Giulia Baroni

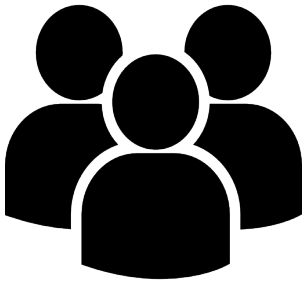
Master Degree Student

Supv.: Prof. Ieva and Prof. Ceri

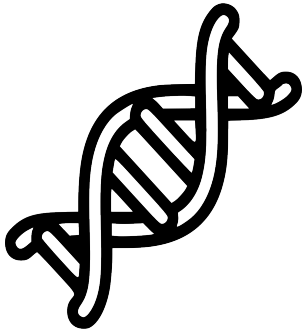
Identification and validation of epigenetic lesions in refractory lymphomas

Aim and Dataset

Upfront identification of refractory classical Hodgkin lymphoma (cHL) patients

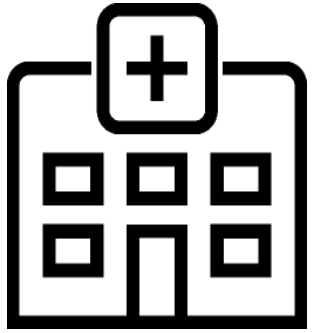


194 patients



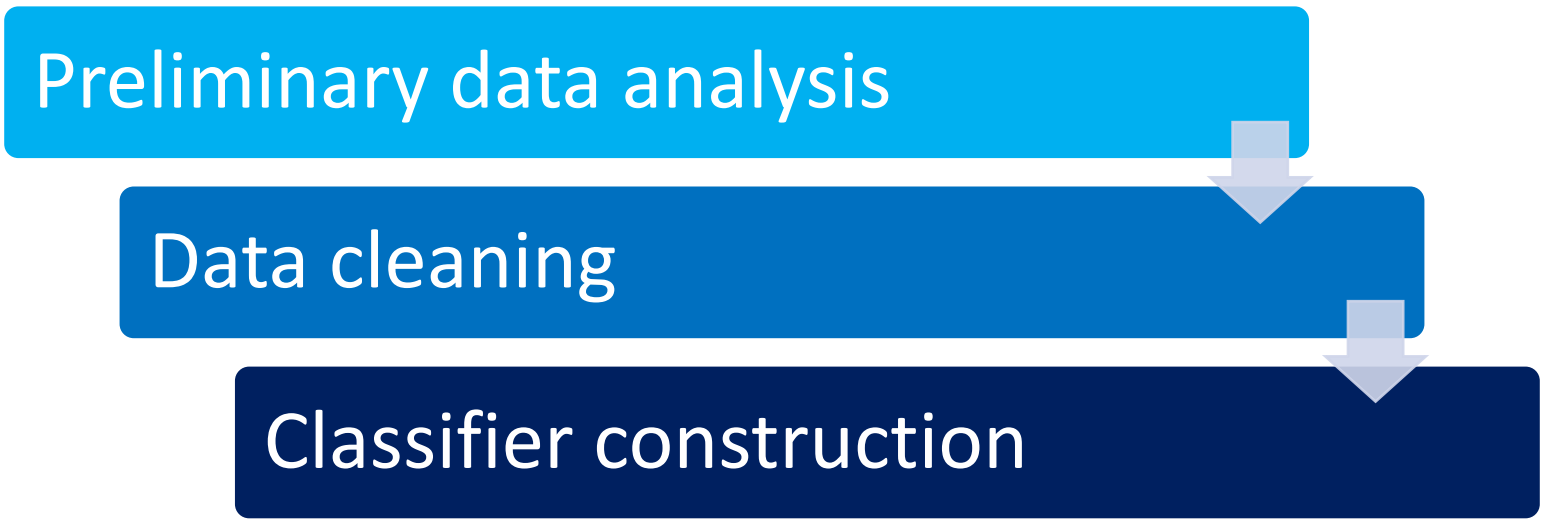
96 RNA expressions

4 clinical factors

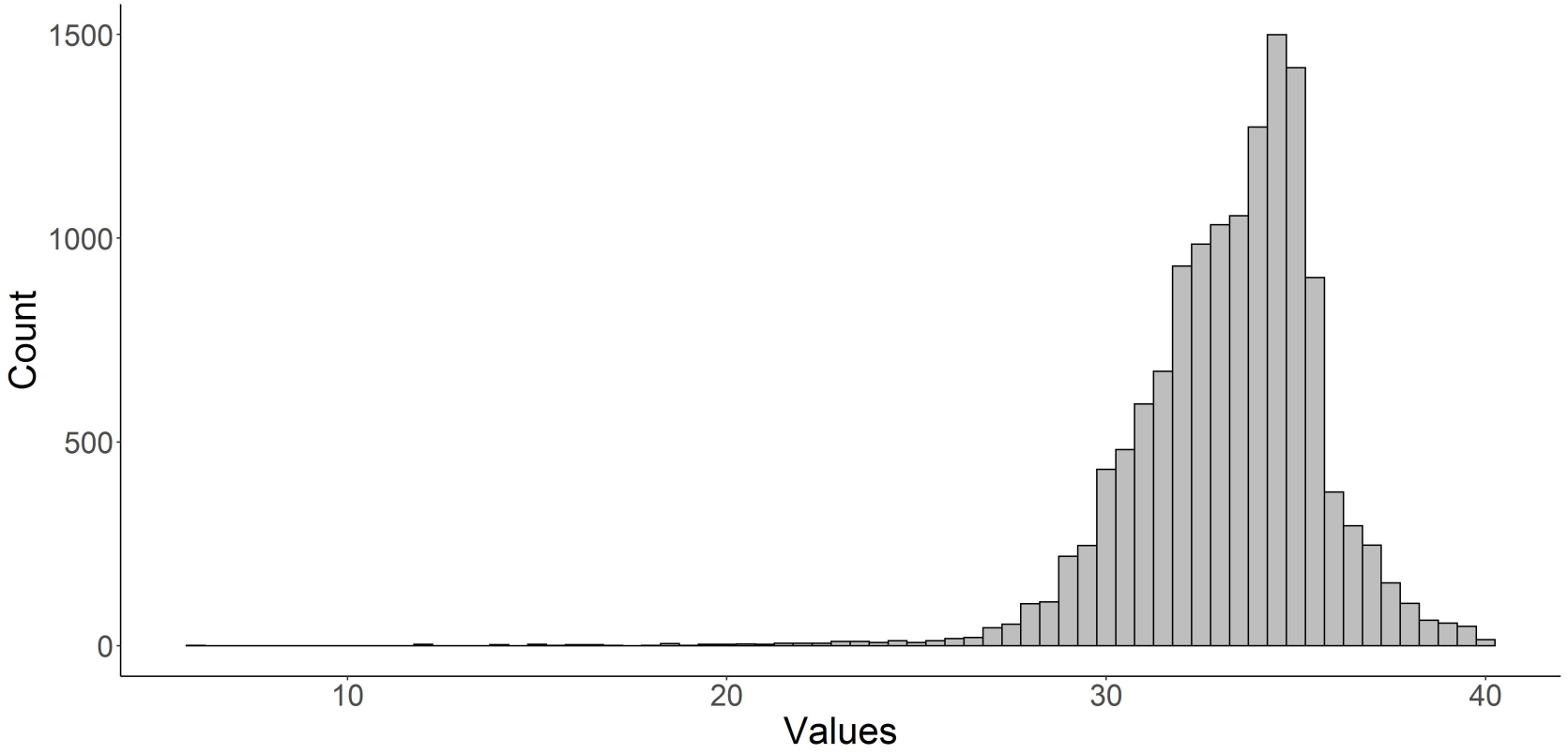


