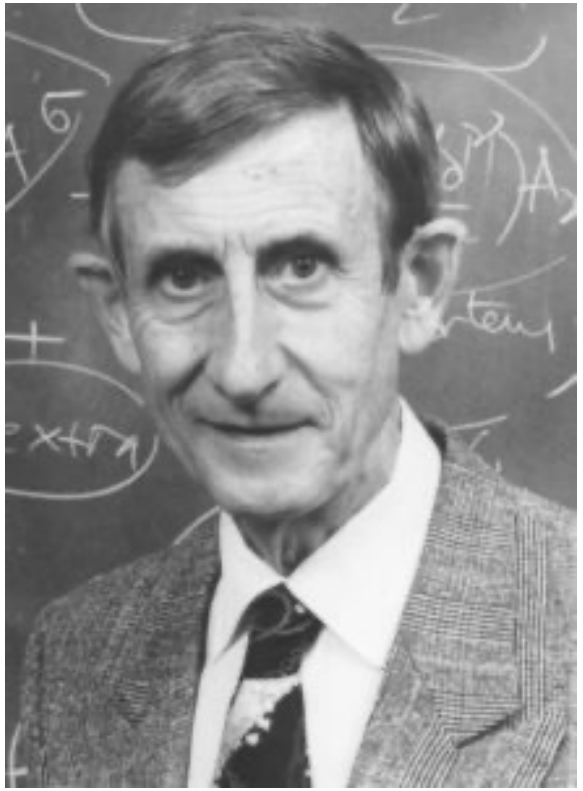
My Bioinformatics Journey

Chongzhi Zang
Center for Public Health Genomics
zang@virginia.edu
zanglab.org

BME 1501 Guest Lecture
November 8, 2021

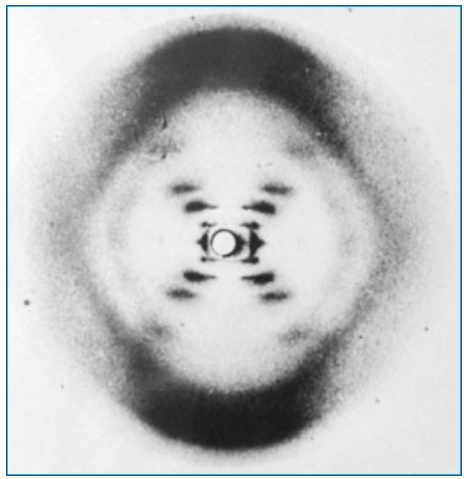


- “It has become part of the accepted wisdom to say that the 20th century was the century of **physics** and the 21st century will be the century of **biology**.”

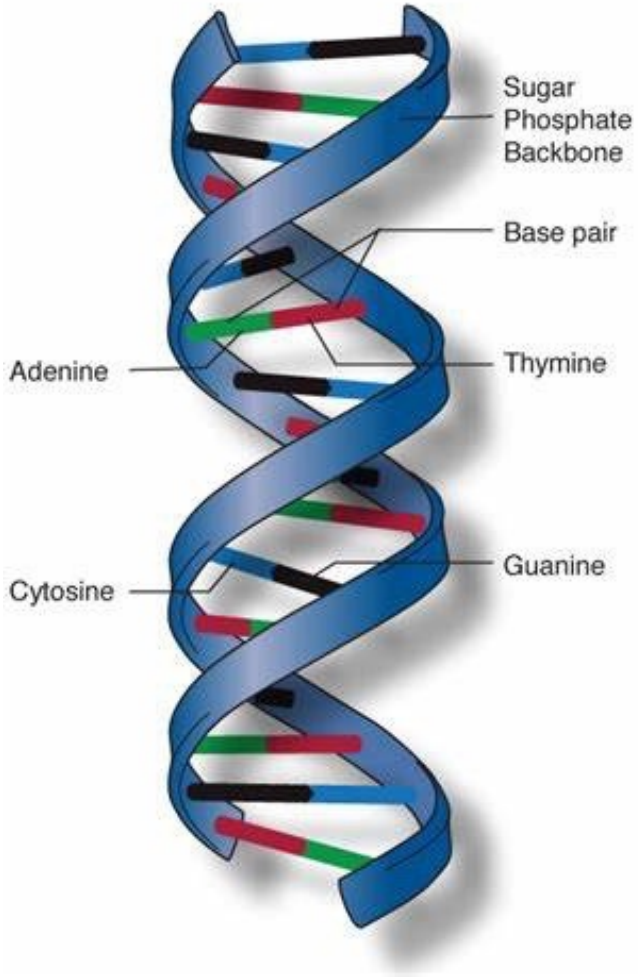
– Freeman Dyson (1923-2020)

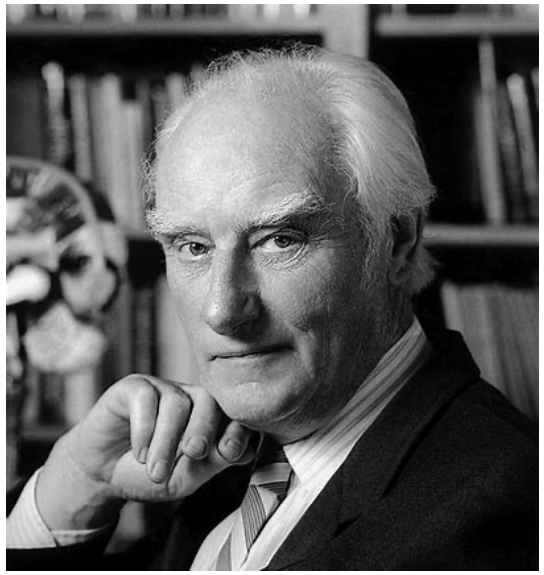
Start of Bioinformatics

- Double helix structure of DNA (1953)

