GMS6850 – Core Concepts in Bioinformatics
Lecture 1, 2022 Jan 10

https://bit.ly/3f5gx2u

Intro to the faculty and the course
Bioinformatics landscape
Review of biological molecules

Faculty: Steve Rozen, Sujoy Ghosh, Enrico Petretto, Lisa Tucker-Kellogg

With guest lectures from Abner Lim, Weng Khong LIM, Owen Rackham, Gavin Smith, Federico Torta, Jacques Behmoaras

Outline

- Faculty introductions
- “What is Bioinformatics?” and aims scope of this course
- Next generation sequencing and bioinformatics
- Review of biological molecules
- Summary

My story

- Undergrad degree in dance
- Got work, but concluded it wasn't viable (and not very interesting, and I wasn't very talented)
- Got a masters in computer science
- Worked on Wall Street for 2 years
- Got a PhD in computer science (the dumb way, part time)
- Was not willing to relocate for a faculty job in computer science for family reasons, so went to work for Eric Lander at a "genome center"

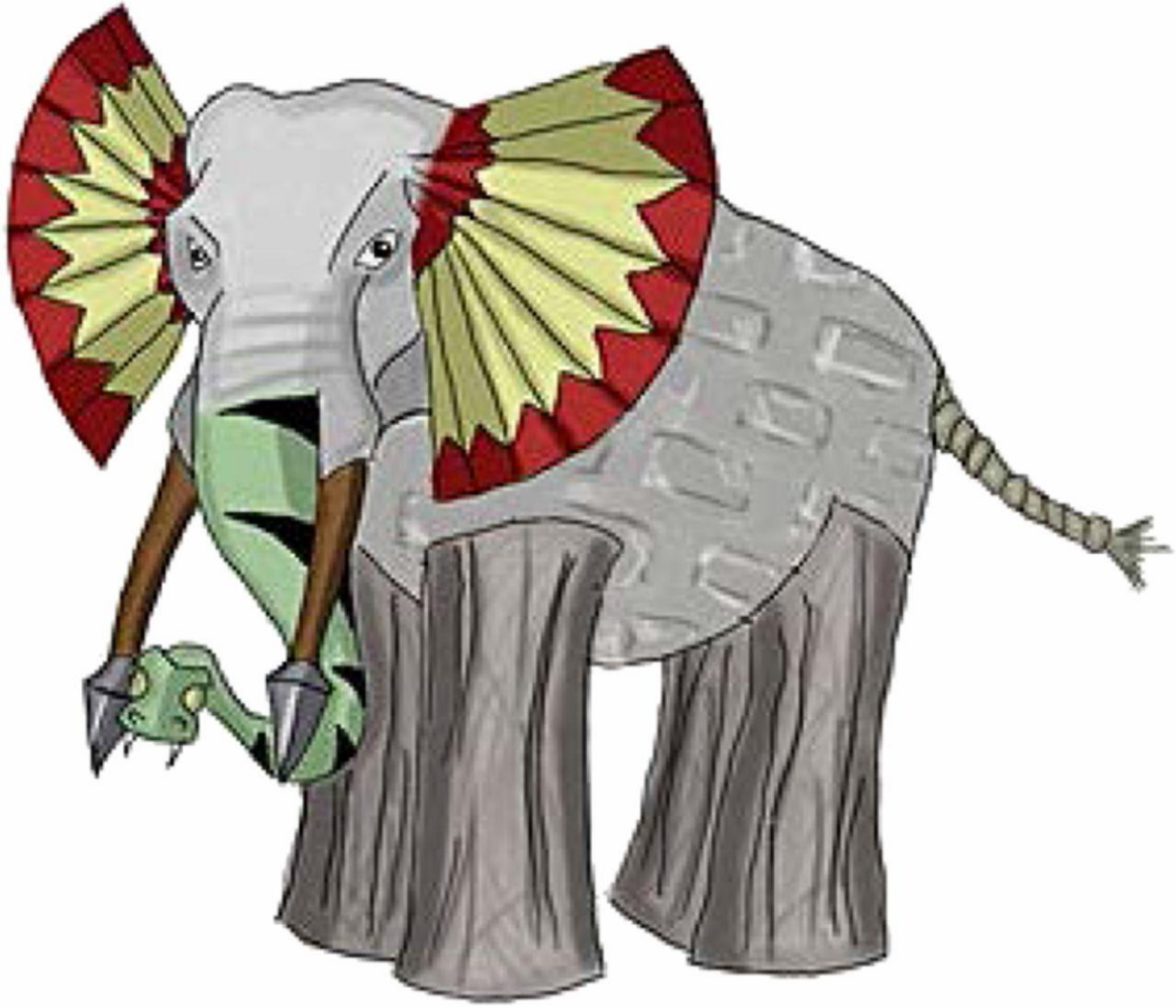
My story

- Helped sequence the human and mouse genomes (small role in huge international effort)
- Wrote the Primer3 software and web interface (with Helen Skaletsky), cited by 20,000 papers
- Went to the lab of David Page and did a mix of lab work and bioinformatics for 8 years
- High profile genetics publications (Nature, Nature Genetics)
- Went to Duke-NUS in 2008, found great collaborators in cancer genomics (Patrick Tan, Bin Teh)
- Multiple high impact papers (AA, gastric cancer, bile duct cancer, mutational signatures)

Outline

- Faculty introductions
- “What is Bioinformatics?” and aims scope of this course
- Next generation sequencing and bioinformatics
- Review of biological molecules
- Summary

What is bioinformatics?



What is ~~bioinformatics~~?

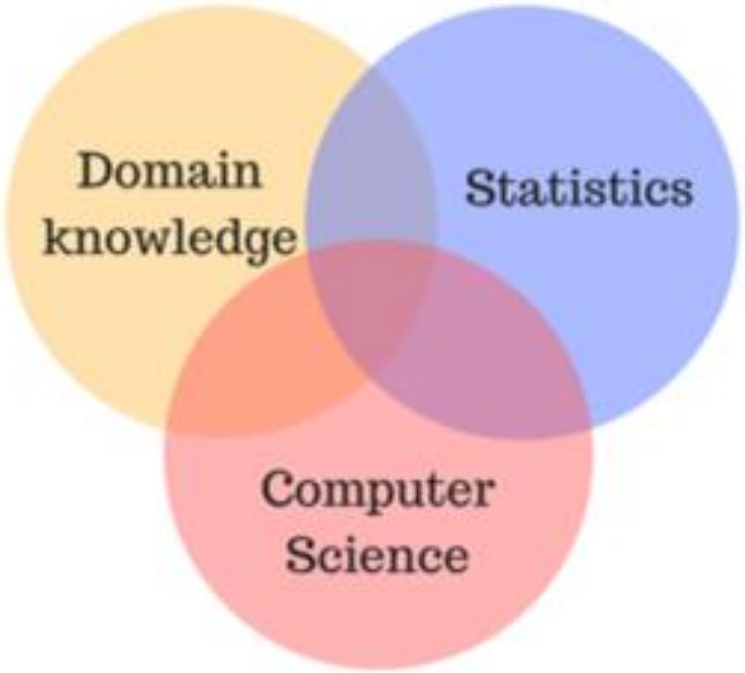
DukeNUS Medical School

Data Science?

What is ~~bioinformatics~~?

- Looking at a lot of data to:
- Show you ads that you will click on
- Detect credit card fraud
- Assemble a winning baseball team (Sabermetrics / Moneyball)
- Predict result of the US November 2018 election

Data Science?



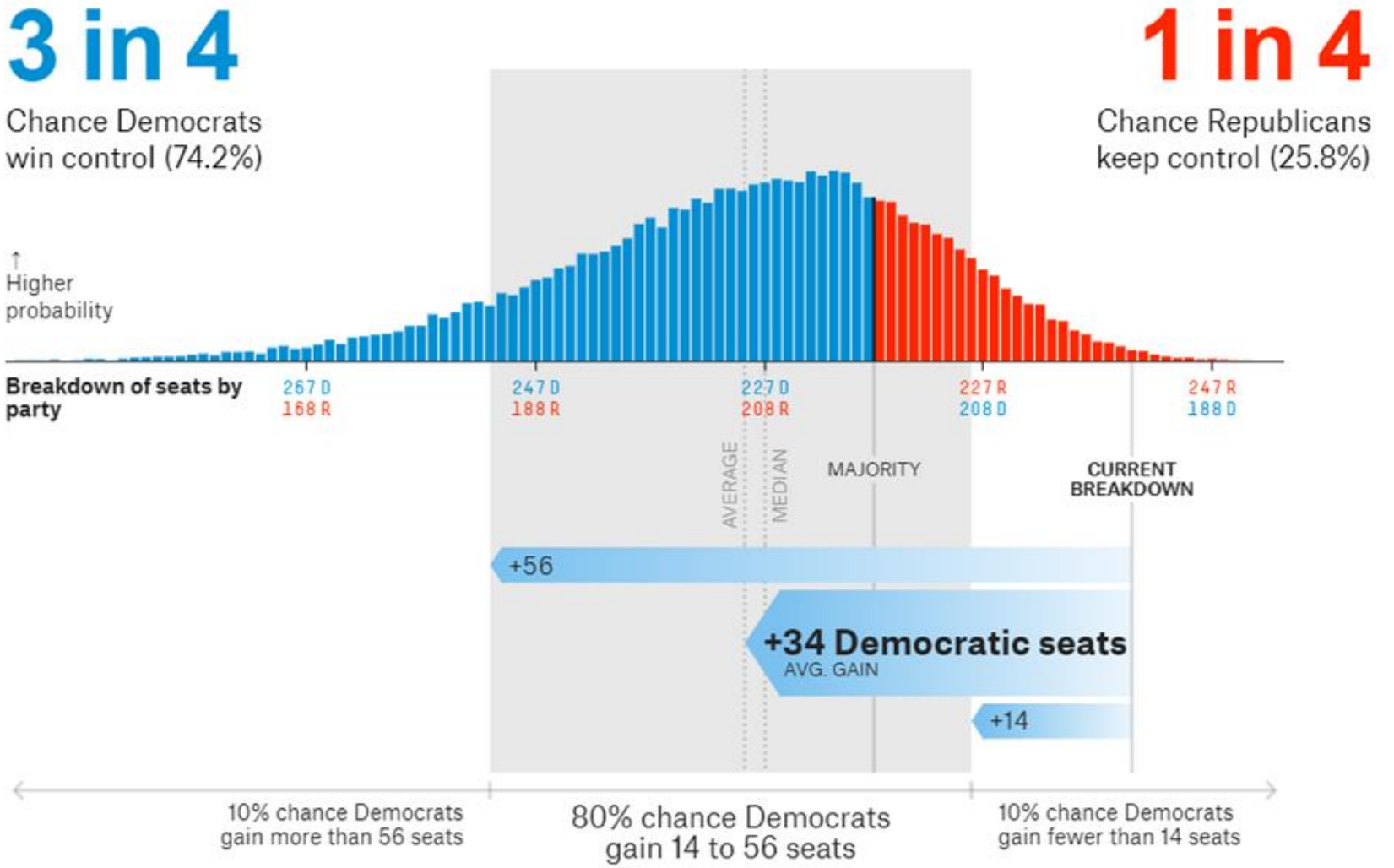
http://omgenomics.com/what-is-bioinformatics/

