Scalable statistical test for identifying disease-associated variants in regulatory DNA

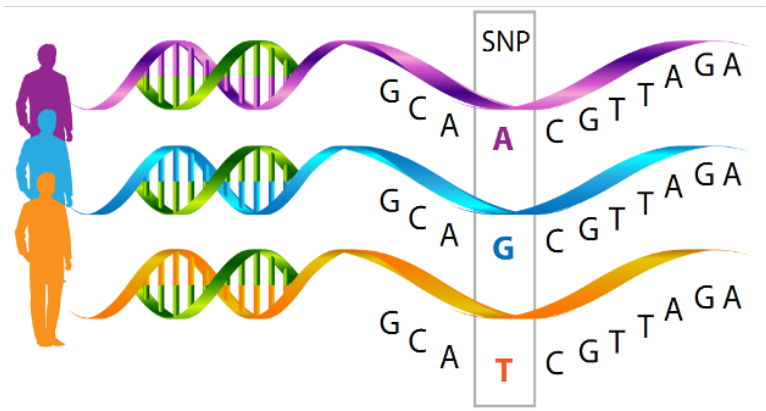
Sunyoung Shin

Department of Mathematics
POSTECH

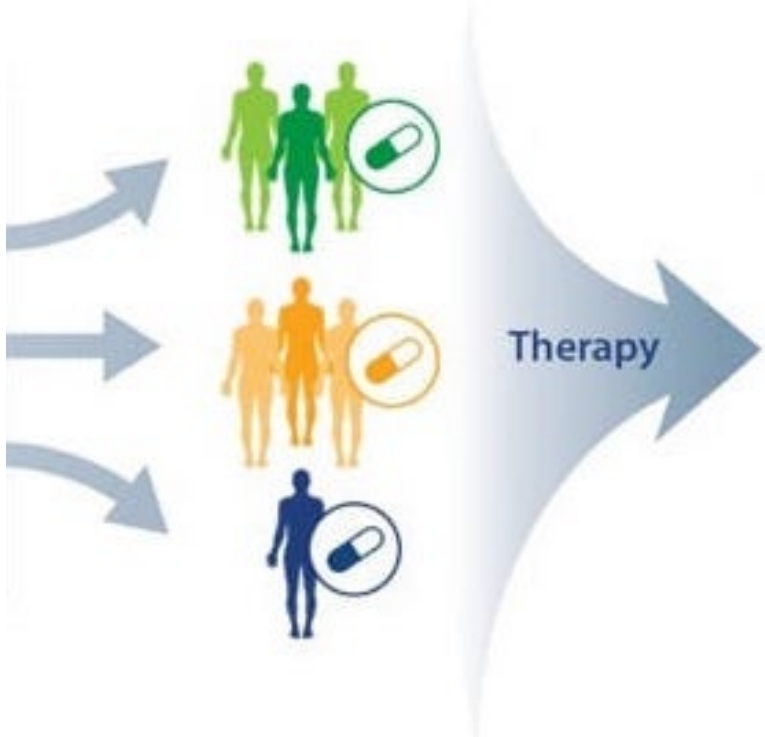
July 3, 2023

Personal Genomics

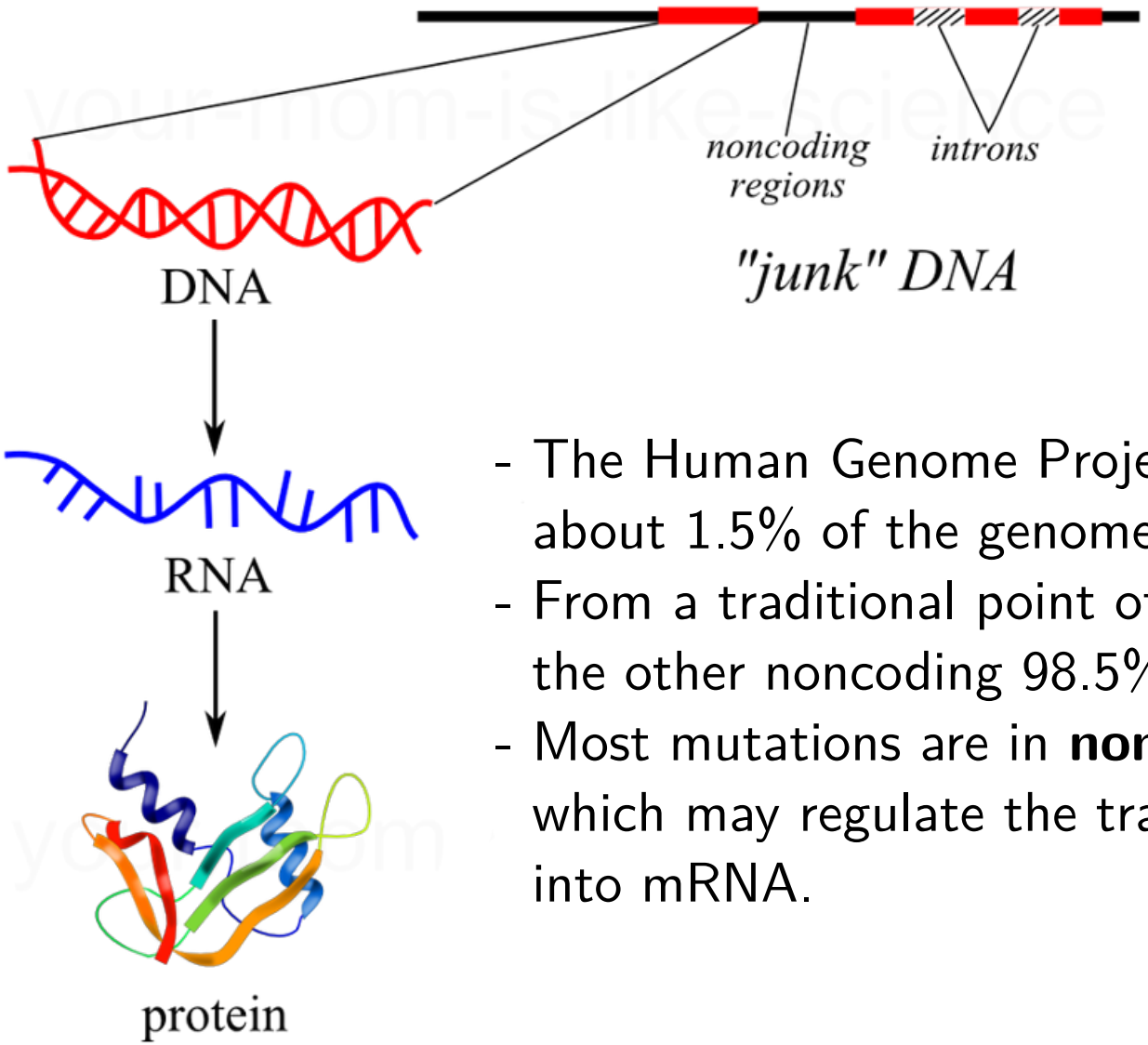
- There are 3 billion letters in the human genome.
- Patients with tumors that share the same genetic mutation receive the drug that targets the mutation.



