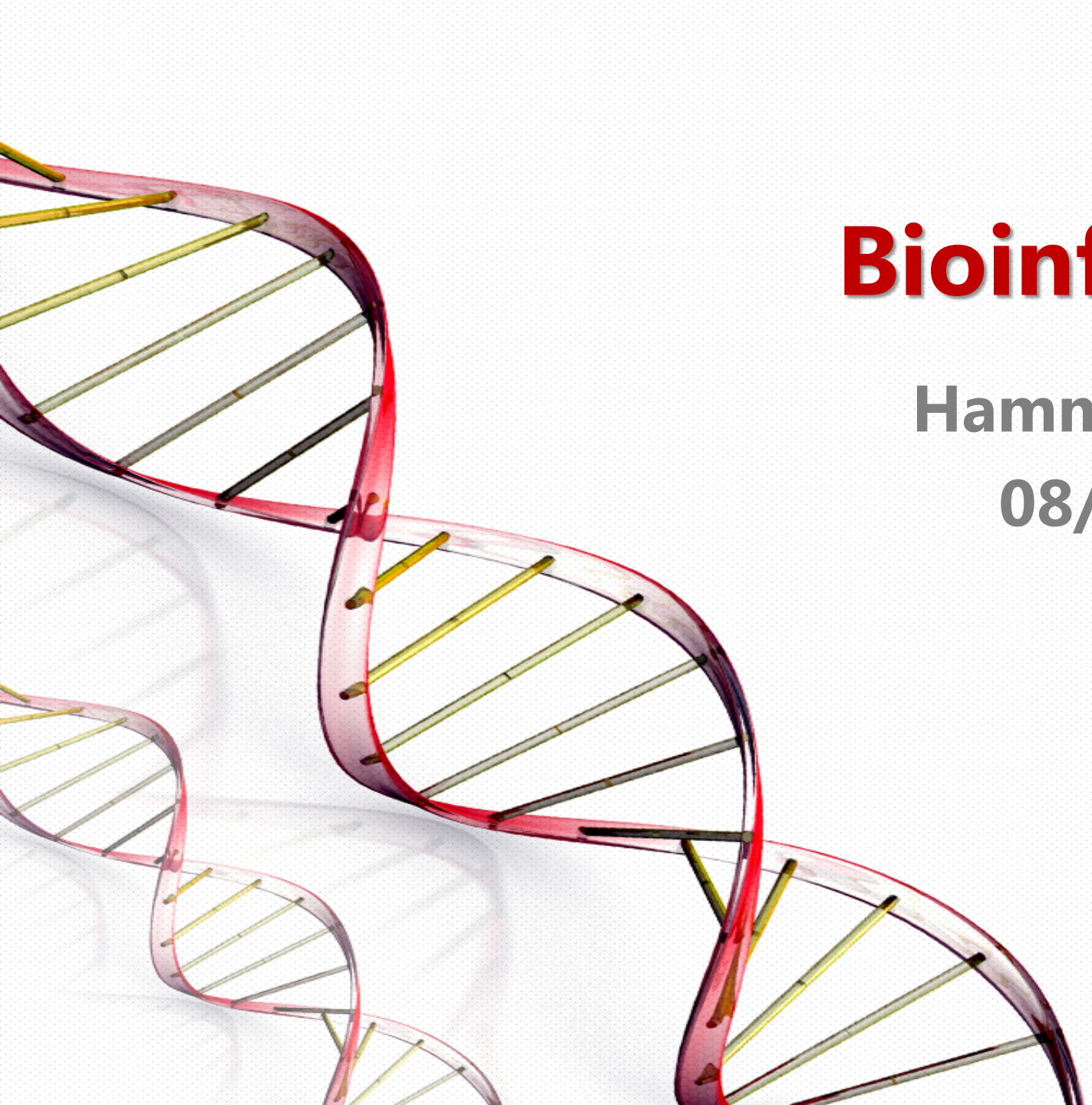


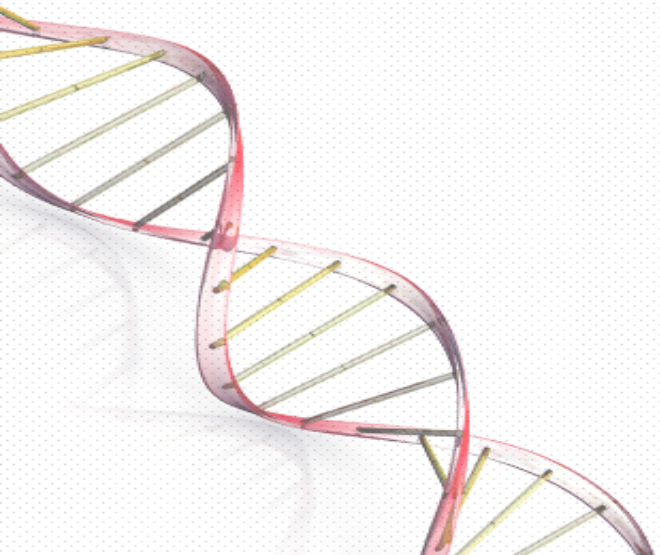
# Bioinformatics

Hammad Naveed

08/21/2023

# Goals for today

- Short survey of interests/background
- Administrivia
- Course Topics



# Your Instructor

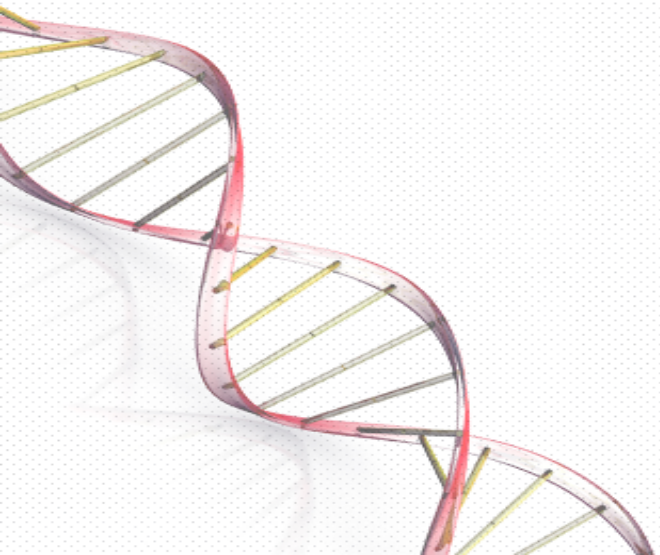
- Email: hammad.naveed@nu.edu.pk
- Office:
- BS(CS) – FAST-NU (2001-2005)
- PhD Bioinformatics – University of Illinois at Chicago (2007-2012)
- Postdocs @ CAS-MPG Partner Institute at Computational Biology and KAUST (2012-2014)
- Research Assistant Professor @ TTIC, University of Chicago (2014-2017)

- Director & Professor in the Department of Computer Science

- Research Interests: Computational modeling of biological systems, heuristics, drug design

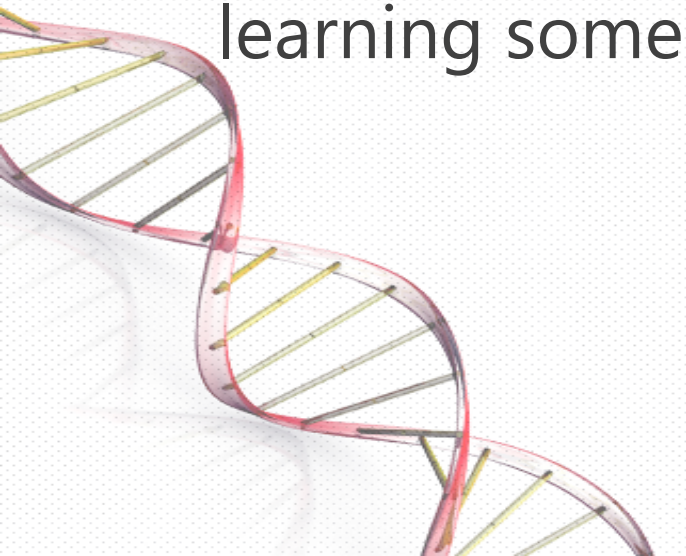
# Office Hours

- Will begin next week
- When I am not taking a class
  - Everyday b/w 11am-1pm
  - Better to email for an appointment
- Course TA: TBA