Forecasting the election of the US House of Representatives in November, 2018

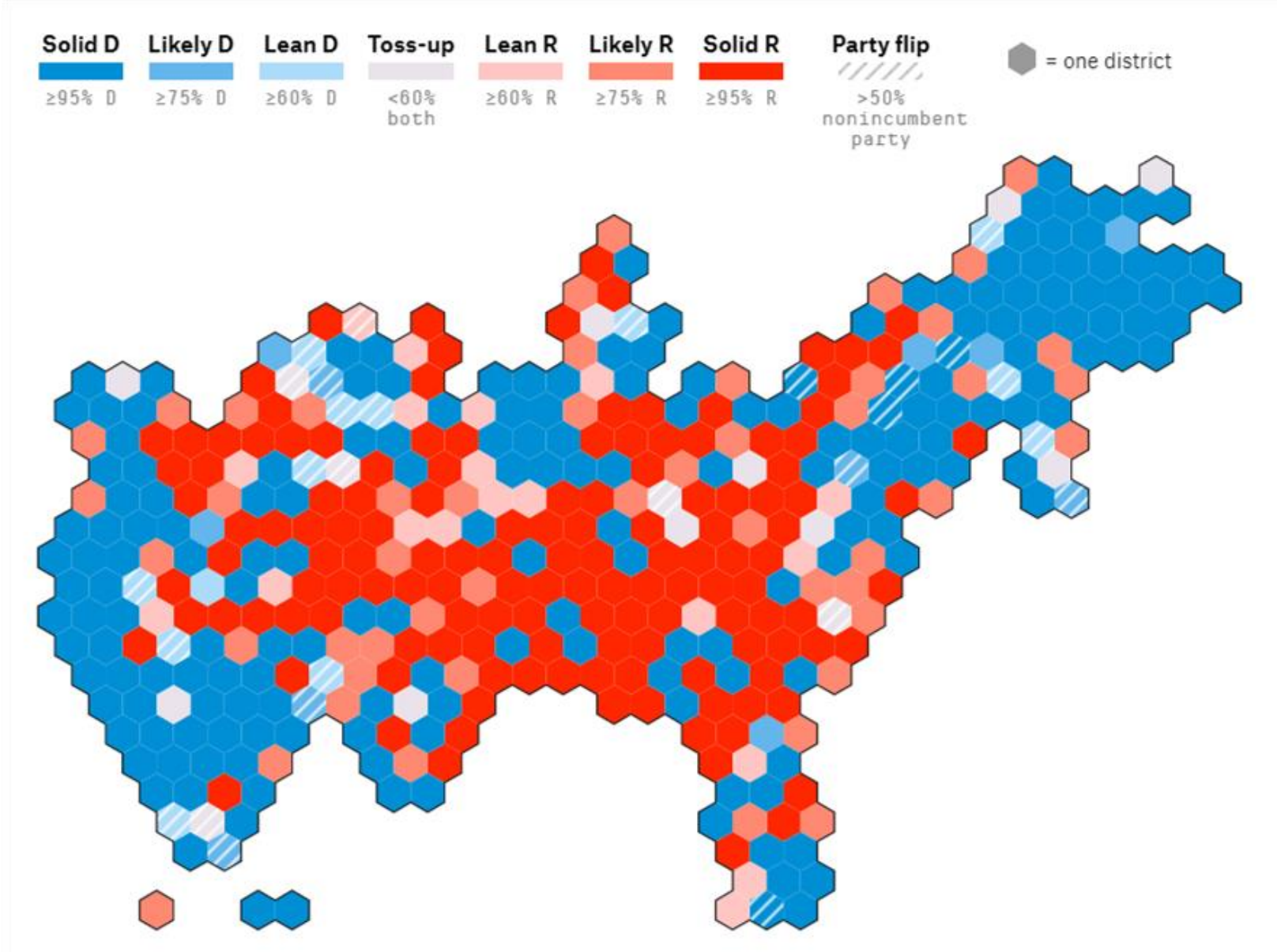
FiveThirtyEight

https://fivethirtyeight.com

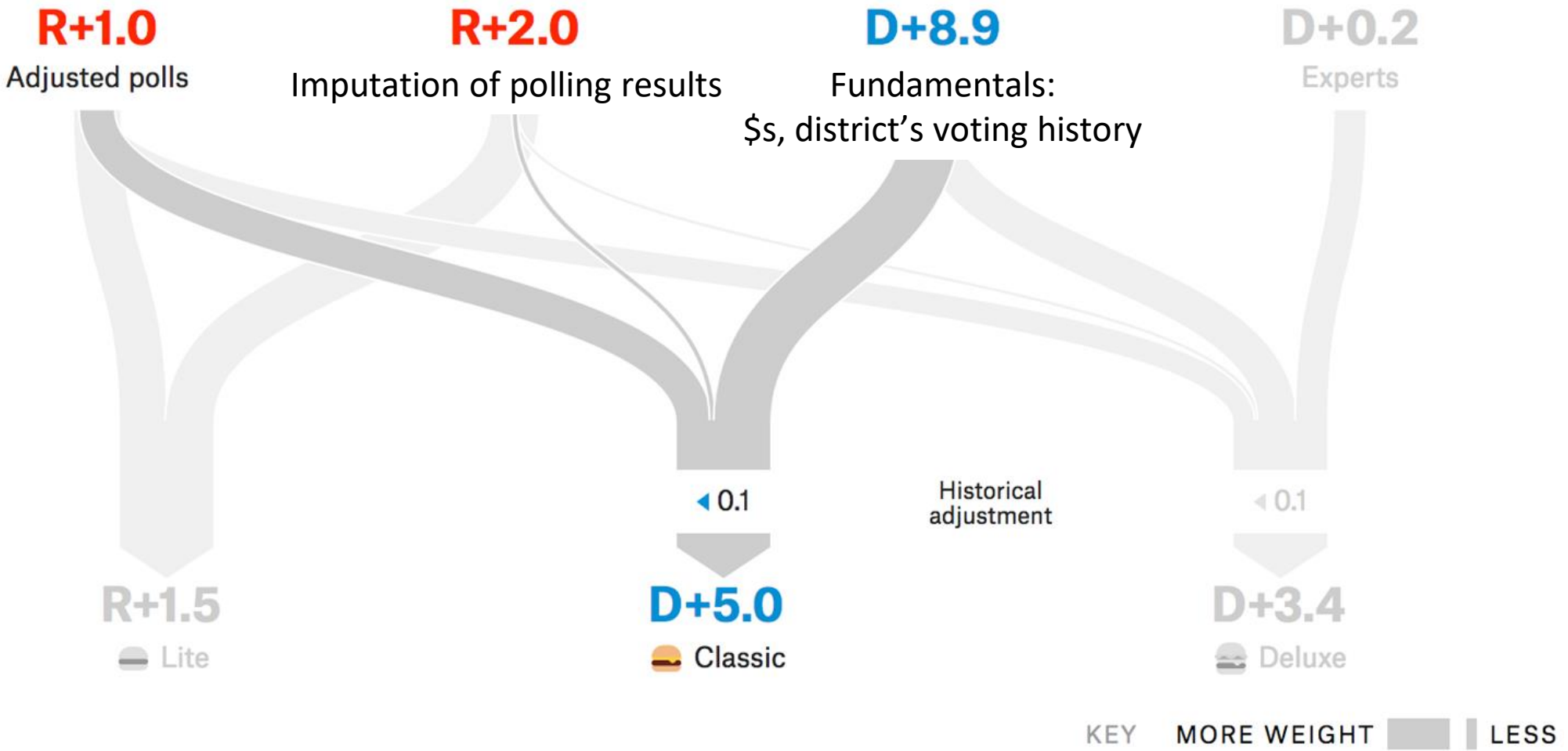
DukeNUS Medical School



All house races



Models for 1 house race

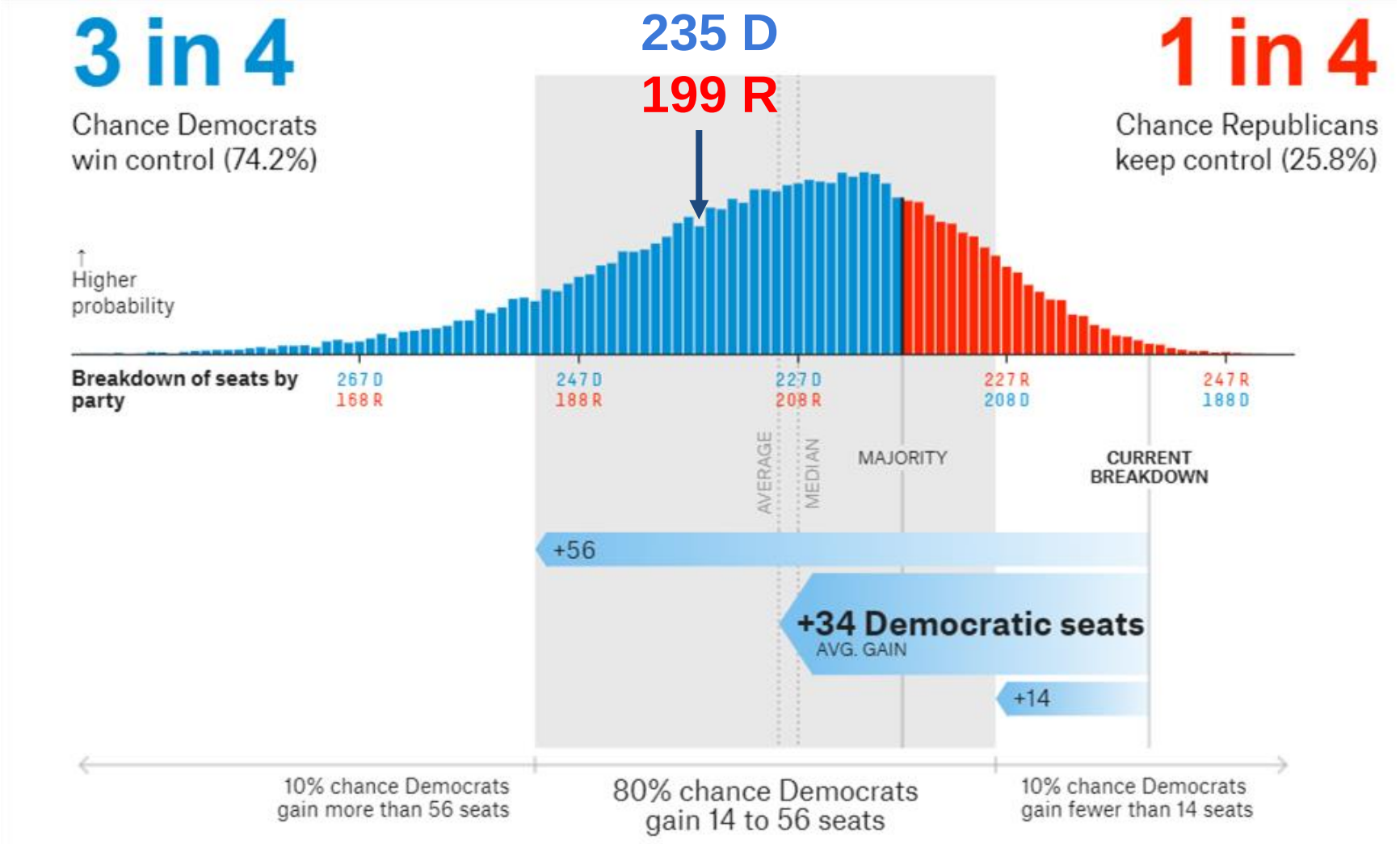


Forecast pretty good:

FiveThirtyEight

https://fivethirtyeight.com

Actual gain, 41 seats

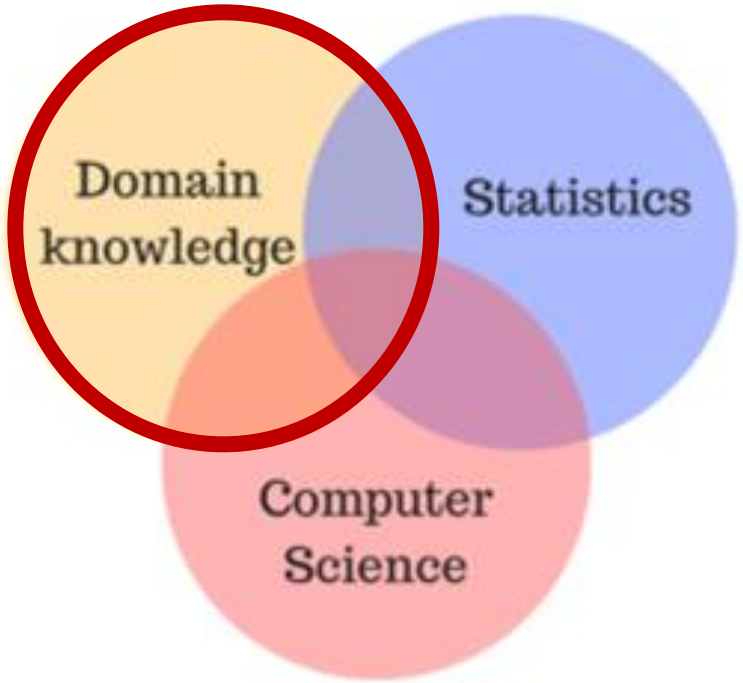


What is ~~bioinformatics~~?

- Statistics with a really large amount of data
- Overlaps with
  - Statistical machine learning
  - Artificial intelligence

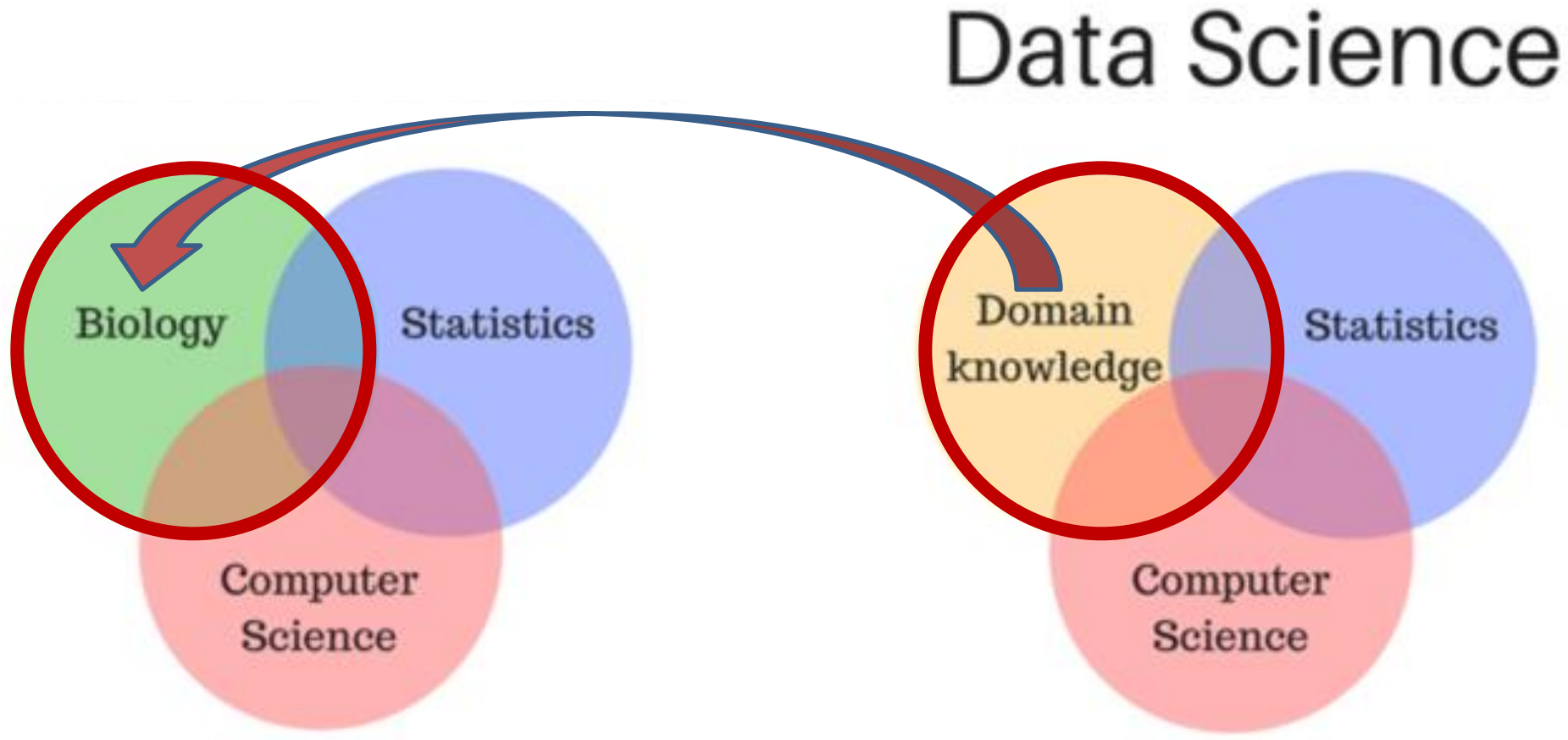
http://omgenomics.com/what-is-bioinformatics/

What is bioinformatics?



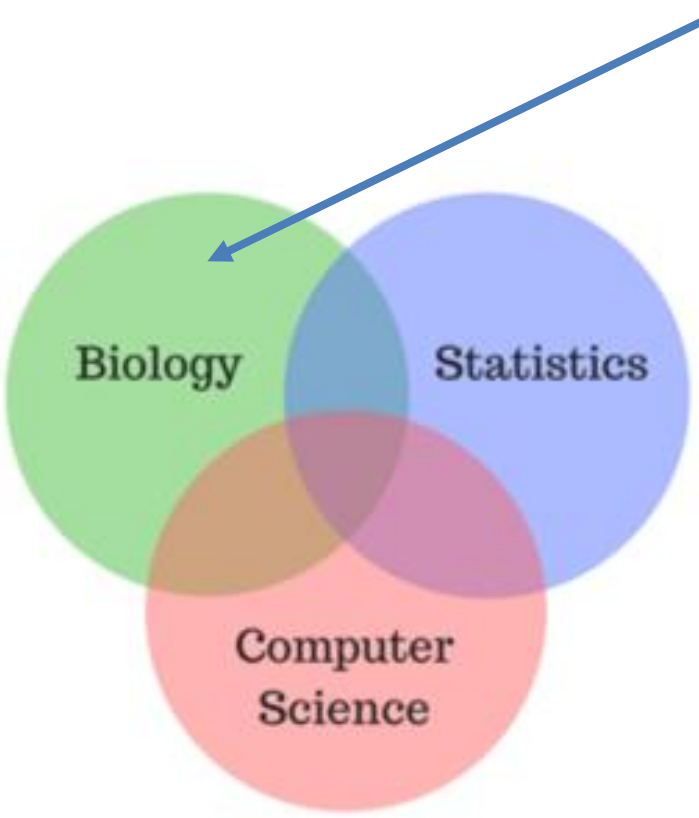
http://omgenomics.com/what-is-bioinformatics/

What is bioinformatics?



http://omgenomics.com/what-is-bioinformatics/

What is bioinformatics?



Specialization within bioinformatics driven by biological focus

Focus of the course will be on "genome informatics" and network biology with some human genetics (e.g. GWAS – genome wide association studies)

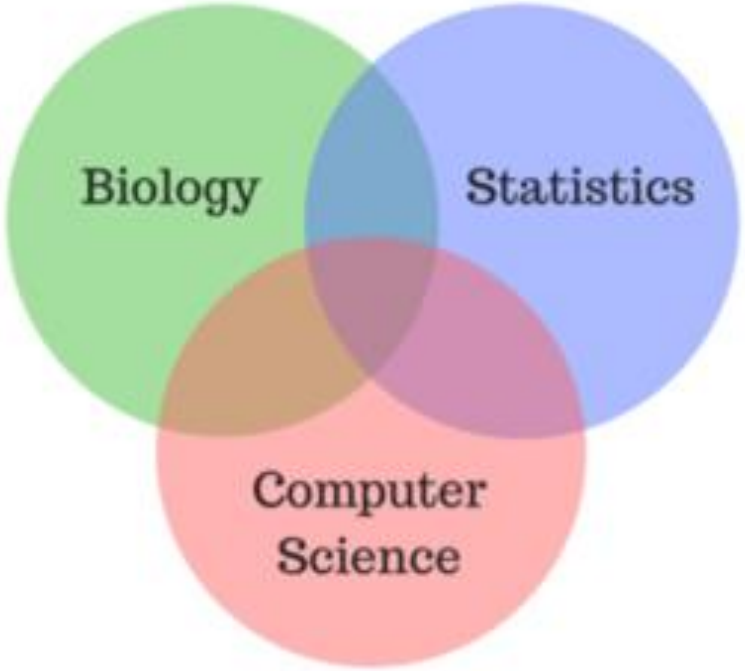
Some other, related areas:

- "structural bioinformatics" https://en.wikipedia.org/wiki/Structural_bioinformatics Alphafold (artificial intelligence to predict protein structure from protein sequence): https://www.youtube.com/watch?v=gg7WjuFs8F4, https://www.youtube.com/watch?v=RmYaM-_7YjY
- Population and evolutionary genetics
- Biological image analysis (from microscopy to medical imaging)
- "Medical informatics" (dealing with health records – not bioinformatics at all (?))

