Gene Set Analysis in Myelodysplastic Syndrome

Aparna Gopal, Lauren Martin, Jewel Ocampo, Sherrie Wang, Xuan Wang

STAT 540 Final Project

BACKGROUND

Myelodysplastic syndromes (MDS) are a heterogeneous group of hematopoietic stem cell (HSC) disorders characterized by defective haematopoiesis and cytopenia, associated with an increased risk of transformation to acute myeloid leukemia (AML). Various classification schemes exist to divide MDS patients into low-risk MDS and high-risk MDS for predicting the risk of transformation to AML, based on the percentage of blasts in the bone marrow or blood, karyotype, and number of cytopenia present [1, 2].

Even low-risk MDS is associated with significant morbidity and mortality, with the median survival for low-risk MDS patients ranging from 5.3 to 8.8 years [3]. The median survival for high-risk MDS patients ranges from 0.8 to 1.6 years [3]. Thus, understanding molecular mechanisms involved specifically in low-risk MDS and high-risk MDS is urgent for development of better treatments for low-risk and high-risk MDS patients, respectively.

Recent evidence suggests that early MDS (low-risk MDS) is characterized by immunodeficiency, apoptosis, and chemokine signalling [4], whereas high-risk MDS showed upregulation of genes related to immature progenitor cells, with deregulated pathways involved in metabolism and DNA repair [5]. Pellagatti et al (2006) [6] demonstrated similarities of MDS gene expression profile to reported interferon-γ-induced gene expression in normal CD34+ cells, and the two most upregulated genes in MDS, are IFIT1 and IFITM1.

We used the raw reads (microarray data) generated by Pellagatti et al (2006) [6] to find differentially expressed genes (DEGs) and the resulting enriched functions in low-risk MDS and high-risk MDS using ErmineJ [7].

OBJECTIVES

1. Find differentially expressed genes in low-risk and high-risk MDS.

2. Perform gene set enrichment analysis to identify enriched functions specifically in low-risk MDS and high-risk MDS.

3. To define a prediction model for the low risk and high risk groups

The data provided from the paper contains microarray data of CD34+ cells from 55 MDS patients, including 37 low-risk MDS and 18 high-risk MDS. Microarray data of CD34+ cells from 11 healthy controls is also included in the study.

METHODS

1. Data was inspected and organized into low-risk MDS and high-risk MDS groups based on information extracted from metadata. Data was normalized and overall data quality was assessed. Data was visualized using heatmaps and dimensionality reduction was performed using PCA.

2. ANOVA analysis was performed to assess differential gene expression between the 3 groups: low-risk, high-risk, and control.

3. Linear regression analysis was performed using limma and the topTable function to obtain the top genes that are differentially expressed between low risk, high risk and control groups. Heatmaps were created for the top differentially expressed genes, sorted by p-value < 0.05 and volcano plots were made to visualize the distribution of the genes.

4. Geneset enrichment analysis was performed using ErmineR and the output from the linear regression analysis to identify different enriched gene sets and pathways among the different groups.

5. A predictive model was generated and cross-validation was performed to test several different methods for the model.

RESULTS

1. Data Quality Control

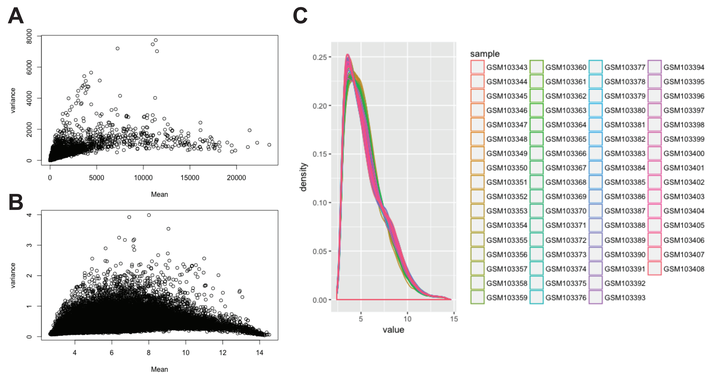

Figure 1: **A**, Scatter plot of Mean vs Variance of untransformed gene expression values. The plot does not display a normal distribution. **B**, Scatter plot of Mean vs Variance of log2 transformed gene expression values. The distribution is now normal. **C**, Density plot showing distribution of gene expression across all samples.

