Serum molecular signature for proliferative diabetic retinopathy in Saudi patients with type 2 diabetes

Jianbo Pan,1 Sheng Liu,1 Michael Farkas,2,3 Mark Consugar,2,3 Donald J. Zack,1 Igor Kozak,4 J. Fernando Arevalo,1,4 Eric Pierce,2,3 Jiang Qian,1,5 Eman Al Kahtani6

(The first and last authors contributed equally to this work.)

1Department of Ophthalmology, Johns Hopkins School of Medicine, Baltimore, MD; 2Ocular Genomics Institute, Massachusetts Eye and Ear Infirmary, Harvard Medical School, Boston, MA; 3Berman-Gund Laboratory for the Study of Retinal Degenerations, Department of Ophthalmology, Massachusetts Eye and Ear Infirmary, Harvard Medical School, Boston, MA; 4Vitreoretina Division, King Khaled Eye Specialist Hospital, Riyadh 11462, Kingdom of Saudi Arabia; 5The Sydney Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, Baltimore, MD

Purpose: The risk of vision loss from proliferative diabetic retinopathy (PDR) can be reduced with timely detection and treatment. We aimed to identify serum molecular signatures that might help in the early detection of PDR in patients with diabetes.

Methods: A total of 40 patients with diabetes were recruited at King Khaled Eye Specialist Hospital in Riyadh, Saudi Arabia, 20 with extensive PDR and 20 with mild non-proliferative diabetic retinopathy (NPDR). The two groups were matched in age, gender, and known duration of diabetes. We examined the whole genome transcriptome of blood samples from the patients using RNA sequencing. We built a model using a support vector machine (SVM) approach to identify gene combinations that can classify the two groups.

Results: Differentially expressed genes were calculated from a total of 25,500 genes. Six genes (CCDC144NL, DYXIC1, KCNHI3, LOC100506476, LOC285847, and ZNF80) were selected from the top 26 differentially expressed genes, and a combinatorial molecular signature was built based on the expression of the six genes. The mean area under receiver operating characteristic (ROC) curve was 0.978 in the cross validation. The corresponding sensitivity and specificity were 91.7% and 91.5%, respectively.

Conclusions: Our preliminary study defined a combinatorial molecular signature that may be useful as a potential biomarker for early detection of proliferative diabetic retinopathy in patients with diabetes. A larger-scale study with an independent cohort of samples is necessary to validate and expand these findings.

Diabetic retinopathy (DR) is a leading cause of blindness in adults [1,2]. Almost 23.7% of the Saudi Arabian population older than the age of 30 has diabetes, and the number of affected individuals is predicted to keep growing due to changes in lifestyle and diet [3]. The overall prevalence of DR is 19.7%; 9.1% have non-proliferative diabetic retinopathy (NPDR), and 10.6% have proliferative diabetic retinopathy (PDR) [4], the more severe form of the disease. Diabetes is also a global health problem. For example, approximately 4.1 million American adults older than 40 years of age have DR [1]. Diabetes duration, level of hyperglycemia, and age are the most significant risk factors for DR [4]. Additional risk factors are nephropathy, neuropathy, insulin use, hypertension, and male gender [5,6]. However, these factors explain only a limited amount of the variance in the risk and severity of DR. Although most patients with longer than 30 years of diabetes exposure show evidence of retinal damage, the onset and severity of DR vary significantly among individual patients [7,8]. As the disease progresses, severe NPDR often enters an advanced or proliferative stage (PDR) when blood vessels proliferate [9]. Ischemia-induced retinal neovascularization in association with the development of fibrovascular epiretinal membranes at the vitreoretinal interface is a specific feature of PDR that often results in severe visual loss due to hemorrhage and/or tractional retinal detachment [10].

Attention to blood glucose, blood pressure, and blood cholesterol levels can help reduce the risk and rate of progression of PDR. However, even with good medical management and strong patient compliance, PDR will continue to develop in many patients. Patients who develop PDR can reduce their risk of blindness by 95% with timely treatment and appropriate follow-up care [10-12]. However, if left untreated, or not treated early enough, PDR can lead to severe vision loss and even blindness. Thus, to limit the development of vision loss and blindness, early diagnosis and treatment of PDR is an important public health goal [10,11].
Currently, DR is usually detected through a comprehensive eye exam that includes visual acuity testing, tonometry, anterior segment observation, and a dilated eye exam. However, all too often, diabetic patients do not receive regular ophthalmologic examinations, and complicating the problem, PDR can develop without symptoms. Therefore, the development of a simple blood-based test for early diagnosis and identification of patients who are at high risk for PDR could significantly improve the visual outcome of patients with diabetes. With this goal in mind, several serum biomarkers of DR (e.g., apolipoprotein AI and B) have been reported [13,14]. However, to date, the markers identified suffer from limited specificity and sensitivity [15-17]. One possible limitation of prior studies is that they were mainly based upon a candidate approach, which is limited by our current knowledge of DR. As an alternative approach, we undertook an unbiased and systematic approach to identify novel blood-accessible molecular signatures for PDR based on genome-wide gene expression profiling.