http://omgenomics.com/what-is-bioinformatics/

Being a bioinformatician

Expertise in one area, competence in others

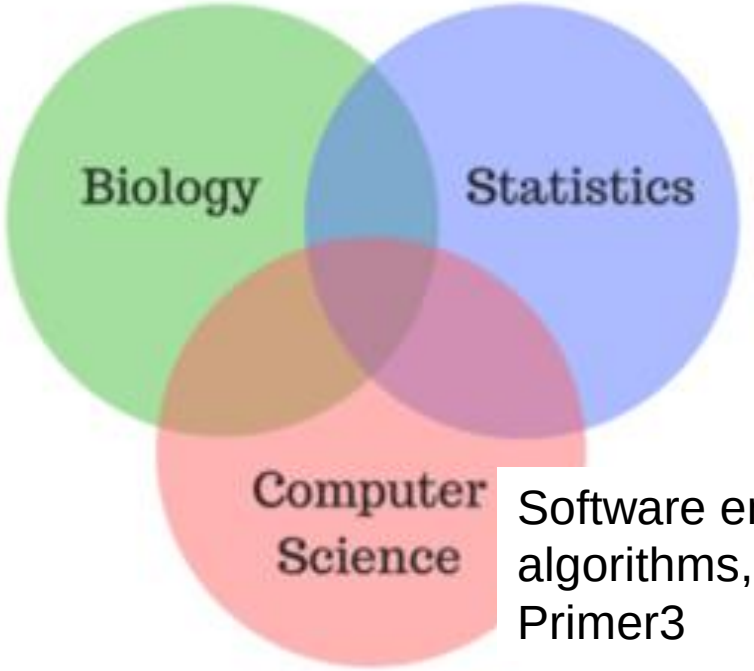


Sometimes unusual career paths

Being a bioinformatician

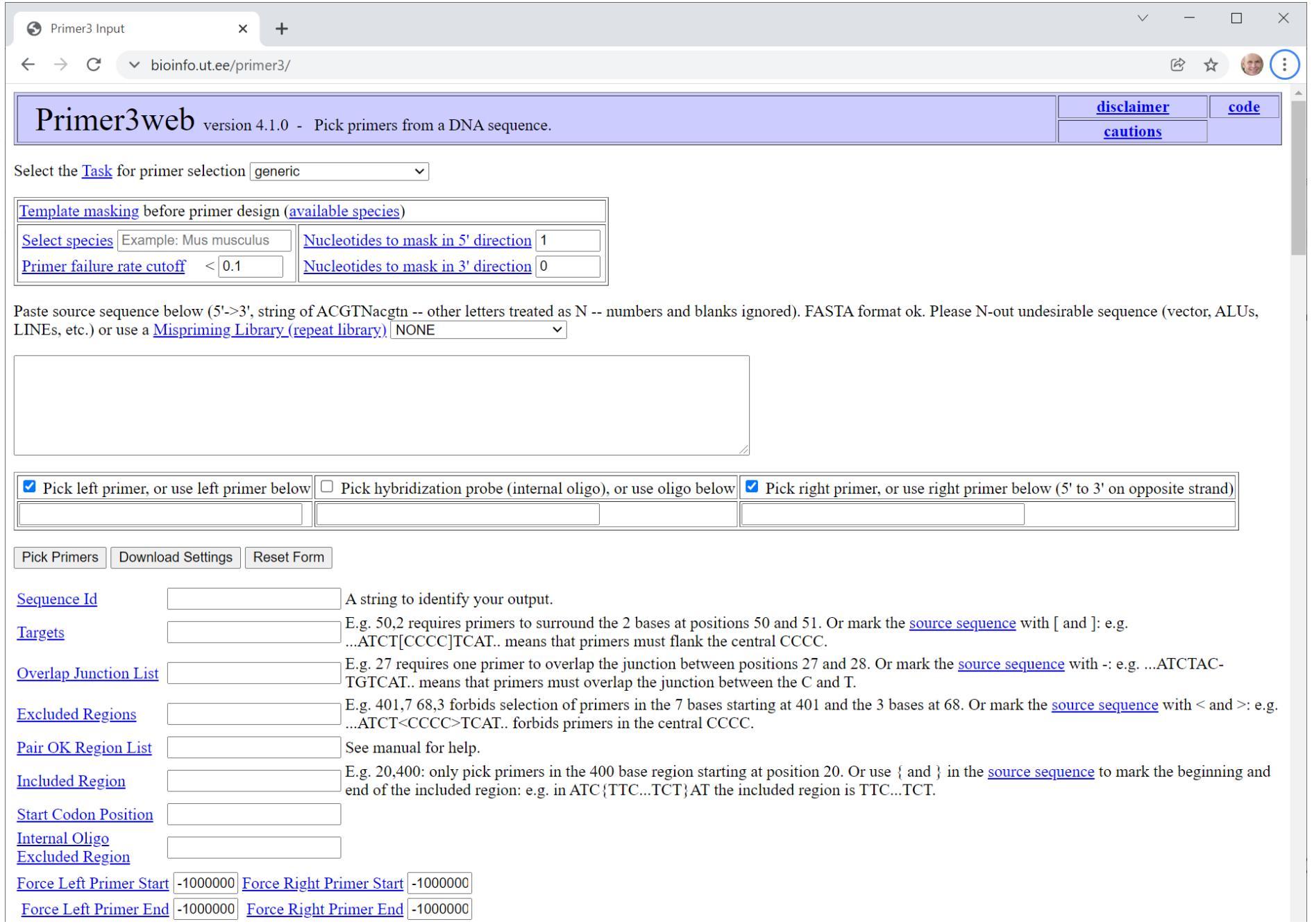
DukeNUS Medical School

Expertise in one area, competence in others

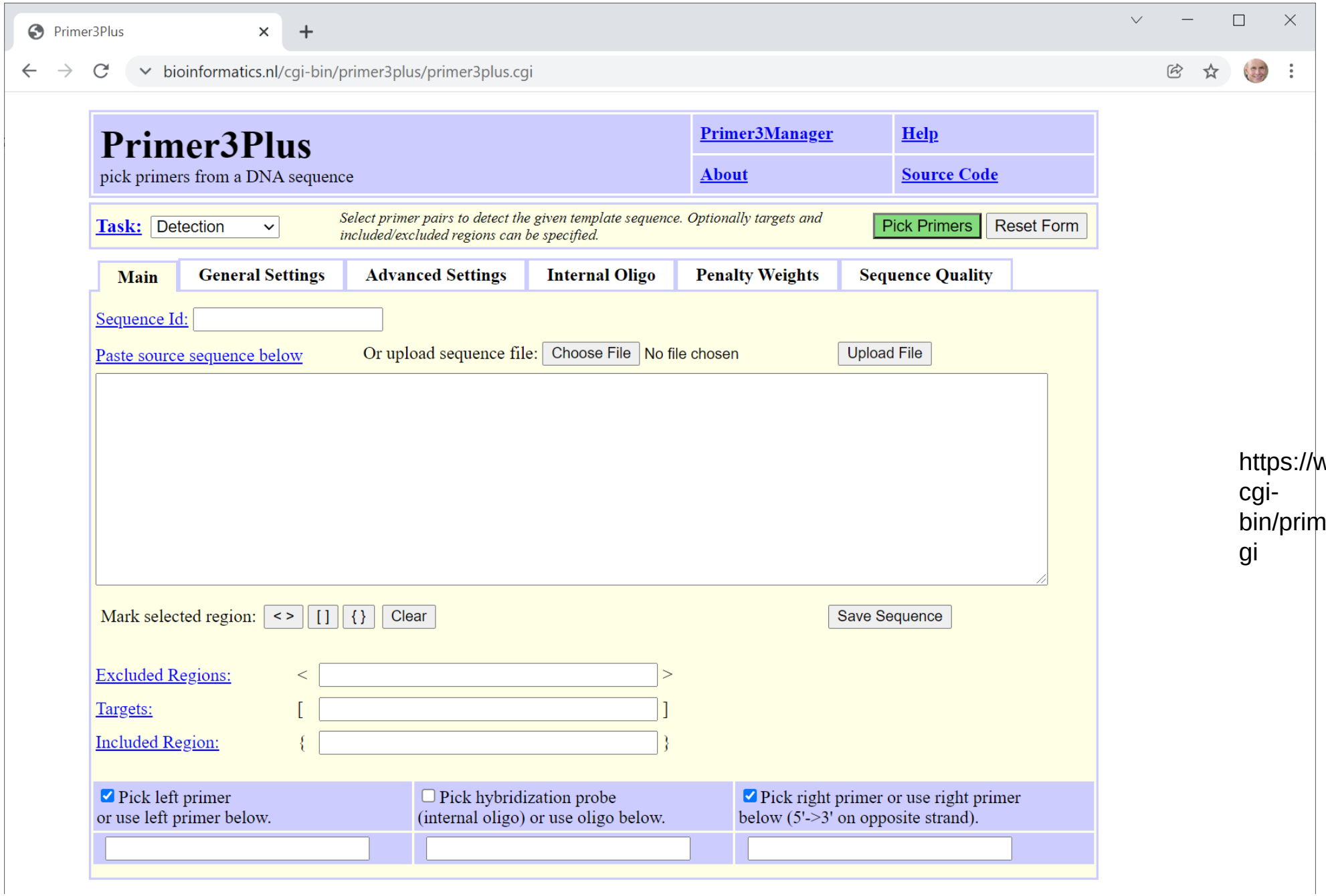


Software engineering and algorithms, for example Primer3

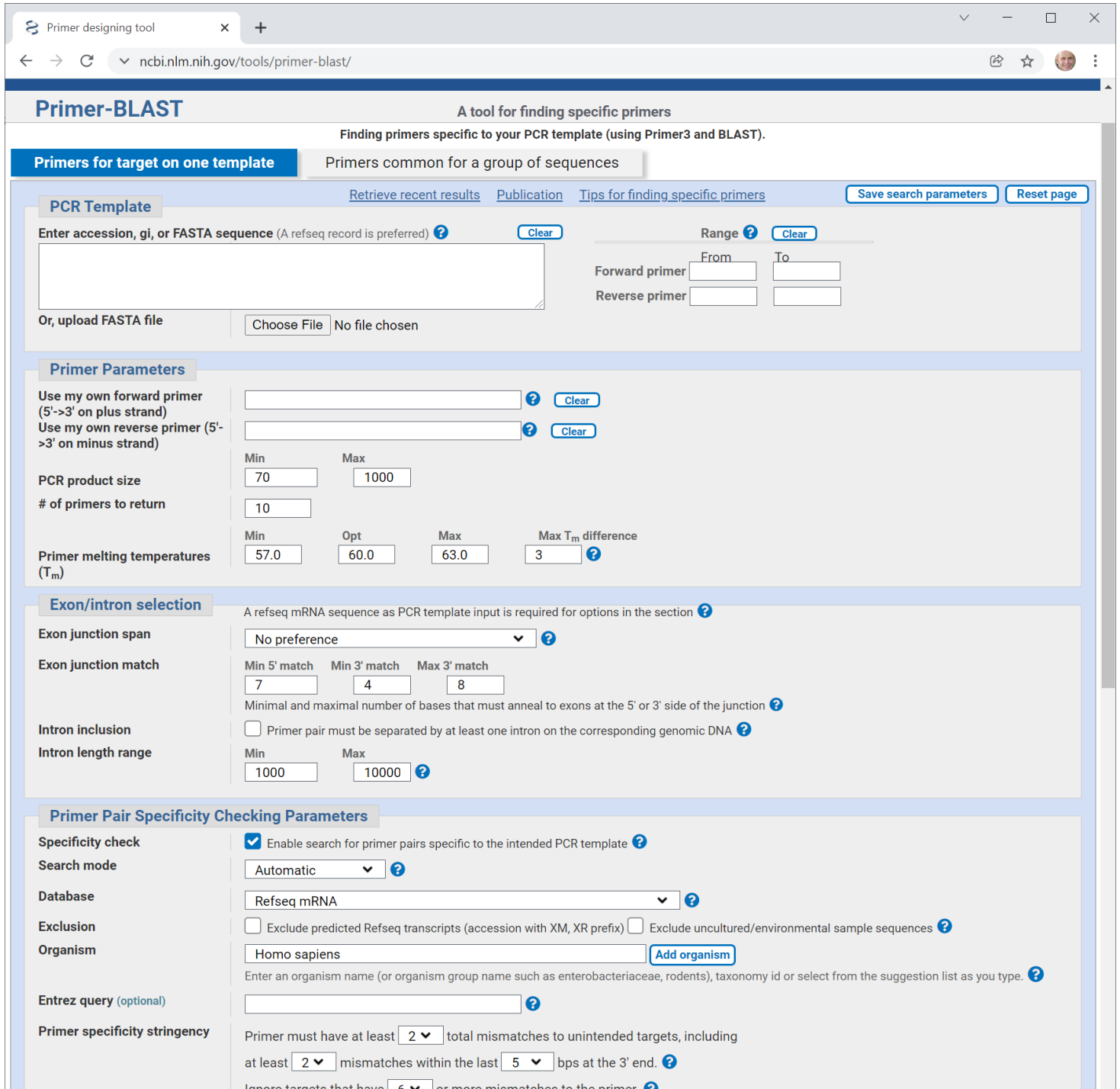
Sometimes unusual career paths



https://bioinfo.ut.ee/primer3/



https://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi

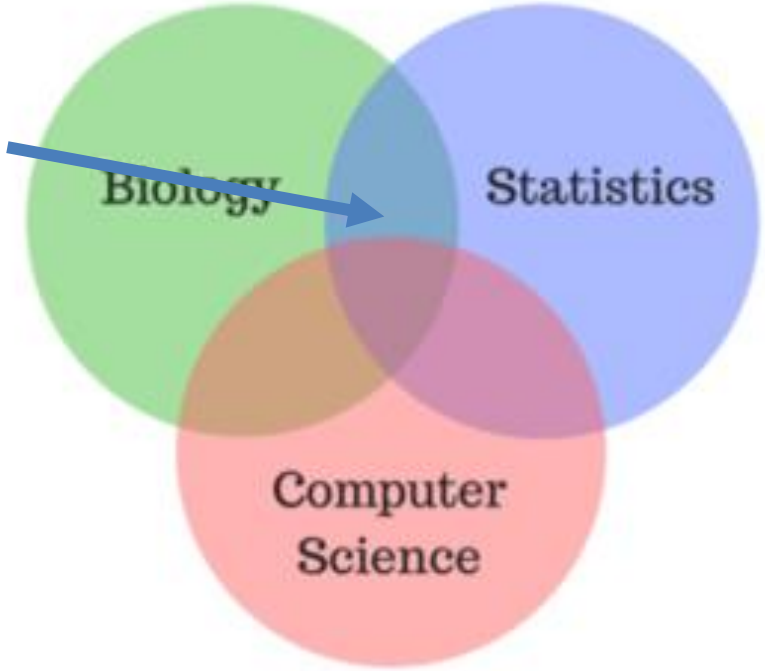


https://www.ncbi.nlm.nih.gov/tools/primer-blast/

Being a bioinformatician

Expertise in one area, competence in others

Naturally occurring aristolochic acid causes many cases of liver cancer



Sometimes unusual career paths

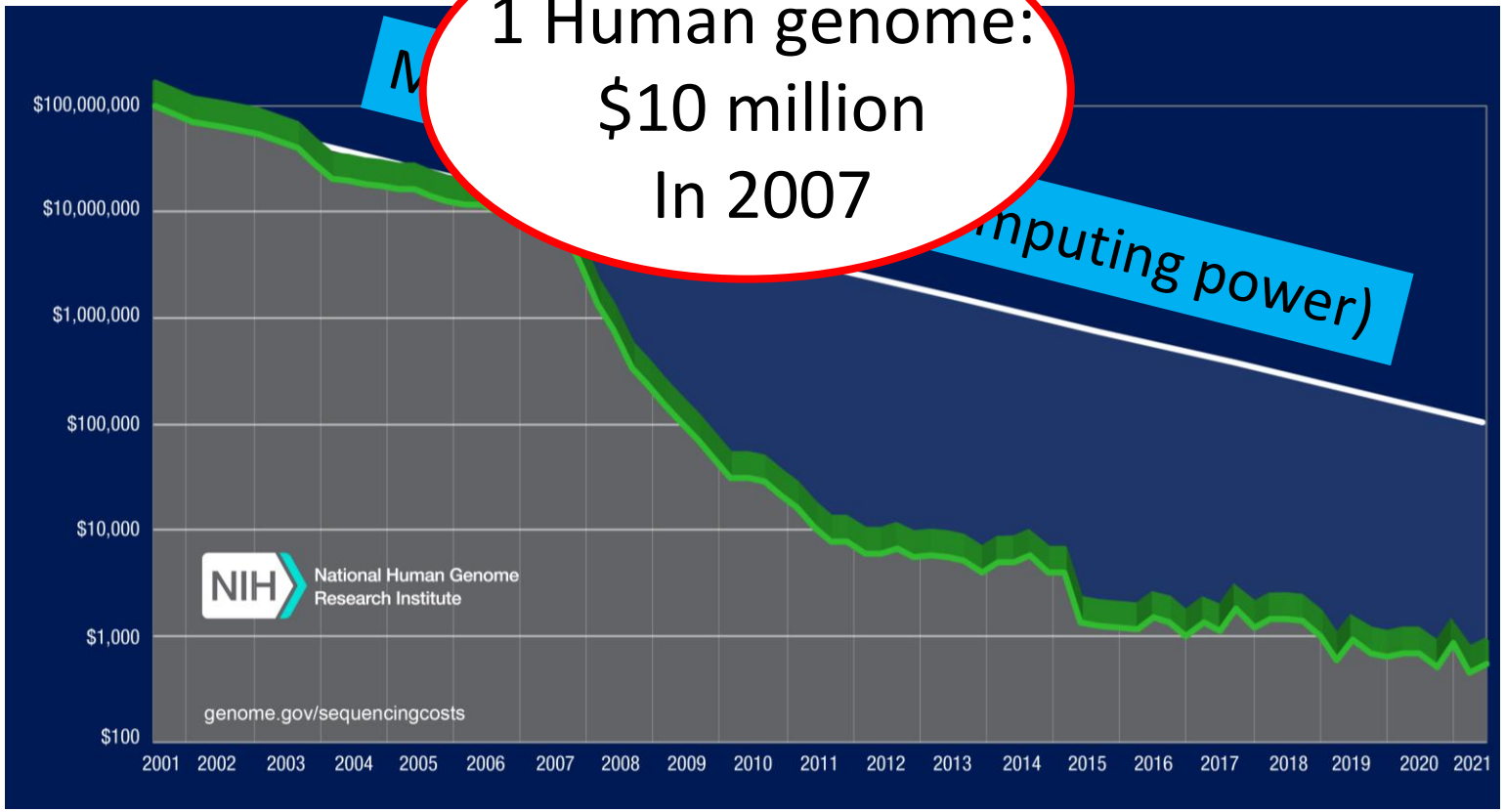
Bioinformatics depends on having a lot of data

- This started to happen in the late 1990s
- At that time it was based on technologies we will not discuss today (microarrays, etc)

Data generation really took off in 2007

Cost to sequence 1 human genome (US$)

Note log scale



10,000 X drop in price

US$ 1,000, today

Year

https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Dominant technology

MiniSeq System

Power and simplicity for targeted sequencing.

MiSeq Series

Small genome and targeted sequencing.

NextSeq Series

Everyday genome, exome transcriptome sequencing, and more.

HiSeq Series

Production-scale genome, exome, transcriptome sequencing, and more.

HiSeq X Series

Population- and production-scale human whole-genome sequencing.

NovaSeq Series

Population- and production-scale genome, exome, transcriptome sequencing, and more.

https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Need for computing infrastructure: $10 million worth of sequencers

DukeNUS Medical School

$> 10 million worth of
computing equipment to analyze data

Other sequencing technology, long reads

https://nanoporetech.com/

https://www.pacb.com/

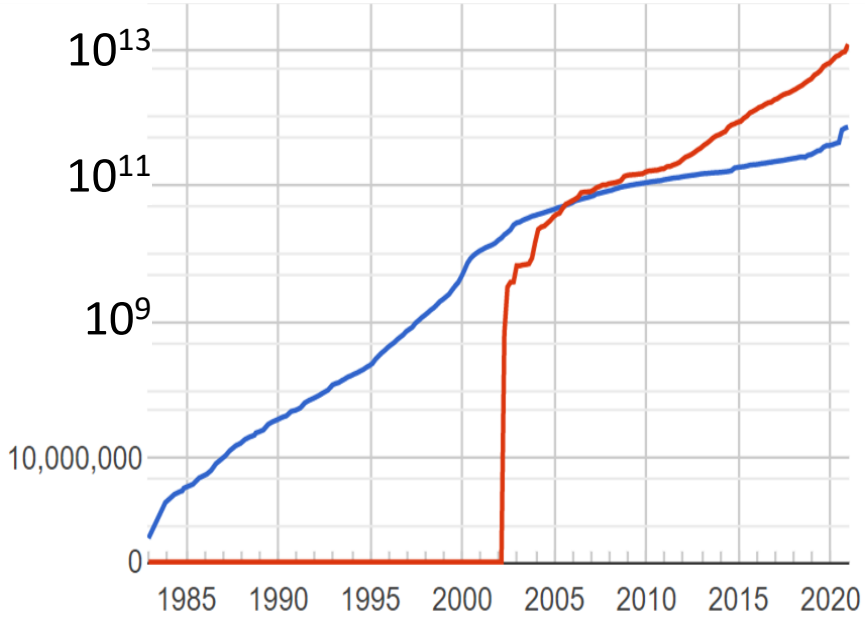
Sequencing is leading the genomics revolution

Sequencing drives perhaps 80% of bioinformatics research

Why is sequencing so important?