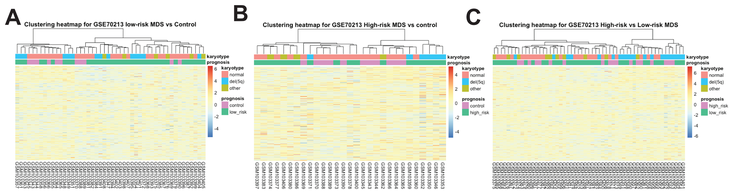

Figure 2: **Clustering heatmaps. A**, Heatmap of gene expression for low-risk MDS samples and control samples. **B**, Heatmap of gene expression for high-risk MDS samples and control samples. **C**, Heatmap of gene expression for low-risk MDS samples and high-risk MDS samples.

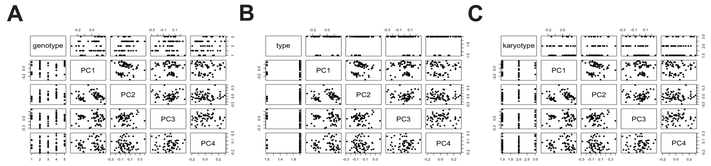

Figure 3: **PCA**. **A**, Scatterplot showing how PC1-PC4 relate to genotype. **B**, Scatterplot showing how PC1-PC4 relate to type (normal or MDS).**C**, Scatterplot showing how PC1-PC4 relate to karyotype. PCA showed that 19% of the variance is explained by PC1, 9.1% is explained by PC2, 5.4% is explained by PC3, 5% is explained by PC4, and the rest of the PCs account for less than 5% of the variance each.

2. Linear Modeling Using Limma

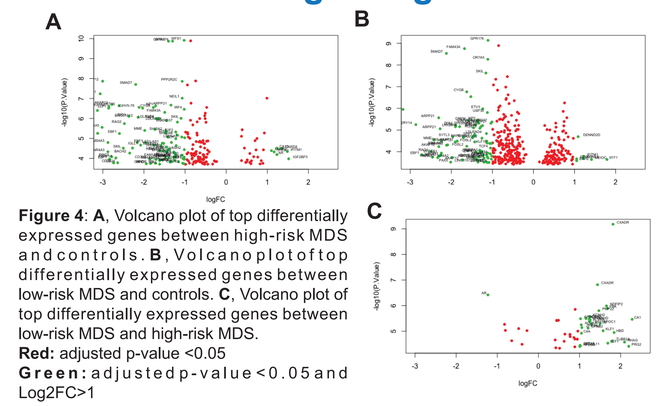

Figure 4: **A**, Volcano plot of top differentially expressed genes between high-risk MDS and controls. **B**, Volcano plot of top differentially expressed genes between low-risk MDS and controls. **C**, Volcano plot of top differentially expressed genes between low-risk MDS and high-risk MDS.
Red: adjusted p-value <0.05
Green: adjusted p-value <0.05 and Log2FC>1

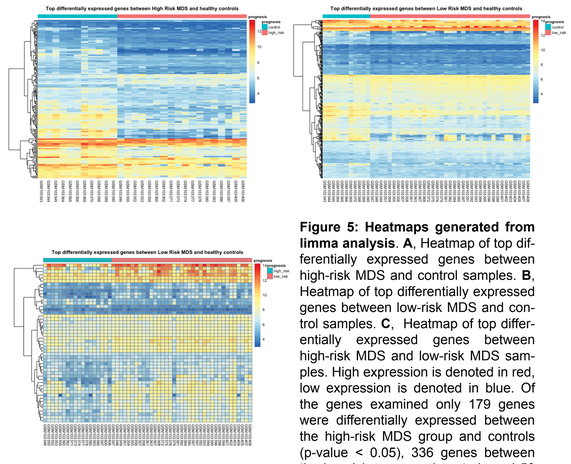

Figure 5: Heatmaps generated from limma analysis. **A**, Heatmap of top differentially expressed genes between high-risk MDS and control samples. **B**, Heatmap of top differentially expressed genes between low-risk MDS and control samples. **C**, Heatmap of top differentially expressed genes between high-risk MDS and low-risk MDS samples. High expression is denoted in red, low expression is denoted in blue. Of the genes examined only 179 genes were differentially expressed between the high-risk MDS group and controls (p-value < 0.05), 336 genes between the low-risk group and controls and 50 genes between the low-risk and high-risk MDS groups.

3. Gene Enrichment Analysis

Name	ID	NumProbes	NumGenes	RawScore	Pval	CorrectedPvalue
immunoglobulin production	GO:0002377	49	49	0.0459047	0e+00	0.0000000
production of molecular mediator of immune response	GO:0002440	67	67	0.0365049	0e+00	0.0000000
humoral immune response	GO:0006959	184	184	0.0452174	0e+00	0.0000000
defense response to bacterium	GO:0042742	200	200	0.0433216	0e+00	0.0000000
cell chemotaxis	GO:0060326	169	169	0.0313467	0e+00	0.0000000

Table 1: Top five enriched genesets in the high-risk MDS samples compared to the control samples based on linear modeling using limma followed by enrichment analysis using ermineR.

