LSM3241 CA1:
Identification of miR-203-driven transcriptomic changes during epithelial-mesenchymal transition (EMT) in a breast cancer model

Andrea Esmeralda Halim A0157554X
Tan Wei Qi A0158131M

1. Introduction

MicroRNA (miRNA), originally discovered in *Caenorhabditis elegans,* is found in most eukaryotes, including humans. miRNAs are small, highly conserved non-coding RNA molecules that binds to messenger RNAs (mRNAs) via complementary base pairing to prevent protein production (Macfarlane, 2010). This is achieved through the assembly of miRNAs into the RNA-inducing silencing complex (RISC) which will cleave the target mRNA after binding, hence resulting in translation inhibition. However, dysregulation of miRNA expression is found to be associated with various human diseases, including cancer. One such miRNA, miR-203, was found to be often repressed during the epithelial-mesenchymal transition (EMT), which is important in the initiation of metastasis in cancer progression. To find out how miR-203 causes cancer, we will perform differential gene expression analysis to identify the key genes that are regulated by miR-203, after which gene pathways can be analysed to further understand the development of cancer.

2. Materials and Methods

Bioconductor is the main tool used to analyse differential gene expression as it contains a wide range of packages tailored for the analysis of genomic data (Gentleman, 2005). We made use of the raw data on the level of expression of genes in SUM159 breast cancer line when miR-203 is repressed and expressed that is publicly available on GEO as a series GSE50697. GEOquery is a package in Bioconductor, that acts as a bridge between Bioconductor and GEO, hence installing it will allow users to download the data on GEO into Bioconductor (Davis, 2007).

As GSE contains both samples (GSM) and platforms (GPL), GSMList, that is available under GEOquery, was used to retrieve the samples in the series. It was found that the GSE contains 6 samples, GSM1226581, GSM1226582, GSM1226583, GSM1226584, GSM1226585 and GSM1226586. Comparing the metadata for the GSE and GSM, it was found that source name and characteristics of each sample differ and these might give us more information on the background of these samples.

Using the characteristics of each sample, it is possible to generate a data frame with each sample and its corresponding characteristic, the presence or absence of miRNA-203, as shown in Table 1. However, the library ‘knitr’ has to be installed to generate the data frame using the function ‘kable’.

	Culture
GSM1226581	control
GSM1226582	control
GSM1226583	control
GSM1226584	miR203
GSM1226585	miR203
GSM1226586	miR203

Table 1. Data frame of sample with its corresponding sample characteristic.

Affy is a package in Bioconductor that contains functions for handling Affymetrix data and CEL file data (Gautier, 2004). The function 'read.affybatch' available in the package Affy will read the CEL files, which in turn can then be used for background correction, normalisation and summarisation of the data with the help of RMA algorithm (Chang, n.d). The steps are performed in order and will convert the AffyBatch object into an ExpressionSet, in which the data size is much smaller.

It is necessary to carry out background correction as the observed signal comes from a combination of real signal and noise, which can originate from imaging and non-specific hybridisation. Normalisation is carried out next, usually via quantile normalisation of the probe level data, making all the samples to have exactly the same empirical distribution hence making comparison of samples possible. Lastly, summarisation is carried out to reduce the size of data for each sample to the number of measured transcripts. In RMA, summarisation is carried out using the median polish algorithm (Chang, n.d).

Using 'plotDensity.Affybatch' and 'plotDensity' function from the Affy package (Gautier, 2004), it is possible to plot the signal densities from the CEL file and the data after RMA respectively to visualize the graph. By comparing Figures 1 and 2, it is clear that after RMA processing, the distribution of 6 data is the same.

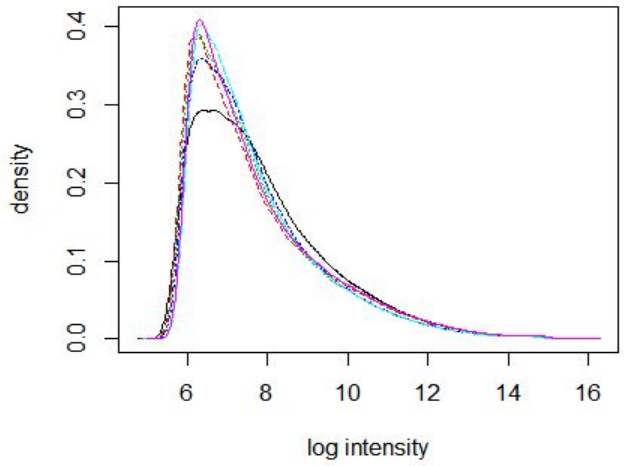

Fig1. CEL files densities before background correction or normalisation.

