CSC 2417 Course Overview
&
Introduction to Genomics

Dr. Jared Simpson
Ontario Institute for Cancer Research
&
Department of Computer Science
University of Toronto

Course Overview

- Course webpage: https://jts.github.io/csc2417/
  - Please sign up for the google group linked there
  - Note that there is no lecture next week - I'm away at a conference
- Course materials:
  - No textbook but we follow Ben Langmead's lectures: http://www.langmead-lab.org/teaching-materials/
- Grading:
  - 60% - three assignments
  - 40% - written course project on topic of your choice

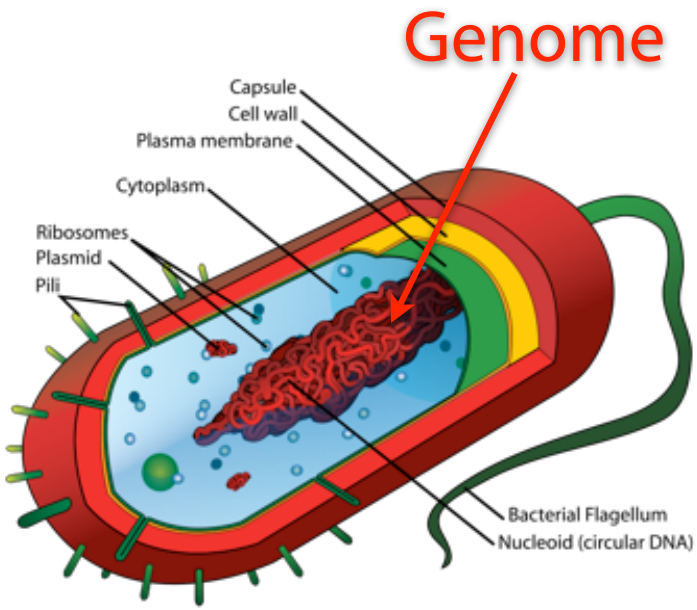
How to find me

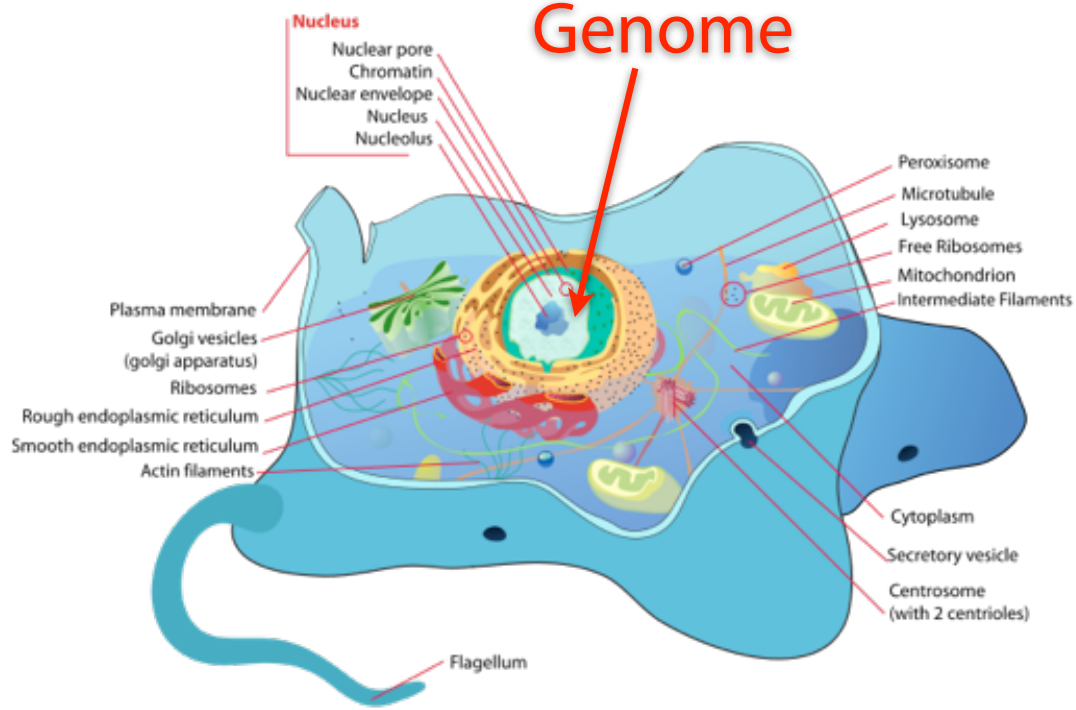
Course Content

- Focus of this course is the methods used to process genomic data
  - Explosion of data in the last decade made efficient computational methods a major issue in the field
  - This touched all areas of biology but mostly in the field of genome sequencing
  - Little to no biology background is needed - this lecture will cover the essentials, wikipedia will cover the rest

Part 1.
What is a genome?

Cells





Prokaryotic cell

A bacterium consists of a single prokaryotic cell

Eukaryotic cell
(pictured: animal cell)

Make up animals, plants, fungi, other eukaryotes

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Pictures: http://en.wikipedia.org/wiki/Cell_(biology)

The central dogma of molecular biology

Short version:

DNA -> RNA -> Protein

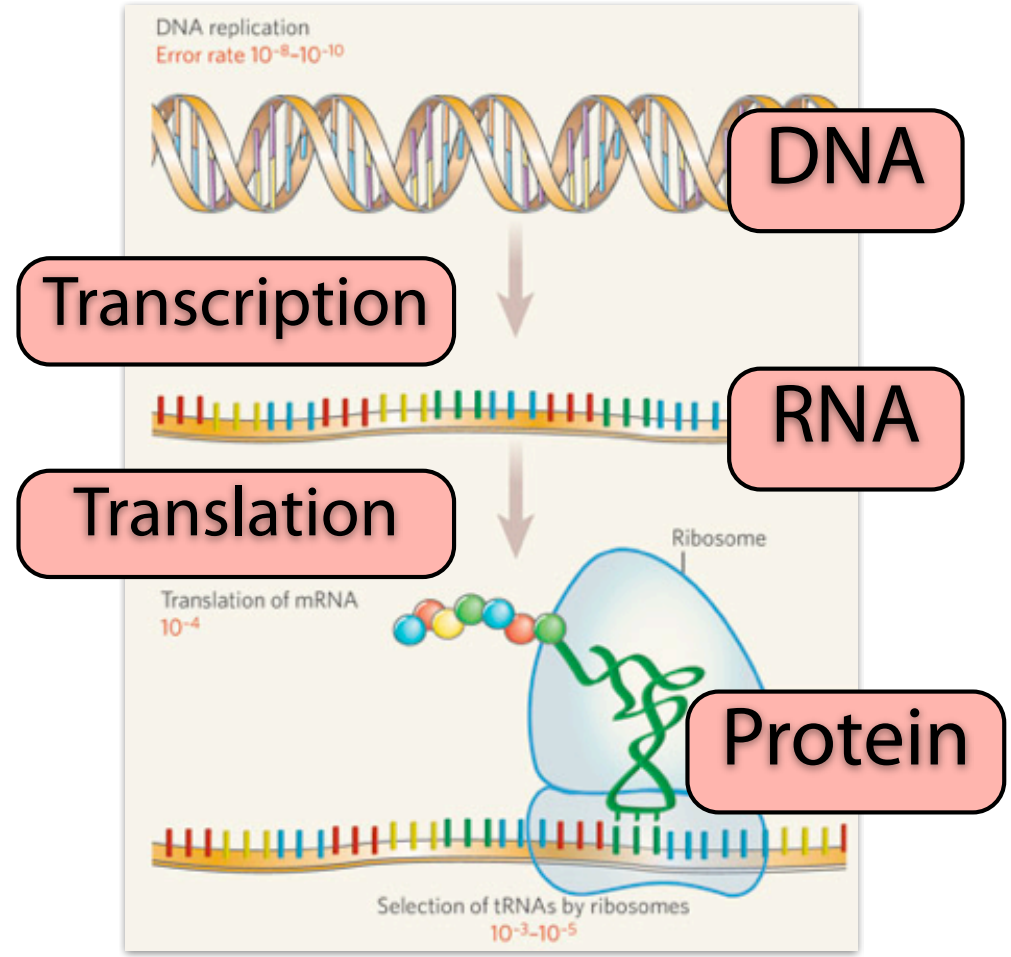
Long version:

DNA molecules contain information about how to create proteins; this information is *transcribed* into RNA molecules, which, in turn, direct chemical machinery which *translates* the nucleic acid message into a protein.

Hunter, Lawrence. "Life and its molecules: A brief introduction." *AI Magazine* 25.1 (2004): 9.

Links genotype and phenotype

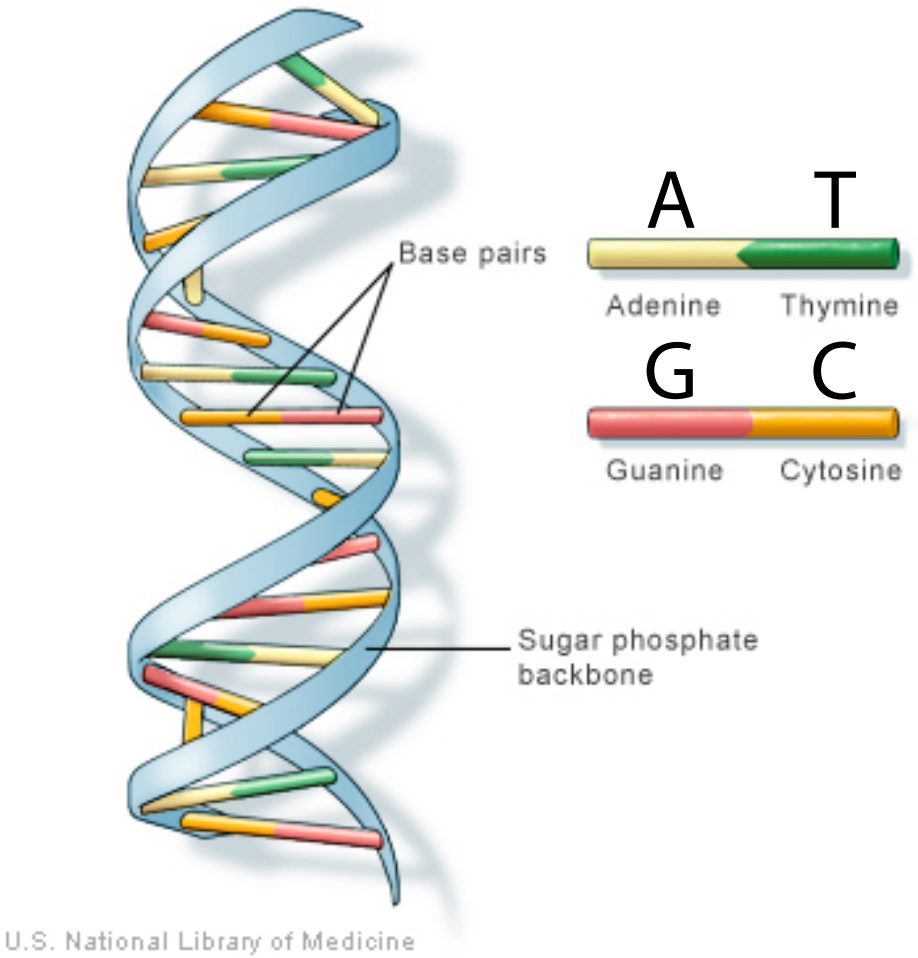
First stated by Francis Crick in 1958



Picture from: Roy H, Ibba M. Molecular biology: sticky end in protein synthesis. Nature. 2006 Sep 7;443(7107):41-2.

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

DNA: the genome's molecule



Picture: http://ghr.nlm.nih.gov/handbook/basics/dna

Deoxyribonucleic acid

"Rungs" of DNA double-helix are base pairs. Pair combines two complementary bases.

Complementary pairings: A-T, C-G

Single base also called a "nucleotide"

DNA fragment lengths are measured in "base pairs" (abbreviated bp), "bases" (b) or "nucleotides" (nt)

Stringizing DNA

DNA has *direction* (a *5′ head* and a *3′ tail*). When we write a DNA *string*, we follow this convention.

When we write a DNA string, we write just one strand. The other strand is its *reverse complement*.

