Genomics 101: It’s All in Your DNA

Nathan Cude, Ph.D.

Senior Scientist, Novozymes

Chestnut Chat - October 15, 2021

Definitions

- Genome (*noun*) - the complete set of genetic material (*e.g.*, DNA) present in a cell or organism
- Genomics (*noun*) - the branch of molecular biology concerned with the structure, function, evolution, and mapping of genomes

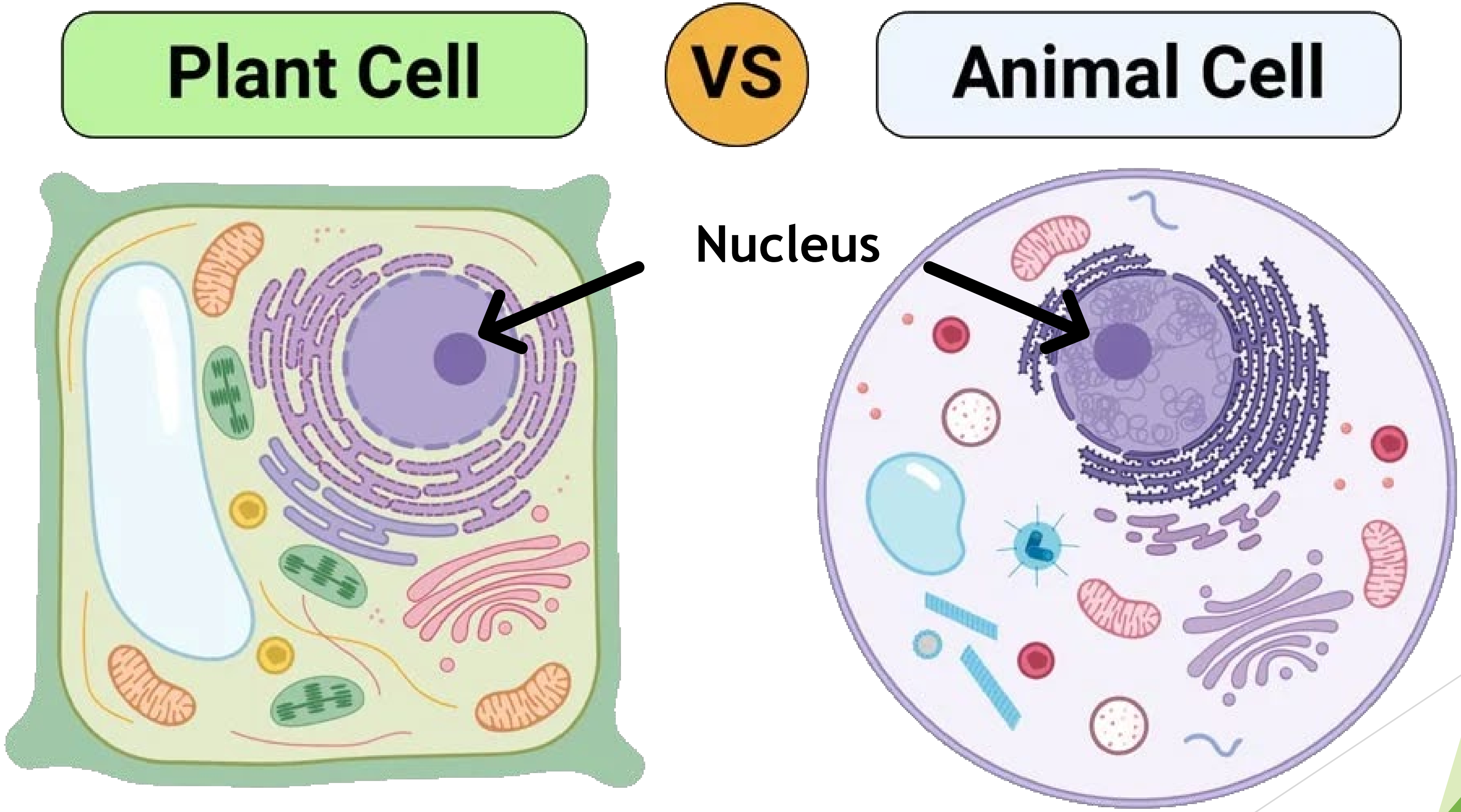
Focus of this Presentation

1. What is DNA?
2. What is the Central Dogma of Biology?
3. How do we sequence a genome?
4. What can we do with a genome sequence?