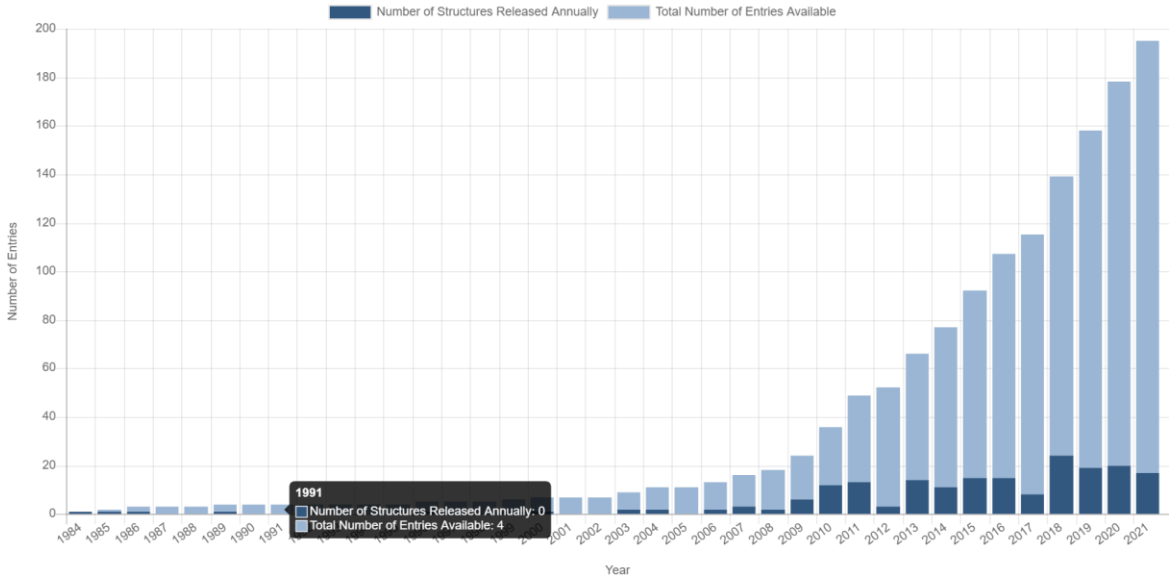
Contrast nucleic acid sequences with protein structures

GenBank sequence database:
13 trillion bases, 1.6 billion sequence entries;
doubling ~every 18 months



https://www.ncbi.nlm.nih.gov/genbank/statistics/

The number of protein structures in Protein DataBank (PDB, https://www.wwpdb.org/)



By Dcbmariano - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=87488007

Outline

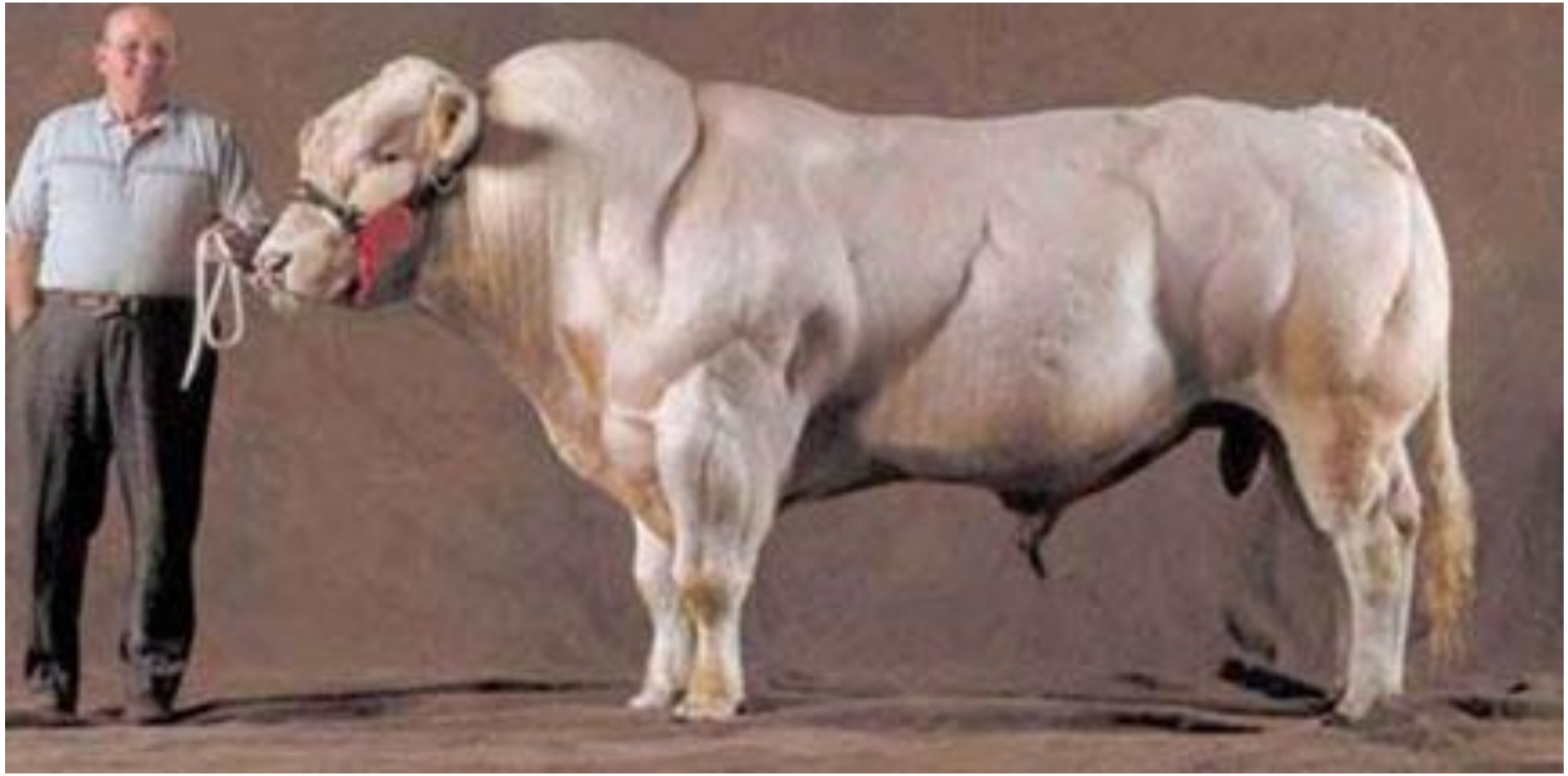
- Faculty introductions
- “What is Bioinformatics?” and aims scope of this course
- Next generation sequencing and bioinformatics
- Review of biological molecules
- Summary

DNA sequencing in context

- Your body contains roughly 30 trillion human cells
- Plus >100 trillion (mostly harmless) bacterial and other microbe cells
- Each cell contains two copies of your genome[1]
- Each copy of the genome is, abstractly, a string of 3 billion DNA letters: A, C, G, T
- 2% of this is “genes”
- Sometimes there are “spelling errors” in genes (“gene variants”)
- Next, two examples of in-born spelling errors…

[1] An oversimplification – some types of cells do not have any copies of the genome and sometimes cells have multiple copies.

Spelling error that inactivates a gene that suppresses muscle growth
(double negative is a positive)



Cleft hand

Sometimes caused by genetic variants in the SEM1 gene (also known as SHFM1)

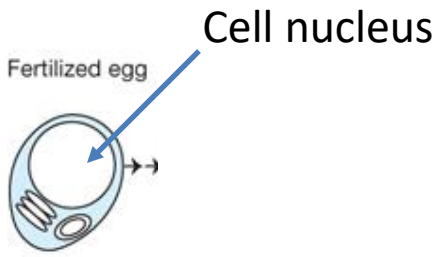
http://omim.org/clinicalSynopsis/119100

DNA mutations and cancer

- First, some vocabulary
- “Oncogenic mutations” – mutations that directly cause cancer
- This can be either because the oncogenic mutation…
  - turns OFF a gene that keeps cells from becoming cancer cells
  - …or…
  - turns ON a gene that encourages the cell to become a cancer cell

Cell divisions from the fertilized egg to a single cell within a cancer ….

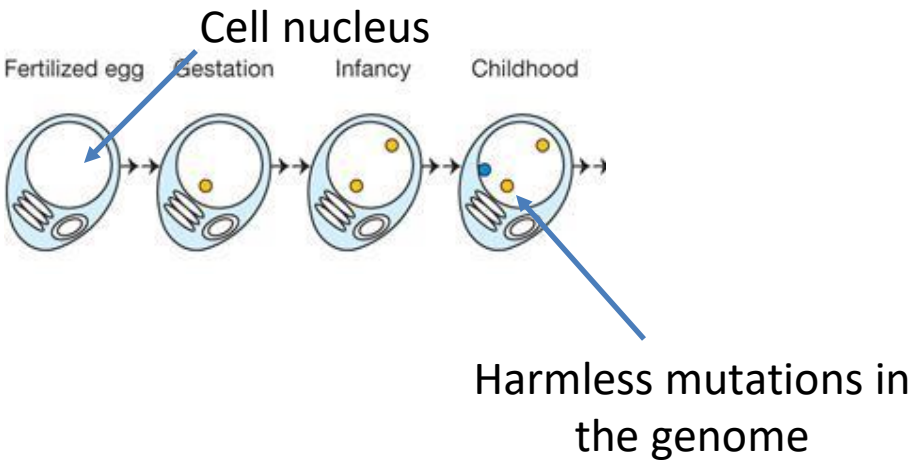
Somatic mutations are acquired by the cells over time



MR Stratton *et al. Nature* **458**, 719-724 (2009) doi:10.1038/nature07943

Cell divisions from the fertilized egg to a single cell within a cancer ….

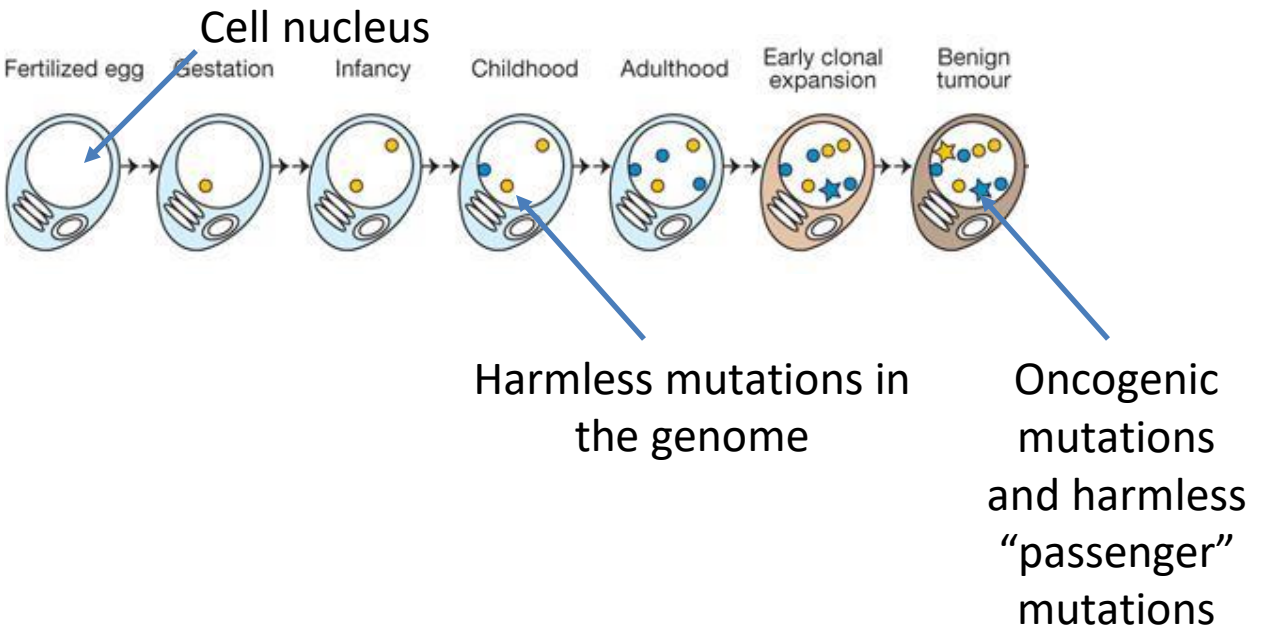
Somatic mutations are acquired by the cells over time



MR Stratton *et al. Nature* **458**, 719-724 (2009) doi:10.1038/nature07943

Cell divisions from the fertilized egg to a single cell within a cancer ….

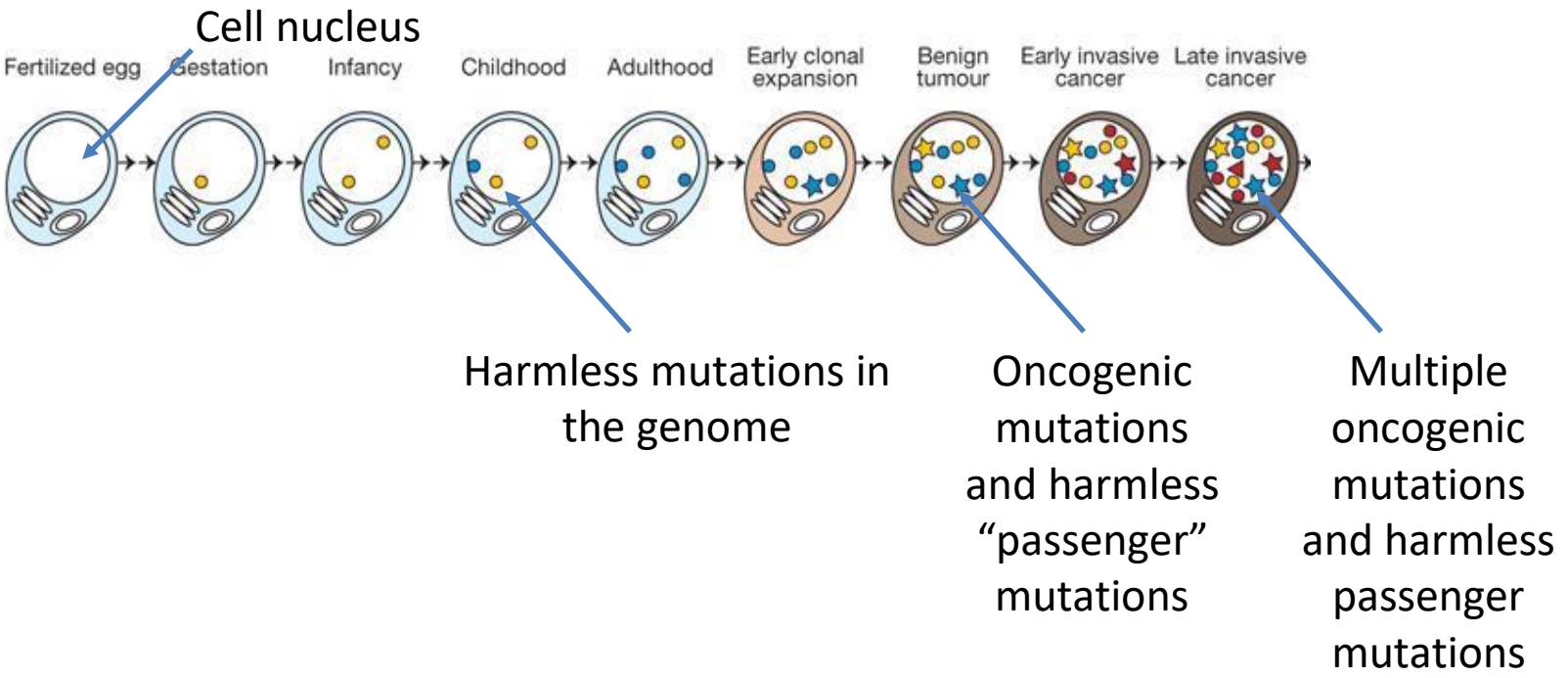
Somatic mutations are acquired by the cells over time



MR Stratton *et al. Nature* **458**, 719-724 (2009) doi:10.1038/nature07943

Cell divisions from the fertilized egg to a single cell within a cancer ….

Somatic mutations are acquired by the cells over time

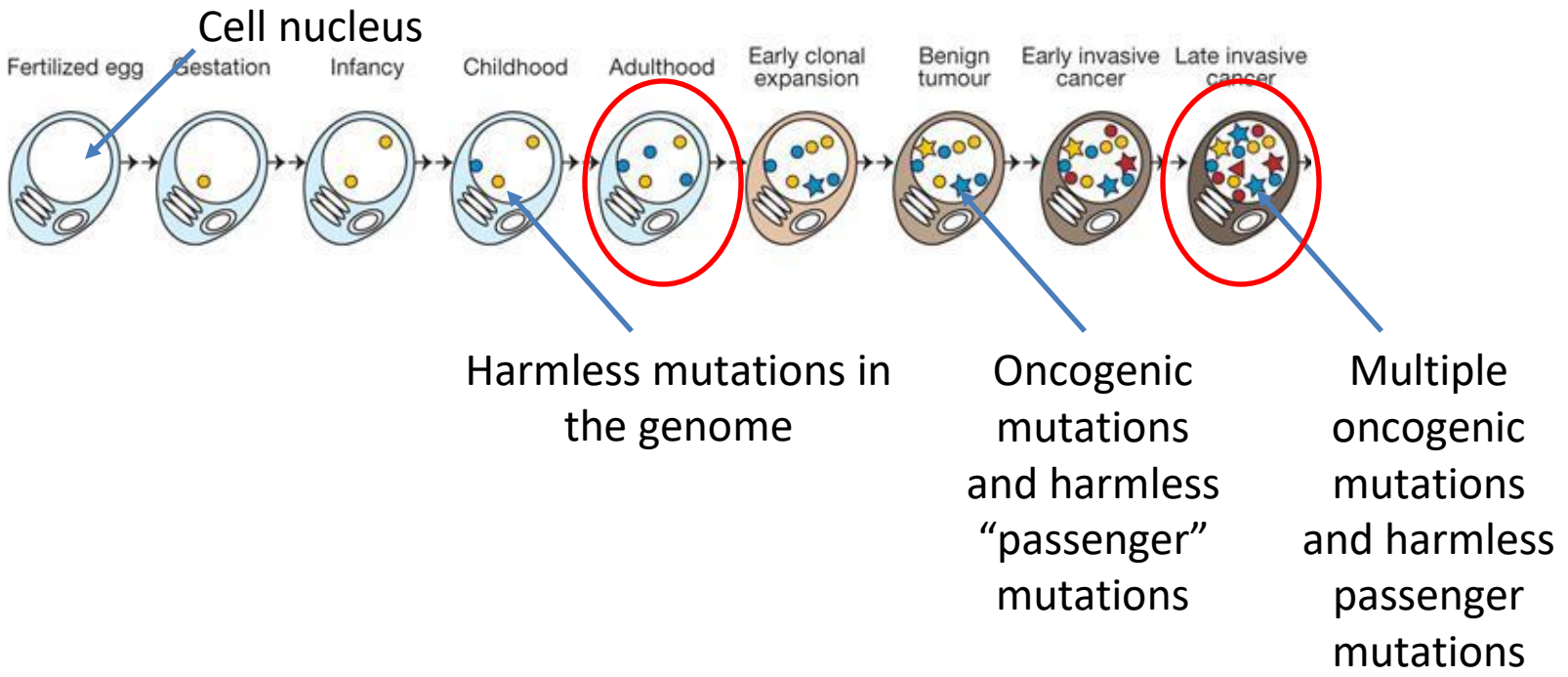


MR Stratton *et al. Nature* **458**, 719-724 (2009) doi:10.1038/nature07943

nature

Cell divisions from the fertilized egg to a single cell within a cancer ….

