Individualized identification of disease-associated pathways with disrupted coordination of gene expression

Hongwei Wang, Hao Cai, Lu Ao, Haidan Yan, Wenyuan Zhao, Lishuang Qi, Yunyan Gu and Zheng Guo

Corresponding author. Zheng Guo, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150086, China; Key Laboratory of Ministry of Education for Gastrointestinal Cancer, Department of Bioinformatics, Fujian Medical University, Fuzhou 350001, China. Tel.: +86451-86620941-116; E-mail: guoz@ems.hrbmu.edu.cn

Abstract

Current pathway analysis approaches are primarily dedicated to capturing deregulated pathways at the population level and cannot provide patient-specific pathway deregulation information. In this article, the authors present a simple approach, called individPath, to detect pathways with significantly disrupted intra-pathway relative expression orderings for each disease sample compared with the stable, normal intra-pathway relative expression orderings pre-determined in previously accumulated normal samples. Through the analysis of multiple microarray data sets for lung and breast cancer, the authors demonstrate individPath’s effectiveness for detecting cancer-associated pathways with disrupted relative expression orderings at the individual level and dissecting the heterogeneity of pathway deregulation among different patients. The portable use of this simple approach in clinical contexts is exemplified by the identification of prognostic intra-pathway gene pair signatures to predict overall survival of resected early-stage lung adenocarcinoma patients and signatures to predict relapse-free survival of estrogen receptor-positive breast cancer patients after tamoxifen treatment.

Key words: relative expression orderings; individualized pathway; stable/reversal gene pair

Introduction

Pathway analysis of high-throughput data can help bridge the gap from individual genes to biological pathways and decipher the molecular mechanisms underlying disease phenomena. To reveal pathways deregulated in a particular disease state, a plethora of pathway analysis tools have been developed, generally classified into three broad categories, including over-representation analysis (ORA), functional class scoring (FCS) and pathway topology (PT)-based analysis [1]. ORA approaches, such as DAVID [2], typically use only the most significant genes and discard others, potentially resulting in insufficient power to detect subtle changes in the expression levels.
of multiple genes within a pathway. Although FCS approaches, such as GSEA [8], obviate the need for differentially expressed gene (DEG) selection, they are likely not sensitive enough to detect drastic changes in the expression levels of several genes within a pathway. Unlike ORA and FCS approaches, PT-based approaches, such as SPIA [4], incorporate additional topology information into the process of detecting deregulated pathways but tend to suffer from either incomplete or inaccurate annotations.

Because many complex diseases, such as cancer, are genetically heterogeneous, dissecting this heterogeneity will be crucial for improving our understanding of the mechanisms underlying the diseases and for developing personalized therapies targeting specific pathways, ultimately paving the way for personalized medicine. Although personalized analysis has practical significance and urgent demand, few of the aforementioned approaches can provide patient-specific information on pathway deregulations. Until recently, several tools, such as PARADIGM [5], Pathifier [6] and iPAS [7], have been developed to characterize deregulated pathways for individual patients. PARADIGM utilizes multi-dimensional data to infer patient-specific pathway activities. Pathifier calculates a deregulation score separately for each pathway in a sample by measuring the deviation of the sample from normal behavior. iPAS takes accumulated normal tissue data as a reference to identify deregulated pathways in individual patients. Nevertheless, the Achilles’ heel of these approaches is their sensitivity to technical artifacts, particularly experimental batch effects. Experimental batch effects can have a widespread and critical effect on high-throughput measurements and downstream analyses, introducing systematic biases and resulting in misleading biological or clinical conclusions [8]. Although various normalization methods have been proposed to remove batch effects, the existing methods are usually ineffective [9]. Particularly when merging data produced by different laboratories, these methods may exacerbate technical artifacts and even distort biological signals [10]. Subsequently, these effects can cause significant problems in translating research findings into clinical practice. For example, predictors built with methods using confounding data tend to produce more variable predictions of clinical outcomes [11], hindering the rate of translation of new signatures into clinical practice. Therefore, it is necessary to find an efficient solution to solve the above problems.

One potential solution would be to make use of the relative orderings of gene expression, considering that the relative expression orderings (REOs) of genes within samples are insensitive to between-chip normalization and robust to systematic batch effects [12–14]. Indeed, we have demonstrated these advantages in our previous work on individualized differential gene expression analysis based on the relative orderings of gene expression [15]. Additionally, our results based on both simulated and real data demonstrated that the amount of information residing in gene ranking profiles was sufficient to reveal differential gene expression at the individual level.

