

Introduction to molecular biology: Gene structure and control



A/Prof Scott Beatson



s.beatson@uq.edu.au



BIOC6000/SCIE2100

DNA as a blueprint

- *genetic* information is stored in the sequence of nucleotides in each DNA molecule (ATCG) → single DNA molecule can contain millions of bases → each cell can have multiple DNA molecules
- DNA content of a cell is called *genome* (usually nuclear DNA is implied)
- in eukaryotic cells, DNA is also stored in mitochondria and chloroplasts (organelles)
- genome size ranges from a few million nucleotides in bacteria to billions in multicellular organisms
- human genome contains 3.2 billion bases divided between 23 pairs of DNA molecules of different length

genomes

- billions of nucleotides → how is this information flows to proteins?
- genes → DNA segments transcribed into RNA → *gene expression*
- genome consists of coding (*genes*) and non-coding parts (also know for a very long time as “junk” DNA)
- more of “junk” DNA found in complex organisms
- human genome contains 3.2 billion nucleotides, but only 5% is protein-coding
- great similarity between people and mice protein-coding regions (85%) → more differences in non-coding areas (up to 50%) → even closer resemblance to chimpanzees and bonobo monkeys (99%)

<https://genome.ucsc.edu/ENCODE/>

<https://www.genome.gov/>

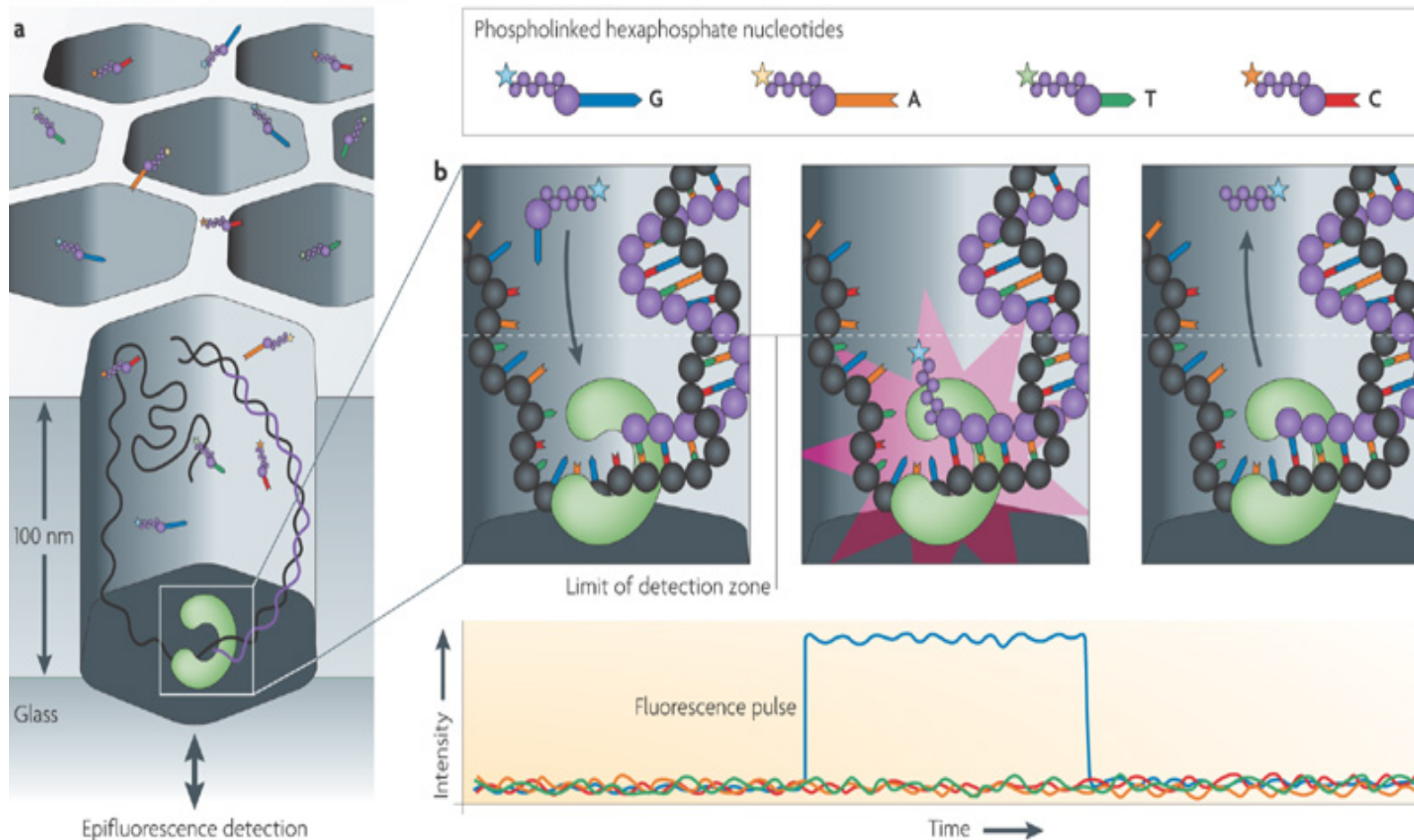


DNA modifications

- damage as a result of hostile environment (UV radiation, aggressive chemicals, etc...)
- changes in the sequence → mutations, insertions, deletions of nucleotides
- cells have elaborate replication and repair machinery to preserve DNA
- over time, changes in the DNA bases accumulate (slowly, as in the case of evolution, or rapidly, as in the case of cancer)
- different mechanisms of changing DNA and increasing the versatility of sequence


