Proposal Report

Human-specific genes

Andrea Tonina, Thomas Sirchi,
Lorenzo Santarelli, Gloria Lugoboni, Sabri Kaci
Group B1

October 2023

Contents

1 Abstract

The term human-specific genes highlights a set of genes that characterize us as humans and cannot be found in our closest relatives, the chimpanzees. The role of human-specific (HS) genes is yet to be completely discovered and understood. Several studies have been carried out to identify and extend the list of known human-specific genes, with the final objective of understanding their linkage with human diseases. This study tries to extend our knowledge of human-specific genes regarding Acute Lymphoblastic Leukemia (ALL) by investigating the associations between HS genes and ALL using quantitative methodologies such as machine learning-based approaches and network gene expansions.

2 Introduction

2.1 Background

2.1.1 Human-specific genes

Humans and chimpanzees separated circa 6 million years ago. From this point over, a rapid evolution and new alterations have been acquired. The main differences can be found at the genetic level, given specifically from aberrations such as rearrangements, duplications, and losses. These mutations resulted in orthologous genes and de-novo ones. The role of these aberrations can be seen at different levels, starting from the diet and immune changes to anatomy (brain and neuroanatomy comparisons, bones, etc). We talk of "human-specific" features, therefore human-specific genes [1].

It was possible to associate human-specific genes with a restricted set of functions. Specifically, the main terms identified were: neural functions, metabolic functions (carbohydrate metabolism, adipogenesis pathway, glutamate biosynthesis, etc), immunological functions (parasitism, host-response, phagocytosis, etc), and functions at the cell level (cytoskeleton organization, motility, transport, protein modifications, and targeting).

These results were obtained via a pipeline that involved the use of:

- GeneTerm Linker (FGNet), an algorithm used to identify associated genes via a process of clustering [2];
- Gene Ontology (GO) Analysis, to perform enrichment analysis on gene sets and identify over-represented (or under-represented) terms [3];
- Ingenuity Pathway Analysis (IPA), a web-based application that allows functional analysis of genes to identify networks or correlations [4];

Human-specific genes are still to be completely uncovered. Correlations have been made between these genes and diseases, identifying so-called human-specific diseases [5].

2.1.2 Pediatric Acute Lymphoid Leukemia (ALL)

In this research project, we decided to focus on Acute Lymphoid Leukemia (ALL) both Pediatric and adult ALL. ALL is one of the most common leukemia in children (80% of patients are children). It is a malignant transformation that causes an abnormal proliferation and differentiation of lymphoid progenitor cells. It correlates with genetic aberrations and complex events such as rearrangement of multiple chromosomes. One known translocation example is the translocation BCR-ABL1, also known as Philadelphia chromosome. In children, it is often observed associated with other syndromes that have a genetic predisposition, such as Down syndrome, Bloom syndrome, and Fanconi anemia. The symptoms are usually non-specific. In some cases, bone marrow failure or involvement of the central nervous system is observed, with cranial nerve deficit or meningismus [6].

3 Materials and Methods

3.1 Databases

3.1.1 GEO

The Gene Expression Omnibus (GEO) database is a public resource containing high-throughput gene expression and other functional genomics data sets [7].
It was founded in 2000, rapidly evolving to contain multiple datasets connected to genome methylation data, chromatin structure, genome–protein interactions, whole-genome sequencing or RNA-sequencing.

3.1.2 cBioPortal

cBioPortal is a database containing cancer genomics data set linked to patients and clinical applications. This database enables an easy and direct viability of raw data to the entire cancer research community [8].

3.1.3 PubMed

PubMed is a database containing citations and abstracts of biomedical papers/literature. It is a searching tool that enables fast and easy retrieval of biomedical and life sciences literature, improving research and viability of information [9].

3.1.4 STRING

The STRING database collects information on protein-protein interactions (PPI). In the current STRING version, 24,584,628 proteins from 5,090 organisms are integrated. Thanks to this tool it is possible to obtain an insight into the interaction partners and networks generated from one or multiple genes [10].

3.1.5 WikiPathways

WikiPathways is a platform containing information on biological pathways and protein interactions underlying biological processes. Thanks to this database it is possible to retrieve information regarding subsets of molecular pathways related to diseases and metabolism. Thanks to the collaborative nature of this database, the information is in continuous development [11].

3.2 Tools

3.2.1 Gene Ontology

A gene ontology is a way to capture biological knowledge for individual gene products in a written and computable form. It has a formal structure and can be defined as a set of concepts and their relationships, arranged in a hierarchy, from a less specific description to a more specific description. In a gene ontology, each component is associated with a specific notion [3].
Three main hierarchies can be defined:

- Molecular Function. This category describes activities that happen at a molecular level, it includes the activities that are involved in an action and do not specify where, when, or in which context the actions happen
- Biological Process. A biological process is a series of events resulting from multiple ordered groups of molecular functions. It can be thought of as a chain of execution.
- Cellular Component. A cellular component is linked to a component of a cell with the condition that is part of a larger object and can be part of an anatomic structure.

3.2.2 EnrichR

Enrichr is an enrichment analysis web-based tool. It is a resource that contains curated gene sets and it can be used as a search engine to visualize and rank enriched terms [12].

3.2.3 Combat and Combat-Seq

In microarray experiments, it is often possible to observe technical and non-technical biases called "batch effects". Combat is a robust tool that exploits an empirical Bayes framework to perform a correction of the batch effects in the data, recalibrating them [13].
Combat-Seq is a tool that exploits a negative binomial regression model to remove batch effects from RNA sequencing data. Combat-Seq is also implemented to reduce the number of false positives in differential expression and recover the biological signal in the data [14].

3.2.4 Differential Gene Expression

A cell's gene expression profile is the snapshot of which genes are expressed in that cell at the time the sample was taken. Knowing which genes are expressed in a cell at a certain moment allows the identification of new genes or transcripts and the comparison of expression profiles between samples (a typical scenario we are interested in). Gene expression profiles are extremely heterogeneous since they vary based on the individual, tissue, condition, and cells of origin. During a differential gene expression experiment the expression profile of genes is compared between samples. Comparisons are usually effectuated over different disease states or differences between healthy and diseased individuals [15].
The main technology to obtain expression data is RNA sequencing. RNA sequencing is a next-generation sequencing approach that sequences the cDNA from the mRNA component. A whole variety of sequencing machines exist. Based on the biological question a specific machine needs to be used. These machines are distinguished on the throughput, the read length, and the coverage [16].

3.2.5 Principal Component Analysis

Principal component analysis (PCA) is a technique used to reduce the dimensionality of large datasets thanks to the identification of so-defined "principal components". This technique is used to increase interpretability and at the same time minimize information loss. Principal components can be seen as variables that maximize the variance among the data, maximizing the corresponding information brought [17].

3.2.6 OneGenE and gene@home project

OneGenE [18] is an evolution of the NES2RA algorithm [19], based on the PC-algorithm, it can expand the Local Gene Network (LGN) using transcriptomic data. Thanks to the gene@home project, a computation project that relies on the use of volunteers' computers to make the computation faster [20]. The term gene network expansion defines a process that aims at finding new genes to expand a given known gene network.
OneGenE uses the gene@home project. The work is divided into work units composed of PC-IM expansions. These are sent to the volunteers, computed, sent back, checked by the validator, and then stored on the server. When the expansion of an LGN is needed, the already computed expansions are retrieved and aggregated and the final ranked list is given to the user [21].

3.2.7 TMM normalization

The TMM (Trimmed Mean Method) is one of the most famous methods to perform intra and intersample normalization. The idea behind this method is to estimate a scaling factor for correction based on the mean of the samples in which outliers have been removed. This method is widely used to work on RNA Sequencing data, an example is the package edgeR that contains multiple commands that can perform this type of normalization [22].

3.2.8 GeneOverlap

GeneOverlap is an R package able to operate a process of overlapping between gene lists. During this process, a contingency table is created and a statistical significance is associated with the overlap. Thanks to this package it is possible to obtain important biological results, that can be used to validate and compare lists of genes obtained via different processes [23].

3.2.9 NetworkX

NetworkX is a Python package able to perform network analysis graph creation [24]. Thanks to this package it is possible to create a different type of object, specifically:

- Undirect graphs with self-loops
- Directed graphs with self-loops
- Undirect graphs with self-loops and parallel edges
- Directed graphs with self-loops and parallel edges

3.2.10 Clustering

Clustering is a type of unsupervised learning. The term clustering is used to indicate a wide group of techniques that aim at identifying subgroups in a dataset. The idea behind this type of learning is to group similar objects, for this reason, it is important to define a criteria for similarity [25].

Hierarchical Clustering Hierarchical clustering is a type of unsupervised learning, a part of machine learning. In hierarchical clustering, the resulting clustering is represented as an observation tree called dendrogram [26]. Two main techniques of hierarchical clustering exist:

- In agglomerative clustering each initial data point is considered as a single cluster, as the algorithm proceeds, the ones less distant are merged.
- In divisive clustering, the initial points are considered all part of the same cluster. As the algorithm proceeds, the data points are separated into more clusters.

PAM - k-medios (PAM) clustering is based on the idea that it is possible to represent the cluster center using a single point. This point is defined by minimizing the distance between all the points that are found in the cluster [27].

3.3 Machine-learning approaches for classification

Machine learning (ML) is a process in which a computer learns the rules and patterns from a large amount of data and uses the mined information to solve problems such as classification or regression [28]. ML approaches for classification are those methods that aim to assign a label or category to input data. Classification is a type of supervised learning, where the algorithm learns from labeled data and then applies the learned rules or patterns to new data of interest. In this project, we tackle a problem of multi-class classification where the goal is to classify the data into more than two classes, such as tumor subtypes. The methods implemented in this project are explained below.

3.3.1 Random Forest

The random forest (RF) method is a well-known and popular method among ensemble classification techniques. Focused on the idea of parallel assembly that can fit several decision trees at the same time to obtain a majority vote or an average to determine the consensus over the final output. The multiple decision trees employed by the algorithms are produced through a process of bootstrap aggregation and by the selection of random features, where the number can change by user input. Therefore, the RF learning model with multiple decision trees is more accurate than a single decision tree-based model [29].

3.3.2 K-nearest neighbors

K-Nearest Neighbors (KNN) is an *instance-based learning* or *non-generalizing learning*, also known as a *lazy learning* algorithm [29]. This approach brings not at the creation of a model itself but at establishing distance between the data points using similarity measurements between the k, hyperparameter, and nearest neighbors. Once k is given as input from the user the algorithm proceeds to compute the classification through majority vote.

3.3.3 Naive Bayes

The naive Bayes algorithm is based on Bayes' theorem with the assumption of independence between each pair of features [29]. The fundamental concept involves determining the features of a sample for pending classification, calculating the a posteriori probability for each label, and subsequently assigning the sample to the label with the highest a posteriori probability [28].

3.3.4 XGBoost

XGBoost (xgb) is an efficient implementation of gradient boosting which is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges [30]. Gradient Boosting, like Random Forests above, is an ensemble learning algorithm that generates a final model based on a series of individual models, typically decision trees [29]. XGBoost method uses a regularized objective function to prevent overfitting and supports various loss functions. It can deal with sparse and missing data and uses efficient algorithms to find the best splits. It stores the data in a compressed column block structure, which allows parallel and out-of-core computation and reduces

disk IO overhead. It is important to note that this open-source package runs on different platforms and supports multiple languages like R used in this project [30].

3.4 Datasets

3.4.1 Expression data matrices

The datasets we used can be found under. A small table is represented under in which we highlight the number of samples for each datasets and their nature (controls or tumor samples).

RNA-Sequencing		
Dataset Names	Number of samples	Data types
GSE181157	173	Pediatric Tumor
GSE133499	42	Pediatric Tumor
GSE227832	330+10	Pediatric Tumor + Control
GSE84445	20	Adult Control
T-ALL_cohort7-8	107	Pediatric + Adult Tumor

Datasets T-ALL_cohort7-8 and GSE181157 were associated with metadata linked to the subtype and clinic relevance.

