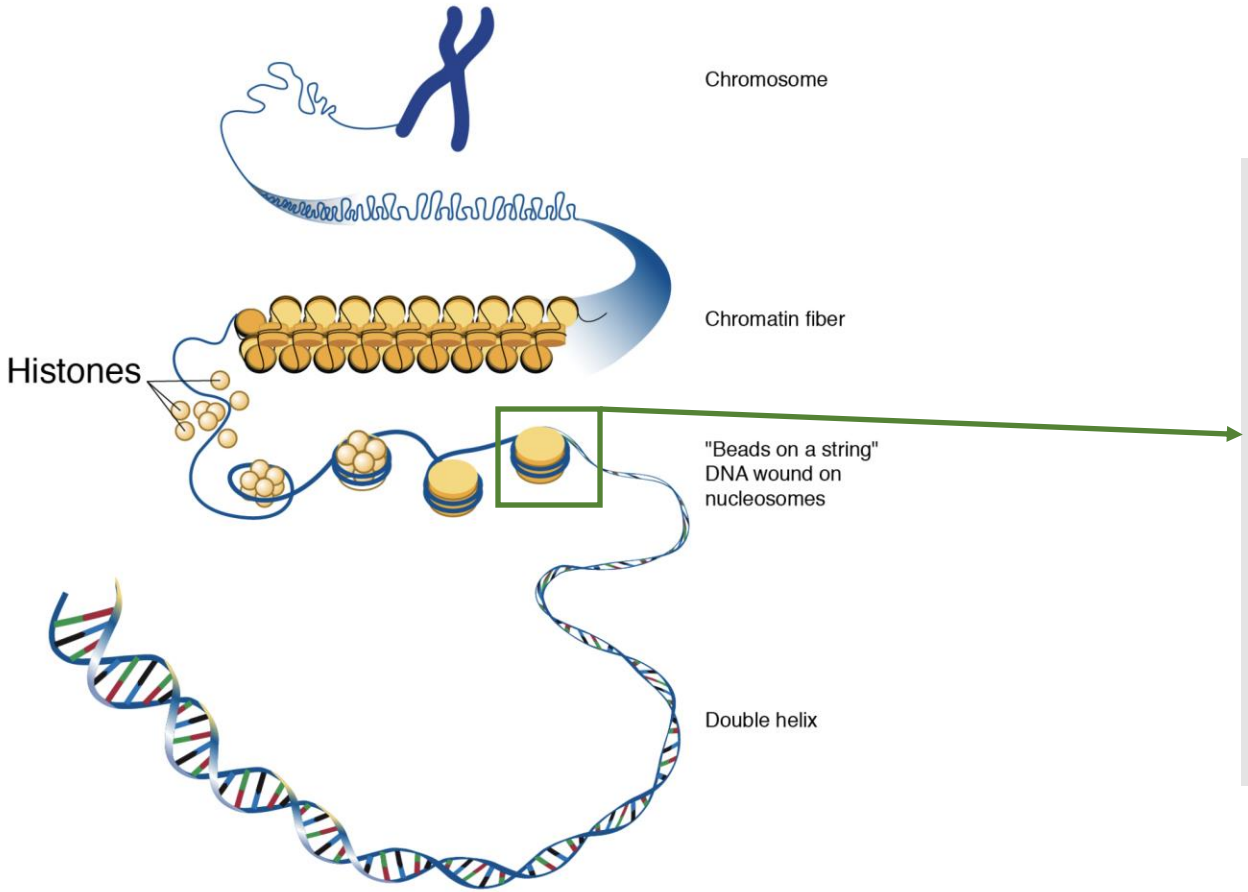
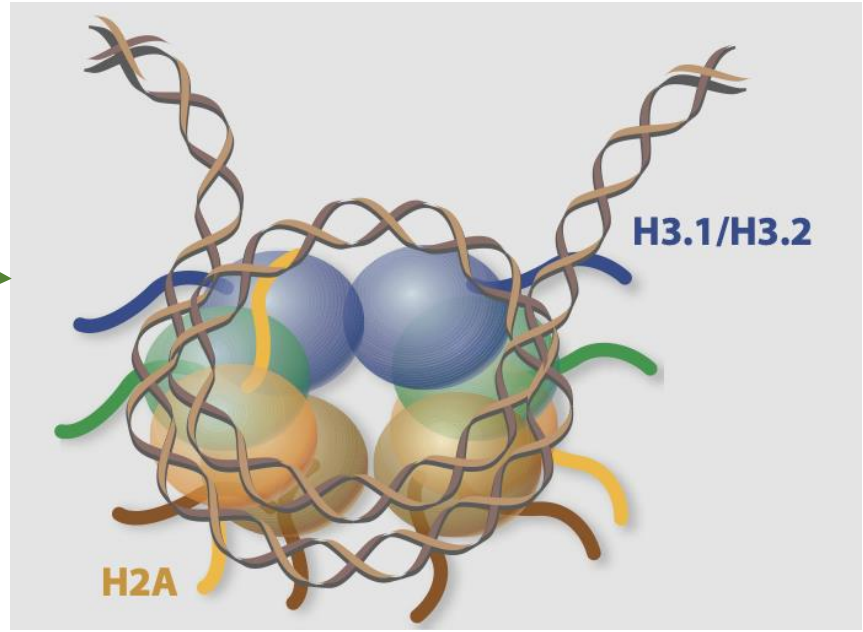
Transcription Factors: What do they do and can we identify potential regulators in our data?

