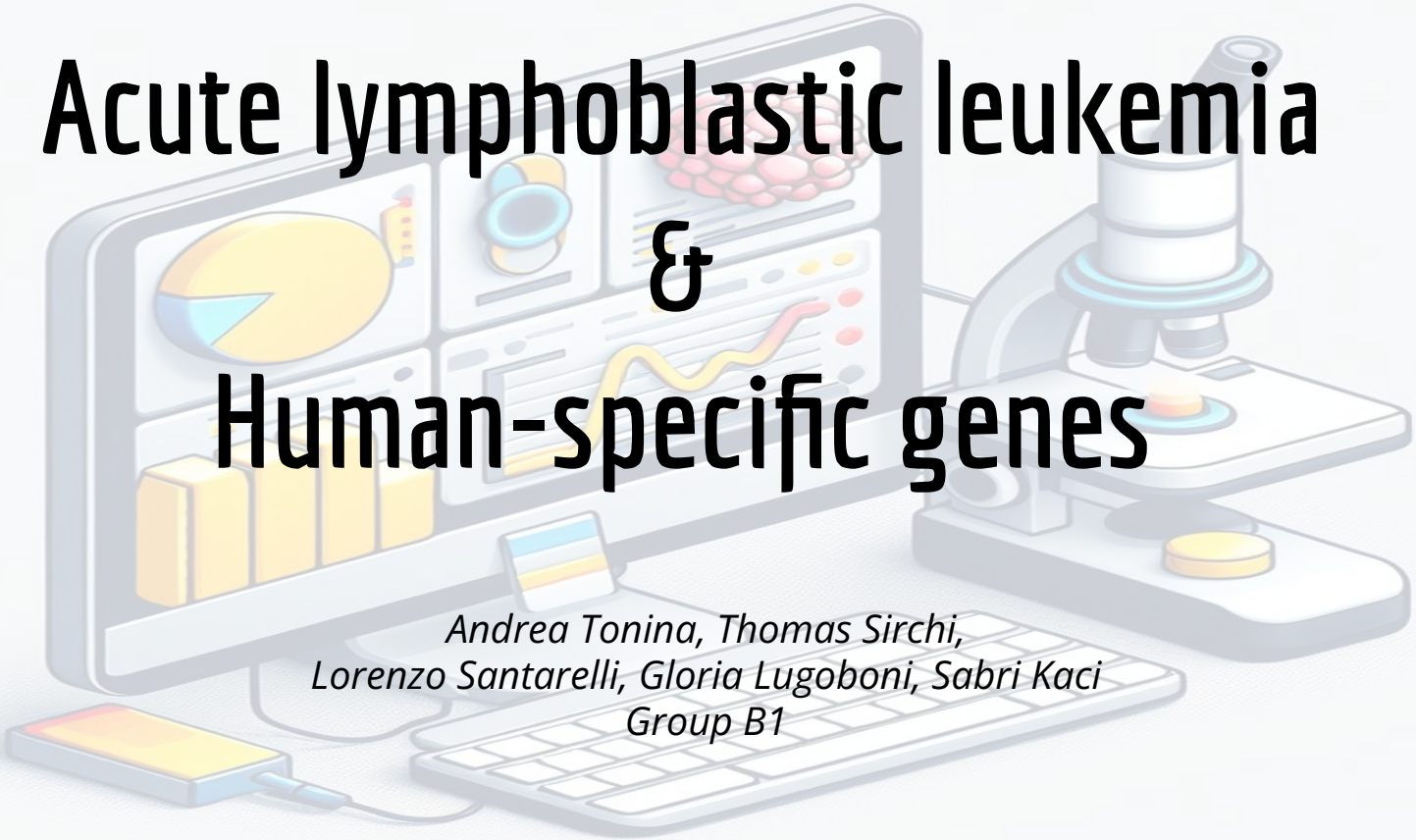
Acute lymphoblastic leukemia & Human-specific genes

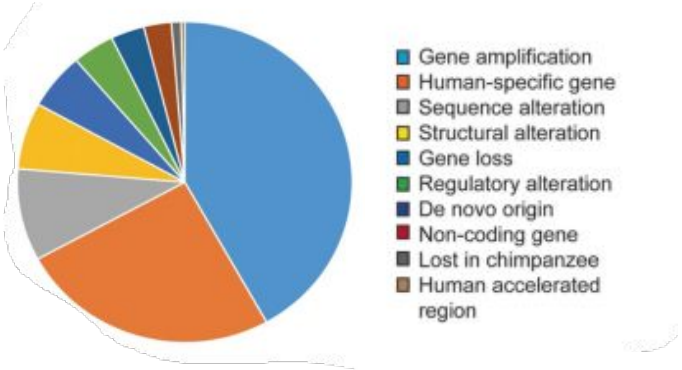


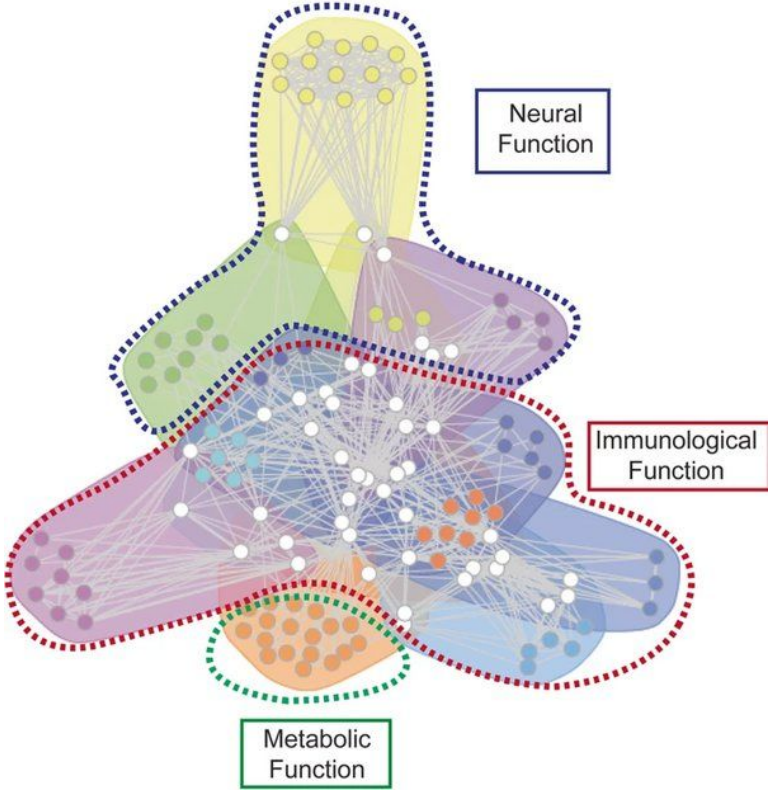
Andrea Tonina, Thomas Sirchi,
Lorenzo Santarelli, Gloria Lugoboni, Sabri Kaci
Group B1

Human Specific genes

Man and chimpanzee separated around 6 million years ago. From that point onwards, rapid evolution and new alterations were acquired, giving rise to orthologous and de-novo genes. It subsequently possible to associate human-specific genes with a restricted set of functions:

- Neural functions
- Metabolic functions
- Immunological functions
- Functions at the cellular level.



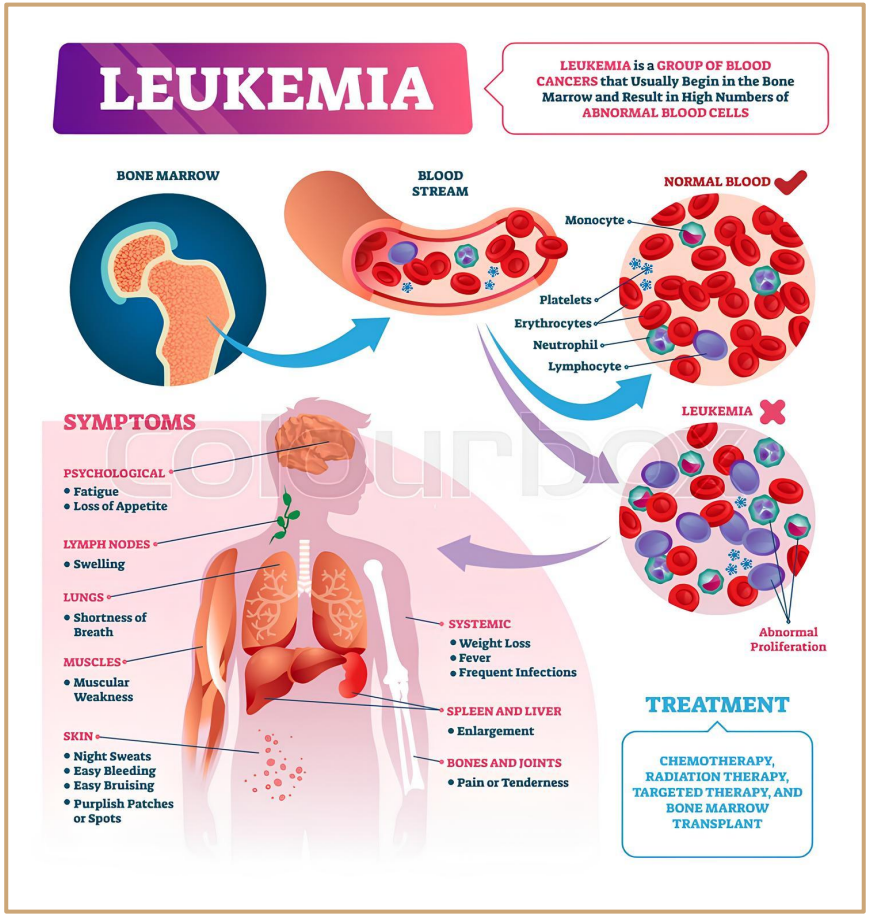


Acute Lymphoblastic Leukemia

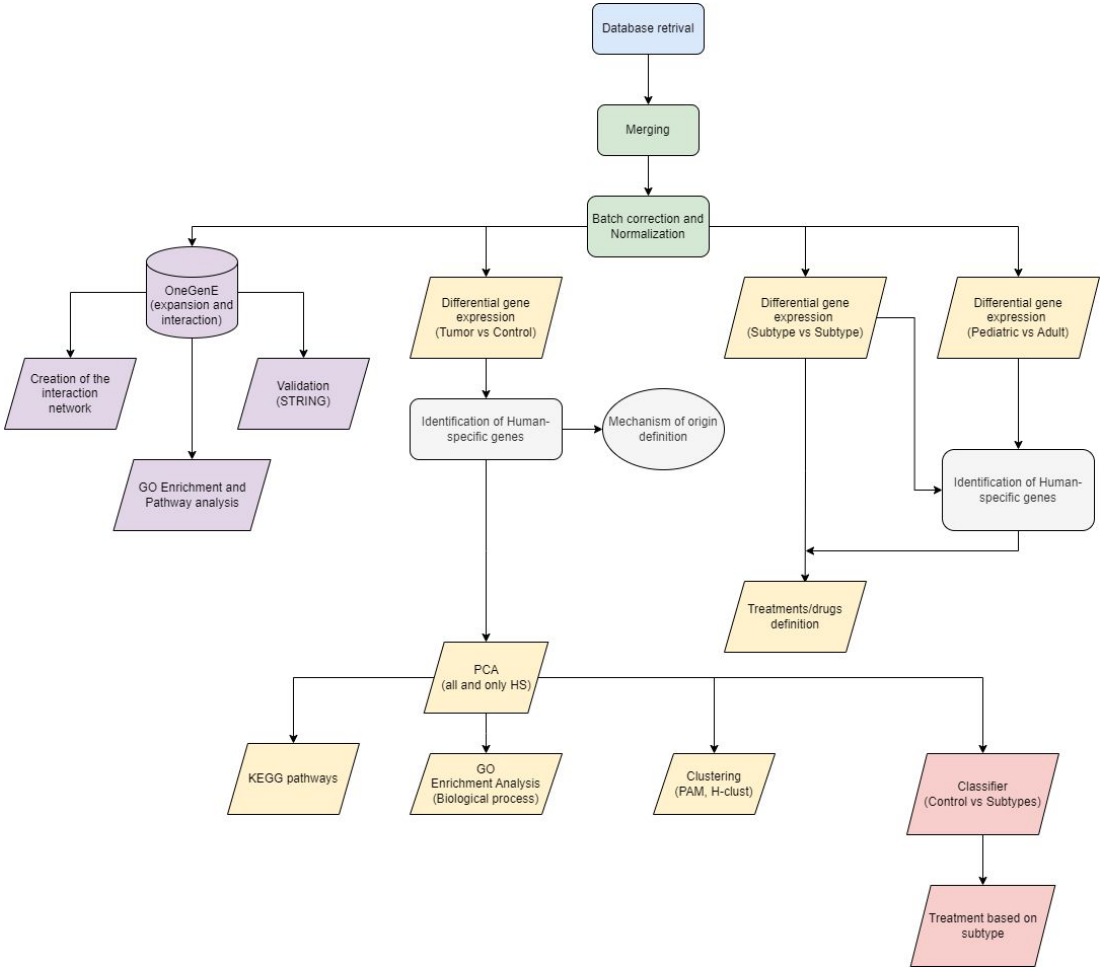
Focus on acute lymphocytic leukemia (ALL), whether pediatric or adult :

