CS273B: Deep learning for Genomics and Biomedicine

Lecture 2: Genomics 101

09/27/2017

Anshul Kundaje, James Zou

Outline

1

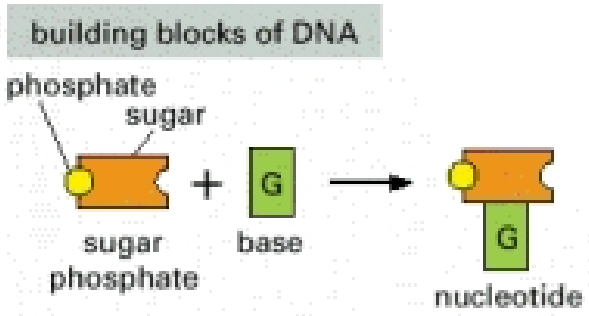
Anatomy of the human genome

2

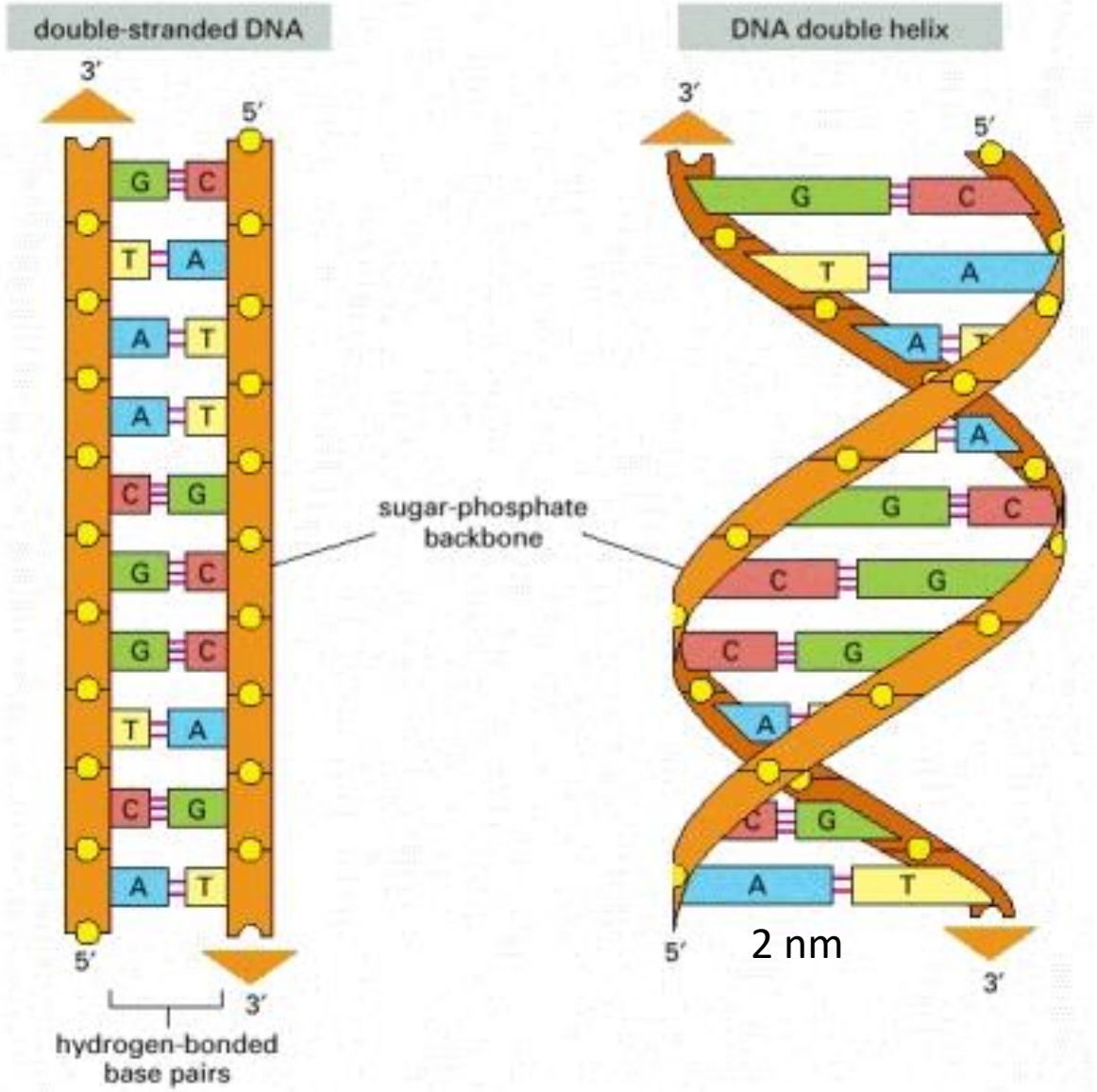
High throughput sequencing

Anatomy of the human genome

DNA: the molecule of heredity



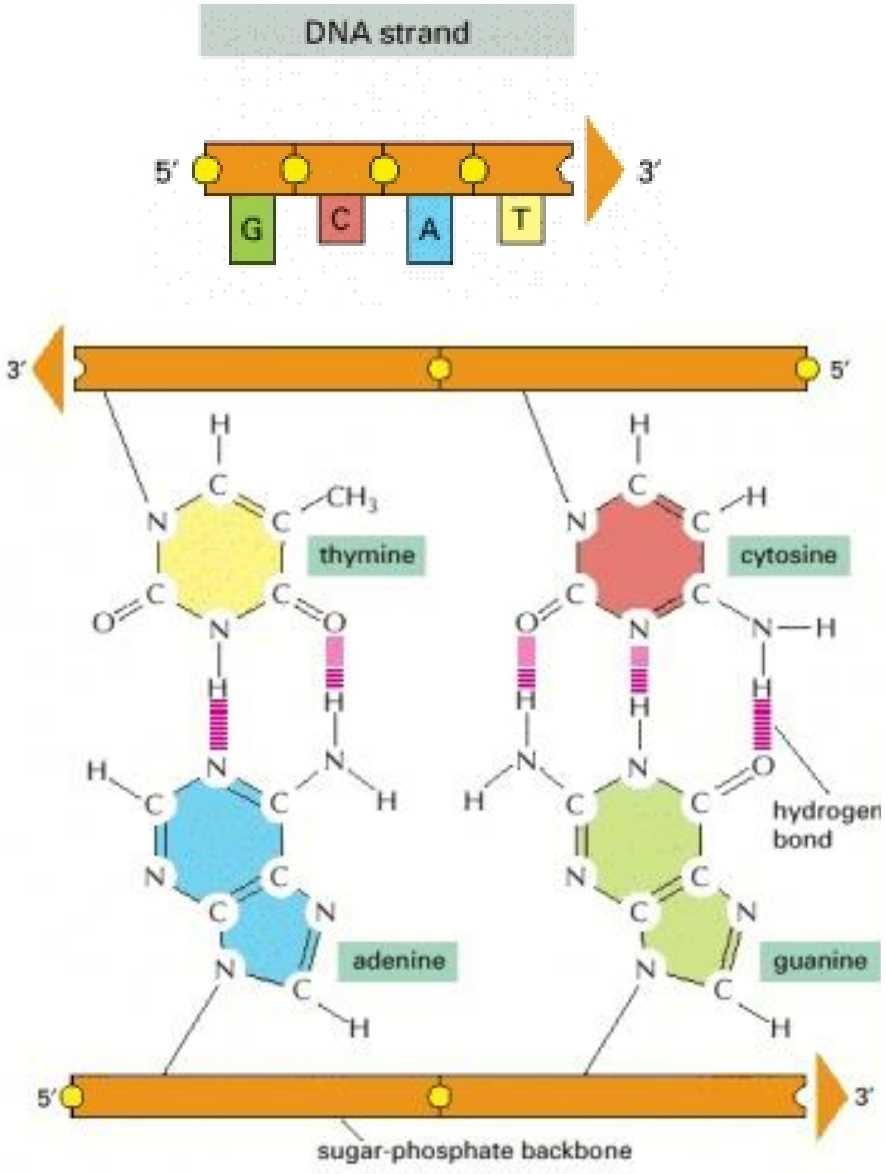
- DNA: Deoxyribose nucleic acid
- Double stranded biopolymer
- Canonical form: right handed double helix
- Phosphate backbone outside
- 4 Bases / nucleotides hidden on the inside (A, C, G, T)
- A pairs with T
- C pairs with G
- 1 helical turn is 10.4 nucleotides



https://www.youtube.com/watch?v=o_-6JXLYS-k

Franklin, Watson, Crick (1953)

DNA is directional and has two complementary strands



- Weak hydrogen bonds hold the two strands together
- This allows low-energy opening and re-closing of two strands

- **Chemical Polarity:** Extension 5’→3’ tri-phosphate coming from newly added nucleotide

Forward strand

5’ → 3’

TGCCAAGCA

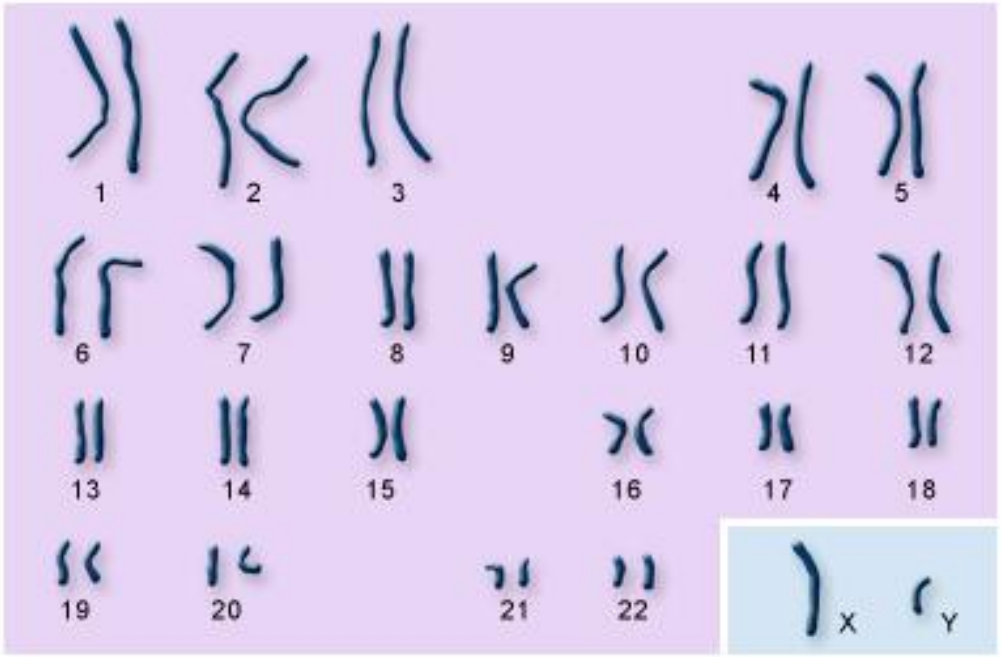
| | | | | | | | | |

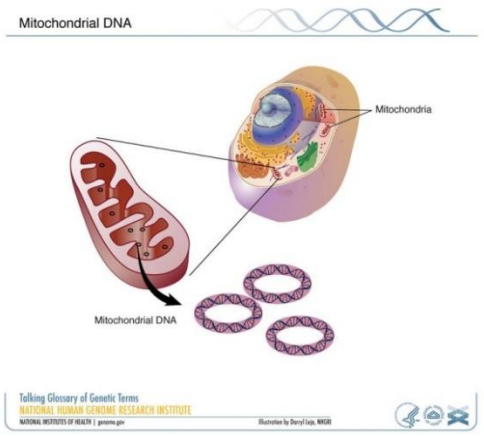
ACGGTTCGT

3’ ← 5’

Reverse complement

Chromosomes in humans





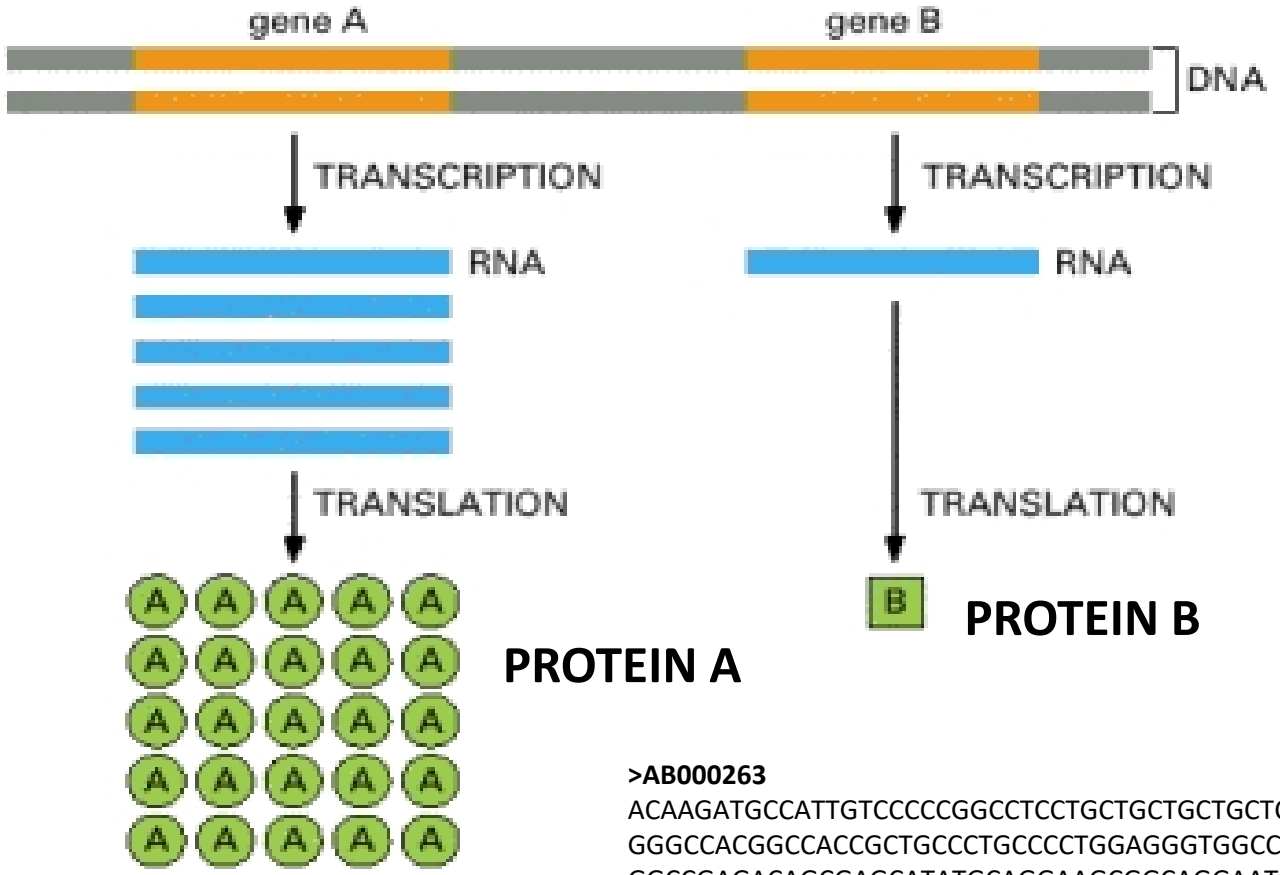
TGCCAAGCA
||||||||||
ACGGTTCGT

TGCCAAGCA
||||||||||
ACGGTTCGT

- Humans are diploid (2 copies of each chromosome)
- 22 pairs of autosomes
- Sex chromosomes: female (X,X) , male (X,Y)
- Mitochondrial DNA (circular, many copies per cell)
- Diploid Human genome = ~3 billion **base-pairs** X 2

