Accurate prediction of functional enhancer-promoter interactions using epigenomic data

Yuchun Guo, Gokul Ramaswami, Swathi Dhanasekaran, Eli Bogart, Bryan Matthews, Jenna Weinstein, Mario Gamboa, Rachana Kelkar, Yun Joon Jung, Brynn Akerberg, Yuting Liu, Alfica Sehgal, David Bumcrot

CAMP4 Therapeutics, Cambridge, MA, USA

Motivation

- Transcriptional enhancers control how genes are expressed in specific cell types.
- Enhancer disruption and misregulation are implicated as disease-driving mechanisms.
- Modalities that specifically target enhancers that control disease-associated genes are being pursued to develop new drugs for a range of indications.
- However, it remains a major challenge to link functional enhancers to their target genes.

Approach

Enhancer-promoter interaction characterization (EPIC) is a machine learning model for predicting functional enhancer-promoter (E-P) pairs.

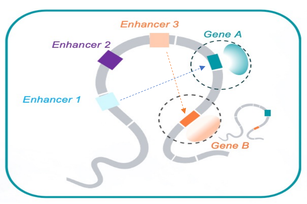
(Gasperini et al., 2019; Xie et al., 2019)

K562 enhancer epigenomic and chromatin interaction features

K562 CRISPR-based enhancer perturbation screening

EPIC: a machine learning model

Infer



Enhancer 1 – Gene A: Yes
Enhancer 2 – Gene A: No
Enhancer 3 – Gene B: Yes
…

Basic features

- HiChIP.AnchorSize: AnchorSize = 5kb, 10kb, 15kb, or 20kb (n=4)
- Assay.Position.WindowSize, where Assay=ATAC, H3K27ac, H3K4me1, H3K4me3, EP300, CTCF, or Input ChIP; Position = Enh or TSS; WindowSize = 300bp, 500bp, 1kb, 2kb, or 4kb (n=7*2*5=70)
- Genomic distance (n=1)

Feature engineering

$APMI = (ATAC.Enh.1kb * EP300.Enh.1kb * H3K4me1.Enh.4kb)^{1/3} * HiChIP.5kb$

Based on APMI, we engineered a new set of features for quantifying the relative contribution of an enhancer ***e*** to a gene ***g*** from the gene perspective or enhancer perspective:

$$fracGene_{eg} = \frac{APMI_{eg}}{\sum_j APMI_{jg}}$$

where j indexes all the enhancers connected to gene ***g***.

$$fracEnh_{eg} = \frac{APMI_{eg}}{\sum_k APMI_{ek}}$$

where k indexes all the genes connected to enhancer ***e***.

In addition, we combined these features to form new features.

$$fracGmE_{eg} = fracGene_{eg} * fracEnh_{eg}$$

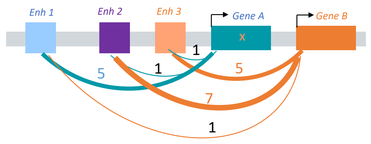
$$fracGpE_{eg} = fracGene_{eg} + fracEnh_{eg}$$