# Expected Background

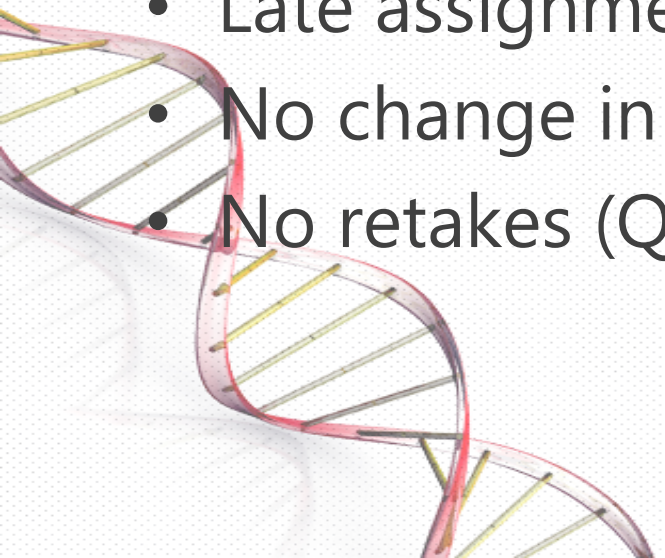
- CS 201 Data Structures
  - Arrays
  - Hash tables
  - Trees
  - Graphs
- Molecular Biology: no knowledge assumed, but an interest in learning some basics is mandatory



# Course Grading

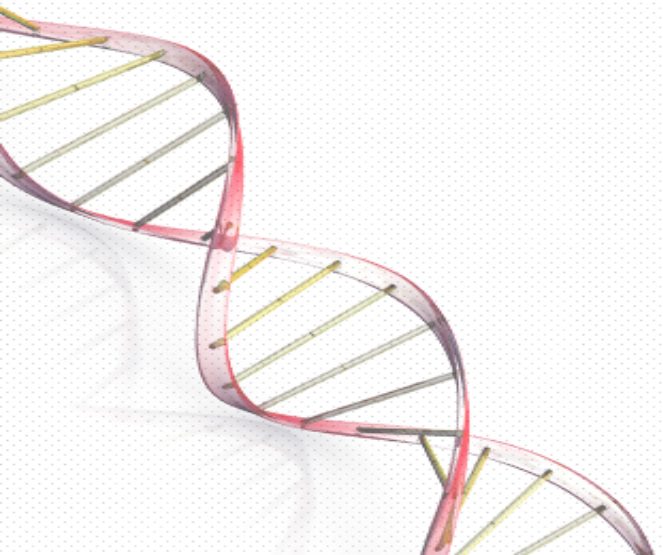
- Quizzes and Assignments: 30%
- Midterm(s) exam: 30%
- Final exam: 40%

- Late assignments will be penalized
- No change in deadlines
- No retakes (Quizzes: Best 5, Assignments: bonus questions)



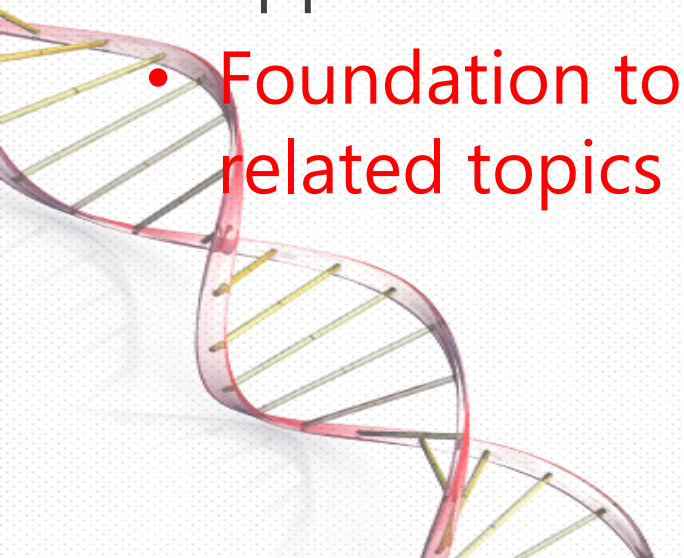
# Plagiarism & Grading

- Copy cases will result in an automatic ‘F’ grade and will be reported to campus disciplinary committee.
- 33% class admitted to copying Fall 2017. This will not be tolerated.

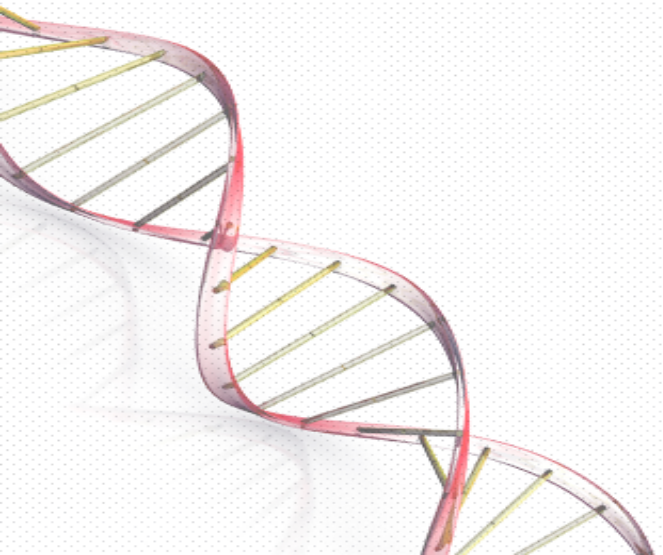


# Learning goals of this class

- Gain an overview of different problem areas in bioinformatics
- Understanding significant & interesting algorithms
- Ability to apply the computational concepts to related problems in biology and other areas
- Ability to understand scientific articles about more cutting-edge approaches
- Foundation to enable independent learning and deeper study of related topics



# Course Feedback



A group of scientists placed 5 monkeys in a cage and in the middle, a ladder with bananas on the top.

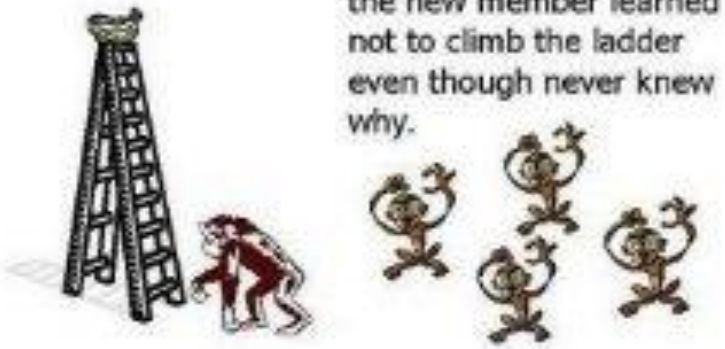
Every time a monkey went up the ladder, the scientists soaked the rest of the monkeys with cold water.

After a while, every time a monkey went up the ladder, the others beat up the one on the ladder.