- 80% of patients are children.
- Malignant transformation that causes abnormal growth of lymphoid cells.
- Symptoms are generally non-specific

Objective : try to extend our knowledge of human-specific genes regarding Acute Lymphoblastic Leukemia (ALL)



The pipeline



Pre-processing, normalization, and batch effect correction

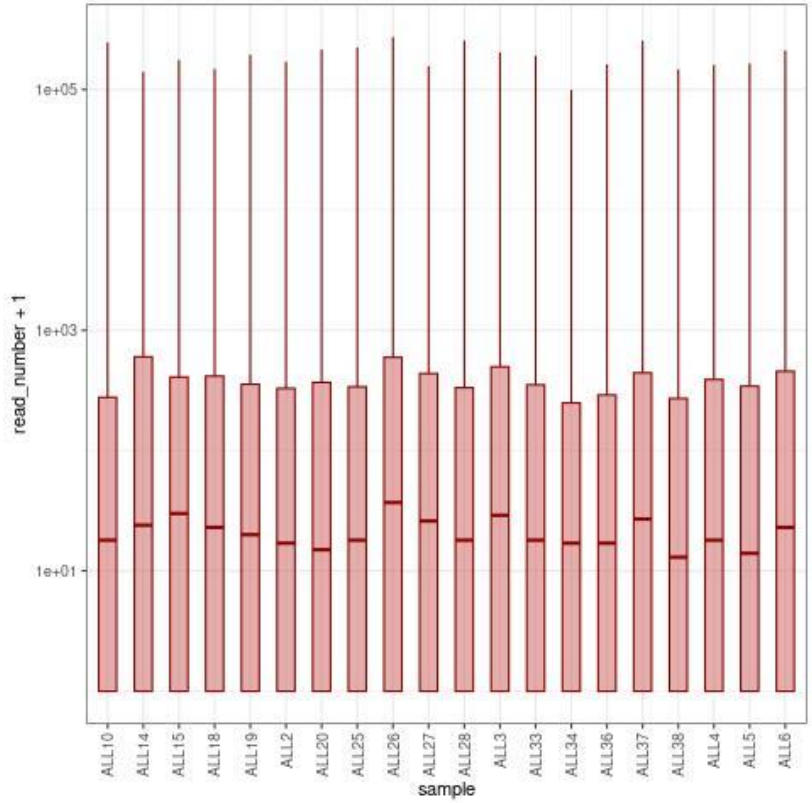
- Pre-processing
- Batch effect correction
- TMM normalization

Pre-processing

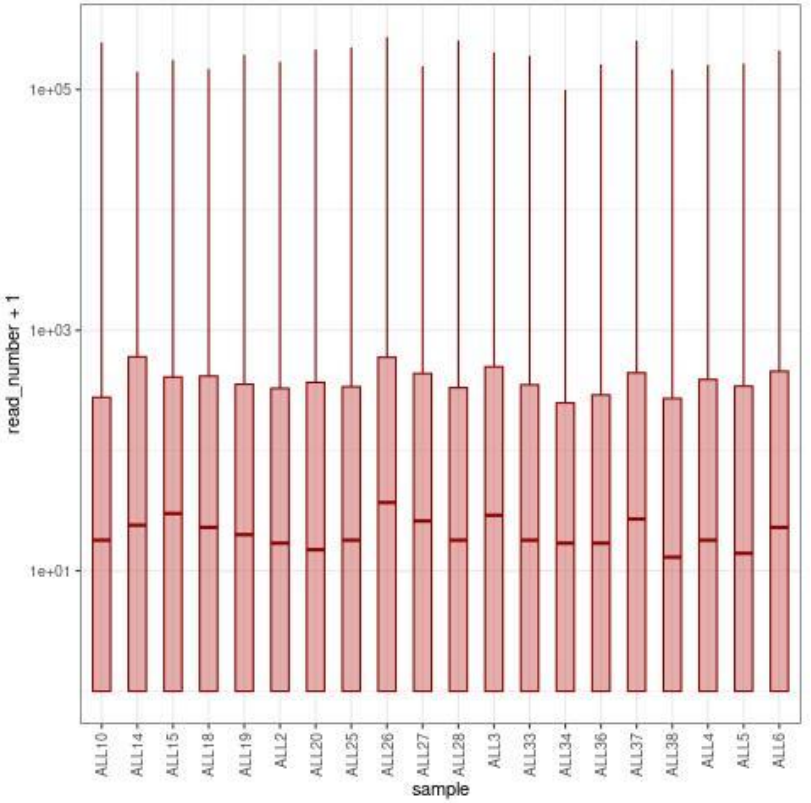
- Hugo Symbol to Ensembl ID
- Filtering for duplicates (less informatives)
- Filtering for low expressed genes

