Regulation of Nuclear Gene Expression Data Analysis in Diabetic Nephropathy and Data Mining

Biotechnology

KEYWORDS : Gene expression; Renal failure; Diabetic Nephropathy, Integrated algorithm

Tejashwini. N	Department of Biotechnology, The Oxford College of Engineering, Bommanahalli, Hosur road, Bangalore – 560 068, Karnataka state, India.
Tanushree Chaudhuri	Department of Biotechnology, The Oxford College of Engineering, Bommanahalli, Hosur road, Bangalore – 560 068, Karnataka state, India.
Kusum Paul	Department of Biotechnology, The Oxford College of Engineering, Bommanahalli, Hosur road, Bangalore – 560 068, Karnataka state, India.

ABSTRACT

Diabetic Nephropathy (DN) is a metabolic disease caused at the end-stage of renal failure that evolved into a progressive fibrosing renal disorder. Recently, a transcriptional and nuclear factor genes show differential expression in type I, Type II and DN. Here we identify differential gene expression signatures related to nuclear genomes of Type I, Type II, Diabetic nephropathy and develop computational classification models of diabetic progression. Gene expression data analysis experiments were performed for 6 sets of glomeruli from normal kidneys and diabetic nephropathy, 18 sets of type 1 and type II diabetes. A series of bioinformatics analysis identified differentially expressed genes that potentially responsible for the progression of diabetic nephropathy. We identified 6 cases of 5727 genes with differential expressions between patient and normal samples. Gene expression data sets for type I, Type II, and DN were analyzed using integrated algorithm. 14 differentially expressed genes correctly classified at the progression status of 92% of patients (P50.001).

Introduction

Diabetic nephropathy (DN) develops in 35–40% of diabetic patients as the result of intra-renal metabolic, hemodynamic and structural changes[1, 2.] DN is a complex phenotype caused by the combined effects of susceptibility alleles and environmental factors which contribute to poor glycemic control and hypertension[3]. The importance of genetic factors in DN is suggested by epidemiological studies and by the familial clustering of nephropathy in both type1 and type 2 DM (Diabetic Mellitus) [4-6]. Nephropathy does not necessarily develop in a significant proportion of diabetic patients, suggesting the entanglement of specific genes. There are several genes is common among type1, type2 and DN. A comprehensive analysis of candidate genes normally associated with type 1 DN, include genes of extracellular matrix material (COL4A1, LAMA4, and LAMC1), matrix metabolism (MMP9 and TIMP3), growth factors/growth factor receptors (IGF1, TGFBR2, and TGFBR3), transcription factors/signalling molecules (HNF1B1/TCF2, NRP1, PRKCB1, SMAD3, and USF1) and those that are believed to be important in the regulation of kidney function[7]. External factors strongly associated with type2 DN genes involve ZNF236, a glucose-regulated Kruppel-like zinc finger gene, Carnosinase 1 gene (CNDP1), Angiotensin-converting enzyme gene (ACE), Manganese superoxide dismutase (MnSOD or SOD2), Apolipoprotein E gene (APOE)[8]. Therefore, our goal is to identify individual gene associates with co-expression in DN in order to develop strategies to prevent or slow disease onset and progression.

Materials and Methods

The RNA sequence datasets of DN with type1 and type2 diabetes were identified manually in the gene expression omnibus (GEO) database on their literature9,10. We are identifying and analyzing multiple independent sources of variation present within multi-dimensional sample datasets, in specific those that are produced by gene microarray expression experiments. The overall approach can be studied as: 1) perform principal components analysis (PCA) of the dataset; for each principal component: 2) identify the most extreme gene probes (those with the highest or lowest weighting) for that principal component; 3) identify and group any conditions in which those extreme probes vary significantly; 4) identify any condition caveats that correlate well with the condition grouping. By extending the apprehension of each principal component from extreme genes (rows) to ordered groups of significant conditions (columns) and further to identify statistically significant correlations with column covariates, we try to make full use of the available data, in an equitable and data-driven way, to analyze and provide meaningful apprehensions of the diverse sources of variation present within the dataset.

RNA Preparation for Microarray

The total RNA seq chip is selected from Affymetrix chip including an on-column deoxyribonuclease. RNA quality and quantity were assessed by micro fluid electrophoresis using an RNA 6000 Pico Lab Chip on a 2100 Bioanalyzer. Fifty samples (14 primary and 36 secondary) with a minimum RNA integrity number of 6.5 was utilized for microarray hybridization. The absolute values of the raw data were used, and then they were normalized by natural logarithm transformation. This pre-processing process was performed by using R statistical software version 2.80.

