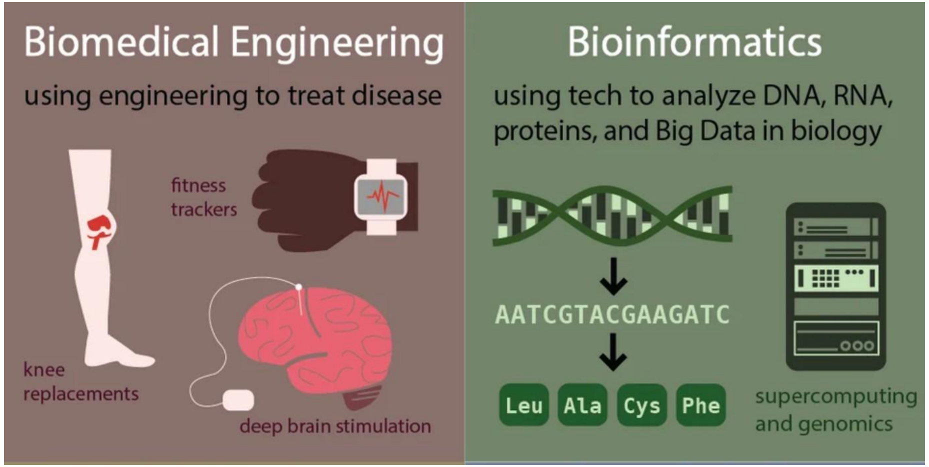
Computational Biology and Bioinformatics

August 5, 2025
Cedric Chan

Biomedical Engineering vs. Bioinformatics vs. Computational Biology vs. Biotechnology

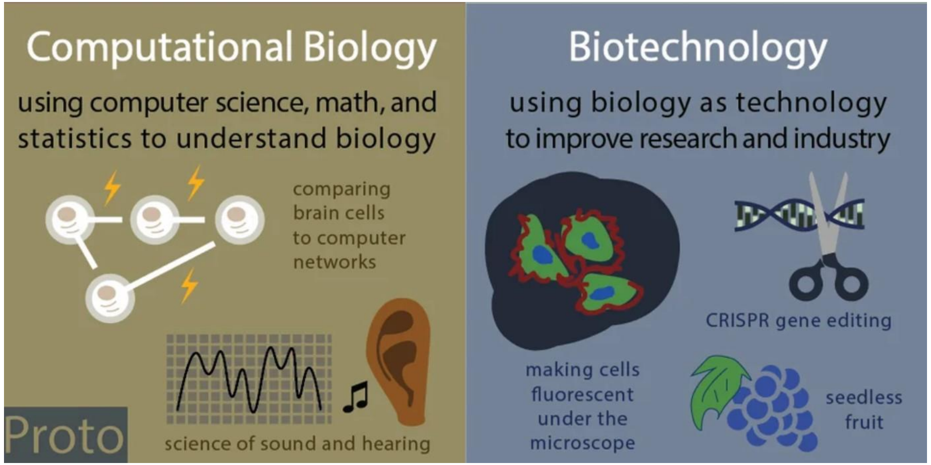


Biomedical Engineering

using engineering to treat disease

Bioinformatics

using tech to analyze DNA, RNA, proteins, and Big Data in biology



Computational Biology

using computer science, math, and statistics to understand biology

comparing brain cells to computer networks
Proto
science of sound and hearing

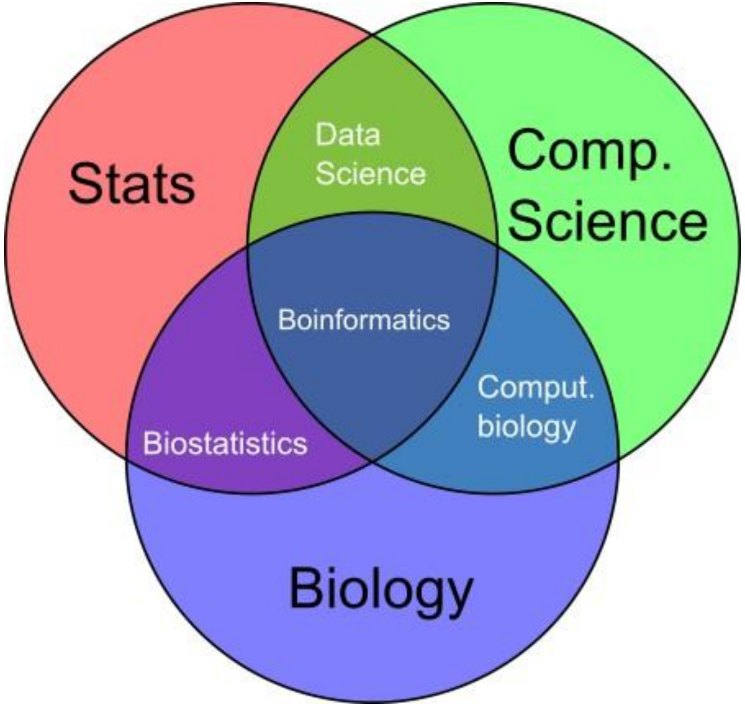
Biotechnology

using biology as technology to improve research and industry

CRISPR gene editing
making cells fluorescent under the microscope
seedless fruit

Today we'll focus on **Bioinformatics** and **Computational Biology**

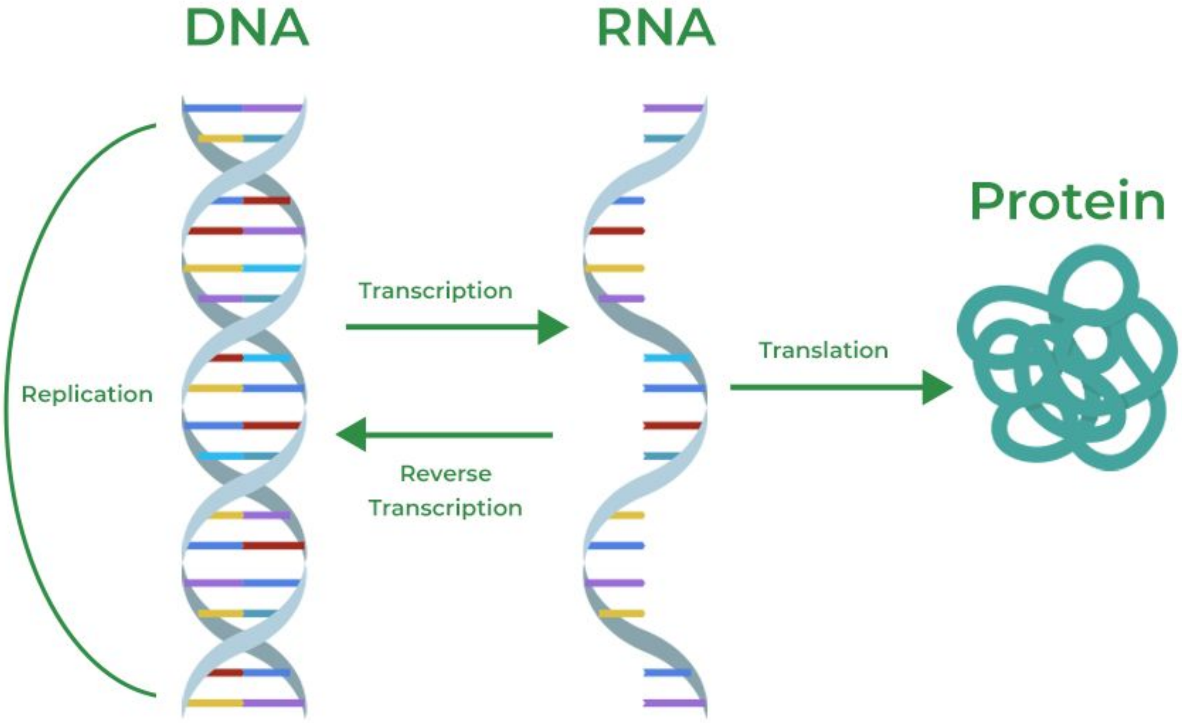
Focus for Today: Bioinformatics



Fields in Bioinformatics

- ***Translational Bioinformatics***– Development of techniques for transforming voluminous biomedical (especially genomic) data to support proactive, predictive, preventive, and participatory health
- ***Clinical Research Informatics***– Development of approaches for enabling the discovery, management, and evaluation of new health knowledge
- ***Clinical Informatics***– Development and application of techniques to improve health care delivery services; clinical informatics is a subspecialty of the American Board of Medical Specialties
- ***Consumer Health Informatics***– Development of information structures and approaches for supporting patient-centric health care needs
- ***Public Health Informatics***– Development of methodologies for supporting public health needs, including surveillance, prevention, preparedness, and health promotion

Central Dogma of Biology



Genome

- **DNA:** string of complex molecules called nucleotides. It contains the genetic information and acts as a set of instructions for how to build and maintain you

- **Genome:** complete set of DNA

- **Gene:** DNA is organized into little chunks of information that each carry a specific set of instructions for how to make a certain aspect of you

Bioinformatics: Genomic Analysis

How does bioinformatics allow us to understand the similarity in genes?

