Reliable pre-eclampsia pathways based on multiple independent microarray data sets

Kaoru Kawasaki¹, Eiji Kondoh¹,*, Yoshitsugu Chigusa¹, Mari Ujita¹, Ryusuke Murakami¹, Haruta Mogami¹, J.B. Brown², Yasushi Okuno², and Ikuo Konishi¹

¹Department of Gynecology and Obstetrics, Kyoto University, Kyoto, Japan ²Department of Clinical System Onco-Informatics, Kyoto University, Kyoto, Japan

*Corresponding address: Department of Gynecology and Obstetrics, Kyoto University, 54 Shogoin Kawahara-cho, Sakyo-ku, Kyoto 606-8507, Japan. Tel: +81-75-751-3269; Fax: +81-75-761-3967; E-mail: kondo@kuhp.kyoto-u.ac.jp

Submitted on April 7, 2014; resubmitted on July 28, 2014; accepted on October 7, 2014

Abstract: Pre-eclampsia is a multifactorial disorder characterized by heterogeneous clinical manifestations. Gene expression profiling of preeclamptic placenta have provided different and even opposite results, partly due to data compromised by various experimental artefacts. Here we aimed to identify reliable pre-eclampsia-specific pathways using multiple independent microarray data sets. Gene expression data of control and preeclamptic placentas were obtained from Gene Expression Omnibus. Single-sample gene-set enrichment analysis was performed to generate gene-set activation scores of 9707 pathways obtained from the Molecular Signatures Database. Candidate pathways were identified by t-test-based screening using data sets, GSE10588, GSE14722 and GSE25906. Additionally, recursive feature elimination was applied to arrive at a further reduced set of pathways. To assess the validity of the pre-eclampsia pathways, a statistically-validated protocol was executed using five data sets including two independent other validation data sets, GSE30186, GSE44711. Quantitative real-time PCR was performed for genes in a panel of potential pre-eclampsia pathways using placentas of 20 women with normal or severe preeclamptic singleton pregnancies (n = 10, respectively). A panel of ten pathways were found to discriminate women with pre-eclampsia from controls with high accuracy. Among these were pathways not previously associated with pre-eclampsia, such as the GABA receptor pathway, as well as pathways that have already been linked to pre-eclampsia, such as the glutathione and CDKN1C pathways. mRNA expression of GABRA3 (GABA receptor pathway), GCLC and GCLM (glutathione metabolic pathway), and CDKN1C was significantly reduced in the preeclamptic placentas. In conclusion, ten accurate and reliable pre-eclampsia pathways were identified based on multiple independent microarray data sets. A pathway-based classification may be a worthwhile approach to elucidate the pathogenesis of pre-eclampsia.

Key words: microarray / pathway / pre-eclampsia

Introduction

Pre-eclampsia is a major cause of maternal and neonatal mortality and morbidity (Young et al., 2010). Pre-eclampsia is a heterogeneous syndrome in which the pathogenesis can be diverse among women (Young et al., 2010). Although the primary role of the placenta in the pathogenesis of pre-eclampsia is undisputed, its precise mechanism has yet to be fully elucidated. Consequently, the only definitive treatment for pre-eclampsia is delivery of the placenta, and no other effective therapy has been developed despite decades of extensive clinical and basic research. Thus, clearly there is an urgent need for clarification of the pathogenesis of pre-eclampsia.

