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

output files:

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