RNAge

Tissue-level Gene Expression Analysis of Aging

https://github.com/edgeslab/cs418-project-RNAge

Team information

Ben Imlay	bimlay2@uic.edu	imlayb
Farid Manuchehrfar	fmanuc2@uic.edu	fmanuc2
Sivaraman Lakshmipathy	slaksh5@uic.edu	sivaramanl
Aditya Ramakrish	aramak24@uic.edu	aditya1793

Problem

Numerous genes contribute to the highly complex process of aging. Our team, RNAge, wants to leverage RNA-sequencing data towards a deeper understanding of the gene expression trends that underpin aging.

- ➔ Current research only focuses on tissue-specific aging or tissue independent aging.
- ➔ Existing knowledge gap in the *comparison* of aging in different tissues.

Goals

- ➔ Characterize and compare the relationship of different tissues with aging.
- ➔ Find significant genes participating in aging across different tissues.
- ➔ Find the differences in aging in healthy and diseased tissues.

Significance

- ➔ Identification of significant genes can aid the study of age-related diseases.
- ➔ Elucidation of the tissue-specific dynamics of aging could change the paradigm of aging.

Data

RNA-sequencing quantifies gene expression from the abundance of the different mRNA in a sample.

GTEx is an initiative to characterize non-diseased tissue from cadaver subjects.

- ➔ The current gene expression data consists of:
  - ◆ 752 subjects comprising 53 tissues.
  - ◆ TSV format → 56,202 (genes) x 11,688 (samples) matrix.
- ➔ Donor metadata: age group, sex, death type
- ➔ Tissue sample metadata: tissue origin.

Data source: *https://gtexportal.org/home/datasets*

Data summary: *https://gtexportal.org/home/tissueSummaryPage*

Data Cleaning

1- Split Data by Tissue Type:

Easier to Use

2- Age Existence:

Remove Samples without Age

3- Tissue Count:

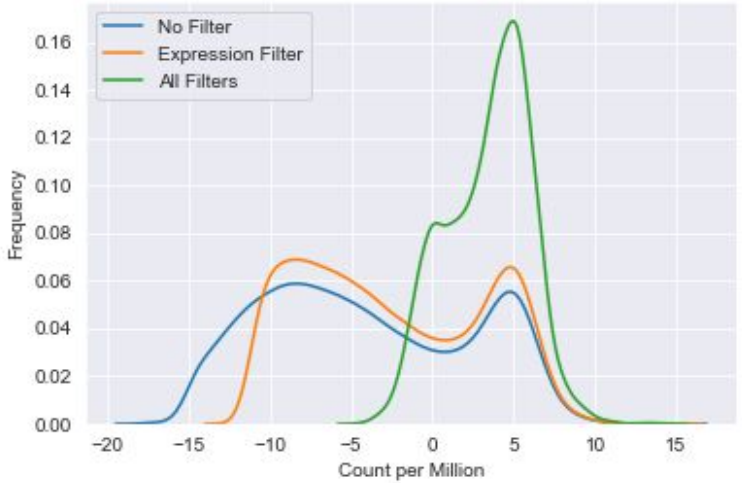
Use Tissues with more than 200 samples

4- Variance and Gene Expression:

Filter Genes with low Variance and low Gene Expression

5- Row Count Filter:

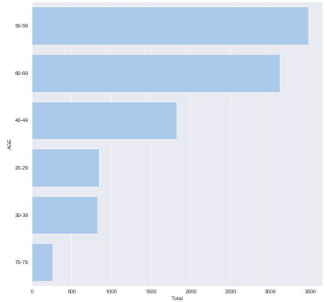
Remove Samples with low row Counts

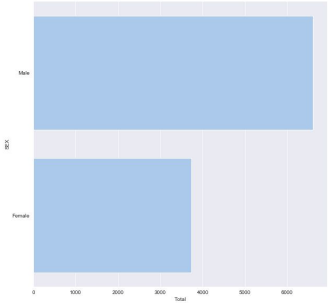


Data EDA

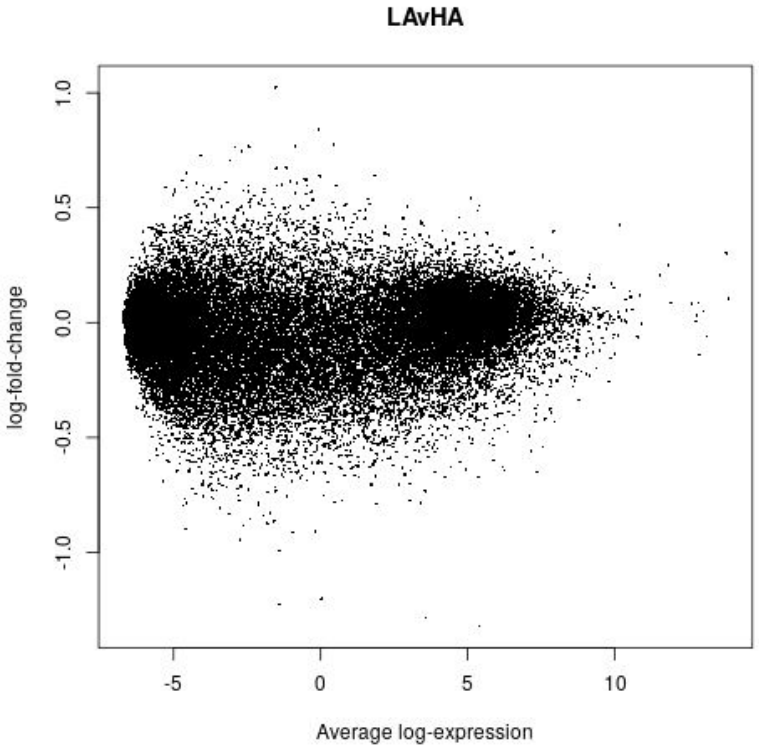
Data Structure	Imported as Pandas DataFrame.
Data Granularity	For each samples and for each tissue we have age group (Ordinal Data), sex (Ordinal), Gene expression (Quantitative), and death hardy (Ordinal) information.
Temporality	Data does not change as it is from dead samples.
Faithfulness	Based on the Data source, it is approved by FDA and is trustful.
Scope	Some data does not have age and death hardy information; it is processed and cleaned.

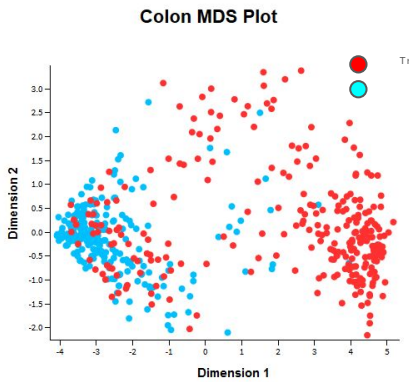
	SAMPID	SMTS	SEX	AGE	DTHHRDY
0	GTEX-1117F-0226-SM-5GZZ7	Adipose Tissue	2.0	60-69	4.0
1	GTEX-111CU-1826-SM-5GZYN	Adipose Tissue	1.0	50-59	0.0
2	GTEX-111FC-0226-SM-5N9B8	Adipose Tissue	1.0	60-69	1.0
3	GTEX-111VG-2326-SM-5N9BK	Adipose Tissue	1.0	60-69	3.0
4	GTEX-111YS-2426-SM-5GZZQ	Adipose Tissue	1.0	60-69	0.0

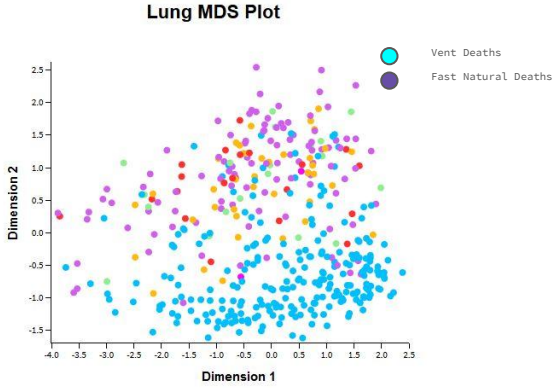




Solution - DGE & Dim reduction







- No differentially expressed genes detected (p=.05 BH cor.)

