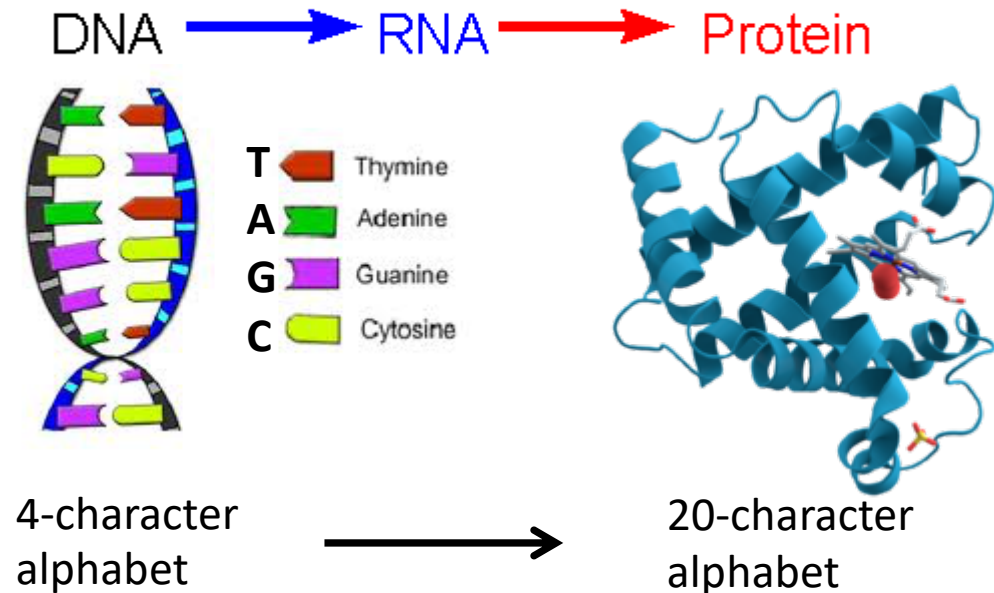# Bioinformatics Wrap-Up

Dr. Garrett Dancik

# What is bioinformatics

- Bioinformatics:
  - Biology + information
  - the study and utilization of methods for storing, retrieving and analyzing biological data
  - Falls under the category of Big Data



4-character alphabet → 20-character alphabet

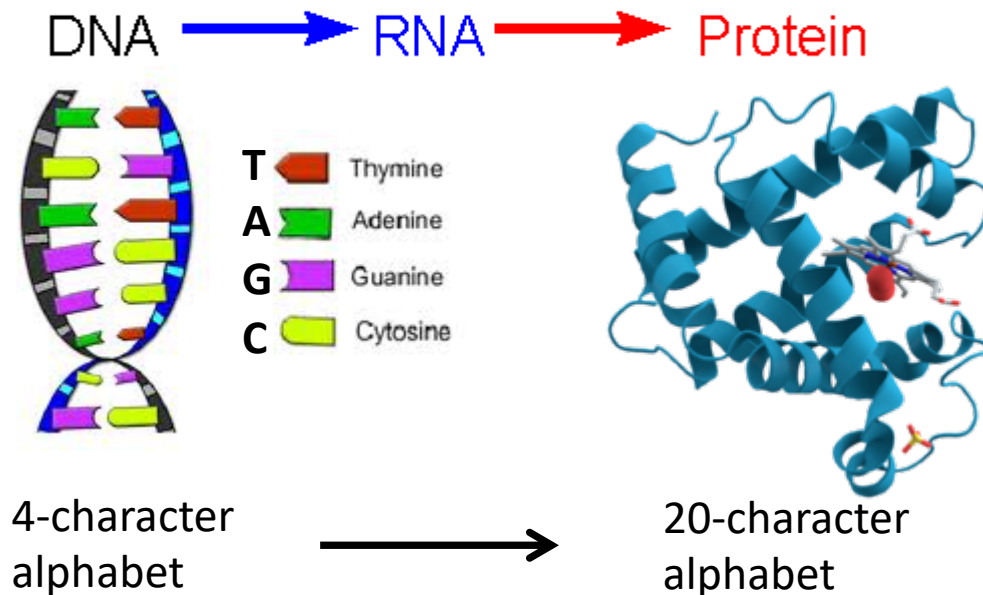- How much information:
  - Human genome: 3 billion nucleotides
  - ~20,000 genes
    - many more when considering "junk DNA" and alternative splicing
  - >10 million sites of DNA variation
  - Countless possible interactions between DNA, RNA, and proteins

# What is this?

print('hello world!')

