In silico characterisation and classification of B-cell maturation to aid precision medicine.

Oliver Dickson

Student ID: 201125222

Supervisor: Professor Westhead

Submitted in part fulfilment of the requirements of the degree of

Master of Science in

Precision Medicine: Genomics & Analytics

Faculty of Biological Sciences,

University of Leeds,

Leeds,

LS2 9JT.

Acknowledgements:

I would like to thank Professor Westhead for the supervision of this research project and Amber Emmett for providing me with the relevant information and data to get me started with the analysis. I would also like to thank John Davies for providing me with the REMoDL-B trial data.

Abstract:

Aggressive lymphomas are the sixth most common cancer type in the UK and form a poor prognosis group, with 10-year survival rates lower than 30%. This, coupled with the fact that only around 4% of cases are considered preventable, makes the need for better more targeted treatments unequivocal. The B-cell lymphoma umbrella covers a wide range of malignancies with huge variations attributable to the stage of B-cell maturation they originate from. These malignancies can often be further subclassified based on the genetic abnormalities present in each tumour. This information has been leveraged to create and provide targeted treatments, enabling significant increases to progression free survival rates. The future of treatment now lies in the development of novel therapeutics using accurate cell line models of each cancer type. This analysis provides a shortcut, circumventing the time and money needed to develop new cell lines, through generation of a classifier which can assign B-cell maturation stages to existing cell lines.

The top 25 B-cell maturation genes were identified through time series differential gene expression, covering 312 hours of lymph node B-cell maturation. Clustering analysis of 65 cell lines from eight types of B-cell lymphomas, using the top 25 genes, identified three major time points of B-cell maturation. These time points and the top 25 genes were then used to train an ensemble classifier with a reported accuracy of 0.989. 3,049 B-cell lymphoma patient samples were then classified, and correlation analysis was used to suggest the most representative cell line for each sample.

Abbreviations:

Abbreviation	*Expansion*
ABC	Activated B-cell like
BL	Burkitt lymphoma
B-NHL.UNC	Aggressive B Non-Hodgkin's lymphoma, unclassified
DAC	DLBCL automatic classifier
DLBCL	Diffuse Lage B-cell Lymphoma
FDR	False discovery rate
GCB	Germinal Centre B-cell like
MAD	Median absolute deviation
MHG	Molecular high grade
MLP	Multi-layer Perceptron
PCA	Principal component analysis
PMBL DLBCL	Primary mediastinal large B-cell lymphoma
RB-CHOP	Rituximab, Bortezomib, Cyclophosphamide, Hydroxydaunomycin, Oncovin and Prednisolone
R-CHOP	Rituximab, Cyclophosphamide, Hydroxydaunomycin, Oncovin and Prednisolone
REMoDL-B	Randomised Evaluation of Molecular guided therapy for Diffuse Large B-cell Lymphoma with Bortezomib
ROC Area	Area under the Receiver Operating Characteristic curve
UNC	Unclassifiable
VDJ recombination	Variable, Diversity and Joining recombination

Table of Contents:

List of Figures:

Introduction:

Results:

List of Tables:

Methods

Results

Introduction:

B-cell Maturation:

B-cells are one of the two classes of the lymphocyte white blood cell subtype, along with T-cells, forming the adaptive immune response. The main differentiating factor between the two being the expression of the antigen binding B-cell receptors expressed on the cell surface membrane of B-cells (Alberts et al., 2002). The B-cell receptors are comprised of immunoglobulin (antibody) molecules which upon contact with antigens, usually on the surface of a pathogen, activate the B-cell, either as a naïve or memory B-cell (Treanor, 2012). Both cell types, when activated, undergo clonal expansion, and differentiate into antibody secreting plasmablasts or plasma cells to combat the infection (Janeway et al., 2001).

The initial stages of B-cell development occur in the bone marrow, beginning with hematopoietic stem cells which differentiate into pro-B-cells (Kondo, 2010). Early pro-B-cells undergo a process of somatic chromosome recombination known as Variable, Diversity and Joining (VDJ) recombination, creating an immunoglobin heavy chain unique to each individual B-cell (Tonegawa, 1983). During the chromosomal recombination, the DNA must be broken and rearranged several times to generate the diversity and specificity seen in the B-cell receptors and that makes them so effective at recognising an unlimited array of different antigens (Roth, 2014). This process is predominantly controlled by the two *RAG* genes causing DNA breaks which are repaired by nonhomologous end-joining (Jung et al., 2006). Aberrant recombination events, however rare, can often lead to some of the translocations seen in B-cell cancers like t(14;18) present in pre-B-cell leukaemia and follicular lymphoma, and can act as a biomarker for the latter (Tsujimoto et al., 1985), (Roulland et al., 2014). Upon successful heavy chain VDJ recombination, the now late pro-B-cell undergoes proliferation and further VDJ recombination to create a unique immunoglobin light chain and progresses into a transitional B-cell, migrating to secondary lymphoid organs, like the lymph nodes. It then matures into a naïve mature B-cell with the unique immunoglobulins forming B-cell receptors (Chung et al., 2003). Once present at the secondary lymphoid organs, B-cell activation and maturation can be initiated in either a T-cell dependent or independent manner (Ng and Chiorazzi, 2021).

T-cell dependent activation begins with stimulation of mature B-cells by a T-cell dependent antigen, occurring in two predominant phases, early and late (Baumgarth, 2000), (Sagaert et al., 2007), (Lenz and Staudt, 2010). The early phase occurs extrafollicularly, in a rapid response, forming antigen experienced B-cells which proliferate into plasmablasts then plasma cells (García De Vinuesa et al., 1999). The late phase involves the creation of a germinal centre and takes around a week to complete (Baumgarth, 2000). After activation

these B-cells migrate to germinal centres becoming centroblasts and localise to the dark zone of the germinal centre (McHeyzer-Williams et al., 2001). Here, they undergo two more mutative stages, somatic hypermutation and class switch recombination, further boosting the affinity of the antibodies. They also undergo intense proliferation and as well as high levels of cell death via apoptosis (Vinuesa et al., 2005). The centroblasts then mature into centrocytes, the follicular helper T-cells then select the centroblasts with antibodies of high affinity, allowing them to differentiate into long-lived antibody secreting plasma and memory cells or resubmit them into the centroblast cycle of mutation, proliferation, and cell death (Vinuesa et al., 2005). These two T-cell dependent pathways can be seen in **Figure 1**. Somatic hypermutation involves the induction of random, predominantly single point mutations within the antigen-binding site of the antibody and class switch recombination involves chromosomal recombination of the switch regions of the antibody, the resultant antibody changing its function. During this process, oncogenes common in lymphomas like the *BCL6* gene can be translocated in place of the desired switch region placing them under the control of strong promotors or enhancers, leading to tumorigenesis (Wilson et al., 1998), (Chaudhuri and Alt, 2004), (Pasqualucci et al., 2001).

T-cell independent activation begins with the naïve mature B-cells encountering pathogenic antigens (Nutt et al., 2015). Upon activation they proliferate extrafollicularly, without the formation of germinal centres, undergoing some class switch recombination and somatic hypermutation (Weller et al., 2008), (Scheeren et al., 2008). These B-cells then mature into short-lived plasma cells, with a small number becoming memory B-cells (Obukhanych and Nussenzweigh 2006). The short life span of these plasma cells thus means the T-cell independent response can be brief compared to that of the T-cell dependent response. As somatic hypermutation and class switch recombination predominantly do not occur to the degree they do in germinal centres, the resultant antibodies of non-germinal centre pathways have much lower affinities when compared to that of the germinal centre antibodies (Tierens et al., 1999).

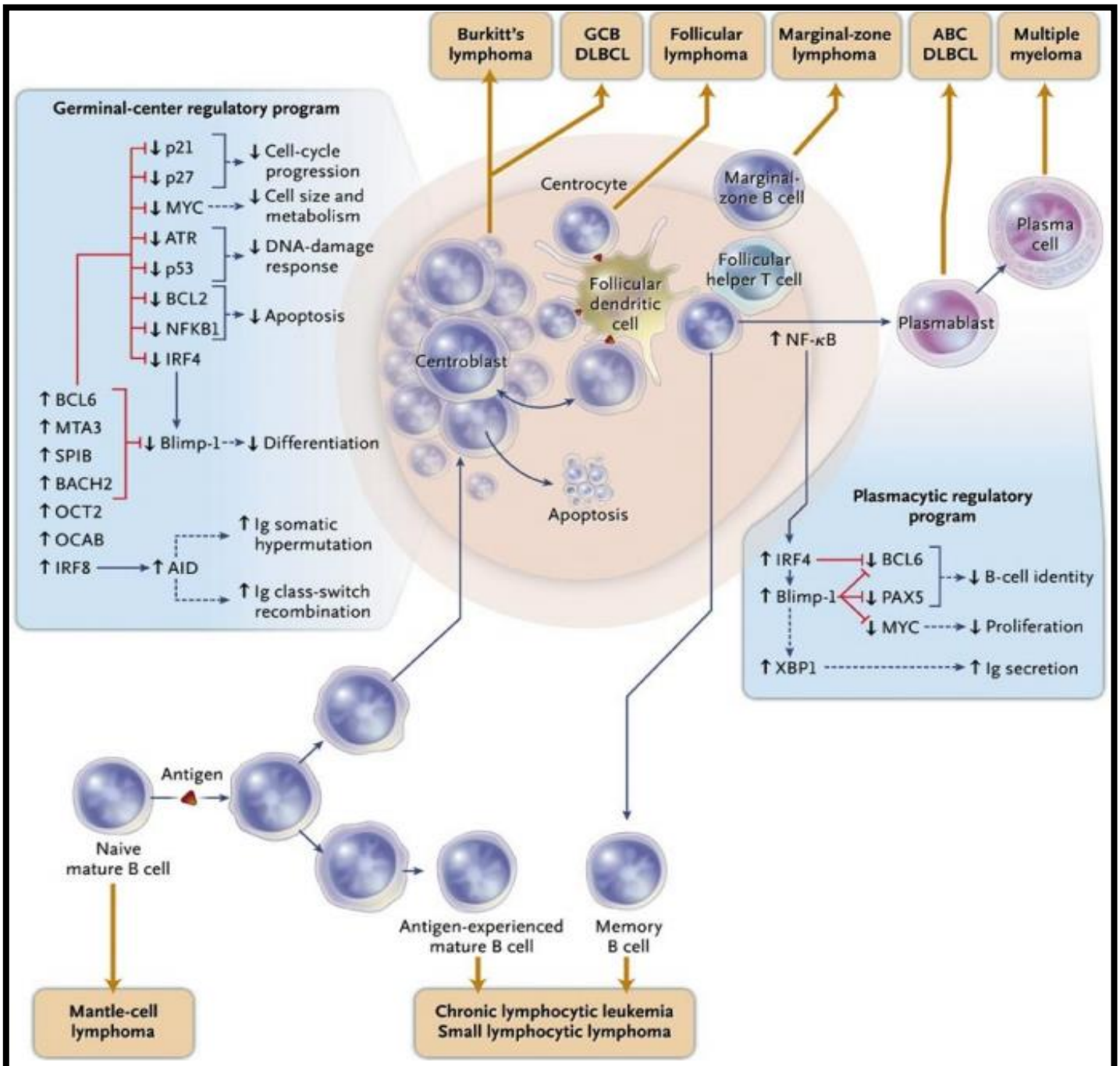


Figure 1: **B-cell Maturation in the Secondary Lymphoid Organs and Tumorigenesis**, taken from Lenz and Staudt, (2010). Shown is the B-cell maturation in the secondary lymphoid organs in response to antigens in both a T-cell dependant and independent manner. Also shown are some of the genes involved in the germinal centre and plasmacytic regulatory programs which aid in controlling the maturation stages and at which point, from which cells, particular subtypes of B-cell lymphomas can arise from.

Aggressive Lymphomas:

As discussed above, immature B-cells during normal maturation undergo three stages of deliberate genetic mutation, VDJ recombination, causing the development of a unique B-cell receptor in the bone marrow, somatic hypermutation and class switch recombination, the latter two taking place in the germinal centre of the lymph node before the B-cells finally differentiate into plasma or memory cells (Harwood and Batista, 2010), (Rajewsky, 1996), (Seifert et al., 2013). Although aberrant VDJ recombination is mainly responsible for leukaemias and falls out of the scope of this analysis. It is these three processes that, when dysfunctional, can act as the main drivers for the B-cell's journey to becoming malignant, exemplifying one of Hanahan and Weinberg's enabling characteristics, genome instability and mutation (Hanahan and Weinberg, 2011). It is also important to remember that lymphomas can arise from the infection of viruses like the Epstein–Barr virus, a common causal factor in Burkitt lymphoma (Seifert et al., 2013). The points in B-cell maturation that some B-cell lymphomas arise from is visualised in **Figure 1**.

In the UK and USA, aggressive lymphomas are responsible for 4% of all new cancer cases each year, making them one of the most common group of cancers (Cancer Research UK, 2020a), (Cancer.net, 2021). Lymphomas can be separated into two main groups, non-Hodgkin's lymphoma, the sixth most common cancer in the UK, which comprises of all lymphomas except Hodgkin's lymphoma (Cancer Research UK, 2020a), (Weniger and Küppers, 2021). Lymphoma types can be continually subdivided down to the stage of B-cell type and the particular cell of origin of the tumour, exemplified by **Figure 1**. Unfortunately, the overall lymphoma 10-year survival rate remains at 55% with many treatments being standardised across many lymphoma types, despite the molecular variety of the disease, providing a poor prognosis for patients. Additionally, only 3% of cases being preventable the only present solution is the development of better therapeutics (Cancer Research UK, 2020b).

A primary example of precision medicine, and the specific targeting of lymphomas, can be seen in Diffuse Large B-Cell Lymphoma (DLBCL). In 2000, DLBCL was successfully broken down into two cells of origin, Germinal Centre B-cell like (GCB) and Activated B-Cell like (ABC) with vastly different survival rates when using the standard Rituximab, Cyclophosphamide, Hydroxydaunomycin, Oncovin and Prednisolone (R-CHOP) treatment, illustrating the need for tailored treatments (**Figure 2A-C**) (Alizadeh et al., 2000). A key regulatory difference between these two cells of origin is the plasmacytic regulatory program, which leads to the upregulation of the NF-κB pathway, present only in the ABC cell of origin (Davis et al., 2001). A new treatment, R-CHOP combined with Bortezomib (RB-CHOP), was created to specifically target the ABC DLBCL subgroup, and was trialed in the 'Randomised Evaluation of Molecular guided

therapy for Diffuse Large B-cell Lymphoma with Bortezomib' (REMoDL-B) trial (Davies et al., 2019). The theory behind the addition of Bortezomib to R-CHOP was that it would prevent the breakdown of a NF-κB inhibitor, through inhibition of the 26S proteasome, therefore increasing inhibition of the NF-κB pathway (Bonvini et al., 2007). Unfortunately, the original trial reported no improvement to survival rates. A retrospective analysis of the same trial by Sha et al. (2019) identified a third DLBCL subgroup, Molecular High Grade (MHG), with drastically worse prognosis. This group was mainly comprised of GCB with double or triple hit phenotypes, namely, *MYC* translocations combined with either a *BCL-2* or *BCL-6* translocation (double-hit) or all three translations (triple-hit) (Pemmaraju et al., 2014). Targeting of the MHG group using RB-CHOP resulted in an increase of around 0.2 in event free proportion (**figure 2D**). As the MHG subgroup mostly encompasses the GCB cell of origin, it is doubtful Bortezomib's mechanism of action is via the NF-κB pathway (Davis et al., 2001).

MHG, in terms of gene expression, is very similar to Burkitt lymphoma, creating a spectrum between DLBCL and Burkitt lymphoma (Momose et al., 2015). Both Burkitt lymphoma and GCB DLBCL originate from the germinal centre, specifically the centroblast cell type (Lenz and Staudt, 2010). All Burkitt lymphoma cases have *MYC* translocations, and the vast majority (80%) of cases also have the t(8;14)(q24;q32) translocation (Molyneux et al., 2012). As mentioned above, Epstein–Barr virus infection is a characteristic of most cases, endemic to Africa (Pannone et al., 2014). Despite similarities to the poor prognosis MHG DLBCL subgroup, progression free survival rate is significantly higher at 90%, this figure is slightly deceptive, however, with the majority of cases occurring in children younger than ten with drastically higher survival rates than adult cases (Molyneux et al., 2012), (Paul and Jonathan, 2019).

Multiple myeloma, which arises the plasma cell, currently has a 10-year survival rate of 29% and only 14% of cases are considered preventable (Lenz and Staudt, 2010), (Cancer Research UK, 2020c). Several features of poor prognosis multiple myeloma include translocations like t(4;14) and non-hyperdiploidy which benefit from targeted therapies (Sonneveld et al., 2016). These therapies, for patients under 65, involve a combination of chemotherapy and a stem cell transplant (Printz, 2016). Provided the stem cells come from a healthy non-cancerous donor, so, not the patient before chemotherapy, it can provide a complete cure. Though highly effective, this treatment also comes with an increased treatment-associated mortality rate (Kyle and Rajkumar, 2004).

Since the early 2000s it has been clear that lymphomas, particularly DLBCL, can be subdivided into separate cells of origin with drastically different molecular characteristics (Alizadeh et al., 2000). The following advent of precision medicine has made large improvements to pre-existing treatments, enabling them to be better targeted and resulting

increased survivability, this is exemplified in the REMoDL-B trial with the further retrospective identification of the MHG subgroup (Sha et al., 2019). Thus, the future of lymphoma therapy relies heavily on the development of novel treatments, and the repurposing of old, to target the molecular specifics of further lymphoma subtypes.

As already discussed, which stage of B-cell maturation the cancer originates from has huge impacts on the molecular phenotype of the cancer. Particularly stages within the germinal centre which plethora lymphomas originate from (**Figure 1**), (Allen et al., 2007). This causes huge ranges in aggressiveness and responsiveness to treatment (**Figure 2**). Consequently, the cell line being used to study the cancer and refine treatments needs to have a high level of specificity to the B-cell maturation stage and molecularity of the cancer being studied. Cell lines have provided huge advances in anticancer drug discovery and act as an ideal model with many advantages including easy availability, enabling continuity between different laboratories, indefinite growth and thus also the provision of limitless cells for testing (Mirabelli et al., 2019), (Gonzalez-Nicolini and Fussenegger, 2005). The curation of such cell lines requires an extensive investment of time and money, taking upwards of two years with considerably low success rates, as demonstrated by the development of the LL-100 cell lines panel by Drexler and Quentmeier (2020), (Drexler et al., 2000).

In recent years it has been common practice to classify B-cell lymphoma tumour samples based on gene expression using machine learning classifiers (Care et al., 2013), (Holmes et al., 2020). These classifiers have their routes in the genes identified through differential gene expression analysis conducted on B-cell maturation enabling characterisation of these processes (Lee et al., 2020), (Stewart et al., 2021). This analysis offers a shortcut to the investment in new cell line models. By studying B-cell maturation, a machine learning classifier can be developed to assign B-cell maturation stages to the B-cell lymphoma cell lines present in the Broad-Novartis Cancer Cell Line Encyclopaedia, as well as acting as a proof of concept for assigning the most similar cell lines for patient tumour samples. This analysis hopes to facilitate the progression of precision lymphoma treatment by removing the need to generate new cell lines.

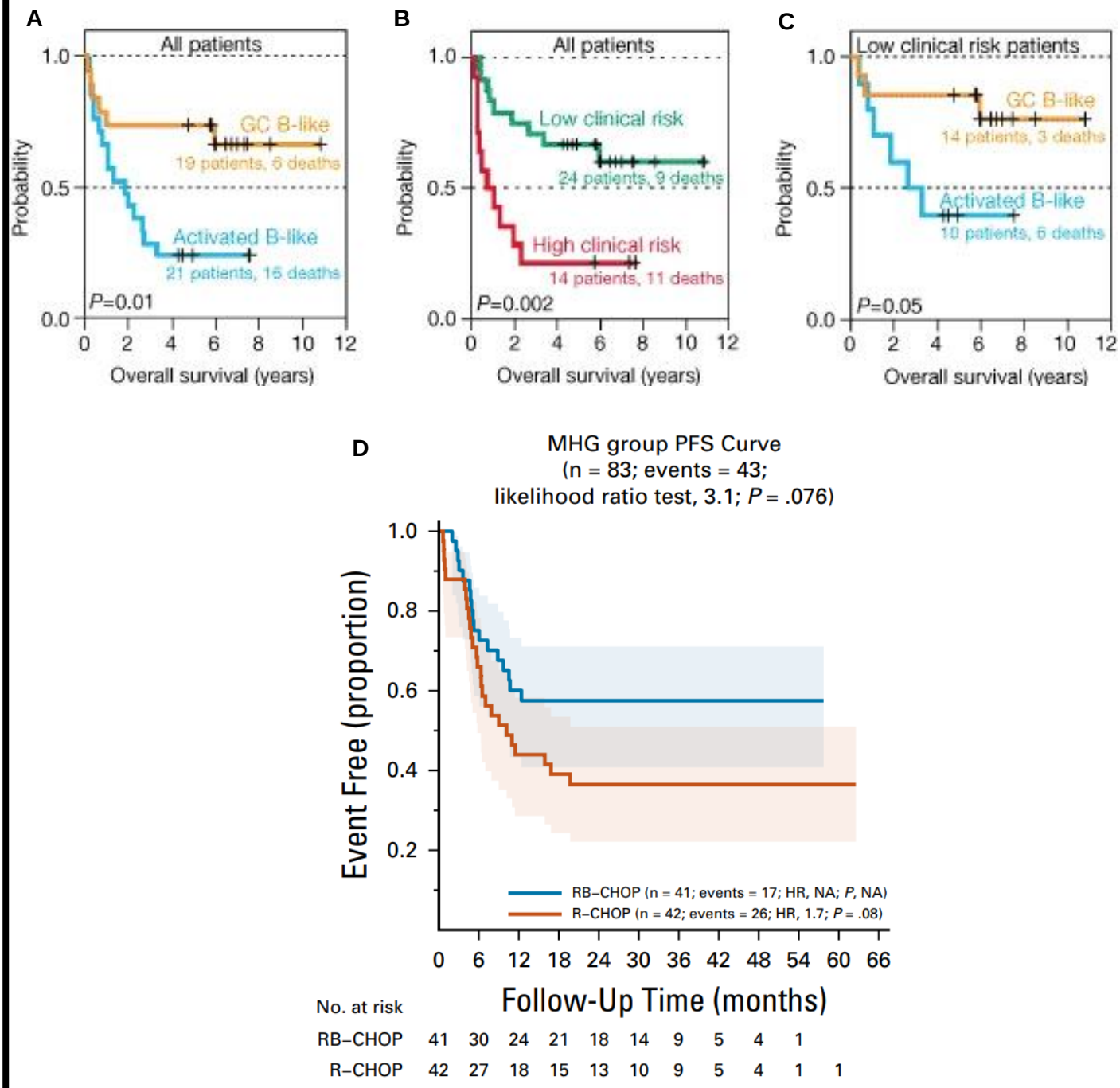


Figure 2: Survival Rates of Lymphoma Patients by Subtype. **A-C,** Survival rates of DLBCL patients when separated into the GCB and ABC DLBCL subtypes. Taken from Alizadeh et al. (2000). **D,** Event free rate of patients in the MHG DLBCL subgroup separated into RB-CHOP and R-CHOP treatment regimens from the REMoDL-B trail. Taken from Sha et al. (2019).

Aims:

The primary aim of this study will be to analyse the B-cell maturation gene expression dataset generated by Coco et al. (2012), with the objective of identifying the genes most associated with B-cell maturation. The dataset covers the maturation of B-cells in the lymph node through samples taken at 10 different time points from 0 to 312 hours, reportedly covering the cell types from mature B-cell through to plasma cell. Time series differential gene expression using the R package DESeq2 will be carried out to identify the top differentially expressed genes across the 312 hours with gene ontology analysis, using the R package TopGO and the String database, also being performed to identify any enriched pathways and networks associated with B-cell maturation (Love et al., 2014), (Alexa and Rahnenfuhrer, 2021), (Szklarczyk et al., 2021).

The B-cell related cancer cell lines gene expression profiles from Barretina et al.'s (2012) 934 human cancer cell line sequencing dataset will then be correlated against the samples from the B-cell maturation dataset, using Pearson's correlation coefficient. This will identify what cell lines and B-cell cancer types are most similar to each time point. Using the hypothesis that a correlation coefficient above 0.9 likely means the two samples belong to the same cell type, the B-cell type for each timepoint can be identified (Koch et al., 2018).

Clustering of the samples, using Principal Component Analysis (PCA), will then identify how many and which maturation stages the B-cell maturation dataset covers, this then will inform the creation of a machine learning classifier. This will be then used to classify cell lines or patient tumour samples into their most likely normal B-cell maturation stage. The classifier will make use of the SciKit-Learn Python 3 library and will be an ensemble stacking classifier containing six base predictors: Logistic Regression, K Neighbours Classifier, Random Forest classifier, Support vector classifier, Naïve Bayes classifier and, Multi-layer Perceptron, the results of each will be combined by a logistic regression metaclassifier (Buitinck et al., 2013). A stacking ensemble classifier will be used as it produces more accurate results than the individual models used on their own (Mohammed et al., 2018).

The final stage of the analysis will be to use the classifier to predict the maturation stages of patient tumour samples obtained from publicly available databases like ArrayExpress and then, through correlation analysis between the tumour samples and the B-cell cancer related cell lines, identify the most similar cell line for each sample analysed, to enable further study of that tumour sample's maturation stage in the laboratory (Parkinson et al., 2007).

