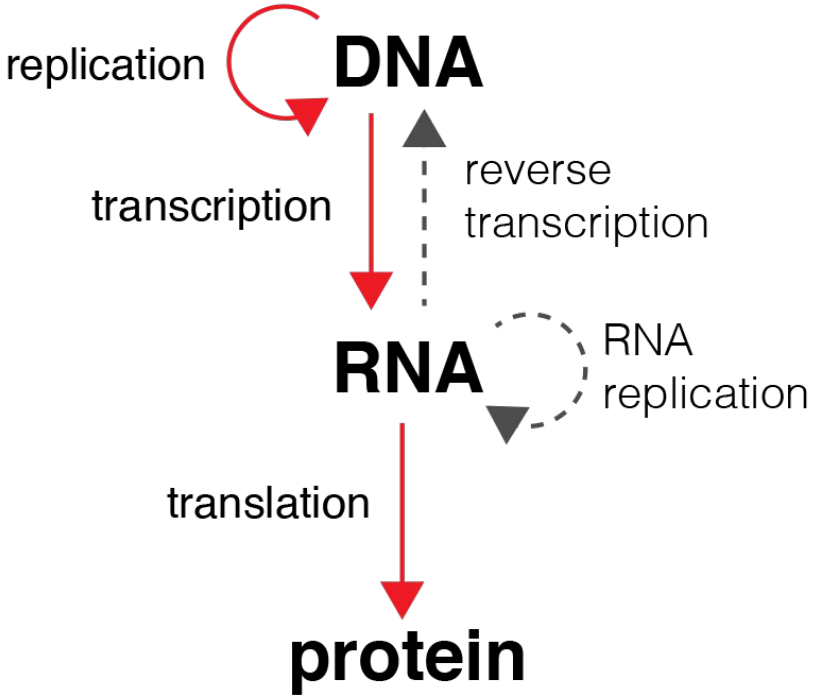
應用資料探勘探討轉錄因子與選擇性剪切的潛在模式

吳旻昇、林子傑

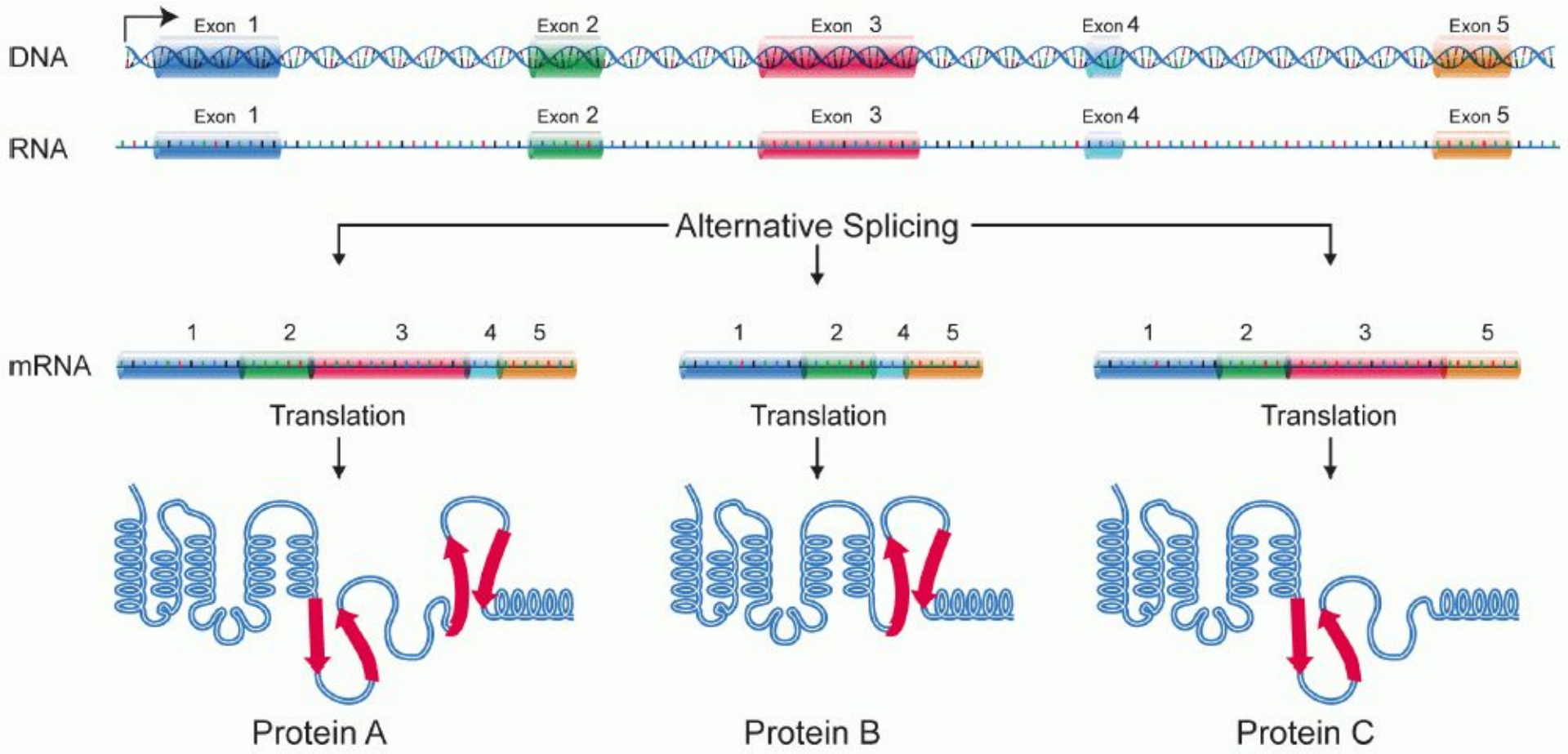
分子生物學的中心教條

- DNA (遺傳訊息) →
  RNA (抄本) →
  蛋白質 (生理功能)
- 高等生物有非常複雜的生理功能
- 基因數量與複雜程度沒有絕對的正相關

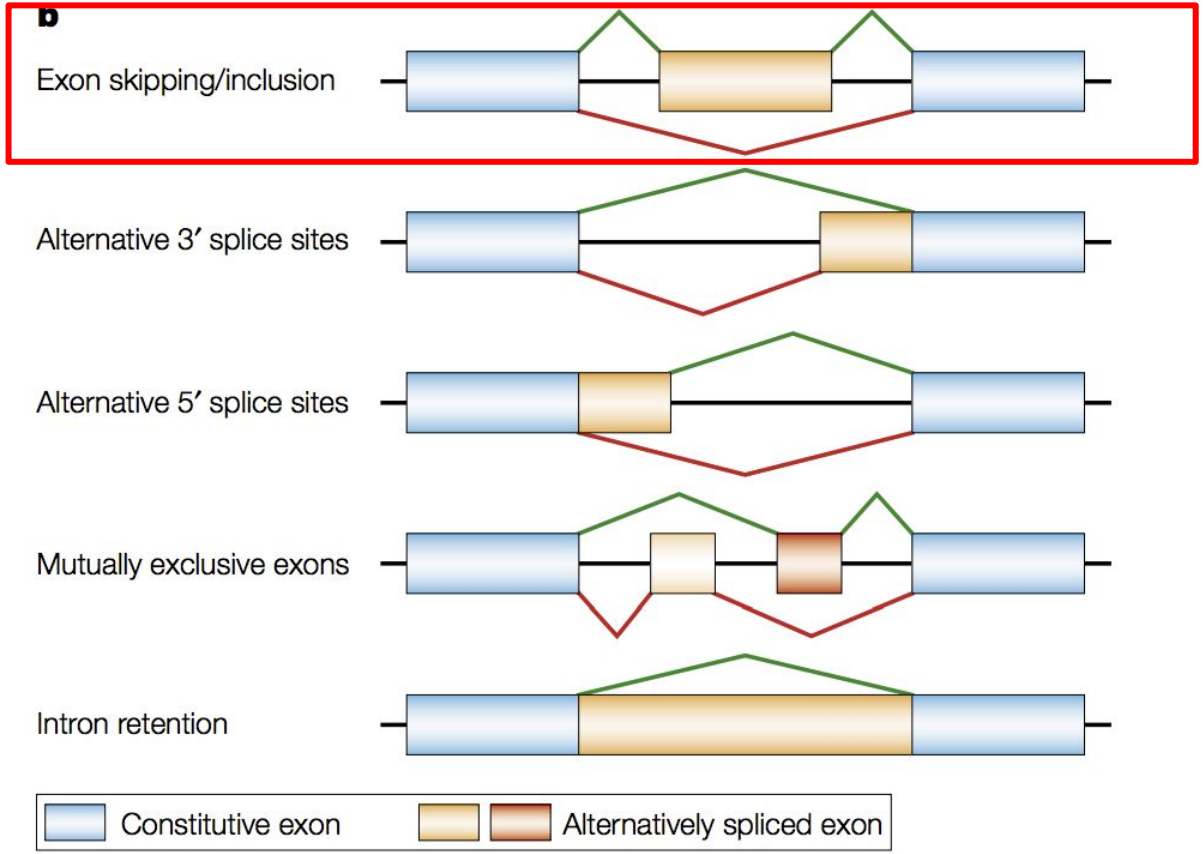


選擇性剪切 (Alternative Splicing)