Genes (DNA -> RNA -> protein)

- **The gene** is a fundamental functional unit of DNA whose sequence codes specific recipes to make other biomolecules (RNA and proteins)
  - In humans each DNA molecule ⇔ 25,000 protein-coding genes
  - Many other gene-like units encode short and long non-coding RNAs



Where to get genome and gene sequences
GENBANK: https://www.ncbi.nlm.nih.gov/genbank/
ENSEMBL: https://www.ensembl.org/index.html

>AB000263
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCG
GGGCCACGGCCACCGCTGCCCTGCCCCTGGAGGGTGGCCCCACC
GGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGG
AAAAGCAGCCTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACC
TCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGGAAGCTCGGG
AGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGC
GCGCCGGGACAGAATGCCCTGCAGGAACTTCTTCTGGAAGACCT
TCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAGT
TTAATTACAGACCTGAA

FASTA format

mRNA: The messenger

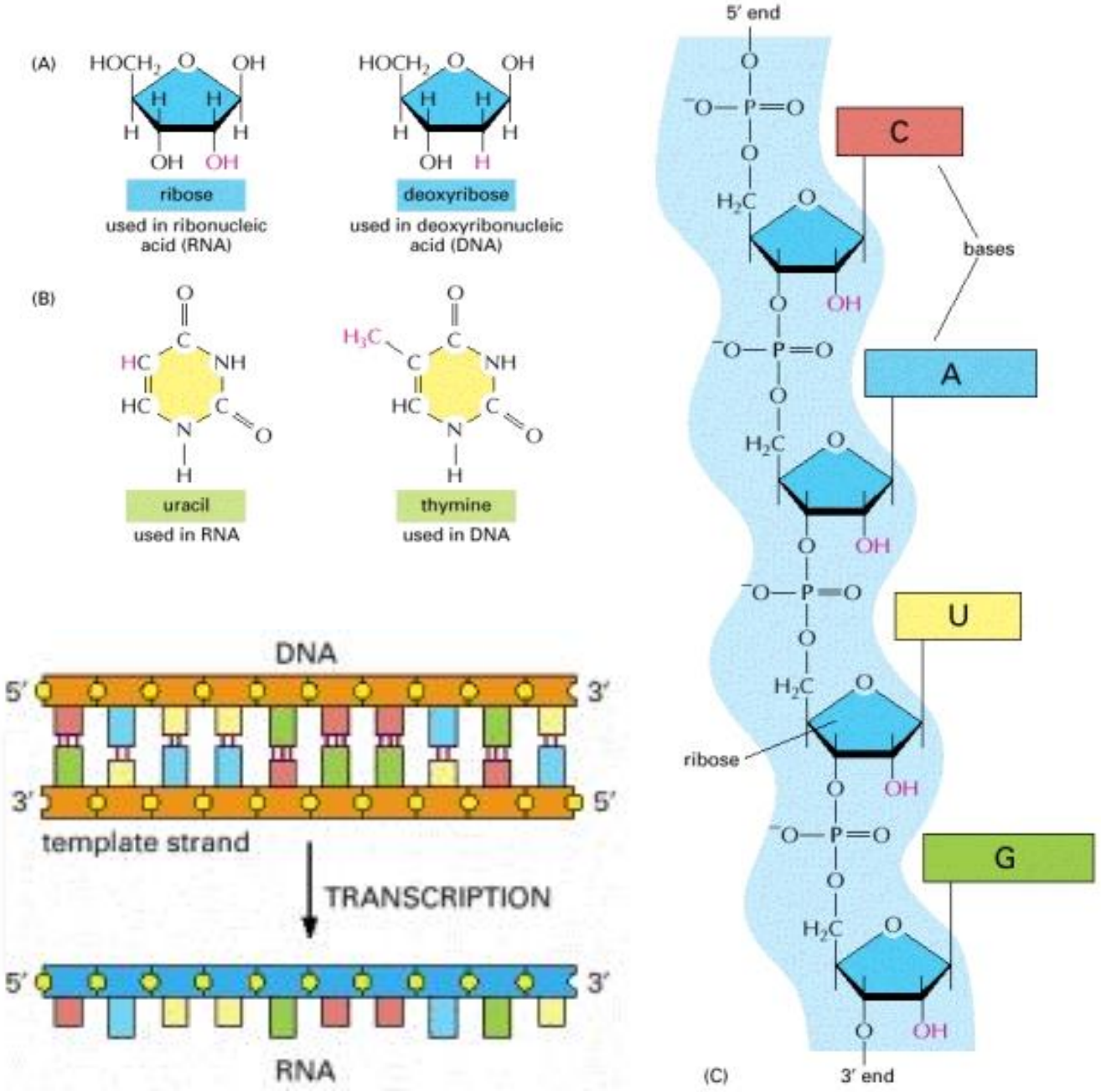
- Information changes medium
  - single strand vs. double strand
  - ribose vs. deoxyribose sugar
  - Uracil (U) instead of Thymine (T)

DNA sequence

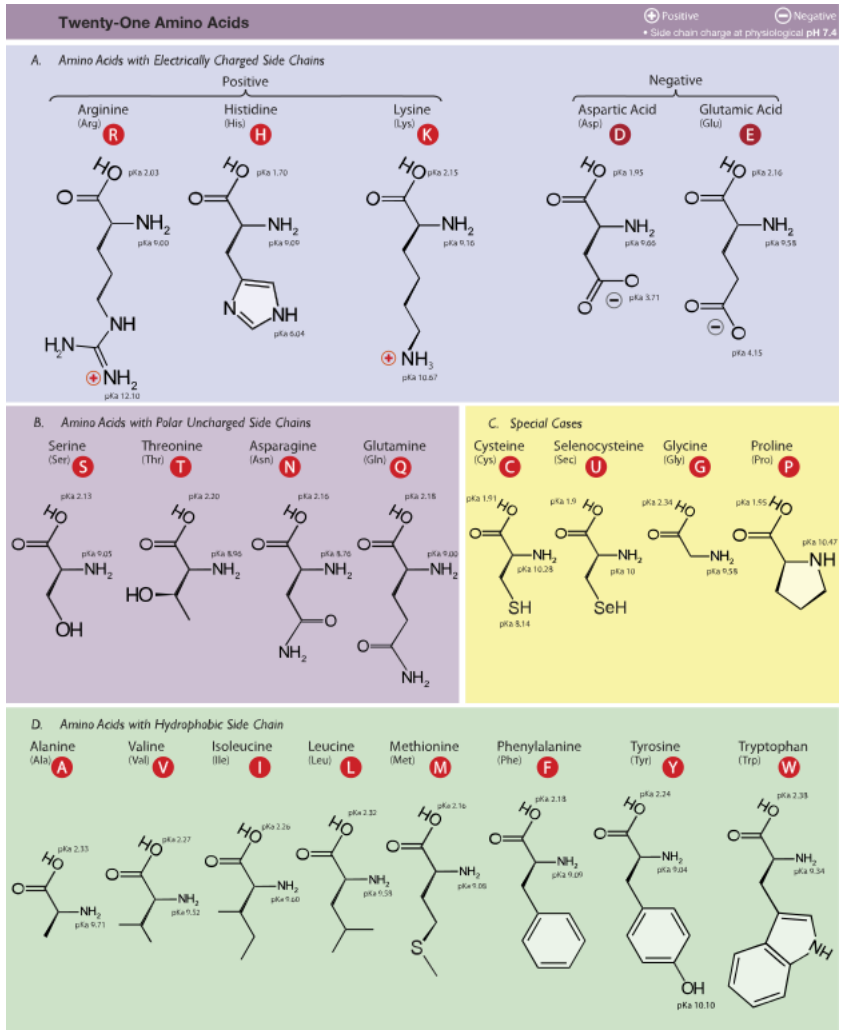
A T T A C G G T A C C G T

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

Corresponding RNA sequence

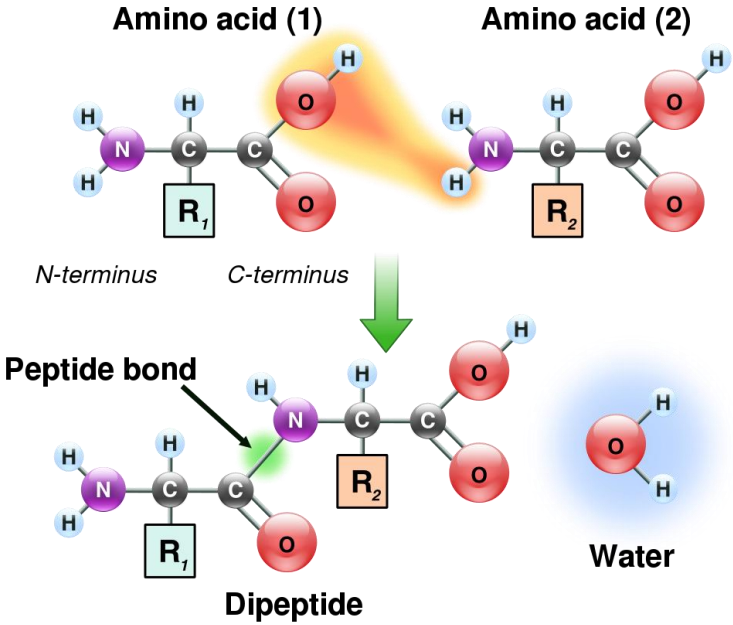


Proteins – chain of 20 (+2) amino acids



http://groundupstrength.wdfiles.com/local--files/amino-acids/amino-acids.gif

Amino acid	3-letter[132]	1-letter[132]
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V



FASTA format

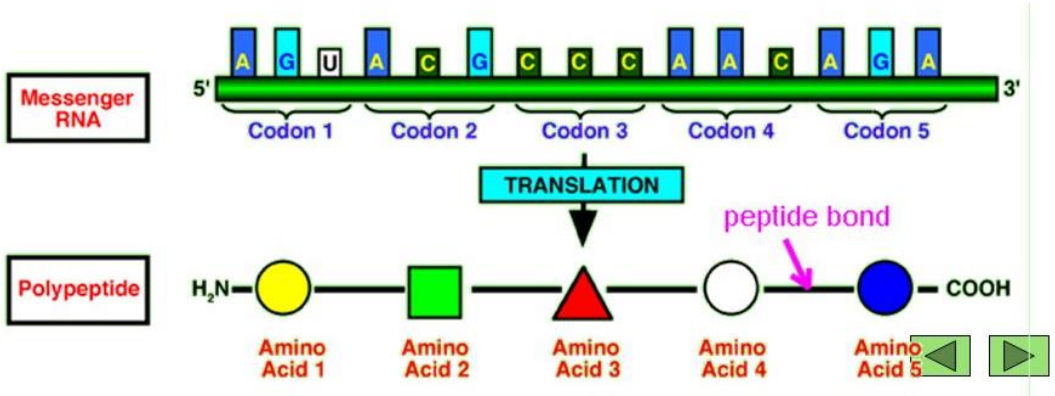
>DROME_HH_Q02936
MRHIAHTQRCLSRLTSLVALLLIVLPMVFSPAHSCGPGRGLGRHRARNLYPLVL
KQTIPNLSEYTNSASGPLEGVIRRDSPKFKDLVPNYNRDILFRDEEGTGADRLM
SKRCKEKLNVLAYSVMNEWPGIRLLVTESWDEDYHHGQESLHEGRAVTIATS
DRDQSKYGMLARLAVEAGFDWVSYVSRRHIYCSVKSDSSISSHVHGCFTPES
TALLESGVRKPLGELSIGDRVLSMTANGQAVYSEVILFMDRNLEQMQNFVQLH
TDGGAVLTVTPAHLVSVWQPESQKLTFVFADRIEEKNQVLVRDVETGELRPQR
VVKVGSVRSKGVVAPLTREGTIVVNSVAASCYAVINSQSLAHWGLAPMRLLST
LEAWLPAKEQLHSSPKVVSSAQQQNGIH WYANALYKVKDYVLPQSWRHD

Where to get protein sequences
UNIPROT: http://www.uniprot.org/

Translation (RNA->protein): The Genetic Code

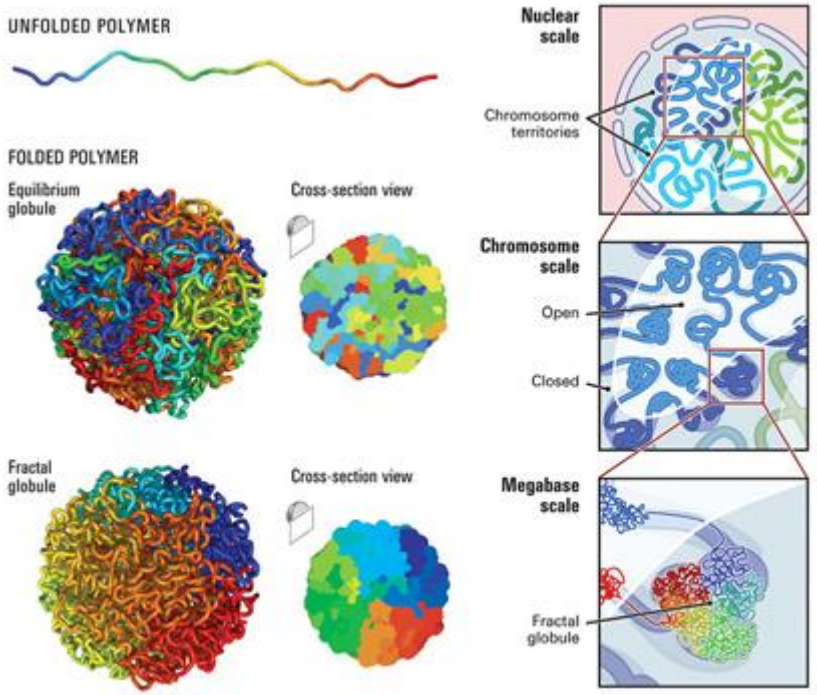
- A *triplet* of nucleic acids (*codon*) codes for one amino acid
- The code is redundant. E.g., both GGC and GGA code for Gly (Glycine)