Name	ID	NumProbes	NumGenes	RawScore	Pval	CorrectedPvalue
humoral immune response	GO:0006959	184	184	0.0300343	0.0000	0.0000000
drug transport	GO:0015893	132	132	0.0297652	0.0000	0.0000000
leukocyte chemotaxis	GO:0030595	120	120	0.0284002	0.0000	0.0000000
defense response to bacterium	GO:0042742	200	200	0.0257991	0.0000	0.0000000
cell chemotaxis	GO:0060326	169	169	0.0322253	0.0000	0.0000000

Table 2: Top five enriched genesets in the low-risk MDS samples compared to the control samples based on linear modeling using limma followed by enrichment analysis using ermineR.

Name	ID	NumProbes	NumGenes	RawScore	Pval	CorrectedPvalue
cell chemotaxis	GO:0060326	169	169	0.0299976	0e+00	0.0000000
sister chromatid segregation	GO:0000819	177	177	0.0298052	0e+00	0.0000000
humoral immune response	GO:0006959	184	184	0.0289836	0e+00	0.0000000
defense response to bacterium	GO:0042742	200	200	0.0266093	0e+00	0.0000000
chemokine-mediated signaling pathway	GO:0070098	70	70	0.0271420	1e-04	0.0396875

Table 3: Top five enriched genesets in the low-risk MDS samples compared to the high-risk MDS samples based on linear modeling using limma followed by enrichment analysis using ermineR.

4. Predictive Model

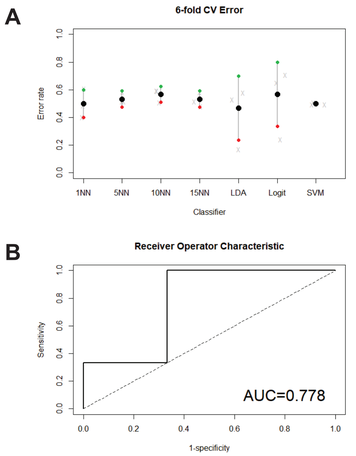

Figure 6: **Predictive model**. **A**, Error rates for 6-fold cross-validation for seven different models: 1NN, 5NN, 10NN, 15NN, LDA, Logit, and SVM. The error rate for our chosen model, 10NN, is 0.5. **B**, ROC curve for 10NN. The AUC is 0.778, which is the highest among the seven different models.

CONCLUSIONS

• Linear analysis using limma allowed identification of differentially expressed genes in low-risk MDS compared to healthy controls (eg: AS1, ZNF782, TSC22D2, PPP2R2C, QDPR, THAP3) as well as in high-risk MDS compared to healthy controls (eg: AS2, RDH10, TERF2, PPP2R2C, CACNB4, BCAS4) and in high-risk MDS compared to low-risk MDS (eg: BICD1, HEATR5B, PTTG3P, FAM129B, CSTF2, ZC3HC1);

• Gene enrichment analysis identified pathways that are enriched in both low-risk and high-risk MDS such as humoral immune responses, cell chemotaxis, and defense responses to bacteria, and also identified pathways that are differentially enriched between low-risk and high-risk MDS, including sister chromatid segregation and the chemokine-mediated signaling pathway;

• The identification of pathways that are differentially enriched between low-risk and high-risk MDS could provide insight into mechanisms that contribute to the higher morbidity and mortality in high-risk MDS patients;

• A prediction model using 10NN was developed to predict whether a patient has low-risk or high-risk MDS based on their gene expression profile, however the model has a high error rate at 0.5, so likely would not be clinically useful at this stage.

REFERENCES

1. Ades et al. Myelodysplastic syndromes, Lancet (2014) 383: 2239-2252. I
2. Corey et al. Myelodysplastic syndromes: complexity of stem-cell diseases, Nature Reviews (2007) 7: 118-129.
3. MDS statistics from American Cancer Society.
4. Pellagatti et al. Deregulated gene expression pathways in myelodysplastic syndrome hematopoietic stem cells, Leukemia (2010) 24: 756-764.
5. Shiozawa et al. Gene expression and risk of leukemic transformation in myelodysplsia, Blood (2017) 130: 2642-2653.
6. Pellagatti et al. Gene expression profiles of CD34+ cells in myelodysplastic syndromes: involvement of interferon-stimulated genes and correlation to FAB subtype and karyotype, Blood (2010) 108: 337-345.
7. Lee et al. ErmineJ: Tool for functional analysis of gene expression datasets. BMC Bioinformatics (2005) 6: 269.