To get reverse complement, reverse then complement nucleotides (i.e. interchange A/T and C/G)

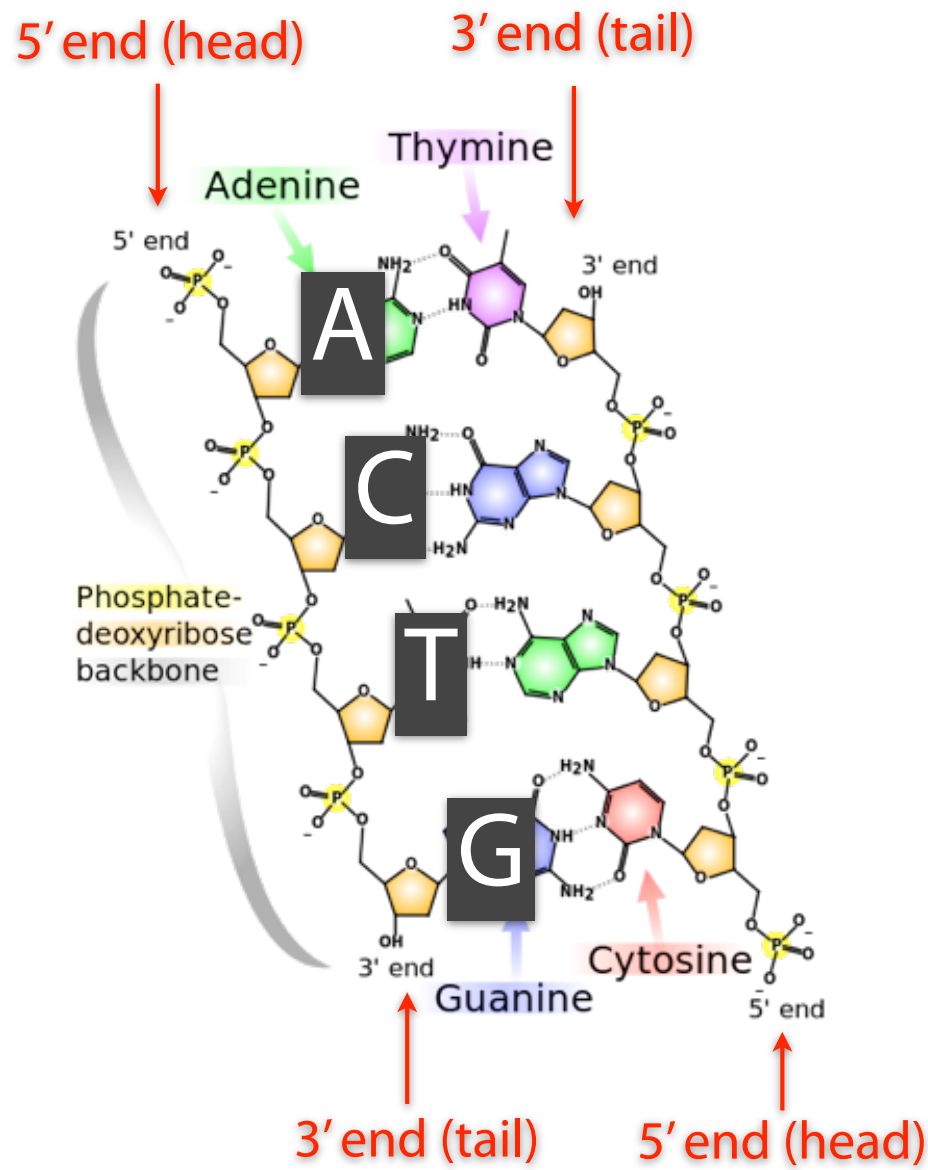
5′ end A C T G *3′ end*

↑

reverse complement

↓

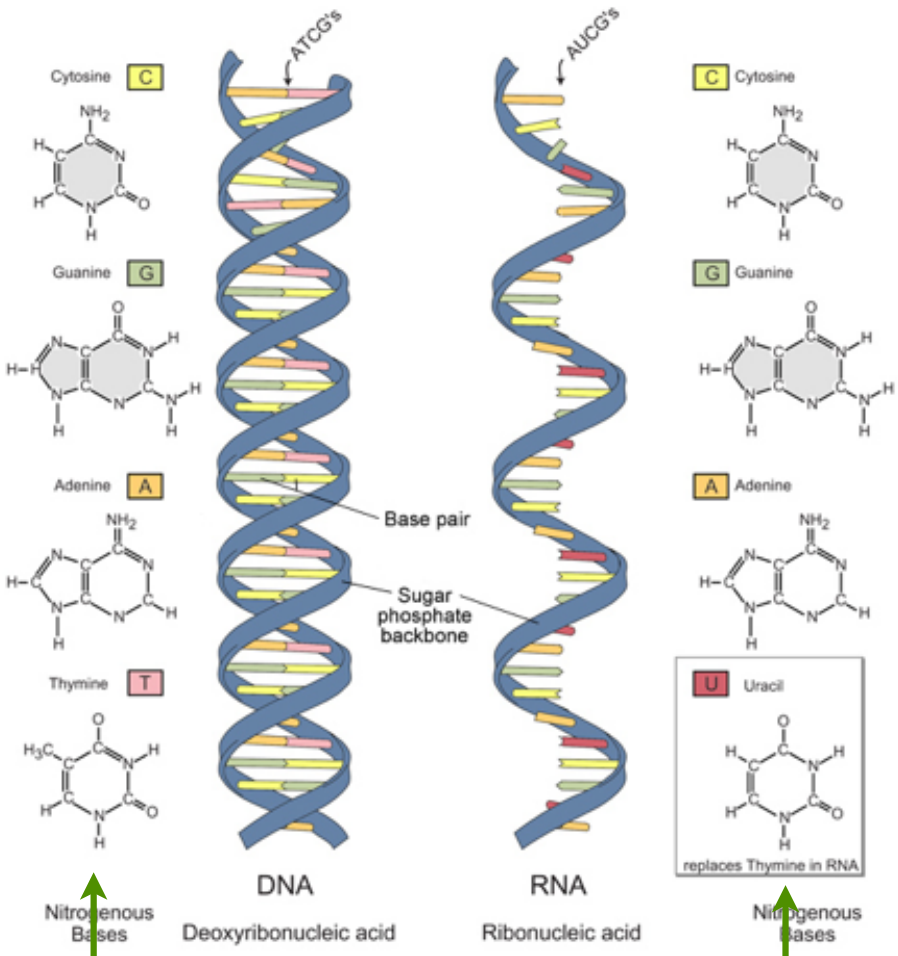
5′ end C A G T *3′ end*



Picture: http://en.wikipedia.org/wiki/DNA

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

RNA

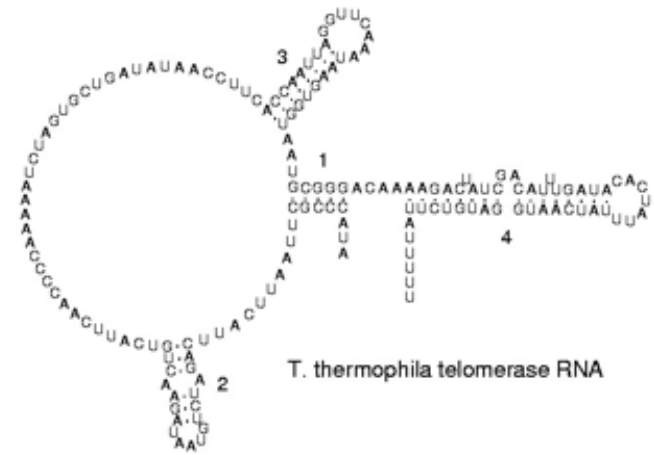


Like DNA but:

Single-stranded

Uses Uracil (U) instead of Thymine (T)

Sugar in the backbone is ribose instead of deoxyribose

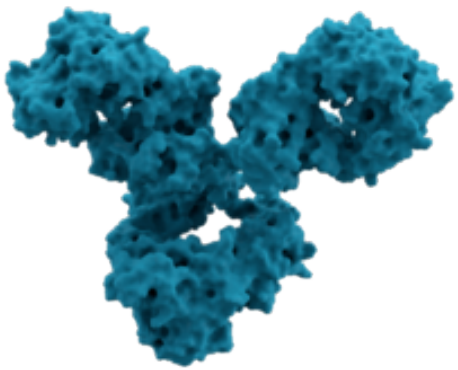


Picture: http://en.wikipedia.org/wiki/Rna

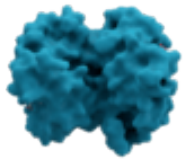
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Proteins

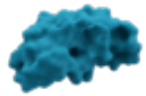
Proteins are typically 100s or 1000s of amino acids long, and fold into exquisitely complicated shapes



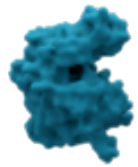
Immunoglobulin



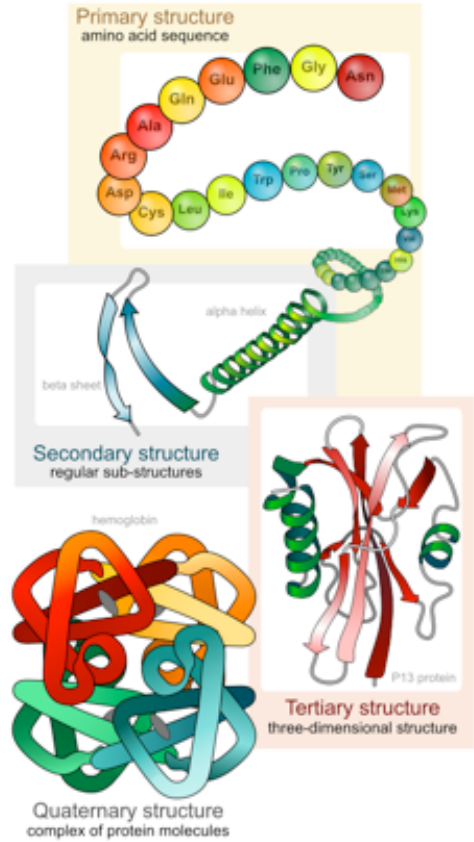
Hemoglobin



Insulin



Adenylate Kinase



Proteins perform a vast array of functions within living organisms: catalyzing metabolic reactions, replicating DNA, transporting molecules from one location to another, etc

Sources: http://en.wikipedia.org/wiki/Protein, http://en.wikipedia.org/wiki/Protein_structure

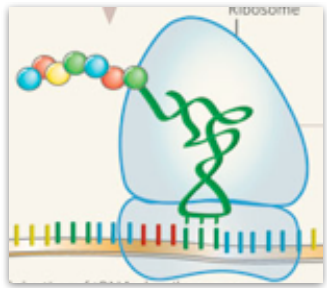
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

The Central Dogma: Genetic code

DNA codes for protein, but DNA alphabet has 4 nucleic acids, whereas protein alphabet has ~20 amino acids

A *triplet* of nucleic acids (*codon*) codes for one amino acid

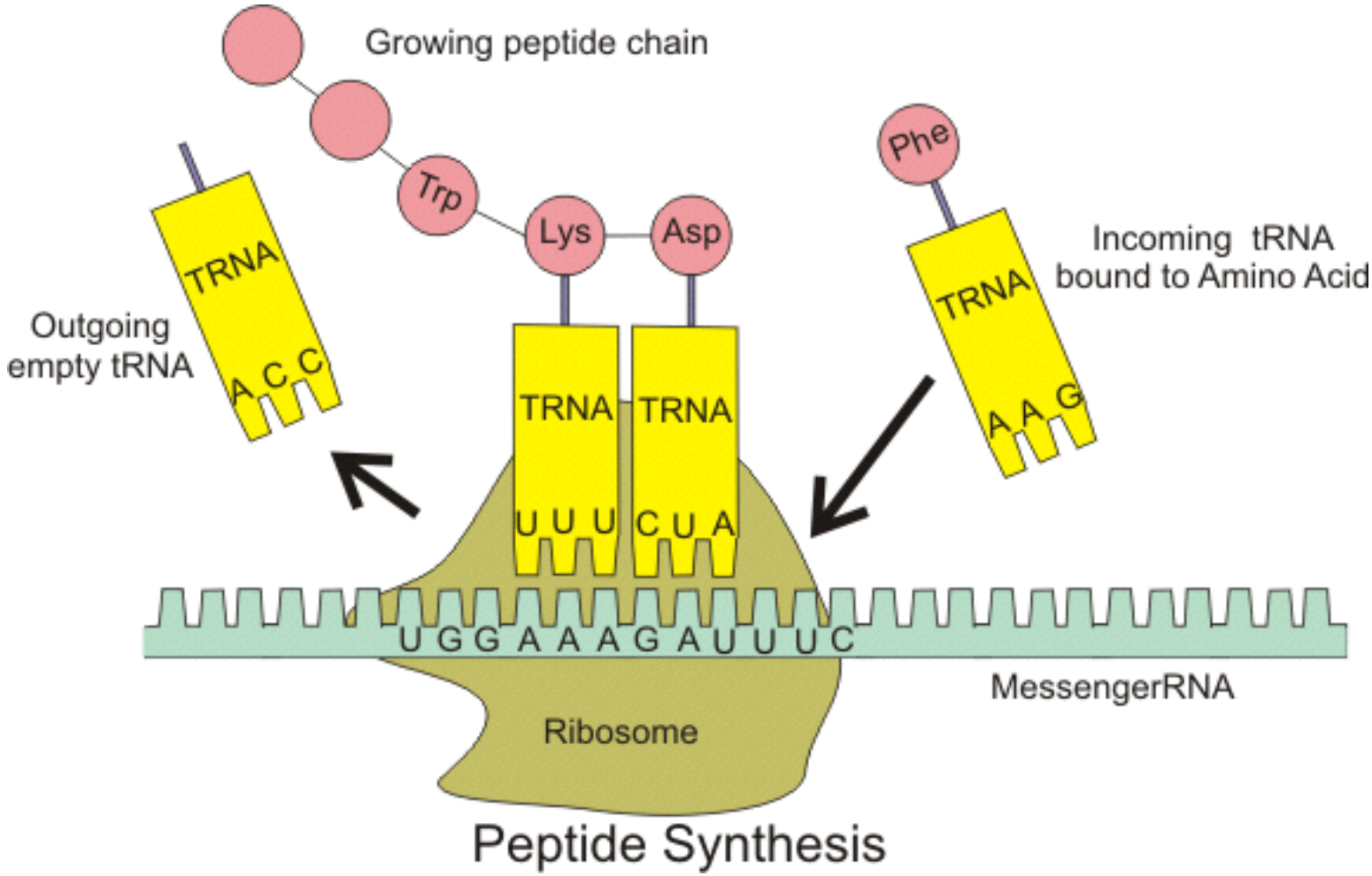
The code is *redundant*. E.g., both GGC and GGA code for Gly (Glycine)