Here, taking advantage of the relative orderings of gene expression, we propose a simple approach, individPath, to identify deregulated pathways for individual disease samples by comparing the REOs of intra-pathway gene pairs between each disease sample and the accumulated normal samples. We demonstrated our approach’s effectiveness for personalized pathway analysis through the analysis of lung adenocarcinoma and estrogen receptor-positive (ER+) breast cancer gene expression profiles. Then, we applied individPath to identify prognostic pathway signatures for lung adenocarcinoma and ER+ breast cancer and discovered heterogeneous deregulation of multiple pathways that were significantly associated with overall survival or relapse-free survival in cancer patients. Specifically, we identified an intra-pathway gene pair signature that can help predict overall survival of resected early-stage lung adenocarcinoma patients and an intra-pathway gene pair signature that can help predict relapse-free survival of ER+ breast cancer patients after tamoxifen treatment.

### Material and methods

#### Data and pre-processing

Multiple gene expression data sets for lung and breast tissues were collected from the GEO database [16] as described in detail in Table 1. Notably, samples of normal lung or breast tissue were collected from data sets for studying different disorders on these tissues. Cancer samples with survival information were used for survival analysis. For overall survival assessment, the 204 resected early-stage lung adenocarcinoma patients without adjuvant chemotherapy from the GSE31210 data set were used as a discovery cohort. The 254 resected early-stage lung adenocarcinoma patients without adjuvant chemotherapy from a pooled data set of GSE50081, GSE29013, GSE30219 and GSE37745 were used as a validation cohort. For relapse-free survival assessment, the 298 ER+ breast cancer patients with tamoxifen treatment from the GSE17705 data set were used as a discovery cohort. The 378 ER+ breast cancer patients with tamoxifen treatment from a pooled data set of GSE4922, GSE12093 and GSE6532 were used as a validation cohort. More detailed clinical information on these cohorts can be found in Table 2.

For each data set, the raw data (.CEL files) were processed using the Robust Multichip Analysis (RMA) algorithm [17] for background adjustment and median polish summarization.

#### Table 1. Description of normal tissue data and cancer survival data used in this study

<table>
<thead>
<tr>
<th>Data type</th>
<th>GEO acc</th>
<th>#Samples</th>
<th>Platform</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal lung tissue data</td>
<td>GSE18842</td>
<td>45</td>
<td>GPL570</td>
<td>Sanchez-Palencia et al., 2011</td>
</tr>
<tr>
<td></td>
<td>GSE19188</td>
<td>65</td>
<td>GPL570</td>
<td>Hou et al., 2010</td>
</tr>
<tr>
<td></td>
<td>GSE19804</td>
<td>60</td>
<td>GPL570</td>
<td>Lu et al., 2010</td>
</tr>
<tr>
<td></td>
<td>GSE27262</td>
<td>25</td>
<td>GPL570</td>
<td>Wei et al., 2010</td>
</tr>
<tr>
<td></td>
<td>GSE31210</td>
<td>20</td>
<td>GPL570</td>
<td>Okayama et al., 2012</td>
</tr>
<tr>
<td></td>
<td>GSE37768</td>
<td>20</td>
<td>GPL570</td>
<td></td>
</tr>
<tr>
<td>Normal breast tissue data</td>
<td>GSE9574</td>
<td>29</td>
<td>GPL570</td>
<td>Tripathi et al., 2008</td>
</tr>
<tr>
<td></td>
<td>GSE15852</td>
<td>43</td>
<td>GPL570</td>
<td>Pau Ni et al., 2010</td>
</tr>
<tr>
<td></td>
<td>GSE16873</td>
<td>12</td>
<td>GPL570</td>
<td>Emery et al., 2009</td>
</tr>
<tr>
<td></td>
<td>GSE20437</td>
<td>42</td>
<td>GPL570</td>
<td>Graham et al., 2010</td>
</tr>
<tr>
<td></td>
<td>GSE21947</td>
<td>30</td>
<td>GPL570</td>
<td>Graham et al., 2011</td>
</tr>
<tr>
<td>Lung cancer survival data</td>
<td>GSE29013</td>
<td>8</td>
<td>GPL570</td>
<td>Xie et al., 2011</td>
</tr>
<tr>
<td></td>
<td>GSE30219</td>
<td>84</td>
<td>GPL570</td>
<td>Rousseaux et al., 2011</td>
</tr>
<tr>
<td></td>
<td>GSE31210</td>
<td>204</td>
<td>GPL570</td>
<td>Okayama et al., 2012</td>
</tr>
<tr>
<td></td>
<td>GSE37745</td>
<td>35</td>
<td>GPL570</td>
<td>Botling et al., 2013</td>
</tr>
<tr>
<td></td>
<td>GSE50081</td>
<td>127</td>
<td>GPL570</td>
<td>Der et al., 2014</td>
</tr>
<tr>
<td>Breast cancer survival data</td>
<td>GSE4922</td>
<td>66</td>
<td>GPL570</td>
<td>Ishihara et al., 2006</td>
</tr>
<tr>
<td></td>
<td>GSE6532</td>
<td>176</td>
<td>GPL570</td>
<td>Lai et al., 2008</td>
</tr>
<tr>
<td></td>
<td>GSE12093</td>
<td>136</td>
<td>GPL570</td>
<td>Zhang et al., 2009</td>
</tr>
<tr>
<td></td>
<td>GSE17705</td>
<td>298</td>
<td>GPL570</td>
<td>Symmans et al., 2010</td>
</tr>
</tbody>
</table>