4 sources

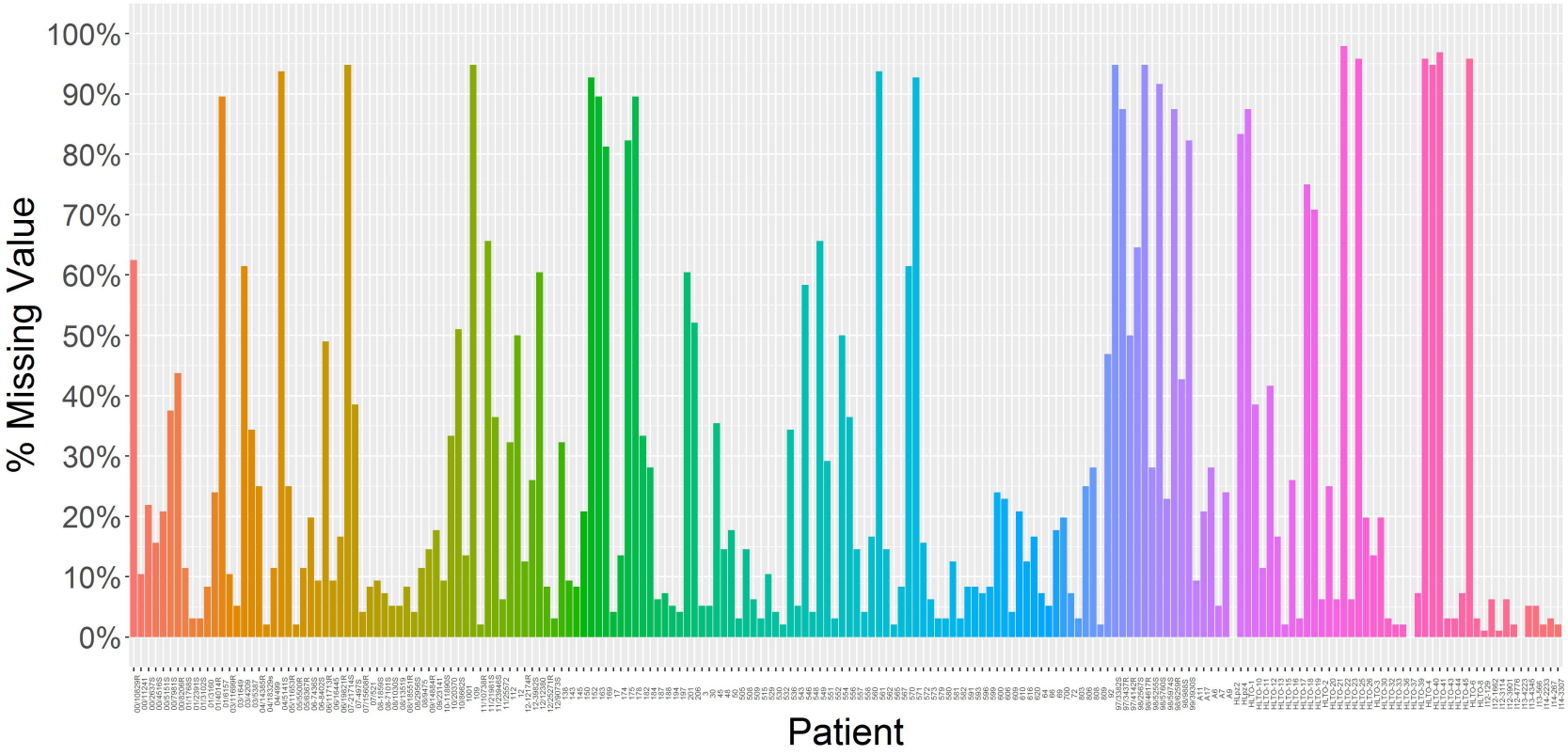
Work schedule on Raw data



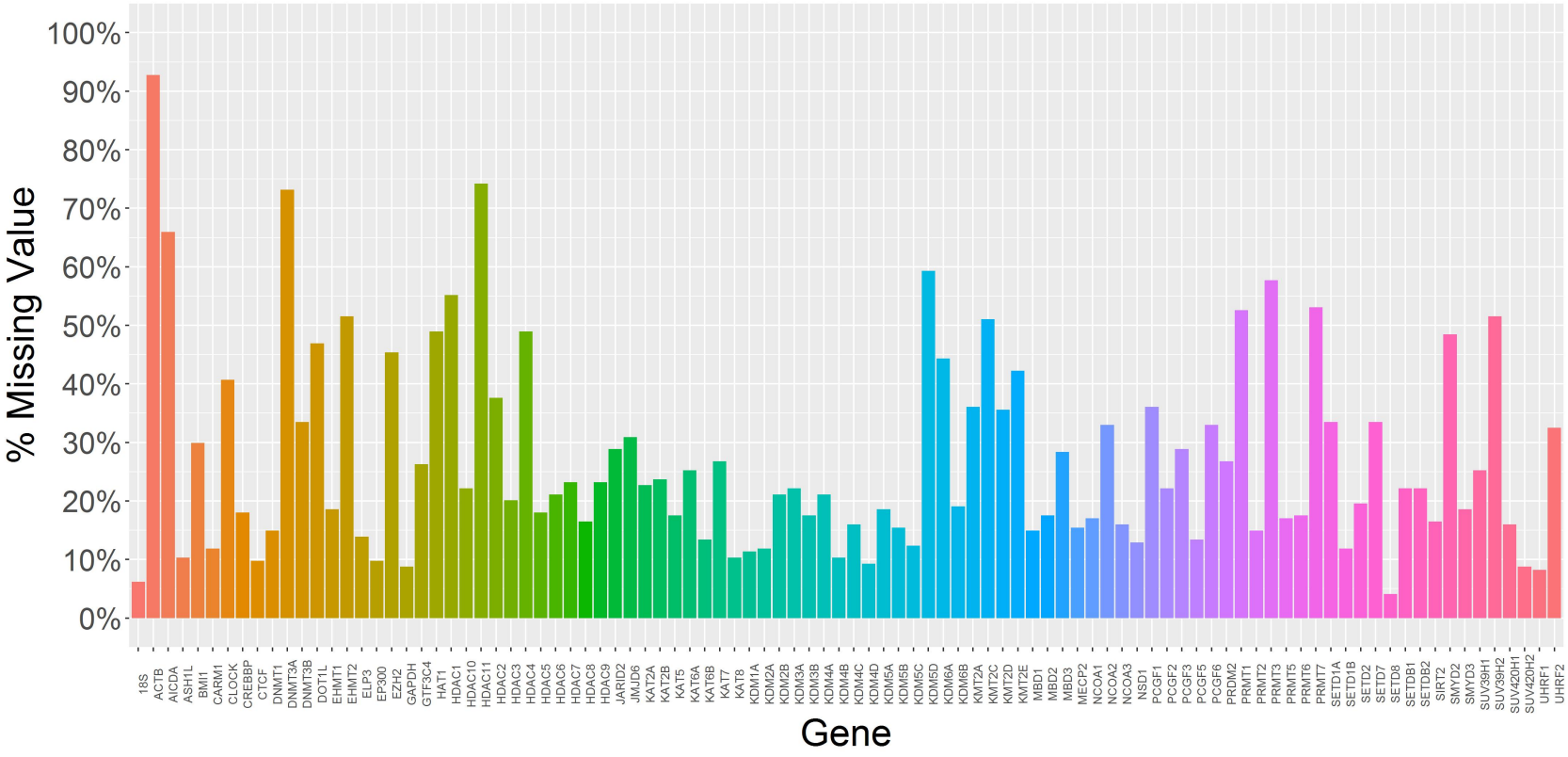
Preliminary Data Analysis:
Data Distribution



