6.874 Recitation 4

Corban Swain

February 27 & 28, 2020

The "Central Dogma" defines information flow in the cell.

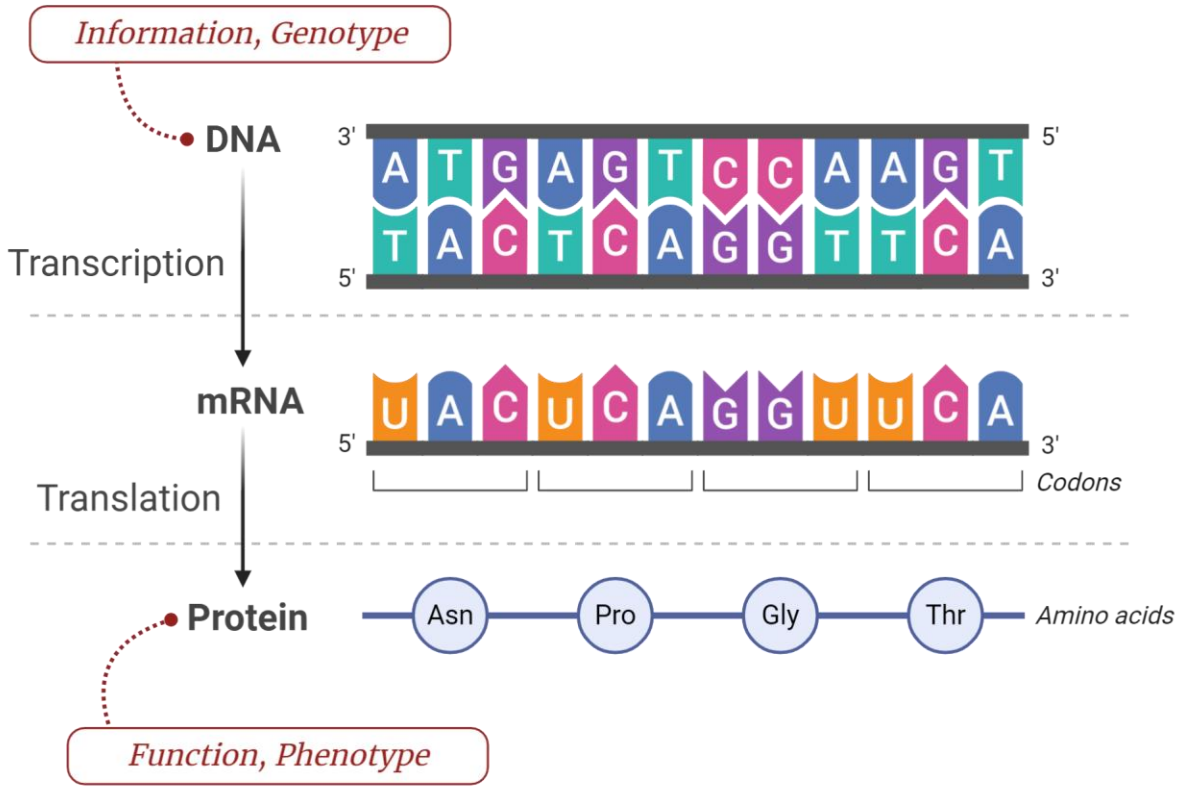
DNA Regulation
DNA Accessibility
DNA structure,
marks on the Backbone
histone presence & modifications
sequence integrity
damage and repair

Transcriptional Regulation
RNA polymerase II binding
transcroption factor binding
enhancer binding
full transcriptional transit along sequence