Epigenomics Data Analysis Workshop 2025

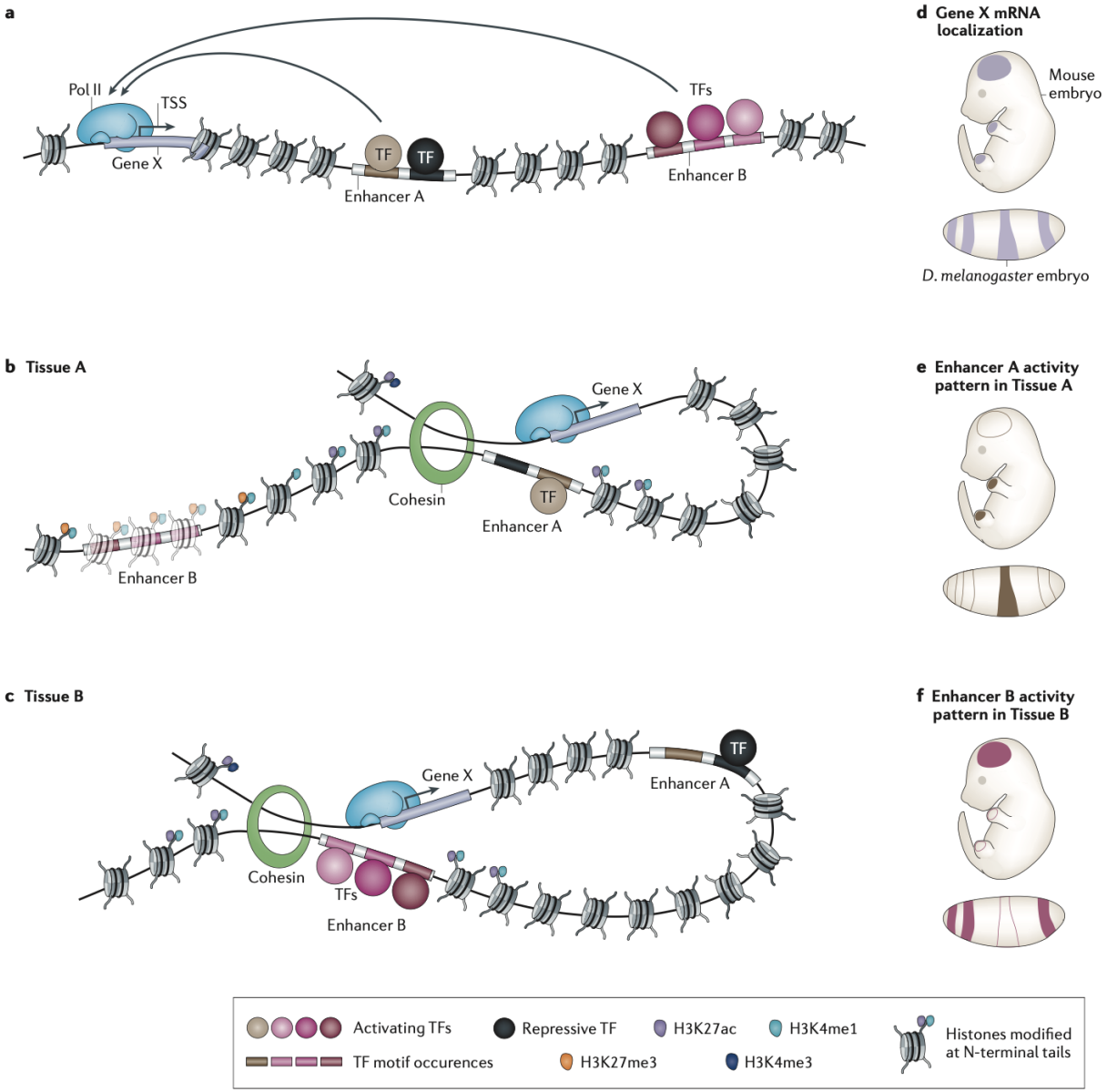
Dania Machlab

Chromatin organization



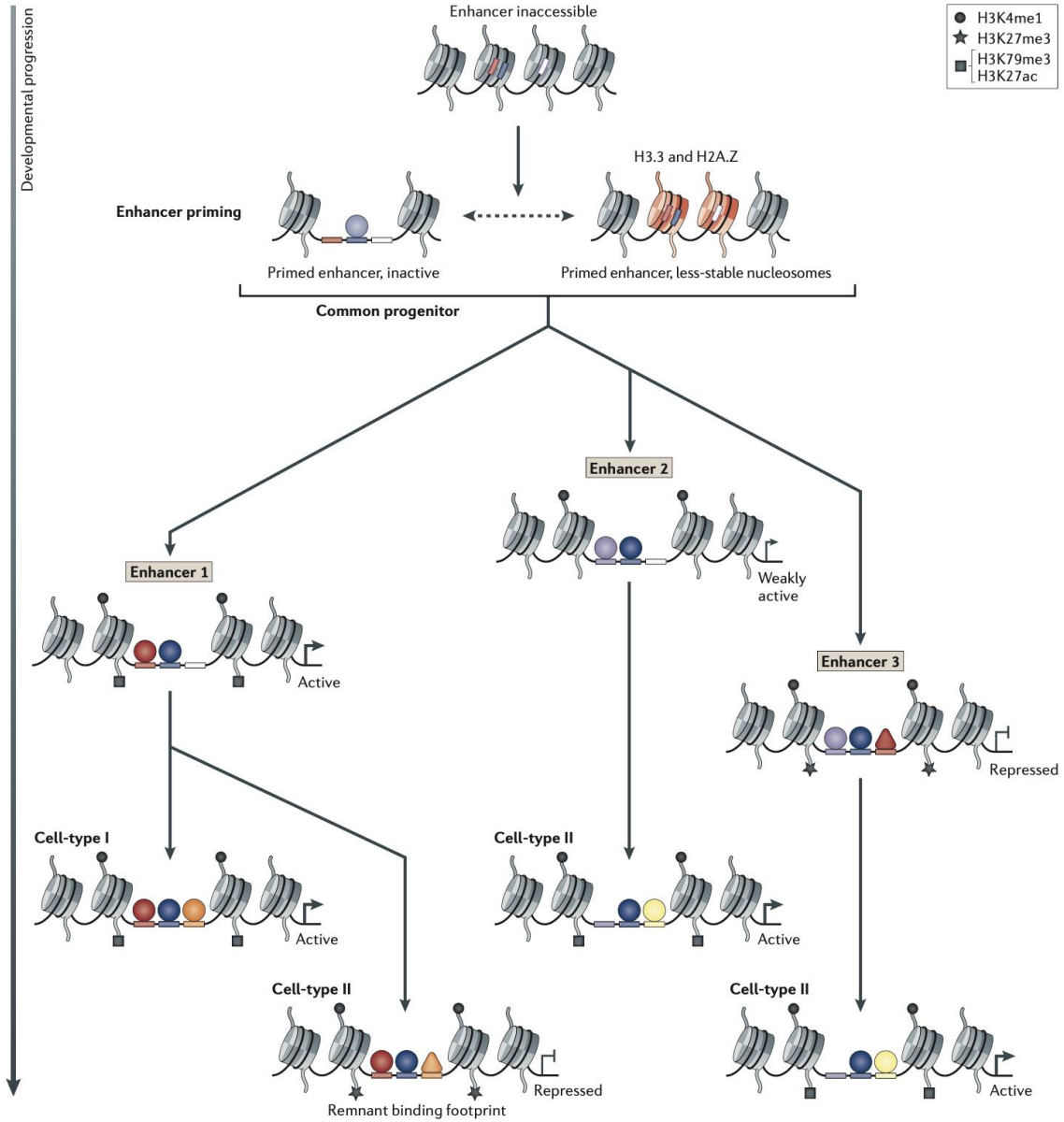


National Human Genome Research Institute
Campos & Reinberg (2009). Annual Review of Genetics



Gene regulation via enhancers/TFs

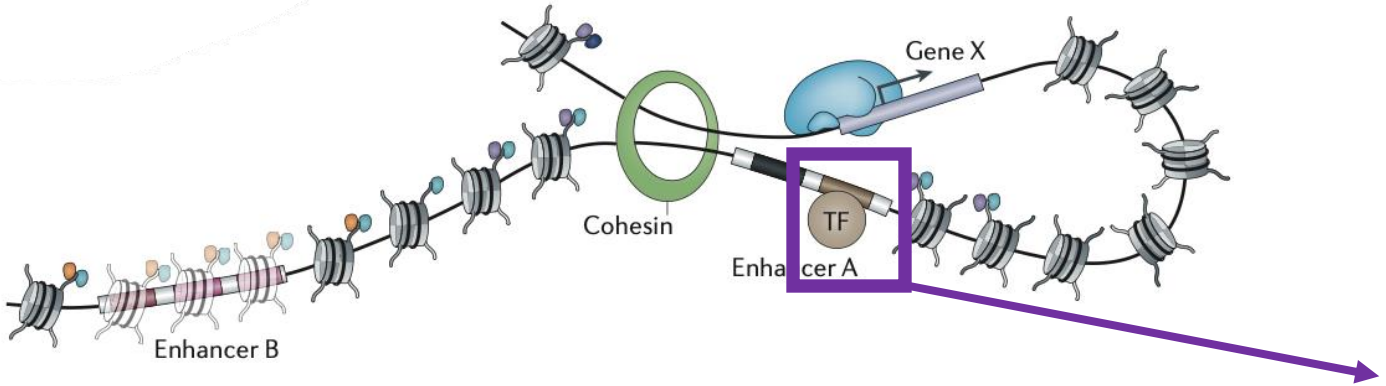
Shlyueva, Stampfel & Stark (2014). *Nat Rev Genet*

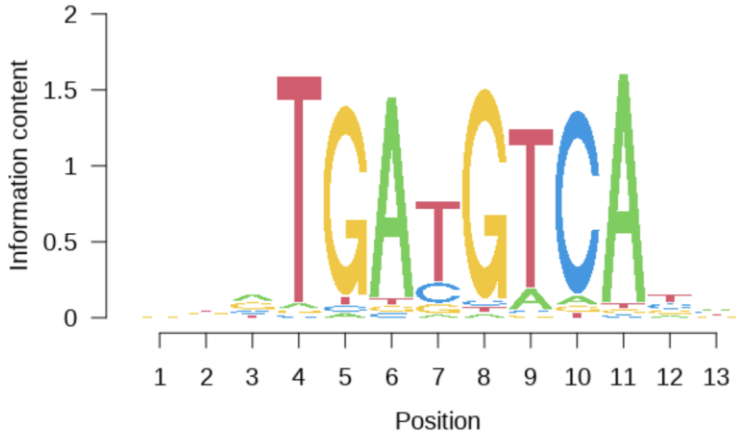


Dynamic changes at enhancers during developmental progression.

Spitz & E. M. Furlong (2012). *Nat Rev Genet*

How can we represent a TFBS or motif?





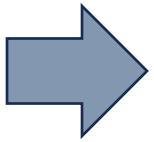
Shlyueva, Stampfel & Stark (2014). *Nat Rev Genet*

Motif databases and PFMs

- Jaspar/Hocomoco

Starting Sequences

#	Sequence
1	AAGAAT
2	ATCATA
3	AAGTAA
4	AACAAA
5	ATTAAA
6	AAGAAT



Position Frequency Matrix (PFM)

Position	1	2	3	4	5	6
A	6	4	0	5	5	4