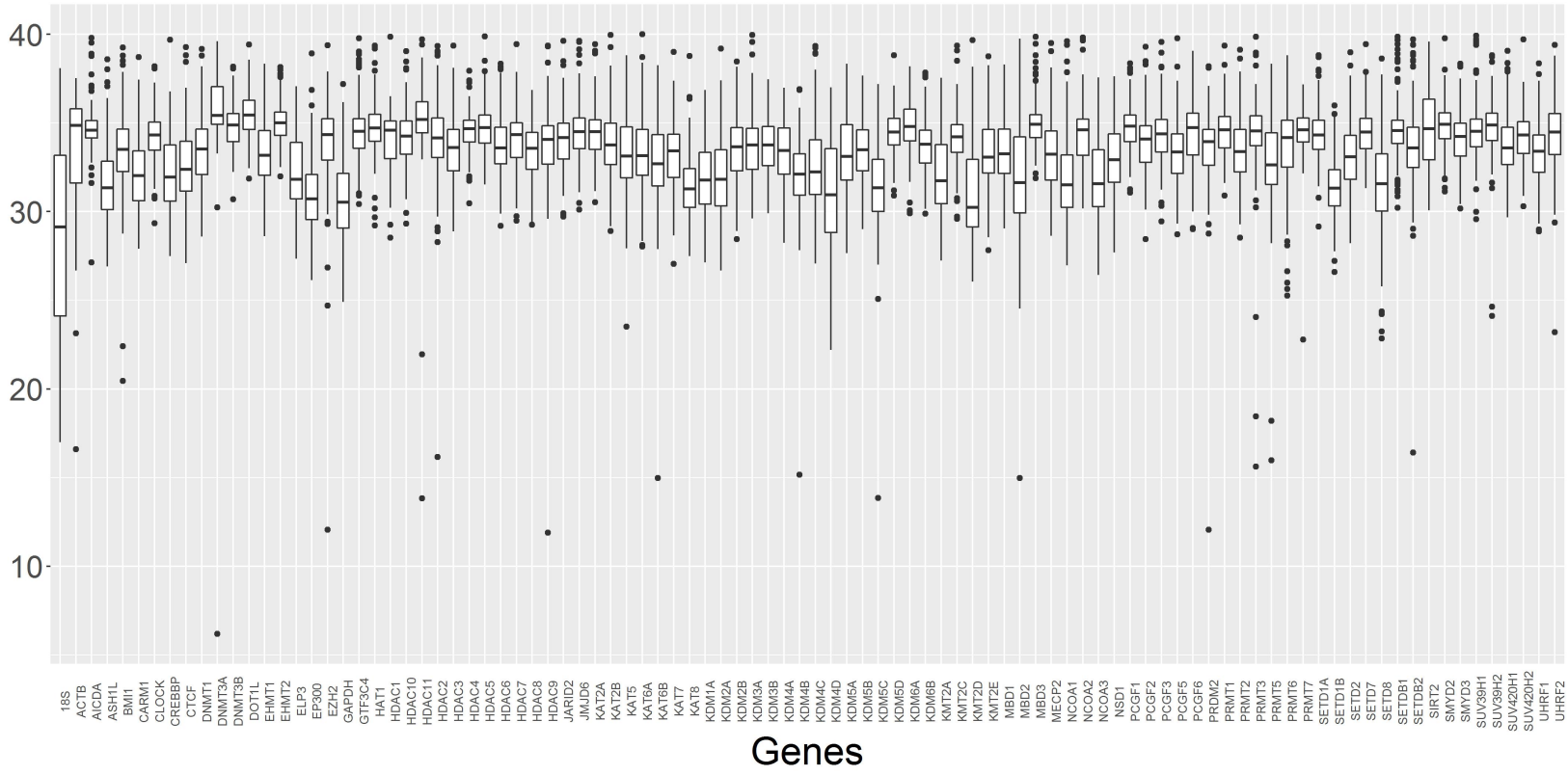
Preliminary Data Analysis:
Missing data for Patient



