Biokemistri
Vol. 33, No. 3, September 30, 2021
Printed in Nigeria

0795-8080/2021 $10.00 + 0.00

http://journals.niseb.org
Also available online at http://www.bioline.org.br/bk

BKR 2021012/33306

In-silico identification of differentially expressed genes in Type 1 diabetes mellitus

O. O. ADEWUMI[1], I. A. TAIWO[2] and E. O. OLADELE[3]

[1,3]Biology Unit, Distance Learning Institute, University of Lagos, [2]Department of Cell Biology and Genetics, University of Lagos, Akoka, Lagos, Nigeria.

Correspondence e-mails: eoladele@unilag.edu.ng, oluadewumi@unilag.edu.ng

(Received August 17, 2021; Accepted September 10, 2021)

ABSTRACT: The incidence of Type 1 Diabetes Mellitus T1DM varies markedly in different geographical populations but seems to be increasing globally. The focus of this research is to screen for T1D-associated differentially expressed genes (DEGs). A meta-analysis was conducted using the Gene Expression Omnibus (GEO) datasets. The datasets included samples from T1DM and normal patients. The Robust Multichip Averaging (RMA) procedure was used for background correction, normalization and summarization to obtain expression level data and to discover differentially expressed genes. Box plots, Density plots, RNA degradation plots and recommended procedures from Affymetrix for quality control were implemented. The DEGs were screened and the exclusively expressed genes were uncovered through the Venn diagrams and heat maps functions in R language. 3,824 genes were classified, as DEGs of which 2,030 were upregulated and 1,794 were downregulated. Seven key genes (TLN1, ANPEP, F13A1, SPARC, SPTBN1, IGHA2 and IGHA1) were exclusively expressed in the whole progression. 58 DEGs were revealed through the Venn diagrams while the Heatmaps showed the differential expression data for 35 genes. IGHA1, IGHA2, IGKV4-1 were significantly expressed and upregulated. Although some of these genes have been previously associated with T1D, many other genes were identified for further studies.

Keywords: Microarray Data, Differentially Expressed Genes, Type 1 Diabetes Mellitus, Downregulated genes, Upregulated genes.

Introduction

Type 1 Diabetes mellitus (T1DM) is a chronic autoimmune disease. It is also known as Insulin-Dependent Diabetes Mellitus (IDDM) or juvenile-onset diabetes. It is a severe condition in which insulin deficiency and hyperglycemia result in the destruction or damage of the beta-cells in the Islet of Langerhans (1). T1DM is regarded as one of the most common genetic diseases and a disorder of glucose homeostasis characterized by susceptibility to ketoacidosis in the absence of insulin therapy (2). T1DM is generally considered as an autoimmune disease affecting millions of people worldwide (3). The main effector function of type 1 diabetes is autoimmunity. The autoimmune breakdown of pancreatic β cells induces type 1 diabetes. However, autoimmunity may not be the primary trigger. Autoimmune disorder occurs when the immune system attacks the tissues and organs of the body itself. The immune system thereby damages the insulin-producing beta cells within the pancreas. Damage to these beta cells impairs the development of insulin which results in signs and symptoms of T1DM. In genetically vulnerable

people, type 1 diabetes precipitates, most possibly due to genetic / environmental risk factors. (4;1;5). A few recent studies on T1D epidemiology have centered on multiple theories including the function of infection, early childhood nutrition, exposure to vitamin D, environmental toxins, obesity, and insulin resistance (3). Research on risk factors for T1D has become an important field of study for detecting hereditary and environmental factors that may possibly be targeted for action. Epidemiological trials play a major ongoing role in researching the diverse triggers, health treatment, prevention, and cure of T1D (6). Genetic experiments have equally played a part in studying the physiology of type 1 diabetes starting with early important observations of the function of the main histocompatibility complex (7). Type 1 diabetes (T1D) happens to be one of the most commonly researched complicated genetic diseases, and it is estimated that the genes in Human Leucocyte Antigens (HLA) account for about 40-50 percent of T1D's family aggregation (8). Significant susceptibility has been shown to occur in the major histocompatibility complex (MHC) on chromosome 6 with other little impacts found in non-MHC loci, although a good number of markers of non-HLA susceptibility genes have been confirmed (9;4). Recent identification of more than 60 loci contributing to the susceptibility of developing type 1 diabetes (T1D) provides a timely opportunity to assess what is currently known of the T1D genetics, and what these findings may tell us about the disease itself (10).

Materials and Methods

Gene Expression Microarray and Data Processing

The gene expression profile (mRNA) datasets for T1DM were obtained from the public functional genomics data repository - Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/). Microarray Dataset 1 (GSM1329616, GSM1329617, GSM1329626 and GSM1329627) contains 4,090,180 samples; Dataset 2 (GSM228562, GSM228563, GSM228582 and GSM228583) contains 1,660,403 samples; Dataset 3 (GSM1329618, GSM1329619, GSM1329628, and GSM1329629) contains 4,090,109 PBMC samples and Dataset 4 (GSM228564, GSM228565, GSM228584 and GSM228585) contains 1,657,165 samples. All samples were from the Peripheral Blood Mononuclear Cell (PBMC) of Humans by Ruijin hospital in China and UTSW Medical Center in Dallas, United States of America. Samples were evaluated for quality assessments. The study was based on Affymetrix GPL570 platform datasets with replicates. Raw data were processed by a robust multichip average algorithm for background correction and normalization.