Methods:

Primary Datasets:

B-cell Maturation Dataset:

The B-cell maturation dataset consists of RNA-seq gene expression profiles of B-cells taken from three donors. These samples were obtained from 'In Vitro Generation of Long-lived Human Plasma Cells', the raw data was in the form of .fastq files (Coco et al., 2012). These samples ranged from 0 to 312 hours in age, totalling 22 individual samples (**Table 1**). Each sample also had a B-cell type attributed to it, although not confirmed experimentally.

		Donor				
Day	Timepoint (hours)	A	B	C	Indicated cell type	Total Samples
0	0	/	/	/	B-cell	3
0	0.5	/	/		B-cell	2
0	2.5	/	/		B-cell	2
0	6	/	/		B-cell	2
0	12	/	/		B-cell	2
1	24	/			B-cell	1
3	72	/	/	/	Activated B-cell	3
6	144	/	/	/	Plasmablast	3
10	240	/		/	Plasma cell	2
13	312	/		/	Plasma cell	2

Table 1: B-cell Maturation Dataset Summary. A table summary of the B-cell maturation dataset including which donors submitted what samples at each time point and the indicated B-cell type associated with each timepoint.

Cancer Cell Line Dataset:

The original dataset comprised of 934 human cancer cell line gene expression profiles from the Broad-Novartis Cancer Cell Line Encyclopaedia, initially sequenced by Barretina et al. in 2012 and last updated in 2017. This dataset was downloaded from ArrayExpress under the code E-MTAB-2770, the whole dataset comprising of all 934 cell lines will be referred to as the E-MTAB-2770 dataset from now on (Parkinson et al., 2007). Of these cell lines, there were 14 B-cell related cancer types, both lymphomas and leukaemias, which totalled 86 individual cell lines. These were used in correlation analysis, and enabled clarification that any correlation observed between lymphomas and the time series was specific to each timepoint, and maturation stage, and not just specific to overall B-cell biology (**Table 2**). This 86-cell line

subset of the Cancer Cell Line dataset is henceforth referred to as the B-Cell Cancer Cell Line Dataset. As the maturation timeline primarily covered the germinal centre and stages

immediately before and after it (lymph node stages), a subset of these 14 B-cell related cancers was created to enable the differential gene expression and classifier to be more specific to genes differentially expressed in the lymph node stages. This subset, henceforth referred to as the Lymphoma Cell Line dataset, totalled eight cancer types and included 65 cell lines. The DLBCL cell lines were classified into cell of origin by the DLBCL automatic classifier (DAC) (Care et al., 2013).

Cancer type	*Number of Cell lines*	*Part of the Lymphoma Cell Line subset?*
Adult B Acute Lymphoblastic Leukaemia	2	No
B-Cell Acute Lymphoblastic Leukaemia	6	No
B-Cell Non-Hodgkin's Lymphoma	2	No
B-Cell Prolymphocytic Leukaemia	1	No
Burkitt Lymphoma	7	Yes
Childhood B Acute Lymphoblastic Leukaemia	7	No
Chronic Lymphocytic Leukaemia	2	No
Diffuse Large B-Cell Lymphoma (ABC)	6	Yes
Diffuse Large B-Cell Lymphoma (GCB)	12	Yes
Epstein-Barr Virus-Related Burkitt Lymphoma	3	Yes
Hodgkin's Lymphoma	8	Yes
Mantle Cell Lymphoma	5	Yes
Multiple Myeloma	24	Yes
Primary Effusion Lymphoma	1	Yes
Total number of cell lines:	**86**	**65**

Table 2: Summary of the Cancer Cell Line Dataset Subsets. A table summary of the cancer types and how many cell lines used in both the B-Cell Cancer Cell Line Dataset and the Lymphoma Cell Line Dataset which were derived from the Cancer Cell Line dataset.

Secondary Datasets:

The secondary datasets were comprised of patient tumour samples and were analysed using the correlation and classification procedures in later outlined stages (**Table 3**). All the datasets, except the REMoDL-B dataset, were downloaded from ArrayExpress under their respective entries (Parkinson et al., 2007). The REMoDL-B dataset was provided by Davies et al. (2017).

Dataset name	*Tumour sample type(s)*	*Number of samples*
REMoDL-B	DLBCL	928
E-GEOD-2658	Multiple Myeloma	559
E-GEOD-4475	DLBCL and Burkitt Lymphoma	221
E-GEOD-4732	DLBCL and Burkitt Lymphoma	303
E-GEOD-31161	Multiple Myeloma	1,038
	Total number of samples:	**3,049**

Table 3: Summary of the Secondary Datasets. A table summary of the secondary datasets analysed by correlation and classification. Including what tumour types the samples originate from and how many samples present.

Dataset Cleaning and Processing:

Before analysis took place the datasets needed to be cleaned and processed to ensure meaningful analysis could be carried out. Raw expression data can vary hugely depending on the experimental sequencing procedure used and the samples assayed and suffers from huge ranges in gene abundances that can be produced in sequencing files (Lovén et al., 2012). This was especially needed as the analysed data came from both sequencing and microarray platforms; thus, all datasets were processed to ensure statistical comparability between all samples.

B-cell Maturation Dataset Differential Gene Expression:

As the B-cell maturation dataset was in the form of raw .fastq sequencing files, which were computationally intensive to process, processing was carried out on the University of Leeds Advanced Research Computing Node 3. It involved a three-stage pipeline of FASTQC for quality checking of the data, followed by TrimGalore to trim the sequencing adaptor sequences off the reads and finally Salmon which mapped and quantified the transcripts producing quant.sf files ready for analysis in the statistical software R (Andrews, 2010), (Krueger, 2021), (Patro et al., 2017), (R Core Team, 2019).

The quant.sf files were then imported into R for differential gene expression using TXimport, this also provided the time series gene expression abundance file, henceforth referred to as the Timepoint dataset, used in the correlation and classification stages (Soneson et al., 2015). Time series differential gene expression on the B-cell maturation dataset was then carried out using DESeq2 in Jupyter Notebook to identify genes that are differentially expressed over the maturation timeline accounting for any variation in donor expression (Love et al., 2014), (Project Jupyter, 2021).

Cancer Cell Line Dataset Differential Gene Expression:

The Cancer Cell Line dataset was only available in a processed FPKM datafile this limited what gene expression can be done as FPKM contains significantly less data than the normalised counts typically used for differential gene expression, due to it being the output of a normalisation procedure (Illumina, 2021) (Zhao et al., 2021). A solution presented was to transform the dataset using formula 1, this avoids any zeros expression values through the +0.1 thus also preventing the log of zero (Nazarov et al., 2017). The differential gene expression was then calculated between the Lymphoma Cell Line Dataset cell lines and all other cell lines in the Cancer Cell Line Dataset to obtain B-cell lymphoma specific genes. This used the eBayes function from the R package Limma (Ritchie et al., 2015).

Formula 1: $$y = log_2(x + 0.1)$$

Correlation and Classification Dataset Pre-Processing:

All datasets, those listed in table 3, the Cancer Cell Line dataset and the Timepoint dataset were transformed using formula 1 and then were Z-score normalised using formula 2. This process standardises all values for each dataset into the same range and magnitude of values, while also keeping the gene expression values in the same context, I.E., higher positive values equate to more expressed genes and larger negative values equate to less expressed genes. As the same process was carried out on each dataset and to produce statistically similar values in each dataset, it enabled the transformed datasets to be comparable to each other (Cheadle et al., 2003). All future mentions of the datasets refer to the datasets that have gone through formula 1 and 2 processing. At this point the two datasets, B-Cell Cancer Cell Line Dataset and Lymphoma Cell Line Dataset, were derived from the Cancer cell line dataset. Probe IDs were converted to gene names using the R package BioMart (Durinck et al., 2009).

Formula 2: $$Z = \frac{x-\mu}{\sigma}$$

Identification of the Top 25 Genes:

Time series differential gene expression using DESeq2 produced 13,206 differentially expressed genes. This was then filtered down to a list of 1,512 genes using the additional filters of log_2(fold-change) standard error < 25% quantile (0.26), and adjusted p-value < 25% quantile (1.72×10^{-17}). This selected only the most significant genes with the least amount of variation in expression and drastically reduced the number of genes to perform analysis on the theory the most constantly and significantly expressed genes will allow for the best characterisation of B-cell maturation.

Limma differential gene expression of the Cancer Cell Line Dataset produced a list of 5,337 genes that were significantly differentially expressed with an absolute log fold change greater than one. This was further filtered down by filtering for an adjusted p-value < 25% quantile ($5.6x10^{-31}$) and an absolute $\log_2$(fold-change) > 25% quantile (1.23). Not only did these filters select the most significantly expressed genes with the largest log fold changes, as these genes are likely the most representative genes of B-cell lymphoma, they also remove general differential expression noise, particularly that caused by low-expression genes (Sha et al., 2015). The genes that appeared in both the Limma and time series differential gene expression results were then ordered by adjusted p-value, and the top 25 were selected.

Gene ontology analysis was performed primarily using the R package TopGO which ranked the gene ontology terms by Kolmogorov–Smirnov test p-values which had been transformed using TopGO's weight01 algorithm (Alexa and Rahnenfuhrer, 2021). Other gene ontology and network analysis was carried out using the String database (Szklarczyk et al., 2021).

Correlation Analysis:

Correlation analysis was performed using Pearson's correlation coefficient in R using the base function cor.test(). The correlation coefficient was calculated between the Timepoint Dataset samples and the B-Cell Cancer Cell Line Dataset. Correlation analysis results were then displayed graphically in heatmaps using the R package Heatmaply (Galili et al., 2018). The correlation p-values were then adjusted for using the Benjamini-Hochberg procedure creating a false discovery rate (FDR) (Benjamini and Hochberg, 1995).

Pearson correlation coefficient was then calculated between each cell line and each B-cell maturation sample using the normalised expression values of the top 25 genes previously identified. The average coefficient for each timepoint was then calculated. Initial Correlation Analysis (1) was performed with all the genes that were present in both datasets. Initial Correlation Analysis (2) was performed with the top 1,508 time series differentially expressed genes. Final Correlation Analysis was performed with the top 25 genes identified from the overlap between the Cancer cell line dataset differentially expressed genes and the time series differentially expressed genes.

Classifier:

Principle component analysis was used to identify the clusters used in the training of the classifier. This was done in R using the base function prcomp() and the graphs were created using the R package Ggfortify (Tang et al., 2016). The Lymphoma cell line dataset was then sorted into the three clusters using K-means clustering as was the Timepoint dataset. These two datasets were then combined to make the complete training dataset.

Multiple classifiers were tested in a 1,000-round cross-validation procedure. For each round of the cross-validation the training dataset was split into testing and training sets with a testing:training ratio of 25:75, each model was trained and tested on this split and the cross-validation metrics true positive rate, false positive rate, precision, recall, F-measure, accuracy, and area under the Receiver Operating Characteristic curve (ROC Area). Finally, each model's average rank for each metric was calculated. This was done in Python 3 using the default settings of the classifiers from the Python 3 module SciKit-Learn (Van Rossum and Drake, 2009), (Buitinck et al., 2013). The classifiers tested were Logistic Regression, K Nearest Neighbours Classifier, Decision Tree Classifier, Random Forest Classifier, Support Vector Classifier, Naïve Bayes Classifier, Multi-layer Perceptron (MLP) and an Ensemble Stacking Classifier combining the results of all the aforementioned classifiers except the decision tree classifier due to its redundancy when using a random forest classifier.

Secondary Datasets Analysis:

First each datasets number of samples and cancer types were visualised by pie charts in R using the module Plotrix (Lemon, 2006). Next correlation analysis was performed with the Timepoint dataset, and the samples were then classified using the ensemble stacking classifier and the results were visualised in pie charts. In order to see the relation between the correlation and classification analysis, the correlation analysis was averaged by time point classification. Finally, for some datasets the most representative cell line for each major group of samples within the dataset was ascertained by correlation analysis between the secondary dataset and the Lymphoma cell lines dataset. The most representative cell lines were deemed to be those with strong correlation levels with the samples.

For the datasets with less than the 25 genes present, the classifier was retrained using just the present genes and the correlation analysis also used just the genes present. Some datasets also had multiple probes for one gene. In these cases, the Median Absolute Deviation (MAD) was used to combine all expression values into one expression value per gene. This is because MAD, as a measure of dispersion, is less effected by outliers than simply taking the mean (Chung et al., 2008).

Results:

Initial Analysis:

Initial Correlation Analysis (1):

Of the 34,357 genes in the Timepoint dataset, 32,775 were also present in the B-Cell Cancer Cell Line dataset. These 32,775 genes were then used in the correlation procedure (**Figure 3**). The results of the analysis showed each cell line was strongly positively correlated with each maturation timepoint, the minimum correlation coefficient being 0.62 with an average correlation coefficient of 0.74. The peak correlation coefficients were seen at the timepoints 0, 72 and 144 hours. There was also limited differences in correlation between cell lines and correlation average by cancer types and all correlation FDR values were significant. As well as the core Lymphoma Cancer Cell Line dataset described in the materials and methods, B-cell leukaemia cell lines were also included to give indications of any correlations between the time points and non-lymph node-based B-cell stages. These cell lines also significantly correlated with all timepoints.

It was hypothesised that differences between B-cell maturation stages hinges on only a few genes being up or down regulated, an effect which is masked by the noise of the correlation analysis being carried out across all the genes present within the dataset. This potentially explains why the various cancer types and the cell lines associated with them had identical correlation results despite a variety of cells of origin.

Time Series Differential Gene Expression:

To significantly reduce the number of genes involved in the correlation analysis, time series differential gene expression was performed on the B-cell maturation dataset using DESeq2. This enabled extreme filtering of the genes and the preferential selection of genes specifically associated with B-cell maturation due to their differential expression across the time series.

Approximately 50% (13,206) of the probes were found to be differentially expressed (0.05 significance level), with 1,512 meeting the additional filtering criteria of $\log_2$(fold-change) standard error < 25% quantile (0.26) and adjusted p-value < 25% quantile (1.72×10^{-17}). This criterion was used to help drastically reduce the number of significantly expressed probes in order to reduce the above-mentioned correlation noise and also had the side effect of reducing the time taken to perform the analysis.

Gene ontology (GO) analysis was performed on the filtered time series differential gene expression results to affirm that the results of the time series differential gene expression had preferentially selected B-cell maturation related genes (the top 10 terms found in **Table 4**).

The GO analysis shows high levels of significance of lymphatic terms. However, only a few terms were specifically B-cell related.

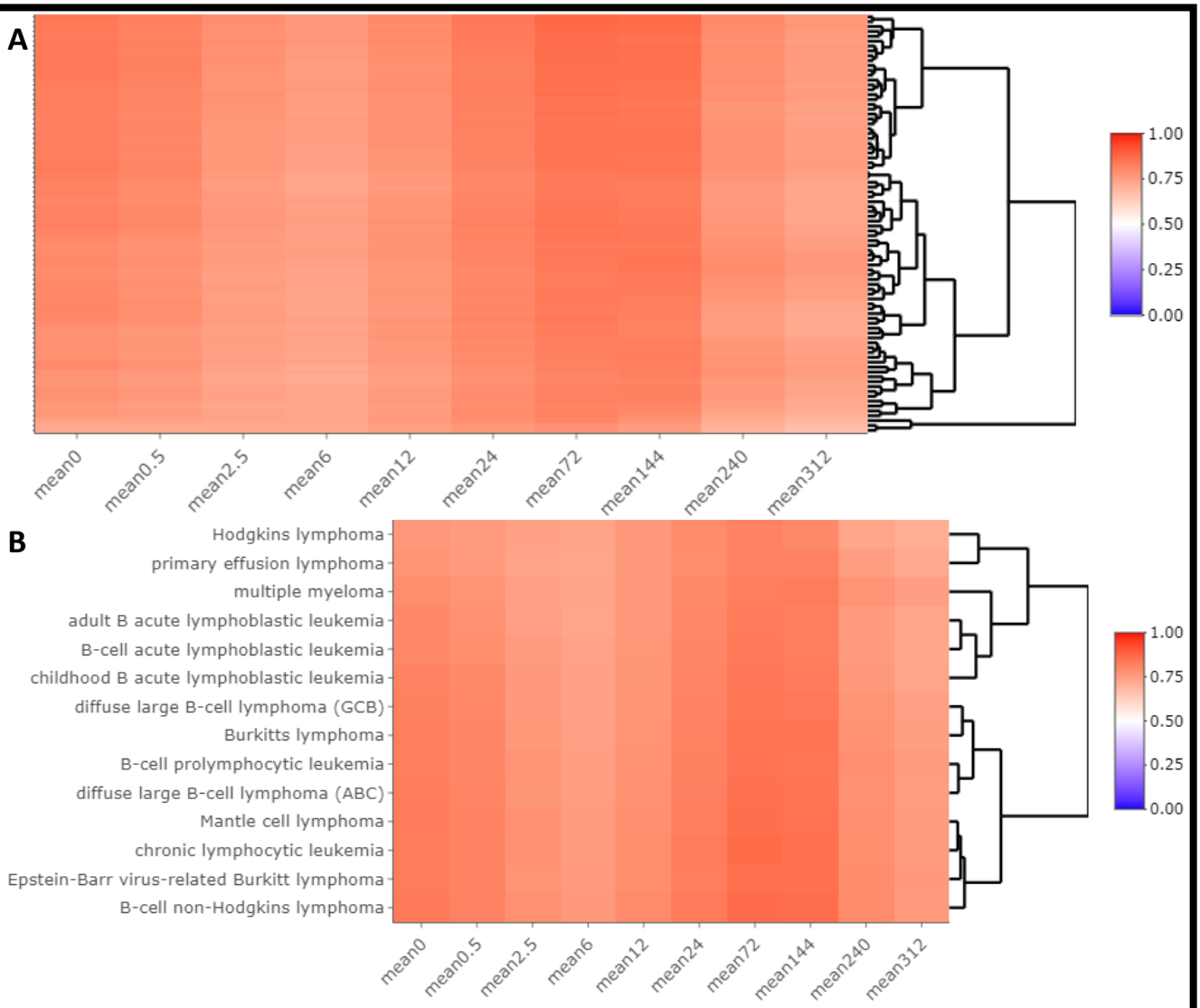


Figure 3: Initial Correlation Analysis (1) Heatmaps. A) Heat map correlation coefficient from the initial correlation analysis (1) between average B-cell maturation timepoint samples and the 86 cell lines that make up the B-cell cancer cell line dataset, using the 32,775 common genes. The y axis represents the individual cell lines, and the x axis represents the mean correlation between the cell line and that timepoint. **B)** Correlation heat map of 86 B-cell cell lines averaged by cancer type. The y axis represents the average cell line correlation for each cancer type, and the x axis represents the average correlation between the cancer type and that timepoint.

GO-ID	*Term*	*Weight.ks*
GO:0006958	Complement activation, classical pathway	$< 1x10^{-30}$
GO:0008228	Opsonization	$< 1x10^{-30}$
GO:0018149	Peptide cross-linking	$3.10x10^{-29}$
GO:0030449	Regulation of complement activation	$9.00\ x10^{-28}$
GO:0006898	Receptor-mediated endocytosis	$6.40\ x10^{-20}$
GO:0050853	B-cell receptor signalling pathway	$1.30\ x10^{-17}$
GO:0002862	Negative regulation of inflammatory response to antigenic stimulus	$2.50\ x10^{-17}$
GO:0050900	Leukocyte migration	$1.80\ x10^{-16}$
GO:0050871	Positive regulation of B-cell activation	$5.80\ x10^{-16}$
GO:0038096	Fc-gamma receptor signalling pathway involved in phagocytosis	$1.00\ x10^{-15}$

Table 4: TopGO Analysis of the DESeq Time Series Differential Gene Expression Results. The top 10 GO terms as identified by TopGO using the DESeq time series differential gene expression results that have been filtered down to 1,512 genes and ranked by Weight.ks p-value. Weight.ks being the result of TopGO's weight algorithm applied to the Kolmogorov–Smirnov test.

Initial Correlation Analysis (2):

Out of the 1,512 genes previously identified, 1,508 were also present in the B-cell Cancer Cell Line dataset. These were used in the second initial correlation analysis (**Figure 4**). This had the desired effect in reducing the correlation noise, however, introduced the unexpected results of moderately strong correlation being found almost exclusively at the 24- and 72-hour time points. It was next hypothesised that this strong signal of correlation could be down to the cell lines correlating at the points where the most proliferation is occurring in the time series, masking the correlation of genes relevant to B-cell biology. This is because not only is each cell line a cancerous cell line, so already has strong proliferative and survivability traits, but they have also been selected as a cell line due characteristics such as being easier to culture and better at surviving in a laboratory environment. This led to a theory that cell lines have a subset of significantly expressed proliferation and survival genes, so called "cell line genes", that enable them to be a good fit as a laboratory cell line. Thus, each cell line is correlating exclusively at the points in the B-cell time series where proliferation is occurring due to these "cell line genes".

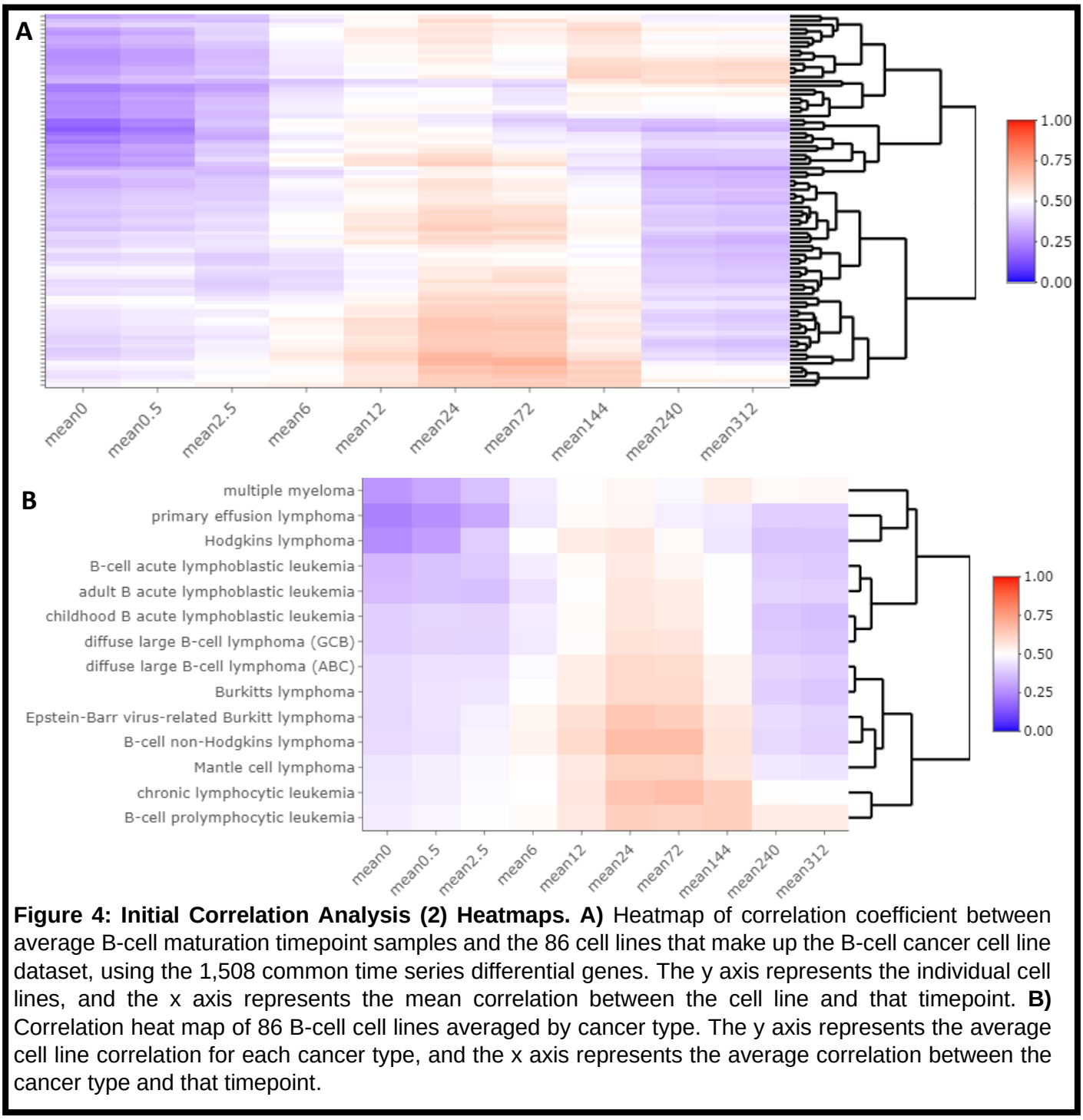


Figure 4: Initial Correlation Analysis (2) Heatmaps. A) Heatmap of correlation coefficient between average B-cell maturation timepoint samples and the 86 cell lines that make up the B-cell cancer cell line dataset, using the 1,508 common time series differential genes. The y axis represents the individual cell lines, and the x axis represents the mean correlation between the cell line and that timepoint. **B)** Correlation heat map of 86 B-cell cell lines averaged by cancer type. The y axis represents the average cell line correlation for each cancer type, and the x axis represents the average correlation between the cancer type and that timepoint.

Top 25 Genes:

To eliminate the “cell line genes” from the analysis, differential gene expression was carried out on the Cancer Cell Line dataset between the Lymphoma Cell Line subset and all other cancer cell lines. This filtered out the “cell line genes” as these would not be differentially expressed between the cancer cell lines.

