BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 1 2015, pages 102–108
doi:10.1093/bioinformatics/btu589

Data and text mining

Advance Access publication September 2, 2014

Over-representation of correlation analysis (ORCA): a method for identifying associations between variable sets

Yotsawat Pomyen[1,2], Marcelo Segura[1], Timothy M. D. Ebbels[1] and Hector C. Keun[1,*]

[1]Department of Surgery and Cancer, Section of Computational and Systems Medicine, Imperial College London, Exhibition Road, London SW7 2AZ, UK and [2]Translational Research Unit, Chulabhorn Research Institute, Bangkok 10210, Thailand

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Often during the analysis of biological data, it is of importance to interpret the correlation structure that exists between variables. Such correlations may reveal patterns of co-regulation that are indicative of biochemical pathways or common mechanisms of response to a related set of treatments. However, analyses of correlations are usually conducted by either subjective interpretation of the univariate covariance matrix or by applying multivariate modeling techniques, which do not take prior biological knowledge into account. Over-representation analysis (ORA) is a simple method for objectively deciding whether a set of variables of known or suspected biological relevance, such as a gene set or pathway, is more prevalent in a set of variables of interest than we expect by chance. However, ORA is usually applied to a set of variables differentiating a single experimental variable and does not take into account correlations.

Results: Over-representation of correlation analysis (ORCA) is a novel combination of ORA and correlation analysis that provides a means to test whether more associations exist between two specific groups of variables than expected by chance. The method is exemplified by application to drug sensitivity and microRNA expression data from a panel of cancer cell lines (NCI60). ORCA highlighted a previously reported correlation between sensitivity to alkylating anticancer agents and topoisomerase inhibitors. We also used this approach to validate microRNA clusters predicted by mRNA correlations. These observations suggest that ORCA has the potential to reveal novel insights from these data, which are not readily apparent using classical ORA.

Availability and implementation: The R code of the method is available at https://github.com/ORCABioinfo/ORCAcode

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 10, 2014; revised on August 13, 2014; accepted on August 26, 2014

1 INTRODUCTION

High-throughput biological techniques such as gene expression microarray or high-throughput sequencing have become a routine for most research laboratories in life sciences. Thousands of biomolecules in different conditions, e.g. mRNAs, microRNAs, proteins or metabolites, can be measured simultaneously. In complex diseases, such as cancer, these biomolecules in a group (such as a pathway or a gene set) are often altered together. Two conventional, but contrasting, approaches for analyzing coordinated responses in molecular profile data are correlation analysis and pathway analysis. In correlation or covariance analysis (and its multivariate extensions such as principal components analysis and canonical correlation analysis), there is a focus on quantitatively estimating the explanatory power of one variable over another to infer some fundamental link between the observed biomolecules. In pathway analysis, however, the objective is to use prior knowledge about interacting or otherwise related sets of biomolecules and to test specific hypotheses about the relationship between a particular set of those biomolecules (the 'pathway') and a given experimental condition. A conventional approach for pathway analysis of high-throughput biological data is over-representation analysis (ORA). To perform ORA, firstly, the biomolecules, such as mRNA, proteins or microRNA, considered 'differentially expressed' in two or more conditions are identified. Secondly, the number of differentially expressed biomolecules in each pathway is determined. Finally, for each pathway, a probability value (*P*-value) of obtaining the number of differentially expressed biomolecules against the background list of all biomolecules measured is calculated using a hypergeometric distribution. The first implementation of ORA was by Tavazoie *et al.* (1999) where transcriptional regulatory subnetworks in yeast were identified by using mRNA microarray profiling. Most implementations of ORA are used to perform pathway analysis based on pathway information, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) or other pathway databases (Cavill *et al.*, 2011), or to infer important biological functions of two different conditions from biological categories, such as Gene Ontology (GO; Beissbarth and Speed, 2004; Zeeberg *et al.*, 2003).

We introduce a novel method called Over-Representation of Correlation Analysis (ORCA) that seeks to combine both conventional approaches for analysis of coordinated biological responses and provide a new means to test for an unexpectedly high prevalence of informative associations between two sets of variables, which could be mRNA, microRNA or proteins, that comprise pathways/groups deemed of importance *a priori*. The method first calculates the correlation coefficients between all the variables in the dataset, in which the variables can be divided into groups according to pre-defined criteria (exclusive groupings). Then, as an analogy to counting 'differentially' expressed

*To whom correspondence should be addressed.

genes in ORA (which might be below or above a certain threshold of *P*-values or other metrics), ORCA defines the number of correlation coefficients that are 'above' a certain threshold within each group. Finally, the probability of association between any two-group pair is calculated in a similar fashion to the calculation of over-representation in the conventional ORA. This probability value is calculated from the number of correlation coefficients that pass a threshold against the background number of correlation coefficients of the two-group pairs and overall correlation coefficients by using the hypergeometric test. The correlation coefficient threshold can be empirically chosen or calculated by a Shannon's entropy-based threshold selection. The method was applied to several biological datasets to demonstrate the concept and implications of ORCA in biological data analysis.