Note: a and b denote the data sets used as training sets and test sets in survival analysis, respectively. p and w denote that partial samples and whole samples were used in this analysis, respectively.
without quantile normalization. Then, each probe-set ID was mapped to an Entrez gene ID with the custom CDF file. If multiple probe-sets were mapped to the same gene, the expression value for the gene was summarized as the arithmetic mean of the values of the multiple probe-sets (on the log2 scale).

Molecular Signatures Database

The 1320 canonical pathways in the C2 collection were downloaded from the Molecular Signatures Database (MSigDB; Version 4.0, updated 31 May 2013) [3], which were curated from several online pathway databases, including BioCarta, KEGG [18], Reactome [19] and others [20, 21]. These pathways were compiled by domain experts as canonical representations of biological processes and covered 8428 unique genes that were used for subsequent analyses.

Individualized Pathway Coordination Analysis (individPath)

As illustrated in Figure 1, the individPath algorithm begins by overlaying expression data onto pathway diagrams. Each pathway is examined by pairwise comparison between intra-pathway genes to identify gene pairs with stable ordering in normal samples, defined as stable gene pairs according to the criteria \( P_{\text{norm}}(G_i > G_j) > 0.99 \) or \( P_{\text{norm}}(G_i < G_j) > 0.99 \). Here, \( P_{\text{norm}}(G_i > G_j) \) (or \( P_{\text{norm}}(G_i < G_j) \)), representing the frequency of normal samples for which the expression intensity of gene \( i \) (\( G_i \)) is greater (or less) than that of gene \( j \) (\( G_j \)), is estimated as follows:

\[
P_{\text{norm}}(G_i > G_j) = \frac{1}{n_1} \sum_{t=1}^{n_1} I[G_{it} > G_{jt}]
\]

Or

\[
P_{\text{norm}}(G_i < G_j) = \frac{1}{n_1} \sum_{t=1}^{n_1} I[G_{it} < G_{jt}]
\]

in which \( n_1 \) denotes the total number of normal samples and \( I \) denotes an indicator function that equals one if the event inside the square brackets is true and zero otherwise.

Taking the pre-defined gene pairs with stable ordering in normal samples as the background, we directly determine the gene pairs with reversal ordering in each disease sample (i.e., \( G_i > G_j \) \( \rightarrow \) \( G_i < G_j \) or \( G_i < G_j \) \( \rightarrow \) \( G_i > G_j \)) and, subsequently, the biological pathways with significantly disrupted ordering of gene expression. The significantly deregulated pathways for each disease sample can be naturally determined by testing whether the frequency of reversal gene pairs observed within each pathway is significantly greater than that expected by chance using the hypergeometric distribution model as follows:

\[
P = 1 - \sum_{i=0}^{k-1} \frac{C_m}{C_i} \frac{C_n}{C_{N-m}}\]

in which \( N \) denotes the total number of all stable intra-pathway gene pairs in normal samples, \( n \) denotes the total number of all reversal intra-pathway gene pairs in a certain disease sample and \( k \) denotes the number of observed reversal gene pairs in a pathway with \( m \) stable gene pairs.

The \( P \)-values calculated for each pathway in individual disease samples are adjusted using the Benjamini–Hochberg (BH) method [22] to control the false discovery rate (FDR).

Identification of disease-relevant pathways

Because disease-irrelevant conditions could contribute to the deregulation of pathways in a disease sample, the deregulated pathways detected for individual patients are further filtered to ensure association with the disease itself. For each pathway, the significance of the association between the deregulation and the disease is determined by a binomial distribution test as follows:

\[
P = 1 - \sum_{i=0}^{n_2} \frac{C_i}{C_{n_2}} (1 - p_0)^{n_2-i} \]

in which \( n_2 \) denotes the total number of disease samples, \( x \) denotes the number of deregulated samples for each pathway and \( p_0 \) is the probability of observing a pathway being significantly deregulated in a disease sample by chance, calculated as the average of the frequencies of deregulated pathways among all investigated pathways for individual disease samples.

