Pete Sheurpukdi & Derrick Liu

https://derrick56007.github.io/miRNA_preT1Diabetes/

5/31/2020

Evaluating miRNA as Biomarkers for pre-Type 1 Diabetes

1. Abstract

Type 1 diabetes is an autoimmune disorder in which the body's immune cells attack its insulin producing cells. The cause for this disease is largely unknown and there is no cure. We are interested in finding what may be the cause of this disease by studying the difference in immune cells in healthy and pre-Type 1 diabetic individuals. Specifically, we will be looking at the difference in micro-RNA (miRNA) levels found in the immune cells. miRNA are small non-coding RNA that regulate whether or not messenger-RNA (mRNA) are translated into proteins. With data from the Australian Red Cross, we found that there are clear differences in the immune cells' miRNA expression levels, and that the levels of miRNA expressed in the prediabetic individuals have been found to be positively correlated with reduced insulin production, a decrease in the protein FOXP3 which is correlated with increased autoimmunity, as well as other miRNA correlated with symptoms of Type-1 diabetes such as changes in vision possibly leading to blindness.

2. Introduction

Micro-RNA (miRNA) are small (~20-25 nucleotides) non-coding strands of RNA that attach themselves to messenger-RNA (mRNA) which codes for proteins. By attaching themselves to mRNA, they prevent the mRNA from being translated and thus are responsible for regulating the production of proteins. The miRNA in the dataset are specifically measured from T-cells, which are immune cells in the human body [(1)].

Type 1 diabetes is an autoimmune disorder which means the immune cells attack the body itself, or cells in the body, and is usually diagnosed before 40 years of age, peaking at 14 years old. Pre-diabetes is just a specific stage of diabetes (and can be associated with both type-1 and 2) in which the individual is asymptomatic but there exist markers (ex: high blood sugar) in the body which likely signal the onset of diabetes [(2)].Because of this we are interested in the different ways T-cells differ between healthy and pre-diabetic patients to understand what caused the change in function.

The data we obtained has values for the number of times the miRNA was read in a cell, which is how many times a sequence of base pairs in a sample of RNA are inferred to belong to a certain miRNA. The counts are separated into different types of T-cells for each individual in the dataset, and the amount of certain miRNA present in an individual's cells can help inform what are causing immune cells to behave differently between healthy vs. pre-T1 diabetic patients.

The types of T-cells we will be looking at are naive T-cells, resting T-regulatory cells (rTreg), and activated T-regulatory cells (aTreg). Regulatory T-cells inhibit the production of normal T-cells and help prevent autoimmunity by suppressing the immune response [(9)]. rTreg cells convert into aTreg to help suppress inflammation and cytokine secretion (which are types of proteins that help signal cells toward sites of inflammation) [(10)] though their precise function still requires more research. Naive cells are what respond to new pathogens and develop other T-cells.

If there exists a higher amount of a specific type of miRNA in a cell-type that is different between groups, we can research what gene may be more highly suppressed or expressed by the higher amount of miRNA. The specific cause and cure of T1-diabetes is unknown, and thus it is useful to research what is making the immune cells attack the body’s own insulin producing cells, as well as what we can do to prevent it for pre-diabetic individuals.

Our goal was to replicate some of the analysis methods from the paper “MicroRNAs in CD4(+) T Cell Subsets Are Markers of Disease Risk and T Cell Dysfunction in Individuals at Risk for Type 1 Diabetes” and compare our results using their data. The main findings of the paper was that the miRNA expressed in pre-diabetic individuals promote an autoimmune response as well as have a miRNA expression profile that are correlated with symptoms of diabetes.

3. Methods

3.1 Data

3.1.1 Data Background

The data is a collection of 80 TSV tables, each representing the miRNA expression levels of a specific T cell (aTreg, naive, nTreg, etc.) in either a healthy or pre-T1 diabetes individual. The data background comes from the paper, “MicroRNAs in CD4(+) T Cell Subsets Are Markers of Disease Risk and T Cell Dysfunction in Individuals at Risk for Type 1 Diabetes,” by Zhang Yuxia, Feng Zhi-Ping, Naselli Gaetano, et al [(6)]. The data has columns miRNA, Chromosome, Position, Strand, Total miRNA reads, and Reads Per Million (RPM). Each row contains values for a specific miRNA. RPM is the basis for many of the calculations using expression levels and calculating change/fold-change and the formula is:

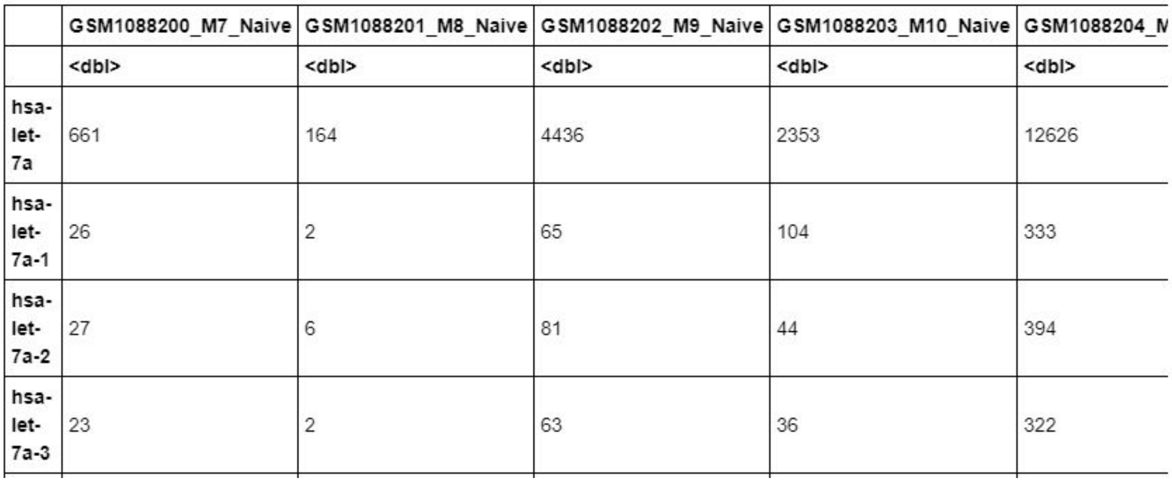
$$\#\ of\ mapped\ reads\ *\ 1,000,000\ /\ \#\ of\ total\ mapped\ reads$$

