Universitat Oberta de Catalunya

Master's Thesis
Data Mining and Machine Learning

Study of the Transcriptional Function of Cyclin D1 in Leukemia

Author: Antonio Milán Otero
Tutor: Carles Barceló
Professor: Jordi Casas Roma

Agenda

1. Problem description
2. Project objectives
3. Methods
4. Results
5. Conclusions
6. Future developments

1- Problem Description

- Leukemia: set of tumor processes that causes an uncontrolled increase in leukocytes in the blood or lymphatic organs.

- Cyclin D1:
  - Oncogene frequently overexpressed in cancer.
  - One of the main regulators of the cell cycle.
  - Role as regulator of transcription is unknown.
  - Binds to the promoter regions of many genes, but the results of its transcriptional activity remains unknown. That activity is believed to be fundamental in the development of leukemia.

- DNA-damage Response (DDR):
  - Process of repairing DNA damage caused by metabolic activities or environmental factors.
  - DNA damage can produce harmful mutations in the cell genome that can lead to a tumor.

1- Problem Description

- Clarification of mechanisms that initiate the process of repairing DNA damage → Improvements in cancer prediction, treatment in early stages and cancer prevention.

- Mantle Cell Lymphoma (MCL):
  - Type of leukemia.
  - Survival average of ~ 3-5 years.
  - Characterized by the overexpression of Cyclin D1.
  - Binding of Cyclin D1 to regions of DNA involved in the regulation of DDR.

1- Problem Description

Why is this project important?

- Analyzing the similarity of the gene expression regulated by Cyclin D1 of MCL with respect to gene expression of DDR would allow to:
  - Explore essential mechanisms of carcinogenesis
  - Focus on important genes in the process of tumor progression
  - Find new biomarkers to help in an early diagnosis (linked to better survival rate)

- This knowledge can improve the prevention, diagnosis and treatment of MCL.

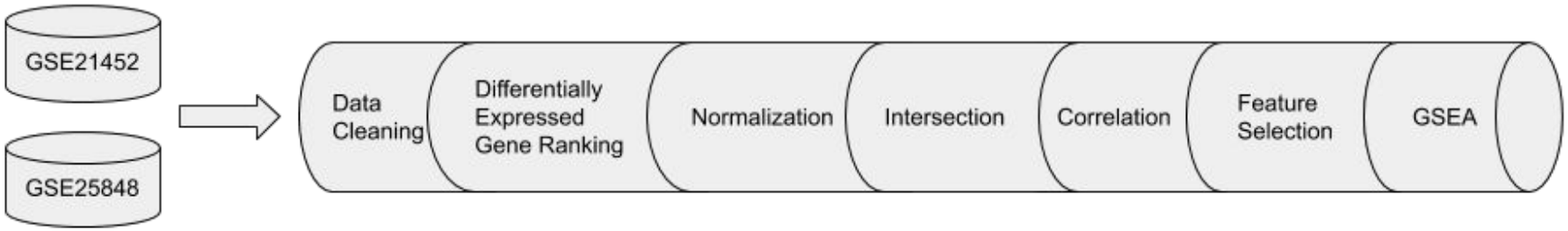
2- Project Objectives

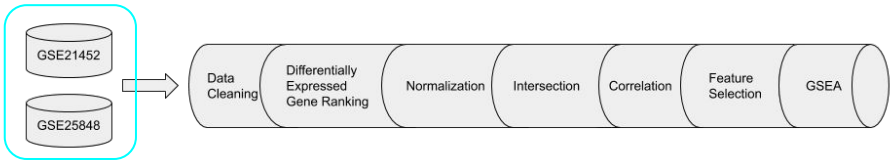
- Analysis of the similarities of the gene expression regulated by Cyclin D1 in MCL and the gene expression of the DDR using Machine Learning and Gene Set Enrichment Analysis.

- Identify significantly enriched genes that can act as therapeutic target and as a biomarker.

- Create Machine Learning models to boost the process of identification of the significantly enriched genes.