DNA -

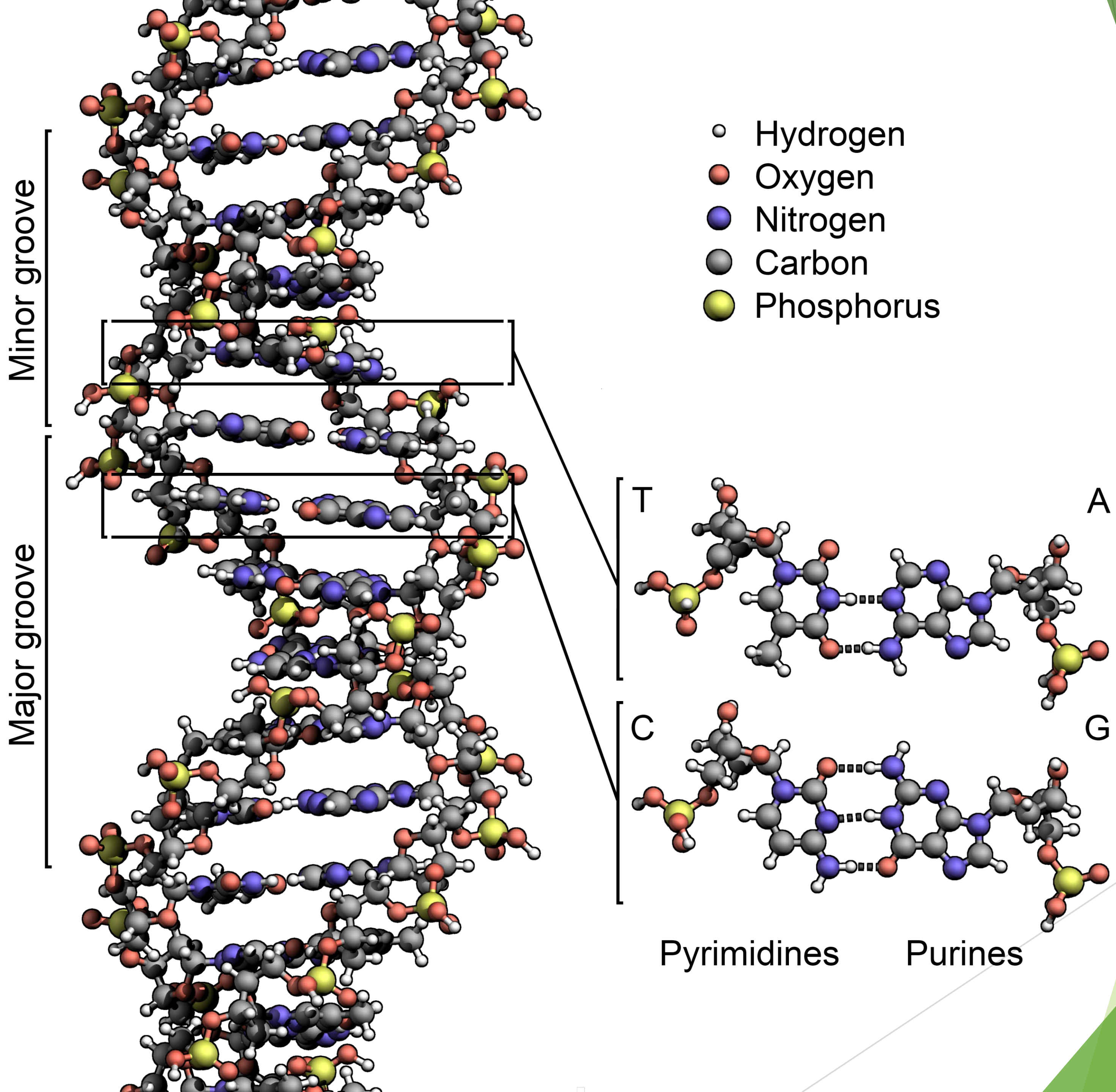
Deoxyribonucleic Acid

DNA in the Eukaryotic Cell



Nucleus

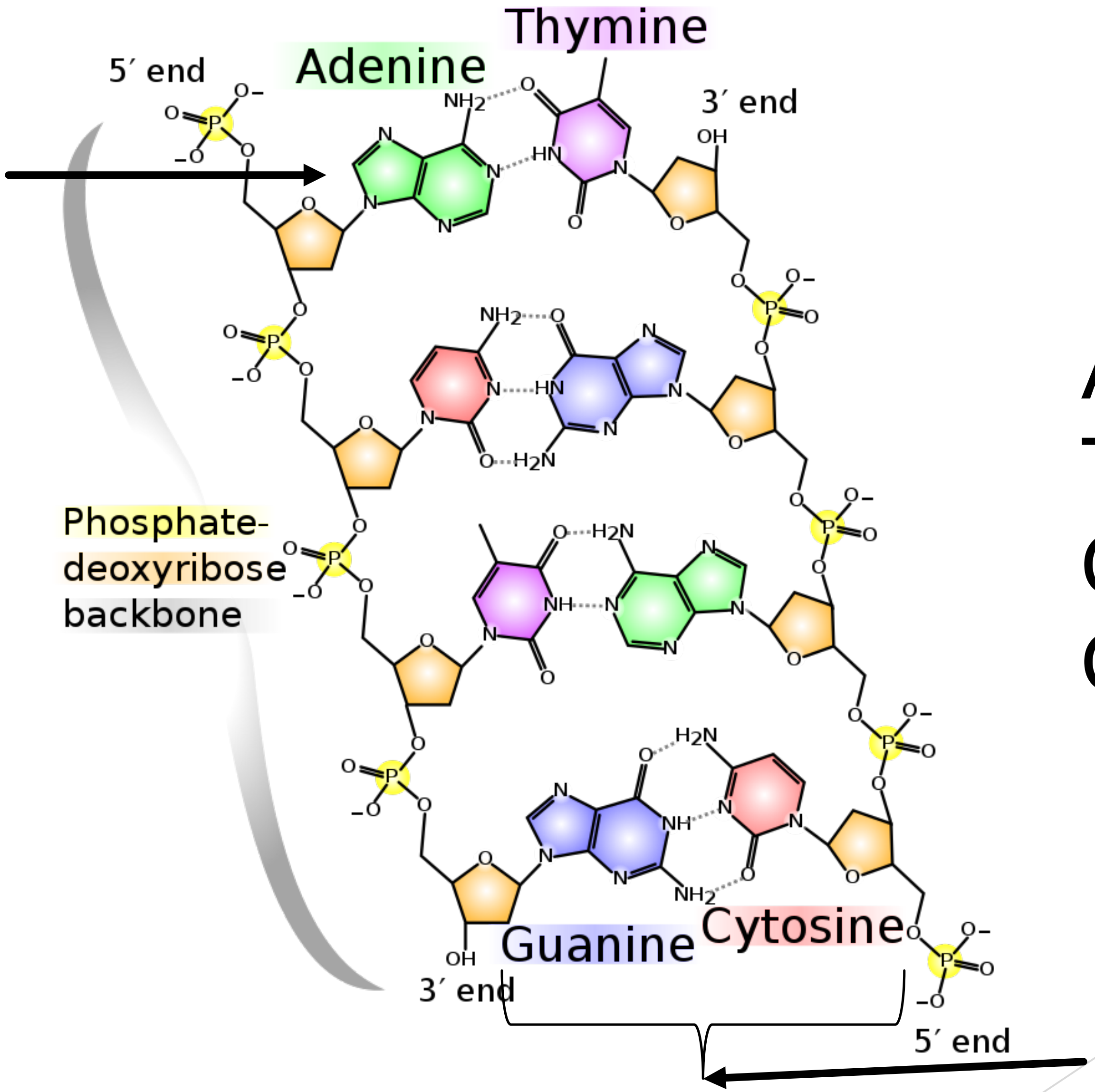
DNA Structure



DNA - Wikipedia

DNA Sequence Simplified

Nucleotide or Base



A→T

T→A

C→G

G→C

Base pair

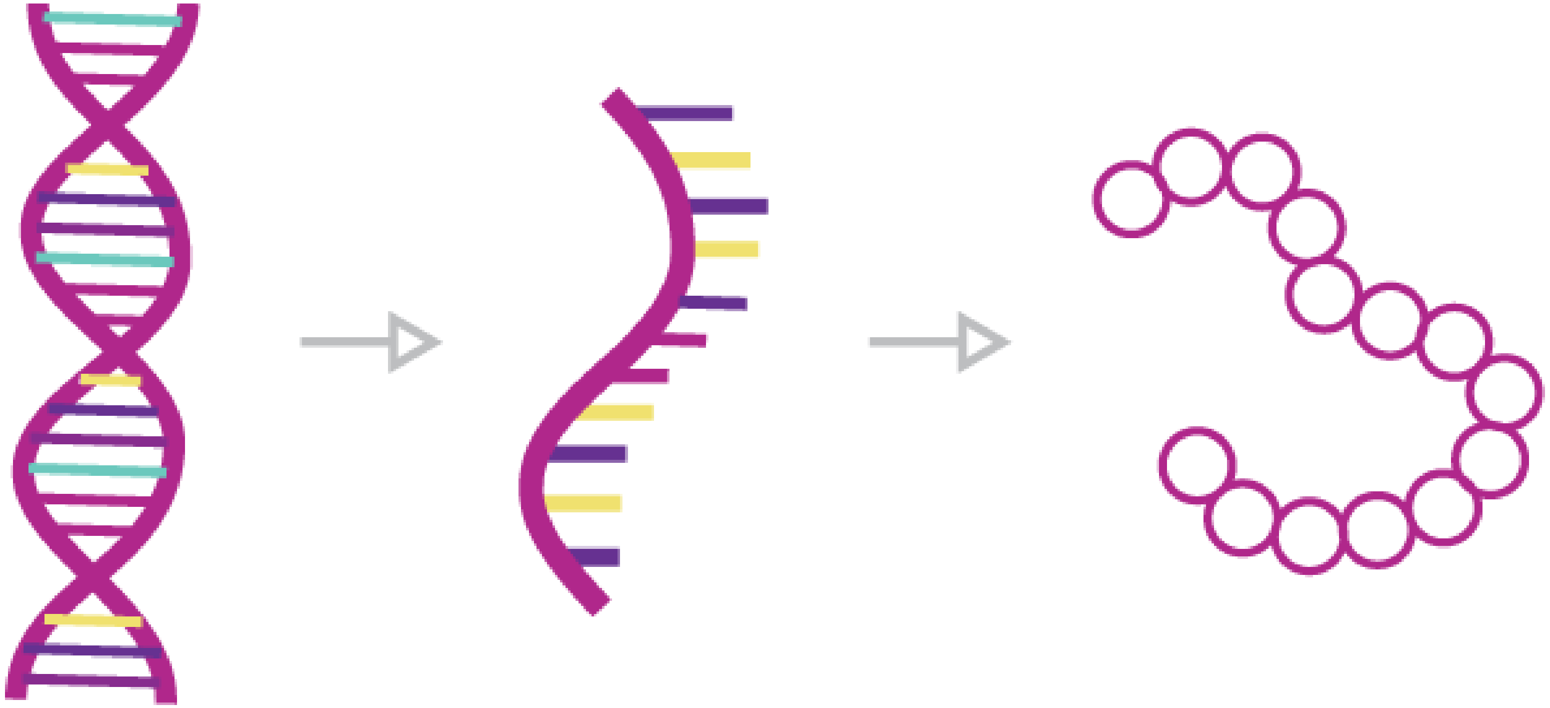
DNA - Wikipedia

DNA Sequence Even More Simplified

```
ATCAAAGAAGAAGATTGGTACCACCATATTAAATGCTCCGCAC
CTAATCAACTGGCCACATTTATGTAGAAATGACACCAGAACAG
TGATATTTTAGTTTACAATTGCTCACAAAGCTTCCAGAAGGTT
TCTAATGGAACAAACATTCATGAGATCACTTACAAGATCAACA
AGCGGAAGGTTGGGATGAAGTATGGAAAAACTATATAAGGAAA
AAGCCTCCATAAAAATTGGGCCTATAATCTTGGGGAAAAAAAA
AACATTTTGTACTTAGACAAAATTTGAGTAAATATAAGAACAA
CTCATCCTTGGATCAGACTAAGGAGTGTCATTCTCAAATTGTG
TTTTCTAACTAGTTTGATCTCTCTCAATTAAGGCCAATATTCG
ATTTACTAAAACCCTATTTCTAGTAAAATCCACTCTCTAATAA
ATATATTGTGTTGGGCTTGTTGGGCCACCATCAATTGTTAGAA
GGTTGAAGTGCAAATAGGTCCTTAAAAAAAGAGCACTTGAAAA
TATAGGCAAAAAGCTATATATAATATATATATAACTTAGAAGC
TAAAGGTTACTACCACACATCCTTGATGCGATTGTCACTCCAC