Mouse

Algorithms will scan past both ends of the matching sequence

. . . ATGCGTAGC CATATCCGA A TCGA . . .

Similarities in sequences: Analyze those genes and see how they translate into similar traits

Differences in sequences: Analyze those genes and see how they translate into different traits

. . . ATGCGTAGC CATATCCGA A CTTT . . .

Human

Bioinformatics: Genomic Analysis

Cinderella

. . . . ATGCGTAGCCACATCCGAATCGA . . .

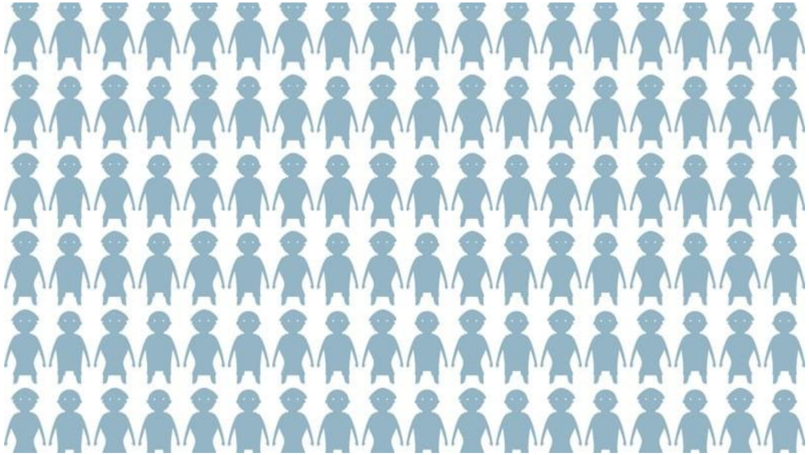
Is this base difference C/T significant for disease?

Belle

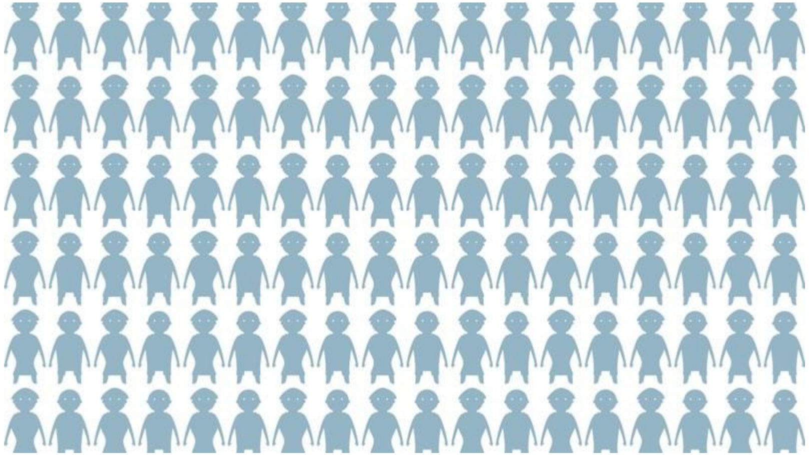
. . . . ATGCGTAGCCATATCCGAATCGA . . .

Conduct a Study

Is this base difference significant for disease?



Group A: 100 Healthy Subjects



Group B: 100 Diabetic Subjects

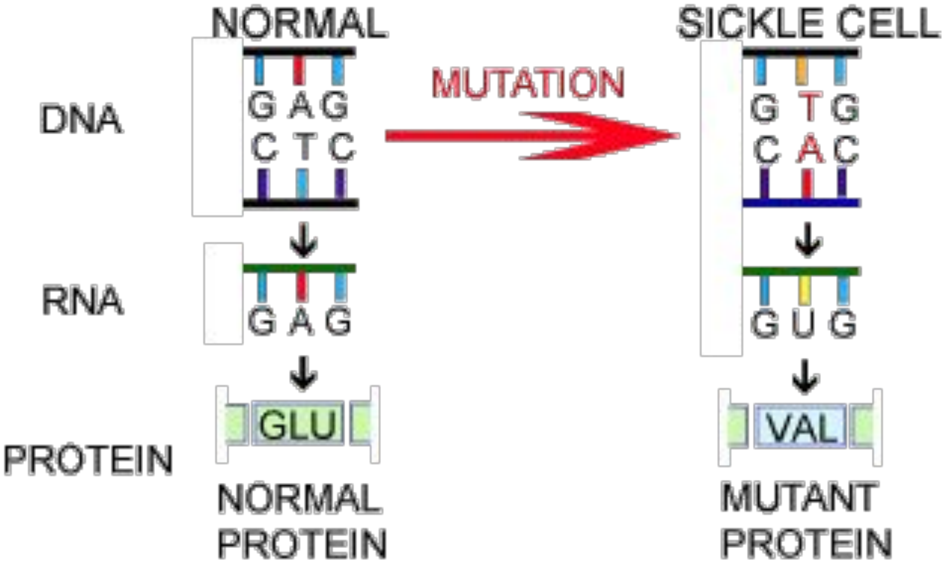
Hypothetical Results: **4/100** of group A have a T and **98/100** of group B have T

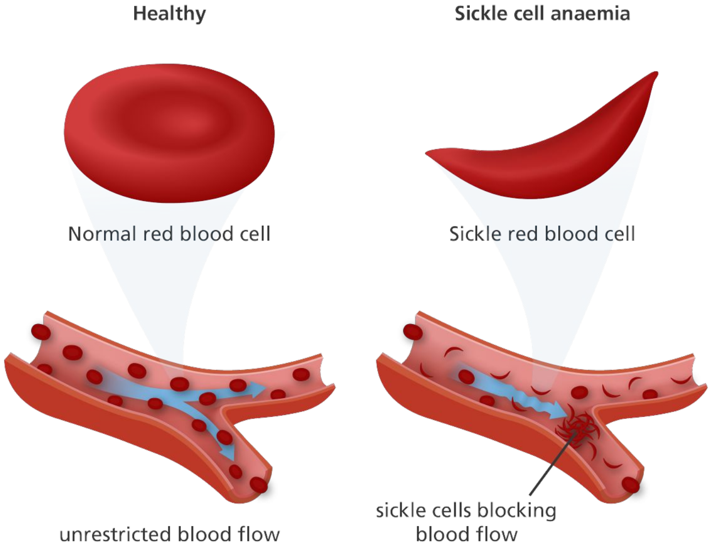
GWAS (Genome-wide association study)
Look at the DNA of many people, and ask: Are certain changes in DNA more common in people with a trait vs without?

Real life example: Base Substitution in Sickle Cell Disease

- Sickle cell disease is an inherited disease in which red blood cells contort into a sickle shape and die early, leaving a shortage of healthy red blood cells

- Discovered through genomic analysis, the genetic basis of sickle cell disease is an **A-to-T transversion** in the sixth codon of the HBB gene



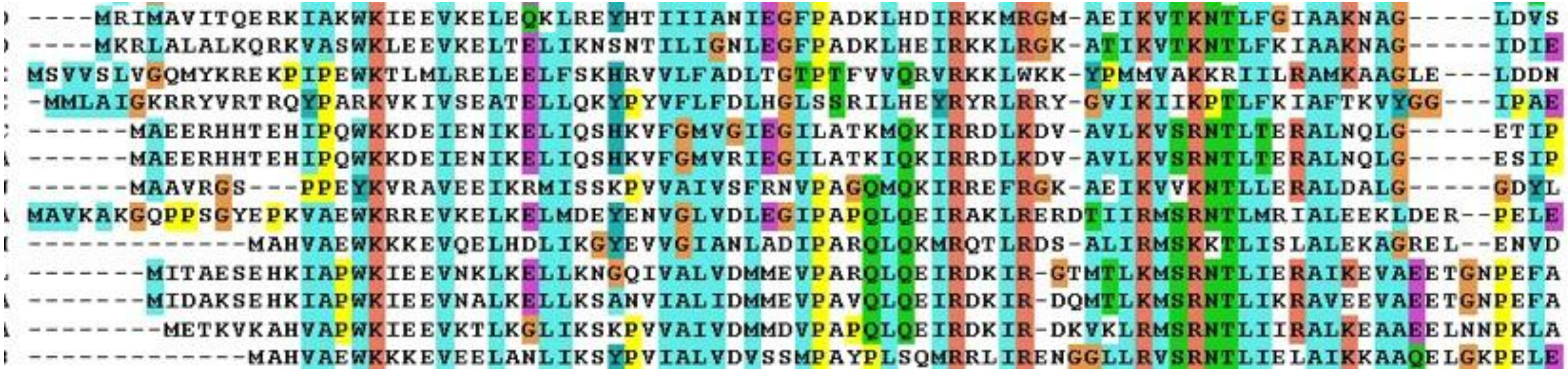


Sequence Alignment

- **Sequence alignment** is a way of arranging the DNA sequences to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences

- Aligned sequences are typically represented as rows within a matrix

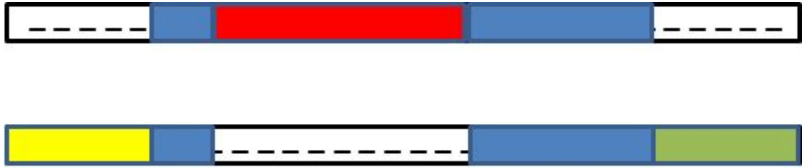
Insulin Gene Sequence Database



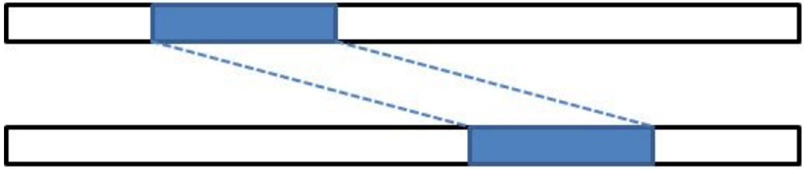
```
Mouse  ----MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG-----LDVS
Rat    ----MKRLALALKQRKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG-----IDIE
       MSVVSLVGQMYKREKPIPEWKTLMLRELEELFSKHRVVLFADLTGTPTFVVQRVRKKLWKK-YPMMVAKKRIILRAMKAAGLE---LDDN
       -MMLAIGKRRYVRTRQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-GVIKIIKPTLFKIAFTKVYGG---IPAE
       ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ETIP
       ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKIQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ESIP
       ------MAAVRGS---PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGQMQKIRREFRGK-AEIKVVKNTLLERALDALG-----GDYL
       MAVKAKGQPPSGYEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPQLQEIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE
       -------------MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLISLALEKAGREL--ENVD
       -------MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQLQEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA
       -------MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVQLQEIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA
       --------METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIRDKIR-DKVKLRMSRNTLIIRALKEAAEELNNPKLA
       -------------MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAQELGKPELE
```

Global & Local Alignment

- The global approach compares one whole sequence with other entire sequences
- The output of a global alignment is a one-to-one comparison of two sequences
    - Used when comparing two genes of similar function
- The local method uses a subset of a sequence and attempts to align it to subset of other sequences
- Local regions are aligned with the **highest level of similarity**
- Looking for conserved patterns in DNA



Global Alignment



Local Alignment

Exercise

First, lets try to implement the naive algorithm for local alignment.

Problem: You are given a reference string ref and a query string query. Return the maximal alignment score of query to ref. Assume the score function properly scores two characters according to an arbitrary scoring scheme.

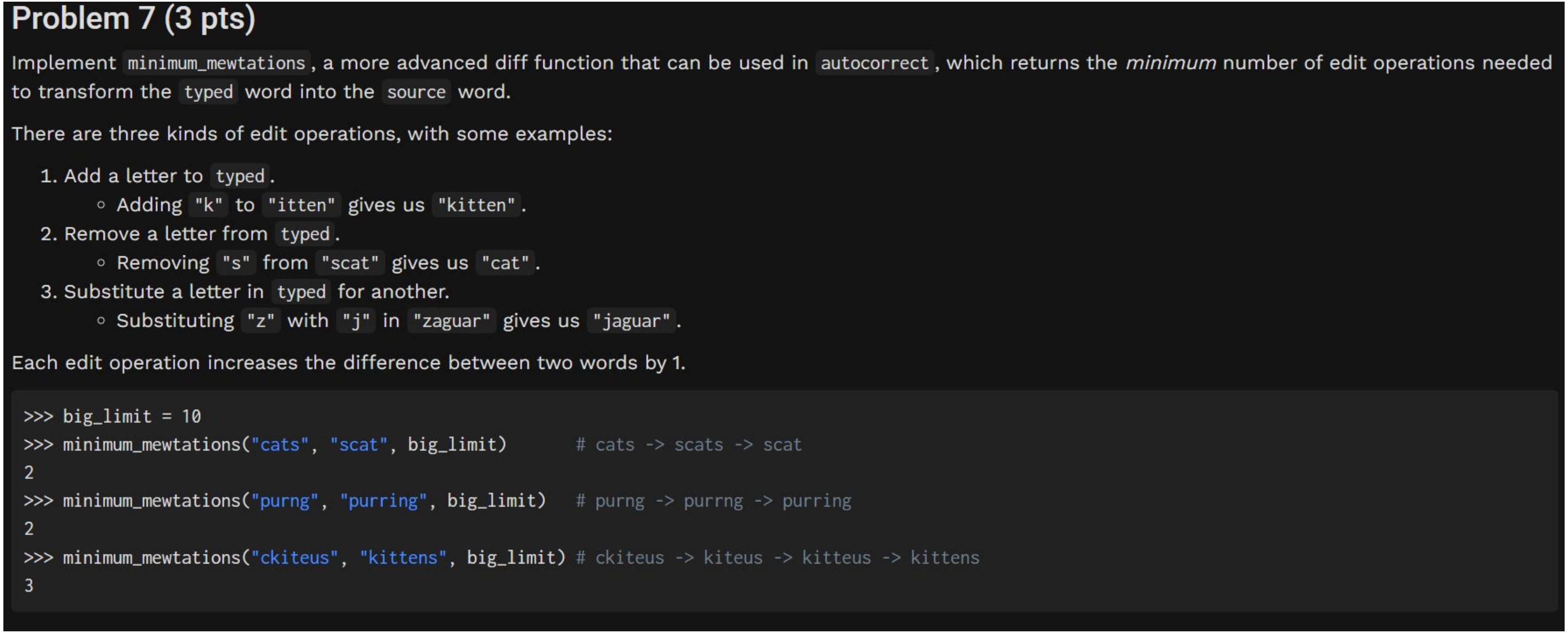
```
def score(c1, c2): #Assume score takes in two characters and returns a score

def local_align(ref, query):

    #IMPLEMENTATION OF BASE CASES NOT SHOWN

    return #YOUR CODE HERE
```

Look familiar?



Problem 7 (3 pts)

Implement `minimum_mewtations`, a more advanced diff function that can be used in `autocorrect`, which returns the *minimum* number of edit operations needed to transform the `typed` word into the `source` word.

There are three kinds of edit operations, with some examples:

1. Add a letter to `typed`.
   - Adding `"k"` to `"itten"` gives us `"kitten"`.
2. Remove a letter from `typed`.
   - Removing `"s"` from `"scat"` gives us `"cat"`.
3. Substitute a letter in `typed` for another.
   - Substituting `"z"` with `"j"` in `"zaguar"` gives us `"jaguar"`.

Each edit operation increases the difference between two words by 1.

```
>>> big_limit = 10
>>> minimum_mewtations("cats", "scat", big_limit)       # cats -> scats -> scat
2
>>> minimum_mewtations("purng", "purring", big_limit)   # purng -> purrng -> purring
2
>>> minimum_mewtations("ckiteus", "kittens", big_limit) # ckiteus -> kiteus -> kitteus -> kittens
3
```

Exercise

First, lets try to implement the naive algorithm for local alignment.