2 METHODS

To apply ORCA, the variables (such as genes, microRNA or proteins) of the dataset must be divided into sets according to relevant criteria, such as GO annotations, KEGG pathways or groups, resulting from unsupervised classification techniques such as hierarchical clustering. There should also be as many data points available for an accurate correlation coefficient as possible between two variables.

All statistical calculations and plots were made using R statistical packages version 2.12. The *P*-values were all adjusted by Benjamini–Hochberg false discovery rate (FDR) procedure (Benjamini and Hochberg, 2009).

2.1 Over-representation of correlation analysis

A schematic representation of ORCA is shown in Figure 1. Firstly, the method starts with calculation of the correlation coefficients between all n variables, yielding a symmetric correlation matrix ($n \times n$ matrix). We chose Spearman's rank correlation coefficients in this study to avoid the assumption of a Normal distribution. Secondly, variables are sorted into sets or groups according to *a priori* information, such as pathways, GO (sets 1–6 in Fig. 1). Correlation coefficients between variables in the same group are called within-group correlations, whereas the ones between different groups are called between-group correlations. Thirdly, all correlation coefficients are labelled as above or below a certain threshold previously established by empirical means or a threshold selection method (see Section 2.2). The association between sets or groups is then quantified by the number of correlation coefficients between those two groups that are above the threshold.

To determine the significance of the association between groups, we introduce ORCA. The total number of correlation coefficients in the correlation matrix of the dataset (N) in the analysis can be categorized in two ways:

(i) correlation coefficients that are above the threshold (X of Venn diagram in Fig. 1)
(ii) correlation coefficients that link members of a particular pair of groups (M of Venn diagram in Fig. 1)

For any association between two groups, these criteria define four different categories of correlation coefficients (see contingency table in Fig. 1). The number of correlation coefficients that are both above the threshold and are members of a particular pair of groups (k in contingency table in Fig. 1) is the determinant of association between the pair of groups being calculated (we used association between sets 1 and 6 as an example in Fig. 1).

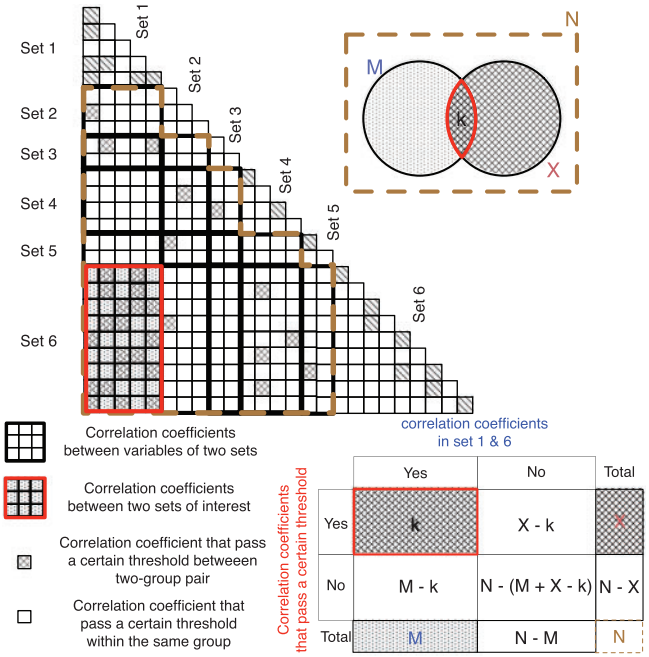

Fig. 1. The concept of over-representation of correlation. A matrix of correlation coefficients calculated from a dataset is divided into sets or groups. Correlation coefficients are represented by rectangles in each group pair. Non-filled and crosshatch-filled rectangles represent correlation coefficients that do not pass and do pass a certain threshold, respectively. The contingency table in the bottom right explains the variables required to calculate the probability value of having a certain number of correlation coefficients that pass the threshold in each group pair by chance alone via the hypergeometric distribution (Equation 1)

Finally, we calculate the *P*-value of obtaining, by chance, the number of correlation coefficients that pass the threshold and are present in each set of between-group correlations (Supplementary Fig. S1). This can be calculated by the hypergeometric distribution:

$$p(n>k) = 1 - \sum_{i=0}^{k} \frac{\binom{M}{i}\binom{N-M}{X-i}}{\binom{N}{X}} \quad (1)$$

where $\binom{t}{u} = \frac{t!}{u!(t-u)!}$ is the binomial coefficient, M is the number of between-group correlation coefficients, N is the total number of correlation coefficients in the correlation matrix (excluding within-group correlations), X is the total number of between-group correlation coefficients that pass the threshold and k is the number of correlation coefficients in the between-group set of interest that pass the threshold.

