BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Consistent Differential Expression Pattern (CDEP) on microarray to identify genes related to metastatic behavior

Lam C Tsoi[1], Tingting Qin[1], Elizabeth H Slate[2] and W Jim Zheng[3*]

Abstract

Background: To utilize the large volume of gene expression information generated from different microarray experiments, several meta-analysis techniques have been developed. Despite these efforts, there remain significant challenges to effectively increasing the statistical power and decreasing the Type I error rate while pooling the heterogeneous datasets from public resources. The objective of this study is to develop a novel meta-analysis approach, Consistent Differential Expression Pattern (CDEP), to identify genes with common differential expression patterns across different datasets.

Results: We combined False Discovery Rate (FDR) estimation and the non-parametric RankProd approach to estimate the Type I error rate in each microarray dataset of the meta-analysis. These Type I error rates from all datasets were then used to identify genes with common differential expression patterns. Our simulation study showed that CDEP achieved higher statistical power and maintained low Type I error rate when compared with two recently proposed meta-analysis approaches. We applied CDEP to analyze microarray data from different laboratories that compared transcription profiles between metastatic and primary cancer of different types. Many genes identified as differentially expressed consistently across different cancer types are in pathways related to metastatic behavior, such as ECM-receptor interaction, focal adhesion, and blood vessel development. We also identified novel genes such as *AMIGO2*, *Gem*, and *CXCL11* that have not been shown to associate with, but may play roles in, metastasis.

Conclusions: CDEP is a flexible approach that borrows information from each dataset in a meta-analysis in order to identify genes being differentially expressed consistently. We have shown that CDEP can gain higher statistical power than other existing approaches under a variety of settings considered in the simulation study, suggesting its robustness and insensitivity to data variation commonly associated with microarray experiments.
Availability: CDEP is implemented in R and freely available at: http://genomebioinfo.musc.edu/CDEP/
Contact: zhengw@musc.edu

Background

Investigating transcription profile by microarray technology has been one of the most promising genomic approaches in the last decade. Thousands of microarray experiments were performed for this purpose and their data made available through databases such as Gene Expression Omnibus, ArrayExpress and Stanford Microarray Database [1-3]. To utilize this massive amount of information, investigators have developed different meta-analysis techniques–parametric approaches such as t-statistic [4,5]; Fisher's inverse Chi-square approach [6]; Bayesian [7-9], and non-parametric approaches [10,11]. However, these approaches still face many challenges in combining data from different sources [12,13]. For example, parametric Bayesian models used in meta-analysis [7,8] are not appropriate due to the small sample size for many datasets, as suggested by Kong et al. [11]. On the other hand, non-parametric methods such as RankProd-based meta-analysis

* Correspondence: zhengw@musc.edu
[3]Division of Bioinformatics, Department of Biochemistry and Molecular Biology, Medical University of South Carolina, 135 Cannon St. Charleston, SC 29425, USA
Full list of author information is available at the end of the article

approach (Meta-RankProd) [14,15] can be significantly influenced by the size of the dataset and hence biased toward genes that are only differentially expressed in a dataset with a large number of samples–an undesirable outcome for studies where the objective is to find genes with differentially expressed patterns common across the datasets. The method of Rhodes et al. [16], which we refer to as Meta-Profile, combines a parametric and non-parametric approach and gives equal weight to each dataset when counting the number of times a gene is identified as differentially expressed in all datasets. However, this is a significant simplification because the resulting power to identify differentially expressed genes and Type I error rate (i.e. false positive rate) vary by dataset according to sample size and proportion of genes truly differentially expressed [17]. The challenges faced by these methods are particularly evident when identifying genes differentially expressed across different cancer types by pooling datasets from various sources. These datasets typically have small sample sizes [18] and the analyses are influenced by cancer-type and/or cancer-subtype specific effects [16,19,20]. In addition, some methods such as Meta-RankProd do not handle varying numbers of differentially expressed genes from different datasets–an issue that needs to be addressed for a meta-analysis approach to be robust.

The objective of this study is to develop a robust meta-analysis approach to identify genes with consistent differential expression patterns across different datasets. In our study, we combined FDR and the non-parametric RankProd approach to estimate the Type I error rate in each dataset. The estimated rates from all datasets were combined using a Bernoulli likelihood to identify genes with common expression pattern. The robustness of this approach in obtaining high statistical power was shown by simulation studies. We then applied the method to analyze different microarray data that compared gene expressions between metastatic and primary cancers and identified a core gene set that is critical to cancer metastasis across different cancer types. Our analysis identified many genes annotated in pathways that are related to metastasis, as well as novel genes that have not been shown to associate with, but may play roles in, metastasis. Further sensitivity analysis indicates that the method is robust and can be applied to other datasets for similar analyses.

Results

Consistent differential expression pattern (CDEP)

The key components of CDEP are the application of: 1) consistent FDR across datasets to identify significant genes [16,21]; and 2) non-parametric rank product (RankProd) approach to identify differentially expressed genes from microarray experiments [10]. By first using a consistent FDR to estimate the Type I error rates in each dataset, CDEP avoids overemphasizing datasets with large sample sizes–a drawback of a previous RankProd-based meta-analysis approach (Meta-RankProd) [14]. CDEP then uses the error rates from all datasets to identify genes with consistent differential expression patterns. Figure 1 shows the workflow of CDEP.

Specifically, let dataset i, $i = 1, 2, ..., D$, consist of gene expression levels for m_i and n_i samples in each of two conditions, respectively (e.g. m_i cases and n_i controls). For dataset i, the geometric mean rank of gene $g = 1, 2, ..., G$ was computed across all $m_i n_i = H_i$ pairwise comparisons for up-regulation:

$$\bar{\gamma}_{gi}^{up} = \left(\prod_{h=1}^{H_i} \gamma_{gih}{}^{up}\right)^{1/H_i} \quad (1a)$$

and down-regulation:

$$\bar{\gamma}_{gi}^{down} = \left(\prod_{h=1}^{H_i} \gamma_{gih}{}^{down}\right)^{1/H_i} \quad (1b)$$

where γ_{gih} is the rank of fold change for gene g in the h^{th} comparison of dataset i, $h = 1, 2, ..., H_i$. Genes with the smallest RankProd values ($\bar{\gamma}_{gi}$) are more likely to be the differentially expressed genes.

We then computed the RankProd p-values and FDRs for up- and down-regulations for each gene in every dataset [10]. Briefly, we used permutation of the sample labels (e.g. case/control) to estimate false positives and

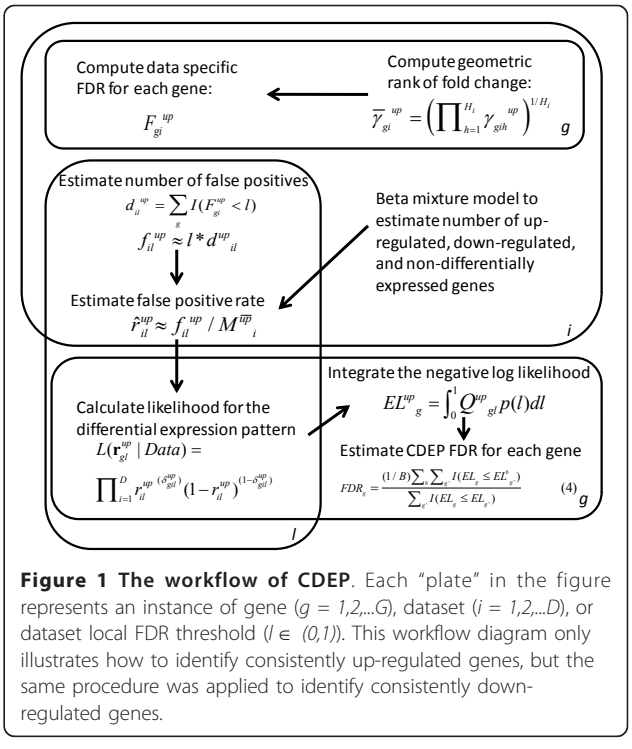

Figure 1 The workflow of CDEP. Each "plate" in the figure represents an instance of gene ($g = 1,2,...G$), dataset ($i = 1,2,...D$), or dataset local FDR threshold ($l \in (0,1)$). This workflow diagram only illustrates how to identify consistently up-regulated genes, but the same procedure was applied to identify consistently down-regulated genes.

the p-value by counting the number of times we observed the permutations' RankProd values smaller or equal to the experiment's RankProd value. The FDR of a gene was then estimated by dividing the p-value by the rank of the RankProd value [10]. Each gene in every dataset was thus associated with an FDR, $F_{gi}{}^{up}(F_{gi}{}^{down})$, for being up(down)-regulated. For genes not present in the platform of a dataset, the median FDR value computed for that dataset was assigned.