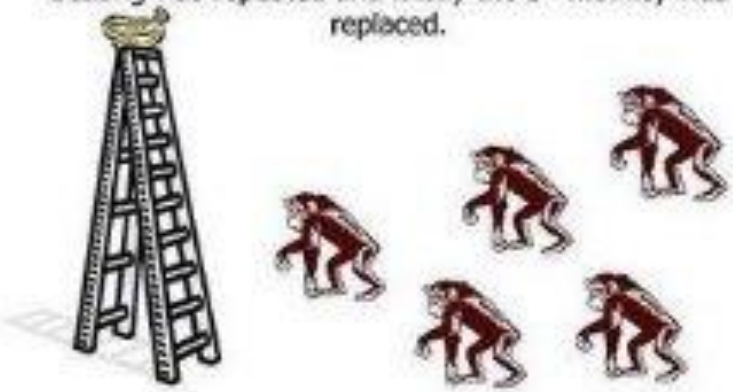
After some time, no monkey dare to go up the ladder regardless of the temptation.

Scientists then decided to substitute one of the monkeys. The 1st thing this new monkey did was to go up the ladder. Immediately the other monkeys beat him up.

After several beatings, the new member learned not to climb the ladder even though never knew why.



A 2nd monkey was substituted and the same occurred. The 1st monkey participated on the beating for the 2nd monkey. A 3rd monkey was changed and the same was repeated (beating). The 4th was substituted and the beating was repeated and finally the 5th monkey was replaced.



What was left was a group of 5 monkeys that even though never received a cold shower, continued to beat up any monkey who attempted to climb the ladder.

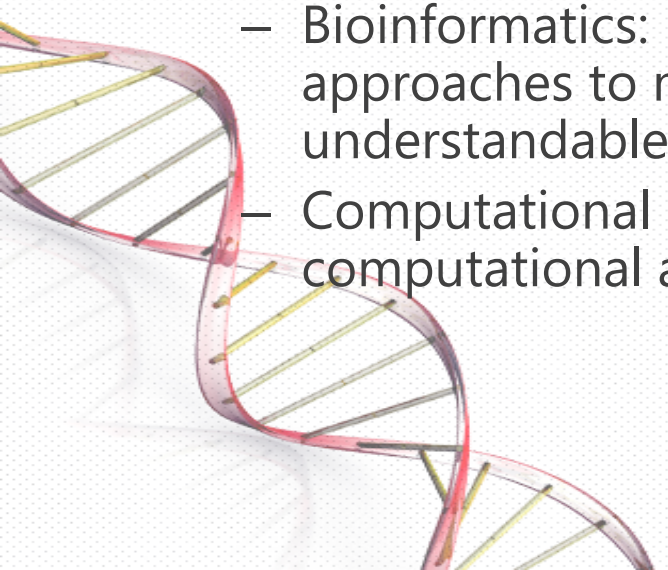
If it was possible to ask the monkeys why they would beat up all those who attempted to go up the ladder.....
I bet you the answer would be....

**"I don't know – that's how things are done around**

Don't miss the opportunity to share this with others as they might be asking themselves why we continue to do what we are doing if

# What is Bioinformatics?

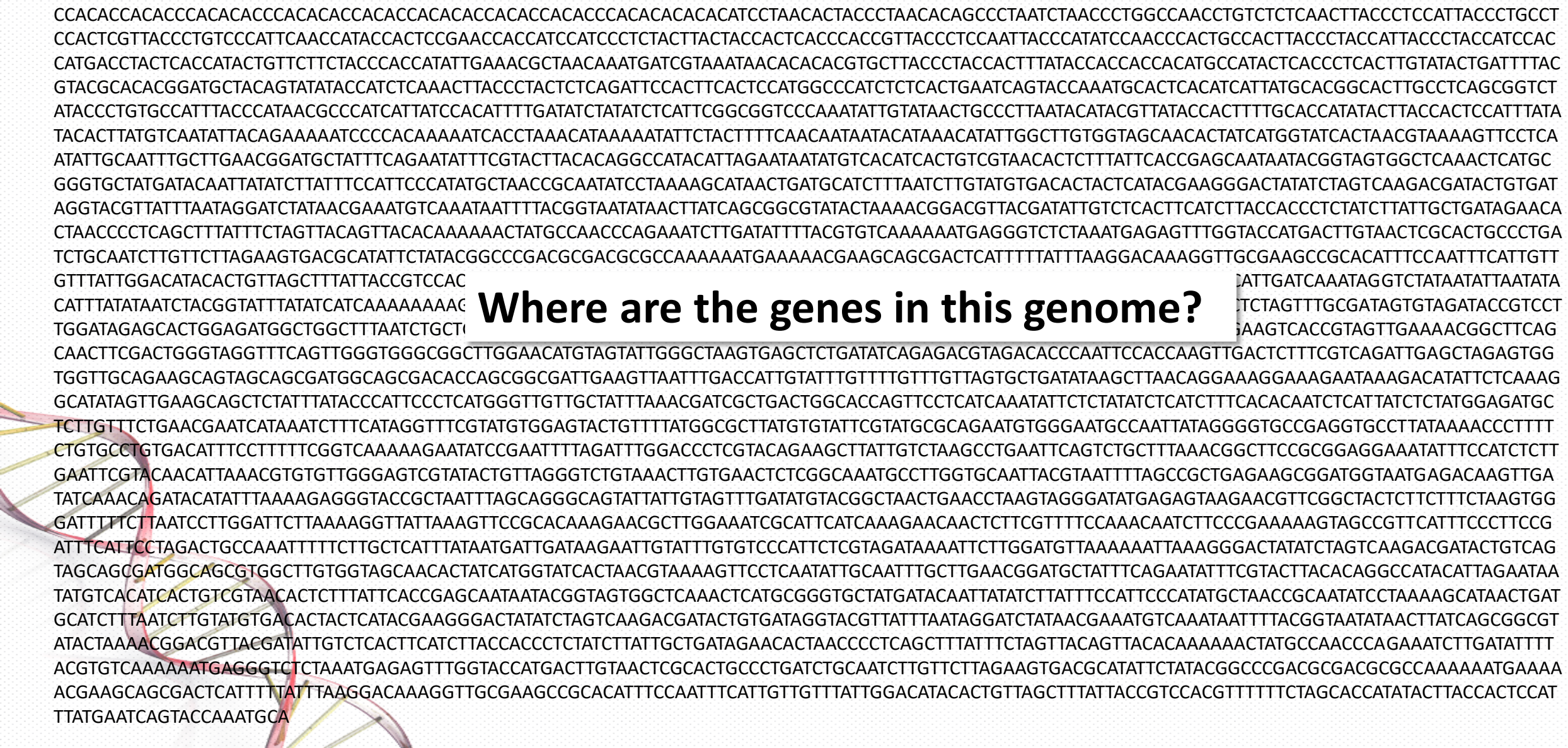
- The term Bioinformatics was coined in the 1970s
- Very close cousin: Computational Biology
- An interdisciplinary field rooted in computer and information sciences and life sciences.
- Draws from other areas such as
  - Math, statistics, machine learning, physics, genetics, evolutionary biology, biochemistry
- Definitions from the National Institute of Health
  - Bioinformatics: Research, development, or application of computational tools and approaches to make the vast, diverse and complex life sciences data more understandable and useful.
  - Computational biology: The development and application of mathematical and computational approaches to address theoretical and experimental questions in biology