The \( P \)-values are adjusted using the BH method to control the FDR.

Table 2. Clinical characteristics of patients with lung adenocarcinoma and ER+ breast cancer

<table>
<thead>
<tr>
<th>Cancer type</th>
<th>Lung adenocarcinoma</th>
<th>ER+ breast cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort</td>
<td>Training</td>
<td>Validation</td>
</tr>
<tr>
<td>GSE31210</td>
<td>GSE17705</td>
<td></td>
</tr>
<tr>
<td>Total patients</td>
<td>( n = 204 )</td>
<td>( n = 254 )</td>
</tr>
<tr>
<td>Average age (years)</td>
<td>60</td>
<td>66</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>109 (52.45%)</td>
<td>104 (40.94%)</td>
</tr>
<tr>
<td>Male</td>
<td>95 (47.55%)</td>
<td>150 (59.06%)</td>
</tr>
<tr>
<td>Follow-up time</td>
<td>Median (mos)</td>
<td>60</td>
</tr>
<tr>
<td>Event</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>30 (14.71%)</td>
<td>121 (47.64%)</td>
</tr>
<tr>
<td>No</td>
<td>174 (85.29%)</td>
<td>133 (52.36%)</td>
</tr>
<tr>
<td>Smoking status</td>
<td>Smoked</td>
<td>99 (48.53%)</td>
</tr>
<tr>
<td>Never</td>
<td>105 (51.47%)</td>
<td>23 (9.06%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>131 (51.57%)</td>
<td></td>
</tr>
</tbody>
</table>

Note. Event represents death even for lung adenocarcinoma and relapse-free survival event for ER+ breast cancer, respectively.
Survival analysis and signature selection

Overall survival was defined as the time from surgery to death or last follow-up contact. Relapse-free survival was defined as the time from initial treatment to relapse or death. To avoid bias in the durations of patient follow-up, lung adenocarcinoma patients with more than 120 months of follow-up were truncated at 120 months and breast cancer patients with more than 150 months of follow-up were truncated at 150 months. Survival curves were estimated by the Kaplan–Meier method [23] and compared with log-rank test [24]. Univariate and multivariate survival analyses were performed using the Cox proportional hazard model [25].

A forward-stepwise algorithm was used to identify the optimal predictors of overall survival or relapse-free survival. Beginning with the intra-pathway gene pair with the highest concordance index (C-index) [26] as the seed signature, candidate intra-pathway gene pairs were added to the signature one at a time until the addition of one did not improve predictive performance. At each step, predictive performance was gauged for all possible additions and evaluated using the C-index to select the optimal addition yielding the largest increase in the C-index value.

Availability

An open-source R statistical analysis package ‘individPath’ is available at http://cran.r-project.org.

Results

Deciphering the texture of REOs in pathways

In a previous work, we demonstrated that REOs are highly stable in many kinds of normal tissues, such as lung tissue and breast tissue, but widely disturbed in their corresponding cancer tissues [15]. Here, we collected a plenty of gene expression data available for lung and breast tissues (Table 1) and used the same strategy to explore the extent of intra-pathway REOs in these two kinds of normal tissues, respectively. To this end, we identified gene pairs with stable REOs for genes in each pathway.

In the 235 pooled normal lung tissue samples, a total of 1,072,753 stable intra-pathway gene pairs were identified according to the criteria described in the Methods section. On average, each pathway contained 1,618 stable intra-pathway gene pairs. To evaluate the reproducibility of the stable gene pairs, we randomly divided the normal samples into two similar groups, one of 117 and one of 118, and identified stable gene pairs separately in each group. The results demonstrated that the two lists of stable intra-pathway gene pairs detected from the two groups significantly overlapped and all overlapping pairs had identical ordering patterns (Table 3), which was highly unlikely to occur by chance (P < 1.0e-16, binomial distribution test). Similar results were also observed for the pooled normal breast tissue samples (Table 3 and Supplementary Table S1).

Taken together, the above results demonstrated that intra-pathway REOs are highly stable in normal tissue, which might...
Capturing patient-specific deregulated pathways
To demonstrate how individPath can help reveal the inherent heterogeneity of pathway deregulation, we applied it to analyze multiple publicly available microarray data sets of lung adenocarcinoma and ER+ breast cancer extracted from the GEO database (Table 1).

