UNIVERSITY OF LEEDS

FACULTY OF BIOLOGICAL SCIENCES

# ELECTRONIC Coursework Coversheet

For use with *individual* assessed work

**Student Identification Number:**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2 | 0 | 1 | 9 | 0 | 8 | 1 | 0 | 9 |

**Module Code: BIOL5327M**

(e.g. BMSC 1101)

**Module Title: Analytical Skills in Precision Med**

**Assessment title: Report**

(e.g. Essay / Lab report)

**Name of marker for this work:**

**Statement of Intent**

By submitting this work I am agreeing to the University Statement on Academic Integrity and confirm that this is all my own work.

**Word count / Page count** **1757 / 6**

Please note that failure to state word/page count, stating and incorrect word/page count or falsifying of word count will result in application of penalties as per the FBS Code of Practice on Assessment

# Investigating Gene Expression Differences in Young and Old Lymphoma Patients Using limma and T-Test Approaches

## Introduction

The prognosis of germinal-center derived B-cell (GCB) lymphomas, including diffuse large B-cell lymphoma (DLBCL), has been shown to vary with age. Children are shown to have a better prognosis than adults. It is still not known what is responsible for this difference. One proposed idea is that the tumour biology is different between children and adults. Salaverria et al (2011) conducted research to identify a subtype of GCB lymphoma that mostly affects children and young adults and has a favourable prognosis. As part of their research, they newly characterised 271 lymphomas and investigated translocations involving the IRF4 gene and looked at how these changes affect the outcome. To further investigate the biological mechanisms behind the differences, the RNA microarray data of the 271 lymphomas from the Salaverria et al (2011) study was used to conduct differential expression (DE) analysis looking at how gene expression differs between the young (<25 years old) and the old (>25 years old). Two different statistical methods have been considered when performing the DE analysis, the standard t-test and linear models for microarray data (limma).

The t-test works by comparing the mean expression of a gene in one group (young) with that of another group (old). The t-test calculates a t-statistic to quantify the difference in means relative to the variance of the data.

Limma works by fitting a linear model for each of the genes to estimate the effect of experimental effects, in this case the age of the individual. Moderated t-statistics are used to test for differential expression. The linear model is central to limma and is represented by:

$$Y = X\beta + \epsilon$$

Y - matrix of gene expression, X - the design matrix which encodes the experimental condition (young or old), β - matrix of coefficients, ϵ - error matrix.

The results of both analyses have been used together to show that there was the downregulation of genes in the adult group. Notably, the gene JAK2, which has a role in causing some child lymphomas, was found to be highly expressed in the young adult samples, this aligns with current knowledge of its role in paediatric lymphomas.

## Methods

The data used in the analysis was generated by Salaverria et al. (2011) and is publicly available GEO accession no. GSE22470. The data contains expression profiles of 271 diffuse large B-Cell lymphoma samples, hybridised to HGU133A Affymetrix GeneChips. The clinical variable age, stored in the column 'characteristics_ch1.1', was recoded so that a value of 0 indicates a sample from a person under 25 years old and a value of 1 indicates a sample from a person over 25. The microarray analysis was carried out in R.

T-test analysis

Raw and experimental data were read into R (R code 1a). Quantile normalisation was performed on the data (R code 2a). The normalised gene expression data was extracted and used to test for DE by using a t-test on each individual gene (R code 3a, 4a). A multiplicity adjustment was applied using Benjamini-Hochberg correction. The top ten most significant genes have been presented in a table (R code 5a).

limma analysis

Processed data was downloaded into R (R code 1b). Probe level normalisation had already been performed using the calibration and variance stabilisation method by Huber et al. (2002). Probe-set summarisation had already been performed using the median polish method on the normalised data (Irizarry et al., 2003) (R code 2b). Principal component analysis (PCA) was performed on the data (R code 2.5b). The gene expression data was extracted, and a design matrix was created to create the sample groups of young adult and Old (R code 3b). Filtered out lowly- expressed genes, calculated 'weights' to define the reliability of each sample, estimated the coefficients to average the replicate arrays for each sample level. Differential gene expression was assessed using the limma software in the context of a linear model (R code 4b). Results of the DE analysis were visualised by a Volcano Plot (R code 4.5b). The top ten most significant genes have been presented in a table (R code 5b).

