**Differential gene expression between non-mBL and Intermediate patients in the E-GEOD-22470 microarray gene expression dataset**

**Introduction**

An acknowledged difficulty in oncology has been establishing a distinction between Burkitt’s lymphoma (BL) and other mature, aggressive, B-cell lymphomas (BCLs) (Salaverria and Siebert, 2011). Recently, gene-expression profiling (GEP) has enabled the development of an index capable of identifying specific gene expression patterns and development features of molecular BL (mBL), allowing classification of BCLs as possessing either an mBL or non-mBL basis. Nonetheless, many patients exhibit BCLs with unclassifiable gene expression profiles with molecular, morphological, and physiological characteristics indicating both BL and other BCLs, particularly Diffuse large B-cell lymphoma (DLBCL). Although these lymphomas reflect the molecular and morphological features of multiple BCLs, traditional methods for treating either lymphoma cannot be applied in a clinical context. These lymphomas are referred to as molecular intermediate lymphomas, or “intermediates”. In the E-GEOD-22470dataset from “Translocations Activating IRF4 Identify a Subtype of Germinal-Center-Derived B-cell Lymphoma Affecting Predominantly Children and Young Adults”, intermediates refer to BCLs with characteristics of mBL alongside either DLBCL or follicular lymphoma (FL), and non-mBL refer to defined DLBCL or FL.

The focus of this investigation was exploring gene expression between 272 patients in the E-GEOD-22470 cDNA microarray dataset based on their molecular diagnosis as either non-mBL (n=229) or intermediate (n=43). The dataset contained differential expression data on 22,283 probe sets (genes) from the Affymetrix hg133a chip. DLBCL and FL are the 2 most common lymphoma subtypes in western countries and account for up to 36% and 32% of all lymphomas in adults respectively. Therefore, identifying significant differences in gene expression between known non-mBLs (FL and DLBCL) and these intermediates may clarify which genes correspond specifically to the mBL portion of intermediates, allowing them to be more precisely classified.

In order to determine differential expression between these samples. Two- sample T-tests were carried out for each gene to determine if a significant difference existed between non-mBL and intermediates. In addition, the Benjamini and Hochberg False Discovery Rate (FDR) procedure was applied to account for the increased type I error rate in multiple testing.

Beyond this traditional approach, additional analysis was conducted via Significance Analysis of Microarrays (SAM). SAM differs from the above models as it determines significance based on how the observed differential expression of a probe sets deviate from an expected differential expression. This expected differential expression of genes is obtained by random permutations of the original data, which create a theoretical distribution (similar to the T-statistic) representing the random fluctuations between the gene’s expression in either intermediate or non-mBL patients. If the absolute difference between the t-statistics obtained from the differential expression and value obtained from random permutations exceeds the user-defined value, delta (di) for gene i, then the expression is considered to be significant. d is selected by the user depending on an appropriate FDR (Chu et al.). SAM does not require that samples are necessarily independent, normally distributed, or there are small yet significant differences between samples (Larsson et al., 2011). nonparametric testing is not constrained by assumptions about population distribution unlike the T-test above.

**Methods**

All analysis was performed in RStudio (version 1.3.1093) and Bioconductor (version 3.12). The data files and experiment description were downloaded from the EMBL-EBI ArrayExpress site. The raw data files were converted to a data frame, background corrected, and normalized via with the Robust Multi-array Average (RMA) method via the *affy* package*.* Normalized gene expression data was extracted, and the differential gene expression between the “non-mBL” and “intermediate” molecular diagnoses were compared using the Student’s t-test, equal variance was assumed for all probe sets. Multiplicity adjustment was performed by application of the Benjamini and Hochberg false discovery rate correction. The respective gene symbols and IDs for the top 10 results were obtained hgu133a.db database and matched via the package, *annotation*.

