Published in IET Systems Biology
Received on 31st December 2008
Revised on 17th June 2009
doi: 10.1049/iet-syb.2008.0183

Special Issue – Selected papers from The 2nd International Symposium on Optimization and Systems Biology (OSB 2008)

ISSN 1751-8849

Pathway level analysis by augmenting activities of transcription factor target genes

H. Jung[1] *E. Lee*[2] *J.-W. Kim*[3] *D. Lee*[2]

[1]*UCSD Bioinformatics Graduate Program, UCSD, La Jolla, California 92093-0653, USA*
[2]*Department of Bio and Brain Engineering, KAIST, Korea*
[3]*Department of Laboratory Medicine and Genetics, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea*
E-mail: dhlee@kaist.ac.kr

Abstract: Many approaches to discovering significant pathways in gene expression profiles have been developed to facilitate biological interpretation and hypothesis generation. In this work, the authors propose a pathway identification scheme integrating the activity of pathway member genes with that of target genes of transcription factors (TFs) in the same pathway by the weighted *Z*-method. The authors evaluated the integrative scoring scheme in gene expression profiles of essential thrombocythemia patients with JAK2V617F mutation status, primary breast tumour samples with the status of metastasis occurrence, two independent lung cancer expression profiles with their prognosis, and found that our approach identified cancer-type-specific pathways better than gene set enrichment analysis (GSEA) and Tian's method using the original pathways [pathways that have TFs from database] and the extended pathways (including target genes of TFs of the original pathways). The success of our scheme implicates that adding information of transcriptional regulation is better way of utilising mRNA measurements for estimating differential activities of pathways from gene expression profiles more exactly.

1 Introduction

Interpretation of biological meaning from genome-wide expression profiles is still challenging. Much of the initial works have concentrated on the identification of differentially expressed genes and verification of their statistical significance. However, in most cases, biological insights cannot be extracted from the identified differentially expressed genes because the interpretation of the large list of genes is daunting works. Another problem of this approach is caused by the use of the cut off threshold value, because the results of this approach are significantly affected by the selected threshold [1]. To deal with this problem, recent efforts have interpreted microarray data by using prior knowledge such as gene ontology (GO) and pathway databases. These researches make it possible to systematically dissect large gene lists in an attempt to assemble a summary of the most enriched and pertinent biology [2].

These methods are uniquely categorised into three major classes, according to their underlying algorithms: singular enrichment analysis (SEA), gene set enrichment analysis (GSEA) and modular enrichment analysis. Here we focus on SEA and GSEA approaches [2]. SEA approach is to take a set of differentially expressed genes and identify distinct GO categories or pathways. The number of differentially expressed genes found in the predefined sets is compared with the number of genes expected to be found in the given predefined sets by chance. In this analysis, the p value can be calculated by with the aid of some common and well-known statistical methods, including chi-square, Fisher's exact test, binomial probability, hypergeometric distribution and so on. However, the limitation of this approach is that only the most significant portion of the gene list is used to compute the statistic, treating the less relevant genes as irrelevant at all [2, 3].

Second category approach considers the distribution of pathway genes in the entire list of genes. The unique idea of this approach is its 'no-cutoff' strategy that takes all genes from gene expression profiles without selecting significant genes unlike SEA approach. Currently, over 20

different tools are available for pathway level analysis [4]. GSEA is one of the most popular methods which use an enrichment score (ES) based on Kolmogorov–Smirnov static as the test statistic [5]. Tian's method is also widely used for identifying differentially activated pathways. *t*-Test is applied to find relationships between the expression levels and then a testing procedure is used to find significant pathways [6]. Dinu *et al.* pointed out problems of GSEA and extended a single gene analysis by significance analysis of microarray (SAM) for pathway level analysis (SAM-GS). Their test statistic was the L2 norm of the vector of the SAM statistics, corresponding to the genes in the pathway of interest [7]. Efron and Tibshirani [8] introduced maxmean statistic for GSEA algorithm (GSA). They concluded that maxmean statistics is the only method with consistently low p values in all situations. Unlike other pathway level analysis tools, the impact factor analysis takes into consideration important biological factors: the magnitude of the expression changes of each gene, the position of the differentially expressed genes on the given pathways, the topology of the pathway that describes how these genes interact and the type of signalling interactions between them [9].

Previous pathway level analysis methods use gene sets from database such as KEGG and BioCarta. Pathway level analysis methods using gene sets from KEGG and BioCarta give relevant results, but in some cases, they are not robust and cannot find altered pathways from microarray data. In particular, these approaches in cancer expression profiling sometimes makes results of cancer-type-non-specific pathways, such as cell cycle pathways and P53 associated pathways. Here, we proposed an extended pathway and a pathway integrative scoring scheme considering the expression levels of target genes of a transcription factor (TF) assuming that the effect of transcriptional regulation following the pathway activation by different types of regulation such as phosphorylation can be directly measured from mRNA expression levels of TF target genes. The extended pathway is defined to include TF target genes of the original pathway (pathways that have at least one human TF from pathway database). The pathway integrative scoring scheme considers two p values each reflecting the differential expression of pathway member genes and TF target genes for each pathway. These two p values are integrated by the weighted Z-transform method (Fig. 1).