Based on the corresponding stable intra-pathway gene pairs pre-determined from the pooled normal lung tissue samples, we identified biological pathways that were significantly deregulated in each cancer sample from the GSE31210 lung adenocarcinoma data set at an FDR of 5%. Some pathways were significantly deregulated in a large fraction of cancer samples (Supplementary Table S2). For example, the ‘KEGG Focal Adhesion Pathway’ was significantly deregulated in 90% of lung adenocarcinoma samples, consistent with the disruption frequently observed in response to various stimuli [27]. Conversely, many other pathways were significantly deregulated in only a small fraction of cancer samples (Supplementary Table S2). Even after defining cancer-relevant pathways that were deregulated in significantly more cancer patients than expected by chance (FDR < 0.05, see the Methods section), many pathways remained to be highly heterogeneous (Supplementary Figure S1 and Table S2). Specifically, 68% of lung adenocarcinoma samples displayed no significant change in ‘KEGG Metabolism of Xenobiotics by Cytochrome p450’, and 62% of lung adenocarcinoma samples displayed no significant change in ‘KEGG Drug Metabolism Cytochrome p450’. Heterogeneous behaviors such as these have been shown to be associated with either cancer susceptibility in different patients or cancer cell sensitivity to therapies [28]. Similar results were also observed in other independent lung adenocarcinoma data sets (GSE29013, GSE30219, GSE37745 and GSE50081; Supplementary Table S2).

By using the similar strategy, the heterogeneous behavior of deregulation of pathways was also observed in the GSE17705 ER+ breast cancer data set (Supplementary Figure S2 and Table S2) and validated in other independent ER+ breast cancer data sets (GSE4922, GSE6532 and GSE12093; Supplementary Table S2). For example, the ‘BioCarta Integrin Pathway’ was significantly deregulated in 29% and 31% of cancer samples separately for the GSE17705 and GSE6532 data sets, and the ‘BioCarta Rho Pathway’ was significantly deregulated in 80% and 75% of cancer samples in each of the above breast cancer data sets, respectively. Overall, each cancer-relevant pathway was significantly deregulated in at least 18% and 20% of breast cancer samples, respectively.

Taken together, these results demonstrated the capacity of individPath to characterize pathway deregulation in individual patients, providing the ability to refine patient classification to improve estimates of prognosis or response to treatment. Obviously, the heterogeneity of pathway deregulation could not be captured by traditional pathway analysis approaches.

Applying individPath to survival analysis
Patients with distinct patterns of pathway deregulation may have underlying differences in survival outcomes. Thus, for each of the cancer-relevant pathways obtained in lung adenocarcinoma and ER+ breast cancer, we explored their prognostic ability by comparing two subgroups of patients with or without a significant change in the pathway.

For early-stage lung adenocarcinoma patients in the GSE31210 data set, we identified 47 pathways that were significantly related to overall survival using the log-rank test (FDR < 0.05; Supplementary Table S3). Notably, the patients with significant changes in each of these pathways resided in the high-risk group with poor overall survival. Considering the clinical applicability, it might be better to obtain a practical signature consisting of a small number of genes with the ability to predict prognosis from these survival-associated pathways. Therefore, we further identified intra-pathway gene pairs from the above survival-associated pathways with reversals that were significantly related to the overall survival of lung adenocarcinoma patients (FDR < 0.05) and used a forward-stepwise algorithm to select a five-gene pair signature consisting of ‘ABL1-BUB1B’, ‘CDK1-POLD4’, ‘E2F4-MAD2L1’, ‘GATA2-KIF4A’ and ‘B3GNT6-B3GNT8’ with an optimized C-index. A detailed description of their REOs in normal lung tissue and patient stratification based on this signature are presented in Supplementary Table S4 and Supplementary methods, respectively. As shown in Figure 2A and B, the patients with at least one gene pair with reversal REOs residing in the high-risk group had significantly poorer overall survival than the other patients (P = 5.62 e-09). The prognostic ability of this signature was further confirmed in an independent pooled cohort of 254 patients (P = 1.08 e-03). Multivariate analysis demonstrated that the designation of high- and low-risk groups remained statistically significant even after adjusting for age, gender and smoking status (Supplementary Table S5).