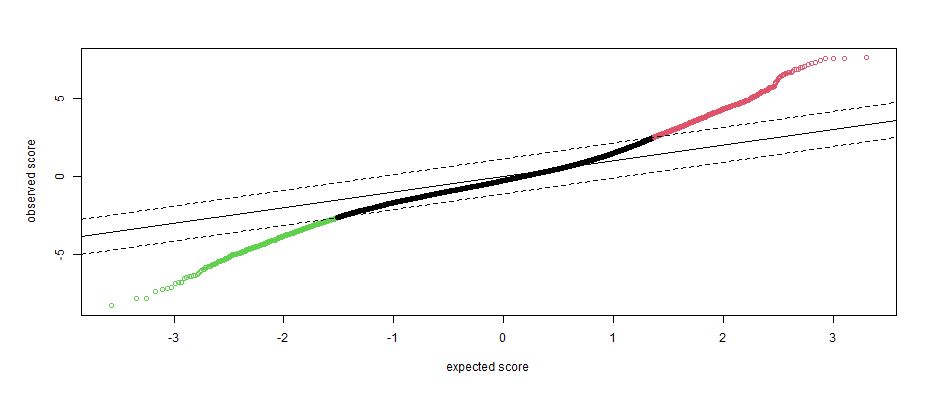
Significance Analysis of Microarrays (SAM) was conducted via the package *samr.* Normalized data and gene IDs were exported to an excel file, and SAM run on this via the SAM user interface. SAM was run with the “array” datatype, “Standard” data analysis and regression method, 100 permutations per gene, and K-nearest neighbors’ algorithm (K=10) run to impute any missing values.

**Results**

The results of the T-test and FDR returned 4,681 significant genes (P-value < 0.05) and 2,027 significant genes (FDR p-value < 0.05) respectively. The results were ranked from most significant to least p-values for the top 10 most differentially expressed genes, and the respective gene names and symbols identified (table 1).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 1: The 10 most differentially expressed genes between patients with non-MBL (n = 229) and intermediate (n =43) molecular diagnoses of BCL:** | | | | | |
| *Probe ID* | *Gene Symbol* | *Gene Name* | *T statistic* | *p-value* | *FDR P-value* |
| 203434\_s\_at | MME | membrane metalloendopeptidase | 8.57 | 7.97E-16 | 1.78E-11 |
| 211862\_x\_at | CFLAR | CASP8 and FADD like apoptosis regulator | -8.11 | 1.79E-14 | 8.01E-11 |
| 215731\_s\_at | MPHOSPH9 | M-phase phosphoprotein 9 | 8.16 | 1.31E-14 | 8.01E-11 |
| 219515\_at | PRDM10 | PR/SET domain 10 | 8.11 | 1.80E-14 | 8.01E-11 |
| 203680\_at | PRKAR2B | protein kinase cAMP-dependent type II regulatory subunit beta | 8.07 | 2.29E-14 | 8.49E-11 |
| 202716\_at | PTPN1 | tyrosine phosphatase non-receptor type 1 | -7.99 | 4.07E-14 | 1.13E-10 |
| 203244\_at | PEX5 | peroxisomal biogenesis factor 5 | 7.99 | 4.07E-14 | 1.13E-10 |
| 204490\_s\_at | CD44 | CD44 molecule (Indian blood group) | -7.82 | 1.18E-13 | 2.20E-10 |
| 213899\_at | METAP2 | methionyl aminopeptidase 2 | 7.82 | 1.19E-13 | 2.20E-10 |
| 222270\_at | PPP4R3B | protein phosphatase 4 regulatory subunit 3B | 7.83 | 1.13E-13 | 2.20E-10 |

SAM analysis was then conducted to determine the most differentially expressed genes. A delta value of 1.12 was selected to produce a 90th percentile FDR of 0.05 (as above) and median FDR of 0.02. At *d* = 1.12, 2083 significant genes were called (fig.1) median percentile for the number of falsely called genes was predicted at 39.14 genes, and the 90th percentile false predicted to 97.1 potential false positives. The results of the SAM analysis are summarized in fig. 1.