Software and Microarray Analysis

The statistical computing and graphics were carried out with R language (http://www.r-project.org). The raw data were loaded into the R statistical environment and processed with the Affymetrix package. Visualization, plotting of the data and quality control were processed with the R-package. Normalization of data was performed to correct for systematic technical differences and variation reduction. Most significant genes were ranked and illustrated with plots and diagrams.

Identification of Differentially Expressed Genes (DEGs)

Genes from diabetic patients showing fold change greater than 0.8 (Upregulated) and less than 0.8 (Downregulated) were compared with normal/control sample genes. The fold change approach was used to find differences. The statistical function in R was used to identify the differentially expressed genes and < 0.05 was considered to be statistically significant.

Data Analysis and Visualization

The Venn diagram function in R was used to identify the common and exclusively expressed genes. It compared and visualized the relationships of the differentially expressed genes between the conditions. The Heatmaps tool was used to visualize the expression of genes across the samples. The normalized expression values of the differentially expressed genes were used to generate the Heatmaps.

Results

Quality Assessment and Normalization of Microarray Data

The performance of the normalization procedure was evaluated using box plots of the normalized data. Before normalization, differences in the form or center of the boxes indicate that the data has to be normalized. There were a lot of varieties among the arrays without normalization. Lower grade arrays have boxes that are substantially higher or spread out than other arrays. The distribution of intensity readings of the arrays is represented by each Box plot (Figure 1). The scale and distribution of data on different arrays are comparable, as seen by the Box plots. None of the samples stood out when they were normalized. The median expression level of the various arrays is the same (or at least extremely similar). In addition, the scale of the boxes is nearly identical, indicating that the intensity levels on the various arrays are spaced out relatedly. After RMA normalization, the shape of the Box plot for all the arrays looked comparable, indicating that there were fewer systematic biases in the data. The biases that were present were removed using the log transformation. After transformation, all intensities were spread evenly.

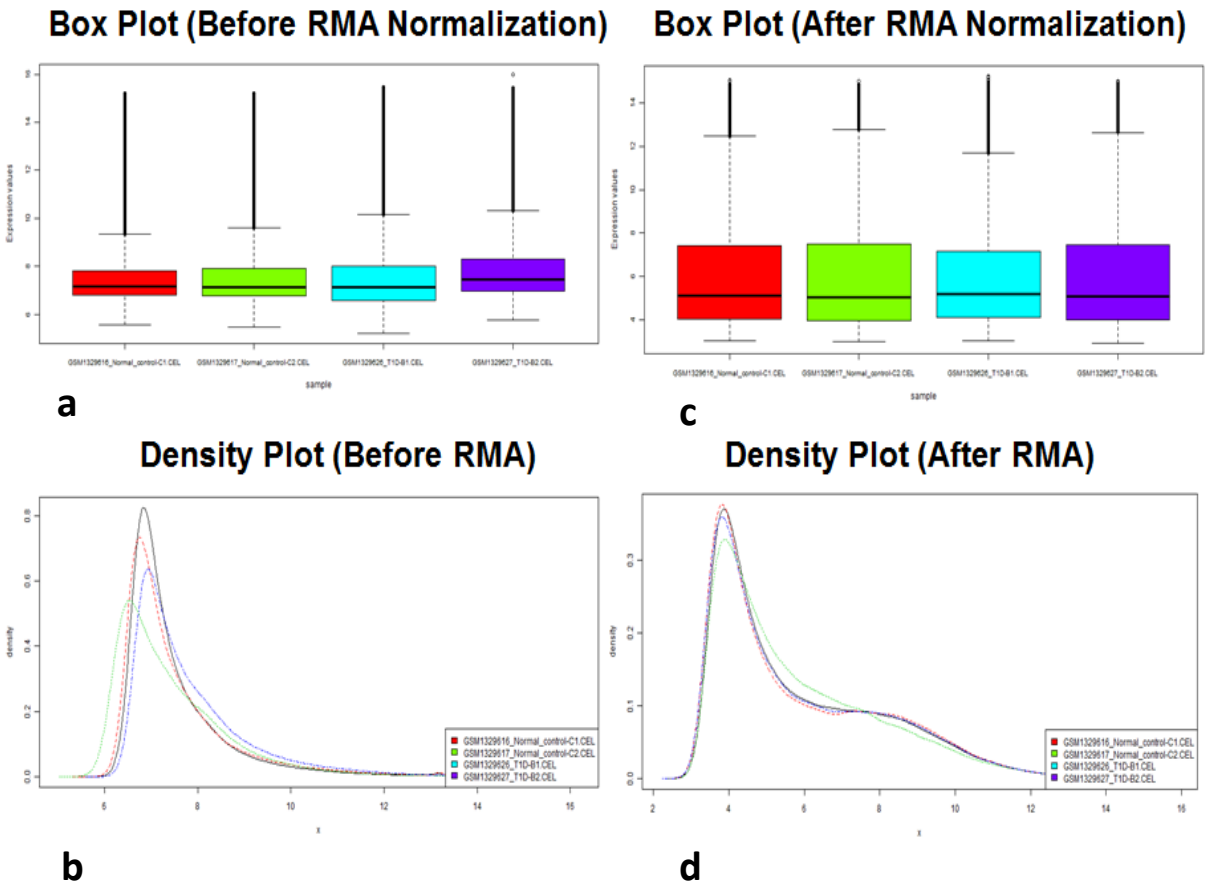

Figure 1. Box and Density Plots for Dataset 1. A and B show Box and Density plots before RMA Normalization. C and D show Box and Density plots after RMA Normalization.