## Results

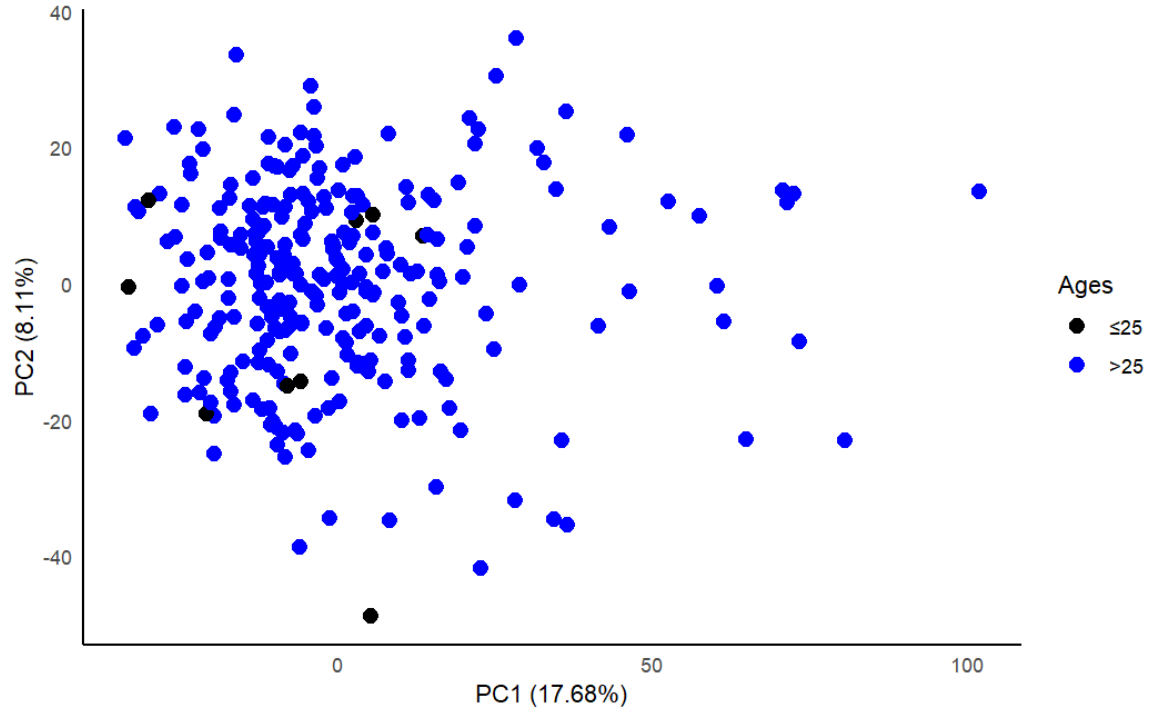


**Figure 1.** PCA plot showing the distribution of gene expression data. The samples are colour-coded based on age groups: black dots represent individuals aged ≤25 years, and blue dots represent individuals aged >25 years. Shows no clear separation of the two age groups along the first two principal components.