Problem: You are given a reference string ref and a query string query. Return the maximal alignment score of query to ref. Assume the score function properly scores two characters according to an arbitrary scoring scheme.

```
#IMPLEMENTATION OF BASE CASES NOT SHOWN

return max( local_align(ref[1:],query[1:]) + score(ref[0],query[0]), #match
            local_align(ref[1:],query) + score(ref[0], ‘-’,), #Skip first ref
            local_align(ref, query[1:]) + score(query[0], ‘-’), #Skip first query
            0 #reset the local alignment)
```

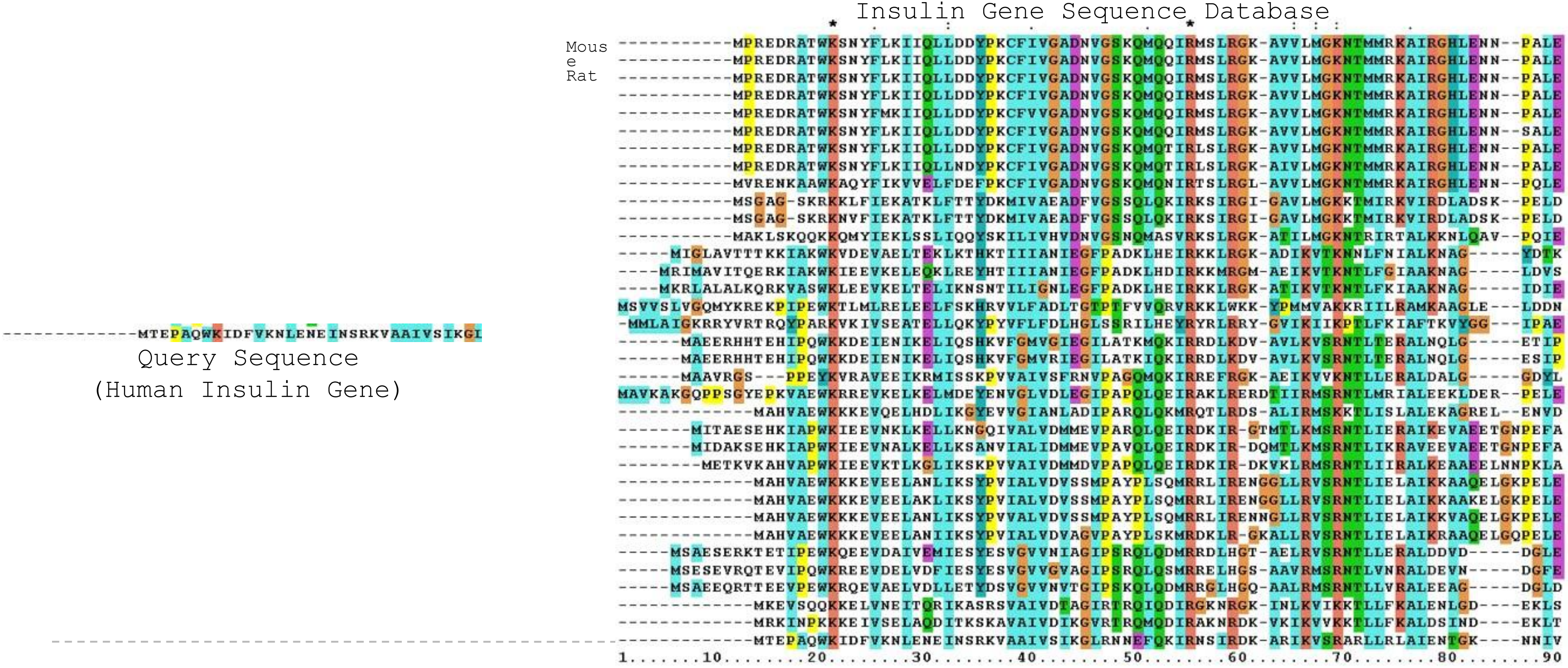
Problem: What’s the runtime? **O(3^n)!!!**

How can we improve this? **Dynamic Programming** improves runtime to **O(nm)**

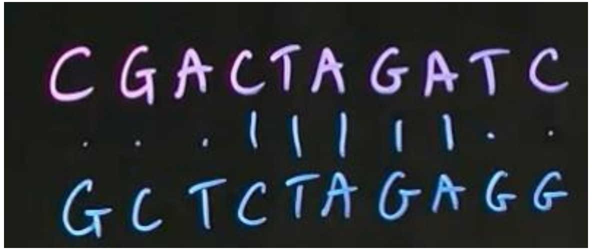
Dynamic Programming (just recursion with caching), recursive call is the same!

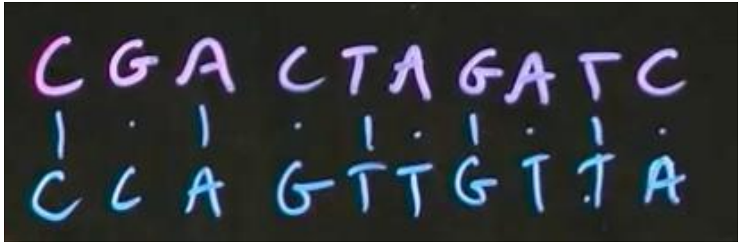
BLAST: Basic Local Alignment Search Tool

- Identifies similarities between sequences by comparing it with database of sequences