(Source: https://medium.com)



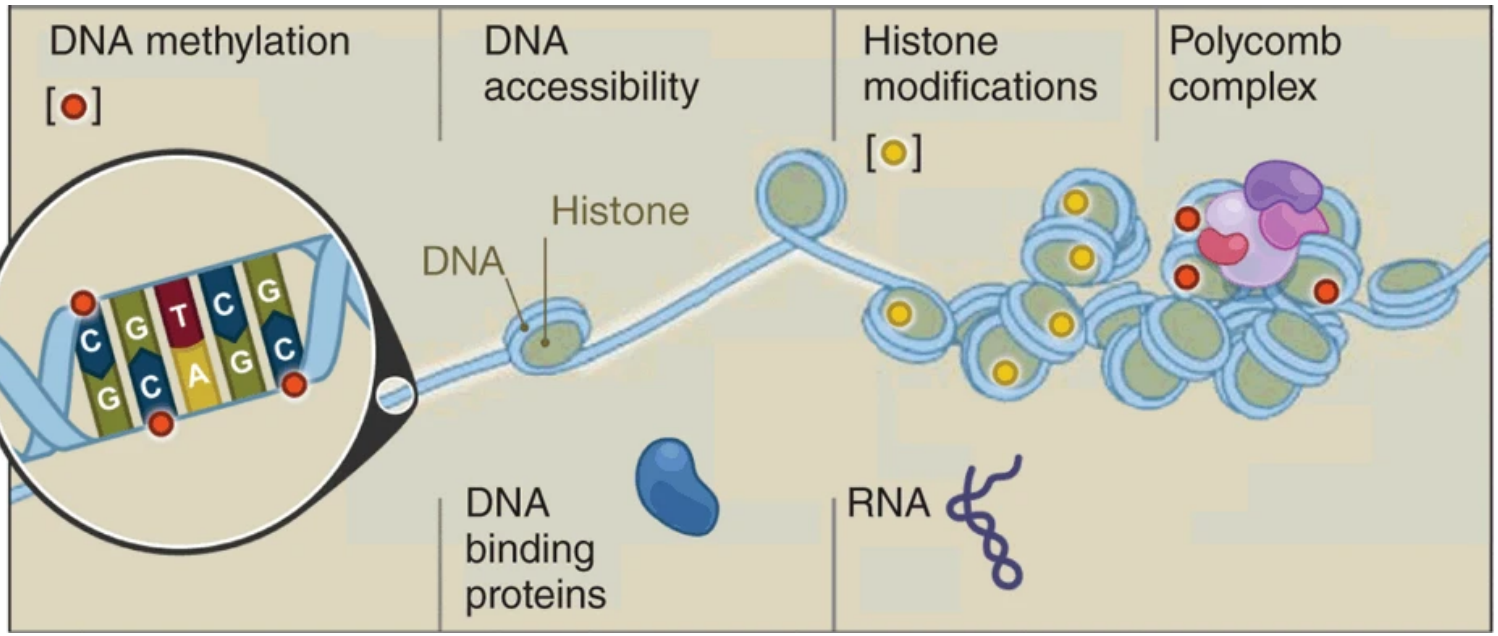
Noncoding variants



- The Human Genome Project found that only about 1.5% of the genome are coding DNA.
- From a traditional point of view, the other noncoding 98.5% was *junk DNA*.
- Most mutations are in **noncoding regions**, which may regulate the transcription of a gene into mRNA.

Epigenomic features

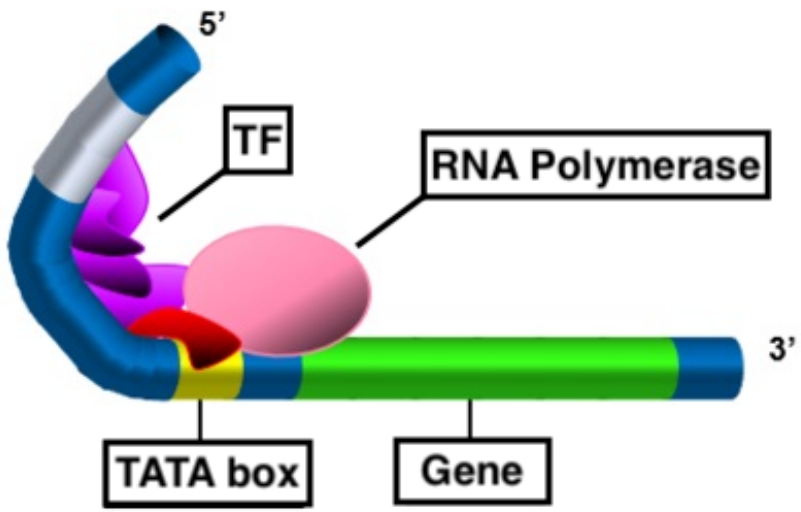
- Some epigenomic features
  - Open/closed chromatin
  - Histone modifications
  - Protein-binding to DNA
  - Protein-binding to RNA
  - DNA methylation
  - DNA looping



Bernstein et al. (2010)

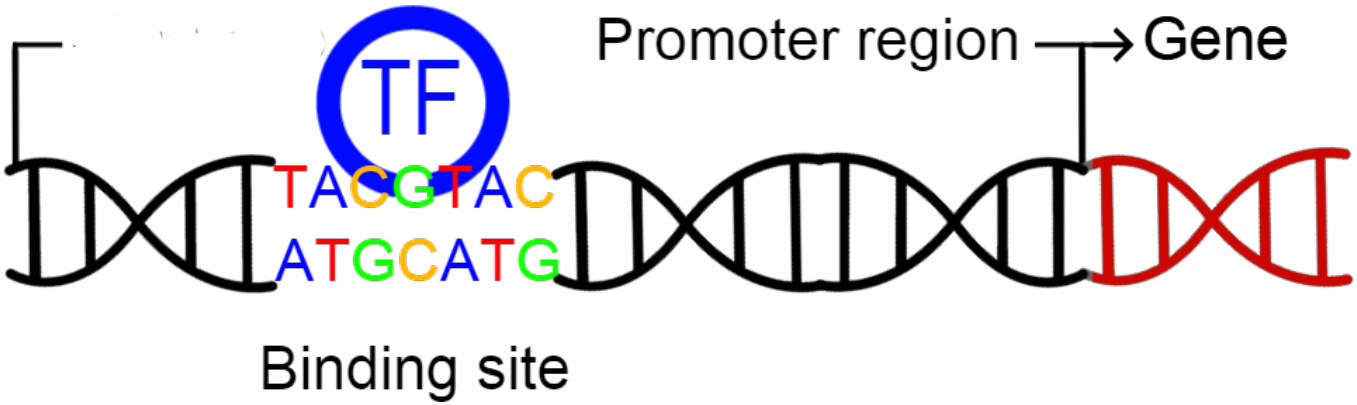
Epigenomic features

- Some epigenomic features
  - Open/closed chromatin
  - Histone modifications
  - Protein-binding to DNA
  - Protein-binding to RNA
  - DNA methylation
  - DNA looping



(Source: `http://pediaa.com`)

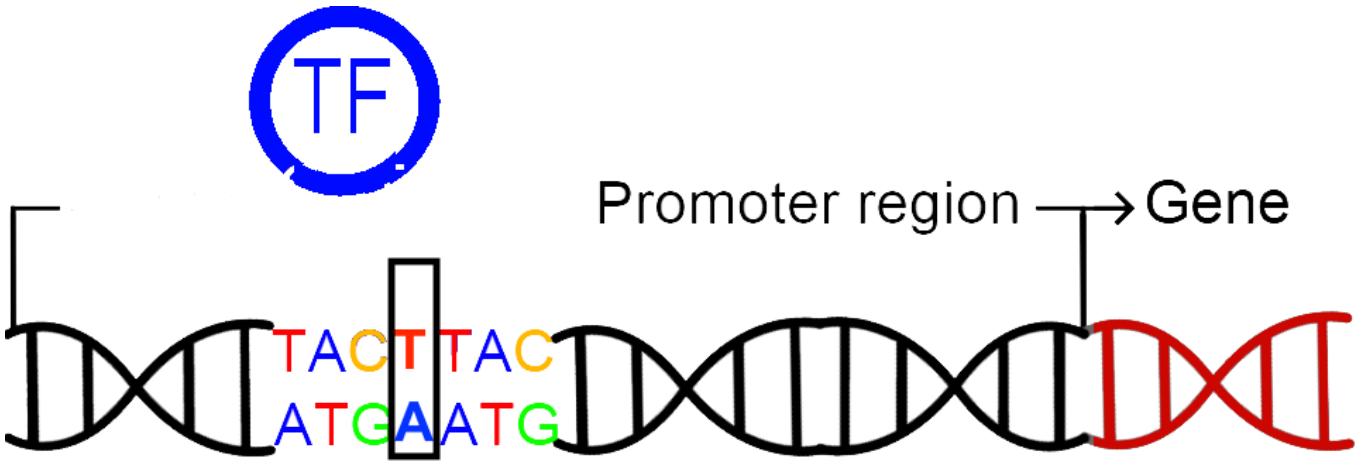
Protein-DNA binding



DNA is stretched out for illustration. (Source: bioinfo.ucc.ie)

- Thousands of proteins may *float in human body cells*.
- Transcription factor (TF) proteins that bind to DNA may control the rate or amount of mRNA (gene expression) copied from DNA.
- TF proteins tend to bind *specific DNA sequences*.