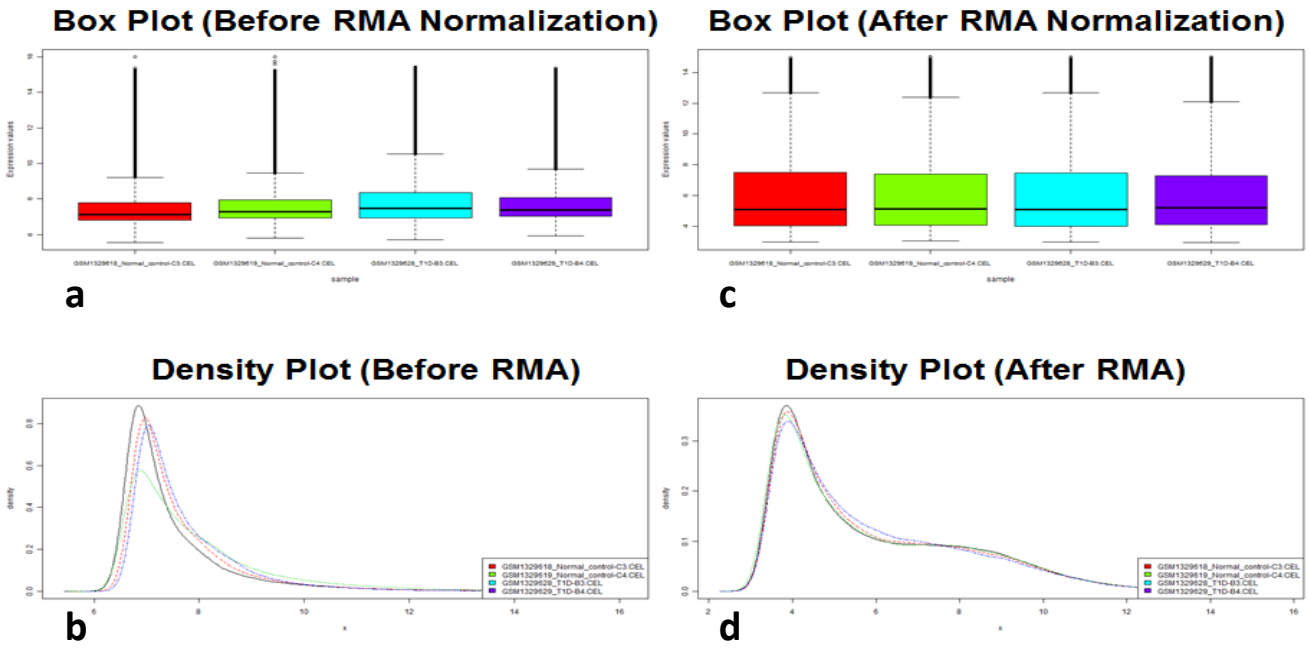

Figure 2. Box and Density Plots for Dataset 2. A and B show Box and Density plots before RMA Normalization. C and D show Box and Density plots after RMA Normalization.

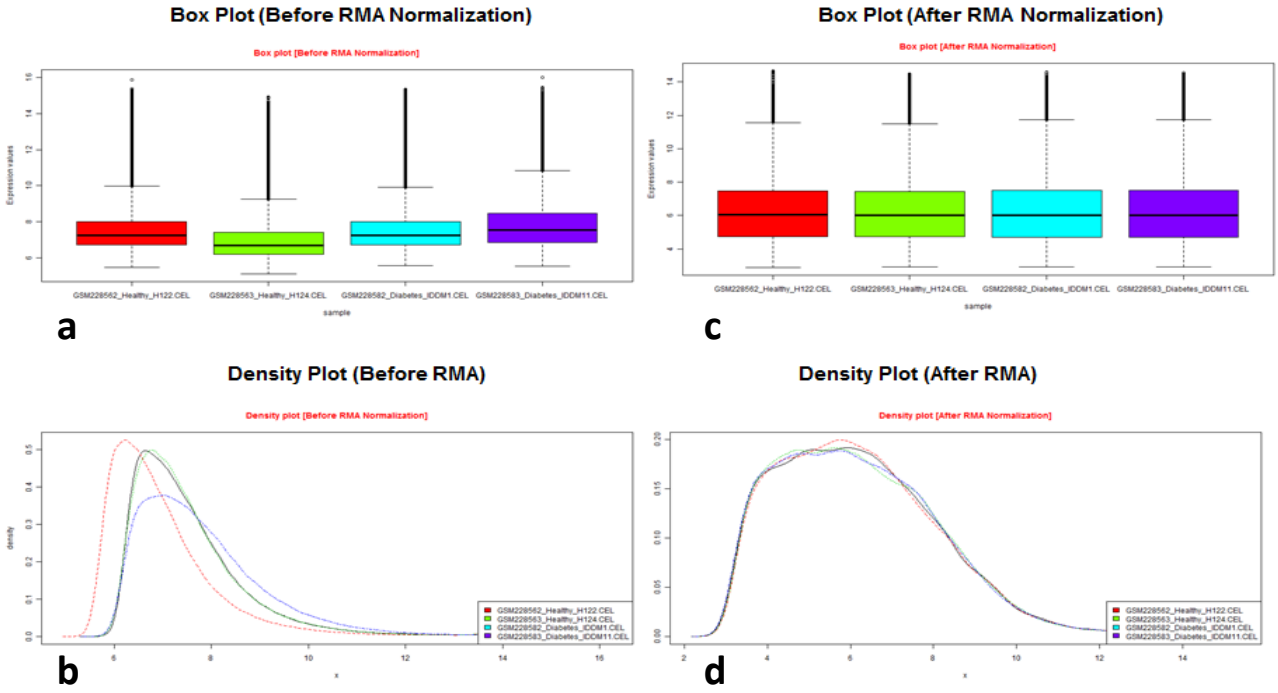

Figure 3: Box and Density Plots for Dataset 3. A and B show Box and Density plots before RMA Normalization. C and D show Box and Density plots after RMA Normalization.