$$LAvs.HA = \frac{(A1+A2+A3)}{3} - \frac{(A4+A5+A6)}{3}$$

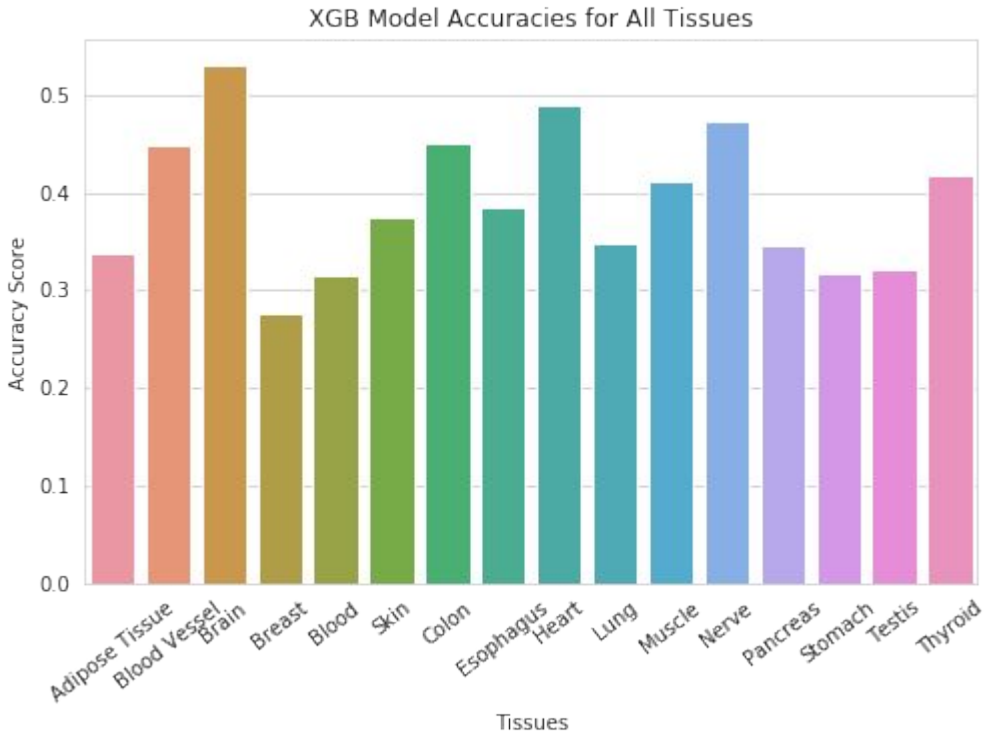
Solution – Machine Learning

Setup:

- Classification problem:
  Classify the age group using the gene expression.
- 6 classes:
  20 - 79 in buckets of size 10.
- 16 tissue types:
  Model trained for each tissue type separately.
- Evaluation:
  Accuracy of the trained model.
- Baseline:
  Majority classifier