Gene expression microarray data are a form of high-throughput genomics data for thousands of genes in each sample. Microarray-based gene expression profiling has provided numerous genes and pathways involved in pre-eclampsia (Siras et al., 2009; Winn et al., 2009; Tsai et al., 2011; Louwen et al., 2012). For example, Maynard et al. conducted gene expression profiling of placental tissue from women with and without pre-eclampsia, and found soluble fms-like tyrosine kinase 1 (sFlt1) (Maynard et al., 2003) to be closely related to the pathogenesis of pre-eclampsia. In addition, angiogenesis and immune-response pathways have been shown to be involved in pre-eclampsia in most microarray data sets (Siras et al., 2009; Winn et al., 2009; Tsai et al., 2011; Louwen et al., 2012). However, the genes and pathways derived from microarray analyses are diverse and even occasionally conflicting in existing studies (Winn et al., 2009; Tsai et al., 2011; Louwen et al., 2012). This might be attributed to sample differences in gestational age, modes of delivery, or experimental artefacts such as types of chips and platform.
as well as heterogeneous aetiologies or clinical manifestations. Thus, a single microarray data set may be insufficient to provide meaningful genes and pathways specific to pre-eclampsia. Indeed, more robust sets of genes and pathways have been provided through multiple independent data sets in a wide range of fields such as cancer research (Sorlie et al., 2003; Rhodes et al., 2004). In the last decade, thousands of microarray data sets have appeared in public databases, which allow other researchers to confirm the results of published papers or to permit novel analyses of the data. Nevertheless, few studies (Moslehi et al., 2013) have been conducted with the use of multiple data sets to seek genes and pathways in preeclamptic placentas. We hypothesized that pathways identified based on multiple independent microarray data sets from studies with large sample sizes were more likely to be functionally relevant to the pathogenesis of pre-eclampsia, and could potentially be new therapeutic targets for pre-eclampsia. The aim of our study was to provide pre-eclampsia-specific pathways using the three largest microarray data sets from four different platforms freely available in a web database.

**Materials and Methods**

**Identification of common pathways overlapping three independent data sets in silico**

In order to identify potentially relevant pathways to pre-eclampsia, gene expression data of control and preeclamptic placentas were obtained from the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/gds/) as serial matrix files. The selection criteria for the data sets were the data sets from the three largest sample sizes (sample size: GSE10588, 43; GSE14722, 23; GSE25906, 60) available at GEO DataSets, because the larger sample sizes can yield more reliable results. A summary of the analysed microarray data sets is shown in Table I (Sitras et al., 2009; Winn et al., 2009; Tsai et al., 2011; Meng et al., 2012; Blair et al., 2013). In GSE14722 study, the same samples were assayed on two different versions of the Affymetrix U133 arrays. The HG-U133A Array includes representation of the RefSeq database sequences and probe sets related to sequences previously represented on the Human Genome U95Av2 Array. In contrast, the HG-U133B Array contains primarily probe sets representing expressed sequence tag clusters. Thus, both of the two different versions of the Affymetrix U133 arrays are meant to be complementary and non-overlapping. HG-U133A and HG-U133B data were therefore combined for further analysis. Different types of microarray platforms have shown significant variability when comparing across platforms. Therefore, the three largest data sets from four different platforms were used for subsequent analysis. Single-sample gene-set enrichment analysis (ssGSEA) was performed to generate gene-set activation scores (Barbie et al., 2009). The ssGSEA script was obtained from GenePattern (http://www.broadinstitute.org/cancer/software/genepattern). According to the instructions described in ssGSEAProjection Documentation, v4 (http://www.broadinstitute.org/cancer/software/genepattern/modules/docs/ssGSEAProjection/4/), GCT files containing the gene expression data were created as input files. Gene sets (8513 pathways) were downloaded from the Molecular Signatures Database v3.1 (http://www.broadinstitute.org/gsea/downloads.jsp), and a 'msigdb.v3.1.symbols.gmt' file, that consisted of all gene set collections named c1, c2, c3, c4, c5 and c6, was used for ssGSEA. We added sets that combined up- and down-regulated sets derived from the same experimental condition or publication (option provided by ssGSEA package). The final total was therefore 9707 pathways. Pathway activation scores in each sample were calculated using