We use two distinct pathway level analysis methods, GESA and Tian's method, to evaluate the extended pathways and compare those algorithms with the integrative scoring scheme. Two chosen pathway level analysis methods using both the original pathways and the proposed extended pathways, and the integrative scoring scheme were evaluated in gene expression profiles of essential thrombocythemia (ET) patients with JAK2V617F mutation status, primary breast tumour samples with the status of metastasis occurrence, two independent lung cancer expression profiles with their prognosis. We found that the integrative scoring method identified more cancer-type-specific pathways than GSEA and Tian's method using the original pathways and the extended pathways.

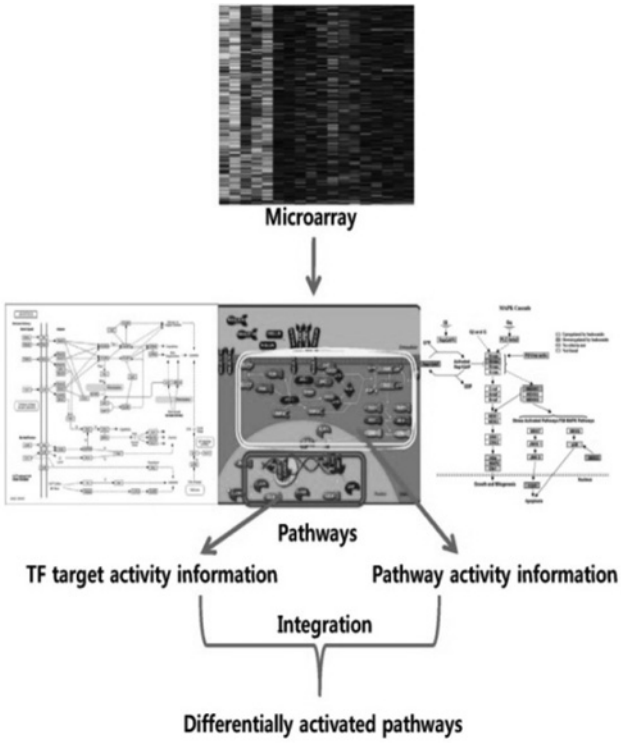

Figure 1 *Schematic overview of pathway identification.*
Differentially activated pathways are identified by integrating gene expression levels of TF target genes with those of pathway member genes by combining their p values through the weighted Z method

2 Materials and methods

2.1 Pathway sets and TF target databases

Firstly, we collected human TF list [10], and downloaded MsigDB canonical pathway gene sets (c2.cp.v2.5.symbols.gmt) for pathway information. Among canonical pathways in MsigDB [5], we selected 248 pathways that have at least one human TF, and collected TF target genes from TRANSFAC database 11.0 [11] and BZIP database [12]. The original pathways represent above 248 pathways and the extended pathways represent the same 248 pathways which include target genes of TFs of the original pathways. Repressed targets by TF were not considered because repressed target information was very limited compared to activated target information. Only activated targets by TFs were considered in this analysis.

2.2 Gene expression datasets

We applied our method to previously published four mRNA expression datasets: expression profiles of 16 ET patients

with JAK2V617F mutation status [13], expression profiles of 295 primary breast tumour samples with the status of metastasis occurrence [14], two independent lung cancer expression profiles of 86 patients from Michigan group [15] and 62 patients from Boston group with their prognosis [16].

Each dataset had two different classes of samples. For the ET study, nine ET patients possessed JAK2V617F mutation ($JAK2V617F^{+}$) and seven ET patients did not ($JAK2V617F^{-}$). For the breast cancer study, 78 of 295 patients had metastasis during follow-up visits within 5 years after surgery, and the remaining 217 of 295 patients did not. For the two lung cancer datasets, 24 of 86 patients had poor outcome and 62 of 86 patients had good outcome in Michigan study, and half of 62 patients from Boston group had poor outcome and the rest half of patients had good outcome.

2.3 Pathway level analysis methods

The size of each gene set and the number of permutations should be the same to make fair comparisons of all methods. We set the minimal pathway size as 10 and the maximum pathway size as 500 in all methods. p values in all methods were computed based on 1000 random permutation of genes. The p value for the original pathway from GSEA and Tian's method was combined with the p value for the target genes of TFs of that pathway from our TF target genes' activity scoring by the weighted Z transform method.

2.3.1 Gene set enrichment analysis: GSEA firstly calculate the ES, which reflects the degree to which a set S is overrepresented at the extremes of the entire ranked list L [5]. For each set S, the distribution of gene ranks from a gene set is compared against the distribution of the rest of the genes by using ES. Statistical significance is established with respect to a null distribution constructed by 1000 random permutation of genes. We utilised signal-to-noise ratio for ranking genes in this analysis.