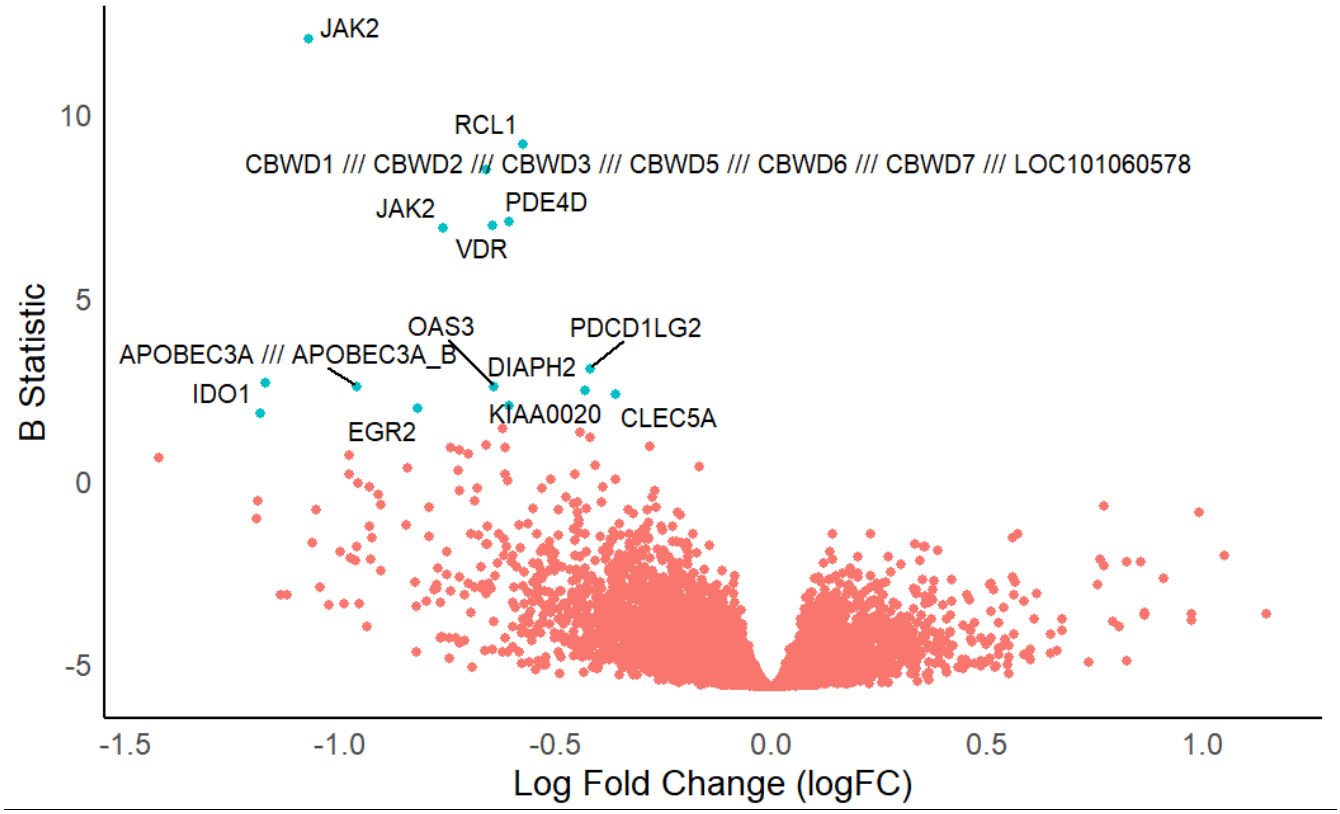


**Figure 2.** Volcano plot showing the results of limma differential gene expression analysis. The x-axis shows the log fold change (logFC) and the y-axis shows the B statistic - the log odds of a gene being expressed. Significant genes are shown in blue and non-significant are shown in red.

**Table 1.** Top 10 significant genes identified using limma analysis.

| ID | Gene Symbol | logFC | AveExpr | t | P.Value | adj.P.Val | B |
|---|---|---|---|---|---|---|---|
| 205842_s_at | JAK2 | -1.07602 | 8.11896 | -6.63698 | 1.73E-10 | 2.76E-06 | 12.062 |
| 218544_s_at | RCL1 | -0.57704 | 8.77007 | -6.02772 | 5.40E-09 | 4.31E-05 | 9.1748 |
| 220175_s_at | CBWD1 /// CB | -0.66192 | 7.84021 | -5.87891 | 1.21E-08 | 6.43E-05 | 8.5003 |
| 204491_at | PDE4D | -0.60818 | 7.96373 | -5.54948 | 6.80E-08 | 0.000221 | 7.0521 |
| 204255_s_at | VDR | -0.64831 | 7.12872 | -5.52649 | 7.65E-08 | 0.000221 | 6.9534 |
| 205841_at | JAK2 | -0.76261 | 7.04145 | -5.51056 | 8.30E-08 | 0.000221 | 6.8852 |
| 220049_s_at | PDCD1LG2 | -0.4223 | 7.38868 | -4.54581 | 8.25E-06 | 0.018819 | 3.0502 |
| 210029_at | IDO1 | -1.1745 | 9.05281 | -4.43712 | 1.33E-05 | 0.023585 | 2.6563 |
| 218400_at | OAS3 | -0.64494 | 7.94564 | -4.41251 | 1.47E-05 | 0.023585 | 2.5682 |
| 210873_x_at | APOBEC3A /// | -0.96165 | 6.63052 | -4.41225 | 1.48E-05 | 0.023585 | 2.5673 |