Out of the 57,538 genes 5,337 genes were significantly differentially expressed with an absolute $\log_2$(fold-change) > 1 . This was further filtered down by filtering for an adjusted p-value < 25% quantile (5.6×10^{-31}) and an absolute $\log_2$(fold-change) > 25% quantile (1.23). GO analysis was then performed again to affirm B-cell related terms had been preferentially selected for and that potential ‘cell line genes’ had been removed. This was done by performing TopGO on both the genes selected in this process and the genes removed in this process (the top 10 terms found in **Table 5** and **6** respectively). The selected for genes contain significant GO terms related to lymphocytes, with some specific to B-cells, indicating some preferential selection B-cell expressed genes. The most significant terms that were filtered out were mostly associated with transcription and translation, giving a slight indication that the removal of the ‘cell line genes’ was successful.

GO-ID	***Term***	***weight.ks***
GO:0002250	Adaptive immune response	1.90×10^{-07}
GO:0050853	B-cell receptor signalling pathway	2.20×10^{-07}
GO:0038096	Fc-gamma receptor signalling pathway involved in phagocytosis	3.10×10^{-07}
GO:0019886	Antigen processing and presentation of exogenous peptide antigen via MHC class II	5.90×10^{-07}
GO:0048010	Vascular endothelial growth factor receptor signalling pathway	2.40×10^{-06}
GO:0090630	Activation of GTPase activity	8.70×10^{-06}
GO:0043547	Positive regulation of GTPase activity	5.80×10^{-05}
GO:0034446	Substrate adhesion-dependent cell spreading	1.40×10^{-04}
GO:2001237	Negative regulation of extrinsic apoptotic signalling pathway	3.30×10^{-04}
GO:0035329	Hippo signalling	6.80×10^{-04}

Table 5: TopGO Analysis of the Selected Differentially Expressed Cancer Cell Line Genes. TopGO analysis of the genes selected for by the additional differential analysis of the Cancer Cell Line dataset. Weight.ks being the adjusted p-value of theTopGO weight algorithm applied to the TopGO Kolmogorov–Smirnov test.

GO-ID	*Term*	*weight.ks*
GO:0002181	Cytoplasmic translation	$1.60x10^{-23}$
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	$2.80x10^{-21}$
GO:0000184	Nuclear-transcribed mRNA catabolic process, nonsens10-mediated decay	$1.60x10^{-17}$
GO:0019083	Viral transcription	$2.30x10^{-18}$
GO:0006413	Translational initiation	$3.90x10^{-18}$
GO:0006364	rRNA processing	$8.10x10^{-21}$
GO:0016032	Viral process	$6.00x10^{-10}$
GO:0000398	mRNA splicing, via spliceosome	$2.30x10^{-09}$
GO:0006406	mRNA export from nucleus	$1.00x10^{-06}$
GO:0090503	RNA phosphodiester bond hydrolysis, exonucleolytic	$2.10x10^{-04}$

Table 6: TopGO Analysis of the Excluded Differentially Expressed Cancer Cell Line Genes. TopGO analysis of the genes removed by the additional differential analysis of the Cancer Cell Line dataset. Weight.ks being the adjusted p-value of theTopGO weight algorithm applied to the TopGO Kolmogorov–Smirnov test.

GO-ID	*Term*	*Strength*	*FDR*
GO:0061903	Positive regulation of 1-phosphatidylinositol-3-kinase activity	2.11	$4.31x10^{-02}$
GO:0071073	Positive regulation of phospholipid biosynthetic process	1.62	$1.79x10^{-02}$
GO:0030889	Negative regulation of B cell proliferation	1.53	$2.74x10^{-02}$
GO:0050855	Regulation of B cell receptor signaling pathway	1.47	$4.40x10^{-03}$
GO:0050853	B cell receptor signaling pathway	1.45	$1.00x10^{-04}$
GO:0030888	Regulation of B cell proliferation	1.26	$7.70x10^{-04}$
GO:0002260	Lymphocyte homeostasis	1.26	$3.80x10^{-03}$
GO:0032715	Negative regulation of interleukin-6 production	1.25	$2.18x10^{-02}$
GO:0038096	Fc-gamma receptor signaling pathway involved in phagocytosis	1.2	$1.30x10^{-03}$
GO:1903727	Positive regulation of phospholipid metabolic process	1.16	$3.79x10^{-02}$

Table 7: String GO Analysis of the Top 102 Filtered Genes. Top 10 strongest enriched GO terms using String analysis on the 102 filtered genes. Strength is a $\log_{10}$ ratio of observed and expected annotated proteins in the network and FDR is the p-value corrected by the Benjamini–Hochberg procedure generating a False Discovery Rate.

Next, the top genes to use in the correlation analysis were identified by filtering for the genes that occurred in both of the filtered differential gene expression results. This left a total of 102 differentially expressed genes upon which GO analysis was performed (**Table 7**). This shows significant overrepresentation of B-cell related terms, showing that B-cell specific genes were successfully filtered for. Additionally, String network analysis generated a p-value of $< 1x10^{-16}$ meaning the network has significantly more interactions than statistically expected.

Rank	***Gene Name***	***Gene description***	***Adjusted P-value***
1	*RCSD1*	RCSD domain containing 1	$2.07x10^{-218}$
2	*SLAMF1*	Signalling lymphocytic activation molecule family member 1	$1.82x10^{-188}$
3	*ITGB7*	Integrin subunit beta 7	$2.65x10^{-187}$
4	*IRF8*	Interferon regulatory factor 8	$7.53x10^{-147}$
5	*CD37*	CD37 molecule	$1.24x10^{-146}$
6	*PIM2*	*PIM-2* proto-oncogene, serine/threonine kinase	$4.10x10^{-136}$
7	*KCNA3*	Potassium voltage-gated channel subfamily A member 3	$5.11x10^{-129}$
8	*TXNDC11*	Thioredoxin domain containing 11	$4.95x10^{-124}$
9	*ST8SIA4*	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 4	$1.78x10^{-123}$
10	*SNX20*	Sorting nexin 20	$8.67x10^{-121}$
11	*P2RX5*	Purinergic receptor P2X 5	$3.66x10^{-119}$
12	*IL2RA*	Interleukin 2 receptor subunit alpha	$1.08x10^{-118}$
13	*FCGR2B*	Fc fragment of igg receptor iib	$6.22x10^{-118}$
14	*SASH3*	SAM and SH3 domain containing 3	$8.76x10^{-118}$
15	*BLK*	*BLK* proto-oncogene, Src family tyrosine kinase	$9.97x10^{-109}$
16	*GNPDA1*	Glucosamine-6-phosphate deaminase 1	$3.72x10^{-97}$
17	*LY9*	Lymphocyte antigen 9	$3.39x10^{-95}$
18	*GNG7*	G protein subunit gamma 7	$6.86x10^{-90}$
19	*MEI1*	Meiotic double-stranded break formation protein 1	$4.05x10^{-85}$
20	*HERPUD1*	Homocysteine inducible ER protein with ubiquitin like domain 1	$9.80x10^{-84}$
21	*STAP1*	Signal transducing adaptor family member 1	$2.16x10^{-82}$
22	*CEACAM21*	CEA cell adhesion molecule 21	$6.93x10^{-79}$
23	*INPP5A*	Inositol polyphosphate-5-phosphatase A	$4.82x10^{-77}$
24	*IFNAR2*	Interferon alpha and beta receptor subunit 2	$1.36x10^{-76}$
25	*RASSF5*	Ras association domain family member 5	$2.32x10^{-76}$

Table 8: Summary of the Top 25 Genes. The top 25 genes selected for use in the correlation analysis ranked by time series differential gene expression adjusted p-value along with their expanded gene names.

Finally, ordering by p-value, the top 25 genes were selected for correlation analysis (**Table 8**). Only the top 25 genes were chosen because correlation analysis with more or less genes created more incoherent time point clusters (**Appendix figure 1**). Their log expression levels relative to the 0-hour time point is visualised in **Figure 5**. The top 25 genes are up- and down-regulated at various time points, crucially all genes between each donor have very similar log fold expression, therefore the genes are expressed at a constant rate between all donors. This indicates that any change in expression is likely to be due to maturation stages and not variation in the donors.

Of the top 25 genes, many are closely associated with B-cells. *SLAMF1* is expressed in B-cells and regulates the activation and differentiation of B-cells and other immune cells (Howie et al., 2002). At 2.5hrs it becomes significantly more expressed compared to the baseline (log_2(fold-change) > 4), as the time series continues however, it decreases to almost baseline levels by 312hrs (**Figure 5**). Unexpectedly, *GNPDA1* follows a very similar expression pattern to *SLAMF1*, however, this gene only seems to have a function in triggering egg activation and early embryo development, suggesting a potentially unknown role in B-cell maturation or this result is an anomaly (Zhang et al., 2003). *IL2RA* also follows this pattern of expression but is a type I transmembrane protein expressed on the surface of regulatory T-cells, not B-cells (Triplett et al., 2012).

PIM2 mediates cell survival and proliferation via action on *MYC*, increasing the stability of MYC via phosphorylation (Fox et al., 2003). It also phosphorylates BAD aiding in the resistance of apoptosis seen in cancer cells. Finally, it is a part of the NF-κB pathway which promotes the germinal centre B-cell to progress into a plasmablast cell (Ayala et al., 2004), (Klein and Heise, 2015). Reflecting this, *PIM2* only becomes significantly expressed from 144hrs onwards, suggesting that the samples 144hrs onwards are potentially post germinal centre B-cells (**Figure 5**).

CD37 has a very low level of expression throughout the time series and is most highly expressed earlier on (**Figure 5**). It is expressed at its highest levels in mature B-cells, indicating that the earliest time points represent antigen naïve mature B-cells (van Spriel et al., 2012). Following a similar pattern to *CD37*, *IFR8* regulates apoptosis and myeloid cell differentiation (Hu et al., 2011). Its continual downregulation could indicate that B-cell samples are becoming more proliferative and thus needing to resist apoptosis more strongly.

KCNA3, is expressed in B-cells in small amounts in naïve and early memory cells and significantly higher amounts in class switched memory B-cells (Wulff et al., 2004). Possibly indicating that at 144hrs onwards the samples could start representing memory cells as this is where *KCNA3* it is most highly expressed (**Figure 5**). Its upregulation at 0.5hrs also supports

the hypothesis that the early timepoints represent mature B-cells. However, *MEI1* also follows this expression pattern and is thought to only be essential in spermatocyte meiotic double stranded break formation (Sato et al., 2006).

FCGR2B can prevent early B-cell activation and can induce apoptosis in mature B-cells potentially explaining why it is under expressed in the early half of the maturation time series (**Figure 5**), (Smith and Clatworthy, 2010).

BLK is active during the very early timepoints of the time series before becoming significantly under expressed. It plays a role in pro-B-cell to pre-B-cell transition and signalling for growth arrest (Saijo et al., 2003). Although expressed at its highest at the beginning of the timeline, before being repressed it begins to increase in expression again after 24hrs, potentially indicating the slowing of proliferation after this time point (**Figure 5**). *STAP1* follows a similar expression pattern to *BLK. STAP1* interacts with *TEC* which in turn acts on *BTK* causing B-cell activation and development, explaining why it is expressed at its highest early on in the time series (Ohya et al., 1999).

Most of the selected top 25 genes are preferentially expressed in B-cell or other lymphatic tissues and have significant B-cell related functional enrichments. However, there are some outliers, namely *GNPDA1* and *MEI1*, which are expressed in gametes. Additionally, *Ly9* and *RASSF5* are expressed specifically in T-cells (Katagiri et al., 2003), (Chatterjee et al., 2012). This shows the differential gene expression analysis, and subsequent filtering of the genes, has been moderately successful in identifying genes related to the B-cell development pathway.

String network analysis was also conducted on the top 25 genes (F**igure 6**). The network was found to have significantly more interactions than statistically expected, with a p-value of 1.2×10^{-4}. This indicated that a significant number of the genes had interactions with each other or belonged to similar pathways. It identified a core network of seven interlinked genes, these are all linked through co-expression, which can be explained by them all having lymphoid-biased expression; however, each has little to no experimental data linking them.

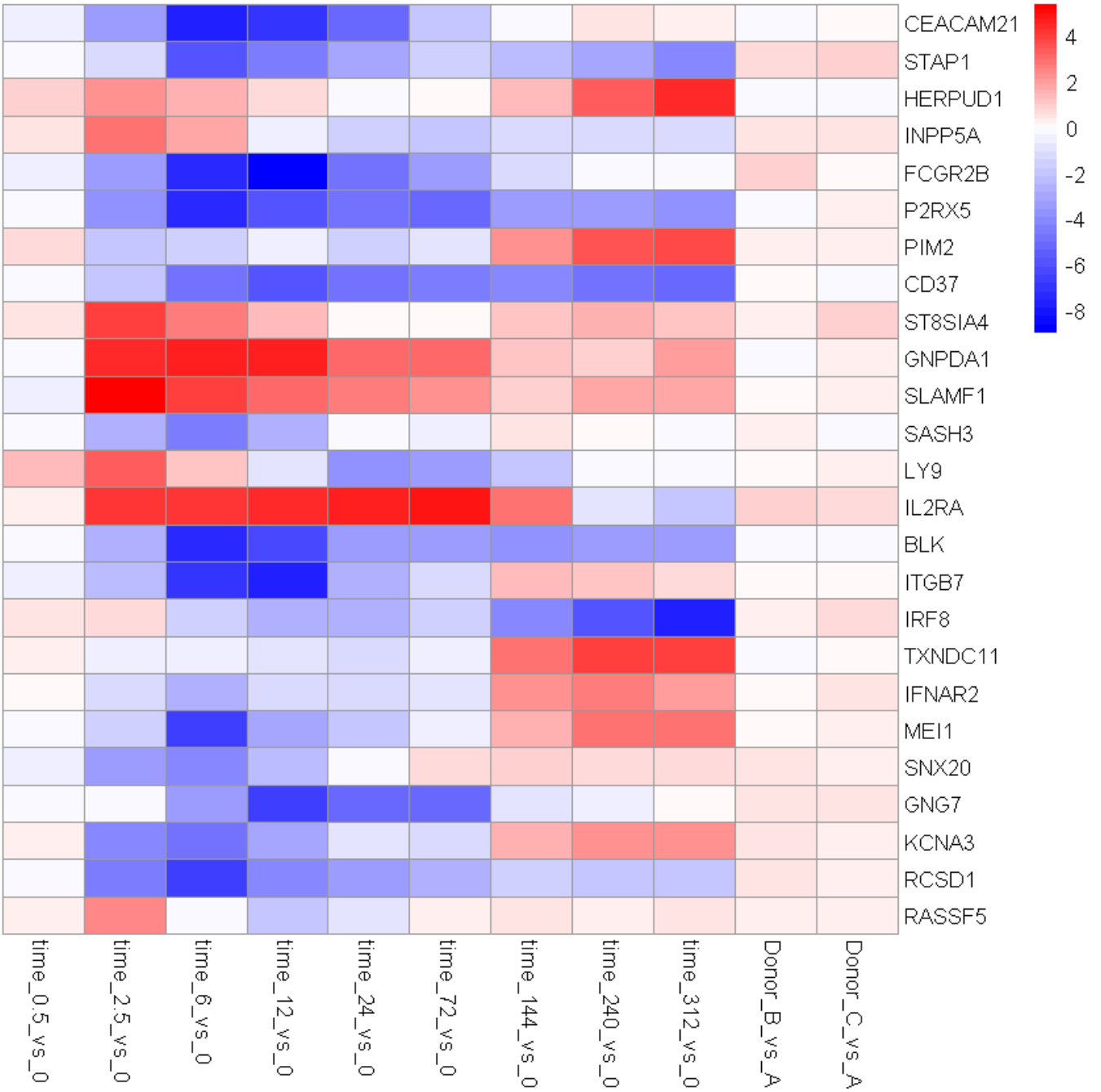


Figure 5: Heatmap of Top 25 Genes Time Series Expression Levels. A heat map of $\log_2$(fold-change) for each timepoint verses timepoint 0 for each of each of the top 25 genes identified in **Table 7** $\log_2$(fold-change) between donors shows how similar the expression of each gene is between the donors. The y axis represents each gene's expression level with dark blue representing a $\log_2$(fold-change) of -8 relative to the timepoint at 0 hours and dark red representing a $\log_2$(fold-change) of +4 relative to timepoint 0.

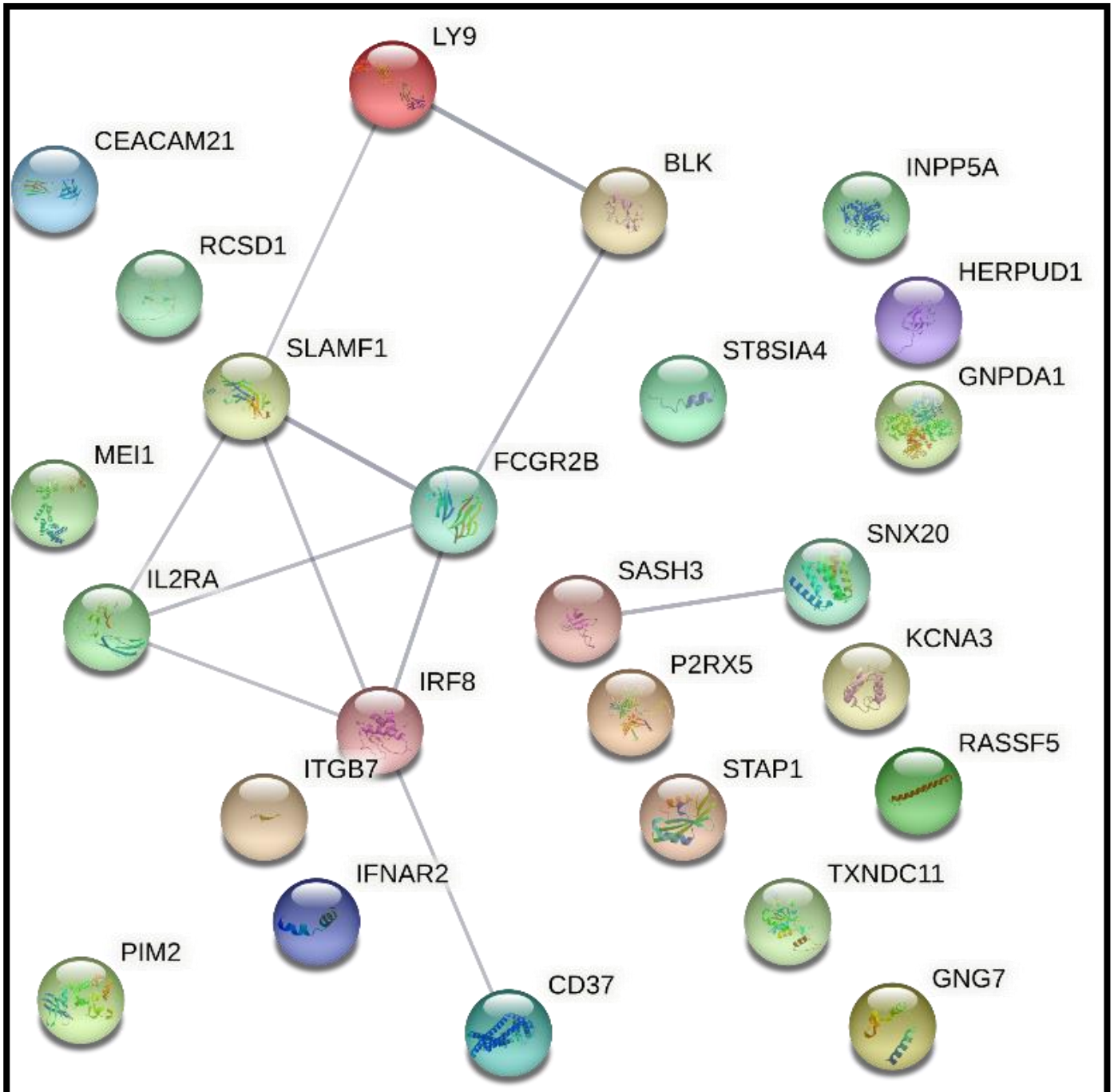


Figure 6: String Network Analysis of the Top 25 Genes. A String network analysis of the top 25 genes found in **Table 7.** The lines connecting the nodes indicate functional and physical associations between the proteins. For this analysis this evidence is in the form of co-expression found in past research contained in the String database.

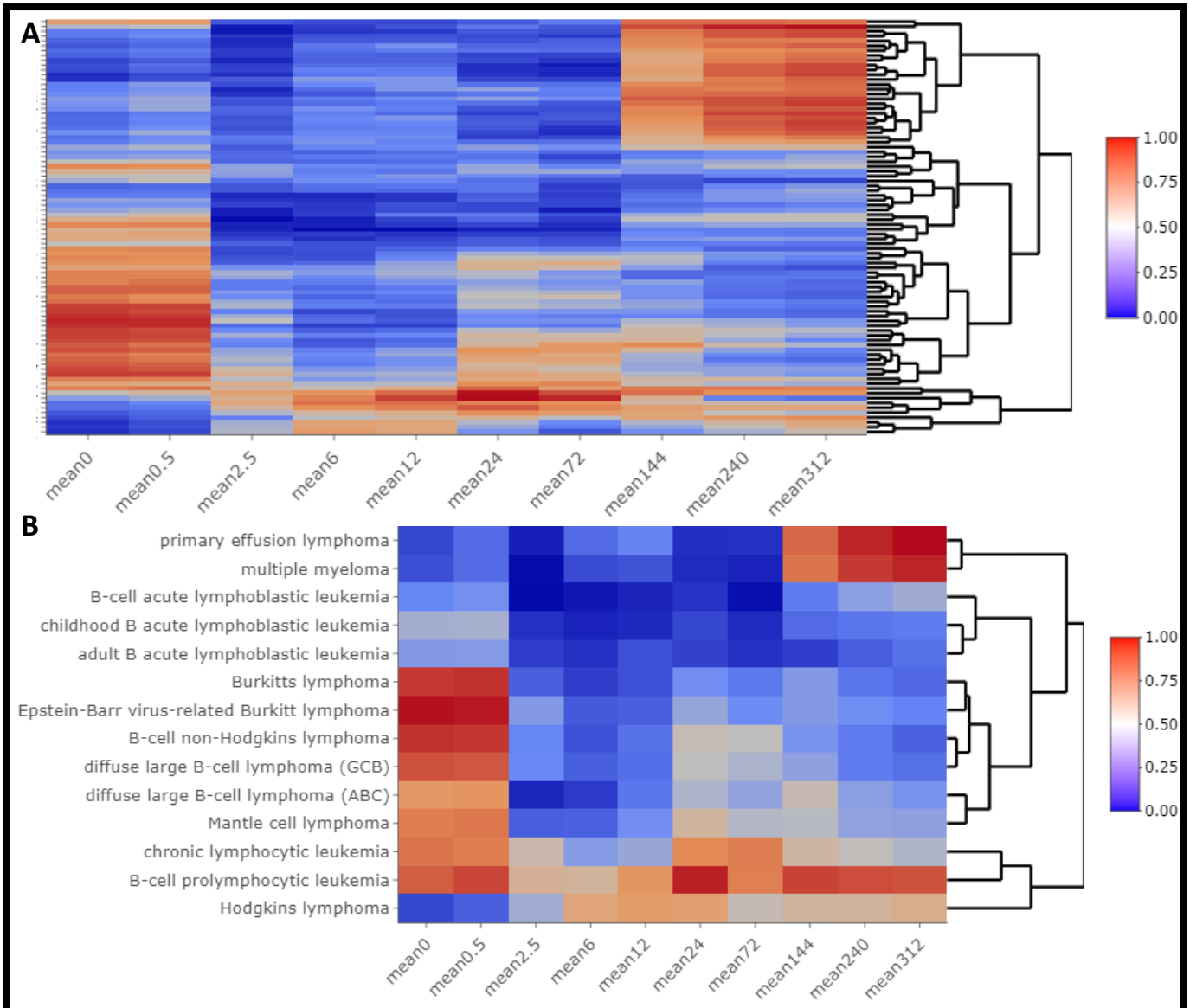


Figure 7: Final Correlation Analysis Heatmaps. A) Heat map of correlation coefficient between average B-cell maturation timepoint samples from the timepoint dataset and the 86 B-cell cell lines from the B-cell Cancer Cell Line dataset using the top 25 differentially expressed genes as identified in **Table 7**. **B)** Figure 7A averaged by cancer type. The x axis represents the average correlation of each timepoint sample for that timepoint and the y axis represents cell line or cancer type.

Final Correlation Analysis:

Correlation analysis was then performed using the top 25 genes on all B-cell related cancers (both lymphomas and leukemias) (**Figure 7**). The vast majority of cell lines that correlate with the timepoint samples only correlate in one of three major timepoints: early (0-0.5 hours), mid (2.5-72 hours) and late (144-312 hours), with some not correlating with any.

The cell lines that have no correlation are all B-cell acute lymphoblast leukemias (**Figure 7B**). These come from blast cells within the bone marrow and occur before the maturation time series dataset, so were unlikely to correlate significantly with any timepoint. B-cell prolymphocytic leukaemia primarily occurs in the bone marrow, before the stages of maturation in the time series data set, arising from mature B-cells (Cross and Dearden, 2019). Chronic lymphocytic leukaemia possibly arises from both pre- and post-germinal centre cells (Ciccone et al., 2014), (Ten Hacken et al., 2017). But, unlike the other leukemias they show strong correlation with most timepoints, potentially explained by Chronic lymphocytic leukaemia's aforementioned two potential cells of origin. This indicates that the time series does not account for many stages either side of the germinal centre.

Non-Hodgkin's lymphoma is a classification of lymphoma containing all lymphomas except Hodgkin's lymphoma, meaning the cell lines categorised as this in **Figure 7B** could belong to one of many lymphomas(Cancer Research UK, 2020). Hence, this gives no precise indication to what stage the early timepoint represents.