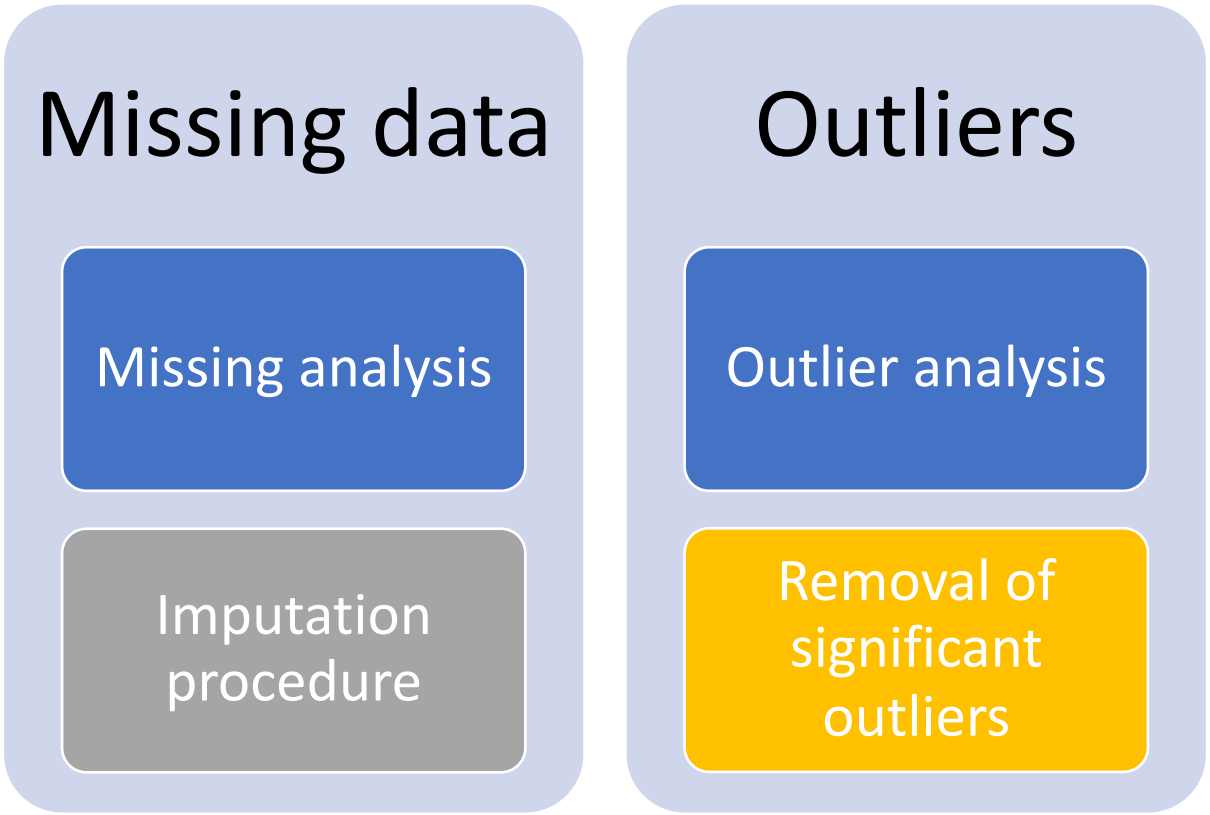
Preliminary Data Analysis:
Missing data for Gene