This study aimed to identify a serum molecular signature for PDR detection. We recruited 40 Saudi patients with type 2 diabetes from King Khaled Eye Specialist Hospital (KKESH) in Riyadh, Saudi Arabia, 20 with and 20 without PDR, obtained blood samples, and used next-generation sequencing technology (RNA sequencing) to obtain comprehensive serum transcriptomes for these patients. Based on the global gene expression profiles we obtained, we proposed and validated a molecular signature that can distinguish between patients with NPDR versus those with PDR.

METHODS

Ethical approval: All aspects of this project were conducted in accordance with the principles of the Declaration of Helsinki, and informed written consent was obtained from all participants. This project was approved by the Institutional Review Boards (IRBs) of the Johns Hopkins University School of Medicine (Baltimore, MD) and KKESH.

Patient recruitment: Two groups of patients with type 2 diabetes were recruited, one group with extensive DR and the other without evidence of DR beyond minimal NPDR. A total of 40 patients were recruited for this study from KKESH. All the subjects are Saudis and have the same ethnic background. Among them, 20 have PDR, and the other 20 have NPDR. The two groups matched in gender, age, and duration of diabetes history (Table 1). Additional clinical information for the patients was provided, including glucose level, antivascular endothelial growth factor (VEGF) treatment history, and systemic medications (Appendix 1). The PDR group had features such as neovascularization at the disc and elsewhere due to widespread capillary non-perfusion of the retinal blood supply, retinal and vitreous hemorrhage, and tractional retinal detachment [18]. The second group had non-proliferative DR (NPDR), which is an early stage of diabetic retinopathy. Visible signs in NPDR are microaneurysm and retinal hemorrhages in one or two quadrants. Whole blood was obtained from the patients, and RNA was extracted from the blood using a PAX-gene tube.

Whole-genome gene expression evaluation: We used RNA sequencing (RNA-seq) to obtain an unbiased gene expression profile for each blood sample. Each samples generated about 63 million reads (Appendix 2), which provided reliable gene expression levels. The RNA-seq reads were mapped against the human reference genome (build 37) using TopHat2 [19]. Quality control analysis was performed and visualized using the QoRTs toolset [20], and no major abnormalities or artifacts were found (Appendix 3). The aligned reads were assembled into transcripts using Cufflinks, and gene expression was calculated using Cuffnorm [21]. The expression level of genes was determined based on the value of reads per kilo base per million (RPKM), which was calculated as the number of reads mapped to the transcripts of one gene divided by the

<table>
<thead>
<tr>
<th>Variable</th>
<th>NPDR</th>
<th>PDR</th>
<th>P value (Paired t-test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, year</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>62</td>
<td>61.5</td>
<td>0.383</td>
</tr>
<tr>
<td>Range</td>
<td>49-75</td>
<td>47-77</td>
<td></td>
</tr>
<tr>
<td>Length, years</td>
<td>18.1</td>
<td>18.2</td>
<td>0.905</td>
</tr>
<tr>
<td>Mean</td>
<td>12267</td>
<td>12114</td>
<td></td>
</tr>
<tr>
<td>Gender, no. (%)</td>
<td>16 (80.0)</td>
<td>16 (80.0)</td>
<td>1</td>
</tr>
<tr>
<td>Male</td>
<td>4 (20.0)</td>
<td>4 (20.0)</td>
<td>1</td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

637
Identification of differentially expressed genes: Differentially expressed genes were identified using the GenePattern platform [23]. The paired t test was chosen to assess the differential expression for each gene between the two classes of matched samples based on the RPKM value. The T score of the test statistic was used to rank the differentially expressed genes and calculated such that

\[
T = \frac{\bar{X}_D - \mu_0}{S_D / \sqrt{N}}
\]

where the RPKM differences between all 20 pairs are calculated, and \(X_D\) is the average of the 20 differences. \(\mu_0\) is the mean difference between paired samples under the null hypothesis, typically 0. \(S_D\) is the standard deviation of the differences, and \(N\) is the number (20) of paired samples. A |T score| of greater than 3 and a p value of less than 0.01, which corresponds to a false discovery rate (FDR) of about 20%, were used as the statistical cutoff thresholds. The FDRs for the genes that were selected for molecular signature were lower than 10%.