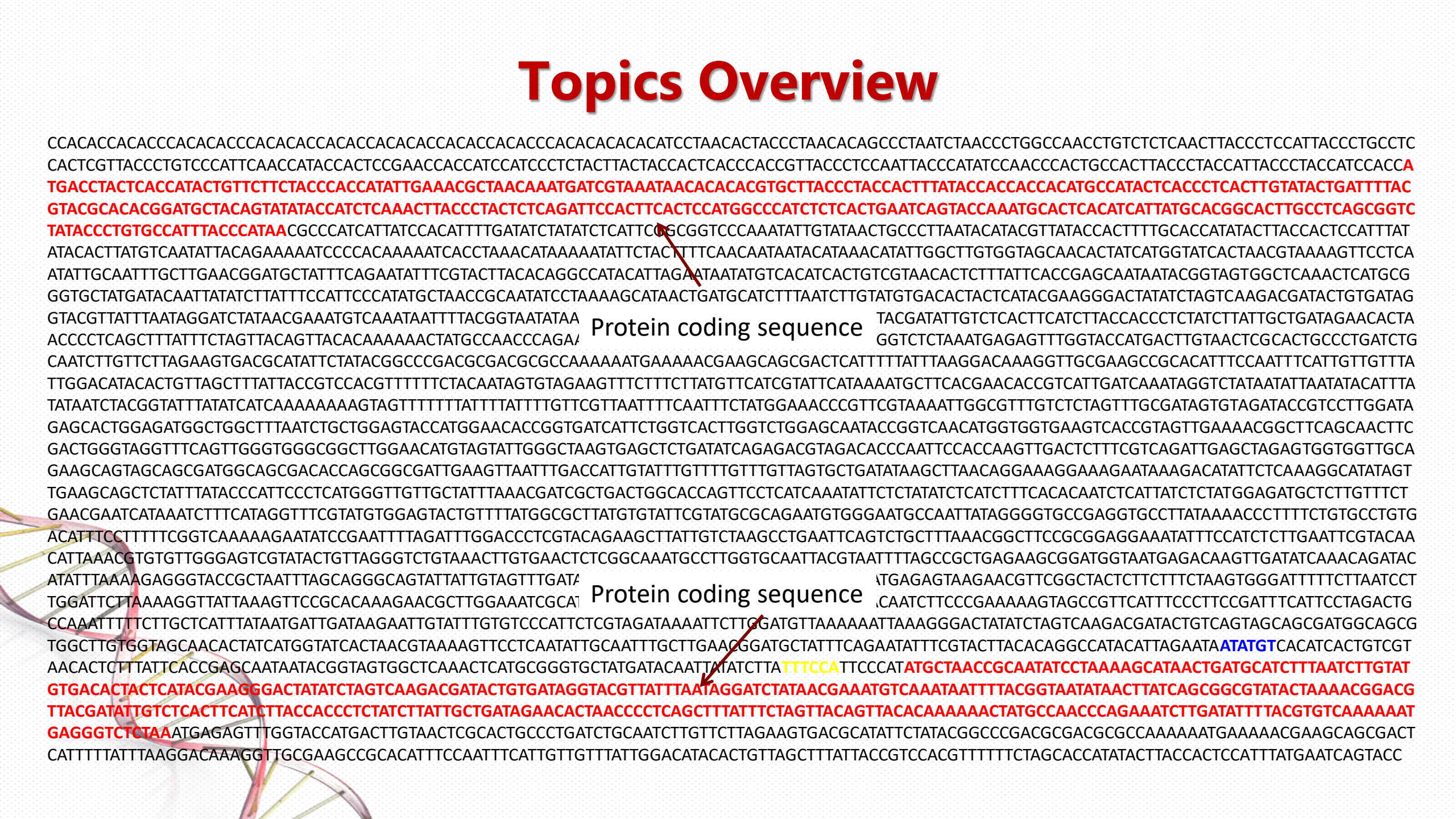
# Topics Overview

- Sequence Assembly

# Topics Overview



# Topics Overview



# Topics Overview

- Sequence Comparison

>gi|224589812:c49547958-49505585 Homo sapiens chromosome 20, GRCh37.p13 Primary Assembly

```
TTGGCTAGGGCAGGTTCGGGGACTCGGTCCGAGGCGCTTGTATTGTCTGCTGAGGGGAGACGCGGGATC
AGCCCCCTCCCCCGCCCGTTGCCGCCGCCGCCTCCGCCGGCCCGGTCTCCCCTCCGCCGCCCGCCGGGAT
CCATGAGCTGAACTCCCGCCCCCCCCCCCCCCGGCCGCCGCCATCTTGTGCCCCCTCCCCCTCCCGCAGGC
AGCGCTAAGGGGGATTTTGGCGTCTTCTCAGACACAGCTCCCTCTCTCGGTCCCCGCTGCTGAGGAGCGA
GAGGAGCCGCGGCCGCCGCCGCCCGCCGCCGGTGAAGCCGCCTCTCCCACACCCTCCGTCTCCTCCCTCCGC
CCCTCCTTTGTCTGCACCGCTCGACGACGCCGCCGCCGCCCGCCGCTCTGCTGCCGCCCGCCGCGCCCCCC
GCAGCACCCGCGGCCGCGGCGCGAGCCGGAGTCCGCCGAGCCGGAGCGCGACGAGGCCCCGGGCGCGCC
CTCCCCGCTGCCGCCACCGCCGTGCCGCCGCCATCCGCCCGCCGCCGCCGCTGTCCGGCCCCCGAGC
ACGCCGGCCCCGCGCGCGCCTCGAGGCCGAGTCAAGGTAAGCCCGGCGCCGCCGCGCGCGGACCCCGGC
CCTGCCCTGCCTGGCTTCCCGACCCCGGCCCGCGCGGCCGCCAGTGCAGCCTGGCCCCGGGCGCGGGCGC
ATTGTTGTTCGCATCGCCGCCCTCCAGCCGCACACACGCTCCTTTGCACACCCCGGCGCCGACGGGTCCC
CCGTCCCGGCACGGACCGCTACCCGACTCTGCTCCGGAGGTCCAGGCGCCTCCTCCTGCAGCCCGACCCC
TTCTCCGCCCTGTGAGCCCACCGAGGCCAGCACTCTGCCTGGCCGCCCGGAGCTTTTGCAGCTAAGGGC
ATCTCCCCCAGCCAGAACCGAGCTTCTCCCATCCTGGAACTAATCCCGTCCGCCCCCCCACCCCCAAAAGG
AGCCCCCCCACGCACACCCCCATGCCTTGAACCAGCCTTCTGCCAGGTCTGCCCACCCTTACCTGGGGGA
GGGGGGACGGGGTTGGGGGGGGCGGGGAGCATGCATGGCATGGAGTGAGTCTTTGCCTCCCCTGTGCAAT
CCTTCTTCCCCAACCTGCCAGCTTTCCTCTTATCCTGACCACAGTTCACCACCATCCCTTTTCCCCAATC
TCAGCTGCCTGGTCAGCACCTCTTTCGAACATTCTGAGCCCCAACTTTGCAGTAAGGTGGTAGAAAATGG
AAATCCAGAGAGCTGCAAAAAGATTTTTTTTTAATGTAATGAATGGGGAGTAGTTTTGTGAACACACTCG
TATTCACCTCTTCTTTCCTCTGGGAGGCCTGCCTGCCTTCCCCTCCCCCCCAACCTTGATCCCTGTCTGC
TGGAAAAGGCGCGATTGCTTCTCTTCCTCTTCTGCCCAAGACCAAGACTCTAGGTCTGATATCCTTTTTA
GCCACACACACACCCCCCACACACGAAGTCTTCTTAATCTTCTTGAATTTACCAGTTACTCTGGACAC
TTCTTGCCATAACCCCATTTCACAGACAAAGGCTTAGGAGGGAAAAAGAACCAAACTGTTTCTGTGTTTG
TCATGTTAATGATGTCCATCCCCACACACCCAGTTCCCATCCCTGATGCTTCTTGTGCCTGTATCTTGAA