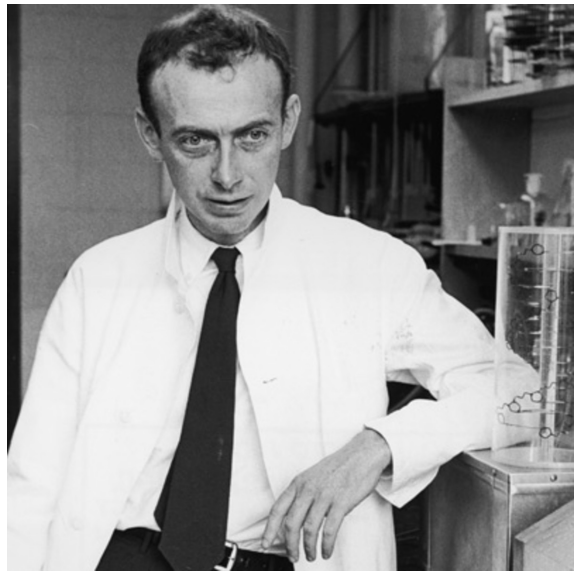


Rosalind Franklin
Maurice Wilkins



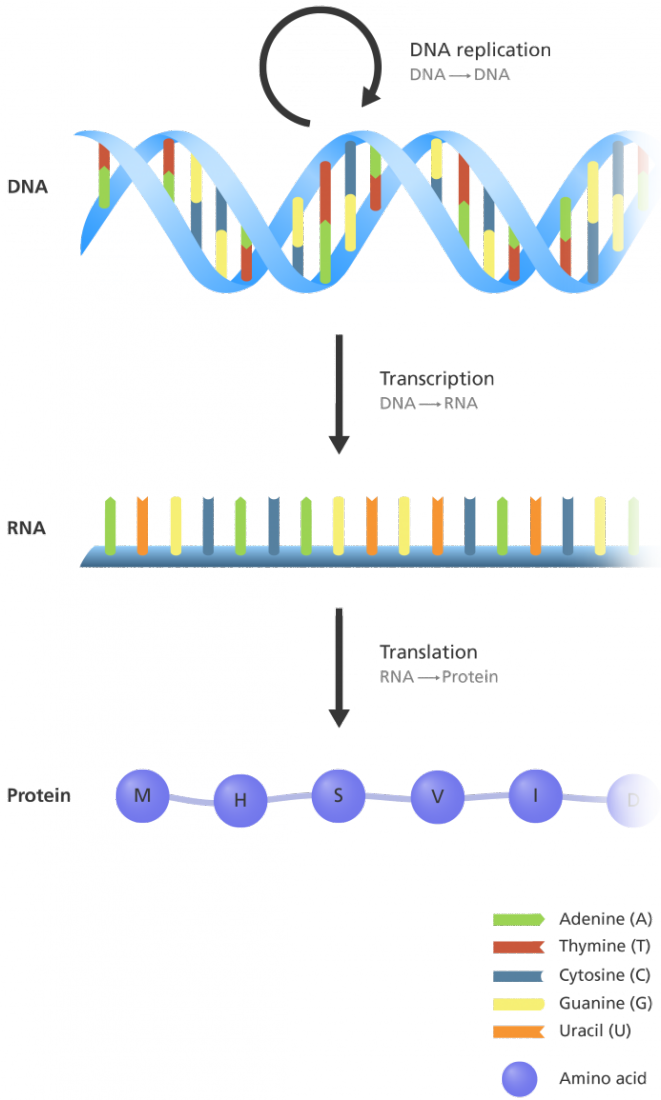


Francis Crick



James Watson

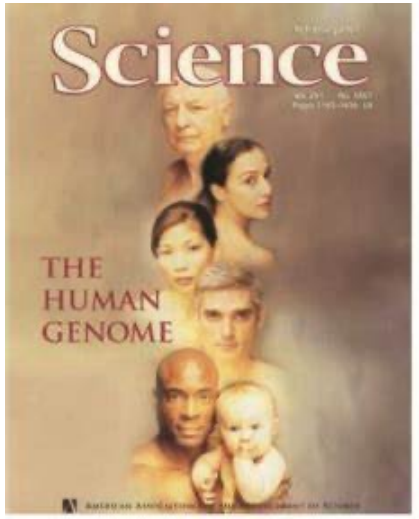
Central Dogma of Molecular Biology

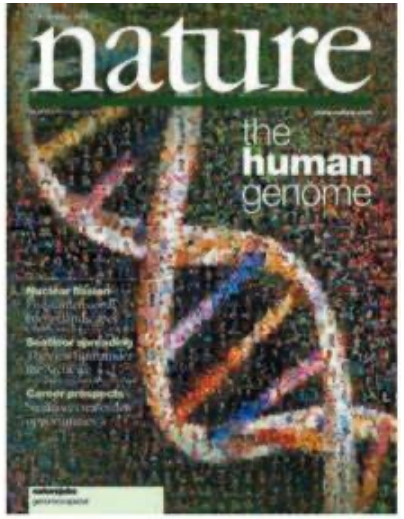


GCATCCATCTTGGGGCGTCCCAATTGCTGAGTAACAAATGAGACGCTGTGGCCAAACTCAGTCATAACTAATGACATTT
CTAGACAAAGTGACTTCAGATTTTCAAAGCGTACCCTGTTTACATCATTTGCCAATTTCGCGTACTGCAACCGGCGGGC
CACGCCCCCGTGAAAAGAAGGTTGTTTTCTCCACATTTCGGGGTTCTGGACGTTTCCCGGCTCGGGGCGGGGGGAGT
CTCCGGCGCACGCGGCCCTTGGCCCCGCCCCCAGTCATTCCCGGCCACTCGCGACCCGAGGCTGCCGCAGGGGGC
GGGCTGAGCGCGTGCGAGGCGATTGGTTTGGGGCCAGAGTGGGCGAGGCGCGGAGGTCTGGCCTATAAAGAGTCGC
GGAGACGGGGTGCGGTTTGCGTCGTAGTCTCCTGCAGCGTCTGGGGTTTCCGTTGCAGTCCTCGGAACCAGGACCTC
GGCGTGGCCTAGCGAGTTATGGCGACGAAGGCCGTGTGCGTGCGAAGGGCGACGGCCCAGTGCAGGGCATCATCAA
TTTCGAGCAGAAGGCAAGGGCTGGGACGGAGGCTTGTTGCGAGGCCGCTCCCACCCGCTCGTCCCCCCGCGCACCT
TTGCTAGGAGCGGGTCGCCGCCAGGCCTCGGGGCCGCCCTGGTCCAGCGCCCGGTCCCGGCCCGTGCCGCCCGGT
CGGTGCCTTCGCCCCAGCGGTGCGGTGCCCAAGTGCTGAGTCACCGGGCGGGCCCGGGCGCGGGGCGTGGACCGA
GGCCGCCGCGGGGCTGGGCCTGCGCGTGGCGGGAGCGCGGGGAGGGATTGCCGCGGGCCGGGGAGGGGCGGGG
GCGGGCGTGCTGCCCTCTGTGGTCCTTGGGCCGCCGCCGCGGGTCTGCGTGGTGCCTGGAGCGGCTGTGCTCGTCC
CTTGCTTGGCCGTGTTCTC

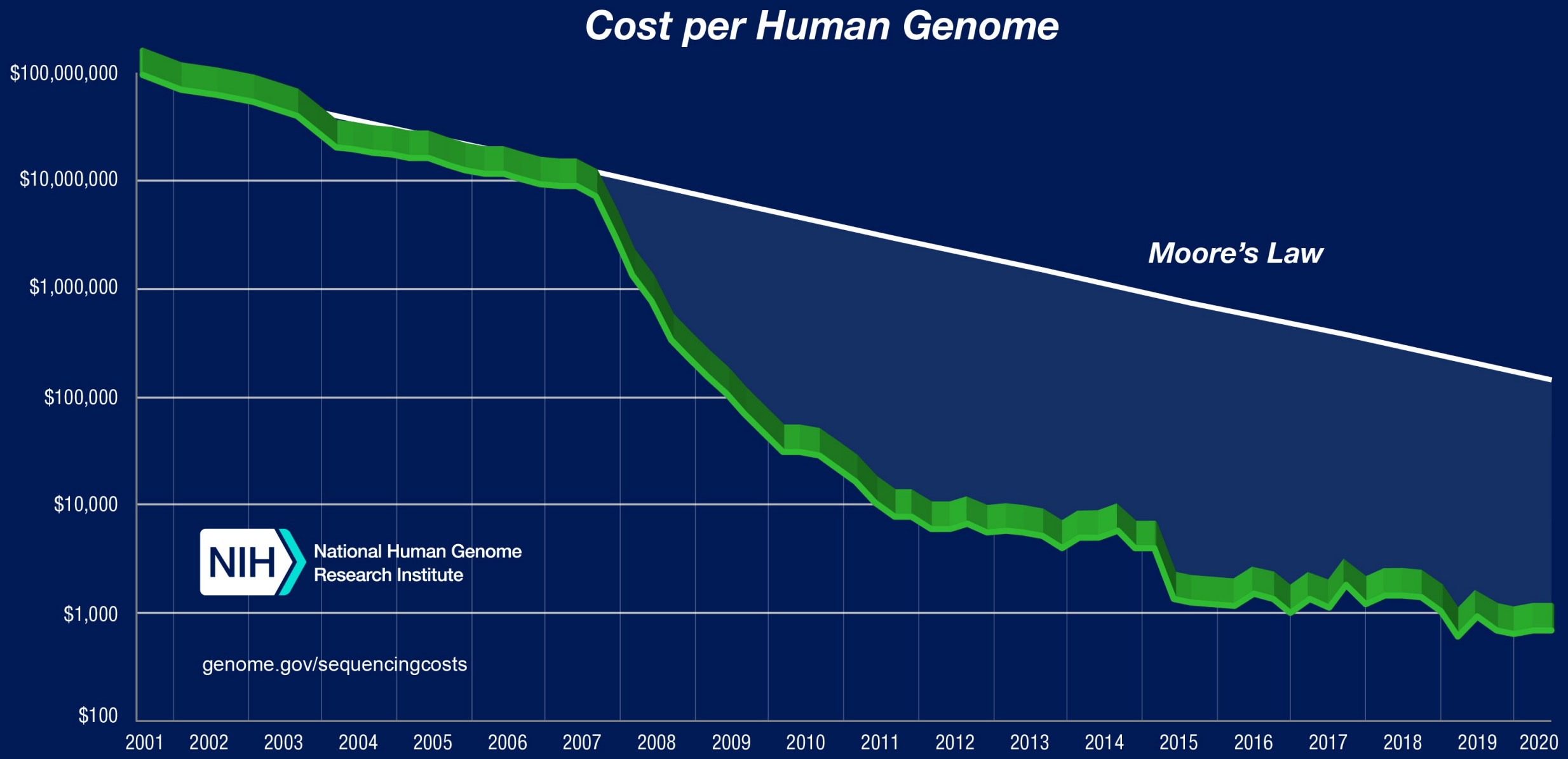
Human Genome Project

- One of the biggest science projects in history
- 13 years, 7 countries, $3 billion: 1 genome
- 3 billion base pairs, 20 thousand protein-coding genes