Affymetrix Microarrays

The Affymetrix and other microarray core facility performed the amplification and hybridization using the Affymetrix GeneChip HG-U95Av2 and additional identically replicated HG-U95Av2 Array. Intensities of each probe set of complete human genomes HG-U95Av2 and 5727 additional genes for analysis of over 5170 transcript variant of DN and 284 series of normal HG-U95Av2 chip sets. All probe sets of HG-U95Av2 are identically replicated on the diseased transcript variant. The sequences from which these probe sets were derived were selected from dbEST, GenBank and RefSeq. The sequence clusters were created from the UniGene database and then refined by analysis and comparison with a number of other publicly available databases.

Data Analysis

Quality Assessment and Data Pre-processing

The Affymetrix CEL files were analyzed using R and Bioconductor. The samples were Robust Multi-array Average normalized using the HG-U95Av2 reference gene set. Microarray quality was assessed using the probe-level modeling and quality metrics provided by the Affy package of BioConductor. Three outlier arrays of type 1, type2 and DN samples that did not cluster with other arrays in principal component analysis results were excluded from further analyses.

Determine the Principal Components of the dataset GDS961, where each row corresponds to a different gene and each column corresponds to one of several different conditions to which the cells were exposed. The GDS961*ith* entry of the matrix contains the *i*th gene's relative expression ratio with respect to a type 1 and typ2 in DN under condition t. To abstinent the influence of gene expression ratios above and below one, we enforced the natural log transform to all ratios. Up-regulated genes thus have a positive log expression ratio, while down-regulated genes have negative log expression ratio. We did comparative expression sets of type 1 and type2 of variance 1 as sometimes recommended when attempting PCA on measurements that are not on a comparable scale. The log ratios included in the analysis are commensurable, no further pre-processing was necessary.

To compute the principal components, the N eigenvalues and their analogous eigenvectors are calculated from the n×n covariance matrix of conditions. Every eigenvector defines a principal component. A component can be considered as a weighted sum of the conditions, while the coefficients of the eigenvectors are the weights.

Identify Extreme Gene Probes for Each Principal Component

The PCA of type 1, type 2 and DN of significant new coordinate systems of rows*columns is effectively rotating the entire dataset. The new principal component describes where the data is to express and projected along each axis of PC. The significant of ranked in one of two ways: by identifying points having a low probability of belonging to a Gaussian fit to the distribution of points along the PCn axis, or by taking a fixed number of n extreme points at each tail of the distribution. These high and low extreme gene point sets are informative in and of themselves because they represent the most extreme of the data points along a principal axis of aberration. As such, the high and low PCEG sets are some of the primary outputs generated by our procedure. We use the term “extreme” in a very familiar sense, in that the points stand out because they are far from the main allocation. By further examining their pattern of expression in the original axes we hope to gain a better understanding of their possible biologic significance and stored all values using Data mining technique.

Results and Discussion

Samples Information

The dataset samples include six DN (T2DN), six diabetic without nephropathy (T2D) and six non-diabetic subjects, using 5727 gene spots of human sequence verified cDNA clone revealed significant differential expression of 416 genes. The significant difference between type 1 and type 2 diabetes groups was changed in nerve fibre density over 4 weeks. Eighty percent of the study participants had type 2 diabetes and 61% were treated with insulin.

Microarray Quality Assessment and Identification of Differentially Expressed Genes

Six samples of type 1, type 2 and DN met the RNA quality (HG-U95av2cdf) criteria and were hybridized to Affymetrix gene expression microarrays. Excluding two outliers and one mislabelled sample identified during the quality assessment process, 5727 probes of genes were used in our analysis. The changes in gene expression are described in changing between primary and secondary treatment groups. In at least one sample, 5727 genes were expressed above background. Among them 455 genes had Bayesian P52.65, while 284 genes had Chip Inspector FDR 48.35%. Only 532 genes deemed as differentially expressed genes by both methods were included for further assays. Real-time reverse transcriptase–polymerase chain reaction demonstrated a positive correlation with the microarray data in all of the eight selected differentially expressed genes. The potential difference of 247 co-expressed genes in both type1 and type2 diabetes includes DN.