First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

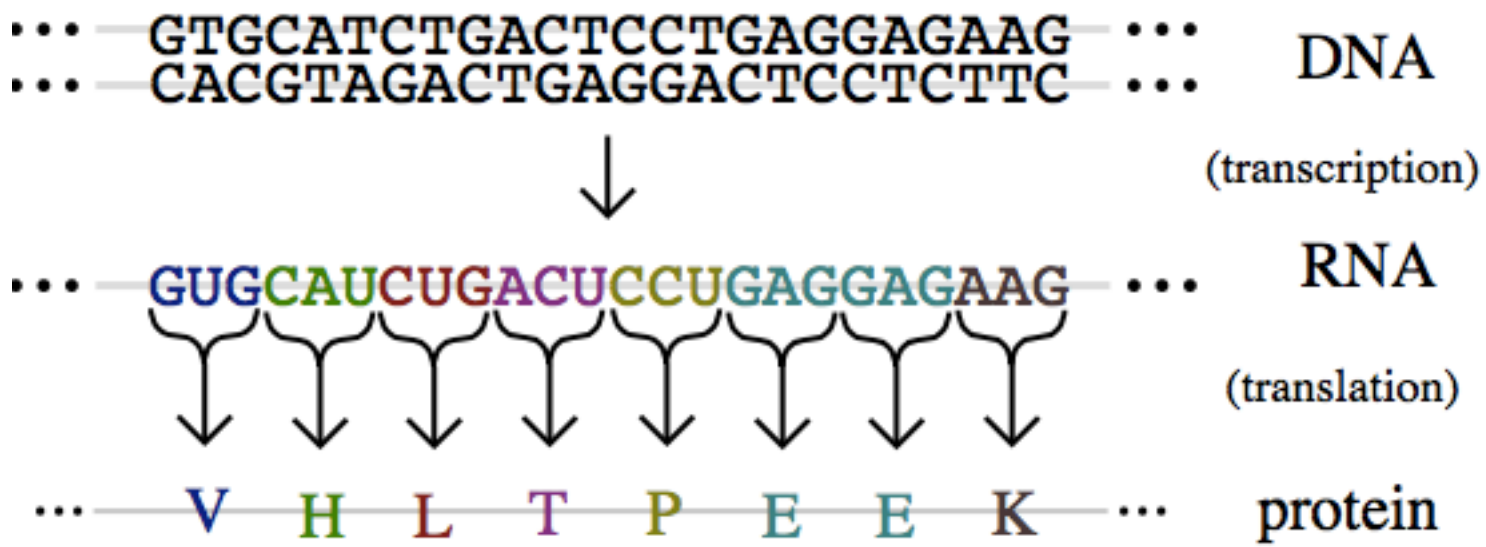
Picture: http://www.mun.ca/biology/scarr/MGA2_03-20.html

Protein Synthesis



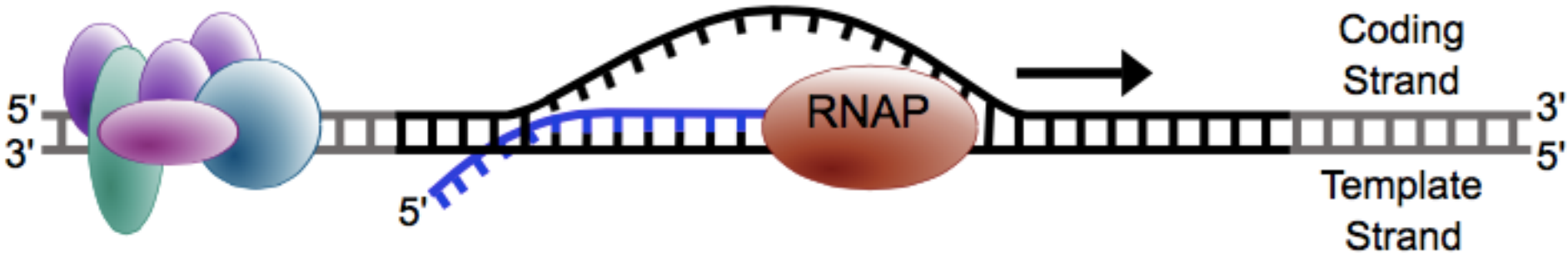
From https://en.wikipedia.org/wiki/Transfer_RNA

Transcription and Translation



From https://en.wikipedia.org/wiki/Gene_Expression

Gene Expression

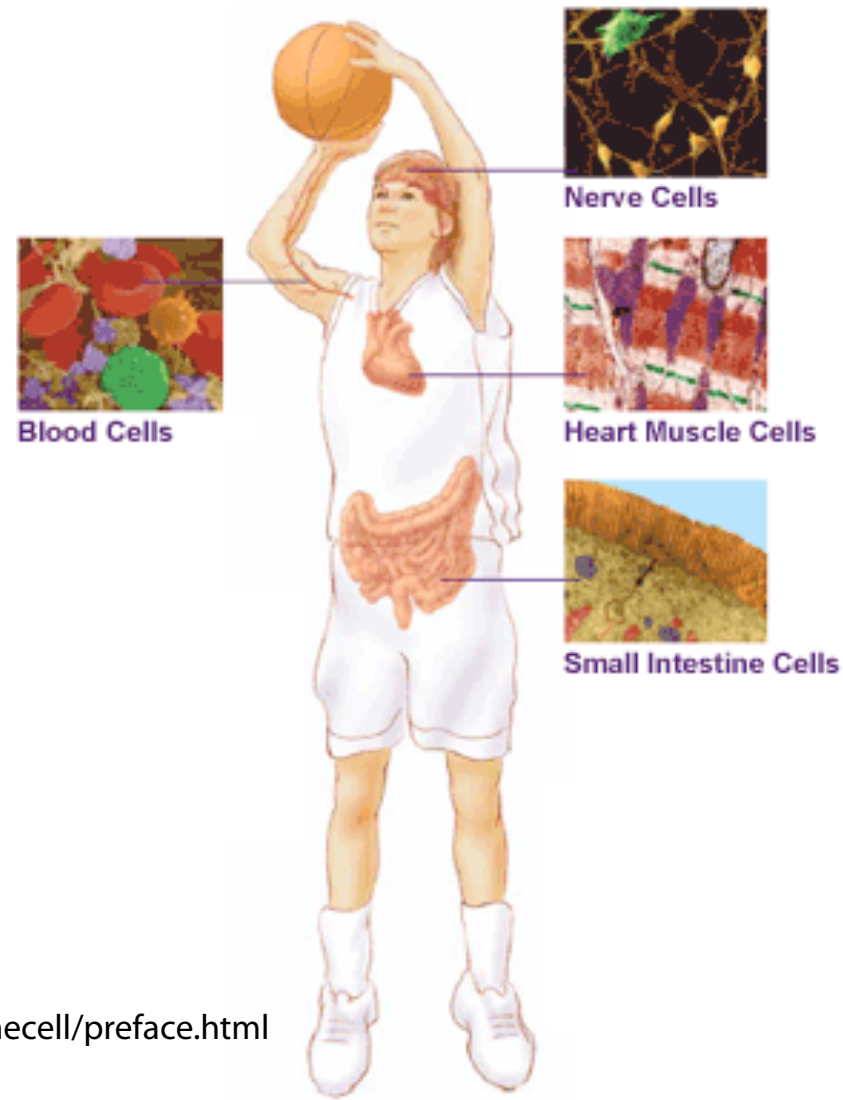


From https://en.wikipedia.org/wiki/Gene_Expression

Cell Types

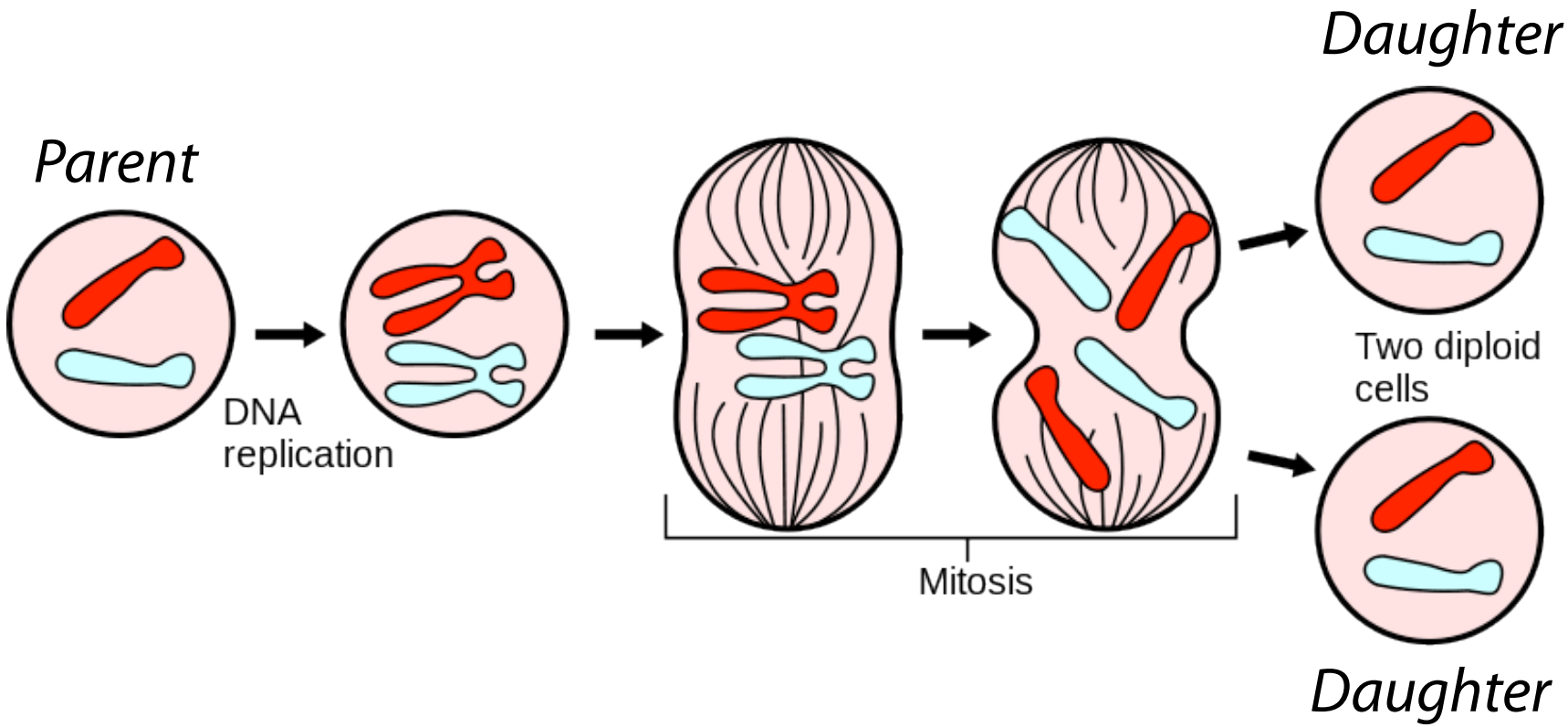
All the trillions of cells in a person have same genomic DNA in the nucleus

Variation in cell types and function is controlled by what proteins are expressed and how much



Picture: http://publications.nigms.nih.gov/insidethecell/preface.html

Cells: division



During cell division (*mitosis)*, the genome is copied

Picture: http://en.wikipedia.org/wiki/Mitosis

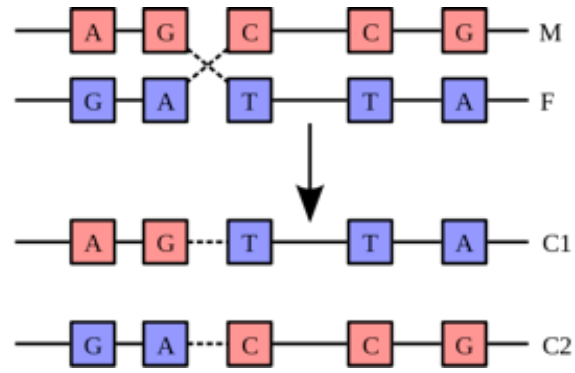
Evolution: why *these* genotypes?

Organisms reproduce, offspring *inherit* genotype from parents

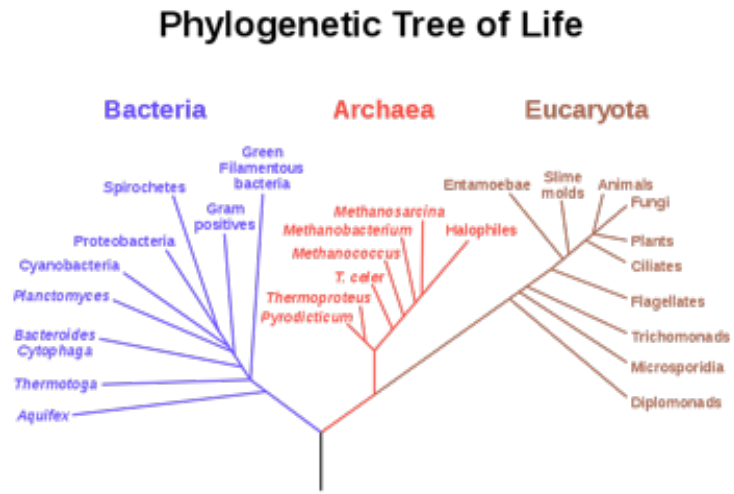
Random *mutation* changes genotypes and *recombination* shuffles chunks of genotypes together in new combinations

Natural *selection* favors phenotypes that reproduce more

Over time, this yields the variety of life on Earth. Incredibly, all organisms share a common ancestor.