- 在真核生物中並非基因的全部都會轉譯 (translate) 成蛋白質
- Exon (外顯子)
  - 最後會轉譯成蛋白的元件
- Intron (內含子)
  - 在轉錄 (transcription) 後的修飾過程中被剪接去除 (mRNA splicing)
- 透過剪切不同的Exon與Intron組合就可以製造出不同的蛋白

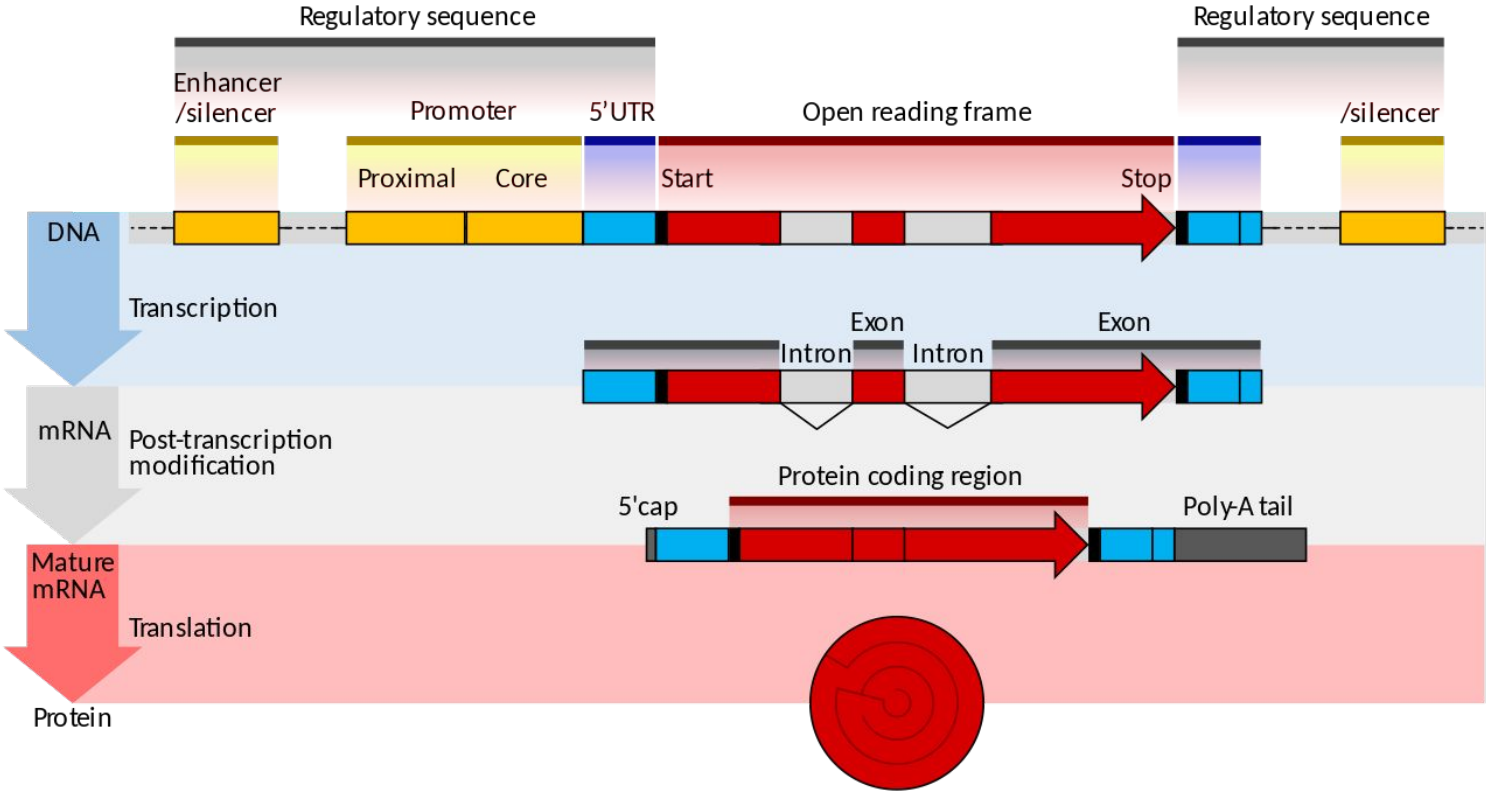

DNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
RNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
Alternative Splicing
mRNA
1 2 3 4 5
1 2 4 5
1 2 3 5
Translation
Translation
Translation
Protein A
Protein B
Protein C

Modes of Alternative Splicing



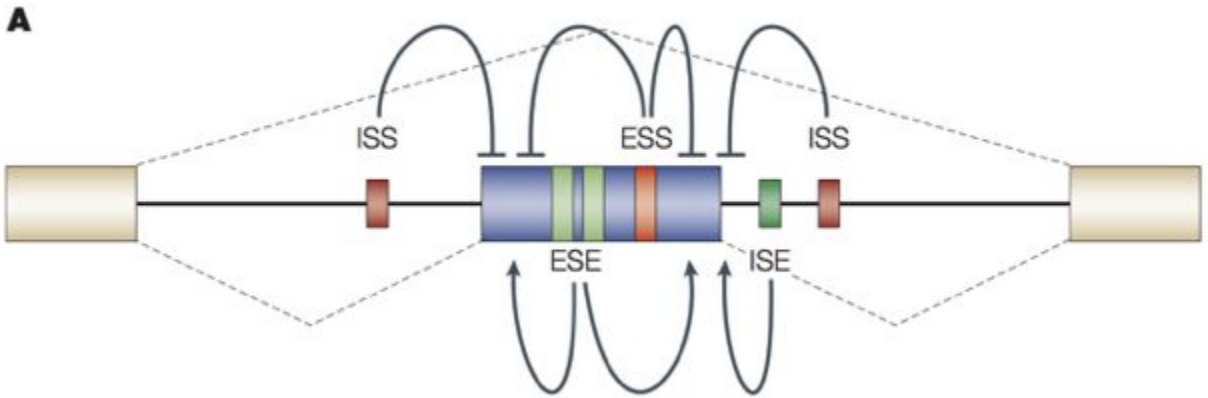
Cartegni, L., Chew, S. L., & Krainer, A. R. (2002). *Nature Reviews Genetics*, 3(4) 285.

