CRISPR-screens and Algorithms for Functional Characterization of Enhancers

Jin Woo Oh
Dept. of Biomedical Engineering and McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University
PI: Michael A Beer

YID 2024

Abstract

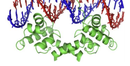
Millions of gene regulatory elements have been mapped across diverse human tissues, revealing that regulatory mutations are associated with common diseases such as schizophrenia, diabetes, and cancer. However, the degenerate sequence structures of regulatory elements and their complex contribution to gene expression pose great challenges to understanding pathogenesis induced by regulatory mutations. To address this issue, we sought to facilitate functional characterization of regulatory elements through three complementary approaches. First, in collaboration with multiple ENCODE consortium labs, we established the largest public database of noncoding CRISPR screens, comprised of more than 100 screens and 540,000 individual perturbations. Noncoding CRISPR screens allow direct and systematic perturbations of regulatory elements, and we believe the public database will be a valuable resource for the scientific and medical community for understanding and treating regulatory diseases. Second, through a CRISPR screen of embryonic stem cell differentiation, we developed quantitative models of regulatory processes that drive cell-state transitions, and proposed a generalizable strategy for detecting enhancers that drive cell-state transitions. The strategy, based on CRISPR screen and quantitative modeling, could discover aberrant regulatory processes that cause organ maldevelopment. Lastly, we developed a novel genome-alignment algorithm that can detect more than 20,000 novel distal enhancers conserved between human and mouse. Putative pathogenic regulatory variants are often tested through their conserved counterparts in mice, but mapping human distal enhancers to mice has long been a computational challenge. Our novel algorithm, *gkm-align*, will facilitate functional characterization of enhancers through model species by mapping conserved enhancers undetectable with conventional algorithms.

Introduction

1. Gene expression is modulated by enhancers and transcription factors.

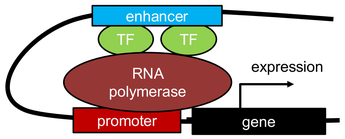
Enhancer

- regulatory DNA element (length ~ 300)
- contains binding sites for transcription factors (TF)



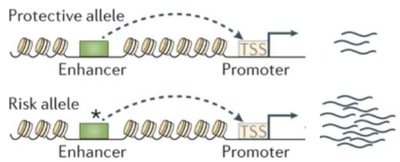
Transcription factor (TF)

- Key modulator of gene expression
- Cell-specific expression
- DNA-binding specificity

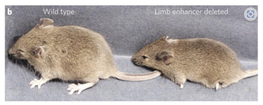


2. Enhancers are enriched with disease-associated variants

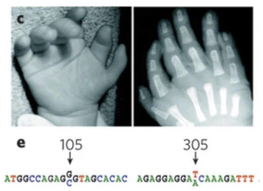
- >90% in non-coding regions (Maurano, Science 2012)
- Cancer, diabetes, etc.



(Sur, Taipale, Nature review cancer 2016)



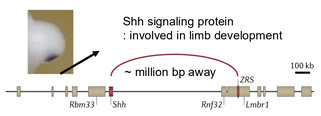
(Visel, Nature 2009)



3. Despite the medical importance, studying enhancer is challenging.

a. Enhancers are functionally complex

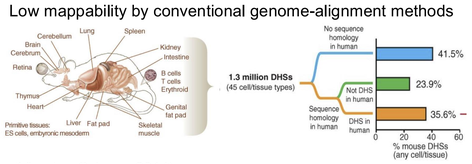
e.g., distal regulation



(Schoenfelder, Nature review genetics 2019)

b. Identifying conserved enhancers is difficult

Low mappability by conventional genome-alignment methods



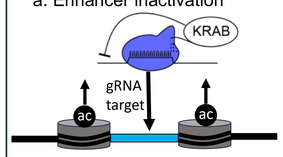
(Vierstra, Science 2014)

Research Goal: Develop quantitative methods and algorithms to facilitate functional characterization of enhancers

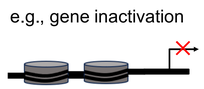
A. Modeling the regulatory dynamics of cell-state transition using CRISPRi screen