**Table 2.** Top 10 significant genes identified using t-test

| ID | SYMBOL | t.stat | pvalue | fdr.pvalue |
|---|---|---|---|---|
| 219701_at | TMOD2 | -5.769 | 2.19E-08 | 0.000489 |
| 213479_at | NPTX2 | -5.527 | 7.68E-08 | 0.000856 |
| 205842_s_at | JAK2 | -5.187 | 4.21E-07 | 0.003128 |
| 220175_s_at | ZNG1A | -5.003 | 1.02E-06 | 0.005696 |
| 220175_s_at | ZNG1B | -5.003 | 1.02E-06 | 0.005696 |
| 220175_s_at | ZNG1E | -5.003 | 1.02E-06 | 0.005696 |
| 220175_s_at | ZNG1C | -5.003 | 1.02E-06 | 0.005696 |
| 220175_s_at | ZNG1F | -5.003 | 1.02E-06 | 0.005696 |
| 205651_x_at | RAPGEF4 | -4.872 | 1.89E-06 | 0.008436 |
| 218793_s_at | SCML1 | -4.679 | 4.57E-06 | 0.016962 |

There were 15 significant genes identified from the limma analysis and 11 from the t-test. All the significant genes were downregulated in the old group. The genes, JAK2, OAS3 and ZNG1*/CBWD* were all shown to be significantly expressed in both differential expression analysis (Table 1,2).

## Discussion

The PCA plot (Figure 1) showed no separation between the young adult and old groups along the first two principal components. PC1 and PC2 account for 17.68% and 8.11% respectively, which suggests that the majority of the variability in the gene expression data was not due to age. The lack of separation is not entirely surprising, tumour biology and molecular subtypes both affect the gene expression in lymphomas so whilst the PCA did not show insight into the role age plays, it should not stop the analysis as it has been shown in previous research that age does have some effect of gene expression.

The differential expression analysis produced by limma and the t-test identified three genes that were significantly downregulated in both, JAK2, OAS3 and ZNG1*/CBWD*, the overlap strengthens the reliability of the findings and increases the confidence that they are involved in age-related differential expression (Table 1,2). Whilst there was a small amount of overlap, each method identified a different number of significantly DE genes, limma identified 15 and t-test identified 11 and the majority were unique to each statistical test. Limma identified RCL1 whilst t-test identified TMOD2, both genes had low adjusted p-values, so it is expected that they should have been identified in both statistical tests. Whilst JAK2 was identified in both, it was ranked differently. The difference in the results is explained by how limma and the t-test estimate variance and adjust for multiple comparisons.

When considering the statistical framework of both the methods it becomes apparent why there is a discrepancy. Limma uses a linear model and empirical Bayes methods, this uses information across all genes in order to stabilise variance estimates. This is important when samples sizes are small, or variances are heterogenous. Whilst the t-test looks at each gene independently and only relies on the sample variance of each gene. When considering my data set in which there was only 9 samples for the young adults, limma's empirical bayes approach provides greater statistical power and more reliable p-values with there being a large number of genes and only a few samples. The t-test is more sensitive to outliers and may lack power to detect genes with smaller but biologically impactful changes in expression. Both methods correct for multiple testing to control the false discovery rate (FDR). This is accounted for in limma as part of the empirical Bayes framework, while t-test results were adjusted using the Benjamini-Hochberg (BH) procedure. These considerations explain why there is a discrepancy between the power of the analysis with the t-test detecting four fewer genes than limma analysis.

Considering limma is more suitable for identifying differential expression patterns in my study, the role of Janus Kinase 2 (JAK2) in lymphomas was chosen for investigation as it was the most significantly expressed gene in the limma analysis and had a relatively high log fold change (Figure 2). JAK2 is a component of the JAK/STAT pathway that regulates cell proliferation, differentiation, and survival. Alterations in JAK2 has been shown to contribute to oncogenesis in acute lymphoblastic leukaemia. Hyperactivation of the JAK/STAT pathway by activating mutations, such as R683G, have been shown to promote leukemogenesis (Hassan et al., 2022).