Gene structure



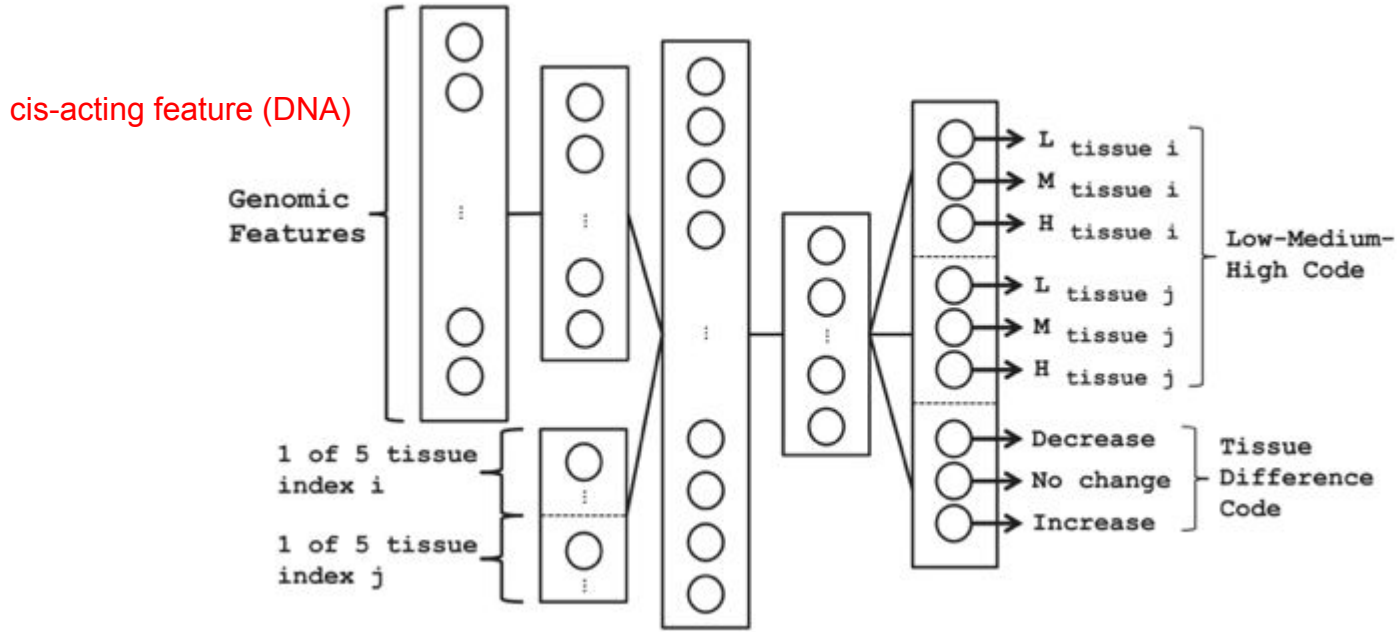
Regulation of Alternative Splicing

- *cis*-acting elements
  - Exon splicing enhancer (ESE)
  - Exon splicing silencer (ESS)
  - Intron splicing enhancer (ISE)
  - Intron splicing silencer (ISS)
- *trans*-acting elements
  - Alternative Splicing Factor
  - Transcription Factor



Matlin, A. J., Clark, F., & Smith, C. W. (2005). *Nature reviews. Molecular cell biology*, 6(5), 386.

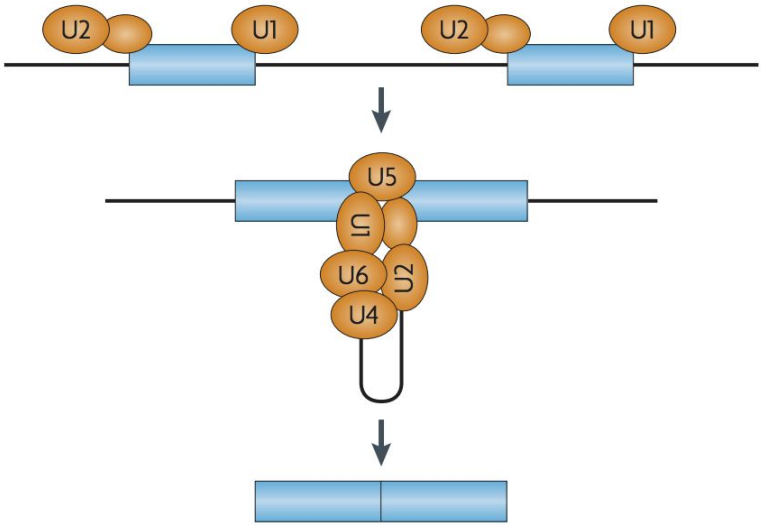
深度學習預測選擇性剪切



Leung, M. K., et al. (2014). *Bioinformatics*, *30*(12), i121-i129.

Regulation of Alternative Splicing

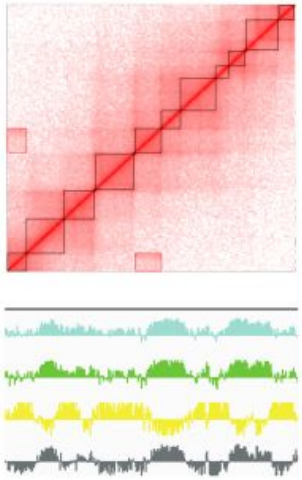
- *cis*-acting elements
  - Exon splicing enhancer (ESE)
  - Exon splicing silencer (ESS)
  - Intron splicing enhancer (ISE)
  - Intron splicing silencer (ISS)
- *trans*-acting elements
  - Alternative Splicing Factor
  - Transcription Factor

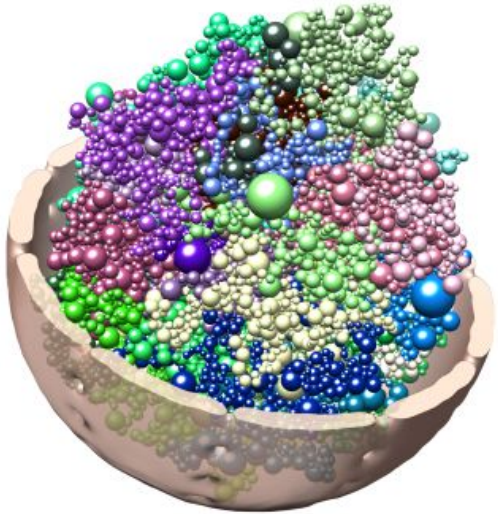


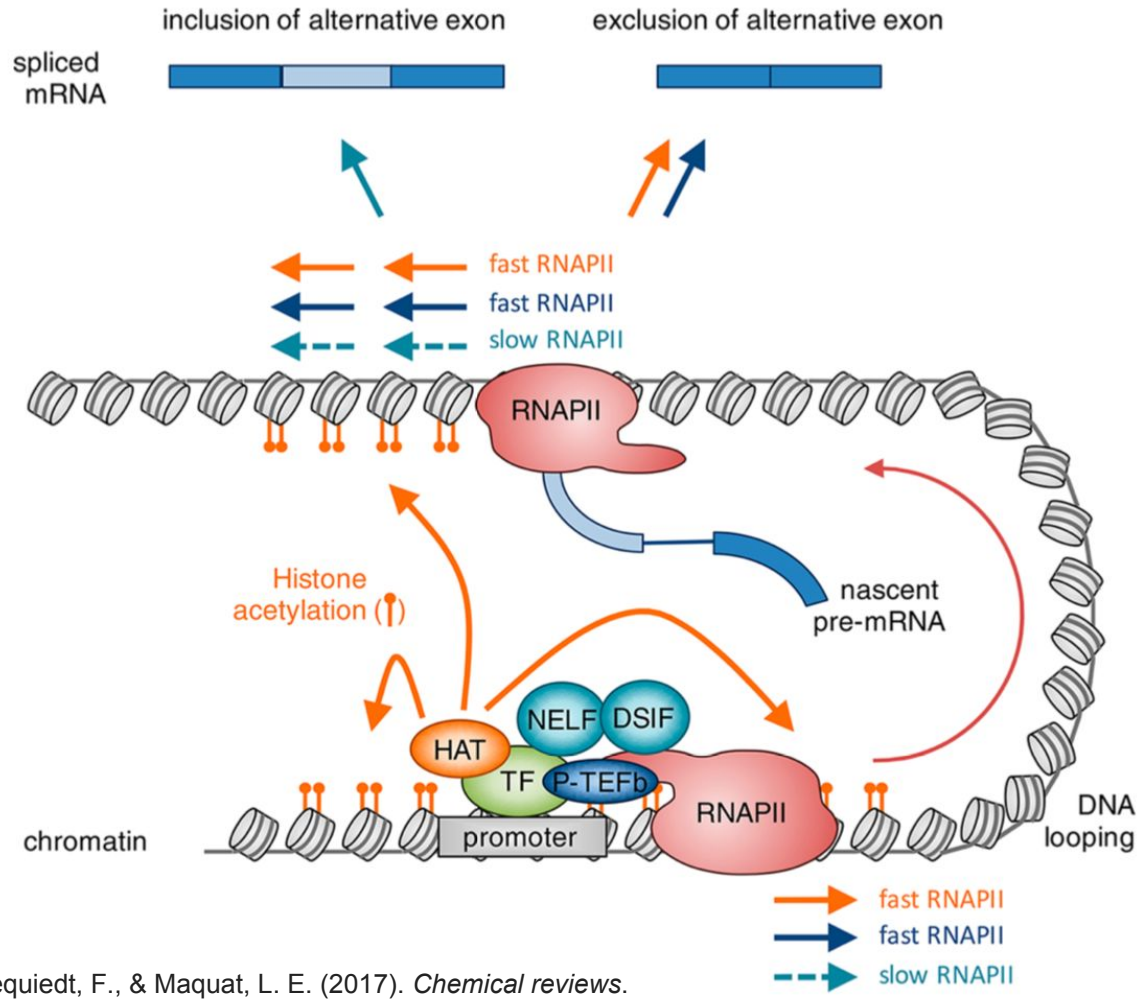
Keren, H., Lev-Maor, G., & Ast, G. (2010). *Nature Reviews Genetics*, 11(5), 345.