C	0	0	2	0	0	0
G	0	0	3	0	0	0
T	0	2	1	1	1	2

Sandelin et al (2004). *Nucleic Acids Research*
Kulakovskiy et al (2017). *Nucleic Acids Research*
https://jaspar.elixir.no
https://bioconductor.org/packages/universalmotif/

Position Frequency Matrix (PFM)

Position	1	2	3	4	5	6
A	6	4	0	5	5	4
C	0	0	2	0	0	0
G	0	0	3	0	0	0
T	0	2	1	1	1	2

Position Probability Matrix (PPM) with pseudo-count of 1

Position	1	2	3	4	5	6
A	0.892	0.610	0.036	0.750	0.750	0.610
C	0.036	0.035	0.320	0.035	0.035	0.035
G	0.036	0.035	0.464	0.035	0.035	0.035
T	0.036	0.320	0.180	0.180	0.180	0.320

Position Weight Matrix (PWM)

Position	1	2	3	4	5	6
A	1.840	1.280	-2.807	1.585	1.585	1.280
C	-2.807	-2.807	0.363	-2.807	-2.807	-2.807
G	-2.807	-2.807	0.893	-2.807	-2.807	-2.807
T	-2.807	0.363	-0.485	-0.485	-0.485	0.363

$$PWM_{ij} = log_2\left(\frac{PPM_{ij}}{B_i}\right)$$

How can we additionally reflect the sequence conservation?