The 80 samples were taken from 9 healthy and 7 pre-T1 diabetes individuals, sorted into 6 subsets: naive, rTreg, aTreg, transitional memory T-cells, central memory T-cells, and effector memory T- cells. However, not all subsets are included for some individuals in our data and some data was missing, which brought the prospective 96 samples (16 individuals * 6 cell types) to 79 samples. The samples from healthy individuals were obtained from the Australian Red Cross Blood Service and from pediatric autoimmune disease (PAID) controls. Those with a T1D first-degree relative and with autoantigens to at least two antigens (insulin, glutamic acid decarboxylase) were defined as being pre-T1D [(6)].

Because the samples came from Australia, we may not have a globally representative population, but the data should still be suitable in understanding Type-1 diabetes. In Australia, ~5% of the population is diagnosed with diabetes, and 10% of that population has Type-1 diabetes [(4)]. The global population with diabetes is estimated to be 9.3% in 2019, with half estimated to be undiagnosed and 10% of the population with Type 1 [(5)]. Australia having a lower amount of estimated diabetes cases than the global average estimate may add bias to our dataset, though it is also the 7th highest in the world for prevalence of type-1 diabetes in children [(7)]. This could mean that it isn’t a representative dataset and that the results may only be applicable to Australians.

3.1.2 Data Schema

This is the schema after transforming the data ourselves by concatenating, pivoting, and normalizing to reads-per-million from the original reads values.



	GSM1088200_M7_Naive	GSM1088201_M8_Naive	GSM1088202_M9_Naive	GSM1088203_M10_Naive	GSM1088204_M
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
hsa-let-7a	661	164	4436	2353	12626
hsa-let-7a-1	26	2	65	104	333
hsa-let-7a-2	27	6	81	44	394
hsa-let-7a-3	23	2	63	36	322

- Columns: Individual ID (healthy/pre-T1D) and the cell type (Naive/aTreg/rTreg)
- Rows: the RPM counts for each Individual+Type for a specific miRNA

3.1.3 Data Ingestion Process

The data is accessed from the Gene Expression Omnibus [8], hosted by the National Center for Biotechnology Information. The ingestion process is simply through an HTTP request, as the 80 files are relatively small, only retaining the read counts, and not the data for each individual read.

3.1.4 Data Privacy

There aren't many ways for the data to be linked back to the people that it was taken from due to the obscurity of the file names. The only way for the data to be linked back to the person it was taken from is if a person who did the sampling can recognize what sample IDs they gave each person.

3.1.5 Data Cleaning

The data we are working with has a single sample that contains inconsistent columns and values compared to the other samples; in order to produce a clean dataset to work with, we kept only the samples that had columns we needed for our analysis (miRNA, total reads, …). Overall, we have noticed that the data contains no obscure values (other than rounded RPM, which we then replace with what we calculate on our own) and requires little to no cleaning. In the data, the pre-calculated reads per million are all non-negative integers which loses the accuracy of the actual reads per million (RPM). We then calculated logRPM on our own by multiplying each read value in a sample by one million, dividing by the sum of reads in the same sample, and finally taking the log base 2.

3.1.6 Filtering Data

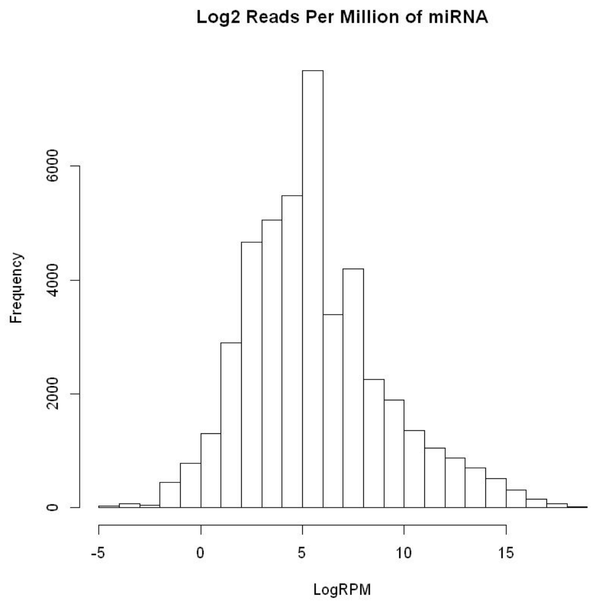
We need to filter out the genes that have low read counts or RPM across all the samples because they do not provide much information about differential expression, and could interfere with any statistical tests used to detect differentially expressed genes. We retained any miRNA that had the sum of its raw read counts across all samples over the total number of samples in the data multiplied by 1.5 (1.5 * 79 = 118.5 read counts) [(6)]. This seemed reasonable because it removed lowly expressing genes, taking the data from 1693 miRNA to 572 and made the logRPM values more uniformly distributed for each sample. Figures 1 and 2 show the resulting distribution of logRPM values.

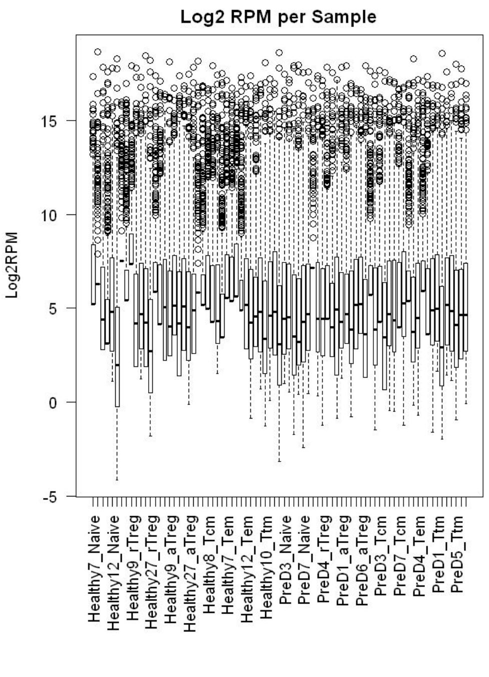
In the beginning, we had it so that we would filter any miRNA that had less than 5 samples with at least 1 read count (the smallest group size is 5). This resulted in around 1081 miRNA left, but the distribution of logRPM values for each sample was less uniform than we would like, so we changed the filter to the former to be more aggressive.

