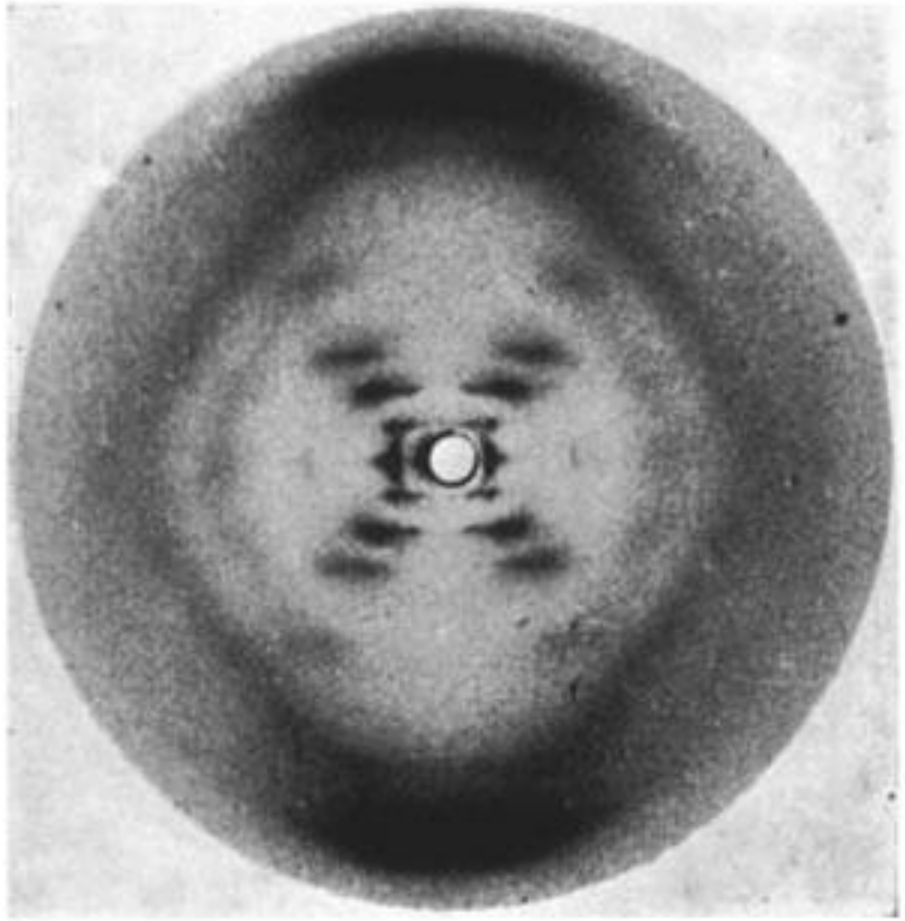
Lecture 0

Class overview & genome basics



GENE/QBS 146

QBS146/GENE146

What? Why?

- Do intro’s here…

QBS146/GENE146

What? Why?

- The goal of the coarse is to cover key methods and approaches in computational biology and bioinformatics.

- “**Computational biology** is the science that answers the question “How can we learn and use models of biological systems constructed from experimental measurements?” ….This field is sometimes referred to as **bioinformatics**, but many scientists use the latter term to describe the field that answers the question “How can I efficiently store, annotate, search and compare information from biological measurements and observations?” - *Robert Murphy, CMU*

- The goal of the coarse is:
  - to cover key methods, including sequencing search, dynamic programming, clustering, and basic graphical models (HMM).
  - Apply those methods to datasets
  - Incorporate modern tools from computational biology into a final project
  - Learn basic methods and try to implement them in Python as a bridge to your future projects

QBS146/GENE146

Goals of the class

- Revamp of existing course this year, with the goals of:
  - Cohesive class structure, not a series of lectures
  - Balance core concepts with practical experiences (2:1 or so)
  - A good book with topical material that enables self-learning
  - Topics that are on-point for computational biology, but are useful for a wider audience (clustering, computational modeling, learning some Python)
- Teaching goals:
  - Put more active learning into the class
  - Focus on practical skills over rote memorization (projects not tests)
  - Take suggestions from *Ungrading*; seek student feedback on class and their grades

QBS146/GENE146

Grades

- How are we going to determine grades:
  - Weekly quiz and write-up (**10% of final grade**): Each week, you'll do a short write-up on the material, answer questions you have, and suggestions for class discussion and improvements, due **11:59 PM on the night before Tuesday's class**.
  - Homework assignments (**40% of final grade**): There will be homework assignments that ask you to implement key algorithms on simulated and real datasets. These will be more substantial than the coding puzzles, focusing on the methods and less on learning to program. Resubmissions will be allowed to improve your grade; see below.
  - Final project (**40% of final grade**): You'll work to use many of the skills learned in class towards a final project. We'll have presentations for the last class periods. More details will be available in the first class.
  - Participation (**10% of the final grade**): We expect you to participate in the class and online discussions, including the peer review of the final projects.
  - We will review final grades with each student at the end of the term; you'll submit your own assessment of your work in the class including the final project, and we'll discuss your final grade in the final write-up.

QBS146/GENE146

Honor code

Our goal for this class is to advance your personal knowledge and skill set in computational biology and bioinformatics.You need to do the assignments yourself: all class work should be done individually unless explicitly noted. If you discuss the general ideas for an assignment with another student please note this in your submission (and detail who you discussed it with), as we will cross-check assignments (both manually and using automated systems). We realize In the era of ChatGPT and StackOverflow that no one works in a vacuum but please liberally document where you got help. **In the end, It is far better to get lower scores on an assignment than to cheat (we also allow resubmissions to raise lower scores).** Given this, we will be aggressive in enforcing the honor code, which is available at the link below. The honor code states, "the faculty may reserve the right to fail the student for the exercise, the course, or both". The instructors and the TAs are available for help, so please seek us out with questions about assignments or the honor code before engaging other students.

QBS146/GENE146

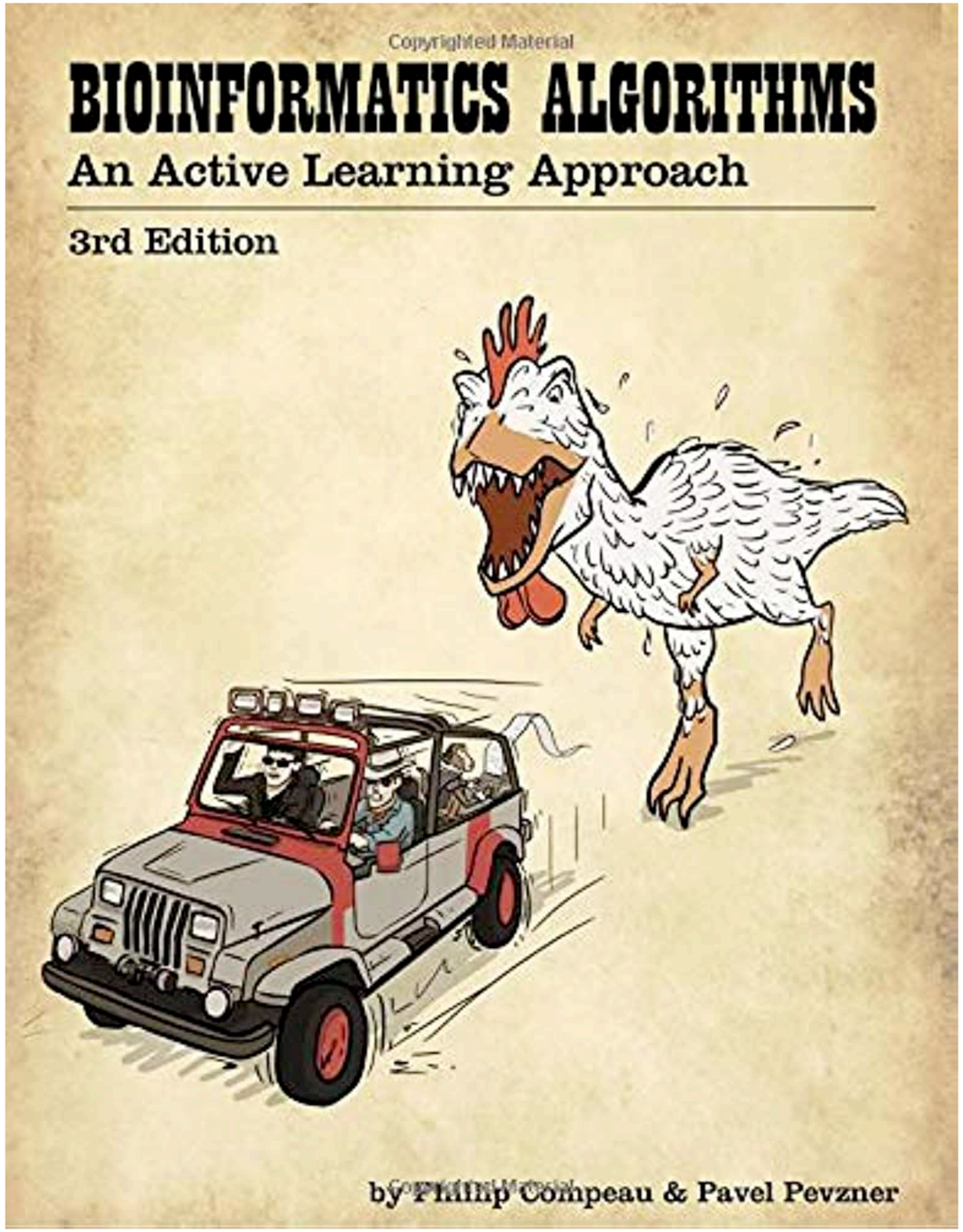
Structure of the class

- Global structure:
  - Canvas and Slack
  - Tuesday lecture and review of homework issues, etc
  - Thursday in-class group exercise and paper dissection
  - Homeworks due before Tuesday’s class
  - Self-assessments and mini-quiz for Thursday classes
  - End-of-term project

QBS146/GENE146

Structure of the class

- Book
  - *Bioinformatics Algorithms: An Active Learning Approach*, by Phillip Compeau & Pavel Pevzner, 3rd Edition (https://bioinformaticsalgorithms.org).



QBS146/GENE146

Structure of the class

- Programming language:
  - We'll use Python
  - Tough choices here: R, Python, command-line, Pipelining with Nextflow or SnakeMake
  - Touch base if you want to try another language (much less support but possible)
  - Jupyter notebooks for assignments

Schedule (draft)

Month	Day	Room	Class #	Content
March	28	-	1	I'm out of town, no class
	30	Kellogg 100	2	Overview of the class, 'the semi-flipped experiment', final project discussions. python notebooks, genomic structure and information content, genome assembly
April	4	Kellogg 100	3	chapter 1: finding enriched sequences, motif finding, kmers, regulatory sequences
	6	Kellogg 100	4	chapter 2: In-class entropy and hidden messages
	11	Kellogg 100	5	chapter 2: probabilistic motif finding, gibbs sampling
	13	**Kellogg 200**	6	chapter 2: In-class motif 'thought experiment'
	18	Kellogg 100	7	chapter 3: How do we assemble genomes?
	20	**Kellogg 200**	8	chapter 3: In class assembly exercise
	25	Kellogg 100	9	chapter 5: Aligning two sequences: Dynamic programming
	27	**Kellogg 200**	10	chapter 5: In-class walking around NYC exercises
May	2	Kellogg 100	11	chapter 8/9: RNA sequencing, read mapping, counting, and enrichment
	4	**Kellogg 200**	12	chapter 8/9: In-class RNA sequencing experimental design exercise
	9	**Chilcott**	13	chapter 8: clustering: RNA to identity
	11	**Chilcott**	14	chapter 8: In-class exploring k-means
	16	**Kellogg 200**	15	chapter 10: probabilistic modeling of hidden states using HMMs
	18	**Vail 120 Auditorium**	16	chapter 10: In-class HMM and CpG island exercise
	23	**Vail 120 Auditorium**	17	chapter 10: HMMs wrap-up, extensions to more complex models
	25	**Chilcott**	18	final project presentations
	30	**Kellogg 200**	19	final project presentations
June	1	Kellogg 100	20	final project presentations (if needed)

Final project

- Final project (**40% of final grade**): You'll work to use many of the skills learned in class towards a final project. We'll have presentations for the last class periods. More details will be available in the first class.

- Class project write-up on Canvas: https://canvas.dartmouth.edu/courses/57831/pages/class-project

- 2-4 students

- First you can attempt to implement an algorithm, or part of an algorithm, to solve a biological problem. To show you've implemented this algorithm, you'll need to apply to at least a toy example of a real-world problem you aim to address.

- The second option is to apply an existing tool to a real-world dataset. Here you'll focus more on the application details of getting the algorithm to work, and comparing the results to both what was shown in the paper as well as compared to what previous groups have done in this space. This could also be analysis on your own dataset, though see the stipulations a few paragraphs down.

- Stipulations: You can use any of your ongoing 'real-world' (lab) projects for this portion of the class, but you must discuss this with your own PI first, from which we will require sign-off and verification. (More on Canvas)

Final project

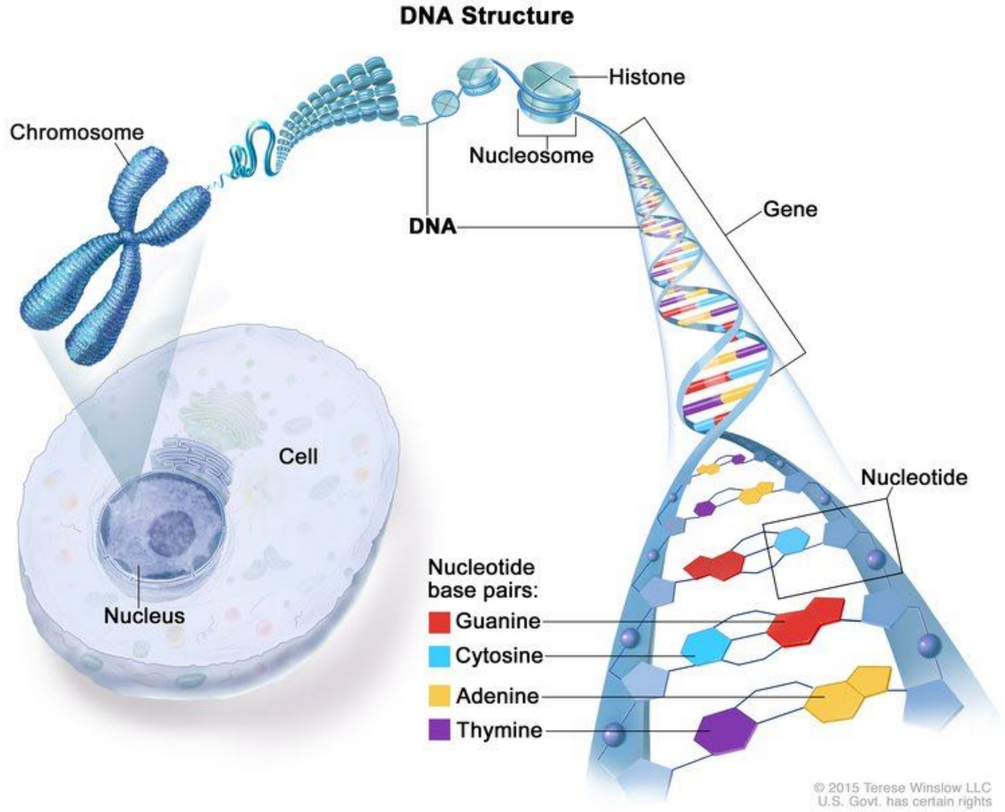
Project deliverables:

- Deliverable 1: Group formation and group contract **- April 6th**
- Deliverable 2: Project Proposal (10%) **- April 20th**
- Deliverable 2: Project Progress Report (10%) **- May 11th**
- Deliverable 3: Project Presentation (25%) **- Starting May 25th**
- Deliverable 4: Final Essay and Explanation of Work (55%) **- June 1st**

Genetics

Our past and present (and a main focus of this class)

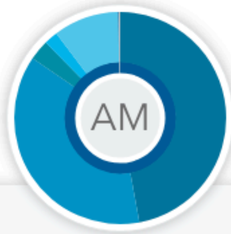
Broad Institute



NIH

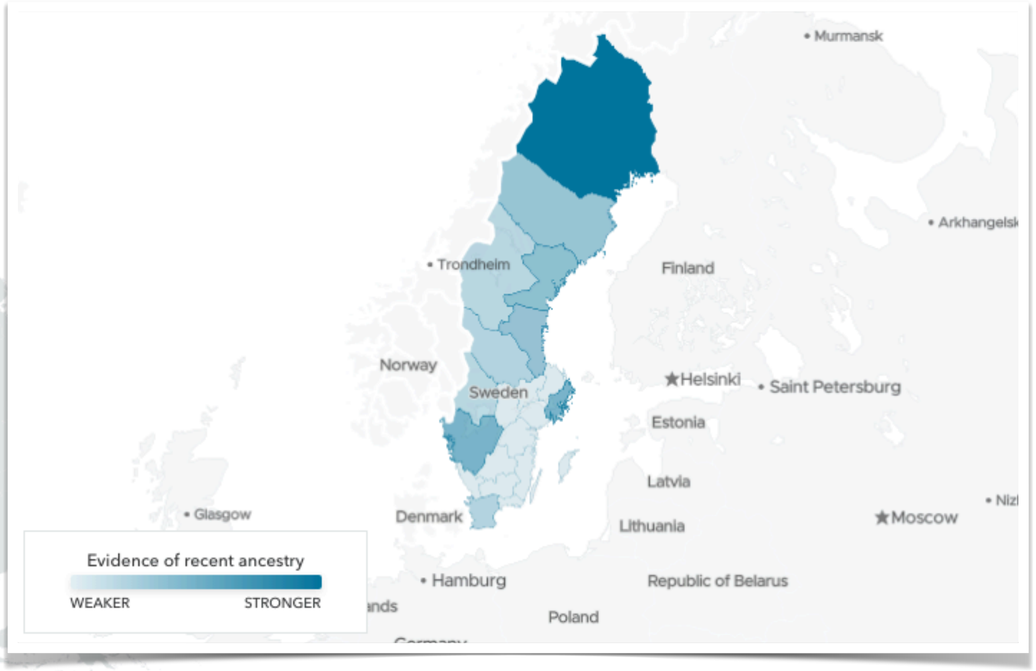
Increasing resolution

From just spitting in a tube + science



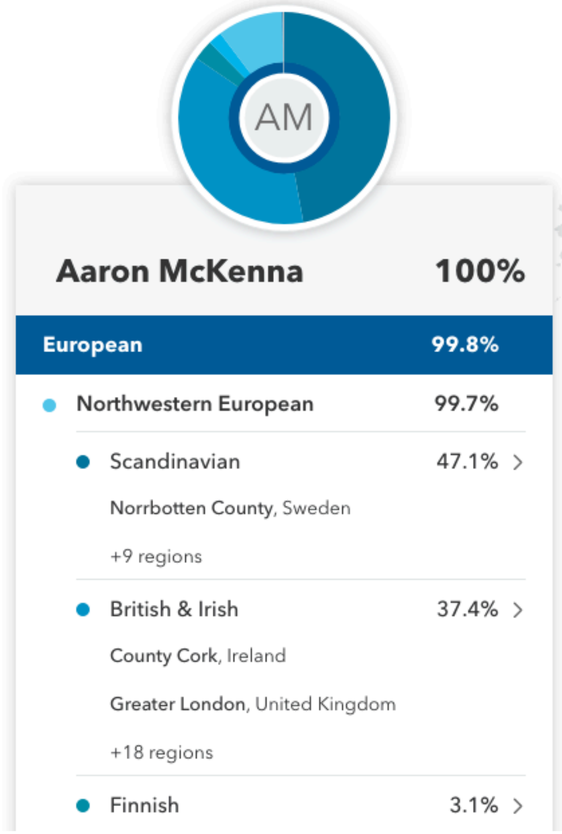
Aaron McKenna **100%**

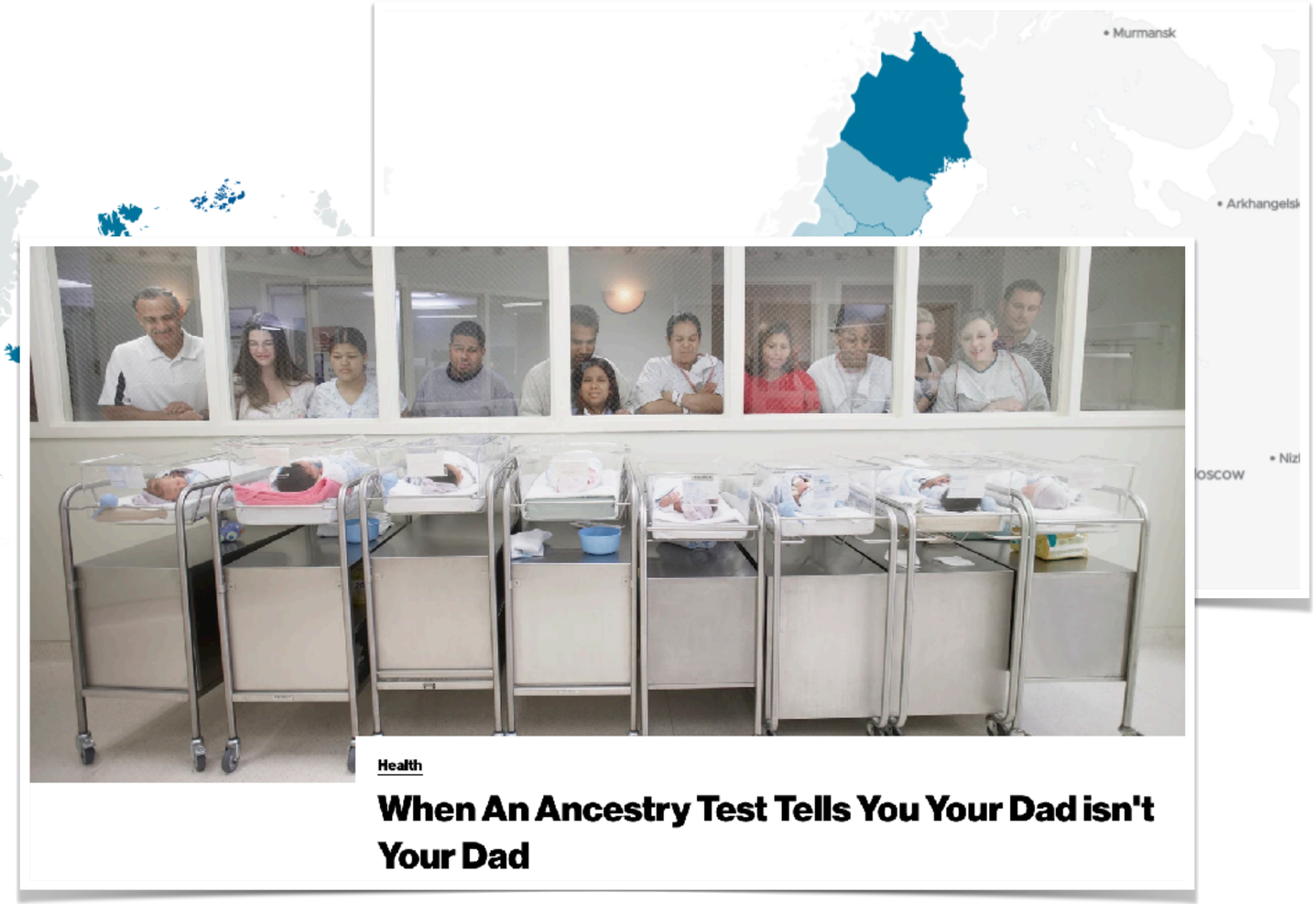
European	**99.8%**
Northwestern European	99.7%
Scandinavian	47.1% >
Norrbotten County, Sweden	
+9 regions	
British & Irish	37.4% >
County Cork, Ireland	
Greater London, United Kingdom	
+18 regions	
Finnish	3.1% >



Increasing resolution

With implications



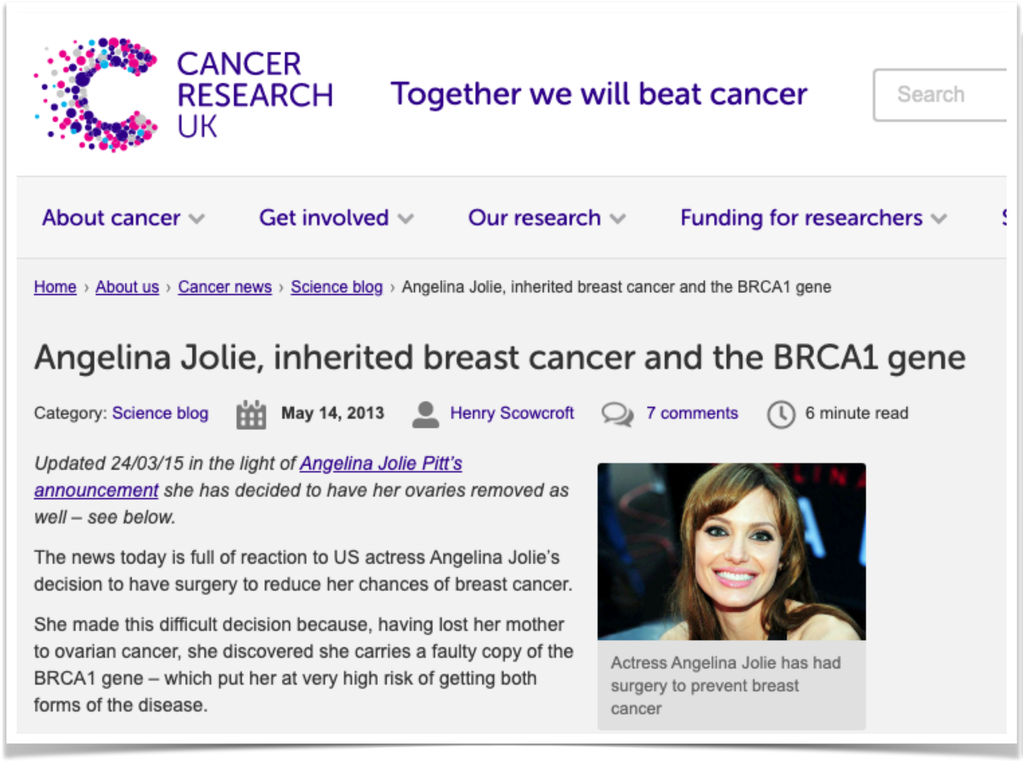


Vice

Rates estimated at about 1%

Increasing resolution

To both family and health

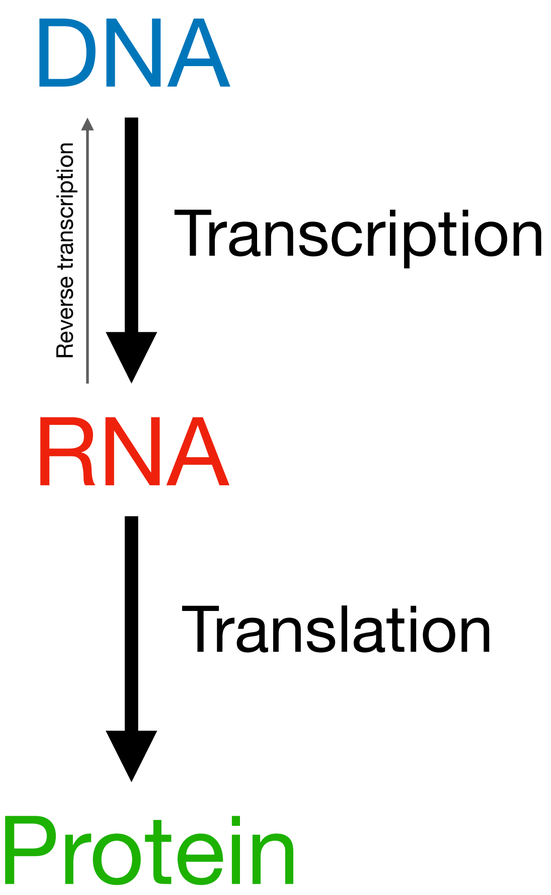


cancer research uk

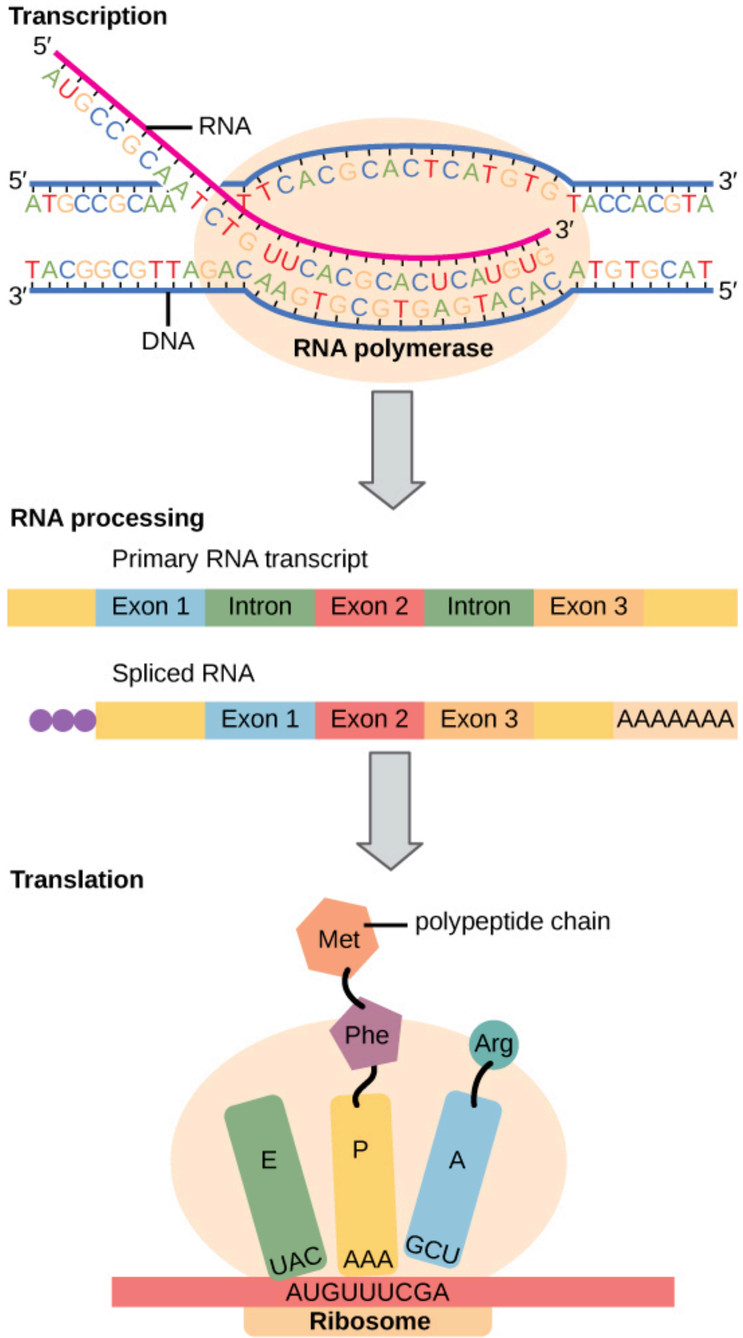
Vice

Central Dogma

Storage to action



- Generally a one-way flow of information, from genome to the correct amount of proteins, the functional units of life
- DNA specifies what and where, RNA regulates how much, and protein does the work
- RNA can be reverse-transcribed into DNA, but done only by viruses, protein can't directly influence what encoded in DNA



