Chromatin architecture and gene regulation

Recitation 4

MIT - 6.802 / 6.874 / 20.390 / 20.490 / HST.506 - Spring 2021

Jackie Valeri

Slides adapted from Corban Swain and previous course materials

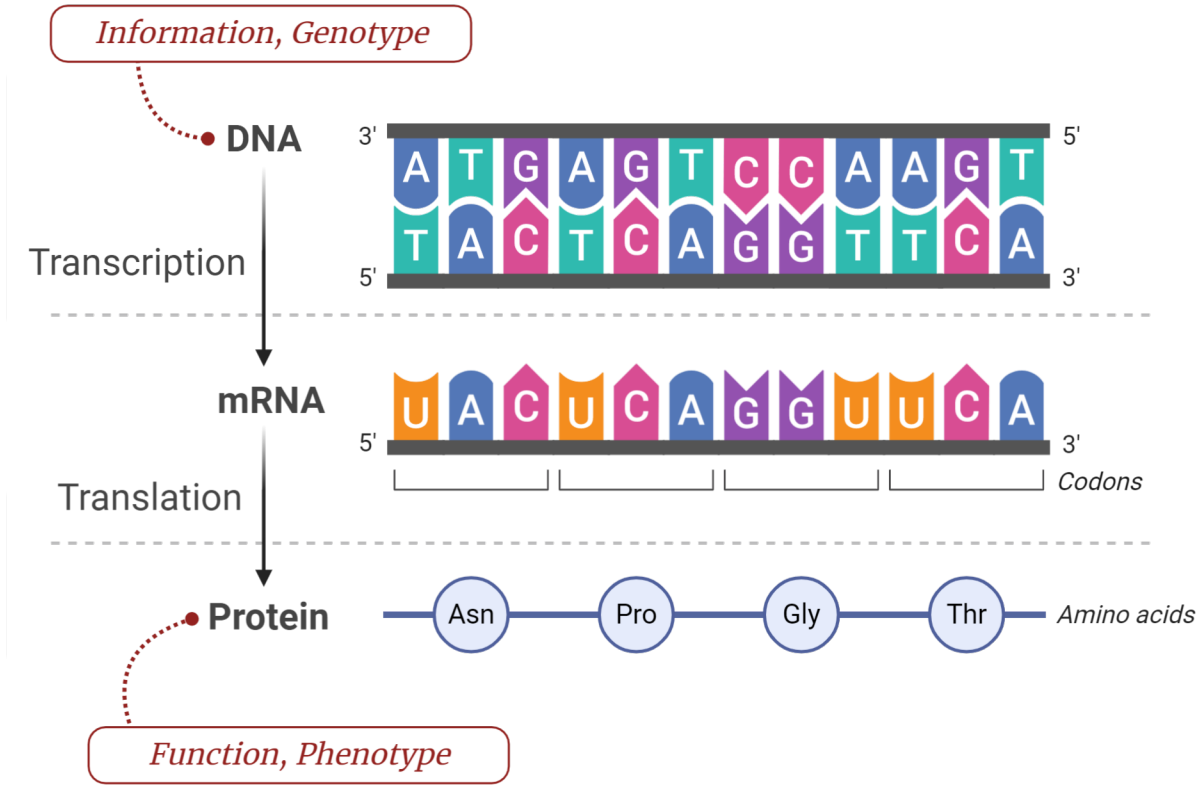
Onto recitation R04!

