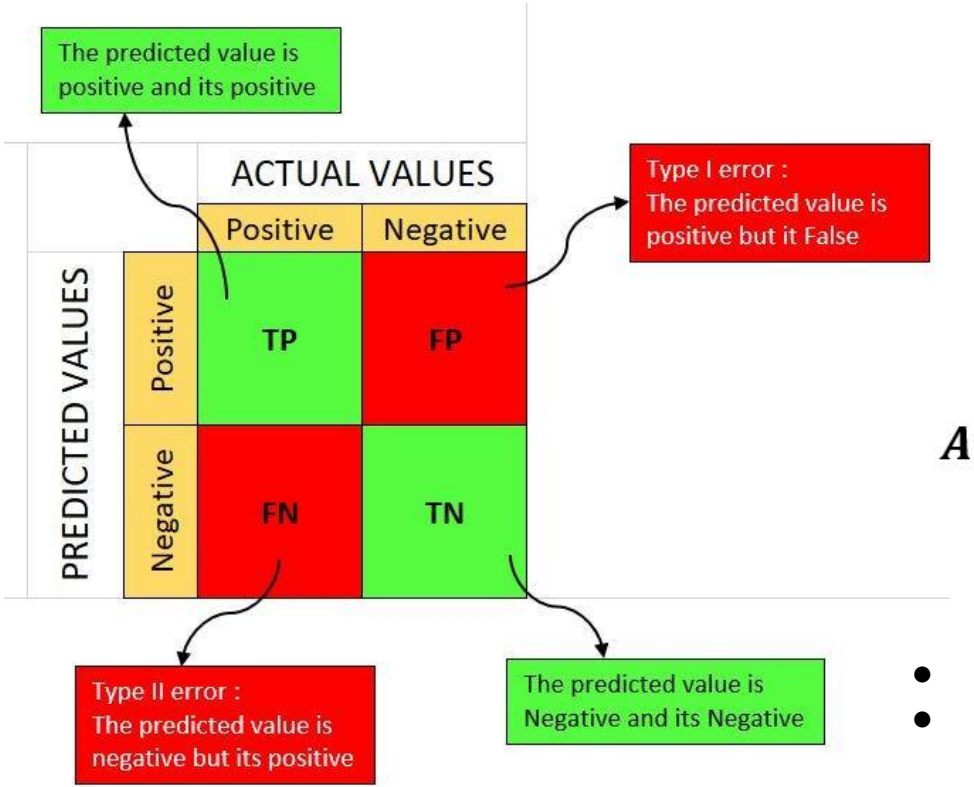

Big_Ocean

METRIC1: Accuracy



$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- the fraction of predictions our model got right.
- Example: out to 0.91, or 91% (91 correct predictions out of 100 total classes).

METRIC2: F1-score

$$F_1\text{−score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

- The F1-score is the harmonic mean between precision and recall
- Evaluation classification result
- Account class imbalance

PyMethylProcess: Accelerating DNA Methylation Data Preprocessing

- preprocess DNA methylation array data
- access traditional differential methylation analyses and machine learning libraries
- Uses both Python and R libraries
- pip-installable command line interface.
- can be used through docker.

Data

The paper used 7 datasets(6 GEO and 1 TCGA).