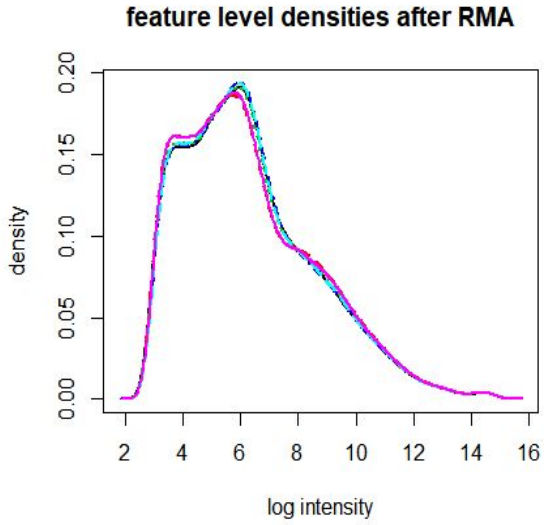

Fig2. Feature level densities after RMA

To identify differentially expressed genes, data from every transcript as a response is fit into a linear model using limma (Ritchie, 2015). However, fitting into a linear model still does not tell us if there are differential expression of the genes in the presence or absence of miRNA-203. Hence, contrasts have to be specified and in this case, since the presence or absence of miRNA-203 differs in each sample, we created a contrast with that as the basis of comparison. To do this, a model matrix must be generated using the function 'model.matrix' to specify the model design. Table 2 below shows the model matrix which shows the two different conditions possible in each sample, which is the control (absence of miRNA-203 in breast cancer cell) and presence of miRNA-203, which is the model design.

	control	miR203
1	1	0
2	1	0
3	1	0
4	0	1
5	0	1
6	0	1

Table 2. Model matrix showing the design and characteristics of each sample.

Following which, a contrast matrix can be generated to specify the comparison of interest, using the function ‘makeContrasts’. Since we are interested in the difference of gene expression between the control and the presence of miRNA-203 in breast cancer cells, we used control minus miRNA-203, with the former being the reference. This means that when the log fold change is positive, gene expression is greater in the control than when miRNA-203 is present, and vice versa.

Contrasts	
Levels	control - miR203
control	1
miR203	-1

Table 3. Contrast matrix where a positive value denotes a higher gene expression in control, while a negative value denotes a higher gene expression in samples containing miRNA-203.

Using both the model and contrast matrix, fitting them into the data is made possible using the function ‘lmFit’. Empirical Bayes correction can then be carried out to shrink the estimated variances of individual probe sets towards the overall distribution among them, using the function ‘eBayes’ available on limma. Using topTable, we can visualize the differentially expressed genes.

To select key genes that might be the cause of cancer, we can use the AnnotationDBI Interface (Pagès, 2018). We used hgu133plus2.db (Carlson, 2016) and selected ‘ProbeID’ as the keytype to see the gene name and other information of the interesting genes. To select the genes that have the most changes in expression due to the presence of miRNA-203, we set the limit using p-value of less than 0.10, and with a fold change of at least two (log fold change at least one). This is because we are interested in the genes that shows a double or half in the expression when miRNA-203 is present.

To visualize the results of differentially expressed genes, we can use a volcano plot, that shows the log fold-change versus the negative log p-value (Goehlmann, 2009). Genes that have larger fold changes usually will have larger negative log p-values that will signify a greater statistical confidence. However, to better view the interesting genes that pass our cutoff, we can change the colour of these plots using R features.

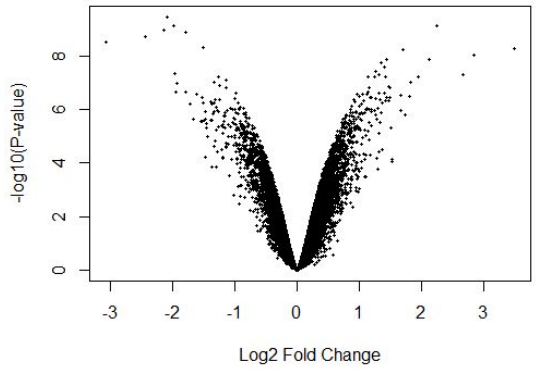

Figure 3. Volcano plot of differentially expressed genes.

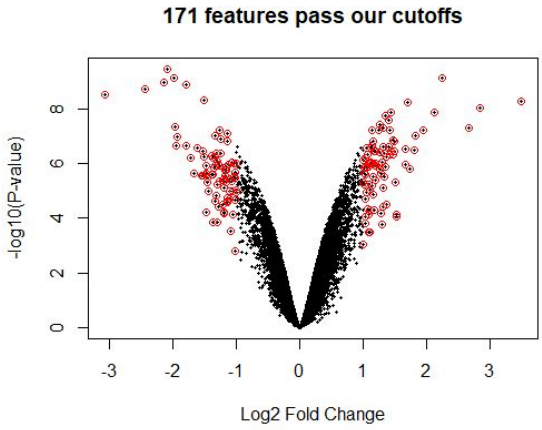

Figure 4. Volcano plot with interesting genes highlighted.

From Figure 3 and 4, we have identified 171 interesting genes that pass the cutoff of p-value less than 0.10 and with a fold change of at least 2. These genes can be further analyzed in terms of the pathways they regulate and how these genes can result in the development of cancer.

Other than volcano maps, heatmaps can also be used to visualize distinct trends in gene expression patterns between different experimental conditions by clustering. Red spots on heatmaps typically means that there is a high expression of the genes. Heatmaps will also generate dendrograms on the top and left that represent the rows and columns ordering which is done using Euclidean distance. However, instead of using Euclidean distance, correlation is more useful for clustering as we only have 2 different types of characteristics of the sample.