3- Methods


GSE21452
GSE25848
Data Cleaning
Differentially Expressed Gene Ranking
Normalization
Intersection
Correlation
Feature Selection
GSEA



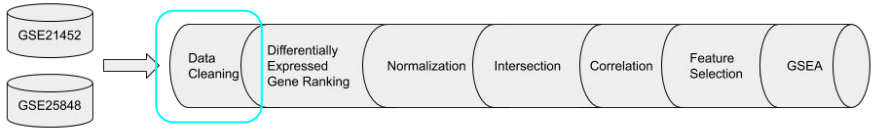
- GSE21452: data from MCL tumors
- GSE25848: DDR data
- Data structure:

Name	Type	Value
gse_21452	S4 [54675 x 64] (Biobase::Expr	S4 object of class ExpressionSet
experimentData	S4 (Biobase::MIAME)	S4 object of class MIAME
assayData	environment [1]	<environment: 0x5564e9259888>
exprs	double [54675 x 64]	10.616 9.274 9.278 10.032 4.263 11.064 9.772 8.997 7.435 9.515 5.075 11. ...
phenoData	S4 [64 x 41] (Biobase::Annotat	S4 object of class AnnotatedDataFrame
featureData	S4 [54675 x 16] (Biobase::Ann	S4 object of class AnnotatedDataFrame
varMetadata	list [16 x 3] (S3: data.frame)	A data.frame with 16 rows and 3 columns
data	list [54675 x 16] (S3: data.fram	A data.frame with 54675 rows and 16 columns
ID	character [54675]	'1007_s_at' '1053_at' '117_at' '121_at' '1255_g_at' '1294_at' ...
GB_ACC	character [54675]	'U48705' 'M87338' 'X51757' 'X69699' 'L36861' 'L13852' ...
SPOT_ID	logical [54675]	NA NA NA NA NA NA ...
Species Scientific Name	character [54675]	'Homo sapiens' 'Homo sapiens' 'Homo sapiens' 'Homo sapiens' 'Homo ...
Annotation Date	character [54675]	'Oct 6, 2014' 'Oct 6, 2014' 'Oct 6, 2014' 'Oct 6, 2014' 'Oct 6, 2014' 'Oct 6, 20 ...
Sequence Type	character [54675]	'Exemplar sequence' 'Exemplar sequence' 'Exemplar sequence' 'Exemplar sequence' ...
Sequence Source	character [54675]	'Affymetrix Proprietary Database' 'GenBank' 'Affymetrix Proprietary Database' 'G ...
Target Description	character [54675]	'U48705 /FEATURE=mRNA /DEFINITION=HSU48705 Human receptor tyrosine kinase DDR ge ...
Representative Public ID	character [54675]	'U48705' 'M87338' 'X51757' 'X69699' 'L36861' 'L13852' ...
Gene Title	character [54675]	'discoidin domain receptor tyrosine kinase 1 /// microRNA 4640' 'replication fac ...
Gene Symbol	character [54675]	'DDR1 /// MIR4640' 'RFC2' 'HSPA6' 'PAX8' 'GUCA1A' 'MIR5193 /// UBA7' ...
ENTREZ_GENE_ID	character [54675]	'780 /// 100616237' '5982' '3310' '7849' '2978' '7318 /// 100847079' ...
RefSeq Transcript ID	character [54675]	'NM_001202521 /// NM_001202522 /// NM_001202523 /// NM_001954 /// NM_013993 /// ...
Gene Ontology Biological Process	character [54675]	'0001558 // regulation of cell growth // inferred from electronic annotation /// ...
Gene Ontology Cellular Component	character [54675]	'0005576 // extracellular region // inferred from electronic annotation /// 0005 ...
Gene Ontology Molecular Function	character [54675]	'0000166 // nucleotide binding // inferred from electronic annotation /// 000467 ...
dimLabels	character [2]	'featureNames' 'featureColumns'
.__classVersion__	list [1] (Biobase::Versions)	List of length 1
annotation	character [1]	'GPL570'
protocolData	S4 [64 x 0] (Biobase::Annotate	S4 object of class AnnotatedDataFrame
.__classVersion__	list [4] (Biobase::Versions)	List of length 4