Molecular signature identification and evaluation: The molecular signature identification process is shown in Appendix 4. With fivefold cross validation, the matched samples were randomly divided into a training set and a test set. The training set included 80% of the samples (i.e., 16 patients with PDR and 16 patients with NPDR), and the test set included the remaining samples (i.e., four patients with PDR and four patients with NPDR). We randomly selected samples for the training or test set. For a given gene set, 1,000 different sets (i.e., training and test sets) were chosen. Among the differential genes, all gene combinations from one gene to six genes were used to build support vector machine (SVM) models using the linear kernel and C-support vector classification [24] and were fitted to the training set. Then the models were used to test the remaining 20% of the samples (i.e., four patients with PDR and four patients with NPDR). For each gene set, the mean area under the receiver operating characteristic (ROC) curve (AUC) was calculated based on the 1,000 sample sets (training and test sets). The gene combination with the highest mean AUC was chosen as a PDR molecular signature. The model was evaluated in terms of sensitivity and specificity. Sensitivity is defined as the true-positive (individuals with PDR who test positive for the condition) expressed as a percentage of all tested individuals who have that disease (total of true positives and false negatives). Specificity describes the true-negative (tested individuals without PDR who test negative for the NPDR) expressed as a percentage of all tested individuals who do not have that disease (the total of true negatives and false positives). They were calculated as

\[
Sensitivity = \frac{TP}{TP + FN}
\]

\[
Specificity = \frac{TN}{FP + TN}
\]

where the true negatives (TN) are the patients with NPDR correctly classified as negative. The true positives (TPs) refer to patients with PDR correctly classified as positive. The false negatives (FNs) are patients with PDR incorrectly classified as negative, and the false positives (FPs) are patients with NPDR incorrectly classified as positive.

RESULTS

Overall study design: The study flow is shown in Figure 1. First, all blood samples were collected and processed, and RNA-seq was performed (Table 1, Appendix 1). Using TopHat [19] and Cufflinks [21], reads were aligned, and then the gene expression levels were calculated. The gene expression data of matched sample pairs were used to identify differentially expressed genes based on the GenePattern paired t test [23]. Then the top differential genes were chosen to build SVM models. The gene combination with the best performance was considered the PDR molecular signature.

Differentially expressed genes between NPDR and PDR: The expression levels of more than 25,500 genes were calculated using TopHat and Cufflinks (Appendix 5). As each sample contains more than 33 million reads (Appendix 4), the gene expression levels could be estimated accurately. Approximately 19,000 genes were expressed in each sample. To assess the quality of the expression profiles, we performed standard quality control analysis for each sample, and the analyses suggested that the sequencing quality was high (Appendix 2). Furthermore, we calculated the correlation coefficient of the gene expression levels between any two samples, and the correlation coefficients were all above 0.93, indicating the overall high quality of the expression profiles (Appendix 6).

We then calculated the differentially expressed genes through paired t test statistics analysis using the GenePattern platform [23]. A total of 155 differentially expressed genes were identified based on the statistical cutoff (|T score| greater than 3.00, p value of less than 0.01, which corresponds to a false discovery rate of about 20%; Appendix 7). Among the 155 genes, 95 were upregulated, and 60 were downregulated. Examples of the differentially expressed genes are shown in
Figure 2. We then performed a Gene Ontology (GO) enrichment analysis using the DAVID tool [25]. The differentially expressed genes were classified into different functional categories such as RNA splicing, protein localization, and tube development (Appendix 8).

Comparison of differentially expressed genes with chemical induced expression changes: To explore the possible mechanisms of these differentially expressed genes in DR, we compared the PDR differentially expressed gene set with lists of genes whose expression has been found to be perturbed by exposure to various chemicals. Connectivity Map (cMAP) is a database that includes the gene expression changes induced by drugs or other chemical components [26,27]. We compared the differentially expressed genes with the expression change profiles in cMAP, which includes expression changes with 1,309 chemicals. Specifically, the differentially expressed genes in PDR were used to query the cMap to find chemicals that induced a consistent and opposite cellular response. Thirty-eight chemicals are significantly ranked with the PDR query (p value of less than 0.01; Appendix 9). Among them, thapsigargin and mitoxantrone had the highest positive mean score and the lowest negative mean score, respectively (Figure 3). Consistent with this finding, it has been reported that mild endoplasmic reticulum stress induced by low concentrations of thapsigargin promoted two critical angiogenic functions, endothelial cell proliferation and migration, and that in vivo administration of thapsigargin accelerates retinal neovascularization in a murine oxygen-induced retinopathy (OIR) model [28]. In contrast, mitoxantrone, an anthracenedione antineoplastic agent, was found to show antiangiogenesis activity and no untoward toxicity in the rat cornea in a previous study [29]. Our findings suggest that the differentially expressed genes identified in PDR might be associated with angiogenesis and relevant to the PDR process.