Batch effect correction Tumors

Pre-Combat

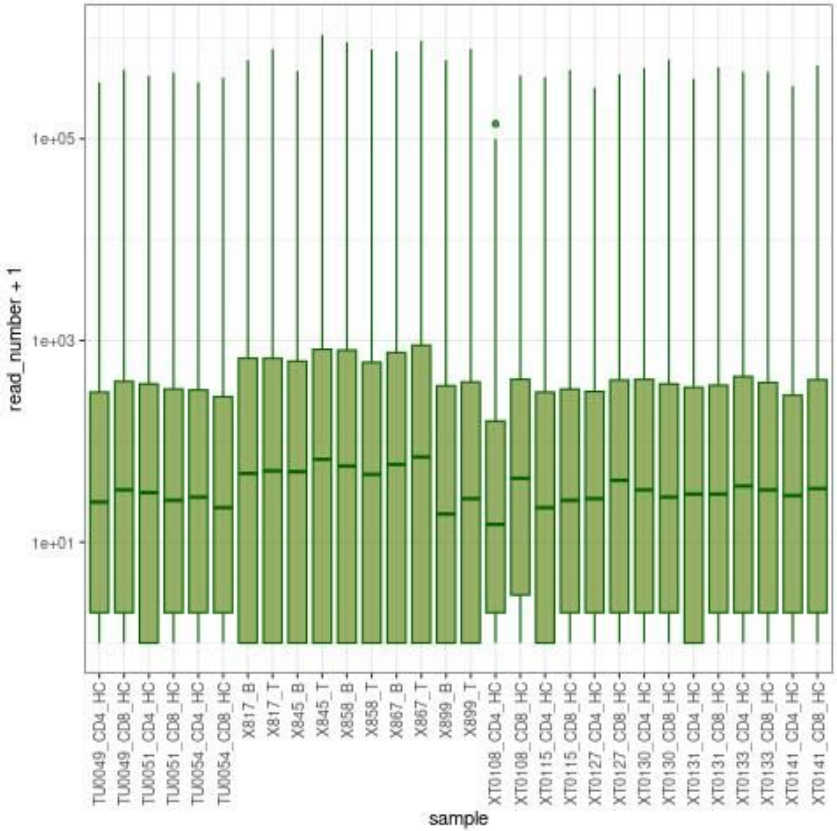


Post-combat

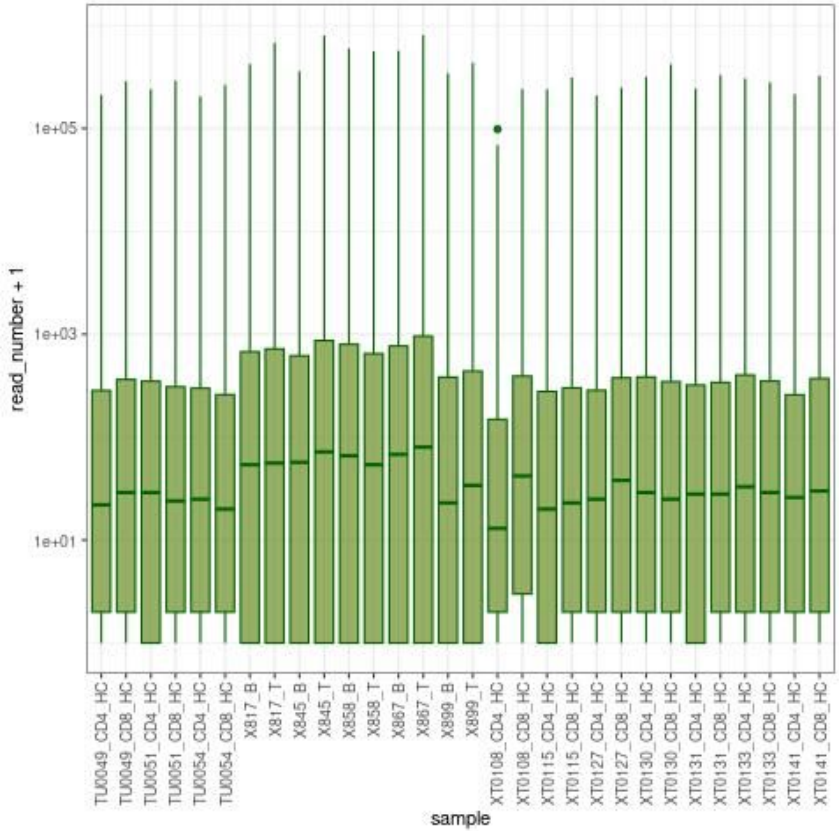


Batch effect correction Controls

Pre-Combat

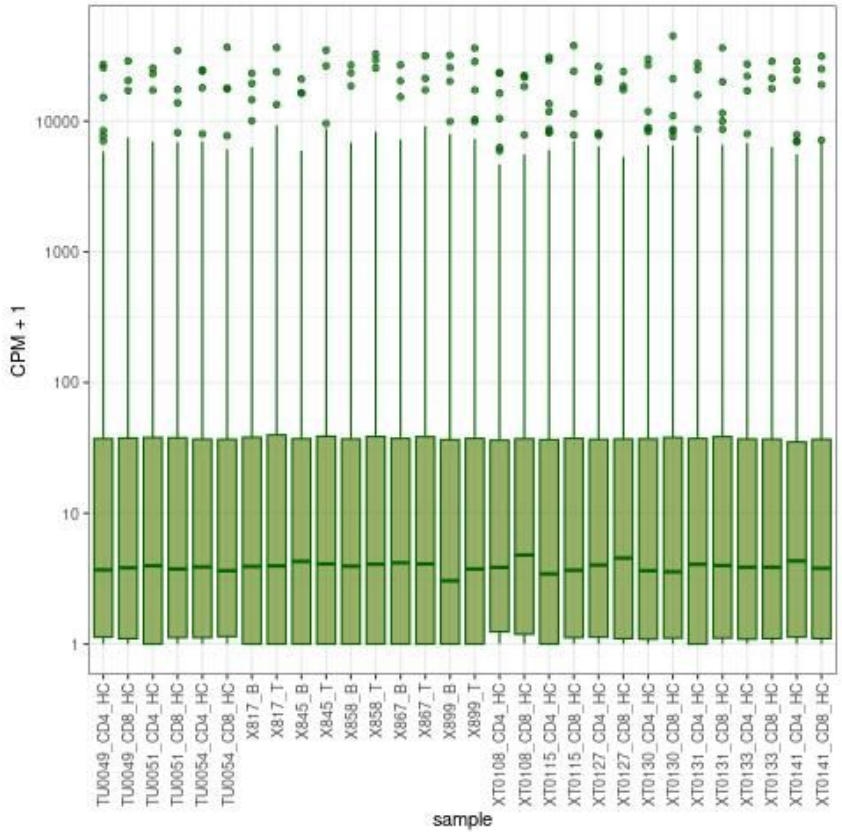


Post-combat

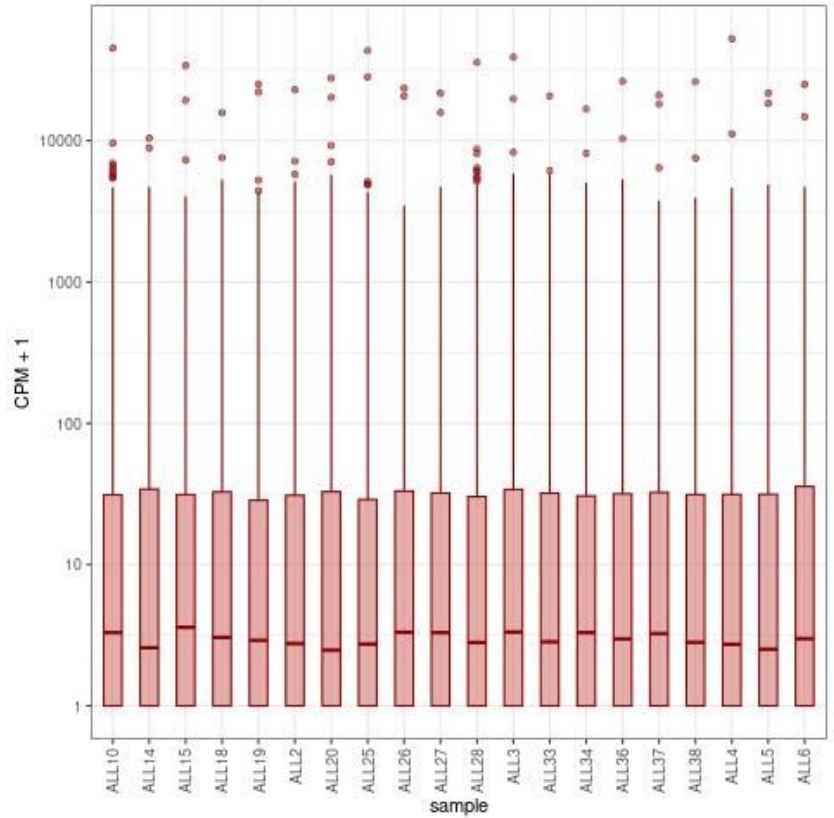


TMM correction

Control



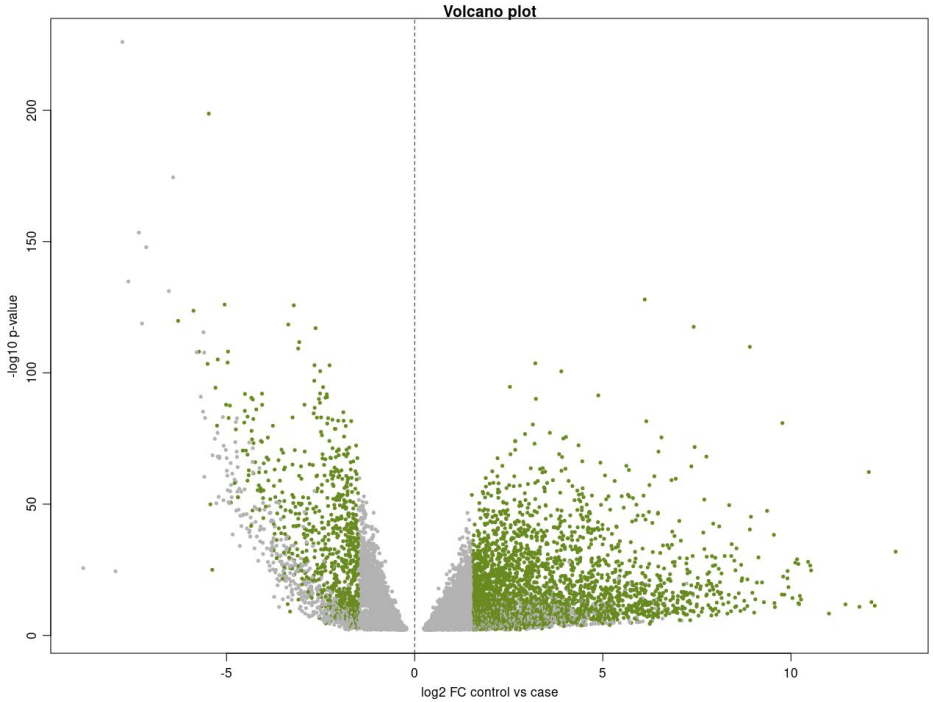
Tumor

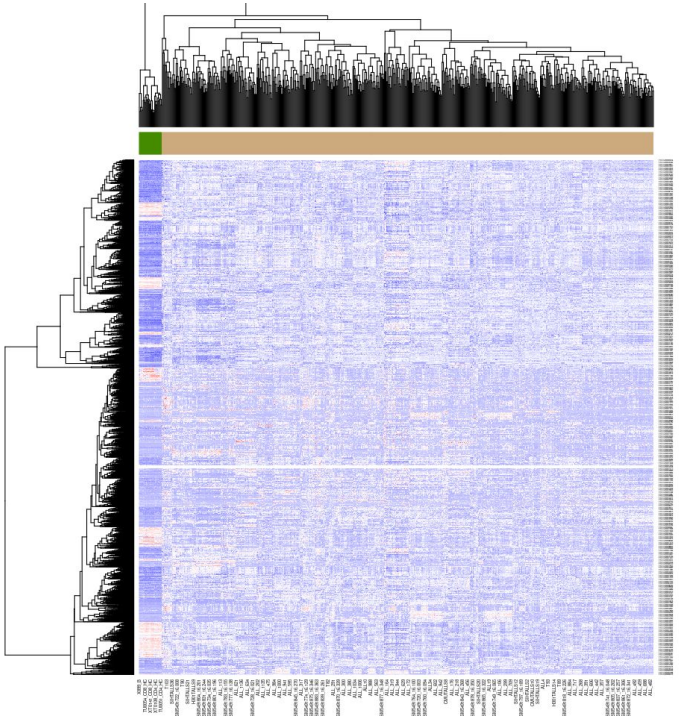


Differential Gene expression

- Tumor vs Control
- Subtype vs Subtype
- Pediatric vs Adult

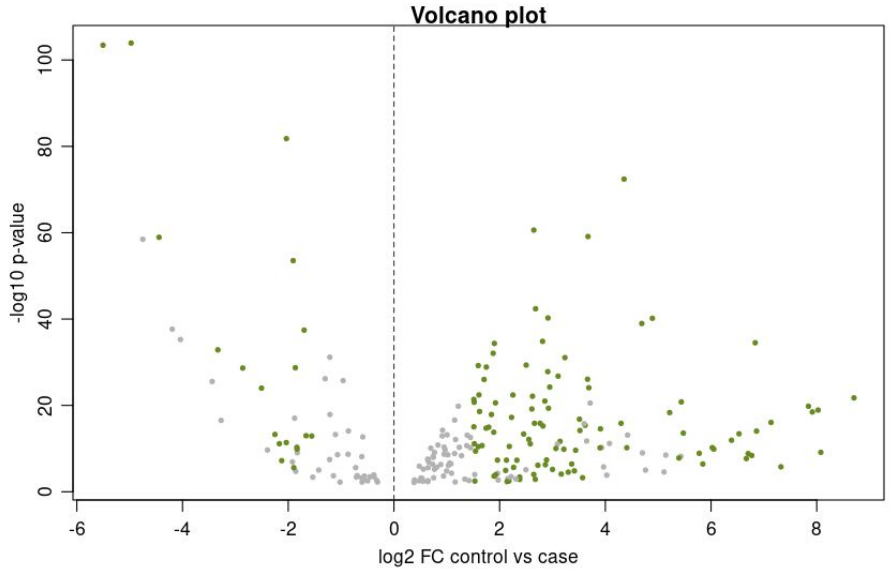
Tumor vs Control

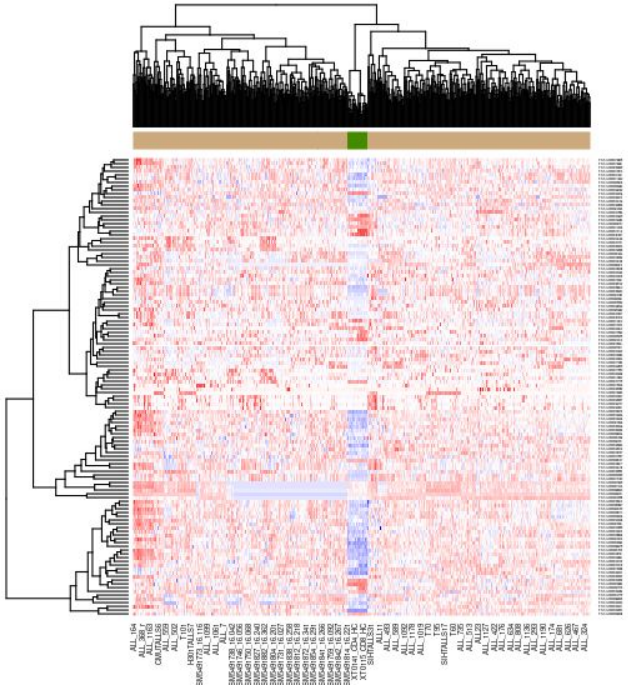




Tumor vs Control

- Human-specific





Subtype vs Subtype

- 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 (PreB has 22 down hs while PreT 23)
- Subtype Pre-T and Subtype T: 2 up-regulated genes in common

Pediatric vs Adult

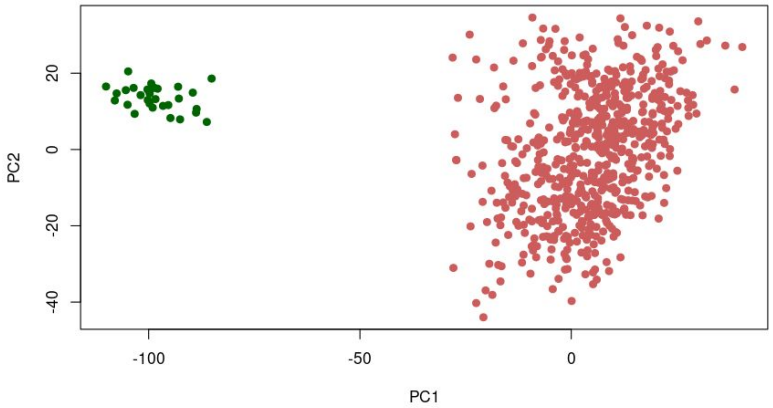
- Pediatric ALL tumors: 55 up-regulated and 108 down-regulated
- Only 10 down-regulated and 6 up-regulated are also human-specific
- Pediatric ALL is genetically different compared to the adult ALL