1. Enhancer characterization with CRISPRi

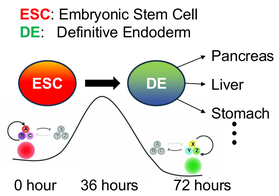
a. Enhancer inactivation

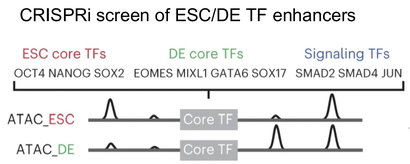


b. Measure downstream effect

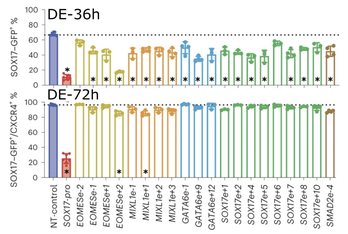


2. Functional screen of putative enhancers that drive ESC→DE transition



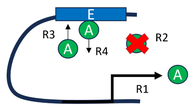


3. Perturbing TF enhancers slows down cell-state transition.

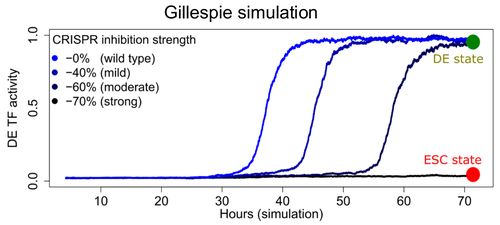


4. Simple discrete stochastic model recapitulates the CRISPRi-induced time-delay effect.

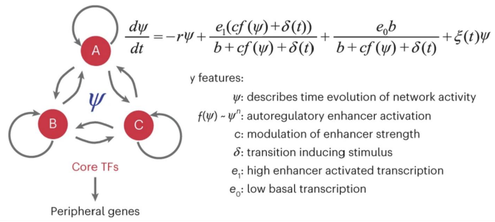
Autoregulatory circuit



Gillespie simulation



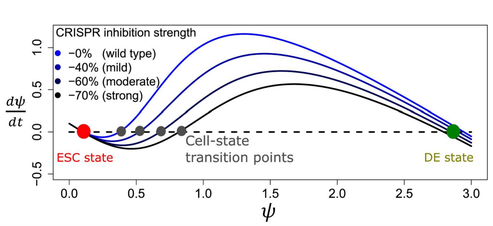
5. CRISPRi enhancer perturbation slows down cell-state transition by increasing the cell-state transition point.



$$\frac{d\psi}{dt} = -r\psi + \frac{e_1(cf(\psi)+\delta(t))}{b+cf(\psi)+\delta(t)} + \frac{e_0 b}{b+cf(\psi)+\delta(t)} + \xi(t)\psi$$

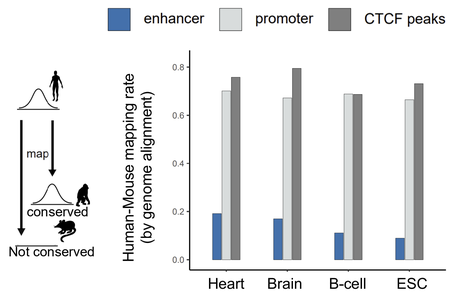
γ features:

- ψ: describes time evolution of network activity
- $f(\psi)$ - ψ^n: autoregulatory enhancer activation
- c: modulation of enhancer strength
- δ: transition inducing stimulus
- e_1: high enhancer activated transcription
- e_0: low basal transcription



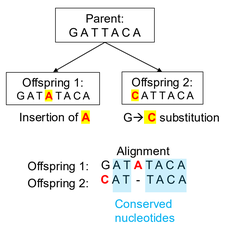
B. Developing novel genome-alignment algorithm for improved detection of conserved enhancers

1. Many human enhancers are unmappable to conserved mouse enhancers using conventional alignment methods.

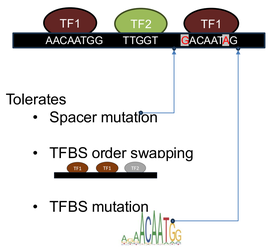


2. Flexible sequence structure and rapid evolution of enhancers make alignment difficult.

a. Alignment example



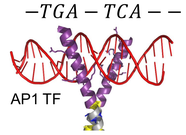
b. Enhancers allow accumulation of mutations