Protein-DNA binding changes due to mutations



TF protein unbinds to DNA due to a point mutation.

- Thousands of proteins may *float in human body cells*.
- Transcription factor (TF) proteins that bind to DNA may control the rate or amount of mRNA (gene expression) copied from DNA.
- TF proteins tend to bind *specific DNA sequences*.

Scalable test of statistical significance for protein-DNA binding changes

- Experiments for all combinations between proteins and mutations are not doable in practice.
  - An enormous number of mutations are in the human genome.
- It is useful to quantitatively evaluate the mutation influence on protein-DNA binding by means of a *statistical modeling*.
- We aim to nominate noncoding mutations that are responsible for diseases and find TFs that plays a role in diseases.
  - Noncoding mutations may regulate expression of disease associated genes through modification of TF-DNA binding.

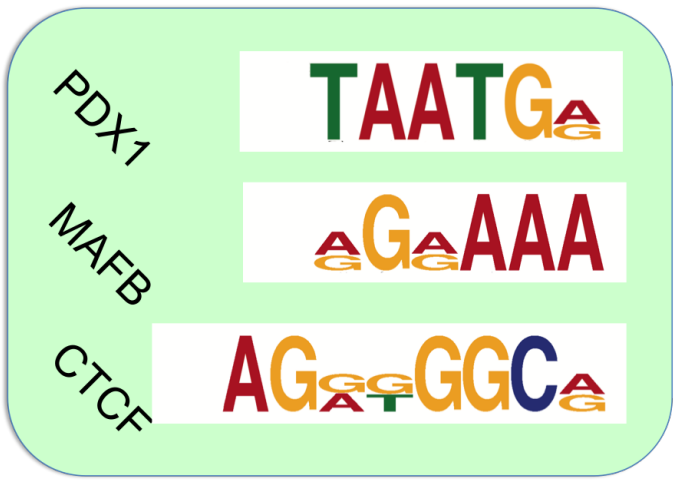
Protein-DNA binding change test for InDels

(Source: www.singerinstruments.com)

- About 15-20% of the mutations in the human genome are considered InDels. (Mullaney et al., 2010)
- Functional impact of InDels can be substantial compared to single nucleotide variants (SNVs).
- Protein-DNA binding change test for InDels
  - is the first novel quantitative approach to find protein-DNA binding changer InDels.
  - works for any lengths of the contrasting sequences, m.
  - easily speeds up the analysis of large-scale InDel mutation data.

Motif library: a large collection of binding sequence patterns

- A large number of TF-binding motifs have been discovered and become publicly available.
  - ENCODE provides at least 2,065 motifs. (Kheradpour and Kellis, 2014)
  - JASPAR provides at least 579 motifs.



Motif library example

Notations

- Consider a $4 \times L$ Position Probability Matrix (PPM) of a given motif of L positions W.
  - Each column contains the four probabilities $W(\cdot, l)$ s.t. $\sum_{k=1}^{4} W(k, l) = 1 \; \forall l \; l = 1, \ldots, L$.
- A contrasting sequence of m nucleotides in the forward strand is $\mathbf{y^c} = (y_L, \cdots, y_{L+m-1})$, $y_i \in \{1, 2, 3, 4\}$, $i = L, \ldots, L + m - 1$.
  - Example: 'TAT' in the insertion, and 'CA' in the deletion
- m and L are fixed.
- Given m and L,
  - $\mathbf{y} = (y_1, \cdots, y_{L-1}, y_L, y_{L+1}, \cdots, y_{L+m-1}, y_{L+m}, \cdots, y_{2L+m-2})$ is the longer sequence in the forward strand, $y_i \in \{1, 2, 3, 4\}$, $i = 1, \ldots, 2L + m - 2$.
  - $\mathbf{y^a} = (y_1, \cdots, y_{L-1}, y_{L+m}, \cdots, y_{2L+m-2})$ is the shorter sequence in the forward strand.
  - $\mathbf{y} = (\mathbf{y^a}_1, \mathbf{y^c}, \mathbf{y^a}_2)$ and $\mathbf{y^a} = (\mathbf{y^a}_1, \mathbf{y^a}_2)$, where $\mathbf{y^a}_1 = (y_1, \cdots, y_{L-1})$ and $\mathbf{y^a}_2 = (y_{L+m}, \cdots, y_{2L+m-2})$.

Markov chain background model

- Background DNA sequences are usually fitted by the Markov chain model as the occurrence of a nucleotide at a given position depends on the previous nucleotides in the sequence (Avery and Henderson, 1999; Menéndez et al., 2011; Reinert et al., 2000).
- The null model for the longer sequences is a stationary reversible first order Markov model with prior probabilities $\pi_0(k) = P(y_l = k),\ k = 1, \cdots, 4$, transition probabilities $a_0(k, n) = P(y_{l+1} = n | y_l = k),\ k, n = 1, \cdots, 4$:

$$f_{\mathcal{H}_0}(\mathbf{y}) = \pi_0(y_1) \prod_{l \in \{1,\ldots,2L+m-3\}} a_0(y_l, y_{l+1}). \qquad (1)$$

Binding score

- Define the binding score for a subsequence of the longer sequence $\mathbf{y}$, which starts at position s with a fixed length of L:

$$C(\mathbf{y}, s) = \sum_{l=1}^{L} \log W(y_{l+s-1}, l). \tag{2}$$

  - $s \in \{1, \ldots, L + m - 1\}$ is the protein binding start position.
- The binding score of the sequence $\mathbf{y}$ is defined as

$$C(\mathbf{y}) = \max_{s \in \{1,\ldots,L+m-1\}} \{C(T(\mathbf{y}), s) : T \in \{I, R\}\}. \tag{3}$$

  - I and R are the forward and reverse strand operators.

$$\begin{cases} I(\mathbf{y}) & = \mathbf{y} \\ R(\mathbf{y}) & = (5 - y_{2L+m-2}, 5 - y_{2L+m-3}, \cdots, 5 - y_1). \end{cases}$$

- Similarly, define the binding score of $\mathbf{y^a}$, denoted as $C^a(\mathbf{y^a})$.
- $\mathbf{y}$ has m more subsequences than $\mathbf{y^a}$.

