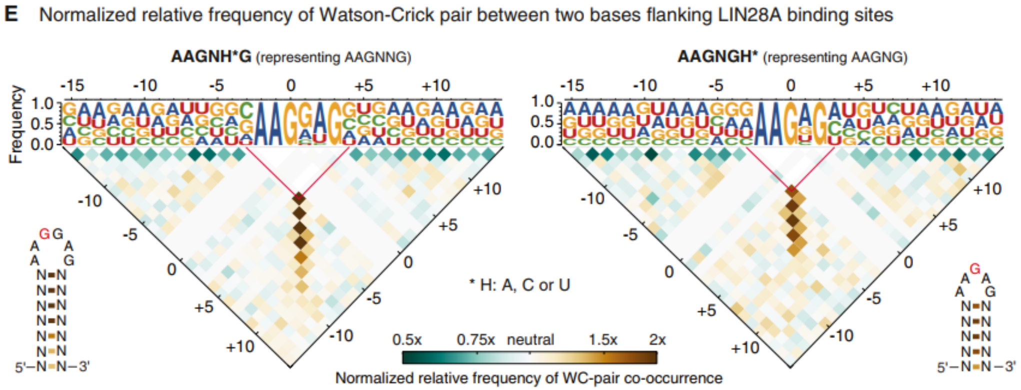
주제: Figure 2E 재현

LIN28A binding site 근처 이차구조 선호도



사용한 데이터

- CLIP-35L33G.bam
- Reference mouse genome (mm39)

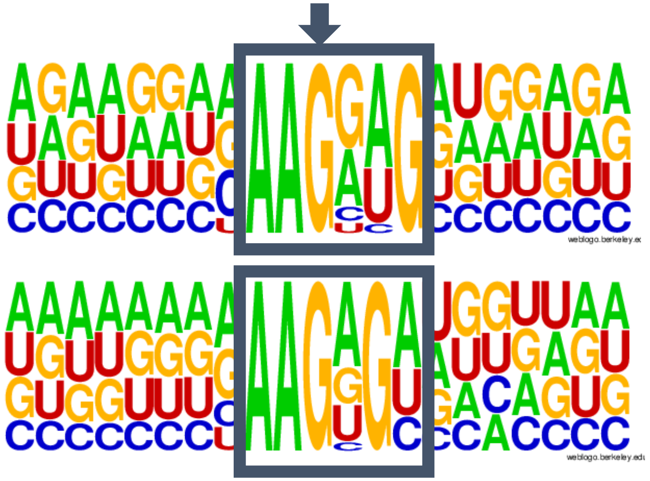
분석과정 (1) LIN28A binding site 찾기

1. **전처리**
   - **Pileup:** CLIP-35L33G.bam → CLIP-35L33G.pileup
   - **Filtering:** chr1~19 & read count > 50
2. **Shannon entropy**
   - $-\sum_n \boldsymbol{p(n)\log_2 p(n)}$ where n is a nt or a del
3. **Binding site**
   - Shannon entropy > 0.8 & read count > 50
   - Strand 구분
   - 예)

chr	pos	ref	strand	entropy	count
chr19	3335353	G	+	1.13	89
chr19	29731408	C	-	2.11	78
chr19	5013598	G	+	2.00	138

분석과정 (2) Binding Site Motif

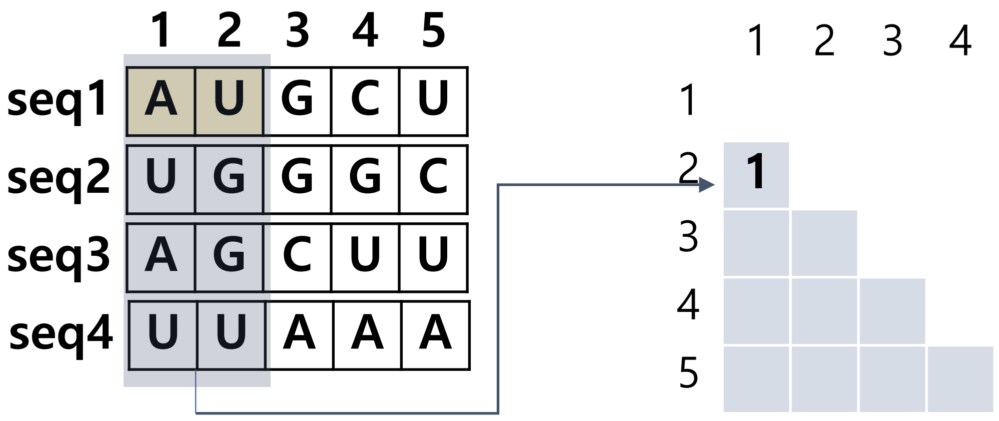
1. **Neighboring sequence 추출**
   - crosslinked bases → centered at zero
   - (-) strand → flip
   - Hexamer (-2 to +3)
2. **Major LIN28A-bound hexamer**
   - AAGNHG / AAGNG(H)
     - ❖ H = A, C or U
   - 21 nt long flanking sequence (-10 ~ +10)
   - WebLogo