Proteins are the product

We’re interested how genetics dictates their function

DNA

Reverse transcription

Transcription

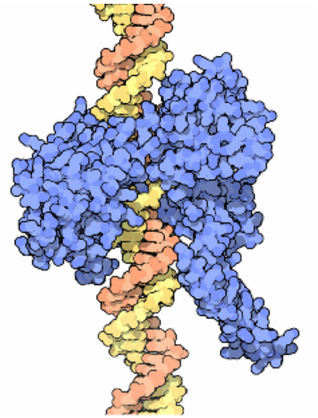
RNA

Translation

Protein

Protein(s):

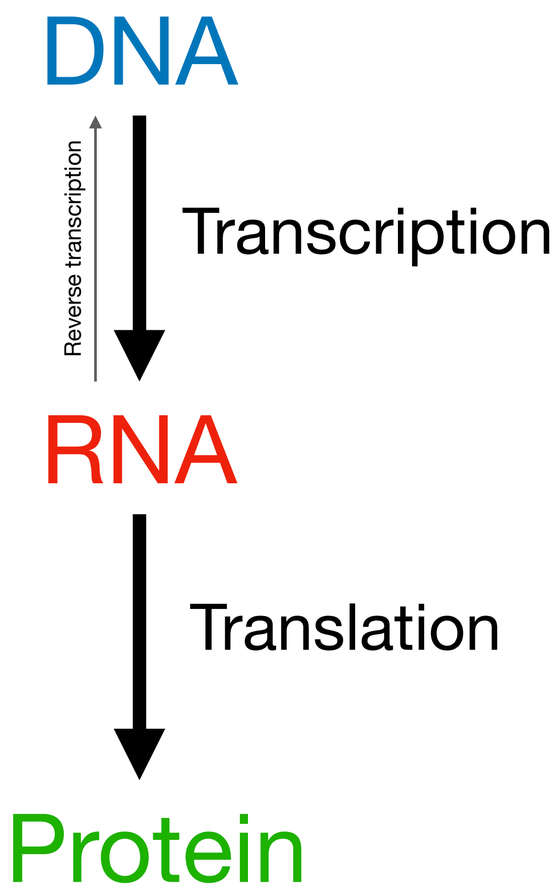
- Linked string of amino acids, folded into the correct shape by the hydrophobicity, synthesis dynamics, thermodynamic properties, and many other factors
- Translated from an RNA message, not reverse translated into RNA or DNA
- Almost everything that you're made of is protein, including the machinery to maintain and propagate DNA.



Topoisomerase, PDB http://pdb101.rcsb.org/motm/73

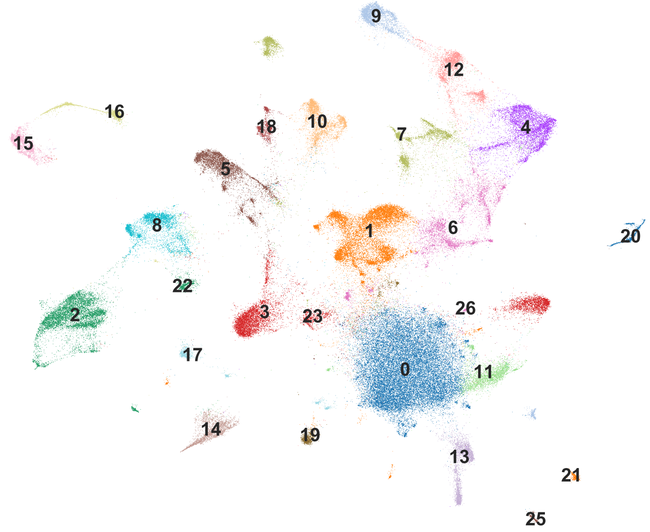
RNA as the messenger

A temporary message and a proxy of state



RNA:

- In human cells most often serves as a message (mRNA) from the genome (DNA) with instructions to build a protein, possibly many times
- The amount and lifespan of RNA is regulated at many levels, is generally single-stranded, and relatively unstable



Single-cell RNA-seq from 100K mouse cells

From DNA to proteins

DNA serves as the long-term storage medium in the central dogma

DNA

Reverse transcription

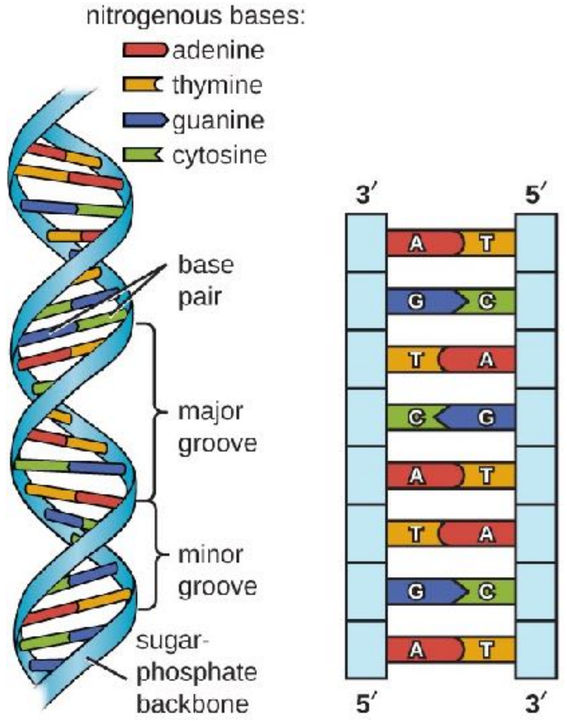
Transcription

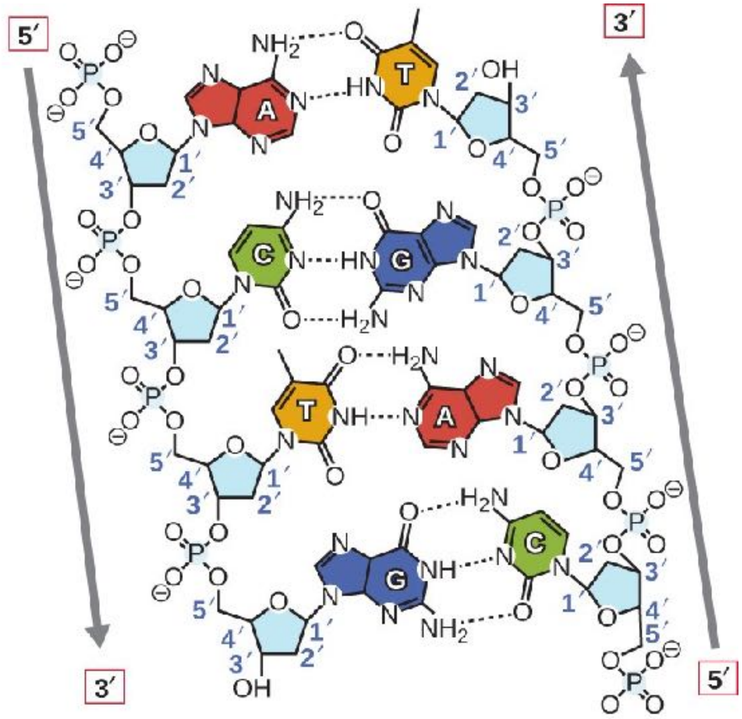
RNA

Translation

Protein

DNA:

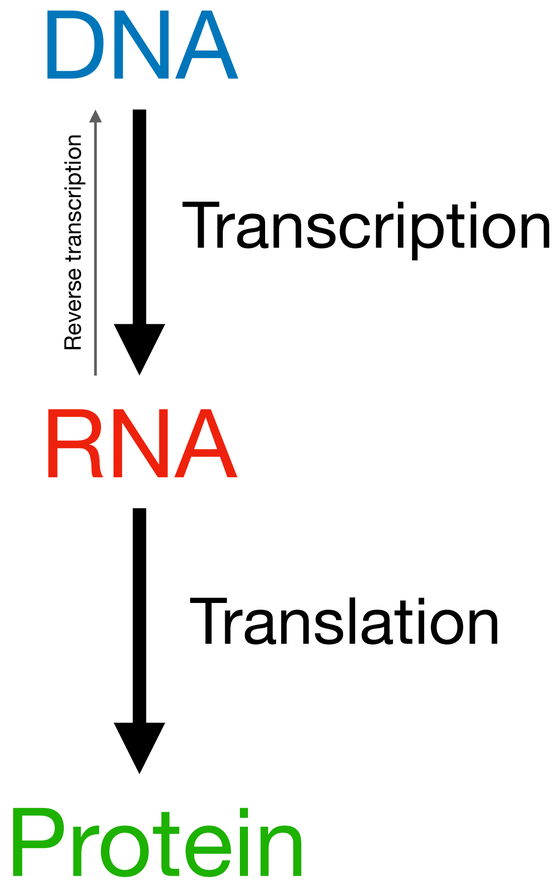




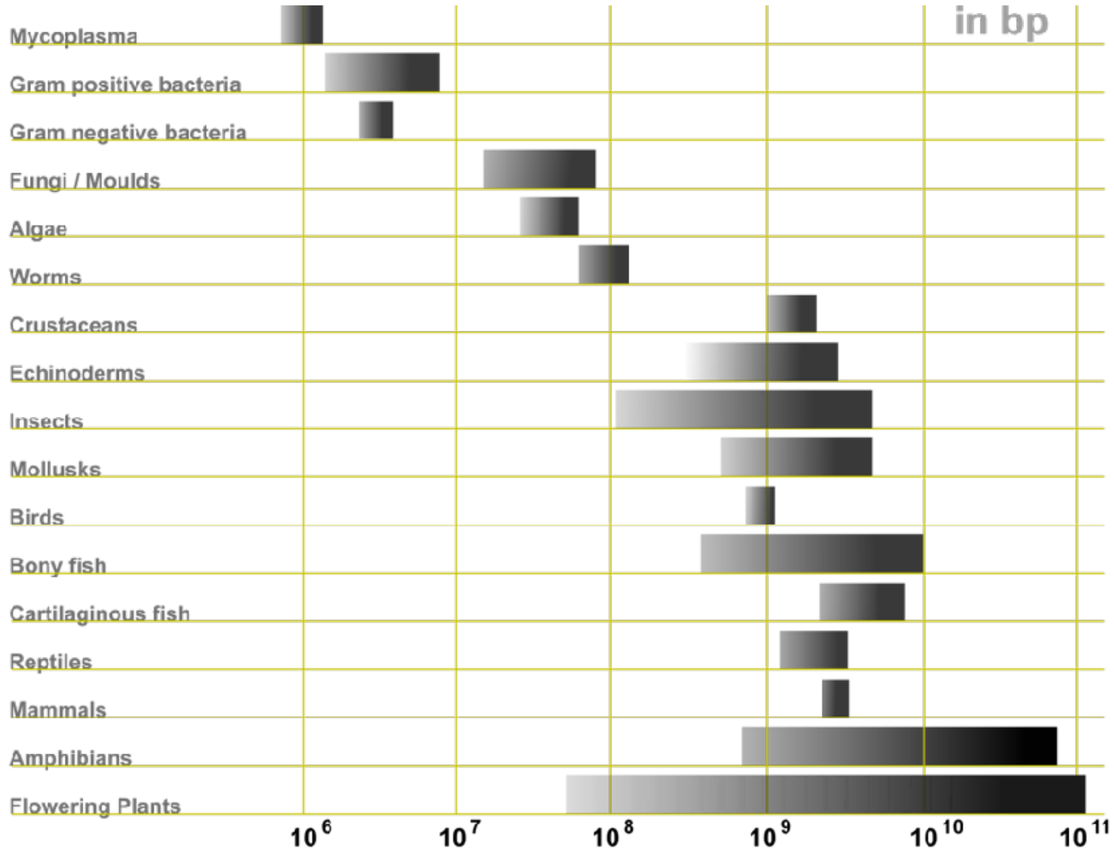
microbenotes.com

From DNA to proteins

DNA serves as the long-term storage medium in the central dogma



DNA:



Wikipedia

From DNA to proteins

DNA serves as the long-term storage medium in the central dogma

DNA

Reverse transcription

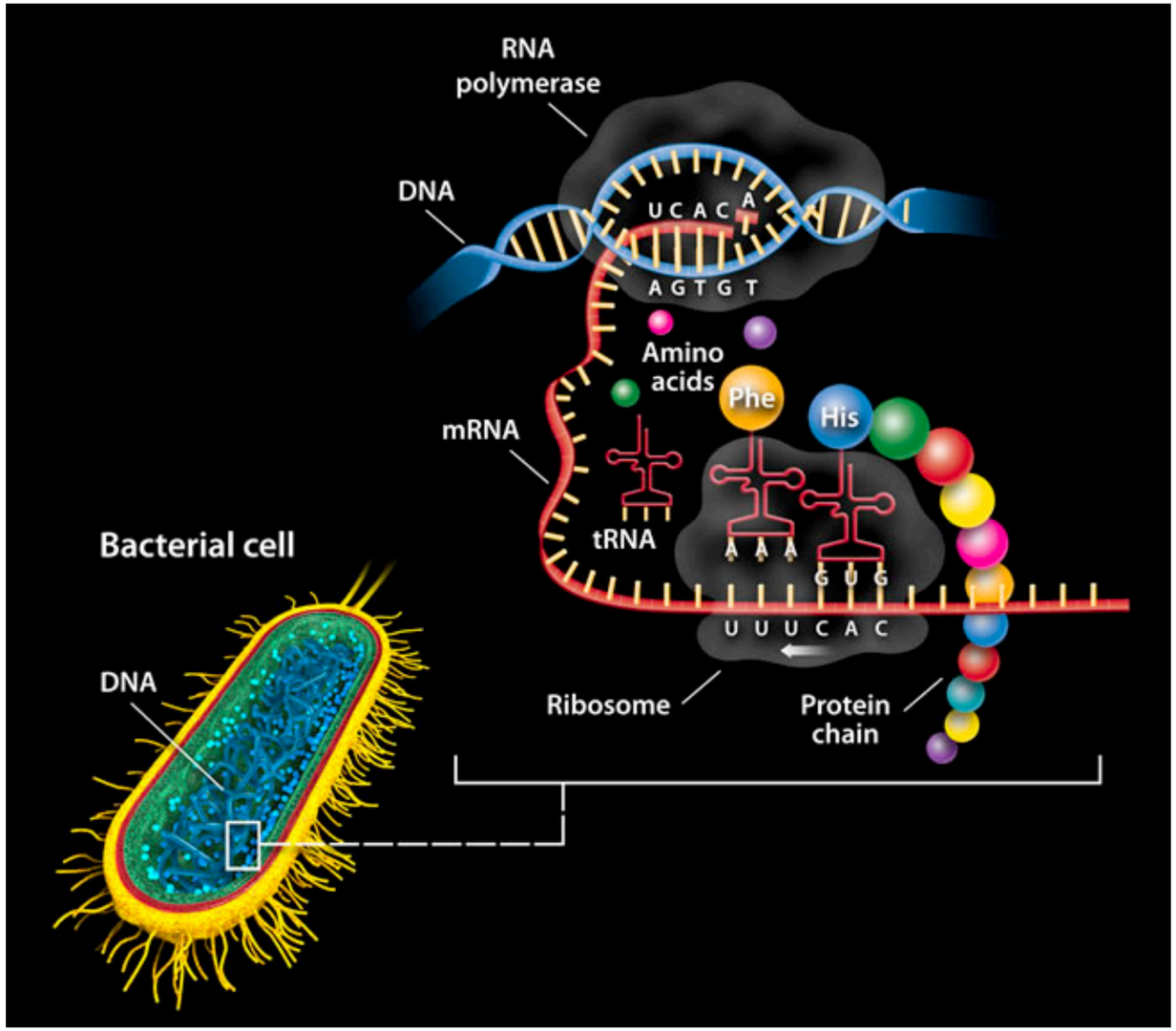
Transcription

RNA

Translation

Protein

DNA:



ornl.gov

From DNA to proteins

DNA serves as the long-term storage medium in the central dogma

DNA

Reverse transcription

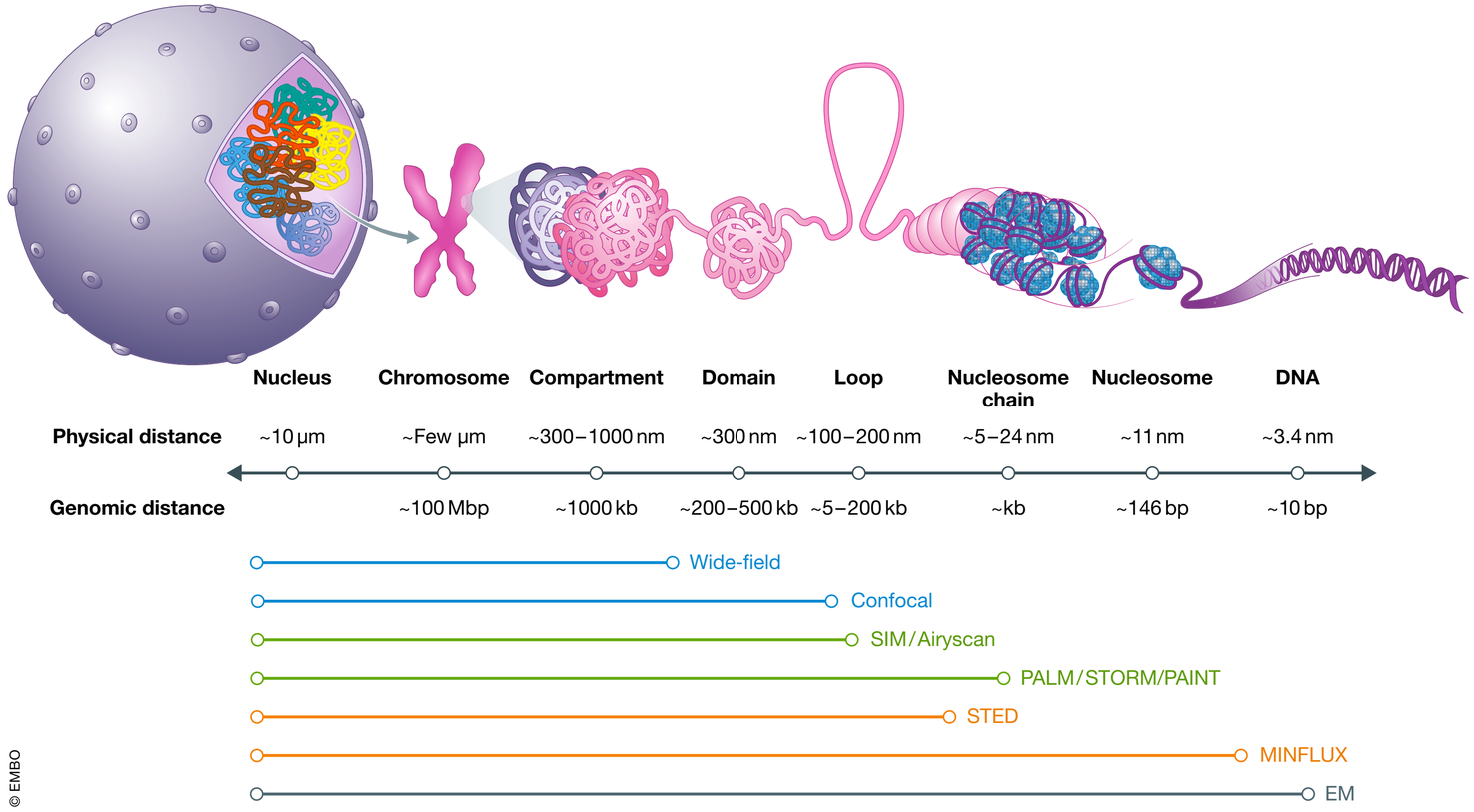
Transcription

RNA

Translation

Protein

DNA:

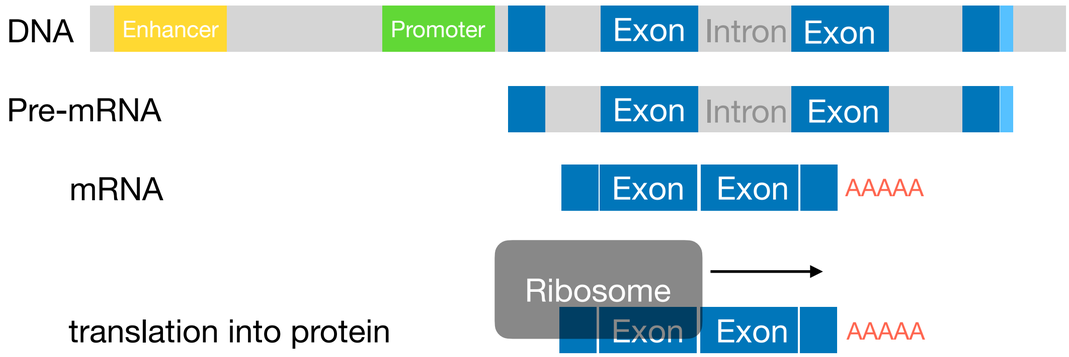


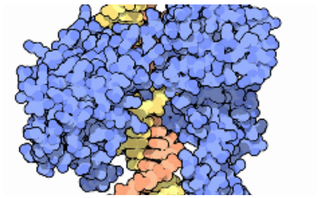
Liu et al, 2021

Encoding proteins in DNA

Genes have a clear structure

DNA:





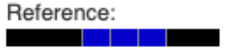
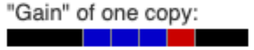
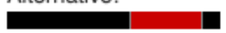
- Contains all of the information to generate every protein, by way of RNA transcription, in the right cell at the right time, in either development or homeostasis
- Redundant and stable, serving primary means of information transfer between generations
- Given the above, when we thing about human history and how this is recorded over time we turn to DNA.
- Nomenclature: We describe changes to DNA in two ways, structurally what the change is *OR* the suspected impact that the change has. This can be confusing.

Describing the type of change

Sequence variants

Type	Description	Example (Reference / Alternative)	
SNP	Single Nucleotide Polymorphism	Ref: ...TTG**A**CGTA...	Alt: ...TTG**G**CGTA...
Insertion	Insertion of one or several nucleotides	Ref: ...TTGACGTA...	Alt: ...TTGA**TG**CGTA...
Deletion	Deletion of one or several nucleotides	Ref: ...TTG**AC**GTA...	Alt: ...TTGGTA...
Indel	An insertion and a deletion, affecting 2 or more nucleotides	Ref: ...TTG**A**CGTA...	Alt: ...TTG**GCT**CGTA...
Substitution	A sequence alteration where the length of the change in the variant is the same as that of the reference.	Ref: ...TTG**AC**GTA...	Alt: ...TTG**TA**GTA...

Structural variants

Type	Description	Example (Reference / Alternative)	
CNV	Copy Number Variation: increases or decreases the copy number of a given region	Reference: 	"Gain" of one copy:  "Loss" of one copy: 
Inversion	A continuous nucleotide sequence is inverted in the same position	Reference: 	Alternative: 
Translocation	A region of nucleotide sequence that has translocated to a new position	Reference: 	Alternative: 

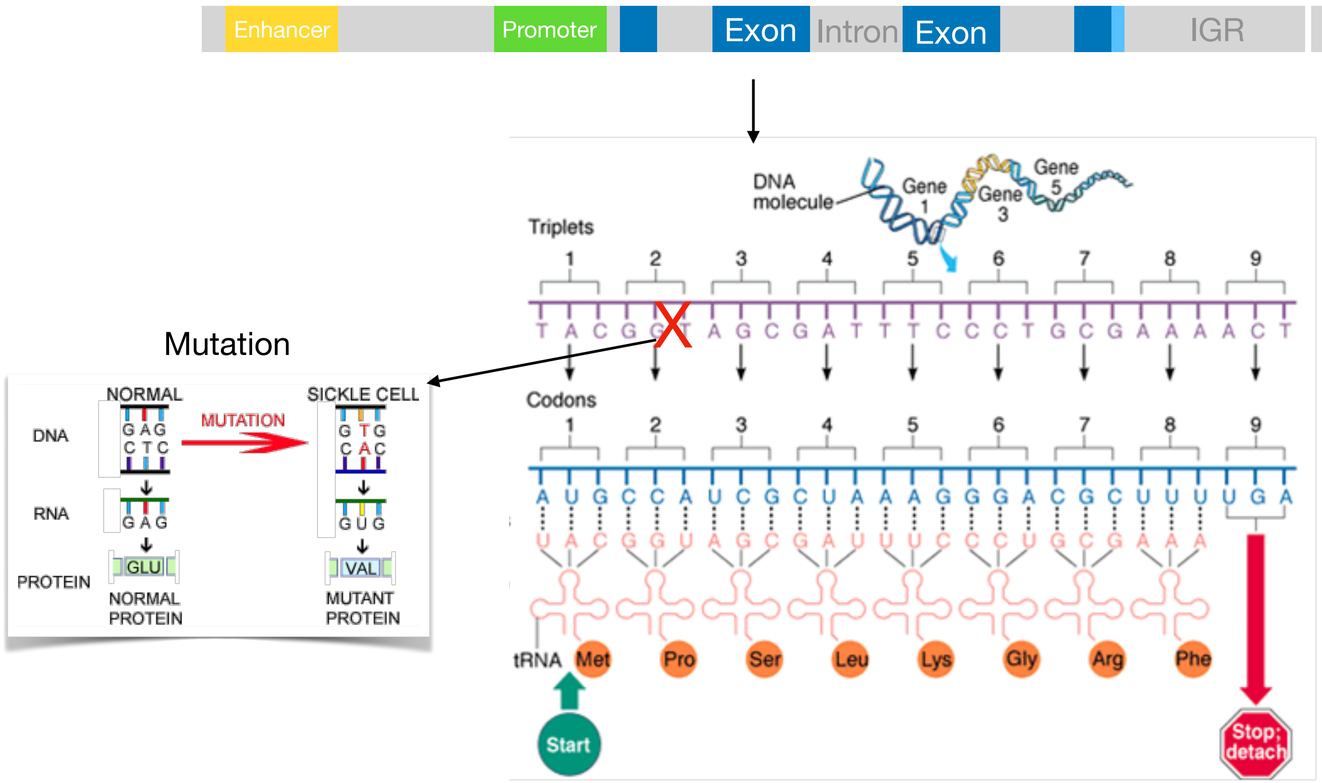
Ensembl

- Focused in what this change does to the underlying sequence, no matter it's location in the genome
- Two overarching classes, small and large
- The breaking point between sequence variants and structural variants is very loosely around 100 bases, but up to who's describing it
- The more complicated structural variants, such as complex inversions and duplications of regions have been harder to detect
- Polymorphism vs mutation vs variant

Changes to coding regions (genes)

Gene structure and amino acid encoding predicts mutation effect

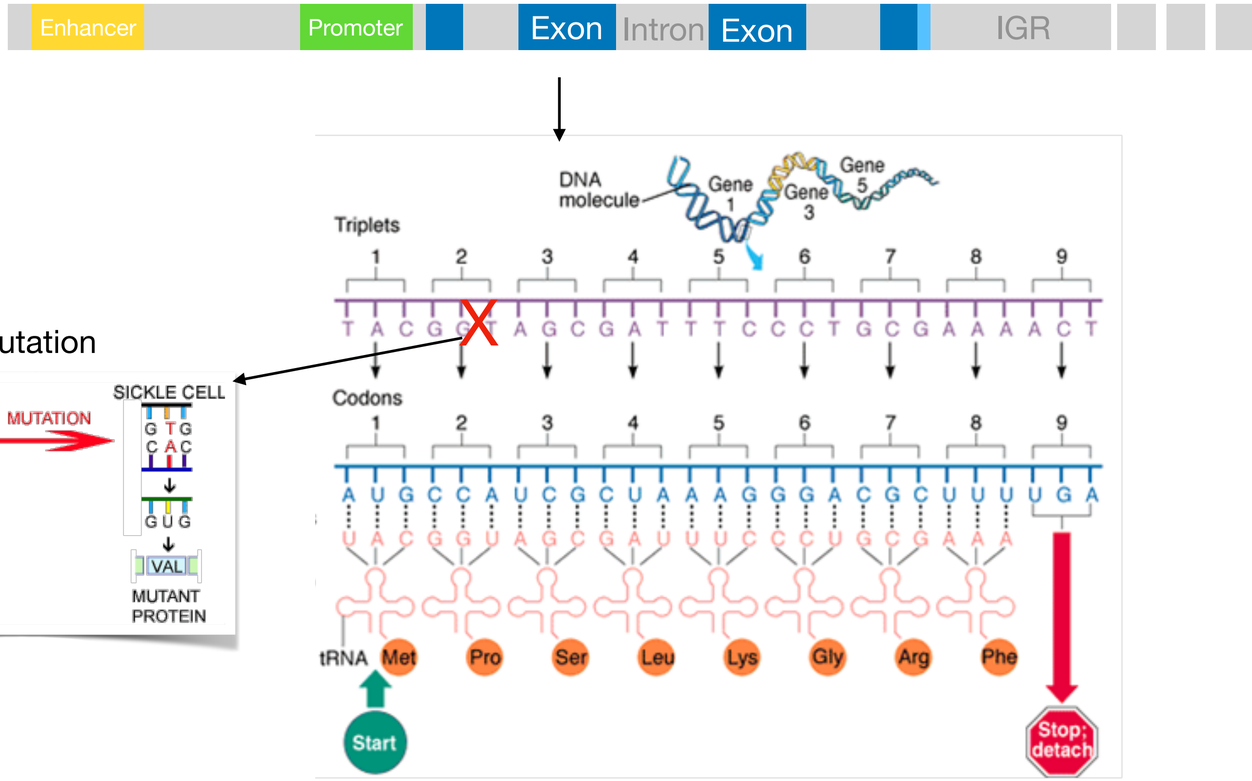
- **Silent** - a change in the nucleotide sequence that does not cause a change in the amino acid sequence
- **Missense** - a change in the nucleotide sequence that results in the amino acid sequence being altered
- **Nonsense** - changes in the nucleotide sequence that cause a premature stop codon to be introduced leading to a truncated protein
- **Elongating** (used less often) - changes that disrupt a stop codon and elongate the protein



Changes to coding regions (genes)