Hodgkin's lymphoma originates from Reed-Sternberg cells which, in turn, originate from germinal centre B-cells, these cells have typically lost characteristic B-cell expression patterns (Weniger and Küppers, 2021). As the Hodgkin's lymphoma cell lines have their peak correlation at the mid timepoint, this could suggest that this represents the germinal centre. However, the peak correlation coefficient is < 0.4, meaning this correlation is not strong (**Figure 7B)**. This could be due to the relatively indirect relation the germinal centre and the aforementioned loss of B-cell expression.

Burkitt lymphoma, commonly caused by the Epstein-Barr virus, primarily effects B lymphocytes in the germinal centre, namely centroblasts (Schmitz et al., 2014). These cell lines have their strongest correlation at the early timepoint indicating this could represent the germinal centre (**Figure 7B**).

DLBCL also represents a large subset of lymphoma types, the vast majority (around 80%) of DLBCL cases are known as 'not otherwise specified' (NOS) (Xie et al., 2015). The two main DLBCL NOS cell of origins, being ABC and GBC, as already mentioned (Alizadeh et al., 2000). The Lymphoma Cell Line dataset DLBCL samples were classified into the two respective cells

of origin using the DAC classifier (Care et al., 2013). GCB samples corelate the strongest at the early timepoint, strengthening the conclusion that the early timepoint could represent the germinal centre (**Figure 7B**). ABC is post-germinal centre in origin, however, still has its peak correlation at the early timepoint with a slight increase in the mid timepoint.

Mantel Cell Lymphoma arises from CD5 positive, antigen-naive pre-germinal centre B-cells, occurring before the germinal centre (Bertoni and Ponzoni, 2007). This likely means that the early and mid-time points both represent the germinal centre with some overlap between the before and after germinal centre stages due to these sample's correlation at these points (**Figure 7B**). However, this is distinct from the post-germinal stages represented by the late timepoint.

Finally, Primary Effusion Lymphoma and multiple myeloma only correlate at the late time point (**Figure 7B**). Primary Effusion Lymphoma is a plasmablast malignancy and multiple myeloma is a malignancy of plasma cells (Chen and Chuang, 2020). Both of these B-cell cancer types occur after the germinal centre stage, suggesting that the late time point is post-germinal cell.

Clustering and Classification:

Clustering:

PCA was carried out in R on both datasets before and after filtering using the top 25 genes previously identified (**Figure 8A-D**). Filtering by the top 25 genes vastly increased the variation accounted for by the first and second principal components, however, filtering has significantly reduced the number of components in the datasets, so this was expected. As seen in **Figure 8A** and **B**, the separate timepoints become more grouped together by their timepoints and suggests the possibility of time point analysis with a higher resolution given more samples.

PCA on the Lymphoma Cell Line dataset revealed two core clusters of cells but the first two principal components only accounted for around 22% of the variability (**Figure 8C**). When filtering by the top 25 genes however, the dataset separated into three clusters, early (0-0.5 hours), mid (2.5-72 hours) and late (144 hours to 312 hours), (using K-means clustering) (**Figure 8D**). These clusters directly represent the three major time points identified in the correlation analysis stage. Again, the variability accounted for by the first two principal components vastly increases.

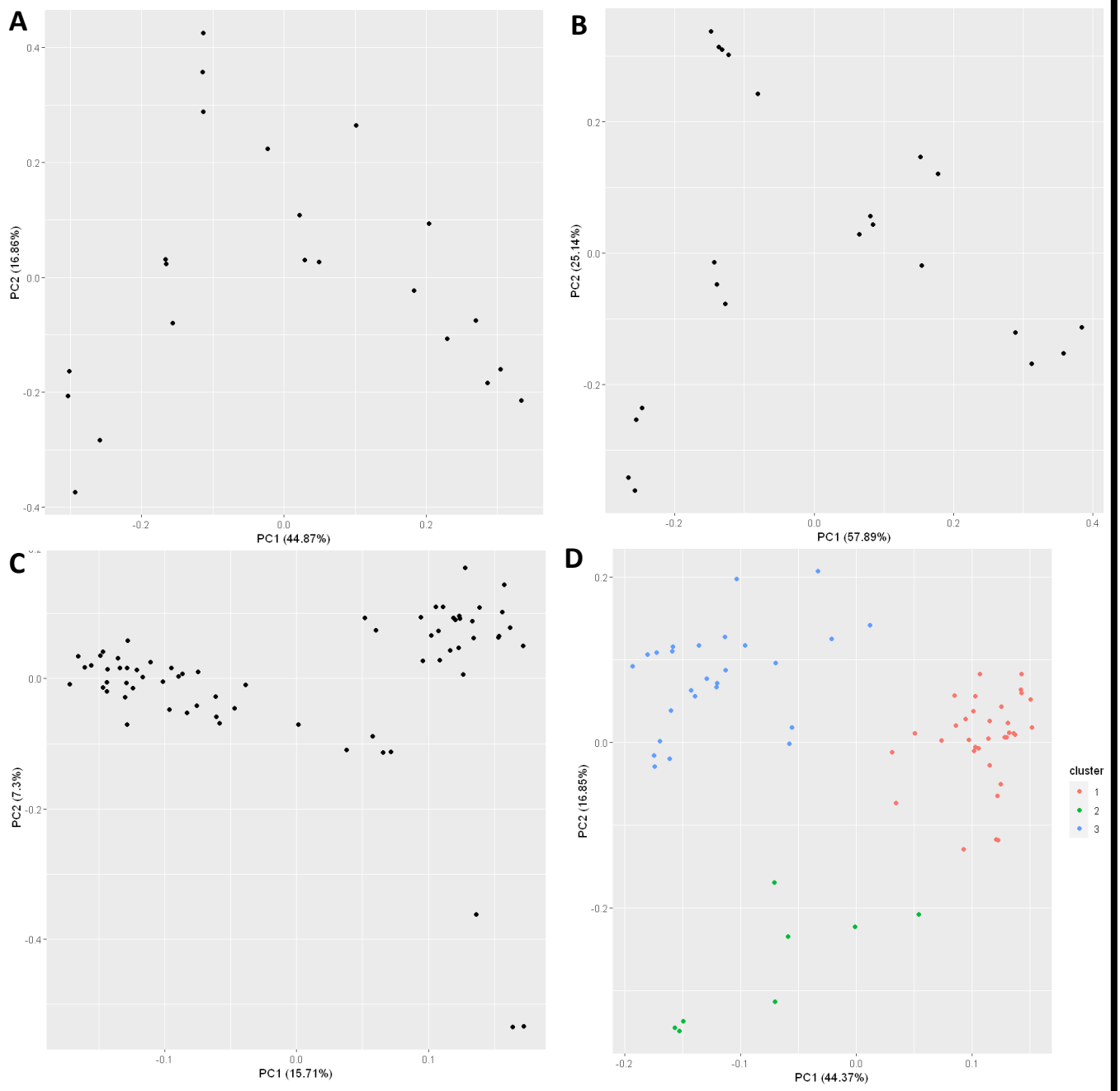


Figure 8: Principal Component Analysis. PCA of the timepoint dataset before (A) and after filtering with the top 25 genes (B) and the Lymphoma Cell Line dataset before (C) and after filtering with the top 25 genes (D). For D, cluster 1 represents the early timepoint, cluster 2 represents the mid time point and cluster 3 represents the late time point.

Classification:

The Cancer cell line dataset was then combined with the time series datasets and the three time point clusters identified in **Figure 8D** were used to define the three classifications for the classifier to use. Within this dataset the number of mid time point is underrepresented compared to the other timepoints, having only around half the number of samples of the other two time points (**Figure 9A**). This could cause potential problems during the cross-validation where splitting the data could prevent any samples from the mid time point being selected for testing.

Initially, a random forest classifier was trained and cross-validated on the combined dataset with an average accuracy of 98.4%, using SciKit-Learn. The 25 genes were then be attributed a feature importance score, showing how important to the classifier each of the genes was (**Figure 9B**). Of the top 5 most significant top 25 genes, only one (*RCSD1*) was in the topmost important classifier genes and the most important classifier gene (*SASH3)* was the 14th significant of the top 25 genes. Analysis of the top 5 important classifier genes, in conjunction with F**igure 5,** show *SASH3* and *PIM2* have a distinct polarised expression. They are both least expressed at the early time point, slightly more expressed during the mid-time point before being strongly expressed during the late stage. The rest of the top five (*STAP1*, *BLK* and *RCSD1*) all show very similar expression levels with very gradual changes through the time series (**Figure 5)**. Interestingly, there does not seem to be any correlation between level of significance in differential expression of the gene and usefulness attributed to the gene by the classifier. Furthermore, there also seems to be no correlation between significant changes in gene expression over the time series and usefulness attributed to the gene by the classifier.

Finally, cross-validation was carried out on several models, all using default settings, including an ensemble stacking classifier which used the results of all the models, except the decision tree, to vote on the classification of a sample (**Table 9**). The decision tree was not used in the stacking classifier due to it being rendered redundant by the random forest classifier. Cross-validation used a testing: training ratio of 25:75 and was carried out 1,000 times.

Out of the 1,000 rounds of cross-validation, unexpectedly, the Logistic Regression model was narrowly proclaimed the best model, with the ensemble stacking classifier a close second. By far the worst classifier was the singular decision tree classifier which scored the worst in each singular metric. However, as there is very little difference between the Logistic Regression model and the Stacking classifier, and the stacking classifier has the lowest false positive rate, thus the stacking classifier was used in future tumour sample analysis.

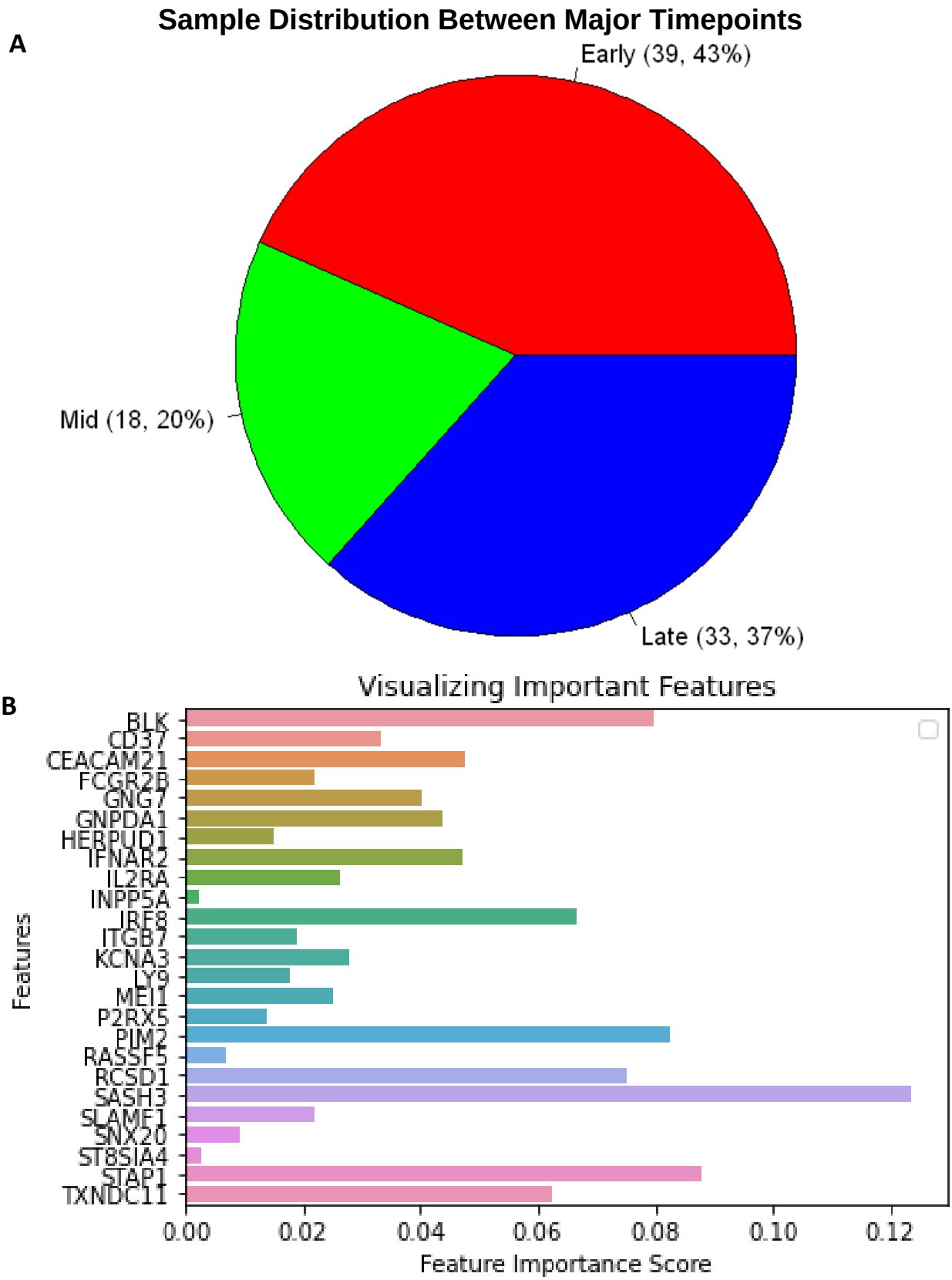


Figure 9: Pre-Classification Analysis. A) Pie chart showing the number and percentage of samples belonging to each major timepoint in the combined Lymphoma Cell Line and Timepoint datasets. **B)** Bar chart showing the feature importance attributed to each of the top 25 gene by the random forest classifier.

Model	TPR	FPR	Precision	Recall	F-measure	Accuracy	ROC Area	Average Rank
Logistic Regression	0.990747	0.005739	0.990747	0.992064	0.990851	0.989783	0.999952	1.6
K-Neighbours Classifier	0.977135	0.012321	0.977135	0.976956	0.973572	0.977565	0.999672	6.6
Decision Tree Classifier	0.882190	0.070140	0.882190	0.878974	0.869371	0.878435	0.905443	8.0
Random Forest Classifier	0.981218	0.009443	0.981218	0.982362	0.979776	0.982739	0.999667	5.1
Support Vector Classifier	0.985868	0.005483	0.985868	0.991626	0.988015	0.989174	0.999930	2.9
Naive Bayes Classifier	0.977578	0.010576	0.977578	0.984549	0.979256	0.980478	0.999142	6.0
MLP Classifier	0.981892	0.005506	0.981892	0.990246	0.984918	0.987957	0.999556	4.1
Stacking Classifier	0.985979	0.005466	0.985979	0.991659	0.988100	0.989217	0.999997	1.7

Table 9: Cross-Validation Results. The average cross-validation results from 1,000 repeats of a testing: training ratio of 25:75 using several different metrics. F-measure is the harmonic mean of precision and recall and ROC area is the area under the receiver operating characteristic curve. TPR – True positive rate and FPR -false positive rate. Accuracy is the proportion of correct predictions out of all predictions and precision is the proportion of true positives out of all positives. Recall is the proportion of true positives out of the total true positives and false negatives.

Tumour Sample Analysis:

REMoDL-B Dataset:

The 928 DLBCL samples come from the REMoDL-B dataset described in the methods (**Figure 10**). These have been pre-classified into cell of origin by the DAC classifier (Care et al., 2013). Along with GCB and ABC there are two more classifications: Molecular High Grade (MHG), which is a molecularly distinct DLBCL subclassification that presents with a double hit phenotype and a very poor prognosis, and the unclassifiable subclassification (UNC) (Sha et al., 2019). An initial correlation analysis was performed on the dataset (**Figure 11**), before being classified by the stacking classifier. All samples correlated at the Mid time point, regardless of what subclassification they belonged to, unlike the Cancer Cell Line dataset samples of the same DLBCL subclassification. Unsurprisingly, the classifier classified all samples as belonging to the Mid time point.

Next correlation analysis was performed comparing the DLBCL GCB and ABC Cancer Cell Line samples to the REMoDL-B dataset samples (**Figure 12**). This identified the cell line Pfeiffer as the most similar to the ABC samples (coefficient of 0.54). The GCB samples also had their highest cell line correlation with Pfeiffer, potentially explained by Pfeiffer being the only DLBCL cell line to correlate at the Mid time point. However, the correlation coefficient was below 0.5 (0.49), so, was not particularly strong. The cell line DB was the only DLBCL GCB cell line to correlate significantly with the REMoDL-B samples and had its maximum correlation with the MHG subclassification (coefficient of 0.54).

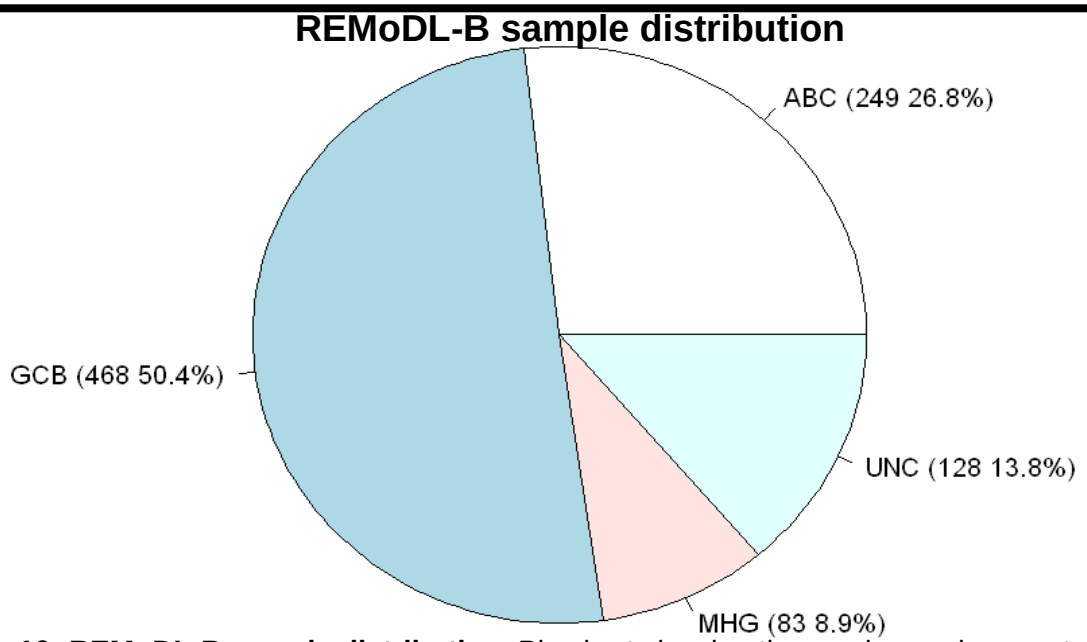


Figure 10: REMoDL-B sample distribution. Pie chart showing the number and percentage of samples belonging to each DLBCL subclassification classification present in the REMoDL-B dataset including the two cells of origins (GCB and ABC) as well as molecular high grade (MHG) and unclassifiable (UNC).

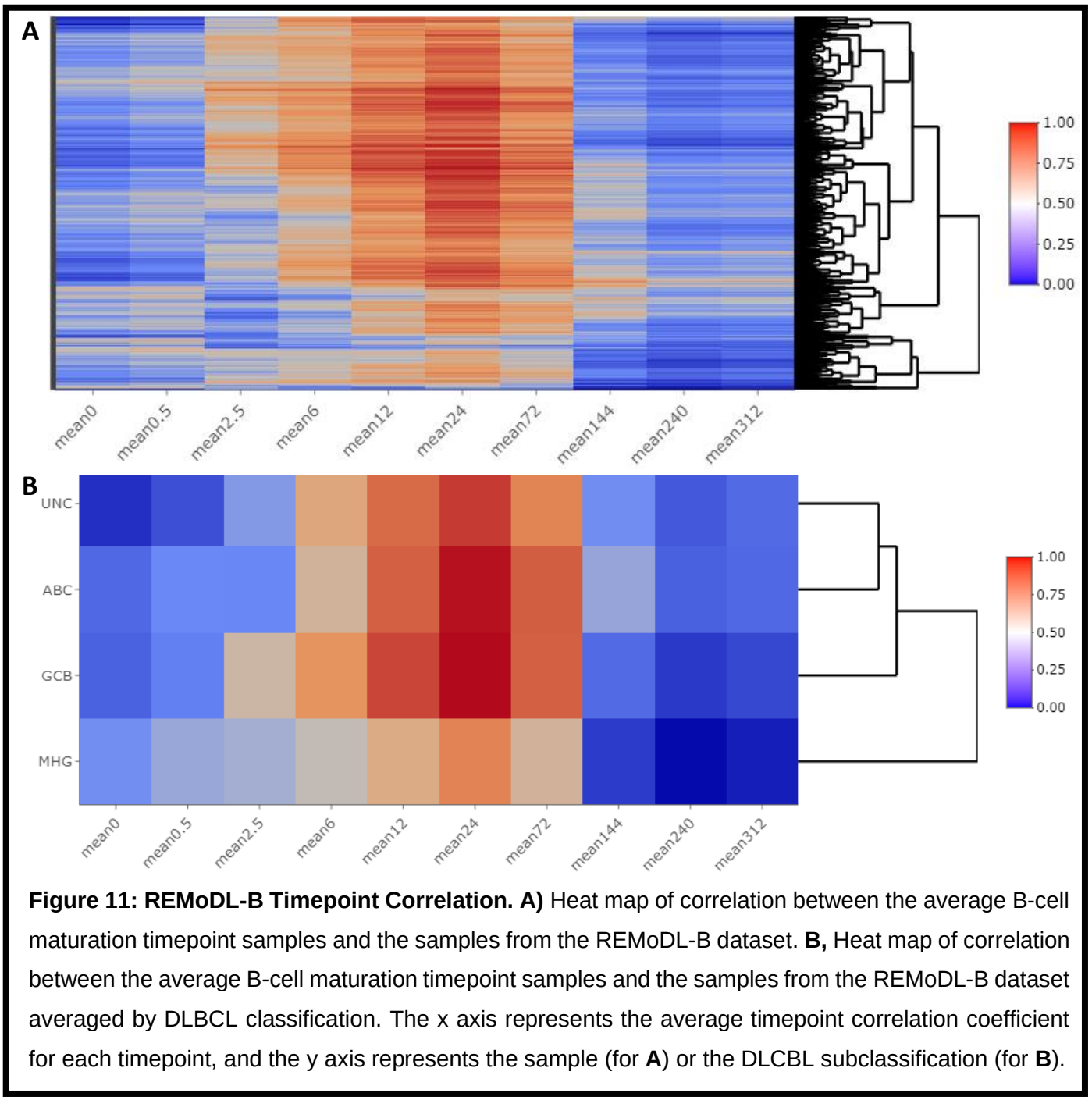


Figure 11: REMoDL-B Timepoint Correlation. A) Heat map of correlation between the average B-cell maturation timepoint samples and the samples from the REMoDL-B dataset. **B,** Heat map of correlation between the average B-cell maturation timepoint samples and the samples from the REMoDL-B dataset averaged by DLBCL classification. The x axis represents the average timepoint correlation coefficient for each timepoint, and the y axis represents the sample (for **A**) or the DLCBL subclassification (for **B**).

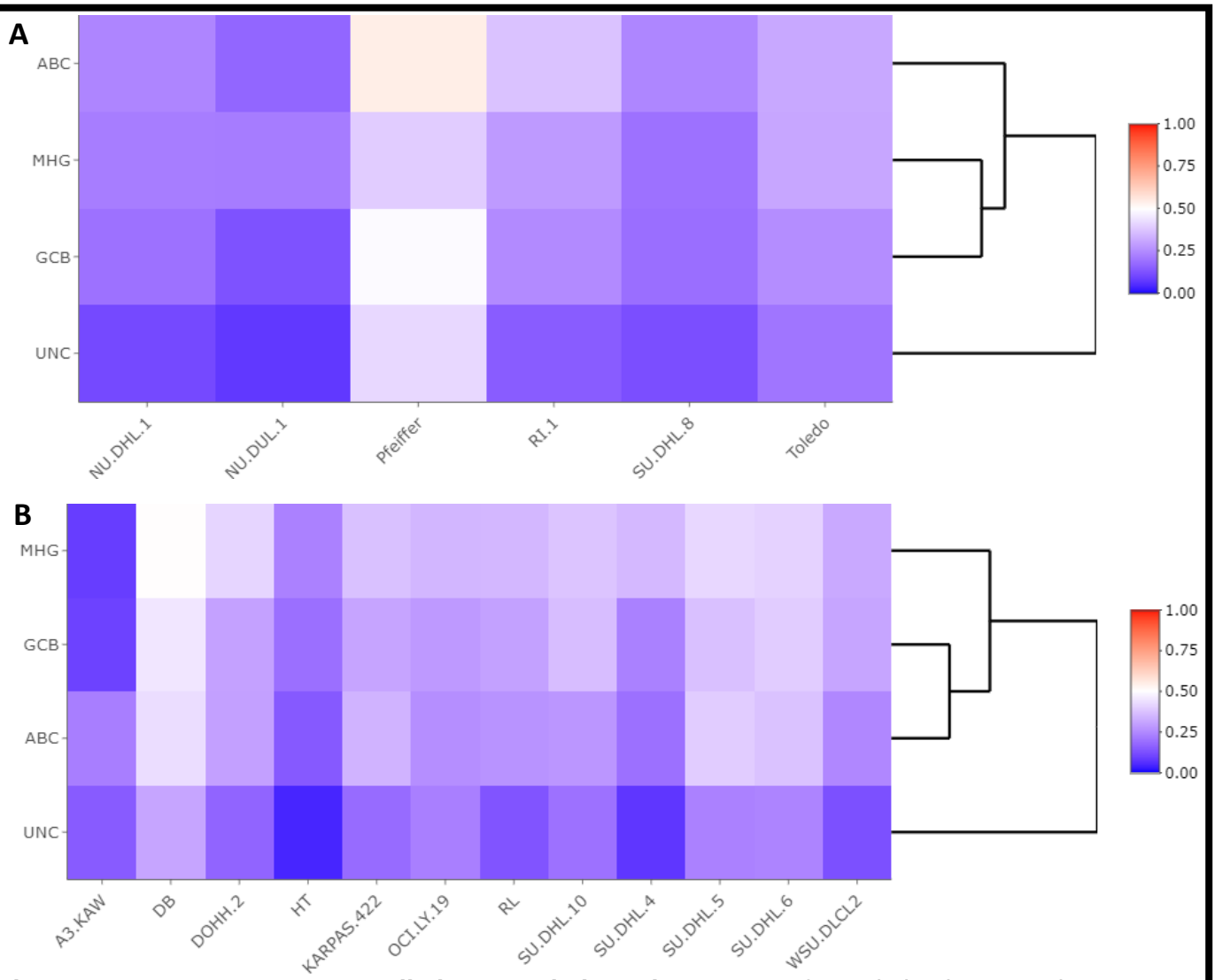


Figure 12: REMoDL-B Cancer Cell Line Correlation. A) Heat map of correlation between the DLBCL ABC classified Cancer Cell Line dataset samples and the samples from the REMoDL-B dataset averaged by DLBCL classification. **D,** Heat map of correlation between the DLBCL GCB classified Cancer Cell Line dataset samples and the samples from the REMoDL-B dataset averaged by DLBCL classification. The x axis represents the cell line correlation coefficient, and the y axis represents the samples correlation coefficient averaged by DLCBL subclassification.