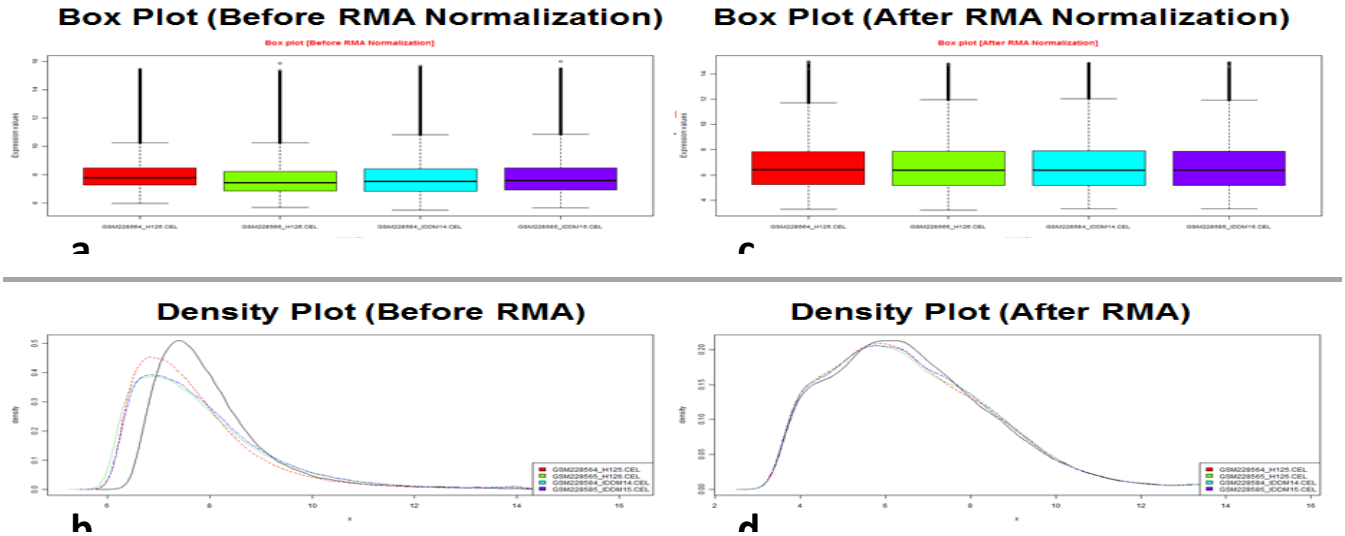

Figure 4: Box and Density Plots for Dataset 4. A and B show Box and Density plots before RMA Normalization. C and D show Box and Density plots after RMA Normalization.

The Quality Control of RNA Samples

The RNA degradation plots illustrate the RNA quality, the quantity of RNA degradation that occurred throughout the RNA preparation, and how well second strand synthesis performed in the sample preparation. The RNA degradation plot shows the average expression of the mRNA from the 5' to the 3' end. Each sample is represented by a single line. The slopes and profiles were discovered to be nearly identical. All of the samples were of good quality. There is a fairly high association between the various arrays in the datasets. There was no outlier among the chips in the RNA degradation plot for the test samples. The RNA Degradation plots for all datasets showed probe intensity decreasing towards 5'. The plots reveal a rising trend in expression levels as the number of probes grows, and the pattern and slopes of the chips were comparable, thus they were all analyzed further.

Legend for RNA Degradation Plots

X axis = Probe Number

Y axis = Expression*

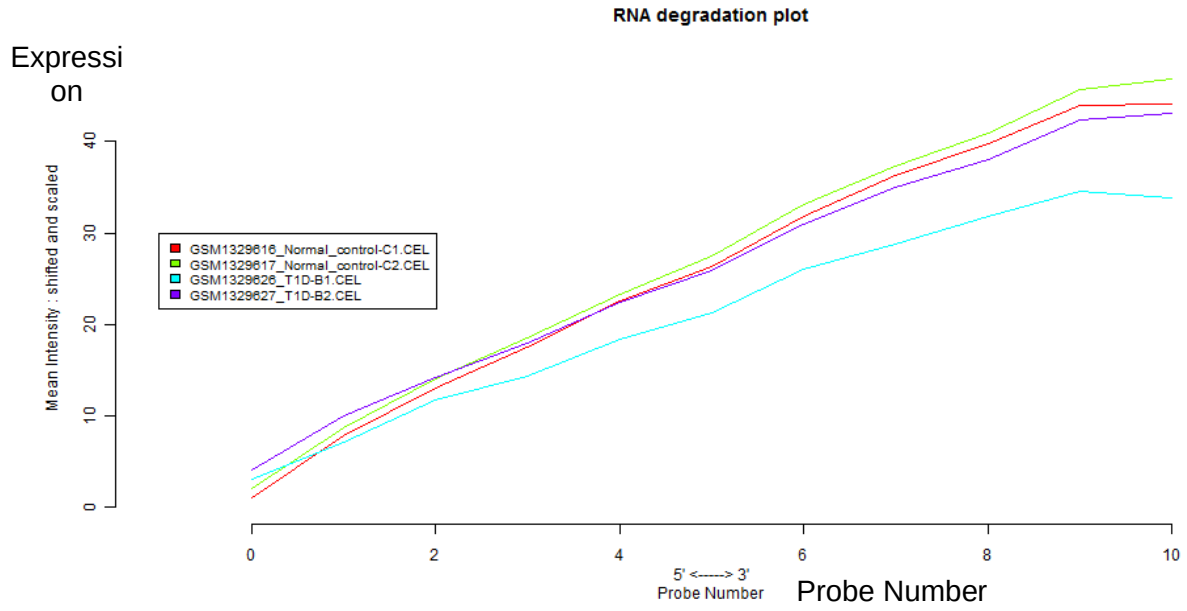

Figure 5: RNA Degradation Plot.