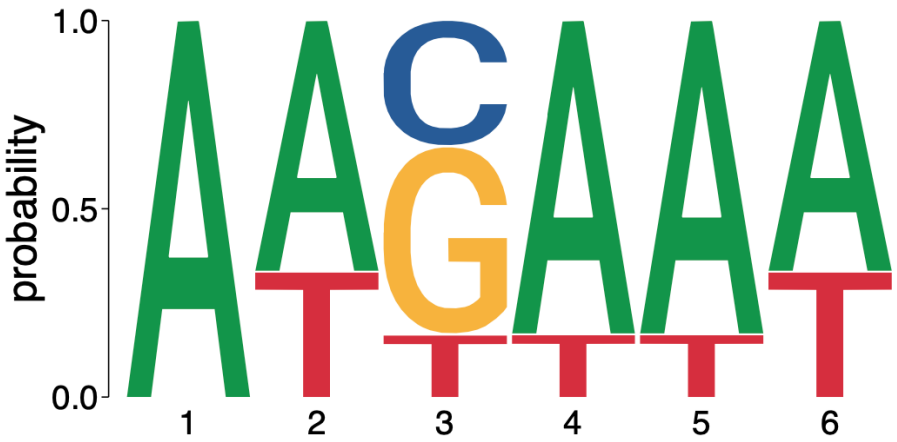


Figure 1: Sequence logo of a Position Probability Matrix

?

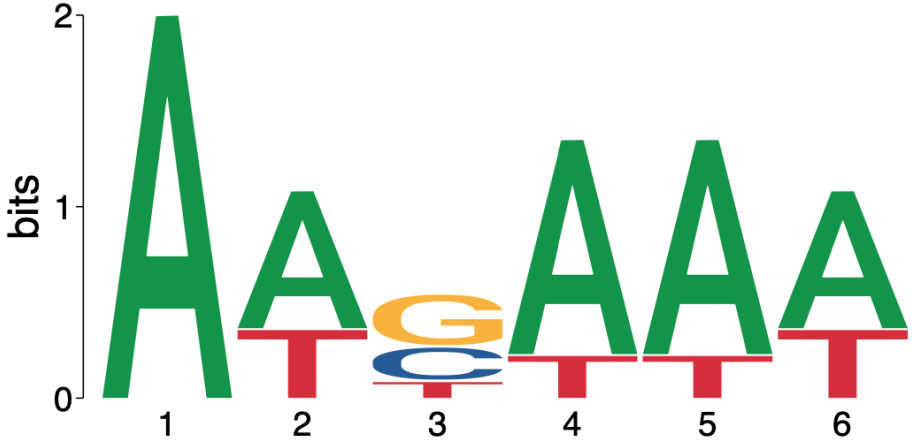


Figure 2: Sequence logo of an Information Content Matrix

i	a_i	p_i	$h(p_i)$
1	a	.0575	4.1
2	b	.0128	6.3
3	c	.0263	5.2
4	d	.0285	5.1
5	e	.0913	3.5
6	f	.0173	5.9
7	g	.0133	6.2
8	h	.0313	5.0
9	i	.0599	4.1
10	j	.0006	10.7
11	k	.0084	6.9
12	l	.0335	4.9
13	m	.0235	5.4
14	n	.0596	4.1
15	o	.0689	3.9
16	p	.0192	5.7
17	q	.0008	10.3
18	r	.0508	4.3
19	s	.0567	4.1
20	t	.0706	3.8
21	u	.0334	4.9
22	v	.0069	7.2
23	w	.0119	6.4
24	x	.0073	7.1
25	y	.0164	5.9
26	z	.0007	10.4
27	-	.1928	2.4
$\sum_i p_i \log_2 \frac{1}{p_i}$			4.1

Shannon Information Content
(can think of it as a measure of surprise)

$$h(x) = \log_2 \frac{1}{P(x)}$$

Entropy (Uncertainty)
(average Shannon Information content)

$$H(X) \equiv \sum_{x \in \mathcal{A}_X} P(x) \log \frac{1}{P(x)}$$

Shannon (1948). *A mathematical theory of communication*
McKay (2005). *Information Theory, Inference, and Learning Algorithms*

Total information per position along the TFBS

Difference in Uncertainty (per position)