Observed score

Expected score

*Figure 1. scatterplot of observed vs expected relative differences (t-statistic) for genes in non-mBL patients compared to intermediate patients***:** Green represents 922 negatively expressed genes and red represents 1162 positively expressed genes in non-mBL compared to intermediate patients. Dashed lines represent the threshold for significance (*d* = 1.12)

The top 10 positively and negatively regulated genes in SAM were identified based on the combination absolute value of the scores, and q-value - the lowest FDR at which the gene is still significant. For the top 10 genes, the Q-value was denoted as 0 by SAM due to their small size (in reality, these are similar to p-values, and are not truly 0). SAM also distinguished if the 10 most differentially expressed genes were positively or negatively in non-MBL compared to intermediate patients. 9 of the 10 original genes were identified in the top 10 negative and positively regulated genes, (appendix table 1) and 6 of the 10 probe sets in the 10 most differentially expressed genes irrespective of positive or negative status (Table 2).

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 2. The top 10 differentially expressed genes from SAM.** Newly identified probe sets identified via SAM are highlighted in red | | | |
| *Probe ID* | *Gene Symbol* | *Regulation (non-mBL compared to intermediate)* | *Score |d|* |
| 203434\_s\_at | MME | Negative | 8.24 |
| 203680\_at | PRKAR2B | Negative | 7.77 |
| 211862\_x\_at | CFLAR | Positive | 7.56 |
| 202716\_at | PTPN1 | Positive | 7.54 |
| 212014\_x\_at | CD44 | Positive | 7.50 |
| 204490\_s\_at | CD44 | Positive | 7.49 |
| 209835\_x\_at | CD44 | Positive | 7.38 |
| 203244\_at | PEX5 | Negative | 7.31 |
| 212063\_at | CD44 | Positive | 7.29 |
| 211316\_x\_at | CFLAR | Positive | 7.21 |

Although 4 different probe sets appeared in the SAM top 10, they did not correspond to new genes, but previously identified gene symbols and names, with 3 new probe sets corresponding to the CD44 gene, and one to the CFLAR gene. The presence of multiple probes sets for one gene indicated that these probe sets identify alternative transcripts of the same genes.

**Discussion and Conclusions**

Although each differentially expressed gene was not evidently linked to distinguishing intermediate and non-mBL, there were several notable discoveries. Firstly, in both T-testing and SAM, MME was identified as the most differentially expressed gene and was negatively expressed in non-mBL with respect to intermediate patients. MME encodes the CD10 marker, the strong expression of which is characteristic of BL’s immunophenotype (Chettiankandy et al., 2016). Although FL and DLBCL also express CD10, FL expressed CD10 in approximately 60% of cases compared to approximately 28% of cases (Dada, 2019, and Ohshima et al., 2001). In this regard, the CD10 expression in intermediates with features of FL and BL may have quantifiable differences in expression levels of CD10 compared BL- DLBCL to intermediates. In a clinical setting, this could help identify which classes of lymphomas constitute intermediates.

CD44, which represented 4 of the 10 most differentially upregulated probes sets. SAM CD44 codes for CD44 transmembrane glycoprotein involved in processes like proliferation, differentiation, and angiogenesis. Evidence has shown transcriptional gene expression profiling of BL generates a molecular signature that includes downregulation of CD44 expression in mBL and mBL-like cancers (Attarbaschi et al., 2007). Moreover. CD44 could be a valuable marker to help classify tumours with morphologic and phenotypic overlap includes BL and DLBCL, as CD44 absence via immunostaining can be used to separate Myc-positive BL from Myc-DLBCL as DLBCL is often strongly CD44 positive (Ayers et al., 2019). In this respect, CD44 could help distinguish DLBCL and BL properties within intermediates and dependent on Myc status.

Lastly, PRDM10, although not directly responsible for the distinction between non-mBL and intermediate pathology, is known to upregulate Bcl-2, a key gene in the development of B-cell lymphomas (Chen et al., 2019). Bcl-2 in particular, Bcl-2 has been proposed as means of distinguishing BL and DLBCL histologically. As PRDM10 is upregulated in non-mBLs compared to intermediates, this tentatively implies increased Bcl-2, which once again could be used as a marker to distinguish the individual contribution of BL and DLBCL to molecular intermediate lymphomas.