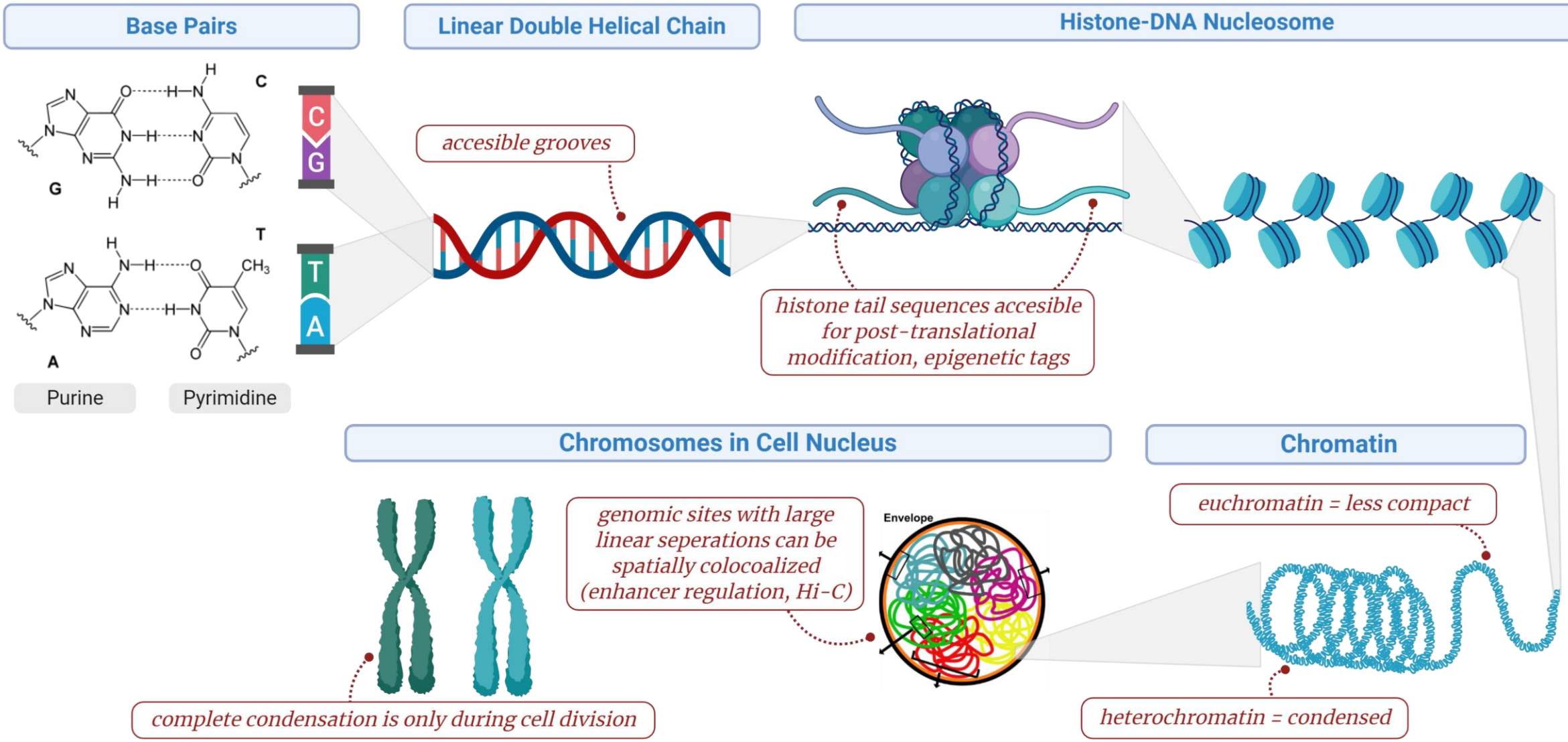
mRNA Regulation
RNA degredation
export from nucleus
RNA processing (e.g. intron excision)
RNA interference

Translational Regulation
Translational machinery
ribosome binding
tRNA avalibility
ribosomal halting