$$apmiGene_{eg} = fracGene_{eg} * APMI_{eg}$$

$$apmiEnh_{eg} = fracEnh_{eg} * APMI_{eg}$$

$$apmiGmE_{eg} = fracGmE_{eg} * APMI_{eg}$$

$$apmiGpE_{eg} = fracGpE_{eg} * APMI_{eg}$$



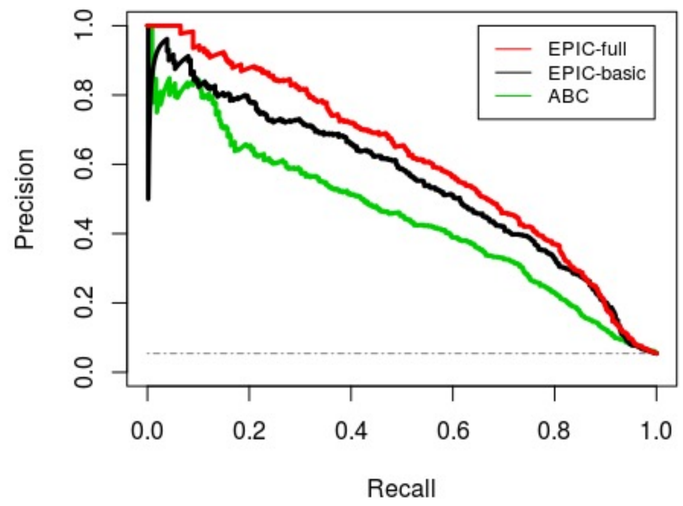
	Enh1 - GeneA	Enh3 -GeneB
fracGene	$\frac{5}{1+1+5} = 0.71$	$\frac{5}{1+5+7} = 0.42$
fracEnh	$\frac{5}{1+5} = 0.83$	$\frac{5}{1+5} = 0.82$

Machine learning model

- Random forest classification model trained on K562 data
- Five-fold cross-validation
- Genetic algorithm for feature selection

Results

EPIC outperforms ABC model in predicting enhancer-promoter pairs (holdout test data)

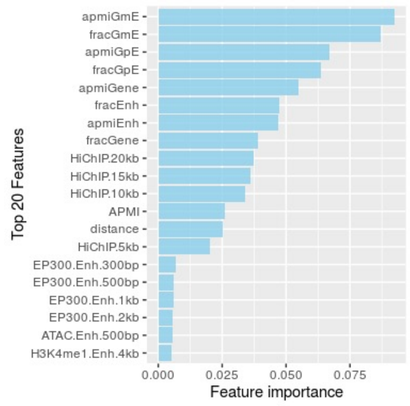


ABC = $(ATAC.Enh.500bp * H3K27ac.Enh.500bp)^{1/2} * HiC.5kb$ (Fulco et al., 2019)

Model	AUPR	AUROC
EPIC-full	**0.613**	**0.918**
EPIC-basic	0.551	0.912
ABC	0.451	0.885

- The area under receiver operating characteristic (AUROC) curve of EPIC-full is significantly higher than that of ABC (p = 1.6e-10) (DeLong, et al., 1988).
- The AUROC of EPIC-full is significantly higher than that of EPIC-basic (p = 0.01), demonstrating the value of feature engineering.

Engineered features rank highest in feature importance.

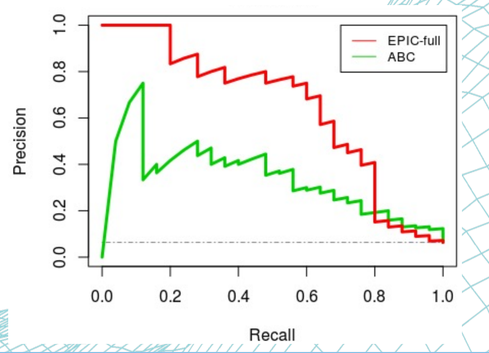


EPIC outperforms ABC model in linking GWAS loci to causal genes in a new cell type

- We generated epigenomic data in human primary hepatocytes and discovered about 30,000 E-P interactions using EPIC.
- Evaluate prediction of causal genes for liver-related GWAS loci using a curated set of "gold standard" locus-gene pairs (Mountjoy, et al., 2021)
  - ➢ Positive: GWAS locus-gene pairs in the gold standard set
  - ➢ Negative: GWAS loci connecting to other genes within 500kb