$$IC_{final} = IC_{total} - U$$
$$= 2 - U$$

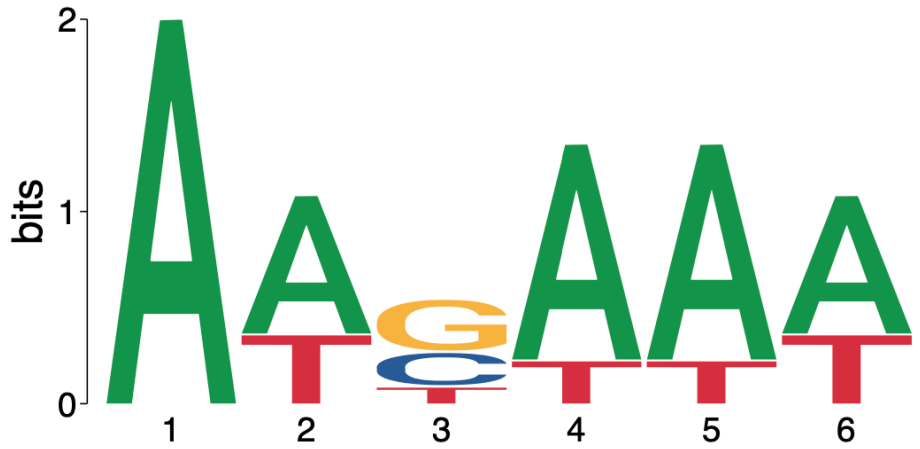
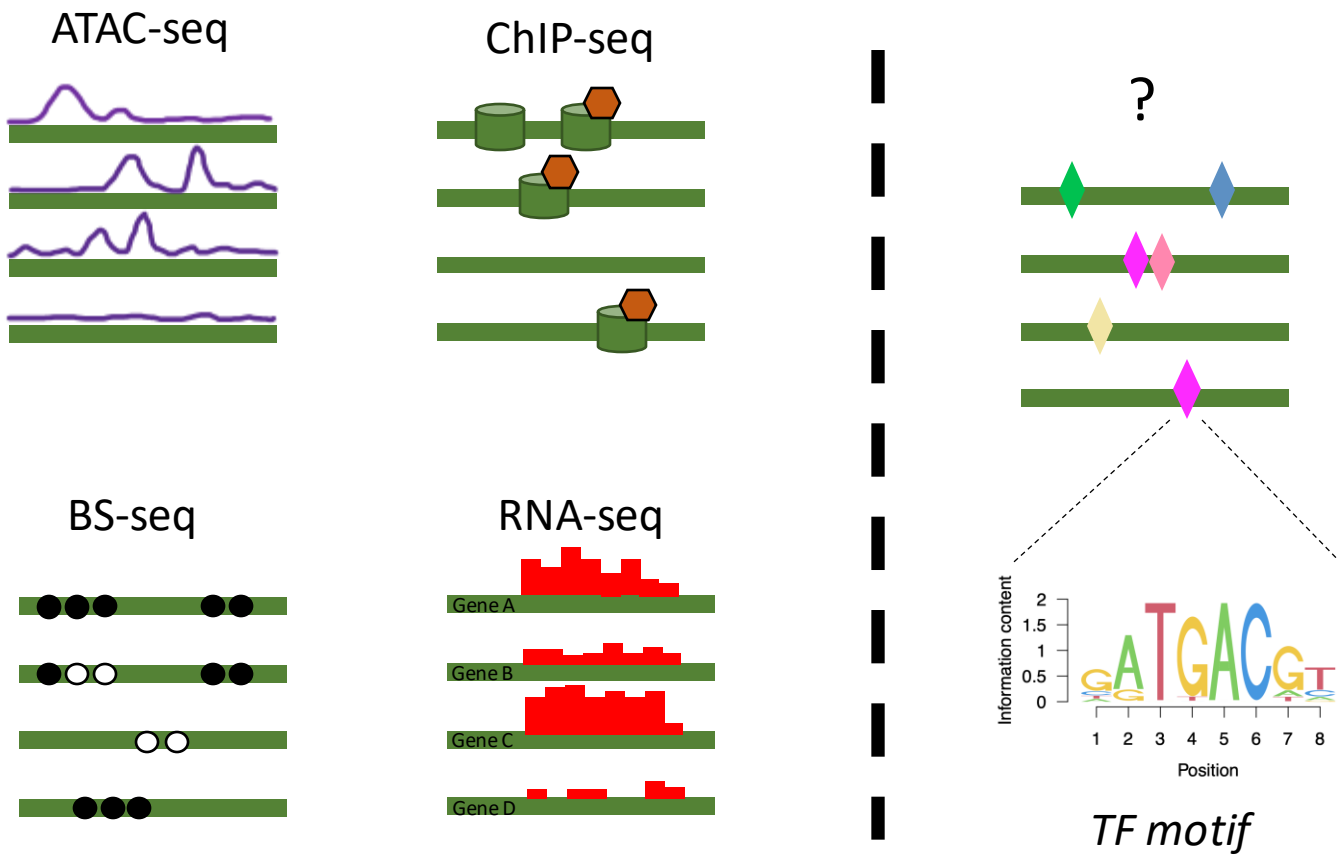


Figure 2: Sequence logo of an Information Content Matrix

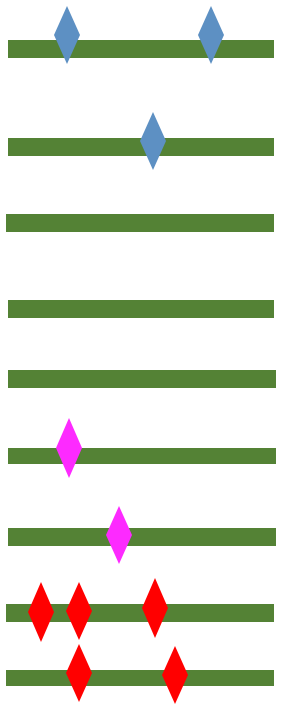
PPM x IC_{final}

Can we find relevant TFs explaining the data?



Can we find relevant TFs explaining the data?