4 Pipeline

The full pipeline is represented in Figure 1

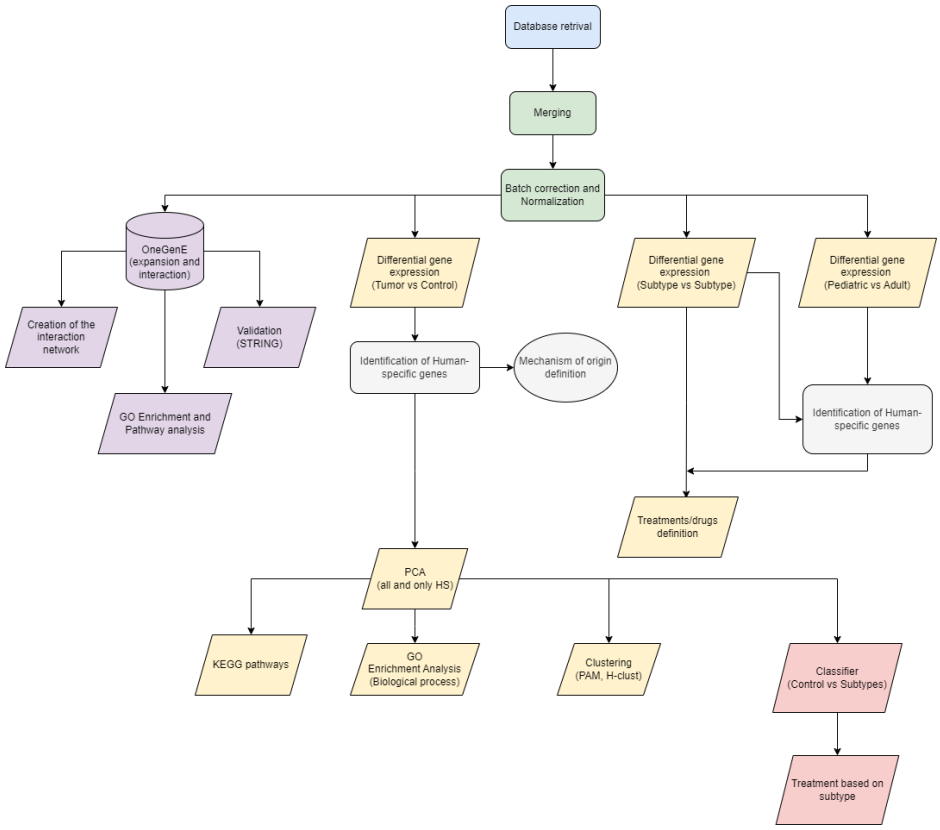


Figure 1: Workflow for the project

4.1 Pre-processing, normalization, and batch effect correction

Replicates were filtered from the datasets to obtain a unique sample for each patient. The Hugo Symbols were converted into Ensembl IDs to facilitate the merging of the datasets. This process is delicate since there isn't a unique mapping 1:1 between the Hugo Symbols and the Ensembl ones (and vice-versa). We resolved this issue by filtering the datasets for duplicates. We also filtered the data for the low-expressed genes in the samples. A batch correction was performed using Combat-Seq (sva package in R), and the vector of the batch was defined considering the merging points of the datasets. The data was normalized using a TMM normalization and the corresponding CPM table was created.

4.2 Differential Gene expression

4.2.1 DGE Control vs Cancer

The Differential Gene expression analysis was effectuated using the R package EdgeR which enable us to model the RNA Seq data as a negative binomial distribution. The samples were labeled as "control" or "tumor" based on the literature information from which the samples were retrieved. We specified the contrast as *'Tumor'-'Controls'*. The differentially expressed genes were filtered using a p-value of 0.01. We highlighted down-regulated genes by setting the logarithm fold change smaller-equal than -1.5 while the up-regulated genes were identified by setting a logarithm fold change major-equal than 1.5. We then extracted the human-specific genes that were found up and down-regulated in tumor samples concerning the control ones.

4.2.2 DGE tumor vs tumor

The same process effectuated for the differential gene expression of the tumor against control was effectuated to study the differentially expressed genes in the different tumor subtypes. Specifically, thanks to the metadata that was associated with the samples, we effectuated a DGE analysis on three subtypes. We labeled the samples as "PreT", "PreB", and "T". We specified the contrasts as *'PreT'-'(PreB+T)/2'*, *'PreB'-'(PreT+T)/2'* and *'T'-'(PreB+PreT)/2'*. We filtered the obtained data using a p-value of 0.01 and defined the up and down-regulated genes setting a threshold for the logarithm fold change (1.5 for up-regulated and -1.5 for down-regulated). We finally extracted the human-specific genes that were found up and down-regulated in the subtypes and compared them to obtain information on the specificity of these genes for the subtypes.

4.2.3 DGE tumor pediatric vs tumor adult

We finally effectuated a DGE analysis comparing pediatric and adult samples. We labeled the samples as "pediatric" and "adult". We specified the contrasts as *'pediatric'-'adult'* and filtered the obtained data using a p-value of 0.01. Also, in this case, we defined the up-regulated genes using by setting a threshold equal to 1.5 on the logarithm fold change, while the down-regulated genes were identified by setting the threshold to -1.5. We finally extracted the human-specific genes.

4.3 PCA

Now on, we focused on the data obtained from the tumor vs control differential gene expression. We effectuated a PCA analysis both on the complete gene expression datasets and the one containing only the human-specific genes. We also decided to effectuate a PCA on tumor-only data, considering both the complete gene set and the human-specific only. We used the information retrieved from the metadata (Risk factor, age, subtype) to investigate the ability of the PCA to stratify the data.

4.4 Clustering

Based on the PCA results, we operated a clustering process. The analysis was effectuated on both the dataset containing tumors and controls and the one containing only tumors. We used two different techniques:

- Hierarchical clustering, in which the distance between clusters was obtained using the average linkage. The number of clusters was defined using the *factoextra* package, obtaining the average silhouette graph and the gap statistics graph.
- PAM (partition around medoids) clustering.
  Also, in this case, the number of clusters was needed.

The analysis was effectuated considering the PCA data retrieved from all the genes and also only considering the human-specific genes.

4.5 Enrichment and Pathway analysis

The analysis was performed using the package *clusterProfiler* in R.

4.5.1 Gene Ontology

GO enrichment analysis was performed on both the up and down DEGs and human-specific DEGs sets. The Ensembl IDs were initially converted to gene symbols and Entrez gene IDs. The enrichment was performed using *enrichGO* and focusing on the Biological Process (BP) sub-ontology.

4.5.2 Pathway analysis

Functional enrichment analysis on WikiPathway was performed using enrichWP. We performed the analysis on both the up and down DEGs set and on the human-specific DEGs set.

4.5.3 Identification of treatments or drugs

Another kind of analysis we performed based on the results of differential expression analysis was a drug enrichment analysis using the *enrichR* tool. Enrichr is a comprehensive resource for curated gene sets and a search engine that accumulates biological knowledge for further biological discoveries[31], we implemented the tool available for R. We started by selecting this dataset in the enrichR ecosystem "DrugMatrix". This dataset provides information related to drugs and can associate combinations of genes, forming a treatment that holds potential effectiveness. To streamline our analysis, we categorized the Differentially Expressed Genes (DEGs) into their specific subtypes. This approach allowed us to conduct an independent analysis for each subtype, offering insights and recommendations regarding the most suitable drugs for use in each case. We will take into consideration the "Combined score" value to determine the best results. The "Combined score" is a metric that takes into consideration various aspects such as p-value and gene overlap and displays a higher number for better results.

4.6 Machine-learning approaching for classification

The objective behind the application of ML to our data was the prediction of metadata regarding the sub-type of tumor of the patients. Following the PCA effectuated on analysis of differential expression we crossed our data to the metadata available and divided the tumor patients into the following subtypes:

- 9836/3 - Pre-B ALL
- 9837/3 - Pre-T ALL
- T Cell
- Unknown

To address the challenge of classifying *Unknown* cases into the specified sub-types, we opted for the utilization of a classification algorithm. We wanted to take advantage of multiple methodologies to establish a consensus, ensuring confidence in the final output. All the implementation was done in the R programming language. In continuation of the prior choice to partition the data into human-specific (HS) and non-human-specific (nonHS) subsets, we opted to implement the machine learning pipeline separately for each of these distinct datasets. The pipeline was designed to keep reusability and ease of use for the user, therefore, we made it salable to satisfy different needs and data without the need to modify the code. In both our training and prediction we always kept the control in to give the model a clear idea of what not to predict and as a metric to further enhance our classification and as a threshold since the correct prediction of all thirty controls will be considered the baseline for the choice of the models.

4.7 Gene expansion

The expansion of the gene set has been done via Gene@Home project expansion using OneGene. The input consists of the expression matrix of our genes and a list of genes of interest to expand, in our cases this list was composed of the up and down-regulated human-specific gene found by previous analyses. Concerning the PC algorithm parameter we set the significance threshold $alpha$=0.05, the size of the tiles in which our gene list will be divided

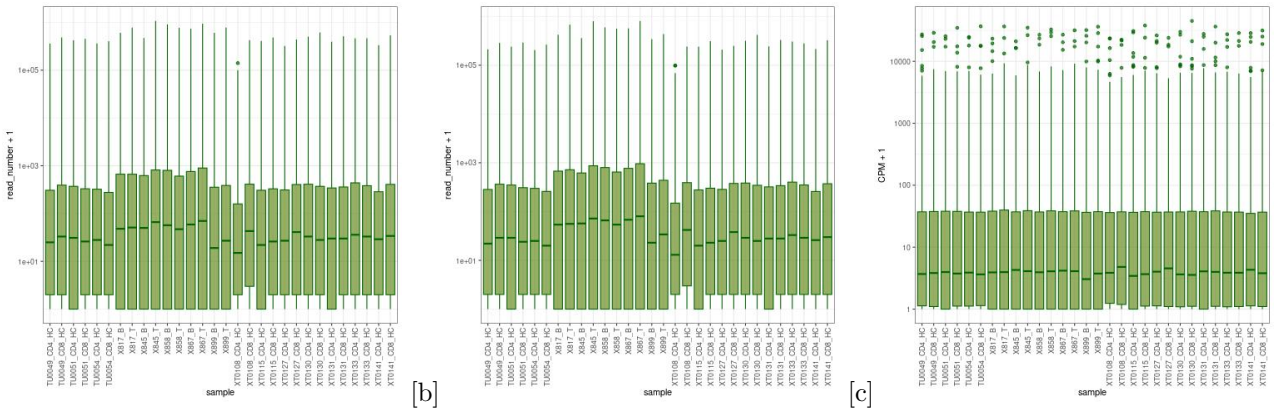


Figure 2: (a) Control data before CombatSeq and before normalization (b) Control data after CombatSeq correction (c) Control data after normalization.

tsize=2000, and finally the number of iterations to be executed *niter*=2000. In particular, the *niter* number is directly connected to the total number of tiles tested, so higher values mean a higher number of gene combinations researched and consequently reduce the false negative rate.

4.7.1 Interaction network

From the expansion files that we retrieved from the TN-Grid platform, we generated a graph using NetworkX in Python. This initial graph was obtained by filtering the data of the expansion by using a threshold of 95% on the absolute frequency. We then obtained a subgraph by maintaining all the nodes with a degree major equal to 3. This way, we tried to focus only on the most connected genes. We then decided to focus on the human-specific genes, we specifically extracted three graphs in which we found human-specific genes highly connected. The subsetting of the node was performed by working on the original graph and retrieving all the neighbors of the identified human-specific genes.

4.7.2 Validation

To validate the expansions obtained via OneGene we compared them with gene interactors obtained from another database. A small number of genes was selected based on the most connected genes from the previous analysis. Those interactors have been retrieved using STRING and subsequently, a comparison between single gene expasions was performed using GeneOverlap. The p-value resulting from the Fisher test of the overlap was used to evaluate the expansions. In those cases where the Overlap was not successful, a comparison using the entire expansion set was performed as a further control.

5 Results

5.1 Pre-processing, normalization, and batch effect correction

After Pre-processing passages, we applied the normalization and batch effect correction, and to show changes in the data some boxplots of the Control and Tumor datasets were generated. We can see in both Figure 2 and Figure 3 that the batch correction and the normalization process were both needed to obtain comparable data.
Since the Tumor dataset is formed by 640 samples, we decided to represent in Figure 3 only a subset of 20 samples them.

5.2 Differential Gene expression

5.2.1 DGE Control vs Cancer

After the normalization part, it was possible to define a set of differentially expressed genes between the controls and the tumor samples. To distinguish up-normal-down regulated genes we used the expression values of the fold change of the transcripts. The selection is based on the log fold change ratio (>1.5 for up-regulated genes and < (-1.5) for down-regulated genes) and a log CPM (>1 for both cases). Specifically, 2693 genes were found up-regulated while

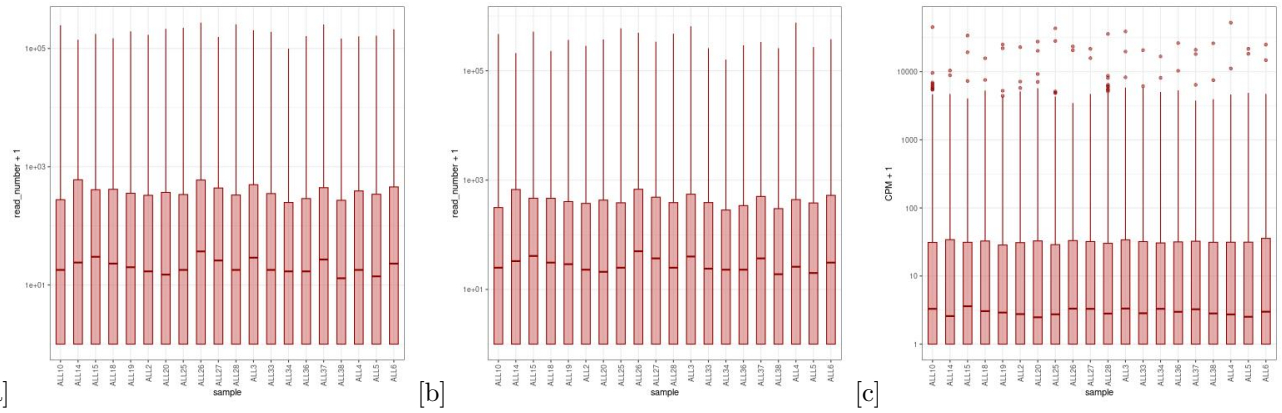


Figure 3: (a) Tumor data before CombatSeq and before normalization (b) Tumor data after CombatSeq correction (c) Tumor data after normalization.