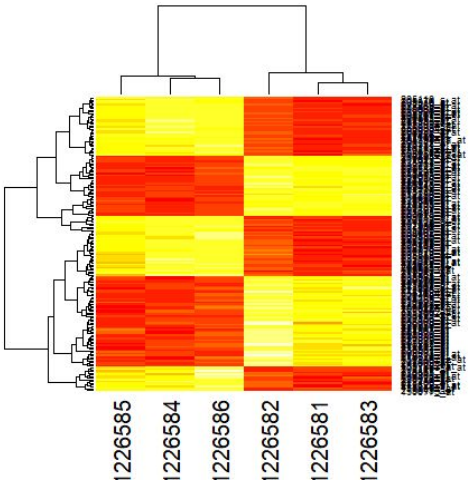

Figure 5. Heatmap generated using Euclidean distance.

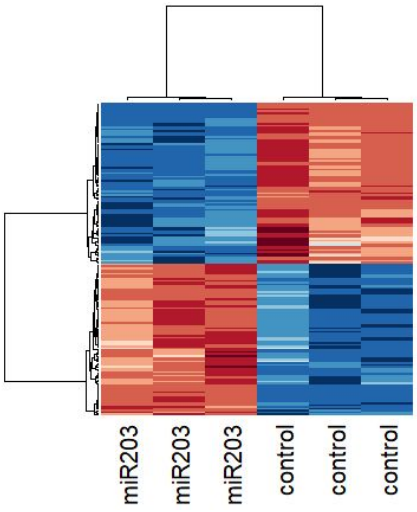

Figure 6. Heatmap generated using correlation.

3. Results and Discussion

From the csv file that we output from R called “recommendations.csv”, we get the list of 171 genes possibly regulated by miR-203. We input the file containing the ENTREZID into DAVID annotations (Huang, 2009) to get the gene orthology of the genes, namely the biological processes, cellular components, molecular functions, biological biochemical image database, BioCarta, and KEGG pathways.

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_BP_DIRECT	chemokine-mediated signaling pathway	RT		5	3.4	1.6E-3	7.7E-1
	GOTERM_BP_DIRECT	positive regulation of T cell migration	RT		3	2.0	2.1E-3	6.3E-1
	GOTERM_BP_DIRECT	wound healing	RT		5	3.4	2.4E-3	5.3E-1
	GOTERM_BP_DIRECT	extracellular matrix organization	RT		7	4.7	2.6E-3	4.5E-1
	GOTERM_BP_DIRECT	immune response	RT		10	6.7	2.9E-3	4.1E-1
	GOTERM_BP_DIRECT	regulation of cardiac muscle cell contraction	RT		3	2.0	3.1E-3	3.8E-1
	GOTERM_BP_DIRECT	induction of positive chemotaxis	RT		3	2.0	4.8E-3	4.7E-1
	GOTERM_BP_DIRECT	cell adhesion	RT		10	6.7	5.1E-3	4.5E-1
	GOTERM_BP_DIRECT	positive regulation of calcium ion import	RT		3	2.0	5.5E-3	4.3E-1
	GOTERM_BP_DIRECT	negative regulation of cell proliferation	RT		9	6.0	6.9E-3	4.7E-1
	GOTERM_BP_DIRECT	positive regulation of leukocyte chemotaxis	RT		3	2.0	7.0E-3	4.4E-1
	GOTERM_BP_DIRECT	intracellular signal transduction	RT		9	6.0	7.6E-3	4.4E-1
	GOTERM_BP_DIRECT	cellular sodium ion homeostasis	RT		3	2.0	7.7E-3	4.2E-1
	GOTERM_BP_DIRECT	regulation of cardiac muscle contraction by regulation of the release of sequestered calcium ion	RT		3	2.0	7.7E-3	4.2E-1
	GOTERM_BP_DIRECT	cell chemotaxis	RT		4	2.7	1.1E-2	5.1E-1
	GOTERM_BP_DIRECT	chemotaxis	RT		5	3.4	1.1E-2	4.9E-1
	GOTERM_BP_DIRECT	regulation of synaptic transmission, glutamatergic	RT		3	2.0	1.1E-2	4.8E-1
	GOTERM_BP_DIRECT	positive regulation of release of sequestered calcium ion into cytosol	RT		3	2.0	1.5E-2	5.7E-1
	GOTERM_BP_DIRECT	positive regulation of peptidyl-tyrosine phosphorylation	RT		4	2.7	2.0E-2	6.5E-1
	GOTERM_BP_DIRECT	regulation of endodermal cell fate specification	RT		2	1.3	2.1E-2	6.4E-1
	GOTERM_BP_DIRECT	cell-cell signaling	RT		6	4.0	3.4E-2	7.9E-1
	GOTERM_BP_DIRECT	MAPK cascade	RT		6	4.0	3.8E-2	8.1E-1
	GOTERM_BP_DIRECT	protein folding	RT		5	3.4	3.8E-2	8.1E-1
	GOTERM_BP_DIRECT	positive regulation of cAMP metabolic process	RT		2	1.3	4.1E-2	8.2E-1