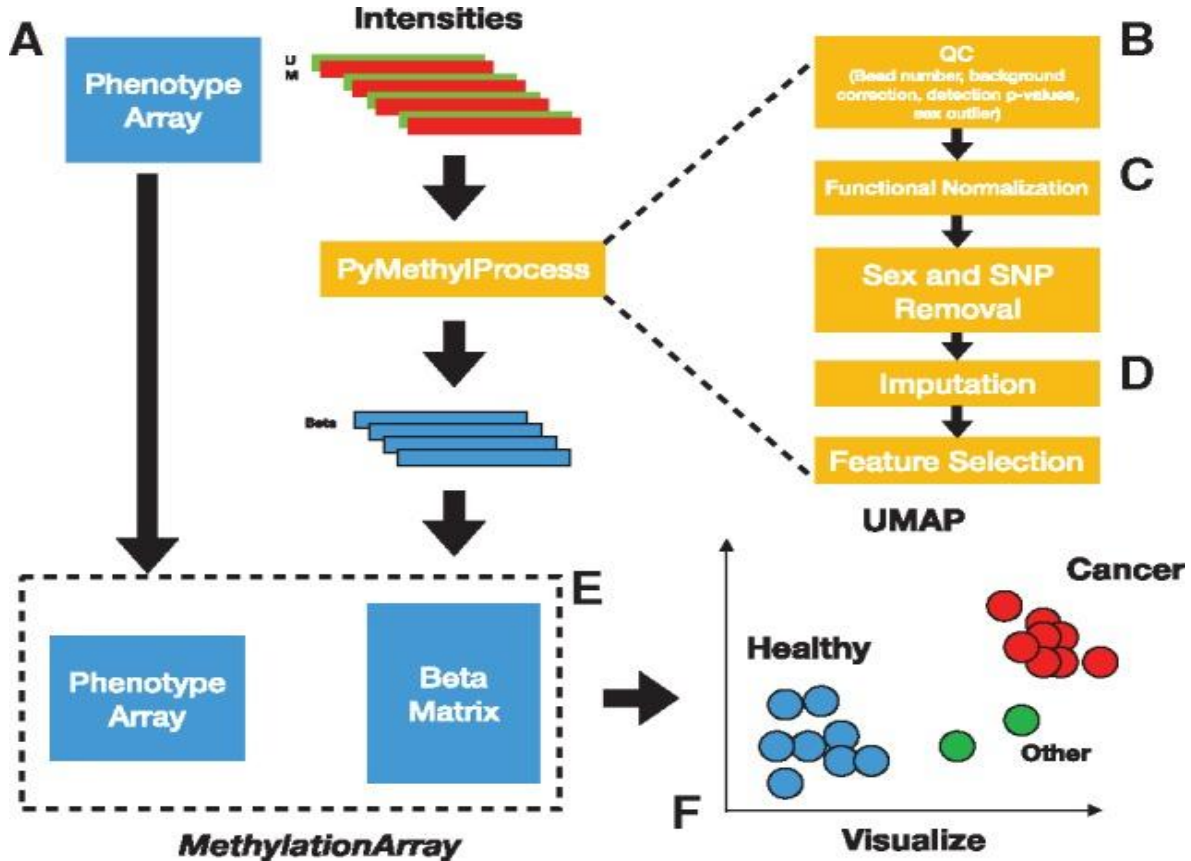
One of mainly focused dataset is :

GSE87571:

- Measured using Illumina HumanMethylation450 BeadChip
- Bisulphite converted DNA from 732 samples 476,366 sites throughout the genome of white blood cells.
- from a population cohort aged 14 to 94 years

From paper: **Johansson,A . et al. (2013) Continuous aging of the human DNA methylome throughout the human lifespan.**

PyMethylProcess: Basic Framework



Installation and Usage

pip install pymethylprocess && pymethyl-install_r_dependencies # python 3.6, R(3.5.1)

Preprocessing

1. **Downloading** the data example:
*pymethyl-preprocess download_geo -g **GEOID** -o **geo_idats/***

2. **Formatting** the Sample Sheet

3. **Running** the Preprocessing Pipeline

4. **Accessing**, Reading and **Writing** MethylationArray Data

Python notebook

5. **Visualization**: UMAP Embed

6. **Supervised models** (Feature extraction)

*yellow lined boxes: we were able to use it, black box: we couldnt yet.
Tried Docker.

