IMPLEMENTAI

DeLEG: Deep Learning for EpiGenomics data to predict phenotype.

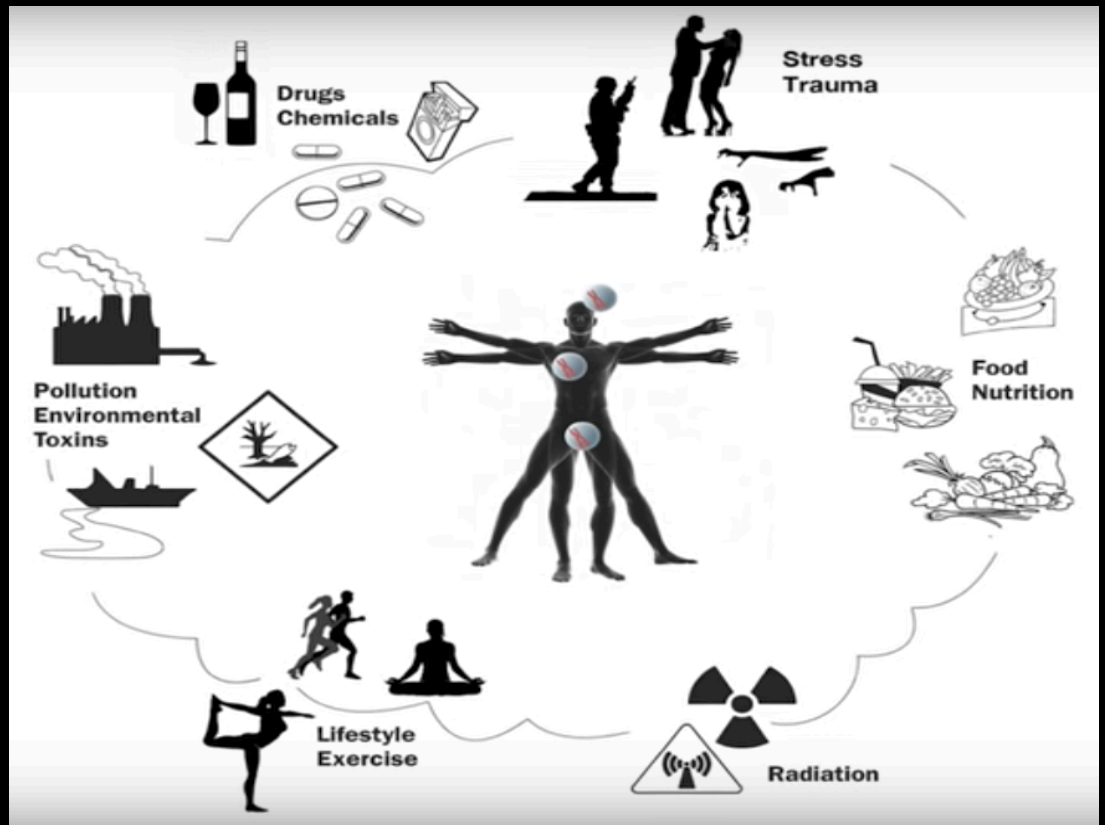
Phenotype, genotype and environment

Phenotype = genotype + environment

P = G + E

Interaction between genetics and environmental factors

P = *f*(G,E)