E-GEOD-2658:

The E-GEOD-2658 dataset exclusively contains 558 multiple myeloma samples, all labelled as originating from plasma cells from the bone marrow. This dataset contained multiple expression values for some of the top 25 genes. MAD was used to create a single expression value for each of these genes. Correlation analysis identified two peak correlations, a minor one at the early timepoint and a major one at the late timepoint which aligns with the correlation analysis of the multiple myeloma cancer cell lines (**Figure 13A**). Correlation analysis between

the samples and multiple myeloma cell lines identified the cell line KARPRAS-620, KMS-12-BM, MM1-S and MOLP-2 as some of the most similar cell line (**Figure 13B**).

The samples were then classified (**Figure 14A**). Surprisingly, only 67% were classified as the expected late timepoint, with the rest being the mid timepoint and one belonging to the early timepoint. The average correlation by timepoint classification shows that all samples, regardless of timepoint classification, have their strongest correlation at the late timepoint (**Figure 14B**). Despite the classification of mid timepoint, the samples belonging to this classification only correlate at the late timepoint. The sample classified as early timepoint has dual peaks at the early and late timepoints.

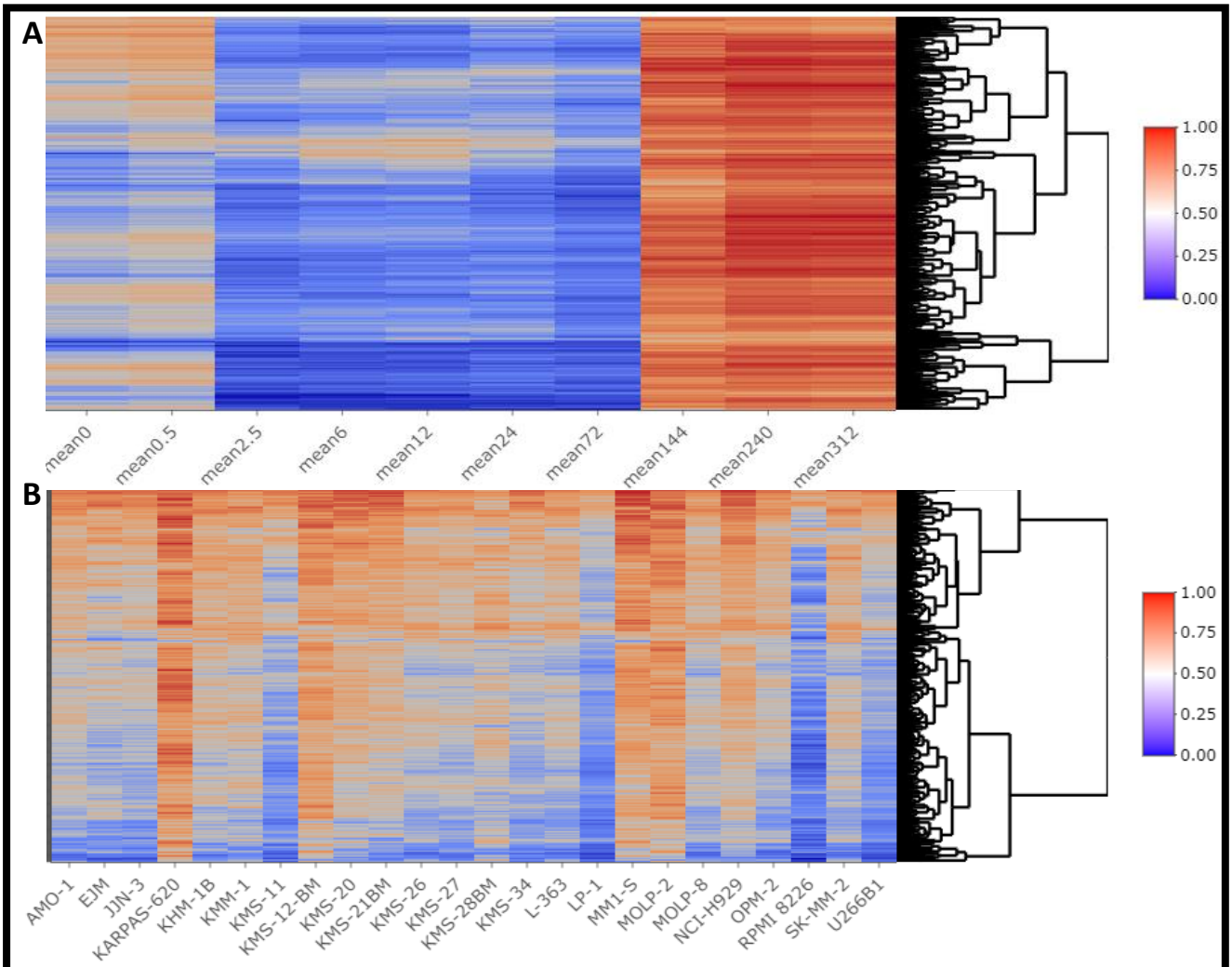


Figure 13: E-GEOD-2658 Correlation Analysis. A) Heatmap of correlation coefficient between the average B-cell maturation timepoint samples and the samples from the E-GEOD-2658 dataset. **B)** Heatmap of correlation coefficient between the multiple myeloma Cancer Cell Line dataset samples and the samples from the E-GEOD-2658 dataset. The x axis represents the average timepoint correlation coefficient for each timepoint (for **A**) or the multiple myeloma cell line (for **B**), and the y axis represents the sample correlation.

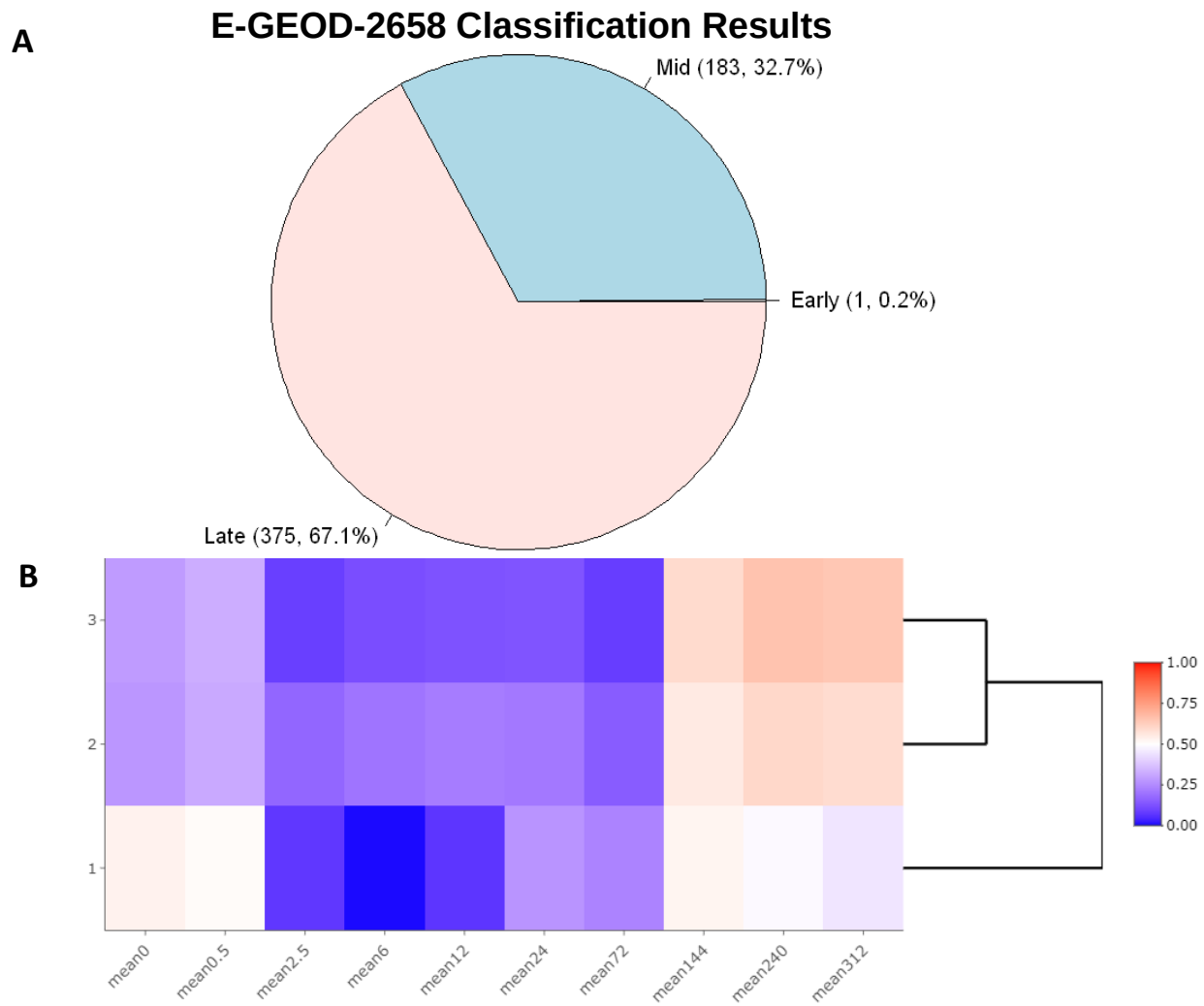


Figure 14: E-GEOD-2658 Classification Results. A) Pie chart of the E-GEOD-2658 dataset classification results. **B)** Heatmap of the timepoint correlation coefficient averaged by timepoint classification (1: Early, 2: Mid, 3: Late). The x axis represents the average timepoint correlation coefficient for each timepoint, and the y axis represents average by classification.

E-GEOD-4475:

The E-GEOD-4475 contains 221 samples primarily from DLBCL and BL cancers (**Figure 15**). Only 20 of the top 25 genes had expression values in this dataset so the correlation analysis was run using only the genes present out of the top 25 genes (**Figure 16**). Of the 20 genes present, some had multiple probes associated with them and single values were calculated using MAD. All samples correlated at the early timepoint regardless of cancer type, like their respective cell lines. A small subset of samples also correlated with the mid timepoint, the strongest correlation at this timepoint was mostly due to the DLBCL-ABC group.

The stacking classifier was then retrained on the 20 genes that were present in the E-GEOD-4475 dataset in order to classify them, cross-validation, using the same procedure as above, resulted in no change in overall accuracy. The samples were then classified using the

retrained classifier (**Figure 17A-F)**. As expected from the correlation analysis all the samples were classified in either the early or mid-timepoint. However, unlike the correlation analysis would have suggested, the majority of samples were classified as the mid timepoint and not early timepoint. Examination of the average correlation by timepoint classification shows that both timepoints have two peaks however the peak is strongest at the early timepoint, even for the group classified as the mid timepoint (**Figure 17G**).

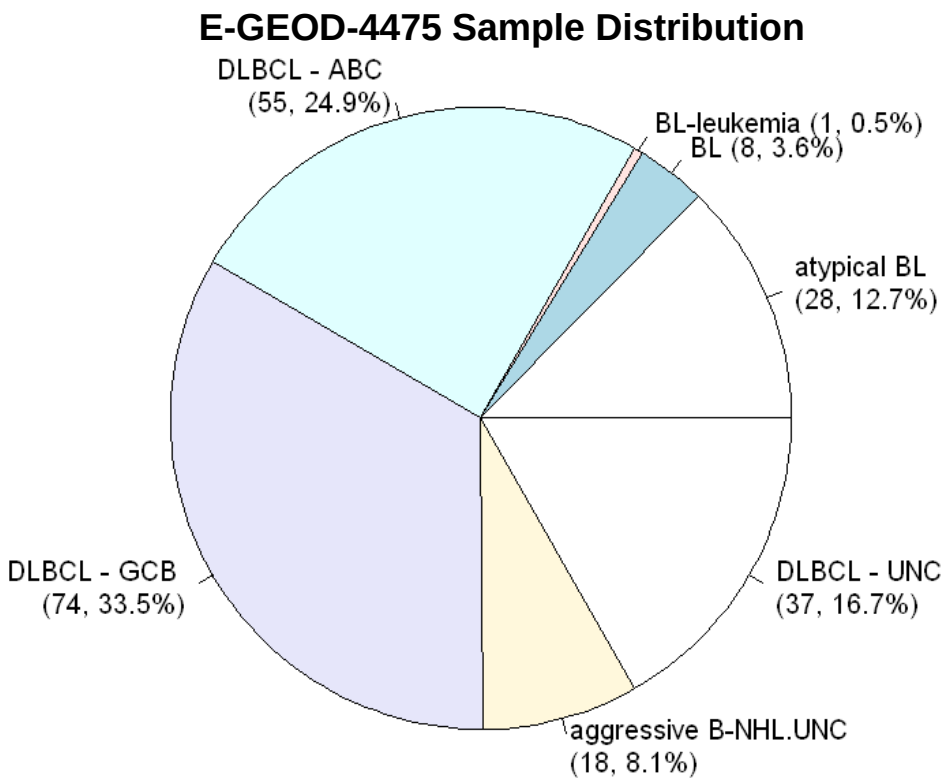


Figure 15: E-GEOD-4475 Sample Distribution. Pie chart showing the number and percentage of samples belonging to each cancer type in the E-GEOD-4475 dataset including three subclassifications of DLBCL (GCB, ABC and UNC), aggressive B Non-Hodgkin's lymphoma unclassified (B-NHL.UNC) and Burkitt Lymphoma (BL).

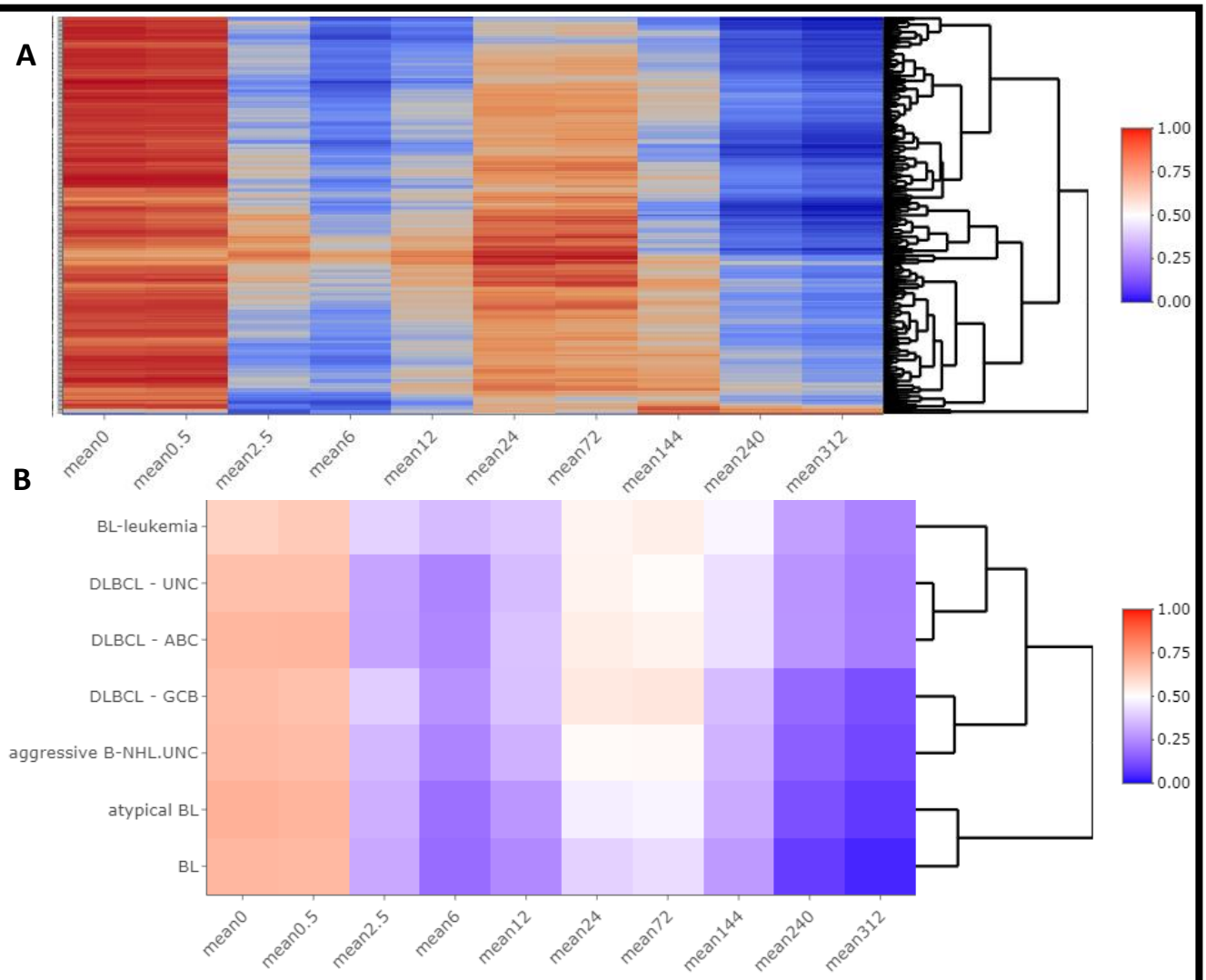


Figure 16: E-GEOD-4475 Correlation Analysis. A) Heat map of correlation coefficient between the average B-cell maturation timepoint samples and the samples from the E-GEOD-4475 dataset. **B)** Heat map of correlation between the average B-cell maturation timepoint samples and the samples from the E-GEOD-4475 dataset averaged by cancer type. The x axis represents the average sample for each timepoint, and the y axis represents samples (for **A**) and samples averaged by cancer type (for **B**). The samples include three subclassifications of DLBCL (GCB, ABC and UNC), aggressive B Non-Hodgkin's lymphoma unclassified (B-NHL.UNC) and Burkitt Lymphoma (BL).

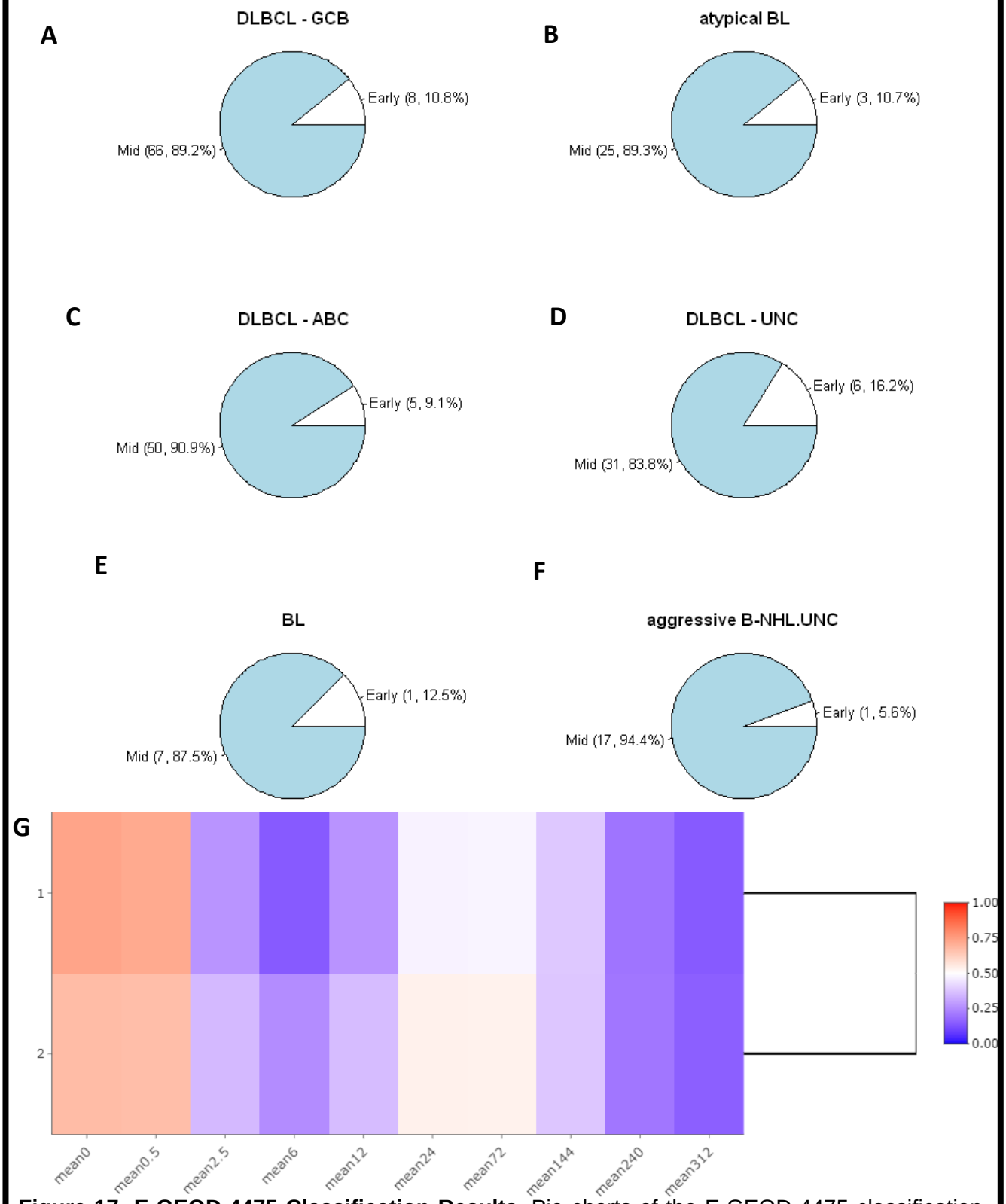


Figure 17: E-GEOD-4475 Classification Results. Pie charts of the E-GEOD-4475 classification results. **A)** DLBCL-GCB samples, **B)** Atypical Burkitt Lymphoma (BL) samples, **C)** DLBCL-ABC samples, **D)** DLBCL-UNC samples **E)** BL samples **F)** aggressive B Non-Hodgkin's lymphoma unclassified (B-NHL.UNC) samples. BL-leukaemia is not listed due to all samples being classified as early timepoint. **G)** Each classifications average correlation results with the timeseries dataset (1: Early, 2: Mid).

E-GEOD-31161:

Like the E-GEOD-2658 dataset, the E-GEOD-31161 dataset also contains exclusively multiple myeloma samples (1038 samples total). It contained all 25 genes, some of which had multiple values, these were combined into one value per gene by MAD. Correlation and classification analysis was performed showing near identical results to the E-GEOD-2658 dataset, as was expected due to all samples belonging to the multiple myeloma cancer type (**Figure 18** and **19**).

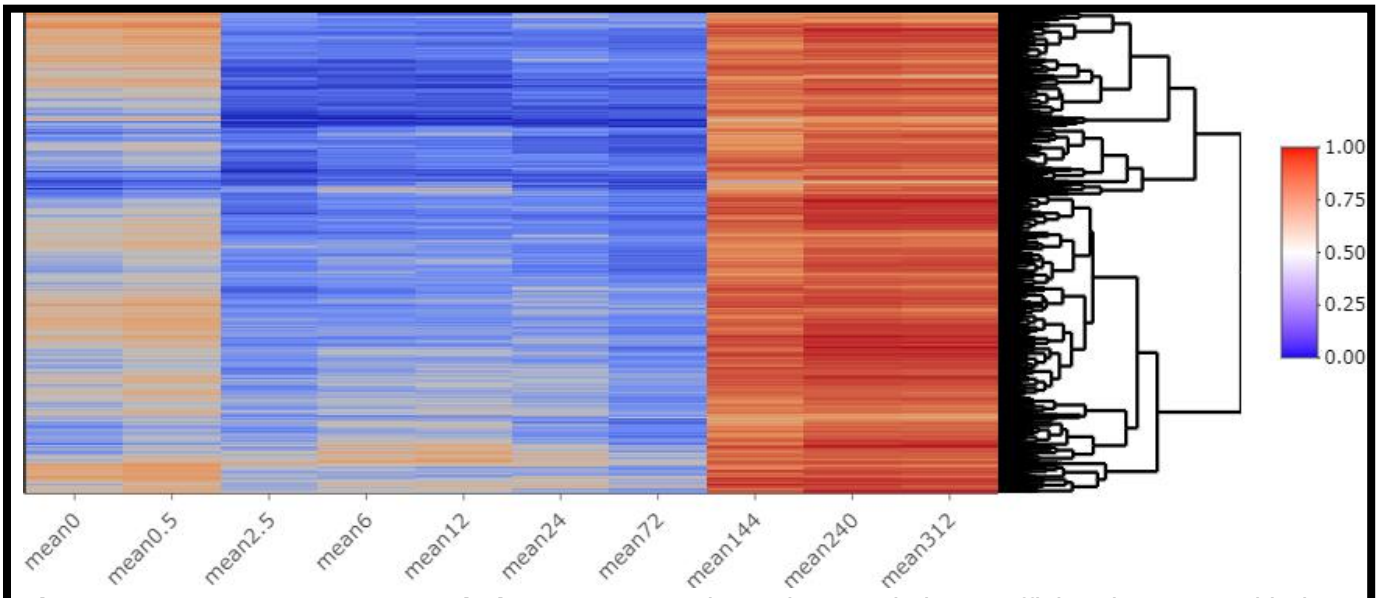


Figure 18: E-GEOD-31161 Correlation Heatmap. Timepoint correlation coefficient heatmap with the E-GEOD-2658 dataset. The x axis represents the average timepoint correlation and the y axis represents the samples.