EPIC
- AUPRC=0.643
- AUROC=0.879
- 80% precision at 50% sensitivity

ABC model
- AUPRC=0.337
- AUROC=0.877
- 40% precision at 50% sensitivity



Linking liver-related GWAS loci to putative target genes

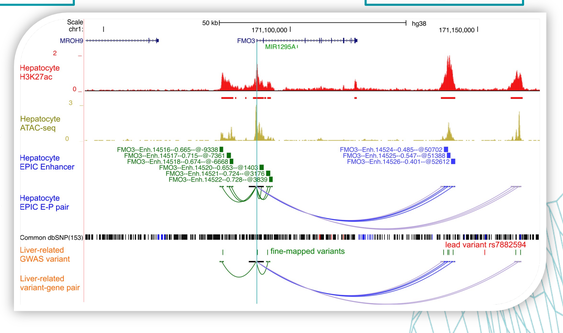
980 EPIC E-P pairs

623 lead variants

1408 fine-mapped variants

464 genes

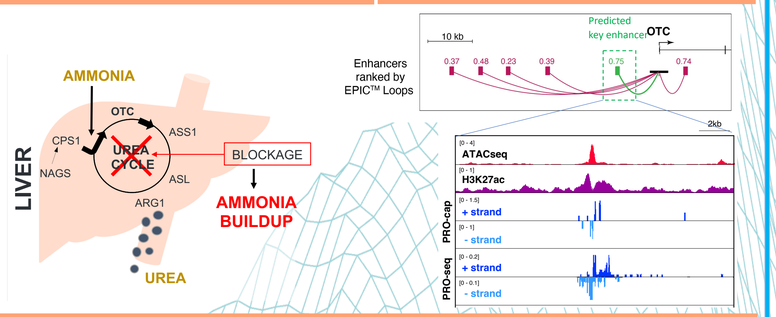
Rs7882954, associated with LDL cholesterol levels, does not overlap enhancers in hepatocyte, but its fine-mapped variants can be linked to FMO3, a liver enzyme for breaking down nitrogen-containing compounds



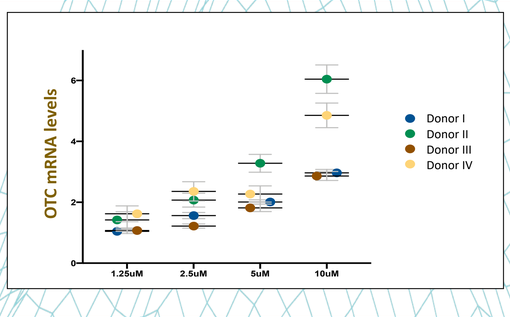
Developing enhancer-RNA-based therapeutics

Partial loss of function mutations cause OTC deficiency

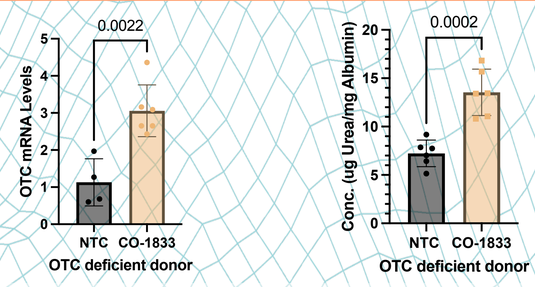
Applied EPIC to predict key enhancer controlling OTC



Lead ASO targeting OTC enhancer RNA shows dose-dependent increase in OTC mRNA in hepatocytes across multiple donors



Upregulation of OTC mRNA increases ureagenesis in human OTC-deficient hepatocytes



OTC c.-106C>A (Allele ID 480410, late-onset OTC deficiency) - pathogenic (dbSNP: rs749748052), leading to decreased OTC mRNA. Variant associated with 10-25% of normal OTC activity

Conclusions

- ❑ EPIC enables accurate cell-type-specific prediction of functional E-P interactions using epigenomic data.
- ❑ EPIC outperforms an established method in predicting E-P interactions and in linking GWAS loci to causal genes in a new cell type.
- ❑ Applying EPIC to diverse human cell types may help discover disease-causing genes and enable development of novel therapeutics that target enhancers of disease-related genes.