Principle Component Analysis of Differential Gene Expressions

PCA plots of the unprocessed, RMA pre-processed and MAS5 pre-processed data intensity levels were plotted. These plots provided further evidence that the array “array3” contains high levels of background noise or otherwise compromised data which cannot be solved by normalization, as becomes clear from its non-clustering with its fellow group members before or after pre-processing as shown (Figure 1).

The results show that PC1 has the highest percentage of deviation 63.3647 and remaining PC2, PC3, PC4, PC5 and PC6 show less. PC1 contains the clusters of arrays which show high deviation. From this PCA plot arrays of genes which are highly expressed and lowly expressed for DN were known.

Functional Enrichment Analysis

The Database for Annotation, Visualization and Integrated Discovery (DAVID) and Concept Gen were used to identify over-represented biological functions and pathways among the differentially expressed genes (Table 1).

Conclusion

In this work with the help of microarray and statistical data analysis gene expression of genes that are common in type1 DM, type2 DM and DN like Myeloid cell nuclear differentiation antigen, KARP-1-binding protein, endosulfine alpha etc. were found along with their functions. Determining common genes will yield novel insights into disease pathogenesis, provide new analytic targets and identify potential DN biomarkers.

Figure 1: Principle Component Analysis of Raw Data.

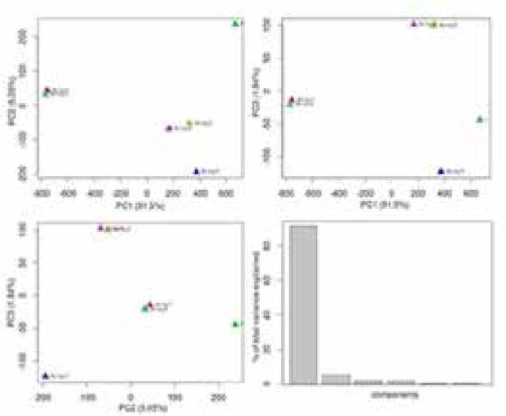


Figure 2:- Principle Component Analysis of Normalized Data.

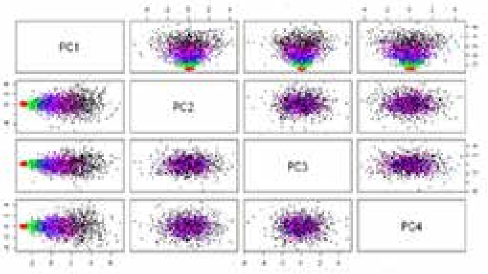
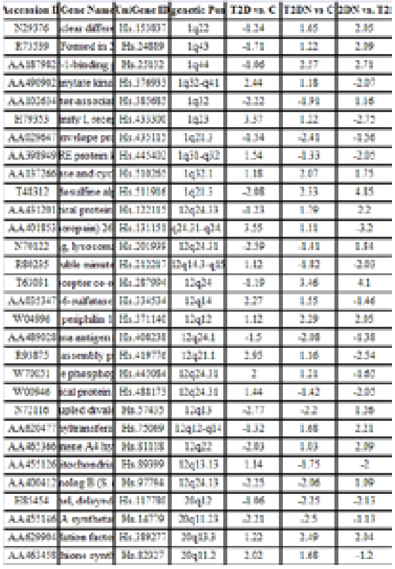


Table 1:- Differential Gene Expression of Common genes.