Results

PyMethylProcess Pipeline Benchmarks

DataSets	Brief Description	Sample Size	Preprocessing Pipeline	# CpGs After Normalization	Principal Components	# Outlier Samples	# CPUs	Memory (Gb)	Runtime (Minutes)	# Sites Removed	Percentage Imputed (%)	Imputation Method
GSE87571	Johannson Aging	732	Minfi: Noob Normalization	482669	NA	13	1	NA	150.0	11233	0.706	K-NN: 15 neighbors
GSE81961	Crohn's Disease	40	Meffil: Functional Normalization	480329	4.0	0	35	10	4.1	11587	0.096	K-NN: 5 neighbors
GSE69138	Stroke	185	Meffil: Functional Normalization	474021	13.0	2	35	NA	11.5	11305	0.166	K-NN: 5 neighbors
GSE42861	Smoking and Arthritis	689	Meffil: Functional Normalization	477482	13.0	8	30[1]	110[2]	38.5[3]	11448	0.190	K-NN: 5 neighbors
GSE112179	Schizophrenia and Bipolar	100	Meffil: Functional Normalization	853772	10.0	2	30	NA	40.0	19278	0.240	K-NN: 10 neighbors
GSE109381	Brain Cancer Subclasses	3897	Meffil: Functional Normalization[4]	320023	10.4[5]	135	30	60	135.0	6791	0.054	Mean
TCGA Pancancer	33 Pan-Cancer subtypes	8891	Meffil: Functional Normalization[4]	378588	14.6[5]	515	30	60	255.0	7869	0.112	Mean

[1] Quality control used 30 CPUs, Normalization Used 14 CPUs
[2] Only for normalization step
[3] 13.5 Minutes for QC, 25 minutes for Normalization
[4] Subclasses Processed in Parallel
[5] Averaged Across Disease Subtypes