This computation in CDEP was performed using the Bioconductor [22] package RankProd [15], as Hong and Breitling [14] indicated that RankProd is more reliable than other existing approaches (see also Additional File 1, Figure S1). The FDR threshold (l) is defined as the proportion of false positives among the genes declared to be positives for each dataset. Given an FDR threshold (l), we counted the number of genes identified to be

up-regulated:

$$d_{il}{}^{up} = \sum_{g} I(F_{gi}^{up} < l) \tag{2a}$$

and down-regulated:

$$d_{il}{}^{down} = \sum_{g} I(F_{gi}^{down} < l) \tag{2b}$$

Therefore, the number of false positives in a dataset could be estimated as $f_{il}{}^{up} \approx l * d^{up}{}_{il}(f_{il}{}^{down} \approx l * d^{down}{}_{il})$. To estimate the proportion of genes that are up-regulated, non-differentially expressed, and down-regulated, we used a Beta mixture to model the genes' p-values for over (under)-expression in each dataset (see details of the Beta mixture model in Methods). We adopted the Beta mixture model because the p-values calculated from our non-parametric approach do not have a uniform distribution for non-differentially expressed genes (Figure 2), in contrast to a previous mixture model based on this assumption [23]. The Beta mixture model and the estimation of the proportion of differentially expressed genes used the Markov Chain Monte Carlo (MCMC) technique implemented in the BUGS program [24]. Our implementation uses WinBUGS [25] on the Windows platform, but OpenBUGS [26] can be used on Linux or Mac platforms (with Wine).

For each dataset, the false positive rate is defined as the probability of a non-up-regulated (non-down-regulated) gene being falsely called as over-expressed (under-expressed):
$(\hat{r}_{il}^{down} \approx f_{il}{}^{down}/M^{\overline{down}}{}_{i})(\hat{r}_{il}^{down} \approx f_{il}{}^{down}/M^{\overline{down}}{}_{i})$, where $M^{\overline{up}}{}_{i}$ ($M^{\overline{down}}{}_{i}$) are the number of genes that are not up (down)-regulated in dataset i respectively, estimated by the Beta mixture model. Based on this rate and using independent Bernoulli distributions, we calculated the likelihood of a gene to be falsely identified as over or under-expressed among the datasets for each FDR threshold l, that is, the likelihood for false positives among the significant genes identified as

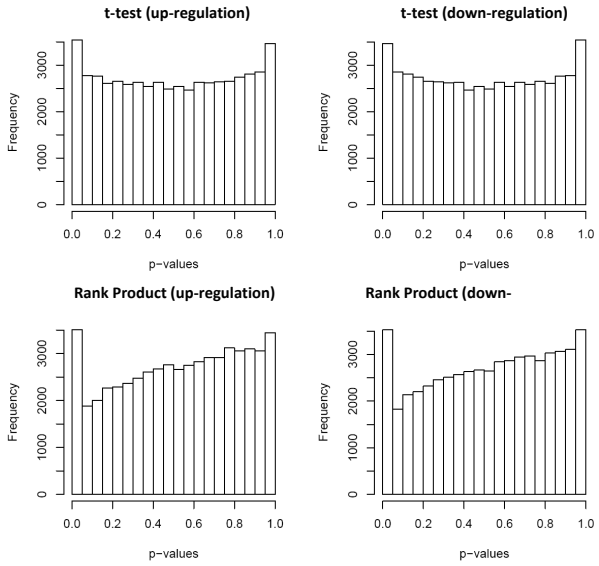

Figure 2 The distributions of p-values computed by t-test and RankProd test. We compared the parametric t-test and the non-parametric rank product approaches to test which of these p-value (computed from one-sided test) calculation is robust for identifying truly differentially expressed genes (p = q = 0.05; |Δ| = 0.5).

up-regulated:

$$L(\mathbf{r}_{gl}^{up}|Data) = \prod_{i=1}^{D} r_{il}^{up\ (\delta_{gil}^{up})}(1 - r_{il}^{up})^{(1-\delta_{gil}^{up})} \tag{3a}$$

and down-regulated:

$$L(\mathbf{r}_{gl}^{down}|Data) = \prod_{i=1}^{D} r_{il}^{down\ (\delta_{gil}^{down})}(1 - r_{il}^{down})^{(1-\delta_{gil}^{down})} \tag{3b}$$

where the binary variable δ_{gil}^{up} (or δ_{gil}^{down}) indicates whether gene g is identified to be up(or down)-regulated in dataset i for threshold l. To prevent underflow during computation, we worked with the minus log likelihood, Q, e.g. for up-regulation $Q^{up}{}_{gl} = -\ln[L(\mathbf{r}_{gl}^{up}|Data)]$. We took into consideration of multiple FDR thresholds l by specifying a probability density function (PDF) for l, $p(l)$, $l \in (0, 1)$ and using the expected value of $Q(l)$ to assess whether the gene is consistently over-expressed among the datasets. In this assessment, low l values were emphasized because low FDR represents a higher proportion of true positives, and we used the linear decreasing function: $p(l) = -2l + 2$. The expected log likelihood across the FDR threshold is: $EL^{up}{}_{g} = \int_0^1 Q_{gl}^{up} p(l) dl$, which was approximated by discretizing the range of FDR value (l) into one hundred bins

with equal width and using the rectangular rule. The same procedure was also performed for down-regulation. The procedure was evaluated by estimating the false discovery rate (FDR_g) of observing the above expected log likelihood. Here, the FDR_g is the proportion of false positives among the genes identified to be consistently differentially expressed. The "null log likelihood" was computed by permuting the $F_{gi}{}^{up}$ and $F_{gi}{}^{down}$ values relative to the genes within each dataset and performing the same above procedures to calculate the expected value of the "null log likelihood" in each permutation b for every gene ($EL^b{}_g$). In b permutations, the FDR_g of a gene could be determined as:

$$FDR_g = \frac{(1/B)\sum_b \sum_{g'} I(EL_g \le EL^b{}_{g'})}{\sum_{g'} I(EL_g \le EL_{g'})} \quad (4)$$

The robustness of CDEP in distinguishing different gene expression patterns is shown in Figure 3 where the minus log likelihood value Q was plotted against l. Genes that are not differentially expressed in all datasets (G_N) have the lowest Q values, while genes that are differentially expressed only in some datasets (G_C) have higher Q values, and genes that are differentially expressed in all the datasets (G_M) have the highest Q values. For G_N, the Q values increase slightly with l. This is because when l increases, the likelihood value decreases as more G_N are falsely called differentially expressed. Moreover, even at high l many G_N are not declared differentially expressed. On the other hand, the Q values of both G_C and G_M decrease when l increases. As l increases, r and the likelihood (L) increase, giving rise to a decreasing Q. Note that the Q values for all 3 types of genes go to zero when l = 1. This is because, in this situation, all genes in the array would be declared as differentially expressed and both r and L have values of one. Figure 4 shows the expected minus log likelihood (EL) for the three types of genes, indicating CDEP is robust in identifying genes that show common differential expression pattern across different datasets. These genes have higher EL values than the other two types of genes.