3. Enhancer sequences are effectively modeled using gapped-kmer sequence feature.

Gapped-kmer:

- Kmer with a fixed number of "gaps".
- Models sequence degeneracy
- e.g., gapped-kmer(10,6):

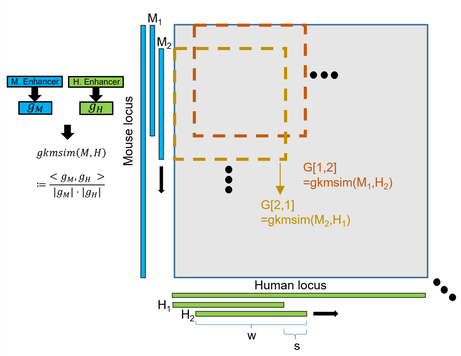


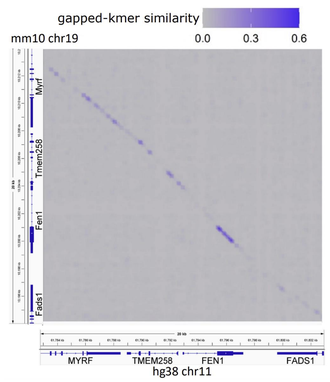
Gapped-kmer (11,7) representation of an enhancer (H)

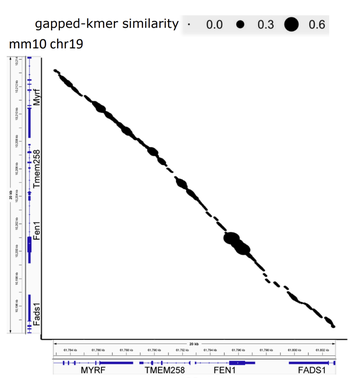
$$g^{H}_{11,7} = \#\ of \begin{pmatrix} AAAAAAA - - - - \\ AAAAAAC - - - - \\ \vdots \\ -TGA - TCA - A - \\ \vdots \end{pmatrix} in\ H$$

Experimental validations
(Yan et al Nature, 2021) , (Lee, Nat. Genetics, 2015), Shigaki, Human Mutation, 2019)

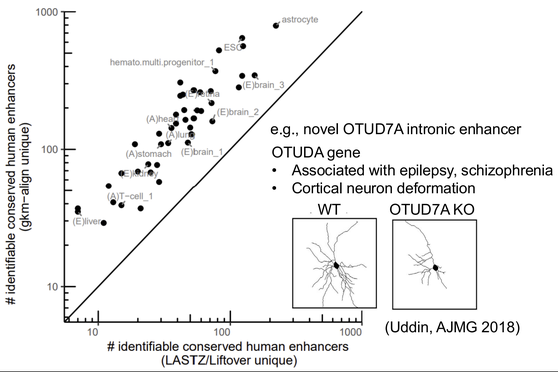
4. The novel *gkm-align* algorithm finds alignment path of maximum gapped kmer similarity to map conserved enhancers.







5. gkm-align discovers thousands of novel conserved enhancers of high biomedical interest.



e.g., novel OTUD7A intronic enhancer

OTUDA gene

- Associated with epilepsy, schizophrenia ..
- Cortical neuron deformation

(Uddin, AJMG 2018)

Publications

[1] **Oh, J.W.**, Yao, D., Tycko, J., Bounds., L, Gosai, S., Lataniotis, L., *et al.* Multicenter integrated analysis of noncoding CRISPRi screens. ***Nature Methods* (2024)**.
[2] Luo, R., Yan, J., **Oh, J.W.** *et al.* Dynamic network-guided CRISPRi screen identifies CTCF-loop-constrained nonlinear enhancer gene regulatory activity during cell state transitions. ***Nature Genetics* (2023).**
[3] **Oh, J.W**. & Beer, M. A. Gapped-kmer sequence modeling robustly identifies regulatory vocabularies and distal enhancers conserved between evolutionarily distant mammals. ***bioRxiv* (2023; under review)**