분석과정 (3) 이차구조 선호도

<u>Watson-Crick (WC) pair co-occurrence frequency</u>

1. **Observed frequency matrix (21 × 21)**



	1	2	3	4	5
seq1	A	U	G	C	U
seq2	U	G	G	G	C
seq3	A	G	C	U	U
seq4	U	U	A	A	A

	1	2	3	4	5
1					
2	1				
3					
4					
5					

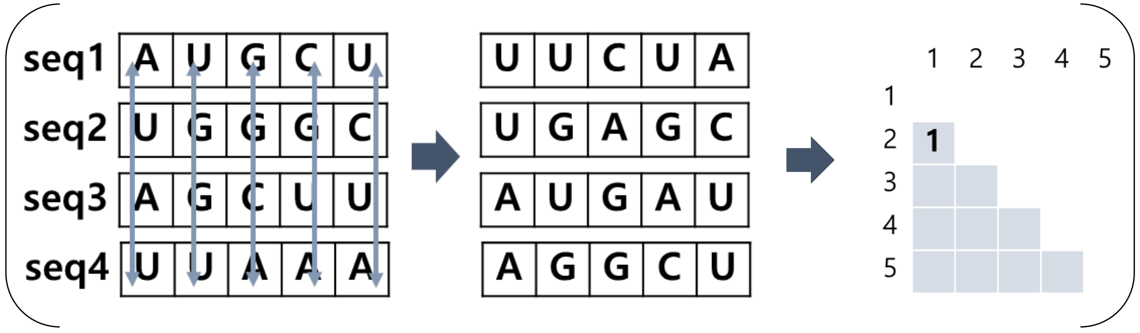
2. **Background by permutation**

분석과정 (3) 이차구조 선호도

2. **Background by permutation**
   - For each iteration

     Shuffle all bases in the same position

     → WC pair co-occurrence freq. matrix
   - ×1,000 times and pooled



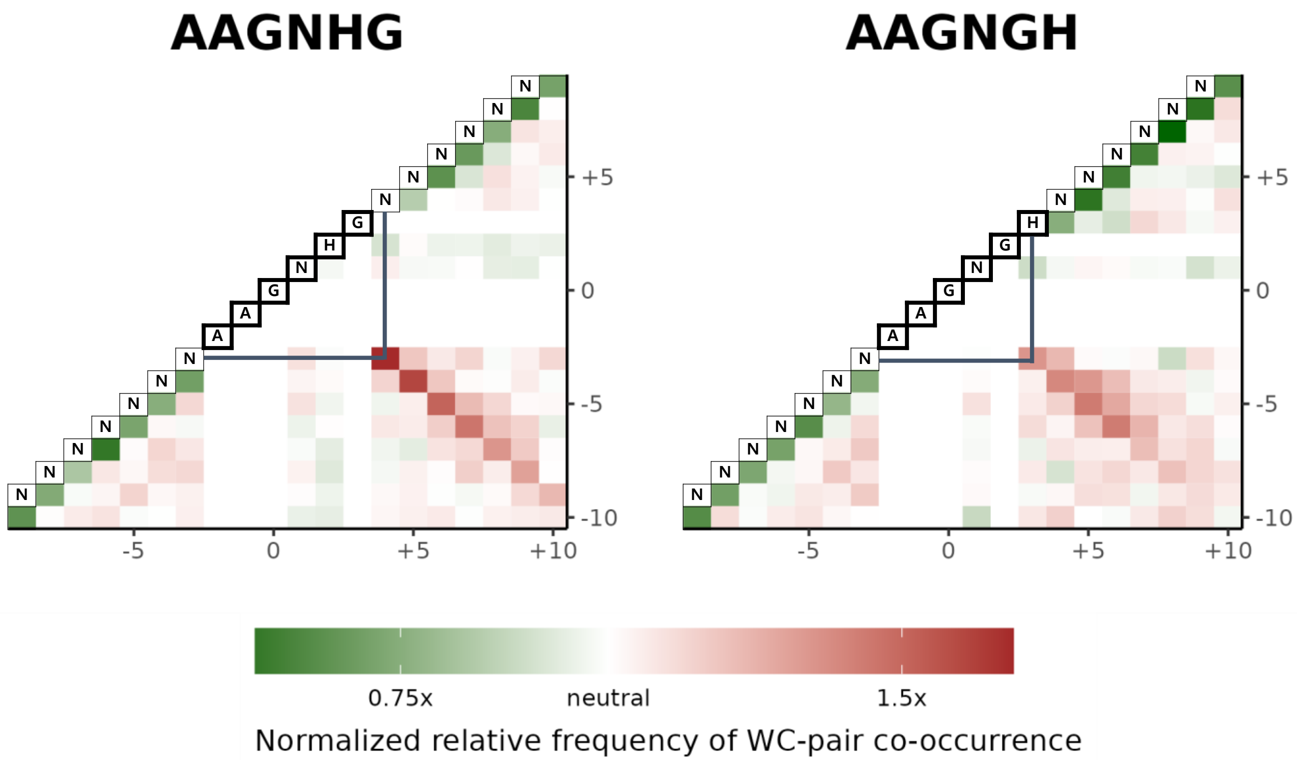
seq1 A U G C U
seq2 U G G G C
seq3 A G C U U
seq4 U U A A A