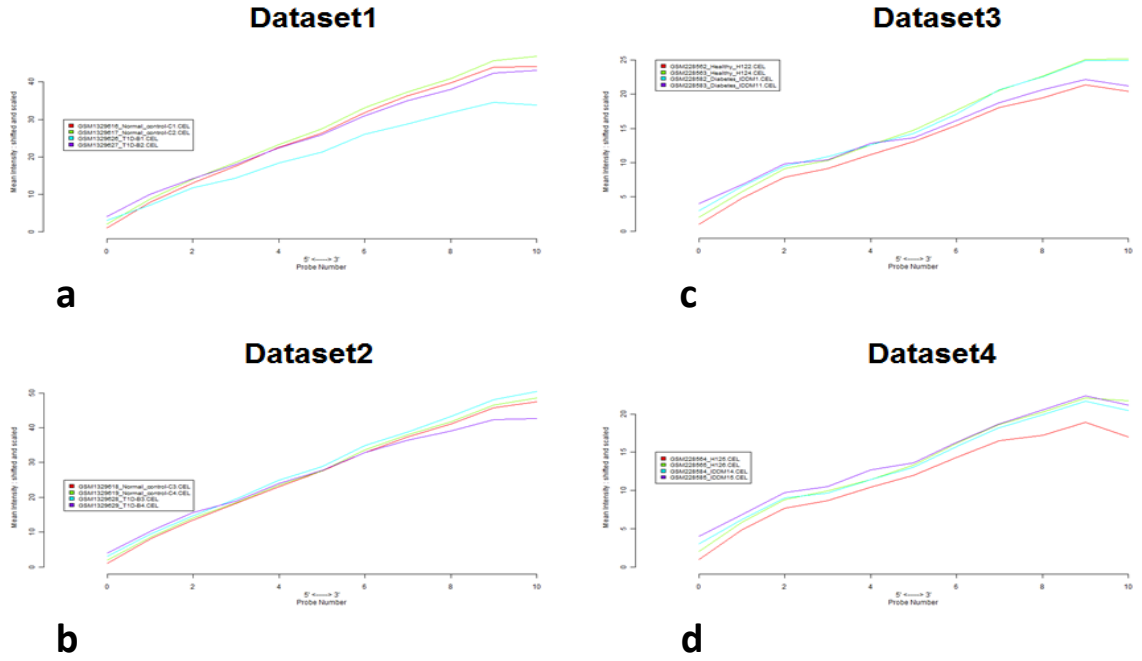

Figure 6: RNA Degradation Plots for all datasets (a, b, c and d).

Identification of Differentially Expressed Genes

The genes that were differently expressed between T1DM patients and controls were obtained. In the meta-analysis, 3,824 genes were identified as DEGs, with 2,030 being upregulated and 1,794 being downregulated. Among other things, most of these genes were involved in DNA binding, actin binding, endonuclease activity, receptor binding, DNA helicase activity, and protein binding. TLN1, ANPEP, F13A1, SPARC, SPTBN1, IGHA2 and IGHA1 were among the differentially expressed genes that overlapped. Only TLN1 and IGHA1 were linked to T1DM among the genes discovered. Gene Ontology annotations related to TLN1 gene include actin binding and insulin receptor binding.

Table 1: Number of Downregulated and Upregulated Genes from all the Samples

	T1D-B1	T1D-B2	T1D-B3	T1D-B4	IDDM1	IDDM11	IDDM14	IDDM15
Down Regulated Genes	668	547	474	525	287	187	170	145
Common Genes	343		250		76		41	
Up Regulated Genes	375	171	348	312	401	462	342	277
Common Genes	122		143		185		116	

Venn Diagrams of Differentially Expressed Genes

The overlap of items between the four datasets was visualized using Venn Diagrams. The Venn diagram distinguished informative modulated genes (exclusively observed for each variable) from non-informative ones, avoiding the reciprocal influence of each variable on the other (intersection with other variables).

The Venn diagram shows upregulated and downregulated genes overlap across the different experimental settings. 58 genes were differentially expressed, including 26 genes between data sets 1 and 2, 10 genes between data sets 1 and 3, 19 genes between data sets 2 and 4, 1 gene between data sets 3 and 4, and 2 genes between data sets 1, 2 and 4.