AAGTATAAGTGTTTGTGGGTTGTGCAGTGACGGACCTACAGTG
GGTAATGACTTCCCCCCACCCCCCCCCCCCCCCCCAAAAAAAA
ACTAGGATATATATATGTGTGTGTGTGTTGAAATTTTTTTAAAT
TTTGGTCTTAATAGCCCCCAATTAAATAAAAAAAAACTGATATA
TATATATATATATTTGGATCTCTTCTGAAATAGCGTTGAATTC
CATTTTGATAATTGTGTGGAAATGTTTAAGAAACACTTATTTT
ATATATTATACTTTATTGACATTTTTATAATATCAAATGTTTA
AGAAACACTTATTTTATATATAGTATTTTGTTGAATATTAAAT
AGATGTTAATTTTTATTTTCATAAAAAAAAAAAAATTGTCTTAA
GCTTTGGTCCCCCCTAGGTTCGAATCCTGATTCCGTCCCTGAG
```

- Human (*Homo sapiens*) genome – 3.2 billion base pairs (6.4 billion bases)
  - 6 feet of DNA per cell
  - 10 billion miles of DNA per person
- American Chestnut (*Castanea dentata*) genome - 793.5 million base pairs
- Chinese Chestnut (*Castanea mollissima*) genome – 785.5 million base pairs

The Central Dogma of Biology

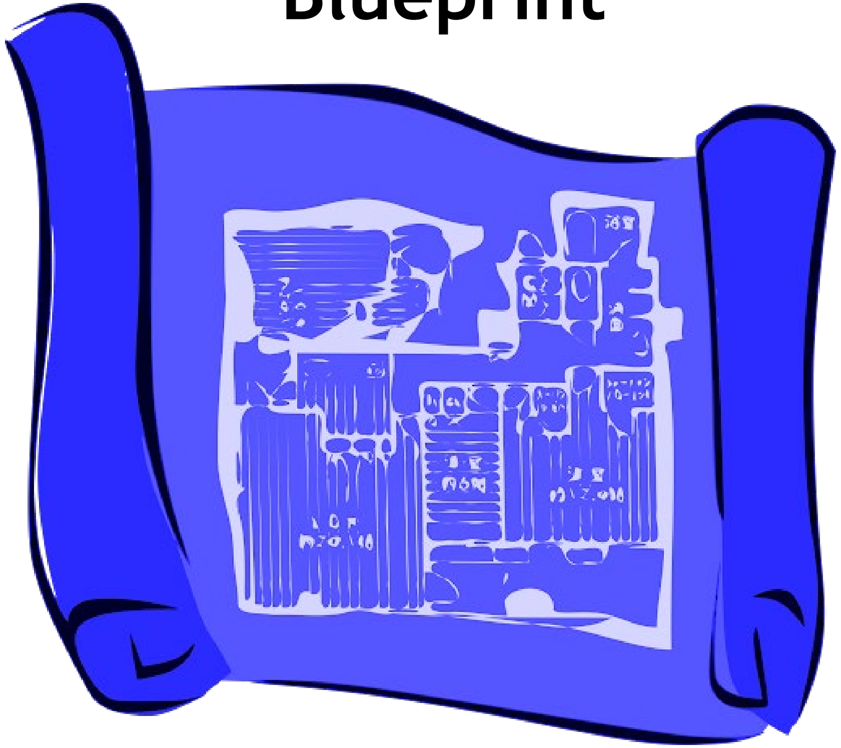


DNA mRNA Protein

How do you build a skyscraper?

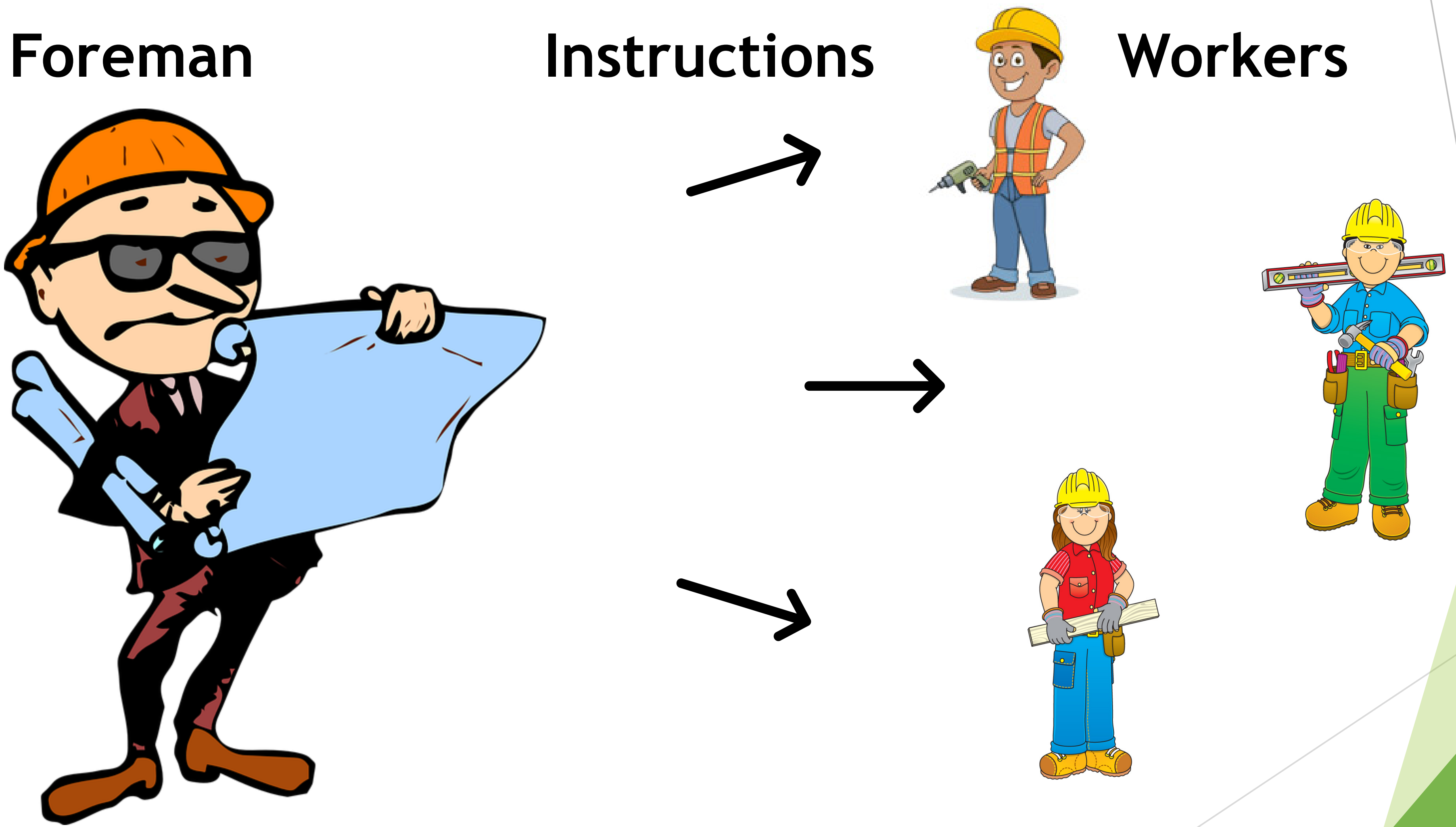
How do you build a skyscraper?

Blueprint



DNA

How do you build a skyscraper?