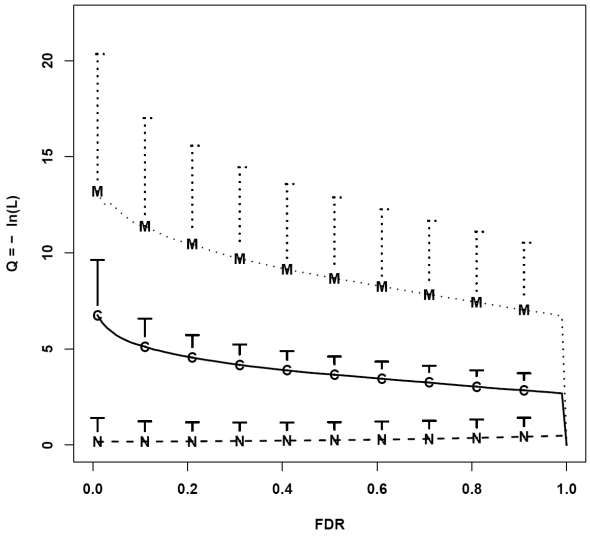

Figure 3 Log likelihood and FDR plot. Minus log likelihood versus the FDR threshold (l) for different genes in one of the simulated data (proportion of cancer-type specific and metastatic related differentially expressed genes: $p = q = 0.1$; degree of effect: $|\Delta| = 1$). FDR is the proportion of false positives among genes declared to be differentially expressed for each dataset. The dotted line represents genes that are consistently differentially expressed, solid line represents genes that are differentialy expressed only in specific dataset, and dashed line represents non-differentially expressed genes. The three lines show the mean, and the vertical bars show the standard deviation of the Q values for the three types of genes at the given FDR. For clarity, only the upper bars are shown.

Comparisons with other approaches

We compared CDEP with Meta-Profile and Meta-RankProd in a simulation study. Briefly, Meta-Profile is based on the number of times a gene is declared differentially expressed among the datasets, and Meta-RankProd uses the rank product among all datasets. Both approaches use permutation to estimate the false discovery rate for the genes as being differentially expressed consistently among the datasets. Simulation scenarios were determined by three key parameters: the proportion of

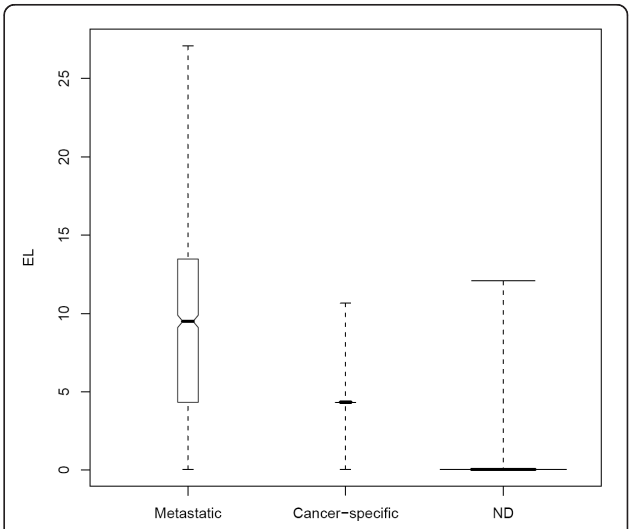

Figure 4 Boxplots for the expected likelihood. The boxplots for the expected likelihood (EL) of the three categories of genes: genes that are consistently differentially expressed, genes differentially expressed in only a certain dataset, non-differentially expressed genes. The ranges and the quartiles are shown. The width of the boxplot is drawn proportional to the square-root of the number of observations. p = q = 0.1, $|\Delta| = 1$.

differentially expressed genes that are dataset-specific (p), the proportion of genes that are consistently differently expressed (q), and the mean difference between the expression values in the two conditions being compared (Δ) (See "Simulation of Microarray data" in Methods and Simulation Section in Additional File 1 for details). Table 1 reports the statistical power and Type I error rate of the three meta-analysis approaches, where statistical power is defined as the sensitivity of detecting genes that have consistent differential expression patterns across datasets. The results show that raising FDR increases the statistical power and Type I error rate for all three approaches. Increasing the mean difference ($|\Delta|$) between the two conditions (e.g. case vs. control) of the differentially expressed genes also improves sensitivity. In addition, the impact of the proportion of G_M on CDEP and Meta-RankProd is obvious: the higher the proportion of G_M (i.e., q), the lower the statistical power and Type I error rate. The reason is that obtaining FDR for these two approaches requires permutation and recalculation of EL_g^b and RP_g^b. After permutation, original G_M genes would act as G_C in CDEP and G_N in Meta-RankProd. As a result, when there is a higher proportion of G_M from the datasets, including these genes to estimate FDR would potentially lead to over-estimation because the variance of these genes is different from the non-differentially expressed genes [27]. Therefore, under the same FDR, the statistical power and the Type I error would be lower for higher q in CDEP and Meta-RankProd, especially when comparing q = 0.1 with q = 0.2. In contrast, Meta-Profile takes a relatively conservative approach, and is insensitive to genes that do not have consistent differential expression patterns. However, the tradeoff is the loss in statistical power. As shown in Table 1, even though the Type I error rate is amongst the lowest of the three approaches, the Meta-Profile method has the lowest statistical power. Overall, CDEP emerges as a robust meta-analysis method that obtains comparably high statistical power while maintaining low Type I error rate under different simulated conditions (see Additional File 1, Section 3: Comparison Between Different Approaches for Genes Appearing in Different Numbers of Datasets. More simulation results can be found in Additional File 2).

Using CDEP to identify a core gene set that is differentially expressed in Metastatic cancer

We used CDEP to investigate the hypothesis that there exists a core gene set differentially expressed consistently in different types of metastatic cancer cells. Six different types of cancer were investigated for this purpose (Table 2) [28-34]. Totally there are 220 samples, of which 126 are from primary and 84 from metastatic cancer, respectively. The diversity of these datasets (i.e. a wide variety of labs, different numbers of samples and probesets for different experiments, etc.) make them

Table 1 The Power and Type I error of CDEP, Meta-Profile and Meta-RankProd from simulation study.