RESEARCH UPDATE: determining the “methylome” with SMRT* sequencing

Pacific Biosciences — Real-time sequencing



Nature Reviews | Genetics

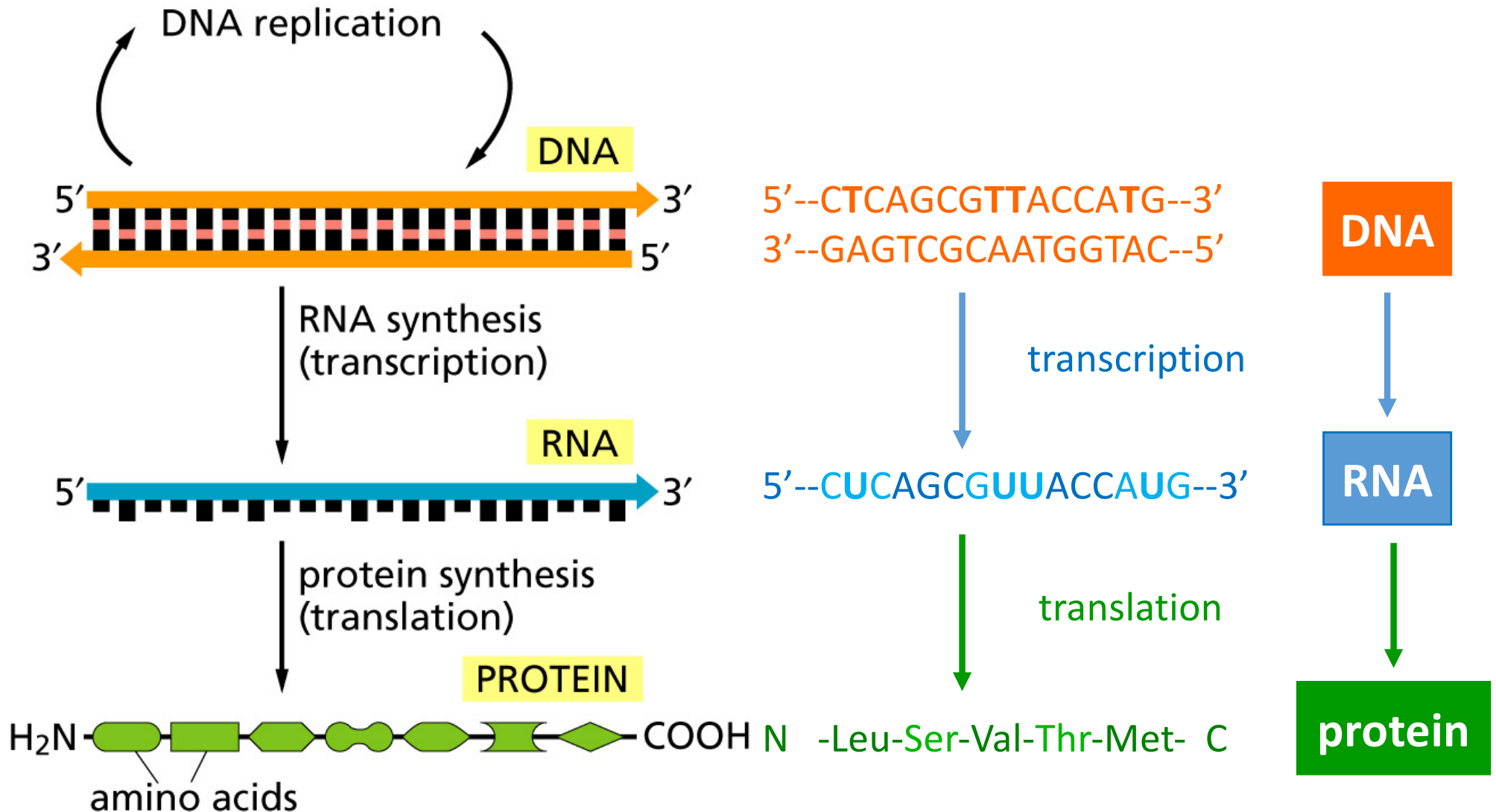
Metzker, *Nature Reviews Genetics* **11**, 31-46 (2010), doi:10.1038/nrg2626
(*Single Molecule Realtime Sequencing with Pacific Biosciences “PacBio”)