First Base	Second Base: U	Second Base: C	Second Base: A	Second Base: G	Third Base
U	UUU Phenyl-alanine F	UCU Serine S	UAU Tryosine Y	UGU Cysteine C	U
	UUC Phenyl-alanine F	UCC Serine S	UAC Tryosine Y	UGC Cysteine C	C
	UUA Leucine L	UCA Serine S	UAA Stop codon	UGA Stop codon	A
	UUG Leucine L	UCG Serine S	UAG Stop codon	UGG Tryptophan W	G
C	CUU Leucine L	CCU Proline P	CAU Histidine H	CGU Arginine R	U
	CUC Leucine L	CCC Proline P	CAC Histidine H	CGC Arginine R	C
	CUA Leucine L	CCA Proline P	CAA Glutamine Q	CGA Arginine R	A
	CUG Leucine L	CCG Proline P	CAG Glutamine Q	CGG Arginine R	G
A	AUU Isoleucine I	ACU Threonine T	AAU Aspargine N	AGU Serine S	U
	AUC Isoleucine I	ACC Threonine T	AAC Aspargine N	AGC Serine S	C
	AUA Isoleucine I	ACA Threonine T	AAA Lysine K	AGA Arginine R	A
	AUG Methionine start codon M	ACG Threonine T	AAG Lysine K	AGG Arginine R	G
G	GUU Valine V	GCU Alanine A	GAU Aspartic acid D	GGU Glycine G	U
	GUC Valine V	GCC Alanine A	GAC Aspartic acid D	GGC Glycine G	C
	GUA Valine V	GCA Alanine A	GAA Glutamic acid E	GGA Glycine G	A
	GUG Valine V	GCG Alanine A	GAG Glutamic acid E	GGG Glycine G	G

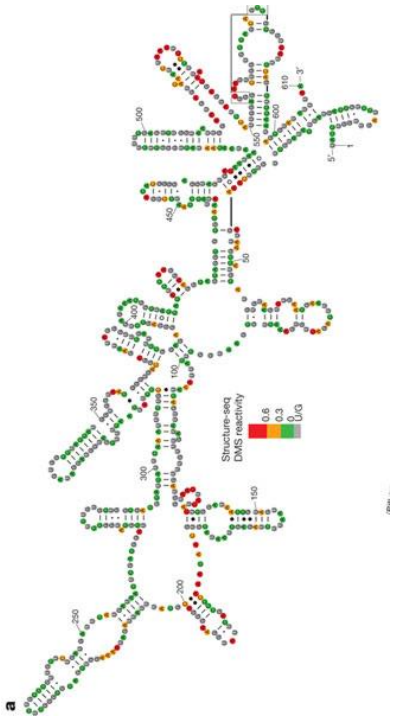


Picture: https://media1.shmoop.com/images/biology/biobook_dna_graphik_22.png

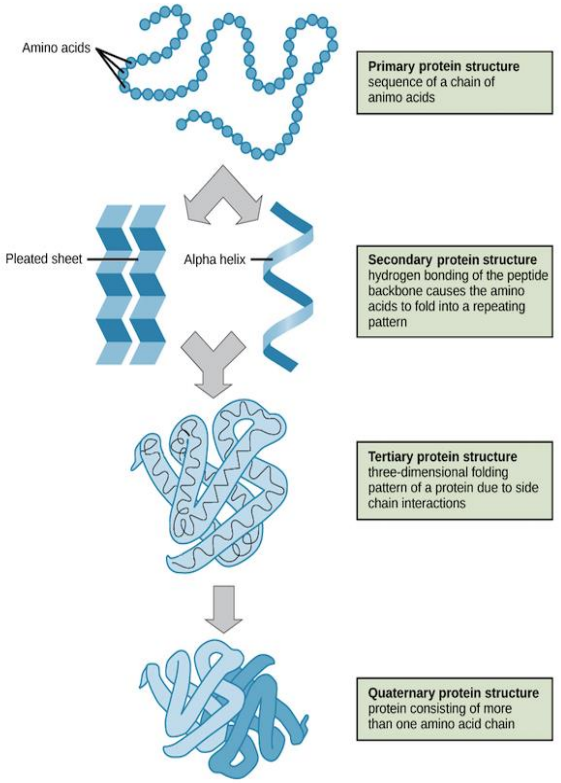
DNA, RNA and proteins form complex secondary and tertiary structures



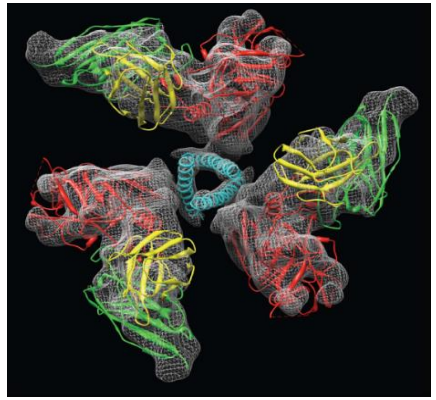
3D genome architecture
(DNA folding)



RNA secondary structure



https://ka-perseus-images.s3.amazonaws.com/71225d815cafcc09102504abdf4e10927283be98.png



https://www.fei.com/uploadedImages/FEISite/Pages/Markets/Life_Sciences/Structural_Biology/Sub-Cellular_Imaging/Sriram_001_465x.png

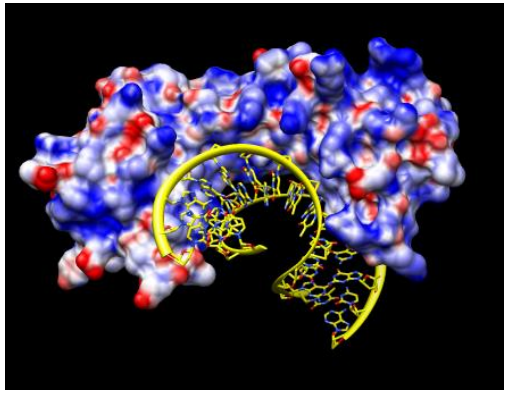
Protein structure

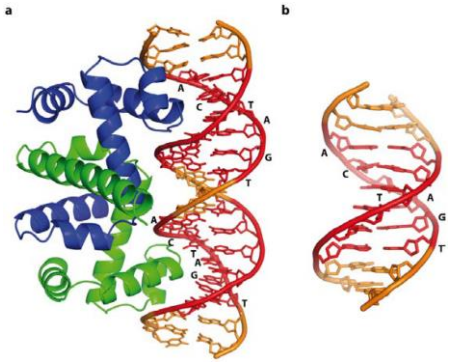
Major ML challenge: Predicting structure

Interactions between proteins, DNA, RNA and small molecules (drugs)

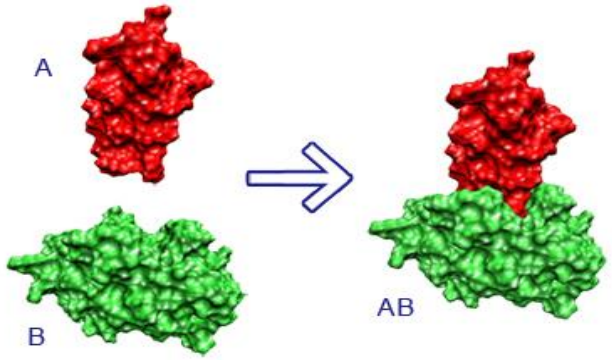
Major ML challenge: Predicting interactions between biomolecules

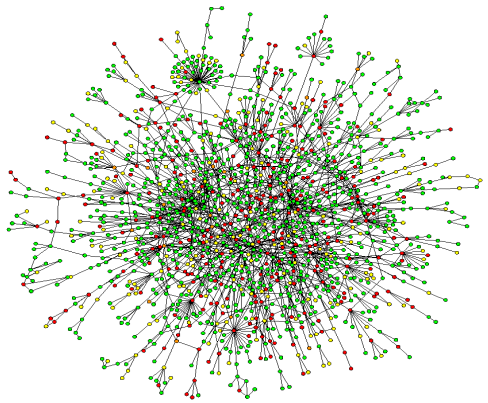
Protein-DNA interactions



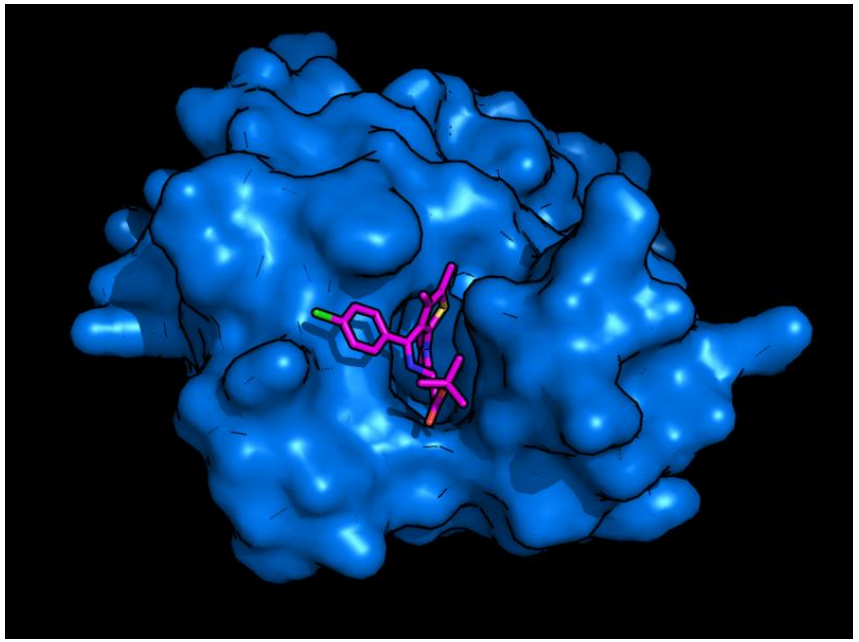


Protein-Protein interactions





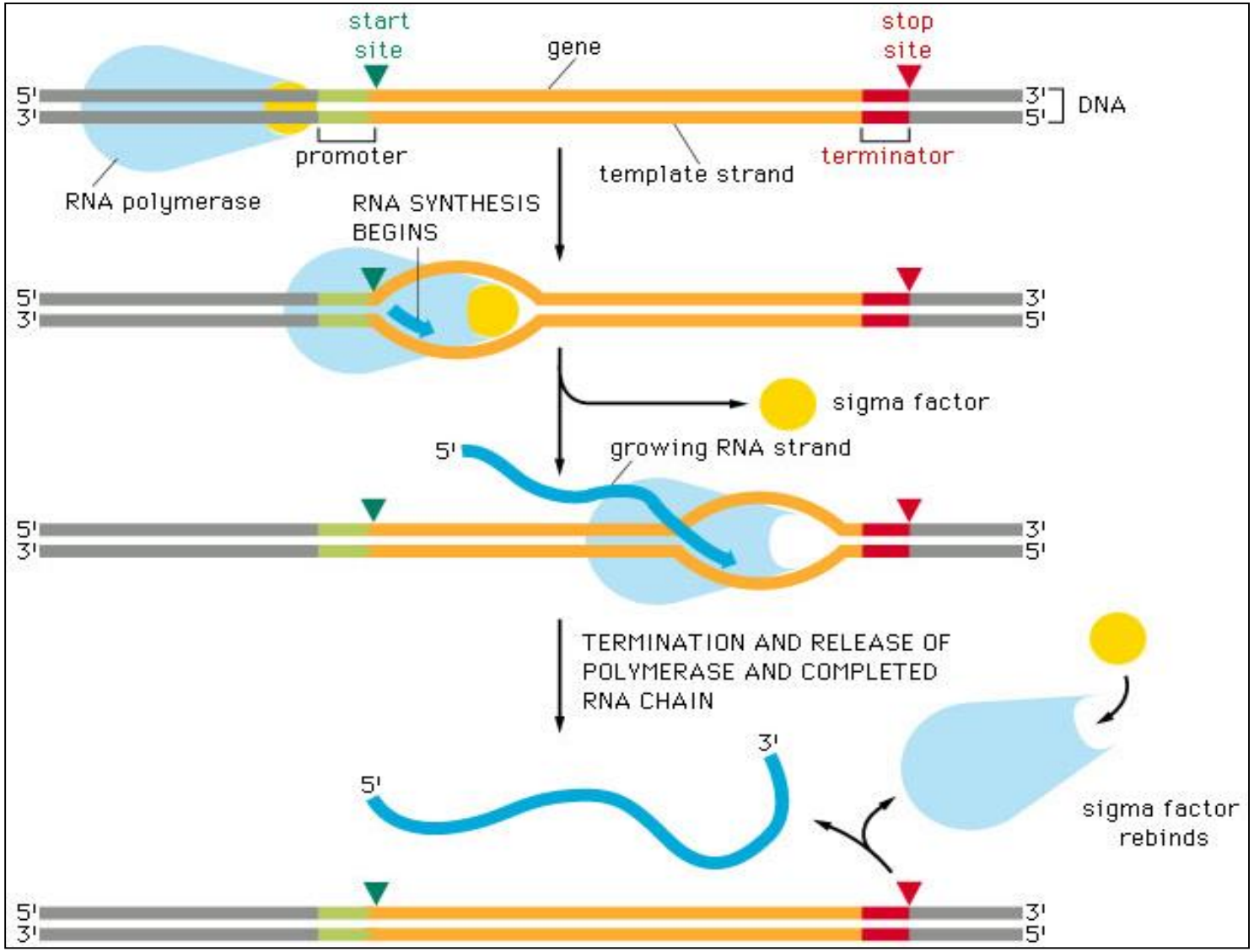
PPI networks



Protein-small molecule interactions

(Drug discovery, cheminformatics, Quantitative structure activity relationships (QSAR)