CCTTCAGAATGAAAAAAAAAAAAAAATCTCTTGGATAGCTTGATACAGTTTCTTTAGAAATCTGCATTGTTG
CTTCTGGTTTATGGCCATGGAGAAAATACCAGCCCCCACCCAAAATGCACCGCAGCCAAGTCGGCTAAAG
GCGATGAGTGTTGGAGTCAGTAGGGGGCGCATTCTCTGCACTGGGTAAGGCTGCAGAGGGGGGAAGGGCC
CCCGAGCTAGAACAGGATGTGCTTCACAGCGCCTTCTACAGACCATGGGAAAATGGGCCCAGCAAGACAC
ACTCGCACTCTTAGTTGGTGAGCTGACAGCAACATGGCGAAGGATTAAGTGAGTCAAGAGCAGTGATTTT
ACTTCTTTTAAAGAGCCCACCCTGCATTTTAGAAGTTTTCTCCTTTTACTTTTTCAAAACTGTTGGCCCC
CAGTCCATATTTGTTCTTGACGAATCACCTTGGCACATCCTTGGTTCGGATTCCCTTCCTCTGGAGTTCT
TCCACCTCAAGTTTTGAATGTGTCACCATATCTTTCCTAAGTCTGGTGGTAAAGGAGAAGGTTTGAAACA
AAGTAACAGGTTTGGGGGCATAGCTGGTTAACTCCCTTCTTCATAGGCCAAATCTTTCCAATTCTTTGAG
AGTTAAAAATCATTCTTGAGAGATGTTGGATTAAGTGTGTAGATGGGTGTCATGATTTGCAAAGCCTTGT
TGGTAAACATCAAGCCCATGTTGGGTTCTTCATTGTCGAACTCAAGTGTTGATTGTAAAAATTTCGTTTC
CCTCCAAATTTGGACTTGTAATTGCCATTCATTATCTCAGCAGCCAACCACCTTTACAAAAATGTTGATA
CAGTCTTCTGTGAGCACCCACCTTCTTAGGGTGTGTTGATTTCCATAGATTTTACTCCTCTTGTTATTT
CATAGGTGTGAGATGCACAATGCGAAACCTAGGCCCCAGCTTTTACACCATGATGCGCAGGGTTGTACT
TTTTGTACTGAACTGATAGGTGGCCTAGTGGTTATGCCCTGTACTACCATTTTGAGGATCTGGACTCCGT
TCCTGCCTTGCTCTTTGGACCACATTGTCAATTCACACCGGTGGGTATATTCATTTTGGAGGGTGGGAA
CTGCAGCAACAGGGAAGCCAGAGACTTCCTTCTCCTTGGTCACCACCAAAACCAACACTTCCAAAATTTC
AGCCTAGCCACATTCCTCAAGGAACCGGTTAGAGCTGGTTGCAGTAGTGGGTTCTGAACACTTATTAACA
AAGAAATTACTTACAGCTTTTTTAAGAGAAATTTTTACAGATGCTAATATTGTAATCATTTGCATTCTA
ACCCCTTTGTATAAAGTATAAGTATTCAGTTACTATTACTTTGATAACCCAATCTCTAGAGTTGCTAAG
TTCTAAAAGCCTTAAATGTCTGATGAAAACATGGTTTTCCATGAACTTGTTCTTGCTACCAAAGATTGA
AGTGGTTGATGTTTGTGGGTACCTTTAAGACTTTGTATGTTACATAAATACACCAAGATAAGTTTATGG
AGAACTGAATGGGTGTATTCTTTGATACCTAGTTAACCAGTTGGTGCTTAACAGGAGAATTTTTCATTGG
AGAATTGCTACCAGGTAACATTTTGGTGAGTGAAAAGGAGGCCCTTTGGTACAGGTATTTGACTAGAGAC
ATTGCTTCTACAATTTCTTAGTCATTGTGAAAAATCACATATTTAAATGGGATTATTTAAATTTTTAAAT
ACTGTATAGTATTTGTAGGCAGGTATAGTTTTTATCTTTATTTGGTGTTTTATATGCAGTTATTTAAGTT
CTGAAAATGACCATATAATTTGAAGACAACTTATTTGACGGAATGTTTGAAATGAAGTTTGTTGAGGGGC
```