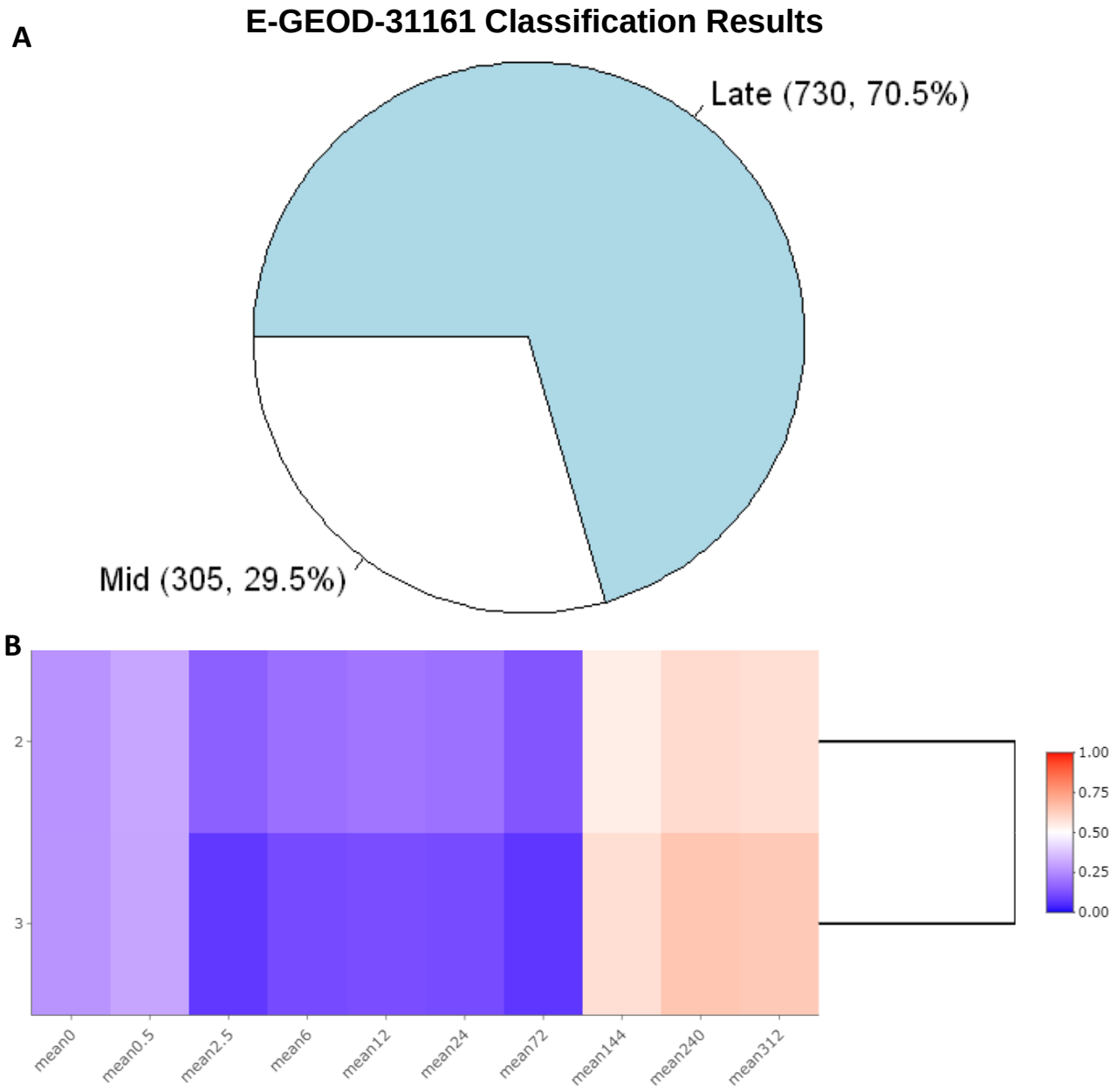


Figure 19: E-GEOD-31161 Classification Results. A) Pie chart of the distribution of timepoint classification in the E-GEOD-31161 dataset **B)** Timepoint correlation coefficient averaged by classification. The x axis representing the average timepoint sample and the y axis representing the average by classification.

E-GEOD-4732:

Finally, the E-GEOD- 4732 dataset contained a mix of DLBCL and BL samples (totalling 303 samples) (**Figure 20**). Included in this was Primary mediastinal large B-cell lymphoma (PMBL DLBCL). Previous *in silico* characterisation of PMBL DLBCL samples identified the cell of origin as GCB like (Duns et al., 2021). Only 13 of the top 25 genes were present in the dataset with multiple probes present, again the probes were combined using MAD. Despite indications that samples belong to the GCB subgroups, all samples correlated the most with the early timepoint (**Figure 21**).

The classifier was retrained on the 13 genes present, causing a small decrease in accuracy to 0.983. The samples were then classified and 77.2% of the samples were classified as the mid timepoint (**Figure 22A-E**). All the MHG subgroups were classified as the early timepoint. The mid timepoint group had a small increase in corelation within the mid timepoint range, however, significantly correlates in the early timepoint range (**Figure 22F**).

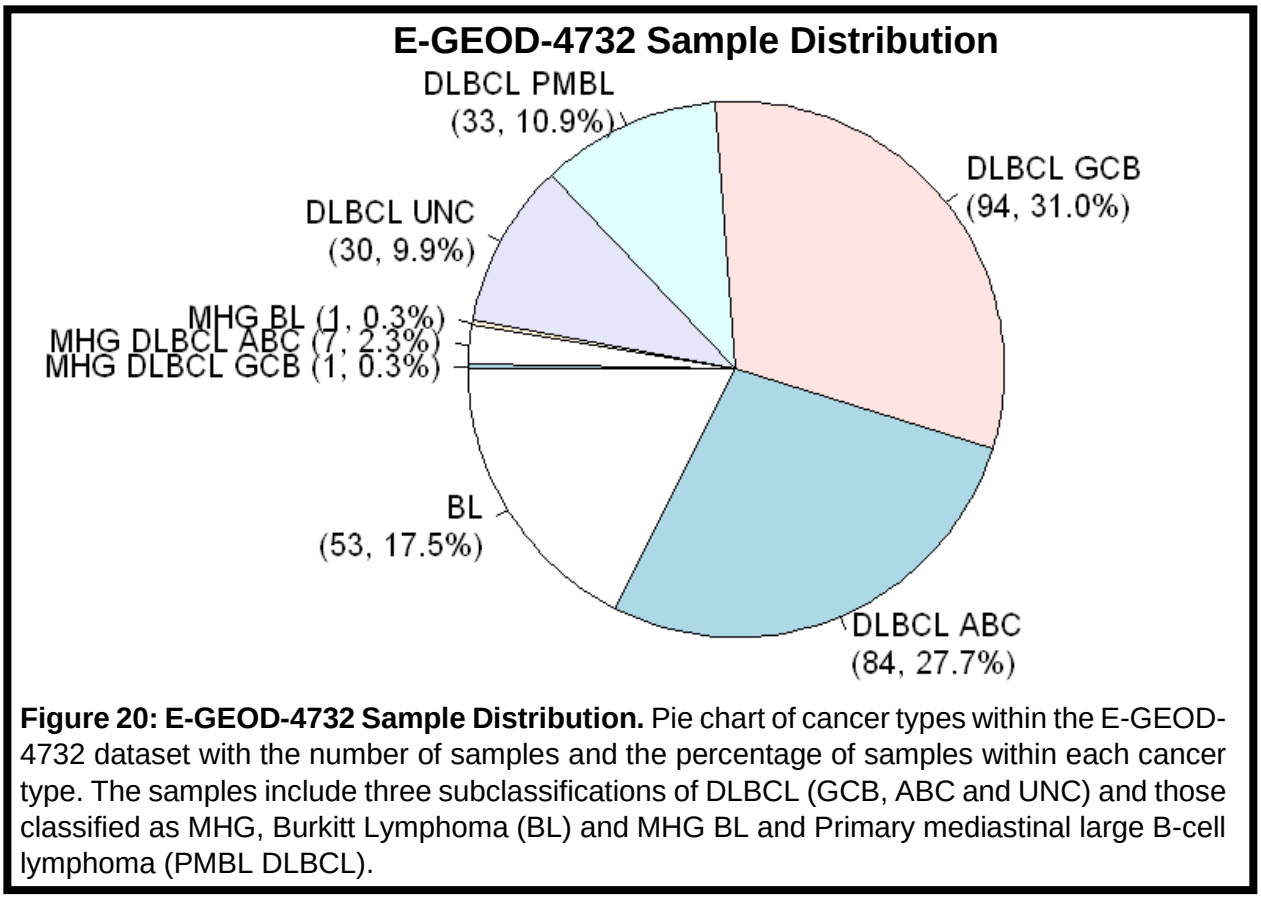


Figure 20: E-GEOD-4732 Sample Distribution. Pie chart of cancer types within the E-GEOD-4732 dataset with the number of samples and the percentage of samples within each cancer type. The samples include three subclassifications of DLBCL (GCB, ABC and UNC) and those classified as MHG, Burkitt Lymphoma (BL) and MHG BL and Primary mediastinal large B-cell lymphoma (PMBL DLBCL).

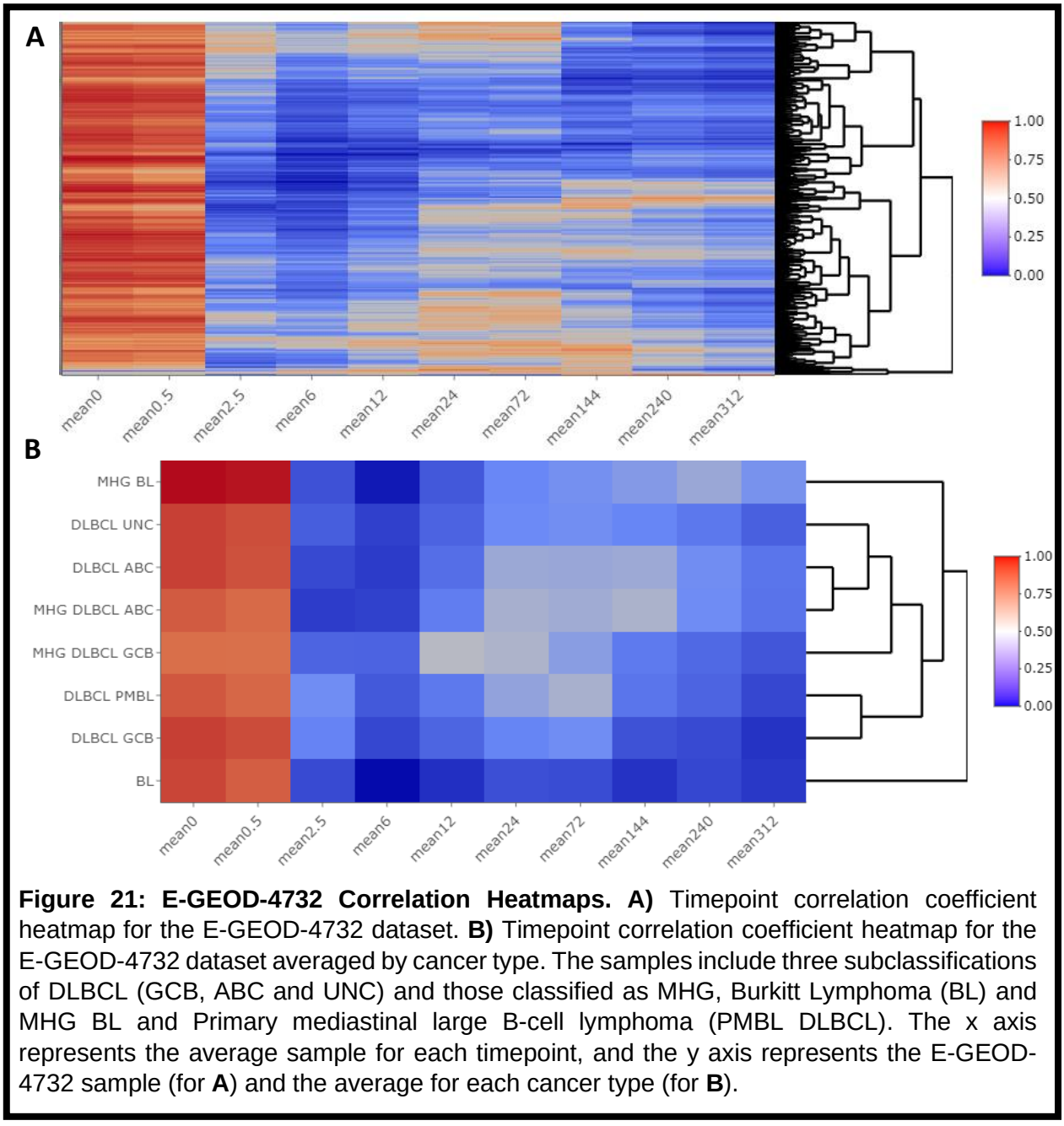


Figure 21: E-GEOD-4732 Correlation Heatmaps. A) Timepoint correlation coefficient heatmap for the E-GEOD-4732 dataset. **B)** Timepoint correlation coefficient heatmap for the E-GEOD-4732 dataset averaged by cancer type. The samples include three subclassifications of DLBCL (GCB, ABC and UNC) and those classified as MHG, Burkitt Lymphoma (BL) and MHG BL and Primary mediastinal large B-cell lymphoma (PMBL DLBCL). The x axis represents the average sample for each timepoint, and the y axis represents the E-GEOD-4732 sample (for **A**) and the average for each cancer type (for **B**).

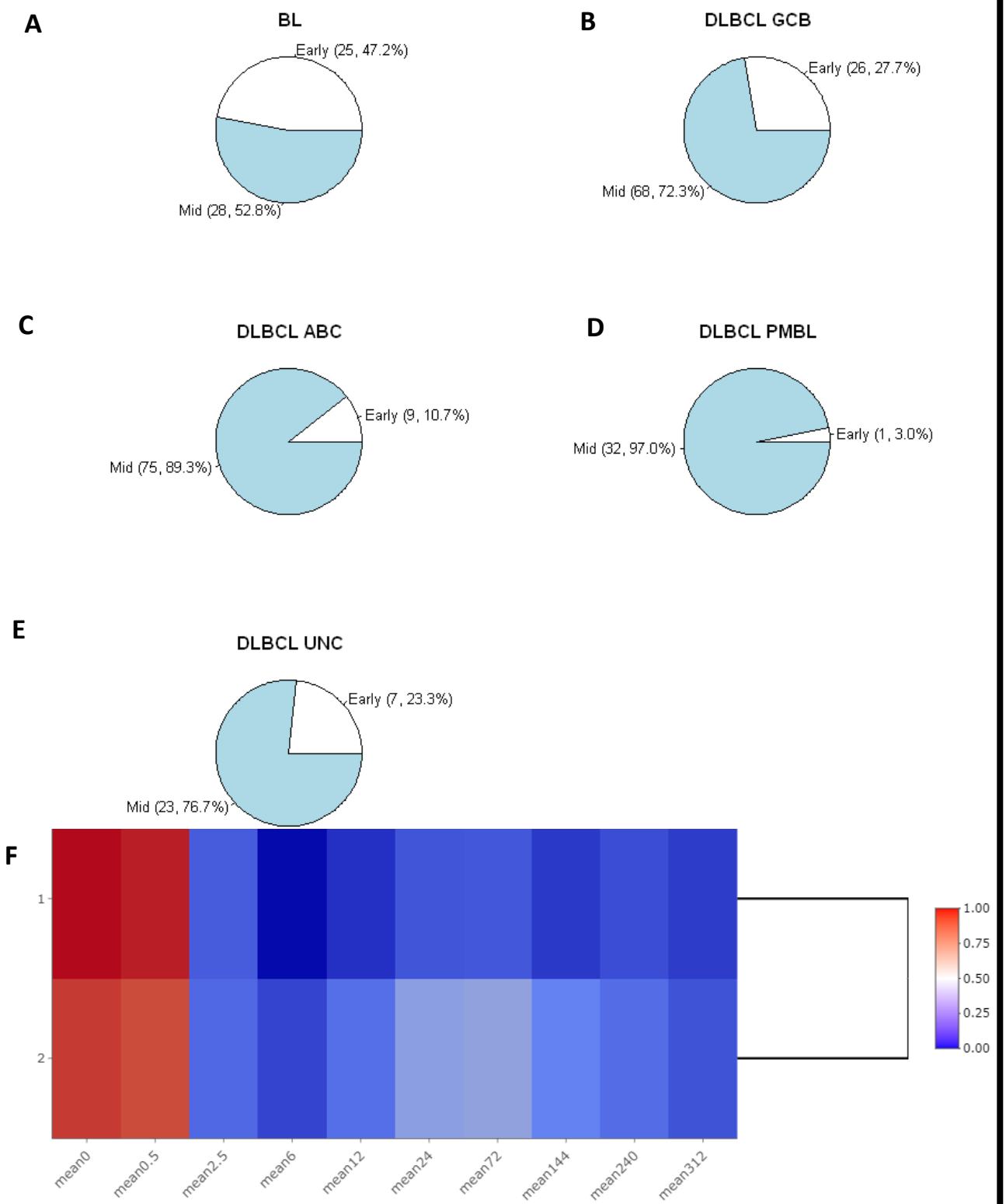


Figure 22: E-GEOD-4732 Classification Results. **A-E** Pie charts of the distribution of timepoints for each cancer type. **A)** Burkitt Lymphoma (BL), **B,** DLBCL GCB, **C,** DLBCL ABC, **D,** Primary mediastinal large B-cell lymphoma (DLBCL PMBL). **E,** DLBCL UNC **F)** Heatmap of timepoint correlation coefficient averaged by classification (1: Early, 2: Mid). The x axis representing the average timepoint samples and the y axis representing the average correlation by classification.

Discussion:

Top 25 Genes:

Overall, gene expression of B-cell maturation is reasonably well characterised (Lee et al., 2020), (Gutiérrez et al., 2007), (Barwick et al., 2016), (Hystad et al., 2007), (Stewart et al., 2021), (Barwick et al., 2018), (Suryani et al., 2010). Specifically relevant to this analysis is Holmes et al.'s single cell analysis of the germinal centre B-cells (2020). Their dataset covers the majority of the cell types identified in the Time series dataset by Coco et al. (2012), primarily the stages from mature B-cell to plasmablast.

25 significant genes covering these stages were identified using time series differential gene expression, followed by several filtering stages (**Table 8**). 25 was selected as a cut of point for use in correlation analysis due to heatmaps with more or less genes forming significantly more incoherent clusters (**Appendix Figure 1**). The top 25 genes acted as the midpoint between statistical confidence in differential expression, enough information for accurate correlation results, and the production of results that made biological sense.

Of the 25 genes selected genes, *SLAMF1*, *PIM2*, *CD37*, and *KCNA3* all represent B-cell biology with significant experimental evidence as already discussed. However, of the 25 selected genes only *BLK* was also identified in the above cited literature, peaking just before the plasmablast stage in Holmes et al.'s (2020) analysis. This was almost the opposite of the time series expression which showed was *BLK* downregulated from the beginning of the time series, the plasmablast stage being at approximately at 144-312 hours in the time series dataset (**Figure 5**).

Although most of the 25 genes are involved with B-cell biology, their lack of appearance in other papers suggests many of these may not be specifically involved in B-cell maturation or that their role in B-cell maturation has not been fully elucidated yet. Also within the top 25 genes were genes with no experimental evidence linking them to B-cells and the experimental evidence that is available linking them to completely different cell types. Namely, *GNPDA1* and *MEI1* which are both expressed in gametes with roles in egg activation and spermatocyte meiosis (Zhang et al., 2003), (Sato et al., 2006). No evidence, however, does not necessarily mean no biological relationship will be found in the future, as well as this, there is the possibility that these two genes are acting akin to reporter genes, like GFP, for some of the maturation stages, being potentially under the control of the same promotor or enhancer, thus when a B-cell maturation gene is expressed so are these genes. Unfortunately, the presence of *Ly9* and *RASSF5* in the top 25 makes these two possibilities unlikely as both these genes are T-cell specific. The more likely explanation for these four genes appearing, and constantly

mentioned genes in the literature not appearing, is that the filtering applied to the time series differential gene expression was not as successful as anticipated (Katagiri et al., 2003), (Chatterjee et al., 2012). This can be made slightly more apparent by the String network analysis, which although found significantly more connections than expected for a normal dataset, the top 25 genes are not a normal dataset as they were curated to characterise B-cell maturation. Therefore, the fact that most of the top 25 genes have no connection, experimental evidence or otherwise, to any other gene strongly indicates filtering has not been as successful as anticipated (**Figure 6**). An obvious caveat to this conclusion, however, is there are still large gaps in our understanding of gene expression and function, meaning that lack of information on function, and even presence of evidence of unrelated functions, still does not exclusively rule out these 'unrelated' genes as having a role in B-cell maturation.

Both differential gene expression analysis stages were carried out using reputable software. The time series differential expression benefited from the whole analysis pipeline being carried out in this study, starting from the raw files, enabling every step to be overseen. One limitation of the time series analysis was the small sample size, only 22 samples were present and most of the time points only had samples from two of the three donors with one timepoint (24 hours) only having one sample. In contrast Holmes et al. (2020) had five donors supplying at least duplicate samples.

Although, it is advantageous to have more samples from more donors, the likely source of the most error will be the differential gene expression analysis of the Cancer Cell Line dataset using the Limma package. Firstly, the data was in the form of an FPKM file, the values in this file have already been through a normalisation procedure which removes integer count information crucial to differential analysis, drastically lowing the accuracy of the analysis (Zhao et al., 2021). The use of formula 1, as suggested by Nazarov et al. (2017), made the dataset usable by Limma, however access to the raw data would be advantageous. Secondly, differential gene expression of the Cancer Cell Line dataset also has some drawbacks aside from datafile format. The hypothesis of "cell line genes", a subset of over expressed proliferative and survival genes, giving the cell lines traits such as replicative immortality and increased proliferation, and their need for removal may be flawed, even though removal of these genes produced more coherent clustering of samples. This is because at several stages of B-cell maturation, within the germinal centre, the cells undergo intense proliferation. Removal of "cell line genes" led to removal of genes like *Myc,* common to many cancers but also identified in Holmes et al.'s (2020) differential analysis and important to the maintenance of germinal centres (Calado et al., 2012). Additionally, due to it being a cancer-based dataset, differential expression analysis may have removed other proto-oncogenes like *Myc*, that are important to B-cell biology, but are also common to many different cancers and therefore not

being significantly differentially expressed just due to their role in B-cell biology. This reduces the effectiveness of using this differential gene expression to identify B-cell specific genes. Although, the B-cell related proto-oncogene *BLK* did make it through the filtering this, however, could have been an exception to the rule in this case.

Correlation Analysis and Major Time Point Identification:

Correlation analysis was performed to first identify major time points (**Figure 7**) before identifying cell lines that most represent the tumour samples. Correlation between gene expression in samples is commonly used, along with PCA to identify outlier samples with a corelation < 0.9 belonging to different cell types (Koch et al., 2018). Correlation has also been used to identify genes associated with disease biomarkers, such as in dilated cardiomyopathy by Witt et al. (2019), though noticeably instead of using correlation coefficient in their analysis, they instead use correlation adjusted p-value. Correlation is also a useful initial stage in biological pathway identification, gene co-expression and relation to disease, or identify genes with potentially similar functions and predict functions of genes, exemplified by R packages such as HGCA and DGCA (Michalopoulos et al., 2012), (Bhuva et al., 2019), (Liesecke et al., 2018), (McKenzie et al., 2016), (Yu et al., 2019). Correlation analysis used to identify clusters and cell types, as well as being used as an alternative to mapping cells to locations manually, is more related to this analysis (Bageritz et al., 2019), (Grün et al., 2015), (Jiang et al., 2018).

Initial correlation analysis (1) was unable to identify particular clusters of cell types but acted similar to Kock et al.'s (2018) description of identifying outliers as it showed no clear outliers with all correlation p-values being significant (**Figure 3**). Although, this approach is more for within dataset outlier identification than comparing datasets. Initial correlation analysis (2), using the time series differential expression results separated multiple myeloma from the rest of the cell lines as it was only expressed from 144-312 hours and all other cell lines were expressed between 12-144 hours (**Figure 4**). Despite the considerable difference in correlation coefficient, the largest adjusted p-value was still highly significant at 3.05×10^{-5}. Finally, correlation using only the top 25 genes established the most significant clusters at three main time points coined as early (0-0.5 hours), mid (2.5-72 hours) and late (144 hours to 312 hours).