From DNA to RNA: Transcription



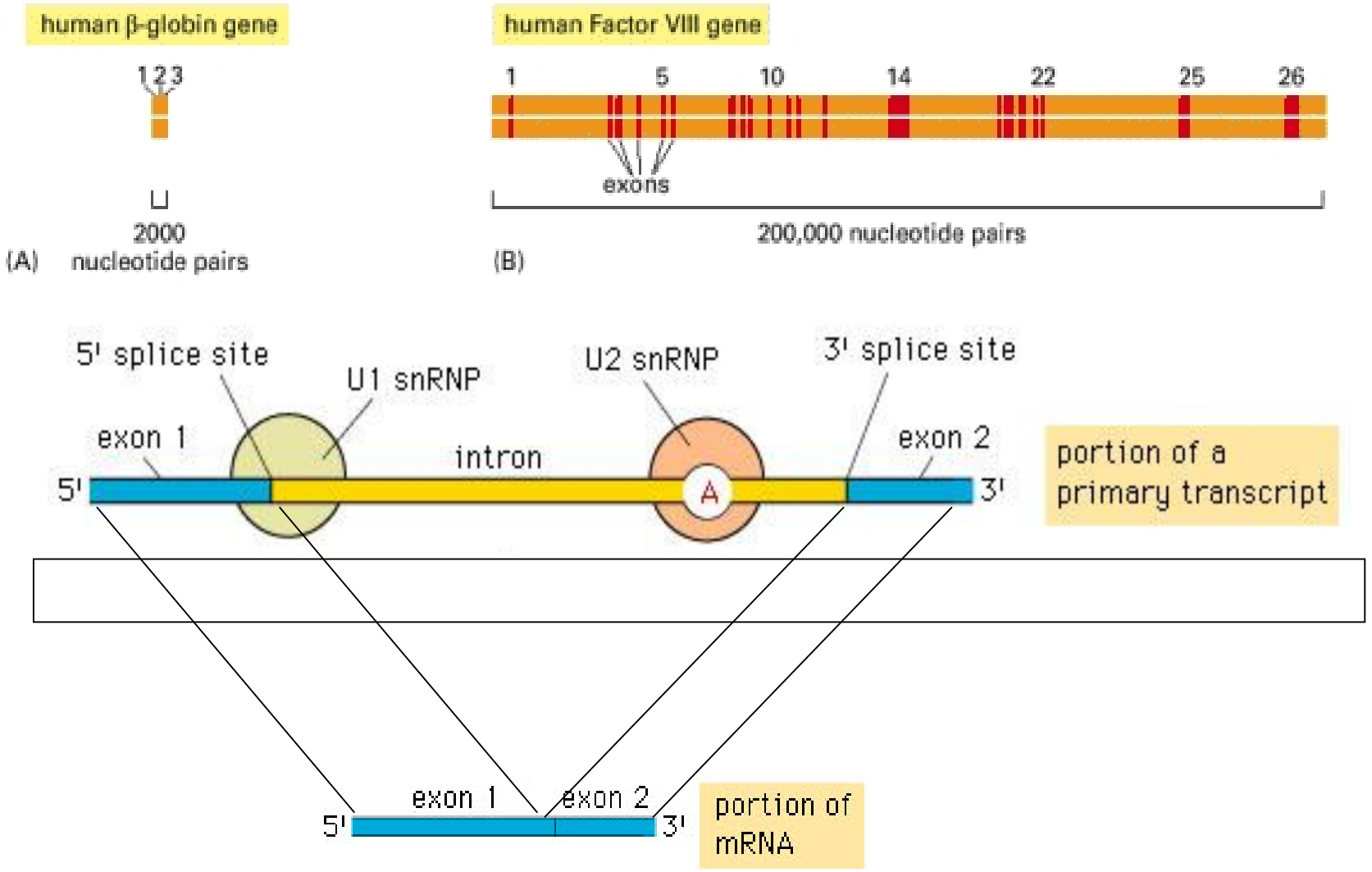
https://www.youtube.com/watch?v=41_Ne5mS2ls

Machine learning problems

1. Predict genes in the genome
2. Predict gene transcription start sites (TSS) and termination sites (TTS)

Parts of a gene: Exons and Introns (Splicing)

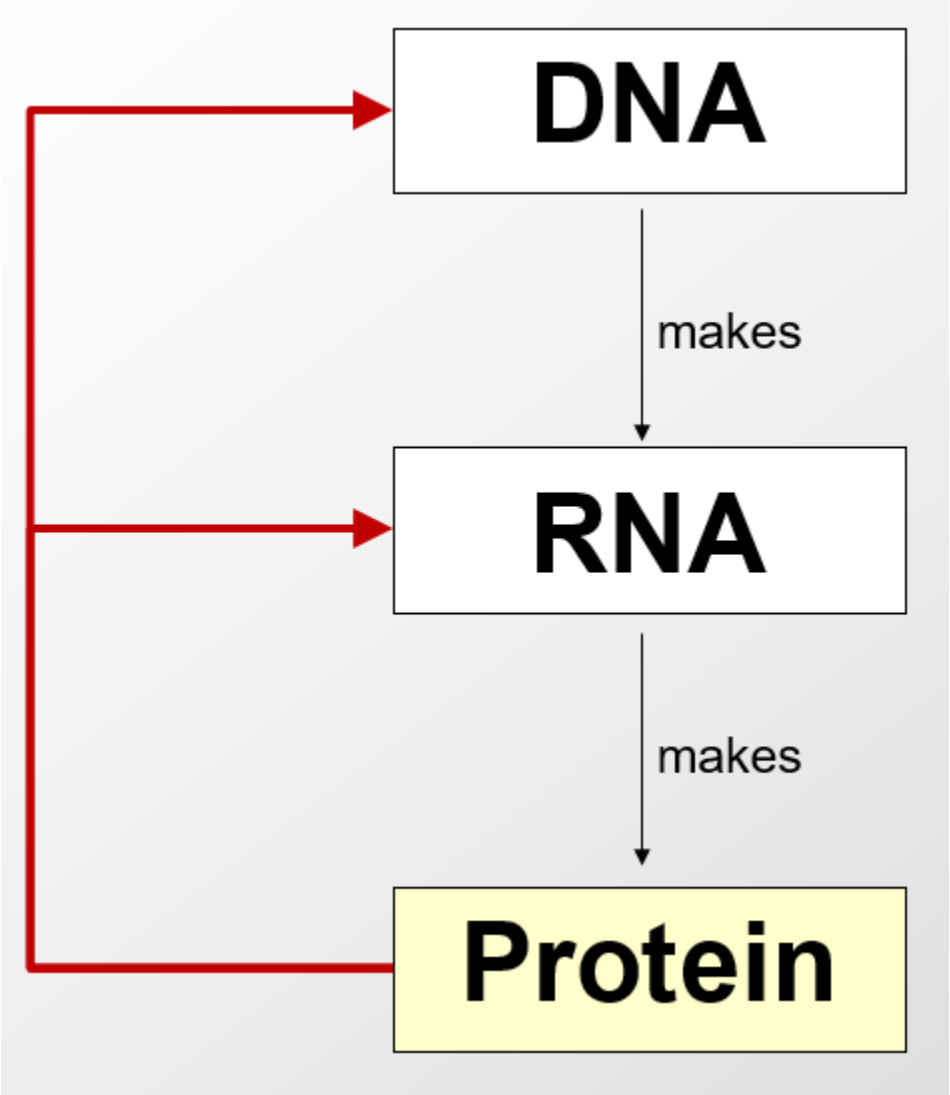
- Not every part of a gene is coding
- Exons (coding) interrupted by non-translated introns
- Introns are **spliced** out
- **Alternative splicing**: different exon subsets for the same gene => different protein products



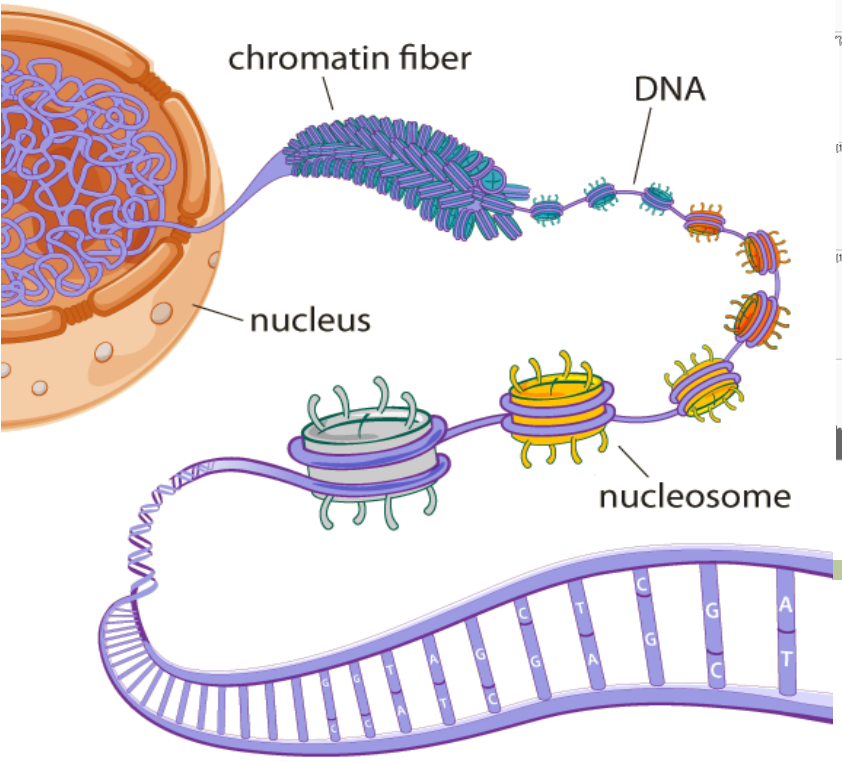
Major ML challenge: Predict slice sites, gene isoforms and alternative splicing events

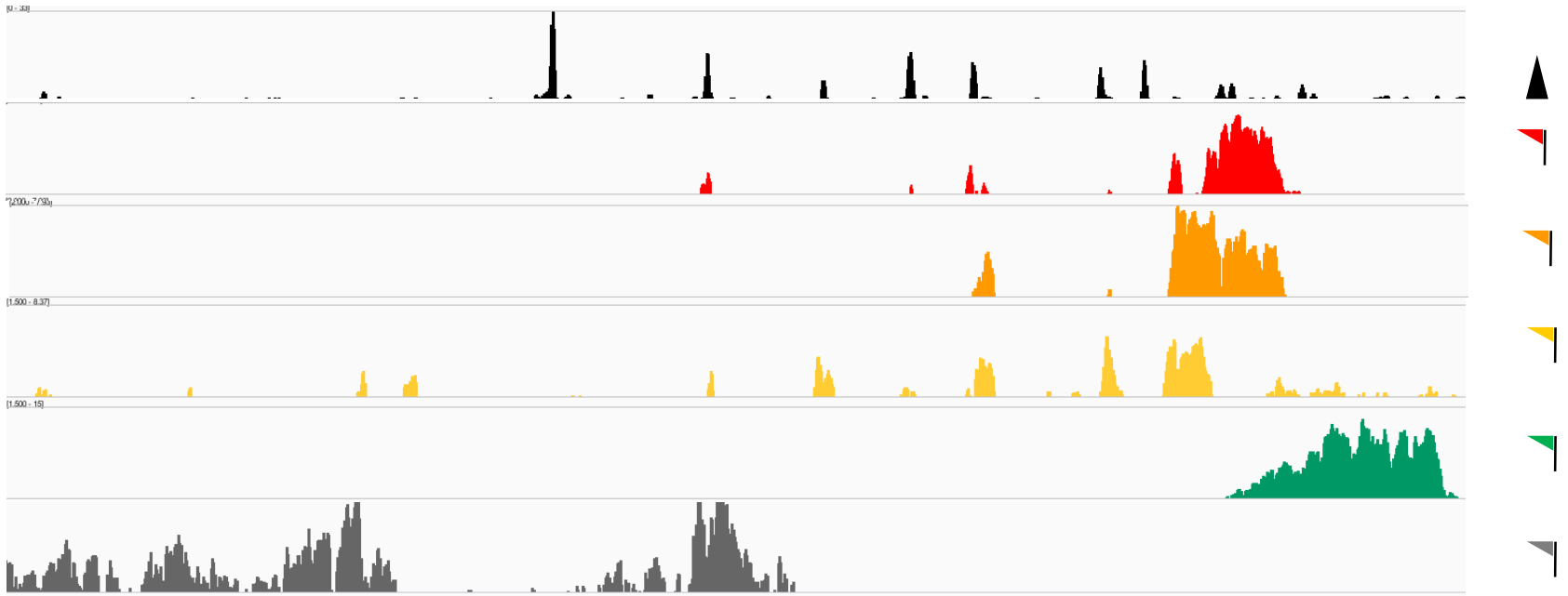
What is gene regulation?

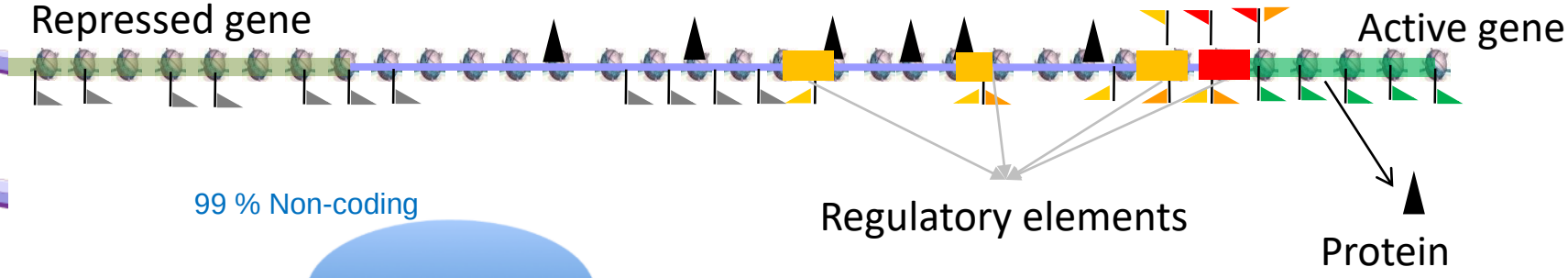
Gene regulation
(When? Where?
How much?)