Human ADNP gene

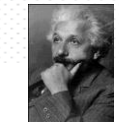
>gi|90093348|ref|NM_009628.2| Mus musculus activity-dependent neuroprotective protein (Adnp), mRNA

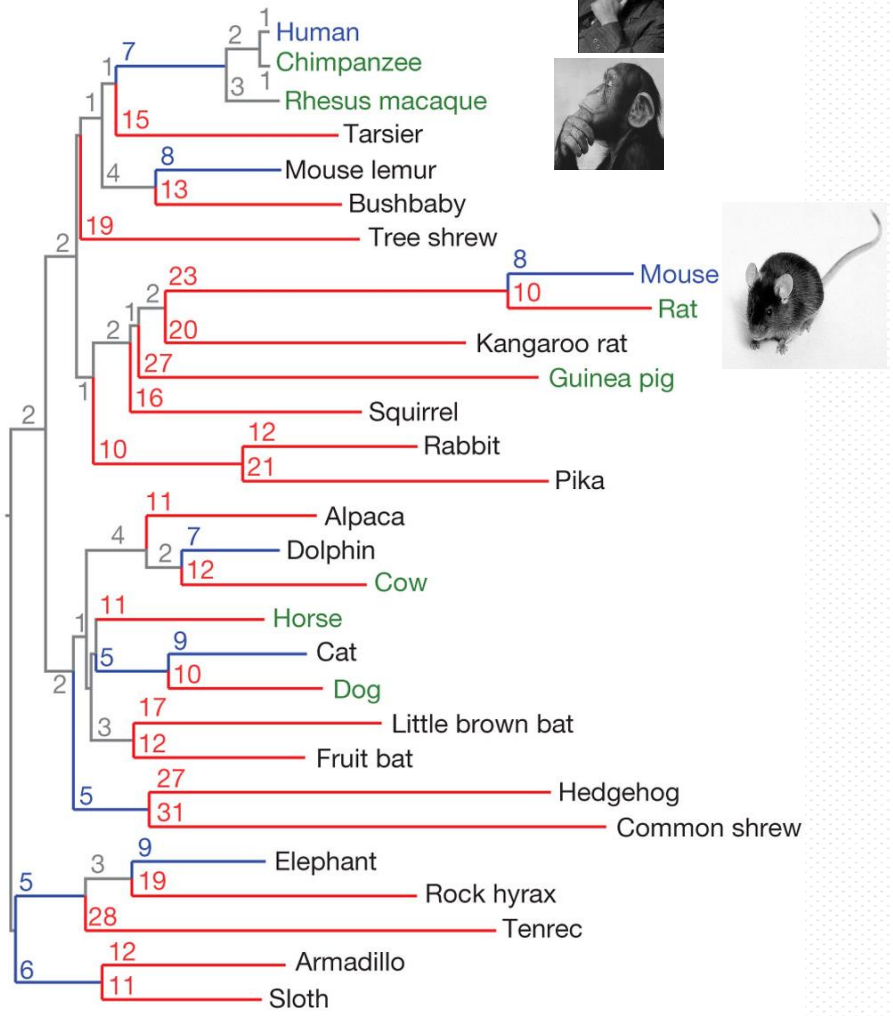
```
CTGAGGGGAGACGCGGGACCAGCCCCCTCCCCCGCCCGCCGCCGCCGCCGCCACCCGTCCGGTCTACCCG
CCGCCGCCGGGACCCATGAGCTGAGCCCCGCCCCGCCCCATCTTGGCCCCCTCCCCCTCCAGCCGCGCT
CGCTCGGGGGGGGATTTGGCGTCGCCTCAGCCACAGCTCCCTCTCTCGGTCTCCGCCGCCGAGGAGCCGCG
CGGCCGTCGCCGCCCCGCCTGCCCGCCCGCGACGGTGAAGCCGCGGCCCCACGCCCTCCGCCGCCCTCCGC
CCCTCCTTTGTCTGCACTGCCCGCCGCCGCCCGCCCGCCGCTCTGCTGCCGCCCGCCGCGCCCCCCGGCAGCA
GCCGCGGCCGCGGCTCGAGCCGGAGCCCGAGCGCGACGAGGCCCCGGGCGCGTCCTCGCCGCCATCGCCG
TGCCGCCGTCCACCCGCCGCCGTGCTCTAGAGCACGCCGGCCCCCGCGCACGCCTCGAGGCCGAGCCAAGA
AACTATGTTCCAACTTCCTGTCAACAATCTTGGCAGTTTAAGAAAAGCCCGGAAAACTGTGAAAAAAATA
CTTAGTGACATTGGGTTGGAATACTGTAAAGAACATATAGAAGATTTTAAACAGTTTGAACCTAATGACT
TTTATTTGAAAAACACTACATGGCAGGATGTAGGACTGTGGGACCCTTCTCTTACGAAAAATCAGGACTA
TCGGACAAAACCTTTTTGCTGCAGTGCTTGTCCGTTTTCCTCAAAATTCTTCTCTGCCTACAAAAGTCAT
TTCCGGAATGTCCATAGTGAAGACTTTGAAAATAGGATTCTCCTTAACTGCCCTTACTGTACCTTCAATG
CAGATAAAAAGACTTTGGAAACACACATTAAAATATTTCATGCTCCAAACTCCAGCGCACCAAGTAGCAG
CCTCAGCACTTTCAAAGATAAAAACAAAAACGATGGCCTTAAACCTAAGCAGGCTGACAATGTAGAGCAA
GCCGTGTATTACTGCAAGAAGTGCACTTACCGAGACCCTCTCTACGAGATCGTCAGGAAGCACATCTACA
GGGAACATTTTCAACACGTGGCAGCACCCTACATAGCAAAAGCAGGAGAAAAATCACTCAATGGTGCAGT
CTCCCTGGGCACAAATGCCCGAGAGGAGTGTAACATCCACTGCAAGCGATGCCTTTTCATGCCCAAGTCC
TATGAAGCTTTGGTACAGCATGTCATTGAGGACCATGAACGGATAGGCTATCAGGTCACTGCCATGATCG
GACACACAAATGTTGTAGTTCCCCGAGCCAAGCCCTTGATGCTGATAGCTCCCAAACCTCAAGACAAAAA
GGGCATGGGACTCCCACCACGAATCAGCTCCCTTGCTTCTGGAAATGTCCGGTCGTTGCCATCACAGCAG
ATGGTAAACCGATTGTCAATACCAAAGCCCAACTTAAATTCAACGGGAGTCAACATGATGTCCAATGTTC
ACCTGCAGCAAAACAACTATGGAGTCAAATCTGTGGGCCAGAGCTATGGTGTTGGCCAGTCAGTGAGGCT
GGGACTAGGTGGCAATGCTCCAGTTTCCATCCCTCAACAGTCTCAGTCCGTGAAACAGTTACTTCCAAGT
GGGAATGGGAGGTCTTTTGGGCTAGGTGCTGAGCAGAGGCCCCCAGCAGCAGCCAGGTACTCCCTGCAGA
CTGCCAACACCTCTCTACCCCCAGGCCAAGTGAAGTCTCCCTCTGTGTCTCAGTCACAGGCATCTAGAGT
ATTAGGTCAGTCCAGTTCTAAACCTCCACCAGCCGCCACAGGCCCTCCTCCAAGCAACCACTGTGCCACT
CAGAAGTGGAAAATCTGTACAATCTGTAACGAGCTTTTCCCTGAGAATGTCTATAGCGTTCACTTCGAAA
AGGAGCATAAAGCTGAGAAAGTCCCAGCCGTAGCTAACTACATTATGAAAATACACAATTTTACTAGCAA
ATGCCCTCTACTGTAATCGCTATTTGCCTACAGATACCCTACTCAACCATATGTTAATTCATGGTCTGTCT
TGTCCGTATTGCCGTTCCACCTTCAATGATGTAGAGAAGATGGCAGCACACATGCGAATGGTTCATATTG
ATGAAGAGATGGGGCCTAAAACGGATTCTACTTTGAGCTTTGATTTGACATTGCAACAGGGCAGTCACAC
CAACATTCATCTCCTGGTGACCACATACAACCTGAGGGATGCCCCCGGCTGAATCAGTTGCTTACCATGCC
CAAAATAATGCCCCAGTTCCTCCAAAGCCACAACCAAAAGTTCAGGAAAAAGCAGATGTCCCGGTTAAAA
GTTCACCTCAAGCTGCAGTGCCCTATAAAAAAGATGTTGGGAAGACCCTTTGCCCTCTTTGCTTTTCAAT
ACTAAAAGGACCCATATCTGATGCACTTGCACATCATTTACGAGAAAGACACCAAGTTATTCAGACAGTT
CATCCGGTTGAGAAAAAGCTAACTTACAAATGTATCCATTGCCTTGGTGTGTATACTAGCAACATGACAG
CCTCAACCATCACTCTGCATCTAGTCCACTGCAGGGGTGTTGGAAAAACCCAGAATGGCCAGGACAAGAC
AAACGCACCTTCTCGGCTCAATCAGTCTCCAGGCCTGGCCCCCTGTGAAGCGCACGTATGAGCAGATGGAG
TTTCCACTGCTAAAAAAGCGGAAGCTGGAGGAGGATGCTGATTCCCCTAGCTGCTTTGAAGAGAAGCCAG
AAGAGCCTGTTGTTTTAGCTTTAGACCCCAAGGGTCATGAAGATGATTCTTATGAGGCTAGGAAAAGCTT
TCTCACAAAGTACTTCAACAAACAGCCCTATCCCACCAGGAGAGAAATTGAGAAGTTAGCTGCCAGTCTA
TGGCTATGGAAGAGTGACATTGCCTCCCATTTCAGTAACAAGAGGAAGAAGTGTGTCCGCGACTGTGAAA
AGTACAAGCCTGGTGTGCTGCTAGGTTTTAACATGAAAGAATTAAATAAAGTCAAACACGAGATGGATTT
TGATGCTGAGTGGCTGTTTGAAAATCACGATGAGAAAGACTCAAGAGTCAATGCTAGCAAGACTGTTGAC
AAAAAGCATAACCTTGGGAAAGAAGATGATAGCTTCTCAGATAGTTTTGAACATTTGGAAGAAGAATCCA
ATGGAAGCGGGAGTCCTTTTGACCCTGTCTTTGAAGTTGAGCCTAAAATTCCCAGTGATAATTTAGAGGA
GCCTGTACCGAAGGTTATTCCGGAAGGTGCTTTGGAATCTGAGAAGCTAGACCAAAAAGAGGAGGAGGAG
GAGGAGGAGGAGGAGGATGGTTCAAAATATGAAACTATCCATTTGACTGAGGAACCAGCCAAATTAATGC
ATGATGCCTCTGATAGTGAGGTAGACCAAGATGATGTAGTTGAAAGATGGTGCTTCACCATCTGA
GAGTGGGCCTGGTTCCCAACAAATCTCAGACTTTGAGGATAATACATGTGAAATGAAACCAGGAACCTGG
TCTGATGAGTCTTCCCAGAGTGAAGATGCAAGGAGCAGTAAGCCAGCTGCCAAAAAAAAGGCTACAGTGC
```

Mouse ADNP gene

# Topics Overview

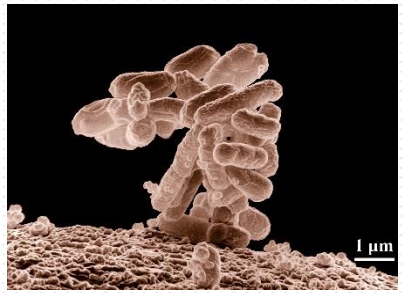
- How are these organisms related?



