RNAge Final Report

https://github.com/edgeslab/cs418-project-RNAge/blob/master/Final_Report.ipynb

In [5]:
```
#Setting up the environment and importing the necessary data
%run Final_source_code/Init.py
%matplotlib inline
```

Introduction

Omics are the holy grail of biological discovery. In short, it is the investigation of a certain class of components in a biological smaples. Genomics refers to the study of the whole genome, transcriptomics the whole of the RNA transcripts, and proteomics the whole of the proteins. These are only a few of the *omics*. This approach enables investigation of cellular processes with greater scope than ever before.

Our team of interdisciplinary students knew from the onset that we wanted to leverage this approach towards biological question. Rather than picking a mainstream topic such as cancer, our team chose to investigate human aging on a transcriptomic level. Our fundamental questions entering this project were as follows:

- Are the gene expressio profiles of different tissues equally associated with aging?
- What genes are involved in aging in different tissues?

These are biologically significant questions as pre-omics research has focused on specific genes of interest or tissue-specific aging. We intend to compare the aging process between tissues and identify associated genes on an omic level.

Data

Source, Acquisition, Structure, and Granularity

Our team used the Genotype-Tissue Expression (GTEx) project for our transcriptome data. Their efforts consist of genotyping and RNA-sequencing 11,688 tissue samples from 752 subjects across 53 tissues. This means our data is of **fine granularity** on a sample level, but the data is somewhat **coarse** on a subject level. Multiple samples are from single subjects, which means each sample is not entirely independent from one another.

We downloaded the data as the final gene-transcript counts, the number of reads experimentally aligned to a given gene. This looks like a **sparse** matrix with 56,202 features (genes). We also obtained sample metadata for the 11K samples and subject metadata for the 752 subjects.

Data source: https://gtexportal.org/home/datasets (https://gtexportal.org/home/datasets)

- Gene counts (https://storage.googleapis.com/gtex_analysis_v7/rna_seq_data/GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_reads.gct.gz): 11,688 x 56,202 TSV
- Sample metadata (https://storage.googleapis.com/gtex_analysis_v7/annotations/GTEx_v7_Annotations_SampleAttributesDS.txt): 11,688 x 63 TSV & Features are related to quality control and experimental protocol.
- Subject metadata (https://storage.googleapis.com/gtex_analysis_v7/annotations/GTEx_v7_Annotations_SubjectPhenotypesDS.txt): 752 x 4 TSV (Subject ID, AGE groups, Sex, Hardy Scale death type)

EDA Principles

- Temporality: Samples range from 2011 to 2015 from a variety of medical centers in the United States.
- Faithfulness: All samples were processed using Illumina TrueSeq RNA sequencing, a polyA selection method. This is an accepted and widely used method.
- Scope: All samples are derived from cadavers of unknown genotype, meaning any analysis is blind to actual disease and other confounding biological effects.

Pre-processing

Our first objective in pre-processing was to merge subject metadata to sample metadata to enable age-based analysis. The following script merges and reorders the metadata according to gene count data. A `merged_meta.tsv` file for further analysis is output to `data`. For simplicity, merged_meta.tsv is packaged with the repository and can be simply copied into `data`.

In [6]:
```
#%run Final_source_code/dataSplit.py
!cp merged_meta.tsv data/
```

Our second objective was to normalize and split the gene count data into single tissue data sets. The raw counts were normalized in `R` using the *trimmed mean of M-values* (TMM) method in the `edgeR` package. This left us with normalized counts in the form of counts-per-million.

The results of normalization and splitting can be obtained here (https://drive.google.com/file/d/1k06AGmZngzlpDBBAV4bOXgoB5iUUiiin/view?usp=sharing). The TSV files in this zip should be placed in `data/tissue-specific`.

In [7]:
```
#Running this script requires the following in R 3.5.2: Plyr, edgeR, limma, Glimma, data.table, stringr, foreach
#!Rscript Final_source_code/GTEx_input_final.R
```

Data Cleaning

Our *Data Cleaning* has four general steps resulting in the reduction of tissues to 16 and features (genes) to ~20,000.

1. Remove samples without age.
2. Choose tissues with counts more than 200.
3. Filter genes with low variance and low raw count sums.