				CDEP		Meta-Profile		Meta-RankProd	
p	q	$\|\Delta\|$	FDR	Power (%)	Type I error	Power (%)	Type I error	Power (%)	Type I error
0.05	**0.05**	**1**	**0.05**	28.7	1.64×10^{-4}	6.40	1.92×10^{-6}	23.7	1.21×10^{-2}
			0.1	31.0	3.52×10^{-4}	9.10	1.92×10^{-6}	24.5	1.23×10^{-2}
			0.2	70.2	2.21×10^{-3}	11.9	1.35×10^{-5}	27.2	1.27×10^{-2}
		2	**0.05**	33.4	2.02×10^{-4}	15.4	1.35×10^{-5}	33.2	1.18×10^{-2}
			0.1	34.3	4.22×10^{-4}	15.6	1.73×10^{-5}	45.0	1.24×10^{-2}
			0.2	74.9	2.28×10^{-3}	18.2	2.31×10^{-5}	56.3	1.38×10^{-2}
0.1	**0.1**	**1**	**0.05**	26.2	2.29×10^{-4}	8.93	2.84×10^{-6}	23.6	2.29×10^{-2}
			0.1	27.8	5.31×10^{-4}	11.4	4.88×10^{-6}	24.4	2.31×10^{-2}
			0.2	33.1	1.84×10^{-3}	13.4	1.12×10^{-4}	26.7	2.36×10^{-2}
		2	**0.05**	32.8	2.99×10^{-4}	15.6	5.90×10^{-5}	27.0	2.34×10^{-2}
			0.1	33.1	6.02×10^{-4}	18.2	1.16×10^{-4}	32.2	2.39×10^{-2}
			0.2	36.8	2.08×10^{-3}	23.6	2.32×10^{-4}	46.4	2.56×10^{-2}
0.05	**0.1**	**1**	**0.05**	28.2	1.61×10^{-4}	8.00	8.12×10^{-6}	24.3	1.16×10^{-2}
			0.1	29.5	3.11×10^{-4}	10.7	1.83×10^{-5}	25.7	1.18×10^{-2}
			0.2	66.5	1.78×10^{-3}	13.0	2.64×10^{-5}	29.7	1.22×10^{-2}
0.1	**0.2**	**1**	**0.05**	21.8	9.36×10^{-5}	10.3	4.35×10^{-5}	23.3	2.33×10^{-2}
			0.1	26.3	4.68×10^{-4}	12.5	9.14×10^{-5}	23.5	2.34×10^{-2}
			0.2	31.7	1.61×10^{-3}	14.0	1.94×10^{-4}	24.3	2.36×10^{-2}

p is the proportion of genes differentially expressed only in a certain dataset, and q is the proportion of consistently differentially expressed genes; Δ is the simulated mean difference between the expression values in case and control condition for the differentially expressed genes. FDR is the proportion of false positives among the genes identified to be consistently differentially expressed across all datasets. The results in the table are the mean values of 10 different simulated datasets. Additional simulation results can be found in Additional File 2.

Table 2 Description of the six microarray datasets used

Cancer Type	Number of samples	Number of Metastatic samples	Affymetrix Platform	Number of probesets	Number of genes
Cervical	33	12	HG-U133 P2	5,4675	20,271
Prostate	90	25	HG-U95Av2	1,2625	9,000
Gastric	22	15	Hu6800	7,129	5,526
Colon	6	3	HG-U133A	22,283	13,069
OSCC*	27	19	HG-U133A	22,283	13,069
RCC#	32	10	HG-U133A	22,283	13,069

Raw data were downloaded from the NCBI GEO database. *OSCC = oral squamous cell carcinoma; #RCC = renal cell carcinoma

ideal for assessing the robustness of CDEP and exploring our hypothesis. We used RMA [35] to pre-process the raw data for each dataset, and the median expression value of the probesets matching to the same Entrez gene id was used as the expression level for the gene.

At FDR ≤ 0.05, CDEP identified 239 genes that are differentially expressed consistently between the primary and metastatic cancer conditions across different cancer types. Out of these 239 genes, 141 were up-regulated and 98 down-regulated (Additional File 3). Table 3 shows the 5 most significantly up- and down-regulated genes identified. Using the same FDR criterion, we also performed meta-analysis by Meta-Profile and Meta-RankProd. Both CDEP and Meta-RankProd recovered the same two significant genes (*BSG and SLC25A1*) identified by Meta-Profile, and 180 genes were identified by both CDEP and Meta-RankProd (Meta-RankProd identified 2,967 significant genes. See Additional File 1, Figure S7 for details). A list of these genes identified by the three methods can be found in Additional File 4. These results further support using CDEP for meta-analysis to select candidate genes: Meta-Profile has insufficient statistical power, and Meta-RankProd tends to have high false positive rates. On the other hand, CDEP has the advantages of maintaining statistical power and keeping low false positive rates for identifying genes that are differentially expressed consistently.

The functional annotation of the 239 genes identified by CDEP is consistent with previous findings that many of them are involved in cancer metastasis. For instance, *GPNMB* overexpresses in a breast cancer cell line that could aggressively metastasize to bone [36], and *SPP1*'s expression level (also called Osteopontin) is elevated in a variety of metastatic cancers [37], including colon cancer [38], hepatocellular carcinoma [39], and gastric cancer [40]. Among the down-regulated genes, the expression level of *SERPINB5* is negatively associated with the depth of invasion, metastasis, and TNM stage in gastric cancer [41]. Interestingly, *SERPINB5* also inhibits invasion and metastasis of epithelial ovarian cancer, which suggests its down-regulation could promote metastasis [42]. *MX1* was also predicted to have an inhibitory effect on tumor cell motility and invasion, an essential attribute for metastatic behavior. While most of these previous findings are specific to different cancer types, analysis from CDEP indicates that these genes could play important roles in metastatic mechanism common to all types of cancers.

The function of these 239 genes were further investigated by DAVID [43] through Gene Set Enrichment Analysis, using all genes present in the microarray platform as background. The results indicate that the up-regulated genes in metastatic cancer cells are enriched in extracellular matrix (ECM) receptor interaction, focal adhesion, and angiogenesis, while down-regulated genes are enriched in genes functioning in immune and inflammatory response (Table 4), and these include laminin, fibronectin, collagen, multimerin, caveolin, etc.. Figure 5 shows the CDEP identified genes mapped to the ECM receptor interaction and focal adhesion pathways. It is widely recognized that these pathways contribute to the aggressiveness and the metastatic behavior of cancer cells [44].

Not only had CDEP identified genes known to be involved in cancer metastasis, but also it discovered novel genes that have not been implicated in metastatic mechanism. For example, *AMIGO2* , a gene identified as differentially expressed by only CDEP out of the three

Table 3 Five most significant genes identified by CDEP as related to common metastatic mechanism across different cancer types by using FDR < 0.05 as threshold

Up-regulated genes	Down-regulated genes
Glycoprotein (transmembrane) nmb (*GPNMB*)	Serpin peptidase inhibitor, clade B (ovalbumin), member 5 (*SERPINB5*)
Secreted phosphoprotein 1 (*SPP1*)	proteasome (prosome, macropain) subunit, beta type, 9 (*PSMB9*)
Transforming growth factor, beta-induced (*TGFBI*)	myxovirus (influenza virus) resistance 1, interferon-inducible protein p78 (*MX1*)
Heat shock 27kDa protein 1 (*HSPB1*)	interferon-induced protein with tetratricopeptide repeats 1 (*IFIT1*)
Mesoderm specific transcript homolog (*MEST*)	ubiquitin D (*UBD*)

Table 4 Gene Set Enrichment Analysis to identify functional groups from CDEP identified genes

Functions (source)	Annotated genes	FDR
ECM-receptor interaction (KEGG)	*COL4A2,COL4A1,TNC,COL3A1,COL5A2,COL5A1,COL6A3,COL1A2,COL1A1,LAMB1,COL11A1,THBS2, FN1,SPP1*	1.84×10^{-8}
Focal adhesion (KEGG)	*CAV1,COL4A2,COL4A1,TNC,COL3A1,COL5A2,COL5A1,MYL9,VEGFC,VEGFA,COL6A3,COL1A2, COL1A1,LAMB1,THBS2,COL11A1,FN1,SPP1*	2.02×10^{-7}
Blood vessel development (GO)	*PLAT,CAV1,IL8,COL3A1,COL15A1,COL5A1,VEGFC,APOB,APOE,CTGF,VEGFA,COL1A2,SEMA3C,LOX,COL1A1,CYR61*	1.10×10^{-5}
Immune response (GO)	*CXCL1,POU2AF1,CCL2,BST2,CXCL3,IGJ,CXCL2,CXCL9,IL32,IFI44L,CCL5,CXCL11,HLA-DMA, HLA-DQA1,PSMB8,CXCL10,PSMB9,CXCL13,TAP1,DEFB1,GBP1*	2.06×10^{-6}
Inflammatory response (GO)	*CXCL1,CCL2,NMI,CXCL3,CXCL2,CXCL9,ANXA1,IDO1,CXCL11,CCL5,CXCL10,FOS,SAA2,CXCL13*	5.95×10^{-5}