3.1.7 Quality of Data

To determine whether or not our dataset is suitable for miRNA differential expression analysis, we used principal component analysis (PCA) to determine whether or not the data encodes the variation/differences between groups (healthy vs unhealthy, or by cell type) that we would want to see from our analysis. This is after filtering for lowly expressed miRNA that added more variation/noise to the data. The PCA bi-plots tell us which principal components contain information about the variation we are interested in (health status, cell type) and how much separation is in the data. We can see from the figures 3a-c that separation between groups is clear, which means the data is suitable for finding the difference in miRNA expression levels between the groups. In addition, the histograms of the miRNA logRPM values show that we have cleaned and normalized the values to be consistent across samples.

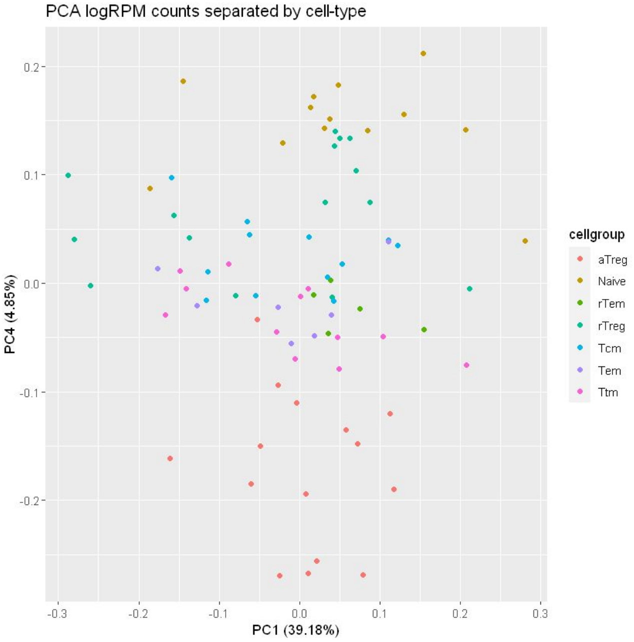
3.1.8 Quality of Data Figures



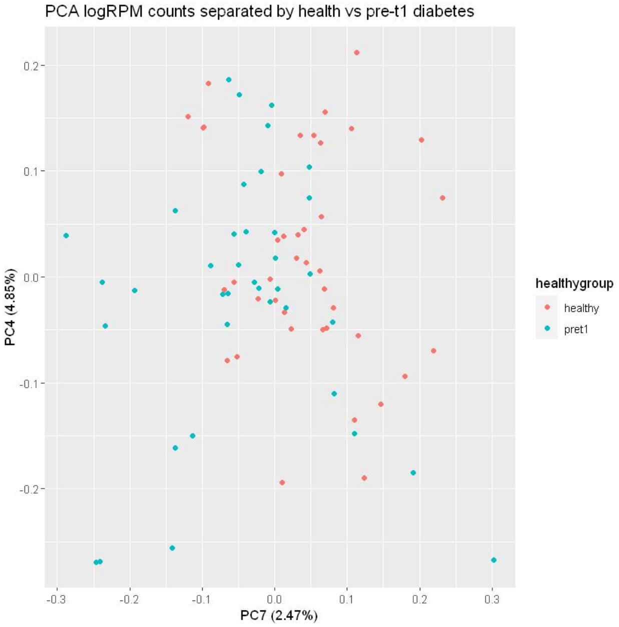


(Figure 1 Above)

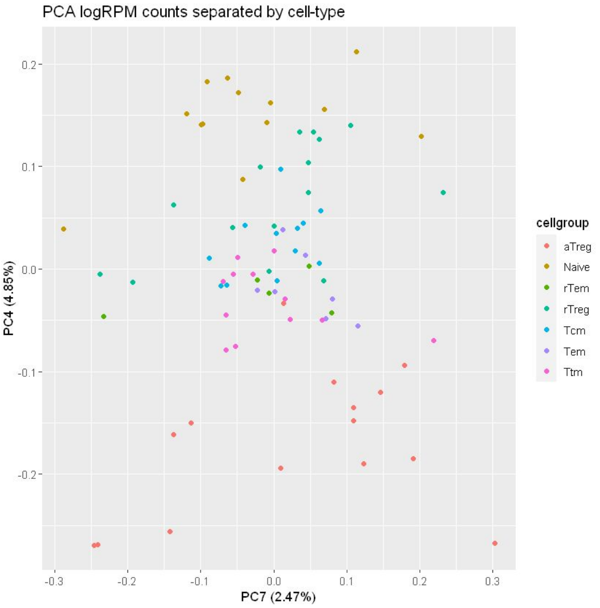
(Figure 2 Above)



(Figure 3a Above)



(Figure 3b Above)



(Figure 3c Above)

Quality of Data Figures Captions

(Figure 1)

We can see that the logRPM of the miRNA is approximately normally distributed across all the samples with a slight right skew. Negative values means the normalized reads per million is between 0 and 1.

(Figure 2)

The boxplot is to help us see the distribution of the logRPM across samples generally. The x-axis doesn't name every sample, but because they are organized by cell types we can get the general idea by group. We see that the distribution is about the same across samples, though there are outliers of highly expressing miRNA for most samples which accounts for the right-skew seen in the histogram.