Similarly, for the ER+ breast cancer data set, we also evaluated the association of cancer-relevant pathways with relapse-free survival of ER+ breast cancer patients. In the GSE17705 data set, we identified two pathways significantly associated with relapse-free survival of breast cancer patients (FDR < 0.05; Supplementary Table S3). The patients with significant changes in each of these two pathways had poor relapse-free survival. To investigate this result further, an optimal three-gene pair signature consisting of ‘ATP6V1C1-JAM2’, ‘CMA1-MMPI’ and ‘EGRF-SRC’ was generated in the training cohort (P = 1.54 e-06; Figure 2C and Supplementary Table S4), and shorter relapse-free survival for patients with reversal REOs in at least two gene pairs was confirmed in an independent pooled cohort of 378 patients (P = 1.91 e-02; Figure 2D).

These findings will enable the identification of high-risk subgroups of patients with poorer overall survival or shorter relapse-free survival, thereby potentially improving their overall survival after curative treatment or tailoring treatment options to achieve maximum therapeutic benefits.

Table 3. Reproducibility analysis of stable intra-pathway gene pairs in normal samples

<table>
<thead>
<tr>
<th>Tissue type</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Overlap²</th>
<th>POC¹</th>
<th>Consistency³</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung</td>
<td>1 119 849</td>
<td>1 080 078</td>
<td>1 044 959</td>
<td>96.75%</td>
<td>100%</td>
</tr>
<tr>
<td>Breast</td>
<td>553 029</td>
<td>540 243</td>
<td>501 448</td>
<td>92.82%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Note: ¹Overlap denotes the number of overlapping stable gene pairs between Group 1 and Group 2.
²POC denotes the percentage of overlap among the list of stable gene pairs from Group 2 (see Supplementary Methods for details).
³Consistency denotes the percentage of overlapping gene pairs that displayed the same ordering patterns across the two groups.

Applying individPath to survival analysis
Patients with distinct patterns of pathway deregulation may have underlying differences in survival outcomes. Thus, for each of the cancer-relevant pathways obtained in lung adenocarcinoma and ER+ breast cancer, we explored their prognostic ability by comparing two subgroups of patients with or without a significant change in the pathway.

For early-stage lung adenocarcinoma patients in the GSE31210 data set, we identified 47 pathways that were significantly related to overall survival using the log-rank test (FDR < 0.05; Supplementary Table S3). Notably, the patients with significant changes in each of these pathways resided in the high-risk group with poor overall survival. Considering the clinical applicability, it might be better to obtain a practical signature consisting of a small number of genes with the ability to predict prognosis from these survival-associated pathways. Therefore, we further identified intra-pathway gene pairs from the above survival-associated pathways with reversals that were significantly related to the overall survival of lung adenocarcinoma patients (FDR < 0.05) and used a forward-stepwise algorithm to select a five-gene pair signature consisting of ‘ABL1-BUB1B’, ‘CDK1-POLD4’, ‘E2F4-MAD2L1’, ‘GATA2-KIF4A’ and ‘B3GNT6-B3GNT8’ with an optimized C-index. A detailed description of their REOs in normal lung tissue and patient stratification based on this signature are presented in Supplementary Table S4 and Supplementary methods, respectively. As shown in Figure 2A and B, the patients with at least one gene pair with reversal REOs residing in the high-risk group had significantly poorer overall survival than the other patients (P = 5.62 e-09). The prognostic ability of this signature was further confirmed in an independent pooled cohort of 254 patients (P = 1.08 e-03). Multivariate analysis demonstrated that the designation of high- and low-risk groups remained statistically significant even after adjusting for age, gender and smoking status (Supplementary Table S5).

Similarly, for the ER+ breast cancer data set, we also evaluated the association of cancer-relevant pathways with relapse-free survival of ER+ breast cancer patients. In the GSE17705 data set, we identified two pathways significantly associated with relapse-free survival of breast cancer patients (FDR < 0.05; Supplementary Table S3). The patients with significant changes in each of these two pathways had poor relapse-free survival. To investigate this result further, an optimal three-gene pair signature consisting of ‘ATP6V1C1-JAM2’, ‘CMA1-MMPI’ and ‘EGRF-SRC’ was generated in the training cohort (P = 1.54 e-06; Figure 2C and Supplementary Table S4), and shorter relapse-free survival for patients with reversal REOs in at least two gene pairs was confirmed in an independent pooled cohort of 378 patients (P = 1.91 e-02; Figure 2D).

These findings will enable the identification of high-risk subgroups of patients with poorer overall survival or shorter relapse-free survival, thereby potentially improving their overall survival after curative treatment or tailoring treatment options to achieve maximum therapeutic benefits.
Comparison with other approaches

We compared individPath with a recently proposed approach, iPAS [7], which consists of five modified pathway analysis technologies, including Average-Z, Fisher, GSEA, Euclidean and Mahalanobis, and with other two approaches allowing for personalized pathway analysis, Barcode [29] and RankComp [15].