All three B acute lymphoblastic leukaemia cell line groupings did not significantly correlate at any time point, as these arise from B-cells before the lymph node stages of maturation (Agraz-Doblas et al., 2019). These samples therefore indicate the time series does not contain B-cells that originate before the lymph node. Chronic lymphocytic leukaemia contrastingly correlates significantly at all time points. This is potentially due to it arising from both pre- and post-germinal centre B-cells, effectively covering the entire length of the time series (Ciccone et al.,

2014), (Ten Hacken et al., 2017). The almost exclusive correlation of primary effusion lymphoma and multiple myeloma cell lines at the late timepoint suggests that this resembles plasmablasts and plasma cells, as these are the cell types both cancers respectively originate from (Chen and Chuang, 2020).

Burkitt lymphoma, DLBCL GCB and Mantel Cell lymphoma's correlation at the early timepoint indicates this could represent the germinal centre stage and stages just before it, due to their respective cells of origin (Schmitz et al., 2014), (Alizadeh et al., 2000). However, DLBCL ABC also correlates at these points and occurs after the germinal centre, although the correlation is less significant. Finally, Hodgkin's lymphoma has its peak correlation at the mid timepoint and also originates from the germinal centre, but the correlation observed was not statistically significant, perhaps due to lack of B-cell specific expression shown by its main cell of origin, the Reed-Sternberg cell (Weniger and Küppers, 2021). This could indicate the early and mid-timepoints represent reasonably similar stages and the time series lacks enough resolution to differentiate between them. Many early timepoint associated cell lines also have smaller correlation peaks in the mid time point range, this could be potentially explained by the cycling between the dark and light zones of the germinal centre undergone by the centroblasts as they alternate between somatic hypermutation, class switch recombination and proliferation and maturation into centrocytes. The time series dataset may simply lack the sample size to effectively differentiate between samples in the centroblast and centrocyte stages given how close the stages are in the maturation process (Pae et al., 2021).

Correlation analysis of the 1,597 multiple myeloma tumour samples also strongly indicated the late timepoint resembles the plasmablast and plasma cells stages by almost exclusive correlation of the multiple myeloma samples at the late timepoint (**Figures 13a** and **18**). Of the 1,452 non-multiple myeloma samples, only a handful also correlated at the late timepoint, these most likely being outliers. These samples were a mix of Burkitt lymphoma and DLBCL, which primarily originate from the germinal centre stages and either directly before or after. Unlike the DLBCL cell lines, the 928 REMoDL-B samples all correlated at the mid timepoint, with the remaining non-multiple myeloma datasets correlating the most at the early timepoint but also having significant correlation at the mid timepoint. This may be because of the aforementioned cycling between different regions of the lymph node the B-cell undergoes as it matures. As a consequence of this analysis, the most reasonable conclusion to be drawn is that as the early and mid-timepoints cannot be reliably distinguished in tumour samples that should originate from the same timepoints, at best this analysis can distinguish between the post-germinal centre plasmablast and plasma cell, cell types and germinal centre cell types.

For some of the datasets, the most similar cell line was also found for each sample through correlation analysis. However, cell lines outside that sample's cancer type also correlated significantly with some of the samples, putting the validity of the recommendation into question. This is likely because the top 25 genes may have been reasonably effective at distinguishing between maturation stages, they may lack the ability to capture the important specifics in gene expression to that specific cancer type. Possibly a more reliable method would be to use differentially expressed genes associated with each cancer type in the correlation, instead of the top 25 genes identified in this method.

Although Grün et al. (2015)'s and Jiang et al. (2018)'s methods involved processing the correlation results using algorithmic clustering, the base idea that significant correlation indicates cell type remains the same. This analysis did also perform algorithmic clustering (K-means) on the data however, not on the correlation results, but verify the observed correlation clusters. Therefore, the observed major timepoints in the aforementioned heatmaps would likely remain the same if passed through clustering, particularly considering the R package, Heatmaply, used to create the heatmaps clustered the results on the heatmaps automatically. However, in identifying similar cell lines to tumour samples the approaches used by Grün et al. (2015) and Jiang et al. (2018) may be better.

Classification:

Since the advent of sequencing classification of tumours using gene expression is extremely common and effective in the field of precision medicine. Of particular relevance to this analysis would be the DAC classifier, differentiating between the GCB and ABC DLBCL subtypes (Care et al., 2013). All forms of machine learning classifier have been used to classify gene expression data, including decision trees, random forests, neural networks, K-nearest neighbour, and support vectors (H. Zhang et al., 2003), (Berrar et al., 2003), (Lyu and Haque, 2018), (Hijazi and Chan, 2013), (Mohammed et al., 2021). The next iterative step in classification came with ensemble learning vastly increasing accuracy and comes in three main forms; boosting, bagging and stacking (Xiong et al., 2021). Stacking benefits from having two layers of classification, the first uses base learners, usually a variety of different types of classifiers, to learn the data, before a second layer metalearner is employed to learn the outputs of the first layer. This can lead to increased performance over the base models alone. This was used to great effect, vastly increasing the accuracy of breast cancer classification by Kwon et al. (2019).

The top 25 genes and a combined dataset made of the B-cell Lymphoma cell lines dataset and the timepoint dataset. This was used to train a stacking classifier consisting of six models, logistic regression, K-nearest neighbour, random forest, support vector, naïve Bayes, and

MLP. These were linked together by the second layer logistic regression metalearner. 90 samples using the correlation results were clustered into the three main timepoints, early, mid and late using K-means clustering (**Figure 9A**). Cross-validation was then used to test the accuracy of the stacking classifier, as well as its constituent classifiers individually (**Table 9**). Surprisingly, the most accurate model was the logistic regression classifier on its own, instead of the expected stacking classifier, although the difference was negligible. In fact, nearly all individual classifiers performed exceptionally well in cross-validation across all metrics. As the difference in performance between the stacking classifier and the logistic regression classifier was near negligible, and the stacking classifier minimalised the false positive rate more, the stacking classifier was used to classify the tumour samples.

The REMoDL-B dataset samples all correlated at the mid timepoint and were all classified as such (**Figure 11**). The other DLBCL and Burkitt lymphoma datasets correlated at the early timepoint (**Figures 16** and **21**). However, when classified had the majority of samples were classified as the mid time point regardless of cancer type and their suggested cell of origin (**Figures 17A-F** and **22A-E**). This might suggest, along with the correlation analysis, the classifier is unable to distinguish between the early and mid-timepoints and the cell types involved reliably. The correlation analysis for each dataset shows that despite their classification, all samples still corelate the most with the early timepoint and with E-GEOD-4475 dataset showing the most correlation with the mid timepoint (**Figures 17G and 21F**). The hypothesis of being able to distinguish between the combined early and mid-timepoints and the late timepoint also seems optimistic, around 30% of the multiple myeloma samples were classified in the mid timepoint despite exclusively correlating with the late timepoints (**Figures 13 and 19**). This led to the conclusion that the extremely high observed accuracy of 0.989 is an artifact of overfitting of the model to the training dataset.

The classifier only had access to a training dataset of 90 samples split across two original datasets, a very small dataset compared to that used to train DAC (2030 samples) (Care et al., 2013). These datasets are also very different in context, one being the Cancer Cell Line dataset and in the format of FPKM with previously mentioned weaknesses and the other, the Timepoint dataset being the main component of the time series analysis. Ideally the training set would have also included tumour samples with known B-cell maturation stages of origin to make it applicable to its use case. In the dataset, the mid timepoint was also vastly underrepresented taking up only 20% of the dataset compared to 37% (late timepoint) and 43% (early timepoint). New samples could be imputed but due to the sheer lack of samples and imprecise definition of the timepoints, the timepoints could potentially to blur together (Bageritz et al., 2019).

Overfitting of a stacking classifier can be a distinct problem due to the presence of multiple models predicting the same results, with the possibility of overfitting rising with the more models that are present (Wolfinger and Tan, 2017). Two ways of avoiding overfitting are 'greedy' selection of criteria and a validation set. 'Greedy' selection, a process of choosing the most effective inputs, removing those that reduce accuracy, was carried out for DAC and could have also been carried out in order to reduce the effect of overfitting by reducing model complexity (Caruana and Freitag, 1994), (Care et al., 2013). This could have been employed in two layers, the first to select the best classifiers to use in the stacking classifier by selecting only the models that increase the overall accuracy of the ensemble model and disregarding those that do not. And the second layer, 'greedy' selection of the genes used in the classifier, as exemplified by **Figure 9B**, which shows there is huge variation in level of importance attributed to each gene, specifically in the random forest model, which could extend to all the other models used. The DAC classifier also introduced the use of an 'unclassifiable' grouping which may have proved useful in avoiding the miss-classification of samples into the timepoint that had the highest probability when that probability did not meet a high enough probability threshold. The final omission was that of a validation dataset which could give a vastly more accurate accuracy rating on the classifier and act to confirm or deny overfitting of the training data. Although, this would only provide a single test of model performance and no indication on average model performance (Wolfinger and Tan, 2017). The primary reason for not employing it in this case was due to it reducing the size of the already small training dataset.

Future Work:

Identification of the B-cell maturation genes using correlation algorithms, such as highest reciprocal rank Pearson correlation coefficient, in combination with the time series differential gene expression may prove fruitful (Liesecke et al., 2018). One specific advantage of this process is the ability to use TPM datasets which are similar to FPKM datasets and was the only other datafile the cancer cell line dataset was available in. Although the ideal analysis would be to obtain the raw count values for the dataset and start the analysis from scratch. Lack of samples was also a common theme and underpowered the analysis, so the ability to obtain more time series samples from multiple donors will prove invaluable for future research. This could increase the resolution to which the analysis could be performed and increase the number of samples available for the classification process to use. Identification and suggestion of the most representative cell line for each tumour sample could also be significantly aided by algorithms like Bageritz et al. (2019)'s which enabled cell mapping and type identification. Finally, improvements to the training of the classifier in the form of 'greedy' selection of models and genes used, and the use of a validation set to vastly reduce the effect of overfitting of the data as well as more samples to use in this process would be beneficial.

Conclusions:

25 key genes have been identified, through time series differential gene expression, that represent the maturation of B-cells to plasma cells in the lymph node over a period of 312 hours. Through correlation and clustering analysis three major timepoints across the 312 hours were identified: early (0-0.5 hours), mid (2.5-72 hours) and late (144 hours to 312 hours). Through analysis of the correlation of 65 cell lines from 8 cancer types and 3,049 samples covering multiple subtypes of DLBCL, multiple myeloma and Burkitt lymphomas the early and mid-time points were attributed to pre-germinal and germinal centre cells and the late time points attributed to plasmablasts and plasma cells. For some of the 3,049 samples, cell lines closely resembling the time point were also recommended. Finally, an ensemble classifier was created and trained on the time points, with an average accuracy of 0.989 over 1,000 rounds of cross-validation. However, due to its inconsistent classification of the 3,049 samples, and its juxtaposition with regards to the correlation analysis, it became evident that the classifier suffered either from overfitting or it is simply most effective at differentiating between cell lines not tumour samples. The classifier may also lack the granularity to differentiate between ABC and GCB cell types like the DAC classifier can.

Considering this, several objectives for future research have been set out. Primarily, the obtainment of more samples from more donors for the time series analysis, as well as the raw expression data for the cell line analysis. Alternative methods for the identification of B-cell maturation and identification of similar cell lines have also been recommended. Finally, several additional stages to the training of the machine learning classifier have also been suggested in order to reduce the effect of overfitting observed.

References:

Agraz-Doblas, A., Bueno, C., Bashford-Rogers, R., Roy, A., Schneider, P., Bardini, M., Ballerini, P., Cazzaniga, G., Moreno, T., Revilla, C., Gut, M., Valsecchi, M.G., Roberts, I., Pieters, R., De Lorenzo, P., Varela, I., Menendez, P. and Stam, R.W. 2019. Unraveling the cellular origin and clinical prognostic markers of infant B-cell acute lymphoblastic leukemia using genome-wide analysis. *Haematologica*. **104**(6), pp.1176–1188.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. 2002. *Lymphocytes and the cellular basis of adaptive immunity*. London, England: Garland Science.

Alexa, A. and Rahnenfuhrer, J. 2021. topGO. *Bioconductor.org*. [Online]. [Accessed 9 August 2021]. Available from: https://bioconductor.org/packages/release/bioc/html/topGO.html.

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J., Jr, Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O. and Staudt, L.M. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. **403**(6769), pp.503–511.

Allen, C.D.C., Okada, T. and Cyster, J.G. 2007. Germinal-center organization and cellular dynamics. *Immunity*. **27**(2), pp.190–202.

Andrews, S. 2010. Babraham bioinformatics - FastQC A quality control tool for high throughput sequence data. *Babraham.ac.uk*. [Online]. [Accessed 9 August 2021]. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Ayala, G.E., Dai, H., Ittmann, M., Li, R., Powell, M., Frolov, A., Wheeler, T.M., Thompson, T.C. and Rowley, D. 2004. Growth and survival mechanisms associated with perineural invasion in prostate cancer. *Cancer research*. **64**(17), pp.6082–6090.

Bageritz, J., Willnow, P., Valentini, E., Leible, S., Boutros, M. and Teleman, A.A. 2019. Gene expression atlas of a developing tissue by single cell expression correlation analysis. *Nature methods*. **16**(8), pp.750–756.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M.F., Monahan, J.E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., Mapa, F.A., Thibault, J., Bric-Furlong, E., Raman, P., Shipway, A., Engels, I.H., Cheng, J., Yu, G.K.,

Yu, J., Aspesi, P., Jr, de Silva, M., Jagtap, K., Jones, M.D., Wang, L., Hatton, C., Palescandolo, E., Gupta, S., Mahan, S., Sougnez, C., Onofrio, R.C., Liefeld, T., MacConaill, L., Winckler, W., Reich, M., Li, N., Mesirov, J.P., Gabriel, S.B., Getz, G., Ardlie, K., Chan, V., Myer, V.E., Weber, B.L., Porter, J., Warmuth, M., Finan, P., Harris, J.L., Meyerson, M., Golub, T.R., Morrissey, M.P., Sellers, W.R., Schlegel, R. and Garraway, L.A. 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. **483**(7391), pp.603–607.

Barwick, B.G., Scharer, C.D., Bally, A.P.R. and Boss, J.M. 2016. Plasma cell differentiation is coupled to division-dependent DNA hypomethylation and gene regulation. *Nature immunology*. **17**(10), pp.1216–1225.

Barwick, B.G., Scharer, C.D., Martinez, R.J., Price, M.J., Wein, A.N., Haines, R.R., Bally, A.P.R., Kohlmeier, J.E. and Boss, J.M. 2018. B cell activation and plasma cell differentiation are inhibited by de novo DNA methylation. *Nature communications*. **9**(1), p.1900.

Baumgarth, N. 2000. A two-phase model of B-cell activation. *Immunological reviews*. **176**(1), pp.171–180.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*. **57**(1), pp.289–300.

Berrar, D.P., Downes, C.S. and Dubitzky, W. 2003. Multiclass cancer classification using gene expression profiling and probabilistic neural networks. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*., pp.5–16.

Bertoni, F. and Ponzoni, M. 2007. The cellular origin of mantle cell lymphoma. *The international journal of biochemistry & cell biology*. **39**(10), pp.1747–1753.

Bhuva, D.D., Cursons, J., Smyth, G.K. and Davis, M.J. 2019. Differential co-expression-based detection of conditional relationships in transcriptional data: comparative analysis and application to breast cancer. *Genome biology*. **20**(1), p.236.

Bonvini, P., Zorzi, E., Basso, G. and Rosolen, A. 2007. Bortezomib-mediated 26S proteasome inhibition causes cell-cycle arrest and induces apoptosis in CD-30+ anaplastic large cell lymphoma. *Leukemia*. **21**(4), pp.838–842.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B. and Varoquaux, G. 2013. API design for machine learning software: experiences from

the scikit-learn project. *arXiv [cs.LG]*. [Online]. Available from: http://arxiv.org/abs/1309.0238.

Calado, D.P., Sasaki, Y., Godinho, S.A., Pellerin, A., Köchert, K., Sleckman, B.P., de Alborán, I.M., Janz, M., Rodig, S. and Rajewsky, K. 2012. The cell-cycle regulator c-Myc is essential for the formation and maintenance of germinal centers. *Nature immunology*. **13**(11), pp.1092–1100.

Cancer Research UK 2020a. Cancer incidence for common cancers. *Cancerresearchuk.org*. [Online]. [Accessed 15 August 2021]. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence/common-cancers-compared.

Cancer Research UK 2020b. Myeloma statistics. *Cancerresearchuk.org*. [Online]. [Accessed 16 August 2021]. Available from: http://cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/myeloma.

Cancer Research UK 2020c. Non-Hodgkin lymphoma statistics. *Cancerresearchuk.org*. [Online]. [Accessed 16 August 2021]. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/non-hodgkin-lymphoma.

Cancer Research UK 2020d. About non-Hodgkin Lymphoma. *Cancerresearchuk.org*. [Online]. [Accessed 6 August 2021]. Available from: https://www.cancerresearchuk.org/about-cancer/non-hodgkin-lymphoma/about.

Cancer.net 2021. Lymphoma - non-Hodgkin - statistics. *Cancer.net*. [Online]. [Accessed 15 August 2021]. Available from: https://www.cancer.net/cancer-types/lymphoma-non-hodgkin/statistics.

Care, M.A., Barrans, S., Worrillow, L., Jack, A., Westhead, D.R. and Tooze, R.M. 2013. A microarray platform-independent classification tool for cell of origin class allows comparative analysis of gene expression in diffuse large B-cell lymphoma. *PloS one*. **8**(2), p.e55895.

Caruana, R. and Freitag, D. 1994. Greedy Attribute Selection *In*: W. W. Cohen and H. Hirsh, eds. *Machine Learning Proceedings 1994*. Oxford, England: Elsevier, pp.28–36.

Chatterjee, M., Hedrich, C.M., Rauen, T., Ioannidis, C., Terhorst, C. and Tsokos, G.C. 2012. CD3-T cell receptor co-stimulation through SLAMF3 and SLAMF6 receptors enhances RORγt recruitment to the IL17A promoter in human T lymphocytes. *The journal of biological chemistry*. **287**(45), pp.38168–38177.

Chaudhuri, J. and Alt, F.W. 2004. Class-switch recombination: interplay of transcription, DNA deamination and DNA repair. *Nature reviews. Immunology*. **4**(7), pp.541–552.

Cheadle, C., Vawter, M.P., Freed, W.J. and Becker, K.G. 2003. Analysis of microarray data using Z score transformation. *The Journal of molecular diagnostics: JMD*. **5**(2), pp.73–81.

Chen, B.-J. and Chuang, S.-S. 2020. Lymphoid neoplasms with plasmablastic differentiation: A comprehensive review and diagnostic approaches: A comprehensive review and diagnostic approaches. *Advances in anatomic pathology*. **27**(2), pp.61–74.

Chen, C., Zhai, S., Zhang, L., Chen, J., Long, X., Qin, J., Li, J., Huo, R. and Wang, X. 2018. Uhrf1 regulates germinal center B cell expansion and affinity maturation to control viral infection. *The journal of experimental medicine*. **215**(5), pp.1437–1448.

Chung, J.B., Silverman, M. and Monroe, J.G. 2003. Transitional B cells: step by step towards immune competence. *Trends in immunology*. **24**(6), pp.343–349.

Chung, N., Zhang, X.D., Kreamer, A., Locco, L., Kuan, P.-F., Bartz, S., Linsley, P.S., Ferrer, M. and Strulovici, B. 2008. Median absolute deviation to improve hit selection for genome-scale RNAi screens. *Journal of biomolecular screening*. **13**(2), pp.149–158.

Ciccone, M., Ferrajoli, A., Keating, M.J. and Calin, G.A. 2014. SnapShot: chronic lymphocytic leukemia. *Cancer cell*. **26**(5), pp.770-770.e1.

Cocco, M., Stephenson, S., Care, M.A., Newton, D., Barnes, N.A., Davison, A., Rawstron, A., Westhead, D.R., Doody, G.M. and Tooze, R.M. 2012. In vitro generation of long-lived human plasma cells. *The journal of immunology*. **189**(12), pp.5773–5785.

Cross, M. and Dearden, C. 2019. B and T cell prolymphocytic leukaemia. *Best practice & research. Clinical haematology*. **32**(3), pp.217–228.

Davies, A., Cummin, T.E., Barrans, S., Maishman, T., Mamot, C., Novak, U., Caddy, J., Stanton, L., Kazmi-Stokes, S., McMillan, A., Fields, P., Pocock, C., Collins, G.P., Stephens, R., Cucco, F., Clipson, A., Sha, C., Tooze, R., Care, M.A., Griffiths, G., Du, M.-Q., Westhead, D.R., Burton, C. and Johnson, P.W.M. 2019. Gene-expression profiling of bortezomib added to standard chemoimmunotherapy for diffuse large B-cell lymphoma (REMoDL-B): an open-label, randomised, phase 3 trial. *The lancet oncology*. **20**(5), pp.649–662.

Davies, A.J., Barrans, S., Maishman, T., Cummin, T.E., Bentley, M., Mamot, C., Novak, U., Caddy, J., Hamid, D., Kazmi-Stokes, S.H., Mcmillan, A., Fields, P.A., Pocock, C.,

Kruger, A., Collins, G., Sha, C., Clipson, A., Wang, M., Tooze, R.M., Care, M.A., Griffiths, G.O., Du, M., Westhead, D.R., Burton, C., Jack, A. and Johnson, P.W. 2017. DIFFERENTIAL EFFICACY OF BORTEZOMIB IN SUBTYPES OF DIFFUSE LARGE B-CELL LYMPHOMA (DLBL): A PROSPECTIVE RANDOMISED STUDY STRATIFIED BY TRANSCRIPTOME PROFILING: REMODL-B. *Hematological oncology*. **35**(S2), pp.130–131.

Davis, R.E., Brown, K.D., Siebenlist, U. and Staudt, L.M. 2001. Constitutive nuclear factor κB activity is required for survival of activated B cell–like diffuse large B cell lymphoma cells. *The journal of experimental medicine*. **194**(12), pp.1861–1874.

Drexler, H.G. and Quentmeier, H. 2020. The LL-100 cell lines panel: Tool for molecular leukemia-lymphoma research. *International journal of molecular sciences*. **21**(16), p.5800.

Drexler, H.G., Matsuo, A.Y. and MacLeod, R.A. 2000. Continuous hematopoietic cell lines as model systems for leukemia-lymphoma research. *Leukemia research*. **24**(11), pp.881–911.

Duns, G., Viganò, E., Ennishi, D., Sarkozy, C., Hung, S.S., Chavez, E., Takata, K., Rushton, C., Jiang, A., Ben-Neriah, S., Woolcock, B.W., Slack, G.W., Hsi, E.D., Craig, J.W., Hilton, L.K., Shah, S.P., Farinha, P., Mottok, A., Gascoyne, R.D., Morin, R.D., Savage, K.J., Scott, D.W. and Steidl, C. 2021. Characterization of DLBCL with a PMBL gene expression signature. *Blood*. **138**(2), pp.136–148.

Durinck, S., Spellman, P.T., Birney, E. and Huber, W. 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature protocols*. **4**(8), pp.1184–1191.

Fox, C.J., Hammerman, P.S., Cinalli, R.M., Master, S.R., Chodosh, L.A. and Thompson, C.B. 2003. The serine/threonine kinase Pim-2 is a transcriptionally regulated apoptotic inhibitor. *Genes & development*. **17**(15), pp.1841–1854.

Galili, T., O'Callaghan, A., Sidi, J. and Sievert, C. 2018. heatmaply: an R package for creating interactive cluster heatmaps for online publishing. *Bioinformatics (Oxford, England)*. **34**(9), pp.1600–1602.

García De Vinuesa, C., Gulbranson-Judge, A., Khan, M., O'Leary, P., Cascalho, M., Wabl, M., Klaus, G.G., Owen, M.J. and MacLennan, I.C. 1999. Dendritic cells associated with plasmablast survival. *European journal of immunology*. **29**(11), pp.3712–3721.

Geraghty, R.J., Capes-Davis, A., Davis, J.M., Downward, J., Freshney, R.I., Knezevic, I., Lovell-Badge, R., Masters, J.R.W., Meredith, J., Stacey, G.N., Thraves, P., Vias, M. and Cancer Research UK 2014. Guidelines for the use of cell lines in biomedical research. *British journal of cancer*. **111**(6), pp.1021–1046.

Gonzalez-Nicolini, V. and Fussenegger, M. 2005. In vitro assays for anticancer drug discovery--a novel approach based on engineered mammalian cell lines. *Anti-cancer drugs*. **16**(3), pp.223–228.

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H. and van Oudenaarden, A. 2015. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*. **525**(7568), pp.251–255.

Gutiérrez, N.C., Ocio, E.M., de Las Rivas, J., Maiso, P., Delgado, M., Fermiñán, E., Arcos, M.J., Sánchez, M.L., Hernández, J.M. and San Miguel, J.F. 2007. Gene expression profiling of B lymphocytes and plasma cells from Waldenström's macroglobulinemia: comparison with expression patterns of the same cell counterparts from chronic lymphocytic leukemia, multiple myeloma and normal individuals. *Leukemia*. **21**(3), pp.541–549.

Hanahan, D. and Weinberg, R.A. 2011. Hallmarks of cancer: the next generation. *Cell*. **144**(5), pp.646–674.

Harwood, N.E. and Batista, F.D. 2010. Early events in B cell activation. *Annual review of immunology*. **28**(1), pp.185–210.

Hijazi, H. and Chan, C. 2013. A classification framework applied to cancer gene expression profiles. *Journal of healthcare engineering*. **4**(2), pp.255–283.

