The scatterplot of DE and DC for TRIM22. Each point in the plot represents a gene. The x-axis represents the absolute value of DC, |Z|, between a gene and TRIM22. The y-axis represents the absolute value of DE, |t|, of a gene. The optimal thresholds for DC and DE are drawn using red dash lines. **(a)** Heat map of the chi-square values (χ^2) for the threshold candidates. **(b)** Gene set "cellular response to type I interferon (GO:0071357)" is best associated to the HDC_HDE partition (Adjusted p-value =2.73×10^{-18}). Genes found in "cellular response to type I interferon" are highlighted using triangles. These genes with different DC and DE values are highlighted in different colours: high DC and high DE (red); high DC and low DE (blue); low DC and high DE (green); low DC and low DE (pink).

Evaluation the stability of DECODE using validation set

To evaluate the reproducibility of DECODE, we analyzed the *validation vs. normal* set of the breast cancer data. Out of all genes, 19302 genes were found to have a significant DC and DE association (adjusted p-value < 0.05). The best associated gene set was then identified for each gene partition of these positive associations. The number of unique best associated gene sets was 88. Additional file 1: Figure S1 showed the top ten best associated gene sets with highest mean minimum FI gain in using high DC and high DE criteria. The detail of the results was shown in Additional file 1: Table S2 in which the top 30 results were included.

Figure 11 showed the venn diagram in comparing the number of selected gene sets from the discovery set and validation set for the high DC and high DE criteria. Considering the HDC_HDE partitions, 72 unique best associated gene sets were commonly identified, which corresponded to 72.7% of the 99 gene sets and 81.8% of the 88 gene sets identified from the *discovery vs. normal* and *validation vs. normal* data sets respectively. Overall, it showed that a substantial number of functional gene