(Figure 3a)

The purpose of using PCA bi-plots was to see if the dataset's logRPM values contained information about the groups we are interested in and to determine whether or not we could use this dataset for our analysis. We found that the values did contain information for cell-type in principal component 4, whereas 2 and 3 did not capture cell-type variation or the health status and were not shown.

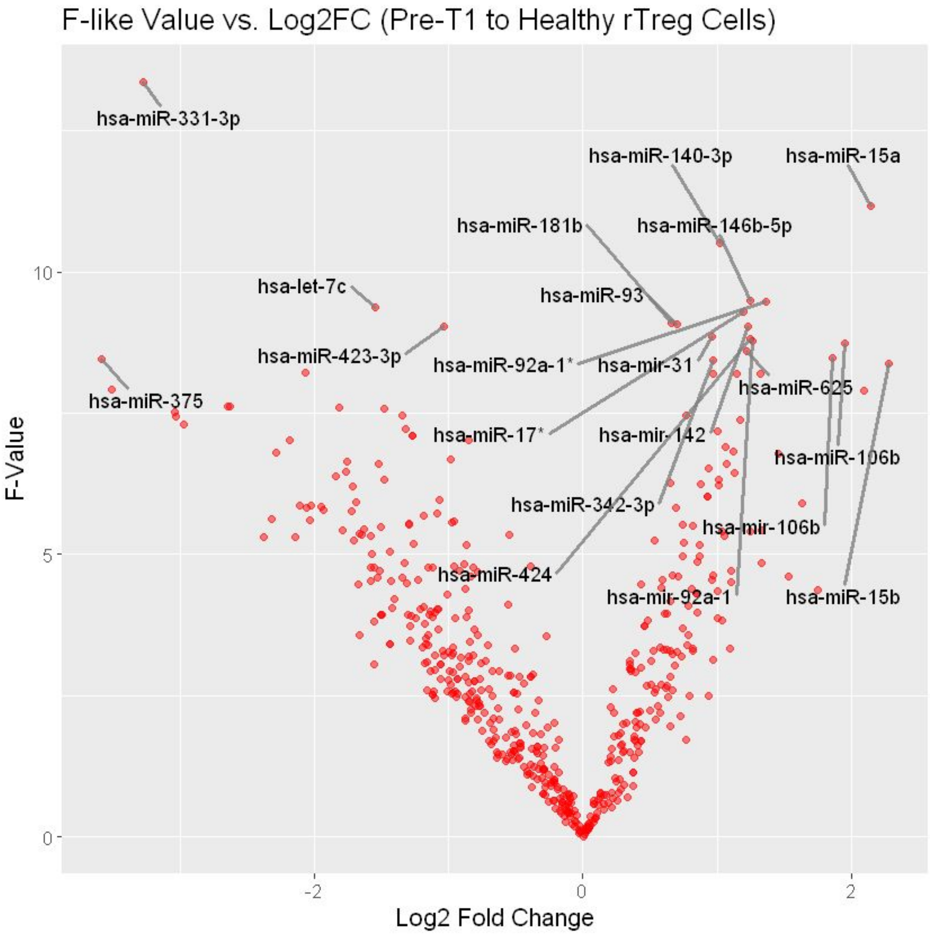
(Figure 3b)

For health status, PC7 encoded the variation best between the healthy and pre-T1 diabetes groups. The figure shows that there is some separation, but still some overlap between healthy and pre-T1 individuals in the miRNA logRPM values. We can infer that the separation is not that extreme because pre-diabetic individuals should be in the early asymptomatic stages of diabetes, and thus may not be as clearly differentiated.

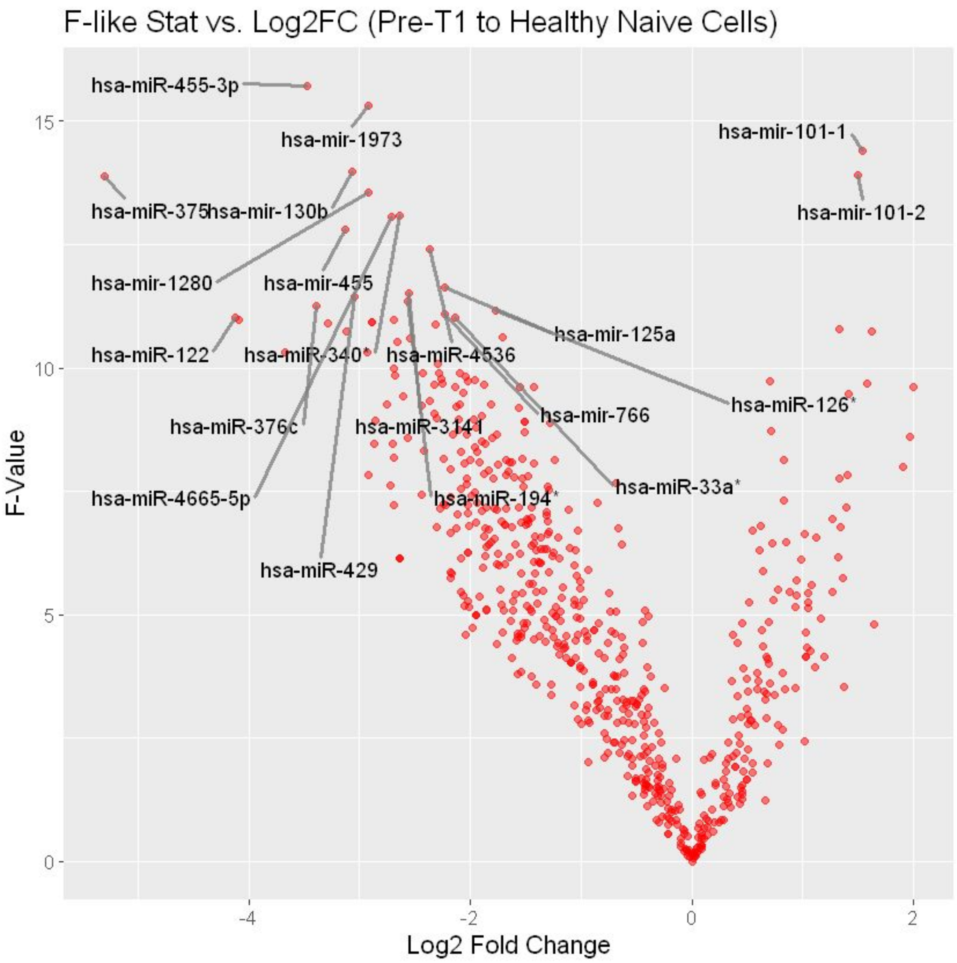
(Figure 3c)

This figure shows figure 3b coloured by cell-type to show the separation is clear even along both groups' axes.