Glance of the BLAST Algorithm





Query Sequence

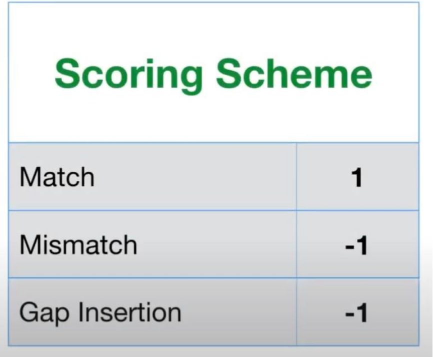
Target Sequence in the Database

Query Sequence

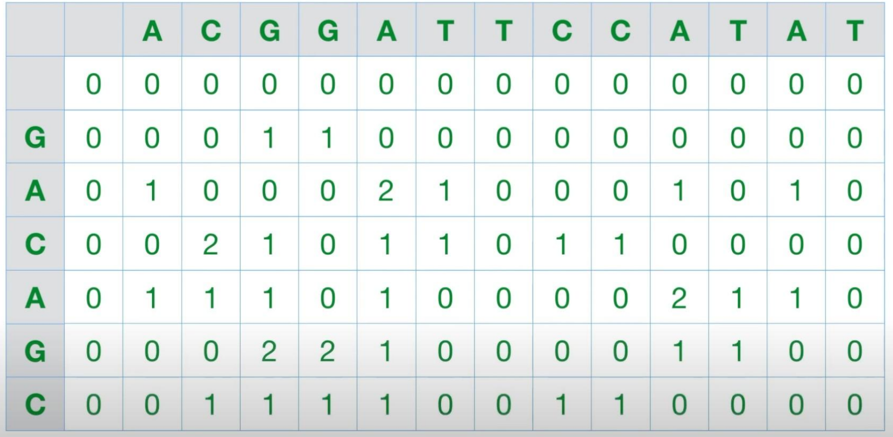
GACAGC

Database Sequence

ACGGATTCCATAT

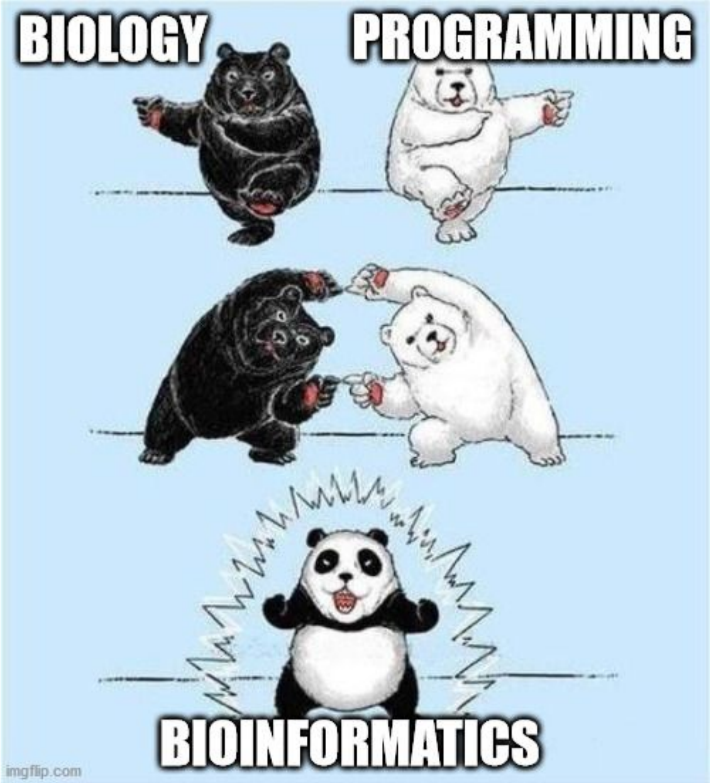


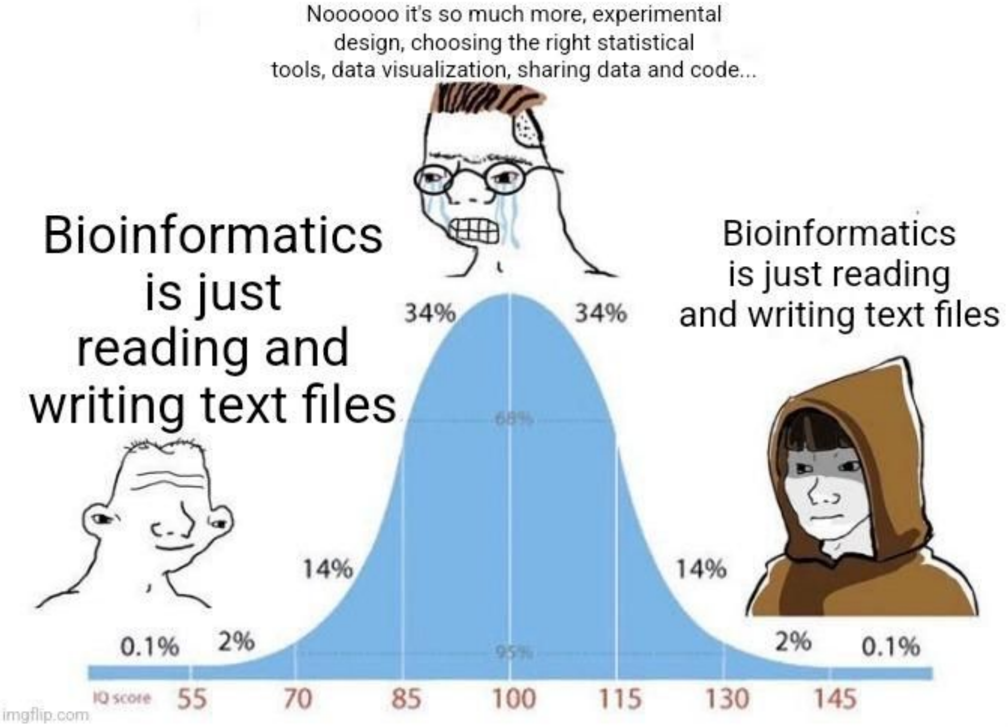
Scoring Scheme	
Match	**1**
Mismatch	**-1**
Gap Insertion	**-1**



		A	C	G	G	A	T	T	C	C	A	T	A	T
	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	1	1	0	0	0	0	0	0	0	0	0
A	0	1	0	0	0	2	1	0	0	0	1	0	1	0
C	0	0	2	1	0	1	1	0	1	1	0	0	0	0
A	0	1	1	1	0	1	0	0	0	0	2	1	1	0
G	0	0	0	2	2	1	0	0	0	0	1	1	0	0
C	0	0	1	1	1	1	0	0	1	1	0	0	0	0

Break


BIOLOGY
PROGRAMMING
BIOINFORMATICS
imgflip.com


Noooooo it's so much more, experimental design, choosing the right statistical tools, data visualization, sharing data and code...
Bioinformatics is just reading and writing text files
Bioinformatics is just reading and writing text files
34%
34%
68%
14%
14%
0.1%
2%
95%
2%
0.1%
IQ score
55
70
85
100
115
130
145
imgflip.com

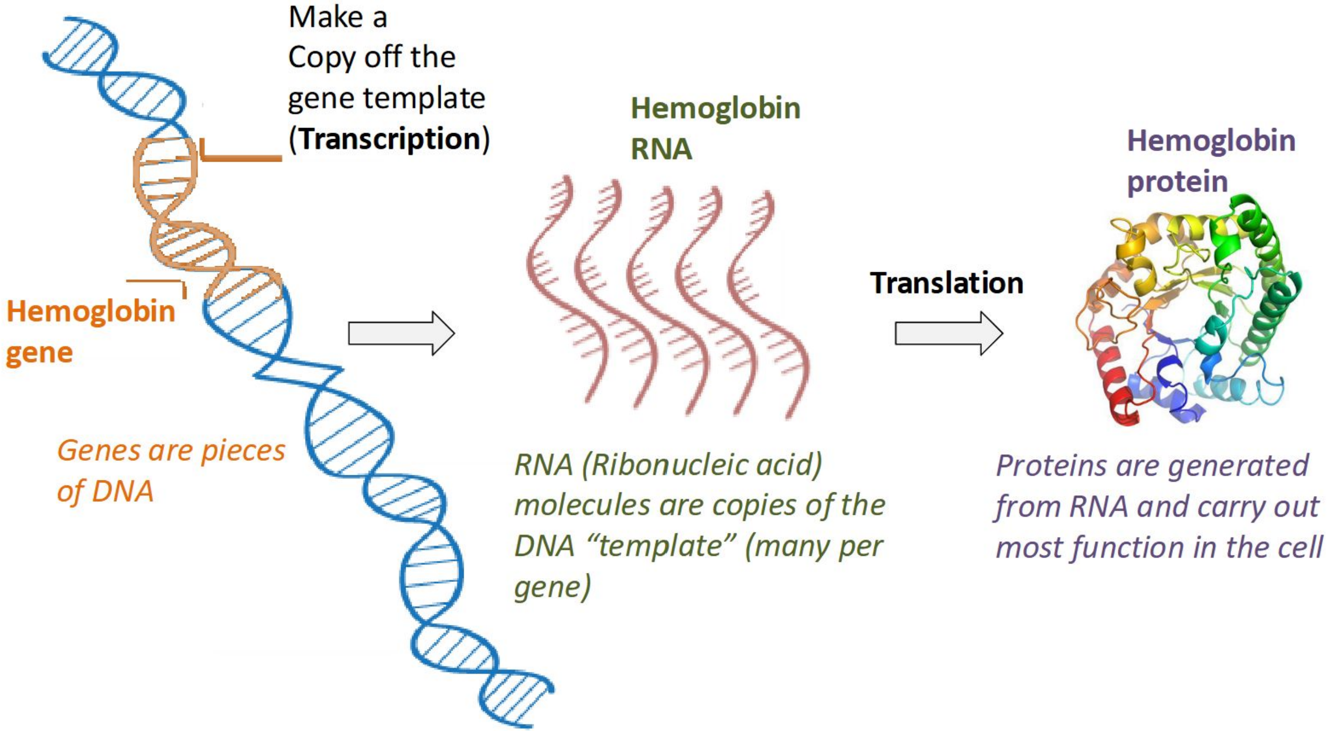
Step back: Cell Types

Your body has many different cell types, but every cell has the same DNA, how is this possible?

In other words, what makes these cells different?

Answer: The **RNA in every cell is different!**

Review: Central Dogma of Biology



Make a
Copy off the
gene template
(**Transcription**)

Hemoglobin gene

Genes are pieces of DNA

Hemoglobin RNA

RNA (Ribonucleic acid) molecules are copies of the DNA "template" (many per gene)