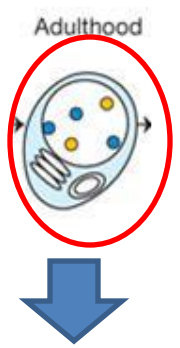
Somatic mutations are acquired by the cells over time



MR Stratton *et al. Nature* **458**, 719-724 (2009) doi:10.1038/nature07943

nature

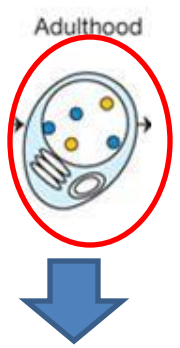
We Identify cancer specific somatic mutations by large-scale sequencing (NGS)



…ACGTCCTAGTCAAAATCGAGCC…

…ACGTCCTAGTCTAAATCGAGCC…

We Identify cancer specific somatic mutations by large-scale sequencing (NGS)



…ACGTCCTAGTC**A**AAATCGAGCC…

…ACGTCCTAGTC**T**AAATCGAGCC…

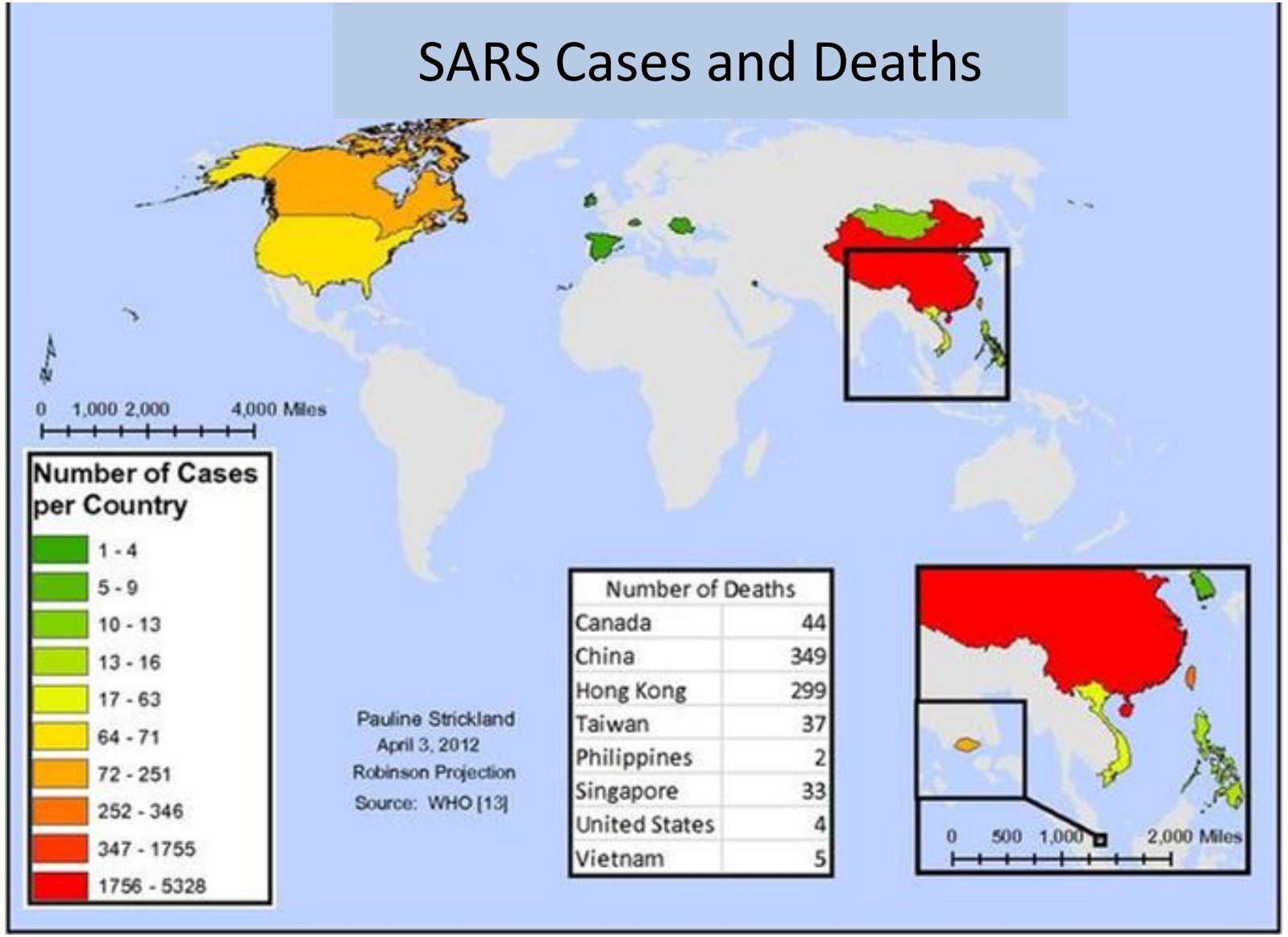
Somatic mutation from **A** > **T**

(in a CAA trinucleotide context, so C**A**A > C**T**A)

What the data actually look like

Shorts lengths of DNA read by sequencer (150 DNA letters long)

Pathogen discovery and evolution



https://en.wikipedia.org/wiki/Asian_palm_civet#/media/File:Asian_Palm_Civet_Over_A_Tree.jpg

https://en.wikipedia.org/wiki/File:Sars_Cases_and_Deaths.pdf

Zoonosis

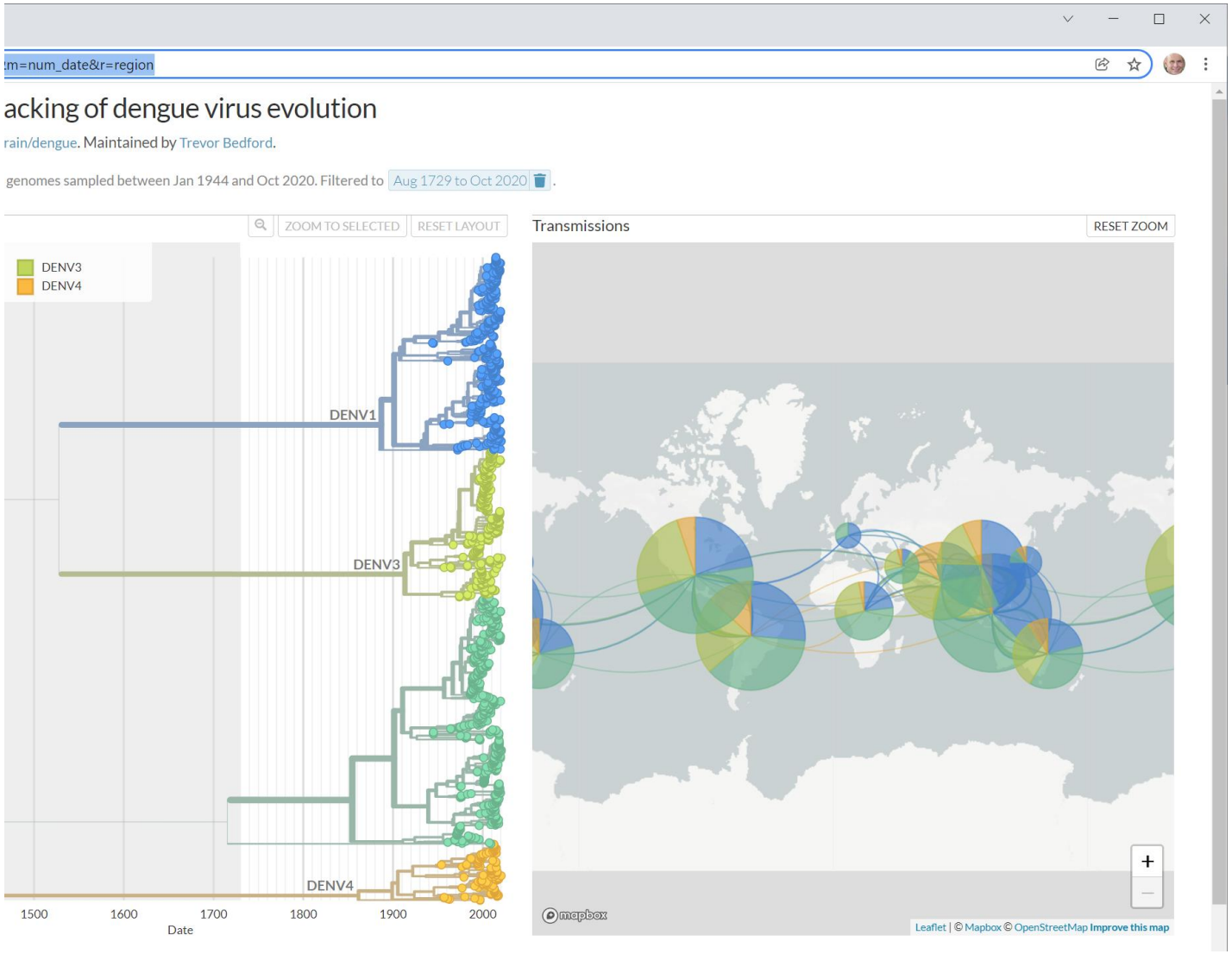
https://www.theborneopost.com/2013/03/31/protected-wildlife-on-the-menu/

https://www.washingtonpost.com/science/2020/04/03/coronavirus-wildlife-environment/

A market where pangolins and other exotic animals are sold in Libreville, Gabon. (Steeve Jordan/AFP/Getty Images)

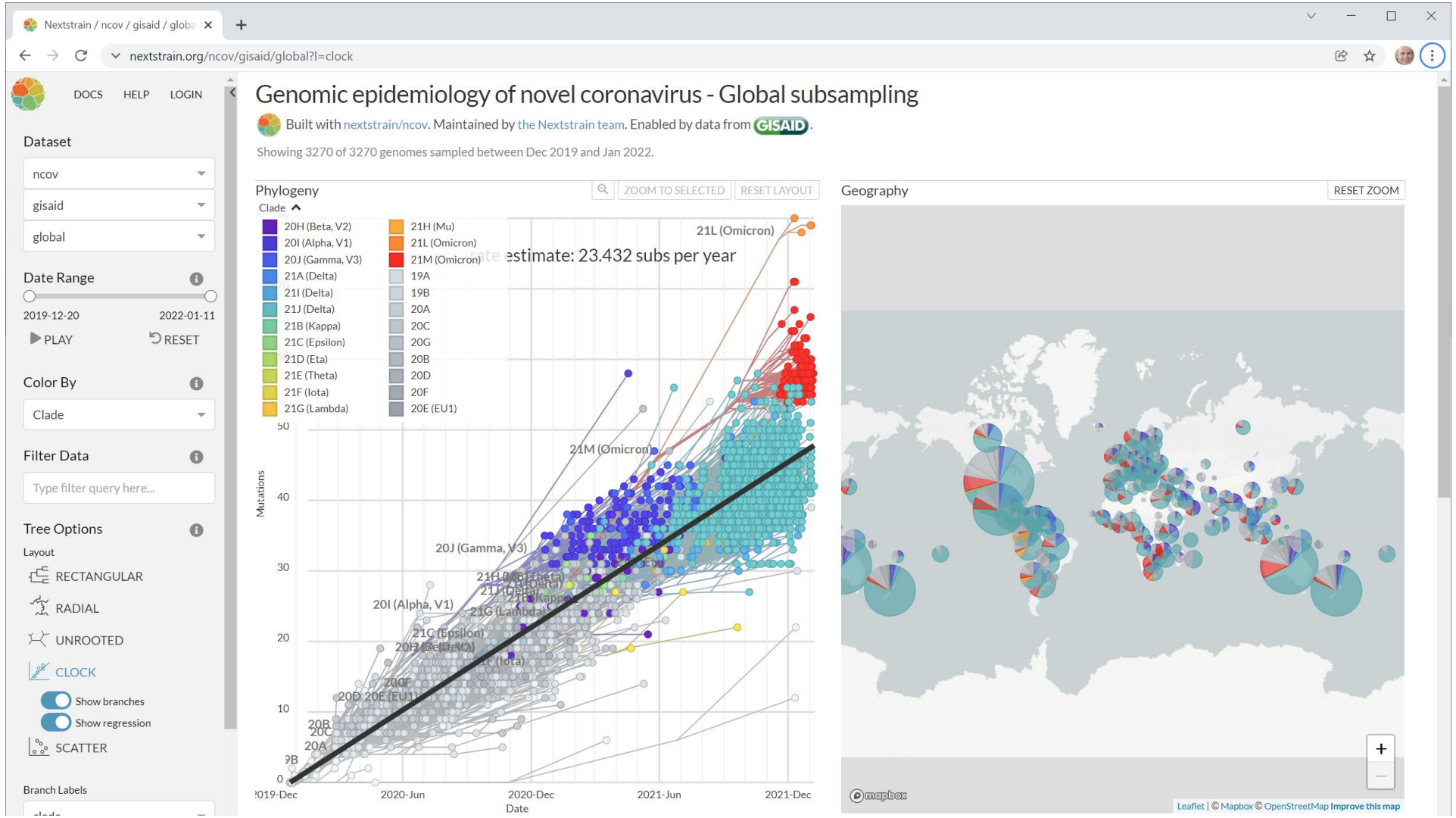
Zoonosis

Dengue virus phylogeny



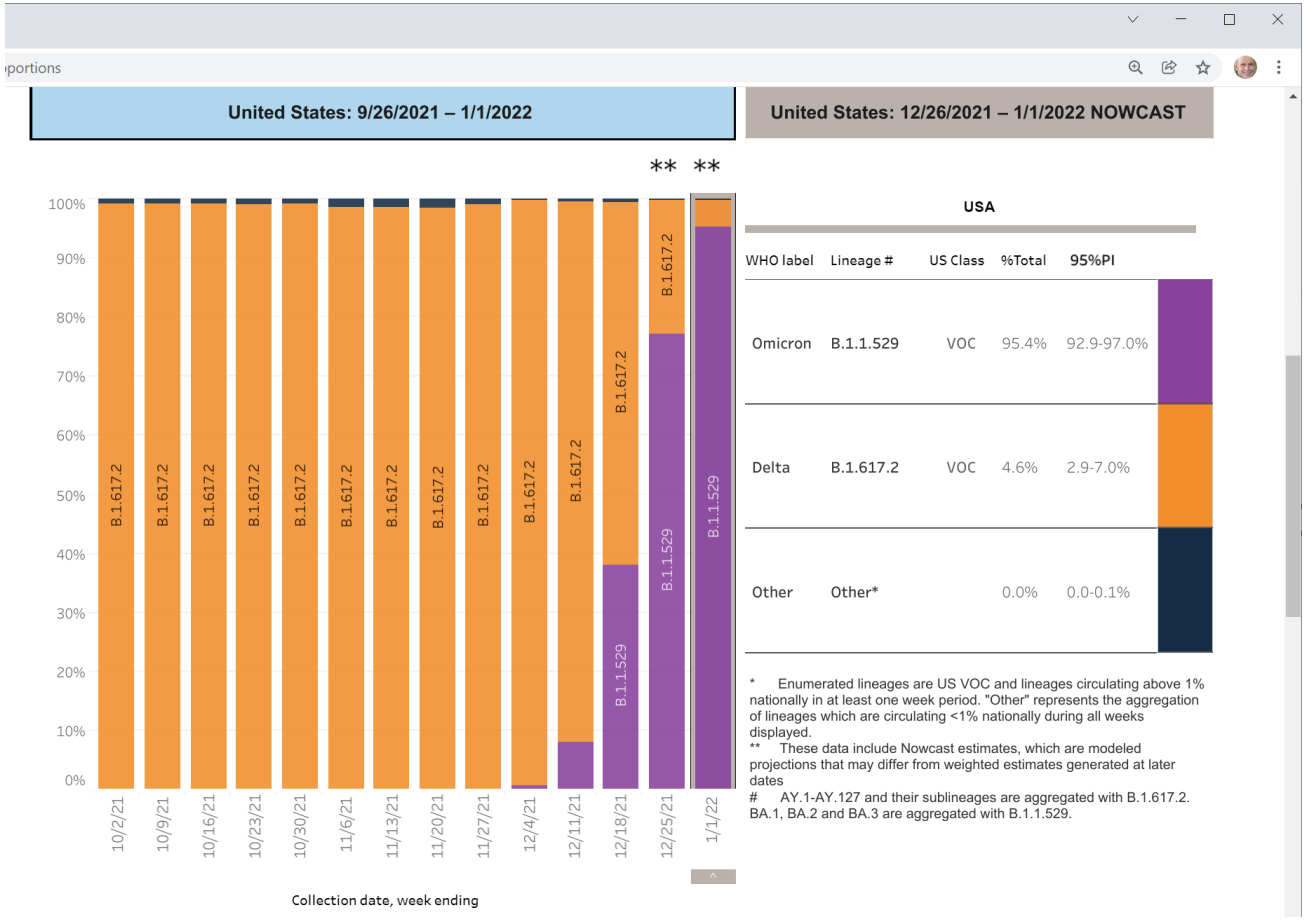
https://nextstrain.org/dengue/all?dmin=1729-08-31&m=num_date&r=region