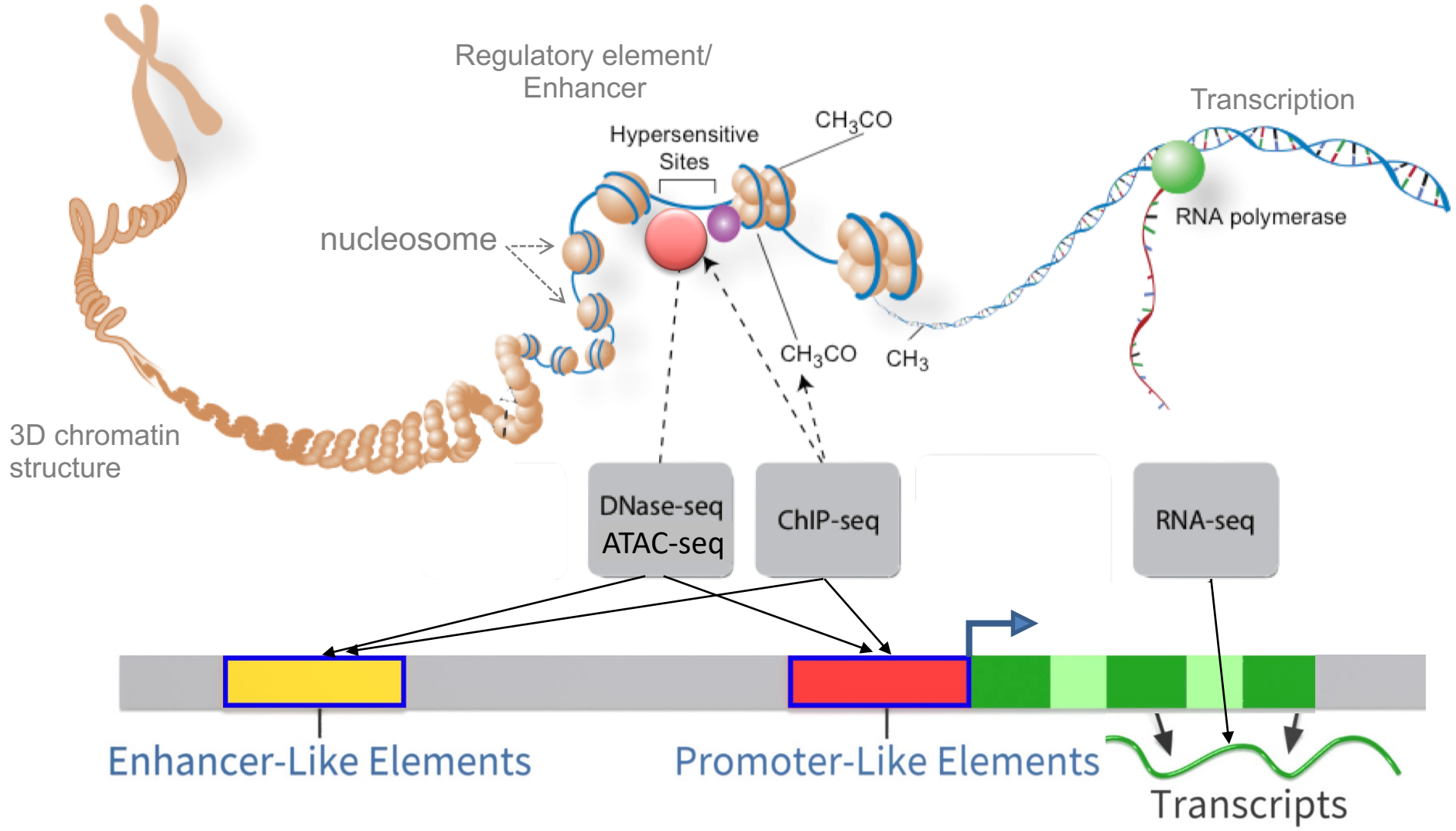





Microhabitats save mammals, but not birds, from warming pp. 553 & 633
Gut microbiota modulate immunotherapy pp. 573, 595, & 602
Physically distanced quantum gates pp. 576 & 614
Science
$15
5 FEBRUARY 2021
sciencemag.org
AAAS
SPECIAL ISSUE
HUMAN GENOME AT
20


Cost per Human Genome
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
$1,000
$100
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts

High-throughput sequencing helps study all elements in the genome, epigenome, and their functions

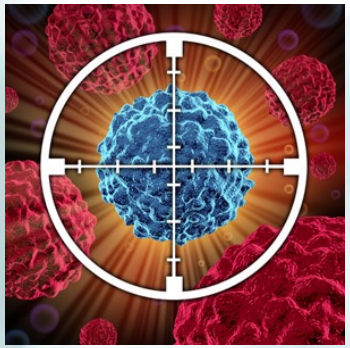


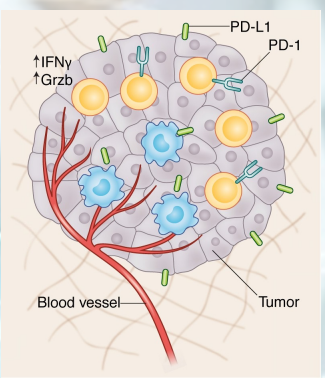
Modified from ENCODE, Darryl Leja (NHGRI), Ian Dunham (EBI)

- “Until we actually understand all the working parts within our **genome**, we won't really be able to practice the most informed **medicine**.”

- Eric Lander

A long journey to fight cancer



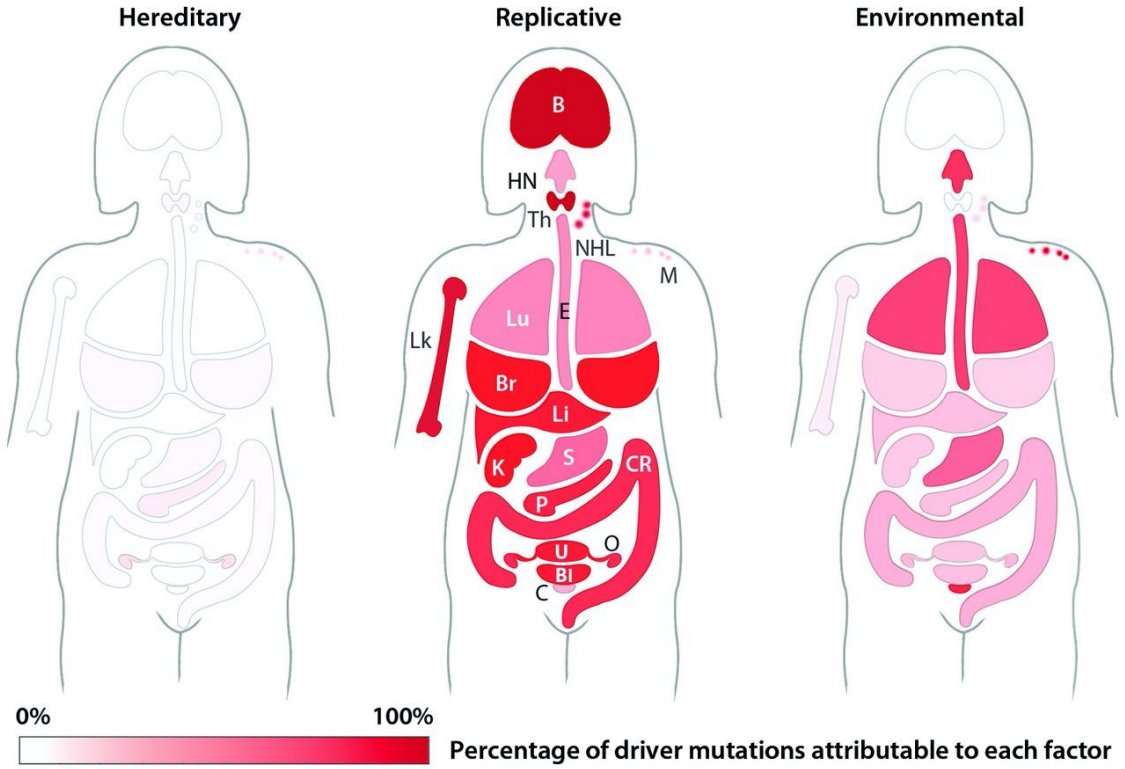


Surgery → Chemotherapy/ Radiotherapy → Targeted Therapy → Immunotherapy

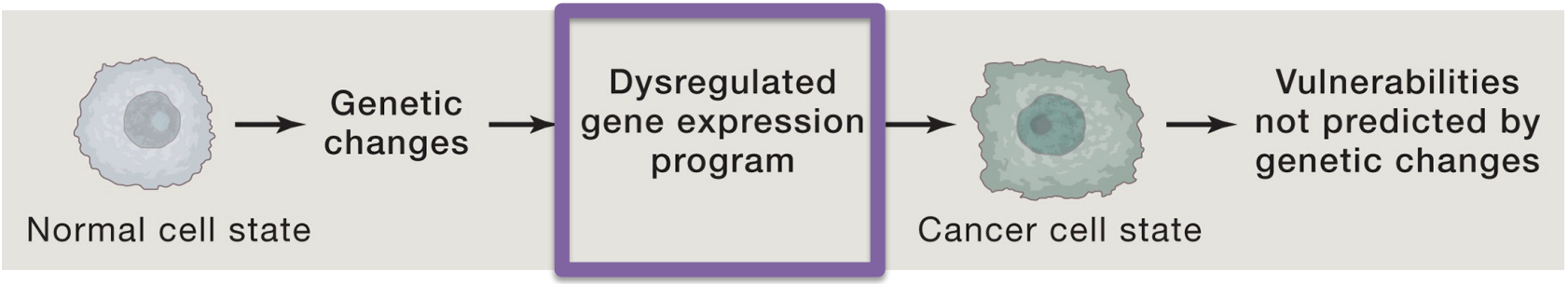
Precision Medicine

- **Epigenetics:** Signatures in chromosomes that define cancer cell identity
- **Immunology:** How immune systems work
- **Single-cell omics:** Precise map of tumor tissues