A. Bio review
   I. Central dogma
   II. Genes as units

B. Chromatin Architecture

C. Quantifying DNA
   I. Next-generation sequencing
   II. DNA + models, math

Bio review: central dogma defines flow of information within a cell



Bio review: central dogma defines flow of information within a cell

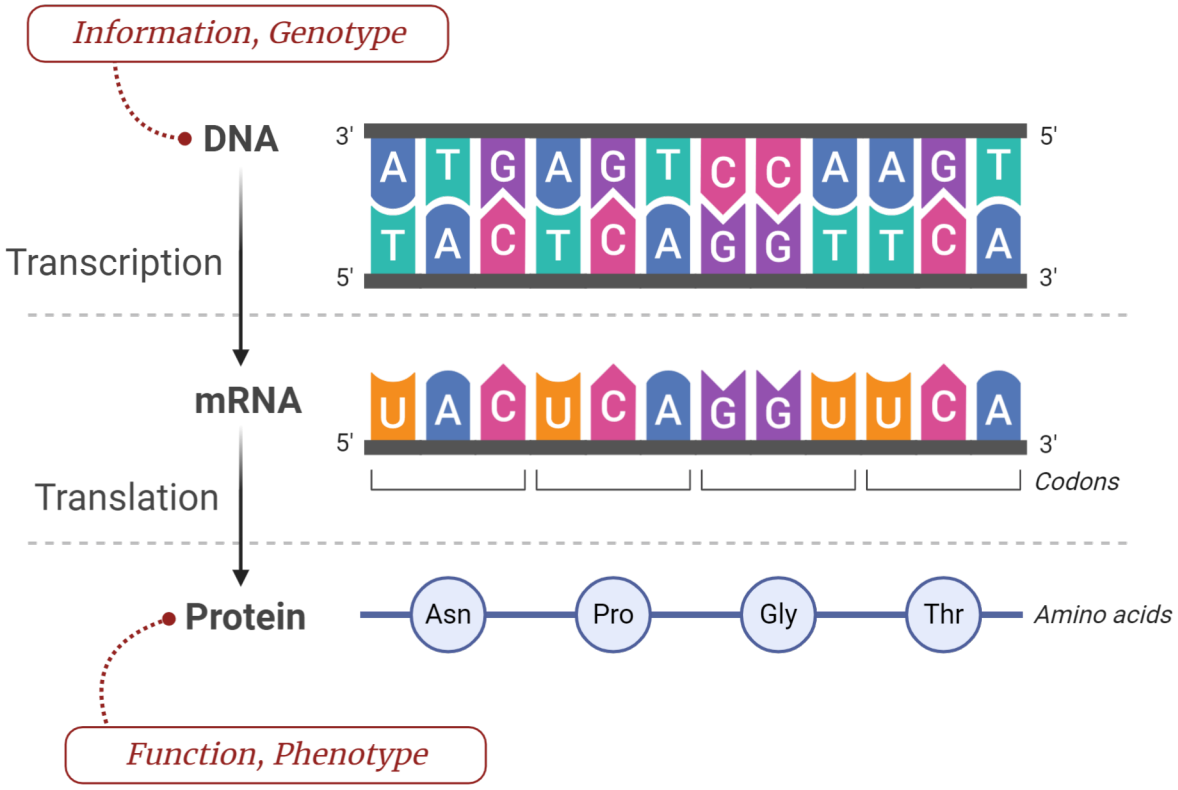
DNA Regulation

Transcriptional Regulation

mRNA Regulation

Translational Regulation

Protein Regulation



Bio review: central dogma defines flow of information within a cell

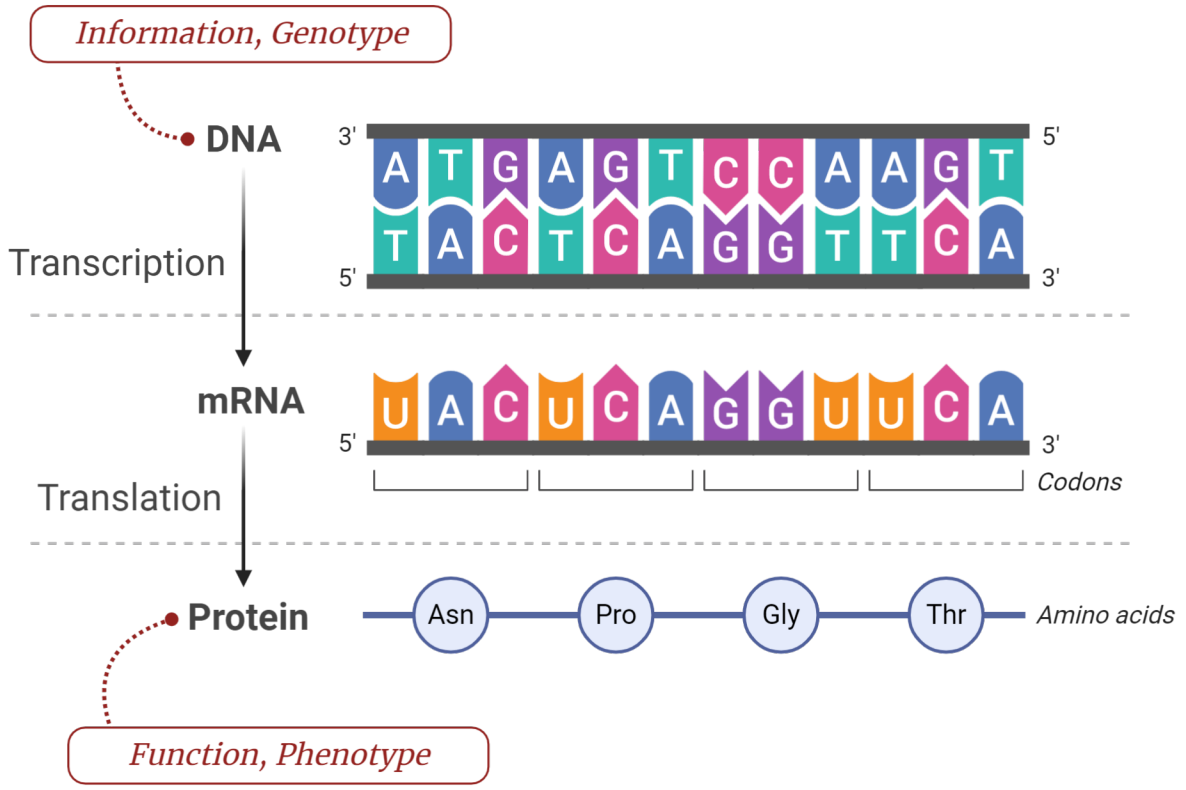
DNA Regulation
DNA Accessibility
DNA structure,
marks on the Backbone
histone presence & modifications
sequence integrity
damage and repair

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

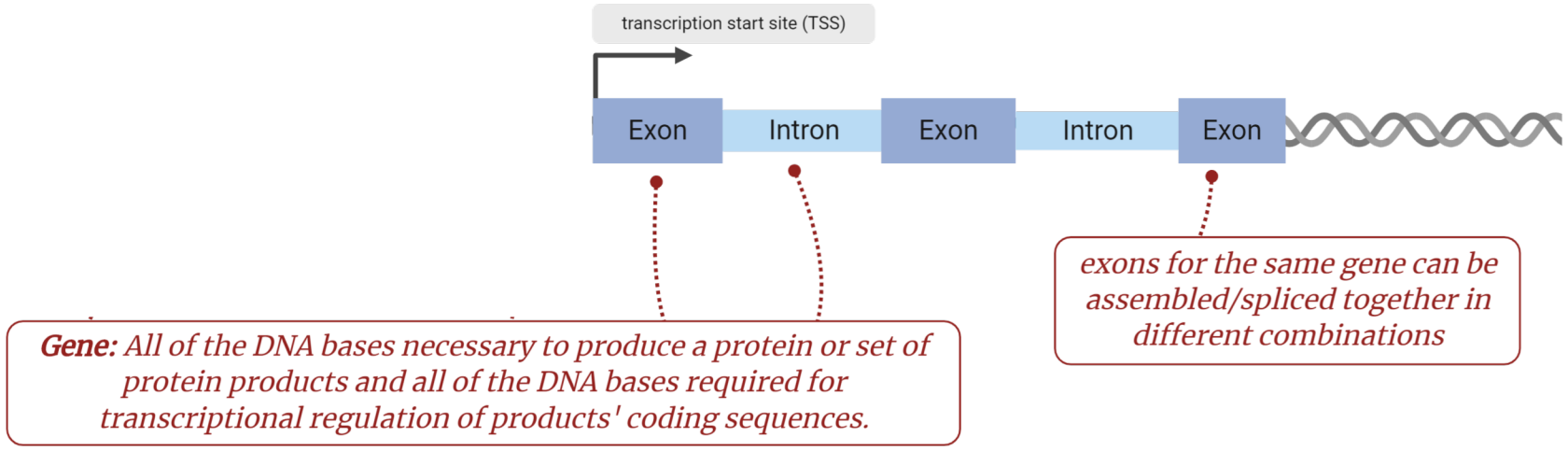
mRNA Regulation
RNA degredation
export from nucleus
RNA processing (e.g. intron excision)
RNA interference

Translational Regulation
Translational machinery
ribosome binding
tRNA avalibility
ribosomal halting

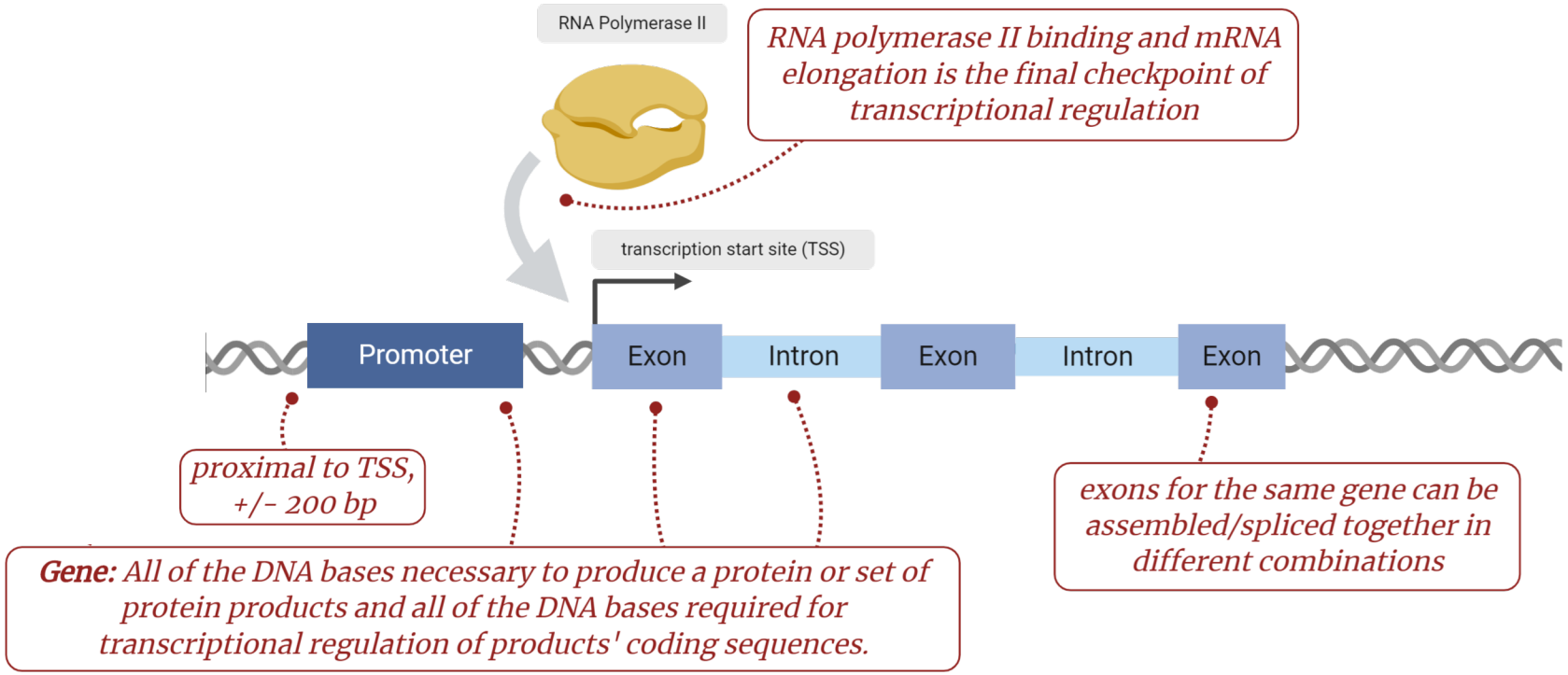
Protein Regulation
post-processing
phosphorylation
degredation tags
export and release into ECM or onto cell surface
miltimerization
affector molecule binding
cofactor binding
intracelluar compartment movement