Binding change score statistic

- The binding changes due to InDels are tested by comparing binding significance on $\mathbf{y}$ to that on $\mathbf{y^a}$.
- The TF binding p-value for a longer sequence $\mathbf{y_0}$ is

$$p_l(\mathbf{y_0}) = P\{C(\mathbf{y}) \geq C(\mathbf{y_0}) | \mathbf{y} \sim f_{\mathcal{H}_0}\}.$$

- The TF binding p-value for a shorter sequence $\mathbf{y_0^a}$ is

$$p_s(\mathbf{y_0^a}) = P\{C^a(\mathbf{y^a}) \geq C^a(\mathbf{y_0^a}) | \mathbf{y} \sim f_{\mathcal{H}_0}\}.$$

- Our BC test statistic for the pair $(\mathbf{y}, \mathbf{y^a})$, named "binding change score", is the difference between the logarithm of the binding p-values to $\mathbf{y}$ and $\mathbf{y^a}$:

$$T \equiv T(\mathbf{y}, \mathbf{y^a}) = \log\{p_s(\mathbf{y^a})\} - \log\{p_l(\mathbf{y})\}. \qquad (4)$$

Binding change score & p-value

- We can determine whether or not the binding is enhanced or disrupted from the sign of T.
- The test statistic T is a function of p-values.
  - Examples of p-value-based test statistics are higher criticism test statistic (Donoho et al., 2004), and its variation for binary regression (Mukherjee et al., 2015).
- For the observed sequence pair $(\mathbf{y_0}, \mathbf{y_0^a})$, define p-value

$$p(\mathbf{y_0}, \mathbf{y_0^a}) = 2 \cdot \min\{P(T \geq t_0 | \mathbf{y} \sim f_{\mathcal{H}_0}), P(T \leq t_0 | \mathbf{y} \sim f_{\mathcal{H}_0})\},$$

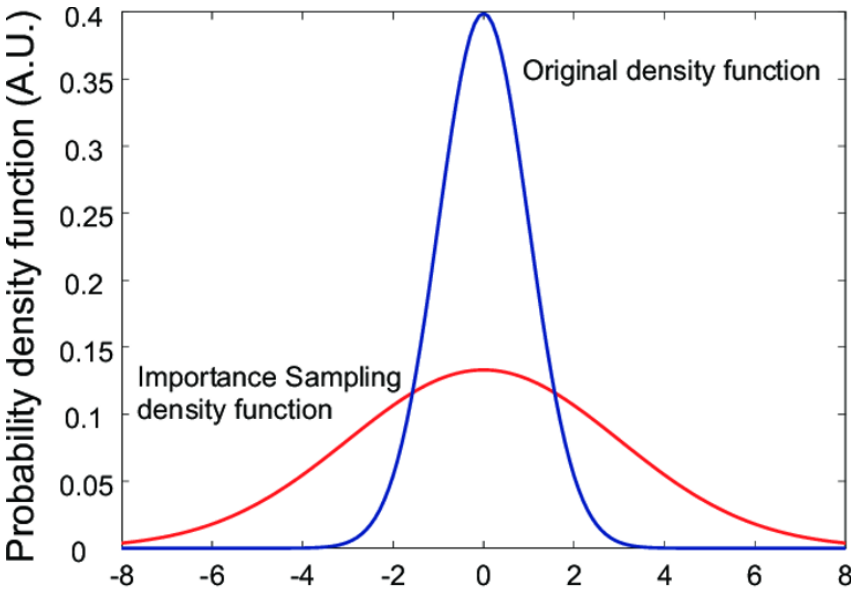
$$t_0 \equiv T(\mathbf{y_0}, \mathbf{y_0^a}) = \log\{p_s(\mathbf{y_0^a})\} - \log\{p_l(\mathbf{y_0})\}.$$

  - The null distribution of T is obtained with sequence pairs from $f_{\mathcal{H}_0}$.

Empirical p-value computation

- A practical challenge is the theoretical calculation of the null distribution of T for distinct values of m, L, W.
- We develop an efficient algorithm for empirical p-value computation of the BC test based on the importance sampling technique, requiring a much smaller number of sequence pairs to be simulated. (Kahn and Harris, 1951; Chan and Zhang, 2007; Chan et al., 2010)
- It is scalable for the BC tests on hundreds of thousands InDel mutations against thousands of binding motifs.
- The algorithm coded in *R* and *C++* is available in R package *atIndel*. (`https://github.com/sunyoungshin/atIndel`)

Importance sampling algorithm



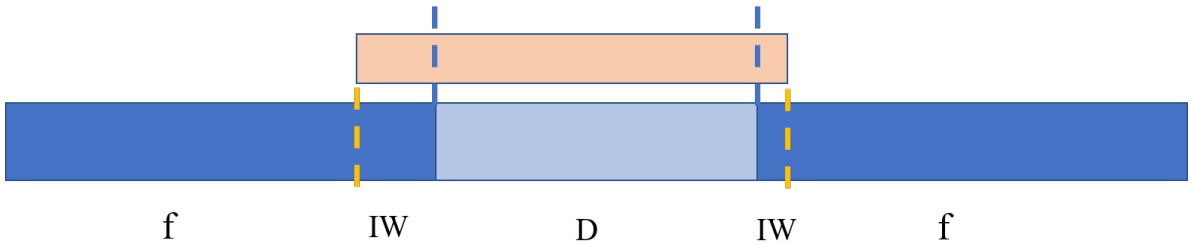
- Original distribution: $f_{\mathcal{H}_0}(\mathbf{y})$
- Importance distribution: $h_\theta(\mathbf{y})$
  - θ is a tilting parameter.

- Importance sampling Monte Carlo algorithm
  1. Sample sequences from the importance distribution $\mathbf{y} \sim h_\theta(y)$.
  2. Compute *binding change p-value* for $(\mathbf{y_0}, \mathbf{y_0^a})$ based on

$$p(\mathbf{y_0}, \mathbf{y_0^a}) = 2 \cdot \min \left[E[1\{T \geq t_0\} \cdot \frac{f_{\mathcal{H}_0}(\mathbf{y})}{h_\theta(\mathbf{y})} | \mathbf{y} \sim h_\theta(\mathbf{y})], \right.$$
$$\left. E[1\{T \leq t_0)\} \cdot \frac{f_{\mathcal{H}_0}(\mathbf{y})}{h_\theta(\mathbf{y})} | \mathbf{y} \sim h_\theta(\mathbf{y})] \right].$$

The p-value is estimated by the weighted frequency.

Conditional importance distribution



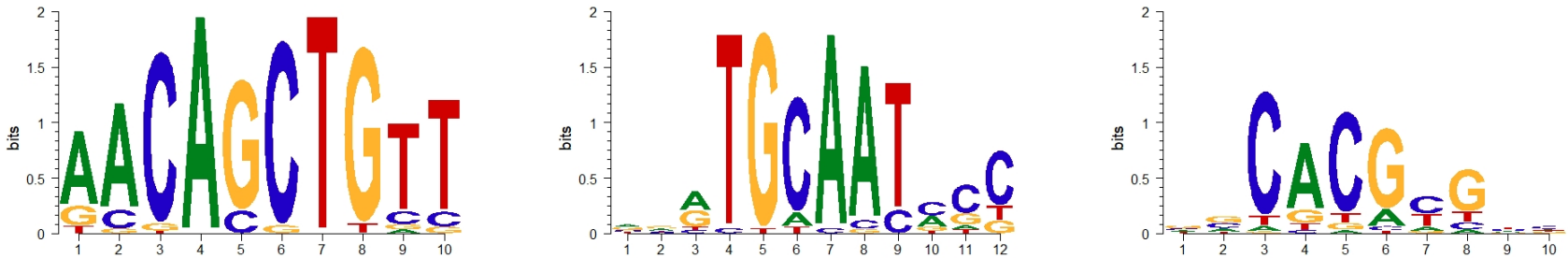
Overlapping example between the protein binding site and the longer sequence

- Conditional importance distribution of $\mathbf{y}$ given starting position s

$$h_\theta(\mathbf{y}|s) = \frac{1}{M_s(\theta)} f(y_1, \ldots, y_{s-1})^{1(s\geq 2)} \left[\prod_{l=s}^{L-1} IW(y_l, l-s+1)\right]^{1(1\leq s\leq L-1)}$$
$$\left[\prod_{c=\max(L,s)}^{\min(s,m)+L-1} D(y_c, c-s+1)^\theta\right] \left[\prod_{l=L+m}^{L+s-1} IW(y_l, l-s+1)\right]^{1(m+1\leq s\leq L+m-1)}$$
$$f(y_{L+s}, \ldots, y_{2L+m-2})^{1(s\leq L+m-2)}$$

- $M_s(\theta)$ is the normalizing constant.