2.3.2 Tian's method: Tian's method tests the significance of a gene set by taking the mean of t-values of genes in the gene set as a test statistic and evaluating its significance by a permutation test. This method regards proper adjustments for correlation structure and multiple testing as critical points [6]. The p values are calculated by 1000 random permutation of genes and the false discovery rates (q value) are computed from the p values for only up-regulated pathways ($NT_K > 0$) to make a fair comparison.

2.3.3 Scoring TF target genes' activity: We applied unpaired two-tailed Student t-test to detect differentially activated genes. The test static for kth TF in the original pathway can be written as

$$\mathrm{TF}_k = \frac{1}{\sqrt{M_k}} \sum_{i=1}^{M_k} t_j$$

where M_k represents the number of downstream target genes of TF_k and t_i represents the t-score of ith downstream target gene of TF_k. After calculating each TF_k in the original pathway, the TF target activities of jth pathway (PTF_j) can be obtained by dividing the sum of TF_k by $\sqrt{N_j}$ in each original pathway. N_j represents the number of TFs in the jth original pathway

$$\mathrm{PTF}_j = \frac{1}{\sqrt{N_j}} \sum_{k=1}^{N_j} \mathrm{TF}_k$$

For example, JAK2/STAT5 pathway has STAT5 and STAT3 TFs. First, the t-scores of the target genes of STAT5 TF are added and divided by the square root of the number of STAT5 target genes. The same procedure is done with STAT3. Next, $\mathrm{PTF}_{\mathrm{JAK2/STAT5}}$ can be obtained by dividing the sum of $\mathrm{TF}_{\mathrm{STAT5}}$ and $\mathrm{TF}_{\mathrm{STAT3}}$ by $\sqrt{2}$. The p value is calculated through 1000 random permutation of genes like GSEA and Tian's method. The p value for the TF target genes' activity (PTF_j) is combined with the p value for the original pathway from GSEA and Tian's method through the weighted Z-transform method.

2.3.4 Integrative scoring using the weighted Z-transform method: In order to combine the two p values from different sources, meta analysis that is a set of classical statistical techniques to combine results from several studies was applied. The Z-transform test is one of the meta analysis methods, and can be used to pool p values into a global p value

$$Z_3 = \frac{\sum_{j=1}^{k} Z_j}{\sqrt{k}}$$

The Z-transform test takes advantage of the one-to-one mapping of the standard normal curve to the p value of a one-tailed test. The Z-transform test converts the one-tailed p values, P_i, from each of k independent tests into standard normal deviates Z_i. The Z_s has a standard normal distribution if the common null hypothesis is true [17]

$$Z_3 = \frac{\sum_{j=1}^{k} w_j Z_j}{\sqrt{\sum_{j=1}^{k} w_j^2}}$$

In the weighted Z-method, each test can be assigned a weight, w_i [18, 19]

$$Z_3 = \frac{1 \times Z_1 + 2 \times Z_2}{\sqrt{1 + 2^2}}$$

The Z_1 is from the p value for the original pathways from

doi: 10.1049/iet-syb.2008.0183

GSEA or Tian's method. The Z_2 is from the p value for the TFs target genes' activity (PTF_j) from our target genes' activity scoring method. Thus, the p value for the integrative scoring of each pathway is from each Z_s. Since the effect of transcriptional regulation following the pathway activation by different types of regulation such as phosphorylation can be directly measured from mRNA expression levels of TF target gene, we gave a weight (weight = 2) on the Z_2 from the p value for the TF target genes' activity from our target genes' activity scoring method. The weighted Z-method was only carried out on the pathways that have both the up-regulated original pathway (from GSEA or Tian's method) and the up-regulated target gene set of TFs of that pathway (from our TF target genes' activity scoring). A gene set whose p value is zero was changed to 0.0001 (the lowest p value in all gene sets in three datasets), because the gene set whose p value is zero cannot be converted to Z_i in the weighted Z-method. We computed false discovery rates from the p values for the integrative scoring of each pathway using the q-value method of Pounds and Morris [20].

We also combined the p values for the original pathways from GSEA and Tian's method with the p values for the target genes of TFs from GSEA and Tian's method through the weighted Z-method. Combining p values for the original pathways from GSEA and Tian's method with the p values for TF target genes' activities from our target genes' activity scoring method showed better results with regard to capturing cancer-type-specific pathways (see supplementary figure1). In order to make a fair comparison, we also treated the p values for target genes of TFs from GSEA and Tian's method in integrating p values by meta analysis. Integrating the p values for pathways genes from GSEA and Tian's method with the p values for target genes of TFs of that pathway from GSEA and Tian's method found less cancer-type-specific pathways than combining the p value for pathways genes from GSEA and Tian's with the p value for target genes of TFs of that pathway from our TF target genes' activity scoring in three datasets.