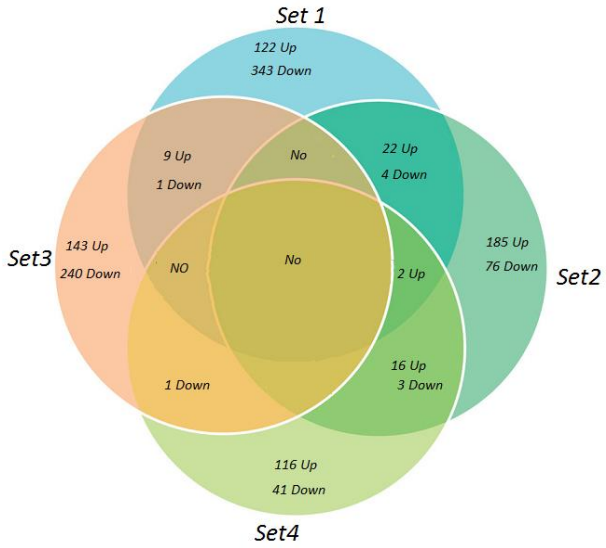

Figure 7. Four-Set Venn diagram comparing differentially expressed genes between analyses. Comparison between significantly co-expressed genes and significantly differentially expressed genes.

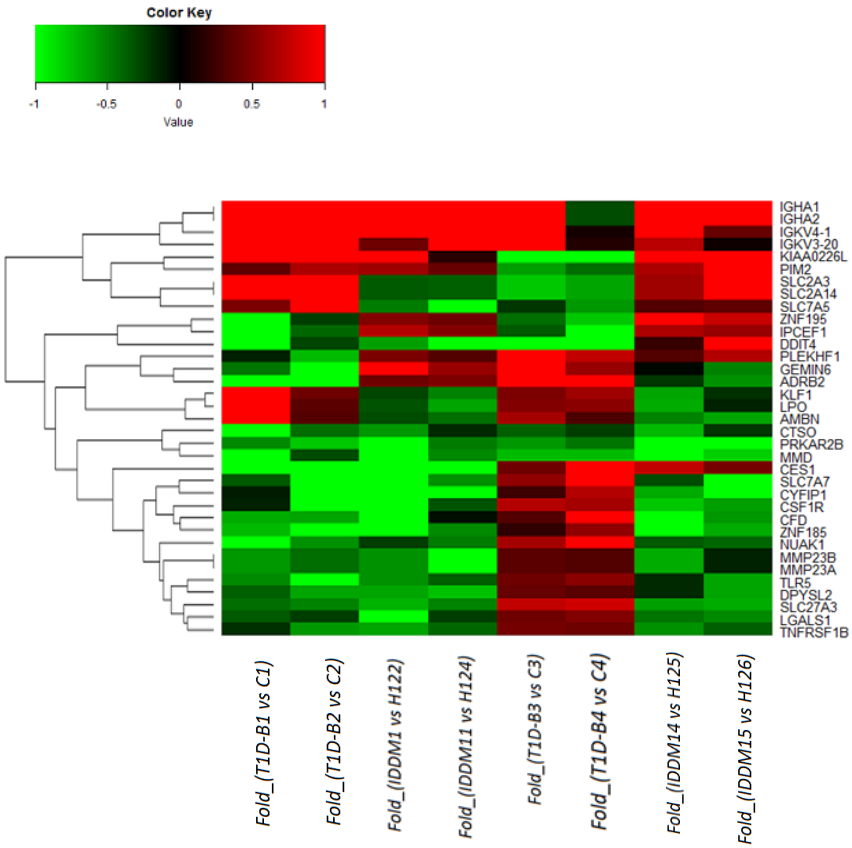

Figure 8. Hierarchical clustering of gene expression profiles for the samples. The colour in each well represents the relative expression of the gene (vertical axis) in each sample (horizontal axis). Red: upregulated genes; Green: downregulated genes; Black: unexpressed genes.

Key:

Column_Headers

ID	*Affymetrix probe Id*
Fold_H125	*Log2 Fold expression value for sample H125*
Fold_H126	*Log2 Fold expression value for sample H126*
Fold_(IDDM14 vs H125)	*Log2 Fold expression value for sample IDDM14 compared to H125*
Fold_(IDDM15 vs H126)	*Log2 Fold expression value for sample IDDM15 compared to H126*
pValue	*t-test pvalue for replicate samples (IDDM14 and IDDM15)*
Annotation	*All annotation were retrieved from NCBI GEO for GPL96 Array*

Discussion

The present study identified a total of 3,824 genes as differentially expressed genes, of which 2,030 were Upregulated and 1,794 were downregulated in T1DM samples studied. TLN1, ANPEP, F13A1, SPARC, SPTBN1, IGHA2 and IGHA1 were exclusively expressed in the blood samples with T1DM. The identification of these biomarkers / differentially expressed genes (DEGs) that are consistently and significantly differentially expressed suggest that these genes are associated with the progression of T1DM in humans (11). Also, (12) in their study on gene expression in children with Type 1 Diabetes Mellitus observed HNRNPD gene to be associated with the progression of T1D. Most of these genes played a role in DNA binding, actin binding, endonuclease activity, receptor binding, DNA helicase activity and protein binding amongst others (12).

In this study, gene expression profiling in peripheral blood cells to identify differences in expression profile of human diabetic patients compared with normal patients has been used. The PCMB was profiled because it is a sample that reflects ongoing pathologic processes such as inflammation and structural abnormalities. Therefore, inflammatory related genes were expected to be modulated in the meta-analysis. In addition, the microarray technology has decreased sensitivity to detect low abundance of genes. In spite of these limitations, the study reveals novel features of human Type 1 Diabetes.

The human leukocyte antigen (HLA) region on chromosome 6p21 was the first known candidate to be strongly associated with T1D in the 1970s (13;14;15). Complex interactions between environmental factors and multiple genes usually influence the risk of developing T1D (16). T1D discovery started as far back as 1974 with five genes being discovered but the arrival of Genome Wide Association Study led to the burst of novel genes that are linked with T1D to the excess of 40 by 2006 and 60 by 2012 (16).