In [20]:
```
%run Final_source_code/DataCleaning.py
#The cleaned data for tissue type = COLON is as shown below
```

```
Original Gene Count: 56202
Expression Filter Gene Count: 44995
+Variance Filter Gene Count: 18572
```

	ENSG00000227232	ENSG00000237683	ENSG00000241860	ENSG00000228463	ENSG00000225972	ENSG00000225630	E
GTEX-14PHX-1126-SM-5YYA5	7.587366	4.917485	0.381412	0.572117	0.871798	207.610493	
GTEX-R53T-1326-SM-48FCQ	8.105863	0.899066	0.442397	0.000000	0.956149	246.786424	
GTEX-18A66-2426-SM-7KFSQ	5.911802	4.944861	0.660947	0.085678	1.897162	298.711234	
GTEX-14JG6-1626-SM-5YYBC	10.808170	1.883867	0.962560	0.165010	1.182573	481.857369	
GTEX-ZY6K-1226-SM-5GZYL	9.580124	1.302486	0.445587	0.257070	2.022280	345.947027	

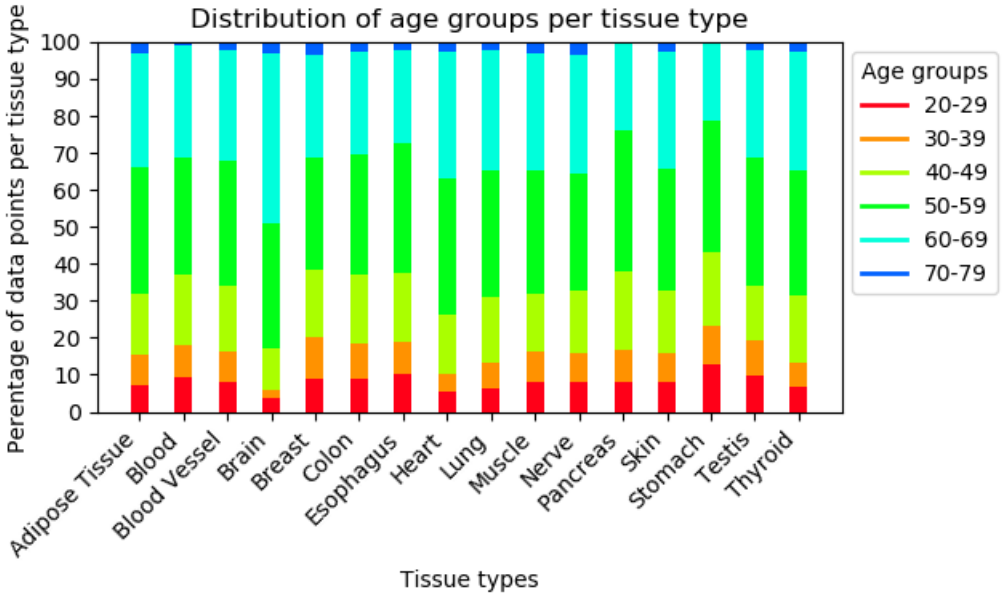
5 rows × 18572 columns

Data Analysis and Exploration

Visualization #1: Distribution of Age

We knew that some tissues had very few samples. To begin testing the relationships of the data, we needed to establish a threshold of minimum number of samples per tissue. This visualization shows us how the distribution of age groups for each tissue is relatively consistent. That is, the 40, 50, and 60 age groups are the most prevalent. This plot helped guide our decision to use only tissues with 200 samples as that always left more than one individual in each age group.

In [8]:
```
import matplotlib as mpl
from matplotlib.lines import Line2D
%run Final_source_code/AgeGroupViz.py
```