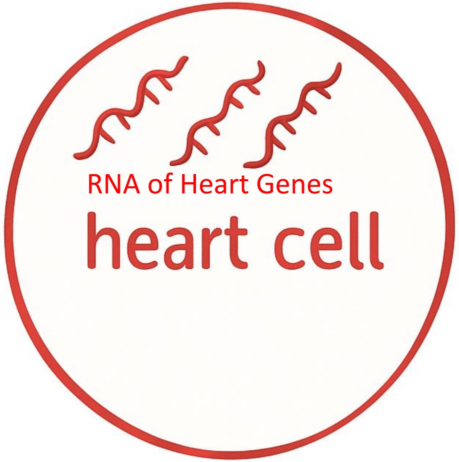
Translation

Hemoglobin protein

Proteins are generated from RNA and carry out most function in the cell

Cell Types continued

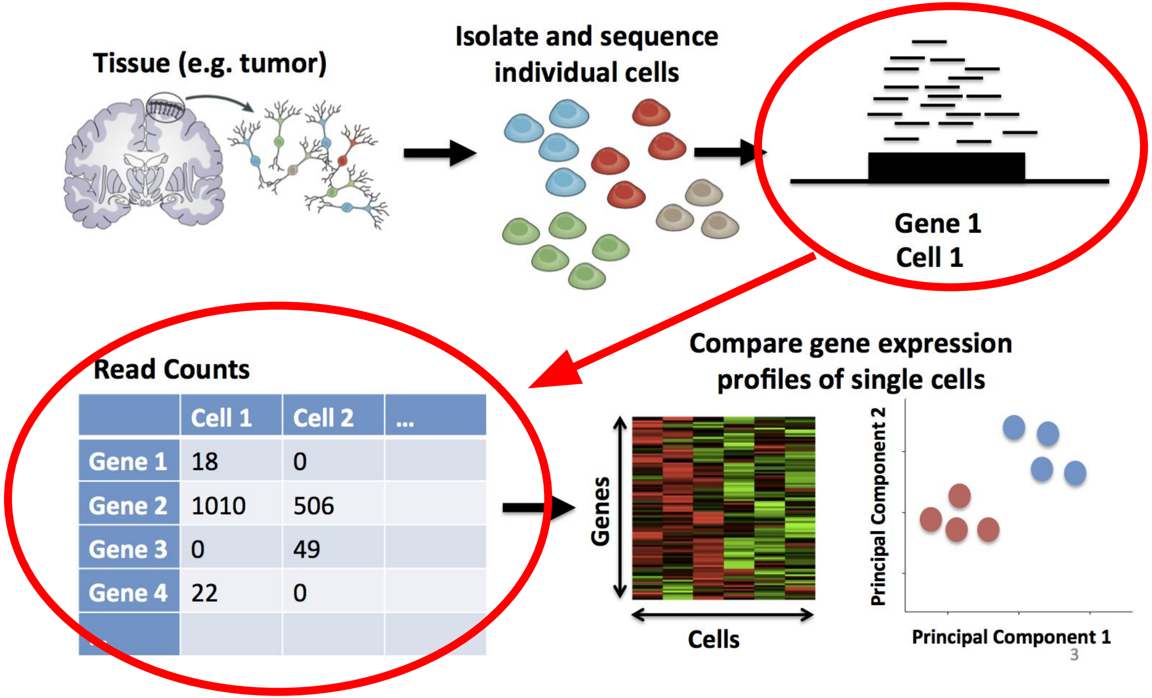
RNA of Skin Genes
Skin cell



Therefore, We know the identity of a cell based on the amount of RNA of specific genes!

	Skin Cell	Heart Cell
RNA of Skin Gene #1	74	0
RNA of Skin Gene #2	17	1
RNA of Skin Gene #3	34	0
RNA of Heart Gene #1	0	20
RNA of Heart Gene #2	3	124
RNA of Heart Gene #3	2	75

Single Cell RNA Sequencing (scRNAseq)


Tissue (e.g. tumor)
Isolate and sequence individual cells
Gene 1
Cell 1
Compare gene expression profiles of single cells
Genes
Cells
Principal Component 2
Principal Component 1

Read Counts

	Cell 1	Cell 2	...
Gene 1	18	0	
Gene 2	1010	506	
Gene 3	0	49	
Gene 4	22	0	

How did this happen?

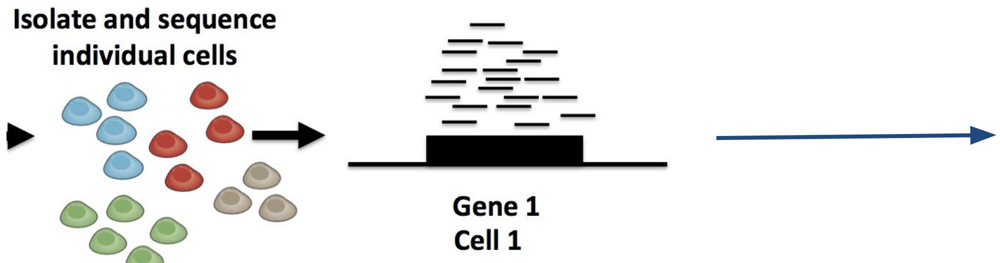
Generating a Gene Expression Matrix

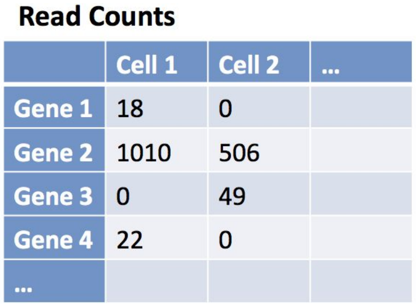
If I know what RNA my cell has, and I also know the sequence of every gene, how can I know what gene my RNA encodes for?

Use the local alignment algorithm from before!

```
>>> local_align(reference transcriptome, RNA #1 from cell #1)
Some score X
```

Not included above, but if you know the score of an RNA, you can also know **where** the score was aligned. This means if RNA X had a maximum alignment score near Gene Y, I know RNA X must code for Gene Y!



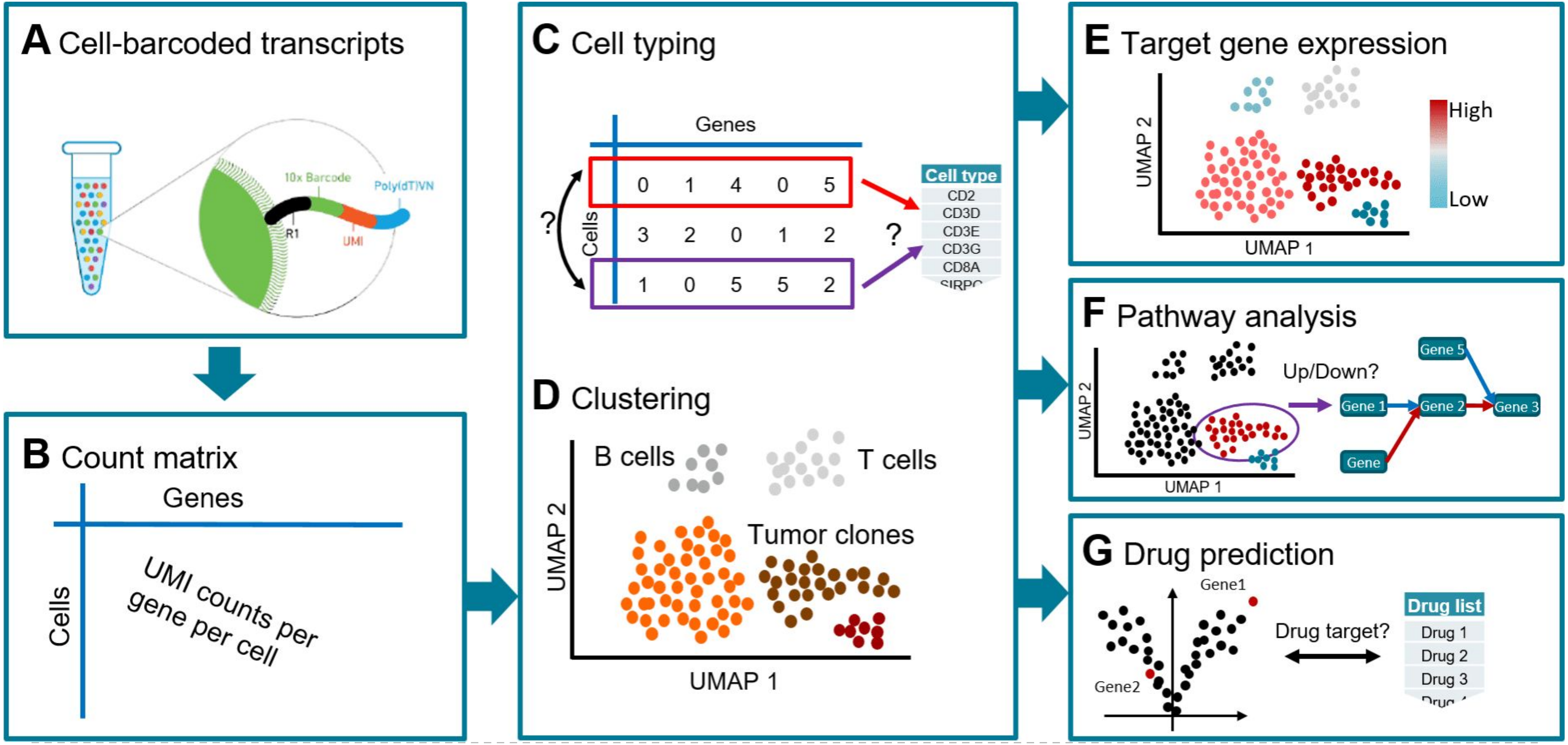


Read Counts

	Cell 1	Cell 2	...
Gene 1	18	0	
Gene 2	1010	506	
Gene 3	0	49	
Gene 4	22	0	
...			