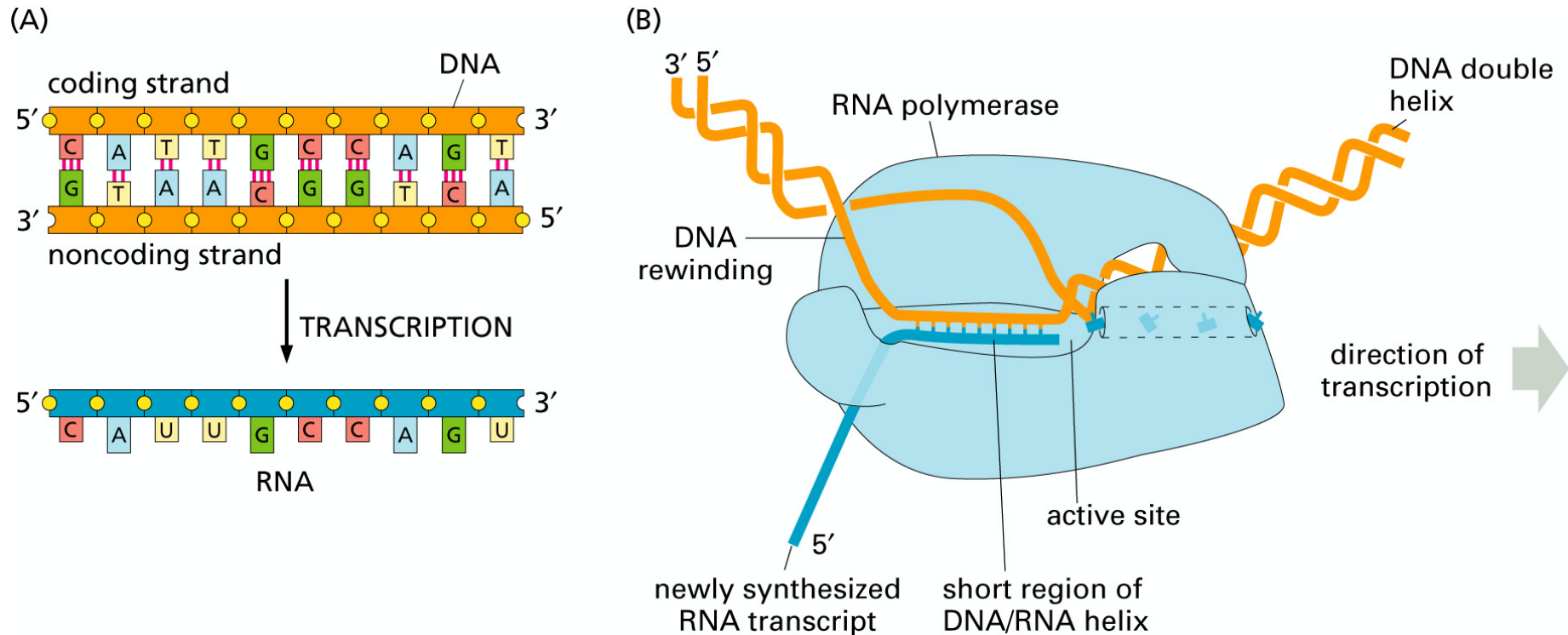
DNA is the information store

- DNA carries the blueprints how to build and maintain the living cell → directions for protein synthesis
- proteins are key ingredients of every cell playing important role in various aspects of cell functioning (*structure, transport, metabolism, catalysis, signalling, etc.*)
- translation of DNA to proteins follows the same basic scheme in ALL the cells → *central dogma of molecular biology*

central dogma of molecular biology



gene expression: transcription of genes to messenger RNA (mRNA)



- transcription → noncoding strand of DNA serves as a template for synthesis of so-called messenger RNA (mRNA) → process catalysed by RNA polymerase
- resulting mRNA has the same sequence as the coding DNA strand, except that thymine (T) is replaced with uracil (U)

Transcription video

<https://www.youtube.com/watch?v=5MfSYnltYvg>



what is a gene?