PCA

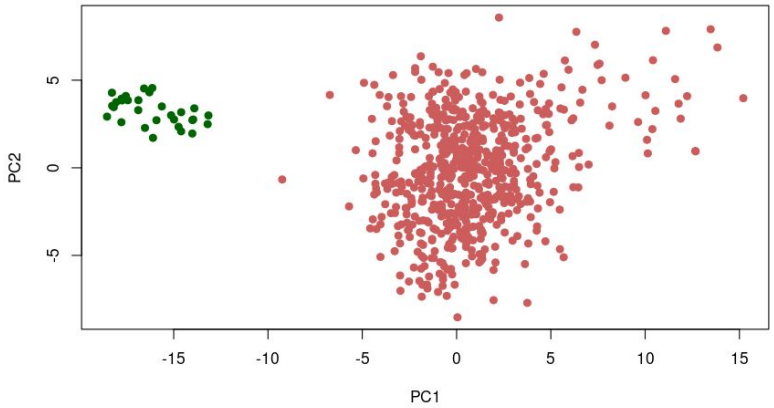
- Tumor - Control
- Subtype - Subtype
- Pediatric - Adult

Tumor - Control

- All

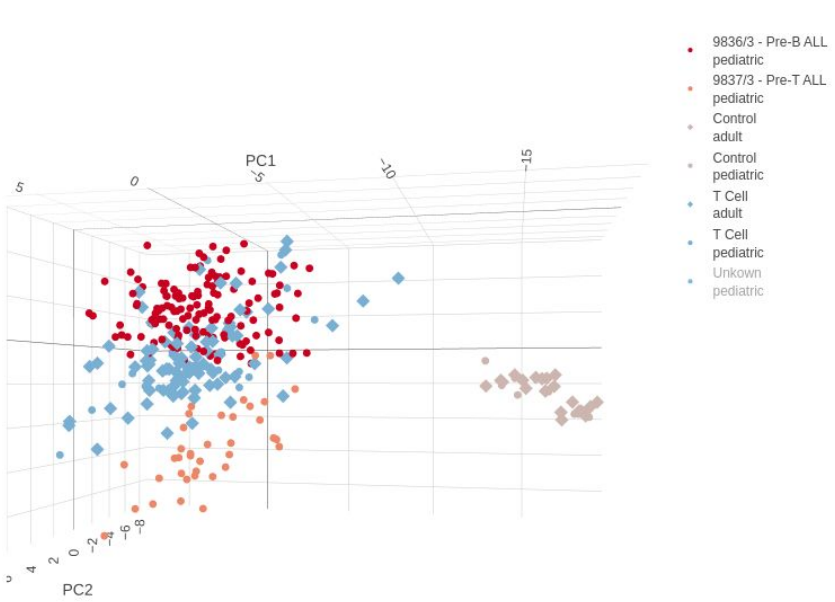


- Human-specific

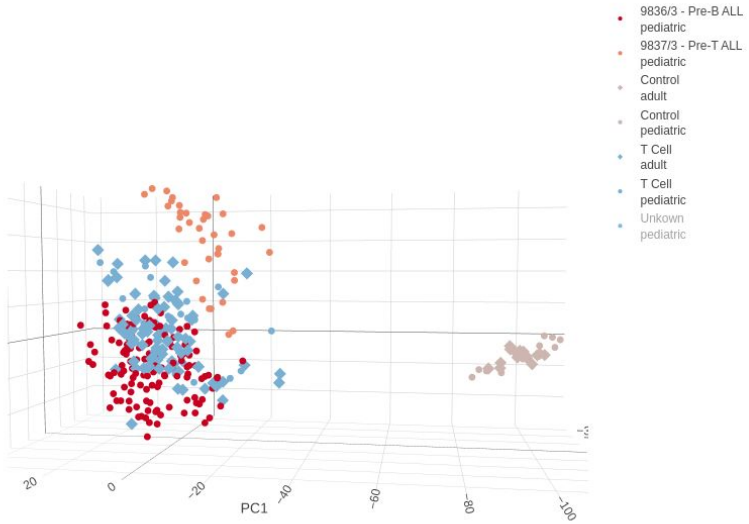


Subtype - Subtype

- All



- Human-specific



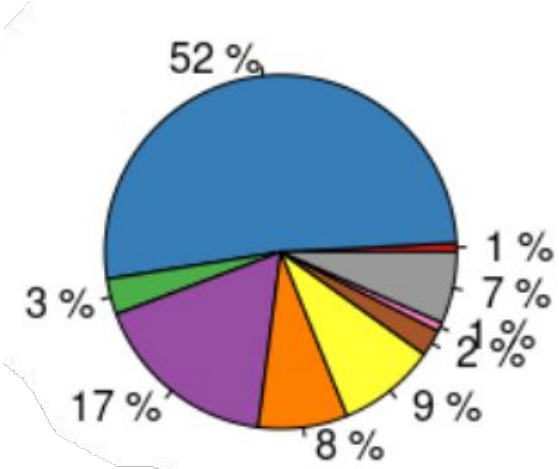
- PCA by utilizing the result data from the DEGs Subtype vs Subtype

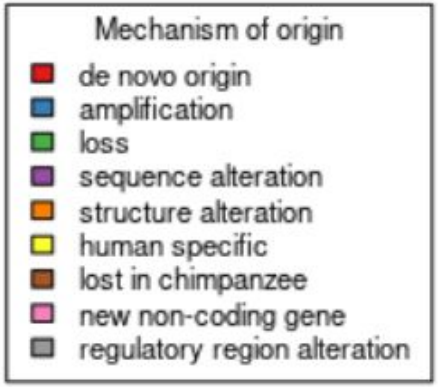
Pediatric - Adult

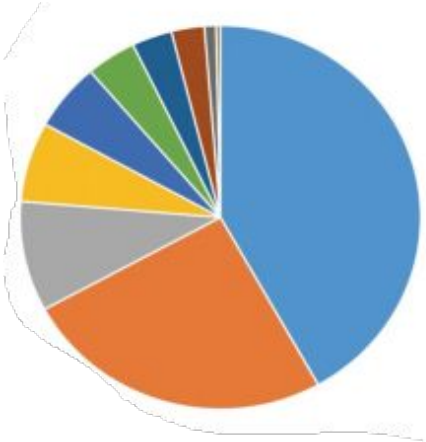
- In this case seems that the Principal Components are defined in such a way as to not capture this information
- PCA by utilizing the result data from the DEGs Pediatric vs Adult

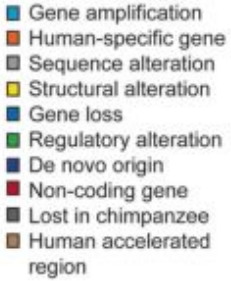
Mechanism of origin Human-specific genes

Literature comparison


52 %
1 %
7 %
1 %
2 %
9 %
8 %
17 %
3 %


Mechanism of origin
de novo origin
amplification
loss
sequence alteration
structure alteration
human specific
lost in chimpanzee
new non-coding gene
regulatory region alteration




Gene amplification
Human-specific gene
Sequence alteration
Structural alteration
Gene loss
Regulatory alteration
De novo origin
Non-coding gene
Lost in chimpanzee
Human accelerated region

Gene set expansion

Tool used

PC algorithm that expand the Local Gene Network (LGN) using transcriptomic data

Use of BOINC to compute on different volunteers computer

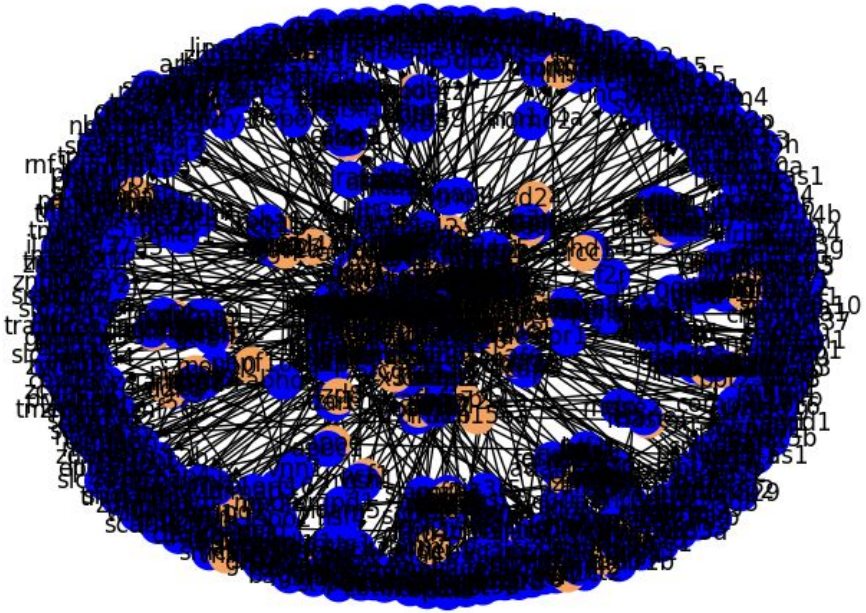
Some parameters:
α= 0,05 (significance threshold)
tsize = 2000 (number of gene in subdivision)
nsize=2000 (number of iterations)

From expansion to graph

Creation of an Interaction Network from the gene expansion using

NetworkX
Network Analysis in Python

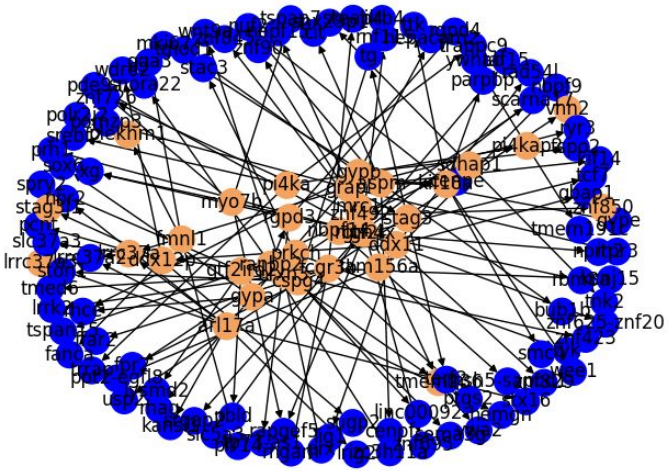
Using absolute frequency as threshold



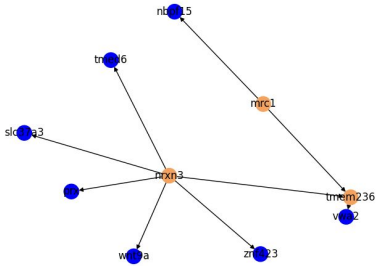
Just too complex to analyse!