A lot of nomenclature around describing coding changes

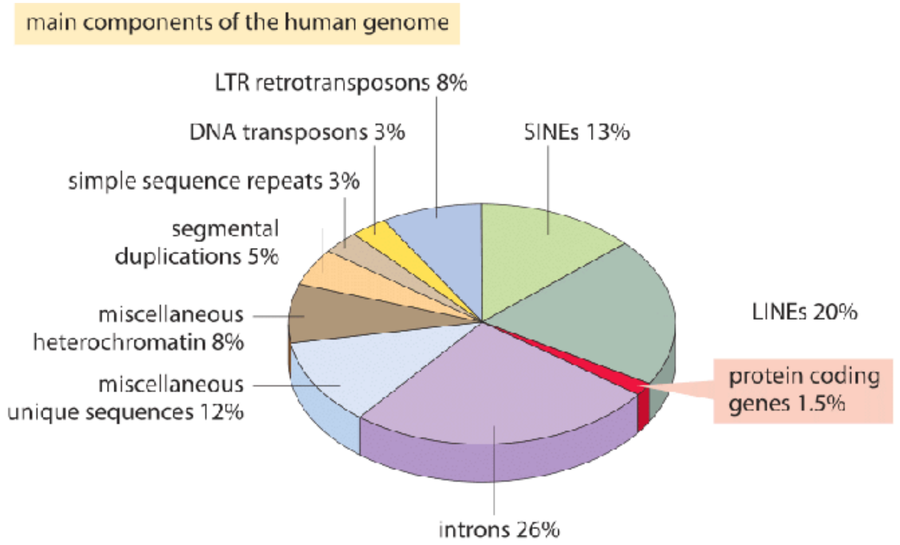
- Nuclear substitutions
  - g.21538G>A in F2/Prothrombin (c.-97G>A)
  - c.621+1G>T or IVS5+1G>T
- Amino Acid substitutions
  - p.R117H or p.Arg117His in CFTR
  - p.E7V (historically p.E6V) in β-globin gene
- Deletions and insertions
  - p.F508del (also known as ΔF508)
  - c.1521_1523delCTT
  - g.409_410insC
- Mitochondrial substitutions
  - m.3460G>A in ND1 (LHON)
  - m.8993T>G in ATP6 (NARP)



Regulatory changes and the rest

Sequences that dictate when genes are expressed, some known motifs

- Intron are removed (spliced) from mRNA, and we know a bit about the signatures needed, and what mutations do
- Our understanding is more loose after that. We know some of the DNA ‘motifs’ regulatory proteins bind to elsewhere in the genome
- but how much a small change matters is less clear
- Our understanding of promoters is slightly better than enhancers
- A large part of the remaining genome is inactivated viruses or retrotransposons
- There are even many super-conserved sequences, with unknown origin, that are being worked out

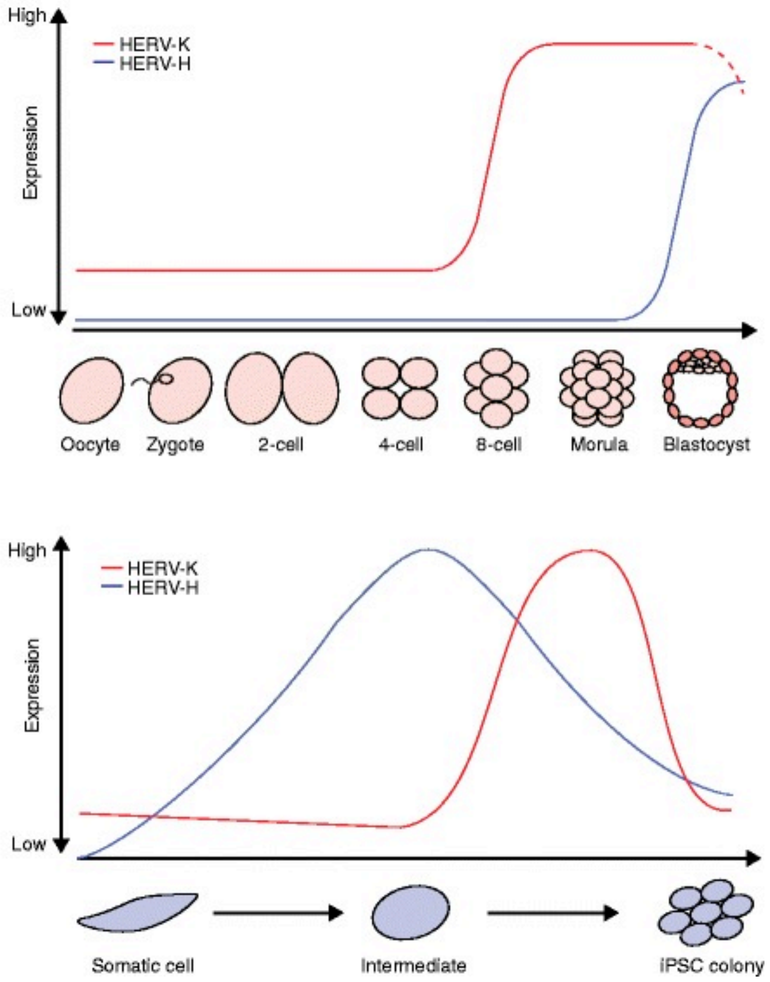


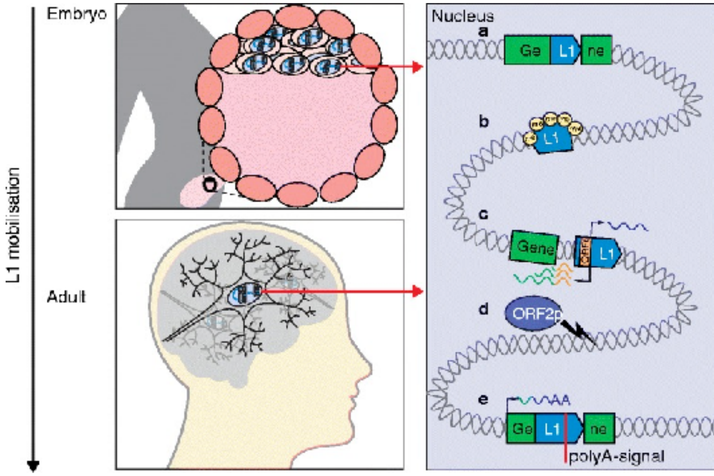
http://book.bionumbers.org/ (sorry for the 3D pie chart)

Regulatory changes and the rest

An aside, the numerous remnants of viruses are fascinating…

- Intron are removed (spliced) from mRNA, and we know a bit about the signatures needed, and what mutations do
- Our understanding is more loose after that. We know some of the DNA ‘motifs’ regulatory proteins bind to elsewhere in the genome
- but how much a small change matters is less clear
- Our understanding of promoters is slightly better than enhancers
- **A large part of the remaining genome is inactivated viruses or retrotransposons**
- There are even many super-conserved sequences, with unknown origin, that are being worked out



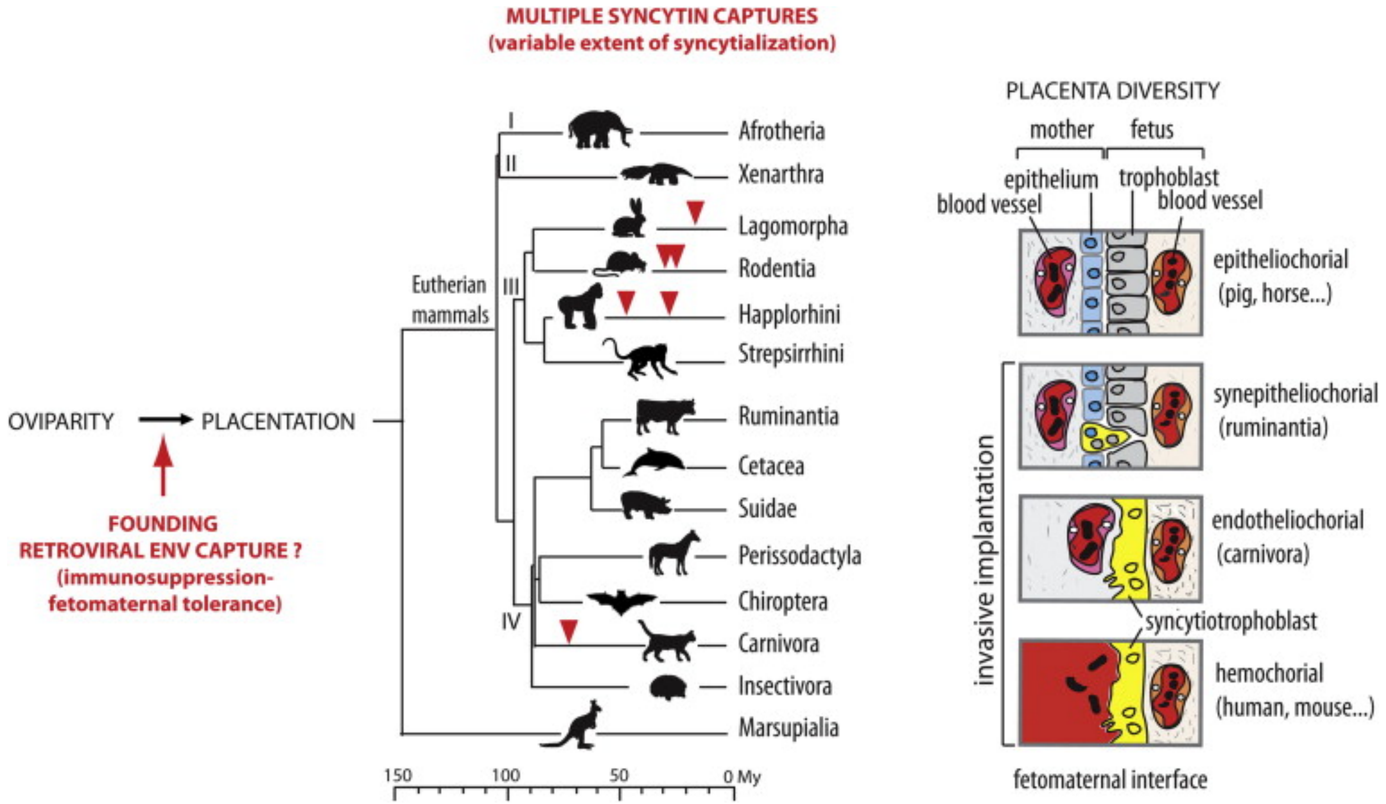


Gerdes, Genome Biology, 2016

Regulatory changes and the rest

An aside, the numerous remnants of viruses are fascinating…

- Intron are removed (spliced) from mRNA, and we know a bit about the signatures needed, and what mutations do
- Our understanding is more loose after that. We know some of the DNA ‘motifs’ regulatory proteins bind to elsewhere in the genome
- but how much a small change matters is less clear
- Our understanding of promoters is slightly better than enhancers
- **A large part of the remaining genome is <u>inactivated</u> viruses or retrotransposons**
- There are even many super-conserved sequences, with unknown origin, that are being worked out



“We speculate that a founding event in the emergence of placental mammals has been the capture of an ancestral retroviral *env* gene that was instrumental in conferring maternal tolerance toward the embryo through immune escape bestowed by the immunosuppressive domain of the envelope glycoprotein, thus allowing development of a “primitive” placental tissue and close contact between mother and fetus (*left*)”

Dupressoir et al. Placenta 2012

Regulatory changes and the rest

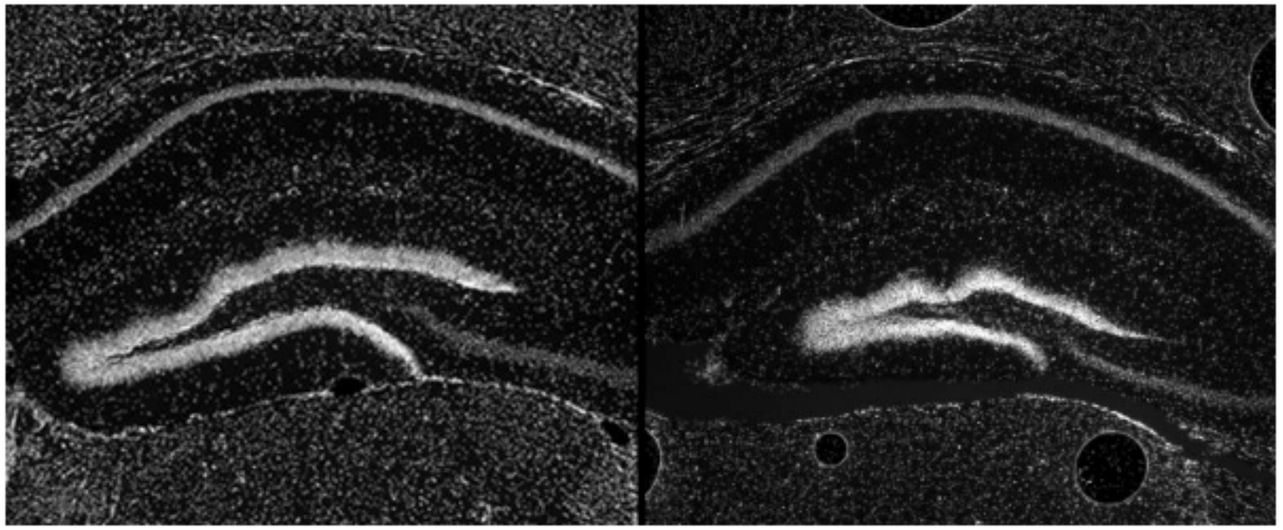
And beyond even viruses

- Intron are removed (spliced) from mRNA, and we know a bit about the signatures needed, and what mutations do
- Our understanding is more loose after that. We know some of the DNA ‘motifs’ regulatory proteins bind to elsewhere in the genome
- but how much a small change matters is less clear
- Our understanding of promoters is slightly better than enhancers
- A large part of the remaining genome is inactivated viruses or retrotransposons
- **There are even many super-conserved sequences, with unknown origin, that are being worked out**

NEWS · 18 JANUARY 2018

‘Dark matter’ DNA influences brain development

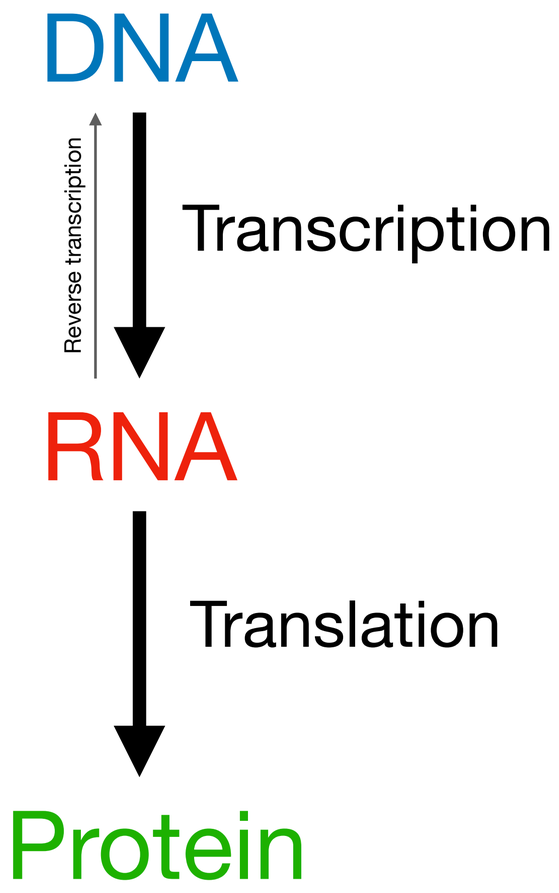
Researchers are finally figuring out the purpose behind some genome sequences that are nearly identical across vertebrates.