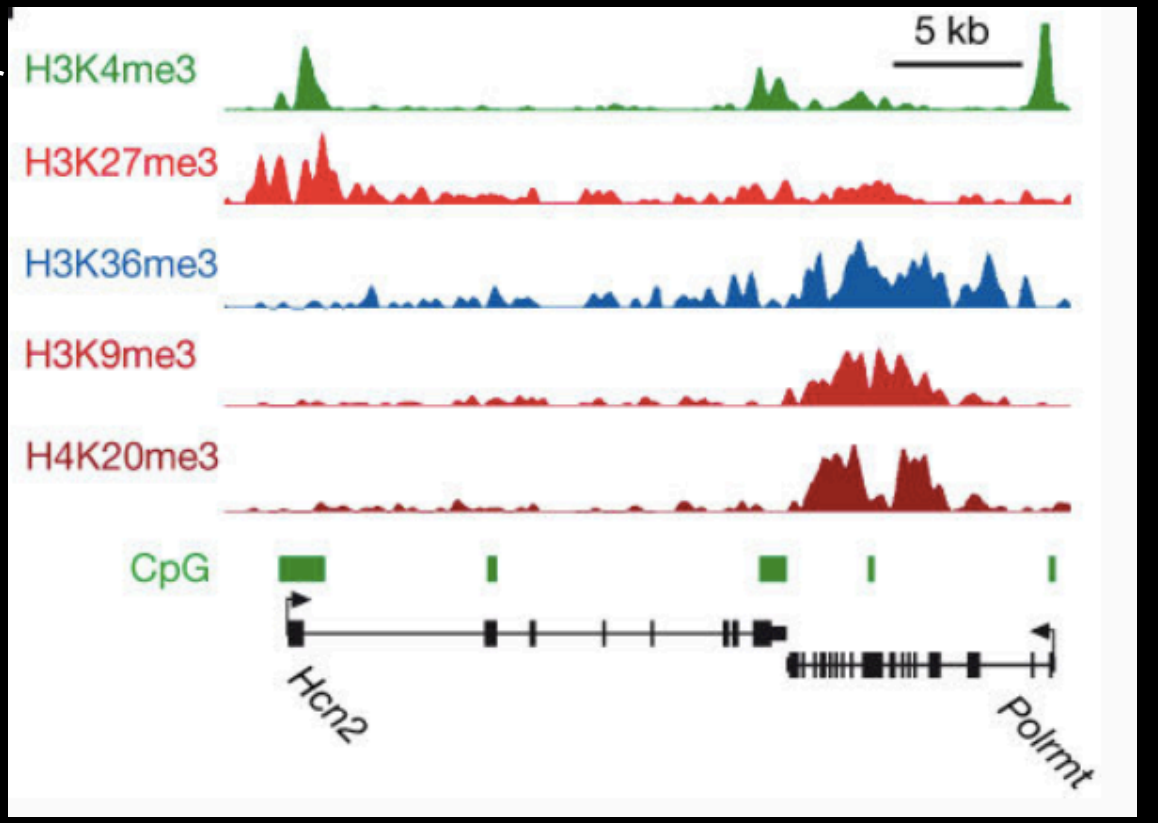
IMPLEMENTAI

Challenges

1. Finding “important” regions in ChIP-seq data
2. Using the “important” regions for prediction, classification and better understanding of Human Epigenome

HEADLINE:

- Ground truth: “Healthy” and “Disease”
- Training set: 14 subjects in each phenotype
- Validating set: 4 subjects in each phenotype
- Testing set: Any new subject not in the set of 36 subjects



doi:10.1038/nature06008

Workflow

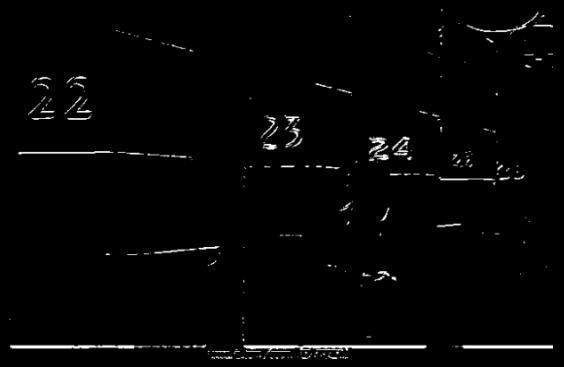
Otsu thresholding for segmentation	Enrichment score of window around TSS	Train a Conv Neural Network for classification
Input: ChIP-seq data	Input: ChIP-seq data and peak regions	Input: Enrichment score of fixed length sequences
Output: "Important" regions of the ChIP-seq data corresponding to peaks	Output: Enrichment score from ChIP-seq data of each bp around TSS for each gene corresponding to peaks	Output: Classification probabilities for the window to lie in each phenotype.
Detailed Method: Otsu thresholding of the ChIP-seq data to remove background and filter out peaks. A post-processing for noise removal is done following thresholding.	Detailed Method: Gene identification from database and window extraction (data manipulation process)	Detailed Method: Training a CNN with windows from 14 subjects in each class, totalling to about 20k windows in each class. Looking back at the learned features -> further insight