Distal DNA Element can influence AS

- Chromatin is organized in a 3D struction
- Distal DNA element may have tightly connection
- Hi-C Sequencing reveal complex 3D chromatin structure in nucleolus

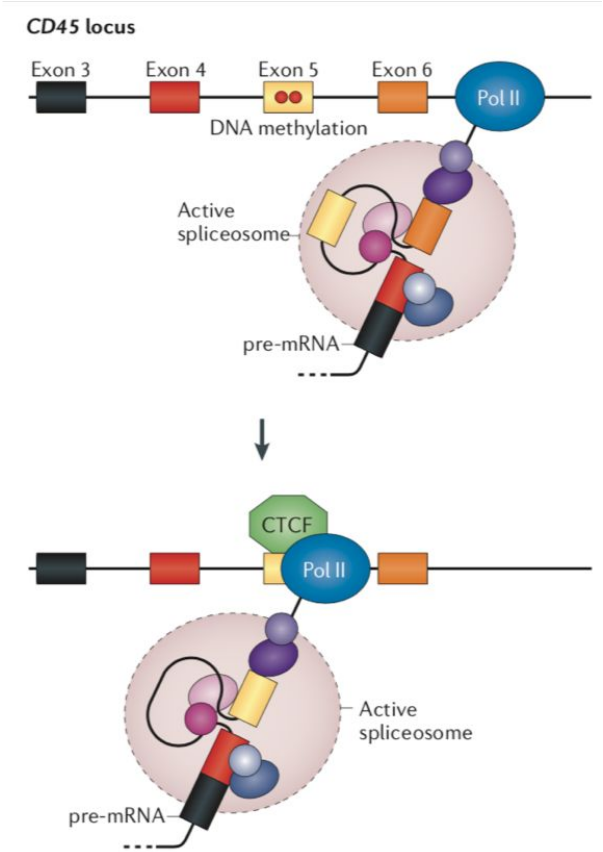






Rambout, X., Dequiedt, F., & Maquat, L. E. (2017). *Chemical reviews*.

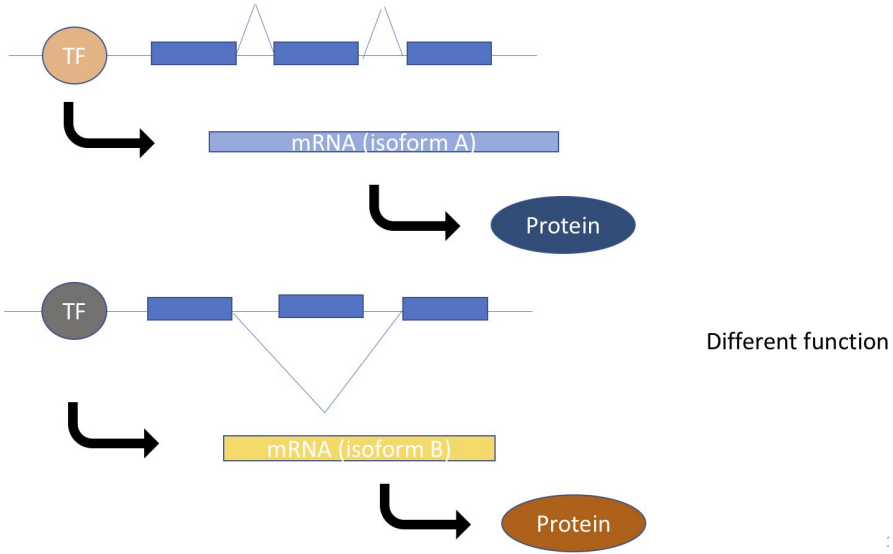
CTCF Promote Alternative mRNA



Ong, C. T., & Corces, V. G. (2014). *Nature reviews. Genetics*, 15(4), 234.

目標

- Gene expression is regulated by TF. How about alternative splicing?



資料來源

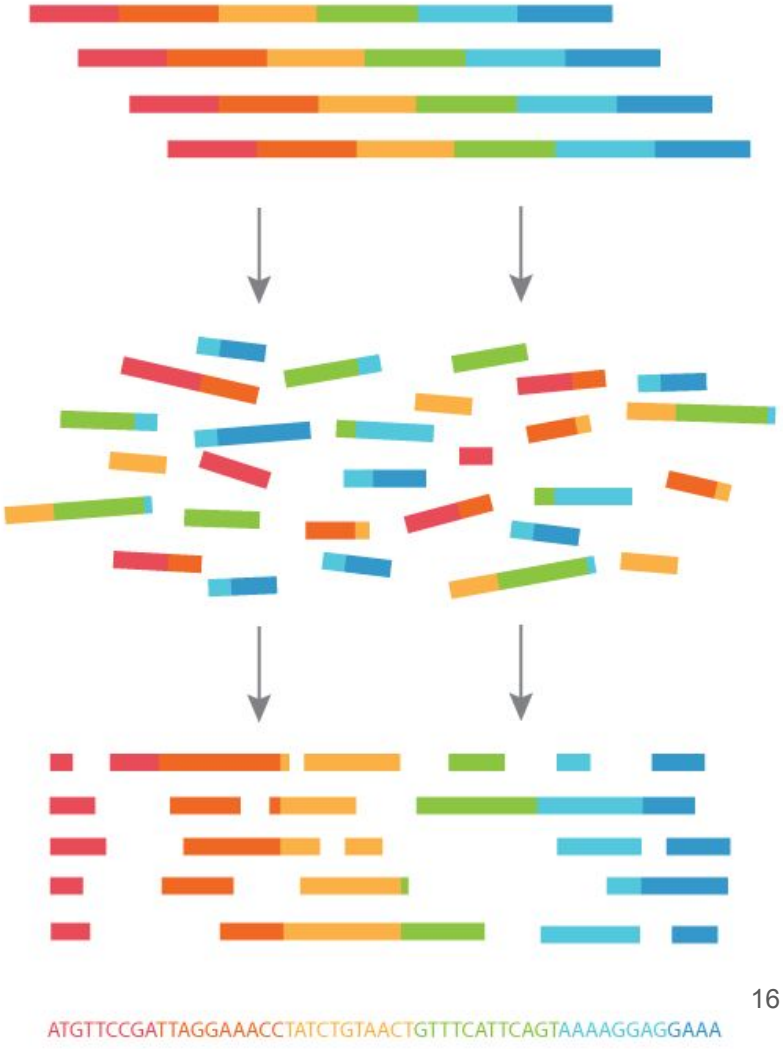
- Drosophila melanogaster (果蠅)
- Genomic data
  - Ensembl BDGP6
  - https://www.ensembl.org/Drosophila_melanogaster/Info/Index
- Annotation
  - BDGP6.90
- FAIRE-Seq & RNA-Seq from NCBI GEO Database
  - GES38727, GSE40739 ,GSE62558
  - https://www.ncbi.nlm.nih.gov/geo/

Next Generation Sequencing

- Procs
  - 高通量 (High-Through)
  - 單位成本低廉
- Concs
  - 短讀長 (short reads)
  - 相對低的準確率
- 需要有效率演算法來處理數據