In a previous study using a set of technical replicates, it has been demonstrated that differences owing to day, module, and storage modality (fresh samples vs. fragmented samples) did not significantly contribute to probe set variation (17). Others have demonstrated that technician and laboratory variation are among the largest sources of gene expression variation (18). In another study, researchers demonstrated that substantial gene expression variation owing to laboratory persisted, but was mitigated after implementing standardized protocols for RNA labeling, hybridization, microarray processing, data acquisition, and data normalization (19). Microarray technology has been used to discriminate differences in gene expression profiles in tissues and PBMCs (20). Correct use of microarray analysis can lead to good adjusted p-values, clustering, gene set enrichment results. However, many important genes can be missed if poor quality arrays are included in the dataset. An outlier array can be interpreted, as being of low quality, and this is the reason why its presence would add noise and impair the statistical and biological significance of the analysis. An array can be detected as an outlier because of a real biological property of the sample or an intentional protocol peculiarity. This makes it difficult (and, in general, not advisable) to automate the removal of outliers, as depending on the context.

Since the sources of noise in microarray experiments may be numerous (21), efforts were geared in this study to minimize the influence of noise through various quality control and normalization procedures. One source of variation is cross-hybridization, which occurs when unintended sequences, along with the intended target, hybridize to the same probe, due to sequence homology and/or physicochemical reasons favoring such hybridization. In the case of Affymetrix microarrays, which use a set of short oligonucleotide probes to target a transcript, hybridization conditions are carefully controlled with the aim of minimizing the effect of cross-hybridization due to non-specific binding (22). Multi-array quality checks methods used in this meta-analysis creates an opportunity to integrate data from different experiments related to the same disease or the same biological process to achieve best possible quality of prediction.

The Robust analysis methods can potentially mitigate many of the common problems observed in microarray datasets (23). On the other hand, there are still scenarios where even the most robust methods cannot recover useful signal from a particular low-quality array. Arrays showing evidence of large spatial artifacts, contamination or other gross errors such as mislabeled samples can rarely be salvaged. A

comprehensive analysis of the mechanism underlying Type 1 Diabetes Mellitus development is of crucial importance for therapeutic intervention. This meta-analysis approach that combines differentially expressed genes (DEGs) from microarray datasets to highlight genes that were consistently expressed with statistical significance was employed. The meta-analysis was used to identify common biological mechanisms involved in the parthenogenesis of Type 1 Diabetes Mellitus.

Molecular biomarkers are useful to improve diagnosis, to predict clinical behavior and to demonstrate new therapeutic efficacy. Since microarray can interrogate expression levels of thousands of genes in human genome simultaneously, it has been widely used in discovery of disease biomarkers (16). Differences in expression profile of Type 1 Diabetes in humans compared with normal patients were investigated. 3,824 genes were identified as differentially expressed genes, of which 2,030 were upregulated and 1,794 were downregulated. These genes identified as being differentially expressed in Type 1 Diabetic versus normal patients may be of significant biological interest. To the best of my knowledge, this is the first time several of these genes become candidates for further investigation in their role in Type 1 Diabetes.

Most of these genes played a role in DNA binding, actin binding, endonuclease activity, receptor binding, DNA helicase activity and protein binding amongst others. The overlapped differentially expressed genes identified in this study were TLN1, ANPEP, F13A1, SPARC, -SPTBN1, IGHA2 and IGHA1. Also, in a study by Bergholdt (2012) on T1DM in humans, eight of the regulated genes (CD83, IFNGRI, IL17RD, TRAF3IP2, IL27RA, PLCG2, MYO1B, and CXCR7) have been identified. These DEGs identified in this study may play critical roles in the initiation of T1DM, and investigation of them may shed new lights on understanding of the molecular mechanism of T1DM (24;25). Among these identified genes, only TLN1 and IGHA1 had been previously associated with T1DM. Gene Ontology annotations related to TLN1 gene include actin binding and insulin receptor binding (24). The other identified genes have not been previously associated with T1DM. This study demonstrated gene expression changes in peripheral blood cells that accurately distinguish patients with T1DM from normal individuals. It is anticipated that understanding the genetic background of T1DM will provide accurate and efficient strategies for prevention, diagnosis and control. The genes screened in this analysis may serve as a genetic marker for diagnosis of T1DM. However, further clinical trials are needed to validate these conclusions. In consequence, this study examined the differentially expressed genes in Type 1 Diabetes Mellitus compared to non-diabetic controls and identified the functions of the DEGs by a computational bioinformatics approach.

Conclusion

This study has described specific procedures for conducting quality assessment of Affymetrix Gene Chip microarray data. The analyses described herein successfully identified differentially expressed genes that may be responsible for Type 1 Diabetes. This lends credibility to and supports the fact that the approach and analytic procedures were useful for identifying differentially expressed genes. These procedures have identified differentially expressed genes that play a role in Type 1 Diabetes Mellitus. The genes presented here will aid in the identification of highly sensitive and specific biomarkers in Type 1 Diabetes Mellitus. In conclusion, by this meta-analysis based on gene expression data of Type 1 Diabetes Mellitus, it has shown the underlying molecular differences between Type 1 Diabetic and normal patients.

Recommendations

To minimize variability owing to tissue acquisition in a multicenter gene expression study, it is imperative that tissue acquisition procedures be standardized among all participating laboratory centers. Moreover, it is essential that standardized protocols for RNA labeling, hybridization, microarray processing, data acquisition, and data normalization be used to achieve reproducible gene expression microarray results.