The reason for excluding within-group correlation coefficients that pass the threshold is that these tend to have high correlation coefficients and will cause an underestimation of the *P*-value calculating from between-group correlation. The *P*-values for within-group correlations were separately calculated using the same equation but including within-group correlations in N, M and X.

2.2 Threshold selection

To select a threshold of correlation coefficient that yields the most information from the data, a selection method based on Shannon's entropy (Shannon, 1948) was developed. Briefly, a score is calculated for a range

of correlation coefficient thresholds based on the *P*-values from ORCA for all group pairs using the equation (2):

$$H(X) = -\sum_{i=1}^{J} p(x_i)\log_e p(x_i) \quad (2)$$

where $H(X)$ is a Shannon entropy-like score, $P(x_i)$ is a *P*-value of a within- or between-group association calculated by hypergeometric test, J is the total number of within- and between-group pairs [if the total number of groups is n then $J = n(n-1)/2 + n$] and $\log_e$ is the natural logarithm. The correlation coefficient that yields the highest Shannon entropy-like score is selected for the ORCA threshold.

2.3 Permutation analysis

As the variables in the correlation matrix are not necessarily independent, permutation analysis was used to determine the null distribution of the *P*-values obtained by the hypergeometric test for a given data matrix. Specifically the group membership of each variable in the dataset was permuted, but the group structure (i.e. the number of groups and the number of variables within each group) together with the overall distribution of correlations was retained. For each dataset, one million permutations were performed, and the empirical *P*-values of each between and within groups were calculated by using the following equation (Davison and Hinkley, 1997):

$$P = \frac{(r+1)}{(n+1)} \quad (3)$$

where P is the empirical *P*-value, r is the number of times that the hypergeometric *P*-values from permutation test are equal or less than the actual *P*-values from ORCA and n is the total number of permutation used in the test.

3 RESULTS

3.1 ORCA reveals an association between the sensitivity of tumour cells to alkylating agents and topoisomerase inhibitors

To demonstrate the concept and potential of ORCA, we examined publicly available phenotypic profiles from the National Cancer Institute cell line panel (NCI-60 panel). This dataset meets the basic requirements of ORCA: the variables can be divided into groups and each variable has multiple data points for correlation analysis. This dataset consists of drug sensitivity data associated with the NCI-60 cell panel: profiles of 58 cancer cell lines treated with a range of drug compounds (Scherf *et al.*, 2000). The effect of a drug on a cell line was represented by the concentration of the drug that lead to 50% growth inhibition (GI_{50}). We selected a collection of well-validated results for 116 chemotherapeutic drugs and divided them into nine groups according to their mechanisms of action, i.e. the molecular targets of the drugs. Our objective was to determine whether there was similarity between different groups of drugs in terms of the sensitivity pattern across the cell lines. The dataset is detailed in Supplementary data files S1 and S2 and Supplementary Figure S2.

Despite a wide range of different mechanisms of action, some of the drugs from different classes have an apparently high degree of similarity in the observed sensitivity profile, as observed in the correlation matrix of the GI_{50} data (Fig. 2). We used ORCA to determine whether any two drug classes, defined by independent modes of action, have more common