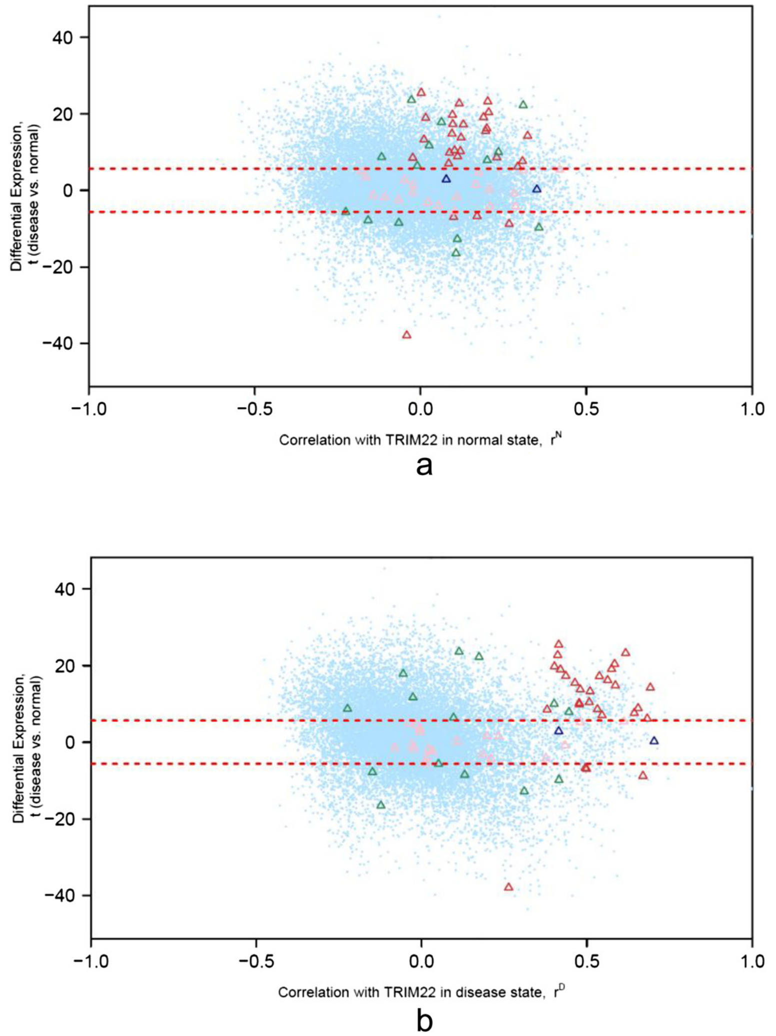

Figure 9 The scatterplot of differential expression (DE) and correlation between TRIM22 and every gene in **(a)** breast cancer state and **(b)** in normal state. Each point in the plot represents a gene. The x-axis in **(a)** and **(b)** represents the correlation coefficients, r^N and r^D, between TRIM22 and every gene in breast cancer state and normal state respectively. The y-axis represents the DE, t, of the gene. A positive t value indicates a higher gene expression in disease state in compare to normal state, and vice versa, a negative t value indicates a lower gene expression in disease state in compare to normal state. The optimal threshold for DE is drawn using red dash line. Genes found in “cellular response to type I interferon” are highlighted using triangles. Similar to Figure 8, these genes with different DC and DE values are highlighted in different colours: HDC_HDE (red); HDC_LDE (blue); LDC_HDE (green); LDC_LDE (pink).

sets identified by the method were reproducible from using two independent data sets of large samples.

DECODE analysis on Malaysian breast cancer data

In additional to the European breast cancer data obtained from European genome-phenome archive, we have performed the same analysis on an independent set of breast cancer data. The Malaysian cancer data of 13210 genes consisted of 43 tumor samples and 43 normal samples was obtained from GEO Database, accession number GSE15852 [67]. By comparing the results obtained from the European and Malaysian data, common re-wired mechanisms in breast cancer among patients from these two nationality groups may be revealed.

By using DECODE, 8780 genes out of all genes in the data have a significant and positive HDC and HDE association (adjusted p-value < 0.05). The best associated gene set was then identified for each gene partition of these positive associations. The number of unique best associated gene sets found was 54. For each unique best associated gene set, the mean minimum FI gains were calculated.

Additional file 1: Figure S2 showed the top 10 best associated gene sets with highest mean minimum FI gain

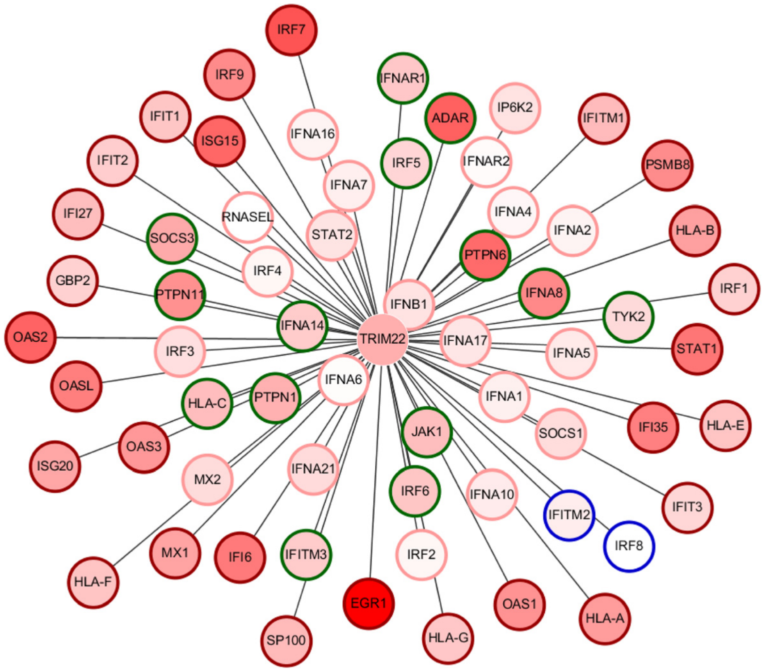

Figure 10 The network betweenTRIM22 and genes of cellular response to type I interferon. Higher DC between a gene and TRIM22 is reflected using longer edge. Genes with higher DE are shaded using deeper red colour. Genes with different DC and DE values are circled with different colours: HDC_HDE (red); HDC_LDE (blue); LDC_HDE (green); LDC_LDE (pink).

for HDC_HDE partitions. Detail of the results were shown in Additional file 1: Table S3. The combining HDC_HDE criteria outperformed both of the individual criteria in ten gene sets, as marked by both red and blue asterisks in Additional file 1: Figure S2. Among the top 10 identified gene sets, cellular lipid catabolic process (GO:0044242), monosaccharide metabolic process (GO:0005996), and glycerolipid catabolic process (GO:0046503) are related to the lipid and glucose metabolism, which is consistent to the results obtained from the European breast cancer data. The finding hence indicate the importance of the lipid and glucose metabolic processes in relation to breast cancer.