---

**Table I** Summary of analysed microarray data sets.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Platform</th>
<th>Number of probes</th>
<th>Sample size</th>
<th>Gestational age (weeks)</th>
<th>Caesarean delivery</th>
<th>Laboured</th>
<th>Fetal gender</th>
<th>Year</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE10588</td>
<td>ABI Human Genome Survey Microarray Version 2</td>
<td>32,878</td>
<td>17 severe pre-eclampsia</td>
<td>34.0</td>
<td>Ⅺ</td>
<td>N/A</td>
<td>11/16†</td>
<td>8/21†</td>
<td>N/A</td>
</tr>
<tr>
<td>GSE14722</td>
<td>Affymetrix Human Genome 4x4表达芯片</td>
<td>22,115</td>
<td>12 control (preterm)</td>
<td>31.0</td>
<td>Ⅺ</td>
<td>N/A</td>
<td>10/12</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>HG-U133A</td>
<td>Microarray Human Genome Affymetrix U133A Microarray</td>
<td>24,477</td>
<td>11 control (preterm)</td>
<td>3.1</td>
<td>Ⅺ</td>
<td>N/A</td>
<td>11/12</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>HG-U133B</td>
<td>Microarray Human Genome Affymetrix U133B Microarray</td>
<td>48,701</td>
<td>23 pre-eclampsia</td>
<td>3.6</td>
<td>Ⅺ</td>
<td>N/A</td>
<td>16 induced/23</td>
<td>6/12</td>
<td>N/A</td>
</tr>
<tr>
<td>GSE15906</td>
<td>Illumina Human-6×2.0 expression beadchip</td>
<td>47,231</td>
<td>6 pre-eclampsia</td>
<td>3.5</td>
<td>Ⅺ</td>
<td>N/A</td>
<td>11/12</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>GSE10186</td>
<td>Illumina Human-6×2.0 expression beadchip</td>
<td>47,231</td>
<td>8 early-onset pre-eclampsia</td>
<td>3.5</td>
<td>N/A</td>
<td>N/A</td>
<td>8 early-onset</td>
<td>6/12</td>
<td>N/A</td>
</tr>
<tr>
<td>GSE44711</td>
<td>Illumina Human-6×2.0 expression beadchip</td>
<td>47,231</td>
<td>8 control</td>
<td>3.9</td>
<td>N/A</td>
<td>N/A</td>
<td>8 control</td>
<td>6/12</td>
<td>N/A</td>
</tr>
</tbody>
</table>

†Data are shown as described in the article.
In order to assess the resulting set of pathways on the ability to distinguish between preeclamptic and control cases, we executed 100 trials of Support Vector Machine (SVM) modelling and prediction, randomly splitting the samples into equal amounts of training and test data, for both end-points for each trial (Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004). Hence, for example, in the case of experiments using the GSE10588, GSE14722 and GSE25906 data sets, 37 of the 74 control cases were randomly selected for use as training data, and the remaining 37 were held out for a prediction test; the 52 preeclampsia cases were handled similarly, and thus a training set contained 63 example cases along with a test set of 63 cases not included in the training data.

In each modelling trial, a SVM model was constructed after an automated parameter grid search using 3-fold cross-validation. This model was then used to predict the preeclampsia or control status of each case in the test data, and the model was evaluated using the accuracy \([\frac{(TP+TN)}{(TP+FP+TN+FN)}]\), Area Under the ROC Curve (AUC) on the test data, and Matthews Correlation Coefficient \([\frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}]\) (MCC) metrics.

In order to handle the per-batch effects of microarrays and resulting ssGSEA scores, two normalization procedures were executed for evaluation of modelling and analysis of results. In both cases, the normalization was done with respect to each pathway (ssGSEA score) using all samples in the batch processed. The first normalization procedure was to scale by using the sample mean and standard deviation, which is also known as the Z-score transformation \([(x - \mu)/\sigma, \mu = \text{sample mean}, \sigma = \text{sample standard deviation}]\). The second normalization procedure was to apply an affine scaling by using the original range of values and scaling to the range \([-1,1]\). It is well known that the SVM algorithm performs better in general when data are scaled, so these two pathway score transformations are appropriate to the data and algorithm used in the study.

In total, four variations of randomized analysis on the reduced pathway set were executed. The reason for this is because we evaluated the statistical performance of modelling using the GSE10588, GSE14722 and GSE25906 data sets as well as when including the GSE30186 and GSE44711 studies (total of five data sets). The two types of combined data sets have nearly identical ratios of preeclampsia to control cases. For each type of combined data set, the two aforementioned normalizations were applied before evaluation.