References

1. Van Belle, T. L., Coppieters, K. T., & Von Herrath, M. G. (2011). Type 1 diabetes: etiology, immunology, and therapeutic strategies. Physiological reviews, 91(1), 79-118.
2. Todd, J. A. (1990). Genetic control of autoimmunity in type 1 diabetes. Immunity Today 11: 122-129.
3. Forlenza, G. P., & Rewers, M. (2011). The epidemic of type 1 diabetes: what is it telling us?. Current Opinion in Endocrinology, Diabetes and Obesity, 18(4), 248-251.
4. Steck, A. K., & Rewers, M. J. (2011). Genetics of type 1 diabetes. Clinical chemistry, 57(2), 176-185.
5. Singh, N., Kesherwani, R., Tiwari, A. K., & Patel, D. K. (2016). A review on diabetes mellitus. The Pharma Innovation, 5(7, Part A), 36.
6. Maahs, D. M., West, N. A., Lawrence, J. M., & Mayer-Davis, E. J. (2010). Epidemiology of type 1 diabetes. Endocrinology and Metabolism Clinics, 39(3), 4887 1-497.
7. Polychronakos, C., & Li, Q. (2011). Understanding type 1 diabetes through genetics: advances and prospects. Nature Reviews Genetics, 12(11), 781-792.
8. Noble, J. A., & Valdes, A. M. (2011). Genetics of the HLA region in the prediction of type 1 diabetes. Current diabetes reports, 11(6), 533.
9. aker, P. R., & Steck, A. K. (2011). The past, present, and future of genetic associations in type Benjamini, Y., & Hochberg .H.Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of Research in Statistics Soc Ser B. 57:289–300.
10. 10.Morahan, G. (2012). Insights into type 1 diabetes provided by genetic analyses. Current Opinion in Endocrinology, Diabetes and Obesity, 19(4), 263-270. National Institute of Diabetes and Digestive and Kidney Disease, NIDDK (2014). Causes of Diabetes.https://www.niddk.nih.gov/healthinformation/diabetes/causeshttps://omim.org/
11. Smyth, G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistics and Applied Genetics Molecular Biology. 3 (1): Article 3.
12. Qian H., Chen Q., Zhang S., & Lu, L (2018) The Claudin Family Protein FigA Mediates Ca(2+) Homeostasis in Response to Extracellular Stimuli in Aspergillus nidulans and Aspergillus fumigatus. Front Microbiol 9:977
13. Singal, D.P. & Blajchman, M.A. (1973). Histocompatibility (hl-a) antigens, lymphocytotoxic antibodies and tissue antibodies in patients with diabetes mellitus. Diabetes, 22, 429– 432.1 diabetes. Current diabetes reports, 11(5), 445.
14. Artists Cooperative Groove Union, ACGU (2008). Food Causing Diabetes. http://www.13.waisays.com/diabetes.htm
15. Cudworth, A.G.; Woodrow, J.C.(1975): Evidence for hl-a-linked genes in “juvenile” diabetes mellitus. Br. Med. J., 3, 133–135.
16. Guttula SV, Allam A, Gumpeny RS. Analyzing microarray data of Alzheimer's using cluster analysis to identify the biomarker genes. Int J Alzheimers Dis. 2012;2012:649456. doi: 10.1155/2012/649456. Epub 2012 Feb 14. PMID: 22482075; PMCID: PMC3296213.
17. Dumur, C. I., Nasim, S., Best, A. M., Archer, K. J., Ladd, A. C., Mas, V. R., ... & Ferreira-Gonzalez, A.(2004). Evaluation of quality-control criteria for microarray gene expression analysis. Clinical chemistry, 50(11), 1994-2002.
18. Irizarry, R.A., Hobbs, B., & Collin, F., (2005). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 4:249–64.
19. Bammler, T., Beyer, R. P., Bhattacharya, S., Boorman, G. A., Boyles, A., Bradford, B. U., ... & Zarbl, H. (2005). Standardizing global gene expression analysis between laboratories and across platforms. Nature methods, 2(5), 1-6.
20. Junta, C. M., Sandrin-Garcia, P., Fachin-Saltoratto, A. L., Mello, S. S., Oliveira, R. D., Rassi, D.M., ... & Passos, G. A. (2009). Differential gene expression of peripheral blood mononuclear cells from rheumatoid arthritis patients may discriminate immunogenetic, pathogenic and treatment features. Immunology, 127(3), 365-372.
21. Harbig, J., Sprinkle, R., & Enkemann, S. A. (2005). A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. Nucleic acids research, 33(3), e31-e31.
22. Wu, Y., Ding, Y., Tanaka, Y., & Zhang, W. (2014). Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention. International journal of medical sciences, 11(11), 1185.
23. NIH (2020) U.S. National Library of Medicine https://www.ncbi.nlm.nih.gov/gtr/conditions/C0011854/

24. Howson, J.M.; Rosinger, S.; Smyth, D.J.; Boehm, B.O.; Todd, J.A. (2011): Genetic analysis of adult-onset autoimmune diabetes. Diabetes, 60, 2645–2653
25. Brorsson, C Palleja, A Lukas A. Berchtold, Tina Fløyel, Claus Heiner Bang-Berthelsen, Klaus Stensgaard Frederiksen, Lars Juhl Jensen, Joachim Storling and Flemming Pociot (2012): Diabetes. 61, 954-962pp