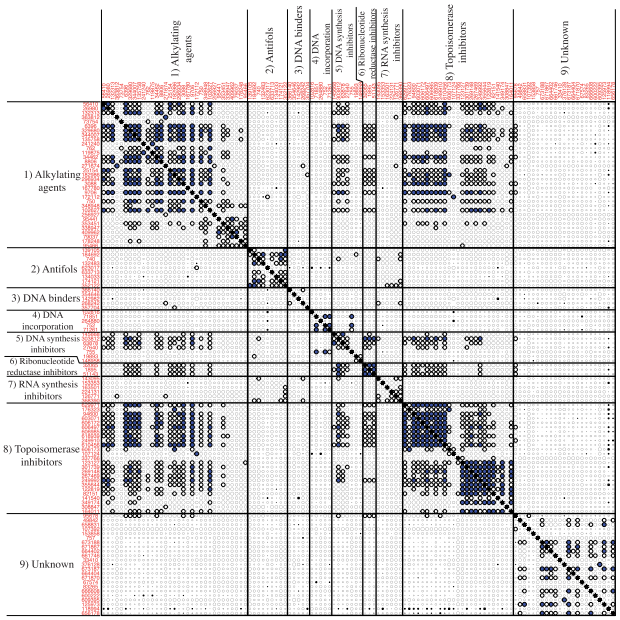

Fig. 2. Correlation matrix of drug sensitivity data from the NCI-60 cell panel. Circles and squares represent positive and negative correlation coefficients between two drugs, respectively. The sizes of circles and squares reflect the strength of correlation. The filled circles and squares are correlation coefficients that pass the threshold of 0.79. The lines divide drugs into their groups according to their mechanisms of actions. The number labels are NSC numbers (NCI's sample accession number) of the drugs. The drug names are given in Supplementary Material. Note the unusually high number of positive correlations between alkylating agents (group 1) and topoisomerase inhibitors (group 8)

correlations that are above an informative threshold than we would expect between them by chance. For this dataset, our threshold selection method identified a correlation coefficient of 0.79 to produce the highest Shannon entropy-like score, i.e. to be the most informative (Supplementary Fig. S3). Using this threshold, ORCA was then applied to generate a matrix of *P*-values for obtaining the observed number of high correlations, between every possible pair of groups of drugs (Table 1). The analysis revealed that the similarity in sensitivity profile across the NCI-60 panel between alkylating agents (group 1) and topoisomerase inhibitors (group 8) was much higher than expected by chance (q = 1e-6, Benjamini-Hochberg FDR). This finding has not previously been reported and could result in part by the fact that both sets of compounds are likely to lead to single and double-stranded DNA breaks in rapidly dividing cells, which leaves the cells reliant on a common set of DNA repair pathways for survival (Bargonetti *et al.*, 2010; Rudolf *et al.*, 2011). Another pair of drug classes exhibiting higher than expected association was DNA synthesis inhibitor (group 5) and ribonucleotide reductase inhibitor (group 6), albeit at a lesser extent than alkylating agents and topoisomerase inhibitors.

3.2 Verification of microRNA cluster significance

The second and third datasets used in this study are microRNA expression profiles of NCI-60 cell panel from Liu *et al.* (2010) and Søkilde *et al.* (2011). The datasets measured baseline

Table 1. (A) Benjamini–Hochberg FDR-adjusted (lower half) and empirical (upper half) *P*-value table for drug sensitivity dataset; (B) *P*-values from diagonal (within-group correlation coefficients)

A)

Drug Group	1	2	3	4	5	6	7	8	9	Drug Group
1		0.94	0.82	0.82	0.29	0.70	0.87	1E-6	0.99	1
2	1.00		0.59	0.59	0.66	0.58	0.60	0.92	0.91	2
3	1.00	1.00		0.58	0.68	0.42	0.64	0.78	0.79	3
4	1.00	1.00	1.00		0.11	0.42	0.64	0.78	0.79	4
5	1.00	1.00	1.00	0.33		0.01	0.62	0.24	0.86	5
6	1.00	1.00	1.00	1.00	0.01		0.47	0.64	0.62	6
7	1.00	1.00	1.00	1.00	1.00	1.00		0.81	0.81	7
8	1E-57	1.00	1.00	1.00	1.00	1.00	1.00		0.99	8
9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00		9
Drug Group	1	2	3	4	5	6	7	8	9	Drug Group

B)

Drug Group	1	2	3	4	5	6	7	8	9
Empirical *P*-value	5E-5	0.03	0.31	0.007	0.21	1E-6	0.11	1E-6	0.01
FDR-adjusted *P*-value	9E-21	0.01	0.41	0.001	0.31	0	0.56	2E-37	0.26

Notes: *P*-values in grey-shaded cells are the group pairs that have the number of correlation coefficients higher than expected that pass the threshold of 0.79 (FDR-adjusted and empirical *P*-value < 0.05).

microRNA expression of the 60 cancer cell lines. The primary objective for using these datasets in ORCA is to verify the groupings (clusters) generated in miRConnect study (Hua *et al.*, 2011). The secondary objective is to identify any interaction between these clusters.