### Validation of potential candidate pathways for pre-eclampsia in silico

In order to assess the resulting set of pathways on the ability to distinguish between preeclamptic and control cases, we executed 100 trials of Support Vector Machine (SVM) modelling and prediction, randomly splitting the samples into equal amounts of training and test data, for both end-points for each trial (Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004). Hence, for example, in the case of experiments using the GSE10588, GSE14722 and GSE25906 data sets, 37 of the 74 control cases were randomly selected for use as training data, and the remaining 37 were held out for a prediction test; the 52 preeclampsia cases were handled similarly, and thus a training set contained 63 example cases along with a test set of 63 cases not included in the training data.

In each modelling trial, a SVM model was constructed after an automated parameter grid search using 3-fold cross-validation. This model was then used to predict the preeclampsia or control status of each case in the test data, and the model was evaluated using the accuracy \([\frac{(TP+TN)}{(TP+FP+TN+FN)}]\), Area Under the ROC Curve (AUC) on the test data, and Matthews Correlation Coefficient \([\frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}]\) (MCC) metrics.

In order to handle the per-batch effects of microarrays and resulting ssGSEA scores, two normalization procedures were executed for evaluation of modelling and analysis of results. In both cases, the normalization was done with respect to each pathway (ssGSEA score) using all samples in the batch processed. The first normalization procedure was to scale by using the sample mean and standard deviation, which is also known as the Z-score transformation \([(x - \mu)/\sigma, \mu = \text{sample mean}, \sigma = \text{sample standard deviation}]\). The second normalization procedure was to apply an affine scaling by using the original range of values and scaling to the range \([-1,1]\). It is well known that the SVM algorithm performs better in general when data are scaled, so these two pathway score transformations are appropriate to the data and algorithm used in the study.

In total, four variations of randomized analysis on the reduced pathway set were executed. The reason for this is because we evaluated the statistical performance of modelling using the GSE10588, GSE14722 and GSE25906 data sets as well as when including the GSE30186 and GSE44711 studies (total of five data sets). The two types of combined data sets have nearly identical ratios of preeclampsia to control cases. For each type of combined data set, the two aforementioned normalizations were applied before evaluation.

### Systematic identification of a reduced set of critical pathways

Next, we executed a further analysis to assess if the focused set of pathways could be further reduced to an even smaller pathway subset which maintains predictive ability. For this purpose, we executed recursive feature elimination (RFE), with a linear kernel Support Vector Machine as the modelling algorithm and feature (pathway) weighting mechanism (Guyon et al., 2002). In RFE, the sample features are assigned weights during the model construction process, and features with lower weight are eliminated; this process is recursively done until the original number of features is reduced to a specified number of features. In this work, we eliminated one pathway per RFE pass.

As in the case with the randomized sample modelling, we executed RFE for four variations of data sets. The number of pathways was reduced by RFE to 10 for each data set. The remaining pathways in each variation were tabulated and considered for their involvement in pre-eclampsia. Further, randomized sample modelling based on the RFE-reduced set of pathways was executed using the same protocol described above. As an additional method of examining the results of RFE, we applied multi-dimensional scaling (MDS) to the further reduced data sets (Kruskal, 1964). In short, MDS automatically derives coordinates for a series of datapoints, given a matrix of distances between each pair of datapoints. For visualization purposes, we calculated a MDS solution in two-dimensional space after transforming the post-RFE matrices to per-patient distances quantified by the standard Euclidean distance metric.

### Patients and placenta samples

Twenty women with normal and severe preeclamptic singleton pregnancies were analysed in this study (\(n = 10\), respectively; Table II). Severe preeclampsia was defined as maternal systolic blood pressure \(\geq 160\) mmHg and/or diastolic blood pressure \(\geq 110\) mmHg in two consecutive measurements at least 6 h apart, and proteinuria \(\geq 2\) g/24 h after 20 weeks of gestation. Small for gestational age was defined as relative birthweight less than the 10th percentile according to Japanese standards. Women with pre-existing chronic hypertension, renal disease, lupus erythematosus, diabetes or gestational hypertension without proteinuria were excluded.