The functional enrichment for the genes identified as consistently differentially expressed between primary and metastatic cancers.

meta-analysis approaches, is involved in anti-apoptosis [45] and cell adhesion [46], and CDEP shows that this gene is up-regulated under metastases conditions. *Gem* is also differentially expressed consistently in metastatic cancer cells, even though no current literature indicates its role in metastasis. However, the gene interacts with microtubule network [47] and regulates actin dynamics [48]. Such activities are highly related to the migratory and invasive properties of cancer metastasis [49]. Another gene, *CXCL11*, was shown to be consistently down-regulated by CDEP analysis. Given that this gene has angiostatic property [50], our results suggest that its down-regulation in metastatic cancer might interrupt the angiostasis process and promote angiogenesis–an important aspect of cancer metastasis.

Discussion

Meta-analysis provides a cost-efficient way to approach biological problems. However, the heterogeneous nature of the data is always a significant challenge. CDEP aims to overcome this hurdle to identify genes that have a common differential expression pattern across different datasets. We illustrated that CDEP can: (i) obtain higher statistical power than existing meta-analysis approaches while maintaining low Type I error rate in the simulation study, and (ii) identify genes that are potentially involved in common metastatic behaviors and relevant biological pathways. CDEP borrows information from each dataset to identify genes differentially expressed consistently–a flexible approach that can be generalized to problems other than metastasis. The high statistical power under diverse sets of parameters considered in the simulation study also suggests robustness of CDEP to the diversity of data sources.

In CDEP, the minus log likelihood Q for different FDR values (l) was used because CDEP does not intend to "filter out" genes in each dataset before performing meta-analysis. This is in contrast to Meta-Profile where genes that only met the threshold (l) in each dataset were used for the meta-analysis. In CDEP, we emphasized low l values to calculate EL and thus employed a linearly decreasing PDF for the log likelihood to: 1) balance the "filtering" behavior that would result from a convex decreasing PDF; and 2) emphasize small l. The PDF used in CDEP outperformed Meta-Profile and Meta-RankProd in obtaining high statistical power and lowering Type I error rate.

CDEP, Meta-Profile, and Meta-RankProd use different permutation approaches to estimate FDR_g. Meta-RankProd permutes gene expression values relative to the gene ID for each array while Meta-Profile and CDEP permute FDR relative to gene ID for each dataset examined. The null distribution produced by Meta-RankProd permutation would lead to RP_g representing genes that are non-differentially expressed in any dataset, while Meta-Profile and CDEP would simply increase the proportion of genes that are differentially expressed in only a single dataset after permutation. Therefore, Meta-RankProd tends to under-estimate FDR_g, as it ignores genes that are only differentially expressed in a single dataset. On the other hand, Meta-Profile and CDEP would over-estimate FDR_g because they have a higher proportion of G_C genes in the null distribution compared to Meta-RankProd. Even though inaccurate estimation of FDR_g is inevitable due to the lack of prior knowledge about the proportion of genes only differentially expressed in one versus multiple datasets, both CDEP and Meta-Profile employed a more conserved approach than Meta-RankProd to obtain high precision.

Conclusion

CDEP is a flexible meta-analysis approach that borrows information from each dataset in order to identify genes that are consistently differentially expressed. CDEP obtains higher statistical power than two existing approaches under a variety of scenarios considered in the simulation study, suggesting its robustness and insensitivity to data variation. By application to metastatic cancer datasets as a case study, CDEP allows identification of genes differentially expressed consistently in different types of metastatic cancer cells. These

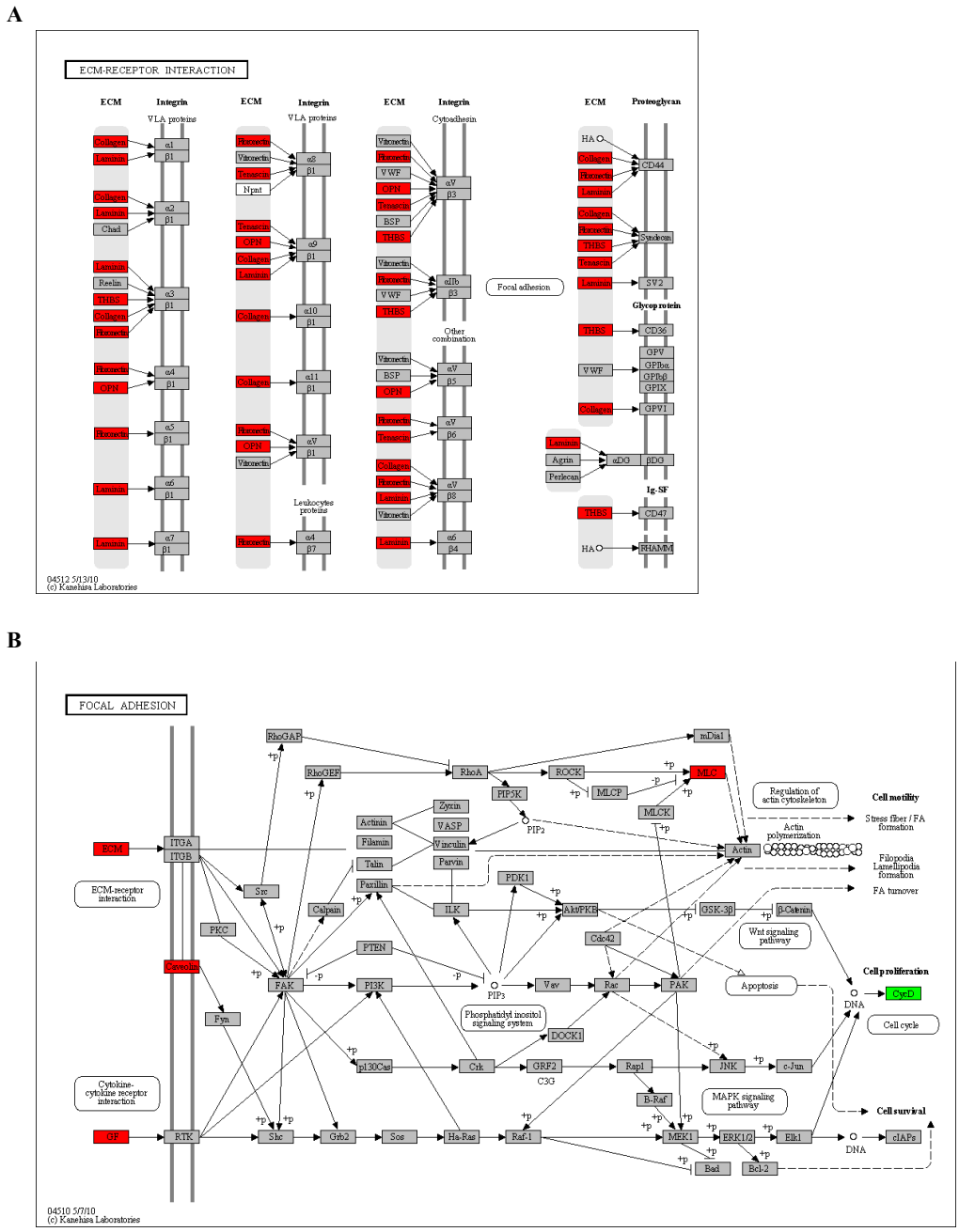

Figure 5 CDEP identified differentially expressed genes map to biological pathways relevant to cancer metastasis. A) ECM-receptor interaction pathway. B) Focal adhesion pathway. Up-regulated genes are annotated with red color, and down-regulated genes are in green.

identified genes could be further developed into universal biomarkers for cancer staging and diagnosis.