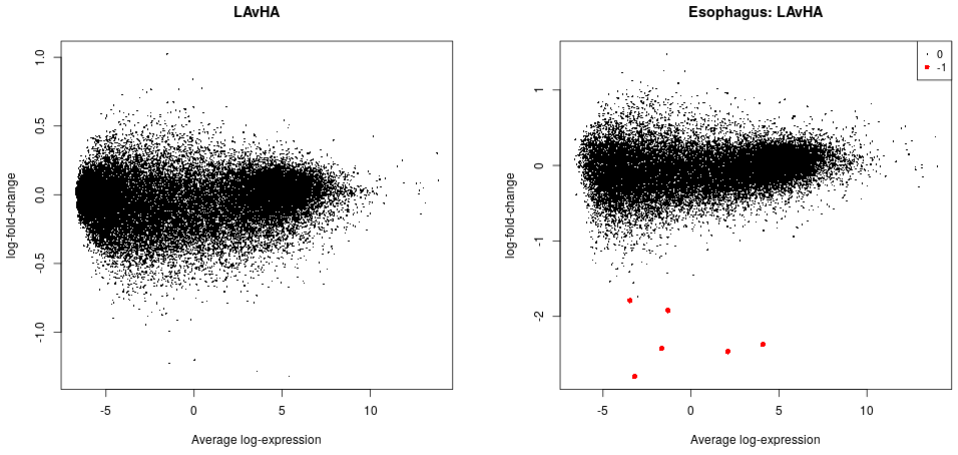


Visualization #2: Differential Gene Expression Analysis

Differential Gene Expression analysis is essentially a mean-difference test with a variance model for each gene. $LAvs.HA = \frac{(A1+A2+A3)}{3} - \frac{(A4+A5+A6)}{3}$ where LA means low-age and HA means high age, and where our null hypothesis for each gene was there is no difference in the gene expression between the two populations.

Unfortunately, there are no tools in Python to accomplish the mean-dispersion estimation/transformation needed to run statistical tests on RNA-seq count data. Instead, we used `R` and the `edgeR`, `Limma`, and `Glimma` packages. DGE results and mean-difference plots in their entirety are located in `DGE_results/`. Full results including interactive MD-plots and multi-dimensional scaling (MDS) plots can be obtained here (https://drive.google.com/file/d/1X4OIkk_C9xLgfcfSap2LhsOcUnaP1sF5/view?usp=sharing). The heart MDS-plot (./progress_plots/Heart/MDS-Plot.html) is included with this repository.

In [9]:
```
#!Rscript Final_source_code/GTEx_DGE_ALL.R
#!Rscript Final_source_code/GTEx_DGE_AGE.R
```

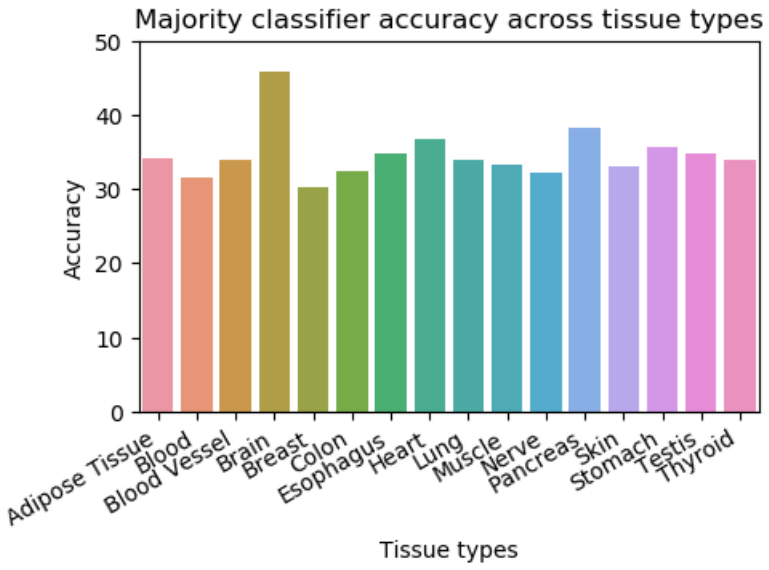


Each point on these mean-difference (MD) plots represents a gene. The X location is the average expression in all samples; the Y location is the log2 difference between low age and high age samples. Thus, more extreme Y locations represent greater differences. The DGE analysis did not yield differentially expressed genes across all samples. However, some tissues exhibited a handful of DEG's, such as the esophagus. These genes are then associated with aging. In the case of the esophagus, there are six genes that are enriched in old age.