⇦ Features (genes) in rows.
Samples in columns

3- Methods

Data Cleaning

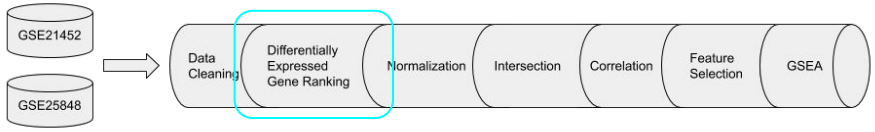


- GSE25848 containing NA values
  - 32443 out of 48803 genes without any data
  - 16360 genes with an expression value.

	GSM634846	GSM634847	GSM634848	GSM634849	GSM634850	GSM634851	GSM634852	GSM634853	GSM634854	GSM634855	GSM634856	GSM634857
ILMN_1343291	5.9666587	5.9632518	5.746952	7.0891942	7.7283965	7.07234597	6.5124040	7.2923316	5.8146148	6.9141257	7.0468403	6.710883
ILMN_1343295	4.6292055	4.7184427	4.600086	5.7745672	6.2601675	5.56834851	5.0913109	5.2818357	4.5288116	5.1151062	5.1583056	4.658634
ILMN_1651199	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
ILMN_1651209	-0.2649516	-0.3468473	-0.390777	-0.2515894	-0.2010714	-0.03880678	-0.1543887	-0.3561458	-0.1362886	-0.2303119	-0.1586851	-0.176603
ILMN_1651210	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
ILMN_1651221	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

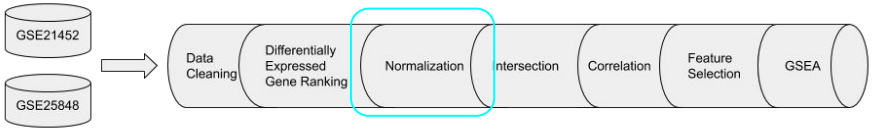
3- Methods

Differentially Expressed Gene Ranking



- Purpose: make a first selection of the top 10000 differentially expressed genes.
- Individually for each data set.
- Ranked by standard deviation of genes.

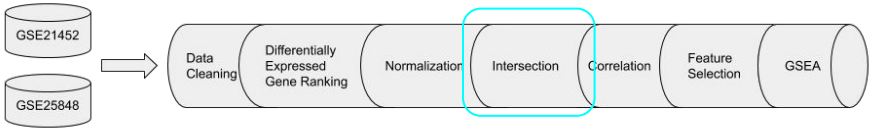
Normalization



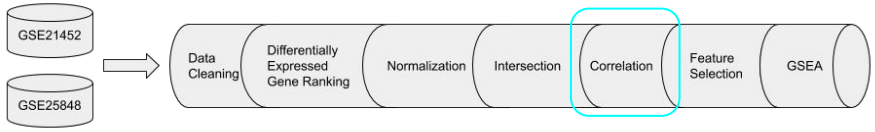
- Data obtained from different platforms and with different pre-processing.
- Data in different scales.
- Z-score normalization applied to each individual data set.
- Data ready to be merged.