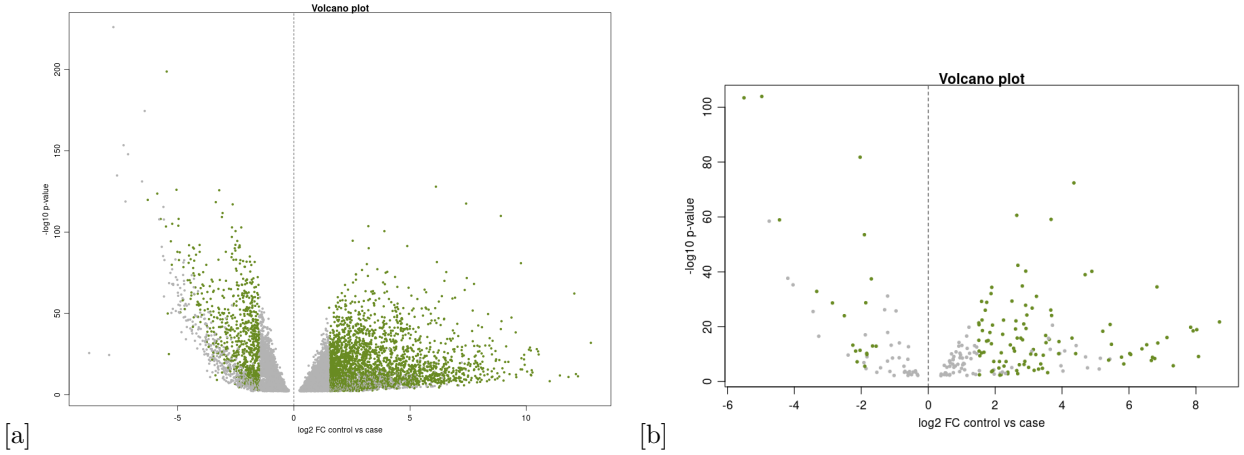


Figure 4: **Vocalno plots**. x-axes: log FoldChange, y-axes: inverse function of the p-value. (a) Volcano plot of DEGs of all ALL genes, (b) Volcano plot of DEGs of human-specific ALL genes.

756 were found down-regulated in tumors with respect to controls. Of these, 122 were human-specific genes, 103 were up-regulated and 19 were down-regulated.

We displayed the result with the utilization of Vulcano plots, Figure 4, the plots diversify the most significant DEGs colored in green, which are genes that surpass a threshold set on both the p-value and the Fold Change. From the figure with all the genes is possible to see that 36% of the genes are differentially expressed, by taking a look at the volcano plot is possible to easily see that there are more DEGs on the right side of the plot, which represent the up-regulated, compared to the left site of the plot, that present the down expressed genes, this confirm the majority are up-regulated.

By taking into account only the subset of the human-specific genes, is possible to see the same pattern where the majority of DEGs are up-regulated by looking at the plot. This is confirmed by the data, indeed 52.6% of the human-specific genes are differentially expressed and 44.4% are up-regulated.

We can also represent the genes, both for all ALL gens and for the subset of human-specific ALL genes, using heatmaps like in Figure 5. It's possible to see that a clustering process is operated and only up or down-expressed genes are plotted using data from both the normalized CPM and the log transformation of the CPM table. It's possible to see that thanks to clusterization, chunks of expressed genes are outlined. Specifically, in the case of the CPM log table, a more refinement clusterization is observed, and minor contamination at the level of the division between "case" samples (brown) and "control" samples (green) is observed. We need to remember that contamination that took place during the collection of samples can influence the clustering process and also the number of genes, indeed for the human-specific genes heatmaps there are some problems in defining distinct branches between tumor and control. From all plots a clear difference in expression between tumor and control samples is observed, indicating that the expression of the genes differs between the two cases, as expected since we are performing a DEGs analysis.

After the identification of the human-specific genes found up and down-regulated between tumors and controls,

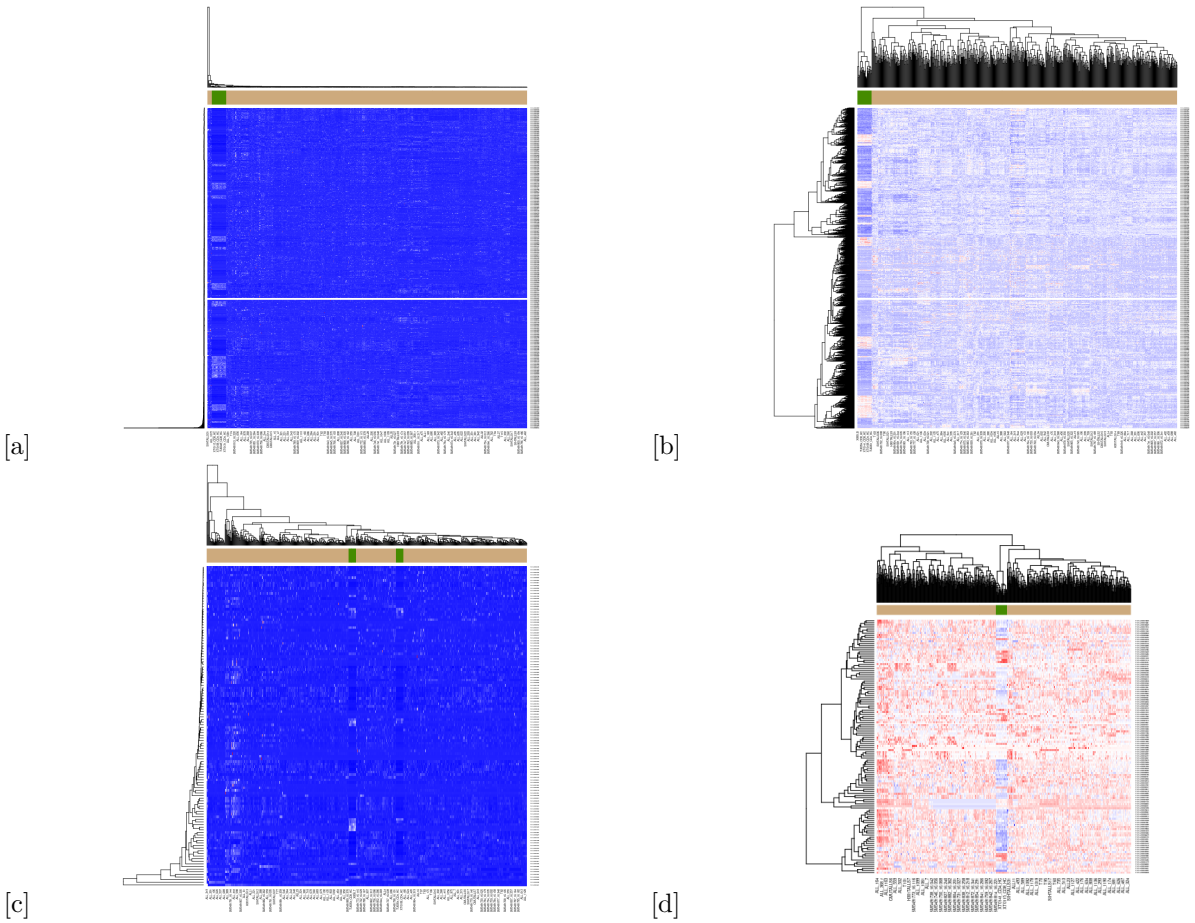
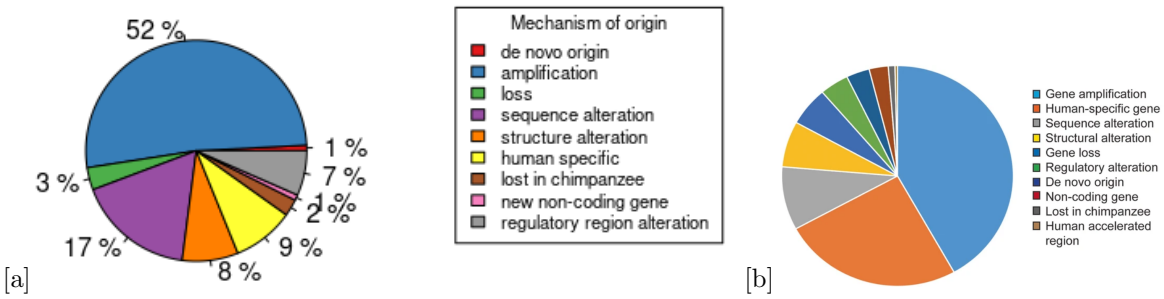


Figure 5: **Heatmap DEGs**. (a)Heatmap of normalized CPM table of all ALL genes,(b) Heatmap of normalized log CPM table of all ALL genes,(c)Heatmap of normalized CPM table of human-specific ALL genes,(d) Heatmap of normalized log CPM table of human-specific ALL genes.



[a] [b]

Figure 6: (a) Pie chart representing the distribution of the human-specific genes found up and down-regulated in tumors; (b) Pie chart illustrating the distribution of the known human-specific genes, from [1]

we investigated their mechanism of origin. We can see in Figure 6 (a) that more than 50% of the ALL-related human-specific genes were generated by copy number alteration. This is indeed similar to what is also found in literature, and we can see from in Figure 6 (b), obtained from [1].

5.2.2 DGE Subtype vs Subtype

After performing the DEGs Tumor vs Control we performed Differential Gene Expression Analysis by taking only into consideration tumor data. For the distinction of the regulation of the genes, we utilized the fold change of the transcript as a discriminative parameter, to be precise we used the log fold change and the log CPM to distinguish if the genes are up, norm, or down-regulated (the threshold are the same as the previous DEG analysis). We found out that for the tumor subtype Pre-B, 462 genes are up-regulated, and 432 are down-regulated, of these, we found that 22 down and 16 up-regulated are human-specific. For the subtype Pre-T, the analysis shows that 482 genes are up-regulated and 848 are down-regulated, where 32 down-regulated and 16 up-regulated genes are human-specific. The analysis for the subtype T resulted in 16 down-regulated genes and 381 up-regulated, the human-specific genes are only 32 that are up-regulated.

Then we made a cross-comparison to see if the DEGs human-specific genes present in a subtype are also present in all the other subtypes of the tumor, we found out that:

- Subtype Pre-B and Subtype T: 2 up-regulated genes in common
- Subtype Pre-B and Subtype Pre-T: 6 down-regulated genes in common
- Subtype Pre-T and Subtype T: 2 up-regulated genes in common

This means that the majority of DEG human-specific genes are not shared among the tumor subtypes, which means that they characterize the ALL subtypes.

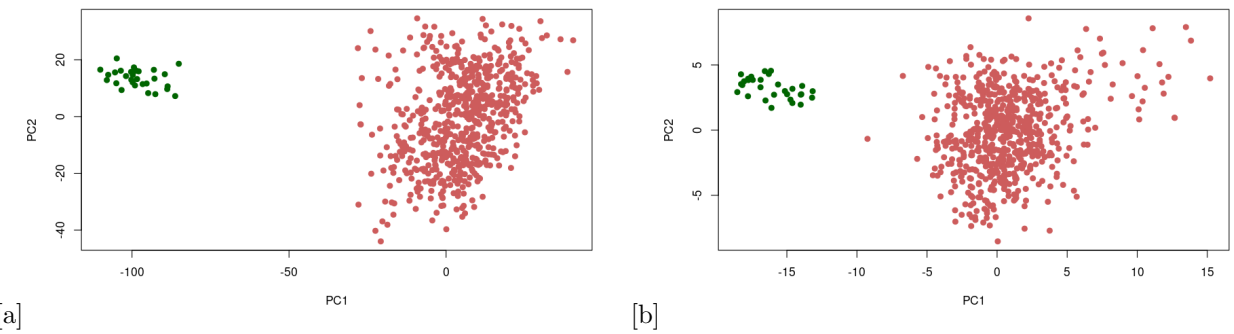
5.2.3 DGE Pediatric vs Adult

After performing the DEGs Subtype vs Subtypes we performed Differential Gene Expression Analysis by taking only into consideration tumor data and concentrating on the possible difference between ALL in pediatric vs adults. Again to distinguish up-normal-down regulated genes we used the expression values of the fold change of the transcripts. The selection is based on the log fold change ratio (>1.5 for up-regulated genes and < (-1.5) for down-regulated genes) and a log CPM (>1 for both cases). We were able to find out that in pediatric ALL tumors, 55 genes are up-regulated and 108 are down-regulated, compared to the adults. We need to note that there could be a bias because we have a lot more pediatric samples compared to adult ones and for adults especially, we were not able to retrieve any data for subtype B. Of these Differential expressed genes only 10 down-regulated and 6 up-regulated are also human-specific.

These findings confirm what is possible to find in the literature, that the pediatric ALL is genetically different compared to the adult ALL.

5.3 PCA

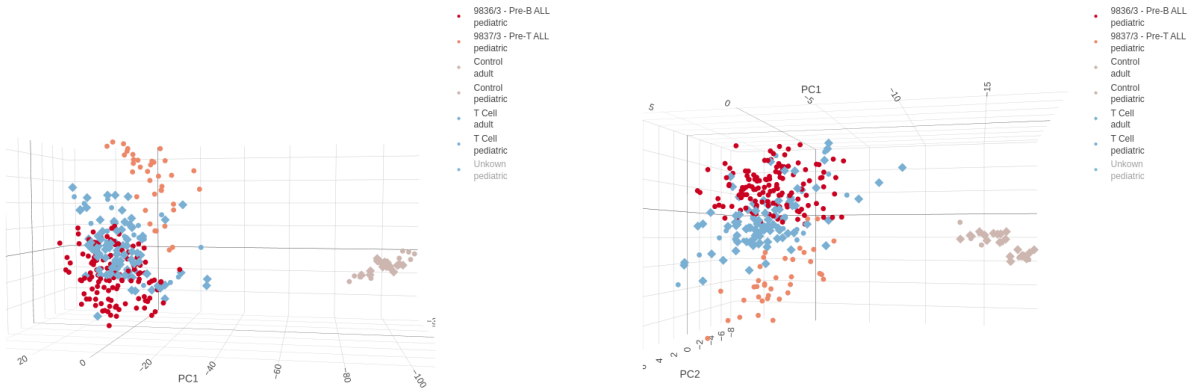
After the Differential Gene expression between Tumor vs Control, we checked if we were able to clusterize our data based on their Principal Component, this approach was used both for all ALL genes DEGS and the human-specific ALL DEGs. The results are presented in Figure 7, we can see that in both situations the identified two principal components can cluster and separate the Tumor samples from the Control and vice versa.