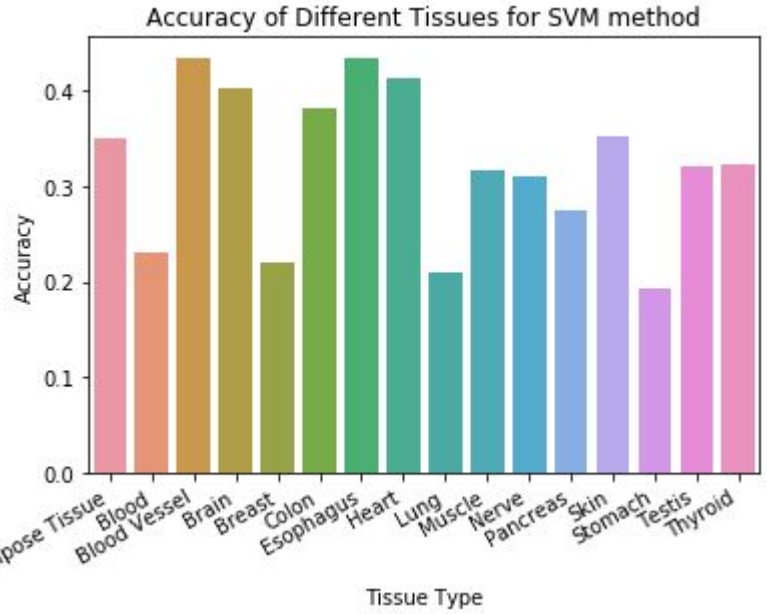
Solution - Machine Learning

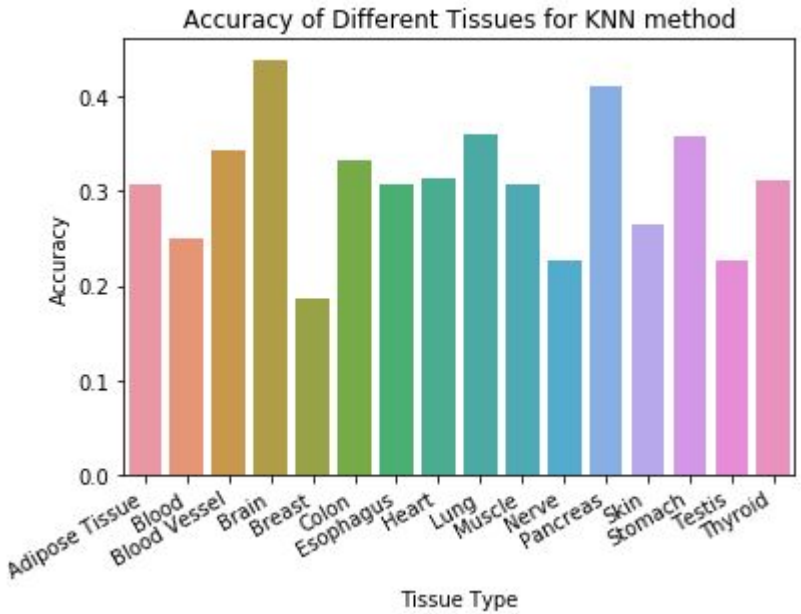
XGBoost



Solution - Machine Learning

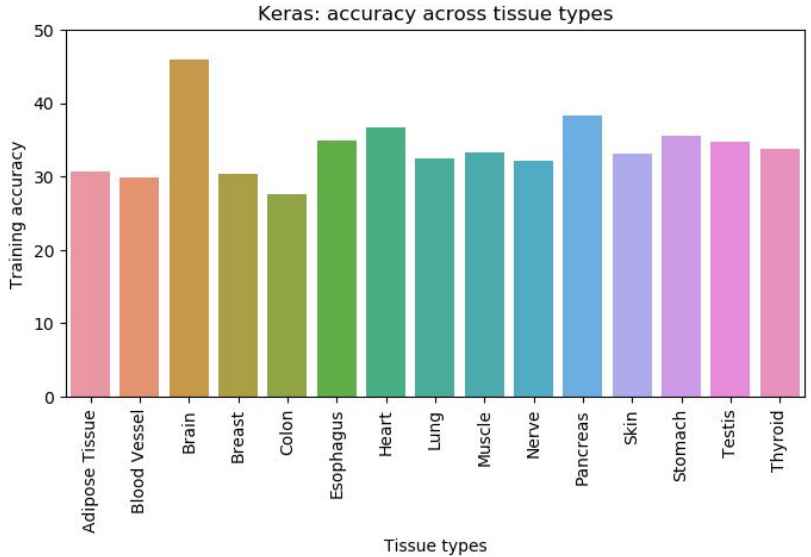
SVM and KNN





Solution - Deep Learning models

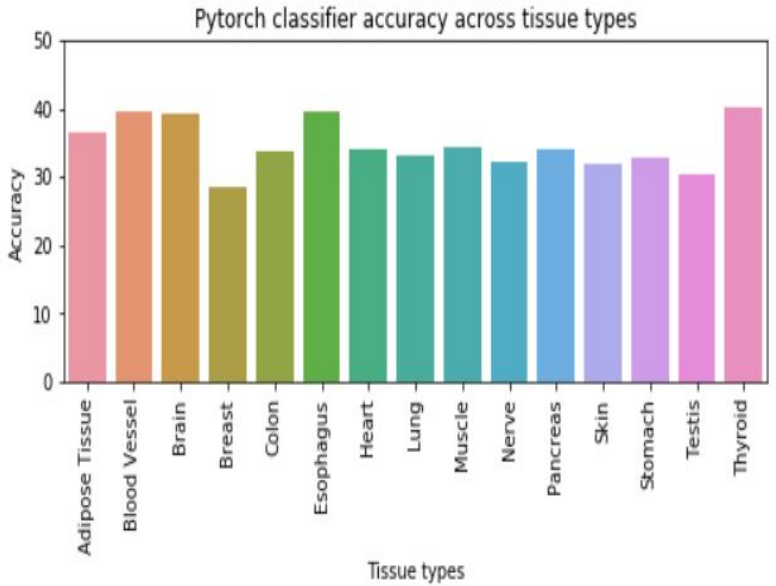
Tensorflow Keras and Pytorch



- 3 hidden layers
  - 1st hidden layer - 1024 neurons
  - 2nd hidden layer - 512 neurons
  - 3rd hidden layer - 256 neurons
- Hyperparameter tuning:
  - Early stopping

Solution - Deep Learning models