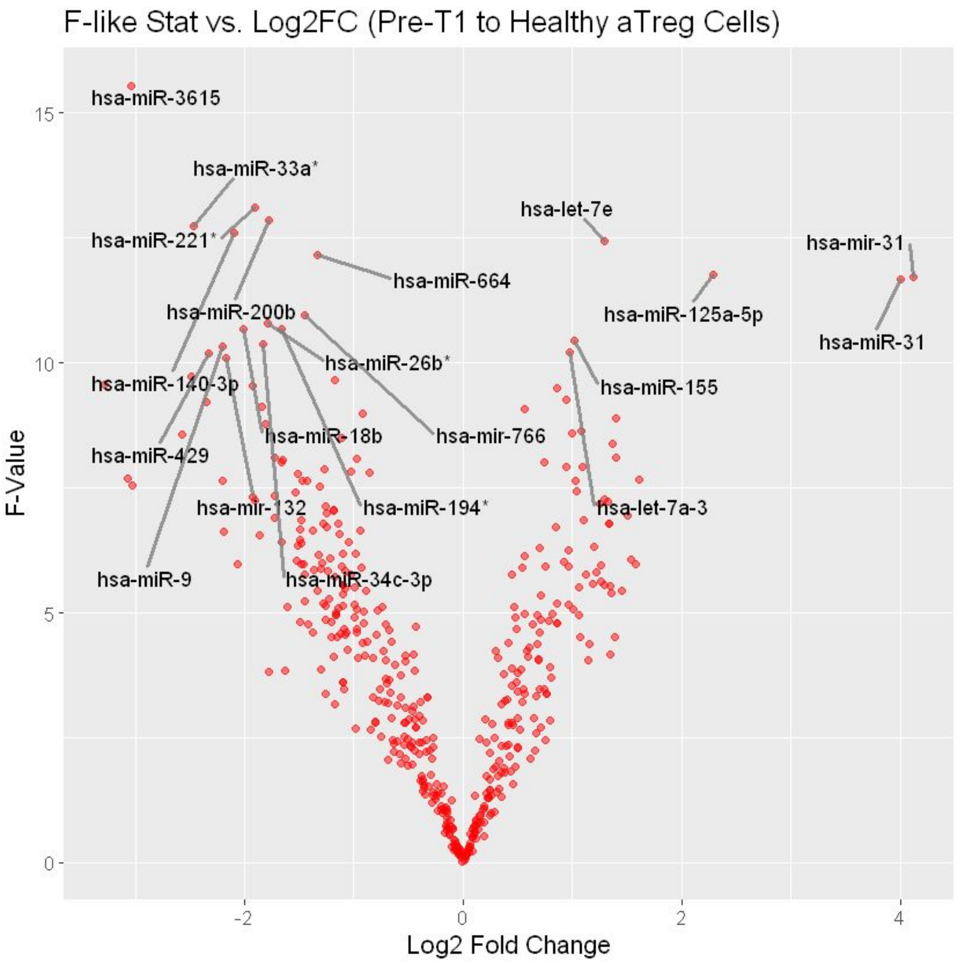
3.2 Analysis Figures



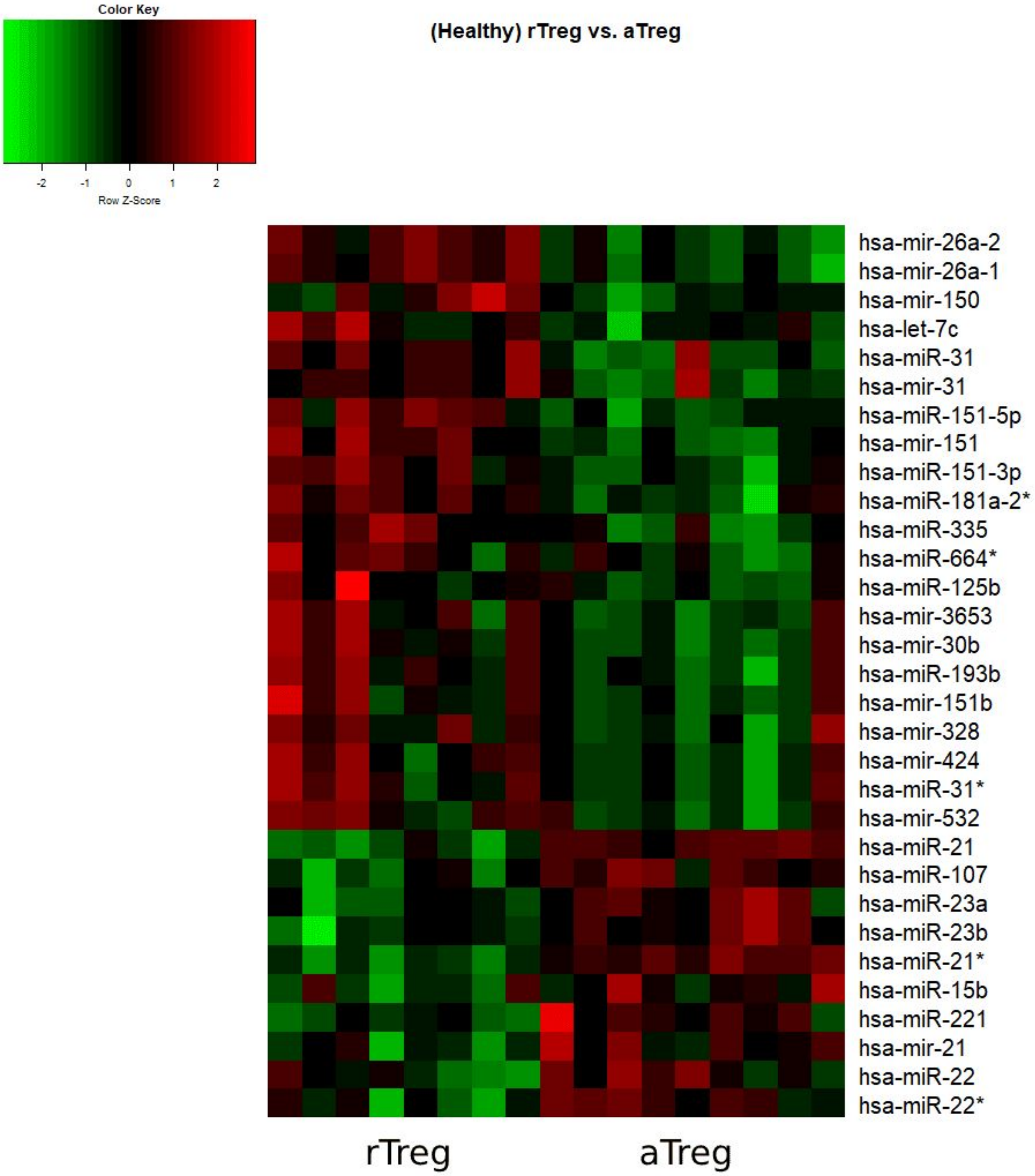
(Figure 4a Above)