- Computer code is a *set of instructions* that tells a computer how to process data and output results
- The genetic code is also a set of instructions, that tells a cell how to produce a molecule (such as a protein)
  - Information flows from DNA → RNA → protein
  - This information determines the structure/function of RNA and protein

# Central Dogma of Molecular Biology

DNA ➡ RNA ➡ Protein

T ◀ Thymine
A ◀ Adenine
G ◀ Guanine
C ◀ Cytosine

4-character
alphabet
➡
20-character
alphabet

- The function of a protein can be predicted from its DNA or protein sequence
- Just like Java (or Python) is a language for computers, genetics is the language of life (DNA is the alphabet)
  - This is  fundamental concept in bioinformatics

# Bioinformatics example

- Let's look briefly at the gene *BRCA1* (breast cancer type 1 susceptibility protein)
  - http://www.ncbi.nlm.nih.gov
  - Search Nucleotide for BRCA1
  - Click on the Transcript reference sequence
- Hopefully this now makes sense!

# Why do we need bioinformatics?

- To identify genetic mechanisms of diseases and other inherited (or acquired) conditions
  - Nature via nurture
- For personalized treatment of disease

# Why do we need bioinformatics?

- BLAST

## Bioinformatics Analysis of the Complete Genome Sequence of the Mango Tree Pathogen *Pseudomonas syringae* pv. syringae UMAF0158 Reveals Traits Relevant to Virulence and Epiphytic Lifestyle

Pedro Manuel Martínez-García [co], Pablo Rodríguez-Palenzuela [co], Eva Arrebola, Víctor J. Carrión,
José Antonio Gutiérrez-Barranquero, Alejandro Pérez-García, Cayo Ramos, Francisco M. Cazorla, Antonio de Vicente ✉

- Sequence Alignment

## A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples

Samia N. Naccache,[1,2] Scot Federman,[1,2] Narayanan Veeraraghavan,[1,2] Matei Zaharia,[3] Deanna Lee,[1,2] Erik
Samayoa,[1,2] Jerome Bouquet,[1,2] Alexander L. Greninger,[4] Ka-Cheung Luk,[5] Barryett Enge,[6] Debra A. Wadford,[6]
Sharon L. Messenger,[6] Gillian L. Genrich,[1] Kristen Pellegrino,[7] Gilda Grard,[8] Eric Leroy,[8] Bradley S. Schneider,[9]
Joseph N. Fair,[9] Miguel A. Martínez,[10] Pavel Isa,[10] John A. Crump,[11,12,13] Joseph L. DeRisi,[4] Taylor Sittler,[1] John
Hackett, Jr.,[5] Steve Miller,[1,2] and Charles Y. Chiu[1,2,14,15]

Author information ▶ Article notes ▶ Copyright and License information ▶

- Group project

# Regulation of developmental rate and germ cell proliferation in *Caenorhabditis elegans* by the p53 gene network

WB Derry[*,1,2,3], R Bierings[1], M van Iersel[1], T Satkunendran[2,3], V Reinke[4] and JH Rothman[*,1]

**p53 consensus binding sites in CEP-1-regulated genes.** In an effort to identify putative direct targets of CEP-1, we analyzed CEP-1-activated genes for p53 consensus DNA-binding sites.[23] Vertebrate p53 binds two copies of the palindromic DNA consensus sequence RRRCWWGYYY, and CEP-1 binds an oligonucleotide containing the human p53 consensus motif *in vivo* and *in vitro*.[14,15] We searched the *C. elegans* genome for pairs of p53 consensus binding sites and found that ~25% of CEP-1-activated genes contain potential p53-binding sites in their promoters and introns (Table S2). These sites were