Simulations under the null model



MSC Ddit3::Cebpa Hes1

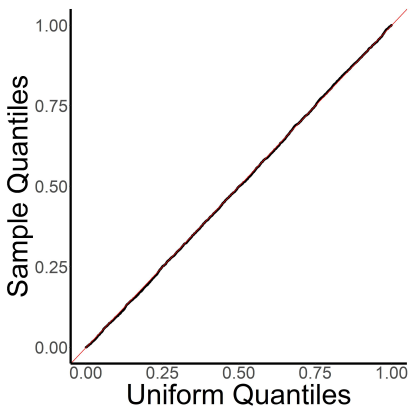
Figure: Three JASPAR motifs

- Obtain π_0, a_0 based on the human reference genome version GRCh37 (hg19).
- 10,000 sample sequences are generated from the first-order Markov model $f_{\mathcal{H}_0}(\mathbf{y})$.
- The length of the sample longer sequences is 28 and the contrasting sequence length is 6 ($L = 12, m = 6$).
- The Monte Carlo sample size for the algorithm is 2,000.

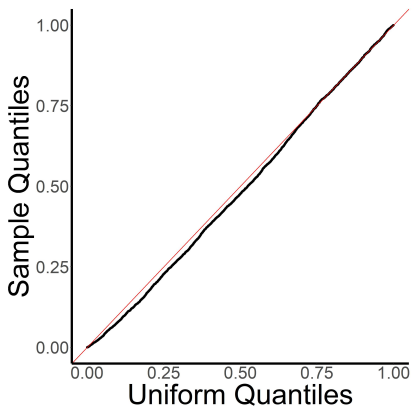
Simulations under the null model

	Empirical rejection probability		
p-value	MSC	Ddit3::Cebpa	Hes1
0.01	0.0138	0.0176	0.0094
0.05	0.0527	0.0684	0.0502
0.10	0.1038	0.1164	0.1008

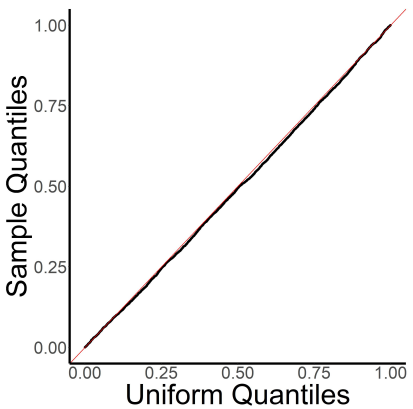
Empirical rejection probabilities of binding changer test



MSC



Ddit3::Cebpa



Hes1

Figure: Q-Q plots of the p-values from the BC tests under the null model

Simulations under a defined set of alternative models

- 2,000 sample sequences are generated from the full alternative model $f_{\mathcal{H}_1}(\mathbf{y})$ such that the longer sequences may have the nucleotide pattern of the motif while the shorter sequences lack the pattern.
- The contrasting sequence length is equal to the motif length ($m = L$).
- The probability mass function of $f_{\mathcal{H}_1}(\mathbf{y})$ is

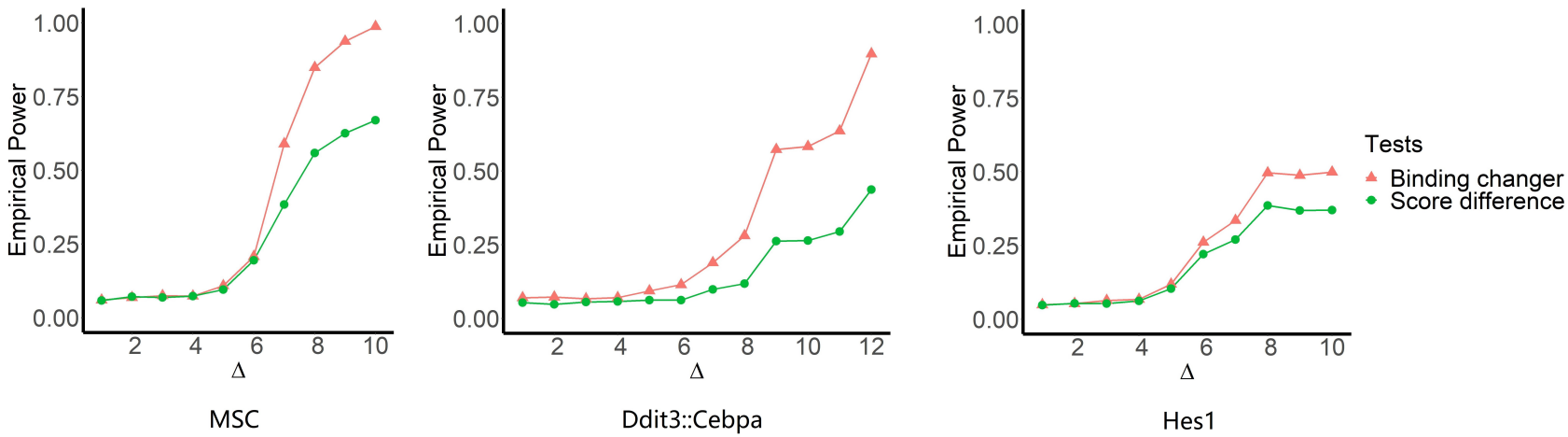
$$f_{\mathcal{H}_1}(\mathbf{y}) = \sum_{\{y_L,\cdots,y_{2L-1}\}\in\{1,2,3,4\}^L} f_{\mathcal{H}_0}(\mathbf{y}) \prod_{l\in\{L,\cdots,2L-1\}} W(y_l, l-L+1).$$

Simulations under a defined set of alternative models

- Further, 2,000 sample sequences are generated from each of the local alternative models $f_{\mathcal{H}_{1\Delta}}(\mathbf{y})$:

$$\sum_{\{y_L,\cdots,y_{L+\Delta-1}\}\in\{1,2,3,4\}^{\Delta}} f_{\mathcal{H}_0}(\mathbf{y}) \prod_{l\in\{L,\cdots,L+\Delta-1\}} W(y_l, l-L+1)$$

  - $\Delta \in \{1, \ldots L\}$ is the number of bases in the contrasting sequence following W.



Power curves evaluated with significance level 0.05

Analysis of AML InDel data

- 5,737 somatic InDels in samples of primary acute myeloid leukemia (AML) reported by Li et al. (2020).
  - Obtained from enhancer regions.