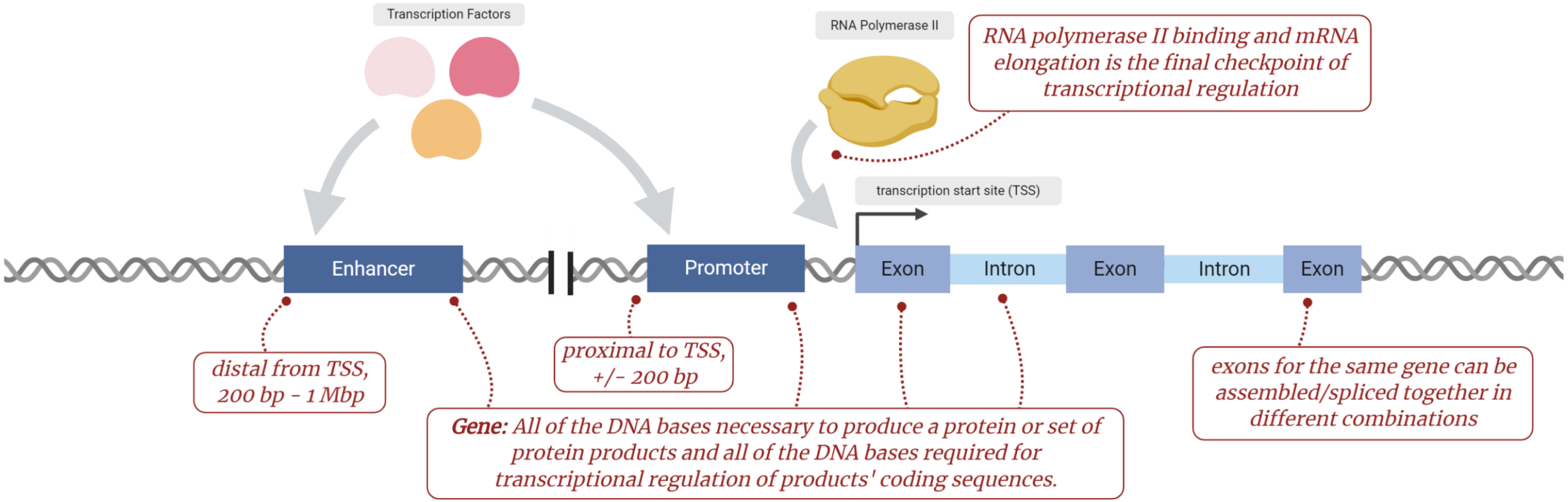
Protein Regulation
post-processing
phosphorylation
degredation tags
export and release into ECM or onto cell surface
miltimerization
affector molecule binding
cofactor binding
intracelluar compartment movement



DNA is structured across many scales.



Genes **are the primary functional units of the** ***Genome.***



Gene Coding Sequences, 1.2%
Exons in the open reading frame

Gene Non-coding Sequences, 40-65%
Introns in the open reading frame
RNA Pol II Binding Site
Promoters
Enhancers
Repressive Domains

Other Sequences
long noncoding RNAs
Repetitious DNA
intergenic regions
telomeres

Different parts of the chromatin can exist in different functional states.

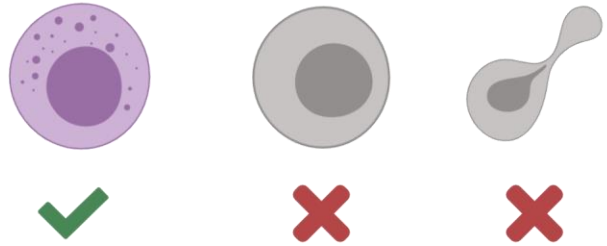
Transcriptional Activation
Transcription Factor-DNA Binding at Promoter & Enhancer
Pioneer Protein Binding
Enhancer - Promoter colocalization
Histon Acetylation (e.g. H4 Lysine)

Transcriptional Inactivation
Protein-DNA Binding at Repressor
TF Degradation
Histone deacetylation
Histone Methylation
HP1 Histone Binding

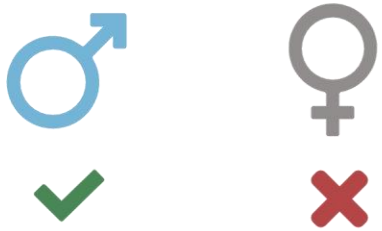
coordinated implementation of transcriptional programs

Tissue Differentiation

Cellular State & Enviornmental Response



Chromosamal Inactivation



Next-generation sequencing technologies enable us to quantify & localize nucleic acid molecules to the genome.