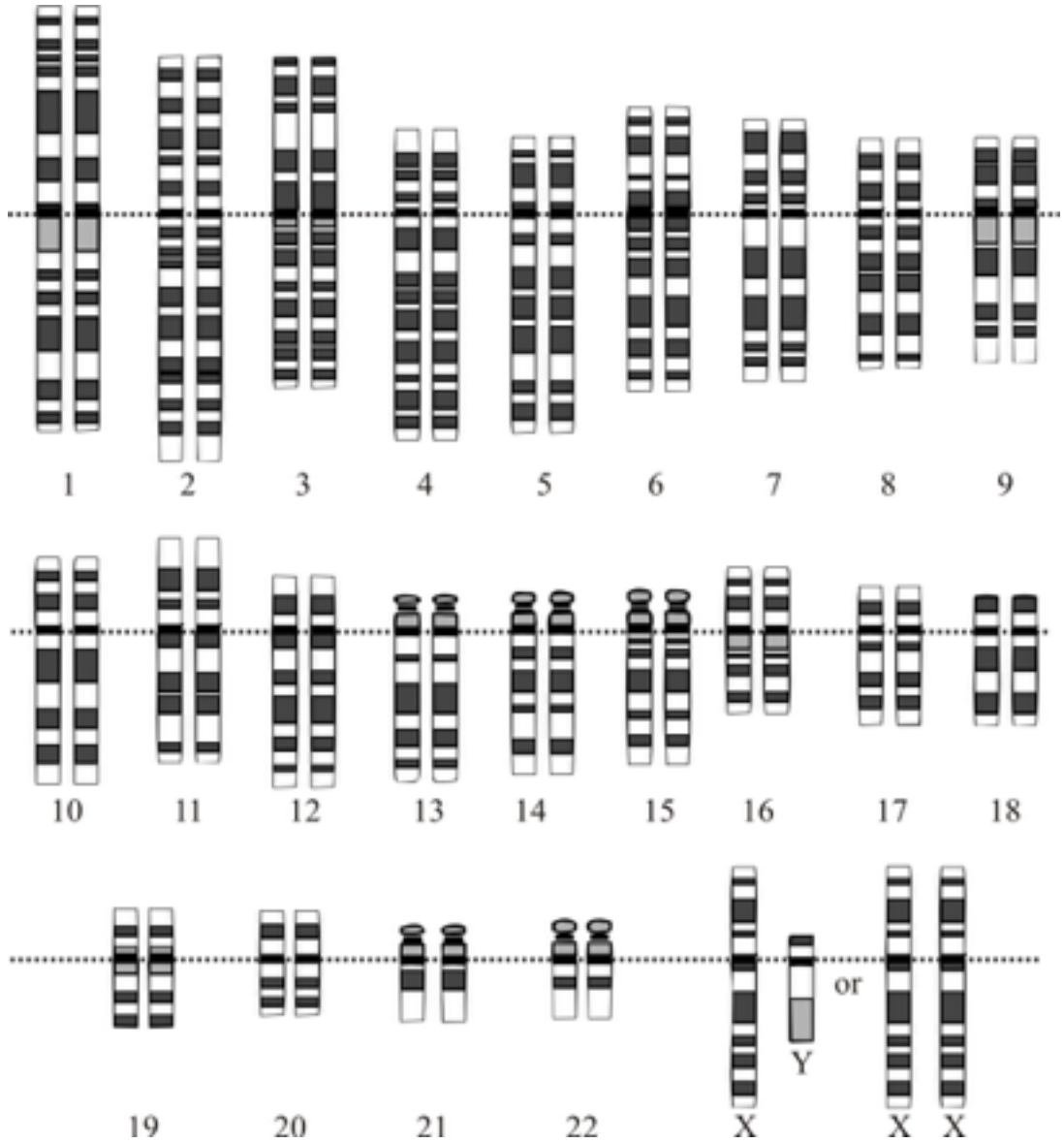
http://en.wikipedia.org/wiki/Genetic_recombination



http://en.wikipedia.org/wiki/Evolutionary_tree

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

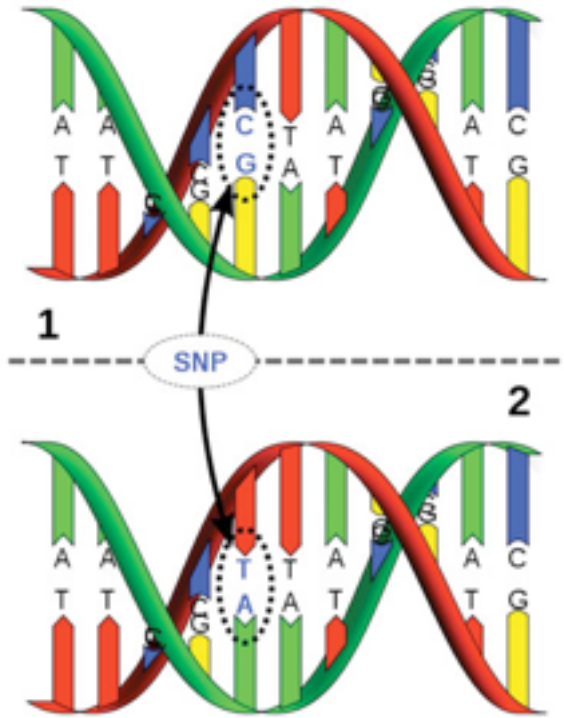
Human Genome



From https://en.wikipedia.org/wiki/Human_genome

The genome: variation

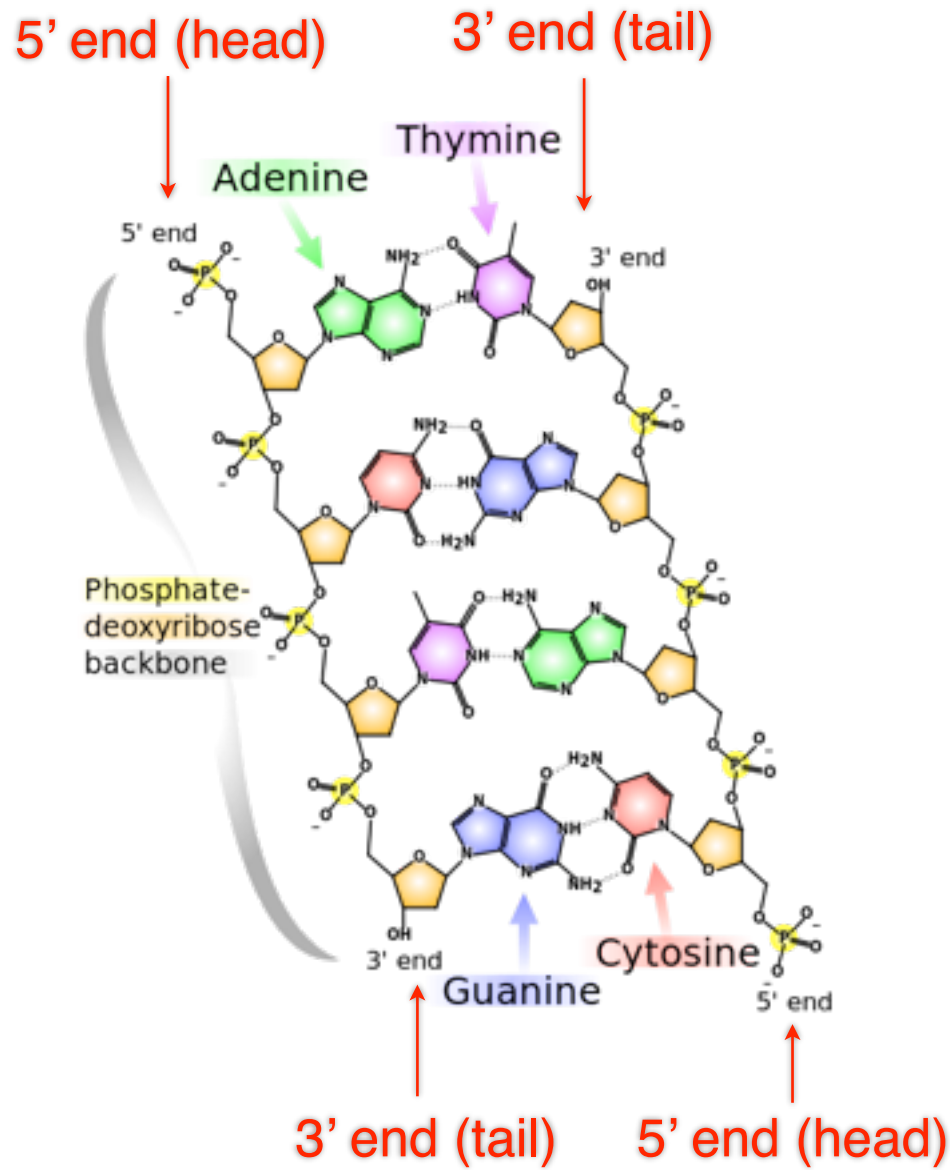
Two unrelated humans have genomes that are ~99.8% similar by sequence. There are about 3-4 million differences. Most are small, e.g. Single Nucleotide Polymorphisms (SNPs).



Summary

- Flow of information: DNA -> RNA -> Protein
- Genome: complete set of genetic information in the cell
- Individuals vary in their genome sequence
  - These differences determine phenotypic “visible” variation
- We want to study these differences by *sequencing* the genome

Part 2.
DNA Sequencing from 1977 to 2016

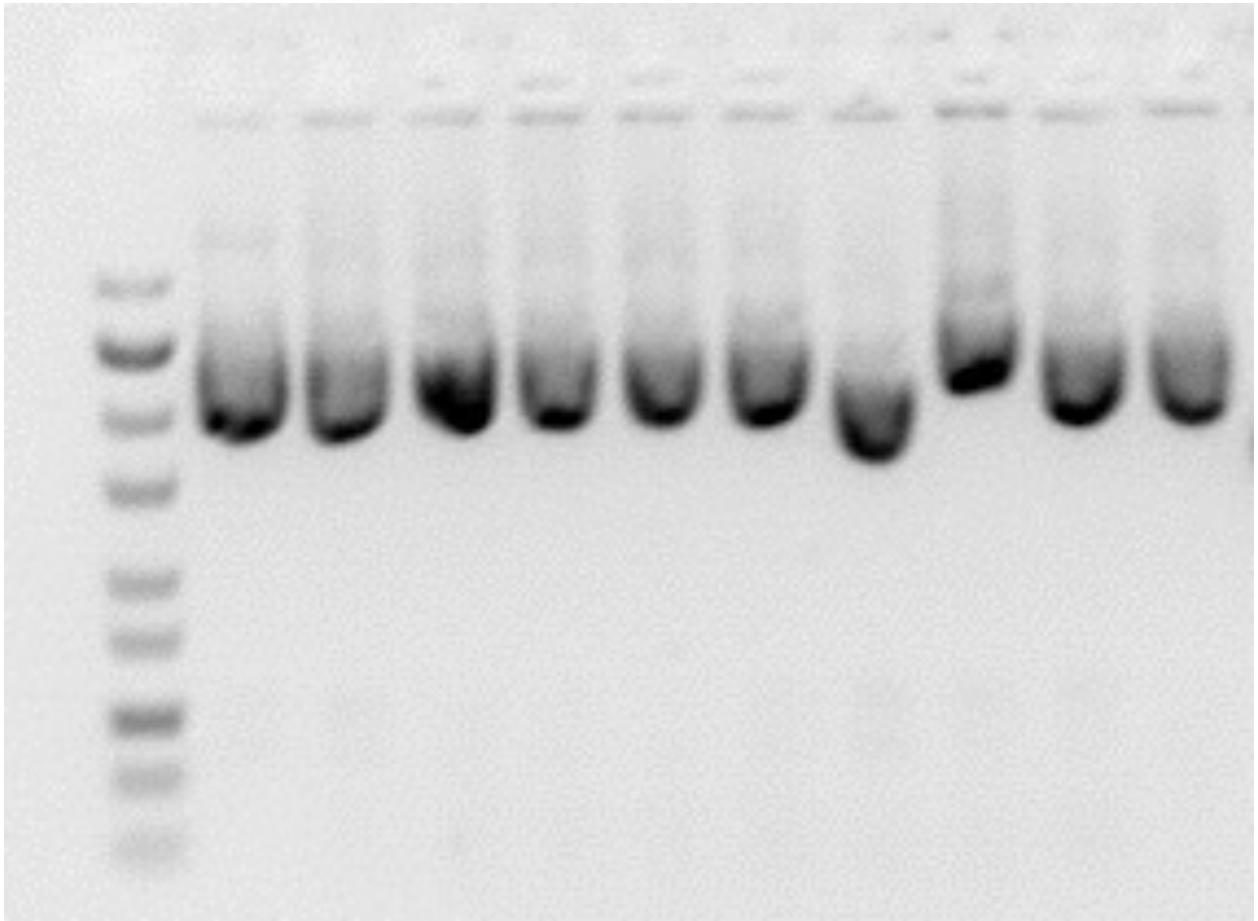


JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Picture: http://en.wikipedia.org/wiki/DNA

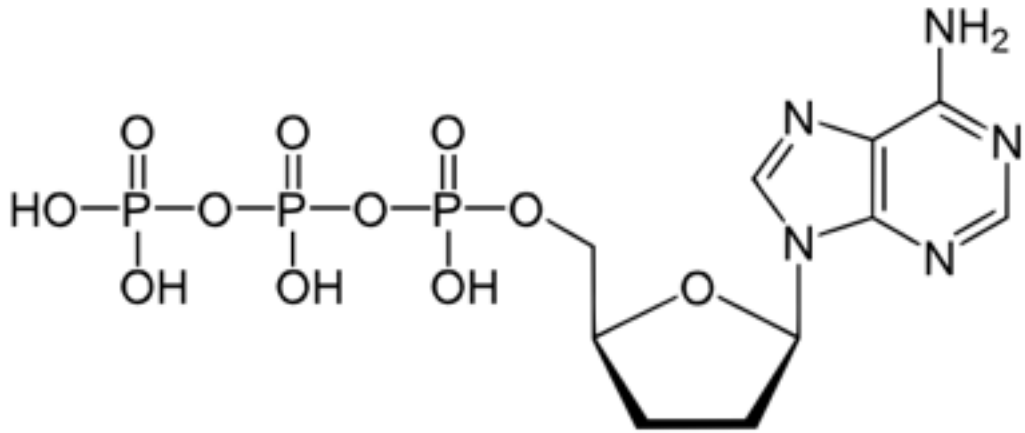
Sanger Sequencing

Gel electrophoresis:
Force DNA to move through an agarose gel using voltage; rate of movement is determined by length of the sequence