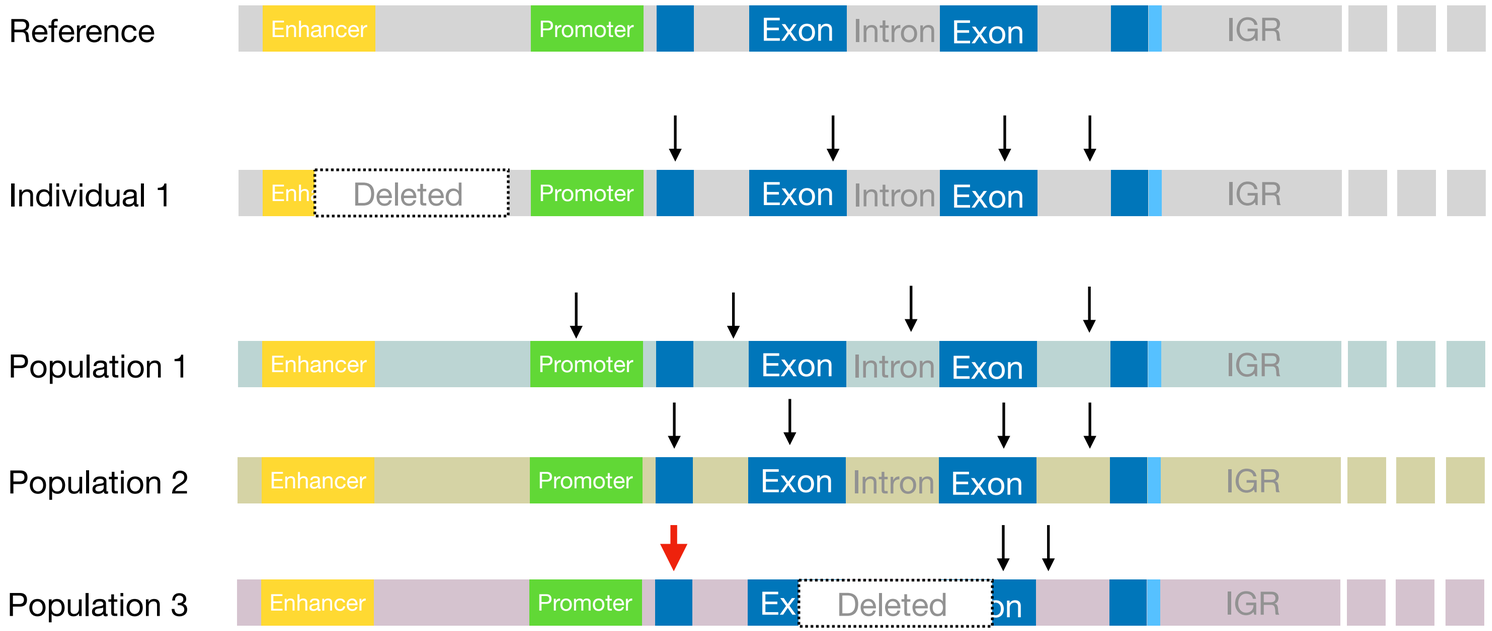
A normal part of the mouse forebrain (left) versus a mutated form (right). Credit: D. Dickel et al., Cell 172, 1-9 Jan. 25, 2018. Elsevier Inc. 2017.

Differences between individuals

We want to know all genetic differences, and what they mean



We were interested in (1) what unique variants individuals have and (2) how these variants differ between individuals

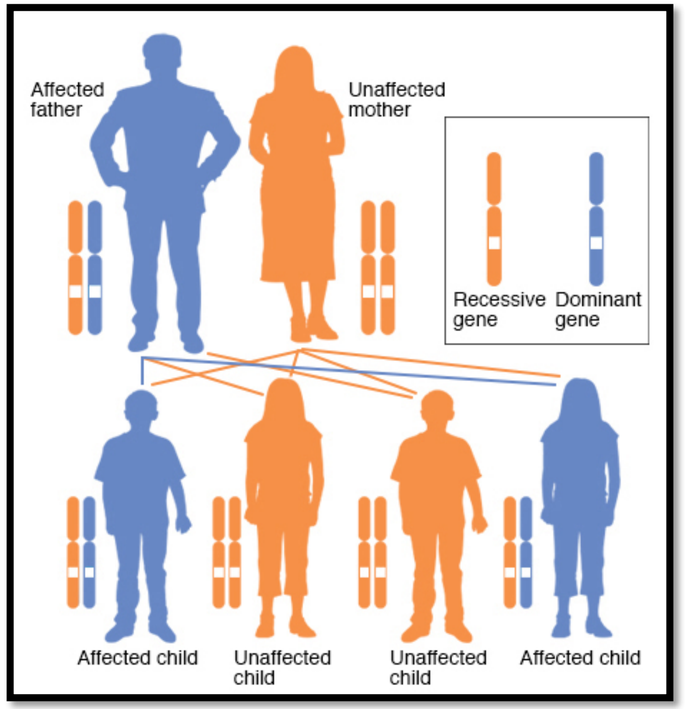


If the red arrow indicates a variant that causes disease, how is it inherited?

Different inheritance pattern

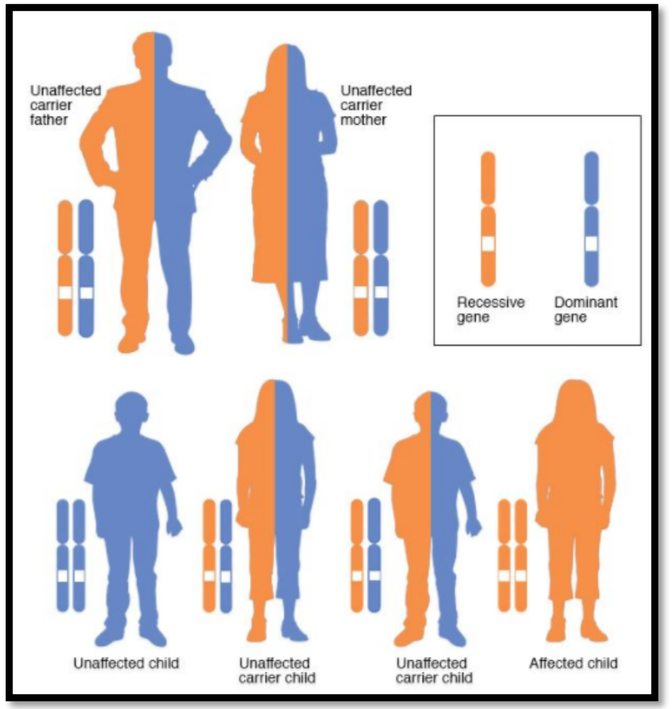
Disease used here instead of traits

Autosomal Dominant Inheritance



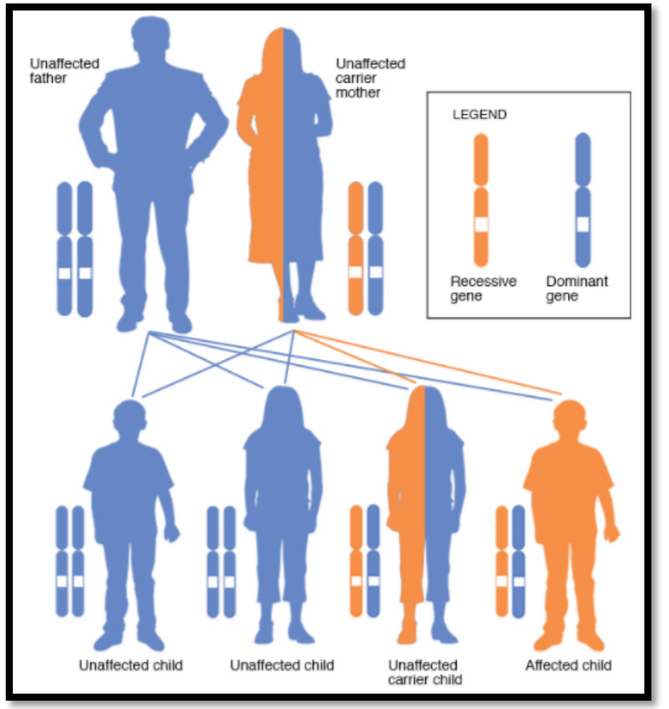
50% Affected, 50% Unaffected
No gender bias

Autosomal Recessive Inheritance



25% Affected, 50% Unaffected Carriers,
25% Unaffected Non-Carriers
No gender bias

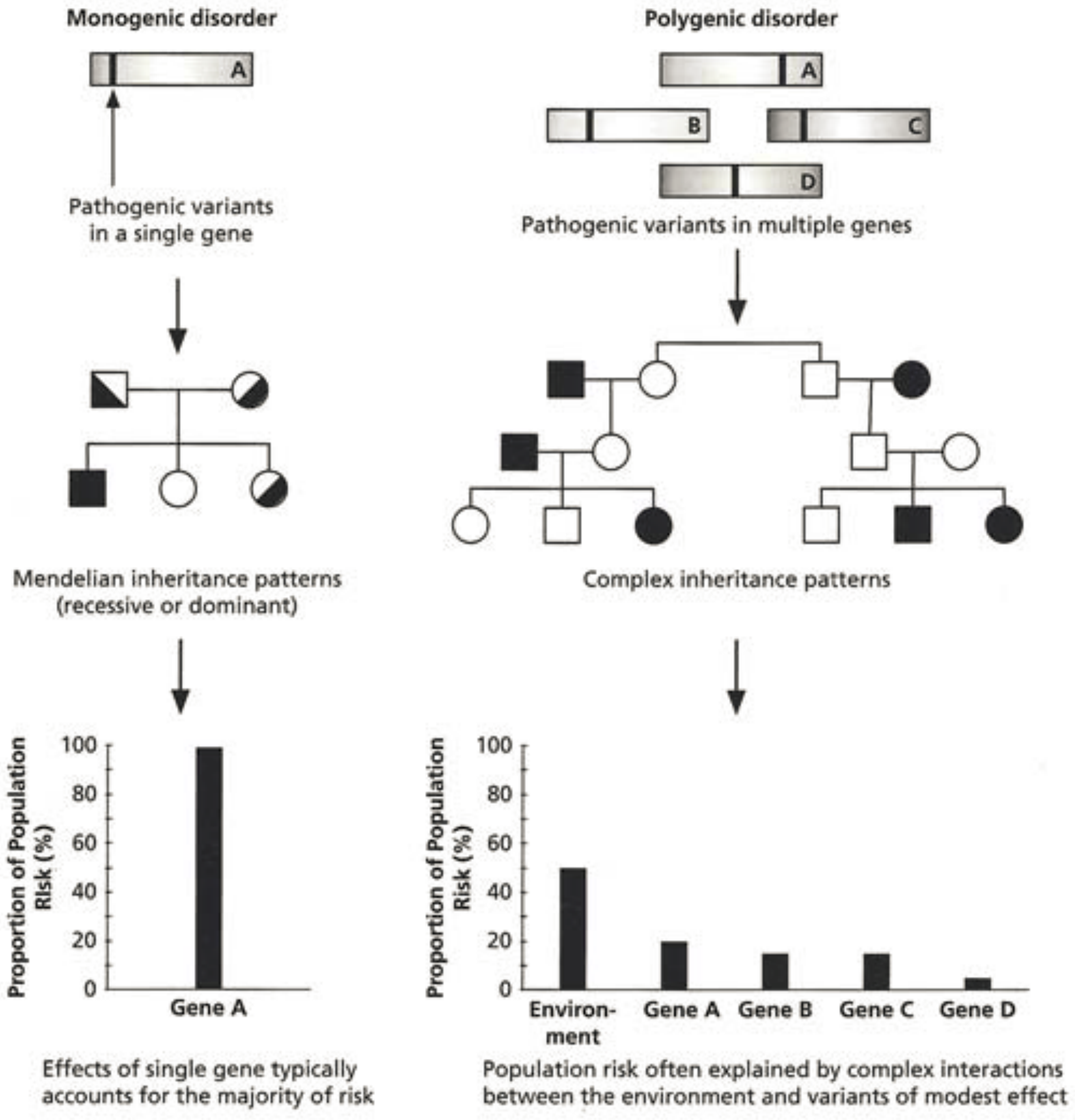
Sex-Linked Inheritance



50% of boys are affected
50% of boys are unaffected
50% of girls are unaffected carriers
50% of girls are unaffected

How many genes are involved

A quick aside into monogenic vs polygenic disorders



How many genes are involved

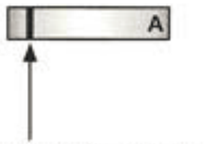
A quick aside into monogenic vs polygenic disorders

HEALTH

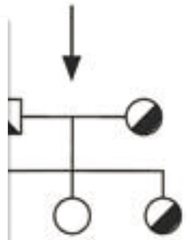
The baby was gravely ill. Rapid sequencing turned around a diagnosis in 13 hours — and pointed to a treatment

By Andrew Joseph July 22, 2021 Reprints

Monogenic disorder



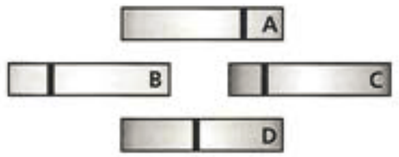
Pathogenic variants in a single gene



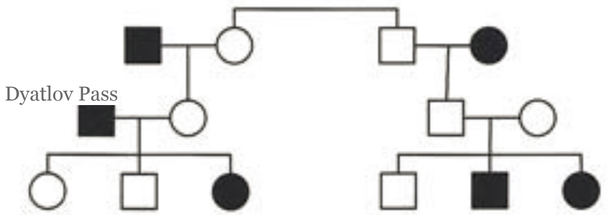
n inheritance patterns
sive or dominant)

Effects of single gene typically accounts for the majority of risk

Polygenic disorder

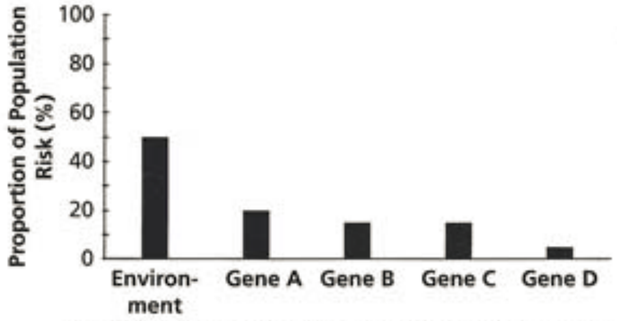


Pathogenic variants in multiple genes



Dyatlov Pass

Complex inheritance patterns



Population risk often explained by complex interactions between the environment and variants of modest effect

Harder

But why computational biology?

Computational Biology

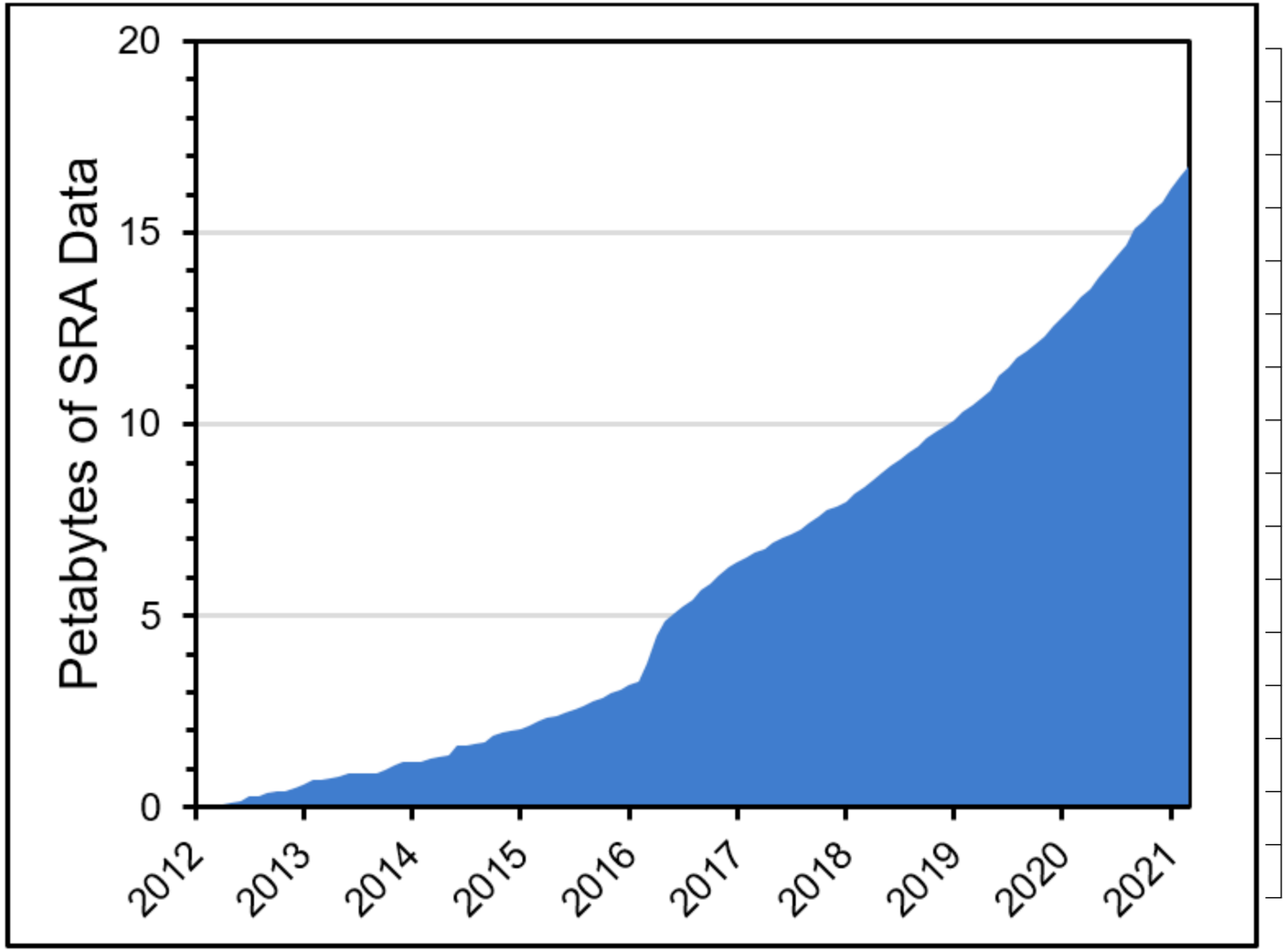
How does computing come in?