- definition of gene changed over the years as the knowledge increases:
 - an abstract concept where gene was defined as a unit of inheritance that ferried a characteristic from parent to child
 - development of biochemistry led to one gene – one protein relationship
 - advances in molecular biology turned genes into real, physical things – DNA sequences which can be converted into RNA, which in turn can lead to protein synthesis
 - a gene is a **heritable** string of nucleotides that can be transcribed, creating a molecule with biological activity
- protein is not necessarily the end product of gene transcriptions → sometimes it is RNA (i.e. “non-coding RNA” plays various roles in the cell)

<http://www.nature.com/nature/journal/v441/n7092/full/441398a.html>

<http://www.sciencemag.org/content/316/5831/1556.full>

<http://scienceblogs.com/digitalbio/2007/01/21/what-is-a-gene-my-definition-i>

overlapping genes

- overlapping genes → compact information packing → occurrence in viral and prokaryotic genomes, mitochondrial DNA, but surprisingly large number found in mammalian genomes



Veeramachaneni et al. Genome Res. 2004;14:280-286

- overlap between three human genes: MUTH, FLJ13949, and TESK2 (grey parts are not translated)
- complex mechanism of control and regulation of gene expression

<http://genome.cshlp.org/content/14/2/280.long>

<http://www.nature.com/nature/journal/v457/n7232/full/nature07728.html>



Coronavirus update: overlapping genes

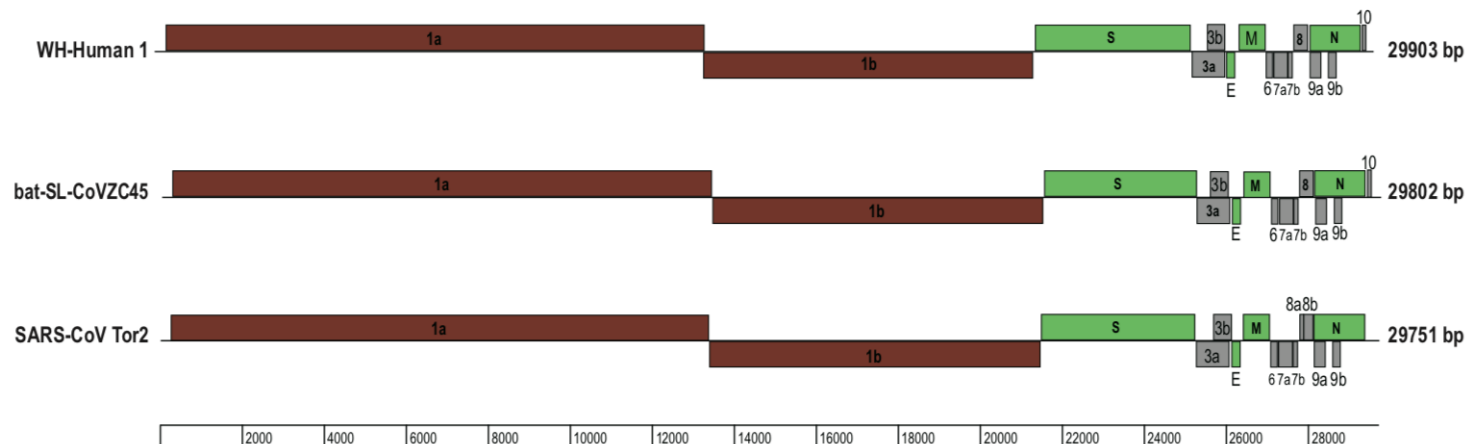


Figure 2. Genome organization of SARS and SARS-like CoVs including Tor2, CoVZC45 and WHCV determined here.

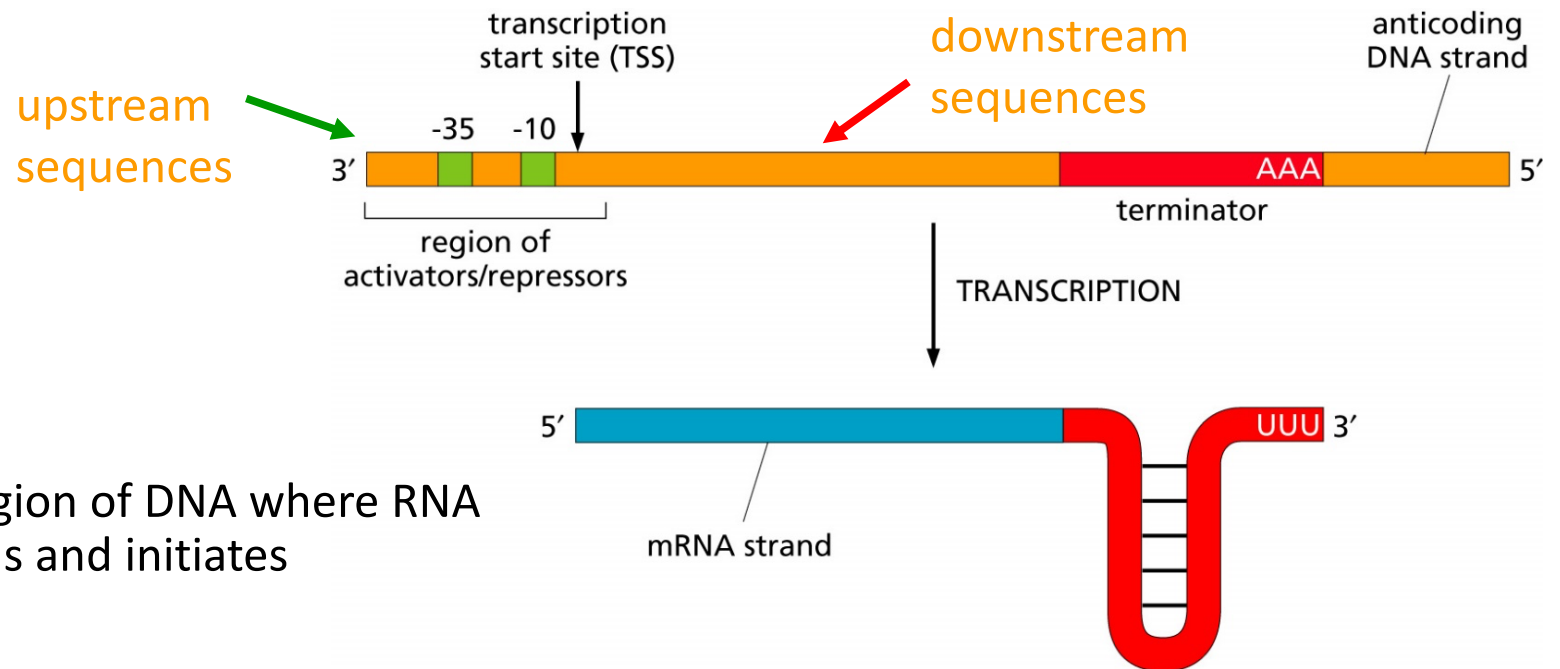
there is
more to
DNA than
just
genes...