Tensorflow Keras and Pytorch



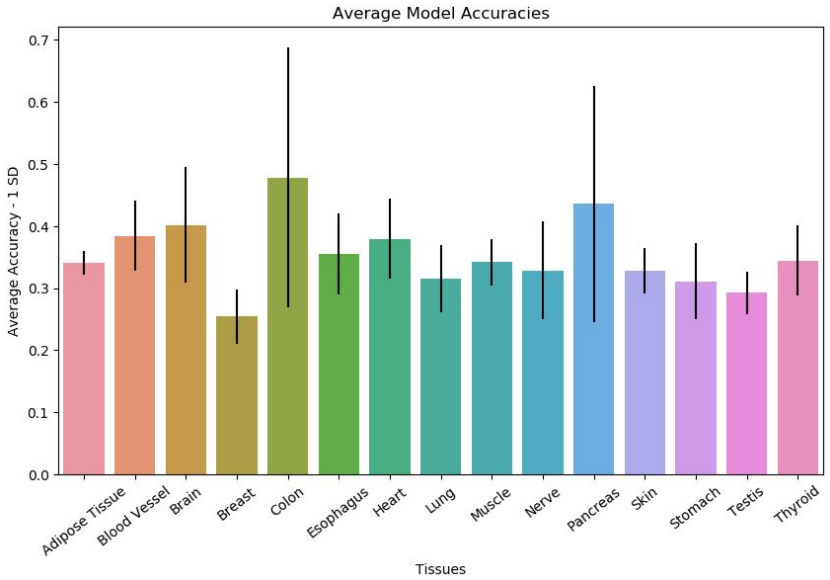
- 3 hidden layers
  - 1st hidden layer - 15 neurons
  - 2nd hidden layer - 10 neurons
  - 3rd hidden layer - 5 neurons
- Hyperparameter tuning:
  - None

Solution - Takeaways

Challenges:

- Imbalanced features vs sample counts:
  - Regularization
  - PCA (Principal Component Analysis)
- Data Normalization:
  - Protein chaining
  - Min max normalization
- Variable accuracy ranges across models:
  - Ensemble (weighted average)