[a] [b]

Figure 7: **Principal Component Analysis**. x-axes: PC1, y-axes: PC2. (a) PCA plot of DEGs of all ALL genes, (b) PCA plot of DEGs of human-specific ALL genes.

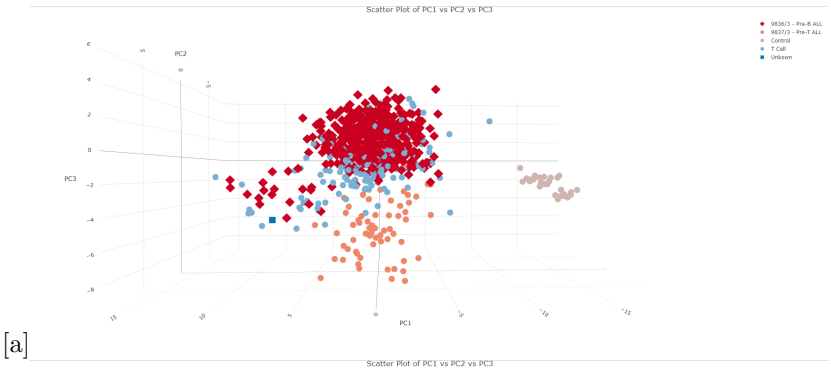
We also wanted to see if our data, both the complete ALL DEGs and the human-specific DEGs ALL, were able to be stratified based on the tumor subtypes and the age of the patients. By looking at Figure 8 it seems that based on the first, second, and third Principal Components the data can be clusterized based on the tumor subtypes (which are distinguished by using different colors), which confirm the results we retrieved from the DEGs subtype vs subtypes. The same cannot be said about the age of the patient (which can be distinguished by the symbols of the dots) because in this case seems that the Principal Components are defined in such a way as to not capture this information. Indeed by doing a PCA by utilizing the result data from the DEGs Pediatric vs Adult.



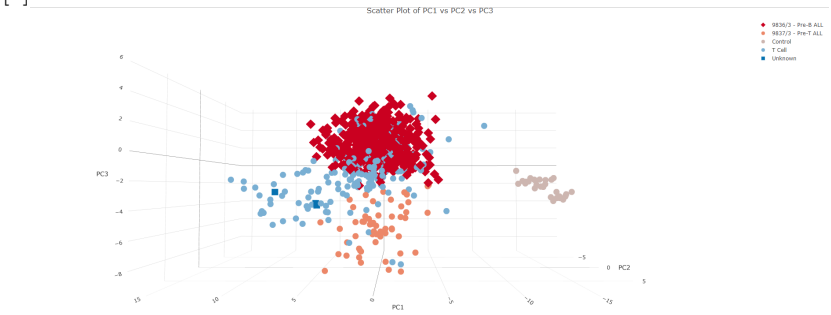
[a] [b]

Figure 8: **Principal Component Analysis 3D**. x-axes: PC1, y-axes: PC2, z-axes: PC3. (a) PCA plot of DEGs of all ALL genes, (b) PCA plot of DEGs of human-specific ALL genes.

Also, after giving the label to the unknown with the utilization of Machine learning classifier data (which will be presented in the next chapters 5.7), the PCA plots, in Figure 9, result in a better stratification of the tumor subtypes based on the Principal Components. Based on the results we can dare to say that human-specific genes are important genes that permit a good Tumor subtype stratification just by considering them.



[a]



[b]

Figure 9: **Principal Component Analysis 3D**. x-axes: PC1, y-axes: PC2, z-axes: PC3. (a) PCA plot of DEGs of all ALL genes, (b) PCA plot of DEGs of human-specific ALL genes. The majority of the unknowns are classified by utilization of machine learning method

5.4 Clustering

Starting from the PCA data generated from the DEG analysis effectuated between tumor and control samples, we operated a clustering analysis. To define the number of clusters in which our samples could be divided, we used the average silhouette graph and the gap statistics graph, as explained in the pipeline section. Thanks to these measurements we could identify that the best number of clusters for the complete dataset was equal to 9, while for the only-tumor dataset was equal to 8, for both the only-human-specific genes and the complete gene set.

- PAM clustering: The samples were more or less homogeneously divided into the different clusters. By plotting the results in the PCA space, we found that the obtained clusters were such that we could not stratify our samples in the three dimensions (see Appendix B - Figure 20)
- Hierarchical clustering: As said above, the samples were more or less homogeneously divided into the different clusters. Also, in this case, the obtained clusters were such that we could not stratify our samples in the three dimensions of the PCA space.

We think that these results could be due to the small number of data points, or also, the possibility that we are using the principal components obtained using the DEG between tumor and control. One possible improvement to the clustering could be to repeat the analysis using principal components calculated specifically on the different tumor subtypes.

5.5 Enrichment and pathway analysis

5.5.1 Tumor - Control

We effectuated the enrichment and pathway analysis on both the only human-specific DEGs and the complete set of DEGs.

- All DEGs.
  From the Gene Ontology enrichment effectuated using the biological process sub-ontology we found that, regarding the up-regulated DEGs non-human-specific, the most significative term is *DNA-templated DNA replication*, with a p-value equal to $\sim 10^{-9}$ (Figure 10 (a)). By looking at the GO Tree View we found that this term has as a parent the term *DNA replication*, which is indeed the second term with the highest p-value.

We expected to find these sorts of terms given the fact that we are dealing with tumor samples. Indeed, two hallmarks of cancer are the limitless replicative potential and the insensitivity to anti-growth signals, we are not surprised to see that the up-regulated genes in ALL relate with these terms. The most significative pathway IDs retrieved from WikiPathway were *Pleural mesothelioma (MPM)*, *Cell cycle*, and *VEGFA VEGFR2 signaling*. From literature [32], it is possible to understand the relationship between the MPM and ALL. Specifically, the MPM pathway, which in our analysis presents to me significative with a p-value of $\sim 10^{-08}$, correlates with the deregulation of different tumor-related genes, like p53, genes related to angiogenesis and cell cycle regulation. Also the second ID (*VEGFA-VEGFR2 signaling*) with a gene ratio of 97/1309 and significative p-value of $\sim 10^{-5}$ results as an important pathway for ALL, indeed VEGFA-VEGFR2 pathway is a major pathway for tumors, from the literature [33], it is known that this pathway activates angiogenesis by inducing multiple activities like the increase of endothelial permeability, proliferation of endothelial cells, their survival, sprouting and migration.

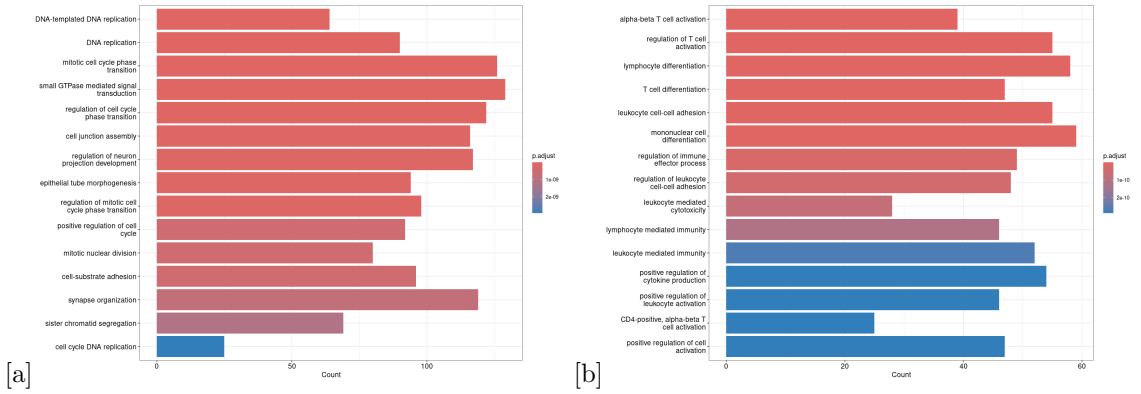


Figure 10: **GO enrichment in all DEGs**. (a) Boxplot representing the first 10 terms ranked by p-value for the up-regulated genes. (b) Boxplot representing the first 10 terms ranked by p-value for the down-regulated genes.

The same can be found for the down-regulated genes non-human-specific. We specifically can see in Figure 10 (b) that the term with the smallest p-value is *alpha-beta T cell activation*. This term has as parents both the second and the third most significative terms, *T cell activation* and *lymphocyte activation*. Alpha-Beta T cells are a class of T cells that express the T receptor alpha/beta. This kind of receptor can link and recognize antigens presented by antigen-presenting cells [34]. ALL is often characterized by chromosomal translocation events and one known hallmark is the translocation of the genes of the T cell receptors [35]. Indeed, in this case, we find that the T cell is characterized by the expression of the T receptors alpha-beta are less activated, this result could be caused by the fact that the T cell receptor could be impaired.

The most significative pathways down-regulated that we were able to retrieve from WikiPathway are: *T cell receptor and co-stimulatory signaling*, *T cell activation SARS CoV 2*. The first ID *T cell receptor and co-stimulatory signaling*, with a p-value of $\sim 10^{-09}$, its down-regulation correlates with our disease ALL. Indeed by searching the literature, we found out that the pathway is necessary for T-cell activities indeed it's important for their proliferation, differentiation, and survival[36]. From the literature, we were able to investigate the *T cell activation SARS CoV 2* pathway [37]. We specifically found that this pathway correlates with the activation of T cells thanks to the interaction of T cell receptors and different molecules like CD45. We believe that the fact that we obtained this pathway down-regulated in ALL could be linked to a more general idea of the inactivation of T cells brought by the tumor, specifically regarding mechanisms of evasion of the immune system.

- Human-specific DEGs.
  From the Gene Ontology enrichment effectuated using the biological process sub-ontology we found that, regarding the up-regulated DEGs human-specific, the most significative term is *NLS-bearing protein input into nucleus* (Figure 11 (a)). It is important to notice that in this analysis, the smallest p-value is around 0.14, we are still going to discuss the obtained results but it is important to remember that the p-value is not under the threshold of 0.05. We think that the large p-value could be because human-specific genes are less studied and there is less knowledge available for them, indeed GO enrichment is based on already known knowledge. Some studies highlighted a possible activation of oncogenes or pluripotency genes derived

from enhanced protein import in the nucleus [38]. Another interesting term found is the *superoxide anion generation*: the accumulation of oxidative stress can lead to degradation of the genome integrity, and this stress condition is frequently observed in cancer cases.

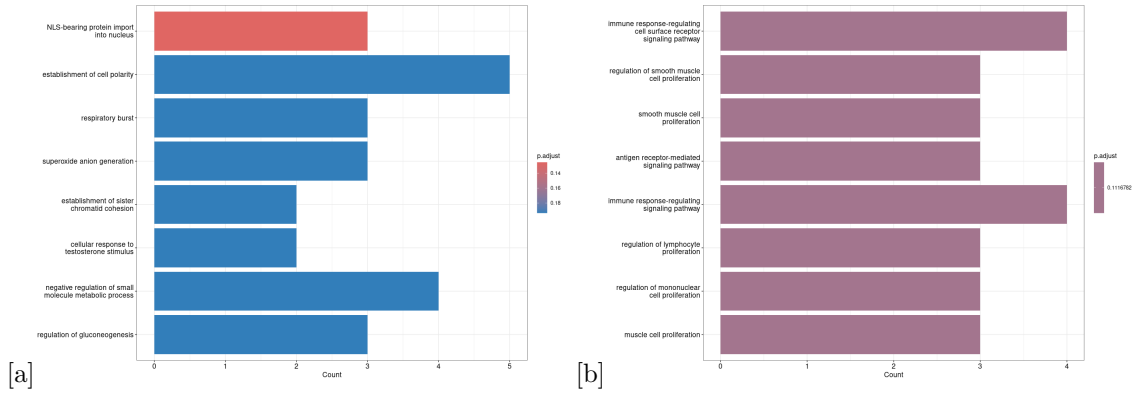


Figure 11: **GO enrichment in human-specific DEGs**. (a) Boxplot representing the first 10 terms ranked by p-value for the up-regulated genes. (b) Boxplot representing the first 10 terms ranked by p-value for the down-regulated genes.

The same information can be found for the down-regulated genes human-specific. We specifically can see in Figure 11 (b) that the term with the smallest p-value is *immune response-regulating cell surface receptor signaling pathway*. This term is a parent of other terms found in the list of the most significative, like, *immune response regulating signaling pathway*. The fact that we find this pathway down-regulated doesn't surprise us, since it is known that an evasive mechanism system operated by ALL is the lack of expression of different immune-activating molecules like CD1 or CD5 expression [39].

For the human-specific DEGs, both up and down-regulated genes, we were not able to retrieve any significative pathway due to the low amount of genes. We did it after the expansion of the data in the following chapter (5.6.3).