3- Methods

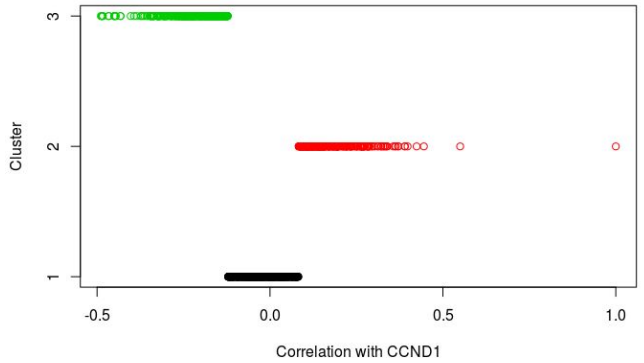
Intersection

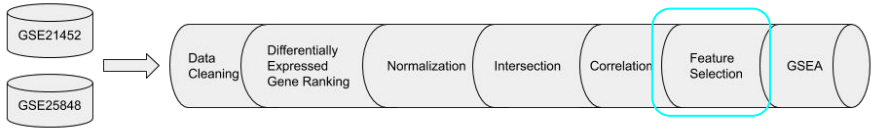


- Intersection between data sets.
- Match done by Gene Entrez ID.
- New data set with only matched genes.

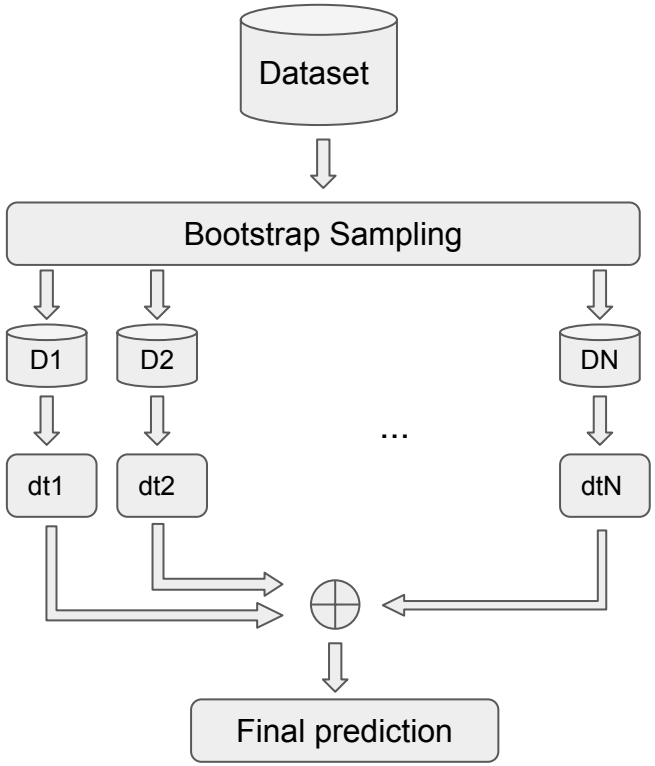


- Correlation between the expression of the genes in the new data set and the expression of Cyclin D1.
- K-Means applied to classify genes in three clusters:
  - Positive correlation
  - No significant correlation
  - Negative correlation
- Two new columns added to the data set:
  - Correlation value
  - Correlation cluster
- Only positive correlation cluster used in the next steps.
- 316 genes selected.



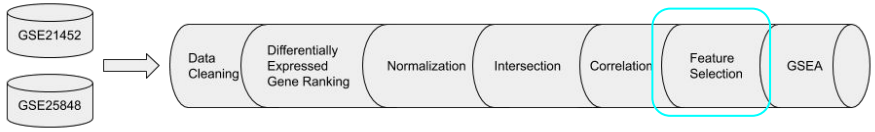


- Random Forest
- Ensemble method (bagging)
- Bootstrap sampling:
  - Random samples with replacement
  - Random feature subsets
- N number of decision trees generated
- Out-of-bag: Estimated error from averaging the partial error in the base classifiers, using the data subtracted from the training dataset (out of bag data)
- Feature importance obtained by the computation of the Gini Impurity and Gini decrease average across the trees in the forest