3 Results

3.1 Identification of pathways perturbed by the JAK2V617F mutation in ET patients

ET is a subtype of myeloproliferative disorders (MPD) which also include polycythemia vera and primary myelofibrosis (PMF) characterised by a clonal expansions of a multipotent haematopoietic progenitor cell. Among the three subtypes of MPD, ET is characterised of increasing bone marrow megakaryocytes, and persistent thrombocytosis [21]. Even though the existence of the JAK2V617F mutation has been reported in a high proportion of MPD patients [22], only the 50% of the ET patients have this mutation. In ET patients with the JAK2V617F mutation, the constitutive kinase activity of JAK2 protein causes cytokine-independent activation of JAK/STAT pathway, whereas JAK2V617F negative ET patients do not have constitutively activated JAK/STAT pathway [13]. Fig. 2 shows the identified pathways from GSEA and Tian's method using the original pathways and the extended pathways, and the integrative scoring.

The significant pathways identified by GSEA using the original pathways and the extended pathways are cancer-type-non-specific pathways and pathways related with muscle cell development (MITRPATHWAY and ALKPATHWAY). G2PATHWAY (activated Cdc2-cyclin B kinase regulates the G2/M transition in the pathway), ATRBRCAPATHWAY (BRCA1 and BRCA2 in the pathway block cell cycle progression in response to DNA damage and promote double-stranded break repair) and RBPATHWAY (RB1 plays a major role in cell cycle entry as it functions as a brake in the cell cycle which is released when external signals inform the cell that it can proceed to S phase) are cancer-type-non-specific pathways which are not directly associated with the JAK2V617F perturbation. Two muscle cell associated pathways identified by GSEA using the extended pathways seem not to have any relationship with JAK2V617F positive ET patients. The majority of identified pathways in $JAK2V617F^+$ ET patients by Tian's method using the original pathways and the extended pathways are cancer-type-non-specific pathways like GSEA. In contrast, the integrative scoring using GSEA and our TF target genes' activity scoring finds significant pathways that are directly associated with the aberration of JAK2 proteins (NO2IL12PATHWAY and HSA04630_JAK_STAT_SIGNALING_PATHWAY). The integrative scoring using Tian's method and our TF target genes' activity scoring also provides ET patients with JAK2 mutation-related pathways (IL22BPPATHWAY, HSA04630_JAK_STAT_SIGNALING_PATHWAY and IL10PATHWAY).

3.2 Identification of pathways associated with metastatic potential in primary breast tumours

Distant metastases are the main cause of death among breast cancer patients [23]. However, breast cancer prognostic standards such as clinical and pathological risk factors fail to classify accurately breast tumours, because breast cancer patients with the same stage of disease can have markedly different treatment responses and overall outcome. An ongoing challenge is to identify new prognostic markers that are more directly related to disease and that can more accurately predict the risk of metastasis in individual patients. In the recent years, many research groups have been trying to predict metastasis status using gene expression profiles. Here, we analysed one of the breast cancer expression profiles (Netherlands dataset) [14] to identify which pathways are differentially expressed in metastatic patients.

We divided breast cancer data into metastatic patients and non-metastatic patients. Fig. 3 shows a comparison of differentially expressed pathways of metastatic patients

Original pathways

Pathway name	P-value	q-value	Type
G2PATHWAY	0.0081	0.4888	N
HYPERTROPHY_MODEL	0.0299	1	N
HSA04110_CELL_CYCLE	0.0369	1	N
CELL_CYCLE_KEGG	0.0771	1	
MEF2DPATHWAY	0.0993	1	

Extended pathways

Pathway name	P-value	q-value	Type
ATRBRCAPATHWAY	0.0216	0.9752	N
CELL2CELLPATHWAY	0.0241	1	N
MITRPATHWAY	0.0260	1	X
ALKPATHWAY	0.0343	0.8349	X
RBPATHWAY	0.0381	1	N

a

Original pathways

Pathway name	P-value	q-value	Type
G2PATHWAY	0.0228	0.9988	N
ATRBRCAPATHWAY	0.0396	0.9988	N
VEGFPATHWAY	0.0584	0.9988	
RBPATHWAY	0.0862	0.9988	
HSA04916_MELANOGENESIS	0.0885	0.9988	

Extended pathways

Pathway name	P-value	q-value	Type
HYPERTROPHY_MODEL	0.0388	0.7505	N
IL10PATHWAY	0.045	0.7505	S
ST_G_ALPHA_I_PATHWAY	0.0458	0.7505	N
HBXPATHWAY	0.0522	0.7505	
IL22BPPATHWAY	0.0531	0.7505	

b

Original pathways

Pathway name	P-value	q-value	Type
HYPERTROPHY_MODEL	0.0266	0.0060	N
NO2IL12PATHWAY	0.0280	0.0246	S
HSA04630_JAK_STAT_SIGNALING_PATHWAY	0.0457	0.0546	S
CARM_ERPATHWAY	0.0471	0.0701	X
HSA04916_MELANOGENESIS	0.0513	0.1368	S