GOTERM_BP_DIRECT	cellular response to tumor necrosis factor	RT	4	2.7	4.2E-2	8.1E-1
GOTERM_BP_DIRECT	positive regulation of cell division	RT	3	2.0	4.3E-2	8.0E-1
GOTERM_BP_DIRECT	cellular response to lipopolysaccharide	RT	4	2.7	4.5E-2	8.1E-1
GOTERM_BP_DIRECT	positive regulation of cell proliferation	RT	8	5.4	4.6E-2	8.0E-1
GOTERM_BP_DIRECT	regulation of cardiac muscle contraction by calcium ion signaling	RT	2	1.3	4.8E-2	8.0E-1
GOTERM_BP_DIRECT	negative regulation of cell division	RT	2	1.3	4.8E-2	8.0E-1
GOTERM_BP_DIRECT	inflammatory response	RT	7	4.7	5.1E-2	8.1E-1
GOTERM_BP_DIRECT	cellular response to cAMP	RT	3	2.0	5.2E-2	8.1E-1
GOTERM_BP_DIRECT	basement membrane organization	RT	2	1.3	5.5E-2	8.1E-1
GOTERM_BP_DIRECT	T cell chemotaxis	RT	2	1.3	5.5E-2	8.1E-1
GOTERM_BP_DIRECT	ion transmembrane transport	RT	5	3.4	6.1E-2	8.4E-1
GOTERM_BP_DIRECT	regulation of apoptotic process	RT	5	3.4	6.3E-2	8.4E-1
GOTERM_BP_DIRECT	cytokine-mediated signaling pathway	RT	4	2.7	6.5E-2	8.4E-1
GOTERM_BP_DIRECT	negative regulation of cytosolic calcium ion concentration	RT	2	1.3	6.8E-2	8.4E-1
GOTERM_BP_DIRECT	positive regulation of T cell differentiation in thymus	RT	2	1.3	6.8E-2	8.4E-1
GOTERM_BP_DIRECT	sequestering of actin monomers	RT	2	1.3	6.8E-2	8.4E-1
GOTERM_BP_DIRECT	response to auditory stimulus	RT	2	1.3	6.8E-2	8.4E-1
GOTERM_BP_DIRECT	creatine metabolic process	RT	2	1.3	7.5E-2	8.6E-1
GOTERM_BP_DIRECT	regulation of release of sequestered calcium ion into cytosol	RT	2	1.3	7.5E-2	8.6E-1
GOTERM_BP_DIRECT	positive regulation of cAMP-mediated signaling	RT	2	1.3	8.1E-2	8.8E-1
GOTERM_BP_DIRECT	negative regulation of bone resorption	RT	2	1.3	8.1E-2	8.8E-1
GOTERM_BP_DIRECT	relaxation of cardiac muscle	RT	2	1.3	8.1E-2	8.8E-1
GOTERM_BP_DIRECT	cell communication by electrical coupling involved in cardiac conduction	RT	2	1.3	8.8E-2	8.9E-1
GOTERM_BP_DIRECT	membrane depolarization during cardiac muscle cell action potential	RT	2	1.3	8.8E-2	8.9E-1
GOTERM_BP_DIRECT	cellular response to interleukin-1	RT	3	2.0	8.9E-2	8.9E-1
GOTERM_BP_DIRECT	negative regulation of keratinocyte proliferation	RT	2	1.3	9.4E-2	9.0E-1
GOTERM_BP_DIRECT	negative regulation of smooth muscle cell migration	RT	2	1.3	9.4E-2	9.0E-1

Figure 7. List of biological processes from the 171 genes

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_CC_DIRECT	extracellular region	RT		26	17.4	7.2E-5	1.2E-2
	GOTERM_CC_DIRECT	extracellular space	RT		23	15.4	1.0E-4	8.2E-3
	GOTERM_CC_DIRECT	intercalated disc	RT		4	2.7	3.5E-3	1.7E-1
	GOTERM_CC_DIRECT	external side of plasma membrane	RT		6	4.0	1.5E-2	4.6E-1
	GOTERM_CC_DIRECT	extracellular matrix	RT		7	4.7	1.6E-2	4.0E-1
	GOTERM_CC_DIRECT	basement membrane	RT		4	2.7	1.7E-2	3.6E-1
	GOTERM_CC_DIRECT	extracellular exosome	RT		29	19.5	2.2E-2	4.0E-1
	GOTERM_CC_DIRECT	T-tubule	RT		3	2.0	2.8E-2	4.3E-1
	GOTERM_CC_DIRECT	keratin filament	RT		4	2.7	3.1E-2	4.3E-1
	GOTERM_CC_DIRECT	apical plasma membrane	RT		6	4.0	4.9E-2	5.5E-1
	GOTERM_CC_DIRECT	membrane	RT		22	14.8	6.8E-2	6.4E-1
	GOTERM_CC_DIRECT	Cul3-RING ubiquitin ligase complex	RT		3	2.0	7.6E-2	6.5E-1
	GOTERM_CC_DIRECT	intrinsic component of the cytoplasmic side of the plasma membrane	RT		2	1.3	9.7E-2	7.2E-1

