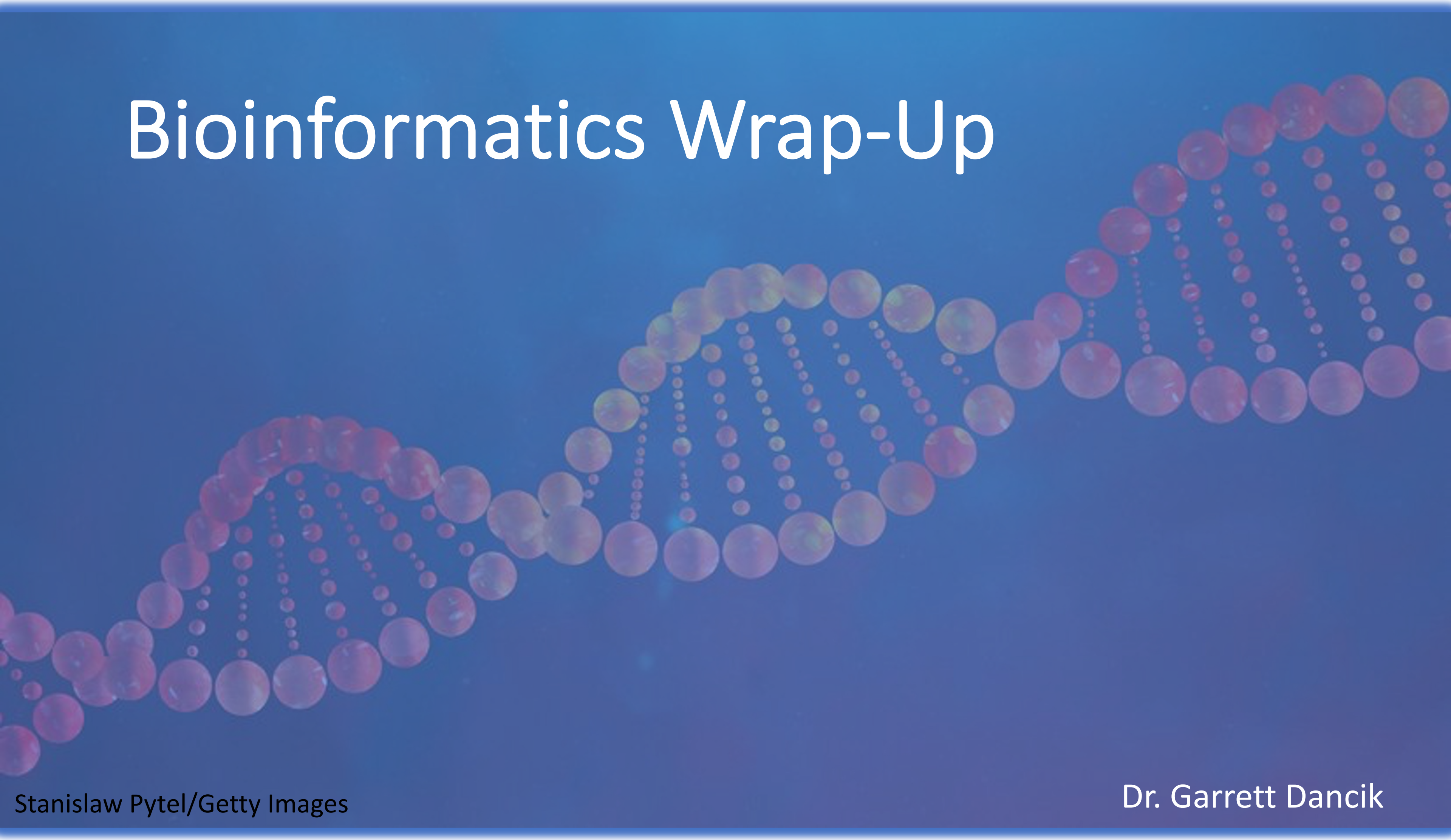
Bioinformatics Wrap-Up

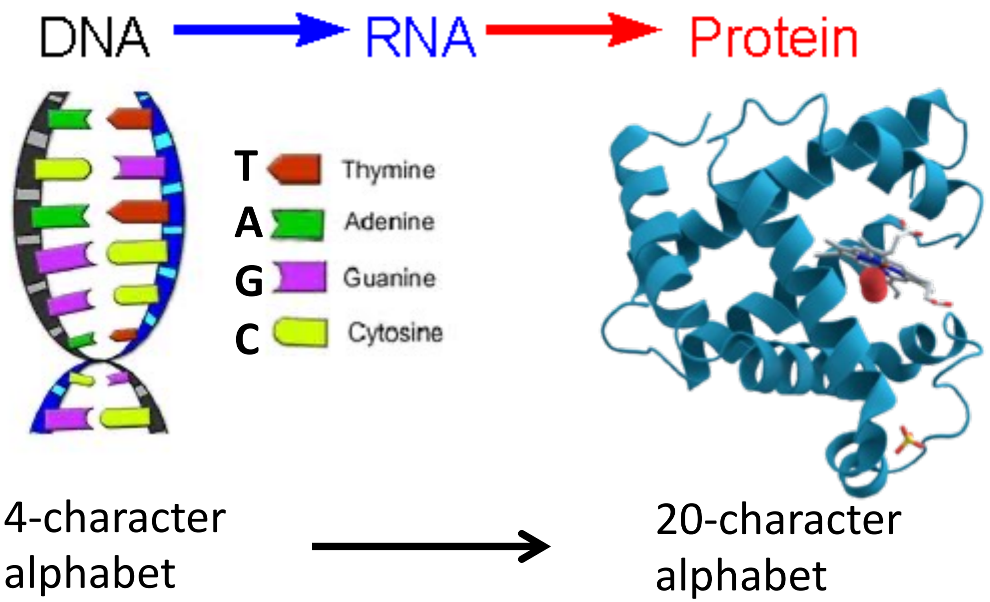


Stanislaw Pytel/Getty Images

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

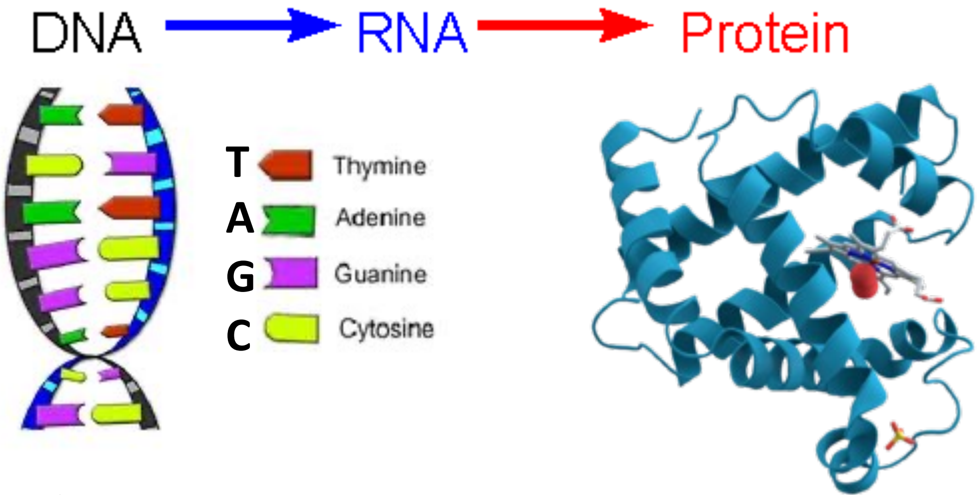
- How much information:
  - Human genome: 3 billion nucleotides
  - ~20,000 genes
    - many more when considering “junk DNA” and alternative splicing
  - >10 million sites of DNA variation
  - Countless possible interactions between DNA, RNA, and proteins

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 protein from a DNA molecule
  - Information flows from DNA → RNA → protein
  - This information determines the structure/function of RNA and protein

Central Dogma of Molecular Biology



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 a fundamental concept in bioinformatics