Methods

Microarray data

We searched the published microarray datasets comparing primary and metastatic cancer samples from the NCBI Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/gds/). To ensure high quality, the included datasets and the associated publications needed to have clear descriptions about primary and metastatic cancer samples, and only single channel oligonucleotide array from Affymetrix was considered. In addition, we selected only datasets that provide raw data so we could apply the same pre-processing method for consistency. Since the goal of this study was to

identify common genes across different cancer types, only one dataset from each cancer type was used to avoid biases. If a cancer type had multiple datasets that met our criteria, the dataset with the most precise and detailed description of the samples was included.

Simulation of Microarray data

We used simulation to evaluate the performance of CDEP. Simulated data were generated to mimic the real experimental datasets retrieved. Specifically, the retrieved raw datasets listed in Table 2 were pre-processed using the Robust Multichip Average (RMA) algorithm [35] to generate the summarized probe readings of the probesets in logarithmic values. The resulting probesets were then matched to the corresponding NCBI Entrez gene id using the annotations from Affymetrix [51]. Because each gene will be matched to multiple probesets, the median expression value of these matched probesets was used for each gene. Overall, for each of the six different cancer types we selected, we generated a simulated dataset for meta-analysis. Each simulated dataset represented comparison of metastatic and primary cancer cells for the corresponding cancer type. The simulation was done using a modified version of a model described in Stevens and Doerge [52]:

$$\begin{aligned} y_{gijk} &= \mu + L_i + G_{gi} + \alpha_{gi}(T_{gjk}) \\ &\quad + \beta_{gij}(M_{gijk}) + \varepsilon_{gijk} \end{aligned} \tag{5}$$

where y_{gijk} is the expression value of gene g in experiment k conducted by laboratory i in cancer condition j (j = 1 for metastatic; j = 0 for primary), μ is the universal background reading, and L_i and G_{gi} are the effects of laboratory i and gene g from laboratory i, respectively. We also incorporated binary variables α_{gj} and β_{gij} to distinguish genes that have common differential expression pattern from genes that are differentially expressed pertinent only to a specific dataset: differentially expression of gene g owing to the cancer-type specific effect is indicated by α_{gj} = 1, and owing to the mechanism of the common metastatic behavior across different cancer types is indicated by β_{gij} = 1. These indicators are used in the model to determine the contributions of cancer type specific and metastatic effects (T_{gik} and M_{gijk}, respectively) to the gene expression value. The detail for the simulation model is as followed:

$$\begin{aligned} &\mu = a_1 \\ &L_i \sim N(0, b_2) \\ &G_{gi} \sim N(0, b_{3i}) \\ &\alpha_{g1} \sim Bern(p) \\ &\beta_{gi1} \sim Bern(q) \\ &T_{gjk} \sim N(\Delta, \psi), \text{ where } \Delta \neq 0 \\ &M_{gijk} \sim N(\Delta, \psi), \text{ where } \Delta \neq 0 \\ &\varepsilon_{gijk} \sim N(0, e_i) \end{aligned}$$

In the simulation, we assumed that different datasets have different numbers of probesets and experiments. In each dataset, genes selected to be involved in cancer-specific effect (i.e. α_{g1} = 1) were randomly assigned as up- or down-regulated to make such behavior dependent on dataset (cancer type). Genes selected to be involved in the metastatic behavior common to all cancer type were also randomly assigned as up- or down-regulated genes to make it independent of dataset (cancer type). However, a gene could only be in at most one of these two categories, i.e. we required that $\alpha_{g1} + \beta_{gi1}$ is contained in the set {0, 1}. The above simulation parameters were estimated from microarray datasets comparing primary versus metastatic cancer cells (Table 2). During simulation, we used different values for the proportion of cancer-type specific (p) and metastatic related differentially expressed genes (q). We also examined different cancer-specific and metastatic-related effects Δ. We then applied CDEP to identify genes involved in metastatic behavior in this simulation study.

Bayesian mixture model for p-value

We used a mixture of beta distributions to model the p-values arising from the RankProd method. Because the p-values correspond to genes that are up-regulated, non-differentially expressed, or down-regulated, we used a 3-component mixture model. The p-value for gene g, y_g, is represented as $y_g = \sum_{k=1}^{3} T_{gk} p_k$, where $T_g = (T_{g1}, T_{g2}, T_{g3}) \sim$ Multinomial(θ, 1) so that exactly one element of T_g is one and the remaining elements are zero. The value p_k arises from the k^{th} component of the mixture: $p_k \sim$ Beta(a_k, b_k), k = 1, 2, 3. We used a Dirichlet prior for θ, $\theta \sim$ Dir(1, 18, 1) and further assigned prior distributions as $a_1 = b_3 = 1$; $a_2 \sim$ Gamma(4, 2); a_3, $b_1 \sim$ Gamma(400, 20), and $b_2 \sim$ Gamma(1, 1), where the Gamma(α, β) is parameterized so that the mean is α/β.

Comparison with other approaches

To assess the robustness of CDEP, we compared it with Meta-Profile and Meta-RankProd [14-16,21]. Meta-Profile is one of the pioneering methods to investigate common cancer signatures at large scale. This approach first identifies a dataset-specific "differential expression signature"–a list of differentially expressed genes for each dataset determined by the pre-defined threshold of FDR (l) [5]. The number of signatures each gene appeared in is then counted and permutation is performed to estimate the false positives of this count. The Meta-RankProd approach is a relatively recent approach that uses the rank product to identify genes differentially expressed between two conditions from multiple datasets. In this method, the rank fold change, γ_{gih}, is

Table 5

Symbol	Range	Annotation
d_{il}	1,2,...,G	Number of genes in a dataset with FDR lower than the threshold l
EL_g	(0,Inf)	Expected value of the log likelihood with respect to the FDR threshold
f_{il}	1,2,..., d_{il}	Number of false positives using the FDR threshold l
F_{gi}	(0,1)	Gene-specific false discovery rate in dataset i: proportion of false positives among the significant calls
FDR_g	(0,1)	Gene-specific false discovery rate for having consistently differentially expressed patterns among the datasets studied
g	1,2,...,G	Index for a gene from the union of gene sets across all datasets
h	1,2,...,H_i	Index for fold change comparison between a case and a control from a dataset, where $m_i * n_i = H_i$
i	1,2,...,D	Index for a gene expression microarray dataset (consists of m_i cases and n_i controls)
l	(0,1)	FDR threshold used to enumerate number of genes with FDR lower than this threshold in a dataset and to estimate the number of false positives under this threshold
$L(\mathbf{r}_{gl}\|Data)$	(0,1)	Gene- and FDR threshold- specific likelihood of observing the differential expressed pattern among the datasets
$M^{\overline{up}}{}_i(M^{\overline{down}}{}_i)$	1,2,...,G	number of genes that are not up(not down)-regulated in dataset i
Q_{gl}	(0,Inf)	Minus log likelihood
$\hat{r}_{il}^{up}(\hat{r}_{il}^{down})$	(0,1)	False positive rate: the probability of a non-up-regulated (non-down-regulated) gene being falsely called as over-expressed (under-expressed)
δ_{gil}^{up} (or δ_{gil}^{down})	0[1]	Binary variable indicating gene g is identified as up(or down)-regulated in dataset i for threshold l
γ_{gih}	1,2,...,G	rank of fold change for gene g in the h^{th} comparison of dataset i

computed as the ranking of gene g in the h^{th} comparison in the i^{th} study, and the rank product for gene g was calculated as the geometric mean across all comparisons. The null rank product was obtained by permuting expression values within each single array. This method was shown to outperform both the parametric t-based modeling approach [53] and the Fisher's inverse Chi-square approach [6] in terms of sensitivity and specificity. CDEP, Meta-Profile and Meta-RankProd were applied to analyze the simulated datasets to evaluate their performances in terms of: i) the statistical power to identify genes with common differential expression pattern across datasets; and ii) Type I error rate of falsely identifying genes without common differential expression. In this analysis, we tested the effect of different proportions of differentially expressed genes attributed to cancer-type specific (p) and metastatic-related (q). We also examined the effects of the detectable difference (Δ) of differential expression. For RankProd and CDEP, genes absent from a dataset were assigned the median rank value of that dataset.