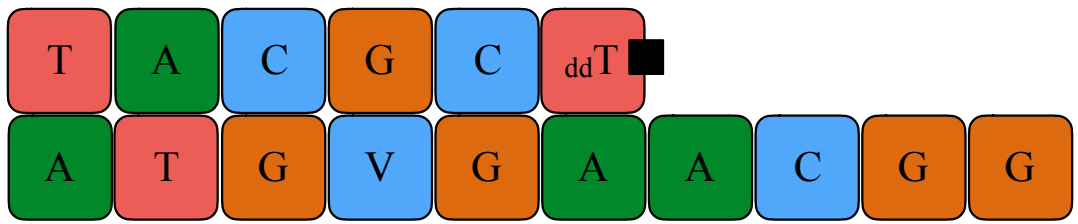


https://en.wikipedia.org/wiki/Gel_electrophoresis_of_nucleic_acids

Sanger Sequencing



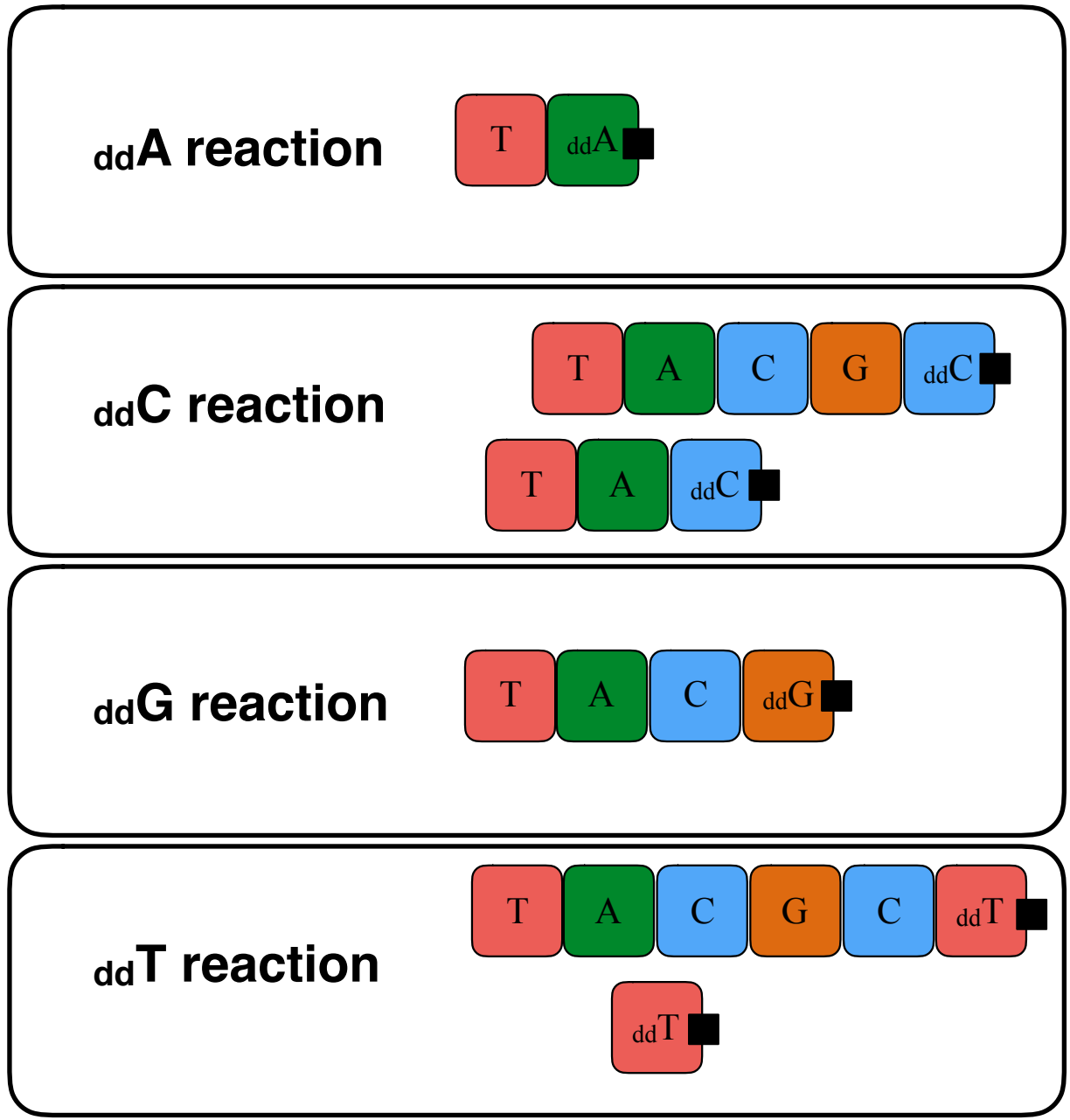
dideoxynucleotides:
Inhibit elongation of a DNA strand after they have been incorporated



https://en.wikipedia.org/wiki/Dideoxynucleotide

Sanger Sequencing

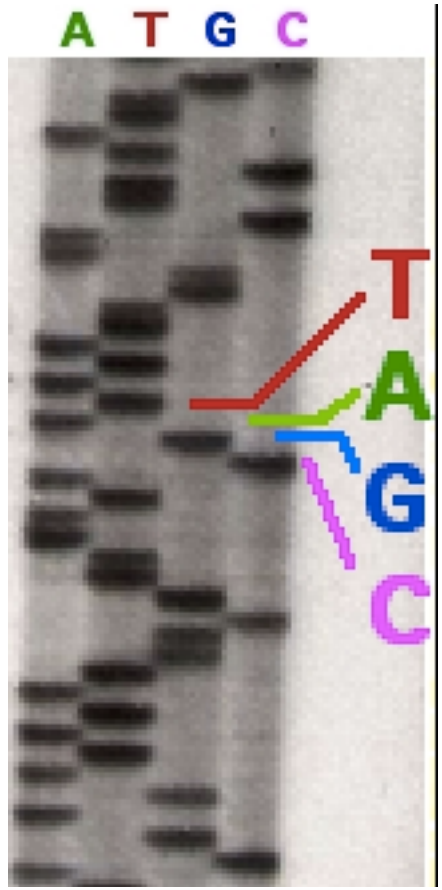
make many copies of the original DNA in four separate reactions, each using a different chain-terminating nucleotide



Sanger Sequencing

Separate each reaction mixture using gel electrophoresis, sorting the molecules by size

Read the sequence by traversing the bands in order of size as the last base of each molecule is known



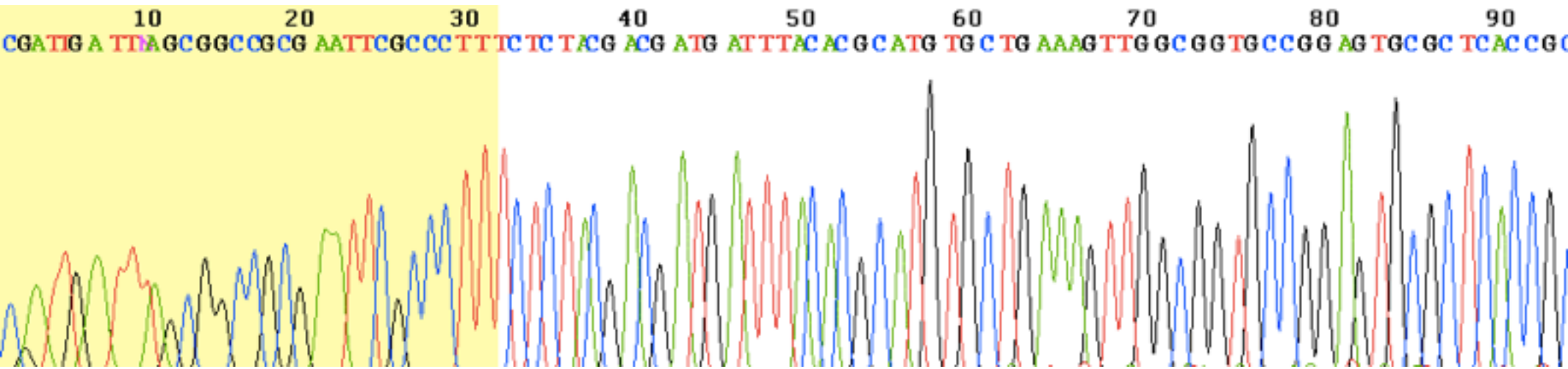
https://en.wikipedia.org/wiki/Sanger_sequencing

Sanger Sequencing

- Frederick Sanger received the Nobel prize for inventing this (shared with Gilbert and Berg for developing a different sequencing method)
  - Sequenced the human mitochondria genome (15,000bp) and virus lambda (48,000bp)
- Manually intensive and rather low throughput - only a few hundred bases could be sequenced at a time

Sanger Sequencing (2)

Improvement:
Use fluorescent labels to simplify reaction



https://en.wikipedia.org/wiki/Sanger_sequencing

Sanger Sequencing (3)

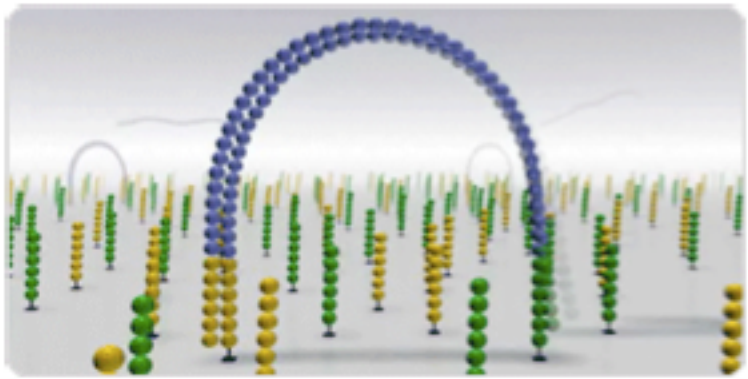
Automation and Parallelisation
improved throughput of Sanger sequencing allowing the 3B bp human genome to be sequenced

Sanger sequencing
1977-1990s

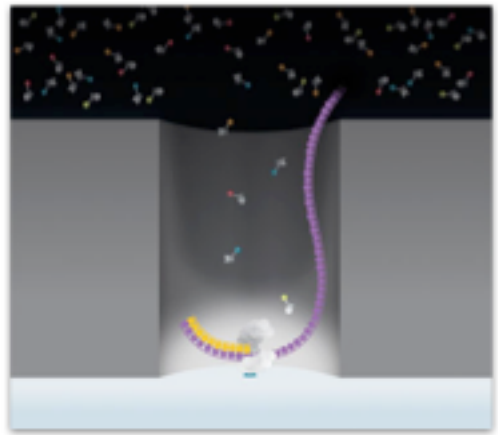
JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

High-throughput sequencing

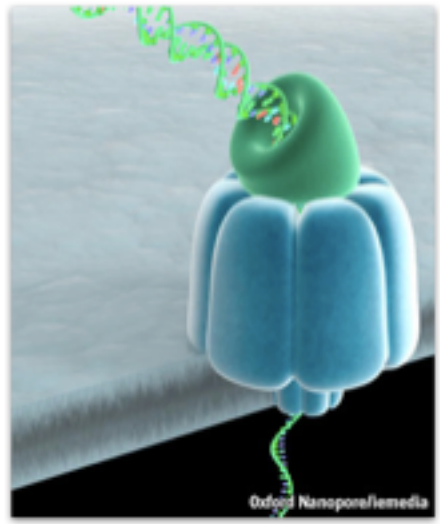
- The human genome was a landmark for genetics but very expensive (~$3B USD)
  - New technology was needed to routinely sequence genomes



Synthesis / ligation



SMRT cell



Nanopore

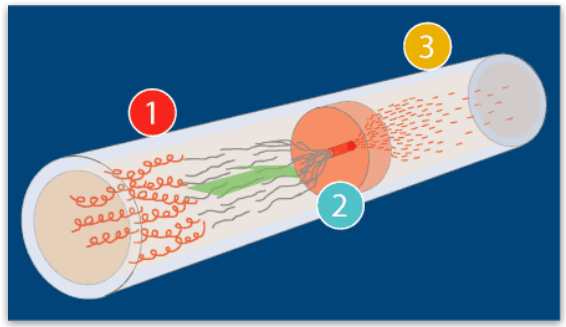
Sequencing by synthesis ("massively parallel sequencing") provides greatest throughput, and is the most prevalent today