Pathways using Tian's method and our TF target genes' activity scoring

Pathway name	P-value	q-value	Type
IL22BPPATHWAY	0.0070	0.0983	S
HSA04630_JAK_STAT_SIGNALING_PATHWAY	0.0250	0.1585	S
HSA04916_MELANOGENESIS	0.0271	0.1632	X
ST_G_ALPHA_I_PATHWAY	0.0316	0.1729	N
IL10PATHWAY	0.0329	0.1754	S

c

Figure 2 *Identified pathways in gene expression profiles of JAK2V617F$^+$ against JAK2V617F$^-$*

a GSEA – Enriched in ET patients with JAK2 mutation
b Tian's method – ET patients with JAK2 mutation VS without JAK2 mutation
c Integrative scoring – ET patients with JAK2 mutation VS without JAK2 mutation
Top five pathways are listed using each method. The pathways marked by green (Type – S) show cancer-type-specific pathways, that is pathways directly related with JAK2 V617F mutation such as JAK/STAT signalling pathway, and the pathways marked by yellow (Type – N) represent cancer-type-non-specific pathways. The integrative scoring method finds the most pathways that are directly associated with the perturbation by JAK2V617F mutation in ET patients

among the identified pathways from GSEA and Tian's method using the original pathways and the extended pathways, and the integrative scoring. GSEA and Tian's approach using the original pathways yield only cancer-type-non-specific pathways such as cell cycle-related pathways, whereas GSEA and Tian's method using the extended pathways discover breast cancer metastatic-related pathways (VEGFPATHWAY). VEGF (vascular endothelial growth factor), a protein is one of the well-known key angiogenesis factors, is released by tumour cells for the generation of new blood vessels to feed the tumour. The tumour cells can be spread to distant organs through these new blood vessels (metastasis) [24]. It is also known that breast cancer metastasis can be suppressed through the inhibition of VEGF-mediated tumour angiogenesis [25]. However, the integrative scoring which combined each GSEA and Tina's method with our TF target genes' activity scoring finds the most cancer-type-specific pathways in this analysis. MTOR is a serine/threonine kinase that has emerged as one of the most important intracellular signalling enzyme regulating cell growth, survival and motility in cancer cells. Furthermore,

doi: 10.1049/iet-syb.2008.0183

Original pathways

Pathway name	P-value	q-value	Type
CELL_CYCLE_KEGG	0	0	N
HSA04110_CELL_CYCLE	0	0	N
G2PATHWAY	0	0.0470	N
UBIQUITIN_MEDIATED_PROTEOLYSIS	0	0.0515	N
G1_TO_S_CELL_CYCLE_REACTOME	0	0.0456	N

Extended pathways

Pathway name	P-value	q-value	Type
CELL2CELLPATHWAY	0	0.0029	N
UBIQUITIN_MEDIATED_PROTEOLYSIS	0	0.0032	N
SA_REG_CASCADE_OF_CYCLIN_EXPR	0	0.0031	N
HSA04110_CELL_CYCLE	0	0.0905	N
VEGFPATHWAY	0.002	0.0712	S

a

Original pathways

Pathway name	P-value	q-value	Type
CELL_CYCLE_KEGG	0	0	N
HSA04110_CELL_CYCLE	0	0	N
G1_TO_S_CELL_CYCLE_REACTOME	0	0.0026	N
UBIQUITIN_MEDIATED_PROTEOLYSIS	0.0005	0.0096	N
MRNA_PROCESSING_REACTOME	0.002	0.0222	N

Extended pathways

Pathway name	P-value	q-value	Type
CELL2CELLPATHWAY	0	0	N
SA_REG_CASCADE_OF_CYCLIN_EXPR	0	0	N
VEGFPATHWAY	0.0002	0.0046	S
UBIQUITIN_MEDIATED_PROTEOLYSIS	0.0011	0.0188	N
MRNA_PROCESSING_REACTOME	0.0014	0.0192	N

b

Pathways using GSEA and our TF target genes' activity scoring

Pathway name	P-value	q-value	Type
VEGFPATHWAY	0	0.0060	S
HSA04150_MTOR_SIGNALING_PATHWAY	0.0005	0.0246	S
HSA05211_RENAL_CELL_CARCINOMA	0.0019	0.0546	X
SA_REG_CASCADE_OF_CYCLIN_EXPR	0.0030	0.0701	N
HSA04720_LONG_TERM_POTENTIATION	0.0101	0.1368	X

Pathways using Tian's method and our TF target genes' activity scoring

Pathway name	P-value	q-value	Type
VEGFPATHWAY	0	0.0042	S
HSA04150_MTOR_SIGNALING_PATHWAY	0.0008	0.0264	S
HSA05211_RENAL_CELL_CARCINOMA	0.0016	0.0387	X
SA_REG_CASCADE_OF_CYCLIN_EXPR	0.0019	0.0430	N
G1_TO_S_CELL_CYCLE_REACTOME	0.0138	0.1397	N

c

Figure 3 *Identified pathways in gene expression profiles of metastatic against non-metastatic primary breast tumours from:*

a GSEA – Enriched in metastatic patients
b Tian's method – Metastatic patients VS non-metastatic patients
c Integrative scoring – Metastatic patients VS non-metastatic patients
Top five pathways are listed using each method. The pathways marked by yellow (Type – N) represent cancer-type-non-specific pathways and the pathways marked by green (Type – S) show cancer-type-specific differentially expressed pathways that are related with breast cancer metastatic potential. Lastly, the pathways marked by red (Type – X) do not have relationship with this disorder. The integrative scoring method discovered more differentially expressed pathways associated with breast cancer metastasis compared to GSEA and Tian's method using the original pathways and the extended pathways

MTOR signalling has been implicated in the development of metastasis in breast cancer. HER2 (ERbB2), a member of the epidermal growth factor receptor, plays a pivotal role in promoting metastasis in breast cancer by enhancing CXCR4 expression through MTOR-mediated pathways [26]. HSA05211_RENAL_CELL_CARCINOMA and HSA04720_LONG_TERM_POTENTIATION identified by the integrative scoring seem not to have any relationships with breast cancer metastatic patients.