Foreman

Instructions

Workers

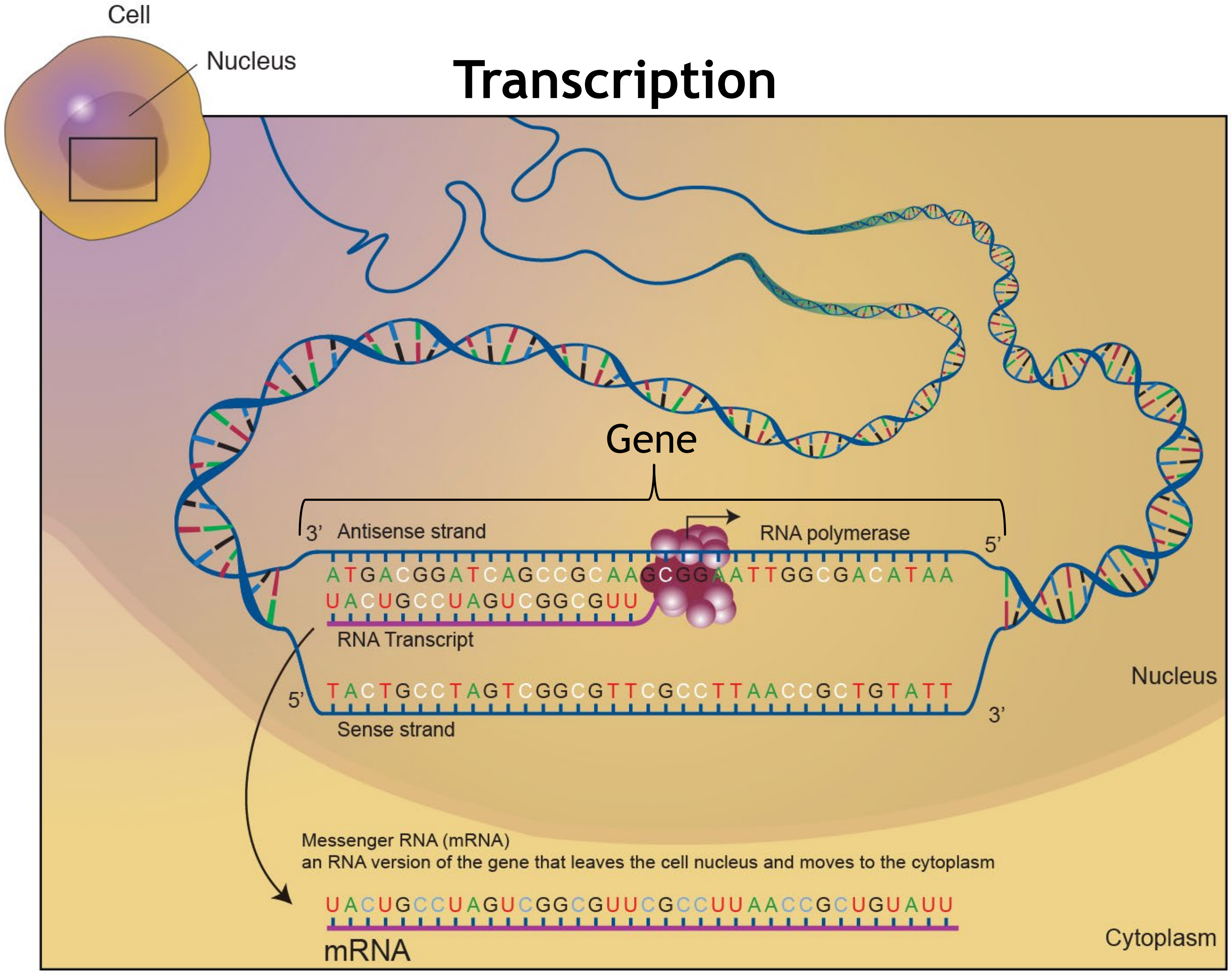
Regulation

mRNA

Proteins

How do you build a skyscraper?

So how does this actually work in a cell?

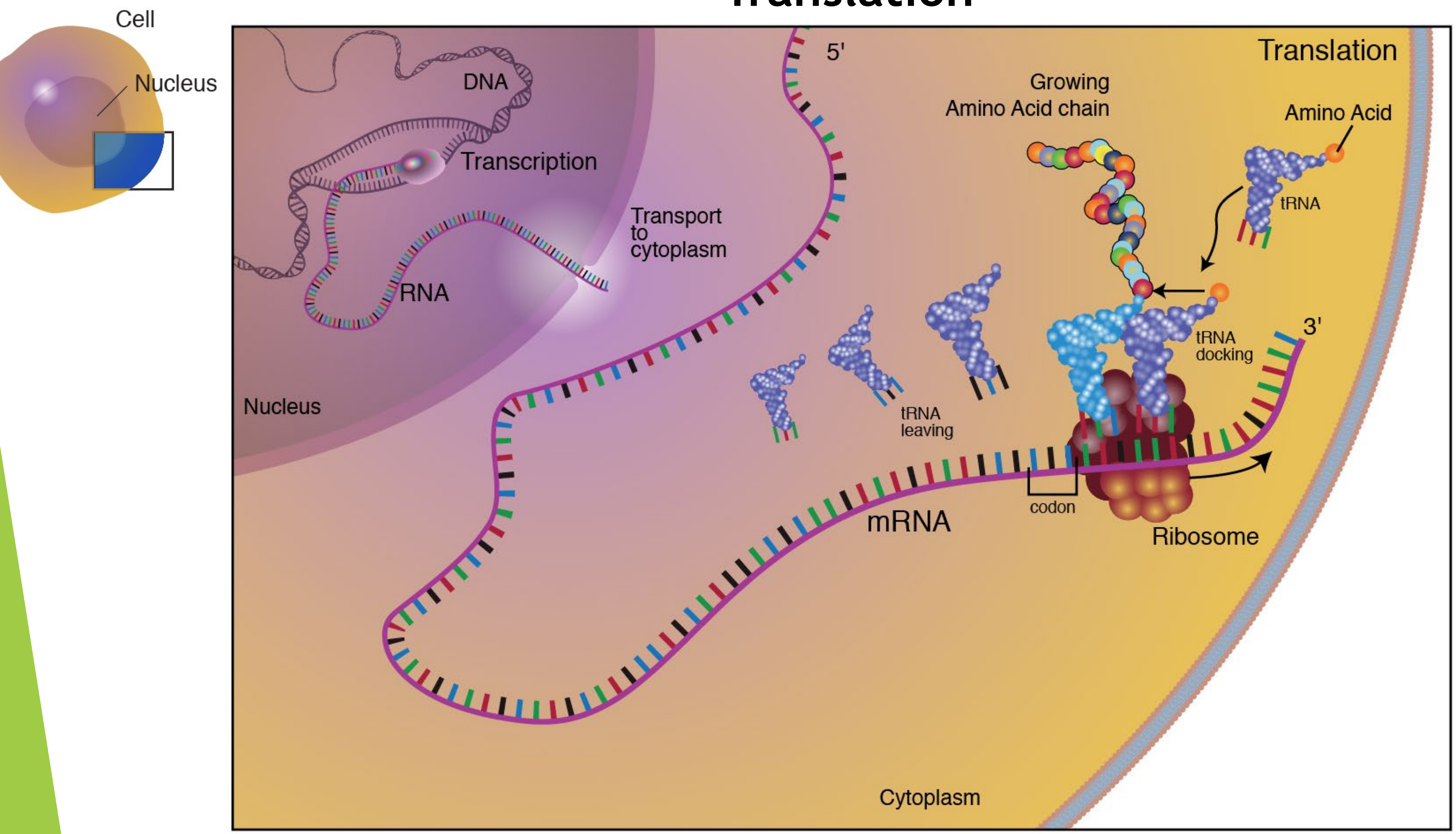


Messenger RNA (mRNA)
an RNA version of the gene that leaves the cell nucleus and moves to the cytoplasm

[Messenger RNA (mRNA) (genome.gov)]

So how does this actually work in a cell?