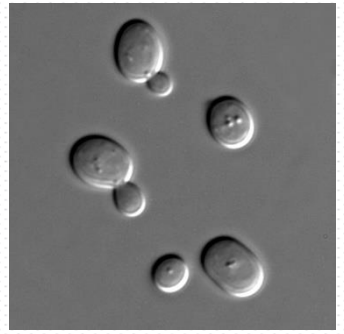





E Coli

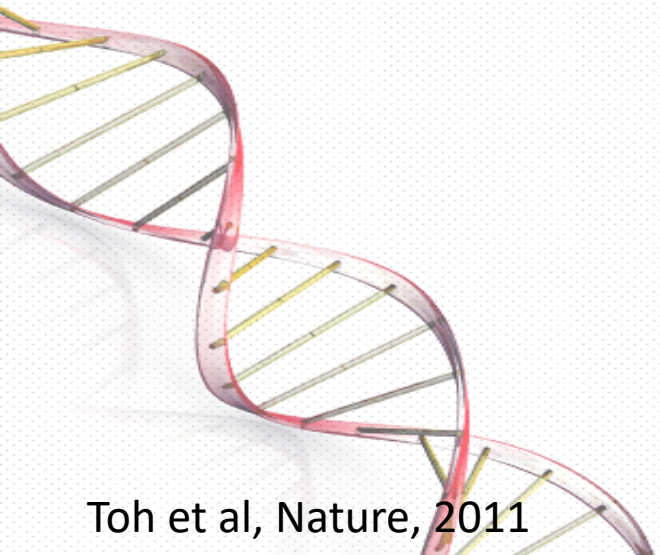



Yeast



Drosophilla

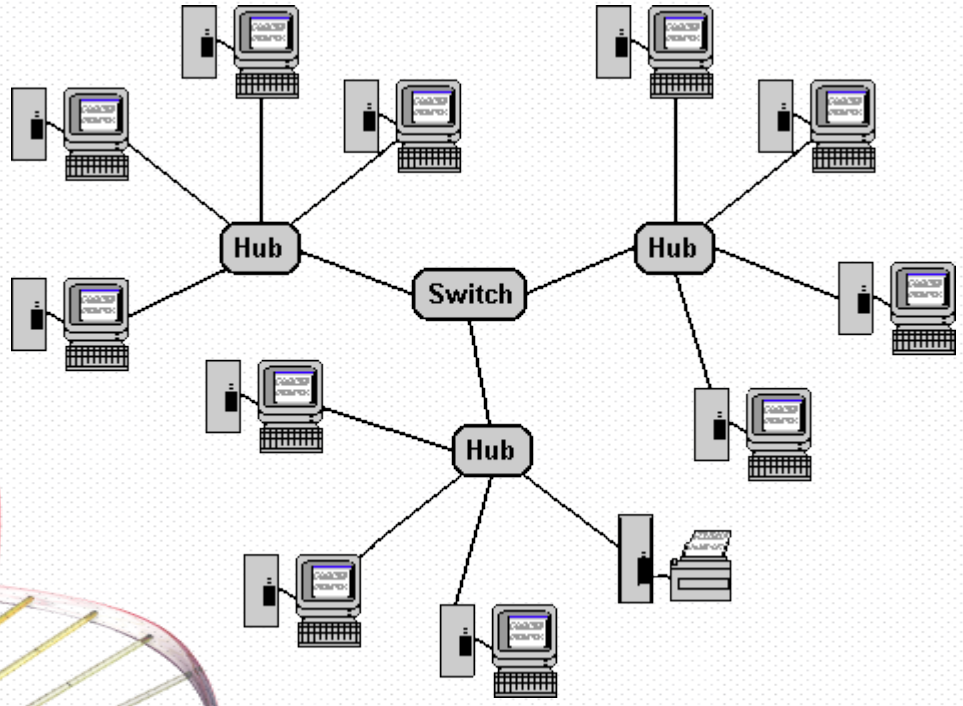
Lab mice

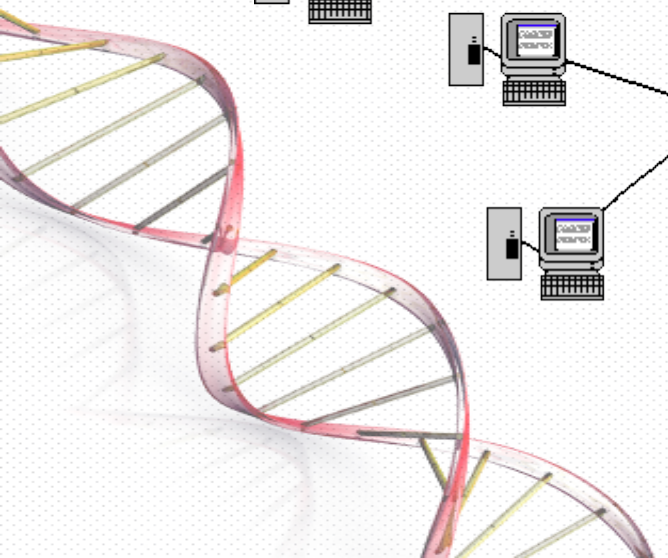


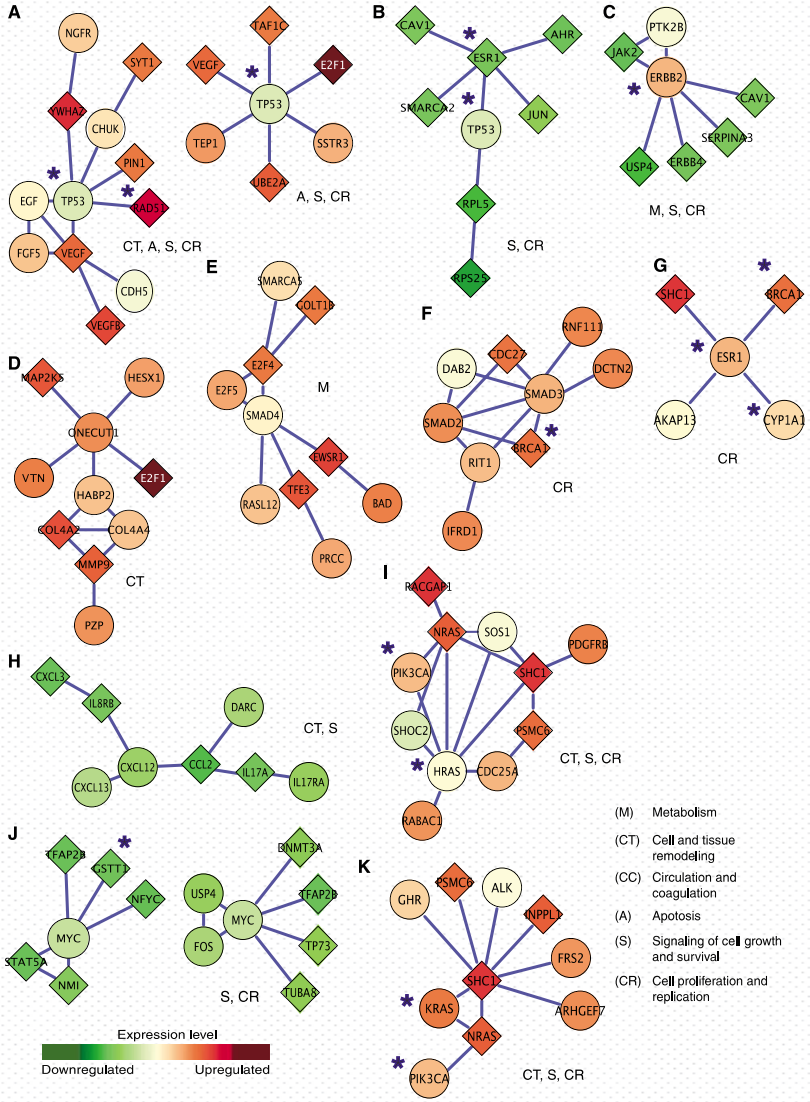
Toh et al, Nature, 2011

# Topics Overview

- Biological Networks

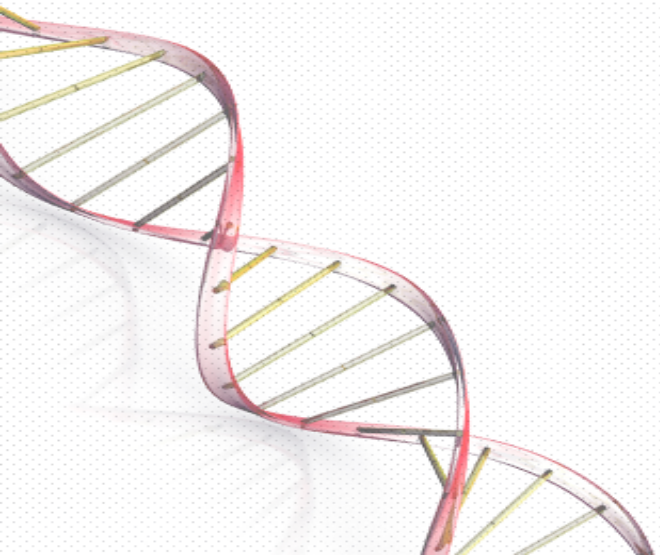




# Topics Overview

- AI in healthcare



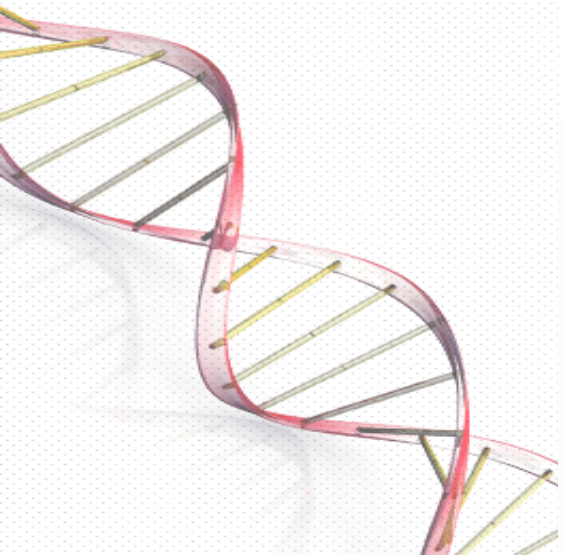
# Human Genome Project

**1990**

The Human Genome Project is launched in the USA, directed by James Watson. The initial target completion date is 2005.

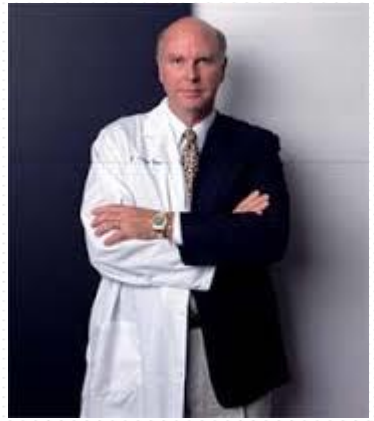
- Several Administrative issues slows down the progress

# Human Genome Project



**1998**

Celera Genomics launches a private venture to finish sequencing the human genome in three years.

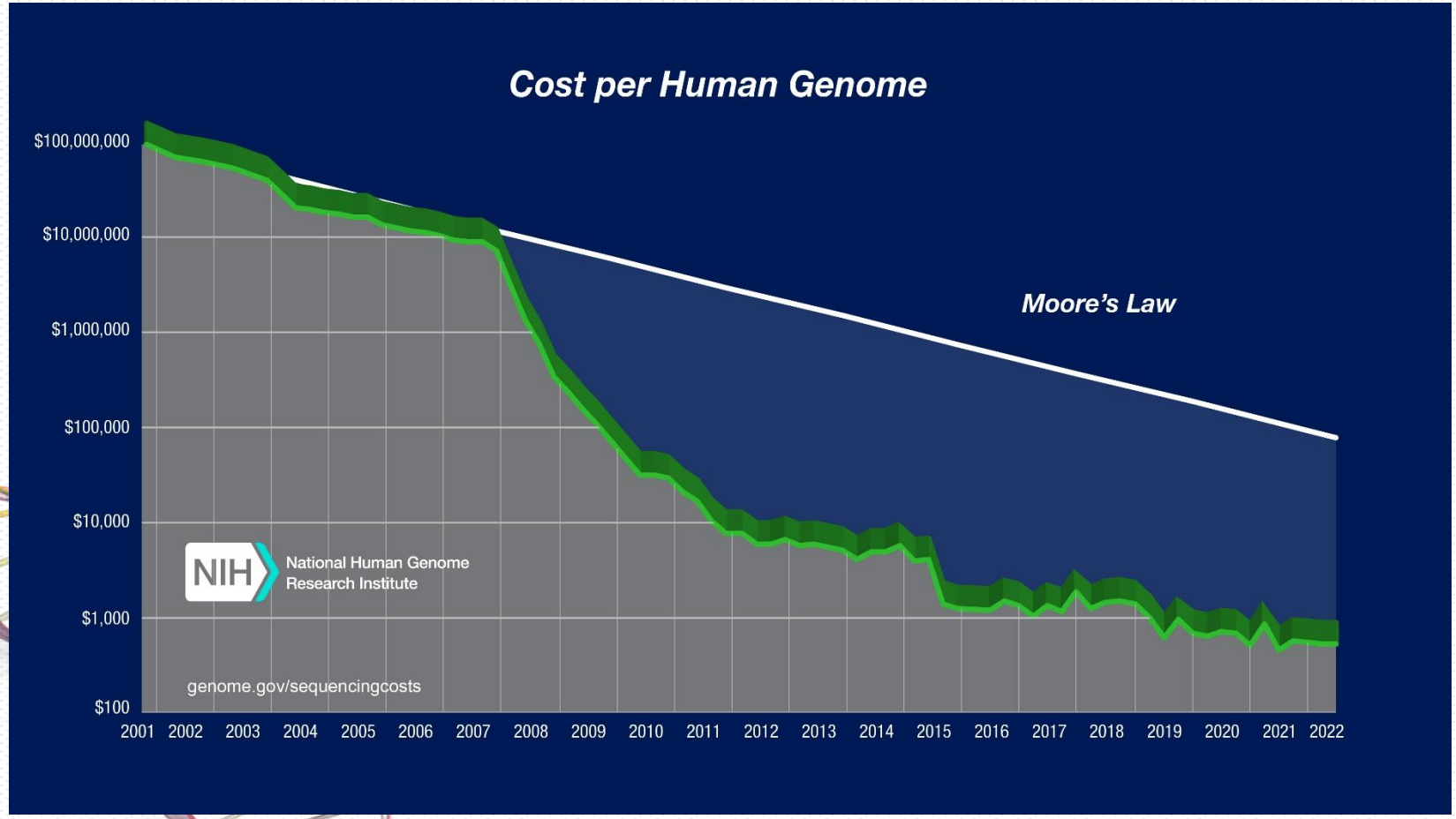


The Wellcome Trust releases more funding allowing the Sanger Centre to raise their contribution to the Human Genome Project from sequencing one-sixth of the human genome to sequencing one-third.

**2000**

The University of California, Santa Cruz (UCSC) launches its Human Genome Browser.

# Human Genome Project




- First human genome cost $2.7 billion
- Now it cost <$1000
- Projected to cost ~$100 in 3-4 years
- Covered by insurance companies and will become a routine procedure
- Opportunity for off-shore analysis

# THE PRECISION MEDICINE INITIATIVE®

## WHAT IS IT?

**Precision medicine** is an emerging approach for disease prevention and treatment that takes into account people's individual variations in genes, environment, and lifestyle.

The Precision Medicine Initiative® will generate the scientific evidence needed to **move the concept of precision medicine into clinical practice.**

# THE PRECISION MEDICINE INITIATIVE®

## WHY NOW?

The **time is right** because of:

Sequencing of the human genome

Improved technologies for biomedical analysis

New tools for using large datasets

# Does it end here????????

- ATCTGCGCGGTTTCGCGGCTGCGGAAAAGCGGGCT