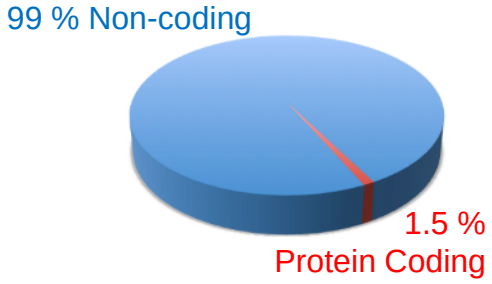


Regulatory control elements and epigenomic marks



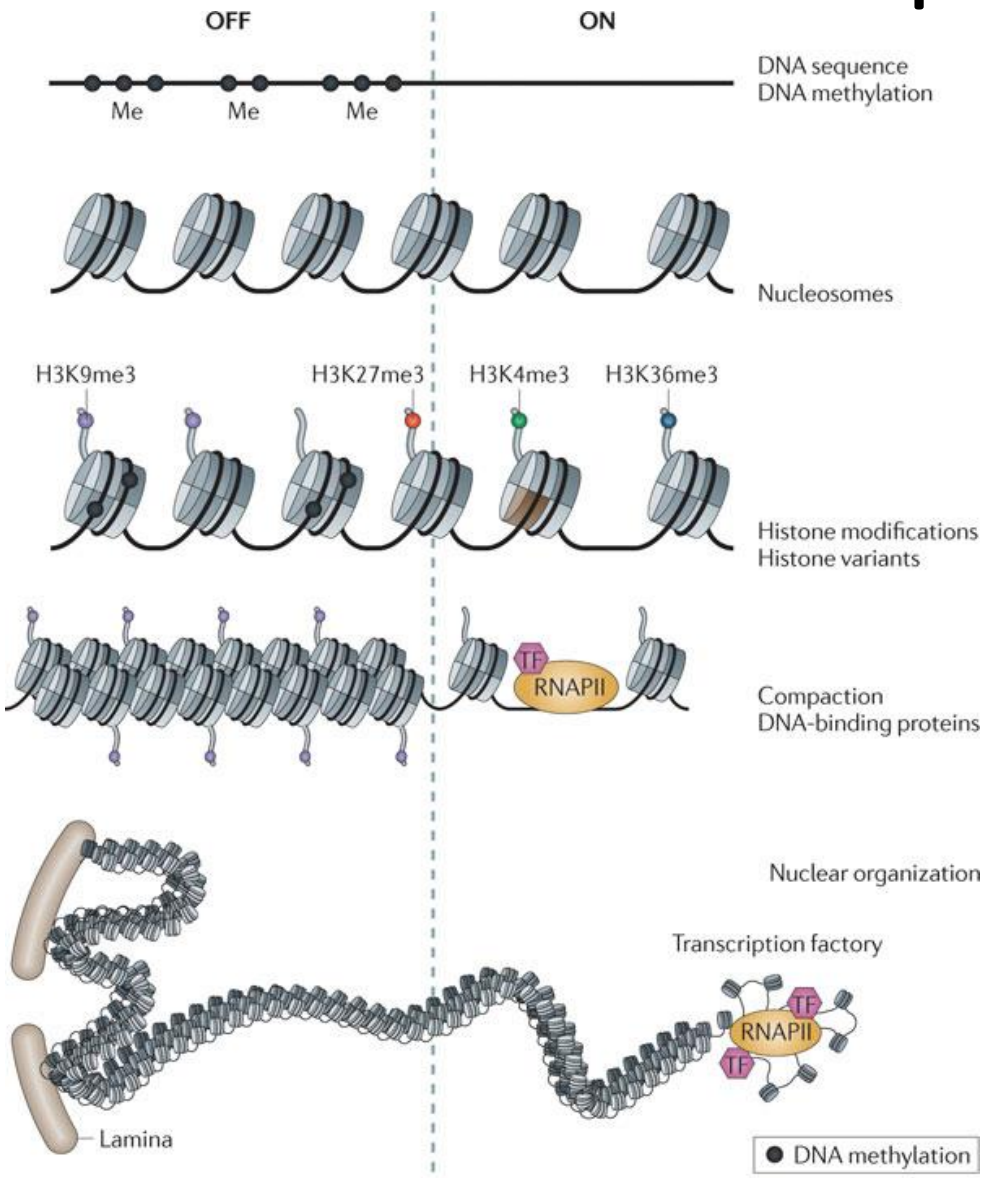






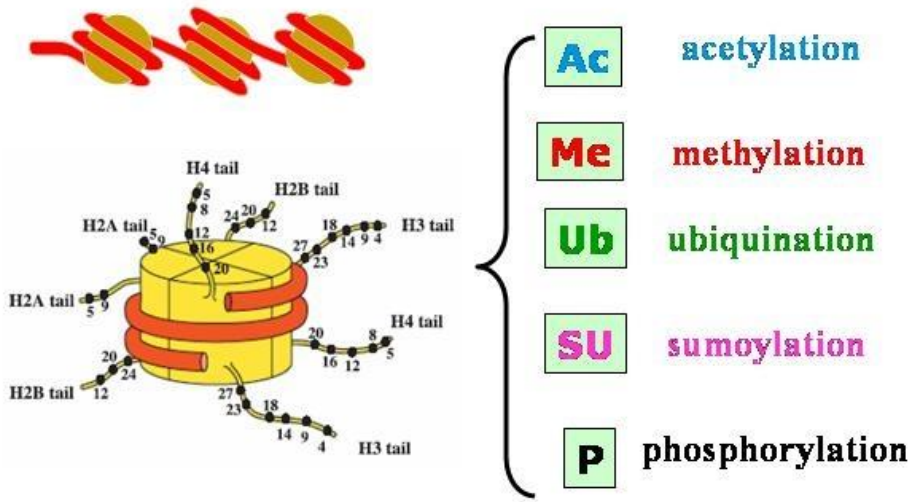
https://www.broadinstitute.org/news/1504

Chromatin and epigenomic modifications



http://www.nature.com/nrg/journal/v12/n1/fig_tab/nrg2905_F1.html

Nature Reviews | Genetics

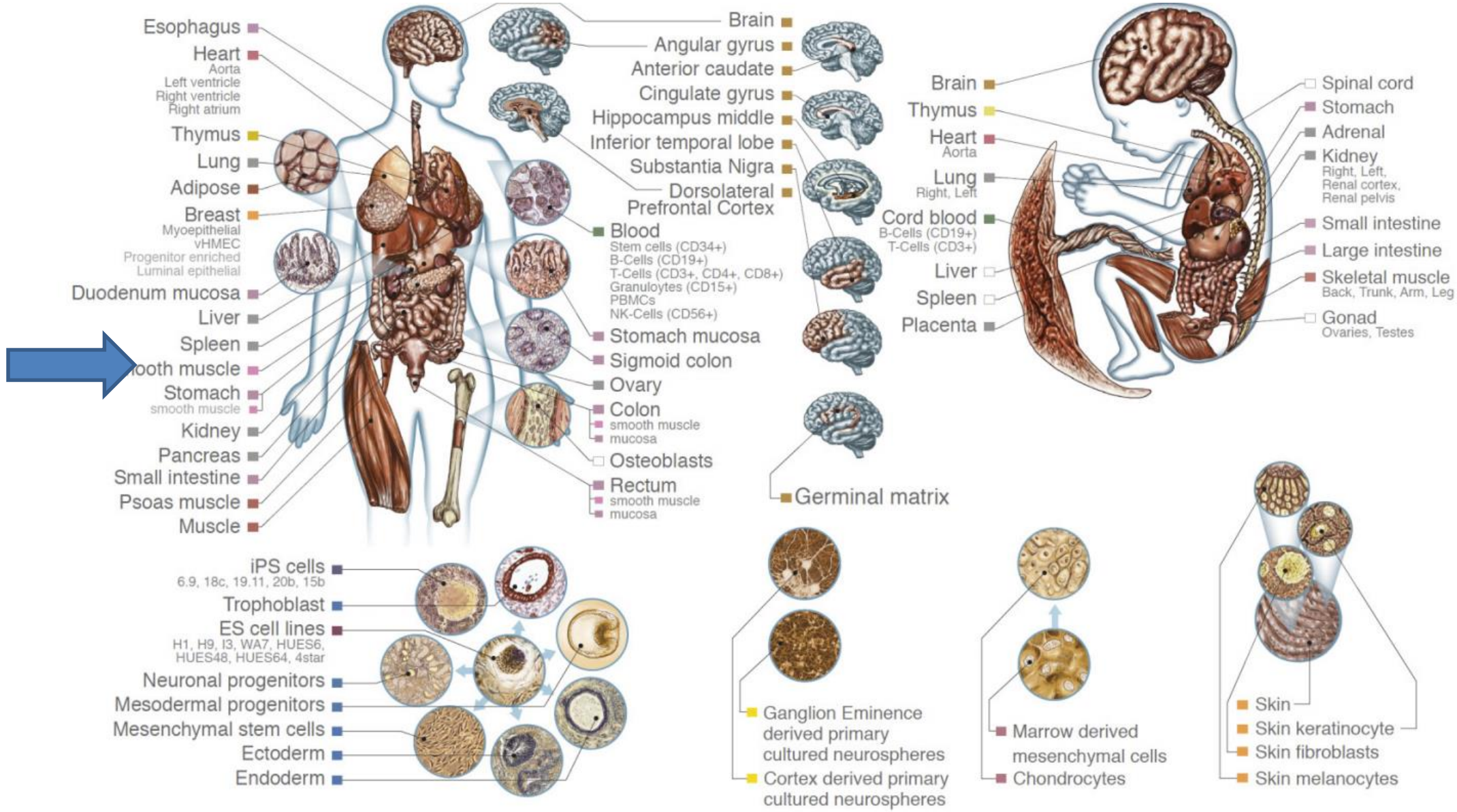


The figure illustrates nucleosome models and major posttranslational modifications which play essential roles in gene expression regulation and disease processes

Major ML challenge: Predict combinatorial epigenomic code

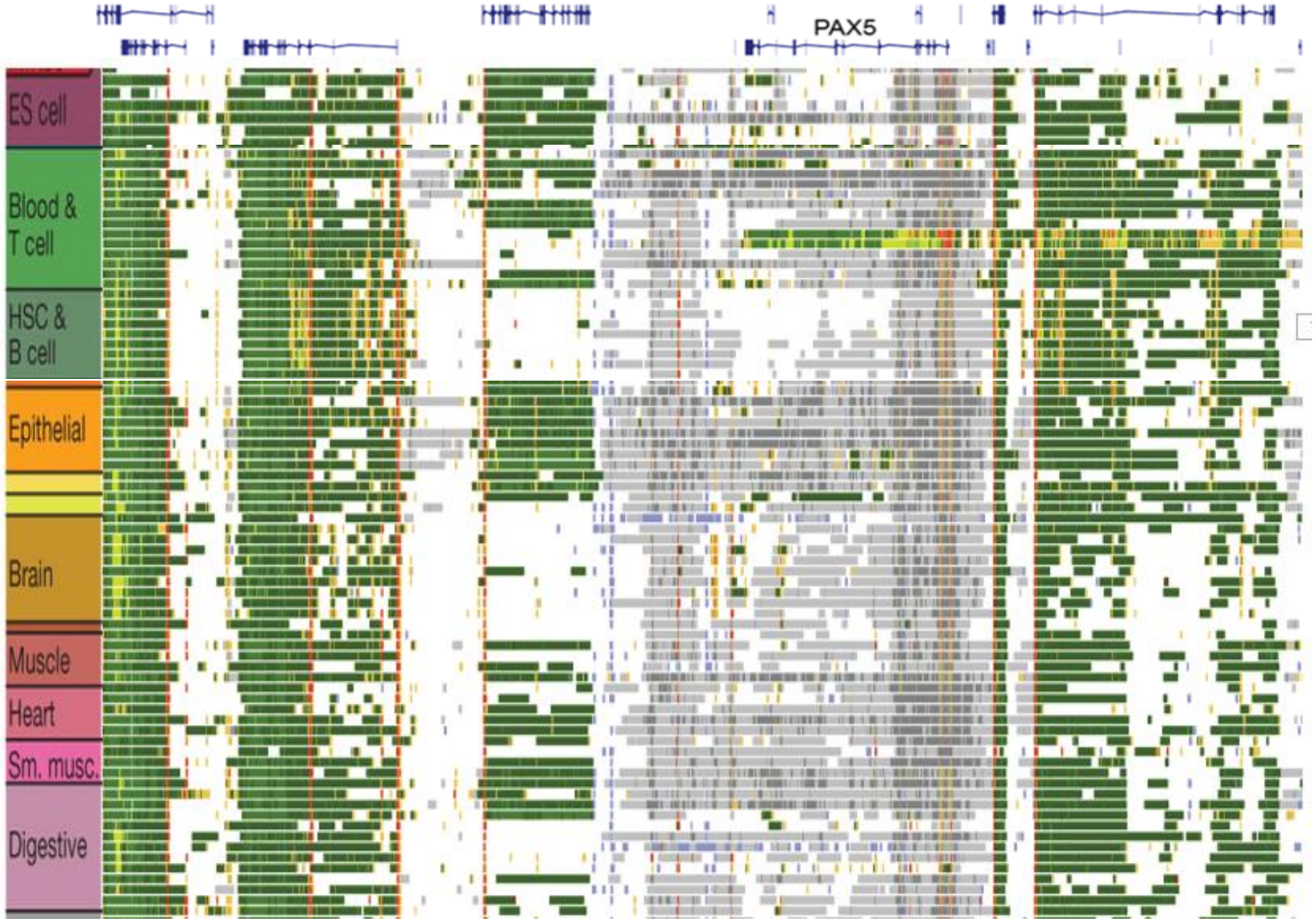
One genome ⇔ Many cell types

```
ACCAGTTACGACGG
TCAGGGTACTGATA
CCCCAAACCGTTGA
CCGCATTTACAGAC
GGGGTTTGGGTTTT
GCCCCACACAGGTA
CGTTAGCTACTGGT
TTAGCAATTTACCG
TTACAACGTTTACA
GGGTTACGGTTGGG
ATTTGAAAAAAAGT
TTGAGTTGGTTTTT
TCACGGTAGAACGT
ACCTTACAAA.....
```



http://www.roadmapepigenomics.org/

Differential activation/repression of control elements and genes defines cell type identity and state



- Active control elements
- Active control elements
- Active genes
- Repressed elements

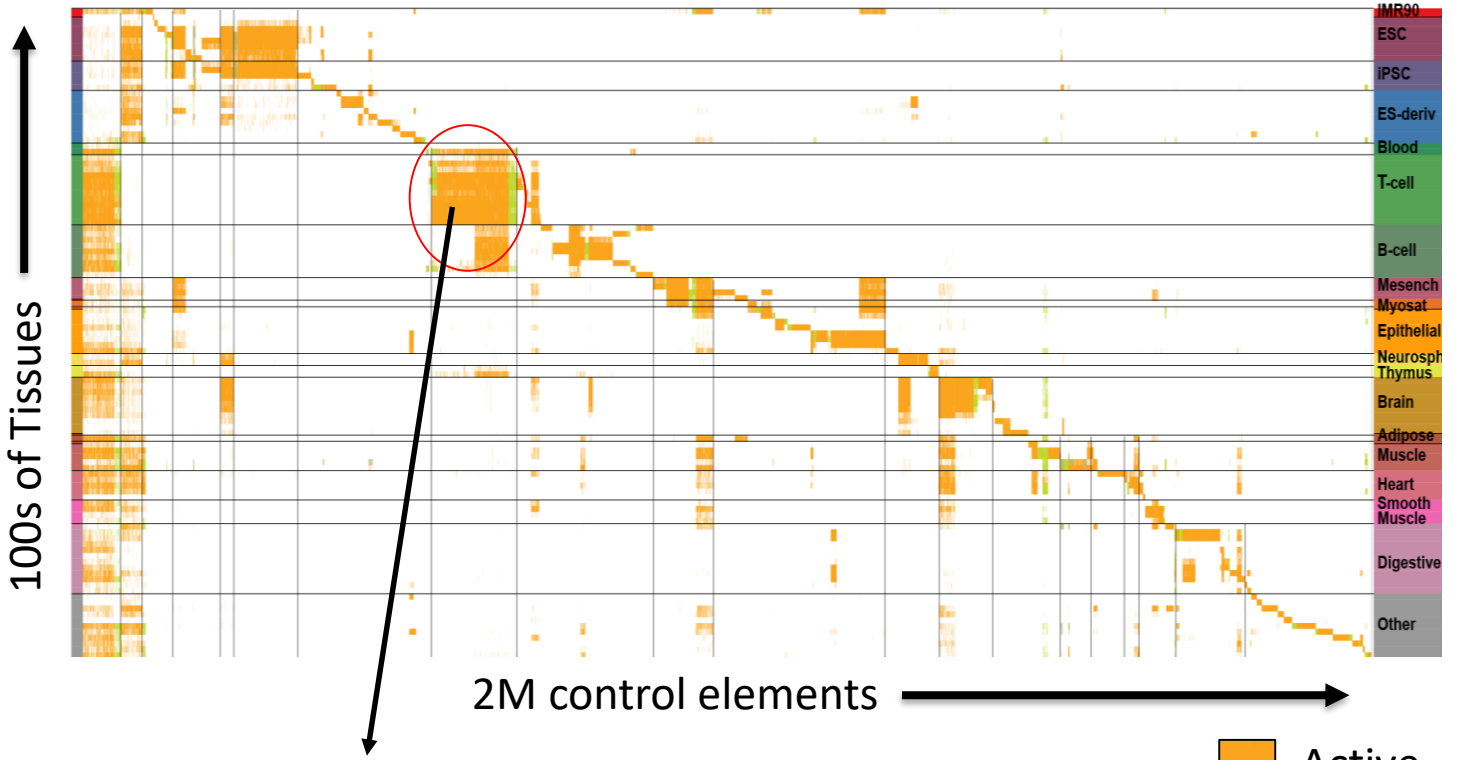
- ~25,000 genes
- ~2 million novel putative control elements!
- cell-type specific usage of elements