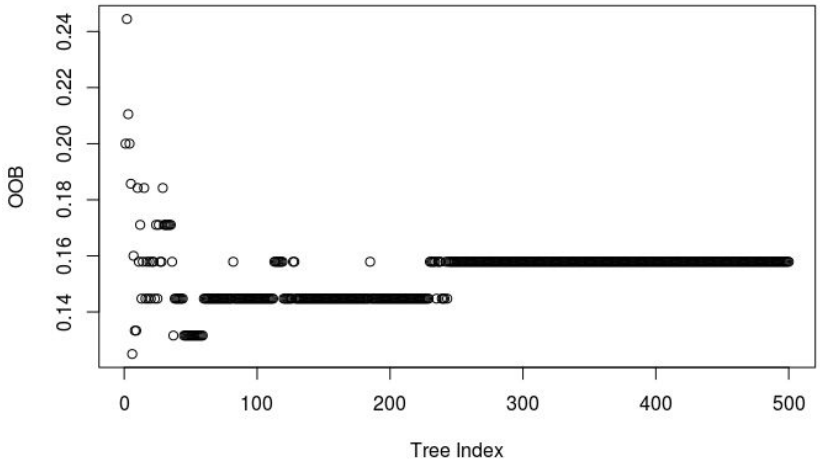


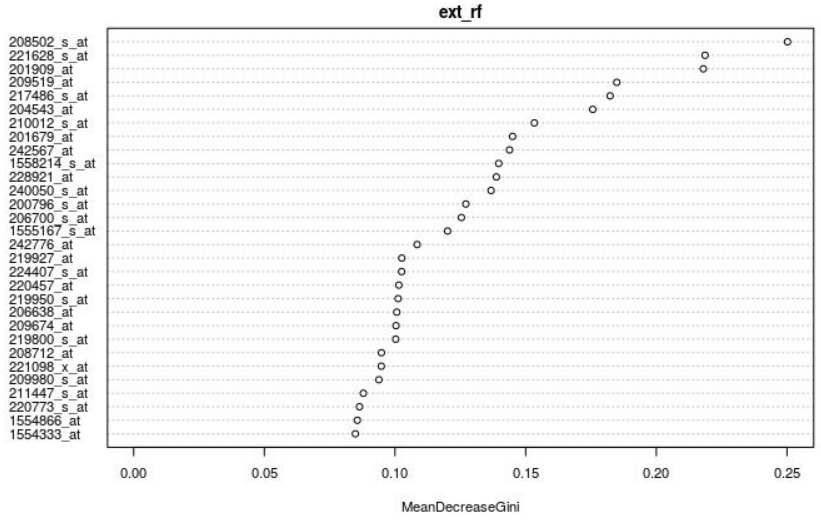
3- Methods

Feature Selection

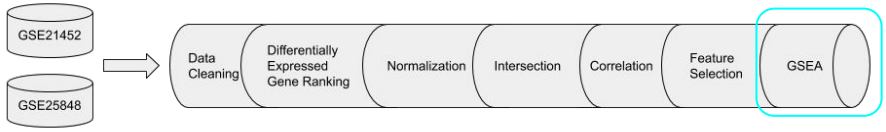


- Unsupervised and Supervised Random Forest applied. Best final results obtained from the supervised method.
- Genes ranked by importance.
- 100 genes selected from the Supervised Random Forest.

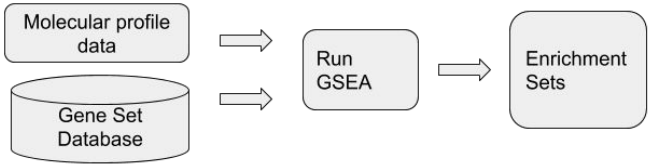




Gene Set Enrichment Analysis



- GSEA receives two inputs:
  - Molecular profile
  - Gene Set Database



- GSEA calculates an Enrichment Score (ES) between phenotypes for each gene contained in the molecular profile.
- Using the ES, GSEA identify which set of genes offers statistical significance.
- Our gene selection used as molecular profile and Cyclin D1 correlation as phenotype.
- Gene set databases used from MSigDB:
  - Hallmark gene set (H)
  - Oncogenic gene set (C6)
  - GO gene set (C5)
  - Curated gene sets (C2)
  - Immunologic gene sets (C7)