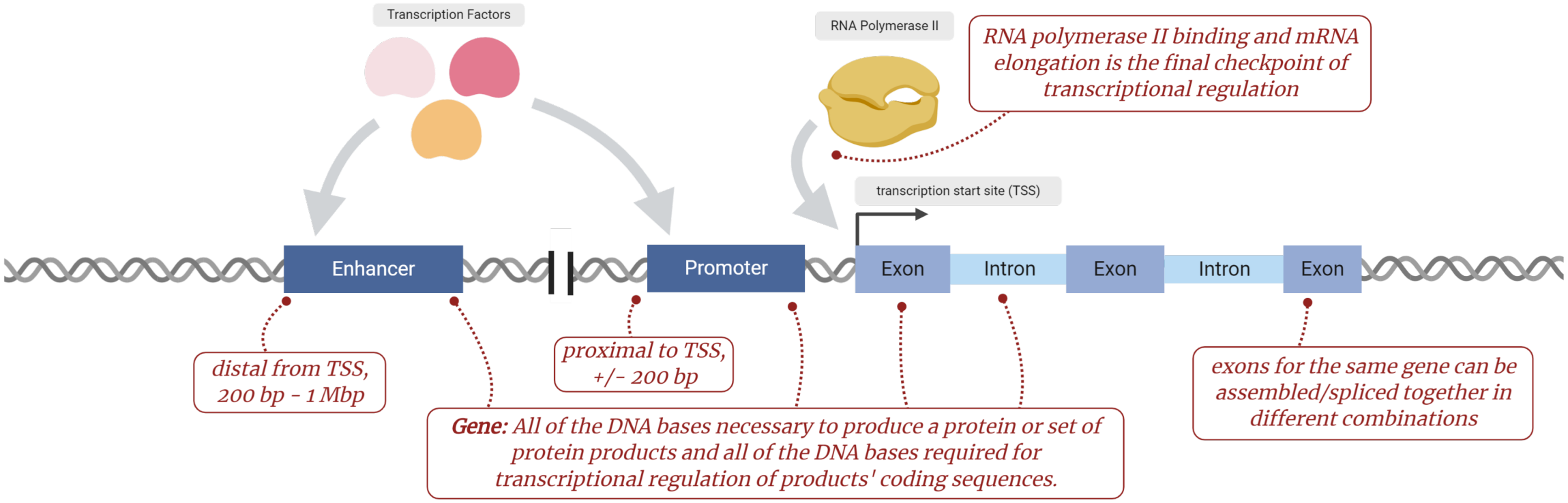
Bio review: genes as the primary functional units of the genome



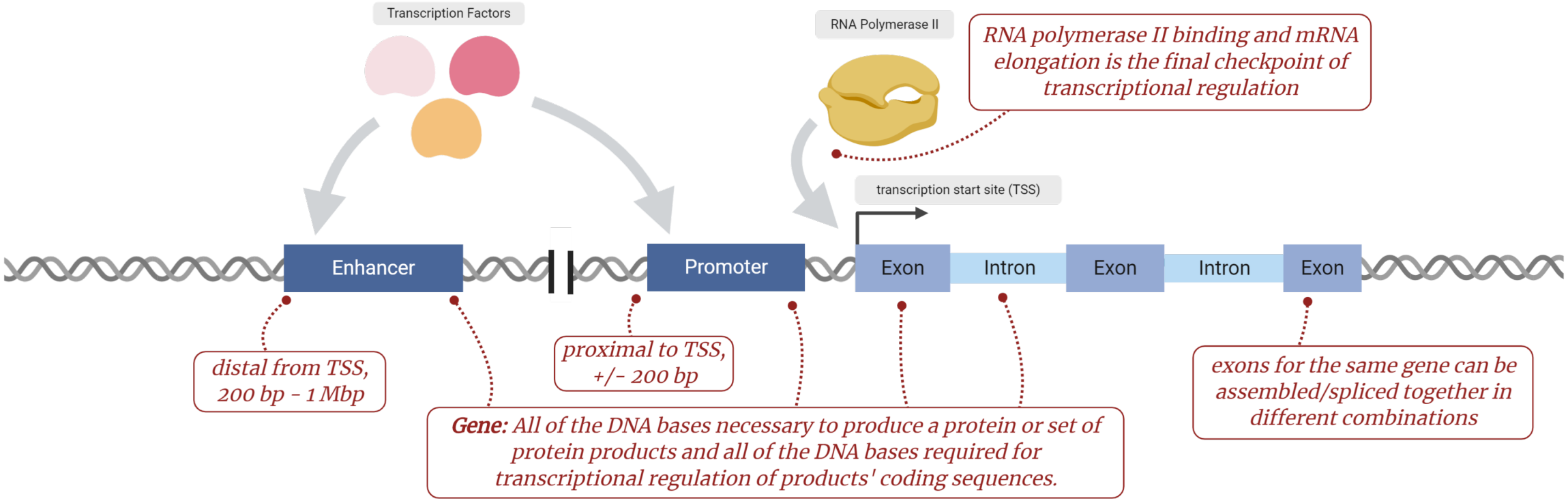
Bio review: genes as the primary functional units of the genome



Bio review: genes as the primary functional units of the genome



Bio review: genes as 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

Onto recitation R04!

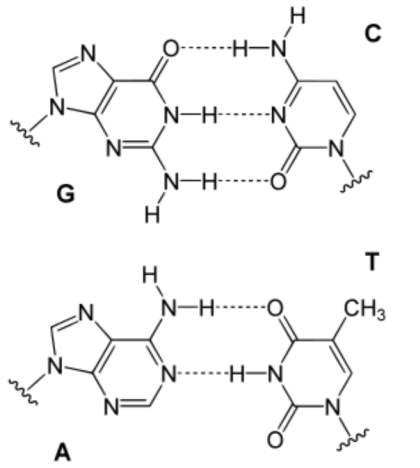
A. Bio review
   I. Central dogma
   II. Genes as units