Epigenetic regulation of gene expression is critical in cancer development

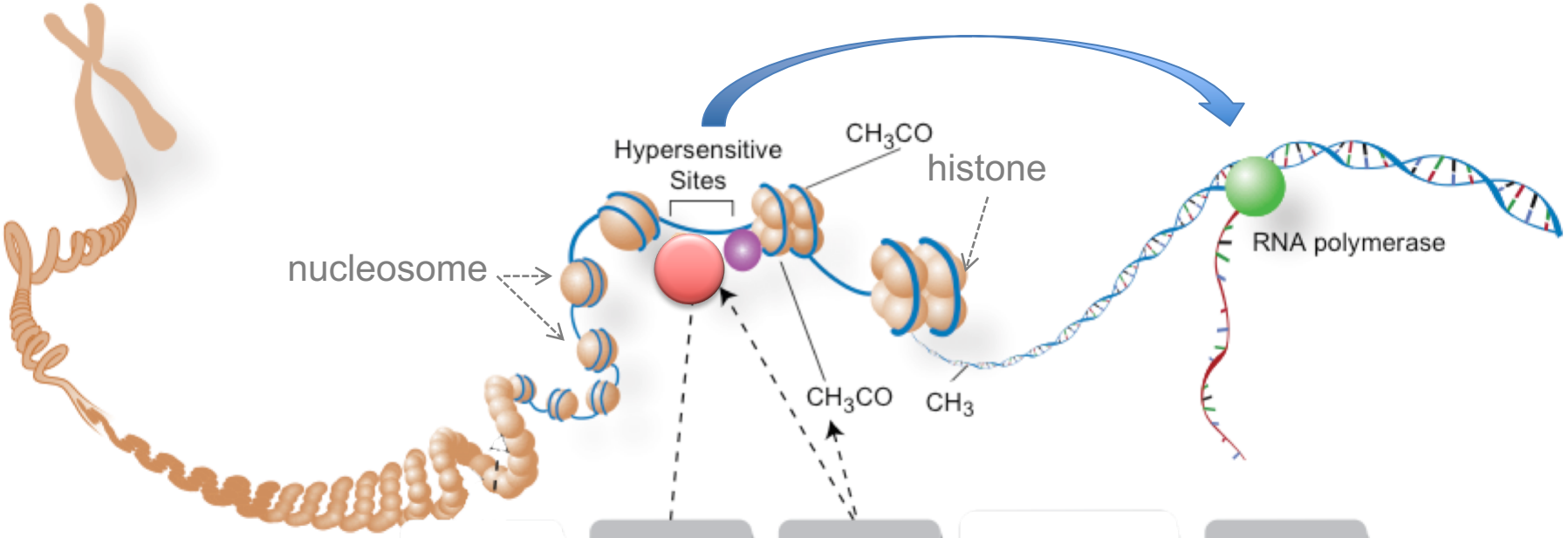


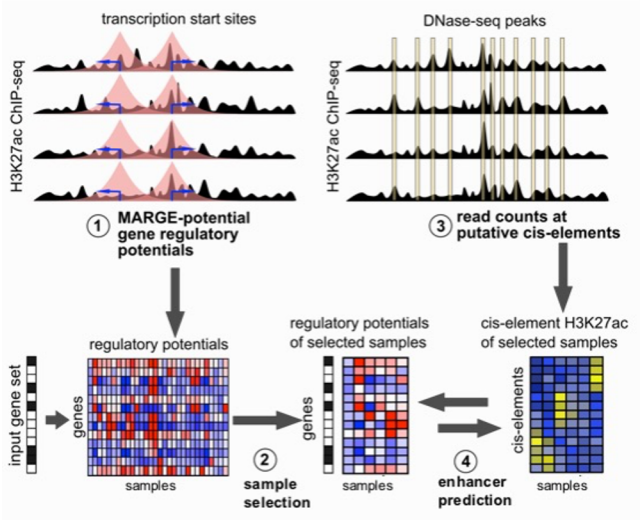
Tomasetti *et al*. *Science* 2017



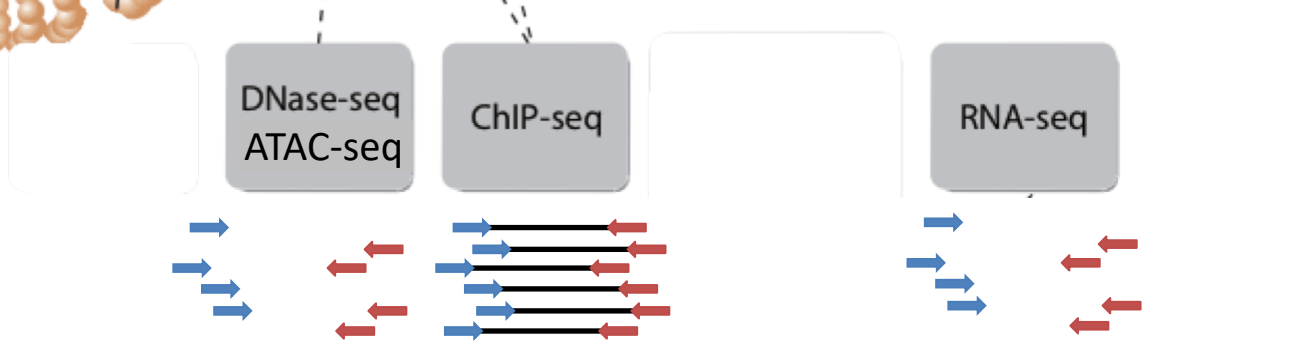
Bradner *et al*. *Cell* 2017

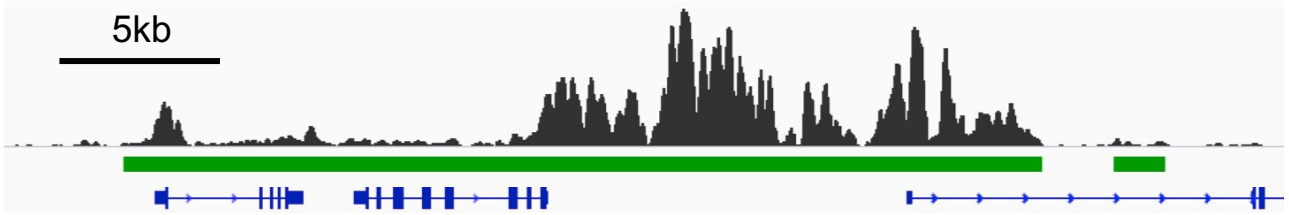
We develop computational methods and use data science approaches to study gene regulation and epigenetics

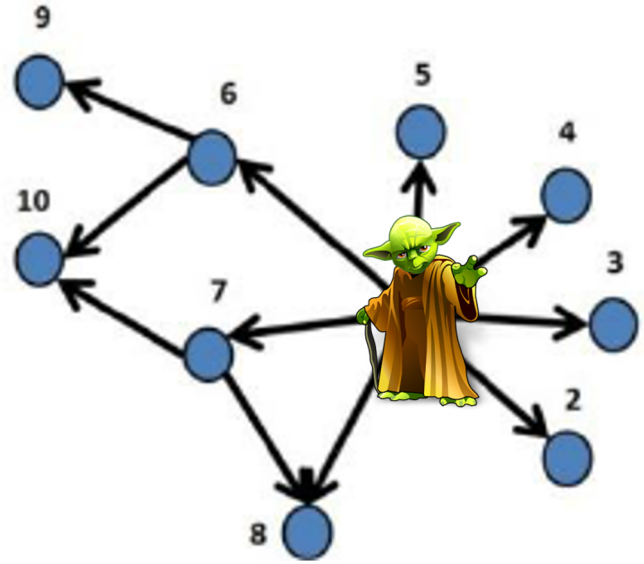





```
@D00372:517:C8AT2ANXX:7:1112:1223:2232 1:N:0:TGACCA
ACTCCAGTGGGTTTAACCTGGTTTTAAAAATCATCTTCCATAAAGGTTGGG
+
BBBBBFFFFFFFFFFFFFFFFFBFFFFFFBFFFFFFFFFFFFFFFFFFFF
@D00372:517:C8AT2ANXX:7:1112:1865:2247 1:N:0:TGACCA
CAAGGCTGTGTAATCAAAGCCCTTTGCACACCAAGAAGCCCATACAAAAGG
+
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@D00372:517:C8AT2ANXX:7:1112:2091:2129 1:N:0:TGACCA
CATCTCCTCTCCCCCTAAAAAAATTTCCATGGAAGATTTTGTGCCCTTGAT
+
BBBBBFFFFFFFFFFFFFFFBFFFFFFFFFFBFFFFFFFFFFFFF/<FFFF<
@D00372:517:C8AT2ANXX:7:1112:2248:2214 1:N:0:TGACCA
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCC
+
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFFFFF
@D00372:517:C8AT2ANXX:7:1112:2472:2150 1:N:0:TTACCA
CCTTAATTATTCTATGACAAGGGACATCAAATAAATGCCCCCAAATGCAAG
+
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@D00372:517:C8AT2ANXX:7:1112:2525:2095 1:N:0:TGACCA
CAGGCAACTAACTGCAGTGTTTTAGCGCCGTCTGGTGTTGGCGCCTAGAAA
+
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@D00372:517:C8AT2ANXX:7:1112:2564:2199 1:N:0:TGACCT
ATTTCAAGTCCCCTGGGCCAGACTTTCCTGGCCTTTGTCGCAGAATTTTTG
+
BBBBB//<FFFB/<<B</F/<FFFFFFFF<<///</B/<F</7F//FFF/F/
@D00372:517:C8AT2ANXX:7:1112:2924:2111 1:N:0:TGACCA
GGTGTTAAGGGACTTATTTCAAGTGTAAATTGGCCATAGATATTGAAGAAG
+
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFB
```