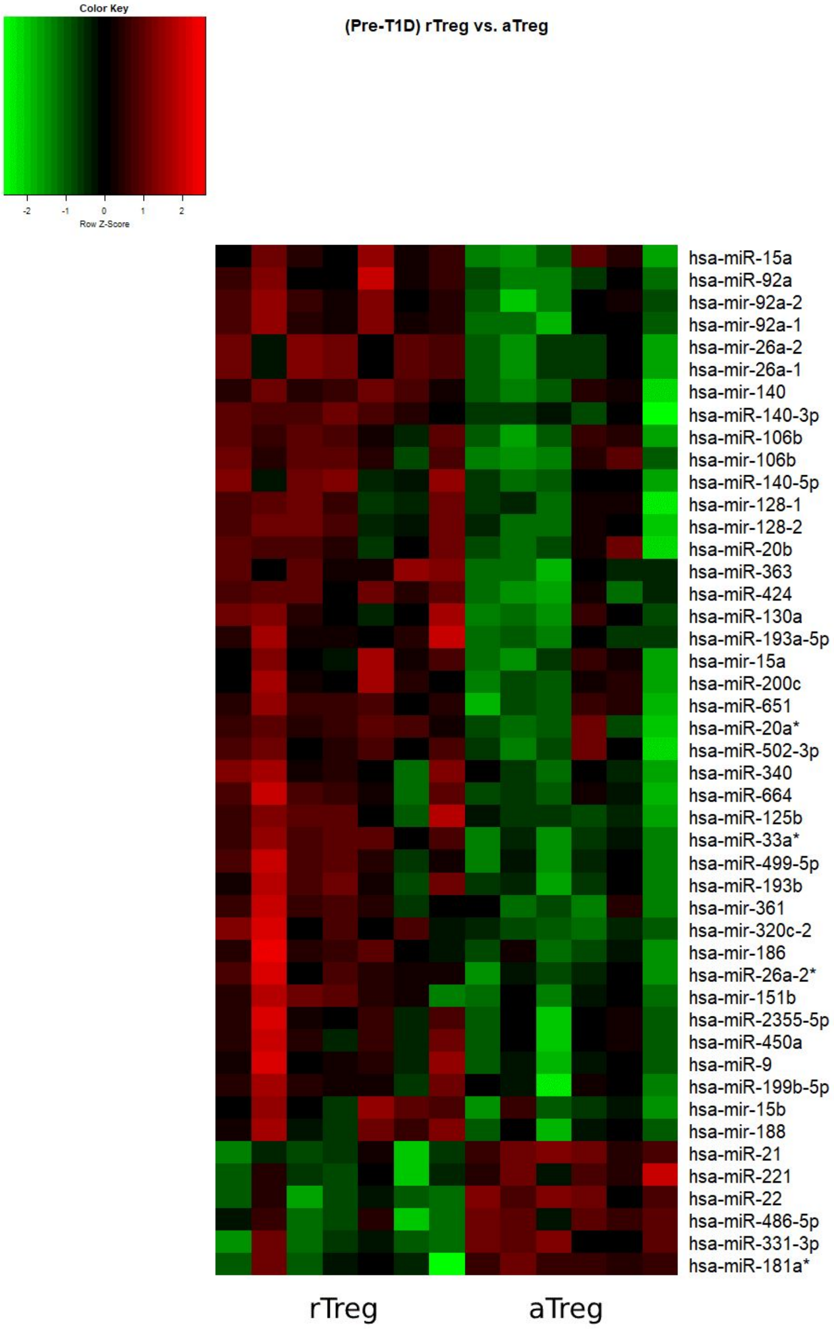
(Figure 4b Above)


F-like Stat vs. Log2FC (Pre-T1 to Healthy aTreg Cells)

(Figure 4c Above)



(Figure 5a, Above)



(Figure 5b Above)

Analysis Figures Captions

Figure 4a/b/c

(Fig. 4a) In rTreg cells, only miR-331-3p, miR-140-3p, and miR-15a were significantly differentially expressed, with >95% of the miRNA having |logFC| less than 2 and F-values less than 9.6.

(Fig. 4b) In Naive cells, there exist >30 significant and differentially expressed miRNA, such as miR-375, miR-455-3p, and mir-101-1/101-2 and more. Most miRNA have decreased expression when going from healthy to pre-Type 1.