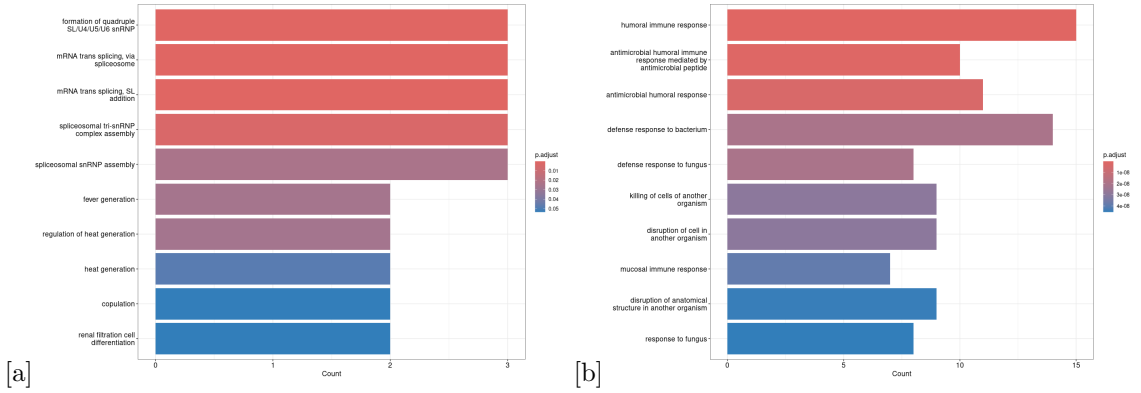
5.5.2 Pediatric - Adult

As done in the case of Tumor vs Control enrichment analysis, we effectuated a GO enrichment analysis and pathway analysis on the genes evidenced as up and down-regulated in pediatric samples.

- All DEGs
  In the up-regulated DEGs non-human-specific, the top 5 significant GO terms all have in common *mRNA splicing, via spliceosome* as parent term (a)). Abnormal regulation of splicing has been observed to accompany the occurrence and development of tumors and in particular the one of hematological malignancy[40]. This upregulation in pediatric cases may have a strong correlation with the major prevalence of ALL in young individuals.
  Regarding the up-regulated pathways, one of the most significant terms is *Small ligand GPCRs* and *GPCRs class A rhodopsin like*: GPCRs belongs to a varied and intricate family of receptors involved in the transduction of numerous pathophysiological activities. Several studies highlighted the role of GPCRs in tumor growth, survival, migration, invasion, and metastasis due to aberrant expression or activation of the receptor. the mechanisms of this involvement of GPCRs in cancer progression are still under investigation.[41]. Also significant is *Thymic stromal lymphopoietin TSLP signaling pathway*. The TSLP pathway promotes lymphocytic proliferation and development affecting immature B-cells and T-cells, thus the overexpression of the pathway could lead to excessive survival and proliferation. Several studies found a strong correlation between this pathway and ALL, in particular in the case of chromosomal translocation or mutation of B-cell development involved genes[42].

  From the Gene Ontology enrichment effectuated using the biological process sub-ontology we found that, regarding the down-regulated DEGs non-human-specific, the most significative term corresponds to humoral immune response with as we can see in Figure 13. This term has as children both the second and the third most significative terms, *anticmicrobial humoral immune response mediated by antimicrobial peptide* and *antimicrobial humoral response*. Is it known that the humoral immune is an adaptive immune response, which

is usually composed of the activation of cytotoxic CD8+ T cells, CD4+ T-cells, and B cells. The immune response is antigene-specific for tumor-associated antigens.[43] We would expect this to be down-regulated in the presence of our disease. The most statistically significative, with corresponding p-values both of order in the $\sim 10^{-4}$, that we were able to retrieve from WikiPathway are: *Complement activation* and *Vitamin D receptor pathway*. The pathway *Complement activation* is down-regulated and correlated with the disease we are analyzing. By searching the literature we were able to find that functions of complement can include the clearance of pathogens and maintenance of homeostasis, and it's possible to be a contributor in the role of anti-tumoral[44]. Which seems right in our case to be down-regulated, we were able to investigate also the *Vitamin D receptor pathway*, which by the literature [45] seems that the low abundance of Vitamin D is usually correlated to a higher risk of contraction of cancer, this because the Vitamin D seem to affect and manage the growth and differentiation, which is usually in a quite growing rate for tumor cells. It seems fair that the Vitamin D pathway is down-regulated in our cancer.



[a] [b]

Figure 12: **GO enrichment in all DEGs pediatric vs adults**. (a) Boxplot representing the first 10 terms ranked by p-value for the up-regulated genes. (b) Boxplot representing the first 10 terms ranked by p-value for the down-regulated genes.

- Human-specific DEGs.
  From the Gene Ontology enrichment effectuated using the biological process sub-ontology we found that, regarding the up-regulated DEGs human-specific, the most significative terms relate with *epithelial cell differentiation* (Figure 11 (a)). From the literature, it was possible to define a correlation between this term and ALL. Specifically, several studies underline the role of the epithelial-mesenchymal transition in the progression of epithelial cancer (like leukemia), drug resistance, and ability to generate metastases [46].
  The only significative pathway ID retrieved from WikiPathway was *Cell lineage map for neuronal differentiation*. It is known that pathways that characterize neural differentiation are the ones linked to MAPK, Wnt/beta-catenin, and Sonic Hedgehog (SHH) pathways. These pathways can be found up-regulated in different types of cancers, specifically hematopoietic malignancies. Deregulation of these pathways correlates with a dysregulated signaling network, and the transformation of healthy hematopoietic stem cells into leukemic stem cells [47].

  The same information can be found for the down-regulated human-specific genes. We specifically can see in Figure 13 (b) that the terms with the smallest p-value are related to *immune response*. These terms correlate to what was found above. Specifically, we know that immune pathways are usually down-regulated in cancer and specifically in leukemia [39].
  Several significative pathway IDs were retrieved from WikiPathway, specifically, we decided to analyze more in-depth the term *Regulatory circuits of the STAT3 signaling pathway*. The STAT3 protein is involved with different cell functions like cell growth, proliferation, migration, and finally, apoptosis. It is also known that STAT3 can activate pathways connected to apoptosis of leukemia cells. Indeed in our analysis, we found that this pathway is under-excessed in ALL.

5.6 Expansion of the Network via OneGene

5.6.1 Interaction network

The first interaction network we generated is shown in Figure 14. As we can see from the image, it is a highly dense network in which the human-specific genes (in orange), can be found at the center of the graph. Given the

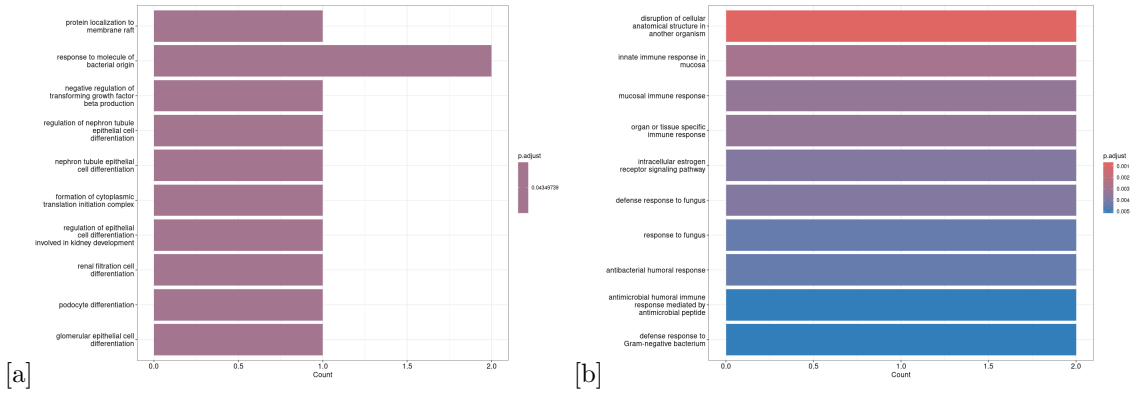


Figure 13: **GO enrichment in human-specific DEGs pediatric vs adult**. (a) Boxplot representing the first 10 terms ranked by p-value for the up-regulated genes. (b) Boxplot representing the first 10 terms ranked by p-value for the down-regulated genes.

complexity of the graph, we decided to reduce the network by sub-sampling only nodes with at least 3 edges. We can see the results in Figure 14. It is now possible to easily distinguish the human-specific genes and the genes found from the expansion (blue nodes).
We finally decided to obtain single networks of human-specific genes, we specifically decided to generate the graphs for the ones "highly" connected. In Figure 14 we have 5 human-specific genes directly connected, in Figure 14 and 14 we have 3 human-specific genes directly connected.

5.6.2 Validation

In the comparison between the OneGene expansion and the STRING, the most interesting results are the p-value of the overlap and the odds ratio. Below the overlap evaluation is reported for LLRC37A, LLRC37A3, ARL17A, and STAG3 genes, the ones highlighted during the interaction network analysis. Each term of the comparison are single gene expansion.

LLRC37A expansion overlap	
Overlapping p-value	5.9e-05
Odds ratio	80.8
Jaccard Index	0.0

ARL17A expansion overlap	
Overlapping p-value	4.7e-08
Odds ratio	220.6
Jaccard Index	0.0

LLRC37A3 expansion overlap	
Overlapping p-value	1.1e-04
Odds ratio	43.7
Jaccard Index	0.0

STAG3 expansion overlap	
Overlapping p-value	0.015
Odds ratio	12.7
Jaccard Index	0.0

As can be seen by the tables for these four genes both p-values and odds ratio confirm the good alignment, but given the difference in the two set dimensions the Jaccard index is not informative.

GRAPL expansion overlap	
Overlapping p-value	1
Odds ratio	0.0
Jaccard Index	0.0

Concerning GRAPL expansion, the overlap scores are not significant. These results may be due to the low number of protein interactions found by STRING expansion and the consequent overlap with the high number of genes in our expansion.

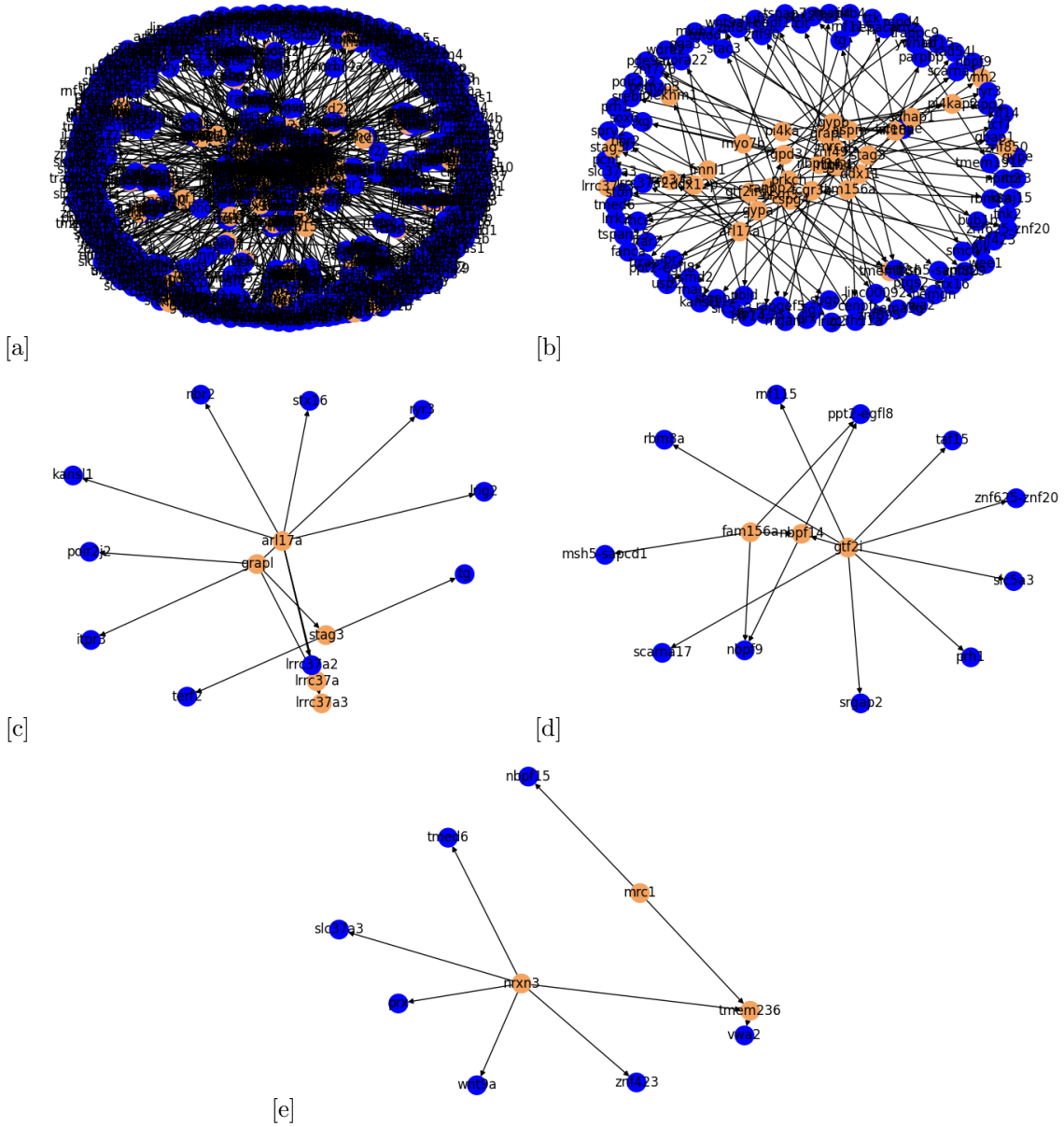


Figure 14: **Graphs of the expansion of the Network via OneGene**. (a) Graph of the network filtered based on the most significative gene correlation with our human-specific genes. (b) Graph simplified of the network filtered based on the most significative gene correlation with our human-specific genes with at least degree 3. (c) Graph that represents 5 human-specific genes correlated between them and other correlated genes. (d) Graph that represents other 3 human-specific genes correlated between them and other correlated genes. (e) Graph that represents other 3 human-specific genes correlated between them and other correlated genes.