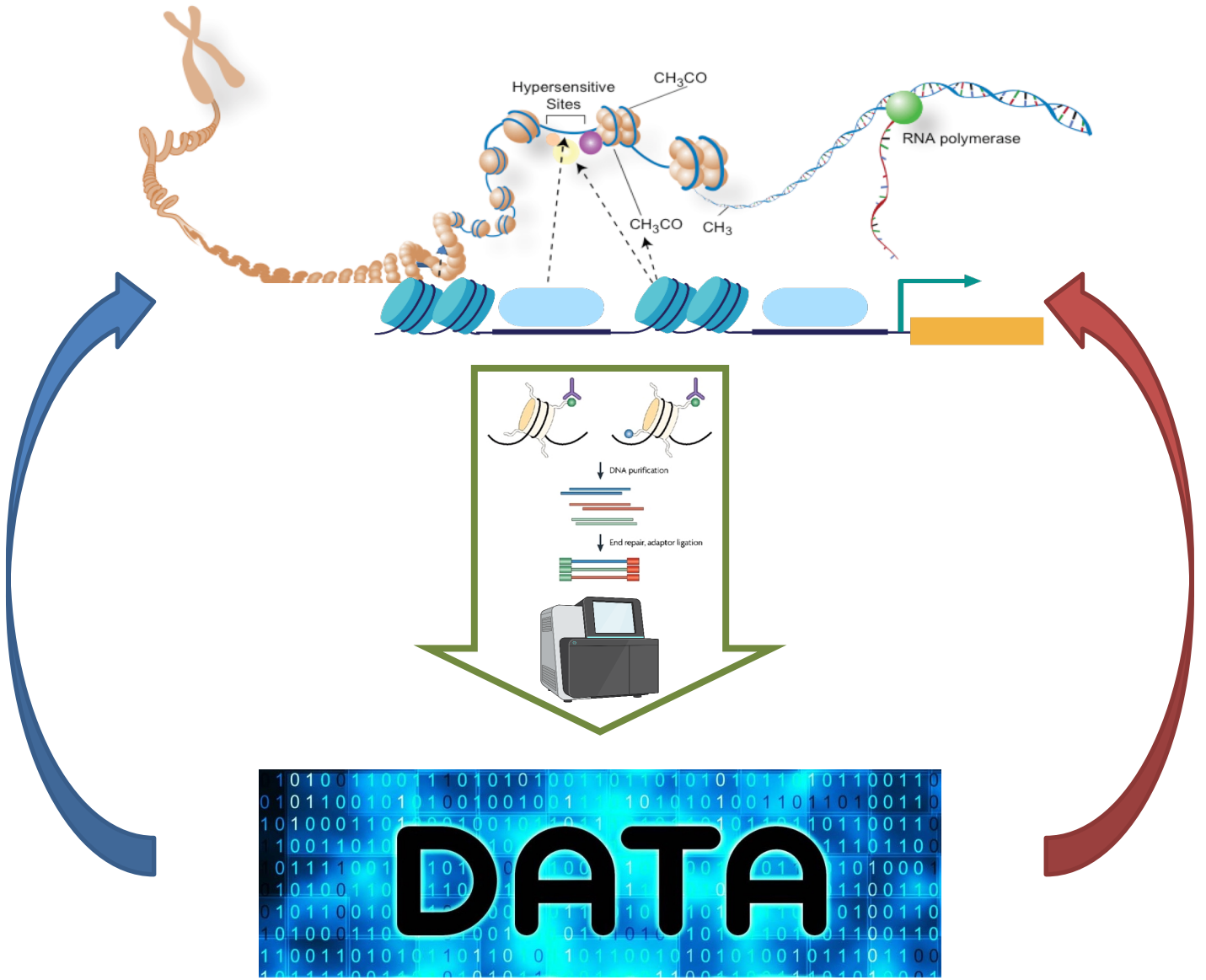




We develop computational methods and use data science approaches to study epigenetics and transcriptional regulation

Method development

- Stat/math/physics /machine learning models
- Algorithms
- Software tools

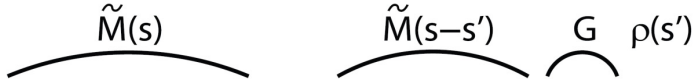


Data science discovery

- Multi-modal data integration
- Innovative models
- New biology

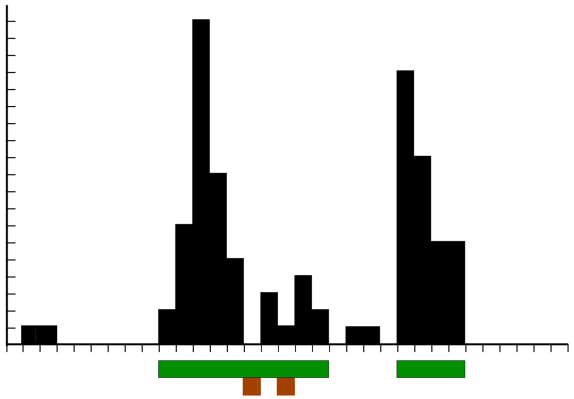
SICER: detecting broad peaks from ChIP-seq data

- **S**patial-clustering **I**dentification of **C**hIP-seq **E**nriched **R**egions