(Fig. 4c) In aTreg cells we identified 20 significantly differentially expressed miRNA. The most significant miRNA were mir-31, and mir-3615.

Figure 5a/b

These heatmaps show differentially expressed miRNA between rTreg and aTreg within groups. The heatmaps are scaled so that the rows have a standard deviation of 1 and a mean of 0. After activation of rTreg to aTreg, red means higher levels of expression while green means lower.

3.3 Differential Expression Analysis

3.3.1 Intro

Our initial analysis was to compare the expression of miRNA between healthy and pre-diabetic groups, and found that while Naive cells and aTreg cells have miRNA that are significantly differentially expressed, the rTreg cells did not have as many miRNA expressing a large fold change between groups (Fig. 4b). This means that there may be something occurring when the resting T-regulatory cells are activated and become activated T-regulatory cells to change the miRNA expression profile (the miRNA that are being down or up-regulated), or that the Naive cells, which play a role in T-regulatory cell development, may be a cause for why the aTreg cells were expressing differently between groups.

We then switched to a within-group analysis to see which miRNA are normally changing when rTreg cells become aTreg cells in healthy individuals, and compared that to the pre-diabetic rTreg to aTreg miRNA expression profile (Fig. 5a-b). We found that the differing miRNA expressed in pre-diabetic individuals aTreg cells compared to healthy individuals is correlated with symptoms of diabetes.

3.3.2 Between-Group Methods

We compared the differential expression of Naive (Fig. 4b), rTreg (Fig. 4a), and aTreg (Fig. 4c) between groups, by calculating the log fold change and plotting it along a value for statistical significance, resulting in a volcano plot. A volcano plot is a scatterplot that plots statistical significance versus a magnitude of change [(15)].

The log fold change (logFC), represented along the x-axes of Fig 4a-c, is a ratio between the average logRPM miRNA expressions of the healthy and pre-diabetic groups. We take the mean for each miRNA across all samples in each group then take the difference. Mathematically, it is LogFC = log2(condition1/condition2), which is identical to LogFC = log2(condition1) - log2(condition2) . In this case we have multiple samples and we need to take the mean, so our equation is LogFC = mean(log2(condition1)) - mean(log2(condition2)). This is equivalent to taking the geometric mean, resulting in Log2FC = log2(geo_mean(condition1) / geo_mean(condition2)), which equates to the original LogFC function [(13)].

To calculate the y-axis value, we used a measure defined as the absF value, an F-like statistic which compares the between-group variability to within-group for each miRNA [(6)].

$$absF = \frac{\sum_{i=1}^{K} n_i \, | \bar{Y}_i - \bar{Y} \, | \, / \, (K-1)}{\sum_{i=1}^{K} \sum_{j=1}^{n_i} | \, Y_{ij} - \bar{Y}_i \, | \, / \, (N-K)}$$

Where

Y_{ij} is the expression level of miRNA in sample *j* and group *i*

K is the number of groups being comparing

n_i is the total number of samples within group *i*

N is the total number of samples in all groups

This means that values with a higher absF value has higher variability between healthy and pre-diabetic, which makes it more likely that the difference in miRNA expression level is due to the difference between groups rather than the calculated mean logFC of the group being inaccurate. An absF value >= 9.6 corresponds to <1% false discovery rate [(6)].

In these figures we provide labels for the 20 miRNA with the highest absF scores. The most prospective to note would be the miRNA with a logFC greater than -2 or 2, and a F-value greater or equal to 9.6, as these are the ones that are statistically significant and most differentially expressed between healthy and pre-diabetic samples.

3.3.3 Within-Group Methods

In the heatmap figures 5a-b, the columns are the different individuals that were sampled. The miRNA included for each group (healthy and pre-T1) were chosen by the highest log fold change and only miRNAs with logFC greater than 2 were included in the figures. The group of individuals on the left side of the plot indicate rTreg samples while the right side indicate aTreg samples. We are interested in miRNAs that are highly differentially expressed after activation from rTreg to aTreg and what the differences are.

The heatmaps in Fig. 5a-b show the miRNA expression levels for rTreg and aTreg within each group. Figure 5a shows the expression levels of 31 miRNAs and Figure 5b shows the expression levels of 43 miRNAs. Between these two heatmaps, we find there are only 8 overlapping miRNAs that exist in both healthy and pre-T1D meaning there is little similarity.

3.3.4 Results and Comparison

The significant differences we found in rTreg between groups were few; from Fig. 5a we saw only miR-15a, which is suggested to have a role in cell cycle regulation (16), and miR-331-3p, which regulates the VHL gene's expression (17), generally functioning as a tumor suppressor by regulating cell division (18). The finding of miR-15a was consistent with the source publication (6) ("MicroRNAs in CD4(+) T Cell Subsets Are Markers of Disease Risk and T Cell Dysfunction in Individuals at Risk for Type 1 Diabetes"), but we did not find that let-7c was significant like they did (though it was in the top 20 most significant miRNA of Fig. 5a). let-7c is important because it is associated with progression to end-stage kidney disease in Type 1 diabetes (6). The overall conclusion that the difference in rTreg between groups is small was consistent however.

In Naive cells, there exist >30 significant and differentially expressed miRNA, such as miR-375, miR-455-3p, and mir-101-1/101-2 and more. Most miRNA are downregulated from healthy to pre-Type 1. The functions of miR-375 is associated with diabetes mellitus and lung cancer susceptibility (19). Up-regulation of miR-455-3p inhibits cell proliferation and suppresses the cell cycle of tumors and is a potential therapeutic choice for melanoma (20). mir-101-1/2 both function as tumor suppressors by targeting oncogenes and antioncogenes (21) as well as drive differentiation of naive T-cells to become Th1 cells that respond to intracellular pathogens (6). By biasing naive T-cells to become Th1 cells, it also reduces the development of natural regulatory T-cells (nTreg), which is known to result in autoimmune disease in humans and mice (6). Our identification of mir-101 is also consistent with the source publication. Though there are some differences in exact miRNA identified, the most significant miRNA found are mutual.