DNA sequences



Experimental read-out

−−−−
−−
−
−
+
++
++
+++
++++++

Binned Approach

Enriched TFs

−−−−
−−
−
−
+
++
++
+++
++++++

Bin 1

Bin 2

Bin 3

Bin 4

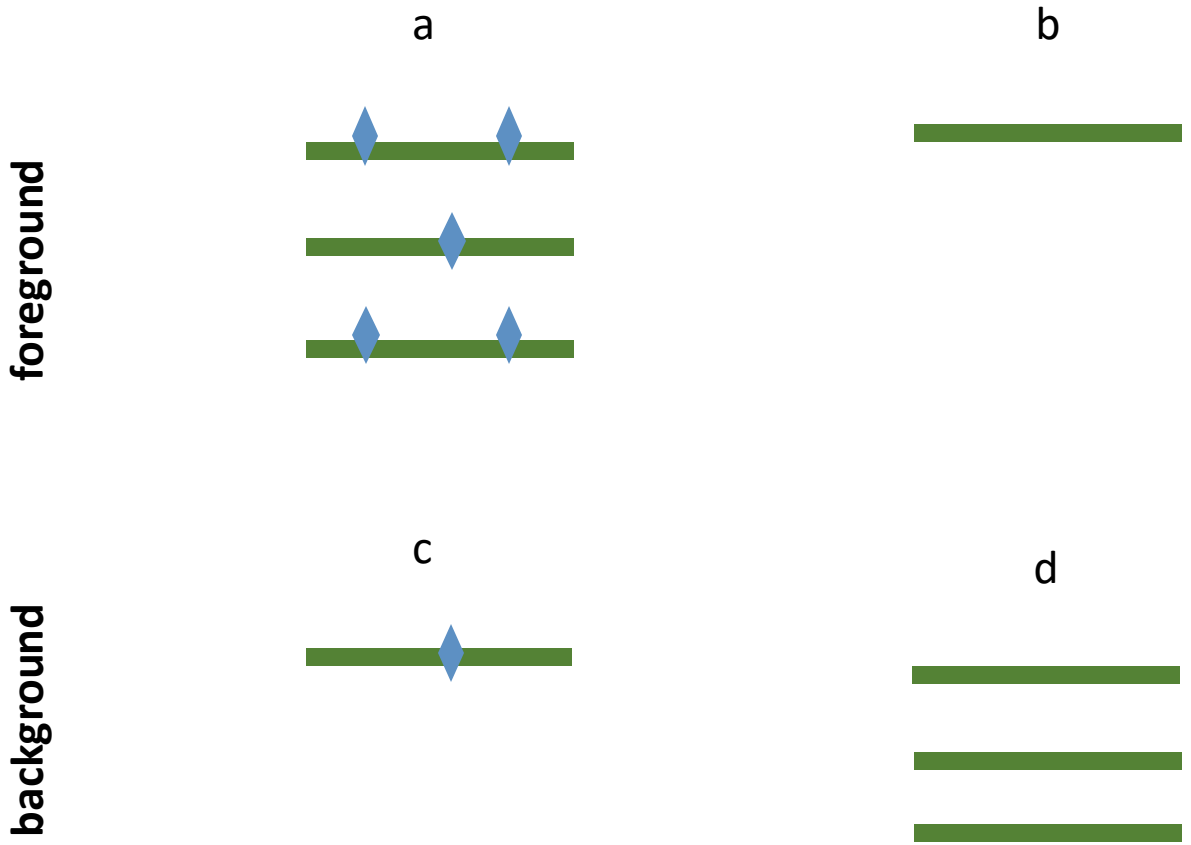
Regression Approach

Selected TFs

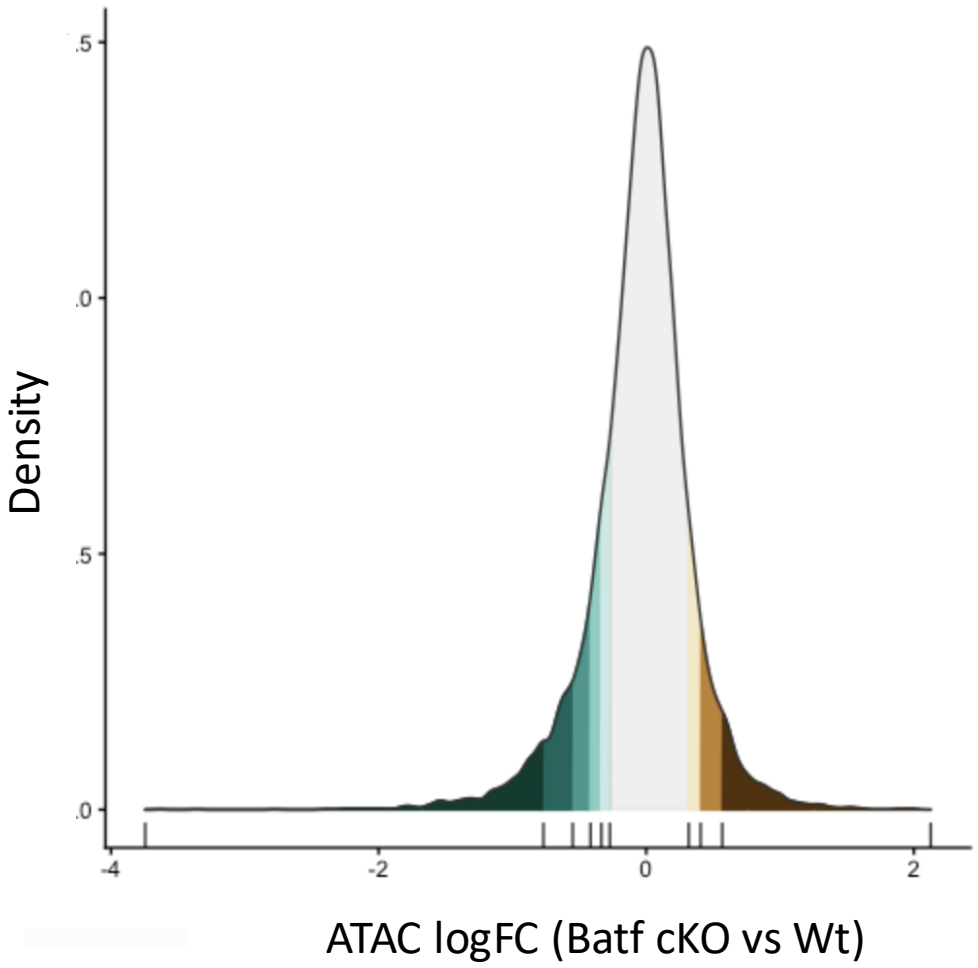
−−−−		2	0	0	0
−−		1	0	0	0
−		0	0	0	0
−		0	0	0	0
+	~	0	0	0	0
++		0	1	0	0
++		0	1	0	0
+++		0	0	3	0
++++++		0	0	2	0

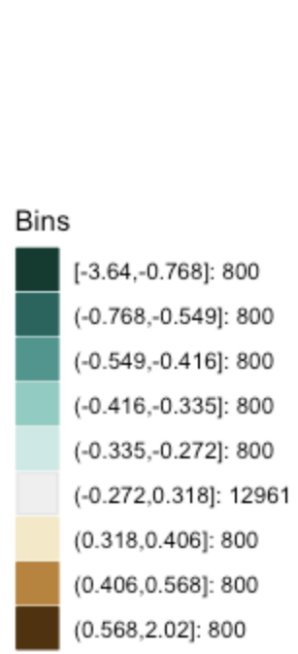
Enrichment Test (Fisher's exact test)