5.6.3 Enrichment and pathway analysis - After expansion

Based on the results of the interaction networks, specifically, given the results that there are certain human-specific genes directly connected, we decided to operate an enrichment and pathway analysis on these subsets. It is important to notice, that as happened above, we found enriched GO terms associated with p-values higher than 0.05. Also, in this case, we think that these values could correlate with the fact that the knowledge of human-specific genes is not complete.

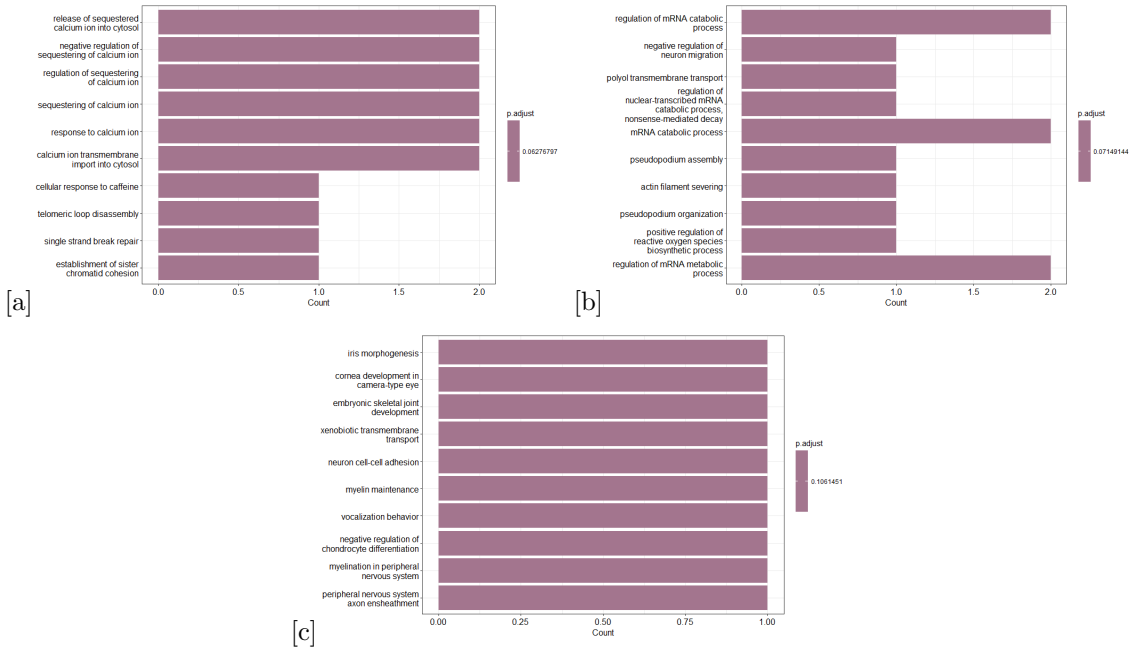


Figure 15: **GO enrichment in human-specific DEGs post-expansion**. (a) Boxplot representing the first 10 terms ranked by p-value for Graph 1 (with 5 human-specific genes). (b) Boxplot representing the first 10 terms ranked by p-value for Graph 2 (with 3 human-specific genes). (c)Boxplot representing the first 10 terms ranked by p-value for Graph 3 (with 3 human-specific genes)

Graph 1 - 5 human-specific genes From the Gene Ontology enrichment analysis, we obtained multiple terms related to the calcium ion. From the literature, it is possible to understand the role of this ion its connection to ALL. Indeed, calcium is involved in regulating cell proliferation and cell cycle progression [48]. From the pathway analysis effectuated by focusing on the biological process, we were able to define significative pathways linked to:

- MFAP5 mediated ovarian cancer cell motility and invasiveness
- Calcium regulation in cardiac cells
- Myometrial relaxation and contraction pathways
- Airway smooth muscle cell contraction

From literature it is possible to find that microfibril-associated protein 5 (MFAP5) is a key element in targeting cancer-associated fibroblasts (CAFs) in leukemia, playing a role as a marker to identify and distinguish subtypes of ALL[49]. The Calcium regulation in cardiac cells pathway leads to leukemia inhibitory factor (LIF) being involved in the cardiac regulation of calcium[50] and a possible agent to alleviate the effects of the deceased. Myometrial relaxation and contraction pathways is a very rare pathway that takes into account the uterine involvement[51], where precursor T-cell from ALL can use the uterus as a possible location of relapsing disease[51]. The last pathway *Airway smooth muscle cell contraction* shows interesting results in the literature, showing the ability of airway smooth muscle cells (ASMC) to stand against ALL as a source of LIF[52]. **Graph 2 - 3 human-specific genes** From the Gene Ontology enrichment analysis, we obtained multiple terms related to the regulation of the

mRNA catabolic process. It is known that mRNA life is influenced by non-coding RNAs and RNA-binding proteins (RBPs). In recent studies it was discovered that three main RBPs are dysregulated in ALL, causing a different regulation of mRNA catabolic process, inducing aberrant cell migration, proliferation, and differentiation [53].
Regarding the pathway analysis, the following pathways have been identified as enriched:

- B cell receptor signaling pathway
- NRF2 pathway

The *B cell receptor signaling pathway* is crucial for B-cell activation, survival, and development. In particular abnormal pathway is strongly related to tumor survival and B-cell aggressiveness, in cases of constitutively activation or gene translocation. This gene is also involved in several treatment resistance[54]. *NRF2* is a transcription factor that enhance the expression of antioxidant response sequence. The protection from oxidative stress may have a protective role in tumor cells, promoting cancer survival[55].

Graph 3 - 3 human-specific genes From the Gene Ontology enrichment analysis, we obtained p-values above 0.1. We think that the obtained terms are such that it is not possible to obtain a good correlation with ALL. We think that these results could be because the genes connected to the graphs are not enough to obtain significant results.
As we did for the other graphs, we found as highly significative pathways the one related to:

- Differentiation of white and brown adipocyte
- Osteoblast differentiation and related diseases
- Disruption of postsynaptic signaling by CNV
- TGF beta receptor signaling

The first pathway *Differentiation of white and brown adipocyte* shows how the adipocytes positioned into the bone marrow is an important factor in ALL. The marrow adipose tissue (MAT) regulation is essential to pinpoint vulnerable processes resulting in malignant transformation that could be exploded by ALL[56]. The next pathway is *Osteoblast differentiation and related diseases* is well known in the ALL context. The osteogenic differentiation of bone marrow mesenchymal stem cells (BMSCs) faces inhibition in the presence of Acute Lymphoblastic Leukemia (ALL) cells[57]. This contributes to the breakdown of the bone marrow microenvironment, adversely affecting its ability to sustain normal hematopoiesis as a result of ALL aggressiveness[57]. The *Disruption of postsynaptic signaling by CNV* brings along the topic of copy number variations (CNVs) into ALL. Literature tells us that CNV "are widespread in both pediatric and adult cases of B-cell acute lymphoblastic leukemia"[58], being a modern and reliable way of detection for the disease. The last pathway found for this graph is *TGF beta receptor signaling* where we find that "the tumor growth factor-beta(TGF-beta)/SMAD signaling pathway is an important mechanism for NK cell immune evasion in childhood B-acute lymphoblastic leukemia (ALL)."[59].

5.7 Machine-learning approaching for classification

The ML methodology employed for categorizing tumor subtypes has demonstrated remarkable efficacy universally. Upon subjecting the algorithms to both not-HS and HS datasets, we observed a consistent pattern in our predictions. This outcome is particularly noteworthy considering the limited data at our disposal. The achievement can be attributed to our thorough data preparation process and our emphasis on precisely fine-tuning hyperparameters based on the characteristics and size of the respective datasets. For each of our datasets, we chose three models from which, after comparing the predictions, drew a consensus on the label to keep. Following the results of our classification.

5.7.1 With all genes

For the dataset with all the genes three models passed the correct control prediction can be seen in Table 1 The models reached a very high consensus, above 80%, as shown in the figure 16. Across all predictions only two samples could reach a consensus and were labeled "Unknown" by us. The resulting classification from the consensus can be seen in table 2. The effect of the new classifications on our PCA can be seen in the with figure 9

Method	F1 Score
RF	0.93
KNN	0.96
XGB	0.98

Table 1: Performance Metrics non-HS

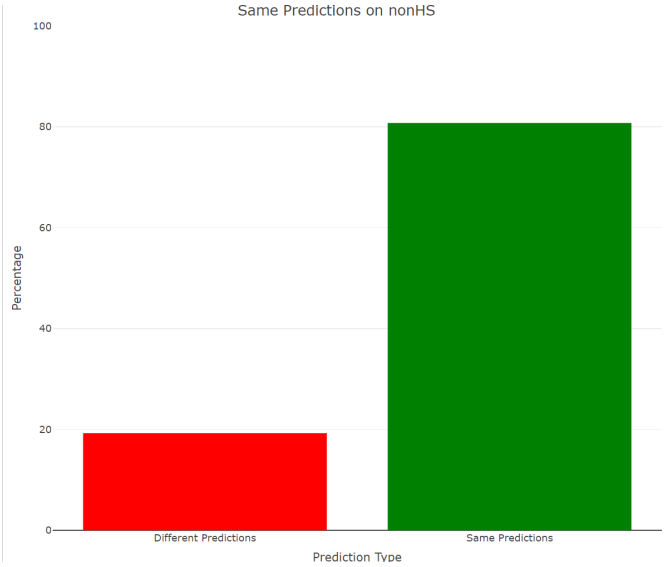


Figure 16: **non-HS consensus**. Barplot of the consensus between the tree models on non-HS.

Subtype	Samples
Pre-B ALL	418
Pre-T ALL	63
T Cell	157
Control	30
Unknown	2

Table 2: Consensus for nonHS

Method	F1 Score
RF	0.97
KNN	0.98
NB	0.73

Table 3: Performance Metrics HS

5.7.2 Only Human-specific genes

For the HS dataset, three models passed the correct control prediction are in table 3. In the HS context, we saw

a good consensus through our predictions despite the small datasets we had in hand for the ML revealing a good classification capacity of around 70% as we can see in figure17. The results of the classification are shown in table 4 below.

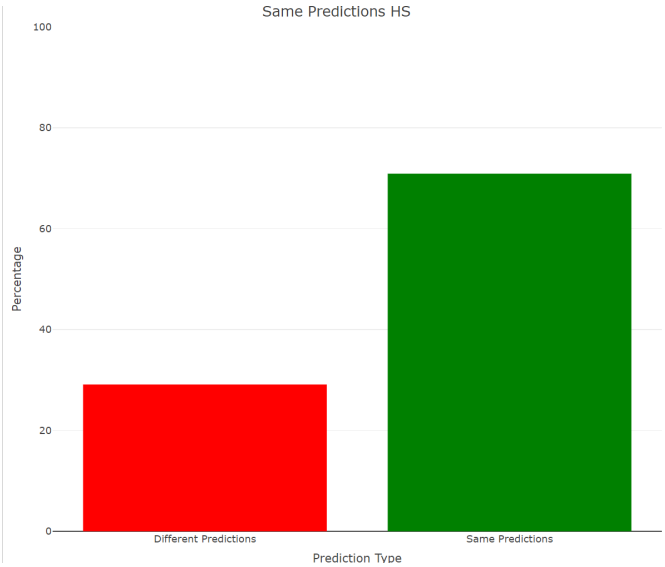


Figure 17: **HS consensus**.Barplot of the consensus between the tree models on HS.

Subtype	Samples
Pre-B ALL	376
Pre-T ALL	56
T Cell	204
Control	30
Unknown	4

Table 4: Consensus for HS

5.8 Identification of treatments or drugs

Thanks to the enrichR tool we obtained an extensive list of potential treatments for each subtype. We decided to keep in consideration the "Combined score" of each drug and select the best five treatments for each subtype in the database. To have a first look at the results, keeping into consideration the score and the gene overlap, we obtained these graphs as shown in figure18.

The DrugMatrix database can link genes to a specific experimental treatment that is correlated with a portion of the genes that are given as input. The terms we retrieved were comprehensive of the drug, the quantity and the experimental condition have been proven. We selected five suggested drugs for each subtype as we can see in figure19 in 7.2. On those, we conducted a validation in the literature on the drugs on table5.

