Xpore

Problem statement

▷ Central Dogma

ยิ่งมาก reads ก็จะมากตาม

จาก DNA (ทั้งสาย) ดูเป็นช่วงๆ (ยีนส์) —Transcription→ mRNA —Translation→ Protein

▷ RNA Sequencing

สมัยก่อน

RNA

↓ Fragmentation

ตัดสาย RNA

↓ convert to cDNA

ไม่สามารถ Detect ลำดับเบสได้โดยตรง

Nanopore ขนาดเล็กสุด คือ Min ION

ปัจจุบัน (ใช้ Nanopore)

สามารถทำ direct RNA sequencing ได้

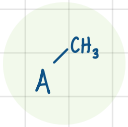
เช็ค RNA modification

หลักการทำงาน

โมเลกุลแต่ละตัวจะทำให้เกิดความต้านทานแตกต่างกัน และใช้ ML ในการอ่านสัญญาณไฟฟ้า เป็นลำดับเบส (Base calling)

▷ RNA modifications

เช่น m^6A



Sample

KO - Knock out : เป็น sample ที่ knock out ยีนส์ที่ผลิต m^6A ออก

WT - Wild type : ปกติ

เปรียบเทียบ KO-WT เพื่อหา modification rate

▷ Research Objectives

<KO> GGACT

<WT> GGm^6ACT

Nanopore Sequencing →

หาตำแหน่ง mod RNA

หาปริมาณการเกิด mod (%) และอัตราการเกิดของ Reads → หา differential

สูง (Sig)

ทดลองต่อเพื่อหาความเกี่ยวข้อง

- Signal-level modification detection methods

EpiNano / MINES (Software)

- สามารถ detect ตำแหน่ง m^6A ได้
- จำเป็นต้องมี data ในการ train (Supervised)

Tombo

- สามารถ detect ตำแหน่งของ mod ได้ทุกประเภท
- ไม่ต้องใช้ data ในการ train (Unsupervised)

Xpore อยู่ประเภทนี้

① Data collection and Preparation (จาก Xpore)

FAST5 file คือไฟล์ สัญญาณไฟฟ้า (pA) ถูกเก็บไว้ในรูปแบบ HDF5 (binary)

FASTQ file คือไฟล์ ลำดับเบสในรูปแบบ text

FASTA file คือไฟล์ ลำดับเบส cdna (rna → dna)

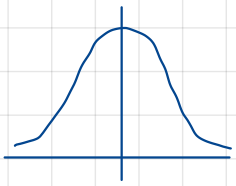
BAM/SAM คือไฟล์ที่จัดเรียง FASTQ และ FASTA
Binary text

Nanopore preprocessing

Direct RNA sequencing (จาก Oxford Nanopore) → Basecalling (FAST5 → FASTQ) ใช้ Guppy → Sequence aligning (FASTQ+FASTA → BAM) ใช้ Minimap2 → Signal event aligning (FAST5 + FASTA + BAM → Event align Op.) ใช้ Nanopolish

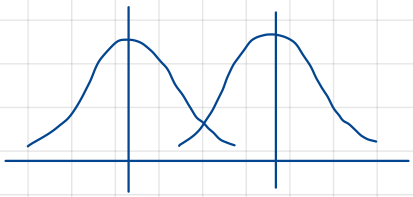
② Bayasian [Multi-Sample] Gaussian Mixture Modelling

Gaussian (normal distribution) => รู้ assumptions ว่า data ถูกสร้างอย่างไร จึงสามารถสร้าง data ขึ้นได้ (Generative AI ตัวแรกๆ)



$$p(x) = \sum_{k=1}^{k} \pi_k N(x|\mu_k, \Sigma_k)$$

2 Gaussian Mixture



มี 2 distribution

Frequentis
Bayasian
} มา parameter

Multi-Sample - เช่น rna_1, rna_2, rna_3 มา train

ในงานวิจัยนี้ใช้ GMM เพราะ - มี 2 distribution เท่านั้น คือ 1) modified (m^6A)
2) unmodified (A)
- รู้ prior จาก nanopolish

③ Evaluation

Cell line - เซลล์ที่เลี้ยงให้เติบโตและเพิ่มจำนวนได้ต่อเนื่อง

▷ Experiment Setup

HEK293T (จากไต) → Wild type (มี m^6A) / KO(METTL3-/-)

Nanopore ทำ direct RNA seq. => ML ทำนาย motif คือ 5-mer ที่เป็น DRACH

องค์ความรู้เดิม → DRACH (fixed!)

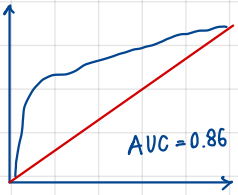
D = {A,G,T}

R = {A,G}

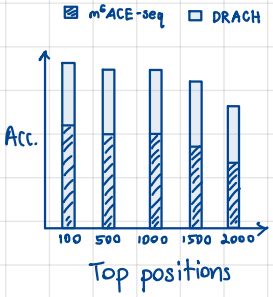
H = {A,C,T}

▷ Validation : m6A calling

ใช้ ROC Curve

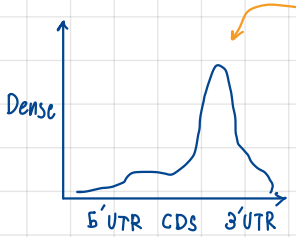


- มีความแม่นยำ 86%



- m^6ACE-Seq + DRACH มี ACC > 95%

จึงจะเกิด m^6A และมักจะเกิดช่วงท้ายๆของ RNA



Confirm องค์ความรู้เดิม!

In Bioinformatics

- fase positive อาจจะไม่ผิดเสมอไป
- จำเป็นต้องทำ Analysis อื่นๆ เพื่อจะได้ Insight มากขึ้น

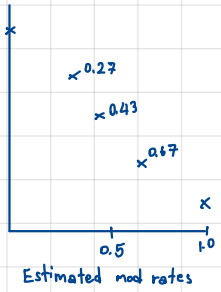
m6A Stoichiometry quantification

RNA Mixture

HEK293-KO + HEK293-WT

0%
25%
50%
75%
100%
(m^6A %)

Modification rates: GGACT



∴ สามารถหา Modification rates ได้

สรุป ML Metrics : ROC curve, Precision-Recall Curve, Accuracy

Analysis : Domain-specific evaluation, Effects of the data size

▷ Applicability : Full Dataset

ทำกระบวนการที่ทำให้คนที่นำไปใช้งาน ใช้ได้ง่ายขึ้น

เช่น เปรียบเทียบหลายๆแบบ (ไต, หัวใจ, ปอด, ฯลฯ) หรือเปรียบเทียบเป็นคนๆ

Keys Takeaway

- Validation - ใช้ metrics ที่เหมาะสม
- Comparison กับของคนอื่น
- Applicability - นำไปใช้ได้จริง