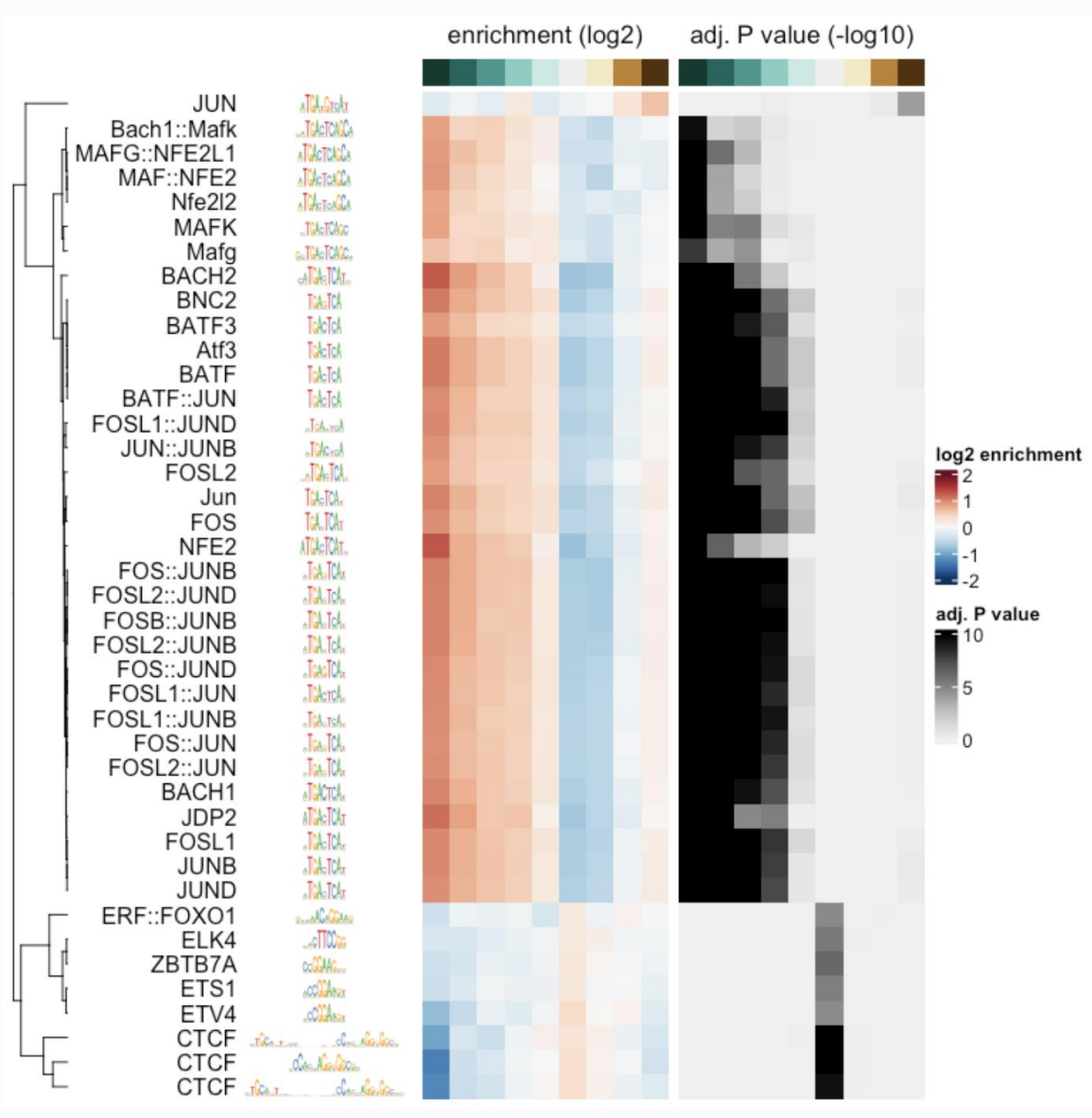
	with TF hit	with no TF hit
foreground	a	b
background	c	d



Binned enrichment approach







Regression approach for TF *selection*

$$y \sim \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

			TF_1	TF_2	TF_p
e_1	$logFC_1$		$\beta_1 * 1$	$\beta_2 * 0$	$\beta_p * 2$
e_2	$logFC_2$	$\sim$	$\beta_1 * 0$	$\beta_2 * 0$	$\beta_p * 0$
.	.		.	.	.
.	.		.	.	.
.	.		.	.	.
e_n	$logFC_n$		$\beta_1 * 0$	$\beta_2 * 0$	$\beta_p * 0$

Randomized Lasso Stability Selection

$$\hat{\beta}^{\lambda,\mathrm{W}} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left(\|Y - X\beta\|_2^2 + \lambda \sum_{k=1}^{p} \frac{|\beta_k|}{W_k} \right)$$