3.3 Identification of pathways associated with bad prognosis in primary lung tumours

To test the robustness of our approach, we reanalysed the lung cancer data that had been previously analysed by GSEA. The aim of our approach, like that of GSEA, is not only to find differentially expressed tumour-specific pathways but also to provide more consistent results than are obtained with single-

Original pathways

Pathway name	Type
BREAST_CANCER_ESTROGEN_SIGNALING	S
VEGFPATHWAY	S
HSA05219_BLADDER_CANCER	X

Extended pathways

Pathway name	Type
VEGFPATHWAY	S
HIFPATHWAY	S
HSA04510_FOCAL_ADHESION	S
EPONFKBPATHWAY	S
P53HYPOXIAPATHWAY	N
HSA05131_PATHOGENIC_ESCHERICHIA_COLI_INFECTION_EPEC	X
HSA05211_RENAL_CELL_CARCINOMA	X
HSA05130_PATHOGENIC_ESCHERICHIA_COLI_INFECTION_EHEC	X

a

Original pathways

Pathway name	Type
BREAST_CANCER_ESTROGEN_SIGNALING	S
VEGFPATHWAY	S
HSA04510_FOCAL_ADHESION	S
HSA04110_CELL_CYCLE	N
HSA04115_P53_SIGNALING_PATHWAY	N

Extended pathways

Pathway name	Type
HSA04510_FOCAL_ADHESION	S
VEGFPATHWAY	S
HIFPATHWAY	S
EPONFKBPATHWAY	S
CELL2CELLPATHWAY	N
HSA05130_PATHOGENIC_ESCHERICHIA_COLI_INFECTION_EHEC	X
HSA05131_PATHOGENIC_ESCHERICHIA_COLI_INFECTION_EPEC	X
HSA05211_RENAL_CELL_CARCINOMA	X

b

Pathways using GSEA and our TF target genes' activity scoring

Pathway name	Type
VEGFPATHWAY	S
HSA04510_FOCAL_ADHESION	S
BREAST_CANCER_ESTROGEN_SIGNALING	S
HSA04530_TIGHT_JUNCTION	S
HIFPATHWAY	S
EPONFKBPATHWAY	S
P53HYPOXIAPATHWAY	N
CELL_CYCLE_KEGG	N
HSA05130_PATHOGENIC_ESCHERICHIA_COLI_INFECTION_EHEC	X
HSA05131_PATHOGENIC_ESCHERICHIA_COLI_INFECTION_EPEC	X
HSA05211_RENAL_CELL_CARCINOMA	X

Pathways using Tian's method and our TF target genes' activity scoring

Pathway name	Type
HSA04510_FOCAL_ADHESION	S
BREAST_CANCER_ESTROGEN_SIGNALING	S
VEGFPATHWAY	S
HIFPATHWAY	S
HSA04530_TIGHT_JUNCTION	S
EPONFKBPATHWAY	S
HSA04110_CELL_CYCLE	N
P53HYPOXIAPATHWAY	N
CELL_CYCLE_KEGG	N
HSA05130_PATHOGENIC_ESCHERICHIA_COLI_INFECTION_EHEC	X
HSA05131_PATHOGENIC_ESCHERICHIA_COLI_INFECTION_EPEC	X

c

Figure 4 *Overlapping pathways from two independent gene expression studies of lung cancer patients with bad against good prognosis in:*

a GSEA – Enriched in poor outcome
b Tian's method – Poor outcome VS good outcome
c Integrative scoring – Poor outcome VS good outcome
The overlapping pathways among 15 top-ranked pathways with p value < 0.05 from Michigan and Boston lung cancer studies are listed. The pathways marked by yellow (Type – N) represent cancer-type-non-specific pathways and the pathways marked by green (Type – S) show cancer-type-specific differentially expressed pathways that are related with the poor outcome of lung cancer. Lastly, the pathways marked by red (Type – X) do not have relationship with this disorder. The integrative scoring method provided the best consistent results with respect to finding pathways related to the poor outcome of lung cancer

IET Syst. Biol., 2009, Vol. 3, Iss. 6, pp. 534–542

doi: 10.1049/iet-syb.2008.0183

gene analysis. GSEA reanalysed data from two studies of lung cancer from the Boston group and the Michigan group. For these two lung cancer datasets, each dataset has two different classes of samples. One class has poor outcome patients and the other class has good outcome patients. Even though GSEA found overlapping pathways in the two datasets, the results by GSEA were cancer-type-non-specific pathways, including cell cycle-related pathways and p53-related pathways [5].