$$\tilde{M}(s) = G(\lambda, l_0, g) \int_{s_0}^{s} ds' \tilde{M}(s-s')\rho(s')$$

$$M(s) = t^{g+1}\tilde{M}(s)t^{g+1}$$



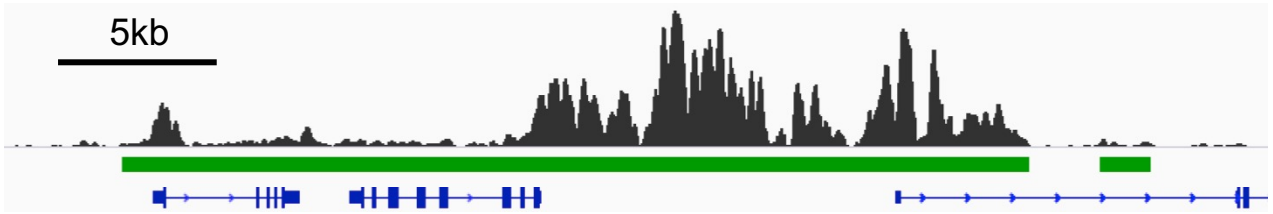
Docs » Quick Start

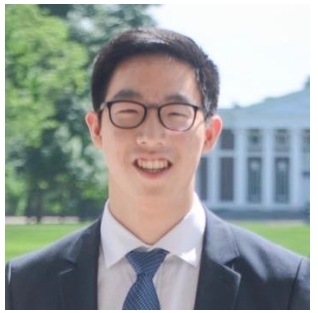
SICER2

Redesigned and improved ChIP-seq broad peak calling tool SICER

build passing

GitHub Repo





Jeffrey Yoo ’20

Zang et al. *Bioinformatics* 2009
Zang et al. *Quantitative Biology* 2020

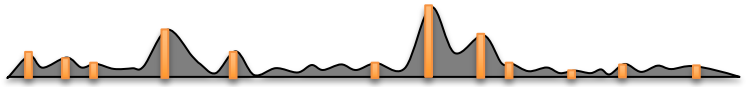
https://zanglab.github.io/SICER2/

Public data can be powerful resources for computational biologists

MARGE

gene set

Adaptive Lasso regression
+ semi-supervised learning



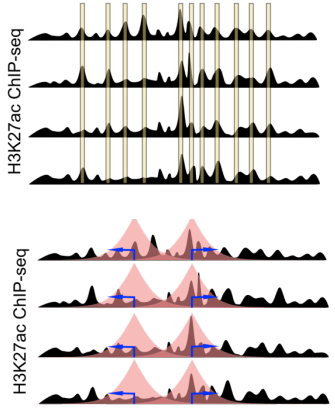
cis-regulatory (enhancer) profile

ROC association analysis,
background adjustment, rank
integration

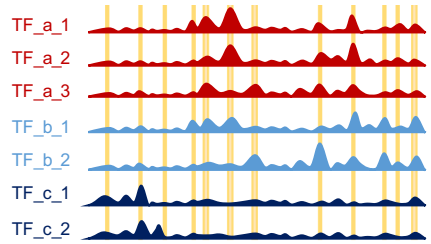
transcription factors

> 500 DNase-seq

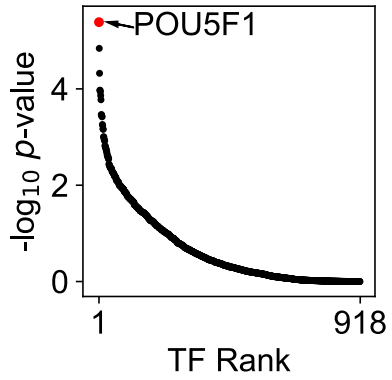
> 1000 H3K27ac ChIP-seq



> 13,000 TF ChIP-seq



BART

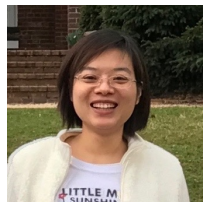


UNIVERSITY of VIRGINIA

SECTIONS

UVAToday

UVA DEVELOPS NEW TOOLS TO BATTLE CANCER, ADVANCE GENOMICS RESEARCH



Zhenjia Wang

Wang *et al. Bioinformatics* (2018)
Ma, Wang *et al. NAR Genomics & Bioinformatics* (2021)
Wang *et al. Bioinformatics* (2021)
Thomas *et al. NAR Cancer* (2021)