Preliminary Data Analysis:
Outliers



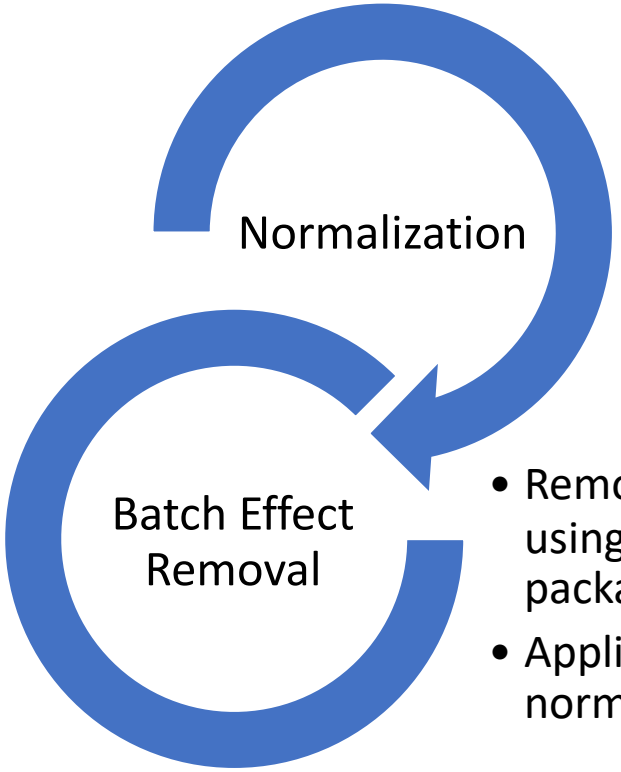
Preliminary Data Analysis



Data Cleaning:
Identification of Normalization Factor (NF)

Literature	• GAPDH • Literature on Hodgkin Lyphoma (GAPDH and ACTB)
Original Data	• Pairwise Comparison Approach (GeNorm) • Model-based Approach (NormFinder) • Pairwise Correlation Approach (BestKeeper)
Combined Ranking	• Pairwise Comparison Approach (GeNorm) • Model-based Approach (NormFinder) • Pairwise Correlation Approach (BestKeeper)

Data Cleaning:
Normalization and Batch Effect Removal



- Normalize genomic data using NF
- Construction of 8 distinct datasets, one for NF

- Removal of the Source effect using *ComBat* function of *sva* package
- Application on all the 8 normalized datasets