Fig. 4 shows a comparison of commonly predicted differentially expressed pathways in both datasets among GSEA and Tian's method using the original pathways and the extended pathways, and the integrative scoring. The integrative scoring method provides robust results, because both GSEA and Tian's method using the pathways with our integrative scoring capture the most lung cancer with poor prognosis associated pathways. Estrogen signalling has been known to promote cell proliferation and suppresses apoptosis, and its role in the late steps of lung metastasis has recently been suggested [27]. VEGF and HIF-1α (hypoxia inducible factor 1) are well-known inducers of angiogenesis. Up-regulation of the angiogenic factor VEGF is crucial in lung cancer metastasis and HIF-1α overexpression is a common event in lung cancer which is related to the up-regulation of the angiogenic factor VEGF [28]. The level of HIF-1α expression has been shown to correspond with tumourigenesis and angiogenesis by activating the expression of VEGF at the transcriptional level. Erythropoietin (Epo) is well documented targets of HIF-1α, and Epo produced by HIF-1α stimulates JAK2 phosphorylation of I-kB, releasing NF-kB to translocate into the nucleus and activate transcription of several genes in EPONFKBPATHWAY [29]. Furthermore, raised Epo is known to associate with reduced survival in lung cancer patients [29, 30]. Their results suggested that an elevated Epo is significant not only in long-term prognosis but also in determining the subsequent resectability of the tumour. Focal adhesion kinase (FAK) is a non-receptor tyrosine kinase linked to the integrin and growth factor receptor-signalling pathways that regulate a number of the biological processes involved in neoplastic transformation, invasion and metastases, such as cell adhesion, migration and apoptosis. Up-regulation of FAK plays a role in the tumourigenesis of invasive lung cancer [31]. In addition, tight junction plays a crucial role in lung cancer invasion and metastasis. Especially, claudin (CLDN) genes that encode a family of proteins important in tight junction formation and function are elevated in lung cancer [32].

4 Conclusion

We have demonstrated that integrating the activities of pathway member genes and those of the transcriptional target genes in each pathway can identify cancer-type-specific pathways better than GSEA and Tian's method using the original pathways and extended pathways in gene expression profiles of ET patients with JAK2V617F mutation status, primary breast tumour samples with the status of metastasis occurrence, two independent lung cancer expression profiles with their prognosis. This might implicate that adding information of transcriptional regulation is better way of utilising mRNA measurements for estimating the pathway activity more exactly though many biological processes are dependent on other types of regulation such as phosphorylation besides transcriptional regulation. Thus, better coverage and quality of human pathway information and more precise identification of TF target genes will enable further identification of pathways specifically associated with various disease phenotypes through gene expression studies. More reasonable selection of weights in the used the weighted Z-method remains as further work.

5 Acknowledgments

This work was supported by National Research Lab. Program (No.2006-01508) from the Ministry of Education, Science and Technology through the Korea Science and Engineering Foundation. We would like to thank CHUNG Moon Soul Center for BioInformation and BioElectronics for providing research facilities.

6 References

[1] KHATRI P., DRAGHICI S.: 'Ontological analysis of gene expression data: current tools, limitations, and open problems', *Bioinformatics*, 2005, **21**, (18), pp. 3587–3595

[2] HUANG DA W., SHERMAN B.T., LEMPICKI R.A.: 'Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists', *Nucleic Acids Res.*, 2009, **37**, (1), pp. 1–13

[3] KHATRI P., DRAGHICI S., OSTERMEIER G.C., *ET AL.*: 'Profiling gene expression using onto-express', *Genomics*, 2002, **79**, (2), pp. 266–270

[4] NAM D., KIM S.Y.: 'Gene-set approach for expression pattern analysis', *Briefings Bioinf.*, 2008, **9**, (3), pp. 189–197

[5] SUBRAMANIAN A., TAMAYO P., MOOTHA V.K., *ET AL.*: 'Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles', *Proc. Natl. Acad. Sci. USA*, 2005, **102**, (43), pp. 15545–15550

[6] TIAN L., GREENBERG S.A., KONG S.W., *ET AL.*: 'Discovering statistically significant pathways in expression profiling studies', *Proc. Natl. Acad. Sci. USA*, 2005, **102**, (38), pp. 13544–13549

[7] DINU I., POTTER J.D., MUELLER T., *ET AL.*: 'Improving gene set analysis of microarray data by SAM-GS', *Bmc Bioinf.*, 2007, **8**, article no. 242

[8] EFRON B., TIBSHIRANI R.: 'On testing the significance of sets of genes', *J. Comput. Theor. Nanosci.*, 2007, **1**, (1), pp. 107–129

[9] DRAGHICI S., KHATRI P., TARCA A.L., *ET AL.*: 'A systems biology approach for pathway level analysis', *Genome Res.*, 2007, **17**, (10), pp. 1537–1545