Placental villous tissues were obtained immediately after Cesarean section in the absence of labour at Kyoto University Hospital, Japan. Villous tissues were collected from the central part of the placenta, and were macroscopically free of infarction or calcification. After brief rinsing in saline, these tissues were stored in RNAlater (Ambion, Austin, Texas) at \(-80^\circ\text{C}\) until RNA extraction. The study protocol was approved by the Ethics Committee, Graduate School and Faculty of Medicine, Kyoto University, and written informed consent was obtained from each patient.

### Quantitative real-time PCR

Total RNA extraction from placental tissues was performed using an RNeasy Mini Kit (QIAGEN, Germantown, Maryland). The quality and quantity of RNA was measured using an ND-1000 spectrophotometer (Nanodrop, Wilmington, North Carolina). Reverse transcription of 1 mg RNA was performed using the ReverTra Ace (TOYOBO, Osaka, Japan). The forward and reverse primers used for cDNA amplification are shown in Supplementary Table II.

<table>
<thead>
<tr>
<th>Table II</th>
<th>Clinical characteristics of study groups.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control ((n = 10))</td>
</tr>
<tr>
<td>Maternal age (year)</td>
<td>34.2 ± 4.2</td>
</tr>
<tr>
<td>Gestational age (weeks)</td>
<td>37.8 ± 0.6</td>
</tr>
<tr>
<td>Nulliparous</td>
<td>5</td>
</tr>
<tr>
<td>Caesarean delivery before onset of labour</td>
<td>10</td>
</tr>
<tr>
<td>Systolic blood pressure at delivery (mmHg)</td>
<td>99</td>
</tr>
<tr>
<td>Diastolic blood pressure at delivery (mmHg)</td>
<td>63</td>
</tr>
<tr>
<td>Small for gestational age</td>
<td>0</td>
</tr>
</tbody>
</table>
Table SII. Quantitative real-time PCR was performed using SYBR premix ExTaqII (Takara Bio, Otsu, Japan) on the LightCycler 480 Real-Time PCR system (Roche Diagnostics, Mannheim, Germany) as previously described (Chigusa et al., 2013).

Results

Pathway analysis based on independent data sets to discover pre-eclampsia-specific pathways

The results of comprehensive analysis of 9707 pathways using t-test-based screening are available at http://clininfo.med.kyoto-u.ac.jp/~jbrown/. Of the top 180 pathways in each data set, only 21 pathways were common to at least two data sets (Supplementary Table SIII). The panel of candidate pathways included well-known pathways involved in pre-eclampsia such as glutathione (oxidative stress), NF-kB (inflammation) and CDKN1C pathways. Moreover, the current study exhibited the emergence of novel pathways (e.g., GABA receptor and Sonic hedgehog) and potential susceptibility loci (3q and 4p15) for pre-eclampsia that have not been reported as being associated with pre-eclampsia. All of the genes involved in pre-eclampsia-specific pathways are shown in Supplementary Table SIV.

Validation of potential candidate pathways for pre-eclampsia in silico

The results of randomized sampling and modelling using the reduced set of 21 pathways are as follows. The smaller combined data set (GSE10588, GSE14722 and GSE25906) had an average accuracy of 84.6 ± 4.9%, AUC-test of 0.980 ± 0.014 and MCC of 0.691 ± 0.092 using affine scaling; using Z-scaling, they had an average accuracy of 83.2 ± 4.3%, AUC-test of 0.975 ± 0.016 and MCC of 0.664 ± 0.085. The larger combined data set (GSE10588, GSE14722, GSE25906, GSE30186 and GSE44711) had an average accuracy of 79.9 ± 4.6%, AUC-test of 0.964 ± 0.022 and MCC of 0.593 ± 0.092 using affine scaling; applying the Z-scale transformation led to average accuracy of 80.8 ± 4.5%, AUC-test of 0.965 ± 0.018 and MCC of 0.616 ± 0.089.