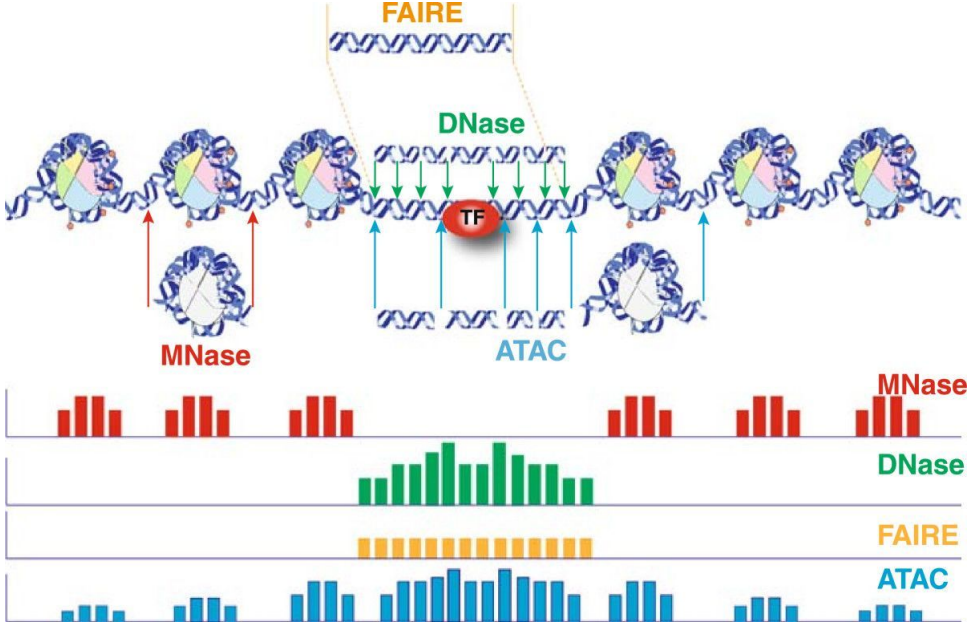
illumina

Roche 454 SEQUENCING



FAIRE-Seq

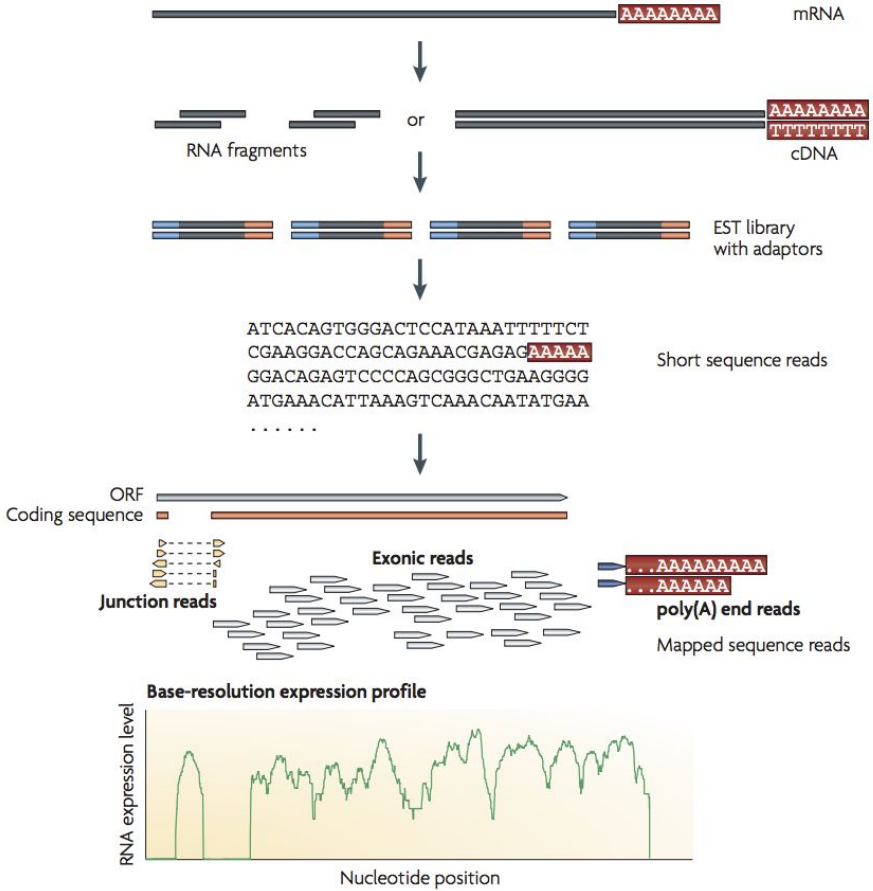
- 與轉錄因子結合的DNA區域，染色質會呈現打開的狀態 (chromatin-accessible)
- 利用formaldehyde固定具有Protein-DNA interaction的區域
- 分離chromatin-accessible的DNA



Tsompana, M., & Buck, M. J. (2014). *Epigenetics & chromatin*, 7(1), 33.

RNA-Seq

- 利用定序來定量以及定性基因的表現量
- 直接定序不同的Isoform
- 可用來研究選擇性剪切

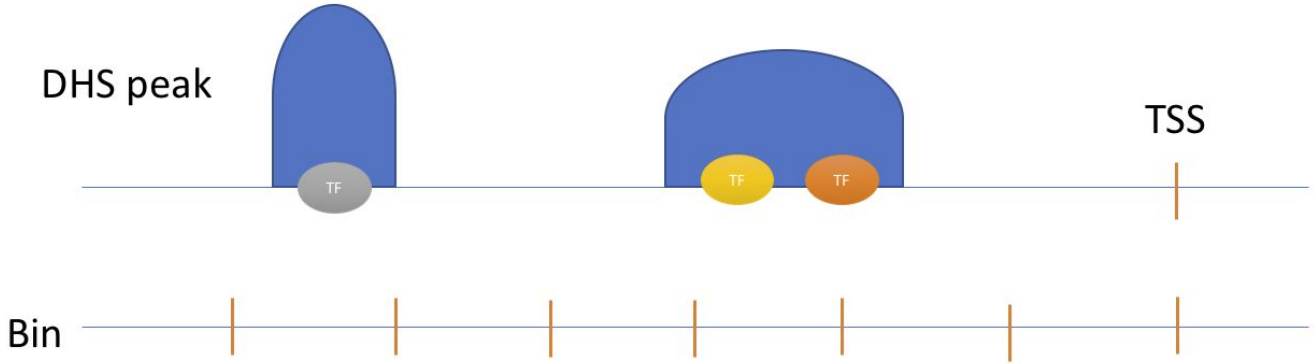


Wang, Z., Gerstein, M., & Snyder, M. (2009). *Nature reviews genetics*, 10(1), 57.

資料前處理

- RAW NGS data
  - Trimmomaitc (QC)
- RNA-Seq
  - HISAT2 (Short Reads Mapping)
  - CATANA (AS events from annotation)
  - MISO (Quantify AS events)
- FAIRE-Seq
  - Bowtie2 (Short Reads Mapping)
  - MACS2 (Peak Calling)
  - CIS-BP (Motif Database)
  - RAST: Matrix-SCAN (Find TF Binding sites in Peak)