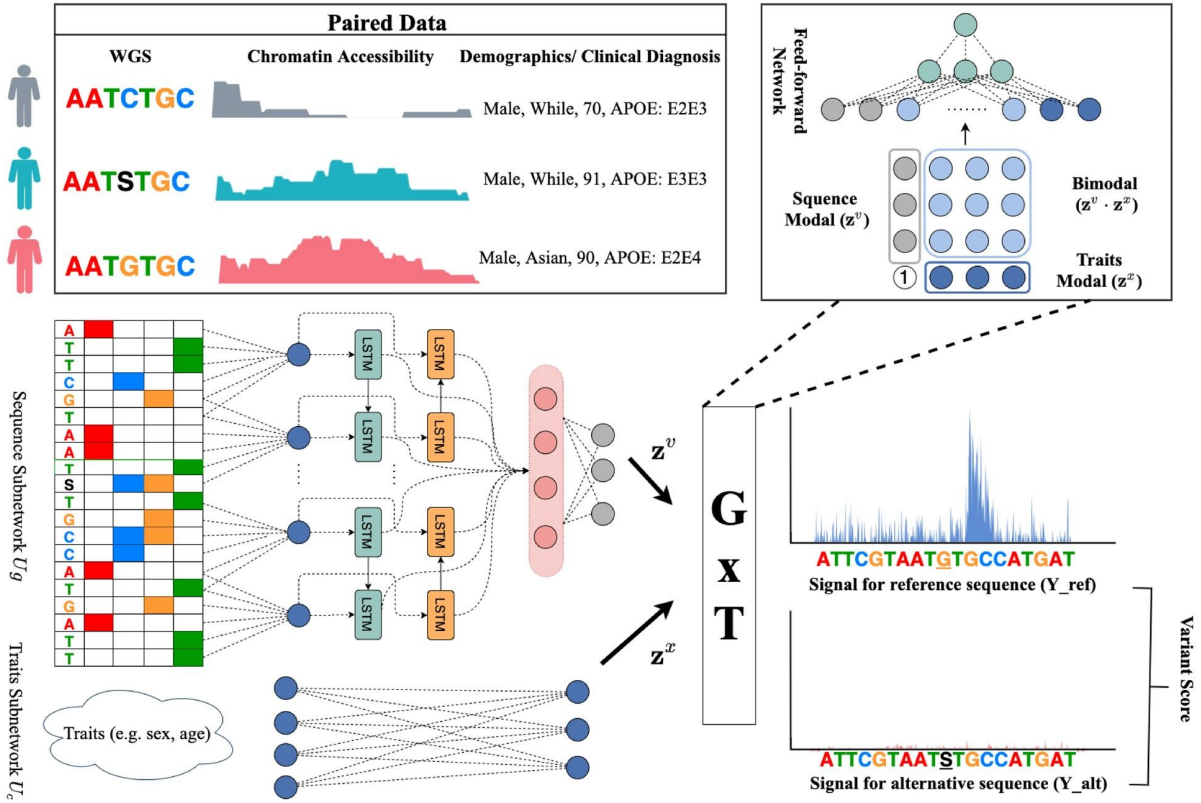
TOOL2: DeepPerVar

- Functional interpretation of genetic variants of personal genome by leveraging paired WGS data and epigenetic functional assays in a population study.
- Quantitatively predict epigenetic signals (e.g. histone modification and DNA methylation)
- Multimodal deep neural network
- Ideas for feature importance extraction:
    - Using feature extraction API from pytorch

Training data

- Genetic data
  - WGS data in the format of genomic VCF files are generated by GATK
- Epigenetic data
  - 439 datasets, which are across 83 cell lines/tissues, 18 tissue classes and 7 core histone marks including H3K4me1, H3K9ac
  - Criteria to pick CHIP-seq peak:
    - ChIP counts are smaller than matched input control counts
    - *P*-value is <0.05 derived from the Poisson test
  - Merge overlapped peaks across all individuals and calculate the normalized read counts for each merged peak by adjusting sequence depth and matched input control. -> 141 807 merged peaks and normalized ChIP counts in each peak
- DNA methylation data
  - 104 DNA methylation datasets generated from three different sequencing technologies
  - adjusted for age, sex and experimental batch, which ends up with methylation ratio at 418 972 CpGs

General method



Installation & Usage

DeepPerVar is implemented by Python3.

- Python 3.8
- numpy >= 1.18.5
- pytorch ==1.7.1
- biopython=1.19.2

Download Reference Genome (hg19), and put them in the DeepPerVar root directory. Download DeepPerVar Models, and put model files in models directory.

```
unzip Models.zip Reference.zip
```

Download DeepPerVar:

```
git clone https://github.com/alfredyewang/DeepPerVar
```

Install requirements.

```
pip3 install -r requirements --user
```

Install Samtools 1.15.1 follow the (instruction)[http://www.htslib.org/download/] .

Input File Format

DeepPerVar takes UCSC Genome Browser BED file. Each line has 5 tab separated fields. The BED fields are:

- The first column: Chromosome name (hg19).
- The second column: Position of SNPs (hg19).
- The third column: The strand information.
- The fourth column: reference allele.
- The fifth column: alternative allele.