[10] MESSINA D.N., GLASSCOCK J., GISH W., *ET AL.*: 'An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression', *Genome Res.*, 2004, **14**, (10B), pp. 2041–2047

[11] WINGENDER E., CHEN X., FRICKE E., *ET AL.*: 'The TRANSFAC system on gene expression regulation', *Nucleic Acids Res.*, 2001, **29**, (1), pp. 281–283

[12] RYU T., JUNG J., LEE S., *ET AL.*: 'bZIPDB: A database of regulatory information for human bZIP transcription factors', *Bmc Genomics*, 2007, **8**, article no. 136

[13] SCHWEMMERS S.H., PAHL H., WILL B., *ET AL.*: 'JAK2V617F-negative ET patients do not display constitutively active Jak/STAT signalling', *Haematol. Hematol. J.*, 2007, **92**, pp. 152–152

[14] VAN DE VIJVER M.J., HE Y.D., VAN'T VEER L.J., *ET AL.*: 'A gene-expression signature as a predictor of survival in breast cancer', *New Engl. J. Med.*, 2002, **347**, (25), pp. 1999–2009

[15] BEER D.G., KARDIA S.L.R., HUANG C.C., *ET AL.*: 'Gene-expression profiles predict survival of patients with lung adenocarcinoma', *Nat. Med.*, 2002, **8**, (8), pp. 816–824

[16] BHATTACHARJEE A., RICHARDS W.G., STAUNTON J., *ET AL.*: 'Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses', *Proc. Natl. Acad. Sci. USA*, 2001, **98**, (24), pp. 13790–13795

[17] WHITLOCK M.C.: 'Combining probability from independent tests: the weighted *Z*-method is superior to Fisher's approach', *J. Evol. Biol.*, 2005, **18**, (5), pp. 1368–1373

[18] MOSTELLER F., BUSH R.R.: 'Selected quantitative techniques' in LINDZEY G. (ED.): 'Handbook of Social Psychology' (Addison-Wesley, Cambridge, MA, 1954, vol. 1), pp. 289–334

[19] LIPTAK T.: 'On the combination of independent tests. Magyar Tud', *Akad. Mat. Kutato Int. Kozl.*, 1958, **3**, pp. 171–197

[20] POUNDS S., MORRIS S.W.: 'Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of *p*-values', *Bioinformatics*, 2003, **19**, (10), pp. 1236–1242

[21] KAUSHANSKY K.: 'Thrombopoietin – the primary regulator of platelet production', *Trends Endocrinol. Metab.*, 1997, **8**, (2), pp. 45–50

[22] JAMES C., UGO V., LE COUEDIC J.P., *ET AL.*: 'A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera', *Nature*, 2005, **434**, (7037), pp. 1144–1148

[23] WEIGELT B., PETERSE J.L., VAN'T VEER L.J.: 'Breast cancer metastasis: markers and models', *Nat. Rev. Cancer*, 2005, **5**, (8), pp. 591–602

[24] LEUNG D.W., CACHIANES G., KUANG W.J., *ET AL.*: 'Vascular endothelial growth factor is a secreted angiogenic mitogen', *Science*, 1989, **246**, (4935), pp. 1306–1309

[25] ZHANG J., LU A., BEECH D., *ET AL.*: 'Suppression of breast cancer metastasis through the inhibition of VEGF-mediated tumor angiogenesis', *Cancer Ther.*, 2007, **5**, pp. 273–286

[26] BENOVIC J.L., MARCHESE A.: 'A new key in breast cancer metastasis', *Cancer Cell*, 2004, **6**, (5), pp. 429–430

[27] BANKA C.L., LUND C.V., NGUYEN M.T.N., *ET AL.*: 'Estrogen induces lung metastasis through a host compartment-specific response', *Cancer Res.*, 2006, **66**, (7), pp. 3667–3672

[28] LIU L.Z., FANG J., ZHOU Q., *ET AL.*: 'Apigenin inhibits expression of vascular endothelial growth factor and angiogenesis in human lung cancer cells: implication of chemoprevention of lung cancer', *Mol. Pharmacol.*, 2005, **68**, (3), pp. 635–643

[29] PAUL I., LAPPIN T.R.J., MAXWELL P., *ET AL.*: 'Pre-operative plasma erythropoietin concentration and survival following surgery for non-small cell lung cancer', *Lung Cancer*, 2006, **51**, (3), pp. 329–334

[30] DAGNON K., PACARY E., COMMO F., *ET AL.*: 'Expression of erythropoietin and erythropoietin receptor in non-small cell lung carcinomas', *Clin. Cancer Res.*, 2005, **11**, (3), pp. 993–999

[31] CARELLI S., ZADRA G., VAIRA V., *ET AL.*: 'Up-regulation of focal adhesion kinase in non-small cell lung cancer', *Lung Cancer*, 2006, **53**, (3), pp. 263–271

[32] HEWITT K.J., AGARWAL R., MORIN P.J.: 'The claudin gene family: expression in normal and neoplastic tissues', *Bmc Cancer*, 2006, **6**, article no. 186

doi: 10.1049/iet-syb.2008.0183