B. Chromatin Architecture

C. Quantifying DNA
   I. Next-generation sequencing
   II. DNA + models, math

Bio review: DNA is structured across many scales

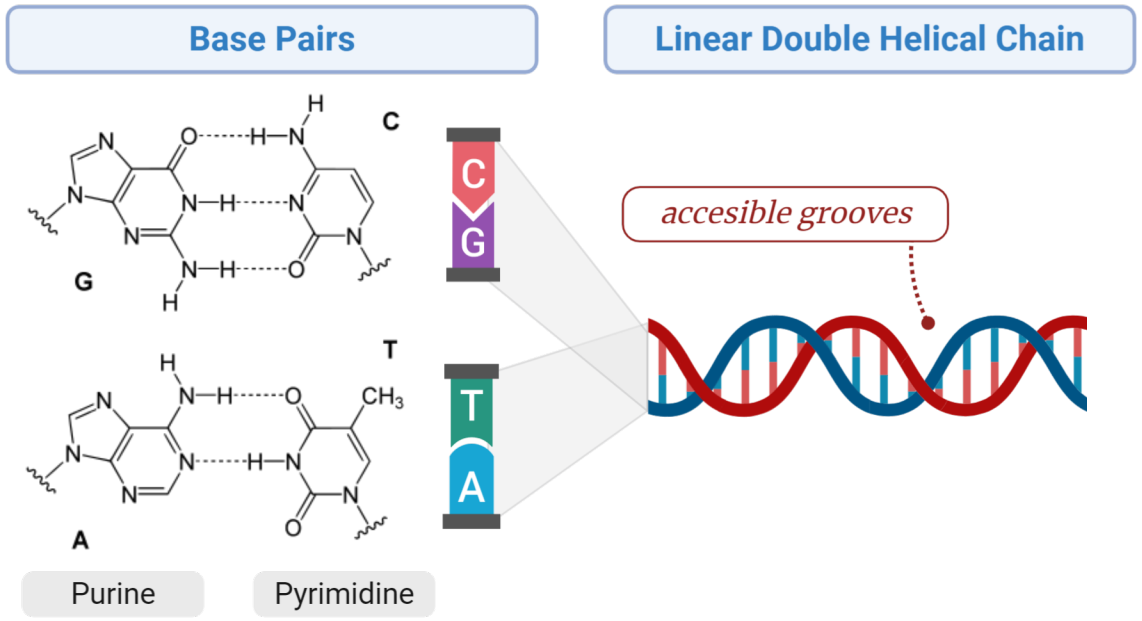
Base Pairs



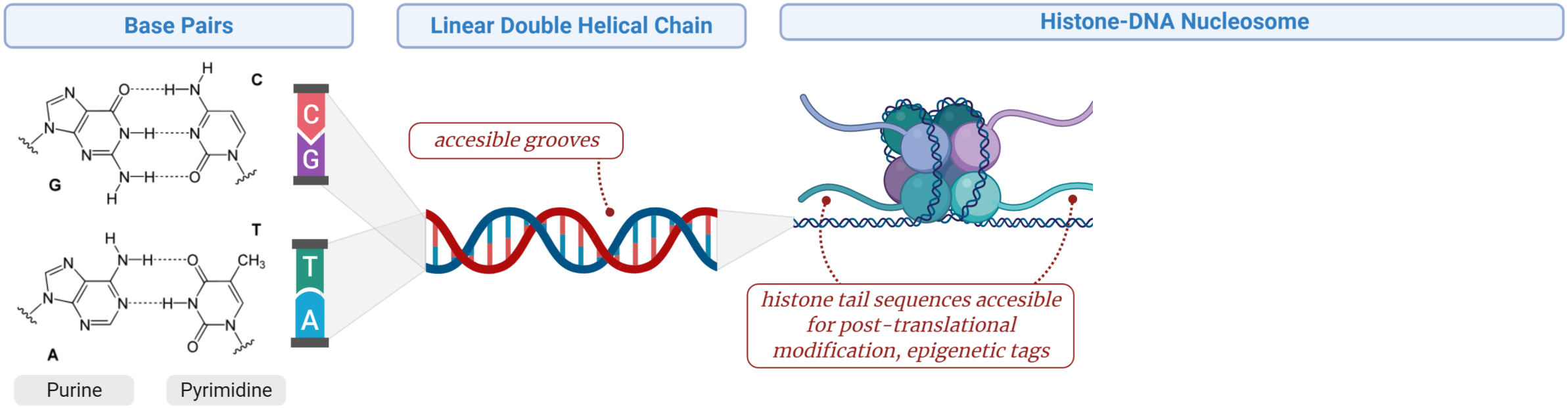
Purine

Pyrimidine

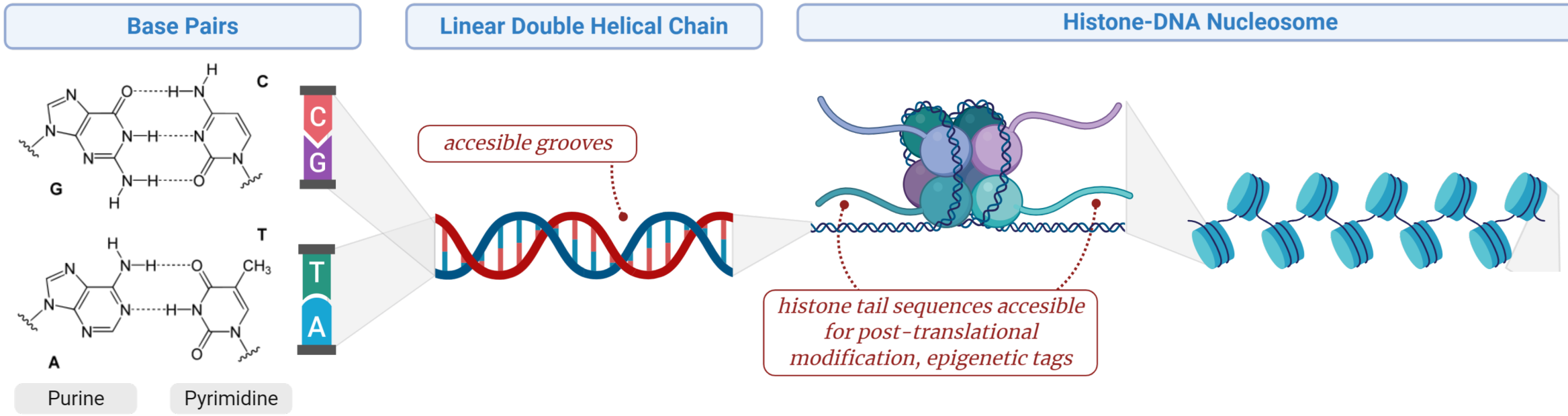
Bio review: DNA is structured across many scales



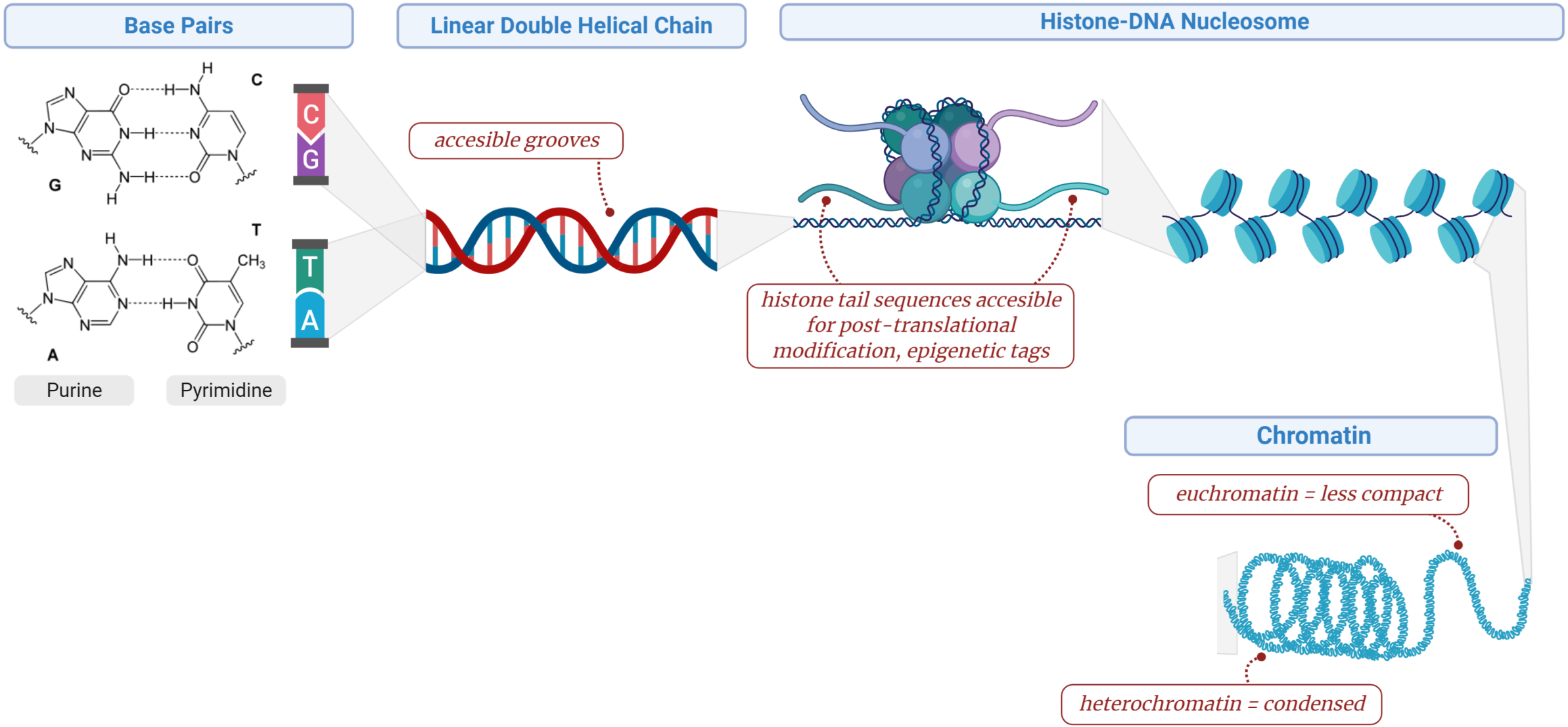
Bio review: DNA is structured across many scales