# The genomic landscape of bladder cancer
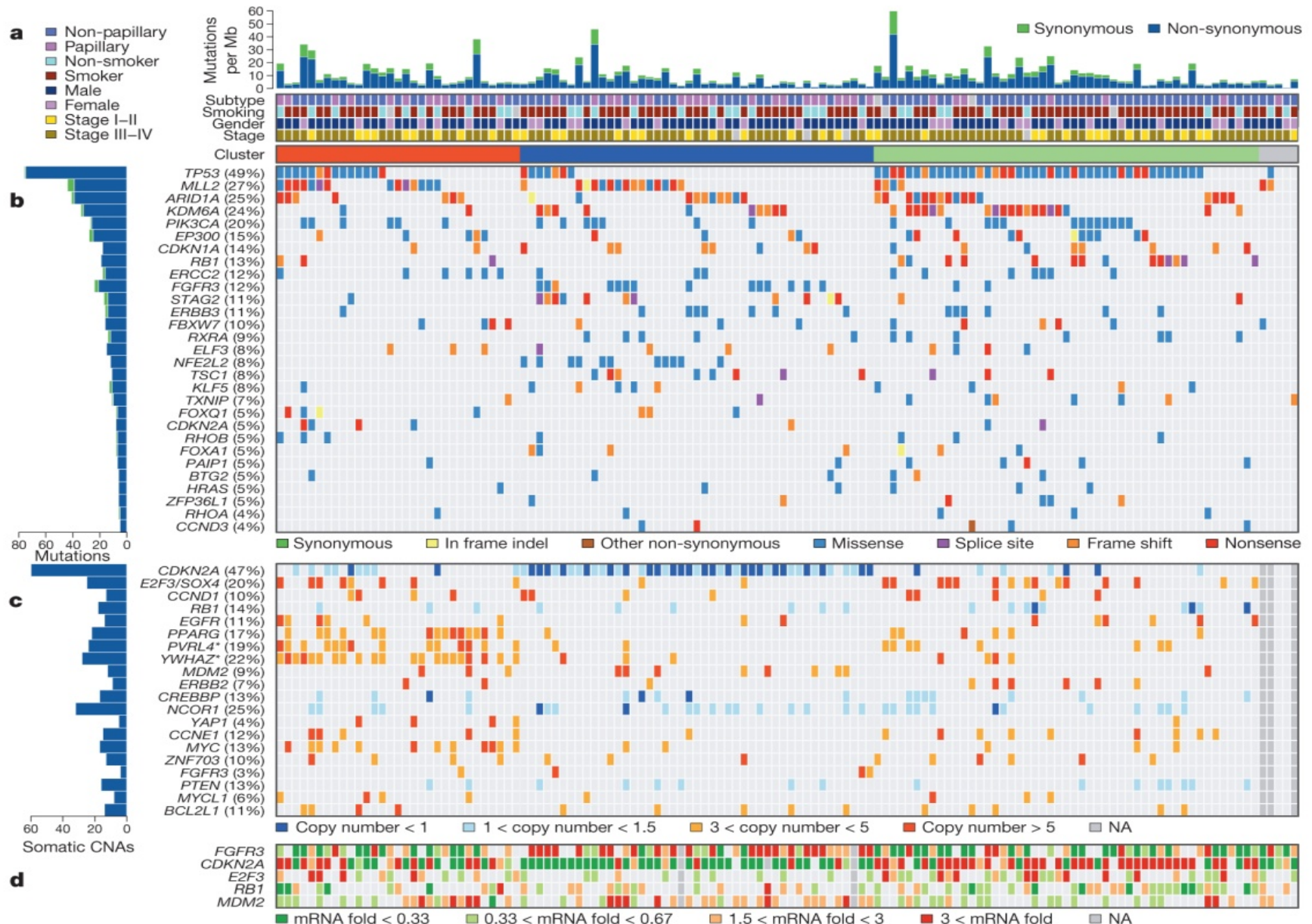
# Bioinformatics and coronavirus

- Coronavirus genomic analysis suggests early spread in the United States
  - https://www.the-scientist.com/news-opinion/coronaviruss-genetics-hint-at-its-cryptic-spread-in-communities-67233
- Advanced bioinformatics rapidly identifies existing therapeutics for patients with coronavirus disease-2019 (COVID-19)
  - https://translational-medicine.biomedcentral.com/articles/10.1186/s12967-020-02430-9
- Combination of Biodata Mining and Computational Modelling in Identification and Characterization of ORF1ab Polyprotein of SARS-CoV-2 Isolated from Oronasopharynx of an Iranian Patient
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7171442/
- Computational Strategies to Combat COVID-19
  - https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa232/5955939

# Additional Databases and Tools

- The Cancer Genome Atlas
  - http://cancergenome.nih.gov
  - Comprehensive genomic characterization of tumors
- cBioPortal for Cancer Genomics
    - http://www.cbioportal.org/
- Genomics of Drug Sensitivity in Cancer
  - http://www.cancerrxgene.org
  - Screen > 1000 cell lines with drug
  - Genomic characterization of cell lines
- And many more (current research)
  - http://bioinformatics.oxfordjournals.org
  - http://www.biomedcentral.com/bmcbioinformatics