With regard to this analysis, there are several points of improvement and future investigation. First, although the false discovery rate has been discovered both in the Hochberg and Benjamini FDR procedure and SAM, there is also the question of false negatives. Estimated miss rates can be given via SAM and describe what percent of false negatives are expected for genes with a particular range of delta values. For example, in this dataset when d = 1.12, there is an estimated miss rate of 82.92% for genes in the score range -2.63 to -1.99, which could lead to excluding meaningful, differentially expressed genes between intermediates and non-mBL patients.

Using SAM, additional features such as the use of the fold-change parameter, which ensures that the significant genes demonstrate a pre-specified change in expression could also be used to determine differences in gene expression. However, although this would highlight large expression differences, genes with large differences in expression are not necessarily biologically meaningful, and specifically in the context of oncogenesis, where small expression changes can lead to pathogenicity. Overall, analysis via T-testing and SAM revealed thousands of significant and differentially expressed genes between non-mBL and intermediate patients which may hold value in distinguishing their molecular basis and treating BCLs more effectively.

**References:**

Attarbaschi, A., Mann, G., Schumich, A., König, M., Pickl, W.F., Haas, O.A., Gadner, H. and Dworzak, M.N. (2007). CD44 deficiency is a consistent finding in childhood Burkitt’s lymphoma and leukemia. *Leukemia*, 21(5), pp.1110–1113.

Chettiankandy, T., Tupkari, J., Kumar, K. and Ahire, M., 2016. B-cell lymphoma, unclassifiable, with features intermediate between diffuse large B-cell lymphoma and classical Burkitt's lymphoma: A case report and review. *Journal of Orla and Maxillofacial Pathology*, 20(2), p.333.

Chu, G., Seo, M., Li, J., Narasimhan, B., Tibshirani, R. and Tusher, V., n.d. *SAM “Significance Analysis of Microarrays”: Users guide and technical document*. Stanford University Department of Biochemistry.

Dada R (June 2019). "Diagnosis and management of follicular lymphoma: A comprehensive review". European Journal of Haematology. 103 (3): 152–163. doi:10.1111/ejh.13271

Larsson, O., Wahlestedt, C. & Timmons, J.A. Considerations when using the significance analysis of microarrays (SAM) algorithm. *BMC Bioinformatics* **6,**129 (2005). https://doi.org/10.1186/1471-2105-6-129

Salaverria, I. and Siebert, R., 2011. The Gray Zone Between Burkitt's Lymphoma and Diffuse Large B-Cell Lymphoma from a Genetics Perspective. *Journal of Clinical Oncology*, 29(14), pp.1835-1843.

Salaverria, I., Philipp, C., Oschlies, I., et al., 2011. Translocations activating IRF4 identify a subtype of germinal center-derived B-cell lymphoma affecting predominantly children and young adults. *Blood*, 118(1), pp.139-147.

Chen, N., Hu, T., Gui, Y., Gao, J., Li, Z. and Huang, S. (2019). Transcriptional regulation of Bcl-2 gene by the PR/SET domain family member PRDM10. *PeerJ*, 7, p.e6941.

Nakamura, N., Nakamine, H., Tamaru, J., Nakamura, S., Yoshino, T., Ohshima, K. and Abe, M. (2002). The Distinction between Burkitt Lymphoma and Diffuse Large B-Cell Lymphoma with c-myc Rearrangement. *Modern Pathology*, 15(7), pp.771–776.