Database	Records	Description
Literature		
PubMed	33 027 761	Scientific and medical abstracts/citations
PubMed Central	7 325 415	Full-text journal articles
NLM Catalog	1 629 799	Index of NLM collections
Bookshelf	892 126	Books and reports
MeSH	348 370	Ontology used for PubMed indexing
Genomes		
Nucleotide	476 054 019	DNA and RNA sequences from GenBank and RefSeq
BioSample	19 473 659	Descriptions of biological source materials
SRA	15 919 320	High-throughput DNA/RNA sequence read archive
Taxonomy	2 492 889	Taxonomic classification and nomenclature catalog
Assembly	1 083 900	Genome assembly information
BioProject	536 242	Biological projects providing data to NCBI
Genome	64 815	Genome sequencing projects by organism
BioCollections	8 468	Museum, herbaria, and biorepository collections

Computational Biology

How does computing come in?

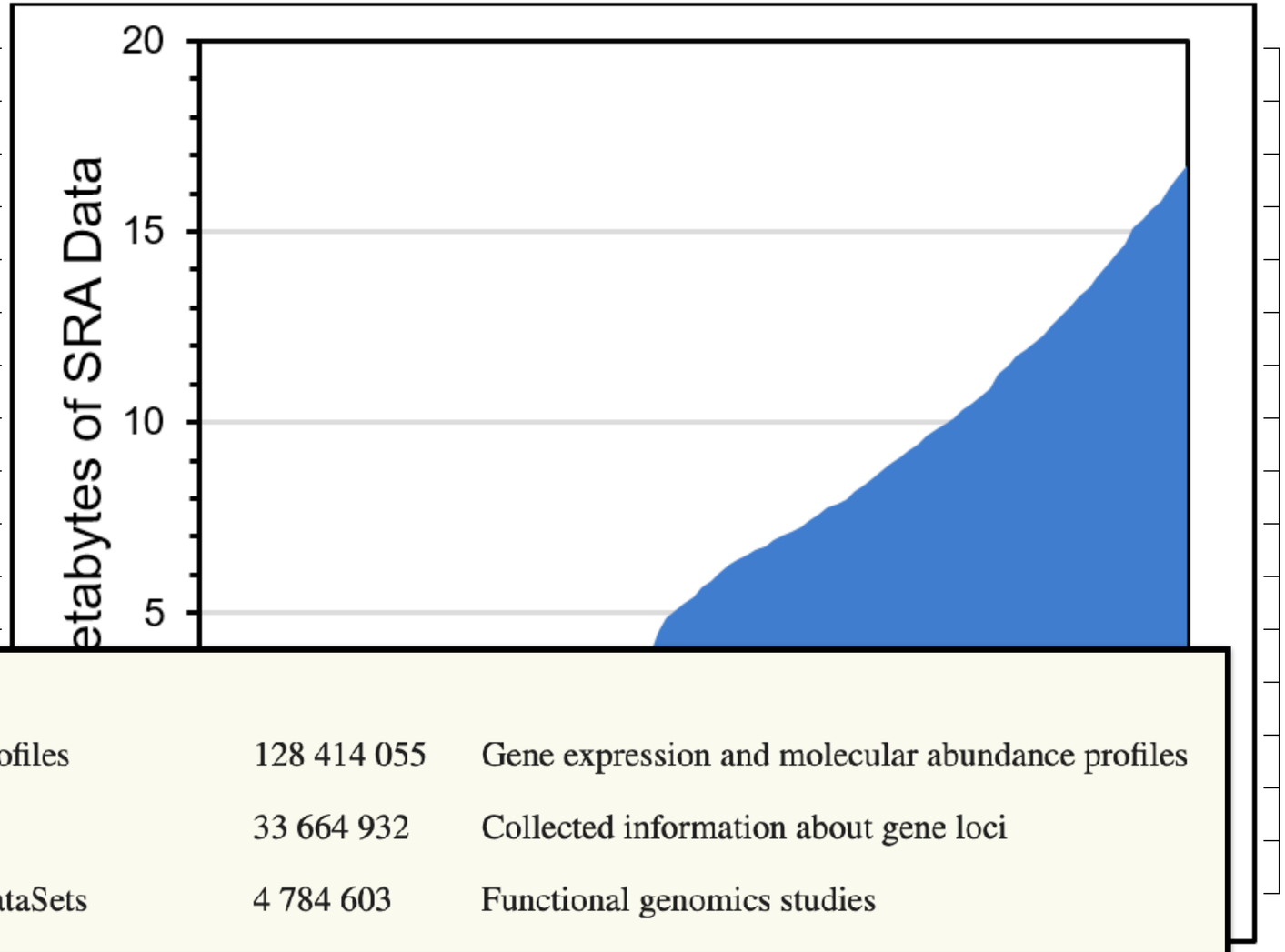
Database	Records
Literature	
PubMed	33 027 761
PubMed Central	7 325 415
NLM Catalog	1 629 799
Bookshelf	892 126
MeSH	348 370
Genomes	
Nucleotide	476 054 019
BioSample	19 473 659
SRA	15 919 320
Taxonomy	2 492 889
Assembly	1 083 900
BioProject	536 242
Genome	64 815
BioCollections	8 468



Computational Biology

How does computing come in?

Database	Records
Literature	
PubMed	33 027 761
PubMed Central	7 325 415
NLM Catalog	1 629 799
Bookshelf	892 126
MeSH	348 370
Genomes	
Nucleotide	476 054 019
BioSample	19 473 659
SRA	15 919 320
Taxonomy	2 492 889
Assembly	1 083 900
BioProject	536 242
Genome	64 815
BioCollections	8 468

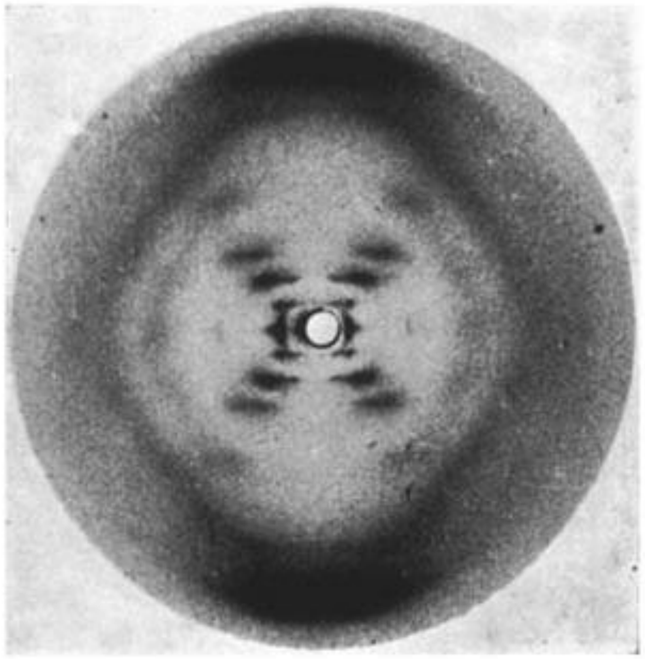


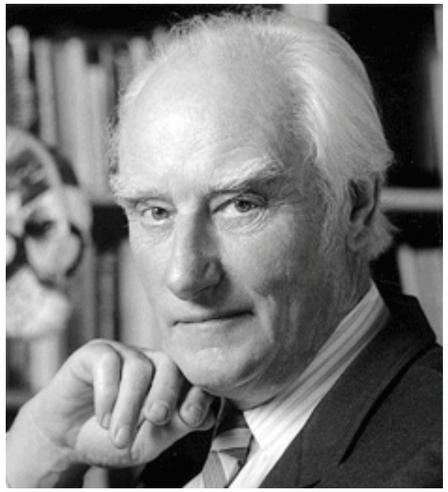
Genes		
GEO Profiles	128 414 055	Gene expression and molecular abundance profiles
Gene	33 664 932	Collected information about gene loci
GEO DataSets	4 784 603	Functional genomics studies

Computational Biology

Follows the history of ‘sequencing’ in general

Franklin’s x-ray diffraction of B form DNA

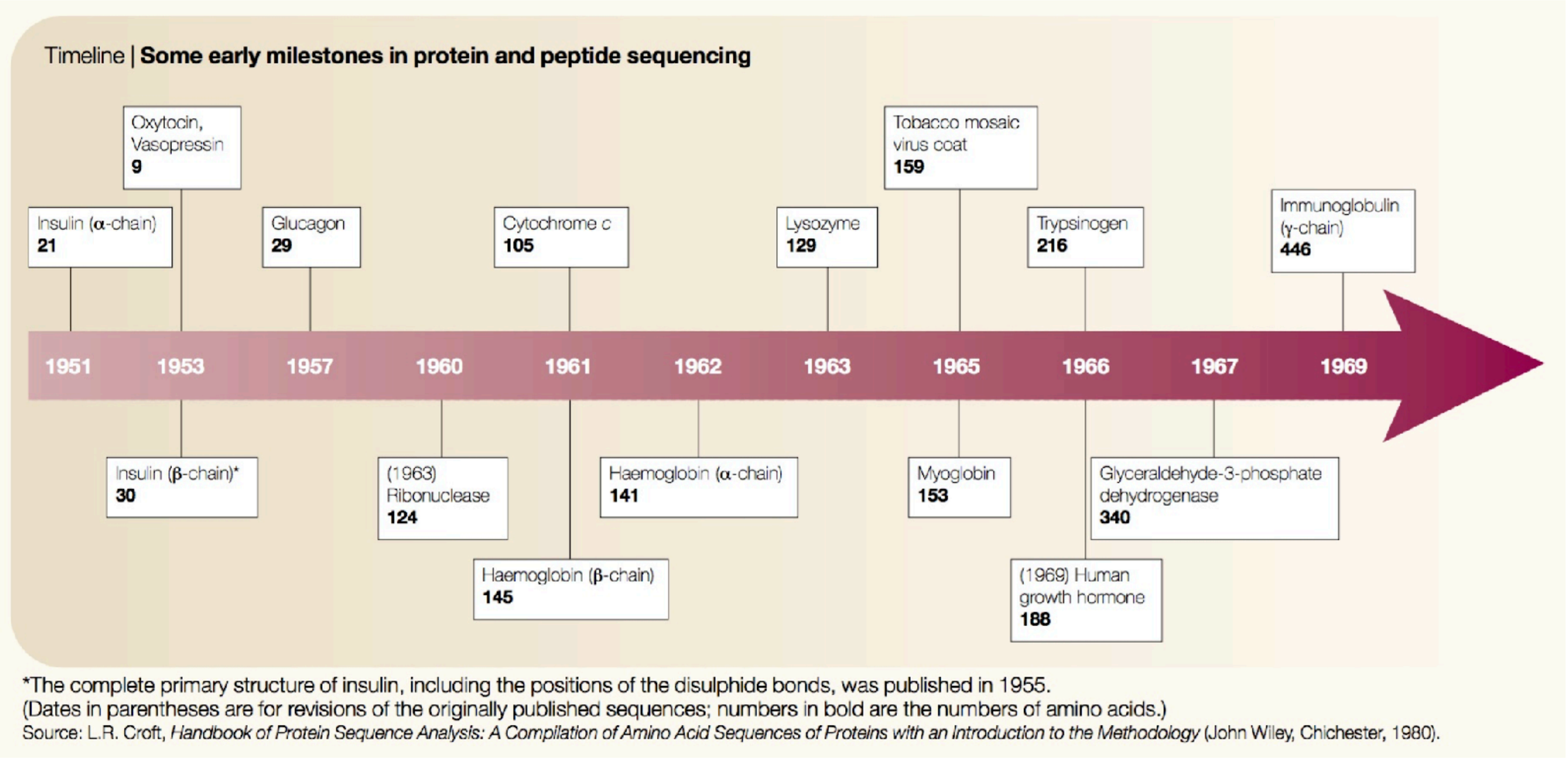




Slide from Aaron Quinlan

Computational Biology

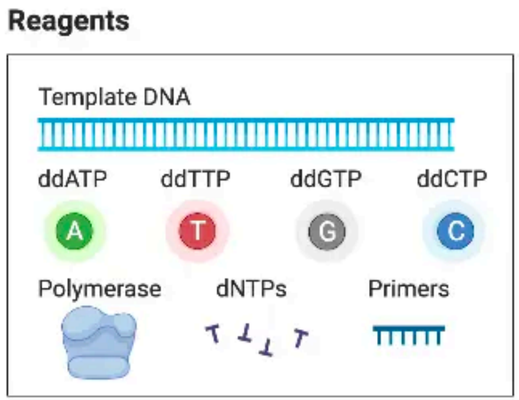
Follows the history of sequencing in general

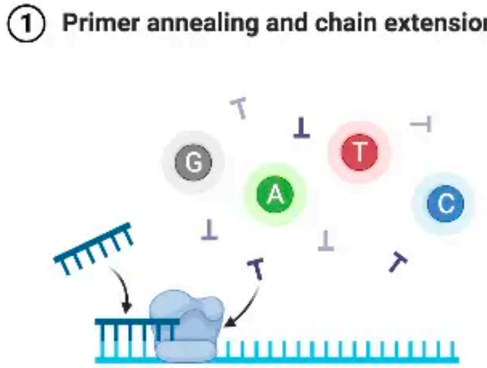


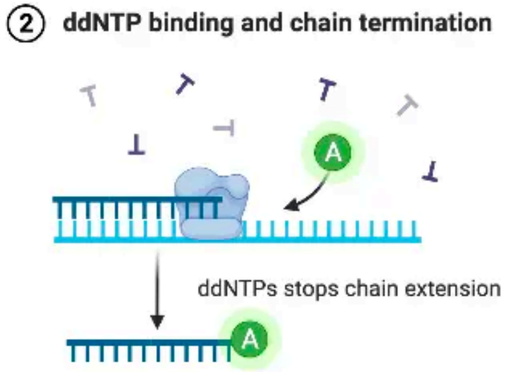
Joel Hagen, “The origins of bioinformatics”, NRG, Dec. 2000