- genomic DNA contains more than just the protein coding sequences:
 - control sequences (promoter sequences, stop codons, etc);
 - coding sequences (actual genes);
 - regulatory sequences (how much of the gene to express)
- ... and we still don't understand the huge part of it
- in humans, only 5% of protein-coding genes and for a lot of them, gene-protein-function relationship is unknown
→ bioinformatics

control and regulation

- gene expression is carefully regulated within a cell → expression dictates the cellular function
- control of transcription and translation rules the amount of every protein in the cell
- difference between prokaryotic and eukaryotic genomes → more complex mechanisms present in the eukaryotic cells
- eukaryotes have more complicated gene structure, transcription takes place within a nucleus physically separated from ribosomes (not the case in prokaryotes), multiple regulatory points and more regulatory proteins involved

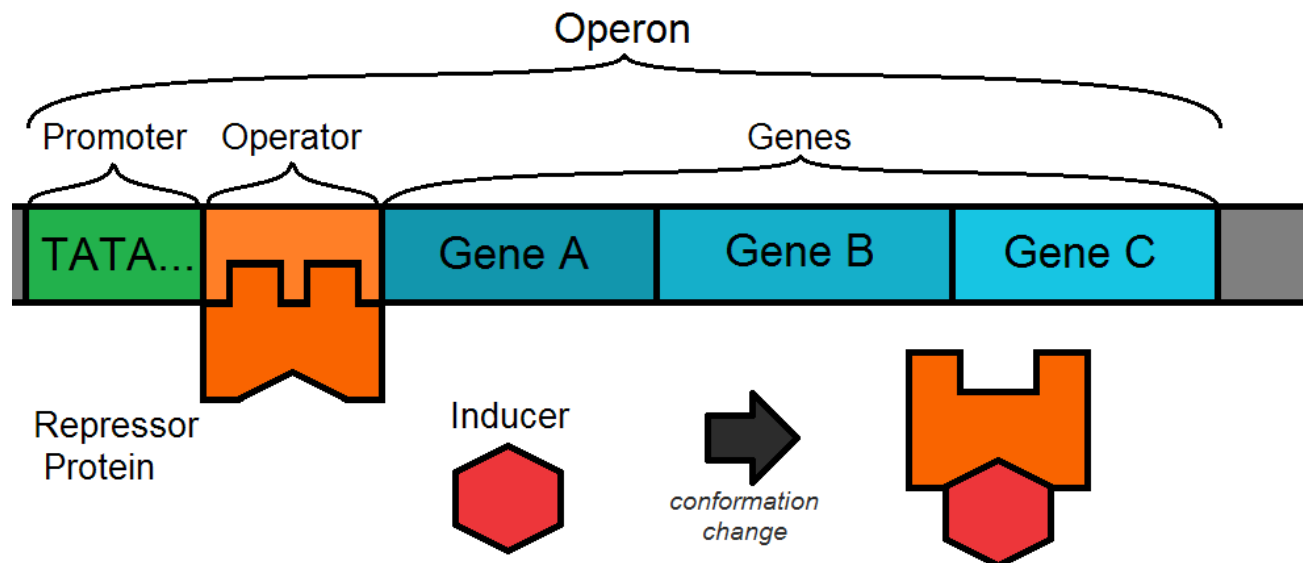
transcription control in bacteria



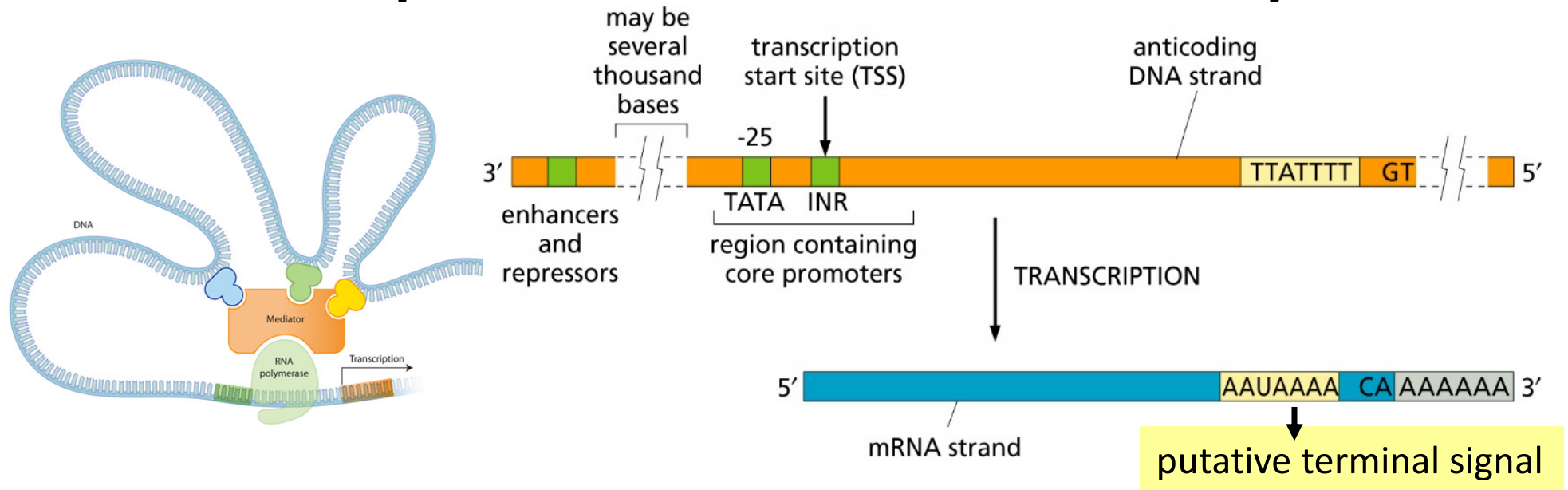
- *promoters* → region of DNA where RNA polymerase binds and initiates transcription
- problem of finding a promoter in DNA sequences due to variations → *consensus sequences* (most frequent bases at these positions) in *E. coli* at -10 TATAAT and at -35 TTGACA → tighter binding of polymerase means more frequently the region is being transcribed
- terminator sequences: two short stretches of complementary bases to form double helix + min. 3 consecutive U nucleotides
- additional controls → *activator* (improve efficiency) and *repressor* proteins (bind to operator site to block promoter sites) → crucial role in regulation
- DNA modifications (e.g. methylation) can prevent activator/repressor binding

Bacterial operons and gene expression

- Shine-Dalgarno sequence (consensus AGGAGGU) is located a few bases upstream from starting codon (AUG) → indicates ribosome binding site
- operons → clusters of functionally related protein-coding sequences transcribed as a single mRNA → specific for prokaryotes, rare in eukaryotes
- specific proteins translated separately → only one control region required to activate expression of several genes (i.e. for metabolic pathways)