In the process of generating DGE results, we also generated MDS plots, which are a form of dimension reduction and visualization. These plots helped reveal

ML/Stats #0: Baseline Classifier

```
In [11]: %run Final_source_code/GrptrendsMajorClassifier
```

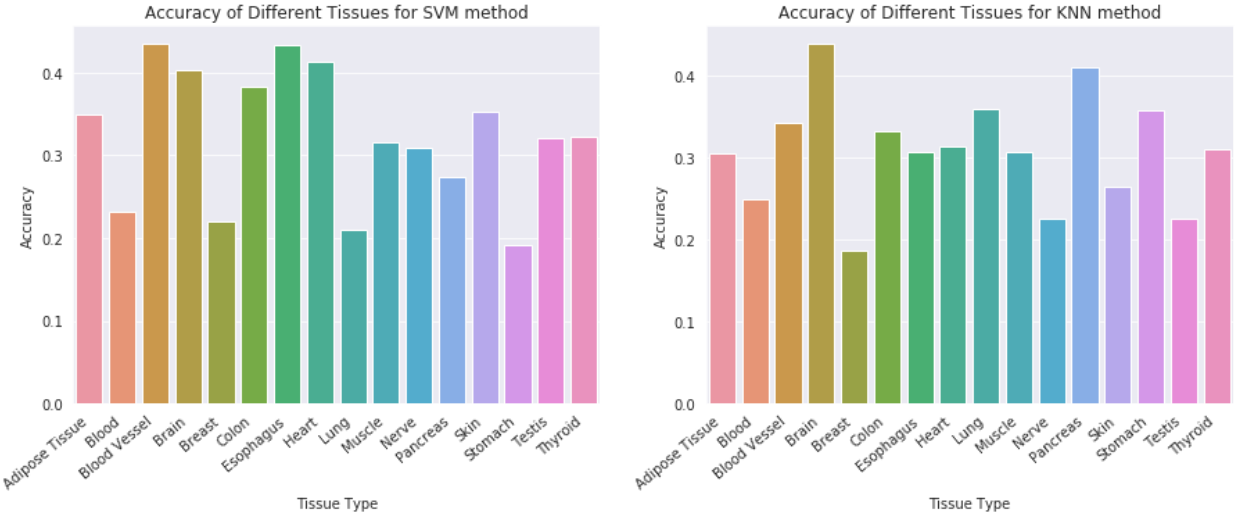


ML/Stats #1: Traditional Classification

We made our classical ML models based on K-nearest neighbor (KNN) and Support Vector Machine (SVM). Here we see the accuracy of these two models for different tissue types. We can see that the accuracy for both models are almost the same and is less than 0.5.')

```
In [16]: %run ClassicML/plots.py
         %matplotlib inline
```

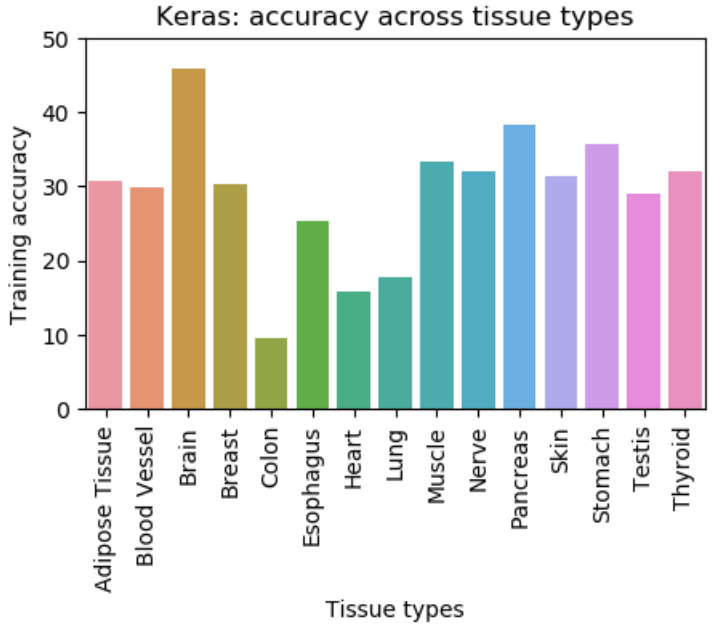
```
Figure(1080x360)
```



ML/Stats #2: Deep Learning

Model: TensorFlow Keras with 3 hidden layers (1024, 512 and 128 hidden neurons in each corresponding layer) with early stopping hyperparameter and Min-max normalization.