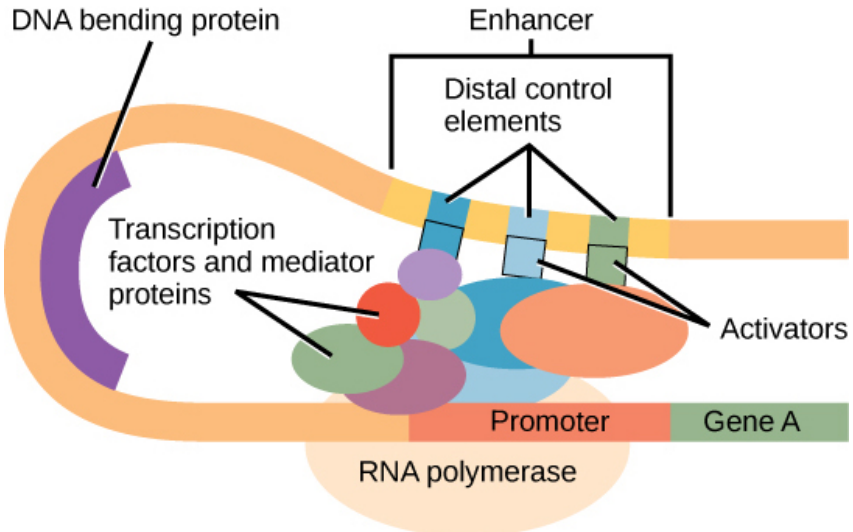


Figure: Enhancers is a region of DNA that can be bound by proteins. (Source: https://www.oercommons.org/courseware/lesson/15017/overview)

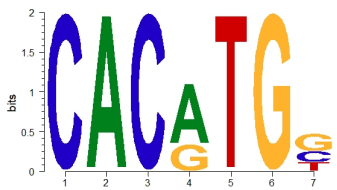
- Used mutation callers Strelka (Saunders et al., 2012) and Scalpel (Fang et al., 2016).

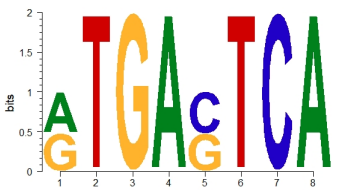
Analysis of AML InDel data

1	2	3	4	5	6 - 10	11 - 57	Total
4,061	858	237	202	69	200	110	5,737
70.8%	15.0%	4.1%	3.5%	1.2%	3.5%	1.9%	100%

Distribution of contrasting sequence lengths (m)

- MYC motifs and negative-control motifs
  - MYC is an important transcription factor and prognostic marker for AML (Salvatori et al., 2011).



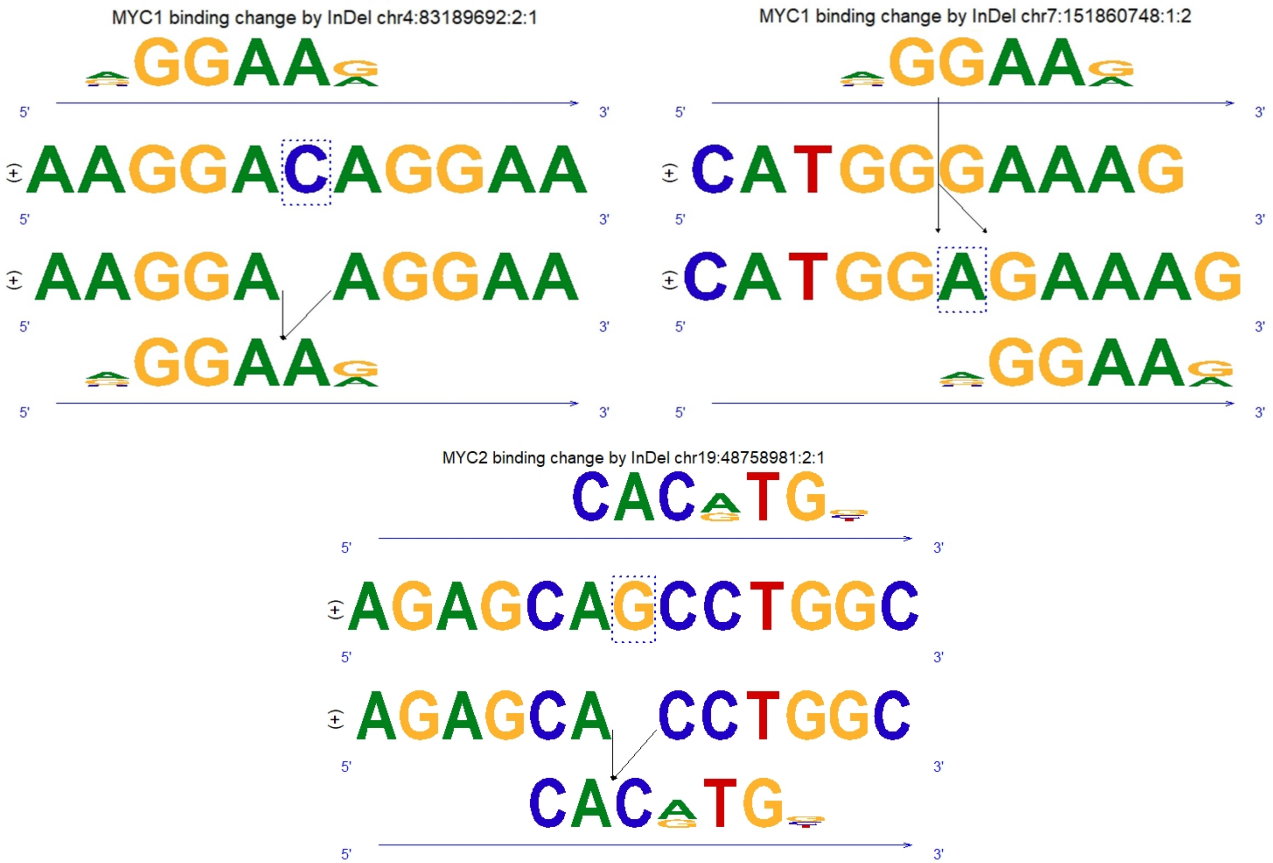


  - SOX2 and OCT4 are negative controls. These two TFs are important for embryonic stem cells, but they are not expressed in leukemia cell lines (Chambers and Tomlinson, 2009)

Binding changer InDels

- 28.685 binding change tests are conducted.
  - Monte Carlo sample size was set to 10,000.
  - *7 InDel mutations* were identified by the following criteria:
    - Benjamni Hochberg adjusted p-value < 0.10
    - At least one of the two binding p-values $(p_l,\ p_s) < 0.05$

Binding changer InDels



- The nearest genes CARD8, ZNF114, RUNX1 are differentially expressed between AML tumors and normal cells (Tang et al., 2019).
- One nearest gene HNRNPD is a MYC target gene from Gene Set Enrichment Analysis (Subramanian et al., 2005).

Future work

- Joint investigation of a collection of mutations that reside in a broader site
- Examination of combinatory TFs that assemble super-enhancers (Huang et al., 2016; Liu et al., 2017, 2020)
- Integrative analysis with scATAC-seq and scRNA-seq data (Suen et al., 2023)

Research Collaborators & Acknowledgement

- Qinyi Zhou, Min Chen at UT Dallas
- Chandler Zuo
- Yuannyu Zhang, Jian Xu at UT Southwestern

National Institutes of Health
Turning Discovery Into Health

CANCER PREVENTION & RESEARCH
INSTITUTE OF TEXAS

Bibliography I

Peter J Avery and Daniel A Henderson. Fitting Markov chain models to discrete state series such as DNA sequences. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(1):53–61, 1999.

Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra, Arthur L Beaudet, Joseph R Ecker, et al. The nih roadmap epigenomics mapping consortium. *Nature biotechnology*, 28(10):1045–1048, 2010.

Ian Chambers and Simon R Tomlinson. The transcriptional foundation of pluripotency. 2009.

Hock Peng Chan and Nancy Ruonan Zhang. Scan statistics with weighted observations. *Journal of the American Statistical Association*, 102(478):595–602, 2007.

Hock Peng Chan, Nancy Ruonan Zhang, and Louis H.Y. Chen. Importance sampling of word patterns in DNA and protein sequences. *Journal of Computational Biology*, 17(12):1697–1709, 2010. doi: 10.1089/cmb.2008.0233.

David Donoho, Jiashun Jin, et al. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

Han Fang, Ewa A Bergmann, Kanika Arora, Vladimir Vacic, Michael C Zody, Ivan Iossifov, Jason A O'Rawe, Yiyang Wu, Laura T Jimenez Barron, Julie Rosenbaum, et al. Indel variant analysis of short-read sequencing data with Scalpel. *Nature Protocols*, 11(12):2529–2548, 2016.