This analysis provides support that age-related gene expression differences exist within lymphomas and identified specific genes that could be used for further investigation into tumour biology. Whilst pervious research agrees with the findings, this study would benefit from larger sample sizes and functional analyse to further explore the biological mechanisms behind these differences.

## References


Hassan, N.M., Abdellateif, M.S., Radwan, E.M., Hameed, S.A., El Desouky, E.D., Kamel, M.M. and Gameel, A.M., 2022. Prognostic significance of CRLF2 overexpression and JAK2 mutation in Egyptian pediatric patients with B-precursor acute lymphoblastic leukemia. Clinical Lymphoma, Myeloma and Leukemia, 22(6), pp.e376–e385.

Huber, W., von Heydebreck, A., Sueltmann, H., Poustka, A. and Vingron, M., 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics, 18(suppl_1), pp.S96–S104. PMID: 12169536. DOI: 10.1093/bioinformatics/18.suppl_1.S96.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B. and Speed, T.P., 2003. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Research, 31(4), p.e15. DOI: 10.1093/nar/31.4.e15.

Salaverria, I., Philipp, C., Oschlies, I., Kohler, C.W., Kreuz, M., Szczepanowski, M., 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. DOI: 10.1182/blood-2011-01-330795.


## R code

### T-test analysis:

**R code 1a:**

```r
batch = read.affybatch(dir(patt="CEL"))
clinical = read.csv("Experiment_info_2.csv", header=T)
```

**R code 2a:**

```r
norm.batch = rma(batch)
```

**R code 3a:**

```r
# Extract column names from dat
matrix_columns <- colnames(dat)

# Extract the Array.Data.File column from clinical
clinical_rows <- clinical$Array.Data.File

# Check they match
all(matrix_columns == clinical_rows)
```

**R code 4a:**

```r
t2 = vector() # vector of t-statistics
pval.t2 = vector() # vector of p-values
group = clinical$group # group labels

for(j in 1:nrow(dat)){
temp = dat[j,]
res=t.test(temp[group==1], temp[group==0], var.equal=T)
t2[j] = res$stat
pval.t2[j]=res$p.val
}
```

**R code 5a:**

```r
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),]
result.table2.sorted[1:10,] # listing the top 10 genes
```

### Limma analysis:

**R code 1b:**

```r
#import the data
library(GEOquery)
my_id <- "GSE22470"
gse <- getGEO(my_id)
```

**R code 2b:**

```r
pData(gse)$data_processing[1]
```

**R code 2.5b:**

```r
# Perform PCA
pca <- prcomp(t(exprs(gse)))

# Join the PCs to the sample information and create the PCA plot
cbind(sampleInfo, pca$x) %>%
  mutate(group = factor(group, levels = c(0, 1))) %>%
  ggplot(aes(x = PC1, y = PC2, col = group)) +
```

**R code 3b:**

```r
# Pull out age column to use for analysis
sampleInfo <- select(sampleInfo, characteristics_ch1.1)
```

```r
# 'group' is a factor with two levels: 0 (young) and 1 (old)
sampleInfo$group <- factor(sampleInfo$group, levels = c(0, 1))
```

```r
design <- model.matrix(~0 + sampleInfo$group)
design
```

**R code 4b:**

```r
## Making comparisons between samples
contrasts <- makeContrasts(old - young, levels = design)

fit2 <- contrasts.fit(fit, contrasts)
fit2 <- eBayes(fit2)
```

**R code 4.5b:**

```r
# Cutoffs
p_cutoff <- 0.05
fc_cutoff <- 1
topN <- 15

# Filter and plot
full_results1 %>%
  mutate(Significant = adj.P.Val < p_cutoff, abs(logFC) > fc_cutoff ) %>%
  mutate(Rank = 1:n(), Label = ifelse(Rank < topN, Gene,"")) %>%
  ggplot(aes(x = logFC, y = B, col = Significant, label = Label)) +
```

**R code 5b:**

```r
anno <- fData(gse)
head(anno)
```

```r
anno <- select(anno,Gene = `Gene Symbol`,ID,GB_ACC)
fit2$genes <- anno

topTable(fit2)
```

### Code Availability

The R code used for the analysis in this report is available on GitHub and can be accessed at: https://github.com/ChristopherLetton/Assessment_2.