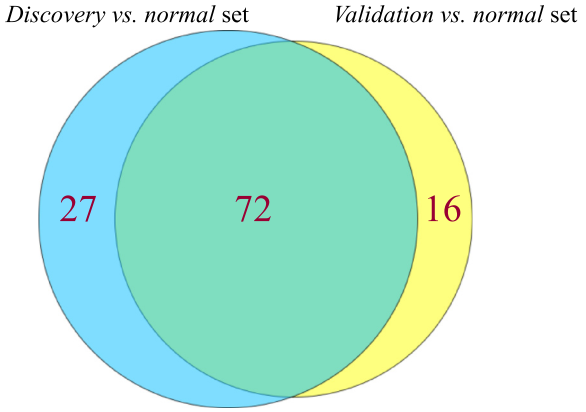

Figure 11 Venn diagrams in comparing the number of selected functional gene sets from the discovery vs. normal set and validation vs. normal set of the breast cancer data using the high DC and high DE criteria. The numbers of common functional gene sets, those found in the discovery vs. normal set, and those found in the validation vs. normal set are shown within the green, blue, yellow shaded areas respectively.

Conclusions

We presented a novel method named DECODE as a mean to integrate the DC and DE analysis. DECODE provides an analytic framework for studying different DC and DE characteristics of the genes. By incorporating dependency between DC and DE, high or low values of the DC and DE variables are systematically defined by selecting optimal thresholds that maximize the chi-square value. In using the optimal thresholds, genes can be divided into partitions with different DC and DE characteristics. The statistical significance of a gene partition can further be evaluated by residual test. Noteworthy, since the identified gene partitions at this stage are not constrained or depended on any predefined functional modules or pathways, they provide the opportunities for the discovery of novel disease related genes.

DECODE is useful for investigating whether the functional information of an identified gene partition using the combining DC and DE criteria is higher than that using individual DC or DE criteria alone. In other words, it may generate critical novel biological insights which may not be easily obtained using individual DC or DE approach. In applying DECODE to the breast cancer data, we demonstrated that it can improve the detection of some immune system, metastasis, lipid and glucose metabolism related gene sets using high DE and high DC criteria. Further investigation on the identified gene partitions and the associated functional pathways provides a more systematic understanding of complex disease mechanism, which in turn yields useful insights in the development of new therapeutic strategies for the disease. In conclusion, in complementing the DC and DE analysis, DECODE is a valuable methodology in identifying functional gene sets exhibiting certain combination of DE and DC characteristics, which serves as a new tool for future gene expression studies.

Additional file

Additional file 1: Supplementary Materials.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Designed and implemented the algorithm: TWHL. Provided critical insights into algorithm design: BYMY NBYT LWCC. Analyzed the data: TWHL. Wrote the paper: TWHL NBYT. Provided critical insights and discussion: CSCW PMS. All authors read and approved the final manuscript.

Acknowledgements
This work was supported by the Project of Strategic Importance funding of The Hong Kong Polytechnic University [1-ZE17].