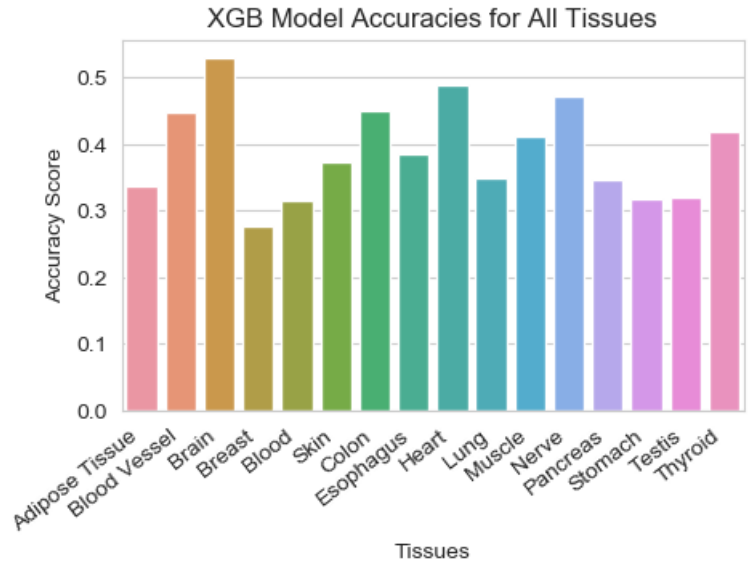
In [11]:
```
%run Final_source_code/dlModel.py
```



ML/Stats #3: XGBoost and Primary Hypotheses Testing

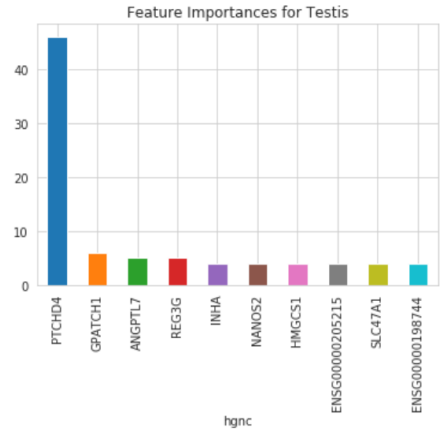
XGBoost utilizes boosted trees to predict continuous or categorical variables. The following accuracies were achieved on a per-tissue basis. Despite general improvement over the majority classifier, the accuracies are not remarkably high.

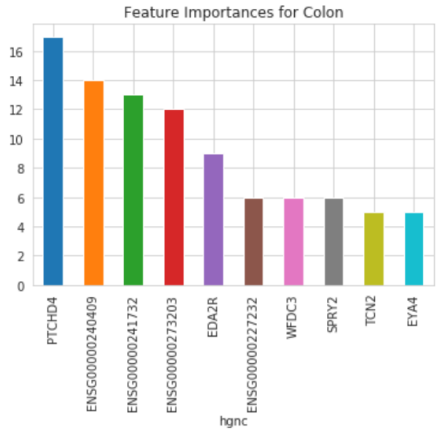
In [19]:
```
%run Final_source_code/XGB_accuracy.py
```

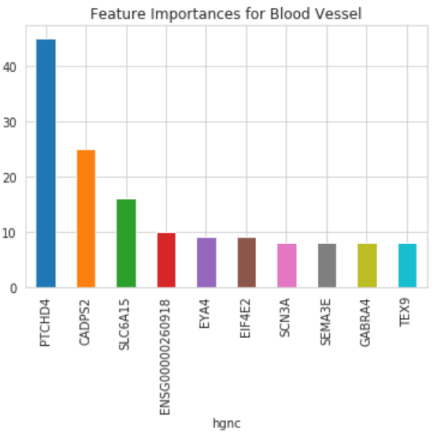


In [19]:
```
#%run Final_source_code/XGB_featImp.py # Generates all feature Importance plots in a row but not formatted as below.
```

Despite weak DGE results, we gleaned some biological insight from feature importance analysis for the XGB model. All of the feature importance plots are output to `plots/`, but PTCHD4 appeared to be important for the classification of such tissues as the testis, colon, and blood vessels.