→ the "raw data" of NGS of technologies are short (≈30 bp) sequence reads

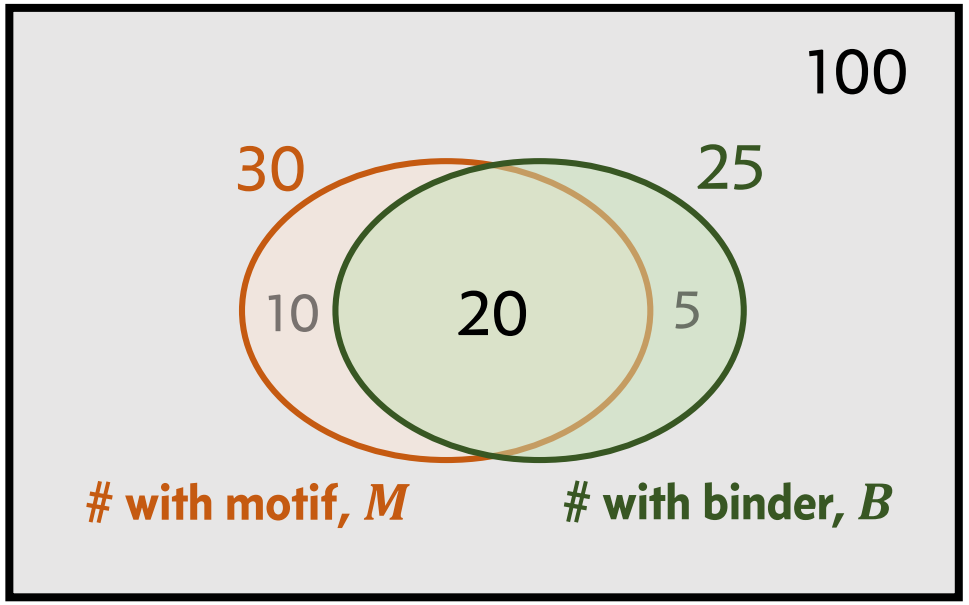
→ Reads Correspond To:

- ChIP-Seq - *fragments pulled down with antibody against a DNA binder*
- DNse-Seq – *fragments accessible to enzymatic cutting by DNase-I*
- ATAC Seq – *fragments accessible to Tn5 Transposase activity*

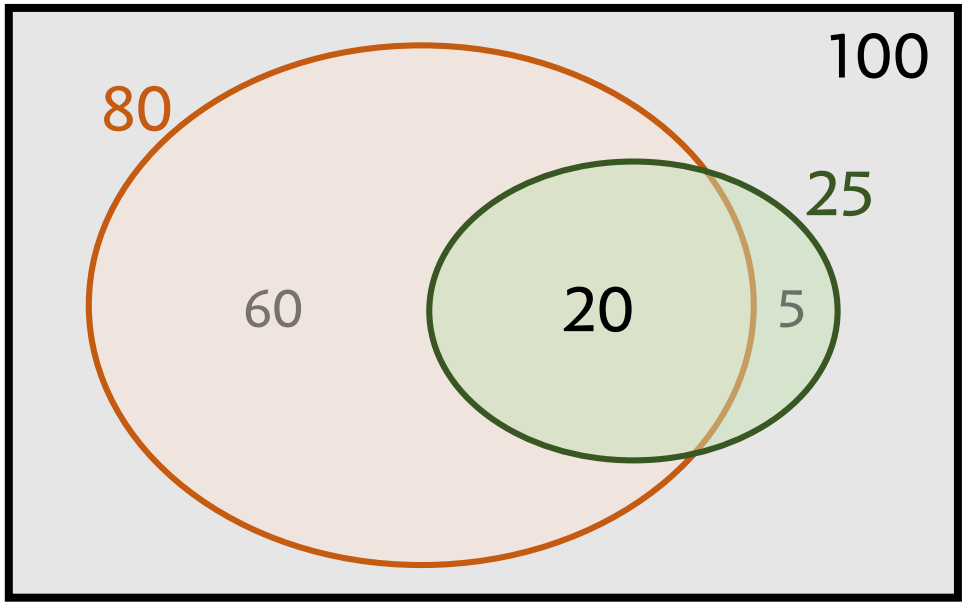
→ Issues

- Reads can map to multiple places
- Base statistics can be poor
- Repetitive elements in the genome could give erroneous results