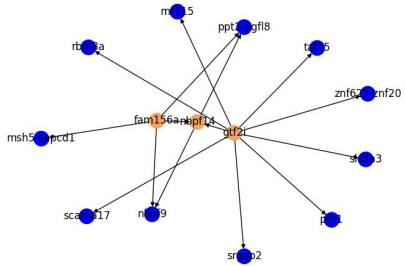
Fixing the mess

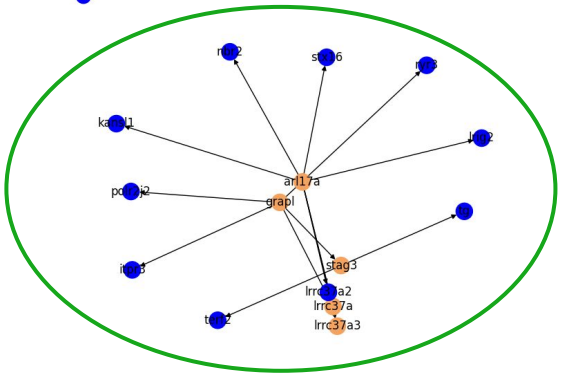
Reducing the nodes
(min 3 edges per node)



Selecting Human specific gene highly connected







Most connected HS gene :LLRC37A, LLRC37A3, ARL17A, STAG3 and GRAPL

Validation of the expansions

Why

Need of a different database to compare the results

How

vs Our gene expansion

Using OverlapGene to define correlation between the sets

Use of contingency table and Fisher Test for correlation

Overlap and Association

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

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

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

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

Overlap and Association

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

The red result

- Difference in the database nature
- Difference in dimension
- Different denominations

Machine learning

Classification

SUBTYPE	COUNT
Pre-B ALL	136
Pre-T ALL	37
T Cell	108
Control	30
Unknown	359

The red is our first threshold

- Correct *prediction of control is essential* to grasp the understanding of the model

The green is our objective

- Classification of all the unknown is what could give us *more insights* on the subtypes

The methods:

01	Random Forest	*Ensemble of decision trees*. It builds multiple decision trees during training and merges them together to get a more accurate and stable prediction.
02	K-Nearest Neighbors	Simple and intuitive algorithm that *classifies a data point based on the majority class* of its k-nearest neighbors.
03	XGBoost	Powerful and efficient implementation of *gradient boosting*. It sequentially adds weak learners to the model, each correcting errors of the previous one.
04	Naive Bayes	*Probabilistic classifier based on Bayes' theorem*. It assumes that the features are conditionally independent given the class label.

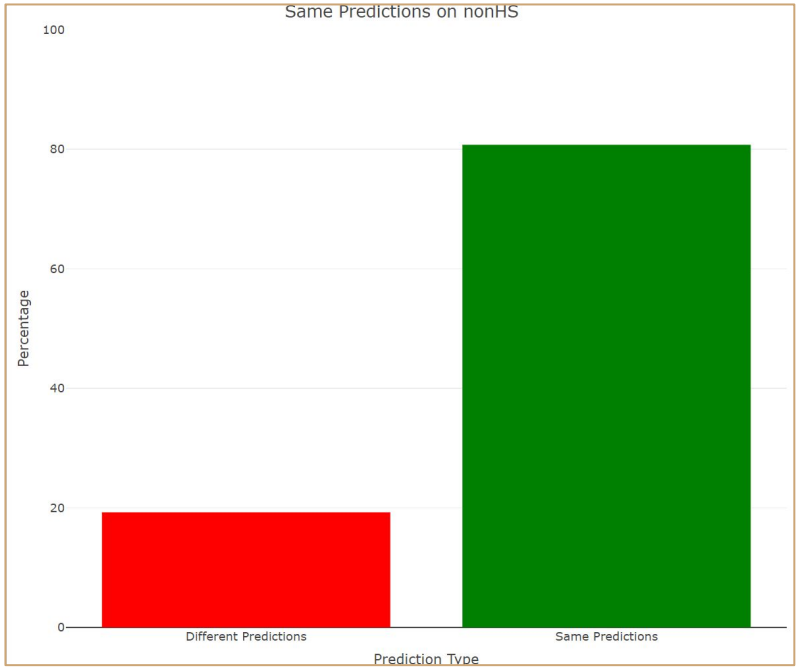
The scores

Non Human specific (nonHS)	
Method	F1 score
Random Forest	0.93
KNN	0.96
XGBoost	0.80

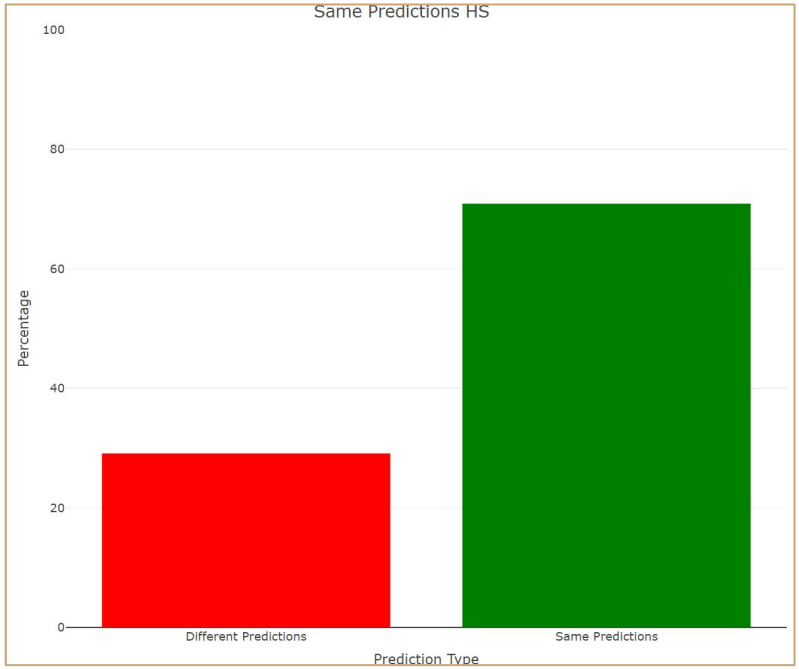
Human specific (HS)	
Method	F1 score
Random Forest	0.97
KNN	0.98
Naive Bayes	0.73

Results

Non Human specific (nonHS)



Human specific (HS)

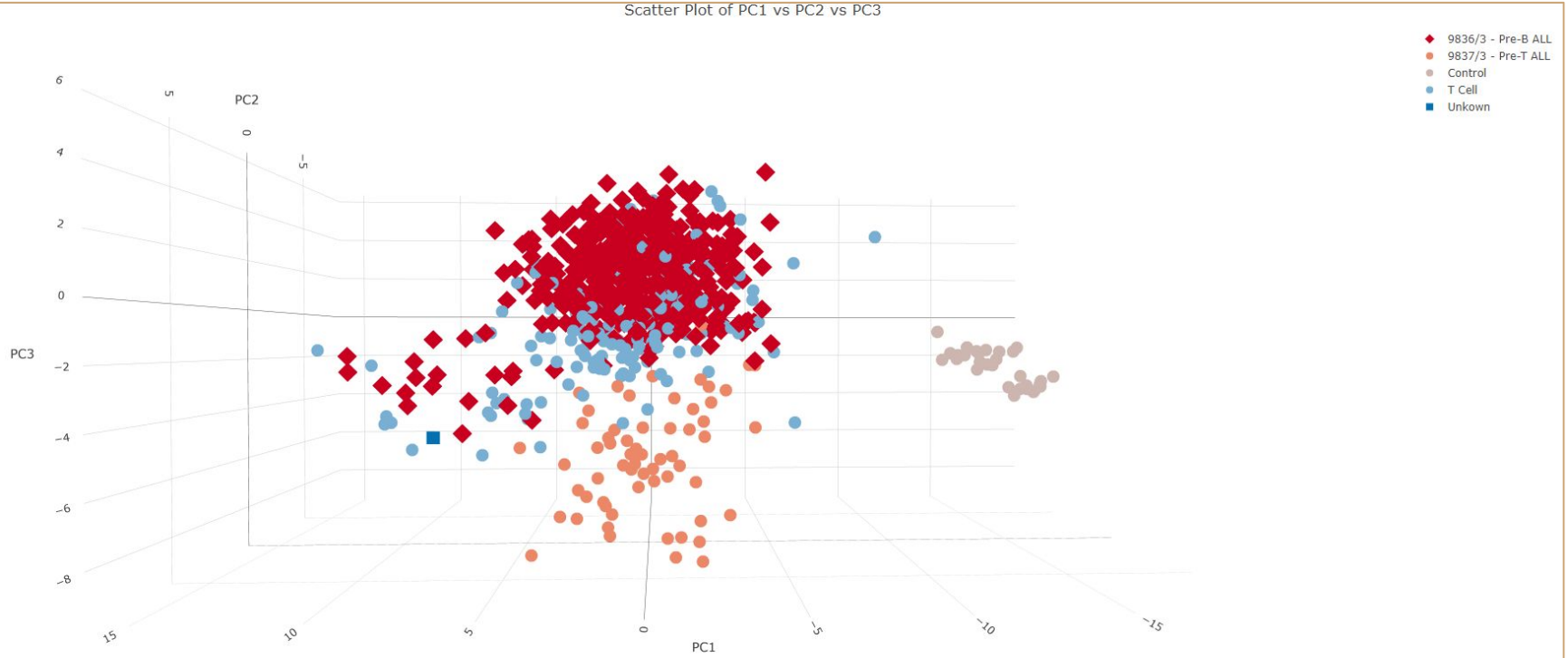


Predictions:

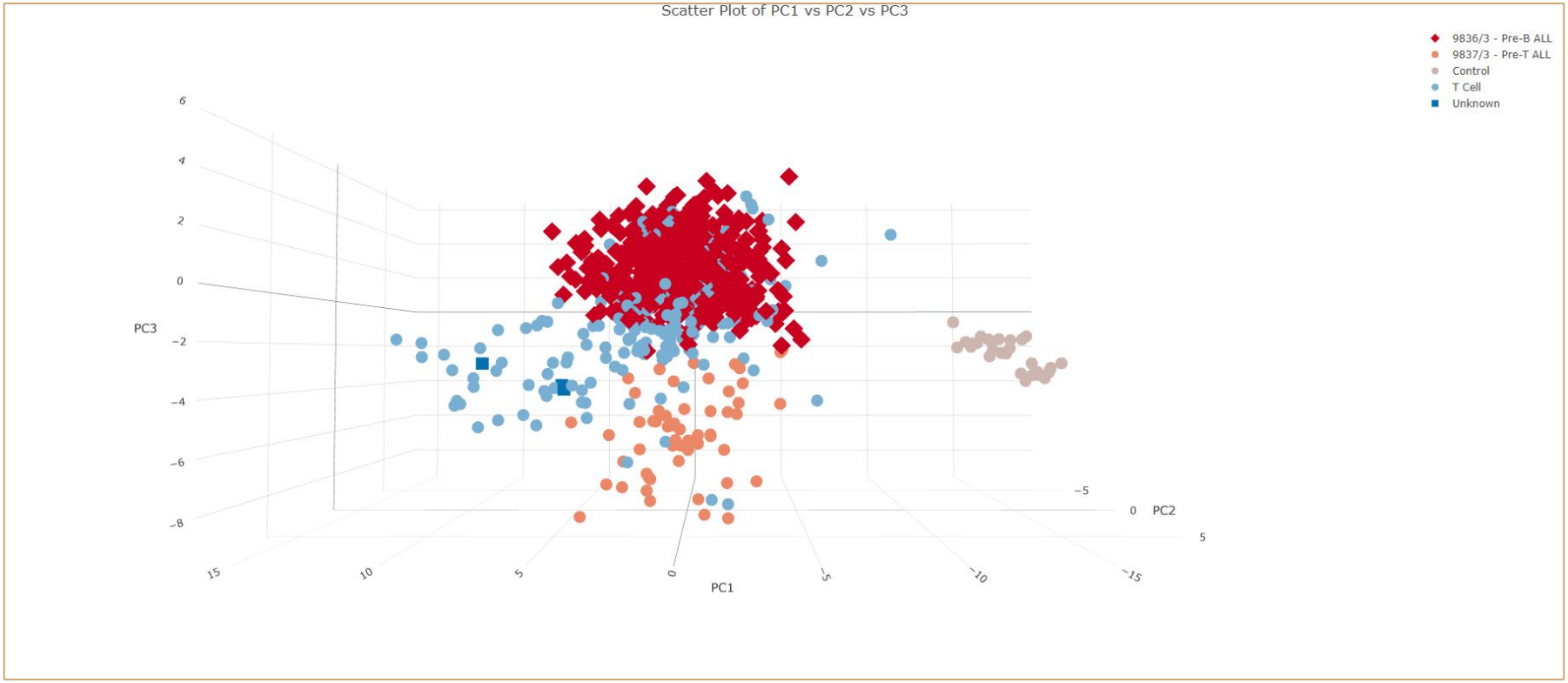
Non Human specific (nonHS)	
SUBTYPE	COUNT
Pre-B ALL	418
Pre-T ALL	63
T Cell	157
Control	30
Unknown	2