IMPLEMENTAI

Learning attention from classification data

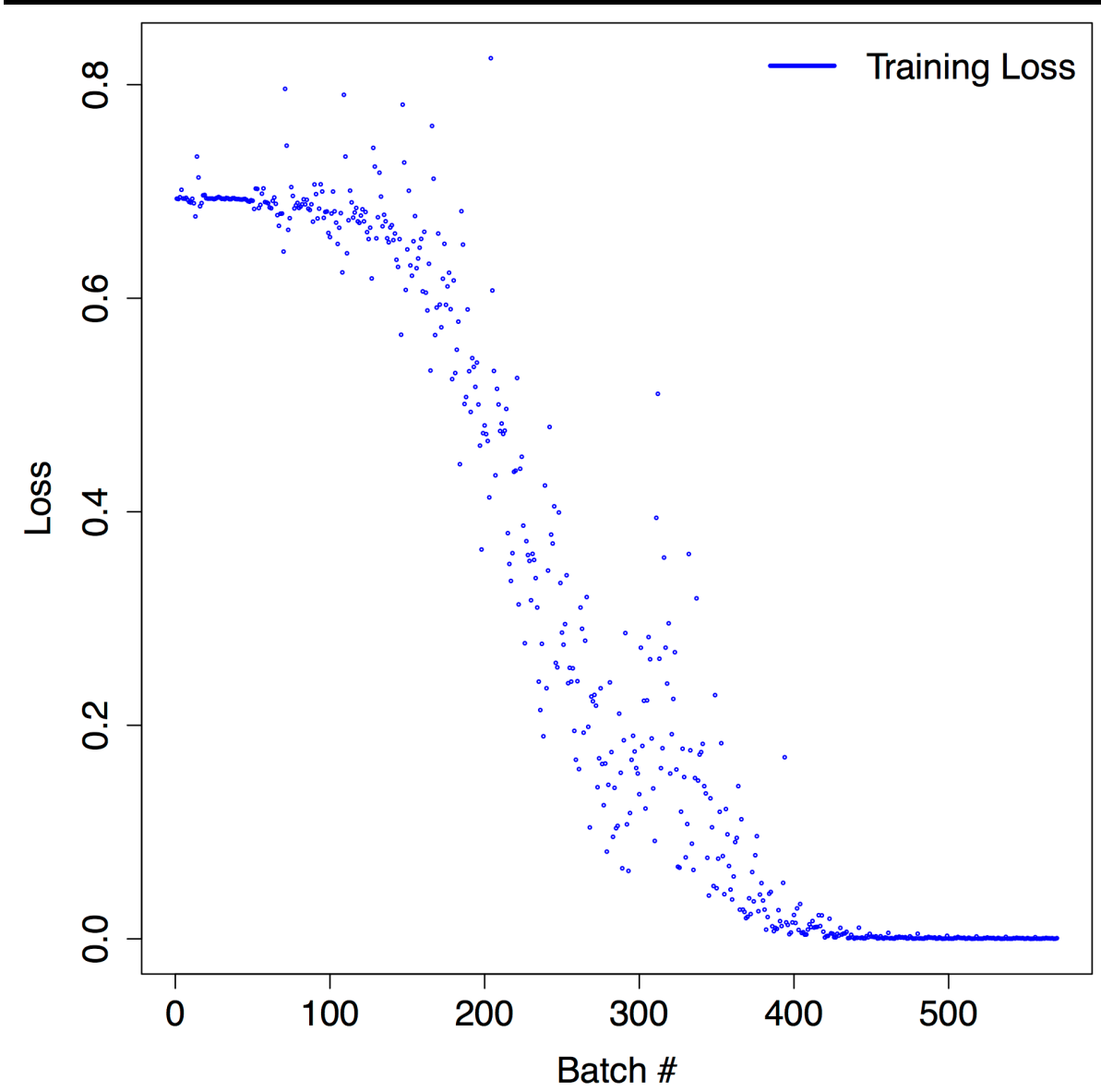
" Those who pay attention *learn*,
Those who don't cram "

- Use Global Average Pooling concept to learn regions in input which caused the network to classify it to a particular class
- Final layer filters know where to look
- Get defining regions - define further research!!