First, we evaluated their performance through personalized pathway analysis in two larger microarray data sets separately for lung adenocarcinoma and ER+ breast cancer patients. As shown in Figure 3A, although some findings from these two approaches were consistent, some results from individPath were not detected by other approaches. Then, for the deregulated pathways identified in each data set, we examined their clinical associations with survival of cancer patients, and found that individPath could better capture clinical relevance, whereas iPAS identified few pathways associated with the survival of cancer patients, with the exceptions for Average-Z and Euclidean only for the analysis of lung adenocarcinoma data (Figure 3B). By reducing spurious association driven by batch effects, the second approach, Barcode, showed comparable numbers of deregulated pathways with individPath, but these findings were poorly associated with patient survival (Figure 3B). The third approach, RankComp, detected only drastic changes in gene expression, followed by a sparse pathway deregulation, thereby leading to poor association with patient survival (Figure 3B). Additionally, the results from the five methods for iPAS displayed large fluctuations in the numbers of detected deregulated pathways. For example, when using the Fisher method, hardly any pathways were identified as deregulated, whereas when using the Mahalanobis method, almost all pathways were identified as deregulated. This result was most likely an artifact resulting from not only the use of different methods with different detection sensitivities but also the existence of batch effects in the analyzed data set. To further demonstrate the potential batch effects, we adopted the same

Figure 2. Kaplan–Meier estimates of overall survival and relapse-free survival. (A) and (B) represent overall survival curves for early-stage lung adenocarcinoma patients. (C) and (D) represent relapse-free survival curves for ER+ breast cancer patients. A colour version of this figure is available at Bib online: http://bib.oxfordjournals.org.
batch-removal strategy as was used in iPAS and then performed a clustering analysis for gene expression data of normal tissue samples measured by different laboratories. As shown in Figure 3C, even after quantile normalization and z-score transformation, gene expression profiles of the normal lung tissue samples from six different experiments still clustered perfectly according to the processing groups, indicating strong laboratory-specific effects. A similar phenomenon was also observed in the normal breast tissue samples from different laboratories (Supplementary Figure S3). Obviously, these substantial batch effects among the data sets from different experiments were not eliminated by z-score normalization. Considering that all direct comparisons between genes only occur within individual samples, individPath has the advantages of insensitivity to systematic batch effects.

Next, we further compared the survival predictive ability of biomarkers provided by these several approaches for lung adenocarcinoma and ER+ breast cancer. The detailed description of biomarkers is presented in the Supplementary Result. As shown in Figure 4, the predictions provided only by both individPath and Barcode were successful with verifiable biomarkers separately for each of these two cancers, which were

---

**Figure 3.** Performance comparison of proposed individPath with other existing approaches. (A) Comparison of the results of individPath with iPAS, Barcode and RankComp for characterizing deregulated pathways. For iPAS, after the empirical P-values calculated for each pathway in individual patients were adjusted by the BH method, the pathways with FDR < 0.05 were identified as significantly deregulated. (B) Comparison of the results of individPath with iPAS, Barcode and RankComp for identifying survival-associated pathways. (C) Hierarchical clustering of gene expression data of normal tissue samples from different experiments normalized by the quantile and z-score methods. Different colors denote different data sets of normal samples as described in Table 1. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.
approximately equal in the term of survival predictive ability, whereas the predictions provided by other approaches were only partially successful. Especially, Fisher and Mahalanobis were unsuccessful without obtaining verifiable biomarker for any one of these two cancers.

Taken together, the above results demonstrated that individPath can provide a better indicator of pathways of deregulation by delivering more robust and relevant results. The biomarkers provided by individPath could potentially serve as a useful tool for predicting survival of cancer patients.

Figure 4. Comparison of survival predictive ability of individPath, iPAS, Barcode and RankComp. The lines shown in the upper and lower part of each panel represent high risk and low risk, respectively. Notably, no biomarker was identified by the Fisher and Mahalanobis approach. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.
Discussion