Human specific (HS)	
SUBTYPE	COUNT
Pre-B ALL	376
Pre-T ALL	56
T Cell	204
Control	30
Unknown	4

Non Human specific PCA



Human specific PCA

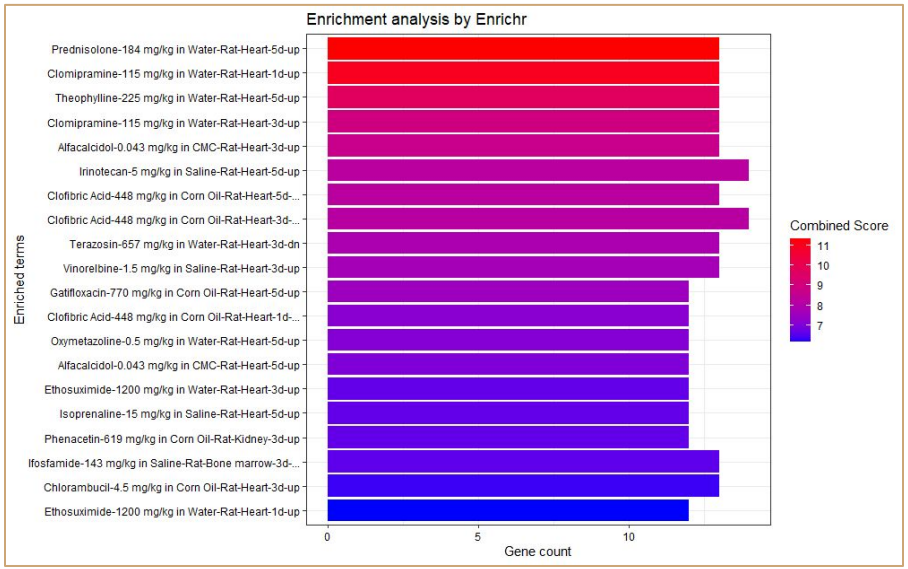


Drug enrichment on subtypes

Enrichr

- Pre B ALL
- Pre T ALL
- T Cell

Pre B ALL



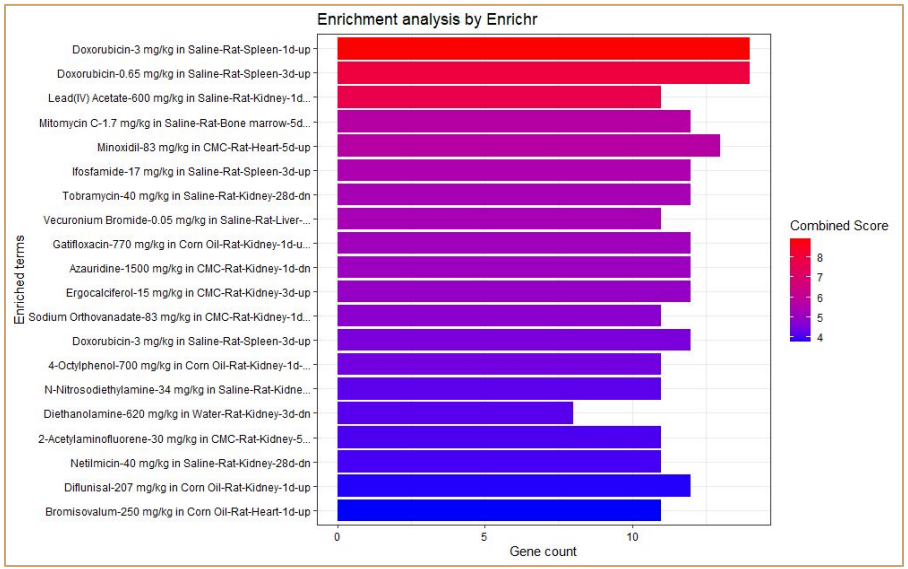
Clomipramine

Known antidepressant, in recent years has been taken under a program of *drug repurposing* and shown promising *anti-cancer activities*

Theophylline

Effective against numerous diseases including leukemia

Pre T ALL



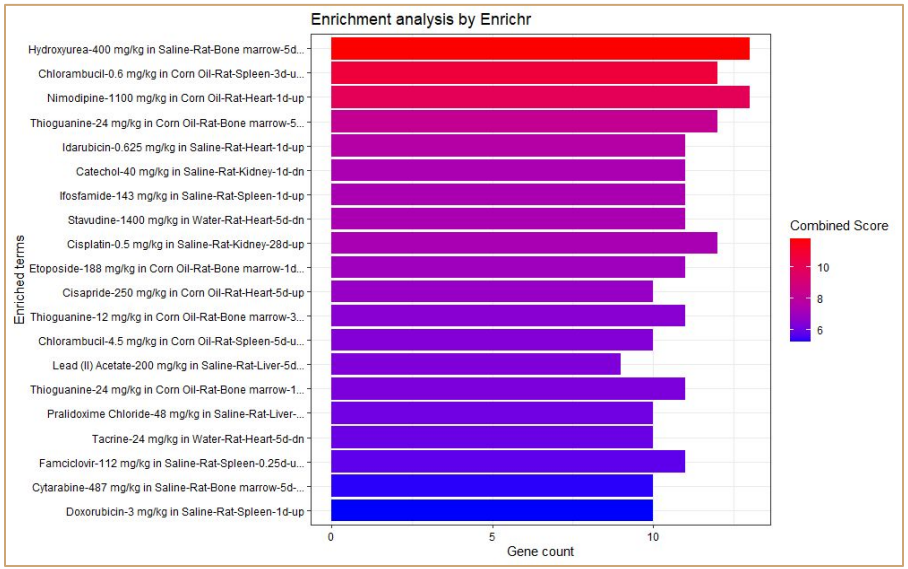
Doxorubicin

Well-known treatment for cancer as one of the most popular chemotherapeutic agents

Mitomycin C

Known treatment against breast cancer, it has shown *potential for leukemia* due to its effect on bone marrow

T Cell



Hydroxyurea

The drug is an *approved treatment* for various forms of cancer and between those the myelogenous leukemia

Chlorambucill

It has been taken into consideration as a *novel chemotherapeutic treatment* in combination with other components to reduce side effects