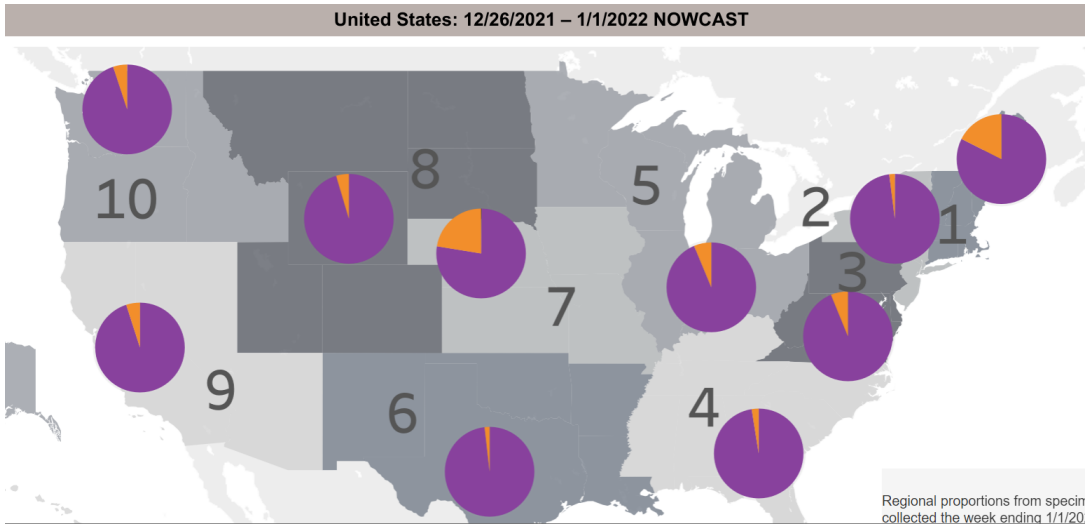
Novel coronavirus phylogeny



https://nextstrain.org/ncov/global

Simplified and more up to date





https://covid.cdc.gov/covid-data-tracker/#variant-proportions

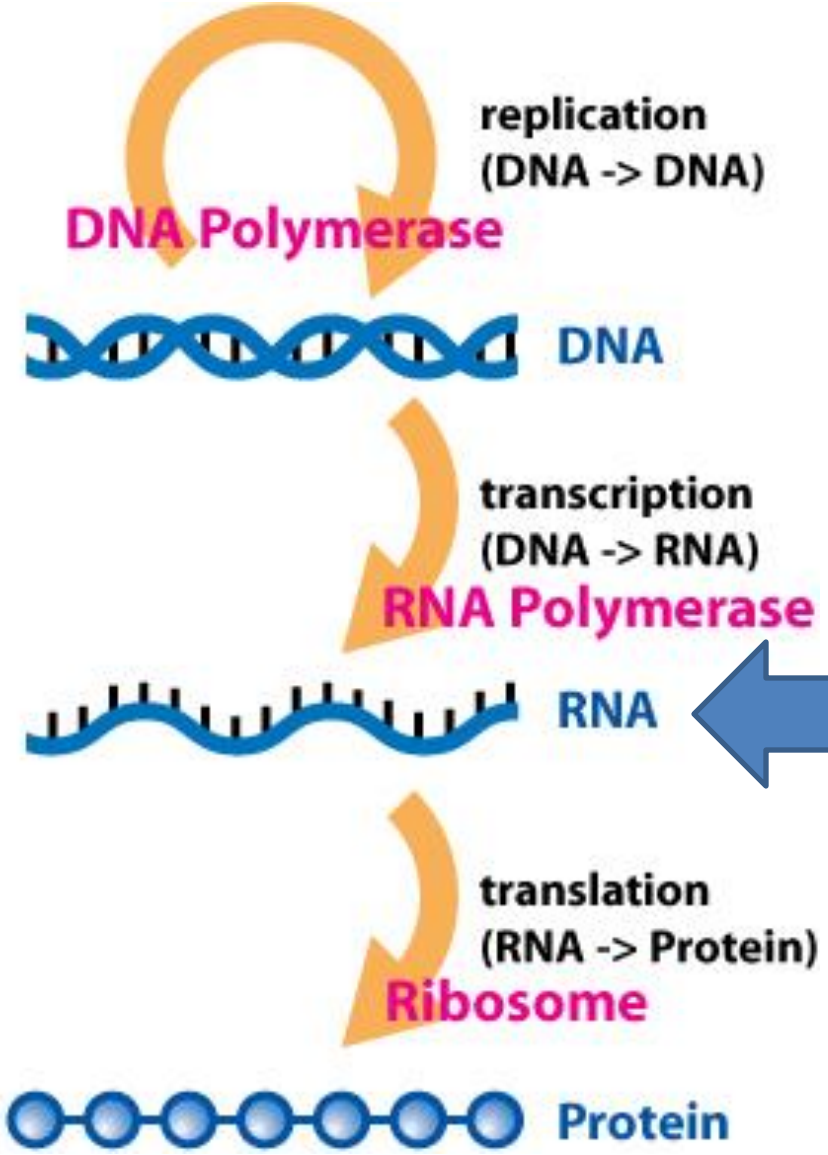
Summary of many applications of genome sequencing

- Human variation / human genetics
- Infectious diseases
- Cancer genetics
- Plant and animal breeding for agriculture
- Sequencing new genomes
- Metagenomics --- sequencing bacteria and other microorganisms in
  - Our bodies
  - Anywhere else on earth (soil, deep sea thermal vents, under ice caps …)

That was genome sequencing

But there’s more….

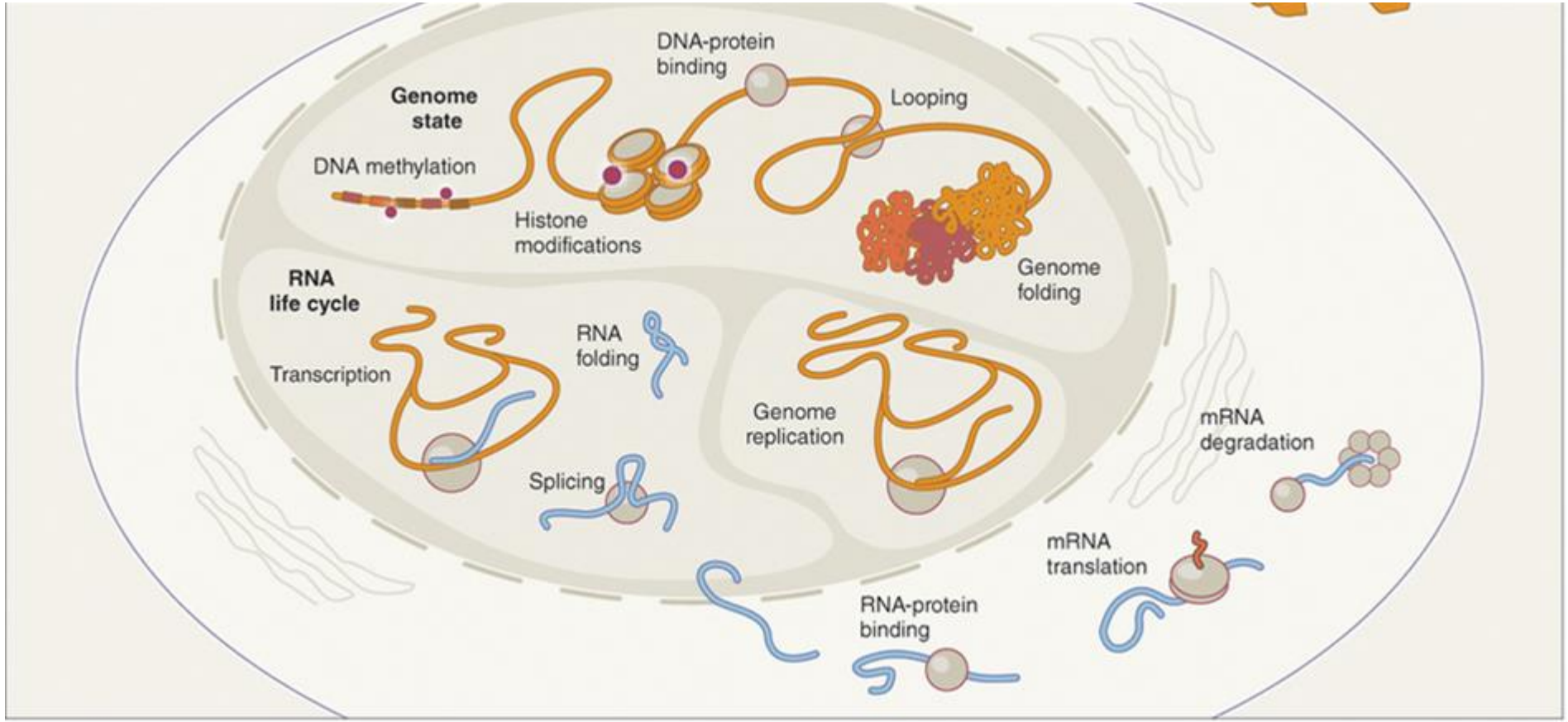
“Central Dogma”



RNA-seq:
Sequence the RNA in a tissue to get a picture of which genes are active (actually, the level of their activity)

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

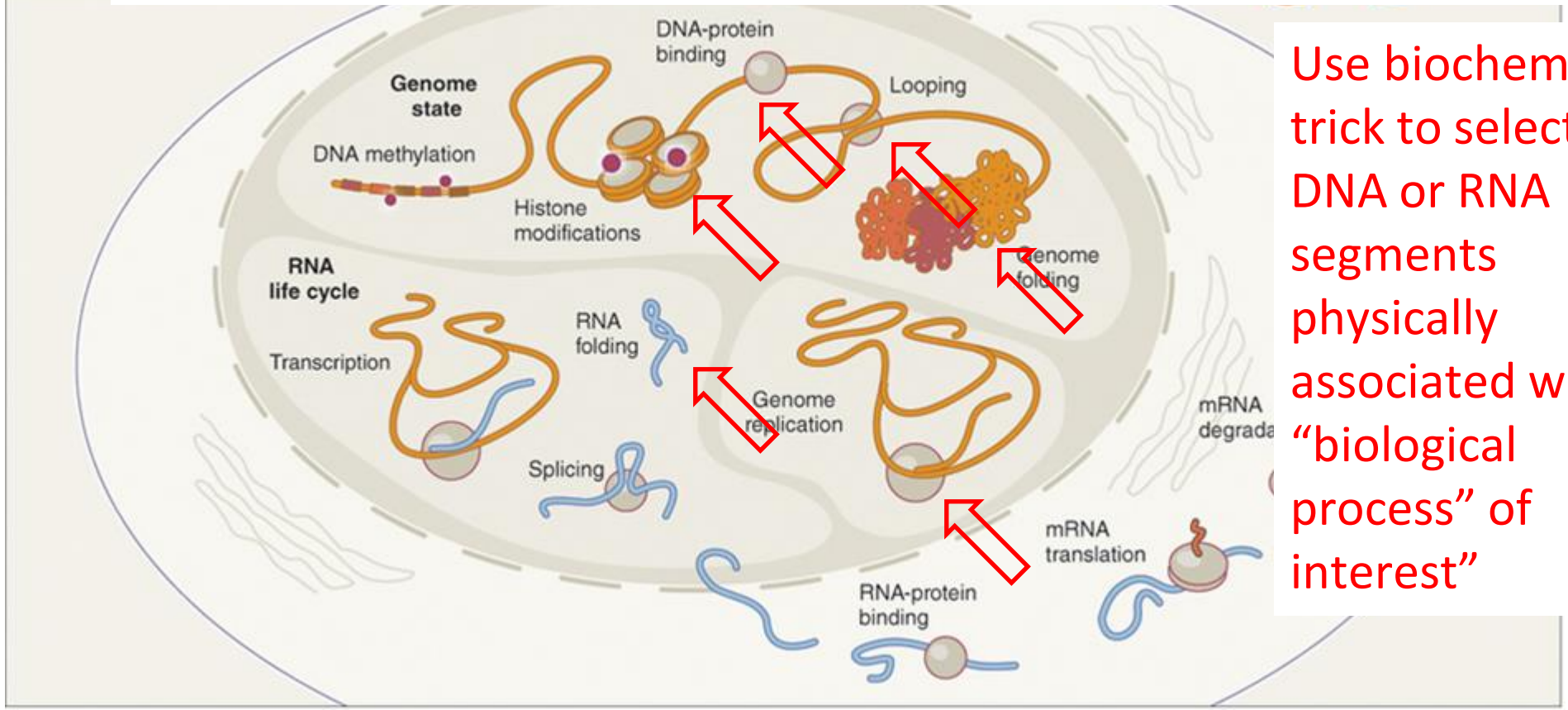
Sequencing is a foundational technology



Shendure & Aiden,
Nat. Biotech. 2012

Sequencing is a foundational technology

Snippets of DNA or RNA are self-identifying (self-barcoded)

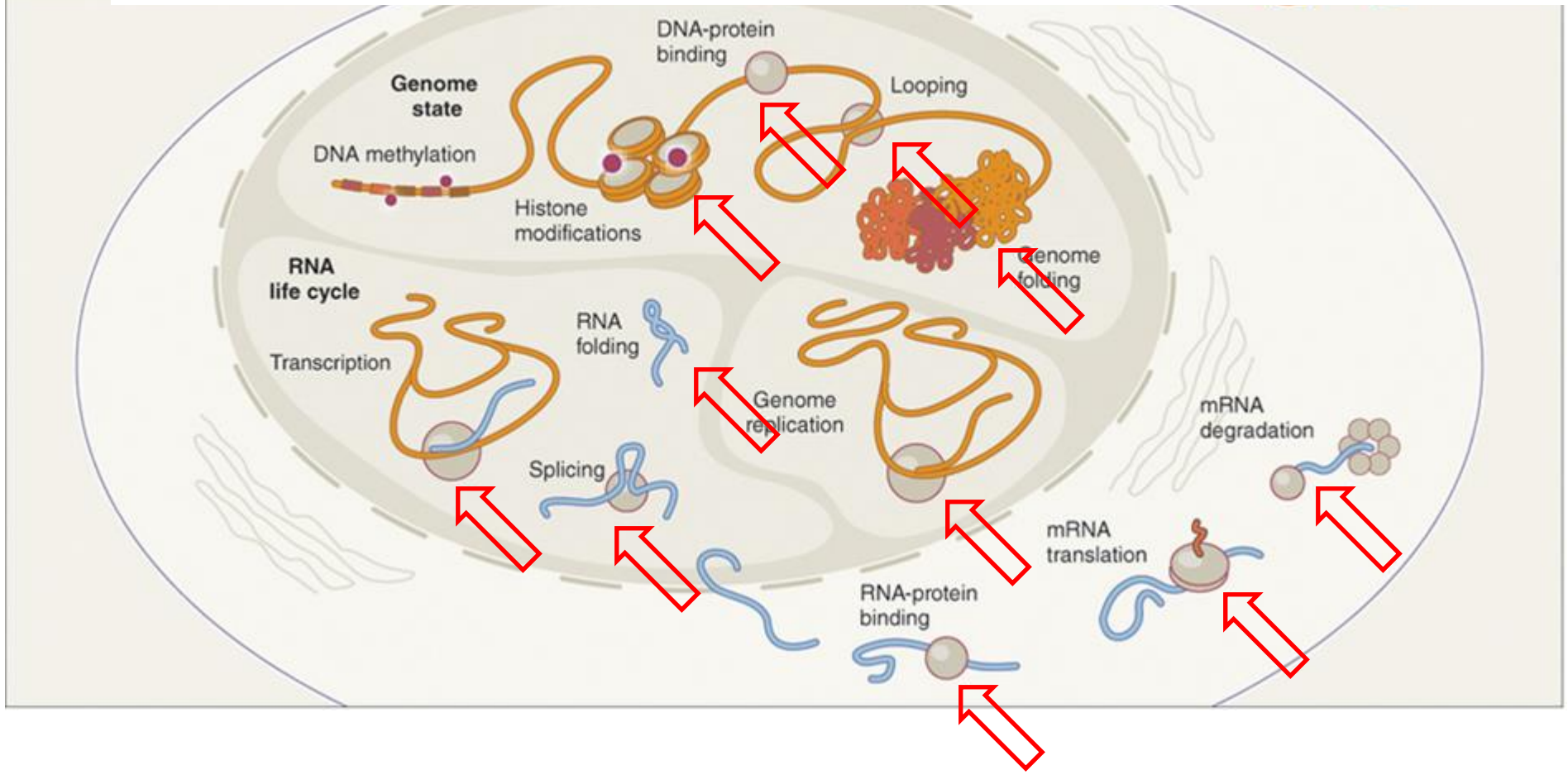


Use biochemical trick to select for DNA or RNA segments physically associated with "biological process" of interest"

Sequencing is a foundational technology

Describ

Snippets of DNA or RNA are self-identifying (self-barcoded)



Applications of NGS

Method	
DNA-Seq	
Targeted DNA-Seq	
RNA-Seq	
Methyl-Seq	
Targeted methyl-Seq	
DNase-Seq, Sono-Seq	
FAIRE-Seq (formaldeh isolation of regulatory	
MAINE-Seq (MNase-a nucleosomes)	
ChIP-Seq	
RIP-Seq (RNA-binding protein immunoprecipitation)	Protein-RNA interactions
CLIP-Seq (cross-linking IP)	Protein-RNA interactions

Many applications

- Human variation / human genetics
- Cancer genetics
- Clinical applications in human genetics and cancer genetics
- Plant and animal breeding for agriculture
- Sequencing new genomes
- Metagenomics
- Pathogen discovery
- Pathogen evolution

Adapted from Shendure and Aiden, Nat. Biotech. 2012

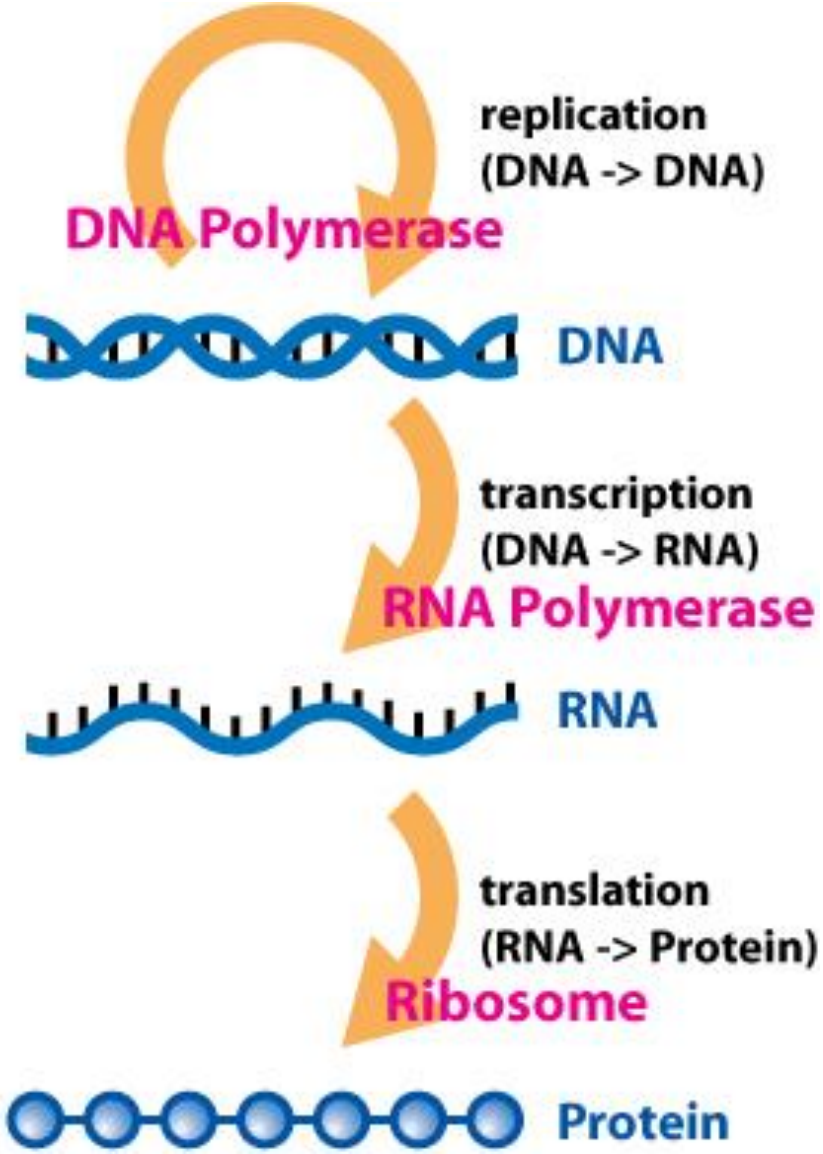
More applications of NGS

Method	To determine…
HITS-CLIP (high-throughput sequencing of RNA isolated by cross-linking IP)	Protein-RNA interactions
FRT-Seq (on-flowcell reverse transcription)	Amplification-free, strand-specific transcriptome sequencing
NET-Seq (native elongating transcript)	Nascent transcription
Hi-C	Three-dimensional genome structure
Chia-PET	Long-range interactions mediated by a protein
Ribo-Seq	Ribosome-protected mRNA fragments (active translation)
TRAP (translating ribosome affinity purification)	Genetically targeted purification of polysomal mRNAs
PARS	Parallel analysis of RNA structure
Synthetic saturation mutagenesis	Functional consequences of genetic variation
Immuno-Seq	The B-cell and T-cell repertoires
Deep protein mutagenesis	Protein binding activity of synthetic peptide libraries or variants
PhIT-Seq (phenotypic interrogation via tag)	Relative fitness of cells containing disruptive insertions in diverse genes