Bioinformatics example

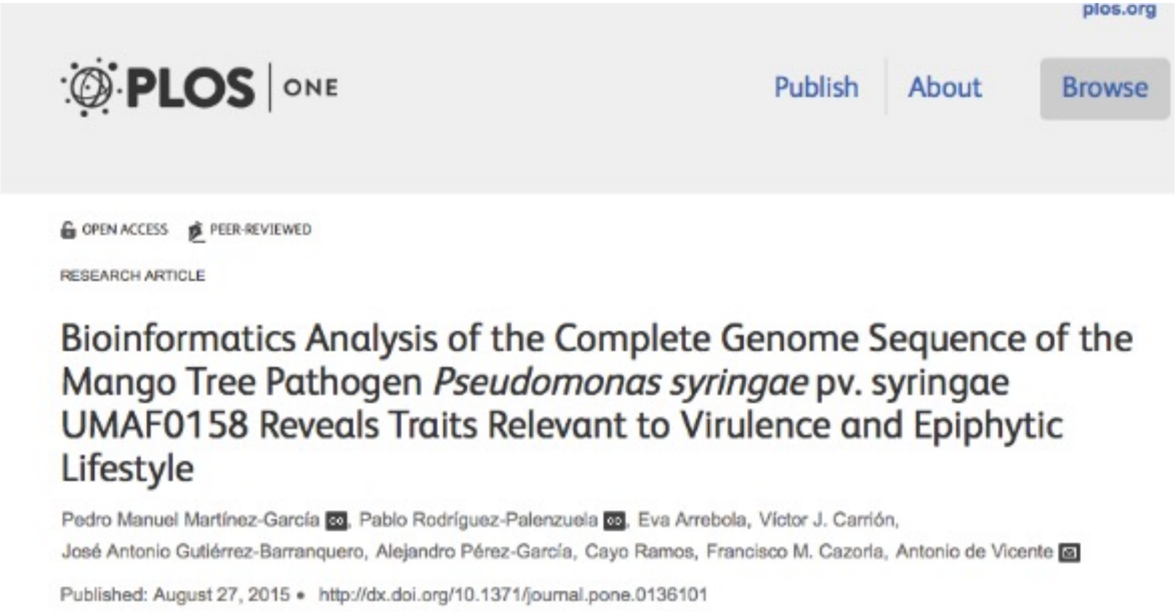
- Let's look briefly at the gene *BRCA1* (breast cancer type 1 susceptibility protein)
  - http://www.ncbi.nlm.nih.gov
  - Search Nucleotide for BRCA1
  - Click on the Genomic reference sequence
- Hopefully this now makes sense!

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



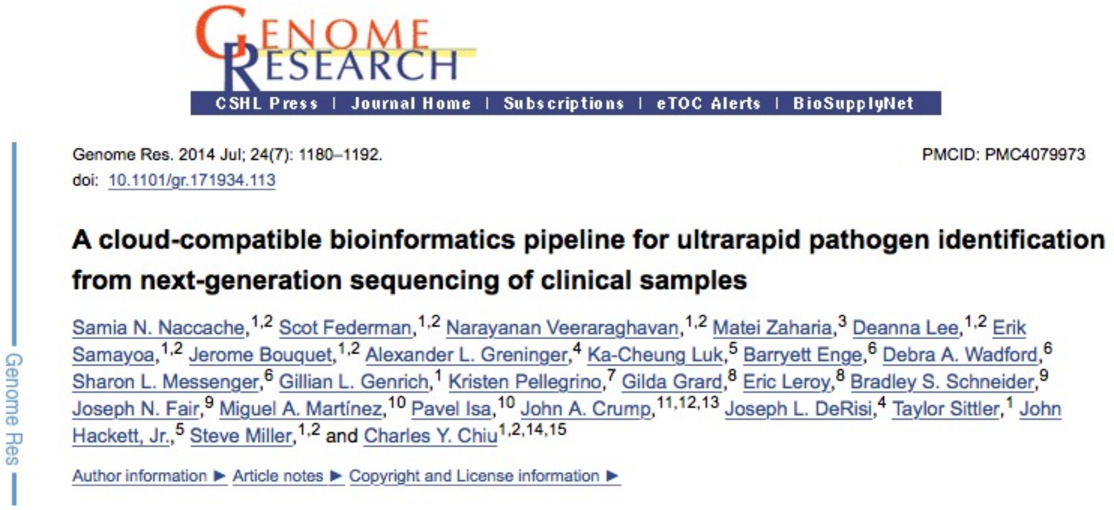
RESEARCH ARTICLE

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, Pablo Rodríguez-Palenzuela, Eva Arrebola, Víctor J. Carrión,
José Antonio Gutiérrez-Barranquero, Alejandro Pérez-García, Cayo Ramos, Francisco M. Cazorla, Antonio de Vicente