Received: 13 October 2014 Accepted: 22 April 2015
Published online: 31 May 2015

References
1. Kulkarni H, Goring HHH, Diego V, Cole S, Walder KR, Collier GR, et al. Association of differential gene expression with imatinib mesylate and omacetaxine mepesuccinate toxicity in lymphoblastoid cell lines. BMC Med Genomics. 2012;5.
2. McCormick KP, Willmann MR, Meyers BC. Experimental design, preprocessing, normalization and differential expression analysis of small RNA sequencing experiments. Silence. 2011;2(1):2.
3. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3).
4. Choi CH, Choi JJ, Park YA, Lee YY, Song SY, Sung CO, et al. Identification of differentially expressed genes according to chemosensitivity in advanced ovarian serous adenocarcinomas: expression of GRIA2 predicts better survival. Br J Cancer. 2012;107(1):91–9.
5. Lucas SM, Heath EI. Current challenges in development of differentially expressed and prognostic prostate cancer biomarkers. Prostate cancer. 2012;2012:640968.
6. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7(3):562–78.
7. Cui XQ, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. Genome Biol. 2003;4(4).
8. Macgregor PF, Squire JA. Application of microarrays to the analysis of gene expression in cancer. Clin Chem. 2002;48(8):1170–7.
9. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL. Hierarchical organization of modularity in metabolic networks. Science. 2002;297(5586):1551–5.
10. Boutros PC, Okey AB. Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data. Brief Bioinform. 2005;6(4):331–43.
11. Do JH, Choi DK. Clustering approaches to identifying gene expression patterns from DNA microarray data. Mol Cells. 2008;25(2):279–88.
12. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998;95(25):14863–8.
13. Oliver S. Guilt-by-association goes global. Nature. 2000;403(6770):601–3.
14. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet. 2003;34(2):166–76.
15. Slonim DK. From patterns to pathways: gene expression data analysis comes of age. Nat Genet. 2002;32(Suppl):502–8.
16. Ben-Shaul Y, Bergman H, Soreq H. Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. Bioinformatics. 2005;21(7):1129–37.
17. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003;34(3):267–73.
18. Ackermann M, Strimmer K. A general modular framework for gene set enrichment analysis. BMC Bioinformatics. 2009;10:47.
19. Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics. 2007;23(8):980–7.
20. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57.
21. Minguez P, Dopazo J. Assessing the biological significance of gene expression signatures and co-expression modules by studying their network properties. PLoS One. 2011;6(3).
22. Nam D, Kim SY. Gene-set approach for expression pattern analysis. Brief Bioinform. 2008;9(3):189–97.
23. Ruan JH, Dean AK, Zhang WX. A general co-expression network-based approach to gene expression analysis: comparison and applications. BMC Syst Biol. 2010;4.
24. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
25. Wei W, Hou J, Alder O, Ye X, Lee S, Cullum R, et al. Genome-wide microRNA and messenger RNA profiling in rodent liver development implicates mir302b and mir20a in repressing transforming growth factor-beta signaling. Hepatology. 2013;57(6):2491–501.
26. Anglani R, Creanza TM, Liuzzi VC, Piepoli A, Panza A, Andriulli A, et al. Loss of connectivity in cancer co-expression networks. PLoS One. 2014;9(1):e87075.
27. Choi JK, Yu US, Yoo OJ, Kim S. Differential coexpression analysis using microarray data and its application to human cancer. Bioinformatics. 2005;21(24):4348–55.
28. Choi Y, Kendziorski C. Statistical methods for gene set co-expression analysis. Bioinformatics. 2009;25(21):2780–6.
29. Watson M. CoXpress: differential co-expression in gene expression data. BMC Bioinformatics. 2006;7.
30. de la Fuente A. From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. Trends Genet. 2010;26(7):326–33.
31. Hudson NJ, Reverter A, Dalrymple BP. A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. Plos Comput Biol. 2009;5(5).
32. Hu R, Qiu X, Glazko G, Klebanov L, Yakovlev A. Detecting intergene correlation changes in microarray analysis: a new approach to gene selection. BMC Bioinformatics. 2009;10:20.
33. Mentzen WI, Floris M, de la Fuente A. Dissecting the dynamics of dysregulation of cellular processes in mouse mammary gland tumor. BMC Genomics. 2009;10:601.
34. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3:Article3.
35. Bockmayr M, Klauschen F, Gyorffy B, Denkert C, Budczies J. New network topology approaches reveal differential correlation patterns in breast cancer. BMC Syst Biol. 2013;7.
36. Cho SB, Kim J, Kim JH. Identifying set-wise differential co-expression in gene expression microarray data. BMC Bioinformatics. 2009;10.

