



Ying-Cheng Chris Lee (leechris), Marcos Prunello (marcosp)

### INTRODUCTION

- CANCER:** diseases characterized by an uncontrolled growth in cells which can extend to different parts of the body, generally caused by changes in DNA
- EPIGENETICS:** variations that affect gene expression without altering the DNA sequence
- CANCER & EPIGENETICS:** changes on top of DNA can activate oncogenes or deactivate tumor suppressor genes
- EPIGENOME PROFILES:** Large multi-cancer data sets are available that profile the epigenome for tumor tissue samples. However, the samples are made up of a mixture of cell types and it's necessary to determine to which extent the samples are composed by tumor cells or by other cell types, such as blood cells
- DNA METHYLATION (MET):** chemical modification of DNA, responsible for cellular differentiation, can be used to distinguish distinct cell lineages. Using methylation signatures it is possible to estimate the distribution of blood cell types in a sample
- GOAL:** Use methylation data to predict the frequency of several types of blood cells present in tumor tissue samples; identify genes whose methylation pattern is important for this task

### DATA

#### TUMOR SAMPLES

- Glioblastoma (GBM): most common and aggressive malignant primary brain tumor
- Gene expression and DNA methylation data downloaded from The Cancer Genome Atlas (TCGA).
- Data was pre-processed to deal with missing values and batch effects.
- Data size: 321 samples and 17812 genes (our predictors)

#### RESPONSE FEATURES

Frequency of each of 7 blood cell subtypes

- B cells
- CD4
- CD8
- Natural Killers (NK)
- Neutrophils (Neu)
- Eosinophils (Eos)
- Monocytes (Mon)

Obtained performing gene expression deconvolution with Cibersort

### FEATURE SELECTION: METHYLMIX

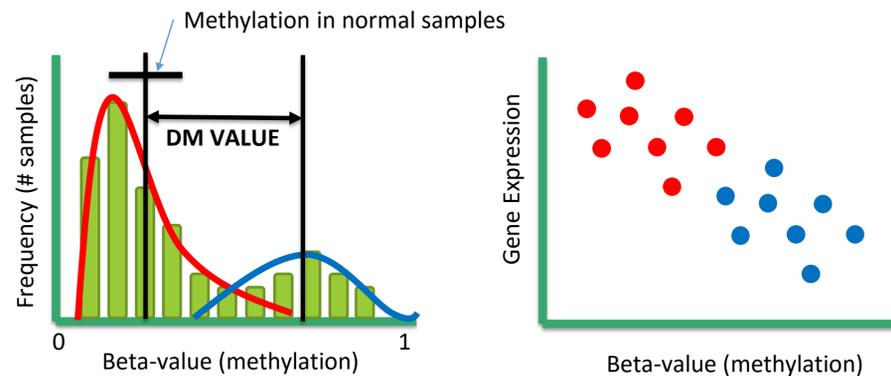
- We reduced our set of 17812 genes using MethylMix, a method that identifies hypo- and hyper-methylated genes which are also transcriptionally predictive, following these steps:

1. A **beta mixture model** is fitted to MET data from tumor samples using the **EM algorithm**, to identify different patterns ("MET states") for each gene

2. Each of these MET states is compared to DNA MET of normal tissue to determine if the gene is hypo- or hyper-methylated. The **DM-value** is calculated as the difference between the cancer and the normal MET state.

3. Genes with differential MET states and whose MET is predictive of its gene expression profile, are called "**MET-driven genes**".

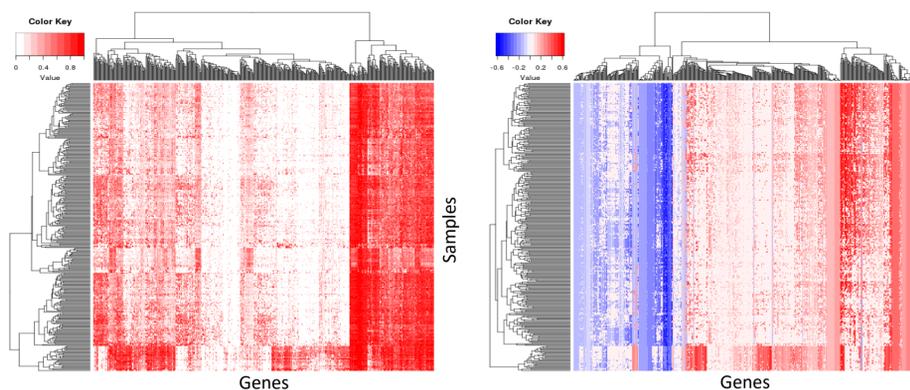
### METHYLMIX ALGORITHM



WE SELECTED FOR OUR PREDICTIVE MODELS THOSE GENES THAT ARE POINTED BY METHYLMIX AS METHYLATION-DRIVEN GENES

- Number of genes: 455
- We trained our models with these genes, and with both raw MET data and DM-values data (to overcome noise in raw MET data)