U U C U A
U G A G C
A U G A U
A G G C U

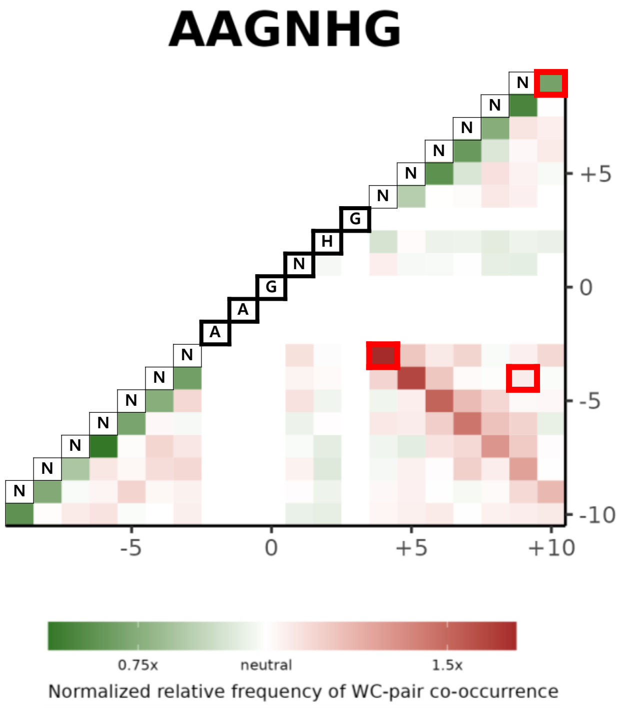
×1,000 times

3. **Enrichment level**
   - $\log_2$ (observed / background)

결과 (1) Figure 2E 재현



결과 (1) Figure 2E 재현

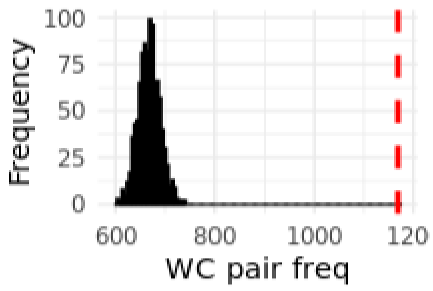
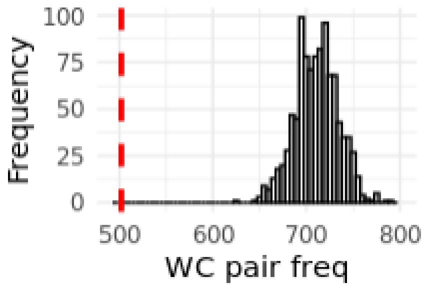
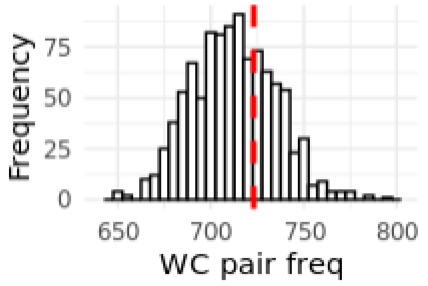


- Enrichment level
  = $\log_2$ (observed / background)

결과 (2) 유의성 검정

1. 가설

- H0: WC pair co-occurrence = background (random)
- H1: WC pair co-occurrence > background

Pair	🟥 (+4, -3)	🟩 (+10, +9)	⬜ (+9, -4)
Dist of Frq (Null)			
Pr(X≥obs)	0	1	0.331

2. Multiple testing

- Pair 별로 test → multiple testing correction 필요