The miRConnect study attempted to cluster microRNAs using a new correlation scheme (summed Pearson Correlation coefficient or sPCC) between baseline microRNA expression profiles and gene expression profiles from the same cell line panel. The study divided 136 microRNAs into 13 clusters according to the correlation patterns between microRNAs and gene expression of selected gene signatures. We focused on the microRNAs, which overlapped in both datasets, resulting in 124 microRNAs for this analysis.

Although the two datasets have the same structure, the data were generated using different microarray technologies, thus resulting in different microRNA profiles (see Supplementary files S3 and S4; Supplementary Figure S4 and S5). The threshold selection method determined correlation coefficient thresholds for Liu *et al.* (2010) and Søkilde *et al.* (2011) to be 0.29 and 0.61, respectively. Figure 3 shows correlation matrices of the two datasets and Tables 2 and 3 show FDR-adjusted *P*-values resulting from ORCA of the correlation matrices corresponding to the correlation matrices for Liu *et al.* (2010) and Søkilde *et al.* (2011) datasets, respectively.

At the thresholds identified previously for the two datasets, ORCA confirmed that several clusters as proposed by Hua *et al.* (2011) contain significant within-group over-representation of correlations at an FDR-adjusted significance level of 0.05 in both datasets, which are clusters I, IV, V, X and XIII.

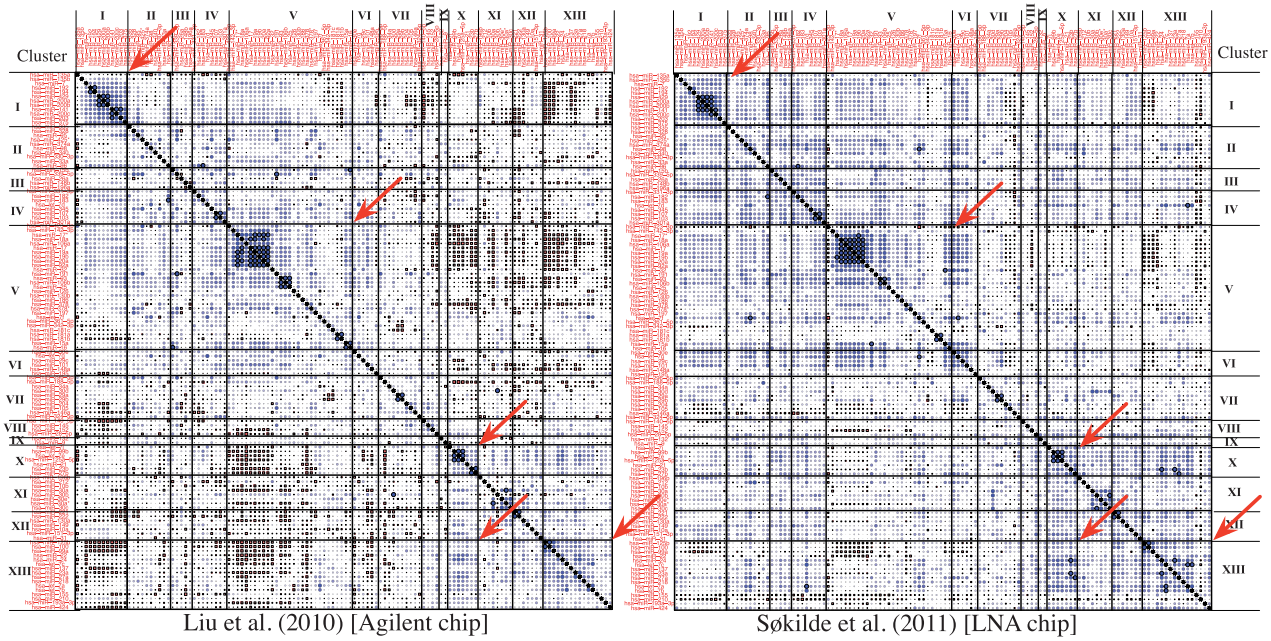

Fig. 3. Correlation matrices of two microRNA datasets characterizing the NCI-60 cell panel. Circles and squares represent positive and negative correlation coefficients between two microRNAs, respectively. The sizes of circles and squares reflect the strength of correlation. The lines divide microRNAs into their clusters according to the miRConnect study. The arrows indicate the clusters and a cluster pair that are significant according to adjusted *P*-values from ORCA in both datasets. Note that the correlation thresholds for two datasets are different

Table 2. (A) Benjamini–Hochberg FDR-adjusted (lower half) and empirical (upper half) *P*-value table for Liu *et al.* (2010) microRNA dataset; (B) *P*-values from diagonal (within-group correlations)

A)