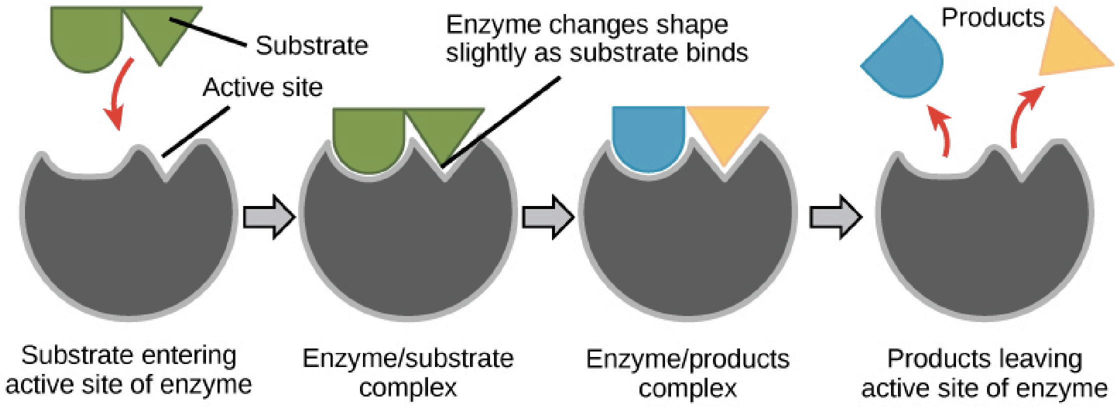
Translation

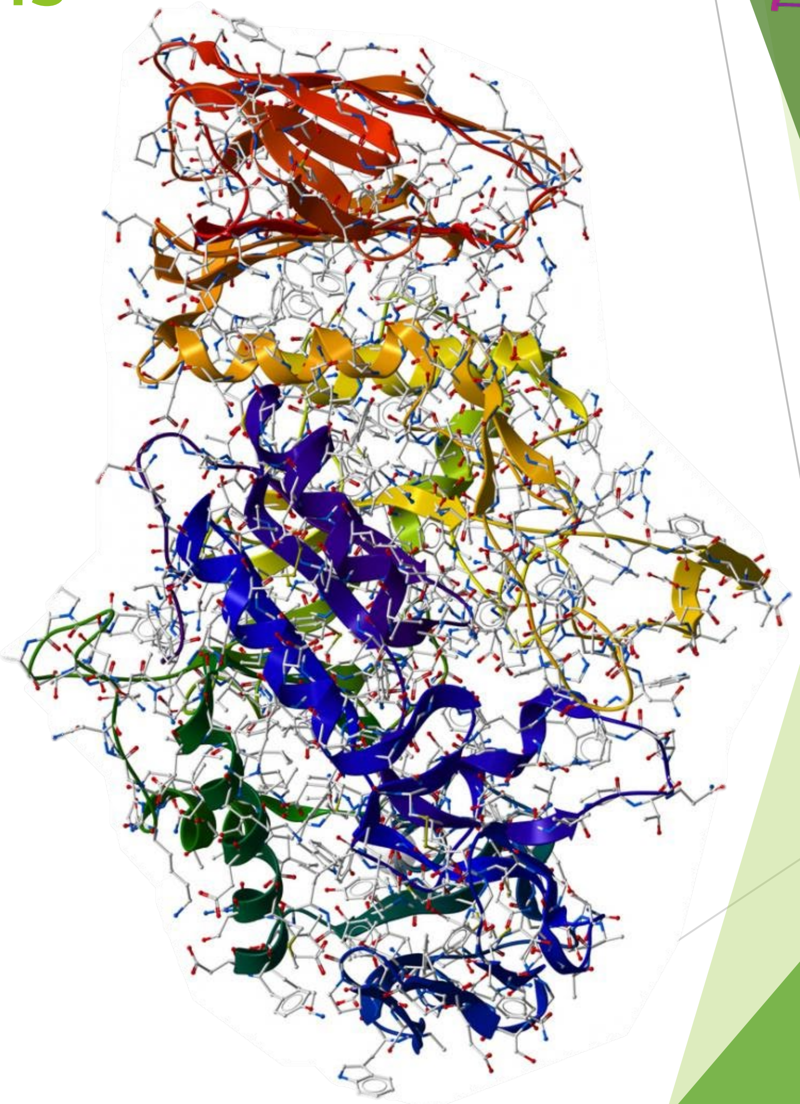


Translation (genome.gov)

A little more about proteins

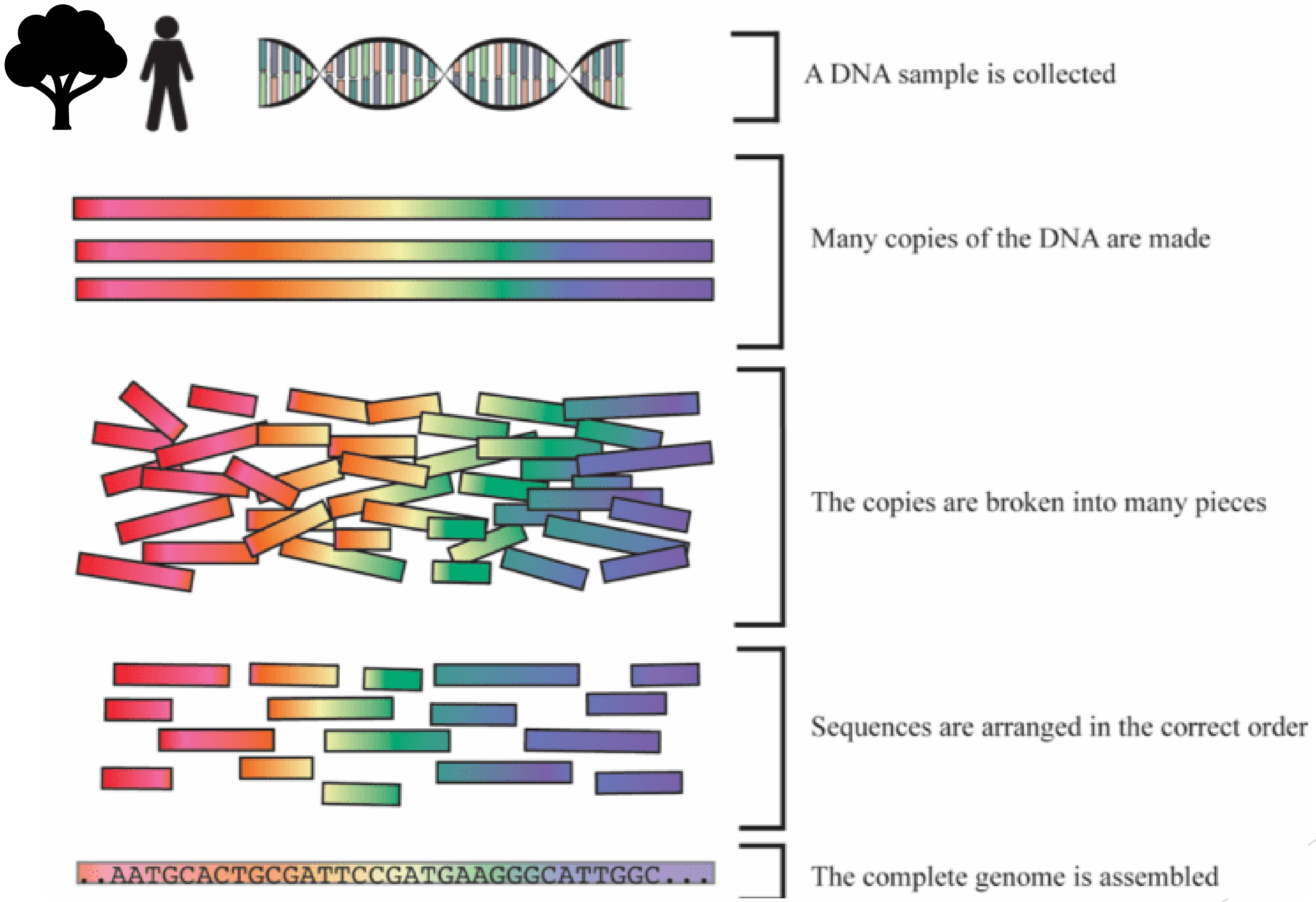
- There are 20(ish) different amino acids encoded by mRNA
- Combinations of amino acids can make incredibly complex molecules
- One primary function of proteins is to catalyze chemical reactions.
- Proteins that catalyze reactions are known as enzymes.


Substrate
Active site
Enzyme changes shape slightly as substrate binds
Products
Substrate entering active site of enzyme
Enzyme/substrate complex
Enzyme/products complex
Products leaving active site of enzyme



Proteinogenic amino acid – Wikipedia, Khan Academy

How do we sequence a genome?