Quantitative real-time PCR for genes in pre-eclampsia-specific pathways

To validate the results obtained from pathway analysis, the expressions of selected genes involved in the glutathione metabolic pathway (GCLC and GCLM), CDKN1C pathway (CDKN1C), and GABA receptor pathway (GABRA3) were analysed by quantitative real-time PCR, respectively. The expression of each of these genes was significantly reduced in the preeclamptic placentas compared with controls (Fig. 2), and these findings reinforce the data of pathway analysis based on independent data sets.

Discussion

Pre-eclampsia has diverse clinical manifestations such as mild or severe pre-eclampsia, early or late onset, and presence or absence of fetal growth restriction. Although previous studies using microarray analysis sought to find differentially expressed genes and pathways in pre-eclampsia, their results have been inconsistent (Sitras et al., 2009; Winn et al., 2009; Tsai et al., 2011; Louwen et al., 2012; Meng et al., 2009; Blair et al., 2013). This may be partly due to small numbers of study participants or differences in the microarray platform. In the current study, we used the independent data sets with the three largest sample sizes from four different platforms available as GEO data sets in order to avoid various biases. Initially, we tried to screen candidate pathways using false discovery rate (FDR). FDR is designed to prevent a large proportion of false positives, and is commonly used in

Table III Pre-eclampsia-specific pathways based on multiple independent microarray data sets.

Pathway

<table>
<thead>
<tr>
<th>Pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>KORKOLA_CHORIOCARCINOMA</td>
</tr>
<tr>
<td>BIOCARTA_SHH_PATHWAY</td>
</tr>
<tr>
<td>ISHIDA_TARGETS_OF_SYT_SSX_FUSIONS</td>
</tr>
<tr>
<td>chr4p15</td>
</tr>
<tr>
<td>REACTOME_GABA_RECEPTOR_ACTIVATION</td>
</tr>
<tr>
<td>GNF2_CDKN1C</td>
</tr>
<tr>
<td>KEGG_BUTANOATE_METABOLISM</td>
</tr>
<tr>
<td>IL2_UP.VI</td>
</tr>
<tr>
<td>CYCLASE_ACTIVITY</td>
</tr>
<tr>
<td>KEGG_GLUTATHIONE_METABOLISM</td>
</tr>
</tbody>
</table>

In Fig. 1, the results of RFE-SVM analysis using the Z-scale transformation on the smaller combined data set are shown. It is evident from the figure that the reduced set of 10 pathways is discriminative for pre-eclampsia, and motivates further study on the individual pathways and their involvement. A further analysis of only the pre-eclampsia patients in which they are clustered using the cosine distance with complete linkage is given as Supplementary Fig. S1. Additionally, the original set of 9707 pathways visualized by means of MDS could not result in clearly distinguishable patient groups, but MDS visual analysis was much more successful with the reduced set of 10 pathways (see Supplementary Fig. S2).
the analysis of a large number of distinct variables in multiple samples. In the current case, there was only a single pathway left (KORKOLA_CHORIOCARCINOMA) common to at least two data sets (FDR < 0.25). Thus, we did not use FDR as a method for screening for candidate pre-eclampsia pathways. Instead, we performed t-test-based screening.

We found that t-test-based screening under the following conditions (180 top-ranked pathways in GSE10588, GSE14722 and GSE25906) of the 9707 pathways yielded only two pathways (IL2_UP.V1 and KORKOLA_CHORIOCARCINOMA) common to all of the three independent microarray data sets, suggestive of the heterogeneous genomic expression in preeclampsic placentas. Nevertheless, the pathway analysis also revealed that a panel of 21 identified pathways, as well as 10 pathways that were narrowed down using computational analysis, discriminates preeclamptic placentas from controls not only in the smaller combined three data sets used to identify the pathways but also in larger data sets including two independent data sets despite various gestational ages, mode of delivery, and presence or absence of labour onset, indicating that these pathways are highly specific to the pathogenesis of pre-eclampsia.