Cluster	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	Cluster
I		0.17	0.94	0.005	0.003	0.74	0.94	0.20	0.95	0.99	0.94	0.35	0.007	I
II	0.33		0.80	**0.04**	0.02	0.71	0.54	0.75	0.67	0.86	0.99	0.94	0.99	II
III	1.00	1.00		0.18	0.06	0.55	0.61	0.20	0.68	0.99	0.48	0.90	0.88	III
IV	4E-4	**0.027**	0.44		0.01	0.35	0.77	0.85	0.87	0.88	0.96	0.64	0.99	IV
V	7E-6	8E-4	0.025	4E-4		**0.04**	0.96	0.65	0.99	0.008	0.46	0.97	0.08	V
VI	1.00	1.00	1.00	0.85	**6E-3**		0.98	0.97	0.93	0.66	0.26	0.67	0.48	VI
VII	1.00	1.00	1.00	1.00	1.00	1.00		0.26	0.82	0.91	0.99	0.54	0.99	VII
VIII	0.44	1.00	0.44	1.00	1.00	1.00	0.69		0.85	0.63	0.97	0.99	0.61	VIII
IX	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00		0.83	0.71	0.83	0.99	IX
X	1.00	1.00	1.00	1.00	1E-4	1.00	1.00	1.00	1.00		0.73	0.28	**3E-6**	X
XI	1.00	1.00	1.00	1.00	1.00	0.67	1.00	1.00	1.00	1.00		0.02	0.46	XI
XII	0.85	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.7	8E-3		0.11	XII
XIII	1E-4	1.00	1.00	1.00	0.051	1.00	1.00	1.00	1.00	**9E-14**	1.00	0.16		XIII
Cluster	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	Cluster

B)

Cluster	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII
Empirical p-value	**3E-6**	0.15	0.21	**0.006**	**1E-5**	0.45	0.57	0.05	0.24	**1E-4**	0.33	0.23	**4E-5**
FDR-adjusted p-value	**4E-15**	0.67	0.83	**4E-3**	**4E-15**	1.00	1.00	0.22	0.83	**4E-7**	1.00	0.87	**3E-11**

Notes: *P*-values in grey-shaded cells are the group pairs that have the number of correlation coefficients higher than expected that pass the threshold of 0.29 (FDR-adjusted and empirical *P*-value < 0.05). *P*-values in bold are the clusters and cluster pair that passed the threshold and overlapped with another microRNA dataset (also correspond to the arrows on left panel in Fig. 3). *P*-values in left-slanted cells correspond to the Figure 4.

Table 3. (A) Benjamini–Hochberg FDR-adjusted (lower half) and empirical *P*-value (upper half) table for Søkilde *et al.* (2011) microRNA dataset; (B) *P*-values from diagonal (within-group correlations)

A)

Cluster	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	Cluster
I		0.02	0.51	0.55	0.84	0.23	0.91	0.72	0.47	0.87	0.92	0.89	0.96	I
II	0.016		0.12	**0.003**	0.15	0.005	0.38	0.66	0.42	0.11	0.27	0.48	0.28	II
III	0.61	0.32		0.024	0.37	0.25	0.74	0.44	0.26	0.31	0.67	0.63	0.86	III
IV	0.79	**1E-3**	0.049		0.27	0.12	0.88	0.59	0.37	0.50	0.82	0.78	0.17	IV
V	0.95	0.28	0.61	0.52		**1E-6**	0.64	0.88	0.70	0.87	0.88	0.95	0.99	V
VI	0.52	1E-3	0.43	0.32	**6E-3**		0.16	0.50	0.30	0.69	0.73	0.69	0.91	VI
VII	0.95	0.61	0.83	0.92	5E-14	0.43		0.66	0.43	0.48	0.15	0.85	0.82	VII
VIII	0.83	0.79	0.61	0.75	0.85	0.68	0.79		0.22	0.55	0.59	0.23	0.79	VIII
IX	0.68	0.61	0.54	0.61	0.95	0.56	0.61	0.51		0.33	0.37	0.33	0.55	IX
X	0.92	0.26	0.51	0.61	0.86	0.79	0.68	0.72	0.59		0.50	0.74	**6E-5**	X
XI	0.93	0.55	0.79	0.86	0.95	0.83	0.32	0.75	0.61	0.61		0.78	0.71	XI
XII	0.92	0.68	0.77	0.85	0.97	0.79	0.89	0.41	0.59	0.83	0.85		0.07	XII
XIII	0.99	0.55	0.92	0.41	0.99	0.93	0.92	0.86	0.75	**6E-8**	0.86	0.13		XIII
Cluster	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	Cluster

B)