4- Results

- GSEA Statistics:
  - Enrichment Score (ES): degree to which a gene set is over-represented by the selected genes
  - Normalized Enrichment Score (NES): ES normalized across the analyzed gene sets
  - False Discovery Rate (FDR): estimated probability that a NES represents a false positive
  - Nominal P Value: statistical significance of the ES

- Identified gene sets ranked by NES value
- Generalized cut-off on FDR at 25%
- Nominal p value cut-off at 1% and 5%

- Enrichment in phenotype for positive correlations with Cyclin D1

Collection	Up-regulated gene sets	FDR <25%	p-value <1%	p-value <5%
Hallmark, H	13/28	7	4	5
Oncogenic, C6	45/107	6	3	5
GO, C5	919/1824	36	60	95
Curated, C2	908/1598	65	141	175
Immunologic, C7	1707/3175	0	41	106

- Enrichment in phenotype for negative correlations with Cyclin D1

Collection	Up-regulated gene sets	FDR <25%	p-value <1%	p-value <5%
Hallmark, H	15/28	0	0	2
Oncogenic, C6	62/107	0	1	7
GO, C5	905/1824	0	10	41
Curated, C2	690/1598	1	37	86
Immunologic, C7	1468/3175	0	37	111

4- Results

Hallmark Gene Set

Collection	Up-regulated gene sets	FDR <25%	p-value <1%	p-value <5%
Hallmark, H	13/28	7	4	5

GS	SIZE	NES	NOM p-val	FDR q-val	LEADING EDGE
HALLMARK_ESTROGEN_RESPONSE_EARLY	3	1.55	0.010	0.071	tags=33%, list=0%, signal=32%
HALLMARK_HYPOXIA	2	1.52	0.006	0.051	tags=50%, list=3%, signal=51%
HALLMARK_ESTROGEN_RESPONSE_LATE	2	1.41	0.043	0.145	tags=50%, list=0%, signal=49%
HALLMARK_APOPTOSIS	3	1.39	0.075	0.126	tags=100%, list=22%, signal=124%
HALLMARK_NOTCH_SIGNALING	1	1.33	0.000	0.186	tags=100%, list=0%, signal=99%
HALLMARK_ANDROGEN_RESPONSE	1	1.33	0.000	0.155	tags=100%, list=0%, signal=99%
HALLMARK_TNFA_SIGNALING_VIA_NFKB	4	1.32	0.132	0.137	tags=75%, list=22%, signal=92%

- Interesting up-regulated gene sets:
  - Hypoxia
  - Apoptosis
  - Notch Signaling

4- Results

Oncogenic Signatures

Collection	Up-regulated gene sets	FDR <25%	p-value <1%	p-value <5%
Oncogenic, C6	45/107	6	3	5

Gene Set	Size	ES	NES	NOM p-val	FDR q-val	Leading Edge
PRC2_EED_UP.V1_DN	3	0.96	1.54	0.004	0.145	tags=100%, list=6%, signal=103%
BMI1_DN.V1_UP	4	0.83	1.54	0.037	0.077	tags=50%, list=3%, signal=49%
BMI1_DN_MEL18_DN.V1_UP	4	0.75	1.43	0.079	0.212	tags=50%, list=3%, signal=49%
MEL18_DN.V1_UP	4	0.75	1.43	0.079	0.159	tags=50%, list=3%, signal=49%
RAF_UP.V1_DN	3	0.79	1.43	0.070	0.129	tags=33%, list=0%, signal=32%
IL2_UP.V1_UP	2	0.93	1.39	0.035	0.161	tags=100%, list=8%, signal=107%

4- Results

Gene Ontology Gene Sets

Collection	Up-regulated gene sets	FDR <25%	p-value <1%	p-value <5%
GO, C5	919/1824	36	60	95