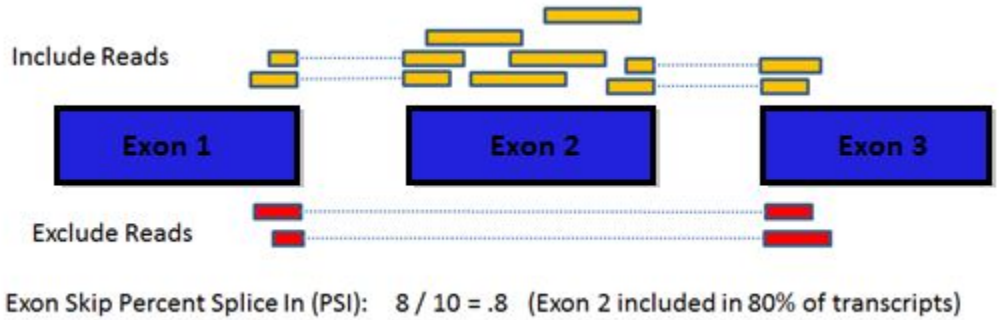
Feature extraction

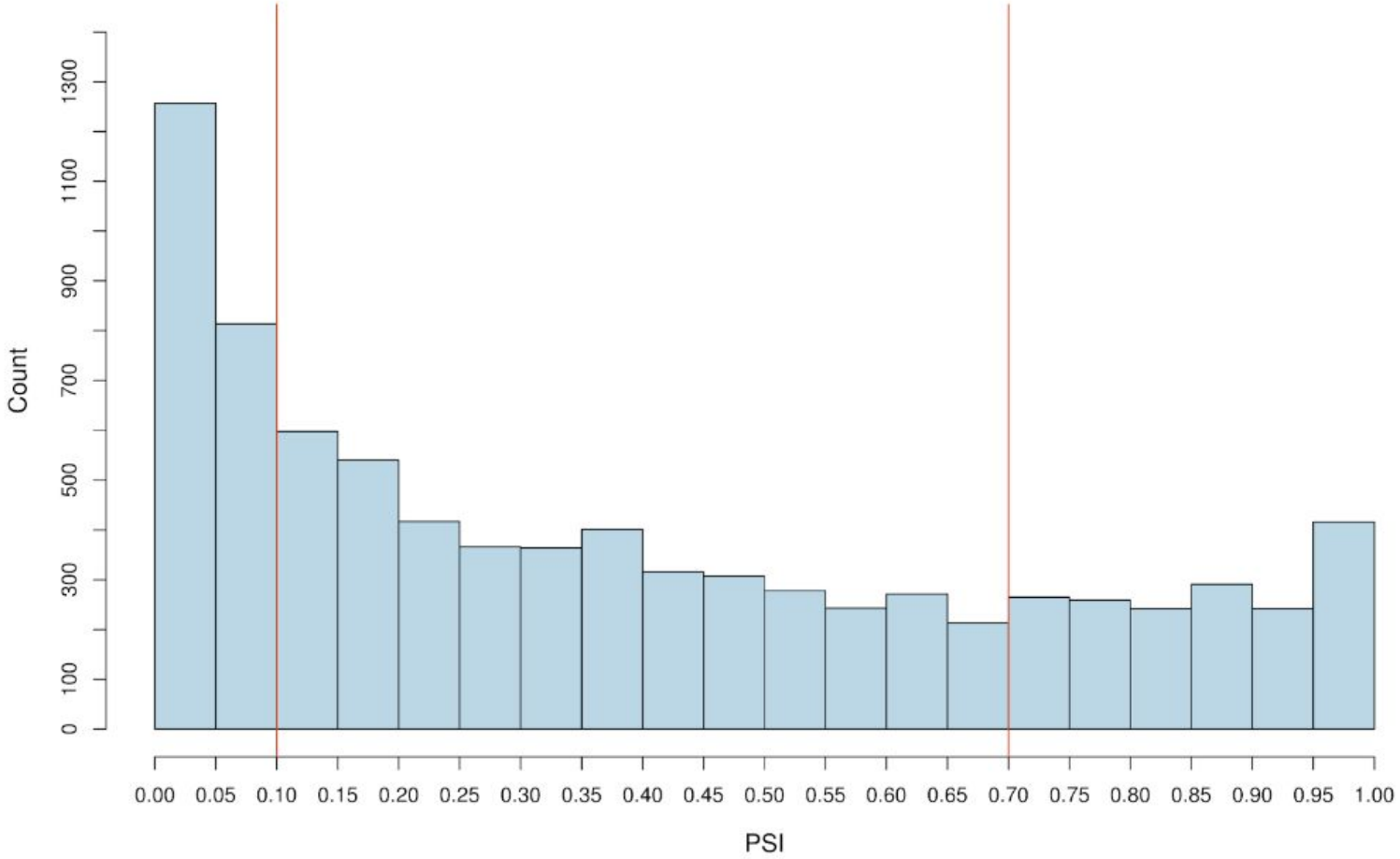


TF1	0	1	0	0	0	0	0	0
TF2	0	0	0	0	1	0	0	0
TF3	0	0	0	0	1	1	0	0
TF4	0	0	0	0	0	0	0	0

Targets

- 只取一個基因的第一個Exon Skipping Event
- 利用MISO計算PSI值
- 依PSI值分成H/L兩類
  - H > 0.7, L < 0.1



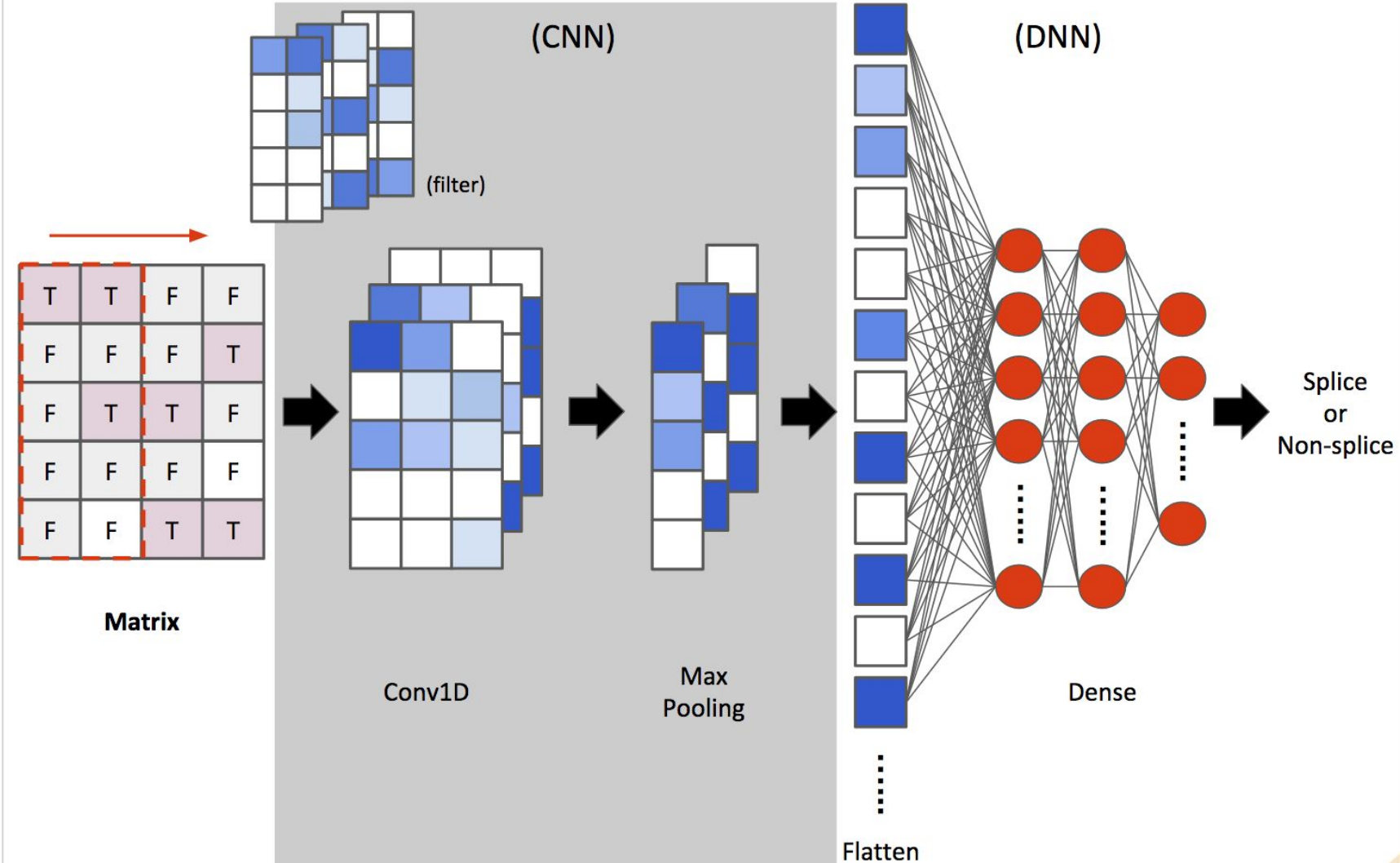


Spliced: 2071

Retained: 1715

Deep Learning Model

- Activation Function: ELU with Batch Normalization
- CNN
  - Conv1D -> Max Pooling -> Flatten -> FC1 -> FC2 -> Output (Softmax)
  - Dropout (rate): 0.5
- DNN
  - Flatten -> FC1 -> FC2 -> FC3 -> FC4 -> Output (Softmax)
  - Dropout (rate)
    - FC1,2: 0.5
    - FC3,4: 0.2

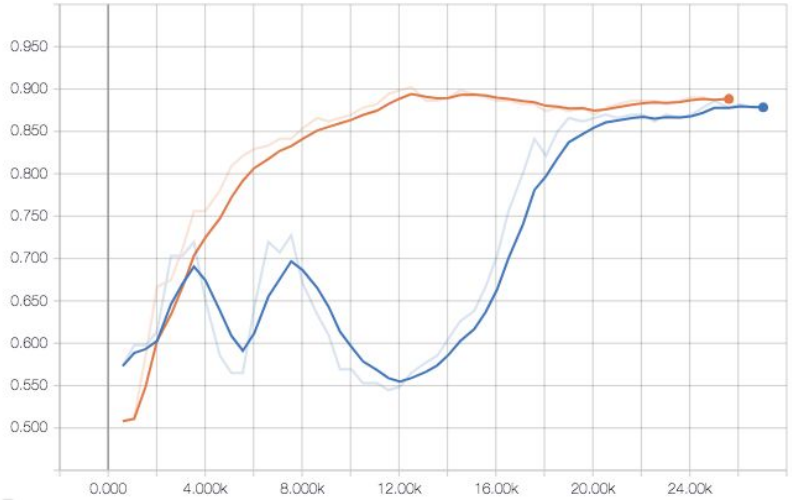

(CNN)
(DNN)
(filter)
T T F F
F F F T
F T T F
F F F F
F F T T
Matrix
Conv1D
Max
Pooling
Dense
Flatten
Splice
or
Non-splice