Figure 8. List of cellular components from the 171 genes

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_MF_DIRECT	chemokine activity	RT		5	3.4	3.9E-4	8.4E-2
	GOTERM_MF_DIRECT	growth factor activity	RT		7	4.7	1.0E-3	1.1E-1
	GOTERM_MF_DIRECT	CXCR3 chemokine receptor binding	RT		2	1.3	3.5E-2	9.3E-1
	GOTERM_MF_DIRECT	transferase activity, transferring phosphorus-containing groups	RT		2	1.3	3.5E-2	9.3E-1
	GOTERM_MF_DIRECT	creatine kinase activity	RT		2	1.3	4.2E-2	9.1E-1
	GOTERM_MF_DIRECT	collagen binding	RT		3	2.0	6.7E-2	9.6E-1

Figure 9. List of molecular functions from the 171 genes

3 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	BBID	109.Chemokine_families	RT		5	3.4	1.1E-3	2.5E-2
	BBID	22.Cytokine-chemokine_CNS	RT		3	2.0	2.2E-2	2.3E-1
	BBID	58.(CD40L)_immnosurveillance	RT		3	2.0	3.9E-2	2.6E-1

Figure 10. List of BBID from the 171 genes

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	BIOCARTA	Cells and Molecules involved in local acute inflammatory response	RT		3	2.0	1.6E-2	5.4E-1
	BIOCARTA	Cytokine Network	RT		3	2.0	2.6E-2	4.7E-1
	BIOCARTA	Cytokines and Inflammatory Response	RT		3	2.0	4.4E-2	5.1E-1

Figure 11. List of BIOCARTA from the 171 genes

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Rheumatoid arthritis	RT		6	4.0	9.6E-4	1.2E-1
	KEGG_PATHWAY	Cytokine-cytokine receptor interaction	RT		8	5.4	4.9E-3	2.8E-1
	KEGG_PATHWAY	Amoebiasis	RT		5	3.4	1.3E-2	4.5E-1
	KEGG_PATHWAY	PI3K-Akt signaling pathway	RT		8	5.4	2.9E-2	6.2E-1
	KEGG_PATHWAY	Non-alcoholic fatty liver disease (NAFLD)	RT		5	3.4	4.2E-2	6.8E-1
	KEGG_PATHWAY	Toll-like receptor signaling pathway	RT		4	2.7	6.5E-2	7.7E-1
	KEGG_PATHWAY	TNF signaling pathway	RT		4	2.7	6.6E-2	7.2E-1
	KEGG_PATHWAY	Chemokine signaling pathway	RT		5	3.4	7.8E-2	7.4E-1

Figure 12. List of KEGG_PATHWAY from the 171 genes

There are 6 main critical endogenous pathways of Epithelial to Mesenchymal Transition (EMT) in breast cancer cells which are the TGF-β pathway, the Wnt pathway, the Notch pathway, the Hippo pathway, the Hedgehog pathway and pathways emanating from receptor tyrosine kinases (Fedele, 2017). Using these known pathways that have an important effect on breast cancer cells during EMT, we were able to narrow down the list of candidate genes to those that directly take part in these pathways.

EntrezID	**Gene Name**	**Pathway**	**Adjusted P-value**
894	cyclin D2(CCND2)	Wnt signalling pathway, Hippo pathway	5.992960e-05
22943	dickkopf WNT signaling pathway inhibitor 1(DKK1)	Wnt signalling pathway	1.411730e-04
8613	phospholipid phosphatase 3(PLPP3)	Wnt signalling pathway	3.437873e-04
50964	sclerostin (SOST)	Wnt signalling	1.341506e-03

		pathway	
1004	cadherin 6(CDH6)	Notch signalling pathway	7.183258e-04

Table 4. Genes that directly take part in pathways in EMT known to be critical in breast cancers, along with their adjusted p-values.

From the R code, we use Padj of 0.10 which means that there is an estimated 1/10 chance of single false positive in the analysis experiment. The 5 genes that we narrowed down have adjusted p-values less than 0.10. False discovery rate (FDR) is a method to conceptualise the rate of type 1 error (false positive) in null hypothesis testing when conducting multiple comparisons of genes. The closer the adjusted p-value is to 0.10, the less significant it is to have transcriptomic changes driven by miR203 during epithelial-mesenchymal transition (EMT) as it is more likely to be a false positive, and vice versa.

In order to rank the 5 genes in terms of how they are significantly linked to breast cancer model, we identified the genes using their ENTREZID and map it back to the list of "differentially expressed" list in the R code containing their respective adjusted p-values as shown in Table 4. The ranking is as follows, 1st being the most significant and 5th being the least significant amongst them:

1. cyclin D2 (CCND2)
2. dickkopf WNT signaling pathway inhibitor 1 (DKK1)
3. phospholipid phosphatase 3 (PLPP3)
4. cadherin 6 (CDH6)
5. sclerostin (SOST)

The Wnt pathway plays a critical role in the development and progression of breast cancer (Lamb, 2013) as it promotes cell migration and EMT. There are 2 branches of the Wnt pathway which are the canonical and the non-canonical pathways (Fedele, 2017). DKK protein, which inhibits the canonical Wnt pathway, are frequently silenced in breast cancer (Suzuki, 2008). Hence, expression of DKK1 is most likely affected by the dysregulation of miRNA-203, making it unable to block the Wnt pathway which promotes cell migration and EMT which leads to breast cancer. Also, it is highly likely that SOST is highly affected by the dysregulation of miRNA-203 as from another research conducted, it was found that DKK1 and SOST have opposing roles in Wnt signalling pathway (Hudson, 2015). Therefore, these genes might be key genes that are differentially regulated by miRNA-203 where the expression of DKK1 and SOST occurs in opposite direction.