Molecular signature identification and analysis: We then selected gene combinations from the differentially expressed gene set that can best separate the NPDR and PDR samples. Due to computational limitations, we tested combinations consisting of up to six genes from the top 26 differentially expressed genes for the modeling. We used fivefold cross validation to evaluate the performance of a given gene combination. In other words, we built a model using an SVM approach based on randomly selected 32 samples as the training set and evaluated the performance of the model using the remaining eight samples as the test set. We repeated the separation between the training and test sets 1,000 times. The performance was assessed with the ROC curve. The AUC represented the balanced measurement of sensitivity and specificity. The overall performance of a given gene set was obtained based on the mean value of AUC from the 1,000 cross validations.

To identify the best molecular signature, we tested a total of 313,911 models, which included all gene combinations (n = 1–6) chosen from 26 differentially expressed genes. When only one gene was used, the best performance yielded AUC = 0.851. The performance improved with the inclusion of more genes (Figure 4). The best model we tested included six genes (CCDC144NL- Gene ID 339184, DYX1C1- Gene
Figure 2. Boxplots of the expression data of the top nine differentially expressed genes in the NPDR and PDR samples. The suffixes “_NPDR” and “_PDR” indicate the reads per kilo base per million (RPKM) distribution of the non-proliferative diabetic retinopathy (NPDR) samples and the proliferative diabetic retinopathy (PDR) samples, respectively.

Figure 3. Two highly ranked perturbagens with the proliferative diabetic retinopathy (PDR) signature.
Figure 4. Performance of all tested gene combinations.

Figure 5. The evaluation of the identified molecular signature for PDR/NPDR prediction. A: Receiver operating characteristic (ROC) curve from 1,000 cross-validations. B: PCA plot of 20 non-proliferative diabetic retinopathy (NPDR) samples and 20 proliferative diabetic retinopathy (PDR) samples.
ID 161582, KCNH3- Gene ID 23416, LOC100506476- Gene ID 100506476, LOC285847- Gene ID 285847, and ZNF80 - Gene ID 7634), which yielded a molecular signature with a mean AUC of 0.945 (Figure 5A). The p value via permutation testing was less than 1.0×10^−6. The corresponding sensitivity and specificity for prediction were 91.7% and 91.5%, respectively. Principal component analysis also showed good separation of the two groups of samples using the expression of these six genes (Figure 5B) [30].

**DISCUSSION**

**Significance and implication:** In this study, we conducted RNA-sequencing analysis of blood samples from 40 patients with diabetes. Based on the 40 samples with matched age range, gender, and duration of diabetes, we identified a combinatorial molecular signature for PDR. We believe that our discovery of the molecular signature has significance in clinical application and basic research. First, early detection of proliferative diabetic retinopathy is crucial for timely treatment and visual loss prevention. Serum molecular signatures can help lighten the clinical burden and accurately detect the disease. Second, the molecular signature could help provide insight into the molecular basis of DR. Our Gene Ontology analysis indicated that some differentially expressed genes might be relevant to the PDR process. For example, DLC1- Gene ID 10395, CAV2-Gene ID 858, MAN2AI-Gene ID 4124, SMAD4- Gene ID 4089, and KDR- Gene ID 3791 were involved in tube development. The early stages of neovascularization, one of the main mechanisms of PDR, involve three main steps: endothelial cell proliferation, migration, and tube development [31,32]. KDR and SMAD4 have been reported to be related to retinopathy. Kinase insert domain receptor (KDR, also known as VEGFR2) is the primary responder to the VEGF signal regulating endothelial cell migration and proliferation [33]. VEGF-A or VEGF-C activation of KDR is involved in the formation of capillary-like tubular structures [31]. Several reports have implicated that KDR plays a role in the etiology of diabetic retinopathy [34,35]. Smad 4 (SMAD4) plays the most important role in transforming growth factor-beta (TGF-β) signal transduction. Increased expression of TGF-β1 and Smad 4 on oxygen-induced retinopathy in neonatal mice indicated that Smad 4 may play an important role in regulation of ocular vascular development [36]. After the PDR signature was queried in the Connectivity Map in this study, thapsigargin and mitoxantrone were found to have strong connections with the signature. Thapsigargin is a non-competitive inhibitor of the sarco/endoplasmic reticulum Ca2+ ATPase (SERCA) and an endoplasmic reticulum (ER) stress inducer [28]. Thapsigargin was observed to have highly consistent cellular responses with PDR, which indicates that the molecular pathways perturbed by this chemical may be related to those involved in PDR. Interestingly, mitoxantrone elicits a nearly opposite gene expression response compared with PDR, suggesting the possibility that mitoxantrone, or a related molecule, may have potential as a treatment for PDR. Consistent with this possibility, mitoxantrone, which is an anthracenedione antineoplastic agent, was found to show antiangiogenesis activity and no untoward toxicity in a rat cornea neovascularization model [29]. The inhibition of angiogenesis by mitoxantrone can be attributed to inhibition of prostaglandin E2 secretion, which is a potent stimulator of retinal VEGF secretion [37,38].