Bibliography II

Jialiang Huang, Xin Liu, Dan Li, Zhen Shao, Hui Cao, Yuannyu Zhang, Eirini Trompouki, Teresa V. Bowman, Leonard I. Zon, Guo-Cheng Yuan, Stuart H. Orkin, and Jian Xu. Dynamic control of enhancer repertoires drives lineage and stage-specific transcription during hematopoiesis. *Developmental Cell*, 36(1):9–23, 2016. ISSN 1534-5807. doi: https://doi.org/10.1016/j.devcel.2015.12.014. URL `https://www.sciencedirect.com/science/article/pii/S1534580715007996`.

Herman Kahn and Theodore E Harris. Estimation of particle transmission by random sampling. *National Bureau of Standards applied mathematics series*, 12:27–30, 1951.

Pouya Kheradpour and Manolis Kellis. Systematic discovery and characterization of regulatory motifs in encode tf binding experiments. *Nucleic acids research*, 42(5): 2976–2987, 2014.

Kailong Li, Yuannyu Zhang, Xin Liu, Yuxuan Liu, Zhimin Gu, Hui Cao, Kathryn E. Dickerson, Mingyi Chen, Weina Chen, Zhen Shao, Min Ni, and Jian Xu. Noncoding variants connect enhancer dysregulation with nuclear receptor signaling in hematopoietic malignancies. *Cancer Discovery*, 10(5):724–745, 2020. ISSN 2159-8274. doi: 10.1158/2159-8290.CD-19-1128. URL `https://cancerdiscovery.aacrjournals.org/content/10/5/724`.

Bibliography III

Xin Liu, Yuannyu Zhang, Yong Chen, Mushan Li, Feng Zhou, Kailong Li, Hui Cao, Min Ni, Yuxuan Liu, Zhimin Gu, Kathryn E. Dickerson, Shiqi Xie, Gary C. Hon, Zhenyu Xuan, Michael Q. Zhang, Zhen Shao, and Jian Xu. In Situ Capture of Chromatin Interactions by Biotinylated dCas9. *Cell*, 170(5):1028–1043.e19, 2017. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2017.08.003. URL `https://www.sciencedirect.com/science/article/pii/S0092867417308917`.

Xin Liu, Yong Chen, Yuannyu Zhang, Yuxuan Liu, Nan Liu, Giovanni A Botten, Hui Cao, Stuart H Orkin, Michael Q Zhang, and Jian Xu. Multiplexed capture of spatial configuration and temporal dynamics of locus-specific 3D chromatin by biotinylated dCas9. *Genome Biology*, 21(1):1–20, 2020.

ML Menéndez, L Pardo, MC Pardo, and Konstantinos Zografos. Testing the order of Markov dependence in DNA sequences. *Methodology and computing in applied probability*, 13(1):59–74, 2011.

Rajarshi Mukherjee, Natesh S Pillai, and Xihong Lin. Hypothesis testing for high-dimensional sparse binary regression. *Annals of Statistics*, 43(1):352–381, 2015.

Julienne M Mullaney, Ryan E Mills, W Stephen Pittard, and Scott E Devine. Small insertions and deletions (indels) in human genomes. *Human molecular genetics*, 19 (R2):R131–R136, 2010.

Bibliography IV

Gesine Reinert, Sophie Schbath, and Michael S Waterman. Probabilistic and statistical properties of words: an overview. *Journal of Computational Biology*, 7 (1-2):1–46, 2000.

Beatrice Salvatori, Ilaria Iosue, Nkerorema Djodji Damas, Arianna Mangiavacchi, Sabina Chiaretti, Monica Messina, Fabrizio Padula, Anna Guarini, Irene Bozzoni, Francesco Fazi, et al. Critical role of c-Myc in acute myeloid leukemia involving direct regulation of miR-26a and histone methyltransferase EZH2. *Genes & Cancer*, 2(5):585–592, 2011.

Christopher T Saunders, Wendy SW Wong, Sajani Swamy, Jennifer Becq, Lisa J Murray, and R Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, 28(14):1811–1817, 2012.

Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

Hoi Ching Suen, Shitao Rao, Alfred Chun Shui Luk, Ruoyu Zhang, Lele Yang, Huayu Qi, Hon Cheong So, Robin M Hobbs, Tin-lap Lee, and Jinyue Liao. The single-cell chromatin accessibility landscape in mouse perinatal testis development. *Elife*, 12: e75624, 2023.

Bibliography V

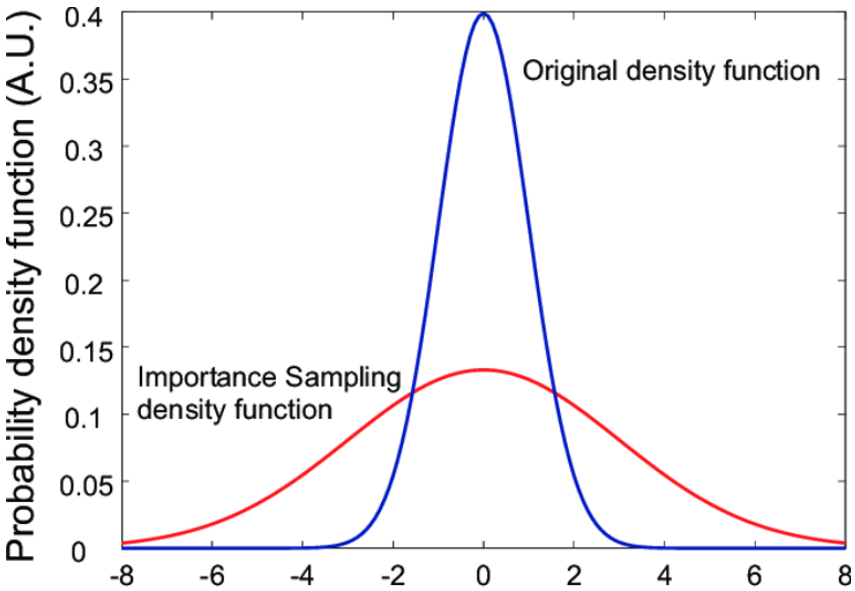
Zefang Tang, Boxi Kang, Chenwei Li, Tianxiang Chen, and Zemin Zhang. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Research*, 47(W1):W556–W560, 2019.

Thank you!

Any questions?

Please feel free to reach out to me at
sunyoungshin@postech.ac.kr.

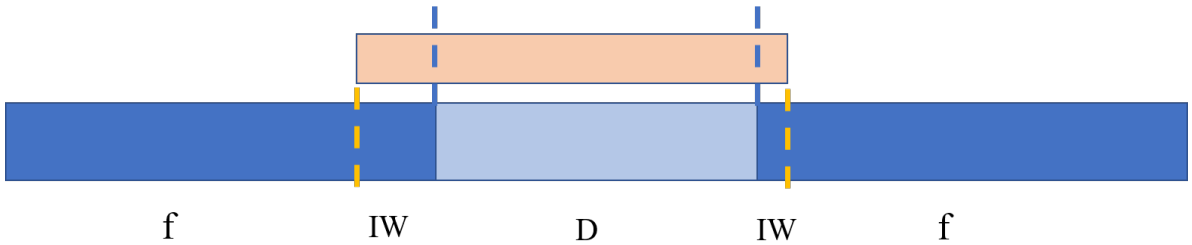
Importance sampling algorithm



- Original distribution: $f_{\mathcal{H}_0}(\mathbf{y})$
- Importance distribution: $h_\theta(\mathbf{y})$
  - θ is a tilting parameter.