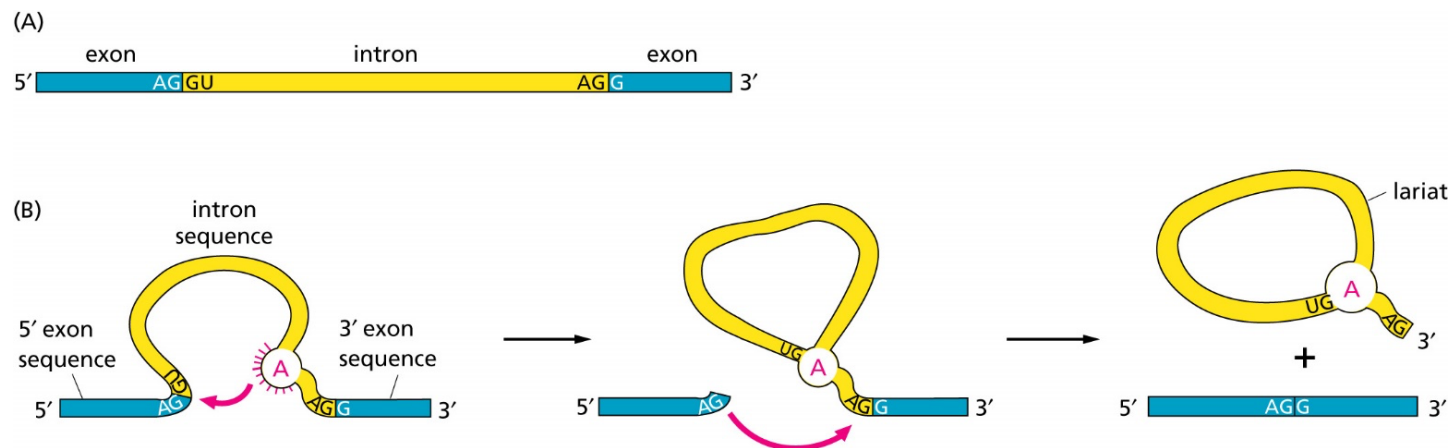
transcription control in eukaryotes



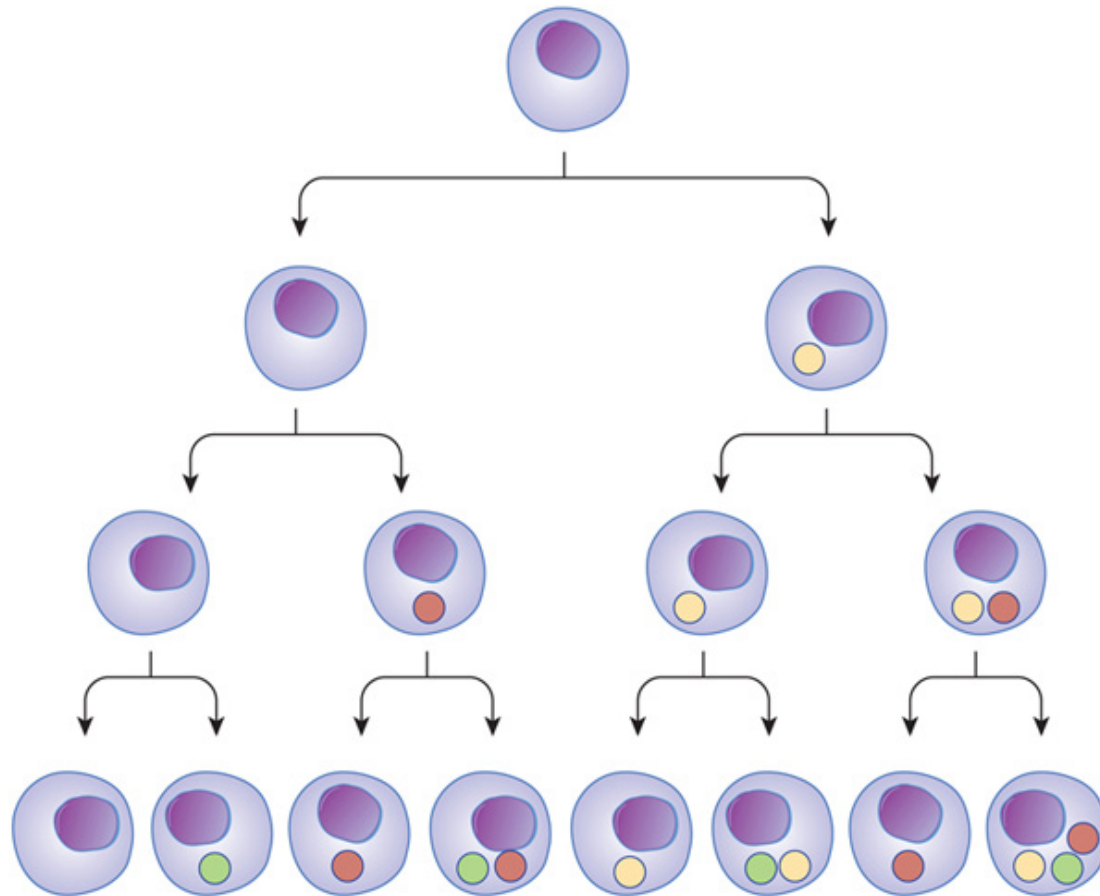
- three types of RNA polymerases → RNA polymerase II catalyses mRNA synthesis (types I and III are related to transcription of genes for tRNA, rRNA and other RNA molecules, regulated with different promoters)
- a set of core promoter signals in TSS region → binding of general transcription initiation factors → TATA box (TATA binding protein, TBP)
- regulatory regions (enhancers and silencers) controlling transcription can be far away from TSS in eukaryotes and both upstream and downstream → DNA loops to bring all the important regulatory proteins together into complex that controls transcription

mRNA modification in eukaryotes

- original mRNA transcript undergoes several modifications before translation in eukaryotes:
 - RNA capping – addition of modified G nucleotide to 5' end (role in ribosome binding and translation)
 - polyadenylation of 3' end after cleaving mRNA triggered by AAUAAA signal (~ 200 polyA chain)
 - RNA splicing – excision of *introns* (noncoding stretches of DNA) and merging *exons* (protein-coding sequences) → process carried out by spliceosome (*small nuclear* RNA + proteins)
- alternative splicing → different ways of merging the same exons resulting in greater protein variability from smaller number of genes