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

Sequencing by synthesis

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

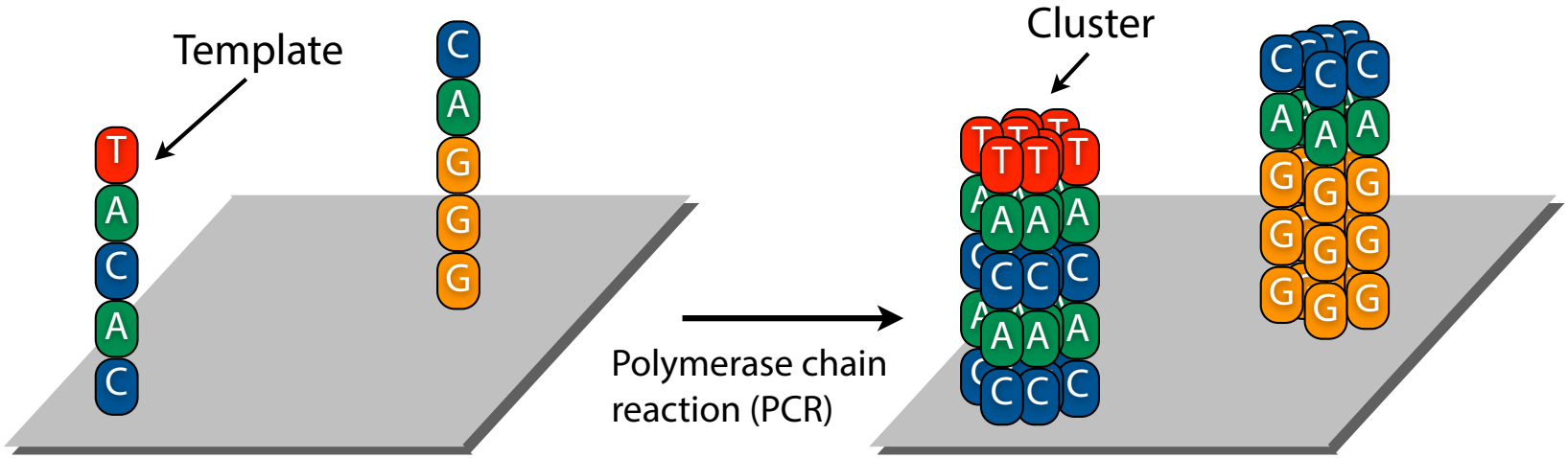
   E.g. with ultrasound waves, water-jet shearing (pictured), divalent cations



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

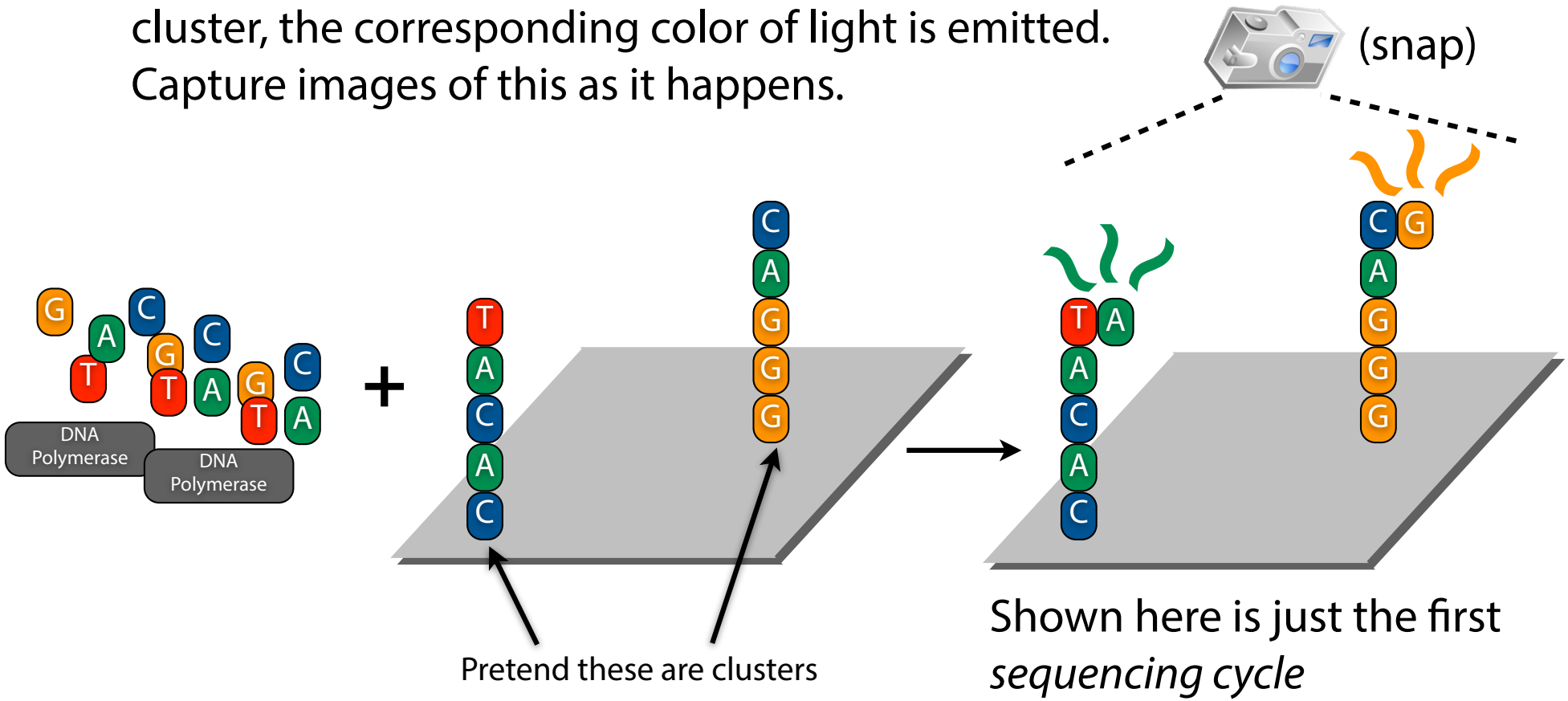
2. Attach templates to a surface

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



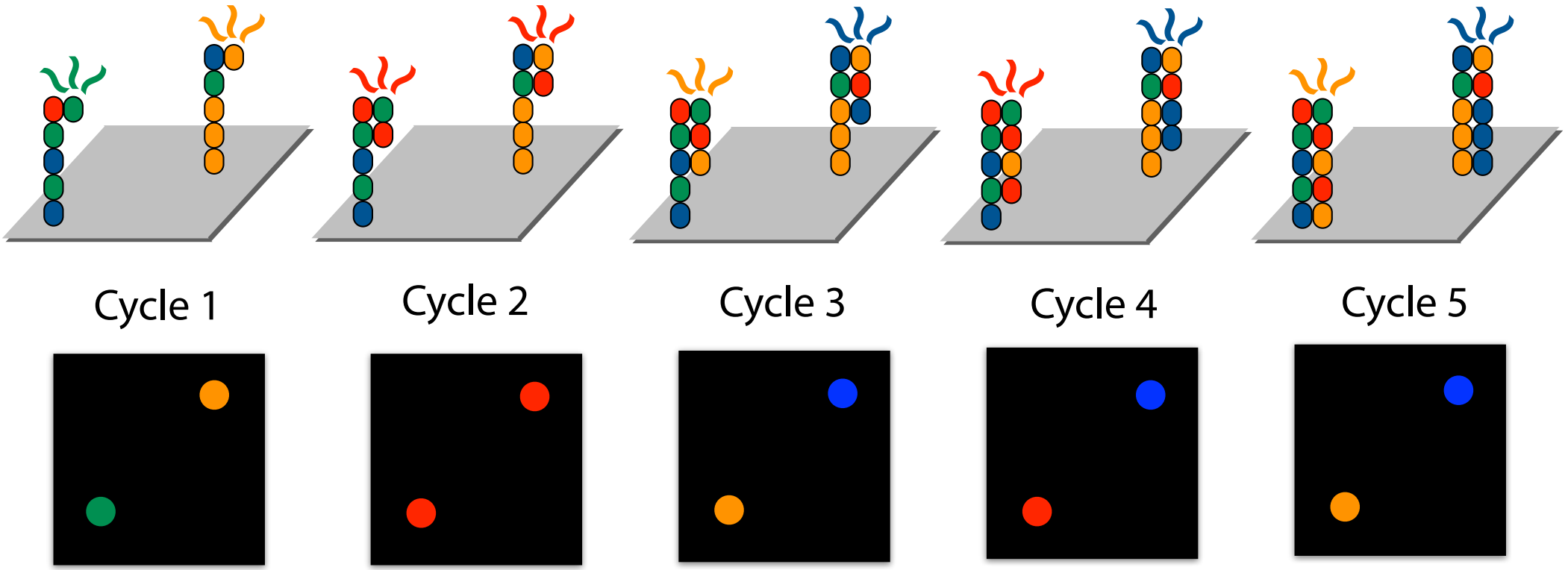
Sequencing by synthesis

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.



Sequencing by synthesis

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



Sequencing by synthesis

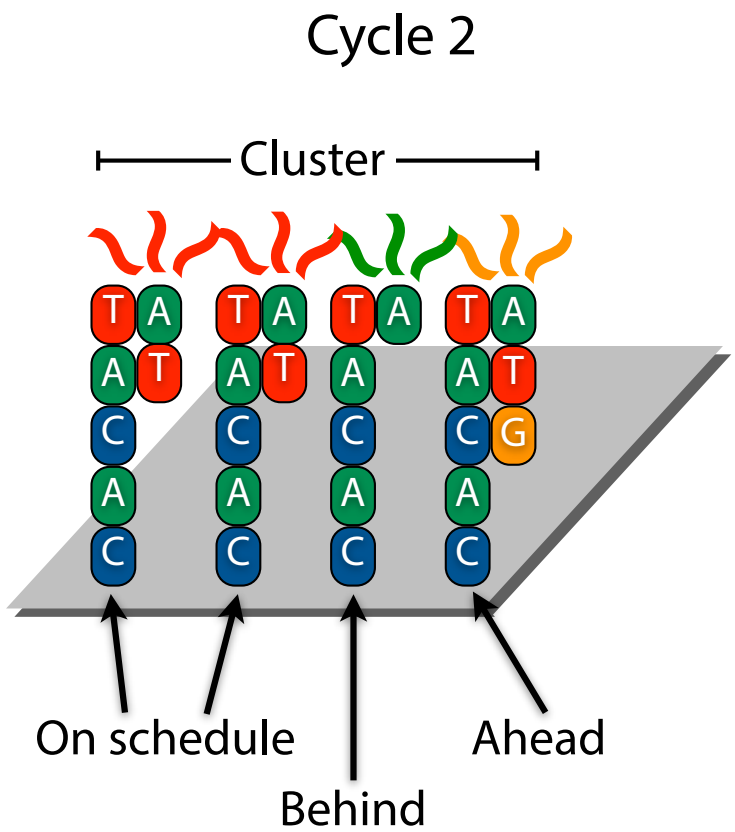
A modern sequencing-by-synthesis instrument such as the HiSeq sequences *billions* of clusters simultanously

A single "run" takes about 10 days to generate about 600 billion nucleotides of data

Cost of the reagents is $5-10K per run; multiplexing (sequencing many samples per run) further reduces cost per genome

Sequencing by synthesis: errors

Errors creep in when some templates get "out of sync," by missing an incorporation or by incorporating 2 or more nucleotides at once



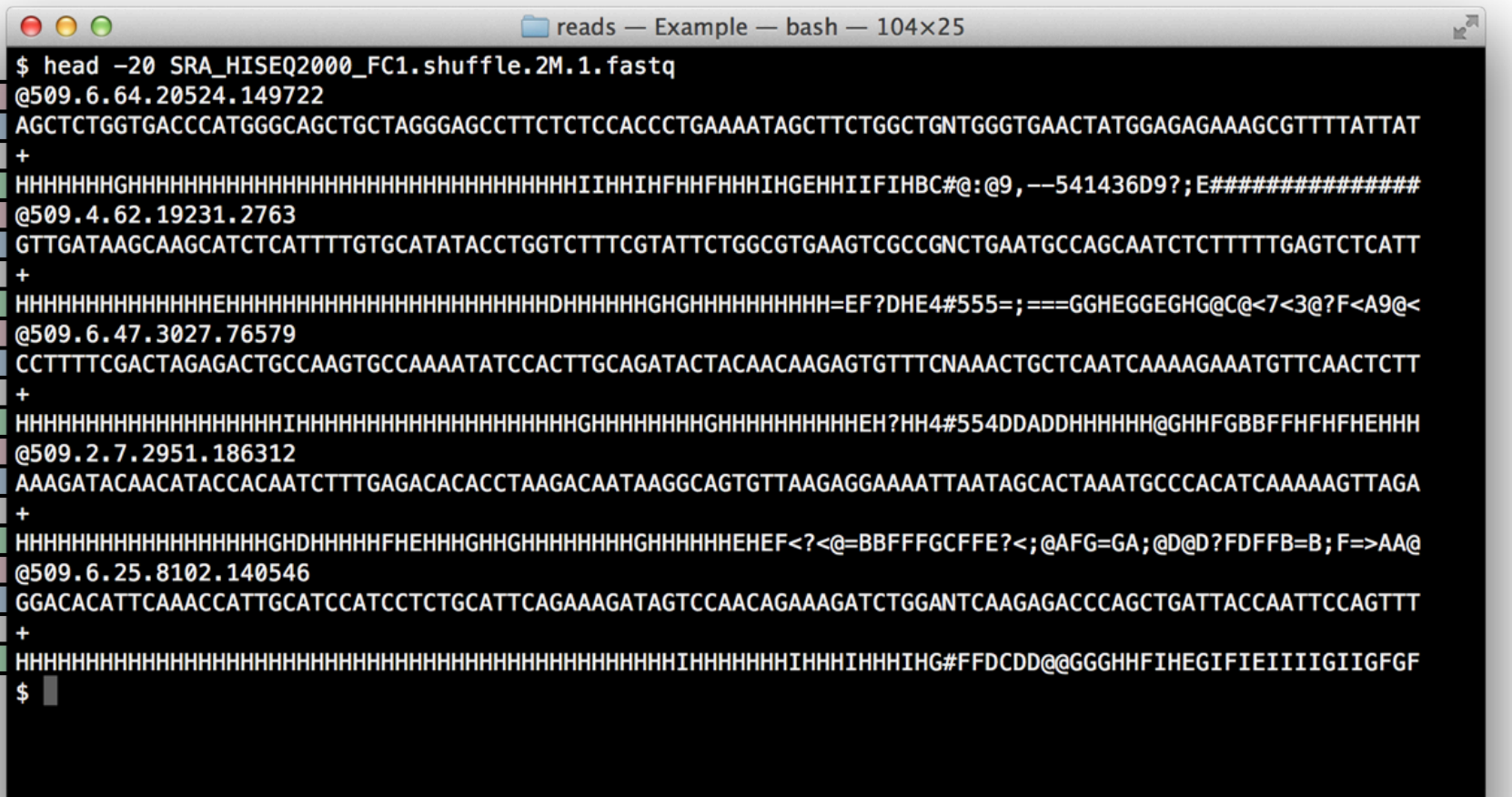
Base caller must deal with this uncertainty. Actual base callers report a *quality score* (confidence level) along with each nucleotide.