Classifier Construction

Best classifier

- NF obtained by GeNorm method
- Classification method: CART

	Patient	Genes
1	194	96

	R	S
R	26	31
S	30	101

	Accuracy	Kappa	AUC
1	0.6546	0.2422	0.7116

	TPR	FPR	Precision	F1
R	0.4643	0.2348	0.4561	0.4602
S	0.7652	0.5357	0.7710	0.7681

Classifier Constructon with Source

Best Classifier

- NF: GAPDH (literature)
- Classification method: CART

	Patient	Genes
1	194	96

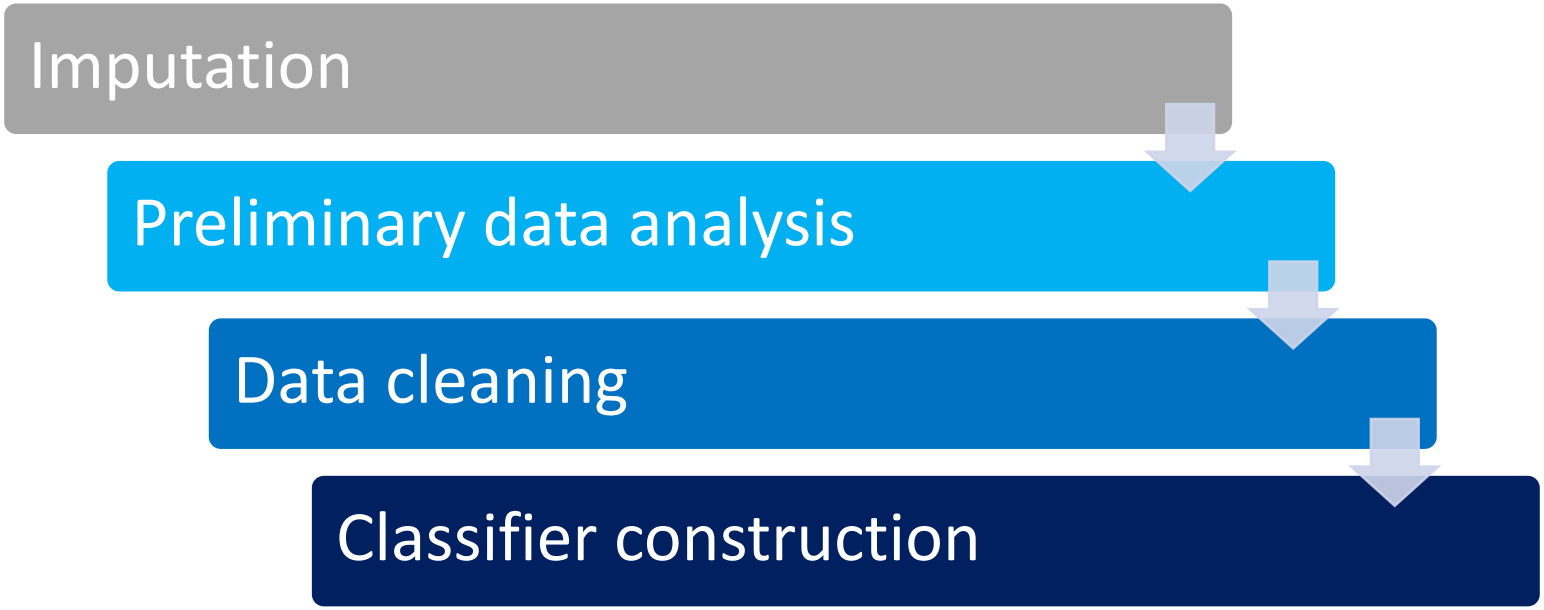
	R	S
R	27	18
S	34	115

	Accuracy	Kappa	AUC
1	0.7320	0.3308	0.6739

	TPR	FPR	Precision	F1
R	0.4426	0.1353	0.6000	0.5094
S	0.8647	0.5574	0.7718	0.8156

Classification methods dealing with group effect

Work schedule:
Missing values issue



Imputation

Random uniform values in [40,45]

- Data concentration at high values
- Generation of fictitious outliers

Fixed value at 40

- Data concentration at 40
- Abnormal trends become regular

Mean of gene

- Data concentration in central band
- Variance reduction

Median of gene

- Data concentration in central band
- Variance reduction

Multiple Imputation Method (*mice* package)

- Imputation with observed data (more realistic)
- Influenced by significant low outliers

Classifier Construction

Best classifier

- Imputation: *mice* package
- NF obtained by BestKeeper method on Final Ranking
- Classification method: RandomForest

	Patient	Genes
1	194	96

	R	S
R	24	8
S	37	125

	Accuracy	Kappa	AUC
1	0.7680	0.3825	0.7309

	TPR	FPR	Precision	F1
R	0.3934	0.0602	0.7500	0.5161
S	0.9398	0.6066	0.7716	0.8475

Classifier Construction with Source

Best classifier

- Imputation: *mice* package
- NF obtained by BestKeeper method
- Classification method: RandomForest

	Patient	Genes
1	194	96

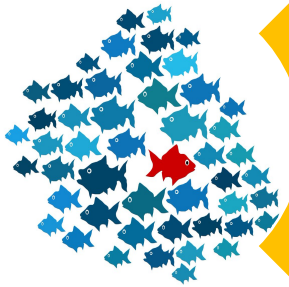
	R	S
R	25	9
S	36	124

	Accuracy	Kappa	AUC
1	0.7680	0.3887	0.7244

	TPR	FPR	Precision	F1
R	0.4098	0.0677	0.7353	0.5263
S	0.9323	0.5902	0.7750	0.8464

Classification methods dealing with group effect

Next Steps



Removal of significant outliers

Classification methods dealing with group effect