Cluster	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII
Empirical p-value	**2E-5**	0.19	0.26	**1E-4**	**3E-6**	0.006	0.38	0.17	0.03	**6E-4**	0.001	0.14	**0.003**
FDR-adjusted p-value	**2E-10**	0.57	0.61	**4E-7**	**2E-17**	0.013	0.77	0.55	0.22	**8E-5**	0.54	3E-4	**3E-11**

Notes: *P*-values in grey-shaded cells are the group pairs that have the number of correlation coefficients higher than expected that pass the threshold of 0.61 (FDR-adjusted and empirical *P*-value < 0.05). *P*-values in bold are the clusters and cluster pair that passed the threshold and overlapped with another microRNA dataset (also correspond to the arrows on right panel in Fig. 3).

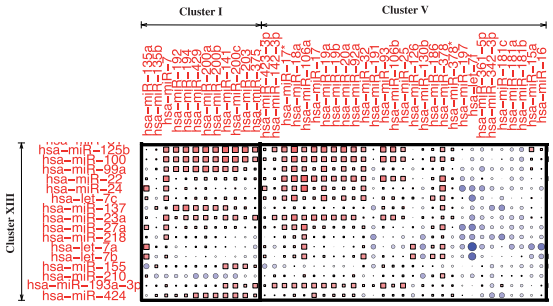

Fig. 4. Parts of the correlation matrix between the expression of selected miRNAs from the Liu et al. (2010) dataset (cluster pairs I/XIII and V/XIII). Circles and squares represent positive and negative correlation coefficients between two microRNAs, respectively. The sizes of circles and squares reflect the strength of correlation. Correlation threshold identified by threshold selection method for this dataset is 0.29. This result is consistent with the finding from Hua et al. (2013) that the microRNAs in cluster I and V are the antagonists of microRNAs in cluster XIII

Considering associations between clusters, ORCA identified several cluster pairs where between-group correlations were overrepresented, but only one cluster pair had over-representation of correlations in both datasets: cluster pair X/XIII.

The finding that clusters I, IV, V, X and XIII contain over-representation of correlation as seen by ORCA supports the hypothesis that miRNAs within these clusters are controlled by the same transcription factors, located at the same chromosomal regions or involved in the same processes or pathways.

The significant association of cluster pair X/XIII could be explained by the fact that several of the miRNAs in the two clusters possess similar seed sequences, i.e. let-7 family, mir-23 family, mir-27 family, mir-125 family and mir-99/100 family (Lewis et al., 2005). This fact was not reported in the original study and might suggest that these two clusters be best considered as a single superfamily of miRNAs.

Other potentially related clusters we identified by ORCA, i.e. cluster pairs II/IV and V/VI, were less easy to be rationalized, as clusters 2 and 6 did not exhibit significant within-group

over-representation of correlation suggest possible misclassification in the original clustering of these miRNA families.

Interestingly, a follow-up study Hua *et al*. (2013) observed that clusters I and V appeared to be functionally antagonistic to miRNAs in cluster XIII, i.e. had the opposite effect on the same set of mRNAs. ORCA was able to identify significant over-representation of (predominantly negative) correlations between clusters I and XIII and between clusters V and XIII in one of the microarray datasets (Fig. 4). This highlighted the potential of ORCA to detect such functional antagonism by analysis of miRNA co-expression alone.

4 DISCUSSION

ORCA can be used as pathway analysis tool, but the research question will be different from existing pathway or gene set analysis methods. Current methods identify pathways that are significantly enriched or depleted with respect to genes associated to a biological condition, while ORCA can identify pathways that are associated through correlations. Although the examples given here were not of typical pathways, ORCA can be applied to any type of pathway or gene set to determine pathways that are associated. In terms of pathway analysis, ORCA addresses some limitations of existing pathway analysis methods that were presented by Khatri *et al*. (2012). First, whereas ORA does not take the actual levels of variables (such as gene expression or metabolite levels) into consideration, ORCA can take these values into account, although indirectly, through the correlation coefficient calculation, which means that ORCA does not weigh the variables equally. Second, by using correlation, ORCA does not assume that the variables are independent, which is usually an important assumption in typical ORA implementations.