Major ML challenges:

1. Predict regulatory elements (REs) and their cell type specific activity
2. Predict which REs regulate which genes in which cell types
3. Predict gene expression from regulatory element activation

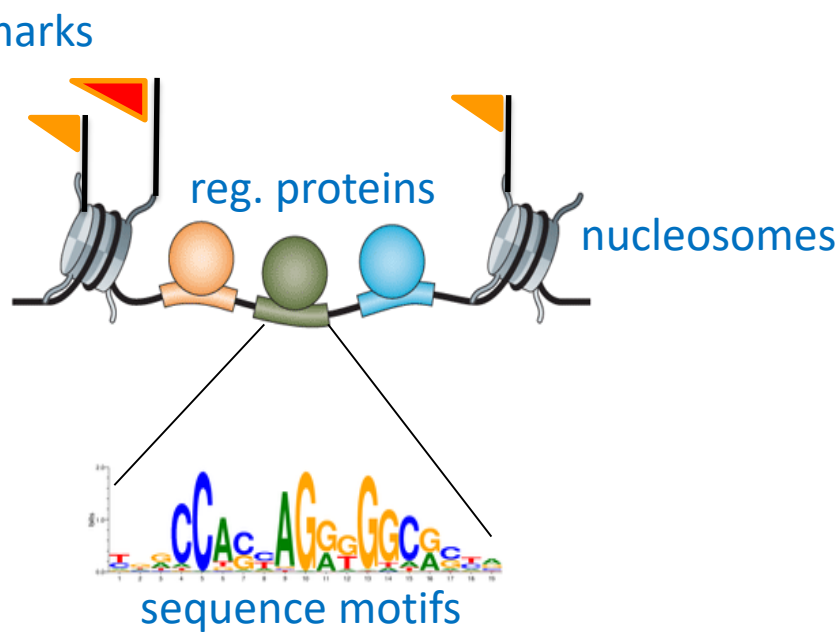
What code in the non-coding DNA controls cell-type specific activation of regulatory elements?

Regulatory proteins bind **DNA words** (landing pads) in control elements!



'Motif Discovery'

The multiple facets of genomic regulatory code



Adapted from Shlyueva et al. (2014) Nature Reviews Genetics.

Major ML challenges:

1. Predict DNA sequence affinity of individual proteins
2. Predict binding landscape of proteins in different cell types and states
3. Predict combinatorial binding patterns of proteins
4. Predict combinatorial regulatory grammars encoded in non-coding regulatory elements

Genetic variation across individuals

Homozygous
(identical alleles)

Heterozygous
(different alleles)

Alleles of the variant
major allele: the more common copy
minor allele: the rarer copy

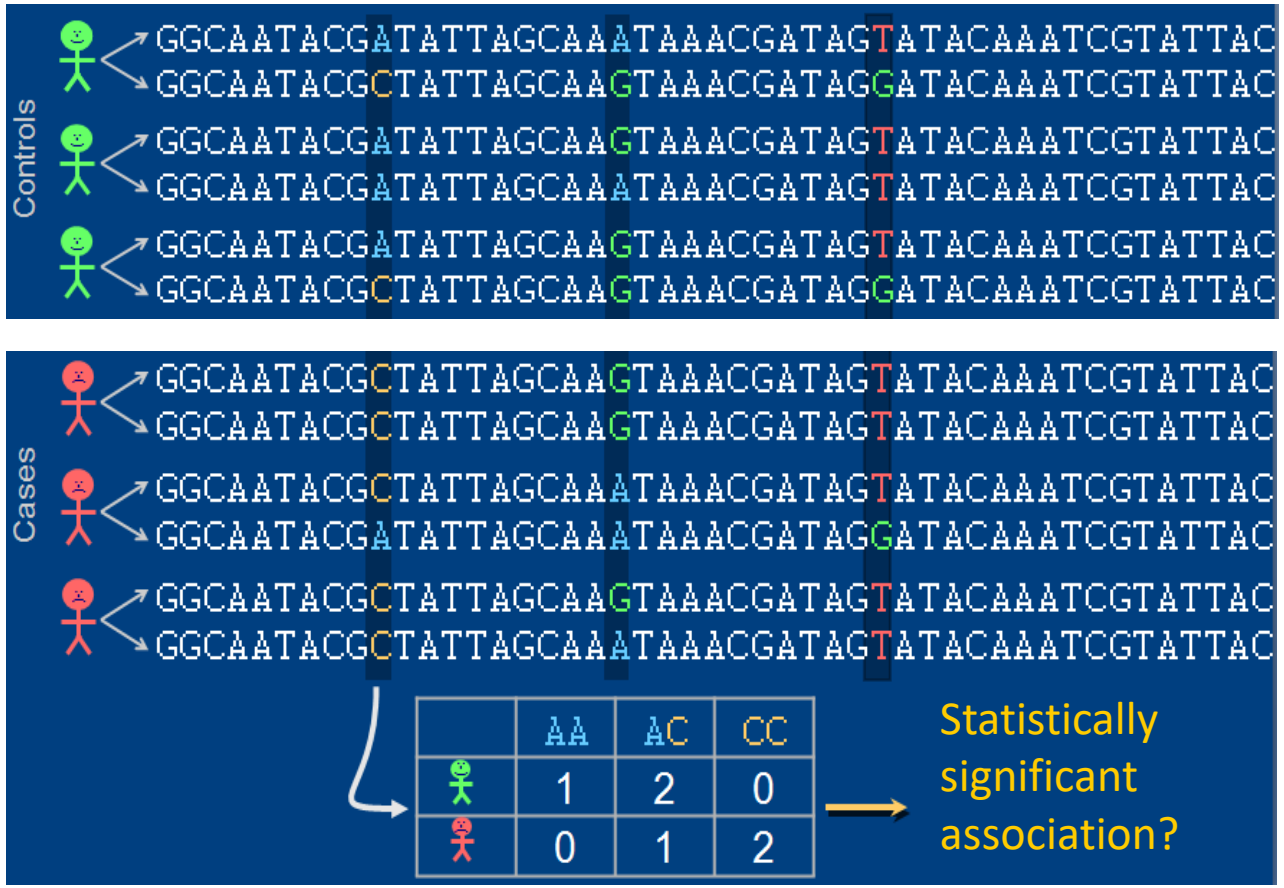
Human genome: ~3 billion bp X <u>2 copies</u>

Types of genetic Variants:

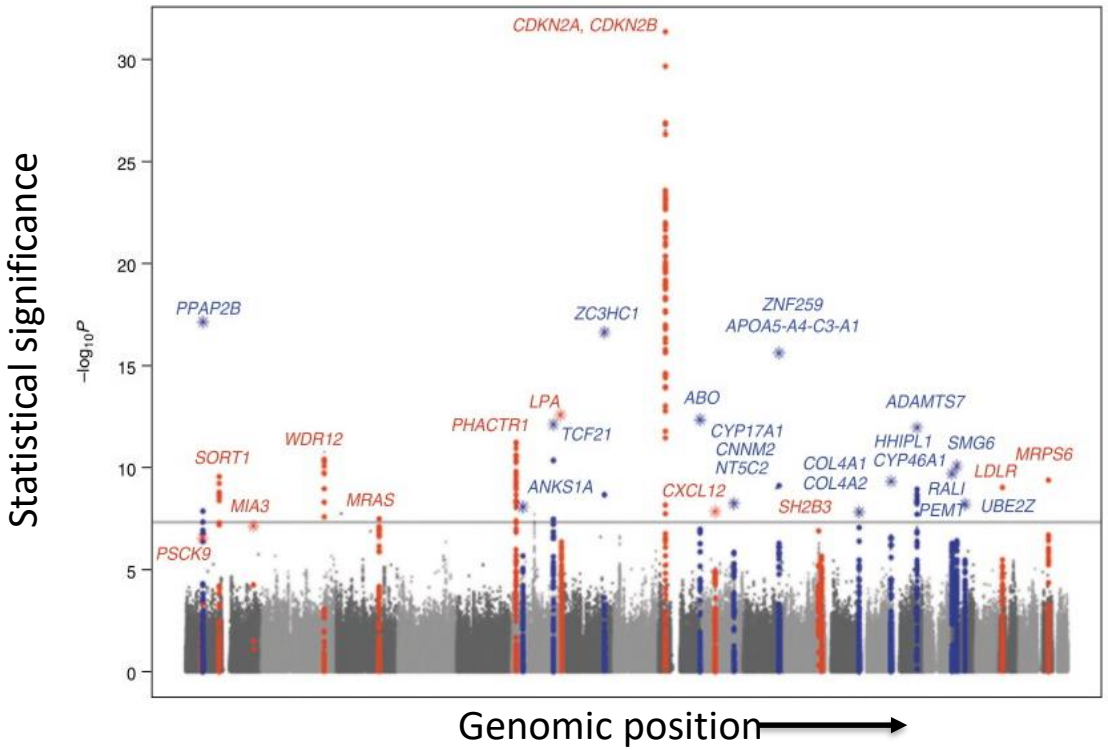
- Single nucleotide polymorphisms (SNPs): substitutions
  - ~3M **<u>common</u>** SNPs (> 2-5% minor allele frequency (MAF) in human population)
  - Rare (< 2-5% MAF frequency) and private SNPs (in a single individual)
- Short and large insertions, deletions, inversions, translocations

<u>Germline vs. Somatic mutations:</u> Genetic alteration acquired by a cell that can be passed to the progeny of the **mutated** cell in the course of cell division. **Somatic mutations** differ from **<u>germ line mutations</u>**, which are inherited genetic alterations that occur in the germ cells (i.e., sperm and eggs)

Case-control studies to identify disease-associated genetic variants

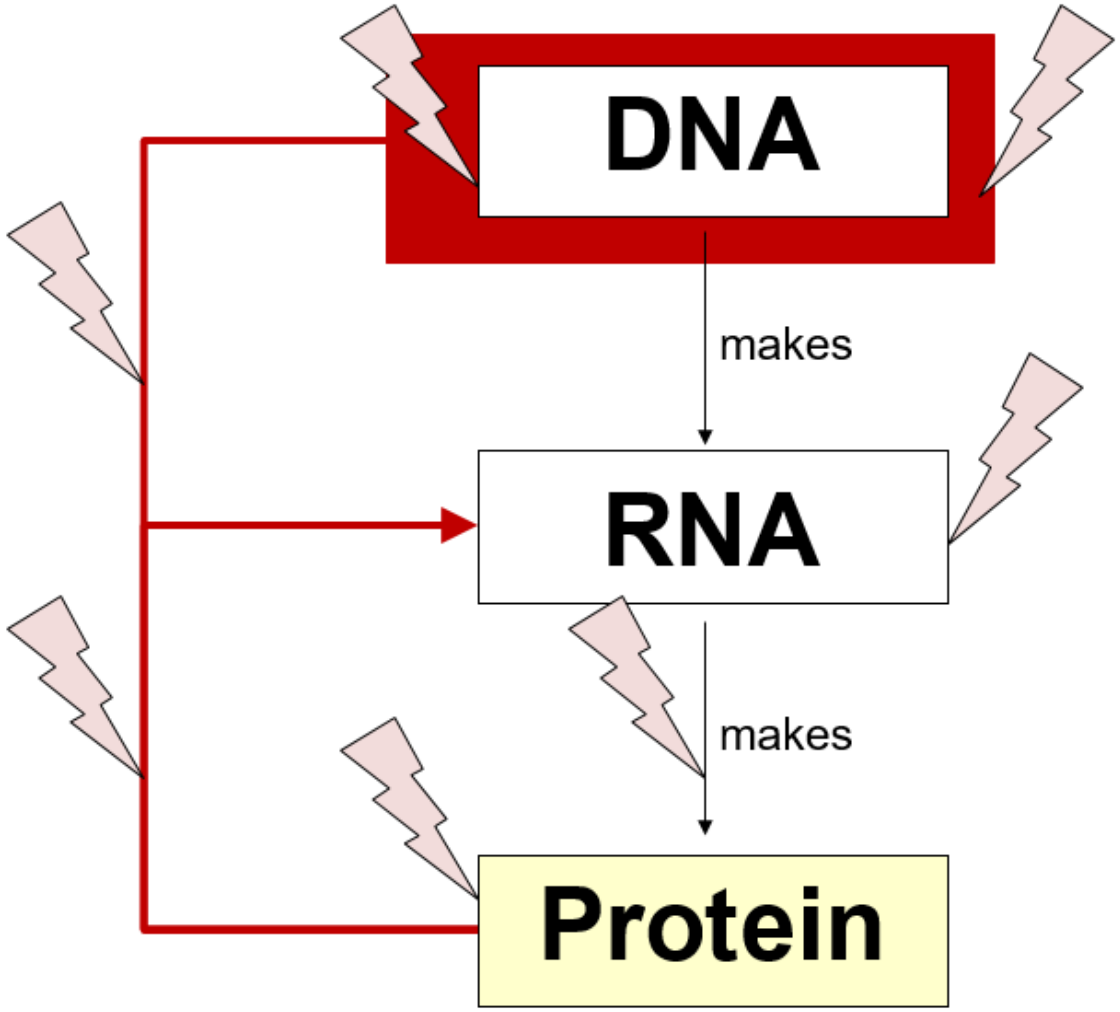


Correlation structure between variants makes it difficult to identify causal variant