BART Cancer: inferring active TFs in each cancer type

TCGA

BLCA

BRCA_1

BRCA_2

COAD_READ

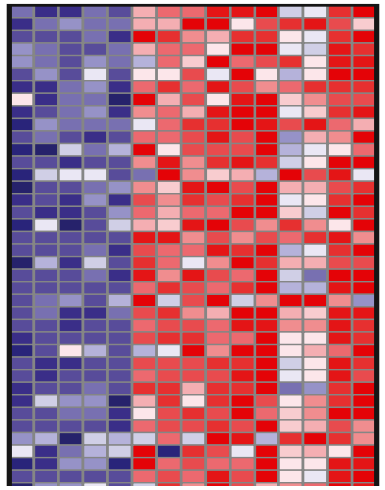
GBM

PRAD

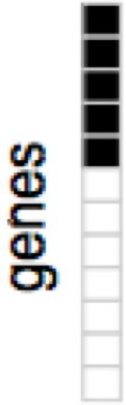
…

TCGA expression

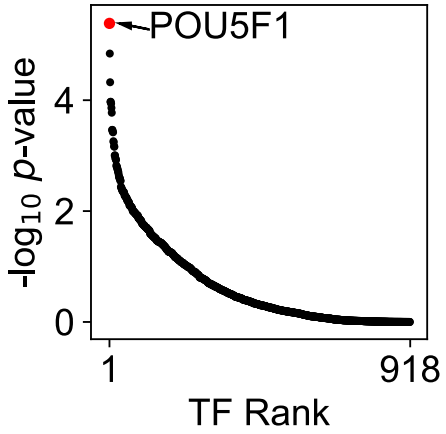
normal cancer



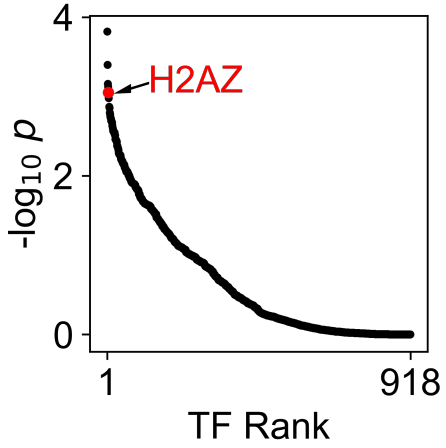
Cancer-specific gene sets

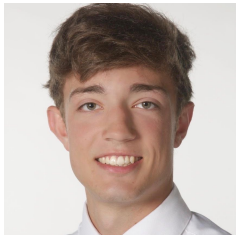


Up-regulated in cancer



Down-regulated in cancer



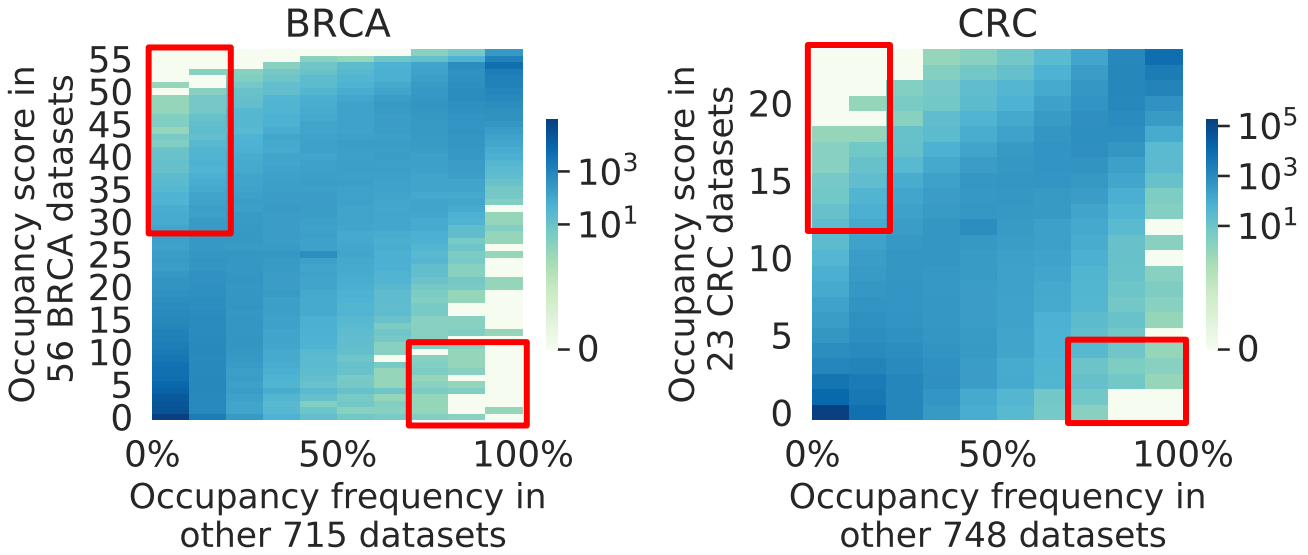


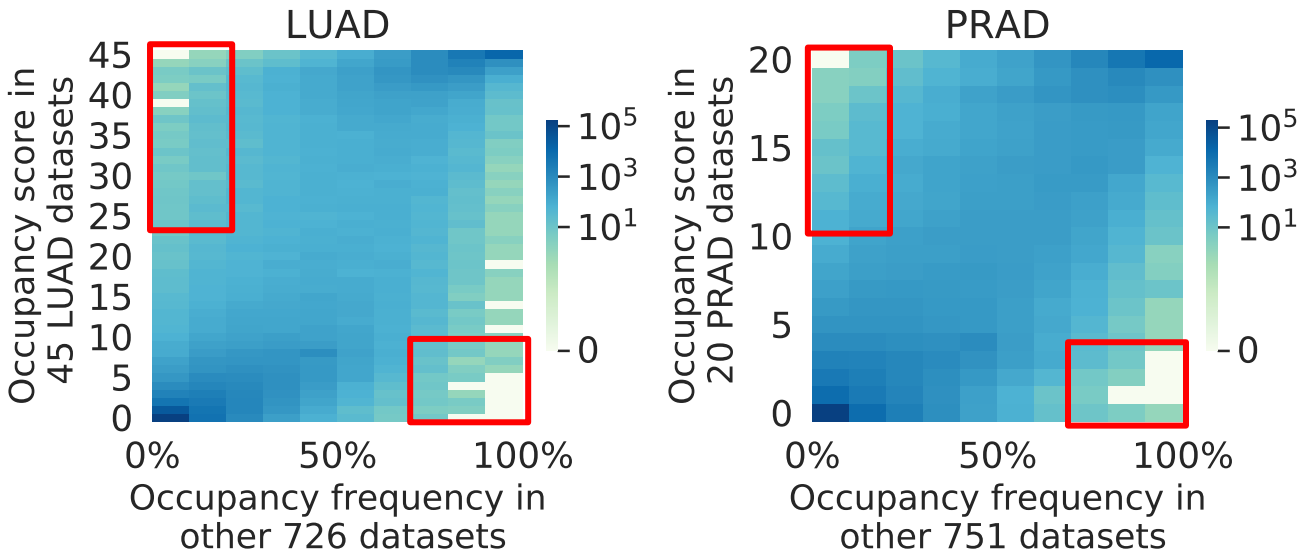
Zack Thomas '21

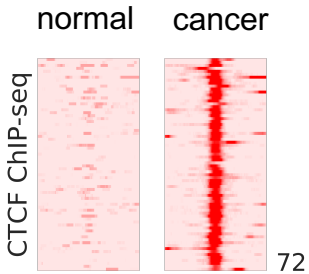
bartcancer.org

Thomas *et al*. *NAR Cancer* 2021

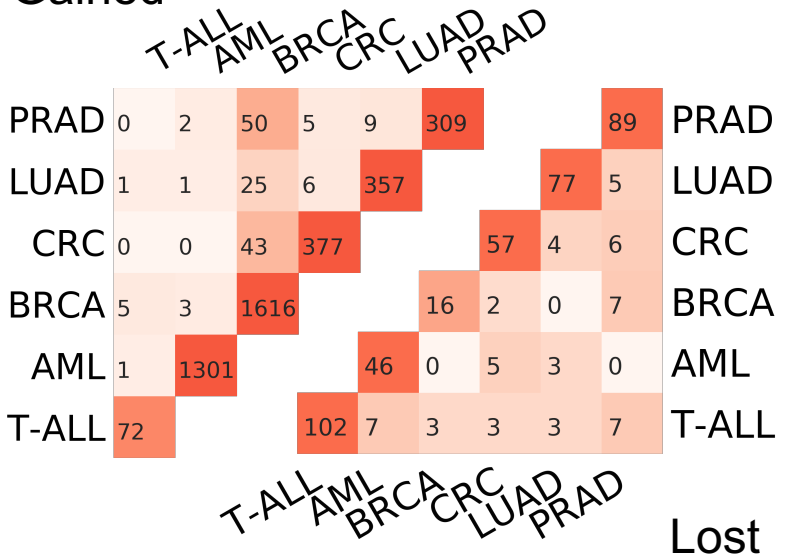
Gained/Lost CTCF bindings occur specifically in each different cancer type





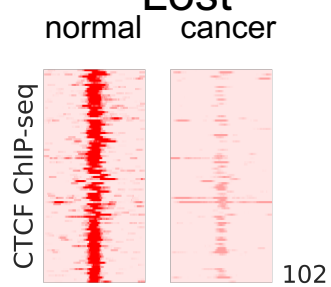


Gained



Emilie Neulen '22

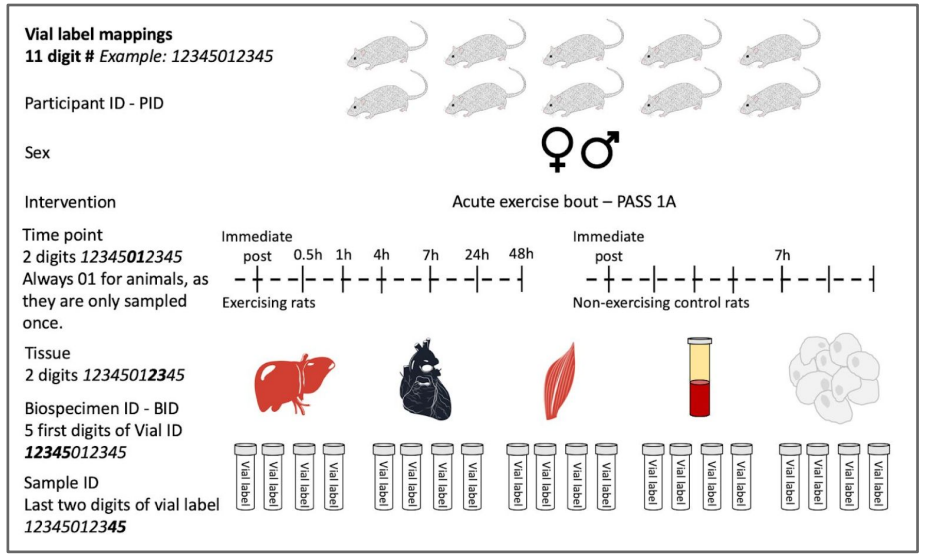
Lost

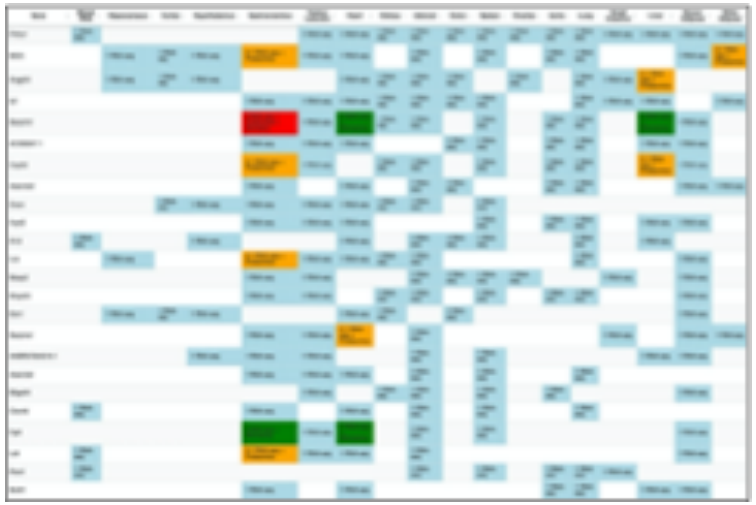


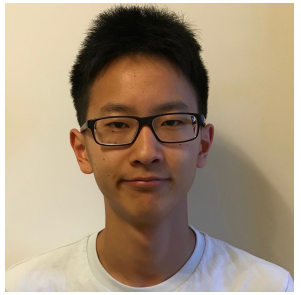
Fang, Wang *et al. Genome Biology* 2020

Molecular mechanisms underlying health of physical activity

- Molecular Transducers of Physical Activity Consortium
  - National research group investigating how physical activity improves and maintains human health and building a map of the molecular responses to exercise
- Preclinical animal data, generated from rats, provided multi-omics information for 19 unique rat tissues at 7 different time points after an acute bout of exercise
- In particular, we are interested in identifying secreted gene products and elucidating the biochemical mechanisms that could elicit pleiotropic effects in other tissues