Adapted from Shendure and Aiden, Nat. Biotech. 2012

Outline

- Faculty introductions
- “What is Bioinformatics?” and aims scope of this course
- Next generation sequencing and bioinformatics
- Review of biological molecules
- Summary



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

The “genetic code”: how cells ***translate*** mRNAs to proteins

Amino acids biochemical properties | nonpolar | polar | basic | acidic

Termination: stop codon

Standard genetic code

1st base	2nd base U		2nd base C		2nd base A		2nd base G		3rd base
U	UUU	(Phe/F) Phenylalanine	UCU	(Ser/S) Serine	UAU	(Tyr/Y) Tyrosine	UGU	(Cys/C) Cysteine	U
	UUC		UCC		UAC		UGC		C
	UUA	(Leu/L) Leucine	UCA		UAA	Stop (*Ochre*) [B]	UGA	Stop (*Opal*) [B]	A
	UUG		UCG		UAG	Stop (*Amber*) [B]	UGG	(Trp/W) Tryptophan	G
C	CUU	(Leu/L) Leucine	CCU	(Pro/P) Proline	CAU	(His/H) Histidine	CGU	(Arg/R) Arginine	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	(Ile/I) Isoleucine	ACU	(Thr/T) Threonine	AAU	(Asn/N) Asparagine	AGU	(Ser/S) Serine	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	AUG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GUU	(Val/V) Valine	GCU	(Ala/A) Alanine	GAU	(Asp/D) Aspartic acid	GGU	(Gly/G) Glycine	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GUG		GCG		GAG		GGG		G

[A] The codon AUG both codes for methionine and serves as an initiation site: the first AUG in an mRNA's coding region is where translation into protein begins.[43]

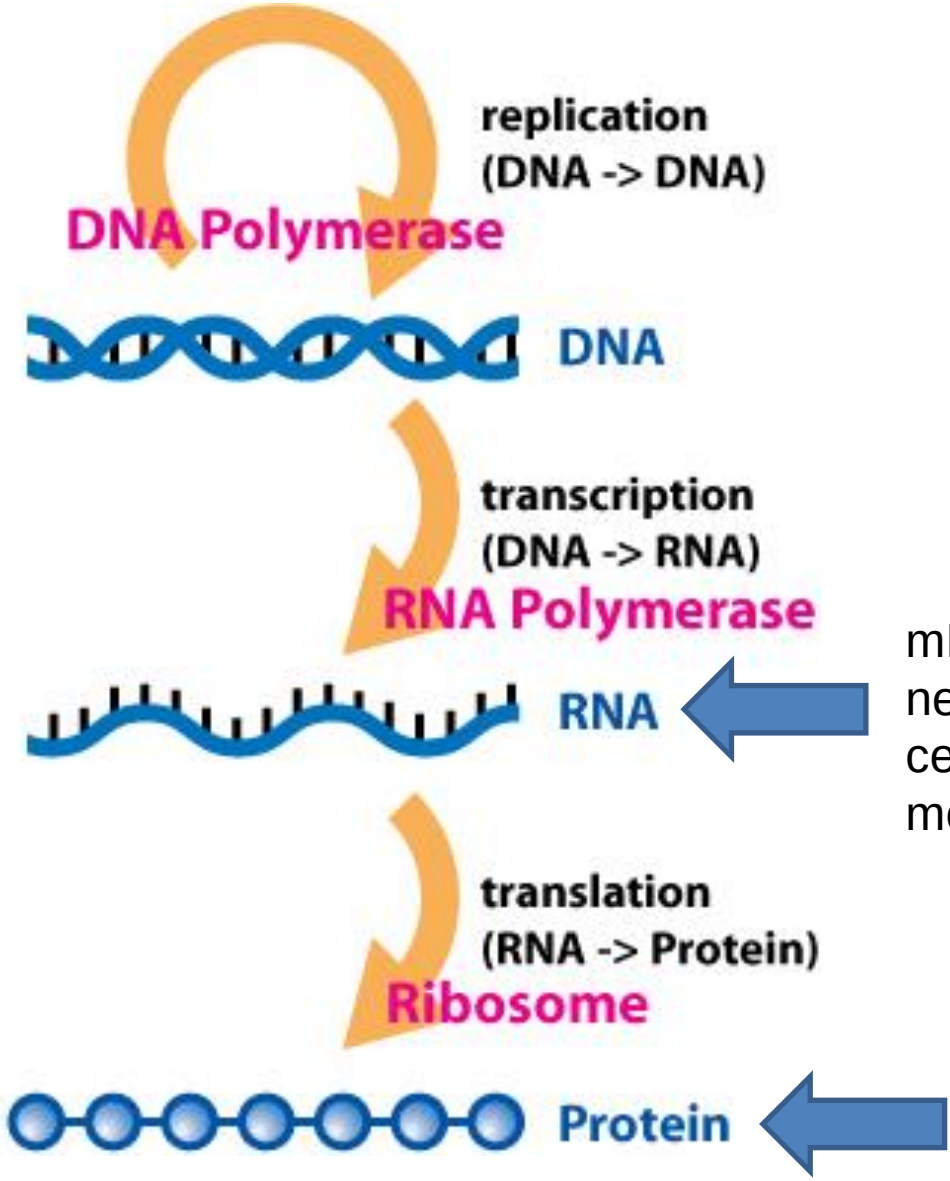
[B] ^ ^ ^ The historical basis for designating the stop codons as amber, ochre and opal is described in an autobiography by Sydney Brenner[44] and in a historical article by Bob Edgar.[45]

Each triplet, e.g. UUU (U = Uracil, which the RNA version of T = Thymine in DNA), is called a **“codon”**

You need to remember that AUG (ATG) encodes methionine and is the “**start codon**” -- every protein starts with methionine.

You also have to remember that 3 codons (UAA, UAG, and UGA) do not code for an amino acid, but rather indicate the end of the protein. These are the **“stop codons”.** No need to memorize which ones they are.

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



mRNA vaccines work by inserting new (modified) RNA into your cells (uracils are replaced by a modified version)

The vaccine mRNA encodes a slightly modified coronavirus spike protein so less likely bind to ACE2

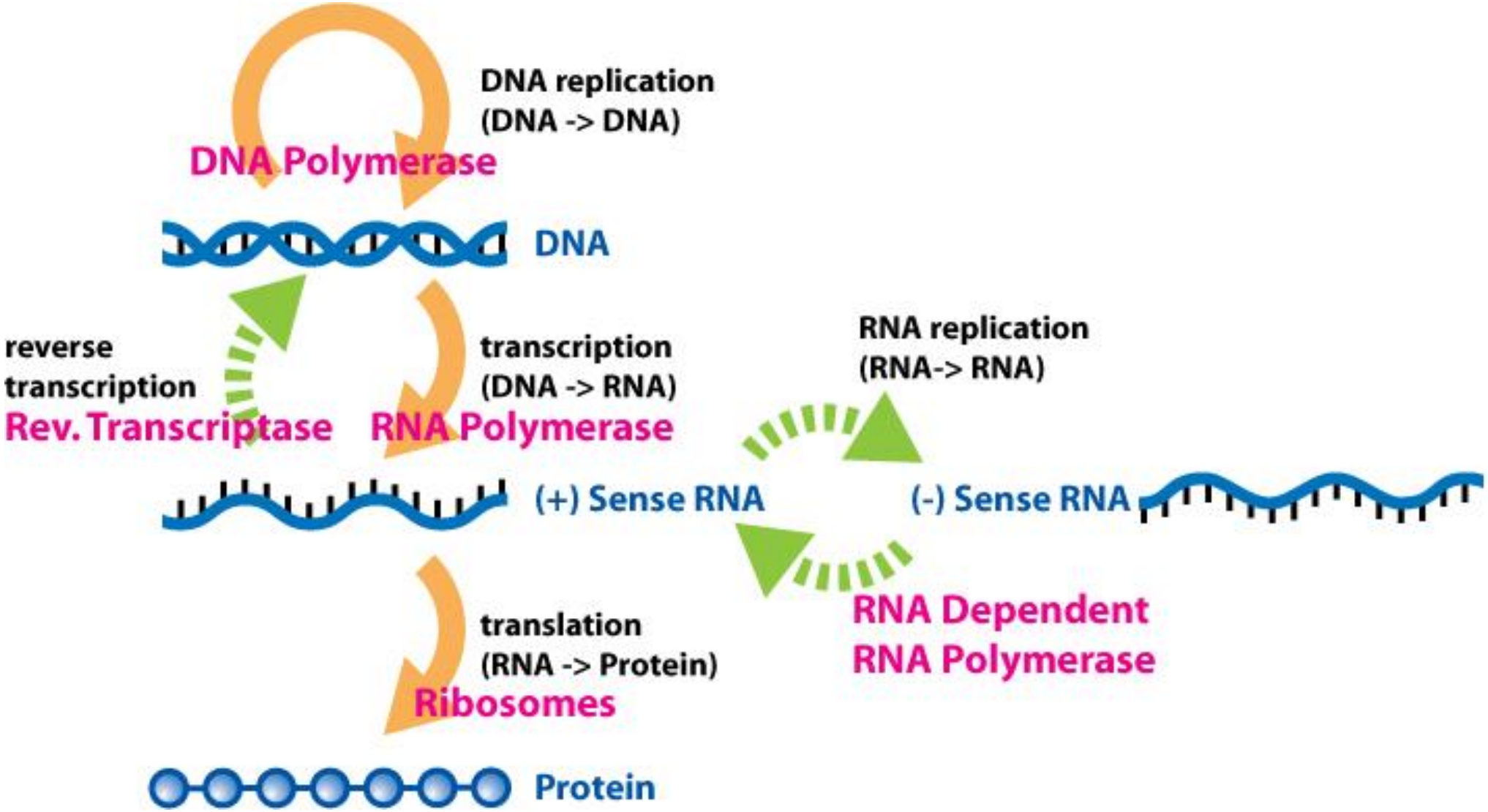
https://www.modernatx.com/pipeline/therapeutic-areas/mrna-therapeutic-areas-infectious-diseases

https://en.wikipedia.org/wiki/Moderna_COVID-19_vaccine

https://en.wikipedia.org/wiki/Pfizer%E2%80%93BioNTech_COVID-19_vaccine

Great summary of SARS-CoV-2 biology:
https://www.nature.com/articles/d41586-021-02039-y

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



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

Biomolecules

- DNA
- RNA
- Proteins (and peptides)
- Lipids
- “Small molecules”
  - Examples from neurotransmitters

Example small molecules (neurotransmitters)

- **Amino acids:** glutamate, aspartate, D-serine, γ-aminobutyric acid (GABA), glycine
- **Gasotransmitters:** nitric oxide (NO), carbon monoxide (CO), hydrogen sulfide (H2S)
- **Monoamines:** dopamine (DA), norepinephrine (noradrenaline; NE, NA), epinephrine (adrenaline), histamine, serotonin (SER, 5-HT)
- **Trace amines:** phenethylamine, N-methylphenethylamine, tyramine, 3-iodothyronamine, octopamine, tryptamine, etc.
- **Peptides:** somatostatin, substance P, cocaine and amphetamine regulated transcript, opioid peptides
- **Purines:** adenosine triphosphate (ATP), adenosine
- **Others:** acetylcholine (ACh), anandamide
- (https://en.wikipedia.org/wiki/Neurotransmitter)

EUKARYOTIC GENOME ORGANIZATION

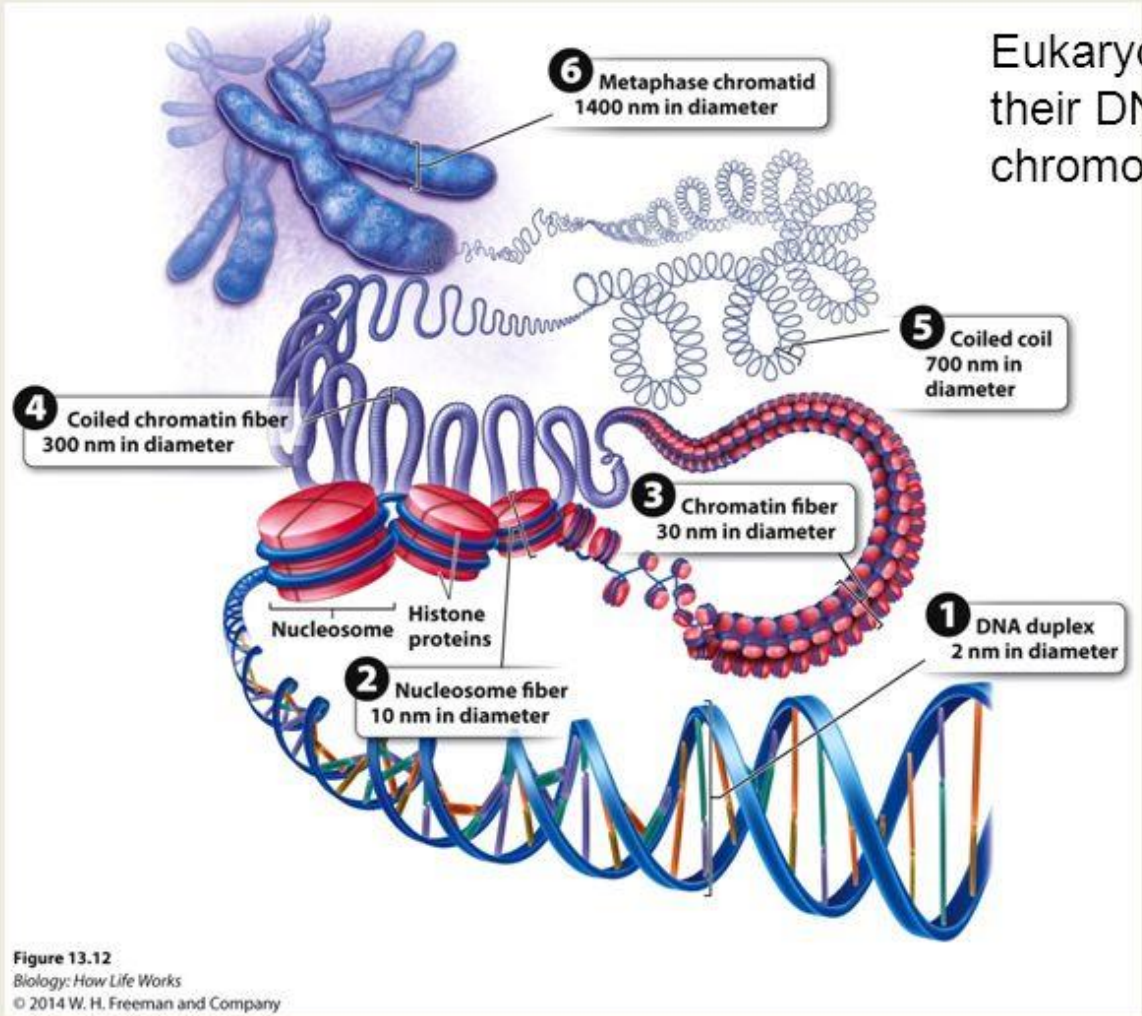


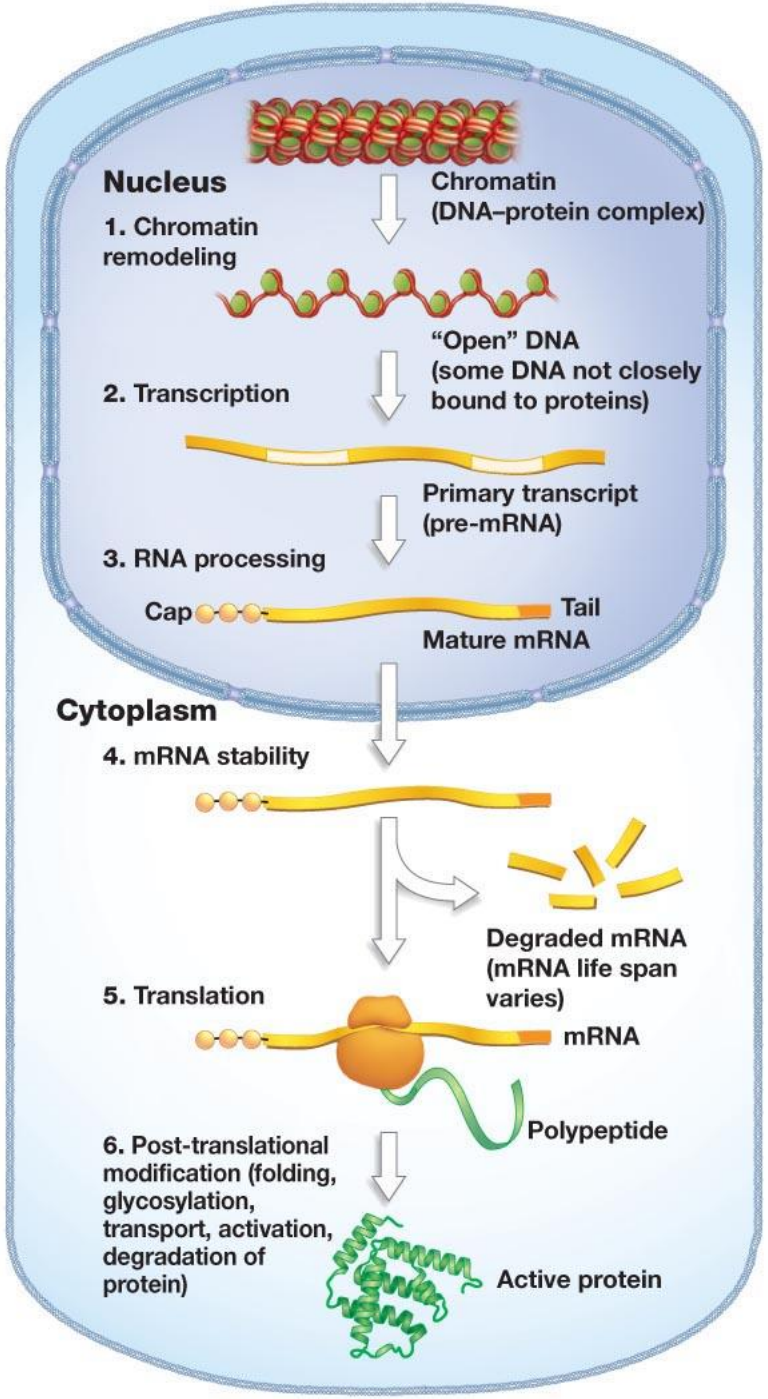
Figure 13.12
Biology: How Life Works
© 2014 W. H. Freeman and Company