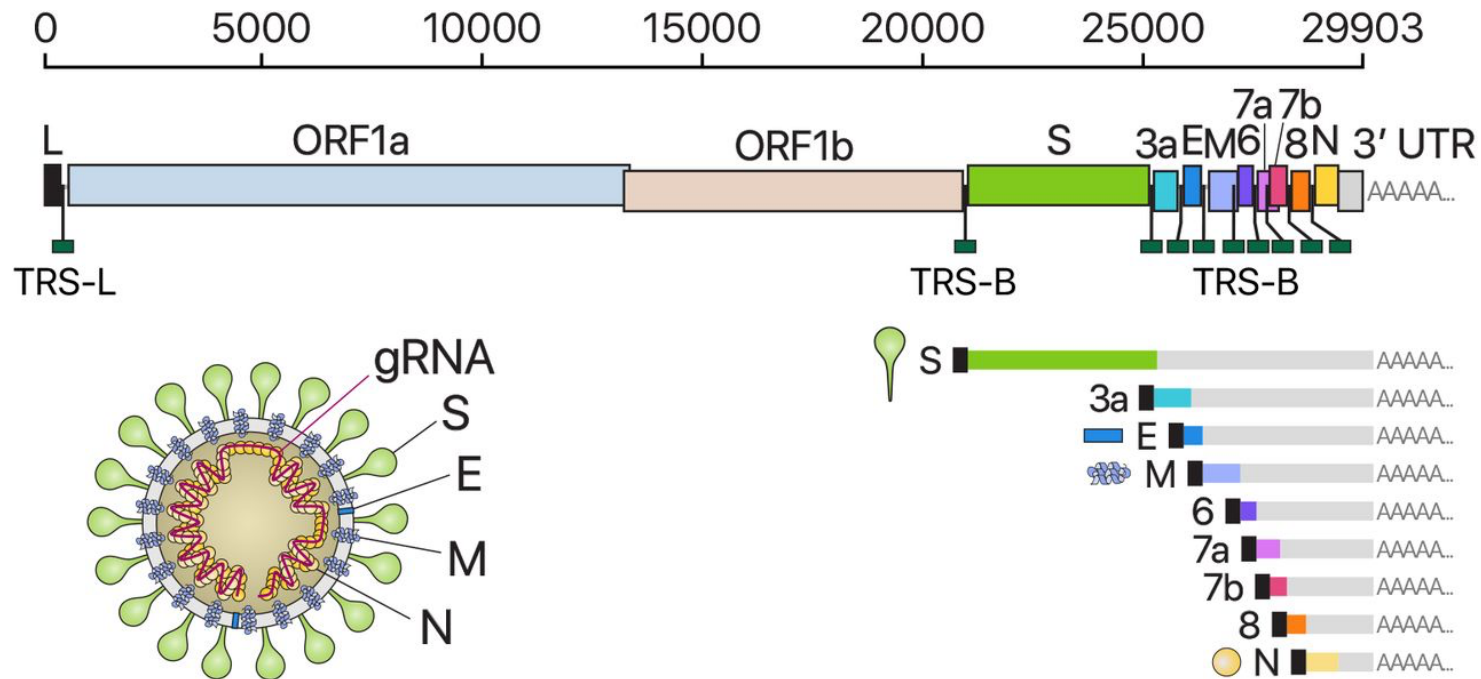


gene expression determines cell type



- the wide variety of cell types in a single organism can depend on different transcription factor activity in each cell type
- different transcription factors can turn on at different times during successive generations of cells resulting ultimately with different cell types

Coronavirus update: gene regulation



Kim et al., Cell 2020, DOI:<https://doi.org/10.1016/j.cell.2020.04.011>

some useful links:

ONLINE TEXTBOOK ON CELL BIOLOGY:

<http://www.nature.com/scitable/ebooks/essentials-of-cell-biology-14749010/contents>

CENTRAL DOGMA ANIMATION VIDEOS:

<http://www.youtube.com/watch?v=J3HVV2k2No>

<http://www.youtube.com/watch?v=ZNcFTRX9i0Y>

TRANSCRIPTION VIDEO:

<https://www.youtube.com/watch?v=5MfSYnltYvg>

Key learning outcomes

- Understand the basic steps in transcription
- Recognize how are genes and genomes are structured.
- Understand the key features of an operon and how it functions in bacterial gene control.
- Appreciate how functional mRNA molecules produced in eukaryote organisms.

Refer to Week 1-2 Study Guide document