Gene Set	SIZE	ES	NES	NOM p-val	FDR q-val	LEADING EDGE
GO_POSITIVE_REGULATION_OF_PROTEIN_METABOLIC_PROCESS	10	0.69	1.82	0.006	0.323	tags=40%, list=9%, signal=40%
GO_MOLECULAR_FUNCTION_REGULATOR	10	0.70	1.81	0.004	0.181	tags=40%, list=6%, signal=38%
GO_POSITIVE_REGULATION_OF_PHOSPHORUS_METABOLIC_PROCESS	8	0.73	1.80	0.000	0.139	tags=50%, list=9%, signal=51%
GO_POSITIVE_REGULATION_OF_PROTEIN_MODIFICATION_PROCESS	8	0.73	1.80	0.000	0.104	tags=50%, list=9%, signal=51%
GO_ENZYME_REGULATOR_ACTIVITY	7	0.80	1.76	0.000	0.134	tags=43%, list=4%, signal=41%
GO_POSITIVE_REGULATION_OF_CATALYTIC_ACTIVITY	11	0.67	1.74	0.010	0.145	tags=36%, list=9%, signal=36%
GO_REGULATION_OF_MULTICELLULAR_ORGANISMAL_DEVELOPMENT	6	0.81	1.71	0.012	0.161	tags=50%, list=10%, signal=52%
GO_REGULATION_OF_HYDROLASE_ACTIVITY	9	0.68	1.69	0.019	0.190	tags=44%, list=9%, signal=44%
GO_POSITIVE_REGULATION_OF_DEVELOPMENTAL_PROCESS	6	0.78	1.68	0.014	0.191	tags=50%, list=10%, signal=52%
GO_POSITIVE_REGULATION_OF_MOLECULAR_FUNCTION	12	0.62	1.68	0.021	0.174	tags=33%, list=9%, signal=32%
GO_POSITIVE_REGULATION_OF_TRANSFERASE_ACTIVITY	5	0.78	1.66	0.004	0.180	tags=40%, list=6%, signal=40%
GO_KINASE_ACTIVITY	7	0.74	1.66	0.024	0.167	tags=43%, list=9%, signal=44%
GO_PROTEIN_KINASE_ACTIVITY	5	0.83	1.66	0.012	0.167	tags=60%, list=9%, signal=63%
GO_REGULATION_OF_MITOTIC_CELL_CYCLE	4	0.87	1.64	0.012	0.176	tags=25%, list=0%, signal=24%
GO_PROTEIN_PHOSPHORYLATION	7	0.71	1.64	0.015	0.170	tags=43%, list=9%, signal=44%
GO_REGULATION_OF_GTPASE_ACTIVITY	7	0.71	1.64	0.023	0.164	tags=43%, list=6%, signal=42%
GO_CELL_DIVISION	4	0.86	1.63	0.008	0.172	tags=25%, list=0%, signal=24%
GO_PHOSPHORYLATION	9	0.63	1.62	0.017	0.169	tags=33%, list=9%, signal=33%
GO_NEGATIVE_REGULATION_OF_CELL_CYCLE_PROCESS	3	0.91	1.60	0.008	0.204	tags=33%, list=0%, signal=32%
GO_NEGATIVE_REGULATION_OF_MITOTIC_CELL_CYCLE	3	0.91	1.60	0.008	0.194	tags=33%, list=0%, signal=32%

- Interesting up-regulated gene sets:
  - Positive regulation of catalytic activity
  - Regulation of multicellular organismal development
  - Regulation of mitotic cell cycle
  - Negative regulation of cell cycle process
  - Negative regulation of mitotic cell cycle

4- Results

Curated Gene Sets

Collection	Up-regulated gene sets	FDR <25%	p-value <1%	p-value <5%
Curated, C2	908/1598	65	141	175