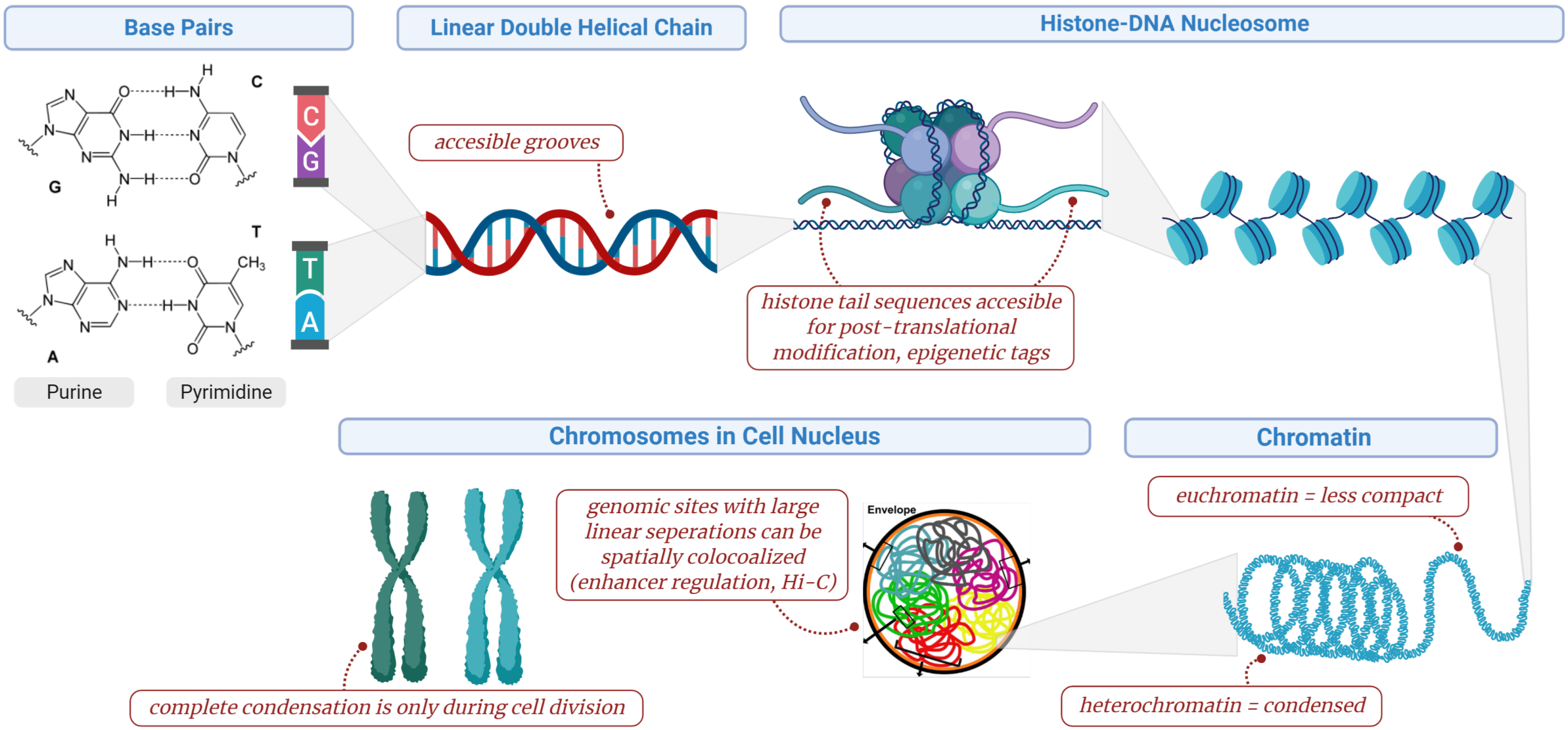
Bio review: DNA is structured across many scales



Bio review: DNA is structured across many scales



Bio review: DNA is structured across many scales



Bio review: 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

Bio review: 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)

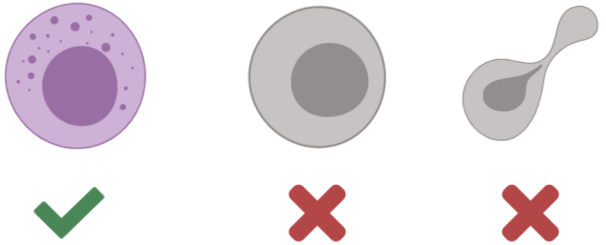
Transcriptional Inactivation

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

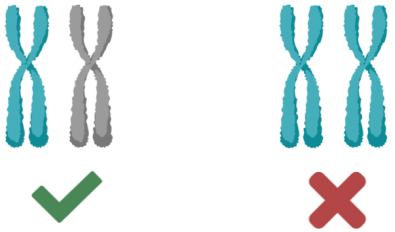
coordinated implementation of transcriptional programs

Tissue Differentiation

Cellular State & Enviornmental Response



Chromosamal Inactivation



Onto recitation R04!

A. Bio review
   I. Central dogma
   II. Genes as units

B. Chromatin Architecture

C. Quantifying DNA
   I. Next-generation sequencing
   II. DNA + models, math

Quantifying DNA: next-generation sequencing

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

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

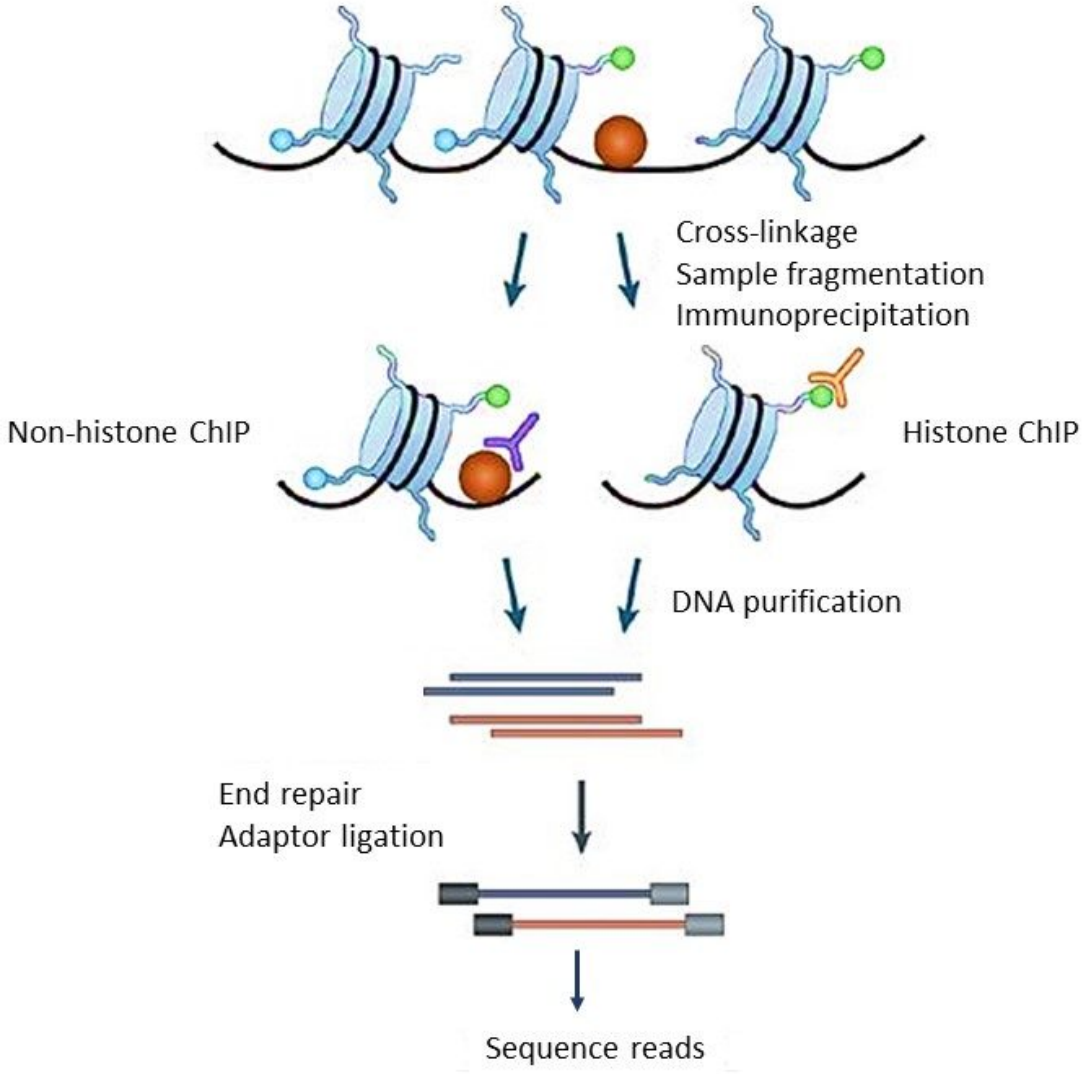
Quantifying DNA: next-generation sequencing