Run local_align on all the RNA from cell 1, repeat for all cells

Analysis of scRNAseq data



A Cell-barcoded transcripts

B Count matrix

C Cell typing

	Genes				
Cells	0	1	4	0	5
	3	2	0	1	2
	1	0	5	5	2

Cell type
CD2
CD3D
CD3E
CD3G
CD8A
SIRPG

D Clustering

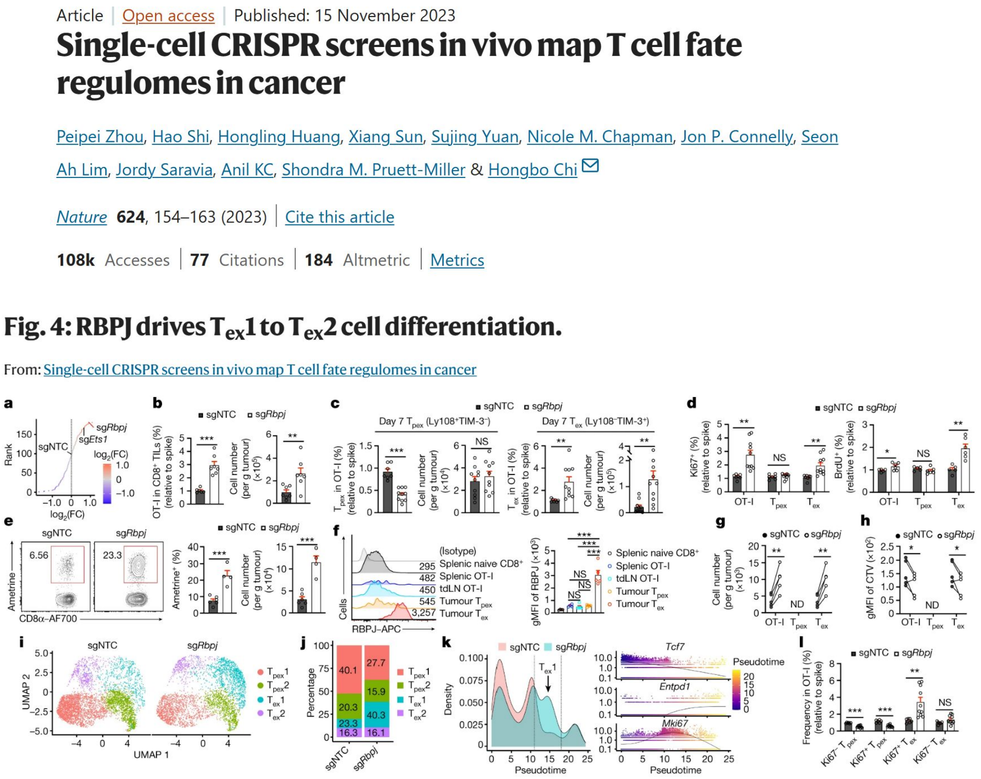
E Target gene expression

F Pathway analysis

G Drug prediction

Drug list
Drug 1
Drug 2
Drug 3
Drug 4

Why use scRNAseq?



Article | Open access | Published: 15 November 2023

Single-cell CRISPR screens in vivo map T cell fate regulomes in cancer

Peipei Zhou, Hao Shi, Hongling Huang, Xiang Sun, Sujing Yuan, Nicole M. Chapman, Jon P. Connelly, Seon Ah Lim, Jordy Saravia, Anil KC, Shondra M. Pruett-Miller & Hongbo Chi

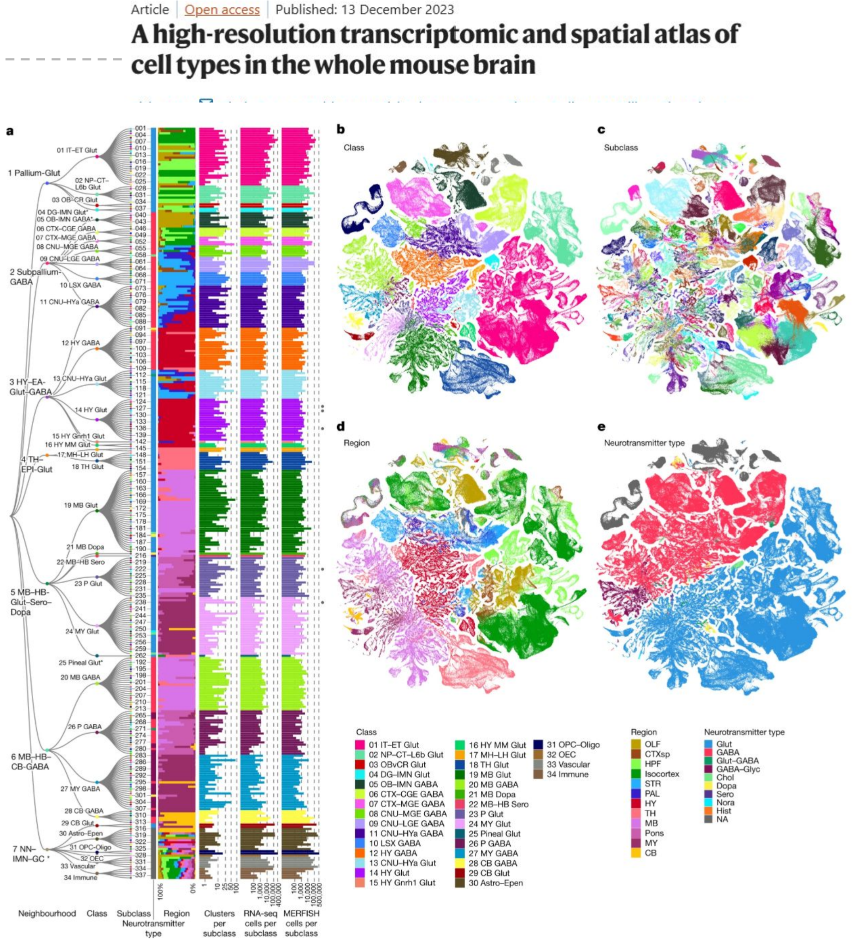
Nature **624**, 154–163 (2023) | Cite this article

108k Accesses | **77** Citations | **184** Altmetric | Metrics

Fig. 4: RBPJ drives $T_{ex}1$ to $T_{ex}2$ cell differentiation.

From: Single-cell CRISPR screens in vivo map T cell fate regulomes in cancer

tldr: knock out RBPJ in cancer killing cell allows it to kill more cancer



Article | Open access | Published: 13 December 2023

A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain

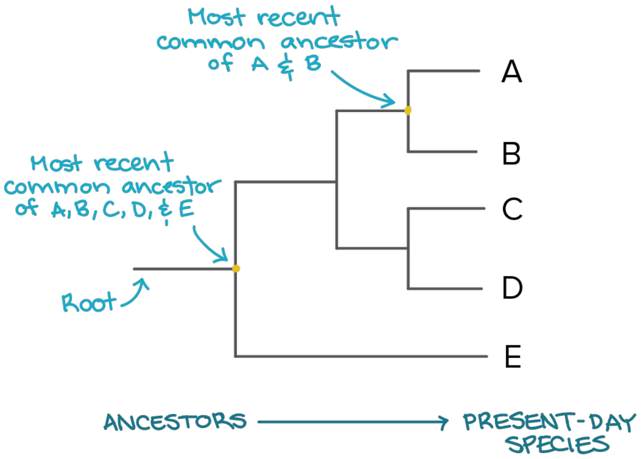
Yao et al., *Nature* 2023

Understanding Evolution: Phylogeny

- How do we track the evolution of a virus? COVID-19 variants, for instance??

- Virus have a VERY HIGH rate of mutation
  - RNA viruses have high mutation rates—up to a **million times higher** than their hosts

- Through genomic analysis of virus samples, we can understand how the sequence of it changes over time
  - Phylogenetic trees allow us to visualize evolution



Most recent common ancestor of A & B
A
B
C
D
E
Most recent common ancestor of A, B, C, D, & E
Root
ANCESTORS
PRESENT-DAY SPECIES

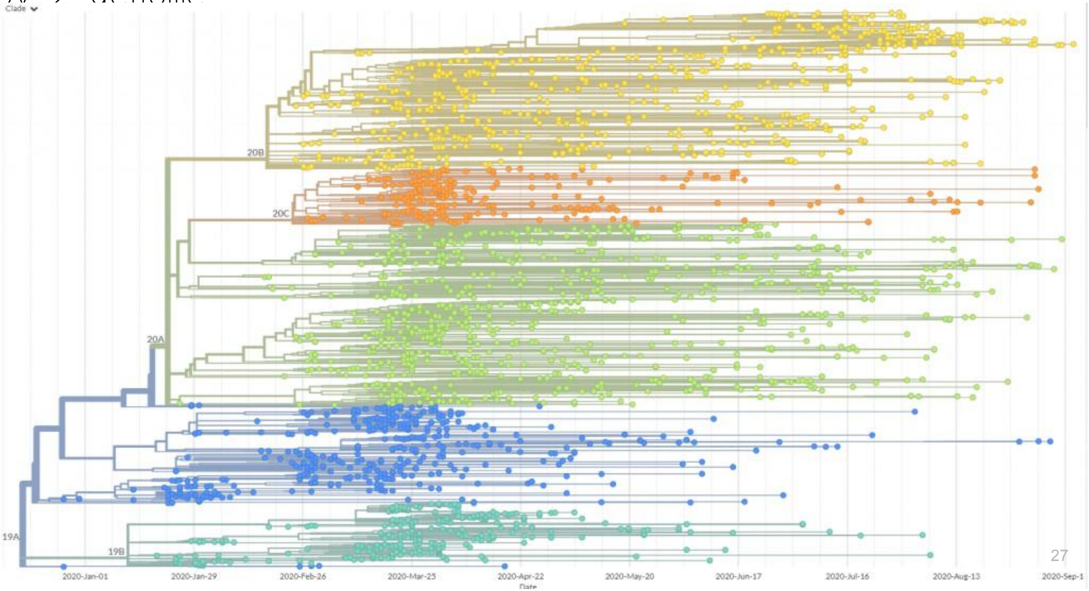
Phylogenetic Trees

- A branching diagram or tree showing the evolutionary relationship among various biological species

