



Making Sense of Biomedical Data Using the Tools of AI

Lior Pachter
Division of Biology and Biological Engineering
Department of Computing and Mathematical Sciences
California Institute of Technology

Cancer Biomarkers AI and Bioinformatics Workshop
August 13, 2024

What is artificial intelligence?

- The term “artificial intelligence” was coined by John McCarthy in **1955**. His aim was to distinguish his research agenda, which emphasized ties between computation and logic (lambda calculus), to that of Norbert Wiener’s vision of intelligent systems based on statistics, control theory, and operations research, as “cybernetics”.
- In a twist of irony, Wiener’s vision is now being realized under McCarthy’s moniker ([Jordan, 2019](#)).
- Jordan argues that we are in the midst of the creation of a new branch of a human-centric engineering discipline. He suggests that we “broaden our scope, tone down the hype, and recognize serious challenges ahead.”

AI Biology Startup EvolutionaryScale Announces \$142M Seed Financing

Bioptimus launches with \$35M to create an AI foundation model for biology

The company was founded by former Google DeepMind and Owkin scientists.

AI at the Intersection: The a16z Investment Thesis on AI in Bio + Health

Vijay Pande

SHARE ▾

Posted June 21, 2023

A Tale of Two Exponentials

AI@HHMI

Over the next decade, HHMI will invest \$500 million to embed AI systems throughout every stage of the scientific process

Moore's Law. These industries have suffered from a decades-long increase in cost, which has reached such extreme levels that healthcare costs are reaching approximately a quarter of US GDP (and rising). Between rising labor costs, clinical trial costs, administrative costs, increasingly hostile payor-provider relations, and more, therapeutics and healthcare costs are following the wrong curve.

\$5 Million Grant Bets on Computational Biology, AI to Change the Future of Cancer

ARTICLE | BY KELLY QUIGLEY

NVIDIA Doubles Down on AI Drug Discovery with \$50M Recursion Investment

July 10, 2023 · 2 min read · Connor Lynch

Cradle raises \$24M Series A and signs partnerships with industry leaders



STEF VAN HOUTEN
NOVEMBER

AI, virtual cells, and a revolutionary platform for biology



Making Cents of Biomedical Data Using the Tools of AI

Lior Pachter
Division of Biology and Biological Engineering
Department of Computing and Mathematical Sciences
California Institute of Technology

Cancer Biomarkers AI and Bioinformatics Workshop
August 13, 2024

What are “the tools of AI”?

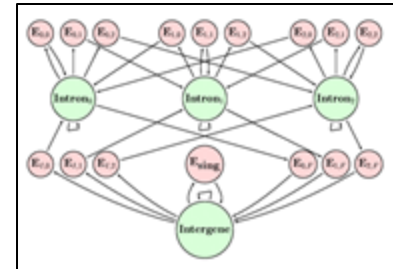
- In the mid-1990s, in the coldest years of the “AI winter”, the state-of-the-art AI in biology was the hidden Markov model, used for gene discovery in newly sequenced genomes.

Observed sequence:

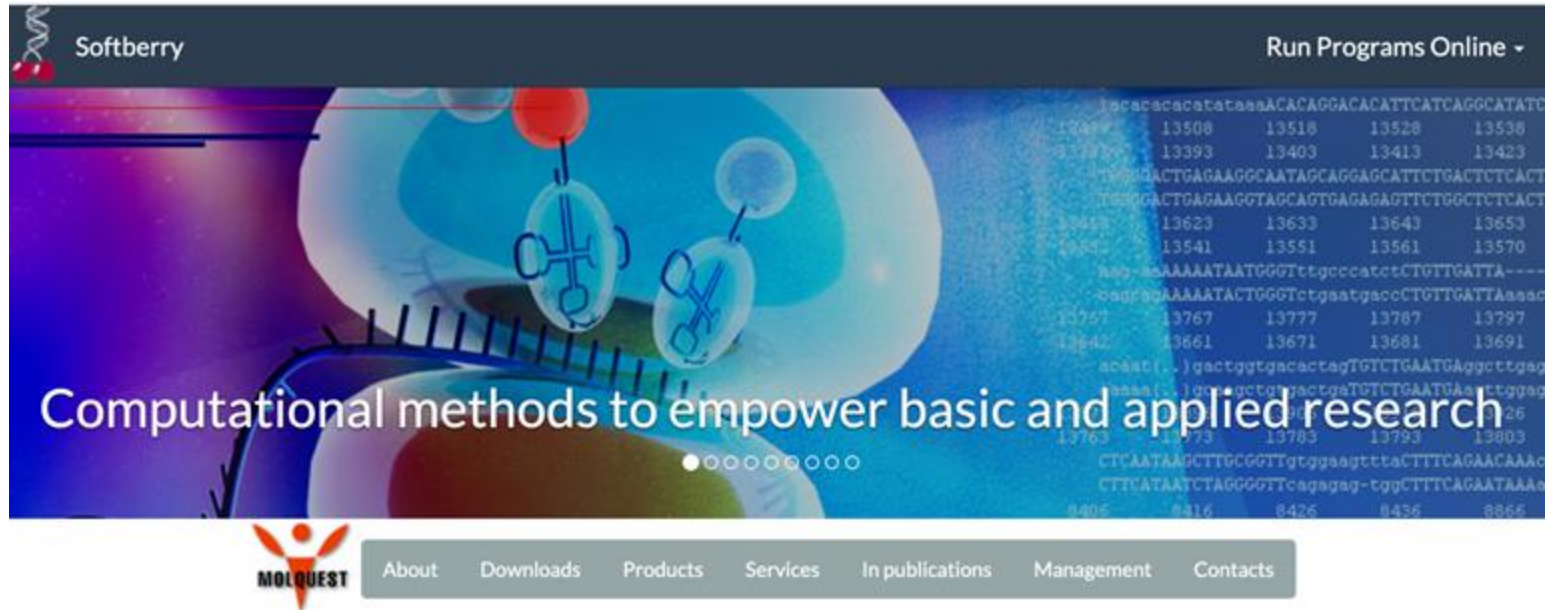
TAATATGTCCACGGGTATTGAGCATTGTACACGGGGTATTGAGCATGTAA TGAA

Predict genes by running the Viterbi algorithm for an HMM:

TAAT ATGTCCACGG GTATTGAG CATTGTACACGGG GTATTGAG CATGTAA TGAA



Making cents of hidden Markov models



The image is a screenshot of the Softberry website banner. The banner features a dark blue background with a stylized illustration of a DNA double helix and a protein structure. The text "Computational methods to empower basic and applied research" is prominently displayed in white. The Softberry logo is in the top left, and a "Run Programs Online" link is in the top right. A navigation bar at the bottom contains links to "About", "Downloads", "Products", "Services", "In publications", "Management", and "Contacts".

Softberry

Run Programs Online ▾

Computational methods to empower basic and applied research

MOLQUEST

About Downloads Products Services In publications Management Contacts

How did we arrive at the AI winter?

In: Proc. 5th annual conf. of the Cognitive Science Society, Rochester NY, May 1983

ANALYZING COOPERATIVE COMPUTATION

Geoffrey E. Hinton
Computer Science Department
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

Terrence J. Sejnowski
Biophysics Department
The Johns Hopkins University
Baltimore, Maryland 21218

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
Washington, D. C., June, 1983

OPTIMAL PERCEPTUAL INFERENCE

Geoffrey E. Hinton
Computer Science Department
Carnegie-Mellon University

Terrence J. Sejnowski
Biophysics Department
The Johns Hopkins University

Proc. Natl. Acad. Sci. USA
Vol. 79, pp. 2554-2558, April 1982
Biophysics

Neural networks and physical systems with emergent collective computational abilities

(associative memory/parallel processing/categorization/content-addressable memory/fail-safe devices)

J. J. HOPFIELD

Division of Chemistry and Biology, California Institute of Technology, Pasadena, California 91125; and Bell Laboratories, Murray Hill, New Jersey 07974