Errors are more common in later sequencing cycles, as proportionally more templates fall out of sync

Sequencing: read format

Below is a FASTQ file. A chunk of 4 lines describes a read. For each read, the 4 lines are (1) read name, (2) nucleotide sequence, (3) (placeholder), (4) quality value sequence.

Read 1	Name
	Nucleotides
	(placeholder)
	Quality values
Read 2	Name
	Nucleotides
	(placeholder)
	Quality values
Read 3	Name
	Nucleotides
	(placeholder)
	Quality values
Read 4	Name
	Nucleotides
	(placeholder)
	Quality values
Read 5	Name
	Nucleotides
	(placeholder)
	Quality values



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Sequencing: qualities

Nucleotides and quality values line up 1-to-1:

A quality value is an ASCII encoding of a number, Q

where $Q = -10 \log_{10} p$

where p is sequencer's estimate of the probability that the nucleotide at that position was called *incorrectly*

$Q = 10$: error probability is 1 in 10
$Q = 20$: error probability is 1 in 100
$Q = 30$: error probability is 1 in 1,000

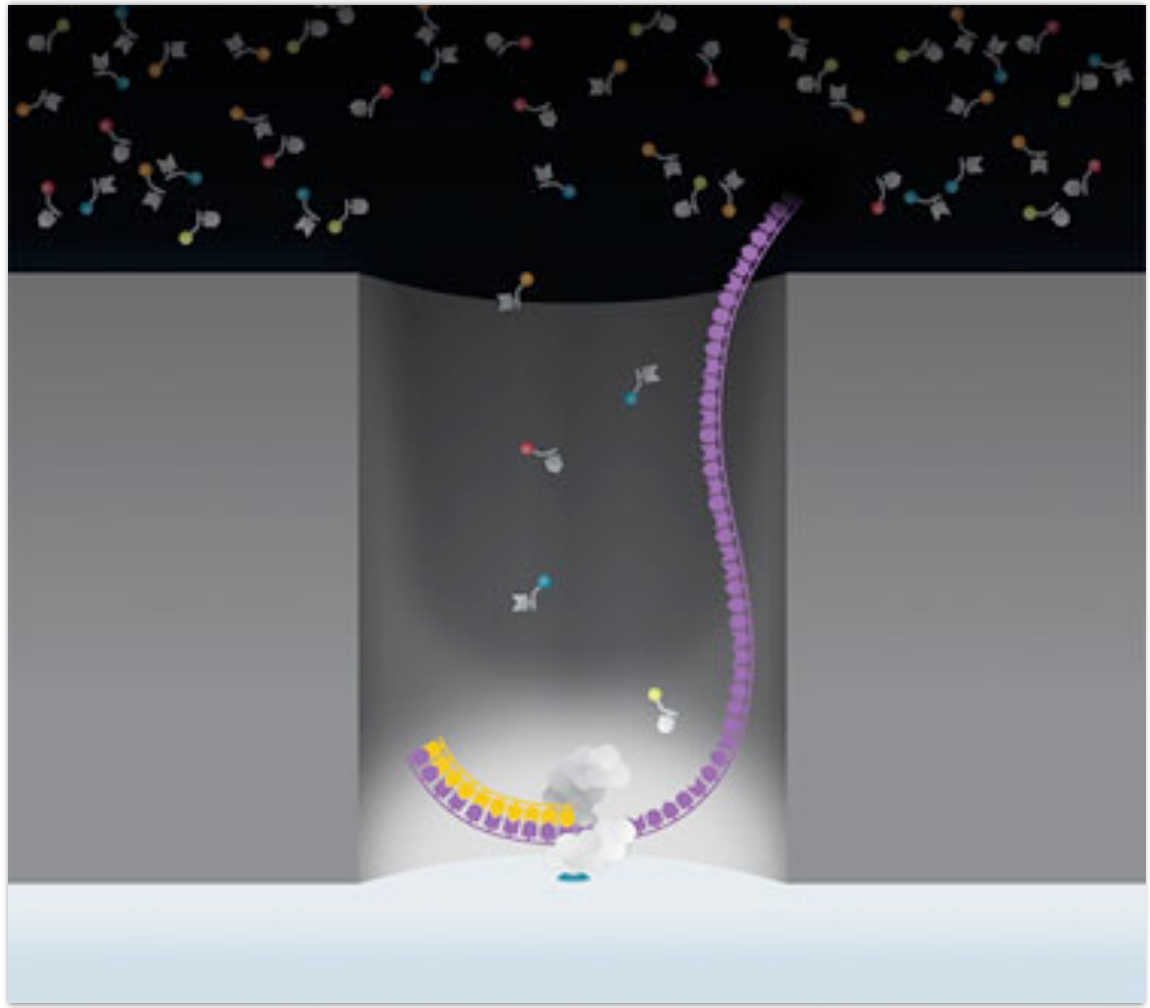
*S*ometimes called "Phred scale"

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

See also: http://en.wikipedia.org/wiki/Fastq

Illumina Sequencing

PacBio Sequencing



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Picture: http://www.genengnews.com/gen-articles/third-generation-sequencing-debuts/3257/

Nanopore Sequencing

Joseph Bore running the MinION in Guinea, Quick et al Nature (2016)

Nanopore Sequencing

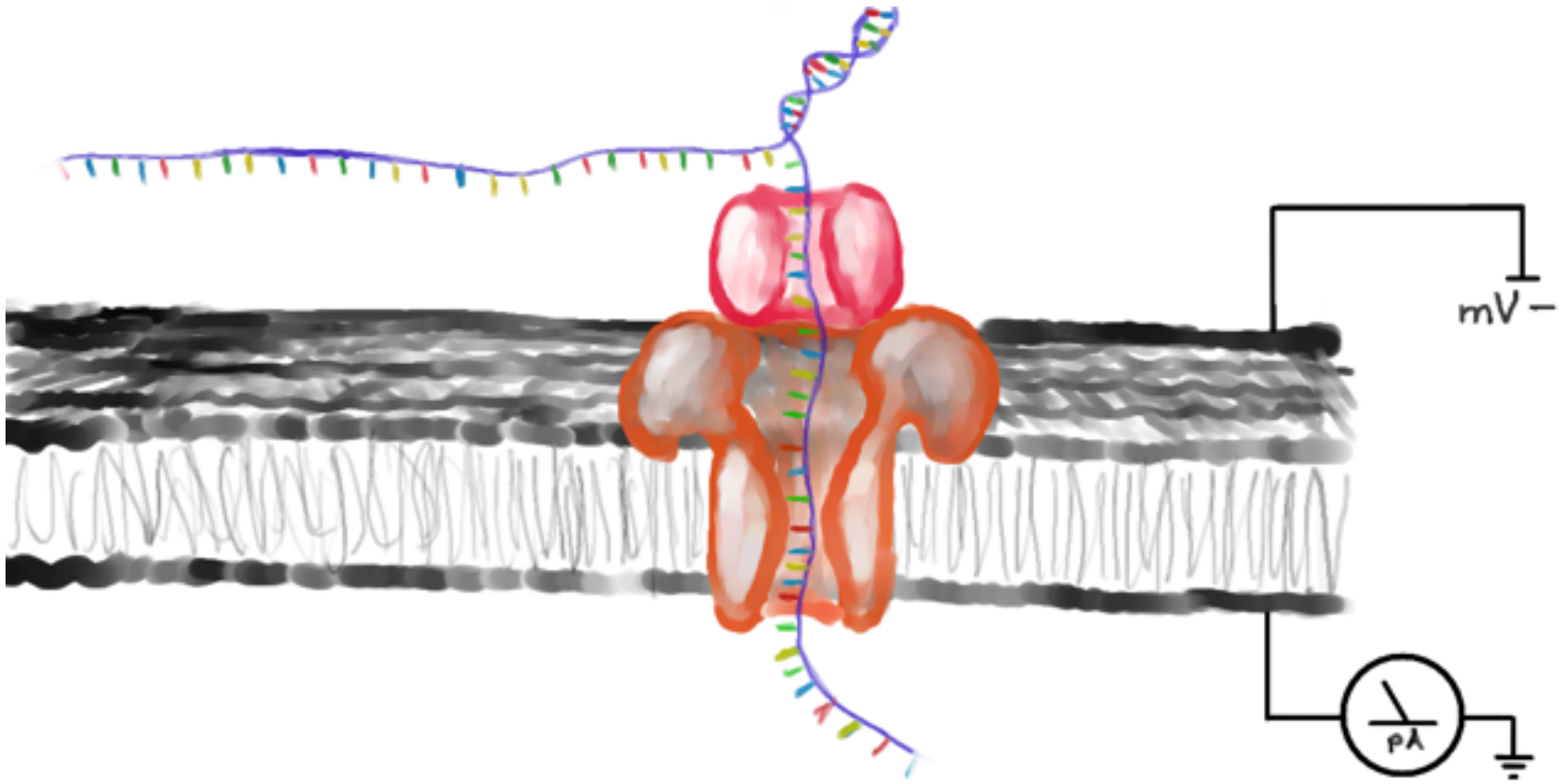
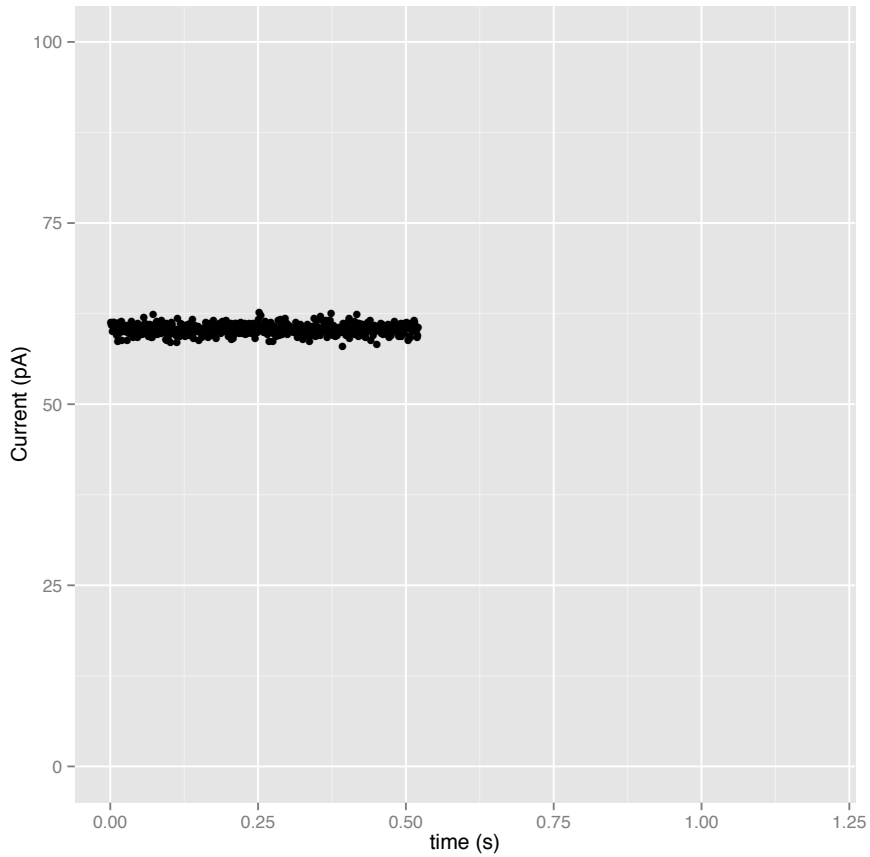
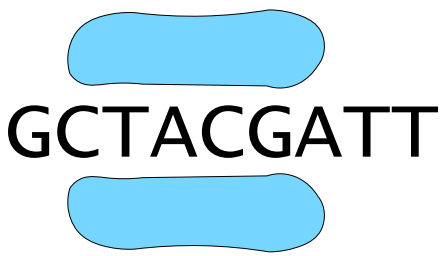