A DNA sample is collected

Many copies of the DNA are made

The copies are broken into many pieces

Sequences are arranged in the correct order

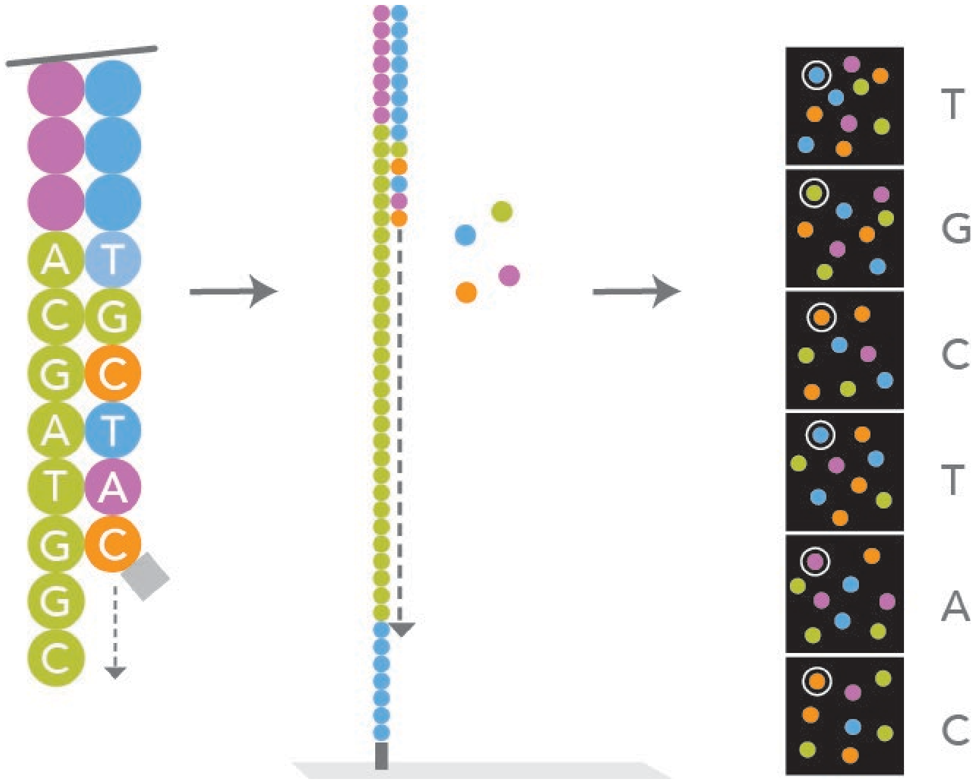
The complete genome is assembled

Two major technologies for DNA sequencing

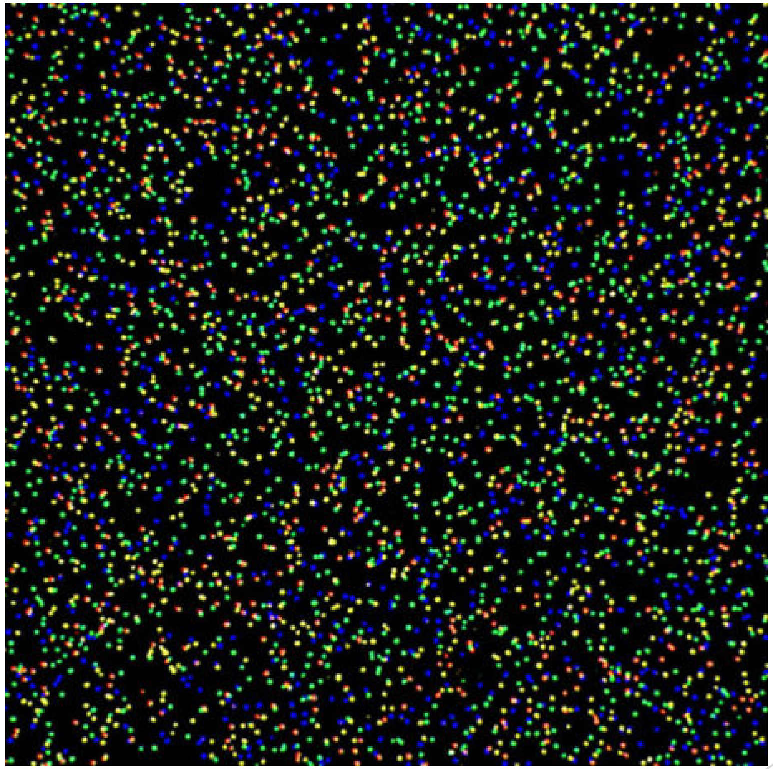
Short read sequencing
(Illumina)

Long read sequencing
(PacBio and Oxford Nanopore)

Short Read Sequencing



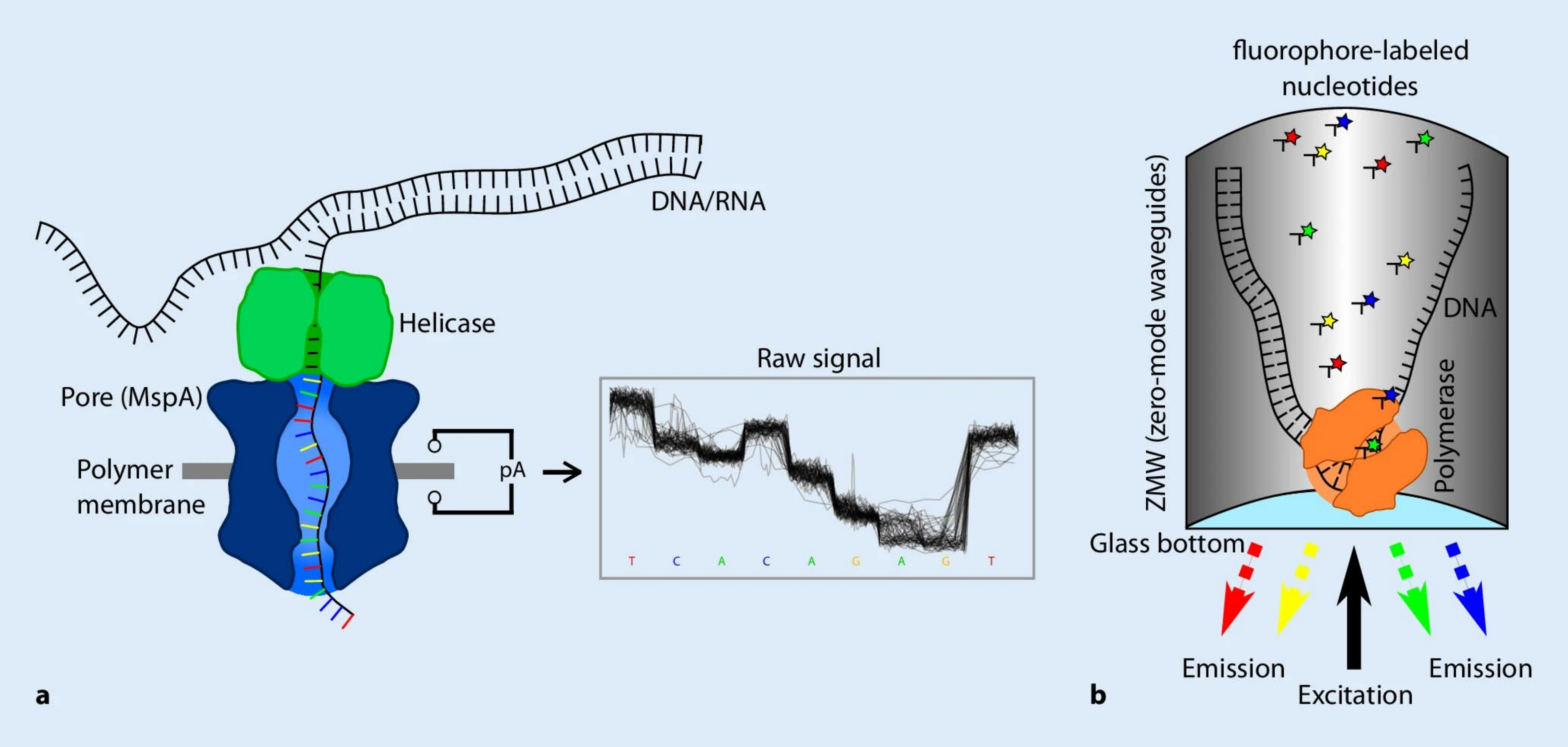
Illumina Sequencing



Generates billions of DNA “reads” that average 200 bases long

Sequencing Technology | Sequencing by synthesis (illumina.com)