**Limitations:** This preliminary study has several limitations. First, we used a limited number of samples (20 patients with NPDR versus 20 patients with PDR) to build the model. A study of more patients, which we hope to perform in the future, will allow us to build a more robust and accurate model to diagnose patients with PDR. Second, the model we developed in this study was not tested using an independent cohort. Although we used cross validation to assess the model, an approach that is considered “data dredging” by some statisticians, we agree that replication using an independent cohort would be a more reliable and rigorous method of assessing the validity of our finding. However, we believe that as the first study to propose a molecular biomarker for PDR, our work provides a new potential signature that is based upon reasonable experimental data, one that can be analyzed by other investigators, and one that future work will either confirm or refute.

**Summary:** A serum molecular signature (CCDC144NL, DYX1C1, KCNH3, LOC100506476, LOC285847, and ZNF80) was proposed for proliferative diabetic retinopathy. The proposed molecular signature can separate patients with PDR and patients with NPDR reasonably well based on gene expression from their blood samples. However, this study is preliminary, and a larger-scale study with an independent cohort of samples will be necessary to validate and expand these findings [39]. Additionally, applying this approach to longitudinal populations of diabetic patients may help determine whether the molecular signature identified here, or other to-be-discovered signatures, can be used as a prognostic factor to predict the development of PDR.

**APPENDIX 1. DETAILED CHARACTERISTICS OF THE 20 NPDR AND 20 PDR PATIENTS.**

To access these data, click or select the words "Appendix 1".
APPENDIX 2. DISTRIBUTION OF THE NUMBERS OF MAPPED READ PAIRS AND EXPRESSED GENES OF THE 40 SAMPLES.
To access these data, click or select the words "Appendix 2".

APPENDIX 3. QUALITY CONTROL OF 40 RNA-SEQ SAMPLES.
To access these data, click or select the words "Appendix 3".

APPENDIX 4. MOLECULAR SIGNATURE IDENTIFICATION PROCESS.
To access these data, click or select the words "Appendix 4".

APPENDIX 5. THE RPKM VALUE OF EACH GENE FOR 40 PAIRED SAMPLES.
To access these data, click or select the words "Appendix 5".

APPENDIX 6. SPEARMAN CORRELATIONS BETWEEN GENE EXPRESSION PROFILES OF 40 SAMPLES.
To access these data, click or select the words "Appendix 6".

APPENDIX 7. IDENTIFIED 155 DIFFERENTIALLY EXPRESSED GENES.
To access these data, click or select the words "Appendix 7".

APPENDIX 8. GENE ONTOLOGY ENRICHMENT OF THE 155 DIFFERENTIAL GENES.
To access these data, click or select the words "Appendix 8".

APPENDIX 9. PERTURBAGENS SIGNIFICANTLY RANKED WITH THE PDR SIGNATURE.
To access these data, click or select the words "Appendix 9".

ACKNOWLEDGMENTS

The study was supported by a grant from KKESH hospital (KKESHJHU01-02). We thank Dr. Zhen Zhang for the discussion. The authors declare that there are no conflicts of interest. Jiang Qian, Ph.D. (jiang.qian@jhmi.edu) and Eman Al Kahtani, MD (ekahtani@kkesh.med.sa) are co-corresponding authors on this manuscript.

REFERENCES