Subtype	Drougs
Pre-B ALL	Clomipramine, Theophylline
Pre-T ALL	Doxorubicin, Mitomycin C
T Cell	Hydroxyurea, Chlorambucil

Table 5: Drugs selected

Starting from Pre-B ALL the drug Clomipramine is a known antidepressant that in recent years has been taken under a program of drug repurposing and in which there are various studies which show that drugs approved for clinical indications other than cancer have shown promising anti-cancer activities[60]. Theophylline has proven its efficacy against numerous diseases, as written in a recent article "theophylline was found to have several immunomodulatory and anti-inflammatory properties...and treatment for leukemias[61]. The drug also showed very promising against breast cancer, showing its ability to induce cytotoxicity together with cell cycle arrest in cancer cells[62]. In the context of Pre-T ALL the drug Doxorubicin is an already well-known treatment for cancer as one of the most popular chemotherapeutic agents. In mice, it has shown the ability to induce splenocardiac cachexia to facilitate defective immuno-metabolism and irreversible macrophage toxicity[63]. In the same context, Mitomycin C is a known treatment against breast cancer, it has shown potential for leukemia due to its effect on bone marrow[64]. In the last subtype, T Cell, the drug Hydroxyurea is an approved treatment for various forms of cancer and

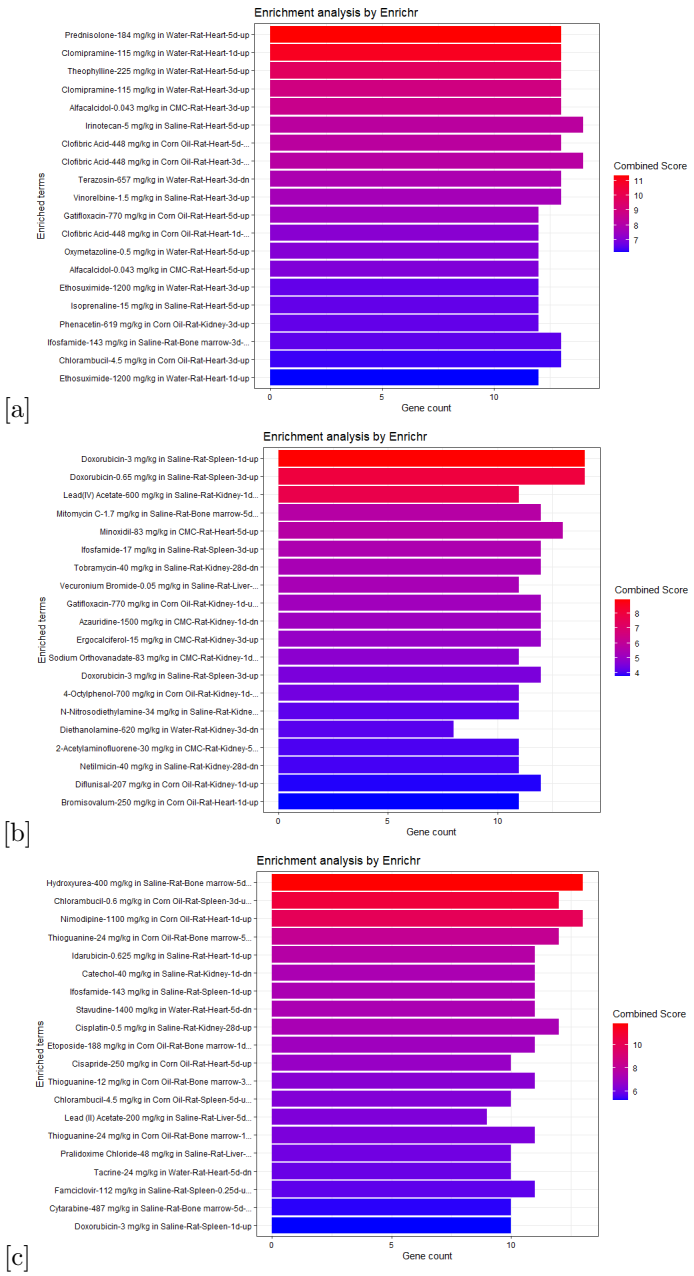


Figure 18: **Drug enrichment**. First 20 terms for the Drug enrichment of Pre-B ALL(a), Pre-T ALL(b) and T Cell(c).

between those the myelogenous leukemia[65] and in animal, the setting has been observed to "decreased circulating leukocytes, erythrocytes, and platelets; decreased cellularity of the thymus, lymph nodes, and bone marrow"[65]. Chlorambucil is the last drug found in T Cell has been taken into consideration as a novel chemotherapeutic treatment in combination with other components. It "has been reported to result in effective anticancer agents with fewer side effects and high therapeutic outcomes"[66].

6 Discussion

Thanks to this project we were able to investigate Acute Lymphoid Leukemia (ALL) in the context of human-specific genes. We initially derived a set of datasets, both from pediatric and adult patients, connected to ALL with corresponding controls. From the expression table, it was possible to define thanks to differential expression analysis, the set of human-specific genes (and not human-specific) that are up and down-regulated in tumors against controls, in pediatric samples against adult samples, and finally in the tumor subtypes. Indeed it was possible to identify human-specific genes (and not human-specific) characterizing three ALL subtypes, specifically, the Pre-B, Pre-T, and T-cell subtypes, which are the ones that we have metadata about. From the differential gene expression analysis we answered an important question, specifically, we wanted to discover if there was a difference in the expression levels between pediatric and adult samples, and thanks to DEG analysis, we defined a gene set of up and down-regulated genes that characterize the pediatric samples. Also from an enrichment analysis, it was found that these genes correlate with biological processes connected to immune response deregulation, differentiation deregulation, and splicing deregulation, as we expected since all these terms correlate with cancer. It was also possible to define the mechanism of origin of the human-specific genes that were found up and down-regulated in tumor samples. The majority of these were generated from events of amplification (copy number alteration), as we knew from literature [1].
From the PCA effectuated on the up and down-regulated genes characterizing tumor samples, it was possible to discover that the first three components are such to stratify the tumor subtypes. Thanks to this data we decided to generate a classifier able to associate a specific tumor subtype based on the gene expression. What is interesting in this case is that the model generated using human-specific DEGs has a high accuracy and is comparable in results to the model obtained by using all DEGs. From the upregulated genes characterizing the three subtypes, we performed an enrichment analysis from which it was possible to define a treatment/drug for a specific subtype. Indeed, for each subtype a different drug is defined, associated with a high significance level. We finally applied a gene expansion to further our knowledge related to human-specific genes. From this expansion, we were able to generate a network graph underlying the interactions between human-specific genes and new genes to us unknown before. We also succeeded in extracting three representative graphs that correlated with cell proliferation, mRNA regulation, and calcium regulation pathways. Thanks to this project we were able to enhance our programming and time managing skills. We were also able to obtain new insight into human-specific genes in the specific context of Acute Lymphoid Leukemia. We know that further analysis and studies are needed, specifically because in this project we were not able to obtain datasets from patients characterized by the B-cell subtype.

6.0.1 Limitations

This project has some main limitations. To start, we weren't able to retrieve datasets from patients with important subtypes, like the B-cell subtype, or by taking into consideration the presence or absence of the Philadephia chromosome (BCR-ABL fusion) which is information that stratifies even deeper the subtypes. This is a major drawback since we left out of the analysis the most common subtype among ALL patients. We still think that our results are significant but we know that the obtained knowledge refers only to a smaller percentage of patients.
During the project, we decided to use machine learning approaches to try to define a classifier able to stratify the patients into specific subtypes using their genetic expression. We know that a major problem connected with our method is the low number of samples that were used to operate the analysis.
In the validation of the gene expansion set, the STRING database was used as a comparison term. This database may not be the best to use because it's about protein interaction rather than regulatory gene networks. We had a problem related to the size difference in terms of the overlap. Another type of database or literature research may be recommendable, but human-specific genes are not widely studied, which can impair the significance of the literature confirmation.

6.0.2 Future Prospectives

We first of all think that, if possible, it would be interesting to operate the analysis effectuated given datasets on B-cell subtype patients. Thanks to this information, we believe it will be possible for us to get further insight into the characterization of human-specific genes correlated to ALL We specifically believe that it will be possible to retrieve a set of human-specific genes characterizing specifically the B-cell subtype, as we found in this analysis for the other subtypes. The same can also be applied to the machine learning classifier, thanks to the new knowledge brought by the B subtype, we believe it will be possible to amplify the quality and significance of the classifier.
After we obtained the machine learning results and almost all our samples were labeled with a specific tumor subtype, we thought of repeating the DEG pipeline and comparing the new results with the one obtained from the

previous analysis. We specifically think that by using the new label data, it will be possible to get better results in terms of significance and hopefully, retrieve new information related to the human-specific genes connected to ALL. During this project we mainly focused on the comparison between tumor samples and control ones. We think that further analysis of the differences (and similarities) between the different ALL subtypes should be performed, the same can be said about the characterization of pediatric and adult samples.
Finally, we think that it could be interesting to operate an enrichment analysis (Gene Ontology and pathway analysis) on the differentially expressed genes associated with the different subtypes. We think that thanks to this analysis it will be possible to associate a specific biological process characterizing the differences observed in the subtypes.

7 Appendix

7.1 Appendix - A

All the code can be found in the GitHub repository Laboratory of Biological Data Mining.

7.2 Appendix - B

Supplementary figures:
- Figure 19: Drug enrichment table
- Figure 20: PAM clustering

References

1. Et al., B. M. Genes with human-specific features are primarily involved with brain, immune and metabolic evolution. *BMC Bioinformatics* (2019).
2. Aibar, S., Fontanillo, C., Droste, C. & De Las Rivas, J. Functional Gene Networks: R/Bioc package to generate and analyse gene networks derived from functional enrichment and clustering. *Bioinformatics* **31,** 1686–1688. `https://doi.org/10.1093/bioinformatics/btu864` (Jan. 2015).
3. Et al., A. M. Gene ontology: tool for the unification of biology. *The Gene Ontology Consortium. Nat Genet.* (2000).
4. Et al.., K. A. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* (2014).
5. Et al., B. The influence of evolutionary history on human health and disease. *Nat Rev Genet 22* (2021).
6. Inaba H, M. C. Pediatric acute lymphoblastic leukemia. *Haematologica.* (2020).
7. E, C. & T., B. The Gene Expression Omnibus Database. *Methods Mol Biol.* (2016).
8. Cerami E Gao J Dogrusoz U, e. a. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* (2012).
9. NCBI. *PubMed* `https://pubmed.ncbi.nlm.nih.gov/about/`.
10. Szklarczyk D, F. A. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* (2015).
11. Martens, M. & Ammar. WikiPathways: connecting communities. *Nucleic Acids Research* **49,** D613–D621. ISSN: 0305-1048 (Nov. 2020).
12. Et al., C. E. Y. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics vol. 14 128* (2013).
13. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.*
14. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics.*
15. Rodriguez-Esteban R & Jiang, X. Differential gene expression in disease: a comparison between high-throughput studies and the literature. *BMC Med Genomics* (2017).
16. Et al., C. A. A survey of best practices for RNA-seq data analysis. *Genome Biol 17* (2016).
17. T., J. I. & Jorge, C. Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc* (2016).
18. Asnicar, F. *et al. OneGenE: Regulatory Gene Network Expansion via Distributed Volunteer Computing on BOINC* in *2019 27th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)* (2019), 315–322.
19. al., P. S. NES2RA: a tool for grapevine transcriptomic data mining. *The First Annual Meeting of COST Action CA17111 INTEGRAPE 2019 - Data Integration as a key step for future grapevine research, Chania, Crete Greece.* `http://hdl.handle.net/10449/54350` (2019).
20. Kalisch, M. & Bühlmann, P. Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *Journal of Machine Learning Research.* `http://jmlr.org/papers/v8/kalisch07a.html` (2007).
21. Asnicar, F. *et al. TN-Grid and gene@home project: volunteer computing for bioinformatics* in *International Conference on High Performance Computing* (2015). `https://api.semanticscholar.org/CorpusID:10481475`.
22. At al., R. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol 11* (2010).
23. Shen, L. *GeneOverlap: An R package to test and visualize gene overlaps* in (2016). `https://api.semanticscholar.org/CorpusID:16979984`.
24. Aric A. Hagberg, D. A. S. & Swart, P. J. Exploring network structure, dynamics, and function using NetworkX. *in Proceedings of the 7th Python in Science Conference (SciPy2008)* (2008).
25. RC, D. Machine Learning in Medicine. *Circulation* (2015).
26. Hartigan, J. Statistical Clustering, International. *Encyclopedia of the Social and Behavioral Sciences* (2001).

27. Erich Schubert, P. J. R. Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. *SISAP* (2019).

28. Wu, Z., Zhang, J. & Hu, S. *Review on Classification Algorithm and Evaluation System of Machine Learning* in *2020 13th International Conference on Intelligent Computation Technology and Automation (ICICTA)* (2020), 214–218.

29. Sarker, I. H. Machine learning: Algorithms, real-world applications and research directions. *SN computer science* **2,** 160 (2021).

30. Chen, T. & Guestrin, C. *Xgboost: A Scalable Tree Boosting System* in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Aug. 2016), 785–794.

31. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research* **44,** W90–W97 (2016).

32. Et al., B. R. *Malignant pleural mesothelioma and mesothelial hyperplasia. A new molecular tool for the differential diagnosis* (Oncotarget, 2017).

33. Claesson-Welsh, L. & Welsh, M. VEGFA and tumour angiogenesis. *Journal of Internal Medicine.* `https://doi.org/10.1111/joim.12019` (6December 2012).

34. T. Leichner, T. K. White Blood Cells and Lymphoid Tissue. *Reference Module in Biomedical Sciences* (2014).

35. Et al, H. R. Translocations and rearrangments in T cell acute leukemia with t(11,14) chromosomal translocations. *Oncogene* (1989).

36. Chen, L. & Flies, D. B. Molecular mechanisms of T cell co-stimulation and co-inhibition. *Nat Rev Immunol.* (2013).

37. Moss, P. The T cell immune response against SARS-CoV-2. *Nat Immunol* (2022).

38. Et al, K. O. Nuclear localization signal in cancer related transcriptional regulator protein NAC1. *Carcinogenesis* (2012).

39. Et al, P. A. Mechanism of immune evasion in acute lymphoblastic leukemia. *Cancers* (2021).

40. Yuanjiao Zhang Jinjun Qian, C. G. Y. Y. Alternative splicing and cancer: a systematic review. *Signal Transduction and Targeted Therapy* (2020).