ccession I	Gene Name	niGene II	genetic Pos	T2D vs. C	T1DN vs C	2DN vs. T2
N29376	clear differe	Hs.153837	1q22	-1.24	1.65	2.05
R71339	Formed in 2	Hs.24889	1q43	-1.71	1.22	2.09
AA187982	-1-binding	Hs.23832	1q44	-1.86	2.57	2.71
AA490902	arylate kina	Hs.376933	1q32-q41	2.44	1.18	-2.07
AA102634	ter-associa	Hs.383683	1q12	-2.22	-1.91	1.16
H79353	inity I, rece	Hs.433300	1q23	3.37	1.22	-2.72
AA029647	nvelope pr	Hs.435115	1q21.3	-1.34	-2.41	-1.56
AA398949	RE protein I	Hs.443402	1q31-q32	1.54	-1.33	-2.05
AA157266	ase and cyc	Hs.310285	1q32.1	1.18	2.07	1.75
T48312	fosulfine alp	Hs.311984	1q21.3	-2.08	2.33	4.85
AA451201	ical protein	Hs.122185	12q24.13	-1.23	1.79	2.2
AA401853	oxopain) 26	Hs.131151	q24.31-q24	3.55	1.11	-3.2
N70122	g, lysosoma	Hs.201939	12q24.31	-2.39	-1.41	1.84
R80235	uble mimetic	Hs.212287	12q14.3-q15	1.12	-1.82	-2.03
T63081	ceptor co-a	Hs.287994	12q24	-1.19	3.46	4.1
AA035347	6-sulfatase	Hs.334534	12q14	2.27	1.55	-1.46
W04896	periplakin 1	Hs.371146	12q12	1.12	2.29	2.05
AA489028	ma antigen	Hs.408238	12q24.1	-1.5	-2.08	-1.38
R93875	assembly p	Hs.419776	12q21.1	2.95	1.16	-2.54
W70031	e phosphog	Hs.443084	12q24.31	2	1.21	-1.62
W03946	ical protein	Hs.488173	12q24.31	1.44	-1.42	-2.05
N72116	upled divale	Hs.57435	12q13	-2.77	-2.2	1.26
AA620477	yltransfera	Hs.75089	12q12-q14	-1.32	1.68	2.21
AA461366	rene A4 hy	Hs.81118	12q22	-2.03	1.03	2.09
AA455126	tochondria	Hs.89399	12q13.13	1.14	-1.75	-2
AA400412	nolog II (S	Hs.97794	12q24.13	-2.25	-2.06	1.09
R81454	ed, delayed	Hs.117780	20q12	-1.06	-2.35	-2.13
AA455168	A synthetas	Hs.14779	20q11.23	-2.31	-2.5	-1.13
AA629901	lation facto	Hs.389277	20q13.3	1.22	2.49	2.04
AA461438	thione synth	Hs.82327	20q11.2	2.02	1.68	-1.2

REFERENCE

1. Wolf G. New insights into the pathophysiology of diabetic nephropathy: from haemodynamics to molecular pathology. EUR J Clin Invest 2004, 34 (12): 785-796. | 2. Rossing P. Prediction, progression and prevention of diabetic nephropathy. The Minkowski Lecture 2005. Diabetologia 2006, 49 (1): 11-19. | 3. Group TDCaCTR. Clustering of long-term complications in families with diabetes in the diabetes control and complications trial. Diabetes 1997, 46 (11): 1829-1839 | 4. Krolewski AS. Genetics of diabetic nephropathy: evidence for major and minor gene effects. Kidney Int 1999, 55 (4): 1582-1596. | 5. Seaquist ER, Goetz FC, Rich S, Barbosa J. Familial clustering of diabetic kidney disease. Evidence for genetic susceptibility to diabetic nephropathy. N Engl J Med 1989, 320(18):1161-1165. | 6. Imperatore G, Knowler WC, Pettitt DJ, Kobes S, Bennett PH, Hanson RL. Segregation analysis of diabetic nephropathy in Pima Indians. Diabetes 2000, 49 (6): 1049-1056. | 7. Moczulski DK, Rogus JJ, Antonellis A, Warram JH, Krolewski AS. Major susceptibility locus for nephropathy in type 1 diabetes on chromosome 3q: Results of novel discordant sib-pair analysis. Diabetes 47: 1164–1169, 1998 | 8. Pradeepa R, Rema M, Vignesh J, Deepa M, Deepa R, Mohan V. Prevalence and risk factors for diabetic neuropathy in an urban south Indian population: the Chennai Urban Rural Epidemiology Study (CURES-55). Diabetes Med. 2008 Apr; 25 (4): 407-12. | 9. Nathalie Vionnet, El Habib Hani, Sophie Dupont, Sophie Gallina, Stephan Francke. Genome wide Search for Type 2 Diabetes–Susceptibility Genes in French Whites: Evidence for a Novel Susceptibility Locus for Early-Onset Diabetes on Chromosome 3q27-qter and Independent Replication of a Type 2–Diabetes Locus on Chromosome 1q21–q24. Am J Hum Genet. 2000 December; 67(6): 1470–1480. | 10. Katalin Susztak and Erwin P. Bottinger. Diabetic Nephropathy: A Frontier for Personalized Medicine. J. Biol. Chem. Aug 5, 2011 286: 27594