37. Mo WJ, Fu XP, Han XT, Yang GY, Zhang JG, Guo FH, et al. A stochastic model for identifying differential gene pair co-expression patterns in prostate cancer progression. BMC Genomics. 2009;10.
38. Yu H, Lin CC, Li YY, Zhao ZM. Dynamic protein interaction modules in human hepatocellular carcinoma progression. BMC Syst Biol. 2013;7.
39. Gayen AK. The frequency distribution of the product–moment correlation coefficient in random samples of any size drawn from non-normal universes. Biometrika (Biometrika Trust). 1951;38(1/2):219–47.
40. Sachs L. Applied statistics, a handbook of techniques, 2nd Edition. Percept Motor Skill. 1985;60(3):1011.
41. Ching JY, Wong AKC, Chan KCC. Class-dependent discretization for inductive learning from continuous and mixed-mode data. Ieee T Pattern Anal. 1995;17(7):641–51.
42. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze. 1936;8:3–62.
43. Benjamini Y, Hochberg Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. J Roy Stat Soc B Met. 1995;57(1):289–300.
44. Agresti A. An introduction to categorical data analysis. New York: Wiley; 1996.
45. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9.
46. Dolinski K, Botstein D. Automating the construction of gene ontologies. Nat Biotechnol. 2013;31(1):34–5.
47. Croft D, O'Kelly G, Wu GM, Haw R, Gillespie M, Matthews L, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 2011;39:D691–7.
48. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40(D1):D109–14.
49. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 1999;27(1):29–34.
50. Fisher RA. On the interpretation of χ2 from contingency tables, and the calculation of P. J Roy Statist Soc. 1922;85:87–94.
51. Suresh K, Chandrashekara S. Sample size estimation and power analysis for clinical research studies. J Hum Reprod Sci. 2012;5(1):7–13.
52. Zar JH. Biostatistical analysis. 5th ed. Englewood Cliffs: Prentice Hall; 2010. p. 393.
53. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346–52.
54. Hervas-Stubbs S, Perez-Gracia JL, Rouzaut A, Sanmamed MF, Le Bon A, Melero I. Direct effects of type I interferons on cells of the immune system. Clin Cancer Res. 2011;17(9):2619–27.
55. Fuertes MB, Woo SR, Burnett B, Fu YX, Gajewski TF. Type I interferon response and innate immune sensing of cancer. Trends Immunol. 2013;34(2):67–73.
56. Slaney CY, Moller A, Hertzog PJ, Parker BS. The role of Type I interferons in immunoregulation of breast cancer metastasis to the bone. Oncoimmunology. 2013;2(1):e22339.
57. Bendas G, Borsig L. Cancer cell adhesion and metastasis: selectins, integrins, and the inhibitory potential of heparins. Int J Cell Biol. 2012;2012:676731.
58. Baenke F, Peck B, Miess H, Schulze A. Hooked on fat: the role of lipid synthesis in cancer metabolism and tumour development. Dis Model Mech. 2013;6(6):1353–63.
59. Carracedo A, Cantley LC, Pandolfi PP. Cancer metabolism: fatty acid oxidation in the limelight. Nat Rev Cancer. 2013;13(4):227–32.
60. Koppenol WH, Bounds PL, Dang CV. Otto Warburg's contributions to current concepts of cancer metabolism. Nat Rev Cancer. 2011;11(5):325–37.
61. Duan ZJ, Gao B, Xu W, Xiong SD. Identification of TRIM22 as a RING finger E3 ubiquitin ligase. Biochem Bioph Res Co. 2008;374(3):502–6.
62. Barr SD, Smiley JR, Bushman FD. The interferon response inhibits HIV particle production by induction of TRIM22. PLoS Pathog. 2008;4(2), e1000007.
63. Singh R, Gaiha G, Werner L, McKim K, Mlisana K, Luban J, et al. Association of TRIM22 with the type 1 interferon response and viral control during primary HIV-1 infection. J Virol. 2011;85(1):208–16.
64. Singh R, Patel V, Mureithi MW, Naranbhai V, Ramsuran D, Tulsi S, et al. TRIM5alpha and TRIM22 are differentially regulated according to HIV-1 infection phase and compartment. J Virol. 2014;88(8):4291–303.
65. Tissot C, Mechti N. Molecular-cloning of a New interferon-induced factor that represses human-immunodeficiency-virus type-1 long terminal repeat expression. J Biol Chem. 1995;270(25):14891–8.
66. Obad S, Brunnstrom H, Vallon-Christersson J, Borg A, Drott K, Gullberg U. Staf50 is a novel p53 target gene conferring reduced clonogenic growth of leukemic U-937 cells. Oncogene. 2004;23(23):4050–9.
67. Pau Ni IB, Zakaria Z, Muhammad R, Abdullah N, Ibrahim N, Aina Emran N, et al. Gene expression patterns distinguish breast carcinomas from normal breast tissues: the Malaysian context. Pathol Res Pract. 2010;206(4):223–8.