Primary Hypothesis

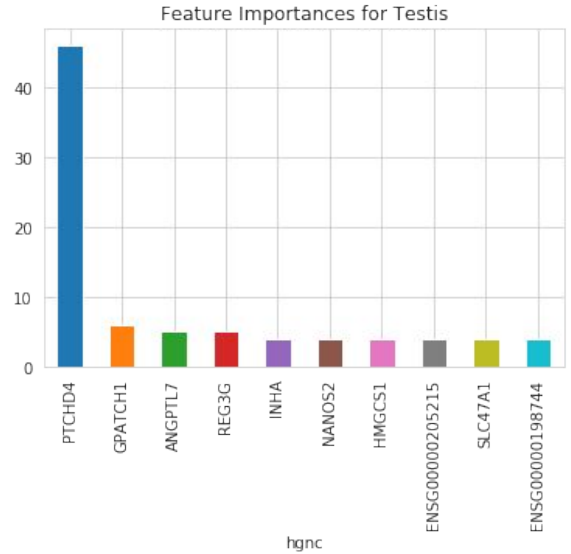


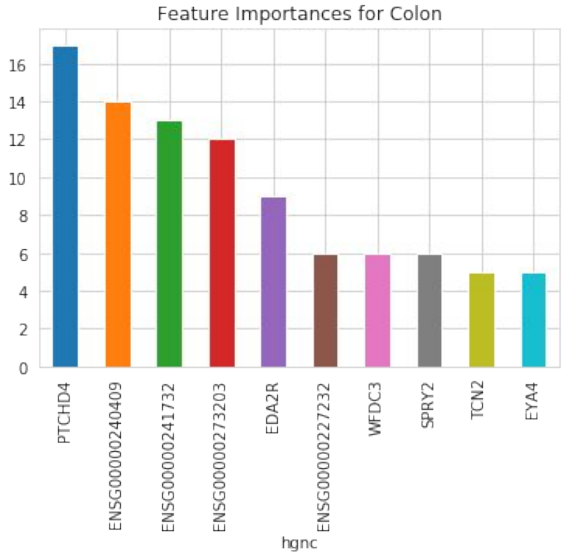
Mann-Whitney U test (Holm Sidak cor.)

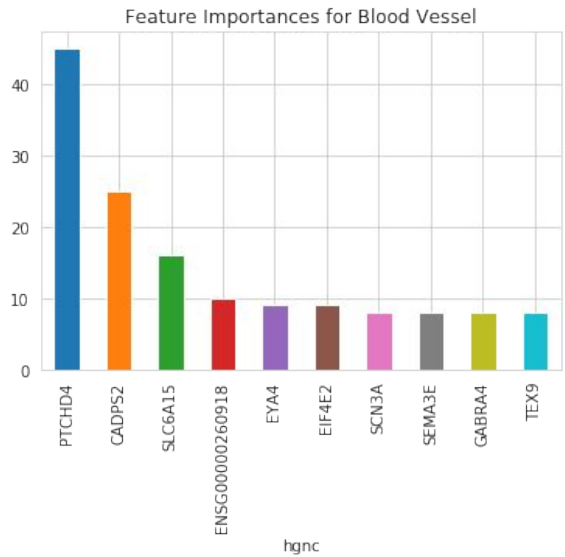
	statistic	p-value	cor_p-value
Heart_Breast	25.0	0.012186	0.724003
Muscle_Breast	25.0	0.012186	0.724003
Adipose Tissue_Breast	25.0	0.012186	0.724003
Testis_Colon	0.0	0.012186	0.724003
Colon_Breast	25.0	0.012186	0.724003
Blood Vessel_Breast	24.0	0.021572	0.887047
Adipose Tissue_Testis	23.0	0.036714	0.975352
Testis_Heart	2.0	0.036714	0.975352
Nerve_Breast	22.0	0.060103	0.997552
Pancreas_Breast	22.0	0.060103	0.997552

Tissue gene expression levels predict age, and different tissues are unequal predictors of age.