IMPLEMENTAI

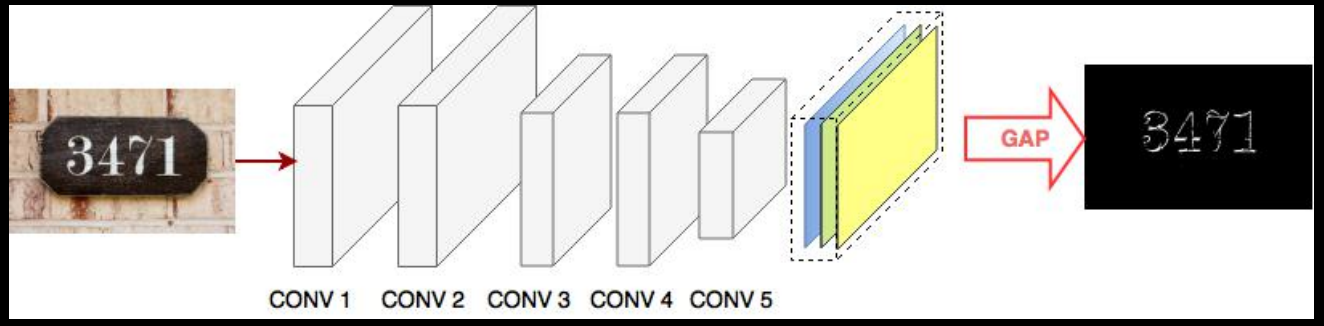
Supplementary Material Results

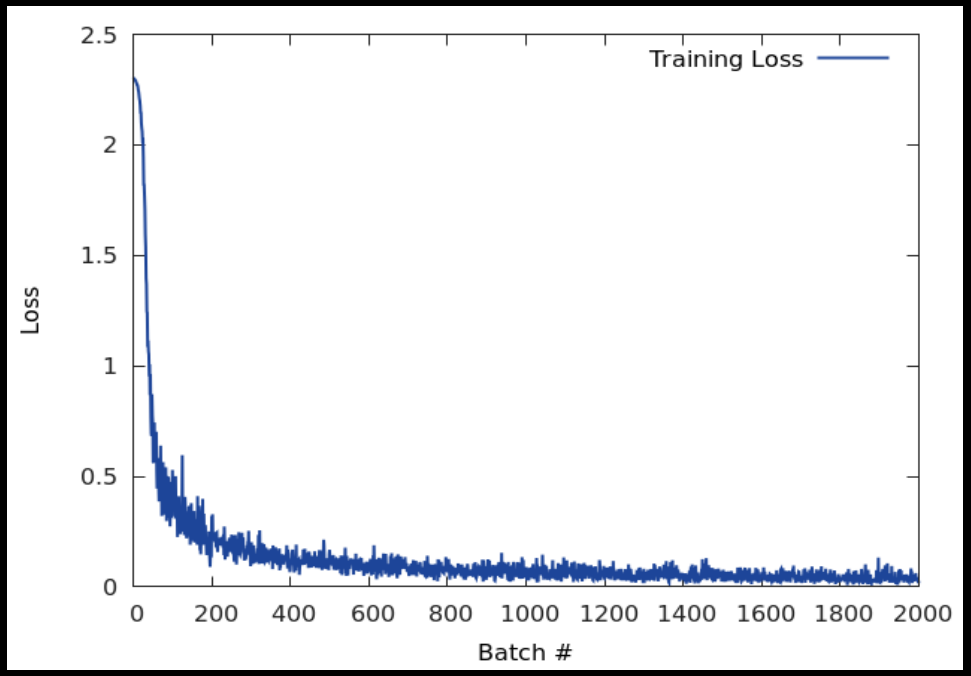



```
Testing accuracy
 percentage correct 81.831664812755 %
Healthy  70.564186426819 %
Disease  91.180461329715 %
```

IMPLEMENTAI

Supplementary material - Results





Class	Accuracy
0	99.59%
1	99.47%
2	99.90%
3	98.51%
4	99.08%
5	98.32%
6	98.23%
7	98.64%
8	97.43%
9	97.92%
Overall	98.73%

IMPLEMENTAI