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

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

→Reads correspond to:

- **ChIP-Seq -** ***fragments pulled down with antibody against a DNA binder***



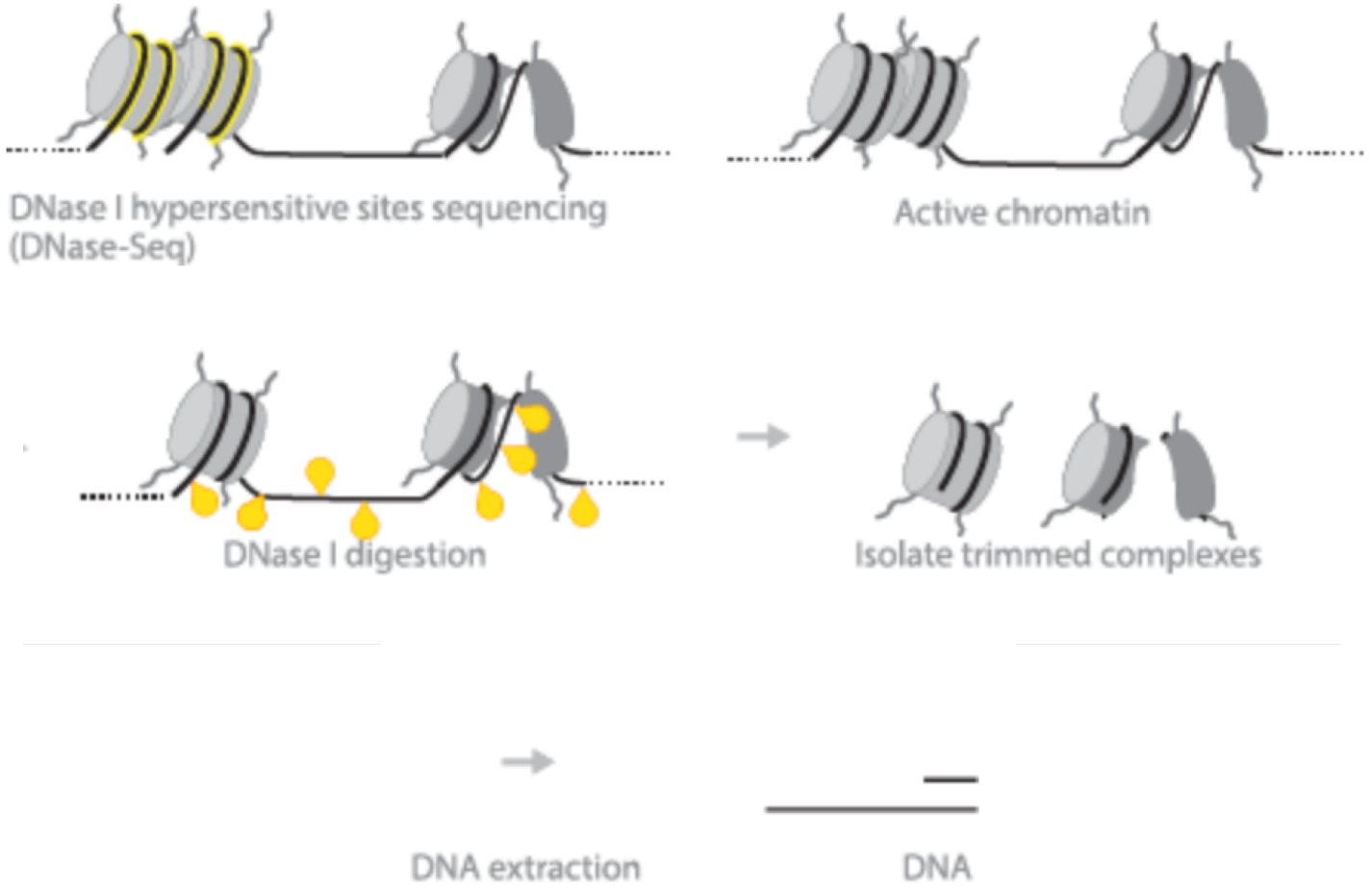
Quantifying DNA: next-generation sequencing

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

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

→Reads correspond to:

- ChIP-Seq - *fragments pulled down with antibody against a DNA binder*
- **DNAse-Seq – *fragments accessible to enzymatic cutting by DNase-I***



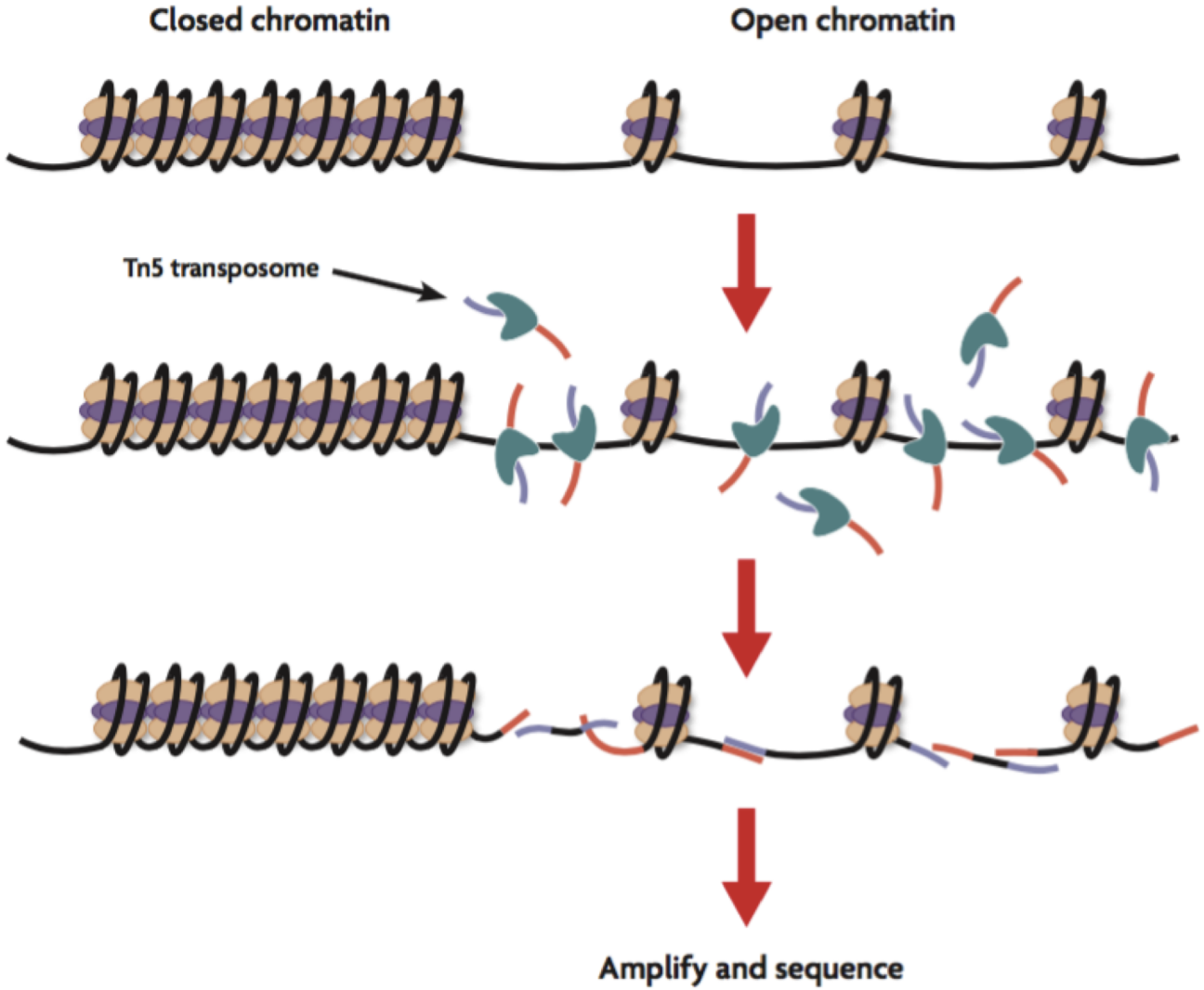
Quantifying DNA: next-generation sequencing

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

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

→Reads correspond to:

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



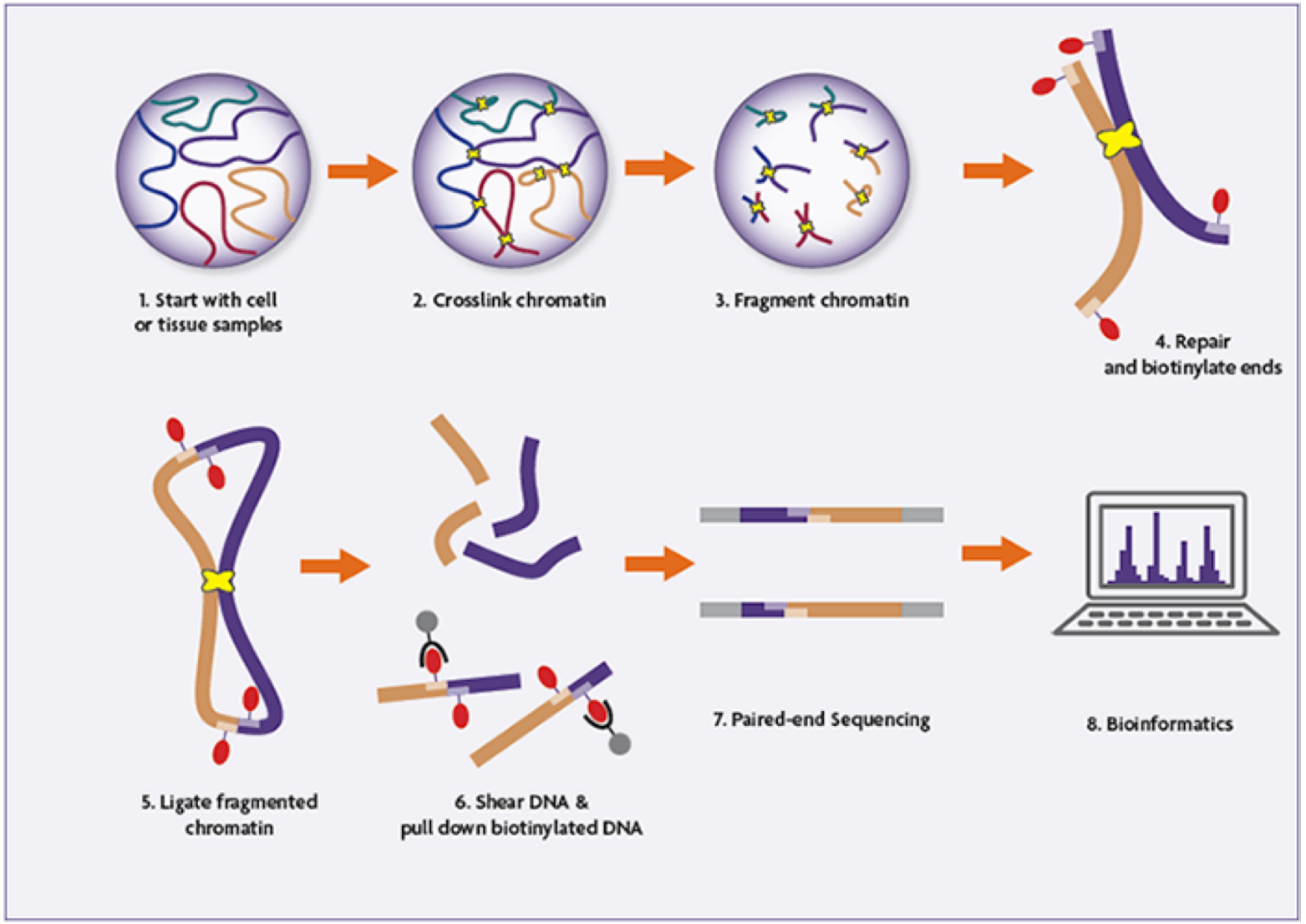
Quantifying DNA: next-generation sequencing

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

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

→Reads correspond to:

- ChIP-Seq - *fragments pulled down with antibody against a DNA binder*
- DNAse-Seq – *fragments accessible to enzymatic cutting by DNase-I*
- ATAC Seq – *fragments accessible to Tn5 Transposase activity*
- **Hi-C or chromatin capture – *fragments close to each other***



Quantifying DNA: next-generation sequencing

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

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

→Reads correspond to:

- ChIP-Seq - *fragments pulled down with antibody against a DNA binder*
- DNAse-Seq – *fragments accessible to enzymatic cutting by DNase-I*
- ATAC Seq – *fragments accessible to Tn5 Transposase activity*
- Hi-C or chromatin capture – *fragments close to each other*

→Issues

- Reads can map to multiple places
- Amplification bias
- Repetitive elements in the genome could give erroneous results

Quantifying DNA: DNA sequences as input to CNNs

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/unbound, chromatin state

Quantifying DNA: DNA sequences as input to RNNs

DNA Sequences can be represented and processed in an "time series" context with RNNs.

Spoken Audio Time Series	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

Quantifying DNA: position weight matrices+more

Table 1: Starting sequences.

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

Quantifying DNA: position weight matrices+more

Table 1: Starting sequences.

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

Table 2: Position Count Matrix.

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

Quantifying DNA: position weight matrices+more

Table 1: Starting sequences.

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

Table 2: Position Count Matrix.

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

Table 3: Position Probability Matrix.

Position	1	2	3	4	5	6
A	1.00	0.67	0.00	0.83	0.83	0.66
C	0.00	0.00	0.33	0.00	0.00	0.00
G	0.00	0.00	0.50	0.00	0.00	0.00
T	0.00	0.33	0.17	0.17	0.17	0.33

N = letter in set [A, T, C, G]
C = counts

$$PPM(N) = \frac{C_N}{\sum C}$$

Quantifying DNA: position weight matrices+more

Table 1: Starting sequences.

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

Table 2: Position Count Matrix.

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

Table 3: Position Probability Matrix.

Position	1	2	3	4	5	6
A	1.00	0.67	0.00	0.83	0.83	0.66
C	0.00	0.00	0.33	0.00	0.00	0.00
G	0.00	0.00	0.50	0.00	0.00	0.00
T	0.00	0.33	0.17	0.17	0.17	0.33

Table 4: Position Probability Matrix with a pseudocount 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

$$PPMp(N) = \frac{C_N + \frac{p}{n}}{\sum C + p}$$

p = pseudocount (usually 1)
n = # of letters

Quantifying DNA: position weight matrices+more

Table 5: Position Weight Matrix.

Position	1	2	3	4	5	6
A	2	1.425	-Inf	1.737	1.737	1.415
C	-Inf	-Inf	0.415	-Inf	-Inf	-Inf
G	-Inf	-Inf	1.000	-Inf	-Inf	-Inf
T	-Inf	0.415	-0.585	-0.585	-0.595	0.415

Table 6: Position Weight Matrix with a pseudocount of 1.

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

$$S(N) = \log_2 \left(\frac{PPM(C_N)}{B_N} \right)$$

B = background frequency matrix --> assume $B_N = 0.25$

Quantifying DNA: position weight matrices+more

Table 5: Position Weight Matrix.

Position	1	2	3	4	5	6
A	2	1.425	-Inf	1.737	1.737	1.415
C	-Inf	-Inf	0.415	-Inf	-Inf	-Inf
G	-Inf	-Inf	1.000	-Inf	-Inf	-Inf
T	-Inf	0.415	-0.585	-0.585	-0.595	0.415

Table 6: Position Weight Matrix with a pseudocount of 1.

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

Table 7: Information Content Matrix.

Position	1	2	3	4	5	6
A	2.000	0.721	0.000	1.125	1.125	0.721
C	0.000	0.000	0.180	0.000	0.000	0.000
G	0.000	0.000	0.270	0.000	0.000	0.000
T	0.000	0.361	0.090	0.225	0.225	0.361

Ask: are some positions more important than others?

Quantify the total amount of information at each position – AKA the level of conservation

https://bioconductor.org/packages/release/bioc/vignettes/universalmotif/inst/doc/IntroductionToSequenceMotifs.pdf

Quantifying DNA: position weight matrices+more

Table 3: Position Probability Matrix.