Genome-wide association study (GWAS)

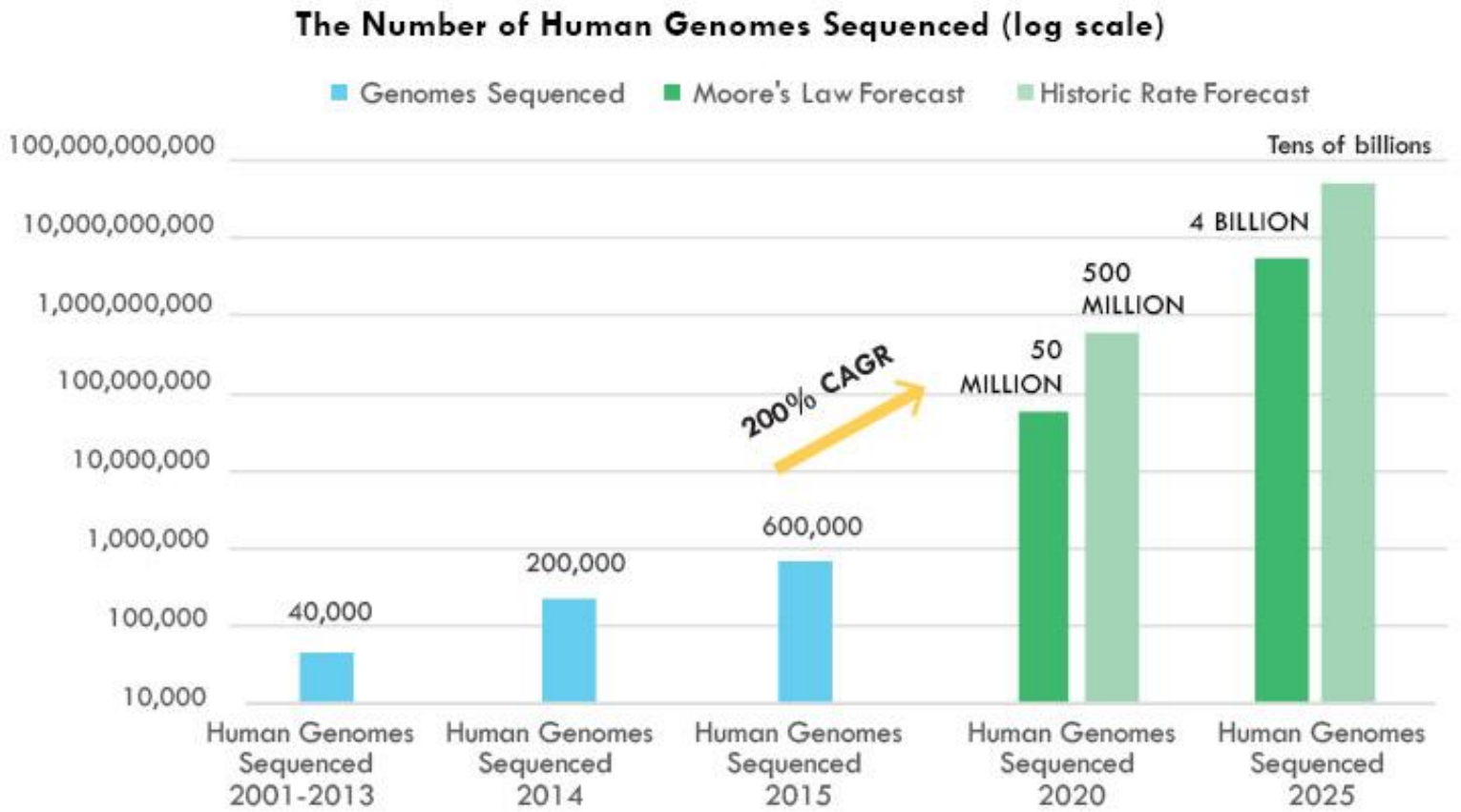
The role of genetic variation in regulation



Major computational challenges:

1. Predict genetic variants from genome sequencing data
2. Which genetic variants are benign vs. harmful i.e. which genetic variants are associated with different phenotypes (diseases/traits)
3. What is the genetic architecture of a disease/trait (relationship between variants to phenotype)
4. Predict the molecular effect of a genetic variant i.e. how it affects cellular function

Introduction to high-throughput sequencing



Source: National Human Genome Research Institute (NHGRI), ARK Investment Management LLC

GA II
1.6 billion bp per day
(2008)

GA IIx
5 billion bp per day
(2009)

HiSeq 2500
60 billion bp per day
(2012)

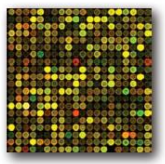
Images: www.illumina.com/systems
Numbers: www.politigenomics.com/next-generation-sequencing-informatics
Dates: Illumina press releases

Oxford Nanopore technology

Sequencing technologies

Sanger DNA sequencing

1977-1990s



DNA Microarrays

Since mid-1990s

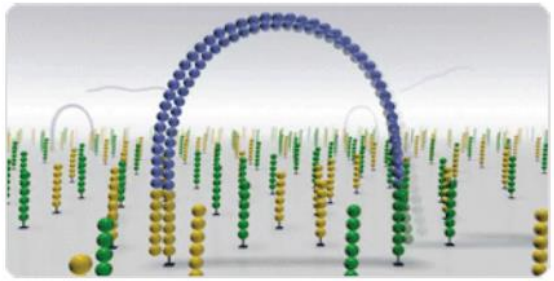
2nd-generation DNA sequencing

Since ~2007

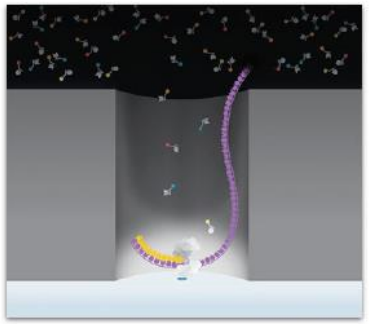
3rd-generation & single-molecule DNA sequencing

Since ~2010

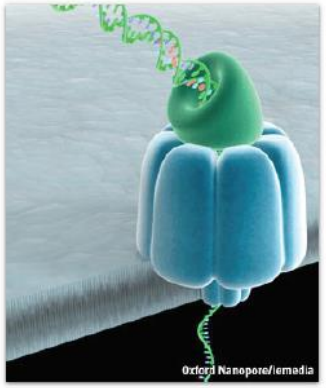
Since 2005, many DNA sequencing instruments have been described and released. They are based on a few different principles



Synthesis / ligation



SMRT cell

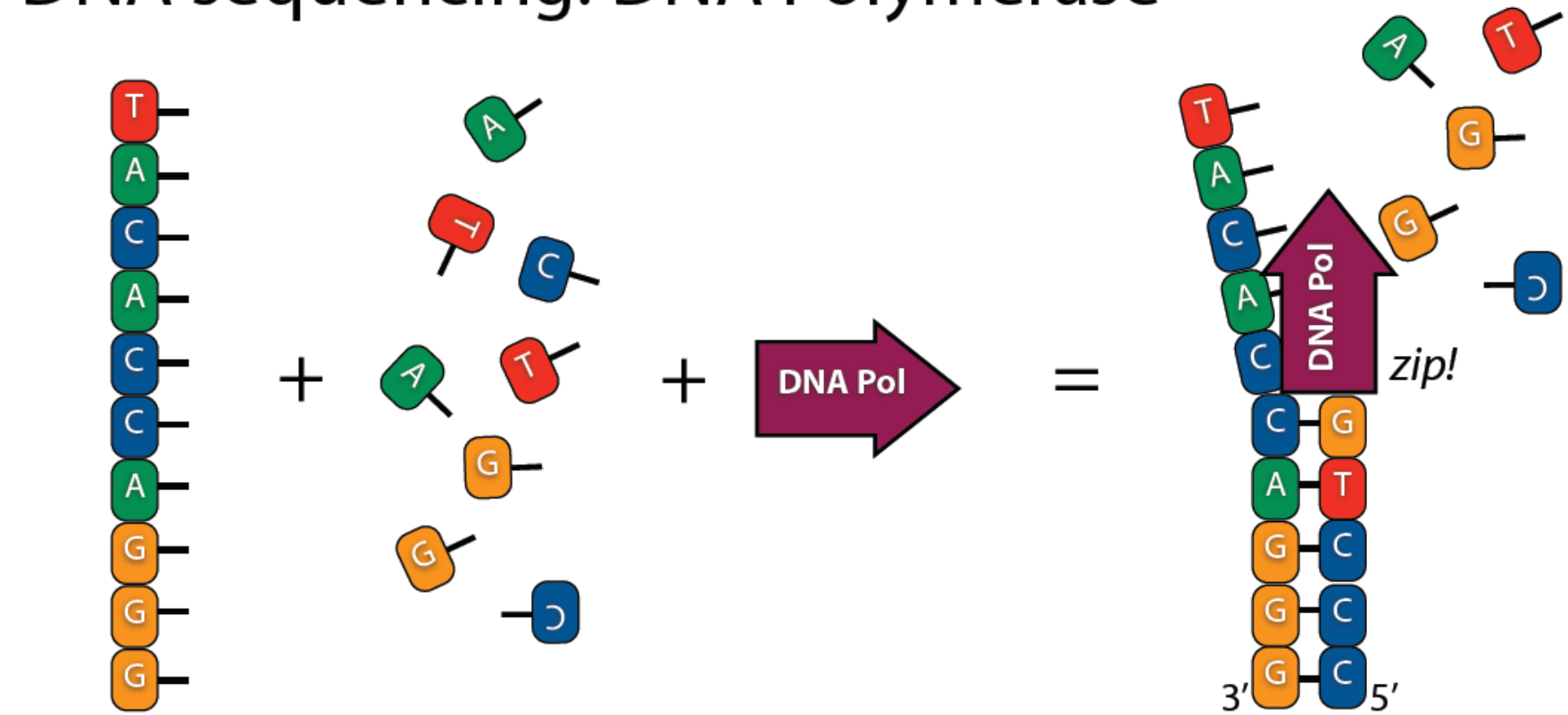


Nanopore

Sequencing by synthesis ("massively parallel sequencing") provides greatest throughput, and is the most prevalent today

Pictures: http://www.illumina.com/systems/miseq/technology.ilmn, http://www.genengnews.com/gen-articles/third-generation-sequencing-debuts/3257/

DNA sequencing: DNA Polymerase



Single-stranded DNA *template* | Free nucleotides | DNA polymerase | Strand synthesis

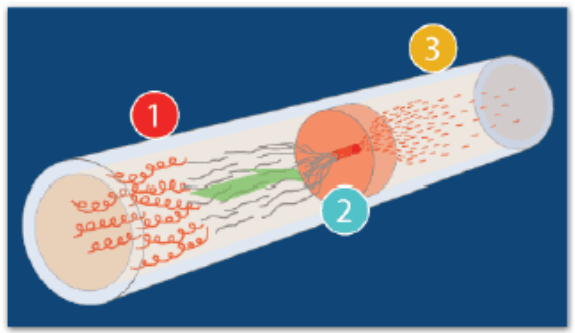
DNA polymerase moves along the template in one direction, integrating complementary nucleotides as it goes

Slides from Ben Langmead

Sequencing by synthesis

1. Take DNA sample, which includes many copies of the genome, and chop it into single-stranded fragments ("templates")

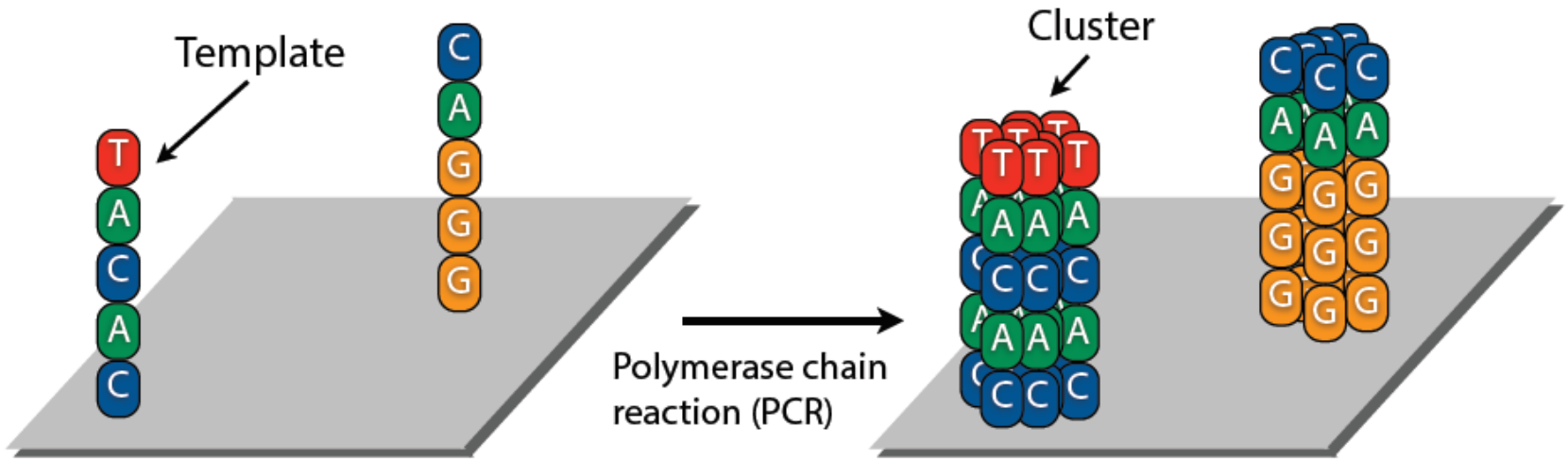
   E.g. with ultrasound waves, water-jet shearing (pictured), divalent cations