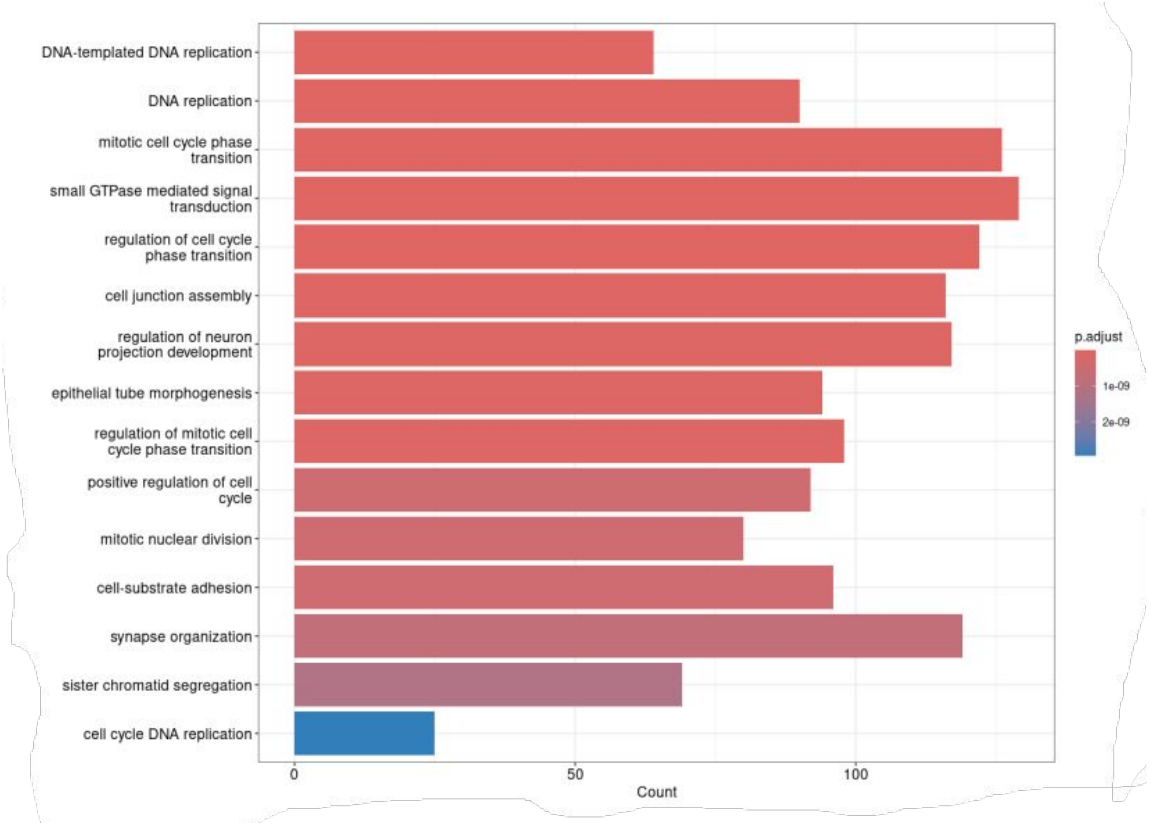
Enrichment and pathway analysis

- Tumor vs Control
- Pediatric vs Adult
- Post-expansion

Tumor vs Control

Non human-specific

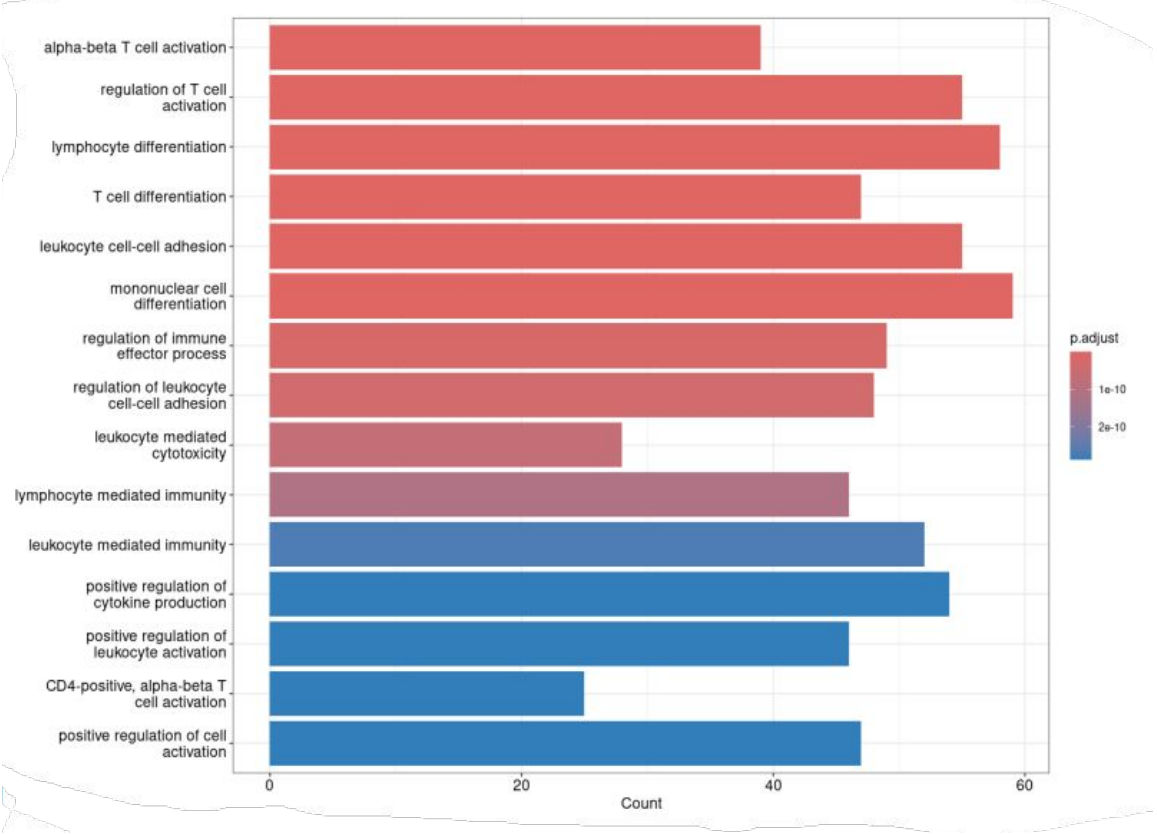
Up-regulated



Tumor vs Control

Non human-specific

Down-regulated



Tumor vs Control

Non human-specific

Up-regulated

- Pleural mesothelioma (MPM)
- Cell cycle
- VEGFA VEGFR2 signaling.

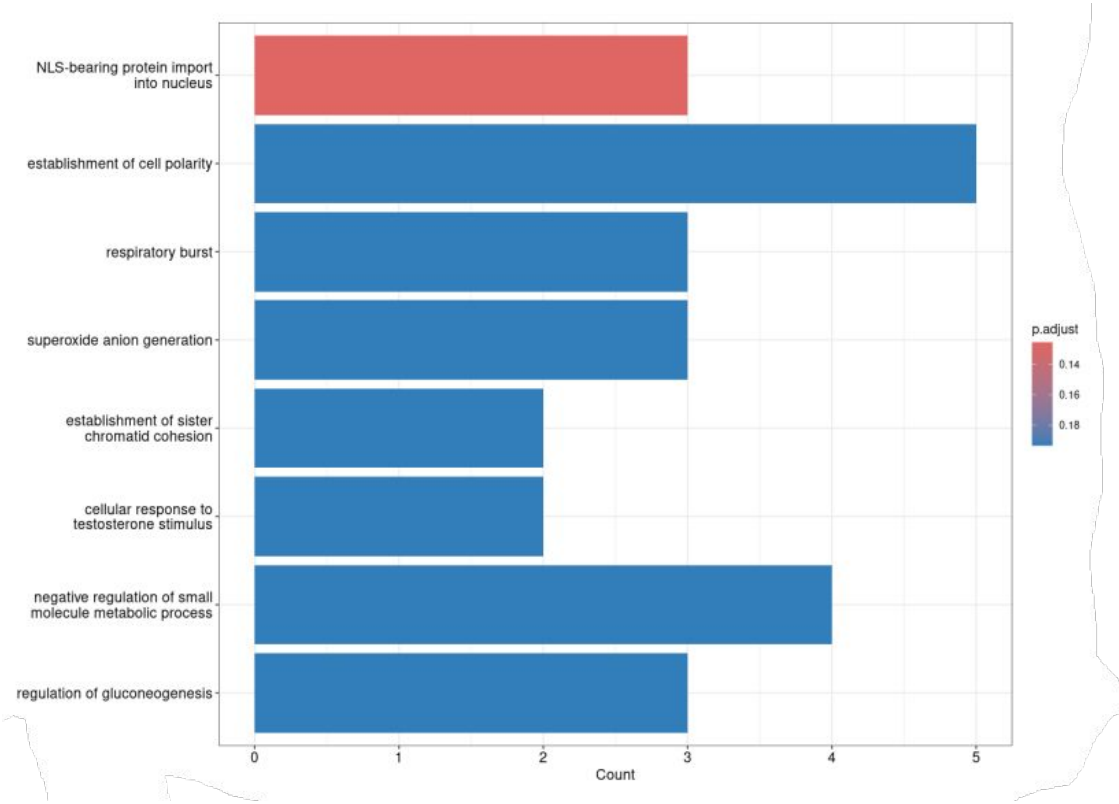
Down-regulated

- T cell receptor and co-stimulatory signaling
- T cell activation SARS CoV 2.