XGBoost Feature Importance - PTCHD4



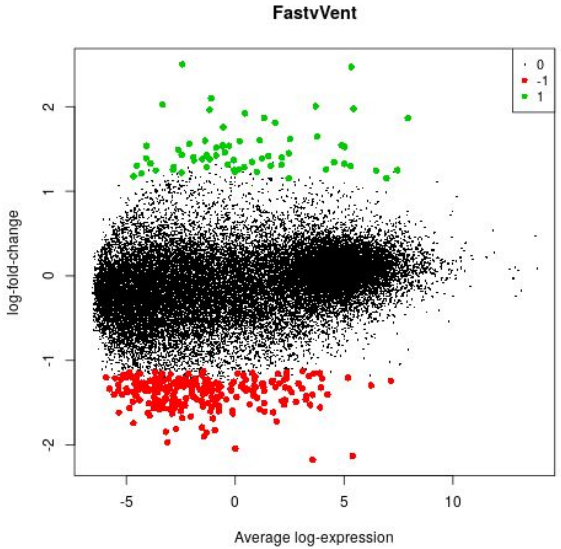


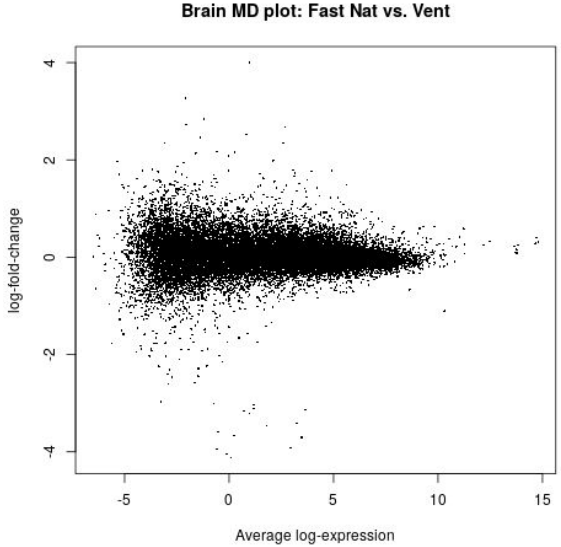


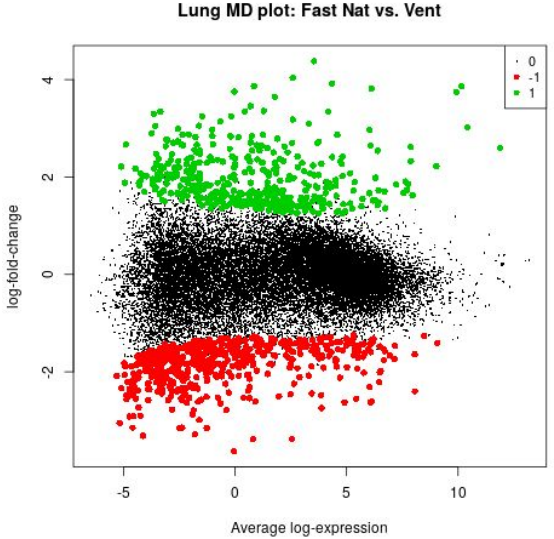
- Repressor of hedgehog signaling → differentiation.

uniprot

Death Type


FastvVent
log-fold-change
Average log-expression
0
-1
1


Brain MD plot: Fast Nat vs. Vent
log-fold-change
Average log-expression


Lung MD plot: Fast Nat vs. Vent
log-fold-change
Average log-expression
0
-1
1

uniprot

Death Type – Biological Insight

	adj.P.Val	hgnc
ENSG00000248187	3.742670e-98	NaN
ENSG00000177575	2.677617e-94	CD163
ENSG00000096060	6.637916e-75	FKBP5
ENSG00000236047	9.670191e-71	NaN
ENSG00000198570	3.300247e-52	RD3
ENSG00000270640	5.394276e-52	NaN
ENSG00000120088	3.588255e-43	CRHR1
ENSG00000188536	2.721721e-40	HBA2
ENSG00000264590	9.257567e-40	NaN
ENSG00000267653	1.807562e-39	NaN
ENSG00000196136	2.009480e-37	SERPINA3
ENSG00000269929	4.766919e-35	NaN
ENSG00000244734	2.703361e-33	HBB
ENSG00000236279	3.818054e-33	CLEC2L

- CD163 induced by intravenous lipopolysaccharide.

Thank you!