Intra-pathway REOs are highly stable in a particular type of normal tissue, such as lung tissue and breast tissue, suggesting that gene expression in biological pathways could be relatively robust against a range of perturbations so that it can continue to coordinate the execution of normal cellular functions to maintain a normal state. When transitioning into a disease state, intra-pathway REOs might often be subject to extensive perturbations, as often observed in complex diseases, such as cancer, which result in widespread gene expression changes. Based on these observations, we have developed a simple approach to individualize the identification of deregulated pathways using information from disrupted intra-pathway REOs. Notably, our previous work on individualized differential gene expression analysis can also find significantly deregulated pathways for each disease sample by performing enrichment analysis of DEGs. Different from that work, our current work has a completely different philosophy and manages to avoid those problems that arise from the pre-selection of DEGs, as mentioned in the Introduction section. Moreover, although both approaches are based on the notion of stable/reversal gene pairs, the pathways detected by these two approaches are quite different: the DEGs-enrichment analysis approach uses only the most significant genes and discards others and thus tends to find pathways with certain proportions of highly DEGs, the latter approach tends to find pathways with significantly disrupted coordination of gene expression. Therefore, these two types of enrichment analysis approaches would be mutually complementary. Besides this, one key advantage of individPath is its insensitivity to systematic batch effects, which enables us to directly combine data from different sources without the need for between-chip normalization. This advantage is particularly important for researchers who are often limited to scarce tissue samples, in particular normal tissue samples, especially normal tissue samples for some important organs such as heart and brain, because of the invasive nature of sample collection. The use of REO information allows us to maximize the reuse of accumulated data, which will greatly facilitate research on human diseases.

Because gene interactions and their dynamic characteristics, as essential components of pathways, underlie the orchestration of biological processes [30], it is reasonable to perform intra-pathway gene pair comparisons between disease and normal phenotypes. The analyses performed in the lung and breast cancer data highlight the ability of individPath to provide effective insights into the deregulation of biological pathways. Beyond contributing to the understanding of the heterogeneous deregulation of pathways, our approach can have important implications for personalized medicine. As illustrated in the clinical utility section, it may provide patient-specific information to tailor more precise risk stratification. This capability is particularly valuable for survival analysis, in which the traditional risk-scoring methods usually make an artificial decision for risk stratification of patients based on the mean or median expression level of gene signatures [31], whereas individPath allows us to directly utilize their deregulation status to stratify patients. Notably, different from our previous work, we here shift our focus from pathway biomarker into intra-pathway gene pair biomarker. As shown in Figure 2, the predicted gene pair biomarkers in lung adenocarcinoma and breast cancer showed promise in outcome prediction of cancer patients. When following the framework presented in our previous work, for example, we have identified a three-pathway biomarker consisting of ‘KEGG Ribosome’, ‘PID RhoA Reg Pathway’ and ‘Reactome G2/M Checkpoints’ for lung adenocarcinoma (in the training cohort: log-rank P = 0.0020; in the validation cohort: log-rank P = 0.0223), whose predictive ability is roughly equal to that using intra-pathway gene pairs (C-index 0.53 versus 0.57). It should be noted that using pathways will increase the stability of biomarker but will require detection of more genes, whereas using intra-path gene pairs will only require detection of few genes, potentially enhancing its clinical applicability. For instance, so as to perform the survival prediction, an average of 68 genes should be measured when using RankComp, while only 10 genes would be required when using predictions provided by individPath.

Despite having comparative advantages, our approach also has certain limitations. First, individPath may be unsuitable for data from cross-platform experiments. Because different platforms have different experimental designs, such as hybridization temperature and duration, REOs could be sensitive to different platforms. Second, individPath may not be sensitive enough to detect those pathways that contain only a handful of genes. Third, individPath can potentially be subjected to the impact of the size of normal samples. But, the sample size requirement will be highly dependent on the characteristics of tissue type and sample quality. Therefore, to estimate the optimal sample size, we suggest adopting the strategy used in our analysis for lung tissue and breast tissue by reproducibility evaluation to determine the minimum sample size needed for obtaining relatively stable gene pairs. If possible, the use of as many normal samples as possible is still highly recommended. Nonetheless, it should be recognized that individPath represents an important step toward ushering in the era of personalized medicine. Moreover, the ambit of individPath, as a simple and intuitive approach, extends beyond the Affymetrix microarray data setting. individPath will be readily usable with data from other platforms and can also be applied to other types of molecular measurements.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- Intra-pathway REOs are highly stable in normal tissue, which might be a requirement for coordinately executing normal cellular functions.
- individPath can efficiently capture patient-specific deregulation information, with the advantages of insensitivity to systematic batch effects and between-chip normalization.
- Beyond contributing to the understanding of the heterogeneous deregulation of pathways, individPath can also provide the ability to refine patient classification for clinical decision making.

Acknowledgement

The authors would like to sincerely thank Yang Li and Libin Chen for assistance in figure preparation, Beibei Chen for assistance in data processing and Ruiping Wang for critically reading the manuscript.
Funding

This work was supported by the Natural Science Foundation of China [grant numbers 81201822, 81372213] and the Research Fund for the Doctoral Program of Higher Education of China [grant number 20112307110011].

References