The hypergeometric distribution allows us to calculate probabilities of enrichment.



total # sequences, T



$$P_{null} = \frac{\binom{M}{x}\binom{T-M}{B-x}}{\binom{T}{B}} = \frac{\binom{30}{20}\binom{100-30}{25-20}}{\binom{100}{25}} = 1.5 \times 10^{-9}$$

$$p = P_{null}(x \geq 20) = 2.0 \times 10^{-9}$$

$$P_{null} = \frac{\binom{80}{20}\binom{100-80}{25-20}}{\binom{100}{25}} = 0.22$$

$$p = P_{null}(x \geq 20) = 0.62$$

DNA Sequences can be represented and processed in an "image" context with CNNs.

Images	CNN Model Features	DNA Sequences
2D grid of pixel values with 1 (monochrome) or 3 (color) channels	Input Representation	*1D array of one-hot encoded DNA sequences*
low-level: edges, shapes *high-level: objects, faces*	Kernel Representations	*low-level: sequence motifs* *high-level: motif combinations & grammar*
probabilities of different object classes	Model Outputs	*predictions of bound or unbound (single- & multi-class), chromatin state*

DNA Sequences can be represented and processed in a "timeseries" context with RNNs.

Spoken Audio Timeseries	RNN Model Features	DNA Sequences
time, evaluating phonemes or words at each time step	Input Axis	*base position, evaluating bases at each sequence-step*
context (within a question, beginning/end of a sentence), vocal profile or accent	Hidden States	*type of DNA region being read (ORF, promoter, etc.); memory of previous motifs*

Responses to asked questions

→ What does it mean to classify data as reproducible and irreproducible in the context of IDR?

- When an observed event (e.g. TF binding to a specific sequence) is due to an event that actually happened, we want to be able to classify it as "reproducible." **When an observed event is due to noise (e.g. non-specific interactions) in the experiment, we want to be able to classify it as "irreproducible" so that we can confidently discard that observation as meaningful.** Once we have made this classification for all observations, we can make prediction about how we would expect the distribution of rank-ordered *p*-values to vary between replicate experiments. This expected distribution can be compared to the real distribution across replicates, and we can use the differences to update the classification scheme. This process is repeated (expectation-maximization algorithm) until the error converges or we choose to stop.

- *Li, Qunhua, et al. "Measuring Reproducibility of High-throughput Experiments." The Annals of Applied Statistics, vol. 5, no. 3, 2011, pp. 1752–79, doi:10.1214/11-AOAS466.*

Responses to asked questions

→ What are the key differences between DNase-seq and ATAC-seq?

- ATAC-seq requires fewer cells (500 – 50 000, vs. > 1 million for DNase-seq) and is recently the more widely used protocol
- ATAC-seq provides lower accuracy compared to DNase-seq.
- ATAC-seq simultaneously fragments and tags DNA with sequencing adaptors *in vitro*, while DNase-seq requires adaptor ligation as an additional step after enzymatic sequence fragmentation.
- The Tn5 preferential cleavage motif is 9 – 13 bp long, while the DNase-I preferential cleavage motif is 5 – 6 bp long.

- *Li, Zhijian et al. "Identification of transcription factor binding sites using ATAC-seq." Genome biology vol. 20,1 45. 26 Feb. 2019, doi:10.1186/s13059-019-1642-2*
- *Hashimoto, Tatsunori, et al. A Synergistic DNA Logic Predicts Genome-Wide Chromatin Accessibility. 2016, doi:10.1101/gr.199778.115.*
- *Buenrostro, Jason D., et al. "Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position." Nature Methods, vol. 10, no. 12, Nature Publishing Group, Dec. 2013, pp. 1213–18, doi:10.1038/nmeth.2688.*