Another gene that is also found to be involved in the Wnt signaling pathway is the PLPP3 gene. A major function of PLPP3 is dephosphorylation of extracellular lysophosphatidic acid, a phospholipid with growth factor-like activity that stimulates tumour cell migration and invasion (Bowlt, 2018). Hence, PLPP3 might be dysregulated by the dysregulation of miRNA-203 that stimulates cancer growth.

The Notch pathway is activated downstream of the Ras and Wnt pathways (Ayyanan, 2006) and it plays an important role in metastasis (Guo, 2011), induces EMT (Leong, 2007) and promotes malignant breast cancer (Li, 2014). Cadherin switch that takes place during EMT in tumour cells changes E-cadherin to N-cadherin induces heightened motility, invasion and metastasis (Hazan, 2004). CDH6 which was found to be involved in the Notch pathway, was also recently found to be TGF-β target (Sancisi, 2013). Hence, CDH6 gene might be a key gene that is regulated by miRNA-203 which affects the Notch pathway and possibly the TGF-β pathway and leads to cancer.

Lastly, the Hippo tumour suppressor pathway consists of a large network of proteins that play important regulatory functions during organ development and regeneration (Fedele, 2017). CCND2 which was found to be involved in the Hippo pathway was also found to play a critical role in cell cycle progression and tumorigenicity of glioblastoma stem cells (GSCs) (Koyama-Nasu, 2013). Hence, CCND2 might also be a key gene that, when miRNA-203 is dysregulated, will promote cancer.

4. Conclusion

In conclusion, from the GSE50697 series, we would recommend 5 candidate genes for qPCR validation in the lab. They are namely cyclin D2(CCND2), dickkopf WNT signaling pathway inhibitor 1(DKK1), phospholipid phosphatase 3(PLPP3), cadherin 6(CDH6), and sclerostin (SOST). It is in decreasing order of how significant they are linked to breast cancer driven by miRNA-203.

These 5 candidate genes were selected via rounds of data selection process. The first filter was via our Rcode where 171 interesting genes passed our cutoff of p-value less than 0.10 with a log fold change of 1, as labelled red in our volcano plot in Figure 4, followed by the use of DAVID annotations to find out their gene orthology for further filtration. From thorough research on EMT, there are 6 vital pathways in breast cancer cells which are TGF-β pathway, Wnt pathway, Notch pathway, Hippo pathway, Hedgehog

pathway, and pathways emanating from receptor tyrosine kinases. These pathways are then used to filter out genes which are involved in the KEGG pathways mentioned above.

Lastly, research was done to analyse how the 5 genes affect the respective pathways which can lead to miRNA-203 driven transcriptomic changes during epithelial-mesenchymal transition (EMT). DKK1 and SOST were found to have opposing roles in Wnt signalling pathway. PLPP3 which is also involved in this pathway stimulates tumour cell migration and invasion, which is dysregulated during cancer. CDH6 is found to be a target of TGF-β target in the Notch pathway, which induce metastasis and EMT, leading to breast cancer. Finally, CCND2 is vital in cell cycle progression and the growth of tumor in stem cells controlled by the Hippo tumour suppressor pathway.

However, there are some limitations to our study. Firstly, due to the false discovery rate, we have false positives in our results and hence, we cannot be very sure that the list of genes that was generated through the analysis are definitely regulated by miRNA-203 which lowers the confidence of the test. Also, another limitation is that we used known critical pathways of EMT in breast cancer to narrow down the list of candidate genes to 5 genes. However, these 5 genes do not cover all 6 known critical pathways and therefore, it is possible that we might miss out other genes that are involved in the pathways that are also regulated by miRNA-203. Furthermore, narrowing down the list of genes is not limited to this method alone, and there are other various methods based on different criterias to produce a different set of genes.

Hence, more extensive studies can be done to ensure that all possible methods of filtering the genes are taken into account during the analysis process. The overlapping genes found from various filtering methods will definitely give us a more accurate list of genes which have higher confidence levels, and thus will be more appropriate as candidate genes for qPCR validation in the lab.

5. Bibliography

Ayyanan, A., Civenni, G., Ciarloni, L., Morel, C., Mueller, N., Lefort, K., … Brisken, C. (2006). Increased Wnt signaling triggers oncogenic conversion of human breast epithelial cells by a Notch-dependent mechanism. Proceedings of the National Academy of Sciences, 103(10), 3799–3804. https://doi.org/10.1073/pnas.0600065103

Bowlt Blacklock, K., Birand, Z., Biasoli, D., Fineberg, E., Murphy, S., Flack, D., … Starkey, M. (2018). Identification of molecular genetic contributants to canine cutaneous mast cell tumour metastasis by global gene expression analysis. PloS One, 13(12), e0208026. https://doi.org/10.1371/journal.pone.0208026

Chang, K.-M., Harbron, C., & South, M. C. (n.d.). An Exploration of Extensions to the RMA Algorithm. Retrieved from http://ludwig-sun2.unil.ch/~darlene/pub.html#Sub

Davis S, Meltzer P (2007). "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor." *Bioinformatics*, **14**, 1846–1847.