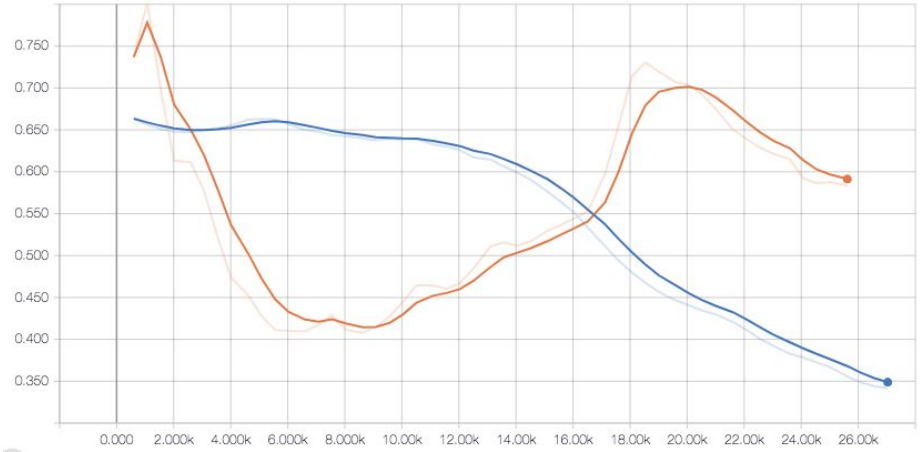
Deep Learning Model

- Mini Batch training
  - Batch size: 128
- Initial learning rate: 0.001
- L2 regularization
  - Wight decay
    - CNN: 0.01
    - DNN: 0.001
- Optimizer: Adam
- Train/Validation/Test: 2832/246/708

Training

- Orange: DNN, Blue: CNN



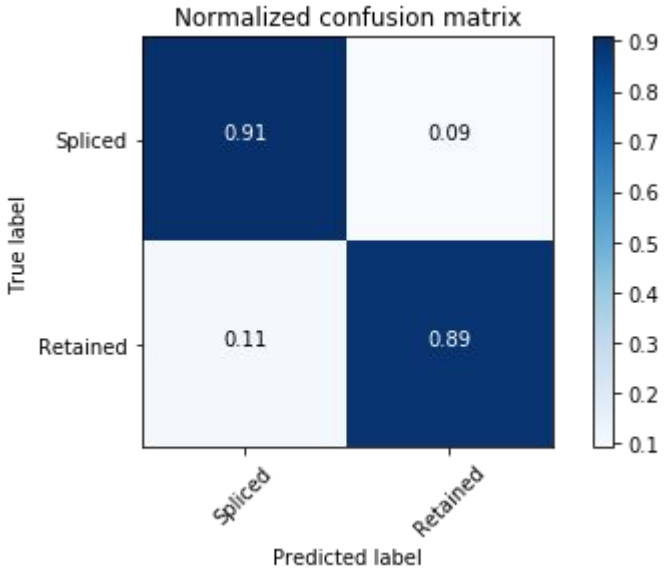


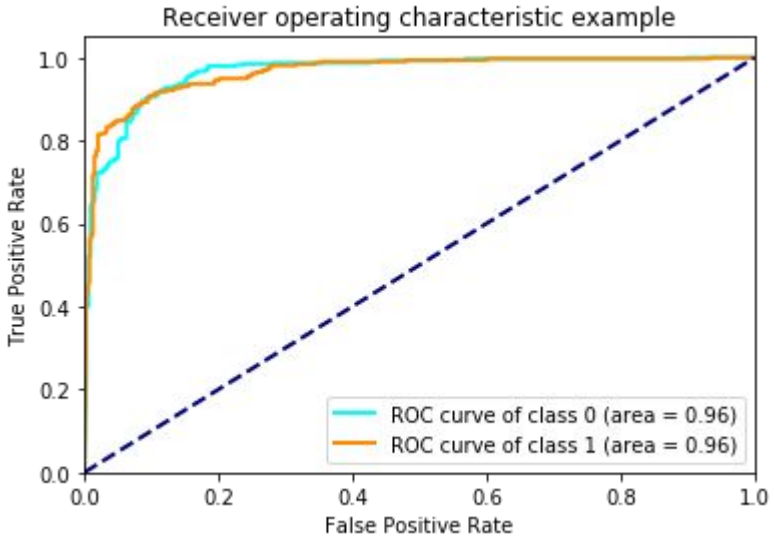
10-fold Cross Validation

- CNN
  - Loss: 0.310 +- 0.032
  - Accuracy: 87.13 +- 1.59%
- DNN
  - Loss: 0.631 +- 0.104
  - Accuracy: 90.16 +- 0.96%

模型評估

- 從PSI轉換成H/L類別 (Spliced/Retained) 時沒有平衡兩個類別
- 所以分類的混淆矩陣兩個類別稍有不平衡





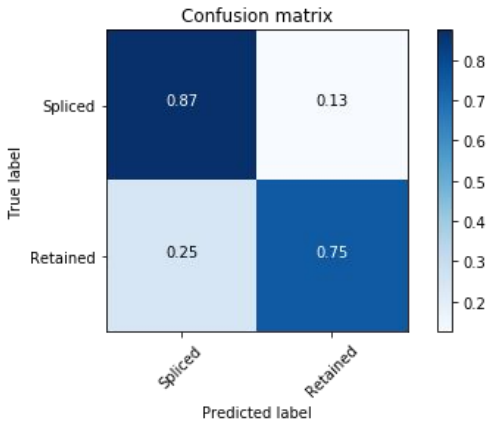
XGBoost

- Deep Learning很強大，但是不能夠解釋究竟那個Feature (TF)是影響選擇性剪切的關鍵
- XGBoost
  - A Scalable Tree Boosting System
  - 實際上除了Boosting，XGBoost同時也做了Bagging
  - 同時利用一階導數及二階導數來優化權重
  - 可以利用Information Gain來找出Feature Importance

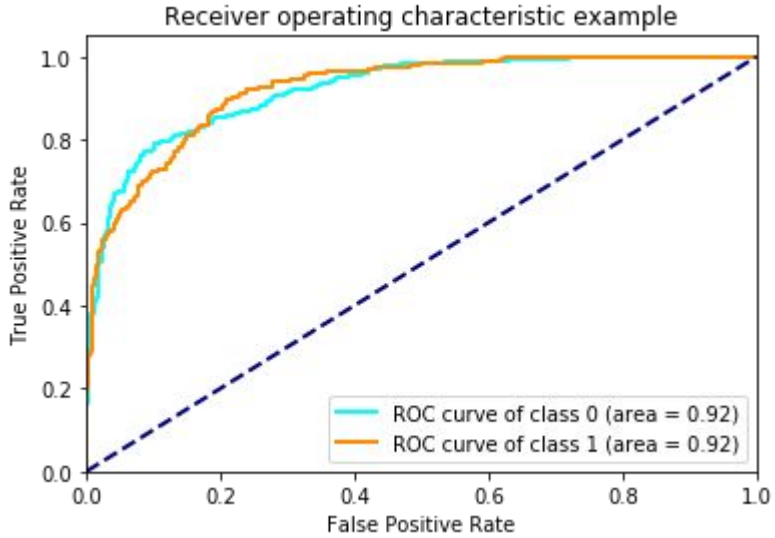
XGBoost

- N_estimators: 500
- Learning_rate: 0.1
- Max_deepth: sqrt(n_TF)=√307
- Train/Test split: 0.8/0.2
- Performance
  - Accuracy: 81.66%
  - AUROC: 0.92