Long Read Sequencing



Oxford NanoPore

PacBio SMRT

Generate hundreds of thousands of reads with average length of 30,000 bases

Kraft and Kurth 2019

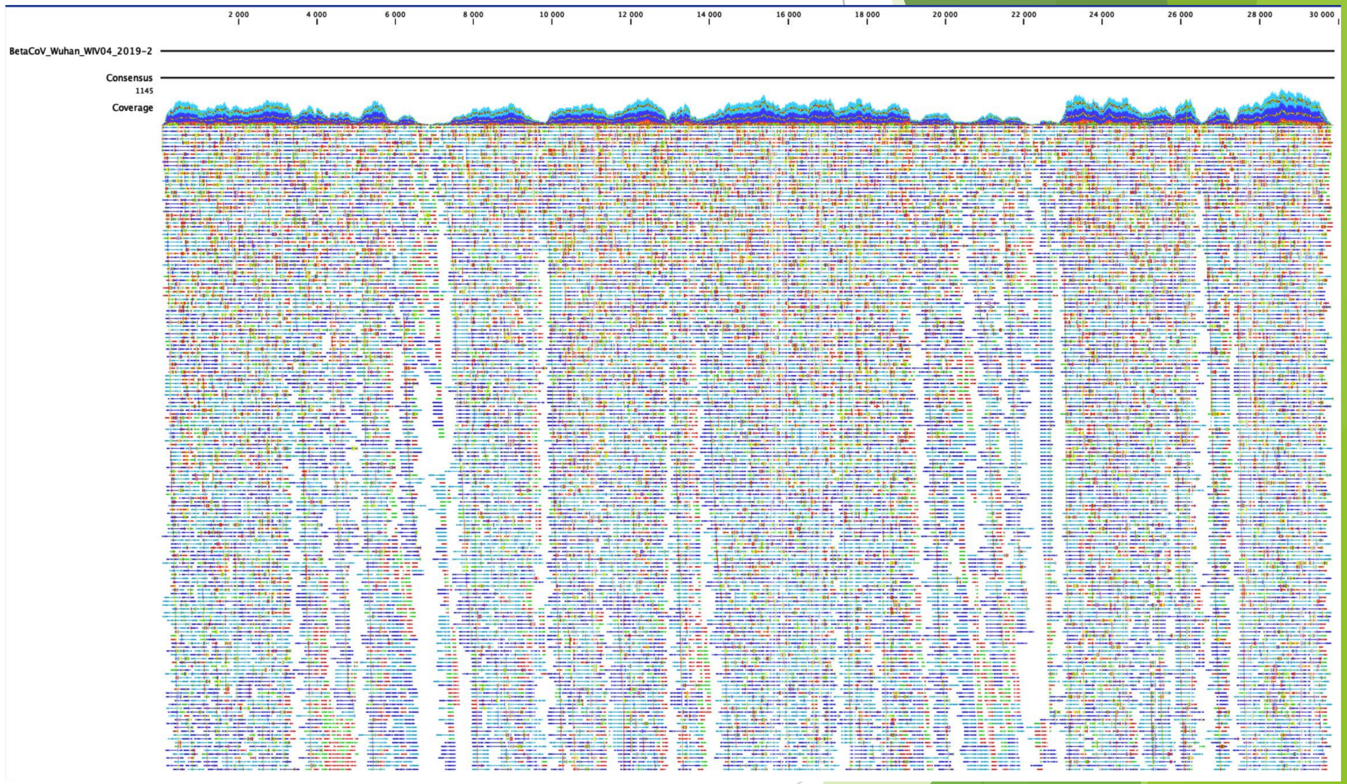
Short read vs. Long read sequencing

- Short read sequencing
  - Highly accurate, 0.1-0.5% error rate
  - Shorter DNA reads requires more reactions to get enough sequence coverage
  - Assembly of the many small reads is more difficult
  - Cannot sequence or assemble repetitive regions (200 As in a row)
  - Less expensive per base pair

- Long read sequencing
  - Less accurate, up to 15% error rate but has been improved with newer versions
  - Can sequence across highly repetitive regions
  - Easier to assemble genomes from long reads but requires more coverage to improve accuracy
  - More expensive per base pair

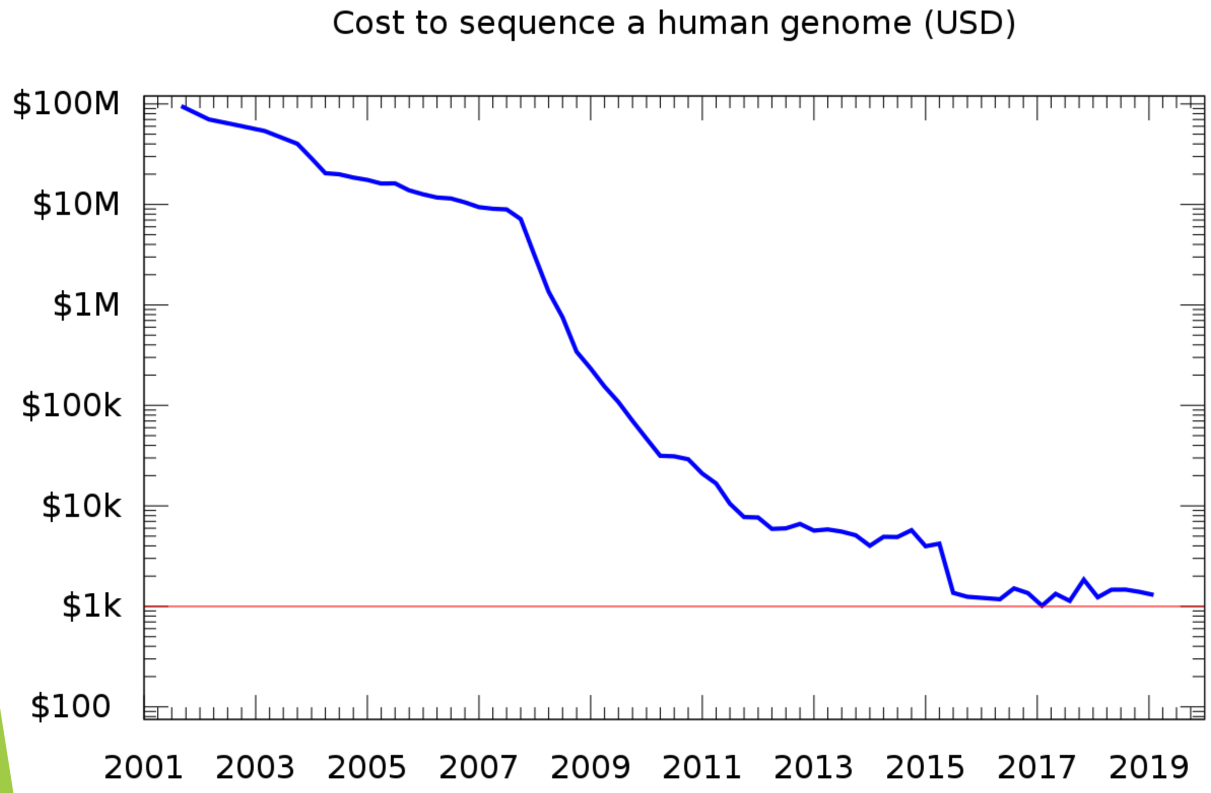
More about genome quality

- Poor quality genomes lead to bad data and frustrated scientists
- Generally related to the amount of sequence generated (coverage)
- The Three C's of Genome Quality:
  - Contiguity – How much of your sequence is assembled into big pieces?
  - Completeness – How much of the genome were you able to sequence?
  - Correctness – How accurate is your genome sequence?



Beyond Contiguity – Assessing the Quality of Genome Assemblies with the 3 C's - PacBio