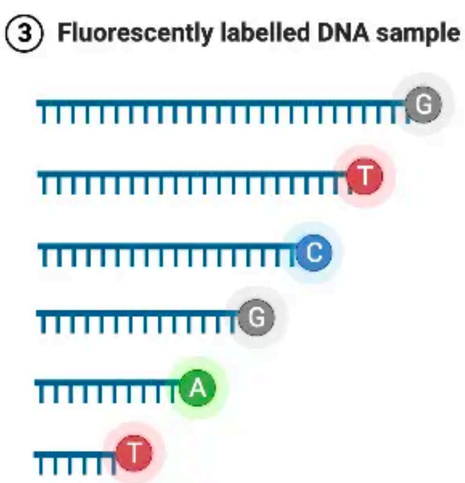
Computational Biology

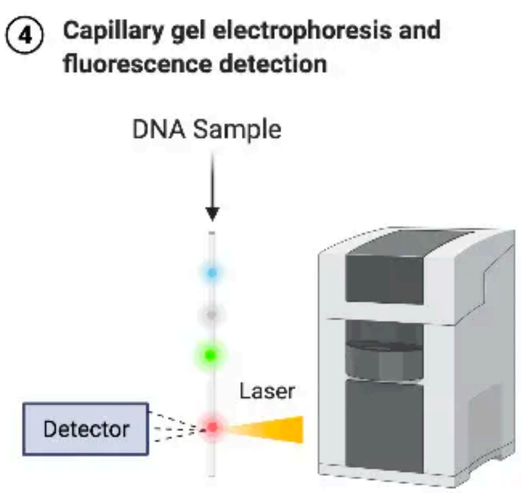
Follows the history of sequencing in general

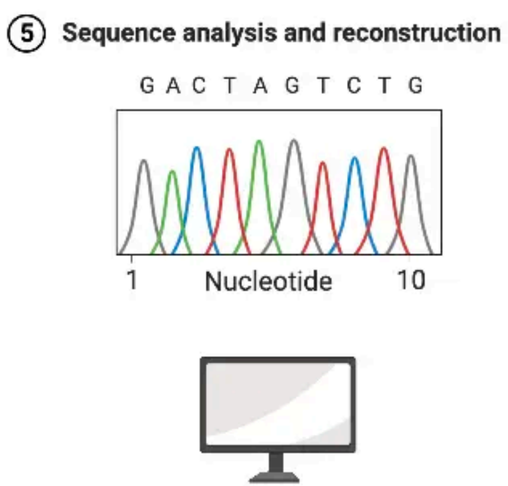


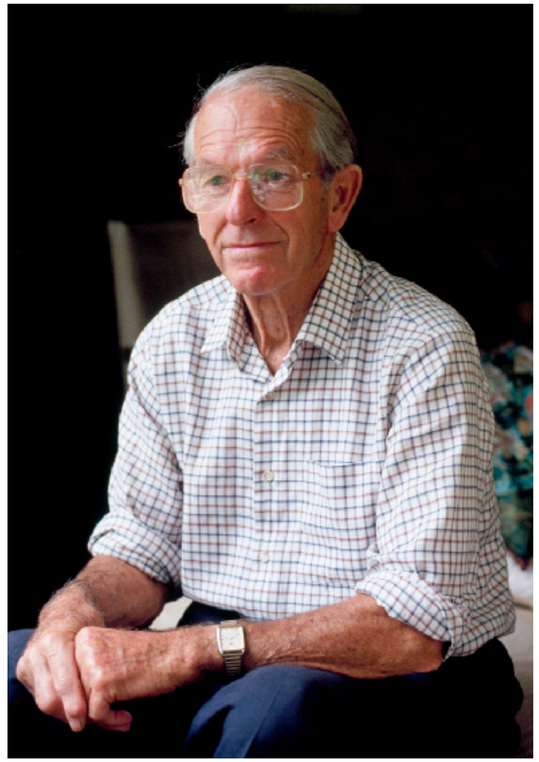










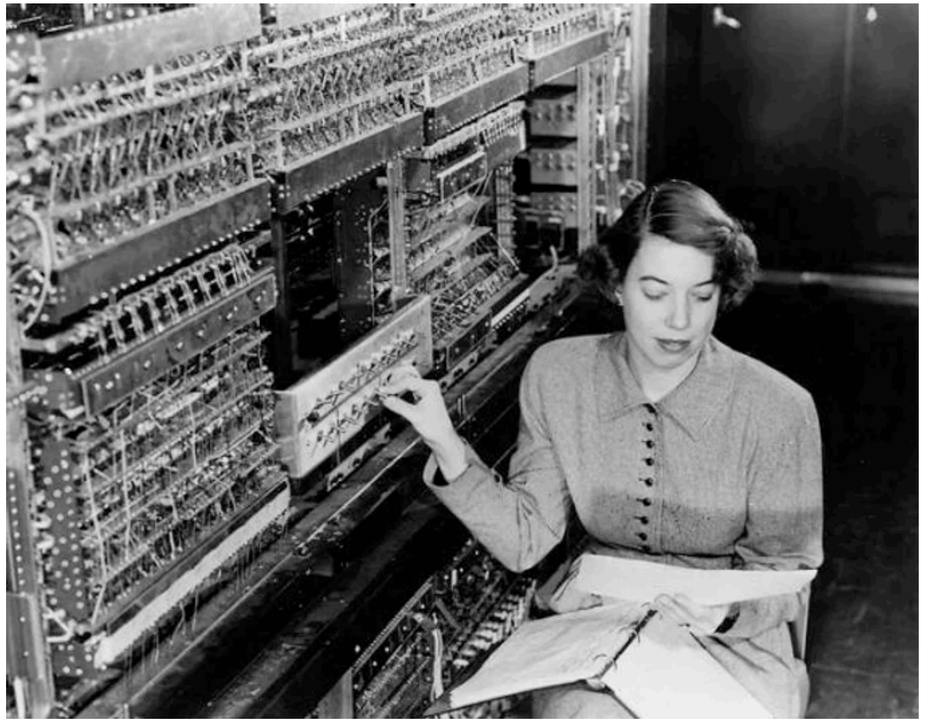


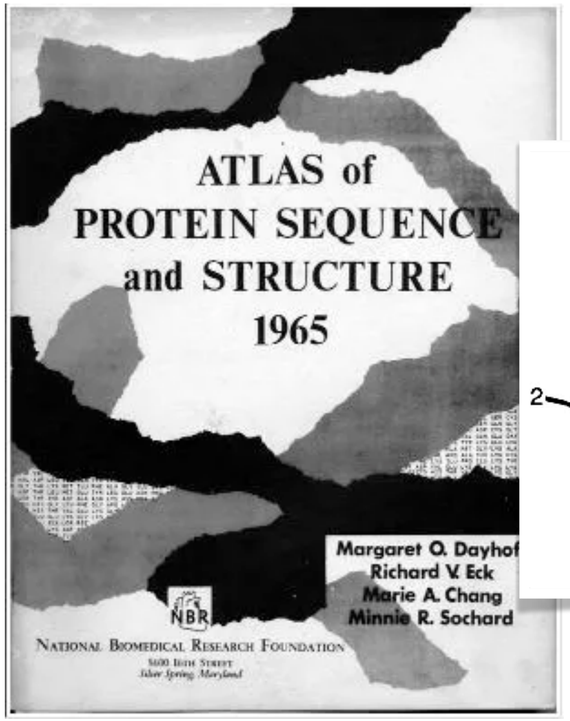
Getty Images

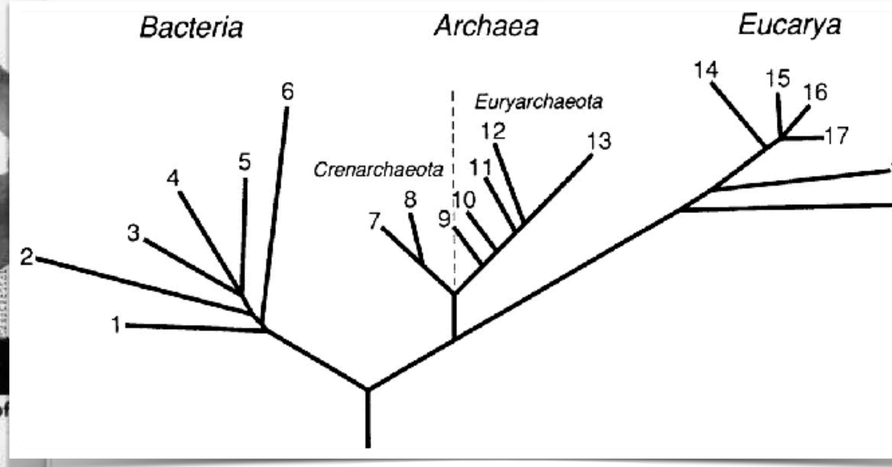
i0.wp.com

Computational Biology

Start asking basic questions about the resulting sequences





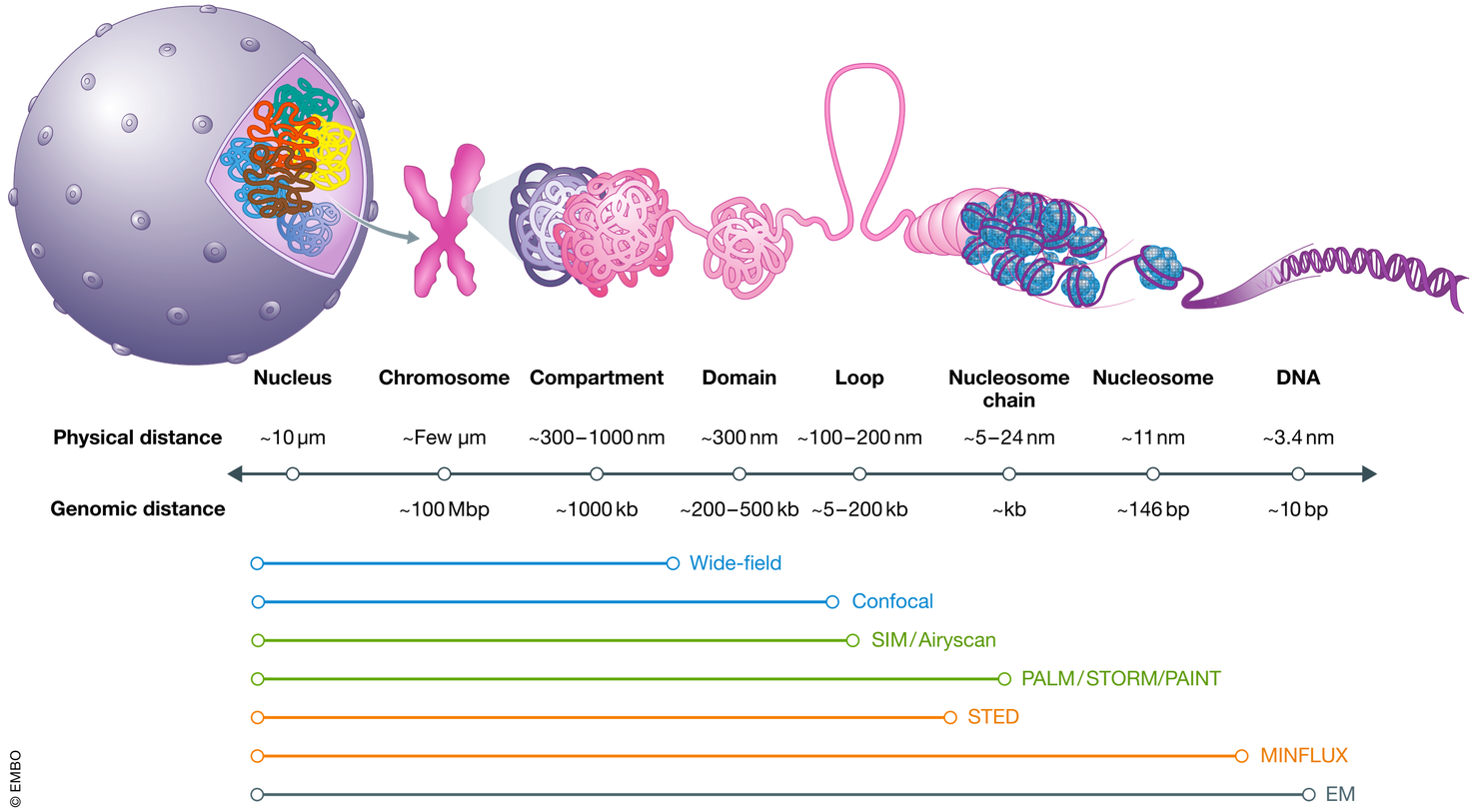


Pace, PNAS 2012

https://annajguzman20.wixsite.com/computationalgenomic/pioneeers

DNA signals

Sequence to function by computational analysis



Liu et al, 2021

TTCCGCTCCGTGCTGCTTTTCGT
GCACGTTCTCTGAGCTGACTACT
AGATTCACGTAGGGTGTGGCGCG
CAAGGCATTTTTTGCGCTTTGTG
CGTTTAGCGTAGAATCTAAGAGT
GGAGGGCCTAATAAATTACACAG
CAGAATACGTTAGTAGTCCGCAC
CGGCCTCGAGCACATCCCTGGGC
CGAACAGCTGCCGCGA

What does this sequence do? A marker to start replication? A gene? Maybe it controls genes expression?

DNA signals

A setup for next time

- Do you see patterns in the sequence to the right?

```
TTCCGCTCCGTGCTGCTTTTCGT
GCACGTTCTCTGAGCTGACTACT
AGATTCACGTAGGGTGTGGCGCG
CAAGGCATTTTTTGCGCTTTGTG
CGTTTAGCGTAGAATCTAAGAGT
GGAGGGCCTAATAAATTACACAG
CAGAATACGTTAGTAGTCCGCAC
CGGCCTCGAGCACATCCCTGGGC
CGAACAGCTGCCGCGA
```

What does this sequence do? A marker to start replication? A gene? Maybe it controls genes expression?

DNA signals

A short aside and setup

- How unexpected is it to see six Ts in a row?
- Longest expected sequence is ~ log4(n), where n is 200 here
- What other about less eye-catching examples?

TTCCGCTCCGTGCTGCTTTTCGT
GCACGTTCTCTGAGCTGACTACT
AGATTCACGTAGGGTGTGGCGCG
CAAGGCA**TTTTTT**GCGCTTTGTG
CGTTTAGCGTAGAATCTAAGAGT
GGAGGGCCTAATAAATTACACAG
CAGAATACGTTAGTAGTCCGCAC
CGGCCTCGAGCACATCCCTGGGC
CGAACAGCTGCCGCGA

What does this sequence do? A marker to start replication? A gene? Maybe it controls genes expression?

DNA signals

A short aside and setup

- This class covers techniques to extract information from underlying sequences, their counts, or their relationships
- Next class we'll start by a easy to describe but hard to solve problem, what sequences is the start signal for replication
- First short write-up due Tuesday on:
  - Your computational biology experience
  - Your interests for final projects
  - Your goals for this class

```
TTCCGCTCCGTGCTGCTTTTCGT
GCACGTTCTCTGAGCTGACTACT
AGATTCACGTAGGGTGTGGCGCG
CAAGGCATTTTTTGCGCTTTGTG
CGTTTAGCGTAGAATCTAAGAGT
GGAGGGCCTAATAAATTACACAG
CAGAATACGTTAGTAGTCCGCAC
CGGCCTCGAGCACATCCCTGGGC
CGAACAGCTGCCGCGA
```

What does this sequence do? A marker to start replication? A gene? Maybe it controls genes expression?

Setup for next time

Jupyter notebook setup

https://jupyter.org/