Additional material

Additional file 1: Supplementary materials for the analysis. Detailed descriptions about: 1) Datasets Used (Suppl. Table 1); 2) The comparisons between p-values computed by the parametric t-test versus the non-parametric RankProd (Suppl. Figure 1); 3) The Bayesian mixture for the p-value distribution (Suppl. Figure 2, Table 2 and Table 3); 4) Comparisons of different approaches for handling genes appearing in different numbers of datasets based on simulation (Suppl. Figure 3, Figure 4, Figure 5 and 6); and 5) Comparisons of the three approaches using the 6 cancer datasets as case study (Suppl. Figure 7).

Additional file 2: Results from the simulation data. The statistical power and Type I error rate are compared for the three meta-analysis approaches on simulation data.

Additional file 3: Metastases-related genes identified by CDEP. Statistically significant genes identified by CDEP as related to metastatic behavior by using *FDR = 0.05*.

Additional file 4: Comparison of Significant Genes Identified by CDEP, Meta-Profile and Meta-RankProd. List of genes that are differentially expressed consistently in metastatic cancer cells as identified by CDEP, Meta-Profile and Meta-RankProd from six data sets used.

List of abbreviations used

Acknowledgements
We thank all researchers who share their valuable microarray data with the public, and especially for those authors who helped us to understand the descriptions of their samples. We acknowledge all the kind and effective suggestions and opinions provided via the Bioconductor mailing list. We thank John Lucas, and Drs. Dennis Watson and Caroline Reed for their valuable comments and suggestions when we retrieved the microarray data. *Funding*: This research was supported by a pilot project from grant NIH/NCRR P20 RR017696-05; NIH/NIGMS R01GM063265-09S1; P20 RR017677-10 (WJZ); PhRMA Foundation Research Starter Grant (W.J.Z); NLM Training Grant (5-T15-LM007438-02 to L.C.T.) NIH/NCI R03CA137805 (E.S.); NSF DMS 0604666 (E.S.); NIH/NCRR P20RR017696 (E.S.). T.Q. was supported by PhRMA Foundation Research Starter Grant, NIH/NCRR 5P20RR017677-10, NIH/NIGMS R01GM063265-09S1 and T32GM074934 07.

Author details
[1]Bioinformatics Graduate Program, Department of Biochemistry and Molecular Biology, Medical University of South Carolina, 135 Cannon St. Charleston, SC 29425, USA. [2]Department of Statistics, Florida State University, 117 N. Woodward Ave. Tallahassee, FL 32306, USA. [3]Division of Bioinformatics, Department of Biochemistry and Molecular Biology, Medical University of South Carolina, 135 Cannon St. Charleston, SC 29425, USA.

Authors' contributions
WJZ conceived the initial idea of the project and worked with LCT on data selection and analysis. EHS advised the statistical method development of

the project. LCT and TQ wrote the R and Winbugs codes for the analysis. LCT drafted and WJZ and EHS finalized the manuscript. WJZ supervised the overall development of the project. All authors have read and approved the manuscript.

Conflict of interests

The authors declare that they have no competing interests.