Hoffman, W., Lakkis, F.G. and Chalasani, G. 2016. B cells, antibodies, and more. *Clinical journal of the American Society of Nephrology: CJASN*. **11**(1), pp.137–154.

Holmes, A.B., Corinaldesi, C., Shen, Q., Kumar, R., Compagno, N., Wang, Z., Nitzan, M., Grunstein, E., Pasqualucci, L., Dalla-Favera, R. and Basso, K. 2020. Single-cell analysis of germinal-center B cells informs on lymphoma cell of origin and outcome. *The journal of experimental medicine*. **217**(10).

Howie, D., Simarro, M., Sayos, J., Guirado, M., Sancho, J. and Terhorst, C. 2002. Molecular dissection of the signaling and costimulatory functions of CD150 (SLAM): CD150/SAP binding and CD150-mediated costimulation. *Blood*. **99**(3), pp.957–965.

Hu, X., Yang, D., Zimmerman, M., Liu, F., Yang, J., Kannan, S., Burchert, A., Szulc, Z., Bielawska, A., Ozato, K., Bhalla, K. and Liu, K. 2011. IRF8 regulates acid ceramidase expression to mediate apoptosis and suppresses myelogeneous leukemia. *Cancer research*. **71**(8), pp.2882–2891.

Hystad, M.E., Myklebust, J.H., Bø, T.H., Sivertsen, E.A., Rian, E., Forfang, L., Munthe, E., Rosenwald, A., Chiorazzi, M., Jonassen, I., Staudt, L.M. and Smeland, E.B. 2007. Characterization of early stages of human B cell development by gene expression profiling. *The journal of immunology*. **179**(6), pp.3662–3671.

Illumina 2021. FPKM Files. *Illumina.com*. [Online]. [Accessed 9 August 2021]. Available from: https://support.illumina.com/help/BS_App_RNASeq_Alignment_OLH_1000000006112/Content/Source/Informatics/Apps/swSEQ_mAPP_RNAseq_FPKM.htm.

Janeway, C.A., Jr, Travers, P., Walport, M. and Shlomchik, M.J. 2001. *Antigen recognition by B-cell and T-cell receptors*. London, England: Garland Science.

Jiang, H., Sohn, L.L., Huang, H. and Chen, L. 2018. Single cell clustering based on cell-pair differentiability correlation and variance analysis. *Bioinformatics (Oxford, England)*. **34**(21), pp.3684–3694.

Jung, D., Giallourakis, C., Mostoslavsky, R. and Alt, F.W. 2006. Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. *Annual review of immunology*. **24**(1), pp.541–570.

Katagiri, K., Maeda, A., Shimonaka, M. and Kinashi, T. 2003. RAPL, a Rap1-binding molecule that mediates Rap1-induced adhesion through spatial regulation of LFA-1. *Nature immunology*. **4**(8), pp.741–748.

Klein, U. and Heise, N. 2015. Unexpected functions of nuclear factor-κB during germinal center B-cell development: Implications for lymphomagenesis. *Current opinion in hematology*. **22**(4), pp.379–387.

Koch, C.M., Chiu, S.F., Akbarpour, M., Bharat, A., Ridge, K.M., Bartom, E.T. and Winter, D.R. 2018. A beginner's guide to analysis of RNA sequencing data. *American journal of respiratory cell and molecular biology*. **59**(2), pp.145–157.

Kondo, M. 2010. Lymphoid and myeloid lineage commitment in multipotent hematopoietic progenitors: Roles of bone marrow microenvironment. *Immunological reviews*. **238**(1), pp.37–46.

Krueger, F. 2021. Babraham Bioinformatics - Trim Galore! *Babraham.ac.uk*. [Online]. [Accessed 9 August 2021]. Available from: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

Kwon, H., Park, J. and Lee, Y. 2019. Stacking ensemble technique for classifying breast cancer. *Healthcare informatics research*. **25**(4), pp.283–288.

Kyle, R.A. and Rajkumar, S.V. 2004. Multiple myeloma. *The New England journal of medicine*. **351**(18), pp.1860–1873.

Lee, R.D., Munro, S.A., Knutson, T.P., LaRue, R.S., Heltemes-Harris, L.M. and Farrar, M.A. 2020. Single-cell analysis of developing B cells reveals dynamic gene expression networks that govern B cell development and transformation. *bioRxiv*. [Online], 2020.06.30.178301. [Accessed 10 August 2021]. Available from: https://www.biorxiv.org/content/10.1101/2020.06.30.178301v1.full.

Lemon, J. 2006. Plotrix: A Package in the Red Light District of R. *R-News*. **6**(4), pp.8–12.

Lenz, G. and Staudt, L.M. 2010. Aggressive lymphomas. *The New England journal of medicine*. **362**(15), pp.1417–1429.

Liesecke, F., Daudu, D., Dugé de Bernonville, R., Besseau, S., Clastre, M., Courdavault, V., de Craene, J.-O., Crèche, J., Giglioli-Guivarc'h, N., Glévarec, G., Pichon, O. and Dugé de Bernonville, T. 2018. Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks. *Scientific reports*. **8**(1), p.10885.

Love, M.I., Huber, W. and Anders, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*. **15**(12), p.550.

Lovén, J., Orlando, D.A., Sigova, A.A., Lin, C.Y., Rahl, P.B., Burge, C.B., Levens, D.L., Lee, T.I. and Young, R.A. 2012. Revisiting global gene expression analysis. *Cell*. **151**(3), pp.476–482.

Lyu, B. and Haque, A. 2018. Deep learning based tumor type classification using gene expression data. *bioRxiv*. [Online], p.364323. [Accessed 11 August 2021]. Available from: https://www.biorxiv.org/content/10.1101/364323v1.

McHeyzer-Williams, L.J., Driver, D.J. and McHeyzer-Williams, M.G. 2001. Germinal center reaction. *Current opinion in hematology*. **8**(1), pp.52–59.

McKenzie, A.T., Katsyv, I., Song, W.-M., Wang, M. and Zhang, B. 2016. DGCA: A comprehensive R package for Differential Gene Correlation Analysis. *BMC systems biology*. **10**(1), p.106.

Michalopoulos, I., Pavlopoulos, G.A., Malatras, A., Karelas, A., Kostadima, M.-A., Schneider, R. and Kossida, S. 2012. Human gene correlation analysis (HGCA): a tool for the identification of transcriptionally co-expressed genes. *BMC research notes*. **5**(1), p.265.

Mirabelli, P., Coppola, L. and Salvatore, M. 2019. Cancer cell lines are useful model systems for medical research. *Cancers*. **11**(8), p.1098.

Mohammed, M., Mwambi, H., Mboya, I.B., Elbashir, M.K. and Omolo, B. 2021. A stacking ensemble deep learning approach to cancer type classification based on TCGA data. *Scientific reports*. **11**(1), p.15626.

Mohammed, M., Mwambi, H., Omolo, B. and Elbashir, M.K. 2018. Using stacking ensemble for microarray-based cancer classification *In*: *2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*. IEEE, pp.1–8.

Molyneux, E.M., Rochford, R., Griffin, B., Newton, R., Jackson, G., Menon, G., Harrison, C.J., Israels, T. and Bailey, S. 2012. Burkitt lymphoma. *Lancet*. **379**(9822), pp.1234–1244.

Momose, S., Weißbach, S., Pischimarov, J., Nedeva, T., Bach, E., Rudelius, M., Geissinger, E., Staiger, A.M., Ott, G. and Rosenwald, A. 2015. The diagnostic gray zone between Burkitt lymphoma and diffuse large B-cell lymphoma is also a gray zone of the mutational spectrum. *Leukemia*. **29**(8), pp.1789–1791.

Nazarov, P.V., Muller, A., Kaoma, T., Nicot, N., Maximo, C., Birembaut, P., Tran, N.L., Dittmar, G. and Vallar, L. 2017. RNA sequencing and transcriptome arrays analyses show opposing results for alternative splicing in patient derived samples. *BMC genomics*. **18**(1).

Ng, A. and Chiorazzi, N. 2021. Potential relevance of B-cell maturation pathways in defining the cell(s) of origin for chronic Lymphocytic leukemia. *Hematology/oncology clinics of North America*. **35**(4), pp.665–685.

Nutt, S.L., Hodgkin, P.D., Tarlinton, D.M. and Corcoran, L.M. 2015. The generation of antibody-secreting plasma cells. *Nature reviews. Immunology*. **15**(3), pp.160–171.

Obukhanych, T.V. and Nussenzweig, M.C. 2006. T-independent type II immune responses generate memory B cells. *The journal of experimental medicine*. **203**(2), pp.305–310.

Ohya, K., Kajigaya, S., Kitanaka, A., Yoshida, K., Miyazato, A., Yamashita, Y., Yamanaka, T., Ikeda, U., Shimada, K., Ozawa, K. and Mano, H. 1999. Molecular cloning of a docking protein, BRDG1, that acts downstream of the Tec tyrosine kinase. *Proceedings of the National Academy of Sciences of the United States of America*. **96**(21), pp.11976–11981.

Pae, J., Ersching, J., Castro, T.B.R., Schips, M., Mesin, L., Allon, S.J., Ordovas-Montanes, J., Mlynarczyk, C., Melnick, A., Efeyan, A., Shalek, A.K., Meyer-Hermann, M. and Victora, G.D. 2021. Cyclin D3 drives inertial cell cycling in dark zone germinal center B cells. *The journal of experimental medicine*. **218**(4).

Pannone, G., Zamparese, R., Pace, M., Pedicillo, M.C., Cagiano, S., Somma, P., Errico, M.E., Donofrio, V., Franco, R., De Chiara, A., Aquino, G., Bucci, P., Bucci, E., Santoro, A. and Bufo, P. 2014. The role of EBV in the pathogenesis of Burkitt Lymphoma: an Italian hospital based survey. *Infectious agents and cancer*. **9**(1), p.34.

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U. and Brazma, A. 2007. ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic acids research*. **35**(Database issue), pp.D747-50.

Pasqualucci, L., Neumeister, P., Goossens, T., Nanjangud, G., Chaganti, R.S., Küppers, R. and Dalla-Favera, R. 2001. Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas. *Nature*. **412**(6844), pp.341–346.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. and Kingsford, C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*. **14**(4), pp.417–419.

Paul, B. and Jonathan, F. 2019. Burkitt lymphoma. *Cancertherapyadvisor.com*. [Online]. [Accessed 16 August 2021]. Available from: https://www.cancertherapyadvisor.com/home/decision-support-in-medicine/hematology/burkitt-lymphoma/.

Pemmaraju, N., Gill, J., Gupta, S. and Krause, J.R. 2014. Triple-hit lymphoma. *Proceedings (Baylor University. Medical Center)*. **27**(2), pp.125–127.

Printz, C. 2016. Study: Stem cell transplant should remain preferred therapy for multiple myeloma. *Cancer*. **122**(19), pp.2937–2937.

Project Jupyter 2021. Project Jupyter. *Jupyter.org*. [Online]. [Accessed 9 August 2021]. Available from: https://jupyter.org/about.

R Core Team 2019. R: A language and environment for statistical computing. *r-project.org*. [Online]. [Accessed 9 August 2021]. Available from: https://www.r-project.org/.

Rajewsky, K. 1996. Clonal selection and learning in the antibody system. *Nature*. **381**(6585), pp.751–758.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*. **43**(7), p.e47.

Roth, D.B. 2014. V(D)J recombination: Mechanism, errors, and fidelity. *Microbiology spectrum*. **2**(6).

Roulland, S., Kelly, R.S., Morgado, E., Sungalee, S., Solal-Celigny, P., Colombat, P., Jouve, N., Palli, D., Pala, V., Tumino, R., Panico, S., Sacerdote, C., Quirós, J.R., Gonzáles, C.A., Sánchez, M.-J., Dorronsoro, M., Navarro, C., Barricarte, A., Tjønneland, A., Olsen, A., Overvad, K., Canzian, F., Kaaks, R., Boeing, H., Drogan, D., Nieters, A., Clavel-Chapelon, F., Trichopoulou, A., Trichopoulos, D., Lagiou, P., Bueno-de-Mesquita, H.B., Peeters, P.H.M., Vermeulen, R., Hallmans, G., Melin, B., Borgquist, S., Carlson, J., Lund, E., Weiderpass, E., Khaw, K.-T., Wareham, N., Key, T.J., Travis, R.C., Ferrari, P., Romieu, I., Riboli, E., Salles, G., Vineis, P. and Nadel, B. 2014. t(14;18) Translocation: A predictive blood biomarker for follicular lymphoma. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*. **32**(13), pp.1347–1355.

Sagaert, X., Sprangers, B. and De Wolf-Peeters, C. 2007. The dynamics of the B follicle: understanding the normal counterpart of B-cell-derived malignancies. *Leukemia*. **21**(7), pp.1378–1386.

Saijo, K., Schmedt, C., Su, I.-H., Karasuyama, H., Lowell, C.A., Reth, M., Adachi, T., Patke, A., Santana, A. and Tarakhovsky, A. 2003. Essential role of Src-family protein tyrosine kinases in NF-kappaB activation during B cell development. *Nature immunology*. **4**(3), pp.274–279.

Sato, H., Miyamoto, T., Yogev, L., Namiki, M., Koh, E., Hayashi, H., Sasaki, Y., Ishikawa, M., Lamb, D.J., Matsumoto, N., Birk, O.S., Niikawa, N. and Sengoku, K. 2006. Polymorphic alleles of the human MEI1 gene are associated with human azoospermia by meiotic arrest. *Journal of human genetics*. **51**(6), pp.533–540.

Scheeren, F.A., Nagasawa, M., Weijer, K., Cupedo, T., Kirberg, J., Legrand, N. and Spits, H. 2008. T cell-independent development and induction of somatic hypermutation in human IgM+ IgD+ CD27+ B cells. *The journal of experimental medicine*. **205**(9), pp.2033–2042.

Schmitz, R., Ceribelli, M., Pittaluga, S., Wright, G. and Staudt, L.M. 2014. Oncogenic mechanisms in Burkitt lymphoma. *Cold Spring Harbor perspectives in medicine*. **4**(2), pp.a014282–a014282.

Seifert, M., Scholtysik, R. and Küppers, R. 2013. Origin and pathogenesis of B cell lymphomas. *Methods in molecular biology (Clifton, N.J.)*. **971**, pp.1–25.

Sha, C., Barrans, S., Cucco, F., Bentley, M.A., Care, M.A., Cummin, T., Kennedy, H., Thompson, J.S., Uddin, R., Worrillow, L., Chalkley, R., van Hoppe, M., Ahmed, S., Maishman, T., Caddy, J., Schuh, A., Mamot, C., Burton, C., Tooze, R., Davies, A., Du, M.-Q., Johnson, P.W.M. and Westhead, D.R. 2019. Molecular high-grade B-cell lymphoma: Defining a poor-risk group that requires different approaches to therapy. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*. **37**(3), pp.202–212.

Sha, Y., Phan, J.H. and Wang, M.D. 2015. Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*. **2015**, pp.6461–6464.

Smith, K.G.C. and Clatworthy, M.R. 2010. FcgammaRIIB in autoimmunity and infection: evolutionary and therapeutic implications. *Nature reviews. Immunology*. **10**(5), pp.328–343.

Soneson, C., Love, M.I. and Robinson, M.D. 2015. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*. **4**(1521), p.1521.

Stewart, A., Ng, J.C.-F., Wallis, G., Tsioligka, V., Fraternali, F. and Dunn-Walters, D.K. 2021. Single-cell transcriptomic analyses define distinct peripheral B cell subsets and discrete development pathways. *Frontiers in immunology*. **12**, p.602539.

Suryani, S., Fulcher, D.A., Santner-Nanan, B., Nanan, R., Wong, M., Shaw, P.J., Gibson, J., Williams, A. and Tangye, S.G. 2010. Differential expression of CD21 identifies developmentally and functionally distinct subsets of human transitional B cells. *Blood*. **115**(3), pp.519–529.

Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P., Jensen, L.J. and von Mering, C. 2021. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*. **49**(D1), pp.D605–D612.

Tang, Y., Horikoshi, M. and Li, W. 2016. Ggfortify: Unified interface to visualize statistical results of popular R packages. *The R journal*. **8**(2), p.474.

Ten Hacken, E., Guièze, R. and Wu, C.J. 2017. SnapShot: Chronic Lymphocytic Leukemia. *Cancer cell*. **32**(5), pp.716-716.e1.

Tierens, A., Delabie, J., Michiels, L., Vandenberghe, P. and De Wolf-Peeters, C. 1999. Marginal-zone B cells in the human lymph node and spleen show somatic hypermutations and display clonal expansion. *Blood*. **93**(1), pp.226–234.

Tonegawa, S. 1983. Somatic generation of antibody diversity. *Nature*. **302**(5909), pp.575–581.

Treanor, B. 2012. B-cell receptor: from resting state to activate: Structure and dynamics of the BCR. *Immunology*. **136**(1), pp.21–27.

Triplett, T.A., Curti, B.D., Bonafede, P.R., Miller, W.L., Walker, E.B. and Weinberg, A.D. 2012. Defining a functionally distinct subset of human memory CD4+ T cells that are CD25POS and FOXP3NEG: Clinical immunology. European journal of immunology. 42(7), pp.1893–1905.

Tsujimoto, Y., Gorham, J., Cossman, J., Jaffe, E. and Croce, C.M. 1985. The t(14;18) chromosome translocations involved in B-cell neoplasms result from mistakes in VDJ joining. *Science (New York, N.Y.)*. **229**(4720), pp.1390–1393.

Van Rossum, G. and Drake, F.L., Jr 2009. *Python 3 Reference Manual: (Python Documentation Manual Part 2)*. Createspace.

van Spriel, A.B., de Keijzer, S., van der Schaaf, A., Gartlan, K.H., Sofi, M., Light, A., Linssen, P.C., Boezeman, J.B., Zuidscherwoude, M., Reinieren-Beeren, I., Cambi, A., Mackay, F., Tarlinton, D.M., Figdor, C.G. and Wright, M.D. 2012. The tetraspanin CD37 orchestrates the α(4)β(1) integrin-Akt signaling axis and supports long-lived plasma cell survival. *Science signaling*. **5**(250), p.ra82.

Vinuesa, C.G., Tangye, S.G., Moser, B. and Mackay, C.R. 2005. Follicular B helper T cells in antibody responses and autoimmunity. *Nature reviews. Immunology*. **5**(11), pp.853–865.

Weller, S., Mamani-Matsuda, M., Picard, C., Cordier, C., Lecoeuche, D., Gauthier, F., Weill, J.-C. and Reynaud, C.-A. 2008. Somatic diversification in the absence of antigen-driven responses is the hallmark of the IgM+ IgD+ CD27+ B cell repertoire in infants. *The journal of experimental medicine*. **205**(6), pp.1331–1342.

Weniger, M.A. and Küppers, R. 2021. Molecular biology of Hodgkin lymphoma. *Leukemia*. **35**(4), pp.968–981.

Wilson, P.C., de Bouteiller, O., Liu, Y.J., Potter, K., Banchereau, J., Capra, J.D. and Pascual, V. 1998. Somatic hypermutation introduces insertions and deletions into immunoglobulin V genes. *The journal of experimental medicine*. **187**(1), pp.59–70.

Witt, E., Hammer, E., Dörr, M., Weitmann, K., Beug, D., Lehnert, K., Nauck, M., Völker, U., Felix, S.B. and Ameling, S. 2019. Correlation of gene expression and clinical parameters identifies a set of genes reflecting LV systolic dysfunction and morphological alterations. *Physiological genomics*. **51**(8), pp.356–367.

Wolfinger, R. and Tan, P.-Y. 2017. Stacked ensemble models for improved prediction accuracy.

Wulff, H., Knaus, H.-G., Pennington, M. and Chandy, K.G. 2004. K+ channel expression during B cell differentiation: implications for immunomodulation and autoimmunity. *The journal of immunology*. **173**(2), pp.776–786.

Xie, Y., Pittaluga, S. and Jaffe, E.S. 2015. The histological classification of diffuse large B-cell lymphomas. *Seminars in hematology*. **52**(2), pp.57–66.

Xiong, Y., Ye, M. and Wu, C. 2021. Cancer classification with a cost-sensitive naive Bayes stacking ensemble. *Computational and mathematical methods in medicine*. **2021**, p.5556992.

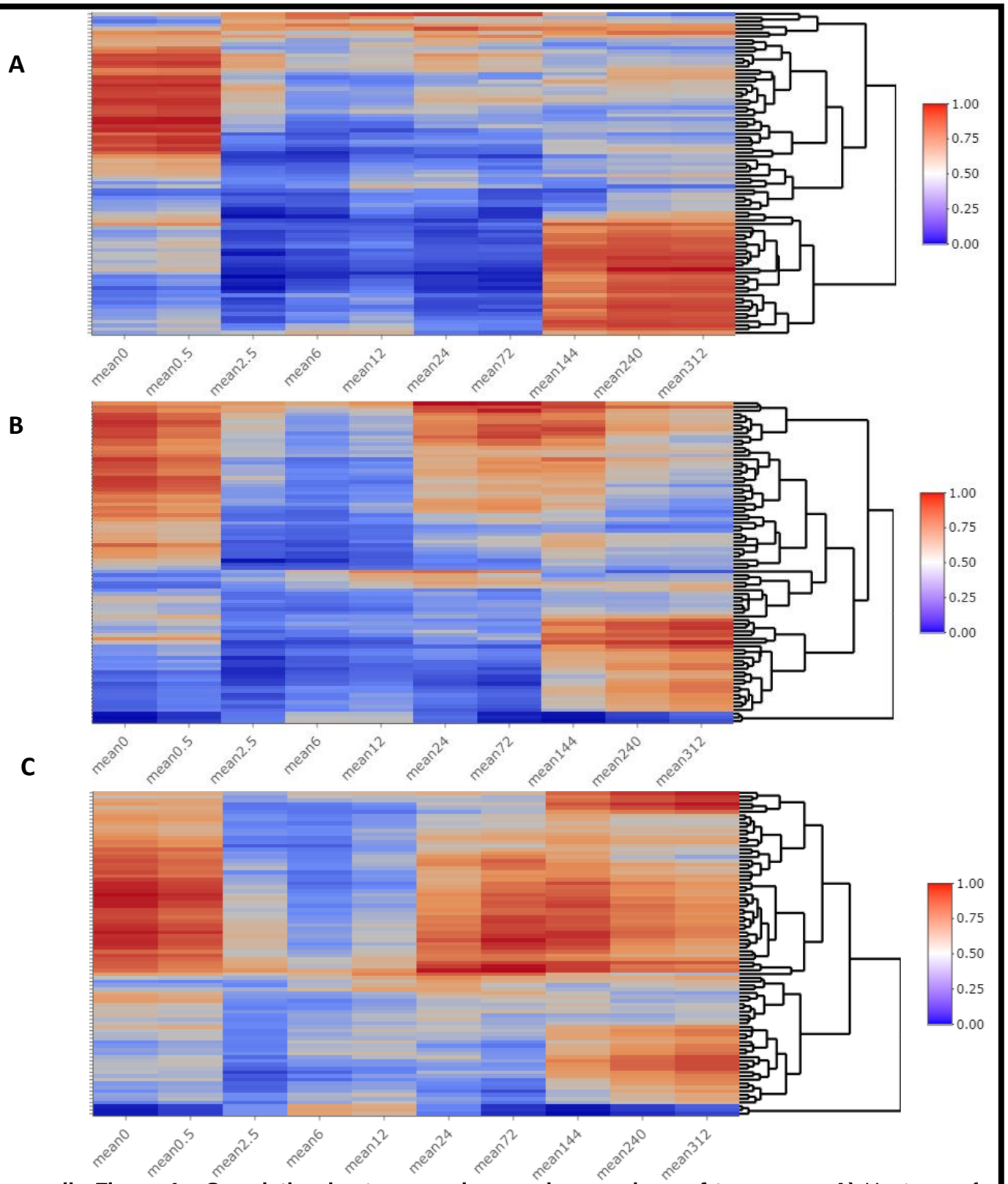
Yu, D., Zhang, Z., Glass, K., Su, J., DeMeo, D.L., Tantisira, K., Weiss, S.T. and Qiu, W. 2019. New Statistical Methods for Constructing Robust Differential Correlation Networks to characterize the interactions among microRNAs. *Scientific reports*. **9**(1), p.3499.

Zhang, H., Yu, C.-Y. and Singer, B. 2003. Cell and tumor classification using gene expression data: construction of forests. *Proceedings of the National Academy of Sciences of the United States of America*. **100**(7), pp.4168–4172.

Zhang, J., Zhang, W., Zou, D., Chen, G., Wan, T., Li, N. and Cao, X. 2003. Cloning and functional characterization of GNPI2, a novel human homolog of glucosamine-6-phosphate isomerase/oscillin. *Journal of cellular biochemistry*. **88**(5), pp.932–940.

Zhao, Y., Li, M.-C., Konaté, M.M., Chen, L., Das, B., Karlovich, C., Williams, P.M., Evrard, Y.A., Doroshow, J.H. and McShane, L.M. 2021. TPM, FPKM, or normalized counts? A comparative study of quantification measures for the analysis of RNA-seq data from the NCI patient-derived models repository. *Journal of translational medicine*. **19**(1), p.269.

Appendix:



Appendix Figure 1: Correlation heatmaps using varying numbers of top genes. A) Heatmap of timepoint correlation coefficient using the top 15 significant timeseries genes. **B)** Heatmap of timepoint correlation coefficient using the top 50 significant timeseries genes. **C)** Heatmap of timepoint correlation coefficient using the top 100 significant timeseries genes. The x axis is the average correlation coefficient for each timepoint, and the y axis represents the B-cell cancer cell line samples correlation coefficient.