To date, a single study alone has been reported with the use of multiple data sets from multiple data sources to seek genes and pathways involved in pre-eclampsia, but their study demonstrated computational analysis alone without sufficient validation (Moslehi et al., 2013). Our study seems to have a number of strengths despite the conceptual similarity between their study and ours. First, our analysis included the largest placenta microarray study in pre-eclampsia (GSE25906). Second, we conducted pathway-based screening using a collection of pre-specified gene sets. Organizing genes into gene sets provides a more intuitive and stable context for assessing deeper biological insights in pre-eclampsia, because gene function is collectively exerted and may vary by environmental stimuli, or disease state. Finally, in order to confirm screening results, we conducted multiple validation through 100 trials of SVM modelling and prediction for both the smaller collection of three GSE data sets and the slightly larger collection of five GSE data sets. We found that results were quite similar regardless of either the collection and/or the method used to normalize data when compensating for per-batch effects. Additionally, we applied RFE to arrive at a further reduced set of pathways that contribute to the discriminative ability of a SVM to distinguish PE from control cases. Via RFE, we selected 10 pathways, and repeated the 100-trial random sampling and evaluation procedure. We found that performance was similar to the initial 100-trial experiment executed, signalling the importance of the pathways selected by RFE. Furthermore, we performed confirmatory quantitative real-time PCR for several selective genes related to candidate pathways using preeclamptic placentas and controls from our own institution. In reality, the results of microarray analysis are quite often unable to be verified in other

**Figure 1** Cluster analysis using a panel of 10 pre-eclampsia pathways. Heatmap of normalized pathway activation scores using combined data set (GSE10588, GSE14722 and GSE25906). The results of RFE-SVM analysis using the Z-scale transformation are shown. In the heatmap, each column represents one pathway, and each row corresponds to a sample of placenta. The relative score of each sample to the pathway is represented by a colour. High and low scores are shown in yellow and blue, respectively.
data sets. Nevertheless, cluster analysis demonstrated that not only the initially reduced data set (21 pathways) but also a further reduced data set (10 pathways) discriminated preeclamptic placentas from controls irrespective of the smaller or larger data set, and irrespective of the pathway score normalization procedure. Taken together, we believe that the panel of 10 pathways can provide deep biological insights into pre-eclampsia because our findings were based on multiple independent microarray data sets and deliberate validation.

Potential candidate functions or pathways that have been reported previously include angiogenesis, immune, inflammatory, oxidative stress, cell proliferation and differentiation, and metabolism (Sitras et al., 2009; Winn et al., 2009; Tsai et al., 2011; Louwen et al., 2012; Meng et al., 2012; Blair et al., 2013). Consistently, we identified ten pre-eclampsia-specific pathways which contained previously described pathways such as glutathione (Mistry et al., 2010), IL2 (Hamai et al., 1997) and CDKN1C (Kanayama et al., 2002) pathways. Furthermore, we also discovered several novel pathways potentially involved in the pathogenesis of pre-eclampsia, such as GABA receptor and Sonic hedgehog pathways. After executing RFE analysis described above, we found that the GABA receptor, Sonic hedgehog and 4p15 pathway were always selected as a relevant pathway. Hence, these newly identified pathways warrant further investigation.

Glutathione, a predominant intracellular antioxidant, is synthesized in the cytosol in a tightly regulated manner. Mistry et al. reported that antioxidant enzyme glutathione peroxidase (GPx) is reduced in preeclamptic placentae (Mistry et al., 2010). In addition, we first found that GCLC and GCLM, both of which are rate-limiting enzymes in the biosynthesis of glutathione, were significantly decreased in preeclamptic placentae. Consistent with this, we previously reported that the activation of Nrf2, a predominant transcriptional factor of both GCLC and GCLM, was reduced in preeclamptic placentae (Chigusa et al., 2012). Furthermore, this is the first report that GABRA3 are suppressed in preeclamptic placentae. GABA receptors are associated with oxidative stress-induced apoptosis (Berntsen et al., 2013), and the activation of GABA receptor signalling reduces oxidative stress-mediated damage in liver (Gardner et al., 2012). These findings support the evidence that an impaired antioxidant defence system in the placenta is related to the pathogenesis of pre-eclampsia.