- Similarities and differences are based upon physical and genetic characteristics

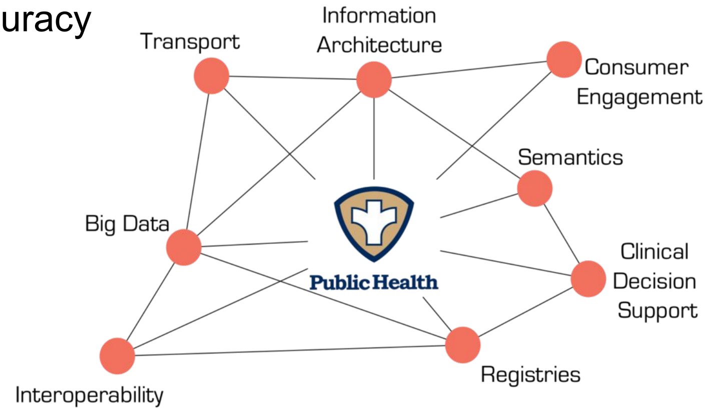
- Two specifies are more related if they have a more recent common ancestor

- The root is the initial Wuhan SARS-CoV-2 genome



Public Health Informatics

- Capturing, managing and analyzing information to improve population-level health outcomes
- Transmit data to public health officials so they can better monitor and prevent disease Providers
- are already using AI algorithms to gain “unprecedented insights into diagnostics, care processes, treatment variability and patient outcomes”
    - 1 in 18 patients getting the wrong diagnosis in the ER department
    - According to the Society for the Improvement of Diagnosis in Medicine (SIDM) between 40,000 and 80,000 individuals die each year due to misdiagnoses
    - “Differential Diagnosis Tool” that had up to 96% diagnostic accuracy

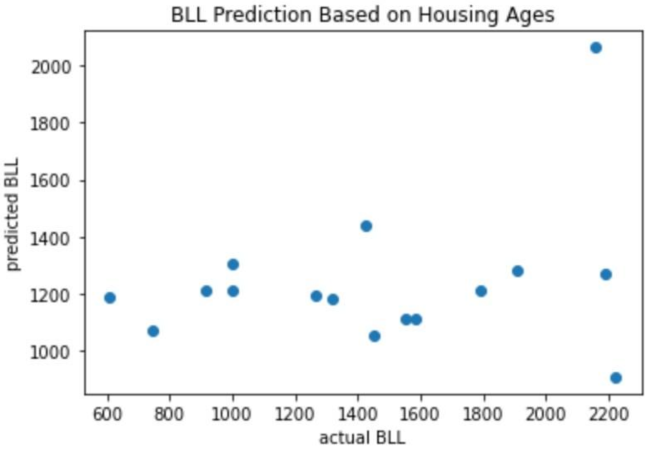


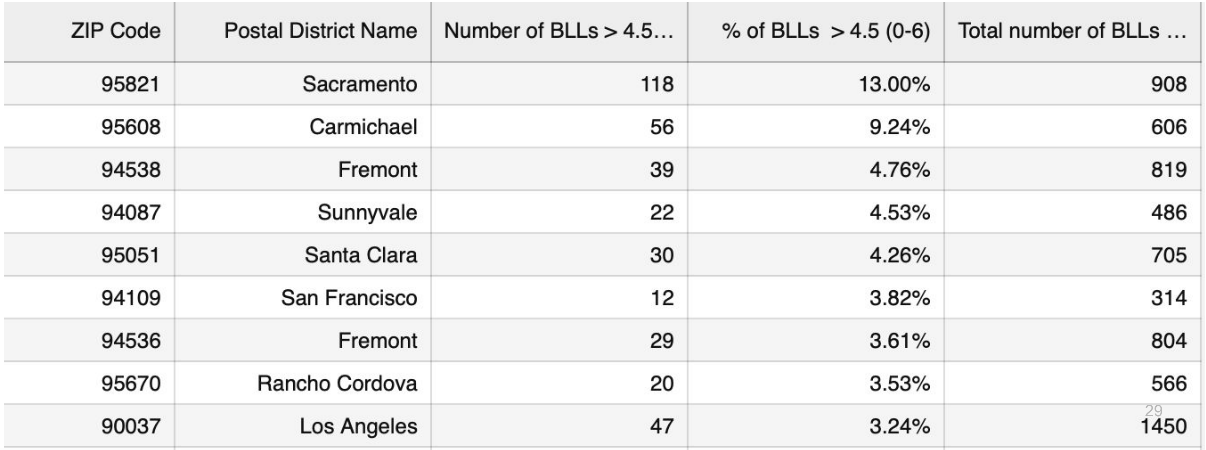
Lead Poisoning Research Project

- There is huge effort for prevention!
- Publicly available data on blood lead levels (BLL) from the Childhood Lead Poisoning Prevention Program (CLPPP)
- Why do some areas have more cases of lead pointing than others?
  - Geographic, demographic, and socioeconomic factors!
  - For instance, I hypothesized there is a positive correlation between the number of severe cases (BLL > 4.5µg/dL) and house age due to likely use of lead paints

Modeling & Testing Hypothesis

  - Geographic, demographic, and socioeconomic factors of a zip code can serve as reasonable features for a multiple regression model to predict number of cases in the future


BLL Prediction Based on Housing Ages



ZIP Code	Postal District Name	Number of BLLs > 4.5…	% of BLLs > 4.5 (0-6)	Total number of BLLs …
95821	Sacramento	118	13.00%	908
95608	Carmichael	56	9.24%	606
94538	Fremont	39	4.76%	819
94087	Sunnyvale	22	4.53%	486
95051	Santa Clara	30	4.26%	705
94109	San Francisco	12	3.82%	314
94536	Fremont	29	3.61%	804
95670	Rancho Cordova	20	3.53%	566
90037	Los Angeles	47	3.24%	1450

Barely scratched the surface!

Translational Bioinformatics- Development of techniques for transforming voluminous biomedical (especially genomic) data to support proactive, predictive, preventive, and participatory health

Clinical Research Informatics- Development of approaches for enabling the discovery, management, and evaluation of new health knowledge

Clinical Informatics- Development and application of techniques to improve health care delivery services; clinical informatics is a subspecialty of the American Board of Medical Specialties

Consumer Health Informatics- Development of information structures and approaches for supporting patient-centric health care needs

Public Health Informatics- Development of methodologies for supporting public health needs, including surveillance, prevention, preparedness, and health promotion

Summary

- Central Dogma of Biology; Each person has their own DNA, transcribed into RNA, translated into proteins
  - Genetic information (DNA, RNA) can be stored as strings of A, T, G, C

- GWAS (Genome Wide Association Study) compares many people and asks, “is there a difference in the DNA of people with/without a certain disease”?

- Local alignment is just altered `minimum_mewtations`, aligns a query string to some reference string
  - Local alignment can help link a RNA to gene

- We can sequence the RNA of single cells (scRNAseq), which allows us to learn many things about individual cells.

- Computational Biology is really cool!

Conclusion

- Bioinformatics is a fast-growing area with lots of exciting opportunities!

- **COMP SCI C176** Algorithms for computational Biology
- Deep dive into some important algorithms in comp bio (phylogeny, HMM, string matching)

- **CS 194-302** Single-Cell Immunology:From Statistic Machine learning to Biomedical Discovery
- More Immunology + Data Science focused class on analysis of Single Cell datasets.

- **CMPBIO C146** Data Science for Biology
- Introductory level class for Biological data manipulation

- **CPH 100** Computational Precision Health 100
- Survey of ML/Data science tools for public health informatics

- **BIO ENG 145** Introduction to Machine Learning for Computational Biology
  - Using machine learning methods for genome-scale experimental data

- **BIO ENG C149** Computational Functional Genomics
- Computational and statistical methods for analyzing genome-scale data, scRNAseq data etc.

- **BIO ENG C131** Introduction to Computational Molecular and Cell Biology
  - Bioinformatics and Computational biology, with an emphasis on alignment, phylogeny, and ontologies