Received: 3 June 2011 Accepted: 11 November 2011
Published: 11 November 2011

References

1. Barrett T, Edgar R: **Mining microarray data at NCBI's Gene Expression Omnibus (GEO)*.** *Methods Mol Biol* 2006, **338**:175-190.
2. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone SA: **ArrayExpress–a public repository for microarray gene expression data at the EBI.** *Nucleic Acids Res* 2003, **31(1)**:68-71.
3. Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, Eisen MB, Spellman PT, Brown PO, Botstein D, Cherry JM: **The Stanford Microarray Database.** *Nucleic Acids Res* 2001, **29(1)**:152-155.
4. Ghosh D, Barette TR, Rhodes D, Chinnaiyan AM: **Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer.** *Funct Integr Genomics* 2003, **3(4)**:180-188.
5. Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM: **Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer.** *Cancer Res* 2002, **62(15)**:4427-4433.
6. Moreau Y, Aerts S, De Moor B, De Strooper B, Dabrowski M: **Comparison and meta-analysis of microarray data: from the bench to the computer desk.** *Trends Genet* 2003, **19(10)**:570-577.
7. Conlon EM: **A Bayesian mixture model for metaanalysis of microarray studies.** *Funct Integr Genomics* 2008, **8(1)**:43-53.
8. Conlon EM, Song JJ, Liu JS: **Bayesian models for pooling microarray studies with multiple sources of replications.** *BMC Bioinformatics* 2006, **7**:247.
9. Shen R, Ghosh D, Chinnaiyan AM: **Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data.** *BMC Genomics* 2004, **5(1)**:94.
10. Breitling R, Armengaud P, Amtmann A, Herzyk P: **Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments.** *FEBS Lett* 2004, **573(1-3)**:83-92.
11. Kong X, Mas V, Archer KJ: **A non-parametric meta-analysis approach for combining independent microarray datasets: application using two microarray datasets pertaining to chronic allograft nephropathy.** *BMC Genomics* 2008, **9**:98.
12. Cahan P, Rovegno F, Mooney D, Newman JC, St Laurent G, McCaffrey TA: **Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization.** *Gene* 2007, **401(1-2)**:12-18.
13. Boedigheimer MJ, Wolfinger RD, Bass MB, Bushel PR, Chou JW, Cooper M, Corton JC, Fostel J, Hester S, Lee JS, Liu F, Liu J, Qian HR, Quackenbush J, Pettit S, Thompson KL: **Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories.** *BMC Genomics* 2008, **9**:285.
14. Hong F, Breitling R: **A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments.** *Bioinformatics* 2008, **24(3)**:374-382.
15. Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J: **RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis.** *Bioinformatics* 2006, **22(22)**:2825-2827.
16. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: **Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.** *Proc Natl Acad Sci USA* 2004, **101(25)**:9309-9314.
17. Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A: **False discovery rate, sensitivity and sample size for microarray studies.** *Bioinformatics* 2005, **21(13)**:3017-3024.
18. Ma S, Huang J: **Regularized gene selection in cancer microarray meta-analysis.** *BMC Bioinformatics* 2009, **10**:1.
19. Rhodes DR, Chinnaiyan AM: **Bioinformatics strategies for translating genome-wide expression analyses into clinically useful cancer markers.** *Ann N Y Acad Sci* 2004, **1020**:32-40.
20. Culhane AC, Quackenbush J: **Confounding effects in "A six-gene signature predicting breast cancer lung metastasis".** *Cancer Res* 2009, **69(18)**:7480-7485.
21. Rhodes DR, Chinnaiyan AM: **Integrative analysis of the cancer transcriptome.** *Nat Genet* 2005, **37(Suppl)**:S31-37.
22. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5(10)**: R80.
23. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proc Natl Acad Sci USA* 2003, **100(16)**:9440-9445.
24. Thomas A, Hara BO, Ligges U, Sturtz S: **Making BUGS Open.** *R News* 2006, **6**:12-17.
25. Lunn DJ, Thomas A, Best N, Spiegelhalter D: **WinBUGS–a Bayesian modelling framework: concepts, structure, and extensibility.** *Statistics and Computing* 2000, **10**:325-337.
26. Lunn DJ, Spiegelhalter D, Thomas A, Best N: **The BUGS project: Evolution, critique and future directions.** *Statistics in Medicine* 2009, **28**:3049-3082.
27. Zhang S: **A comprehensive evaluation of SAM, the SAM R-package and a simple modification to improve its performance.** *BMC Bioinformatics* 2007, **8**:230.
28. Bachtiary B, Boutros PC, Pintilie M, Shi W, Bastianutto C, Li JH, Schwock J, Zhang W, Penn LZ, Jurisica I, Fyles A, Liu FF: **Gene expression profiling in cervical cancer: an exploration of intratumor heterogeneity.** *Clin Cancer Res* 2006, **12(19)**:5632-5640.
29. Chandran UR, Ma C, Dhir R, Bisceglia M, Lyons-Weiler M, Liang W, Michalopoulos G, Becich M, Monzon FA: **Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process.** *BMC Cancer* 2007, **7**:64.
30. Hippo Y, Taniguchi H, Tsutsumi S, Machida N, Chong JM, Fukayama M, Kodama T, Aburatani H: **Global gene expression analysis of gastric cancer by oligonucleotide microarrays.** *Cancer Res* 2002, **62(1)**:233-240.
31. O'Donnell RK, Kupferman M, Wei SJ, Singhal S, Weber R, O'Malley B, Cheng Y, Putt M, Feldman M, Ziober B, Muschel RJ: **Gene expression signature predicts lymphatic metastasis in squamous cell carcinoma of the oral cavity.** *Oncogene* 2005, **24(7)**:1244-1251.
32. Provenzani A, Fronza R, Loreni F, Pascale A, Amadio M, Quattrone A: **Global alterations in mRNA polysomal recruitment in a cell model of colorectal cancer progression to metastasis.** *Carcinogenesis* 2006, **27(7)**:1323-1333.
33. Yu YP, Landsittel D, Jing L, Nelson J, Ren B, Liu L, McDonald C, Thomas R, Dhir R, Finkelstein S, Michalopoulos G, Becich M, Luo JH: **Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy.** *J Clin Oncol* 2004, **22(14)**:2790-2799.
34. Jones J, Otu H, Spentzos D, Kolia S, Inan M, Beecken WD, Fellbaum C, Gu X, Joseph M, Pantuck AJ, Jonas D, Libermann TA: **Gene signatures of progression and metastasis in renal cell cancer.** *Clin Cancer Res* 2005, **11(16)**:5730-5739.
35. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4(2)**:249-264.
36. Rose AA, Pepin F, Russo C, Abou Khalil JE, Hallett M, Siegel PM: **Osteoactivin promotes breast cancer metastasis to bone.** *Mol Cancer Res* 2007, **5(10)**:1001-1014.
37. Denhardt DT, Mistretta D, Chambers AF, Krishna S, Porter JF, Raghuram S, Rittling SR: **Transcriptional regulation of osteopontin and the metastatic phenotype: evidence for a Ras-activated enhancer in the human OPN promoter.** *Clin Exp Metastasis* 2003, **20(1)**:77-84.
38. Takami Y, Russell MB, Gao C, Mi Z, Guo H, Mantyh CR, Kuo PC: **Sp1 regulates osteopontin expression in SW480 human colon adenocarcinoma cells.** *Surgery* 2007, **142(2)**:163-169.
39. Takafuji V, Forgues M, Unsworth E, Goldsmith P, Wang XW: **An osteopontin fragment is essential for tumor cell invasion in hepatocellular carcinoma.** *Oncogene* 2007, **26(44)**:6361-6371.

40. Tang H, Wang J, Bai F, Zhai H, Gao J, Hong L, Xie H, Zhang F, Lan M, Yao W, Liu J, Wu K, Fan D: **Positive correlation of osteopontin, cyclooxygenase-2 and vascular endothelial growth factor in gastric cancer.** *Cancer Invest* 2008, **26(1)**:60-67.
41. Lee DY, Park CS, Kim HS, Kim JY, Kim YC, Lee S: **Maspin and p53 protein expression in gastric adenocarcinoma and its clinical applications.** *Appl Immunohistochem Mol Morphol* 2008, **16(1)**:13-18.
42. Ma Y, Peng ZL: **[Expression of maspin and its relation to tumor vascularization in epithelian ovarian cancer].** *Sichuan Da Xue Xue Bao Yi Xue Ban* 2009, **40(2)**:223-227.
43. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4(1)**:44-57.
44. Weinberg RA: **The biology of cancer.** New York: Garland Science; 2007.
45. Ono T, Sekino-Suzuki N, Kikkawa Y, Yonekawa H, Kawashima S: **Alivin 1, a novel neuronal activity-dependent gene, inhibits apoptosis and promotes survival of cerebellar granule neurons.** *J Neurosci* 2003, **23(13)**:5887-5896.
46. Kuja-Panula J, Kiiltomaki M, Yamashiro T, Rouhiainen A, Rauvala H: **AMIGO, a transmembrane protein implicated in axon tract development, defines a novel protein family with leucine-rich repeats.** *J Cell Biol* 2003, **160(6)**:963-973.
47. Piddini E, Schmid JA, de Martin R, Dotti CG: **The Ras-like GTPase Gem is involved in cell shape remodelling and interacts with the novel kinesin-like protein KIF9.** *EMBO J* 2001, **20(15)**:4076-4087.
48. Splingard A, Menetrey J, Perderiset M, Cicolari J, Regazzoni K, Hamoudi F, Cabanie L, El Marjou A, Wells A, Houdusse A, de Gunzburg J: **Biochemical and structural characterization of the gem GTPase.** *J Biol Chem* 2007, **282(3)**:1905-1915.
49. Hall A: **The cytoskeleton and cancer.** *Cancer Metastasis Rev* 2009, **28(1-2)**:5-14.
50. Lasagni L, Francalanci M, Annunziato F, Lazzeri E, Giannini S, Cosmi L, Sagrinati C, Mazzinghi B, Orlando C, Maggi E, Marra F, Romagnani S, Serio M, Romagnani P: **An alternatively spliced variant of CXCR3 mediates the inhibition of endothelial cell growth induced by IP-10, Mig, and I-TAC, and acts as functional receptor for platelet factor 4.** *J Exp Med* 2003, **197(11)**:1537-1549.
51. Affymetrix: **Statistical Algorithms Description Document.** Santa Clara: Affymetrix; 2002.
52. Stevens JR, Doerge RW: **Meta-analysis combines affymetrix microarray results across laboratories.** *Comp Funct Genomics* 2005, **6(3)**:116-122.
53. Choi JK, Yu U, Kim S, Yoo OJ: **Combining multiple microarray studies and modeling interstudy variation.** *Bioinformatics* 2003, **19(Suppl 1)**:i84-90.

doi:10.1186/1471-2105-12-438
Cite this article as: Tsoi *et al.*: Consistent Differential Expression Pattern (CDEP) on microarray to identify genes related to metastatic behavior. *BMC Bioinformatics* 2011 **12**:438.