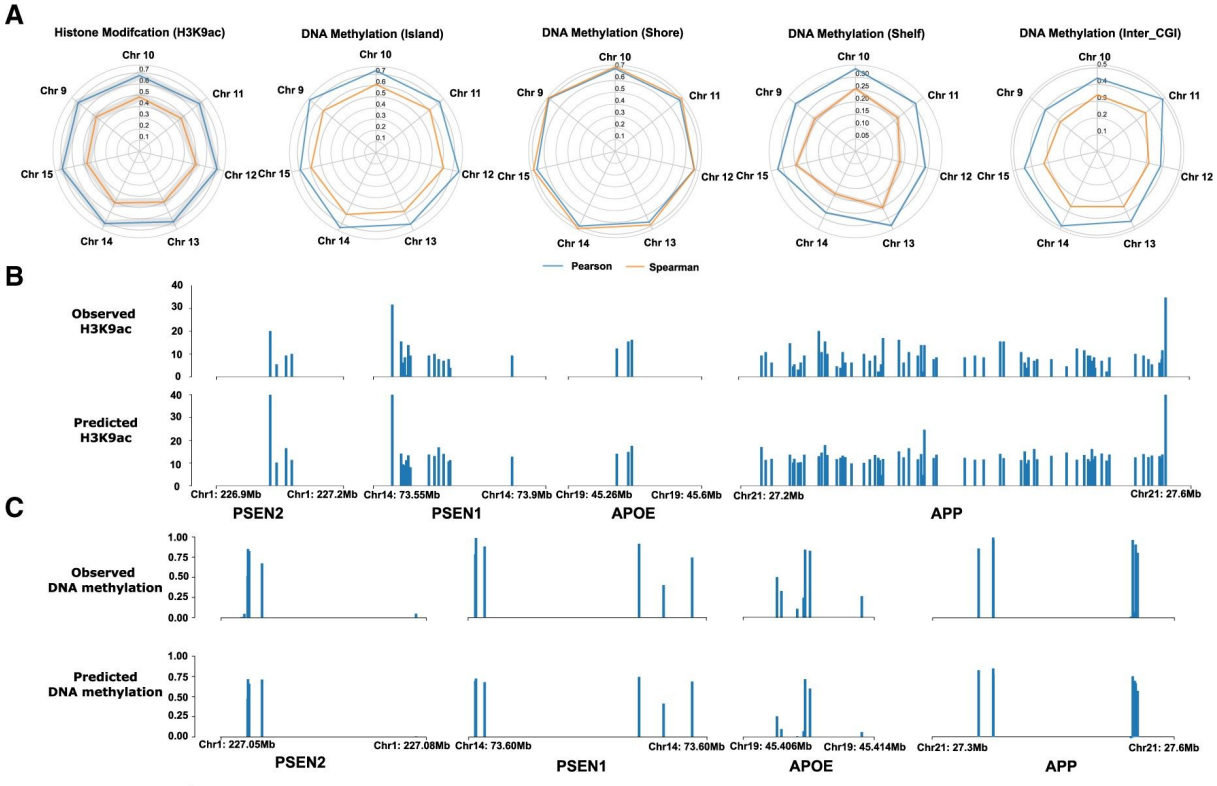
H3K9 Example

```
python3 src/DeepPerVar.py --prediction --epigenomics H3K9 --bed data/snps.bed --res_dir res --model_
```

Results will be save into res/Results_histone.csv

chr	pos	strand	ref	alt	H3K9AC_REF_Pred	H3K9AC_ALT_Pred	DELTA_H3K9AC
1	1265154	-	T	C	18.415241	18.509096	0.093854904
1	1265460	-	T	A	17.707266	17.64615	-0.061115265
1	2957600	-	T	C	10.322433	10.464524	0.1420908
1	3691528	-	A	G	16.85876	16.950903	0.092142105
1	8021919	-	C	G	82.27526	82.20313	-0.072128296
1	8939842	-	G	A	42.205887	42.33795	0.13206482
1	10457540	-	T	C	13.674403	13.556186	-0.11821747
1	11072117	-	C	T	57.86567	56.590023	-1.2756462
1	11072691	-	G	A	37.507782	37.999027	0.49124527
1	11083408	-	G	A	16.937225	15.624798	-1.3124275

Paper's result



GEO dataset (GSE97362)

- About Neurodevelopmental syndromes (CHARGE and Kabuki syndrome)
- Infinium HumanMethylation 450K + EPIC
- Gene-specific DNA methylation signatures
- 285 cases across 14 syndromes, 650 controls

Samples

CG id

	GSM2562699	GSM2562700	GSM2562701
cg00000029	0.519920000	0.59989020	0.565073500
cg00000108	0.944431600	0.93048600	0.920178500
cg00000109	0.848503500	0.87495590	0.849404200
cg00000165	0.201246000	0.22542840	0.277427300

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE97362