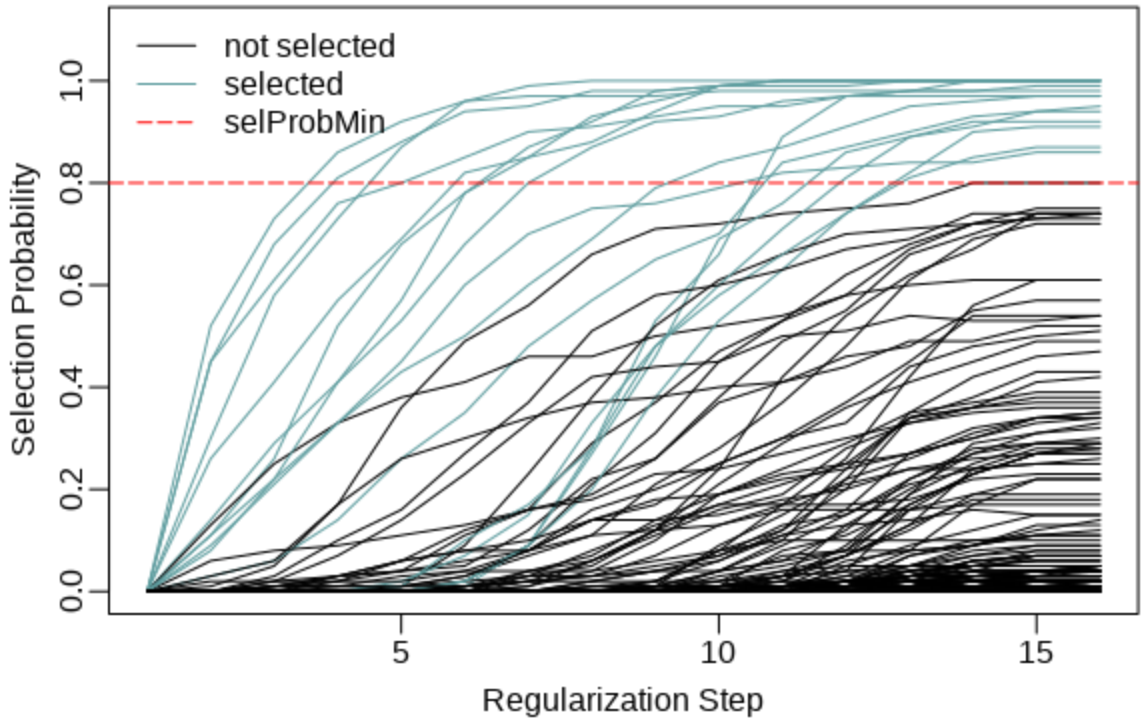
Penalty term to *select* TFs

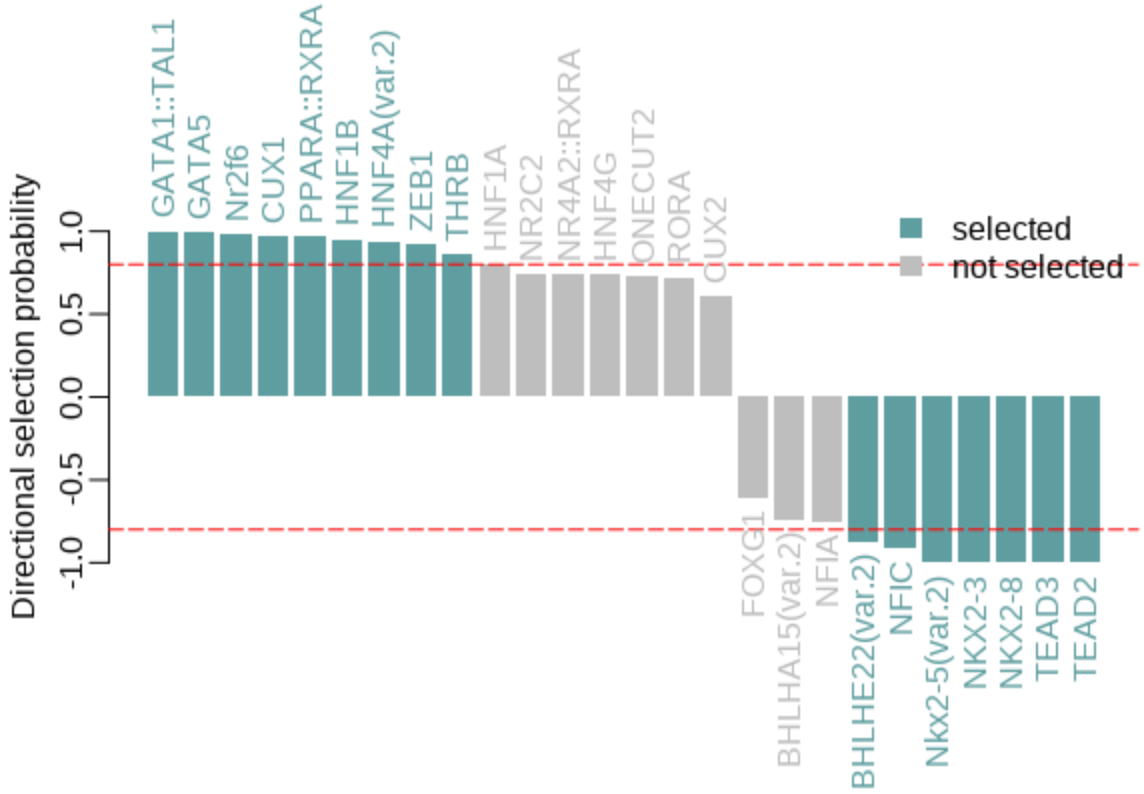
$logFC$

TF binding sites

Meinshausen & Bühlmann (2010). *Journal of the Royal Statistical Society Series B: Statistical Methodology*

Selected TFs



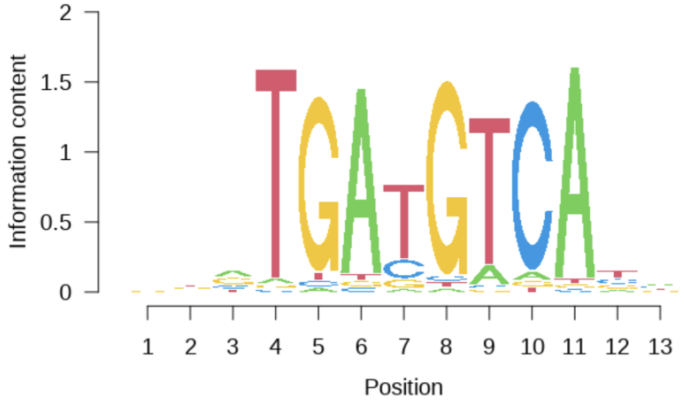


References and resources

- **Benchmarking papers:**
  - Gerbaldo, F. E., Sonder, E., Fischer, V., Frei, S., Wang, J., Gapp, K., Robinson, M. D., & Germain, P.-L. (2024). On the identification of differentially-active transcription factors from ATAC-seq data. *PLOS Computational Biology, 20*(10), e1011971.https://doi.org/10.1371/journal.pcbi.1011971
  - Santana, L. S., Reyes, A., Hoersch, S., Ferrero, E., Kolter, C., Gaulis, S., & Steinhauser, S. (2024). Benchmarking tools for transcription factor prioritization. *Computational and Structural Biotechnology Journal, 23*, Article 1274-1287. https://doi.org/10.1016/j.csbj.2024.03.016
- **Jaspar website:** https://jaspar.elixir.no
- **Bioconductor packages:**
  - https://bioconductor.org/packages/universalmotif/
  - https://bioconductor.org/packages/monaLisa/
- **More on Information Theory:** David McKay's book "Information Theory, Inference, and Learning Algorithms"
- **Stability selection paper:** Meinshausen, N., & Bühlmann, P. (2010). Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(4), 417–473. https://doi.org/10.1111/j.1467-9868.2010.00740.x
- **monaLisa paper:** Machlab, D., Burger, L., Soneson, C., Rijli, F.M., Schübeler, D., Stadler, M.B. (2022). monaLisa: an R/Bioconductor package for identifying regulatory motifs. *Bioinformatics*, *38*(9), 2624-2625. https://doi.org/10.1093/bioinformatics/btac102

Exercises

- Part 1: representing motifs



- Part2: finding relevant TFs