41. Preeti Kumari Chaudhary, S. K. An Insight into GPCR and G-Proteins as Cancer Drivers. *Cells* (2021).

42. Ziegler, E. L. K. S. F. Thymic Stromal Lymphopoietin and Cancer. *The Journal of Immunology* (2014).

43. Miriam Reuschenbach Magnus von Knebel Doeberitz, N. W. A systematic review of humoral immune responses against tumor antigens. *Cancer Immunp* (2009).

44. Et al, M. R. Complement system: Promoter or suppressor of cancer progression. *antibodies* (2020).

45. Carsten Carlberg, A. M. An update on vitamin D signaling and cancer. *Elsevier* (2022).

46. Chen SC Liao TT, Y. M. Emerging roles of epithelial-mesenchymal transition in hematological malignancies. *J Biomed Sci.* (2018).

47. Et al, C. F. The Role Played by Wint beta-catenin Signaling Pathway in Acute Lymphoblastic Leukemia. *Int J Mol Sci* (2020).

48. Et al, L. C. Put in a CA2 ll to Acute Myeloid Leukemia. *Cells* (2022).

49. Gu, L., Liao, P. & Liu, H. Cancer-associated fibroblasts in acute leukemia. *Frontiers in Oncology* **12,** 1022979 (2022).

50. Murata, M. *et al.* Leukemia Inhibitory Factor, a Potent Cardiac Hypertrophic Cytokine, Enhances L-type Ca2+ Current and [Ca2+] iTransient in Cardiomyocytes. *Journal of molecular and cellular cardiology* **31,** 237–245 (1999).

51. Lyman, M. D. & Neuhauser, T. S. Precursor T-cell acute lymphoblastic leukemia/lymphoma involving the uterine cervix, myometrium, endometrium, and appendix. *Annals of Diagnostic Pathology* **6,** 125–128 (2002).

52. Fayon, M. *et al.* Increased secretion of leukemia inhibitory factor by immature airway smooth muscle cells enhances intracellular signaling and airway contractility. *American Journal of Physiology-Lung Cellular and Molecular Physiology* **291,** L244–L251 (2006).

53. Et al, S. K. RNA-Binding Proteins in Acute Leukemias. *Int J Mol Sci* (2020).

54. Et al, N. P. P. Regulation of B Cell Receptor Signaling and Its Therapeutic Relevance in Aggressive B Cell Lymphomas. *Cancers* (2022).

55. Zimta, A.-A. *et al.* The role of Nrf2 activity in cancer development and progression. *Cancers* (2019).

56. Zinngrebe, J., Debatin, K.-M. & Fischer-Posovszky, P. Adipocytes in hematopoiesis and acute leukemia: friends, enemies, or innocent bystanders? *Leukemia* **34,** 2305–2316 (2020).

57. Yang, G.-C., Xu, Y.-H., Chen, H.-X., Wang, X.-J., *et al.* Acute lymphoblastic leukemia cells inhibit the differentiation of bone mesenchymal stem cells into osteoblasts in vitro by activating notch signaling. *Stem Cells International* **2015** (2015).

58. Song, Y., Fang, Q. & Mi, Y. Prognostic significance of copy number variation in B-cell acute lymphoblastic leukemia. *Frontiers in Oncology* **12,** 981036 (2022).

59. Rouce, R. H. *et al.* The TGF-β/SMAD pathway is an important mechanism for NK cell immune evasion in childhood B-acute lymphoblastic leukemia. *Leukemia* **30,** 800–811 (2016).

60. Kumari, P. & Dang, S. Anti-cancer potential of some commonly used drugs. *Current Pharmaceutical Design* **27,** 4530–4538 (2021).

61. Vassallo, R. & Lipsky, J. J. *Theophylline: recent advances in the understanding of its mode of action and uses in clinical practice* in *Mayo Clinic Proceedings* **73** (1998), 346–354.

62. Tapadar, P., Pal, A., Dutta, S. & Pal, R. Enhanced expression of death receptor 5 is responsible for increased cytotoxicity of theophylline in combination with recombinant human TRAIL in MDA-MB-231 breast cancer cells. *Journal of Cancer Research and Therapeutics* **18,** 754–759 (2022).

63. Jadapalli, J. K. *et al.* Doxorubicin triggers splenic contraction and irreversible dysregulation of COX and LOX that alters the inflammation-resolution program in the myocardium. *American Journal of Physiology-Heart and Circulatory Physiology* **315,** H1091–H1100 (2018).

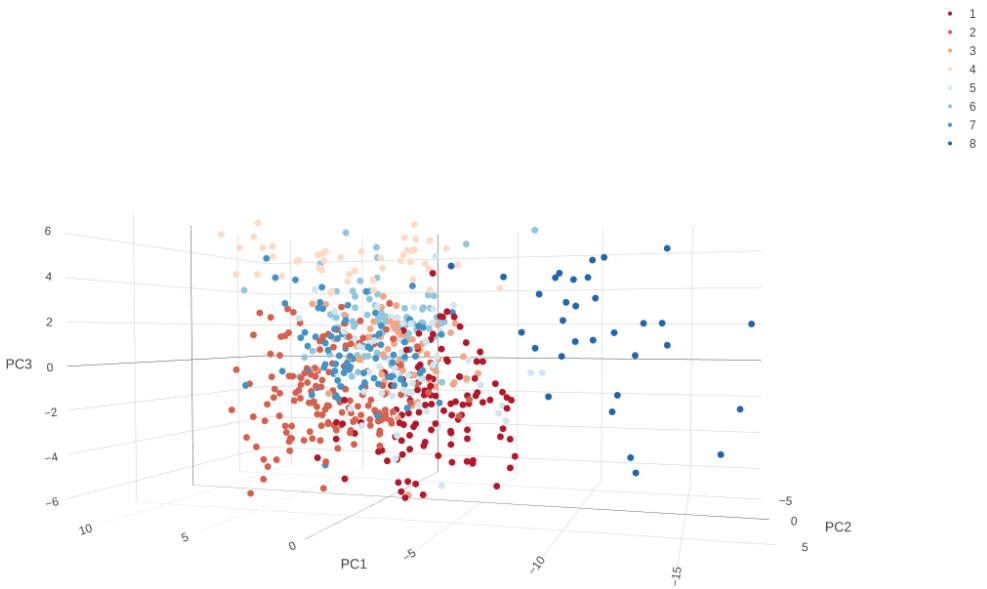
64. Varshosaz, J., Sarrami, N., Aghaei, M., Aliomrani, M. & Azizi, R. LHRH targeted chonderosomes of mitomycin C in breast cancer: an in vitro/in vivo study. *Anti-Cancer Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Anti-Cancer Agents)* **19,** 1405–1417 (2019).

65. Morton, D. *et al.* Toxicity of hydroxyurea in rats and dogs. *Toxicologic pathology* **43,** 498–512 (2015).

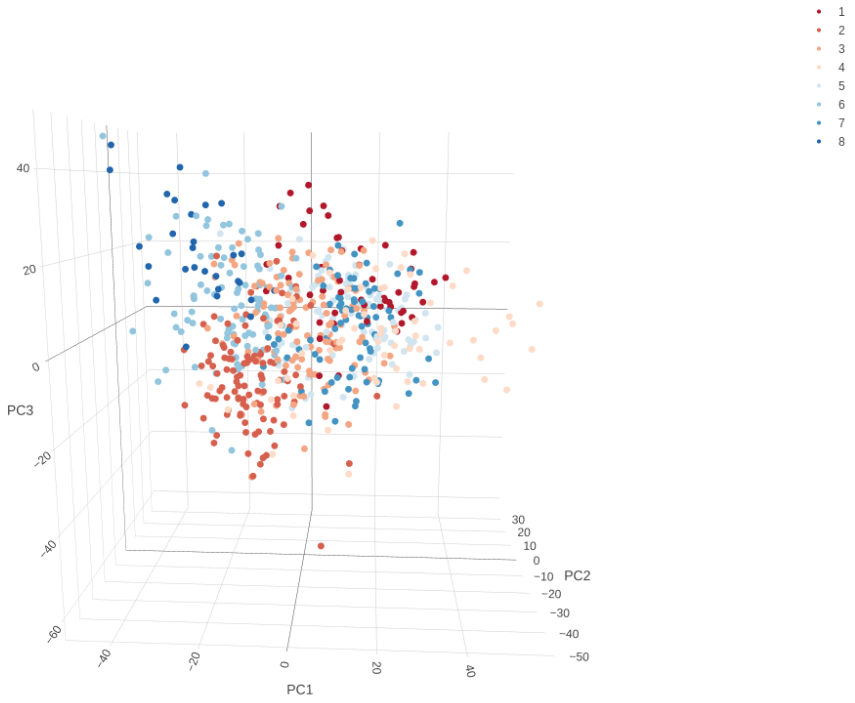
66. Peter, S. & Aderibigbe, B. A. Chlorambucil-Bearing Hybrid Molecules in the Development of Potential Anti-cancer Agents. *Molecules* **28,** 6889 (2023).

Term	Combined Score	Genes
Pre B ALL		
Prednisolone-184 mg/kg in Water-Rat-Heart-5d-up	11.35067929	CLIC5;ZCCHC7;GLDC;CHD7;SOCS2;KLF9;PDK4;SERPINI2;PDE4B;TCF4;SCN4A;KCNK3;RAI14
Clomipramine-115 mg/kg in Water-Rat-Heart-1d-up	11.00082375	ZCCHC7;GLDC;CHD7;SOCS2;EDNRB;KLF9;PDK4;SERPINI2;PDE4B;TCF4;SCN4A;KCNK3;RAI14
Theophylline-225 mg/kg in Water-Rat-Heart-5d-up	9.720564989	CDKN1A;GLDC;HTRA3;KCNA5;GPC1;KLF9;PDK4;SERPINI2;CSPG4;TCF4;SCN4A;KCNK3;RAI14
Clomipramine-115 mg/kg in Water-Rat-Heart-3d-up	8.961926643	CLIC5;ZCCHC7;CDKN1A;GLDC;CHD7;SOCS2;KLF9;PDK4;SERPINI2;TCF4;SCN4A;KCNK3;RAI14
Alfacalcidol-0.043 mg/kg in CMC-Rat-Heart-3d-up	8.69497075	CDKN1A;CHD7;PDLIM1;EDNRB;COL4A1;KLF9;PDK4;SERPINI2;PDE4B;TCF4;SCN4A;KCNK3;RAI14
Pre T ALL		
Doxorubicin-3 mg/kg in Saline-Rat-Spleen-1d-up	8.940195624	SPON2;ANXA1;TCF7;FGL2;DSTN;HMGA2;ETS1;GJA1;CD8A;LCK;ACP5;PTPN7;LAT;MYO1G
Doxorubicin-0.65 mg/kg in Saline-Rat-Spleen-3d-up	8.030479736	SPON2;ANXA1;DSTN;HMGA2;ETS1;GJA1;SBK1;CD8A;LCK;CD5;ACP5;PTPN7;LAT;MYO1G
Lead(IV) Acetate-600 mg/kg in Saline-Rat-Kidney-1d-dn	7.75373012	DPP4;ENPEP;IGFBP5;DST;FXYD2;EPHX2;TCEA3;MGST1;HMGA2;SLC4A4;XYLB
Mitomycin C-1.7 mg/kg in Saline-Rat-Bone marrow-5d-up	5.730843137	CPA3;SPON2;ANXA1;VWF;KYNU;ALDH1A2;TCF7;FGL2;MGST1;ARHGEF3;PTPN7;LAT
Minoxidil-83 mg/kg in CMC-Rat-Heart-5d-up	5.730636774	ANXA1;IGFBP5;DST;VWF;FGL2;DSTN;ATP1B1;FSTL1;ETS1;FBLN2;GJA1;FXYD2;SCN7A
T Cell ALL		
Hydroxyurea-400 mg/kg in Saline-Rat-Bone marrow-5d-up	11.77946516	EIF4A1;SPON1;FST;EEF1G;HLA-DMB;S100A16;COL4A1;RASSF5;FXYD2;NPY;CCL5;TXNIP;ADAMTS9
Chlorambucil-0.6 mg/kg in Corn Oil-Rat-Spleen-3d-up	10.70211265	EIF4A1;CLIC5;ORM1;HLA-DMB;RAB4B;RASSF5;FXYD2;NPY;EBF1;PRG2;TXNIP;SCN4A
Nimodipine-1100 mg/kg in Corn Oil-Rat-Heart-1d-up	9.931174551	EIF4A1;CLIC5;ORM1;S100A1;NPR1;RASL10B;SCHIP1;COL4A1;P4HA2;TXNIP;SCN7A;BVES;KCNK3
Thioguanine-24 mg/kg in Corn Oil-Rat-Bone marrow-5d-up	8.234050376	EEF1G;EIF4A1;HLA-DMA;HLA-DMB;COL4A1;RASSF5;FXYD2;NPY;PRG2;LCN2;TXNIP;PGLYRP1
Idarubicin-0.625 mg/kg in Saline-Rat-Heart-1d-up	7.703457059	EIF4A1;CLIC5;PHOSPHO1;S100A1;COL4A1;TXNIP;GMPR;SCN7A;BVES;GNAI1;KCNK3

Figure 19: **Drug enrichment**.The best five drugs with a higher Combined score for each subtype.



[a]



[b]

Figure 20: **PAM clustering on tumor-only data** (a) Clustering obtained considering only human-specific genes. It is possible to observe that there are 8 subtypes (b) Clustering obtained considering all genes.