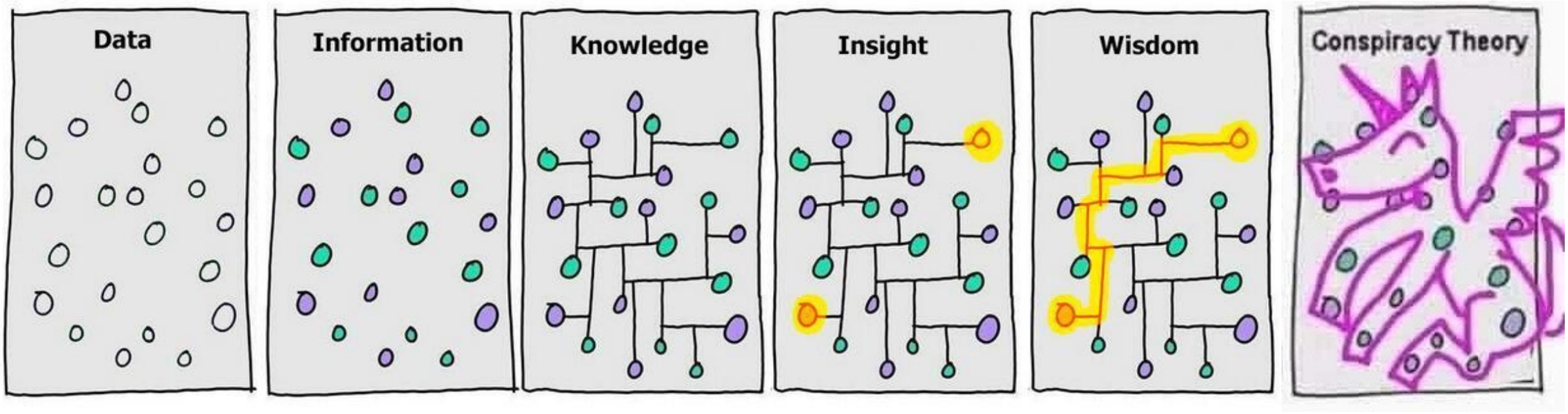






Ben Ke ’22

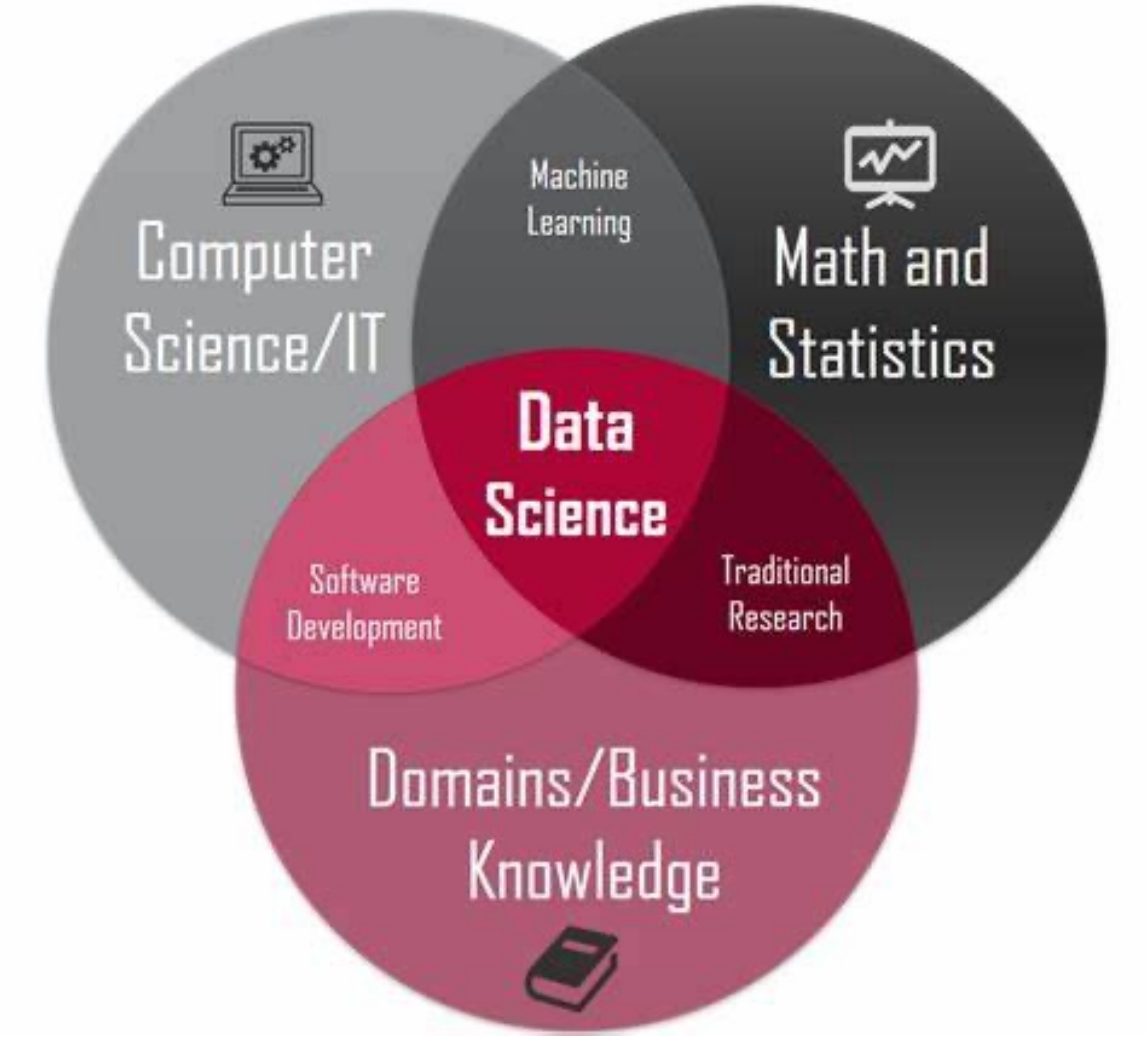
Data Science: Data + innovative computational modeling → new methods and/or scientific discoveries



RNA-seq ChIP-seq DNase/ATAC-seq Hi-C Public databases …	Gene expression Protein factors Chromatin 3D genome Multi-omics …	Transcriptional regulation, Chromatin organization, …	New insight?	New biology!	Overfitting, overinterpretation...

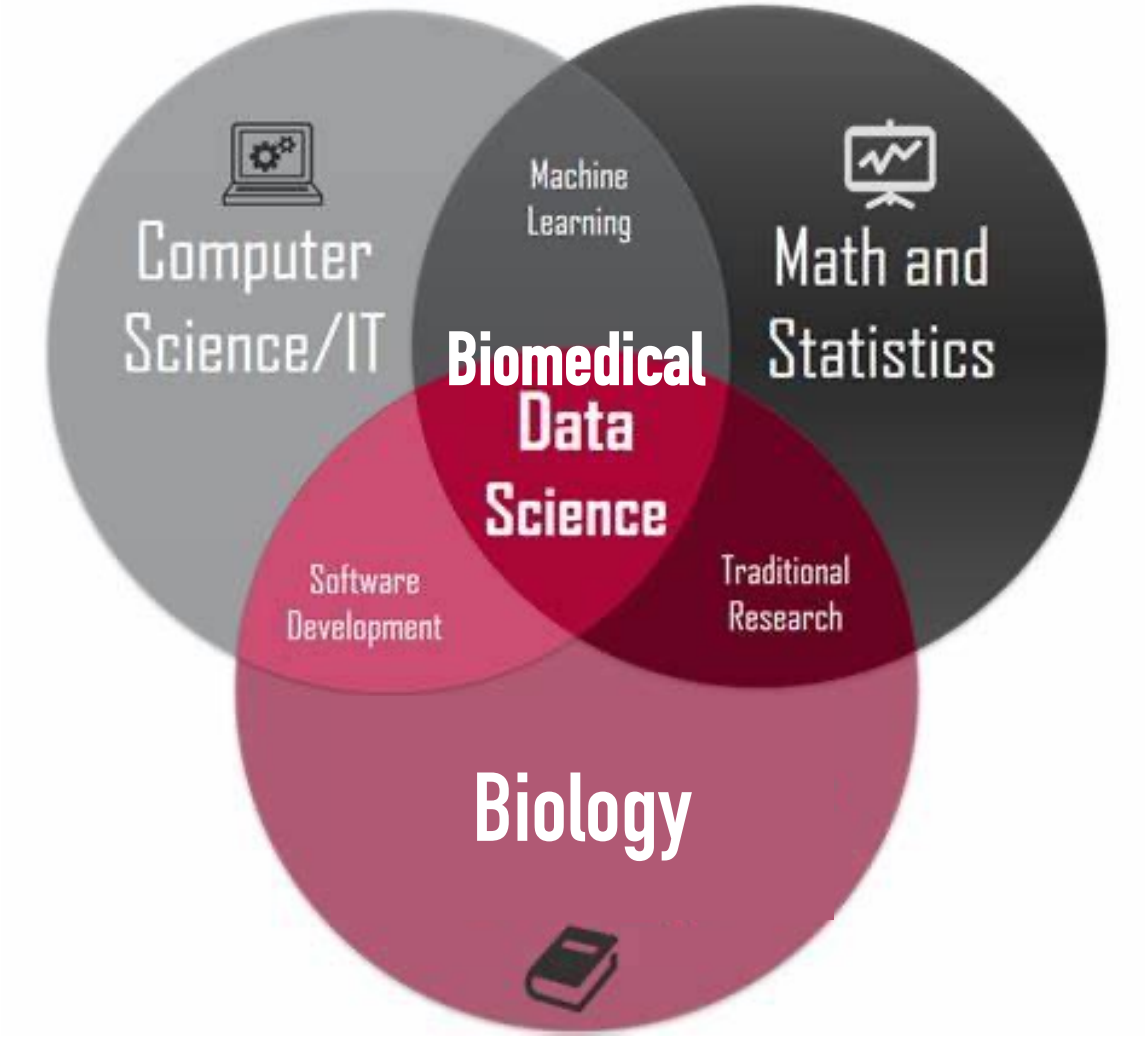
Data Science

zanglab.org



Biomedical Data Science (Bioinformatics/Computational Biology)

zanglab.org



Levels of Bioinformatics Research (by Shirley Liu)

- **Level 0**: “modeling for modeling’s sake”
- **Level 1**: applying existing software tools to analyze new data and to make novel findings
- **Level 2**: developing new methods (models/algorithms/tools) for a general data science problem for biomedical research
- **Level 3**: integrating existing big data in a smart way to make new biological discoveries
- **Level X**: providing key integration and modeling to massive data from big consortia, leading a team science effort combining Levels 1-3

“The best minds of my generation are thinking about how to make people click ads. That sucks.”

-- Jeff Hammerbacher

Co-founder of Facebook

Computational Biologist

Data Scientist

Embrace your passion!

Acknowledgements

Zang Lab

Collaborators

UVA
- School of Medicine
- Cancer Center
- Biochemistry and Molecular Genetics
- Biomedical Engineering
- Computer Science
- Statistics
- School of Data Science

Northwestern University

New York University

Children's National Medical Center

Hackensack University Medical Center

Dana-Farber Cancer Institute

zanglab.org

Welcome to join us!

R35 GM133712
R01 AI112579

CHRB