Published: August 27, 2015 • http://dx.doi.org/10.1371/journal.pone.0136101

- Sequence Alignment



Genome Res. 2014 Jul; 24(7): 1180–1192.
doi: 10.1101/gr.171934.113

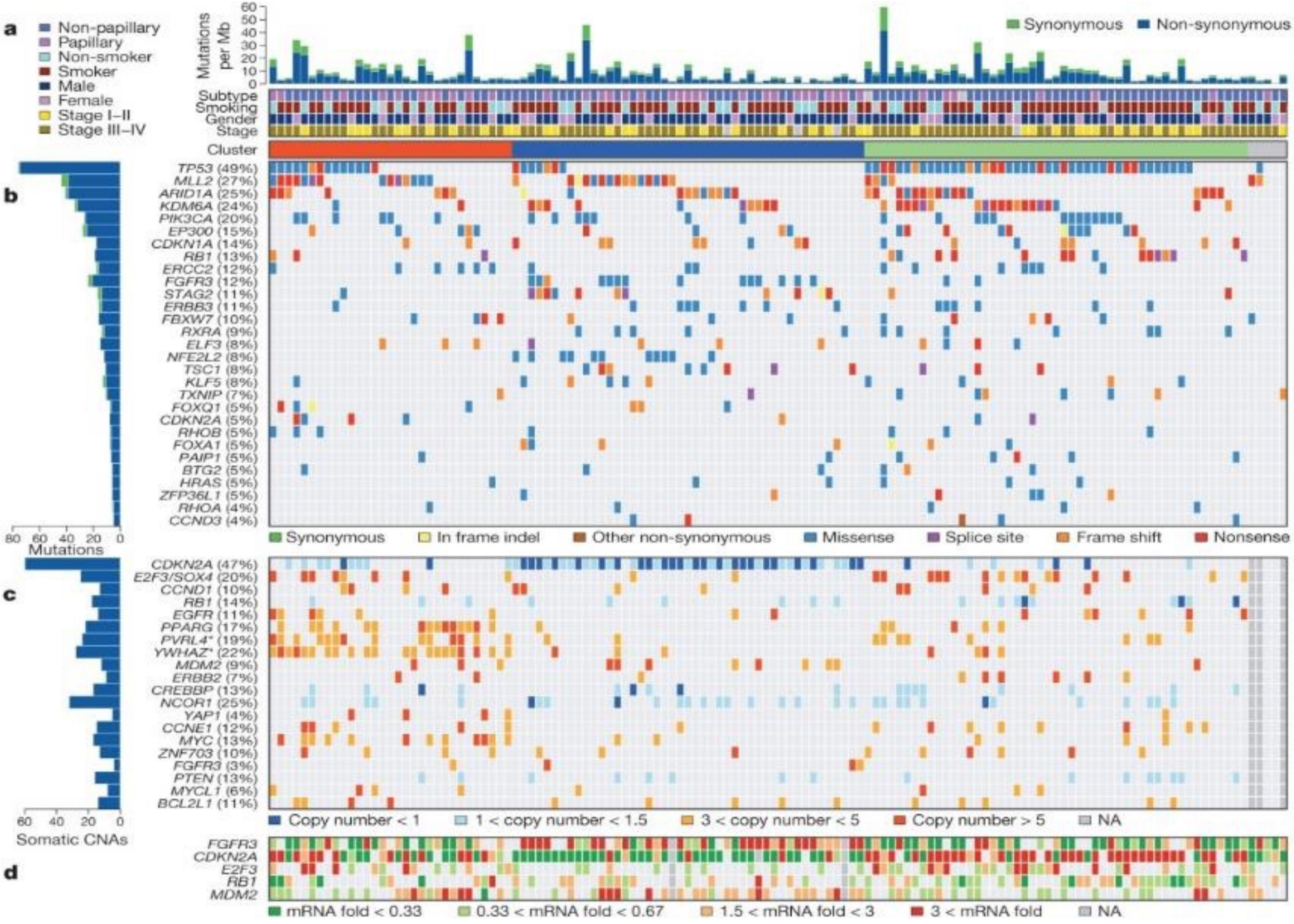
PMCID: PMC4079973

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 ▶

The genomic landscape of bladder cancer



The Cancer Genome Atlas Research Network *Nature* **507**, 315-322 (2014) doi:10.1038/nature12965

More bioinformatics applications and research

- 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
- Species identification using DNA barcoding and BLAST
  - https://www.sciencedirect.com/science/article/abs/pii/S0956713521008902
- Identification of mutation hotspots using Biopython
  - https://www.mdpi.com/1422-0067/25/12/6612
- Topic modeling (summarizing the literature) using Biopython
  - https://academic.oup.com/bib/article/23/5/bbac194/6590367
- Creating custom tracks for UCSC Genome Browser
  - https://www.biorxiv.org/content/10.1101/2024.12.23.629822v1.abstract

Additional Databases and Tools

- The Cancer Genome Atlas
  - http://cancergenome.nih.gov
  - Comprehensive genomic characterization of tumors
- cBioPortal for Cancer Genomics
    - https://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