Position	1	2	3	4	5	6
A	1.00	0.67	0.00	0.83	0.83	0.66
C	0.00	0.00	0.33	0.00	0.00	0.00
G	0.00	0.00	0.50	0.00	0.00	0.00
T	0.00	0.33	0.17	0.17	0.17	0.33

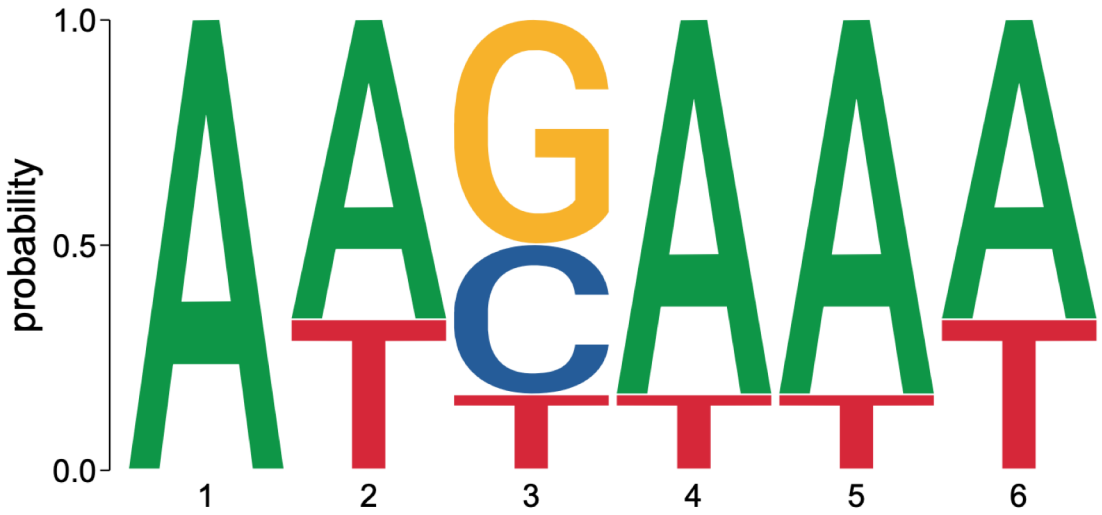


Figure 1: Sequence logo of a Position Probability Matrix

Table 7: Information Content Matrix.

Position	1	2	3	4	5	6
A	2.000	0.721	0.000	1.125	1.125	0.721
C	0.000	0.000	0.180	0.000	0.000	0.000
G	0.000	0.000	0.270	0.000	0.000	0.000
T	0.000	0.361	0.090	0.225	0.225	0.361

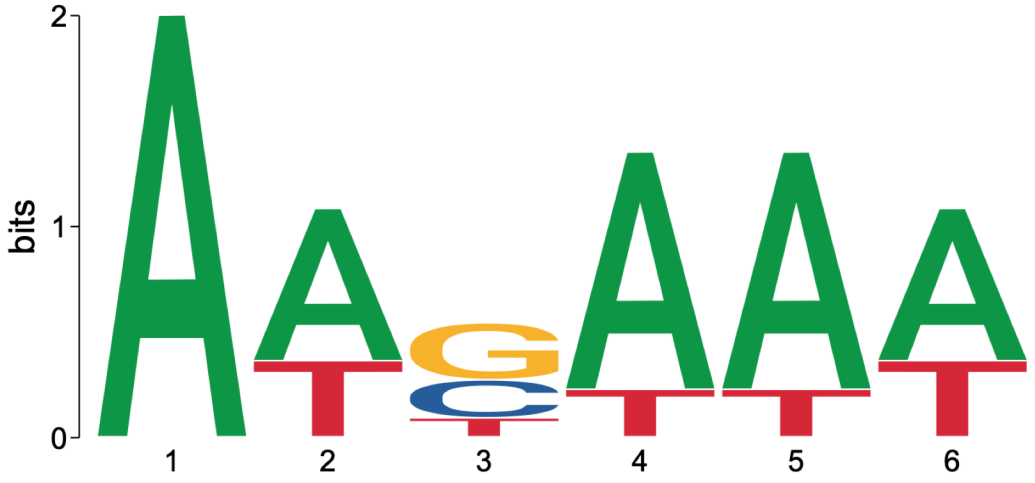


Figure 2: Sequence logo of an Information Content Matrix

Next week

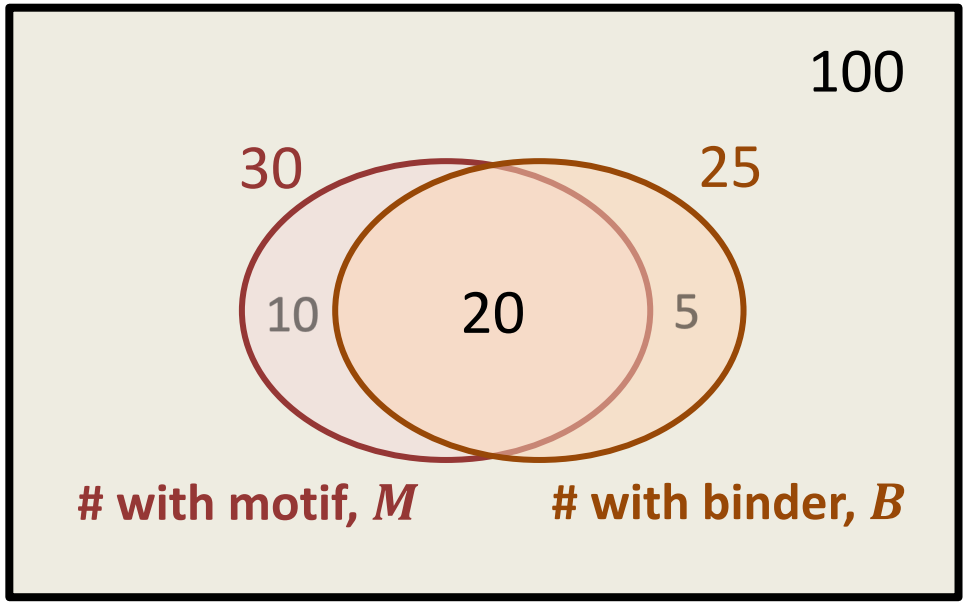
Transcription factors

DNA methylation

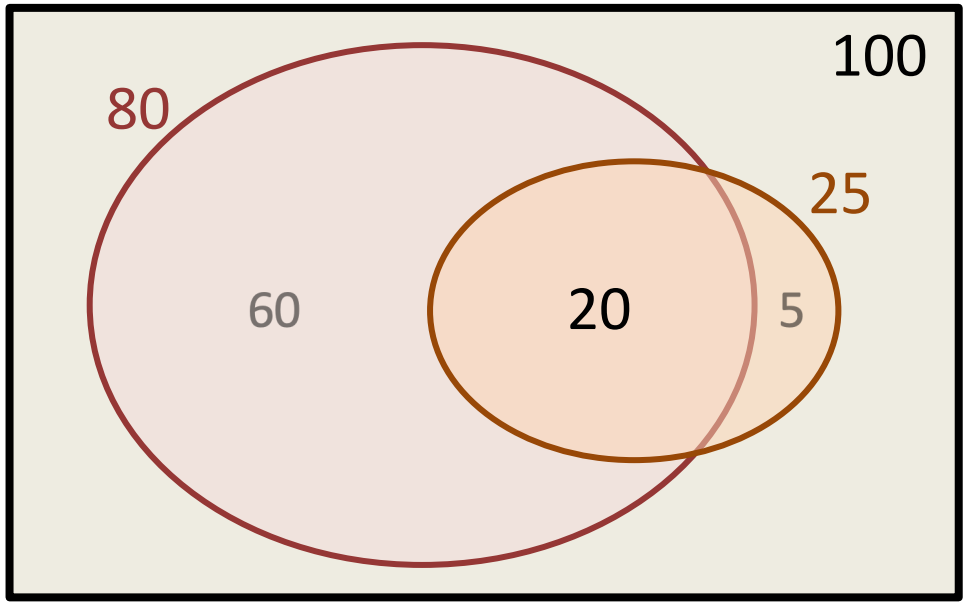
Gene expression & splicing

Quantifying DNA: hypergeometric distribution

The hypergeometric distribution allows us to calculate probabilities of enrichment.



total # sequences, T



$$P_{null} = \frac{\binom{M}{x}\binom{T-M}{B-x}}{\binom{T}{B}} = \frac{\binom{30}{20}\binom{100-30}{25-20}}{\binom{100}{25}} = 1.5\times10^{-9}$$

$$p = P_{null}(x \geq 20) = 2.0\times10^{-9}$$

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

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