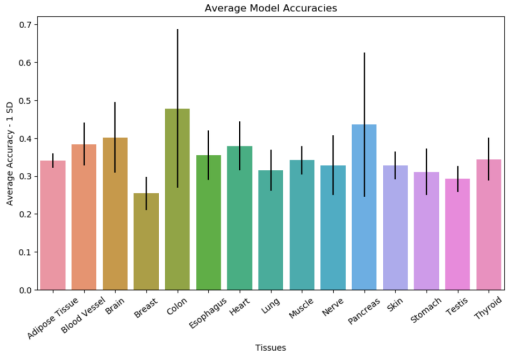




In [76]:
```
#%run Final_source_code/all_model_evaluation.py # Generates same output as below but not formatted sid
e by side.
```

The Mann-Whitney U test is a nonparametric test whether a random sample from one population is equally likely to be greater or less than a random sample from another population. We chose this test since we were not confident in assuming normal distribution for our model accuracies. We compared the accuracies for all models for all tissues with one another where each tissue was an accuracy population. This amounted to 105 comparisons. We applied the Holm-Šídák correction for multiple hypothesis testing.

Though we observed differences in overall accuracy between some tissues, we did not find statistically significant differences. This means we fail to reject our null primary hypothesis that different tissue are equal predictors or age. The gene expression in tissues as they age are all about equally strongly associated with age. This makes sense given that we did not find many genes associated with aging in general.



	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

Conclusion

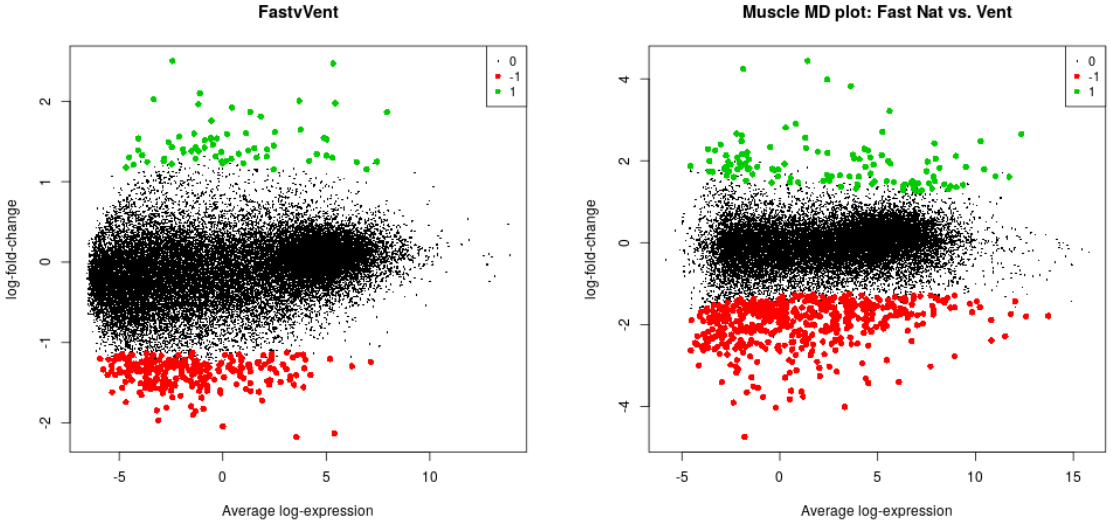
Our analysis represents a novel approach to understanding how the body ages.

The results of our primary hypothesis and model performances impact how we view aging in the human body. The body does not undergo fundamental changes in the gene expression of its cells. It makes sense that a single cell is essentially agnostic to the year age of the parent organism. And though some tissues are more associated with aging as evidenced in their higher age prediction accuracies, these differences are not significant.

The genes that we did find in association with aging were reasonably related to proliferation and senescence. And though there were relatively few genes detected, the DGE results could significantly guide future molecular genetics research into aging as some detected genes are not well characterized.

Future Directions

In the process of studying aging, we detected that death type appears to play a major role in the gene expression in many tissues. These DGE results are packaged in `DGE_results` as well. The MD-plot on the left shows that there are numerous differentially expressed genes between fast natural deaths and ventilator deaths. And there are even more differences on a per-tissue basis as seen in the MD-plot for the muscle tissue. Positive fold-change is enrichment in vent deaths, and negative fold-change is down-regulation in vent deaths.



In [20]:
```
#!Rscript GTEx_DGE_DEATH_ALL.R # Scripts needed to generate these DGE results
#!Rscript GTEx_DGE_DEATH.R
```