Third, classical ORA only uses variables, such as mRNAs or microRNAs, that are deemed most differentially expressed, while ORCA uses all the data in the calculation. Fourth, pathway analysis tools based on ORA use multiple testing corrections that assume independence of each pathway. ORCA, on the other hand, assumes the opposite and looks for the association between two sets of variables. Because our method is based on the hypergeometric test, it could be argued that ORCA violates the assumptions of the independence of each data point by using correlation coefficients as the source data. This may lead to inaccurate *P*-value calculations from the test. Goeman and Bühlmann (2007) have shown in simulated data that when hypergeometric test was used in correlated datasets, the calculated *P*-values will be underestimated. Possible remedies for the underestimated *P*-values could be multiple comparisons procedures, such as Bonferroni correction or Benjamini–Hochberg FDR correction (which was applied in this study). A table showing the nominal alpha level for correlated data from Goeman and Bühlmann (2007) can also be used to select a suitable alpha level according to a correlation threshold. A limitation of this approach is that in the simulation experiment mentioned above, all the data points have the same correlation coefficient. An alternative test that could be used instead of the hypergeometric test is the Wilcoxon rank-sum test, where the comparison is between the number of correlation coefficients in one group-pair and all other group-pairs.

ORCA can be used in pathway analysis in the same way as tools based on ORA. However, the information that will be derived from ORCA relates specifically to the pathway interactions. For pathway or gene set analysis, ORCA can calculate the correlation coefficients between genes or metabolites and then find the pathway-pairs that have more correlation coefficients that pass a certain threshold (determined by the threshold selection method or by other means) than expected.

Our version of ORCA requires group membership of variables to be mutually exclusive. Therefore, it cannot yet be used with pathway data where group members overlap, i.e. variables can not belong to more than one group. This is the subject of future work.

Recently gene co-function networks were used to identify cross-category association between different GO classes (CroGo, Peng *et al*., 2013). However, CroGO does not explicitly include gene expression values into the analysis and, therefore, could miss actual association between GO categories in the real biological context. In this regard, ORCA can be a crucial downstream analysis to CroGO to highlight associations between GO categories that are of greatest importance, using gene expression to complement the associations identified by CroGO.

In conclusion, ORCA is a new method that combines analysis of correlation with ORA, and has the potential to reveal otherwise obscured associations between sets of variables, whether they are genes, proteins, metabolites or other molecular signals, in a wide variety of biological datasets. Although the method has clear application in '-omics' data analysis, ORCA can be profitable in any circumstance where an association network can be constructed between variables that can be classified into meaningful sets.

ACKNOWLEDGEMENTS

We would like to thank Dr Rachel Cavill for the discussions during the early development of this method.

Funding: Yotsawat Pomyen is funded by Ministry of Science and Technology, Royal Thai Government.

Conflict of Interest: none declared.

REFERENCES

Bargonetti,J. *et al*. (2010) Differential toxicity of DNA adducts of mitomycin C. *J. Nucleic Acids*, **2010**, 6.

Beissbarth,T. and Speed,T.P. (2004) GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.

Benjamini,Y. and Hochberg,Y.B. (2009) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Cavill,R. *et al*. (2011) Consensus-phenotype integration of transcriptomic and metabolomic data implies a role for metabolism in the chemosensitivity of tumour cells. *PLoS Comput. Biol*., **7**, e1001113.

Davison,A.C. and Hinkley,D.V. (1997) *Bootstrap Methods and Their Application*. 9th edn. Cambridge University Press, New York, NY.

Goeman,J.J. and Bühlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.

Hua,Y. *et al*. (2013) miRConnect 2.0: identification of oncogenic, antagonistic miRNA families in three human cancers. *BMC Genomics*, **14**, 179.

Hua,Y. *et al*. (2011) miRConnect: identifying effector genes of miRNAs and miRNA families in cancer cells. *PLoS One*, **6**, e26521.

Khatri,P. *et al.* (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, **8**, e1002375.

Liu,H. *et al.* (2010) mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities. *Mol. Cancer Ther.*, **9**, 1080–1091.

Peng,J. *et al.* (2013) Identifying cross-category relations in Gene Ontology and constructing genome-specific term association networks. *BMC Bioinformatics*, **14** (Suppl. 2), S15.

Rudolf,E. *et al.* (2011) Camptothecin induces p53-dependent and -independent apoptogenic signaling in melanoma cells. *Apoptosis*, **16**, 1165–1176.

Scherf,U. *et al.* (2000) A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.*, **24**, 236–244.

Shannon,C. (1948) A mathematical theory of communination. *Bell Syst. Tech. J.*, **27**, 379.

Søkilde,R. *et al.* (2011) Global microRNA analysis of the NCI-60 cancer cell panel. *Mol. Cancer Ther.*, **10**, 375–384.

Lewis, B.P. *et al.* (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.

Tavazoie,S. *et al.* (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.

Zeeberg,B.R. *et al.* (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28.