Encode Enhancer Challenge

Input Data

file download from http://cistrome.org/db/#/	ChIPTF	GEO Accession Number	Reference genome	tissue/cell ty
54499_H3K27ac_sort_peaks.narrowPeak.gz	H3K27ac	GSM851284	mm10	Embryonic E
5060_OLIG2_sort_peaks.narrowPeak.gz	OLIG2	GSM766058	mm10	Embryo
60947_H3K4me2_sort_peaks.narrowPeak.gz	H3K4me2	GSM632045	mm10	Embryo
55119_SOX2_sort_peaks.narrowPeak.gz	SOX2	GSM1033096	mm10	Embryo
62993_H3K27ac_sort_peaks.narrowPeak.gz	H3K27ac	GSM1264370	mm10	Heart
68244_H3K4me1_Ren_sort_peaks.narrowPeak.gz	H3K4ME1_E14.5	GSM1000136	mm10	Heart
56691_POLII_sort_peaks.narrowPeak.gz	POLR2A	GSM1163129	mm10	Cardiomyocy
68115_EP300_Ren_sort_peaks.narrowPeak.gz	P300_ADULT-8WKS	GSM918747	mm10	Heart
53322_H3K4me1_sort_peaks.narrowPeak.gz	H3K4me1	GSM851281	mm10	Embryonic E
5097_HOXC9_sort_peaks.narrowPeak.gz	HOXC9	GSM766061	mm10	Embryo
54562_H3K27ac_sort_peaks.narrowPeak.gz	H3K27ac	GSM1039565	mm10	Hindlimb Au
58262_SPI1_sort_peaks.narrowPeak.gz	SPI1	GSM878650	hg38	Fetal Brain
1491_DNase_sort_peaks.narrowPeak.gz	DNase	GSM595926	hg38	Fetal Brain
1932_EP300_sort_peaks.narrowPeak.gz	EP300	GSM602299	hg38	Neuroectode
53421_H3K4me1_sort_peaks.narrowPeak.gz	H3K4me1	GSM772785	hg38	Neuron
61864_H3K27ac_sort_peaks.narrowPeak.gz	H3K27ac	GSM956008	hg38	Embryo
54525_H3K27ac_sort_peaks.narrowPeak.gz	H3K27ac	GSM910557	hg38	Right Atrium
58256_SPI1_sort_peaks.narrowPeak.gz	SPI1	GSM878630	hg38	Fetal Heart
1545_DNase_sort_peaks.narrowPeak.gz	DNase	GSM665811	hg38	Fetal Heart
61702_H3K9ac_sort_peaks.narrowPeak.gz	H3K9ac	GSM706849	hg38	Heart
58318_DNase_sort_peaks.narrowPeak.gz	DNase	GSM1027324	hg38	Fetal Renal
58242_SPI1_sort_peaks.narrowPeak.gz	SPI1	GSM878662	hg38	Fetal Renal

Method 1(using ChIPseq signal as features)

- Select relative ChIPseq dataset based on Enrichment in VISTA regions and prior knowledge
- Liftover hg19 coordinates of VISTA regions to hg38 and mm9 coordinates of VISTA regions to mm10
- Annotate VISTA region with overlaping ChIP-seq dataset peak score. – At this step, hg38 regions associate with scores only from hg38 chipseq peaks, and similarly mm10 regions only associate with mm10 chipseq peak
- Annotate regions highly conversed across human and mouse with both peaks score from two species. – Highly conserved regions are defined bsaed on UCSC liftOver 0.95 conserved – At this step, we have a feature matrix with missing values: each row is VISTA region and each column is one ChIP-seq signal feature for both mm10 and hg38
- Impute the missing value using R package “mi”
- Train 3 binary classification problems based on the imputed feature matrix and label of each VISTA region: brain, heart, other enhancer
- Build logistic regression model using R package “glmnet”
- Predict LBNL tested regions: construct imputed feature matrix for tested regions as stated above, apply the trained logistic model to predict the probability of each types of enhancer in the given tested regions.
- Predict genome-wide regions: construct imputed feature matrix for human DHS region with GWAS SNPs, apply the train model above to predict three probabilities for each region, finally convert coordinate to mm10

output files:

- 240 LBNL test regions prediction: PredictionUsingChIPSeq.txt
- Regions likely to function in e11.5 mouse embryo : GenomeWide5kPredictionUsingChIPSeq.txt

Method 2(using Kmer frequency as features)

- Two types of Kmer features are used: 8mer frequency allowing 3 mismatches , 5mer pair allow 1 mismatch and 0-30bp gap in between two 5mers
- Train: for each VISTA region sequence, extract Kmer frequency vector, build logistic regression model using R package "glmnet" for 3 enhancer classification problems: brain, heart, other enhancer
- Predict LBNL tested regions: extract Kmer frequency feature vector for each test regions, and apply the trained logistic model to predict the probability of each types of enhancer in the given tested regions
- Prioritize other genomics regions: only use top5k regions predicted by chipseq feature method, and apply the trained logistic model(Kmer model) to predict the probability of each types of enhancer in the given 5k regions, and report the predictive probability of those highly positive regions.

output files:

- 240 LBNL test regions prediction: PredictionUsingKmer.txt
- Regions likely to function in e11.5 mouse embryo : GenomeWide5kPredictionUsingKmer.txt

output file format:

- first row: indication of types of enhancer for last 3 columns
- column 1:mm10 coordinates of predicting region
- column 2: probability of the given region is a enhancer in any tissue of e11.5 mouse
- column 3: probability of the given region is a enhancer in forebrain tissue of e11.5 mouse
- column 4: probability of the given region is a enhancer in heart tissue of e11.5 mouse

Submission Files

- Under submission folder, there are two sets of files : set1_xxx(prediction from method1) and set2_xxx(prediction from method2)
- setX_file1,2,3,4: follow the order of submission guideline:
  - File 1 of prediction - corresponding to heart regions being tested (120 regions) - bed5 format: first 3 columns correspond to chromosome, start, end of prediction. 4th column is prediction probability of being active in heart, 5th column is prediction probability of being active in any tissue.
  - File 2 of prediction - corresponding to forebrain regions being tested (120 regions) - bed5 format: first 3 columns correspond to chromosome, start, end of prediction. 4th column is prediction probability of being active in forebrain, 5th column is prediction probability of being active in any tissue.
  - File 3 of prediction - genome-wide heart predictions (<= 5k regions) - bed4 format : first 3 columns correspond to chromosome, start, end of prediction. 4th column is prediction probability of being active in heart.
  - File 4 of prediction - genome-wide forebrain predictions (<= 5k regions) - bed4 format : first 3 columns correspond to chromosome, start, end of prediction. 4th column is prediction probability of being active in forebrain.