Pre-eclampsia is a multifactorial systemic vascular disorder affecting 5–8% of all pregnancies. It has been suggested that immunologic factors cause failure of the trophoblast to sufficiently invade and remodel maternal uterine arteries at the fetomaternal interface (Redman and Sargent, 2005), and that some are linked to a multifactorial polygenic inheritance with a genetic component (Arngrimsson et al., 1999; Lachmeijer et al., 2001; Redman and Sargent, 2005). A familial predisposition to pre-eclampsia has been demonstrated through previous studies which identified susceptibility loci for pre-eclampsia on 2p, 4q, 9p, 10q, 11q and 22q (Arngrimsson et al., 1999; Lachmeijer et al., 2001; Laiwuri et al., 2003). In the present study, the loci on chromosome 3q and 4p15 were newly identified as candidate loci for pre-eclampsia.

Preeclamptic placenta and cancer share a number of common pathways including angiogenesis, immune, inflammatory, oxidative stress, cell proliferation and differentiation, and metabolic pathways (Louwen et al., 2012). Although most cancer is quite heterogeneous in clinical...
phenotype as well as pathological findings, a pathway-based classification discovered subtypes that reflect specific histological properties and clinical outcomes in breast and lung cancer (Gatza et al., 2010; Nevinns., 2011). We anticipate that this is also the case with pre-eclampsia. In the current study, some of the 10 pathways showed seemingly opposite directions and four subtypes may exist in preeclamptic cases (Supplementary Fig. S1). For example, the heatmap of normalized pathway activation scores demonstrated that the Sonic hedgehog pathway or the glutathione pathway was down-regulated in most, but not all, samples from preeclamptic placenta.

These findings are probably due to the heterogeneity of pre-eclampsia, and suggest that the pathway-based classification is likely to be worthwhile approach to elucidate the pathogenesis of pre-eclampsia, and that pre-eclampsia could be categorized into clinically meaningful subtypes, including early/late onset, mild/severe pre-eclampsia, presence/absence of severe proteinuria, and coincident or not with fetal growth restriction, based on multiple distinct pathways. If detailed vital information could be obtained in each data set analysed in the current study, subpopulations of patients with common clinical manifestations might be identified using the panel of 10 pathways. The present study may be valuable in the understanding of the heterogeneity of pre-eclampsia and for providing a framework to develop rational therapeutic strategies according to pathway-based subtypes. On the other hand, the major limitation of the study is that this is basically an in silico study using a limited number of data sets including different modes of delivery, and presence or absence of labour onset.

In conclusion, ten accurate and reliable pre-eclampsia pathways were identified based on multiple independent microarray data sets.

**Supplementary data**

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

**Authors’ roles**

E.K. designed this study, K.K., E.K., R.M., J.B.B. and Y.O. analysed and interpreted the data. K.K., E.K., R.M., J.B.B. and Y.O. collected and assembled the data. I.K. finally approved the version to be published.

**Funding**

This work was supported in part by Grants-in-Aid for Scientific Research from the Ministry of Education, Science, Culture and Sports, Japan (No. 21592096, 22591822 and 23791383) and by a grant from the Smoking Research Foundation.

**Conflict of interest**

None declared.

**References**


Chigusa Y, Tatsuki K, Kondoh E, Fujita K, Nishimura F, Mogami H, Konishi I. Decreased lectin-like oxidized LDL receptor 1 (LOX-1) and low Nrf2 activation in placenta are involved in preeclampsia. *J Clin Endocrinol Metab* 2012; 97: E1862–E1870.


Mistry HD, Kurlak LO, Williams PJ, Ramsay MM, Symonds ME, Broughton Pipkin F. Differential expression and distribution of placental glutathione...
peroxidases 1, 3 and 4 in normal and preeclamptic pregnancy. Placenta 2010;31:401–408.