Picture: http://www.jgi.doe.gov/sequencing/education/how/how_1.html

2. Attach templates to a surface

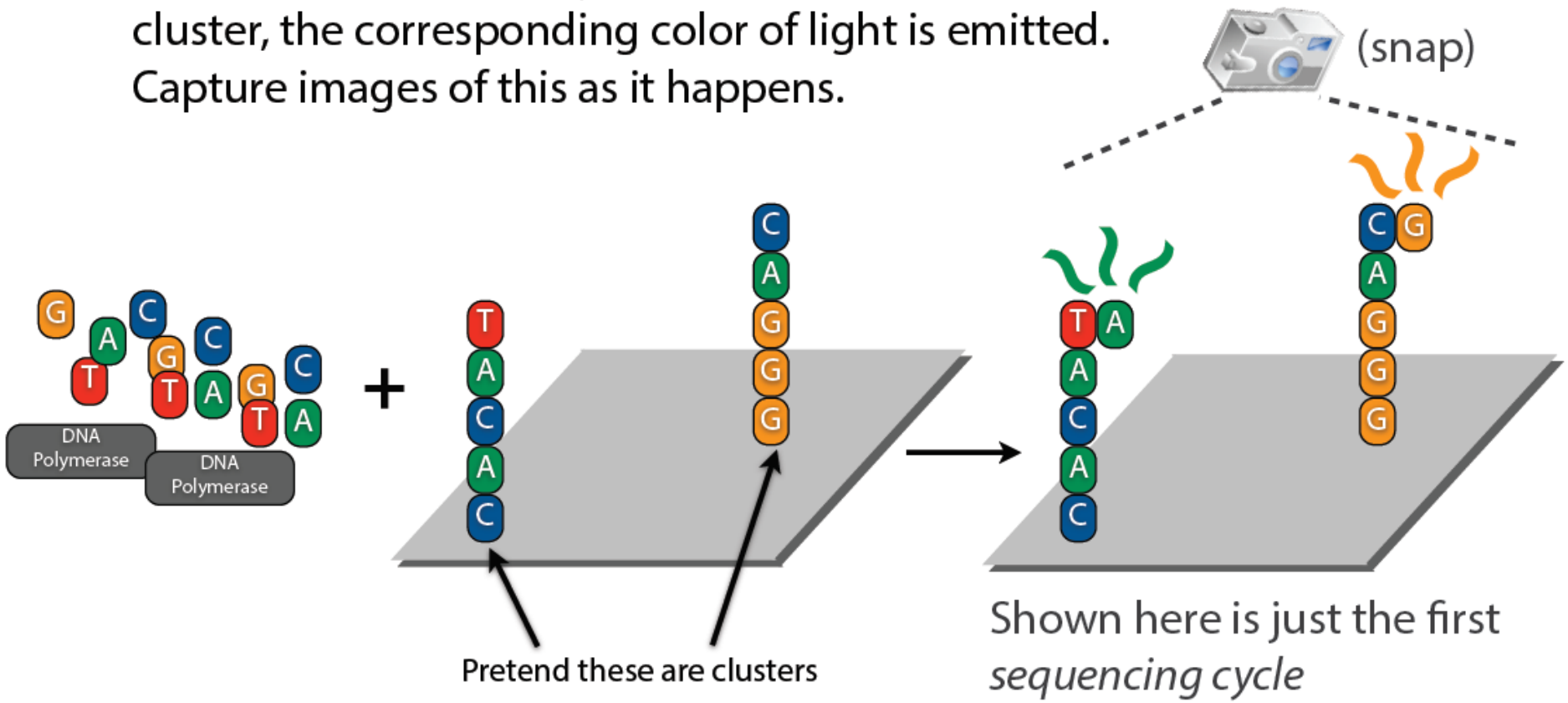
3. Make copies so that each template becomes a "cluster" of clones



Slides from Ben Langmead

Sequencing by synthesis

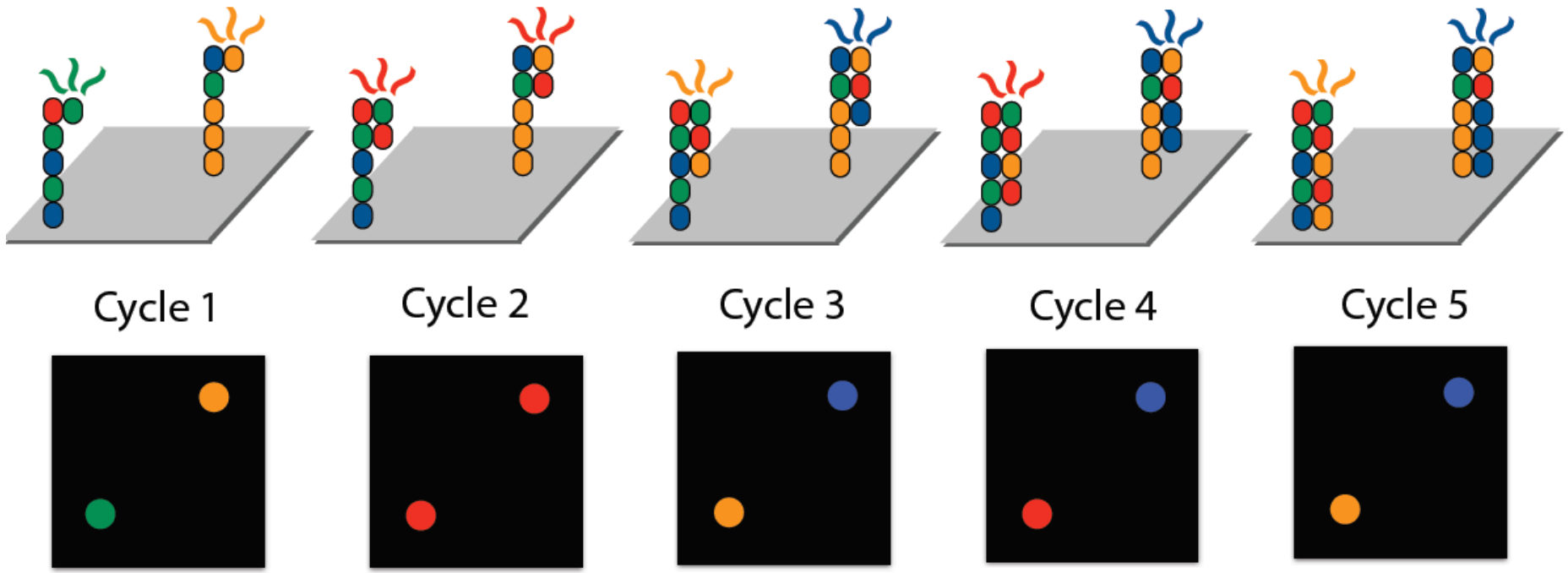
4. Repeatedly inject mixture of *color-labeled* nucleotides (A, C, G and T) and DNA polymerase. When a complementary nucleotide is added to a cluster, the corresponding color of light is emitted. Capture images of this as it happens.



Slides from Ben Langmead

Sequencing by synthesis

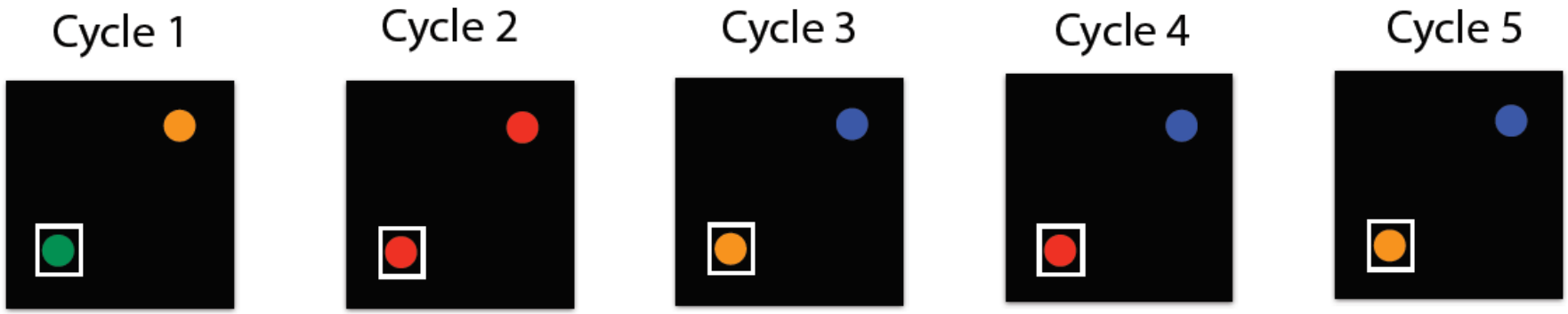
5. Line up images and, for each cluster, turn the series of light signals into corresponding series of nucleotides



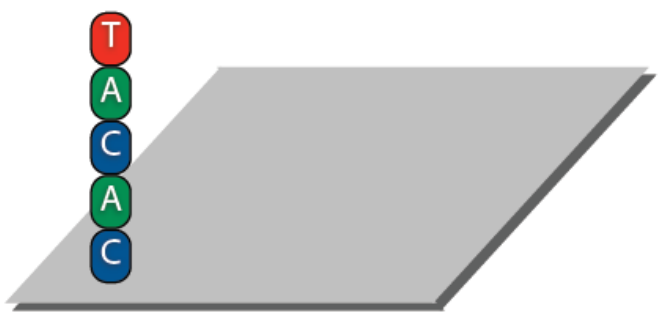
Slides from Ben Langmead

Sequencing by synthesis

5. Line up images and, for each cluster, turn the series of light signals into corresponding series of nucleotides



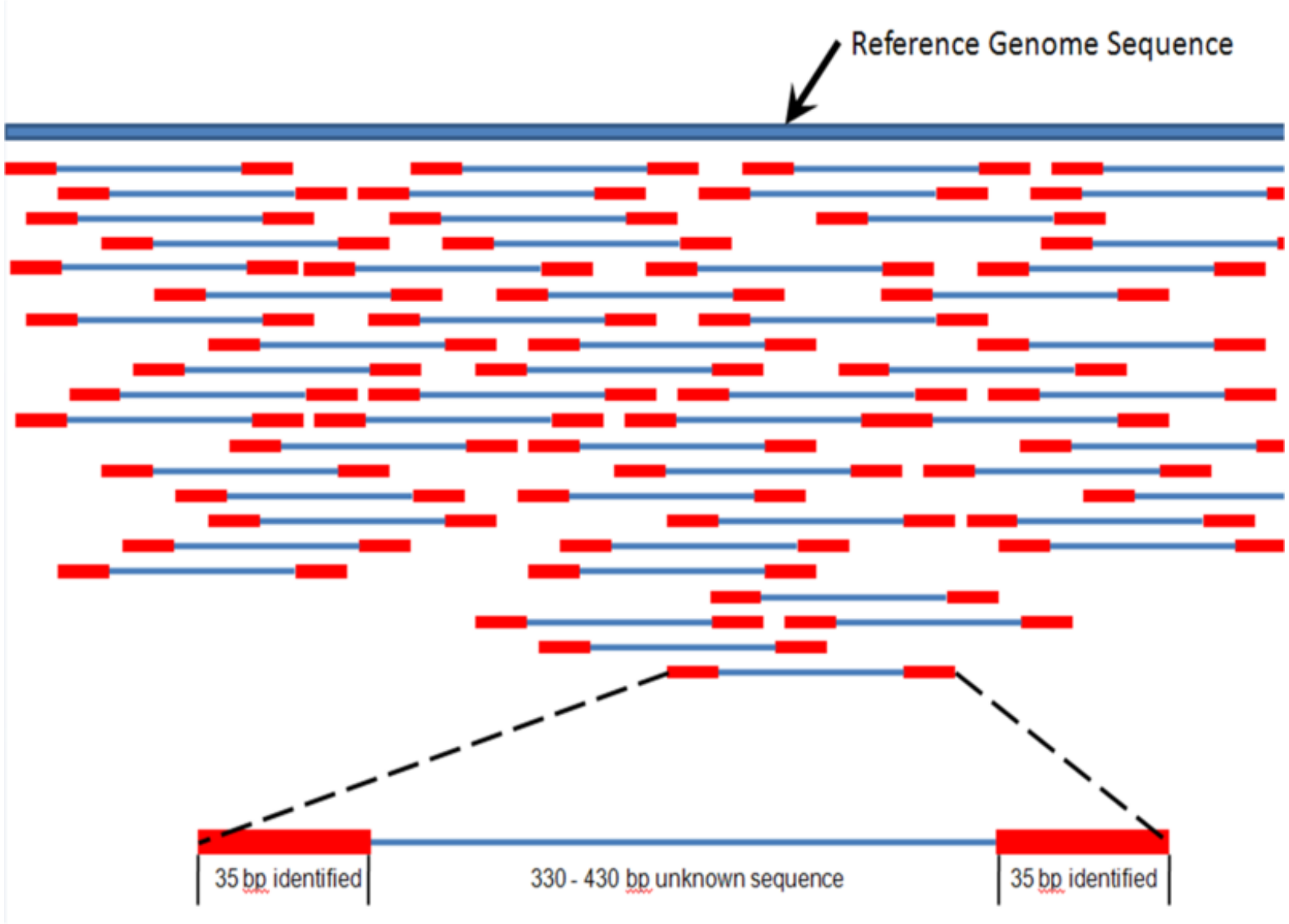
"Base caller" software looks at this cluster across all images and "calls" the complementary nucleotides: TACAC, corresponding to the template sequence



TACAC is a "sequence read," or "read." Actual reads are usually 100 or more nucleotides long.

Slides from Ben Langmead

Genome assembly



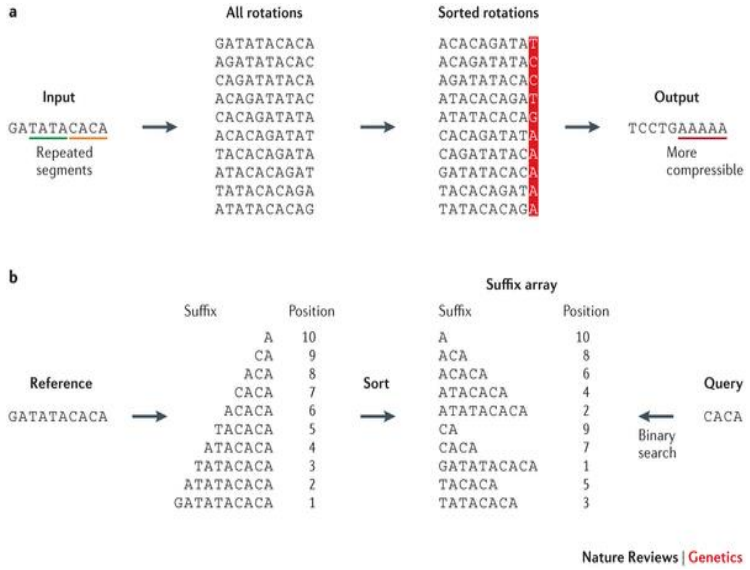
Mapping reads to reference genome

Naïve method

- Scan whole genome with every read
- Problem: Too slow

Indexing + Alignment approach

- Create a compressed reference 'genome index'
  - a map of where each short subsequence of length 'k' hits the genome
- Map reads using index via smart alignment algorithms and data structures (e.g suffix array)
- Allow for errors: insertions, deletions, mismatches in alignments



http://www.nature.com/nrg/journal/v14/n5/box/nrg3433_BX2.html

Run times for indexing alignment

- Indexing human genome ~ 3 hours
- Alignment speed: 2 million 35 bp reads on 1 processor ~20 mins
- Alignment speed depends on error rate