- For small p-values, naive Monte-Carlo based on $f_{\mathcal{H}_0}(\mathbf{y})$ needs a large number of samples.
- Importance sampling based on $h_\theta(\mathbf{y})$ generates many larger scores in the target region. The number of simulations needed is significantly reduced.

Conditional importance distribution



Overlapping example between the protein binding site and the longer sequence

- How to specify $h_\theta(\mathbf{y}|s)$?
  - $D(k,l) = \exp\left\{ \sum_{j=1}^{4} \pi_0(j) \Big(\log W(k,l) - \log W(j,l) \Big) \right\}$
  - $IW(\cdot,l) = \{W(\cdot,l) + 1/4\}/2$
  - $f(\cdot)$ is the first-order Markov chain with π_0 and a_0.

Importance distribution

- The marginal distribution of s is a multinomial distribution with probability mass function

$$h_\theta(s) = \frac{M_s(\theta)}{M(\theta)}, \ s = 1, \ldots, L + m - 1$$

  - $M(\theta) = \sum_{s=1}^{L+m-1} M_s(\theta)$.
- Importance distribution of $\mathbf{y}$

$$h_\theta(\mathbf{y}) = \sum_{s=1}^{L+m-1} h_\theta(\mathbf{y}, s) = \sum_{s=1}^{L+m-1} h_\theta(\mathbf{y}|s) h_\theta(s).$$

Tilting parameter choice

- Under the importance sampling distribution with the optimal tilting parameter, it is not rare to obtain the observed score difference.
- Estimation of the tilting parameter θ
  - Consider $\mathbf{y}' = (y'_s, \ldots, y'_{s+L-1})$ that follows zero-order Markov model with $\pi(\cdot)$
  - Set the observed binding score change $C(\mathbf{y_0}) - C^a(\mathbf{y_0^a})$ to be equal to the expected score difference between $(y_s, \ldots, y_{s+L-1})$ and $\mathbf{y}'$.

$$
\begin{aligned}
& E_{\mathbf{y},\mathbf{y}',s}\left\{ \sum_{j=1}^{L} \left(\log W(y_{j+s-1}, j) - \log W(y'_{j+s-1}, j) \right) \right\} \\
=& C(\mathbf{y_0}) - C^a(\mathbf{y_0^a})
\end{aligned}
$$

  - Solve for θ.

Tilting parameter choice

Proposition 1 (Expectation of score difference)

Suppose that a random vector of the sequence and the binding start position $(\mathbf{y}, s)$ *follows the importance distribution* $h_\theta(\mathbf{y}, s)$*. Further, suppose that* $\mathbf{y}' = (y'_s, \ldots, y'_{s+L-1})$ *is independent of* $(\mathbf{y}, s)$ *and follows zero-order Markov model with* $\pi(\cdot)$*. The expected binding score difference between the binding subsequence* $(y_s, \ldots, y_{s+L-1})$ *and the subsequence* $\mathbf{y}'$ *is as follows:*

$$
\begin{aligned}
& E_{\mathbf{y},\mathbf{y}',s}\left\{ \sum_{j=1}^{L} \left(\log W(y_{j+s-1}, j) - \log W(y'_{j+s-1}, j) \right) \right\} \\
&= \sum_{s=1}^{L+m-1} \frac{M_s(\theta)}{M(\theta)} \Bigg[\sum_{i<\max(L,s) \text{ or } i\geq\min(s,m+L)} \left\{ \sum_{k=1}^{4} \left(IW(k, i-s+1) - \pi(k) \right) \log W(k, i-s+1) \right\} \\
&\quad + \sum_{i=\max(s,L)}^{\min(s,m+L-1)} \frac{\sum_k D(k, i-s+1)^\theta \log D(k, i-s+1)}{\sum_k D(k, i-s+1)^\theta} \Bigg].
\end{aligned}
$$

Importance sampling Monte Carlo algorithm

1. Generate N independent longer sequences $\{\mathbb{Y}_t : t = 1, \ldots, N\} \sim h_\theta(\mathbf{y})$ and obtain shorter sequences $\{\mathbb{Y}_t^a : t = 1, \ldots, N\}$ by removing the contrasting sequences.
2. Compute the empirical estimators of p_l and p_s:

$$\widehat{p}_l(\mathbb{Y}_t) = \frac{1}{N-1} \sum_{s \neq t} 1\{C(\mathbb{Y}_s) \geq C(\mathbb{Y}_t)\} \frac{f_{\mathcal{H}_0}(\mathbb{Y}_s)}{h_\theta(\mathbb{Y}_s)}$$

$$\widehat{p}_s(\mathbb{Y}_t^a) = \frac{1}{N-1} \sum_{s \neq t} 1\{C(\mathbb{Y}_s^a) \geq C(\mathbb{Y}_t^a)\} \frac{f_{\mathcal{H}_0}^a(\mathbb{Y}_s^a)}{h_\theta^a(\mathbb{Y}_s^a)},$$

   $f_{\mathcal{H}_0}^a$ is chosen to be a first order Markov model with π_0, a_0.
3. Compute the empirical p-values of the observed $(\mathbf{y_0}, \mathbf{y_0^a})$:

$$\widehat{p}_l(\mathbf{y_0}) = \frac{1}{N} \sum_{s=1}^{N} 1\{C(\mathbb{Y}_s) \geq C(\mathbf{y_0})\} \frac{f_{\mathcal{H}_0}(\mathbb{Y}_s)}{h_\theta(\mathbb{Y}_s)}$$

$$\widehat{p}_s(\mathbf{y_0^a}) = \frac{1}{N} \sum_{s=1}^{N} 1\{C^a(\mathbb{Y}_s^a) \geq C^a(\mathbf{y_0^a})\} \frac{f_{\mathcal{H}_0}^a(\mathbb{Y}_s^a)}{h_\theta^a(\mathbb{Y}_s^a)}.$$

Importance sampling Monte Carlo algorithm

4. Obtain N binding changer test statistics corresponding to the simulated sequences:

$$\mathbb{T}_t \equiv \log\{\widehat{p_s}(\mathbb{Y}_t^a)\} - \log\{\widehat{p_l}(\mathbb{Y}_t)\}.$$

5. Obtain the observed test statistic for $(\mathbf{y_0}, \mathbf{y_0^a})$:

$$T_0 \equiv \log\{\widehat{p_s}(\mathbf{y_0^a})\} - \log\{\widehat{p_l}(\mathbf{y_0})\}.$$

6. Compute our target p-value of the observed $(\mathbf{y_0}, \mathbf{y_0^a})$:

$$\widehat{p}(\mathbf{y_0}, \mathbf{y_0^a}) = 2 \cdot \min\left[\frac{1}{N}\sum_{t=1}^{N} 1\{\mathbb{T}_t \geq T_0\}\frac{f_{\mathcal{H}_0}(\mathbb{Y}_t)}{h_\theta(\mathbb{Y}_t)}, \frac{1}{N}\sum_{t=1}^{N} 1\{\mathbb{T}_t \leq T_0\}\frac{f_{\mathcal{H}_0}(\mathbb{Y}_t)}{h_\theta(\mathbb{Y}_t)}\right]$$

Simulations under misspecification

- We examine if the Type I error is well controlled even when the background sequences are not from the first-order Markov chain model.
  - The null model is independent multinomial distribution.

	Empirical rejection probability		
p-value	MSC	Ddit3::Cebpa	Hes1
0.01	0.0135	0.0172	0.0093
0.05	0.0523	0.0691	0.0504
0.10	0.1027	0.1218	0.0996

  - The null model is the fifth-order Markov model.

	Empirical rejection probability		
p-value	MSC	Ddit3::Cebpa	Hes1
0.01	0.0185	0.0170	0.0078
0.05	0.0602	0.0635	0.0390
0.10	0.1135	0.1106	0.0843