Tumor vs Control

Human-specific

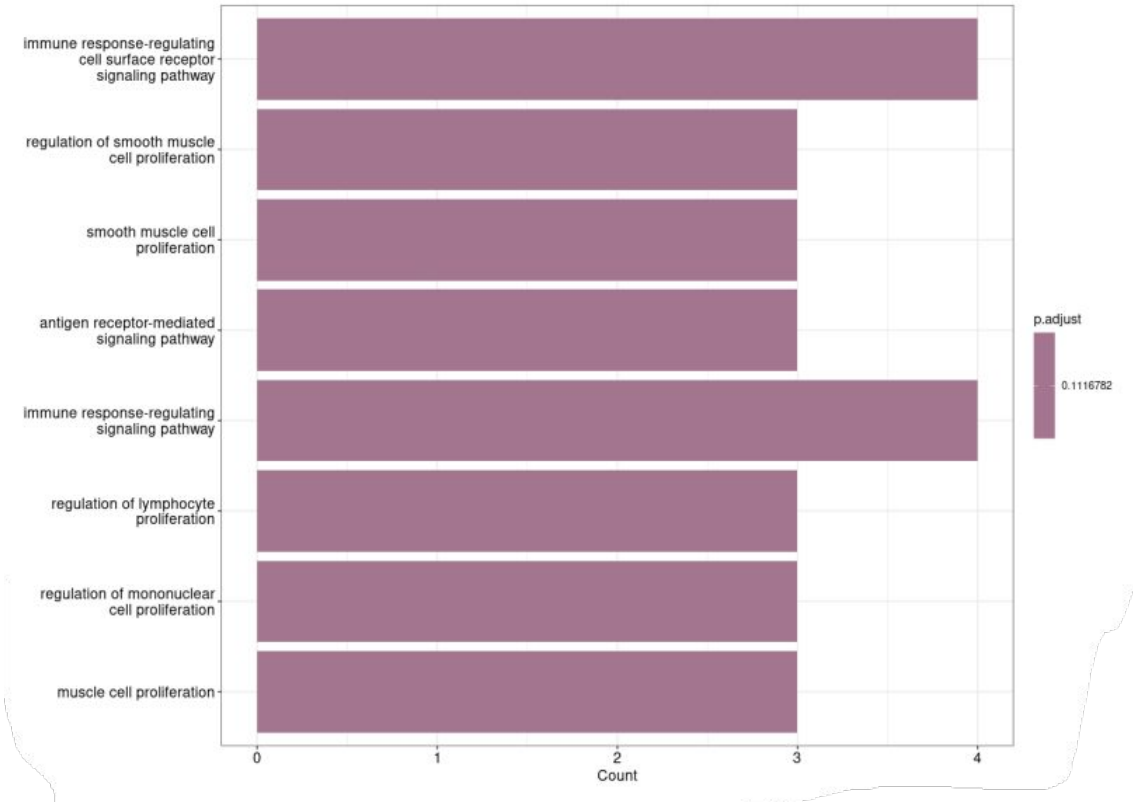
Up-regulated



Tumor vs Control

Human-specific

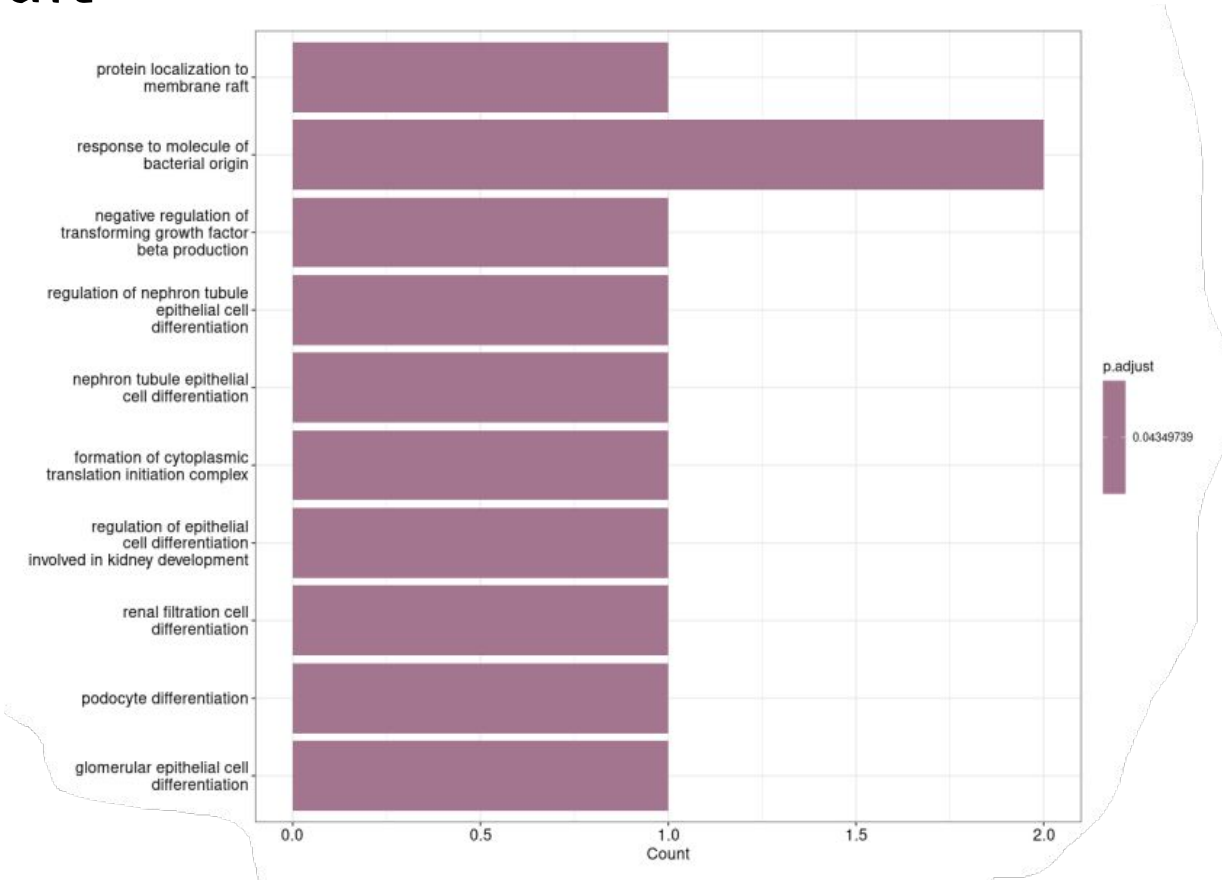
Down-regulated



Pediatric vs Adult

Human-specific

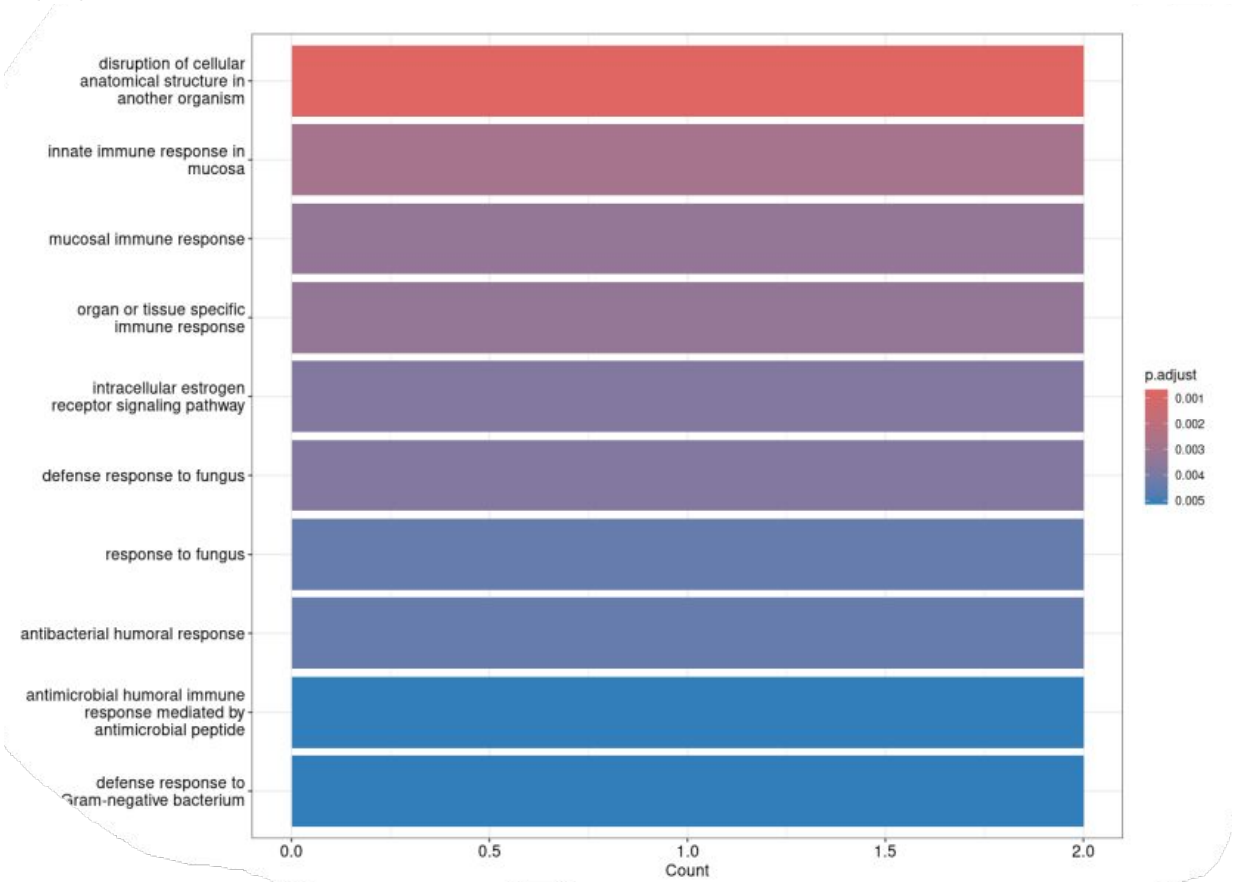
Up-regulated



Pediatric vs Adult

Human-specific

Down-regulated



Pediatric vs Adult

Human-specific

Up-regulated

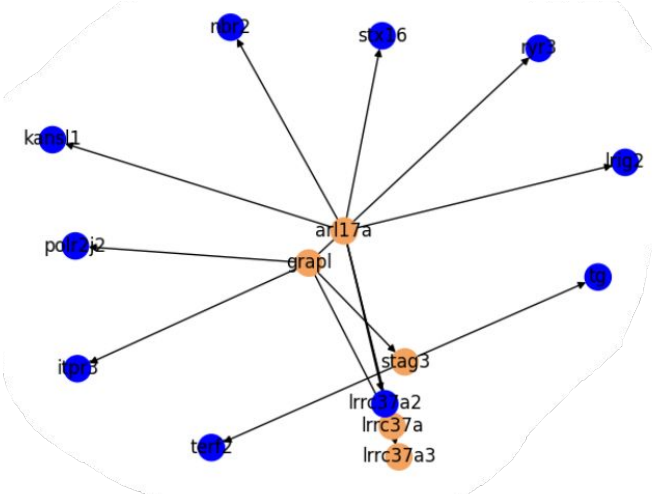
- Cell lineage map for neuronal differentiation

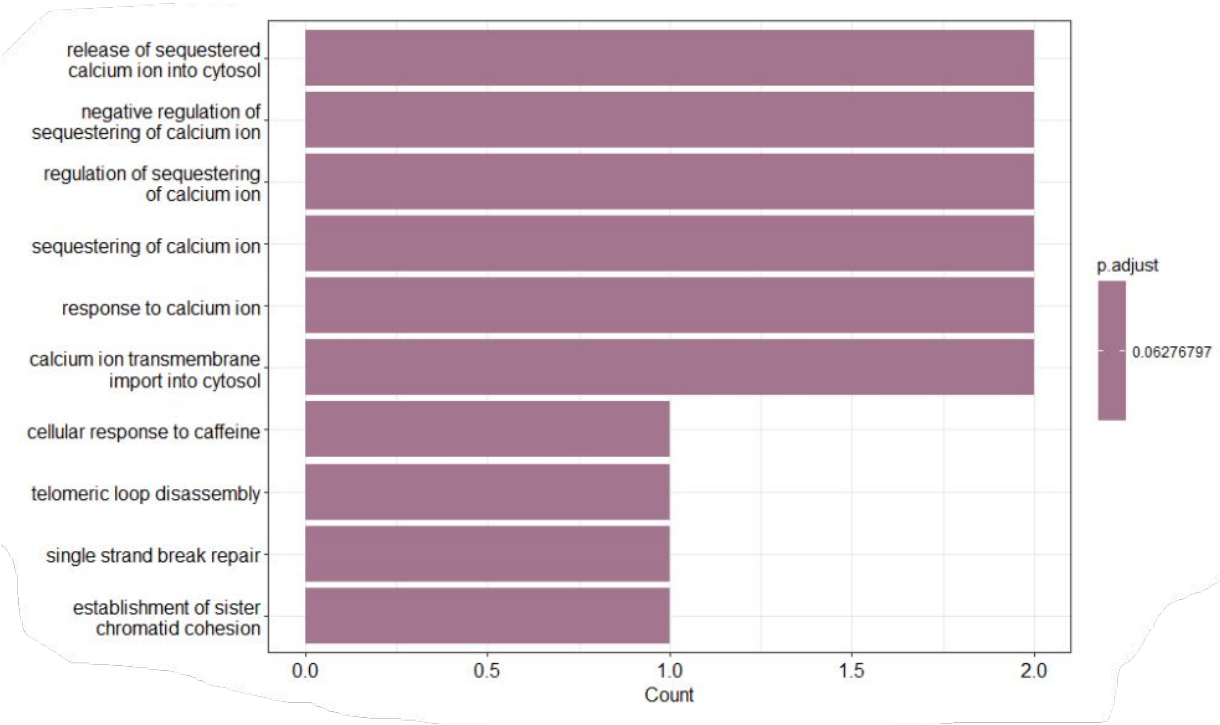
Down-regulated

- Regulatory circuits of the STAT3 signaling pathway

Post-expansion Tumor vs Control

Graph 1





Limitations and Future Perspectives

Limitations

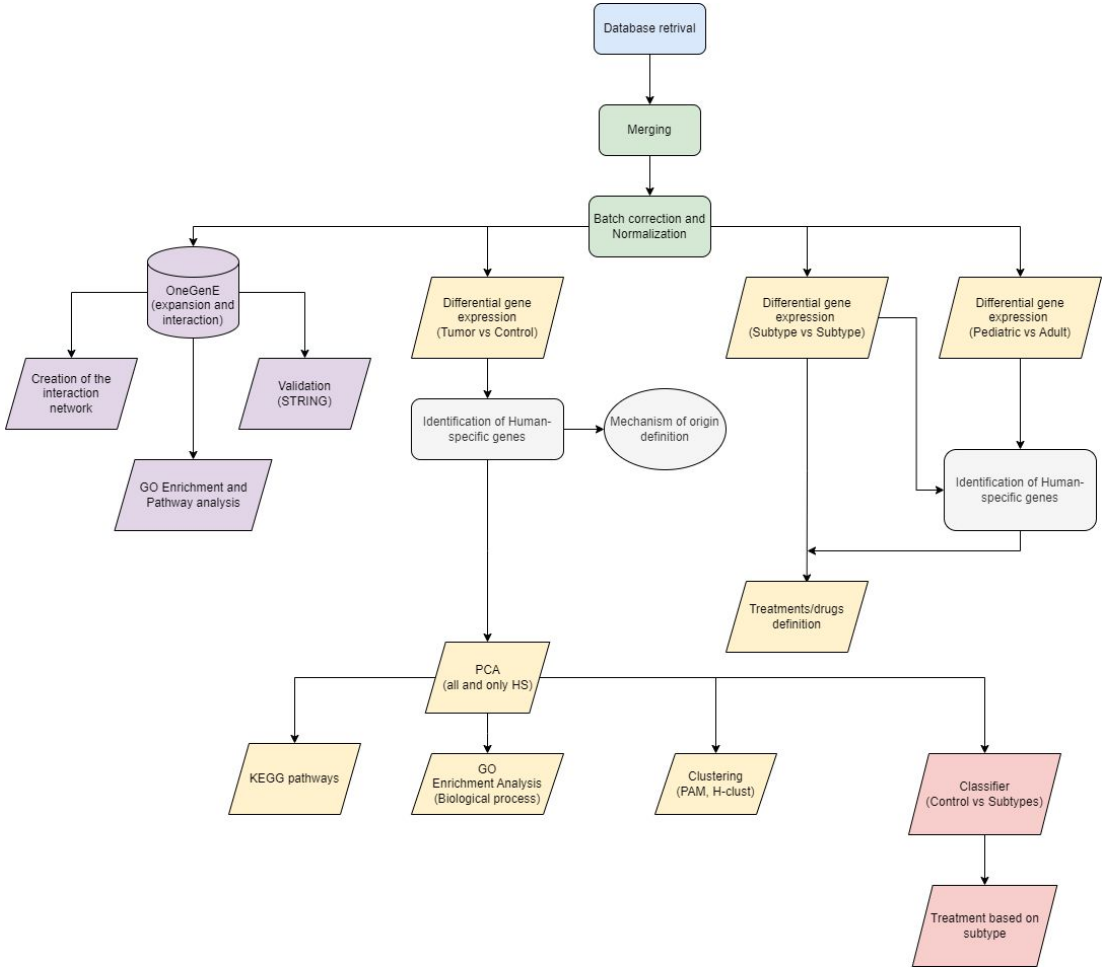
- Lack of datasets from patients with important subtypes (B-cell subtype, Philadephia chromosome (BCR-ABL fusion))
- Low number of samples that were used to create the machine learning classifier

Future Perspectives

- Operate the analysis effectuated given datasets on B-cell subtype patients.
- Using the new labeled data to repeat the subtypes DEG pipeline
- More in depth analysis of the differences (and similarities) between the different ALL subtypes, and pediatric vs adult samples.

Work subdivision

- Preprocessing: Lorenzo, Thomas, Andrea, Gloria
- Gene expansion, validation, enrichment and networks analysis: Lorenzo, Andrea, Gloria
- Differential gene expression analysis, PCA, Enrichment and Clustering: Andrea and Gloria
- Machine learning, consensus, treatment enrichment: Thomas and Lorenzo
- Report writing: Andrea, Thomas, Lorenzo, Gloria, Sabri



Thank you for the attention!