Fedele, M., Cerchia, L., & Chiappetta, G. (2017). The Epithelial-to-Mesenchymal Transition in Breast Cancer: Focus on Basal-Like Carcinomas. Cancers, 9(10). https://doi.org/10.3390/cancers9100134

Gautier, L., Cope, L., Bolstad, B. M., & Irizarry, R. A. (2004). affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics, 20(3), 307–315. https://doi.org/10.1093/bioinformatics/btg405

Gentleman, R. (2005). Bioinformatics and computational biology solutions using R and Bioconductor. Springer Science+Business Media.

Goehlmann, H. and W. Talloen (2009). Gene Expression Studies Using Affymetrix Microarrays, Chapman \& Hall/CRC, pp. 148 - 153.

Guo, S., Liu, M., & Gonzalez-Perez, R. R. (2011). Role of Notch and its oncogenic signaling crosstalk in breast cancer. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer, 1815(2), 197–213. https://doi.org/10.1016/j.bbcan.2010.12.002

Hazan, R. B., Qiao, R., Keren, R., Badano, I., & Suyama, K. (2004). Cadherin switch in tumor progression. Annals of the New York Academy of Sciences, 1014, 155–163. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15153430

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols, 4(1), 44–57. https://doi.org/10.1038/nprot.2008.211

Hudson, B. D., Hum, N. R., Thomas, C. B., Kohlgruber, A., Sebastian, A., Collette, N. M., … Loots, G. G. (2015). SOST Inhibits Prostate Cancer Invasion. PLOS ONE, 10(11), e0142058. https://doi.org/10.1371/journal.pone.0142058

Koyama-Nasu, R., Nasu-Nishimura, Y., Todo, T., Ino, Y., Saito, N., Aburatani, H., … Akiyama, T. (2013). The critical role of cyclin D2 in cell cycle progression and tumorigenicity of glioblastoma stem cells. Oncogene, 32(33), 3840–3845. https://doi.org/10.1038/onc.2012.399

Lamb, R., Ablett, M. P., Spence, K., Landberg, G., Sims, A. H., & Clarke, R. B. (2013). Wnt Pathway Activity in Breast Cancer Sub-Types and Stem-Like Cells. PLoS ONE, 8(7), e67811. https://doi.org/10.1371/journal.pone.0067811

Leong, K. G., Niessen, K., Kulic, I., Raouf, A., Eaves, C., Pollet, I., & Karsan, A. (2007). Jagged1-mediated Notch activation induces epithelial-to-mesenchymal transition through Slug-induced repression of E-cadherin. The Journal of Experimental Medicine, 204(12), 2935–2948. https://doi.org/10.1084/jem.20071082

Li, L., Zhao, F., Lu, J., Li, T., Yang, H., Wu, C., & Liu, Y. (2014). Notch-1 Signaling Promotes the Malignant Features of Human Breast Cancer through NF-κB Activation. PLoS ONE, 9(4), e95912. https://doi.org/10.1371/journal.pone.0095912

Macfarlane, L.-A., & Murphy, P. R. (2010). MicroRNA: Biogenesis, Function and Role in Cancer. Current Genomics, 11(7), 537–561. https://doi.org/10.2174/138920210793175895

Pagès H, Carlson M, Falcon S, Li N (2018). *AnnotationDbi: Annotation Database Interface*. R package version 1.44.0.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). “limma powers differential expression analyses for RNA-sequencing and microarray studies.” *Nucleic Acids Research*, **43**(7), e47.

Sancisi, V., Gandolfi, G., Ragazzi, M., Nicoli, D., Tamagnini, I., Piana, S., & Ciarrocchi, A. (2013). Cadherin 6 Is a New RUNX2 Target in TGF-β Signalling Pathway. PLoS ONE, 8(9), e75489. https://doi.org/10.1371/journal.pone.0075489

Suzuki, H., Toyota, M., Caraway, H., Gabrielson, E., Ohmura, T., Fujikane, T., … Tokino, T. (2008). Frequent epigenetic inactivation of Wnt antagonist genes in breast cancer. British Journal of Cancer, 98(6), 1147–1156. https://doi.org/10.1038/sj.bjc.6604259

6. Appendix

Full Rcode with results and diagrams:
https://docs.google.com/document/d/1WfvKdyixAuBND8gf61fSylBUW_Da78q0c03MgzuiYBQ/edit?usp=sharing

Gene_of_interest.csv:
https://drive.google.com/file/d/1HNRM7lhTnhsUT4yFaz5EAkFSJU5rCOD9/view?usp=sharing