Eukaryotic cells package their DNA as 1 molecule/linear chromosome

6 levels of chromosome packing:

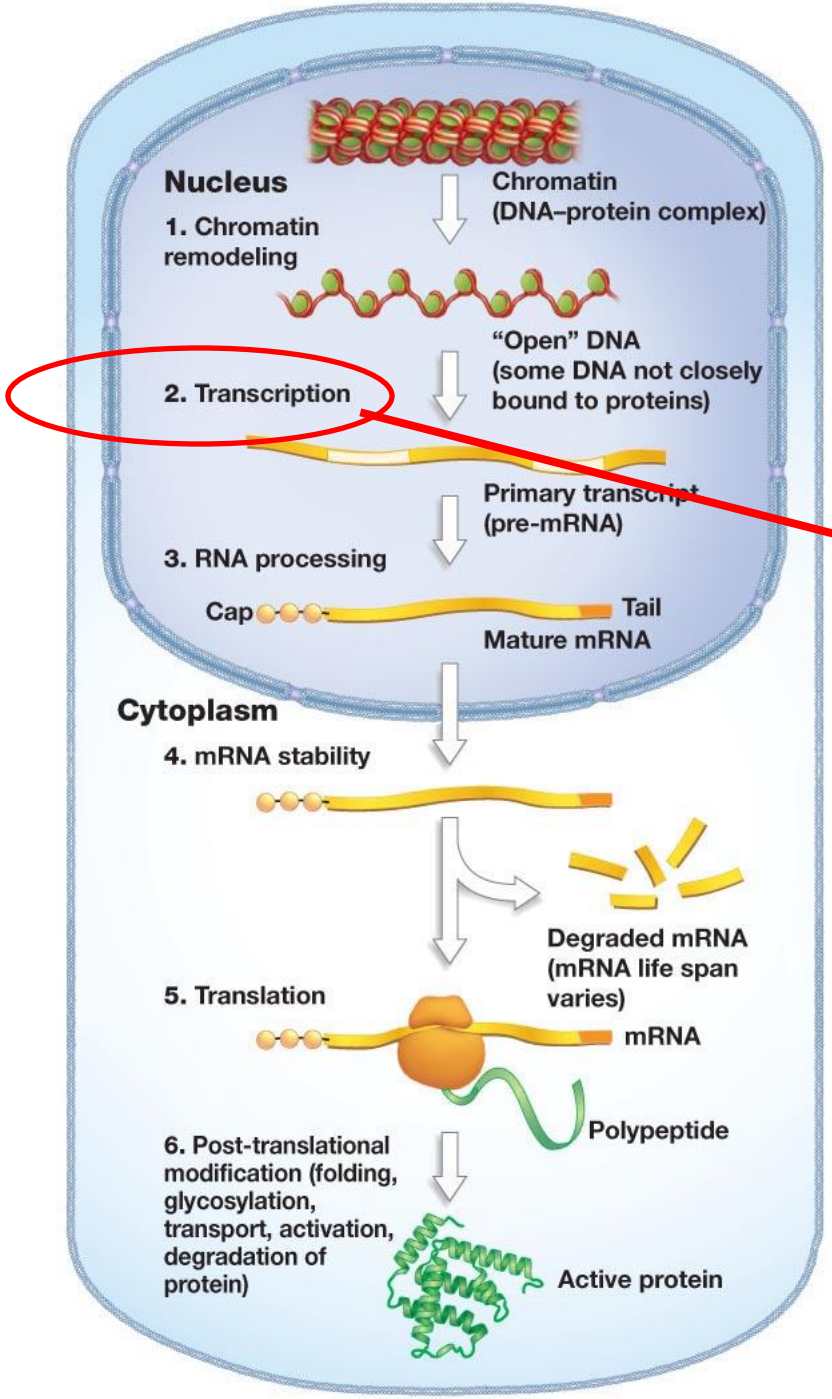
1. DNA duplex (2 nm)
2. Nucleosome fiber(10 nm)
   - Beads on a string
   - DNA with histone proteins
3. 30 nm chromatin fiber
4. Coiled chromatin fiber
5. Coiled coil
6. Metaphase chromosome

Interpreting the genetic program

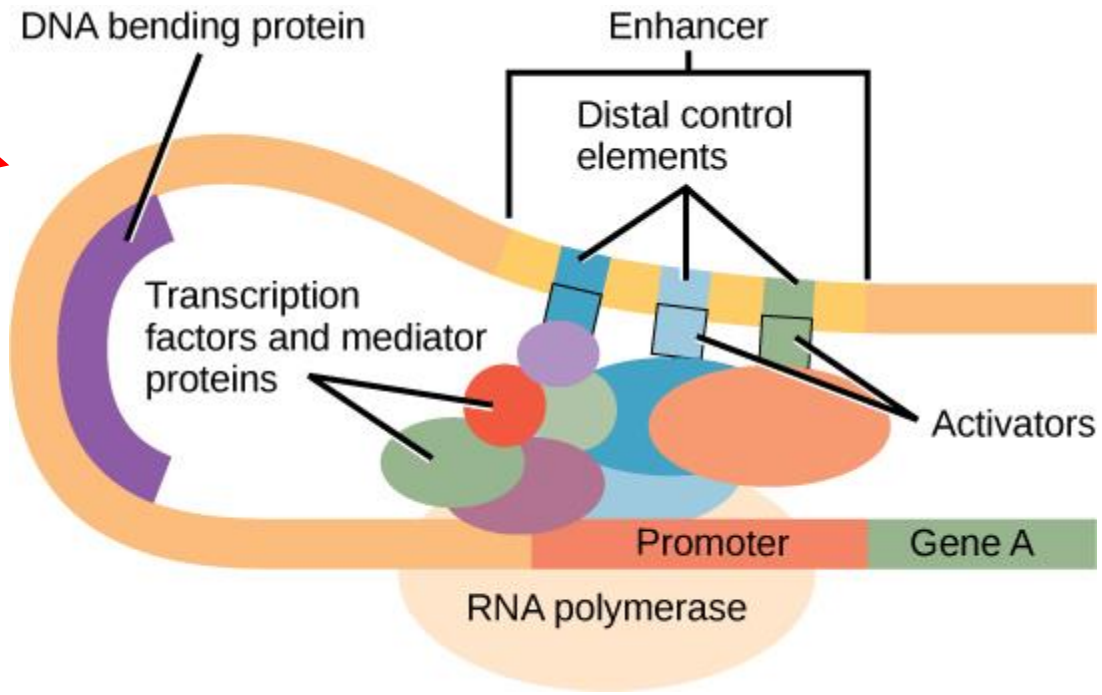


© 2011 Pearson Education, Inc.

Interpreting the genetic program



© 2011 Pearson Education, Inc.



https://courses.lumenlearning.com/wmopen-biology1/chapter/eukaryotic-gene-regulation/

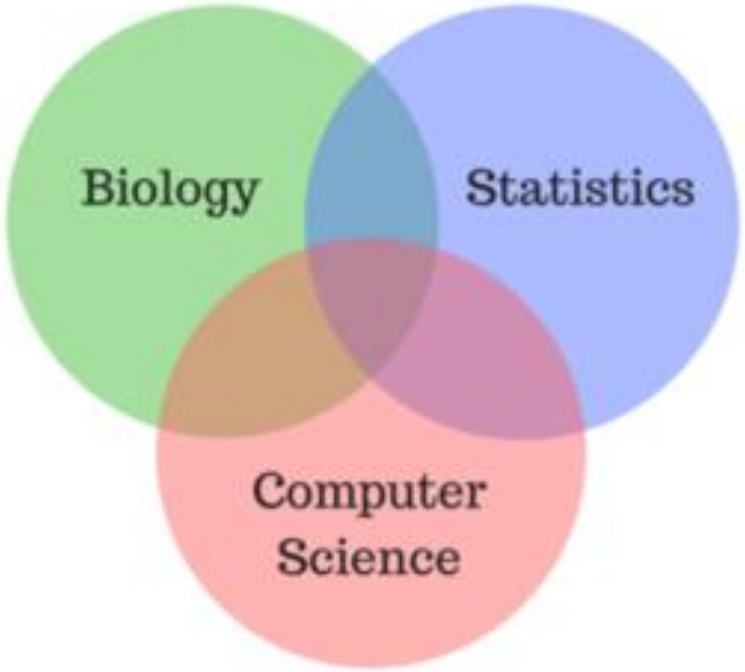
Outline

- Introductions, faculty and students
- The professional histories of the faculty in their words
- “What is Bioinformatics?” and aims scope of this course
- Next generation sequencing and bioinformatics
- Review of biological molecules
- Summary

Summary

- Introductions, faculty and students….
  - Students have varied backgrounds (to be expected)
  - Rozen will provide some students with supplementary reading on the basics of molecular cell biology
- The professional histories of the faculty….
  - Variety of backgrounds – some from bench biology, some from computation
- “What is Bioinformatics?” and aims scope of this course….
  - Several students pointed out that Nattestad’s video seemed to cast bioinformatics in a service role; faculty responded with approaches for establishing intellectual ownership of research questions
- Review of biological molecules….
  - Source for more info (fairly high level) https://courses.lumenlearning.com/wmopen-biology1/
- Next generation sequencing and bioinformatics….

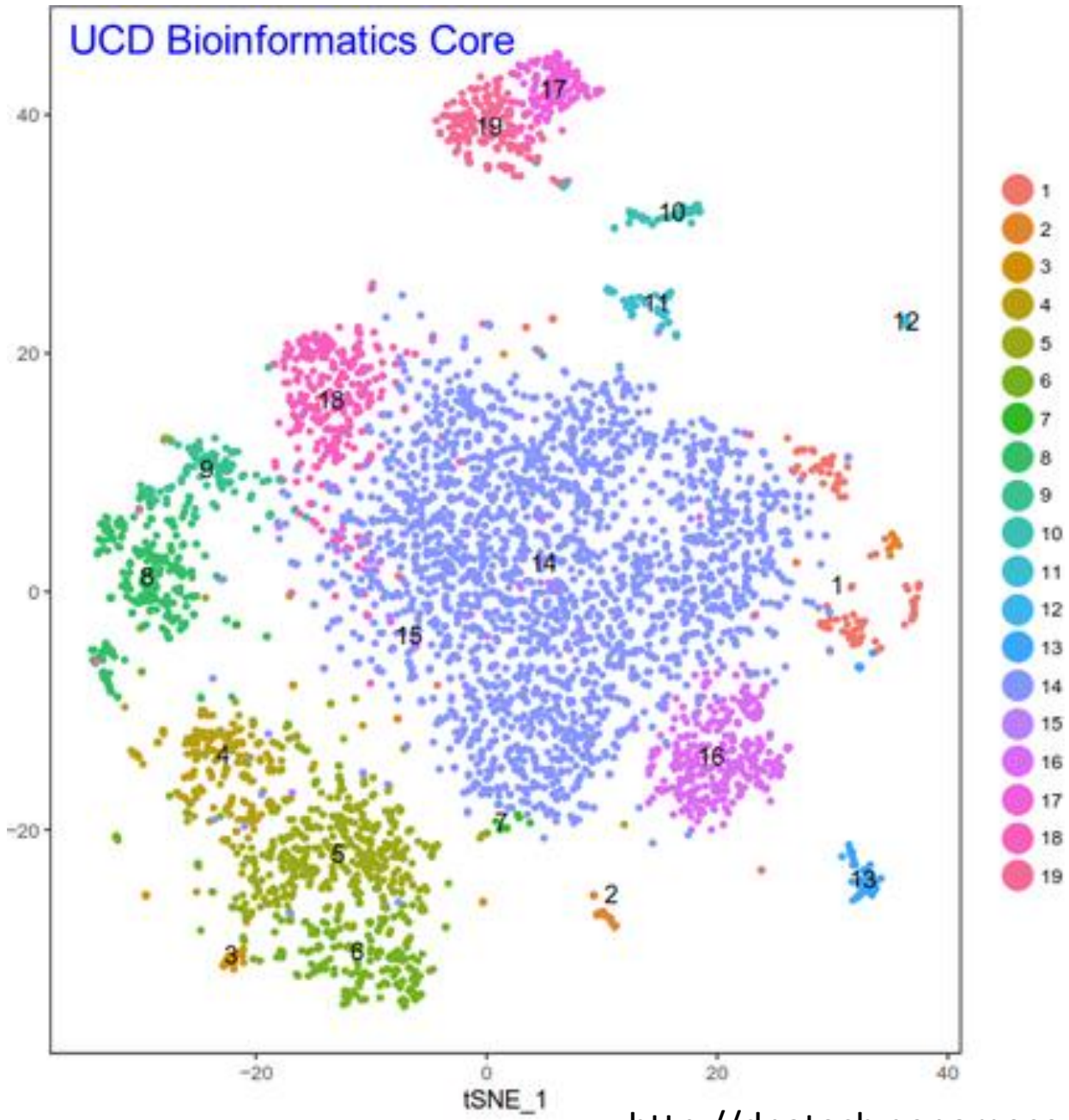
The future of bioinformatics will be driven by data



The future of bioinformatics

- Data generation will continue to improve rapidly
- Will generate more and new kinds of data
  - Example Oxford Nanopore: aiming for $1,000 human genome based on long reads
  - Better information on the small molecules and proteins in cells
- More medical applications – first in severe genetic diseases and selection of therapies for cancer
  - Both already realities, but will expand
- Personalized medicine based on genetics and biomarkers

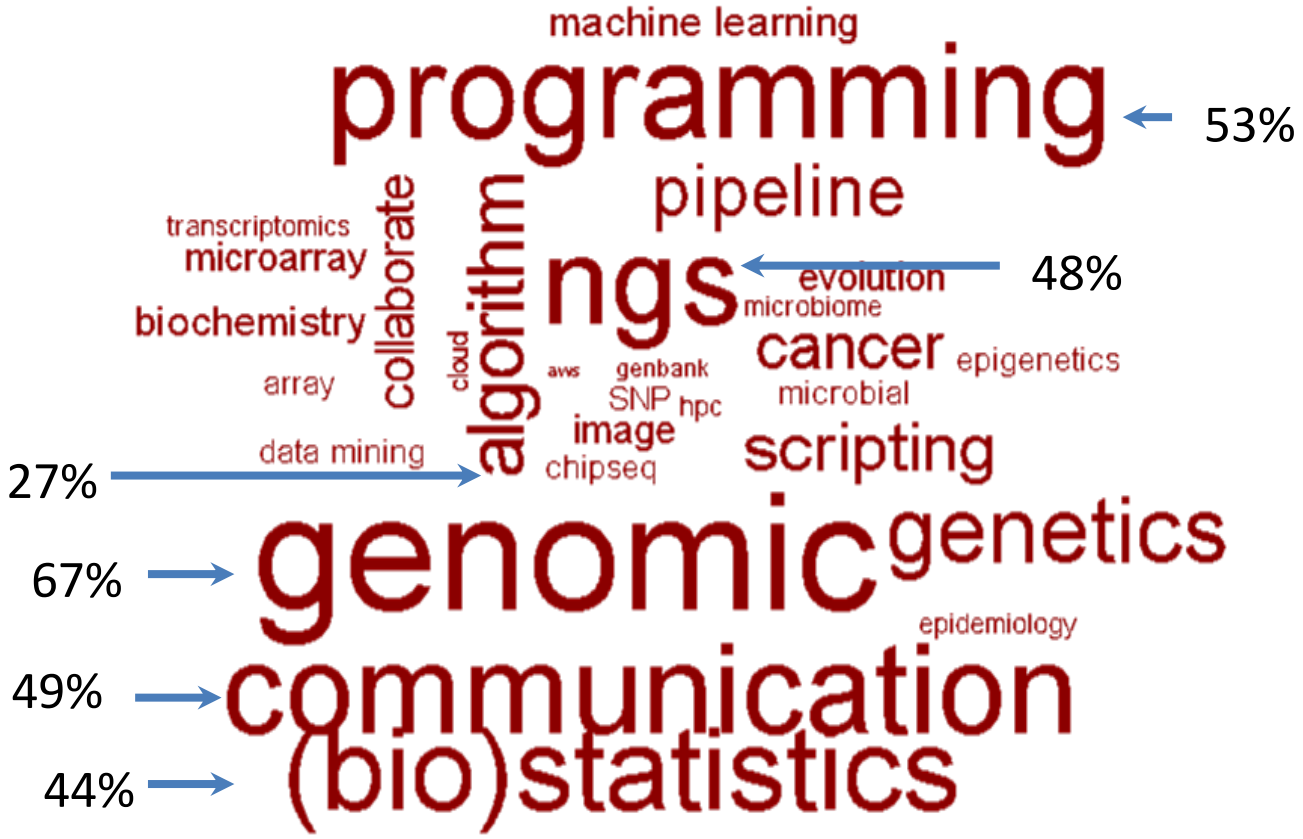
Hot technology: Single cell RNA-seq



- Tissues contain multiple cell types
- Even cancer tumours have multiple cell types in addition to the cancer cells
- The RNA populations of individual cells can be measured, then the cells can be clustered based on their RNA population profiles
- We can usually identify which cluster corresponds to which cell type based presence of RNAs from a few characteristic genes

http://dnatech.genomecenter.ucdavis.edu/single-cell-analyses/

Genome bioinformatics is important
439 bioinformatics.org job ads



(Now quite old)