### HEATMAPS OF DATA



### METHODS FOR PREDICTION

- We divided our data sets in a training set and test set.
- We applied several machine learning algorithms to predict each blood cell frequency in both raw MET and DM values training sets.
- We tuned the parameters required for each method with 10 fold cross-validation, and re-fitted the model with the best performance in the whole training set.
- We predicted the responses in the validation set, and used mean squared error to evaluate the performance.

Linear regression with L1 penalization (LASSO)	K Nearest Neighbors	Basic Regression Trees
Bagging (for regression trees)	Random Forest	Boosting (for regression trees)
Support Vector Regression: No Kernel	Support Vector Regression: Radial kernel	Support Vector Regression: Polynomial Kernel

x 7 cell subtypes x 2 data types

### RESULTS

- For each cell subtype, we selected the best performing model according to the Mean Square Error (MSE) in the test set.
- Looking at the top results for each cell subtype, there is no general trend regarding to which is the method that works better in general, or which among raw data or DM values is better.

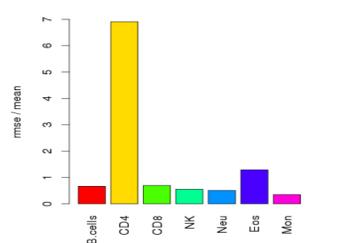
#### BEST MODEL FOR EACH SUBTYPE ACCORDING TO MSE TEST

Subtype	Obs Mean	Data	Model	MSE Test	MSE Train	Cor Test	Cor Train
B cells	0.0941	Raw	SVR Linear	0.0043	0.0020	0.5450	0.7915
CD4	0.0045	Raw	Lasso	0.0007	0.0003	0.3907	0.8812
CD8	0.1733	DMvalues	Random Forests	0.0181	0.0032	0.3948	0.9877
Eos	0.1404	Raw	Boosting	0.0045	0.0039	0.1286	0.5590
Mon	0.2449	DMvalues	SVR Linear	0.0095	0.0089	0.2716	0.6221
Neu	0.0547	Raw	SVR Radial	0.0143	0.0076	0.4331	0.8581
NK	0.2882	Raw	SVR Polynomial	0.0053	0.0038	0.6236	0.7769

#### IMPORTANT GENES

- We identified for each of the models which are the genes that most contribute to the prediction
- There were no genes identified as important genes for all the subtypes simultaneously.
- For each cell subtype, the top 5 most important genes are:

MSE RELATIVE TO RESPONSE MEAN



Subtype	Top 5 important genes				
B cells	HSD17B14	ZNF329	PLA2R1	MT2A	PDGFRA
CD4	ABCA5	ABHD8	ACAA2	ACN9	ACOT4
CD8	TNFRSF1A	KLHL26	PSMB8	CTS2	OAS1
Eos	CRYBB1	DDX43	C17orf76	ACSBG1	ANKRD9
Mon	KLHL34	SMOC1	RDH5	CABYR	ZNF135
Neu	NETO1	STAG3	ME1	FOXD1	CTHRC1
NK	ZNF549	BCL11A	C9orf95	ZNF781	DUSP23

### DISCUSSION & CONCLUSION

- We built models that are able to **predict** the **frequency** of each blood cell **subtype** in samples from tumor tissue using only **methylation** data.
- Using DMvalues showed no improvement over using raw MET data.
- From the inspection of the results, it is clear that there is wide opportunity for improvement of the performance of this prediction task.
- Some suggestions that should be addressed:
  - Feature selection: implement other algorithm not independent of the response, exploring all the 17812 genes.
  - GBM data set is small, with only 321 samples, evaluate performance in larger data sets.
  - Incorporate in the models the constraint that the response is a proportion can only take values from 0 to 1.
  - Predict all the frequencies together with a multivariate response, since they are all related to each other