Illustration by David Eccles

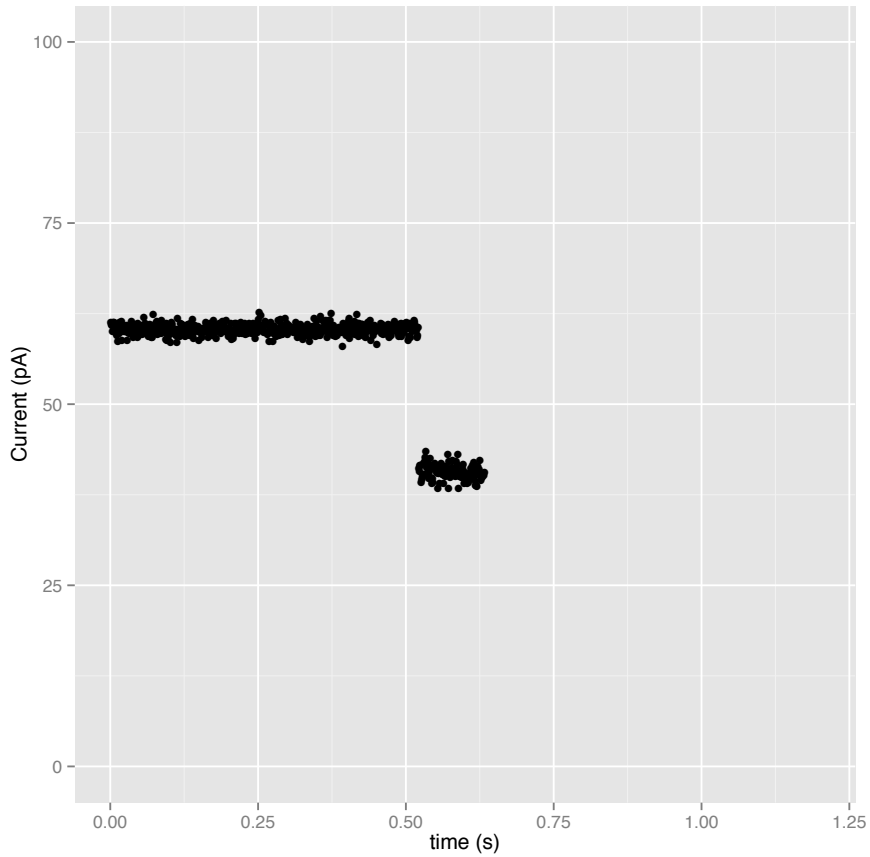
Nanopore Sequencing



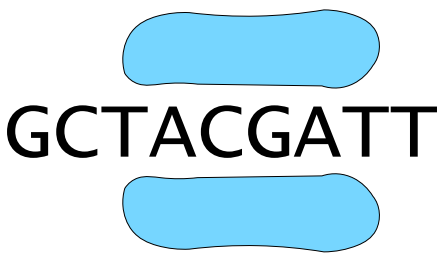
Nanopore Sequencing



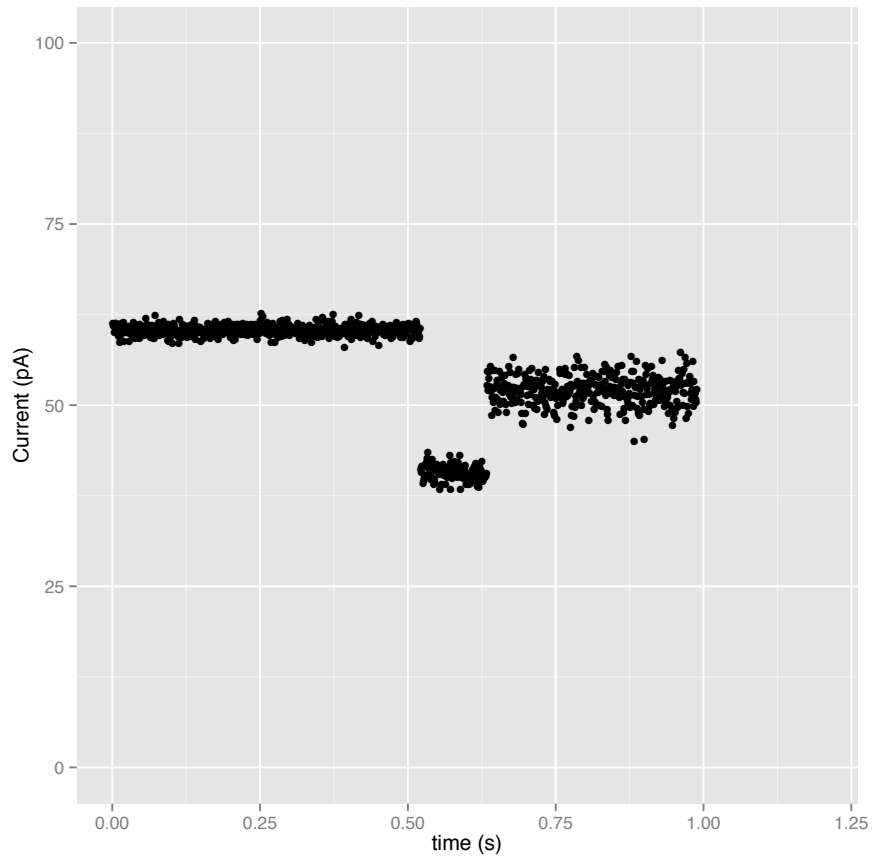
Sample Current



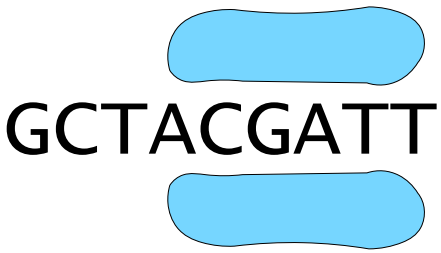
Nanopore Sequencing



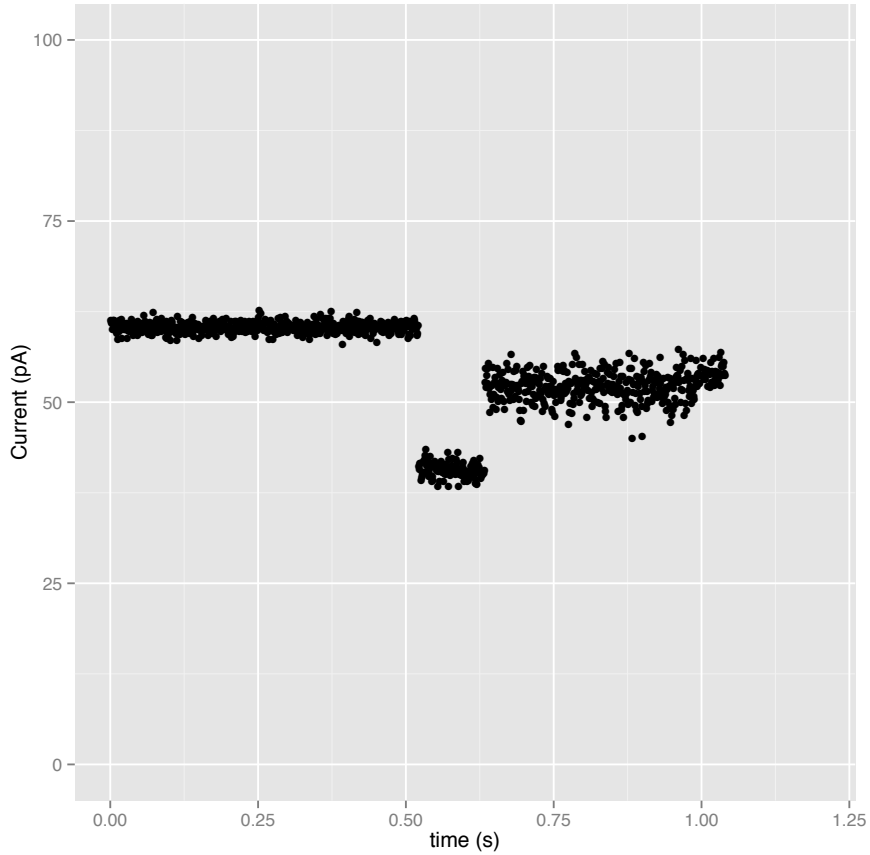
Sample Current



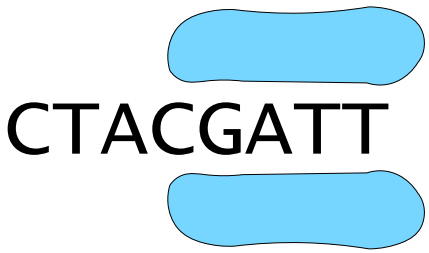
Nanopore Sequencing



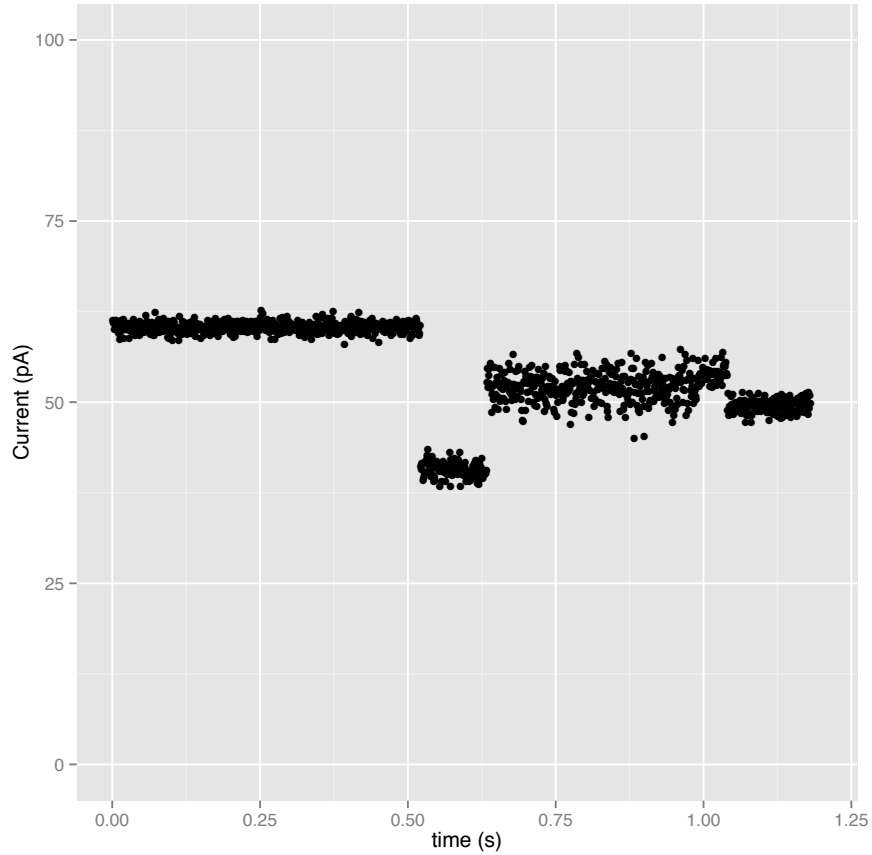
Sample Current



Nanopore Sequencing

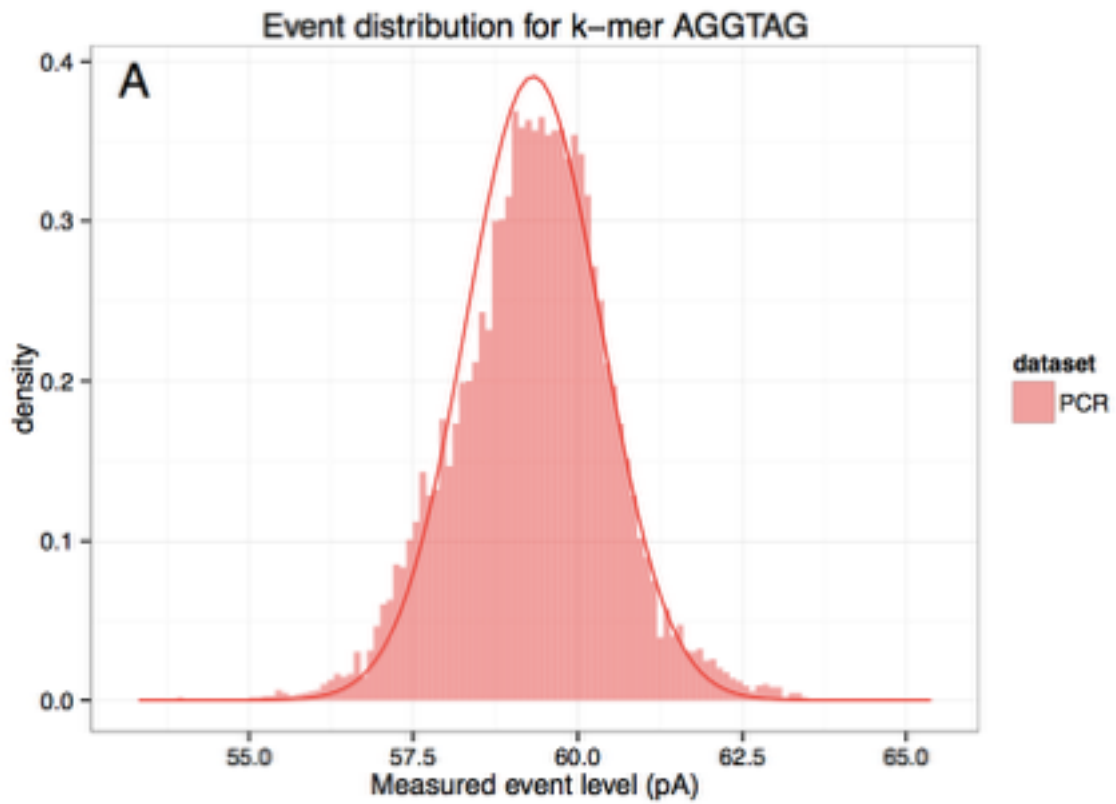


Sample Current



Oxford Nanopore

- Use probabilistic models (like HMMs that we will learn later) to predict the most likely DNA sequence from the measured current samples



Sequencing Technology Summary

- Illumina:
  - 100-200bp reads
  - Up to 600Gbp per run
  - Very low error rate (<1% bases miscalled)
- Pacbio/Oxford Nanopore:
  - Single molecule sequencing (no amplification)
  - >10,000 bp reads
  - Up to 1Gbp per run
  - Higher error rate (5-15%)

Sequence Analysis Problems

- The central focus of this course will be describing how we interpret this data
  - Can we reconstruct the complete genome sequence from the read sequences?
  - How do we detect variation within human genomes?
  - How do we do this efficiently?
- Next lecture: algorithms on strings and sequences