In aTreg cells, we found about 20 miRNA significantly differentially expressed, with the most significant being mir-31 and mir-3615. A significant increase in mir-31 for pre-Type 1 aTreg samples was also identified in the source publication and is associated with impaired FOXP3 expression which is necessary to the development of nTreg cells, thus leading to autoimmune disease (6).

We did not see mir-3615 discussed in the source publication in our differential expression analysis in aTreg, but we have found that it is associated with the genetic diseases retinitis pigmentosa and Usher syndrome (22). Retinitis pigmentosa has been reported to rarely have a co-occurrence with diabetes mellitus due to recessive inheritance. This co-occurrence can result in four rare syndromes: Bardet-Biedl syndrome, Alström syndrome, Kearns-Sayre syndrome, and Wolfram syndrome (23). The association of this miRNA with diabetes and retinitis pigmentosa may provide further evidence for the genetic cause of Type 1 diabetes, as well as provide guidance for genetic studies to help diagnose individuals and carriers, or for genetic counseling.

Because rTreg between groups had mostly similar miRNA expression profiles, and aTreg had 20 or more significantly different miRNA, we wanted to research what is happening in the activation from rTreg to aTreg to cause similarly expressing cells to become different. There are 31 significantly differentially expressed miRNAs in the healthy group (Fig. 5a) and 43 in the pre-Type 1 group (Fig. 5b) from rTreg to aTreg. Between these two lists of differentially expressed miRNAs, there are only 8 that overlap between both healthy and pre-T1D. Our findings were very similar to that of the paper; they showed 36 differentially expressed miRNA in Pre-T1D, 21 in healthy, with only 4 overlapping.

Some examples of differences we found were: a decrease of miR-23a/b expression in the healthy group with an absence of any significant differential expression in the pre-diabetic group. Up-regulated mir-23a/b (what we see in the pre-diabetic group) is known to be correlated with inflammation and cancer, although their roles still need to be researched [24]. We also saw that while there was a decrease in miR-31 in the healthy group, there was not in the pre-diabetic group. mir-31 is known to negatively regulate the production of FOXP3, which is necessary in suppressing autoimmunity by developing T regulatory cells [6]. The specific reason why rTreg is activating differently between these groups requires further research, but we would start by researching Naive cells more in depth, since do know that Naive cells play a role in developing these T regulatory cells, and we saw that the pre-diabetic Naive cells had an miRNA expression profile biased toward developing Th1 T-cells which would harm the development of the regulatory T cells.

5. Discussion and Conclusion

Our goal was to replicate some of the analysis methods from the paper “MicroRNAs in CD4(+) T Cell Subsets Are Markers of Disease Risk and T Cell Dysfunction in Individuals at Risk for Type 1 Diabetes” and compare our results using the same data. The main findings of the paper was that the miRNA expressed in pre-diabetic individuals promote an autoimmune response as well as have a miRNA expression profile that are correlated with symptoms of diabetes. Using between group and within group methods we were able to see similar miRNAs in pre-T1D indicative of possible T cell dysfunction in T1D. Profiling of miRNA expression shows differences in expression due to rTreg activation that could serve as a marker to the dysfunction of T cells in pre-T1D individuals. We were also able to discover a significant miRNA (mir-3615) that is correlated with genetic diseases that when co-occurring with diabetes, lead to four rare syndromes. Identifying this relationship can help prevent the occurrence of these syndromes by genetic counseling and identifying carriers within families. Further research would be necessary to understand the cause of the differential miRNA expression, however we believe that identifying the markers of dysfunctional T cells will be helpful in diagnosis and furthering scientific research.

6. References

1. https://www.frontiersin.org/articles/10.3389/fendo.2018.00402/full
2. https://www.canr.msu.edu/news/prediabetes_can_be_associated_with_both_type_1_and_type_2_diabetes
3. https://www.biolegend.com/en-us/treg
4. https://static.diabetesaustralia.com.au/s/fileassets/diabetes-australia/e7282521-472b-4313-b18e-be84c3d5d907.pdf
5. https://www.diabetesresearchclinicalpractice.com/article/S0168-8227(19)31230-6/fulltext
6. https://www.ncbi.nlm.nih.gov/pubmed/26786119

7. https://static.diabetesaustralia.com.au/s/fileassets/diabetes-australia/e7282521-472b-4313-b18e-be84c3d5d907.pdf
8. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE44639
9. https://www.ncbi.nlm.nih.gov/pubmed/20672742
10. https://www.mayocliniclabs.com/test-catalog/Clinical+and+Interpretive/89318
11. http://dx.doi.org/10.1038/ni.2868
12. http://refhub.elsevier.com/S0896-8411(15)30060-3/sref33
13. https://www.biostars.org/p/342756/
14. https://cellero.com/blog/ask-scientist-whats-difference-naive-memory-t-cells/
15. https://galaxyproject.github.io/training-material/topics/transcriptomics/tutorials/rna-seq-viz-with-volcanoplot/tutorial.html#:~:text=A%20volcano%20plot%20is%20a,the%20most%20biologically%20significant%20genes.
16. https://www.researchgate.net/profile/Andrea_Lauri/publication/280735892_Bianchessi_JMCC_2015_and_supp_mat/links/55c46ca308aea2d9bdc1db35.pdf
17. https://ojs.ptbioch.edu.pl/index.php/abp/article/view/1808
18. https://ghr.nlm.nih.gov/gene/VHL
19. https://www.genecards.org/cgi-bin/carddisp.pl?gene=MIR375
20. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5937487/
21. https://www.ajtr.org/files/ajtr0087112.pdf
22. https://www.genecards.org/cgi-bin/carddisp.pl?gene=MIR3615
23. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2850186/#:~:text=The%20combined%20occurrence%20of%20diabetes,in%20an%20autosomal%20recessive%20fashion.
24. https://onlinelibrary.wiley.com/doi/full/10.1002/eji.200838509