Gene Set	SIZE	ES	NES	NOM p-val	FDR q-val	LEADING EDGE
BERENJENO_TRANSFORMED_BY_RHOA_UP	6	0.85	1.87	0.000	0.037	tags=33%, list=4%, signal=33%
KRIGE_RESPONSE_TO_TOSEDOSTAT_6HR_DN	8	0.77	1.78	0.002	0.089	tags=50%, list=18%, signal=56%
KRIGE_RESPONSE_TO_TOSEDOSTAT_24HR_DN	8	0.77	1.78	0.002	0.059	tags=50%, list=18%, signal=56%
CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_UP	5	0.89	1.78	0.000	0.045	tags=40%, list=5%, signal=40%
ONKEN_UVEAL_MELANOMA_UP	4	0.93	1.71	0.000	0.093	tags=75%, list=8%, signal=78%
WAMUNYOKOLI_OVARIAN_CANCER_LMP_UP	3	0.98	1.69	0.000	0.113	tags=33%, list=0%, signal=32%
BLALOCK_ALZHEIMERS_DISEASE_INCIPIENT_UP	6	0.78	1.68	0.006	0.108	tags=83%, list=24%, signal=102%
NUYTTEN_NIPP1_TARGETS_DN	5	0.82	1.68	0.004	0.103	tags=60%, list=13%, signal=66%
BLALOCK_ALZHEIMERS_DISEASE_UP	15	0.64	1.67	0.008	0.093	tags=60%, list=24%, signal=66%
MARTORIATI_MDM4_TARGETS_NEUROEPITHELIUM_UP	3	0.96	1.62	0.004	0.168	tags=67%, list=5%, signal=68%
MEISSNER_BRAIN_HCP_WITH_H3K4ME3_AND_H3K27ME3	5	0.88	1.61	0.020	0.184	tags=80%, list=8%, signal=83%
KRIEG_HYPOXIA_NOT_VIA_KDM3A	4	0.83	1.61	0.004	0.171	tags=50%, list=6%, signal=51%
SWEET_LUNG_CANCER_KRAS_UP	4	0.85	1.60	0.018	0.170	tags=25%, list=0%, signal=24%
BENPORATH_SOX2_TARGETS	3	0.90	1.59	0.012	0.167	tags=33%, list=0%, signal=32%
PENG_GLUCOSE_DEPRIVATION_DN	4	0.85	1.59	0.012	0.157	tags=50%, list=10%, signal=53%
REACTOME_CELL_CYCLE	3	0.91	1.58	0.006	0.168	tags=33%, list=0%, signal=32%
REACTOME_CELL_CYCLE_MITOTIC	3	0.91	1.58	0.006	0.158	tags=33%, list=0%, signal=32%
CHESLER_BRAIN_QTL_CIS	2	1.00	1.57	0.000	0.174	tags=50%, list=0%, signal=49%
YAGI_AML_WITH_T_8_21_TRANSLOCATION	4	0.85	1.56	0.018	0.172	tags=25%, list=0%, signal=24%
PUJANA_BREAST_CANCER_LIT_INT_NETWORK	3	0.87	1.56	0.018	0.173	tags=33%, list=0%, signal=32%

Immunologic Signatures

Collection	Up-regulated gene sets	FDR <25%	p-value <1%	p-value <5%
Immunologic, C7	1707/3175	0	41	106

- No gene set passes the cut-off of FDR

5- Conclusions

- Hypoxia and apoptosis resistance as a fundamental mechanism of tumor progression. Better understanding of this two conditions might lead to better treatments for MCL.

- Blocking Notch signaling pathway may be considered as a potential therapy for MCL treatment.

- Notch inhibitors may improve chemotherapy response, being a great promise for cancer control.

Great interest for future studies:

- Targeting Notch pathway
- Studying potential common mechanisms of hypoxia and apoptosis resistance

4- Future Developments

- Add several Feature Selection algorithms. Execute them in parallel and extract the most common genes selected.
- Optimization of the machine learning models.
- Consider the addition of more data sets.
- Study the negative correlated genes with Cyclin D1.
- Validate this in silico analysis with further experimental studies.

Thanks for Your Attention

- Repositories
  - LaTeX document: https://github.com/amilan/Thesis-DS
  - Code developed: https://github.com/amilan/Thesis-DS-dev