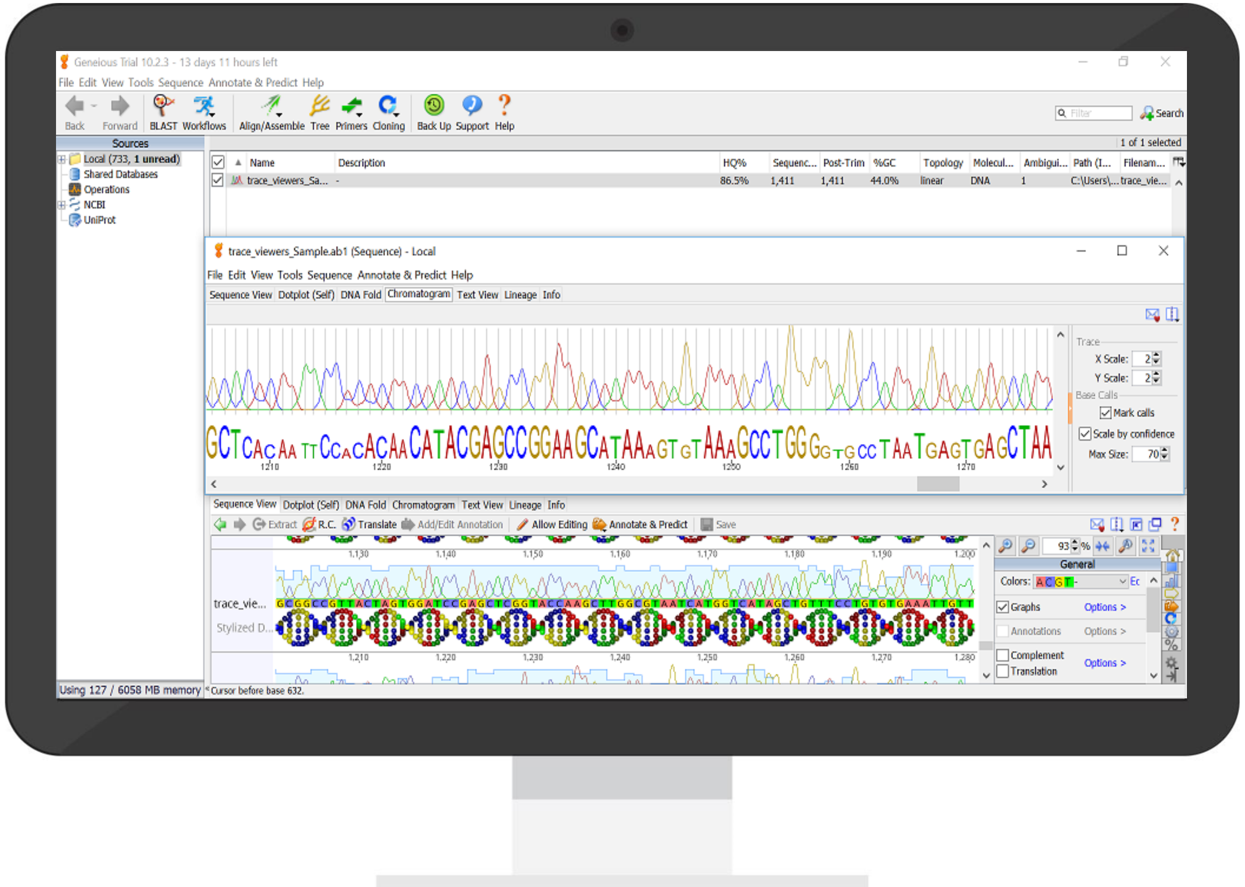
Progress in sequencing technologies has been rapid



- Cost of sequencing continues to go down
- New technologies are coming on to the market with increased accuracy and lower prices
- New computational tools are developed all the time to increase what we can predict from a genome sequence

What can we do with a genome sequence?

A genome sequence allows researchers to make new hypotheses to test in the lab.



Geneious Trial 10.2.3 - 13 days 11 hours left
File Edit View Tools Sequence Annotate & Predict Help
Back Forward BLAST Workflows Align/Assemble Tree Primers Cloning Back Up Support Help
Search
Sources
Local (733, 1 unread)
Shared Databases
Operations
NCBI
UniProt
1 of 1 selected
trace_viewers_Sample.ab1 (Sequence) - Local
Sequence View Dotplot (Self) DNA Fold Chromatogram Text View Lineage Info
Trace
Base Calls
Extract R.C. Translate Add/Edit Annotation Allow Editing Annotate & Predict Save
General
Colors: ACGT
Options >

☑	Name	Description	HQ%	Sequenc...	Post-Trim	%GC	Topology	Molecul...	Ambigui...	Path (I...	Filenam...
☑	trace_viewers_Sa...	-	86.5%	1,411	1,411	44.0%	linear	DNA	1	C:\Users\...	trace_vie...

X Scale: 2
Y Scale: 2
☑ Mark calls
☑ Scale by confidence
Max Size: 70

☑ Graphs
□ Annotations
□ Complement
□ Translation

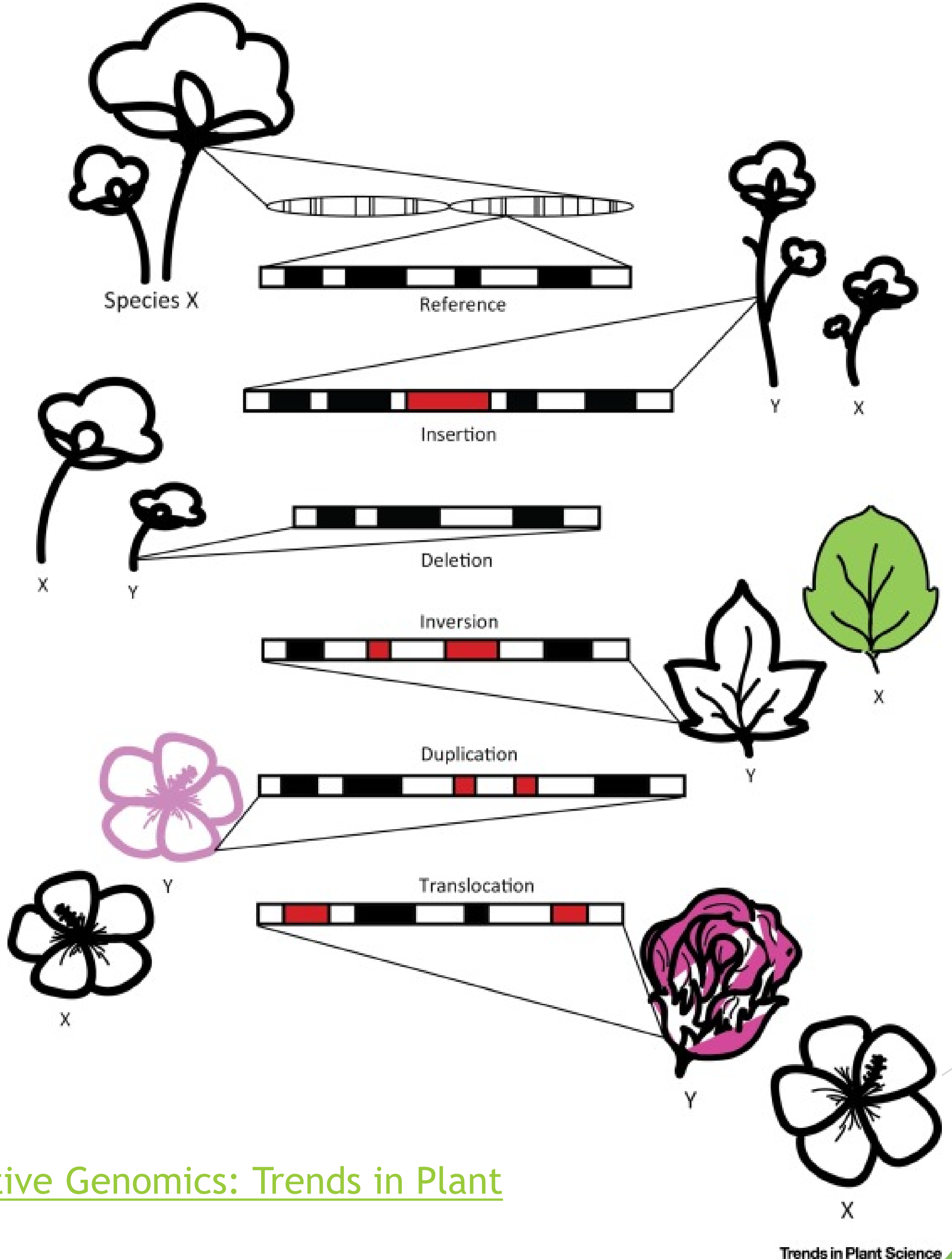
Using 127 / 6058 MB memory

Cursor before base 632.

- Bioinformatics (*noun*) - the science of collecting and analyzing complex biological data such as genetic codes
- Computational tools allow researchers to analyze DNA sequences to identify genes and predict protein sequences
- We can also predict regulation, what turns a gene or function on or off from DNA
- There are still a lot of gaps in our knowledge, and our models are far from perfect.

Comparative Genomics Hypotheses

- Comparing genomes of closely related species allow us to identify unique regions
- Experiments in the lab can be designed to test the hypotheses predicted by the genome



[Genome Mapping in Plant Comparative Genomics: Trends in Plant Science (cell.com)]

Review

1. What is DNA?
   - DNA is a large molecule that encodes the blueprint of living organisms
2. What is the Central Dogma of Biology?
   - The Central Dogma of Biology is the sequential transfer of biological information from DNA to mRNA to Protein.
3. How do we sequence a genome?
   - There are multiple technologies that “read” the individual bases of DNA and assemble the pieces into large contiguous genomes.
4. What can we do with a genome sequence?
   - Genome sequences allow researchers to identify proteins encoded in genes, predict organism functions, make connections between species, and generate new hypotheses to test in the lab.

Thank you!