**Relevant Code**

#3- Probes to Gene ID's and Names

BiocManager**:**install**(**"hgu133a.db"**)**

library**(**"hgu133a.db"**)**

library**(**"annotate"**)**

GENEIDs **<-** select**(**hgu133a.db,

c**(**result.table2.sorted**$**ID**)**,

c**(**"SYMBOL","ENTREZID", "GENENAME"**))**

GENEIDs

# 4- SAM

install.packages**(**"samr",

repos**=** "https://statweb.stanford.edu/~tibs/SAM", type **=** "source"**)**

install.packages**(**"writexl"**)**

library**(**writexl**)**

library**(**samr**)**

write\_xlsx**(**samrIDs, "filepath"**)**

dat **<-** as.data.frame**(**dat**)**

write\_xlsx**(**dat, "filepath”)

runSAM**()**#open the SAM interface

#1 - normalization

setwd**(**"filepath"**)**

batch **=** read.affybatch**(**dir**(**patt**=**"CEL"**))**

IRF4sdrf **=** read.csv**(**file **=** "filepath", header**=**T**)**

norm.batch **=** rma**(**batch**)**

dat **=** exprs**(**norm.batch**)**

#2 - T test and FDR code

t2 **=** vector**()**

pval.t2 **=** vector**()**

group **=** IRF4sdrf**$**Characteristics.molecular.diagnosis. **for(**j **in** 1**:**nrow**(**dat**)){**

temp **=** dat**[**j,**]** res**=**t.test**(**temp**[**group**==**"intermediate"**]**, temp**[**group**==**"non-mBL"**]**, var.equal**=**T**)**

t2**[**j**]** **=** res**$**stat

pval.t2**[**j**]=**res**$**p.val

**}**

adj.pval.t2 **<-** p.adjust**(**pval.t2, "BH"**)**

result.table2 **=** data.frame**(**ID**=**rownames**(**dat**)**, t.stat**=**t2, pvalue**=**pval.t2, fdr.pvalue**=**adj.pval.t2**)**

result.table2.sorted **=** result.table2**[**order**(**adj.pval.t2**)**,**]**

results **<-**result.table2.sorted**[**1**:**10,**]**

**Appendix**

|  |  |  |  |
| --- | --- | --- | --- |
| **Appendix Table 1: The top 10 differentially expressed genes from two-sample t-testing and FDR and their results in SAM.** | | | |
| Gene ID | *Gene Symbol* | Score |d| | Positively or negatively regulated (non-mBL compared to intermediate) |
| 203434\_s\_at | MME | 8.24 | Negative |
| 211862\_x\_at | CFLAR | 7.56 | Positive |
| 215731\_s\_at | MPHOSPH9 | 7.17 | Negative |
| 219515\_at | PRDM10 | 7.22 | Negative |
| 203680\_at | PRKAR2B | 7.77 | Negative |
| 202716\_at | PTPN1 | 7.54 | Positive |
| 203244\_at | PEX5 | 7.31 | Negative |
| 204490\_s\_at | CD44 | 7.49 | Positive |
| 213899\_at | METAP2 | 7.11 | Negative |
| 222270\_at | PPP4R3B | 5.45 | Negative |

|  |  |  |
| --- | --- | --- |
| **Appendix Table 2: Estimated miss rates when score (d) = 1.12** | | |
| *Quantiles* | *Score (lower bound, upper bound)* | *Miss Rate (%)* |
| 0.00 - 0.05 | -2.63, -1.99 | 82.92 |
| 0.05 - 0.10 | -1.99, -1.61 | 67.27 |
| 0.10 - 0.15 | -1.61, -1.37 | 61.58 |
| 0.15 - 0.20 | -1.37, -1.15 | 50.49 |
| 0.20 - 0.25 | -1.15, -0.98 | 45.68 |
| 0.25 - 0.75 | -0.98, -0.52 | 10.47 |
| 0.75 - 0.80 | 0.52, -0.73 | 0 |
| 0.80 - 0.85 | 0.73, -0.98 | 0 |
| 0.85 - 0.90 | 0.98, -1.29 | 3.67 |
| 0.90 - 0.95 | 1.29, -1.74 | 31.09 |
| 0.95 - 1.00 | 1.74, -2.49 | 72.38 |