模型評估



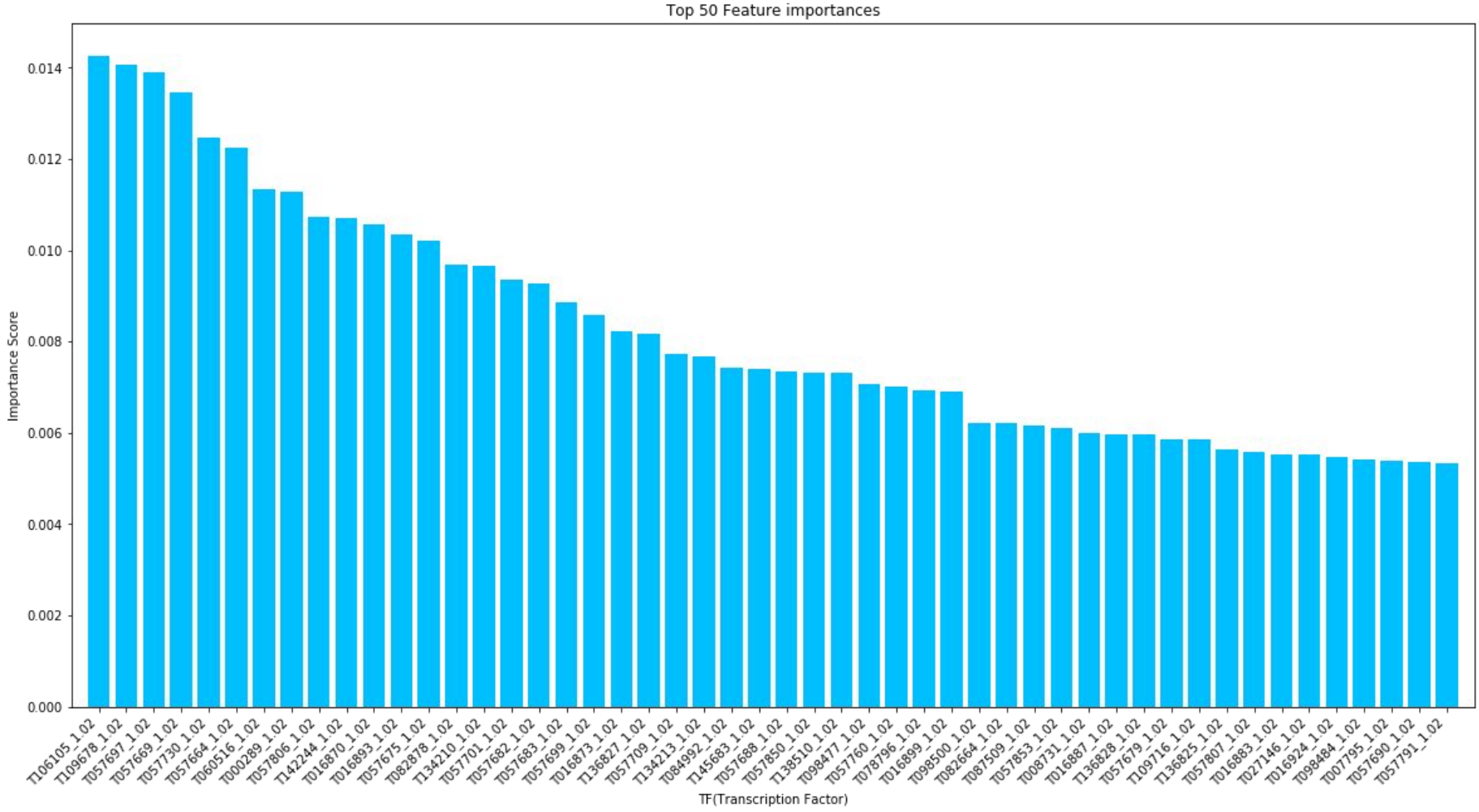
稍微Unbalance



5-fold Cross-Validation

- Accuracy: 83.44 +- 0.93%
- AUC: 0.924 +- 0.009

- 因爲TF在每個Fold中的Feature Importance有時變化很大
  - 利用5個Fold的Rank做平均
  - 避免訓練時Stochastic的影響


Top 50 Feature importances
Importance Score
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
T106105_1.02
T109678_1.02
T057697_1.02
T057669_1.02
T057730_1.02
T057664_1.02
T060516_1.02
T000289_1.02
T057806_1.02
T142244_1.02
T016870_1.02
T016893_1.02
T057675_1.02
T082878_1.02
T134210_1.02
T057701_1.02
T057682_1.02
T057683_1.02
T057699_1.02
T016873_1.02
T136827_1.02
T057709_1.02
T134213_1.02
T084992_1.02
T145683_1.02
T057688_1.02
T057850_1.02
T138510_1.02
T098477_1.02
T057760_1.02
T078796_1.02
T016899_1.02
T098500_1.02
T082664_1.02
T087509_1.02
T057853_1.02
T008731_1.02
T016887_1.02
T136828_1.02
T057679_1.02
T109716_1.02
T136825_1.02
T057807_1.02
T016883_1.02
T027146_1.02
T016924_1.02
T098484_1.02
T007795_1.02
T057690_1.02
T057791_1.02
TF(Transcription Factor)

Top-10 TF

1. T106105_1.02 (0.014262) (Hsf)
2. T109678_1.02 (0.014078) (Adf1)
3. T057697_1.02 (0.013909) (lola)
4. T057669_1.02 (0.013463) (hb)
5. T057730_1.02 (0.012457) (jim)
6. T057664_1.02 (0.012262) (Cf2)
7. T060516_1.02 (0.011338) (rn)
8. T000289_1.02 (0.011273) (br)
9. T057806_1.02 (0.010726) (CTCF, CCCTC-binding factor)
10. T142244_1.02 (0.010704) (Mad)

Previous study

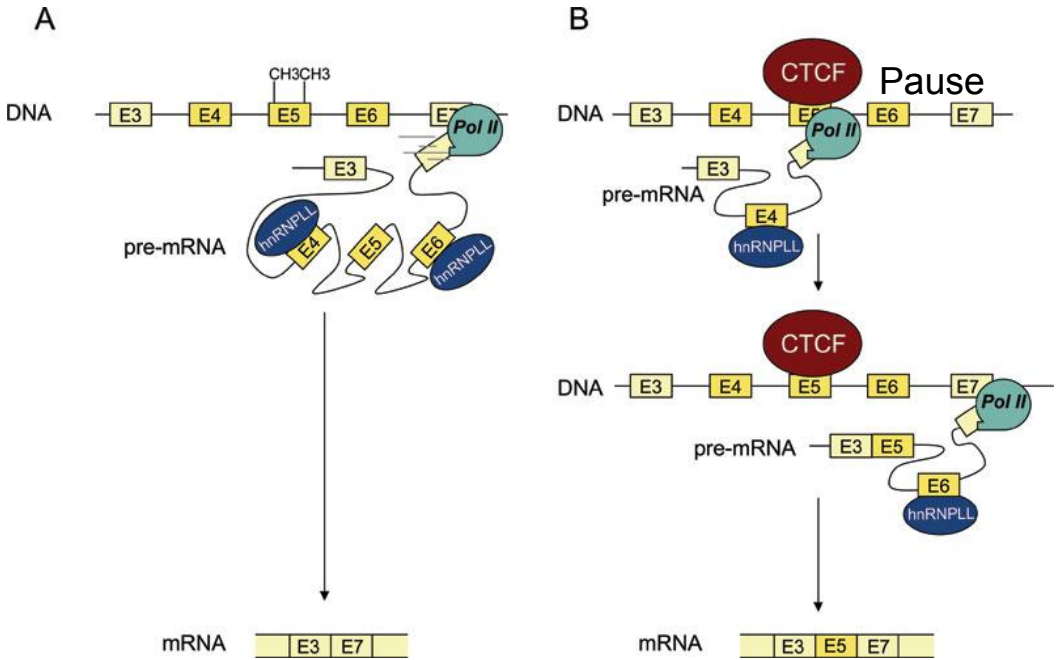
ARTICLE

doi:10.1038/nature10442

CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing

Sanjeev Shukla[1], Ersen Kavak[2,3], Melissa Gregory[1], Masahiko Imashimizu[4], Bojan Shutinoski[1], Mikhail Kashlev[4], Philipp Oberdoerffer[1], Rickard Sandberg[2,3] & Shalini Oberdoerffer[1]

Mammalian-model (哺乳類)



Shukla S et al. (2011) Nature. 479:74–79

Kornblihtt AR (2012) Cell Res. 22(3):450-2

結論

- 透過深度學習以及XGBoost發現轉錄因子與選擇性剪切存在潛在的Pattern
- 轉錄因子極有可能是調控選擇性剪切的因子之一
  - CTCF已經被證實能夠調控Exon Skipping (哺乳類)
  - CTCF在果蠅都發現可能存在調控 AS功能
  - TF調控AS的機制具有廣泛性, 因此可能在很早期的真核生物就已經存在 (待更多證據)
- 沒有少數重要的TF, 所以Combiorial Effect是需要被考慮的
- 資料探勘協助發現潛在的生物問題與機制

生物問題

產生資料與結果

生物實驗

Data-Driven Biological Research

資料探勘

資料科學家

生物學家

給予生物學家 Inslight

重新分析