171 genes recommendations.csv:
https://drive.google.com/file/d/1-MBG9ljSTigjV5TiT5Hll6P7tnEsKtMc/view?usp=sharing

```
#LSM3241 CA1 R Code
#install Bioconductor
if (!requireNamespace("BiocManager"))
  install.packages("BiocManager")
BiocManager::install()

#install Bioconductor packages
BiocManager::install('foo')

#install relevant packages needed
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("GEOquery", version = "3.8")

BiocManager::install("affy")
BiocManager::install("limma")
BiocManager::install("hgu133plus2.db")
BiocManager::install("org.Hs.eg.db")

#load packages installed
library(affy)
library(limma)
library(hgu133plus2.db)
library(org.Hs.eg.db)

#check if packages are loaded
sessionInfo()
```

```
#Calling GSE50697
library(GEOquery)
#downloading the gse file
gse <- getGEO('GSE50697',GSEMatrix = FALSE)

#retrieving whole GSE
names(GSMList(gse))

#Getting the raw data for the series using GEOquery
filePaths <- getGEOSuppFiles('GSE50697')
#generating a vector containing the name of all the CEL files.
list.celfiles('GSE50697_RAW')

#finding info from meta data of gse
names(Meta(gse))

#creating GSM object names
gsm <- GSMList(gse)[[1]]

#find information we need from the metadata of gsm
names(Meta(gsm))

#extract metadata of gsm not in gse
names(Meta(gsm))[!(names(Meta(gsm)) %in% names(Meta(gse)))]
Meta(gsm)[!(names(Meta(gsm)) %in% names(Meta(gse)))]

#From the metadata information, we decided to use the following elements
#source_name_ch1 and characteristics_ch1
#getting sample growth conditions into a data frame
culture_medium <- function(gsm) {
  Meta(gsm)[['characteristics_ch1']][2]
}
sapply(GSMList(gse),culture_medium)
pd <- data.frame(culture=as.factor(sapply(GSMList(gse),culture_medium)))
pd

#simplifying our dataframe, convert columns to 2 values only
pd$culture <- as.factor(pd$culture)
levels(pd$culture) <- c("control","miR203")
#to enable the kable function, we need to install the knitr package
install.packages("knitr")
library(knitr)
kable(pd)
```

```
#Reading in the CEL files with the phenoData
celfiles <- paste0('GSE50697_RAW/', list.celfiles('GSE50697_RAW/'),'.')
affydata <- read.affybatch(celfiles,phenoData = new("AnnotatedDataFrame",pd))
phenoData(affydata)

#pseudo images of chips
image(data[,1])
image(data[,2])
image(data[,3])
image(data[,4])
image(data[,5])
image(data[,6])


#CEL file densities before normalisation or background correction
plotDensity.AffyBatch(affydata)
#perform RMA normalisation
eset <- rma(affydata)
plotDensity(exprs(eset),xlab='log intensity',main="feature level densities after RMA",lwd=2)
#phenotype data for each samples after rma is retained
pData(eset)
#generate model matrix
model <- model.matrix( ~ 0 + eset$culture)
#rename the model columns to correspond to the different growth conditions
colnames(model) <- levels(eset$culture)
model

#look at when the growth conditions differ, we create contrast
contrasts <- makeContrasts(control - miR203, levels=model)
contrasts

#fit model and contrast matrix into a data
fit <- lmFit(eset, model)
fit


#calculate test statistics via performing eBayes correction
fitted.ebayes <- eBayes(fitted.contrast)
fitted.ebayes

#extracting differentially expressed genes
topTable(fitted.ebayes)
```

```
#Get a limited number of probesets
ps <- rownames(topTable(fitted.ebayes))
ps

#The AnnotationDbi interface
#look at available columns for our chip
columns(hgu133plus2.db)

#Look at which that can be used as keys
keytypes(hgu133plus2.db)

#realise all can be used as keys, choose "PROBEID" as most suitable
head(keys(hgu133plus2.db,keytype="PROBEID"))

AnnotationDbi::select(hgu133plus2.db,ps,c("SYMBOL","ENTREZID","GENENAME"),keytype="PR
OBEID")

#Restrict to upregulated genes
differentially_expressed <- topTable(fitted.ebayes,number=Inf,p.value=0.1,lfc=1)
differentially_expressed

upregulated <- differentially_expressed[differentially_expressed$logFC > 0,]
genes_of_interest <- AnnotationDbi::select(hgu133plus2.db,
                                           keys=rownames(upregulated),
                                           columns=c("SYMBOL","ENTREZID","GENENAME"),
                                           keytype="PROBEID")
genes_of_interest
#save data into a csv file
write.csv(genes_of_interest,'genes_of_interest.csv')

#volcanoplots
volcanoplot(fitted.ebayes)

interesting_genes <- topTable(fitted.ebayes,number=Inf,p.value = 0.1,lfc=1)
volcanoplot(fitted.ebayes, main=sprintf("%d features pass our cutoffs",nrow(interesting_genes)))

points(interesting_genes[['logFC']],-log10(interesting_genes[['P.Value']]),col='red')

#heatmaps
#normalise expression value
```

```
eset_of_interest <- eset[rownames(interesting_genes),]
heatmap(exprs(eset_of_interest))

#fix and beautify heatmap
install.packages("RColorBrewer")
library(RColorBrewer)
eset_of_interest <- eset[rownames(interesting_genes),]
heatmap(exprs(eset_of_interest),
      labCol=eset$culture,labRow=NA,
      col       = rev(brewer.pal(10, "RdBu")),
      distfun   = function(x) as.dist(1-cor(t(x))))
#Results
recommendations <- AnnotationDbi::select(hgu133plus2.db,
                                  keys=rownames(interesting_genes),
                                  columns=c("SYMBOL","ENTREZID","GENENAME"),
                                  keytype="PROBEID")
write.csv(recommendations,'recommendations.csv')
```