Contributed by John J. Hopfield, January 15, 1982

Precursors of neural networks

- These papers popularized two key ideas in machine learning (AI):
 - Network structure: probability distributions structured according to a graph whose nodes could be understood as representing “neurons”
 - Learning: procedures for estimating parameters for models based on training data.
 - The primitives of Boltzmann Machines are key to current neural networks (via the Restricted Boltzmann Machine)

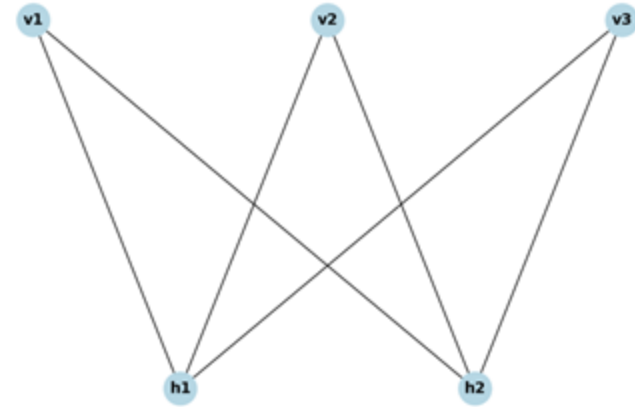
The Boltzmann Machine

- The energy function is given by

$$E = - \left(\sum_{i,j} w_{ij} v_i h_j + \sum_i b_i v_i + \sum_j c_j h_j \right)$$

weight

bias



- The states for the hidden and visible nodes are from the set $\{0,1\}$.
- For an n node Boltzmann Machine, the number of possible states is 2^n . This is the *state space*.
 - Every state configuration in the state space has an “energy”.

A probability distribution for state configurations

- The Boltzmann Machine can be viewed as an approach to defining a (structured) probability distribution over the state space.
- This is accomplished by defining the probability of a state configuration

$$P(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{v}, \mathbf{h})}$$

where Z is the *partition function*, defined to be

$$Z = \sum_{\mathbf{v}, \mathbf{h}} e^{-E(\mathbf{v}, \mathbf{h})}$$


Note that $P(\mathbf{v}, \mathbf{h})$ is *log-linear*. It is also known as a *Boltzmann distribution*

The point of a Boltzmann Machine

- With a Boltzmann machine (i.e. estimated parameters yielding a probability distribution of states), we can assess the probability of a change in hidden states, say from a 0 to a 1:

$$P(h_j = 1|\mathbf{v}) \propto e^{-E(\mathbf{v}, h_j=1)}$$

$$P(h_j = 0|\mathbf{v}) \propto e^{-E(\mathbf{v}, h_j=0)}$$

The energy function is $E(\mathbf{v}, \mathbf{h}) = -\sum_i b_i v_i - \sum_j c_j h_j - \sum_{i,j} v_i w_{ij} h_j$
so a difference in states results in

$$\Delta E = E(\mathbf{v}, h_j = 0) - E(\mathbf{v}, h_j = 1) = c_j + \sum_i w_{ij} v_i$$

The point of a Boltzmann Machine

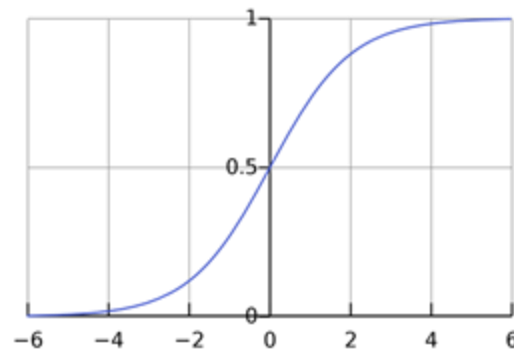
- Now the probability that a specific hidden state is 1 is

$$P(h_j = 1|\mathbf{v}) = \frac{e^{-E(\mathbf{v}, h_j=1)}}{e^{-E(\mathbf{v}, h_j=1)} + e^{-E(\mathbf{v}, h_j=0)}}$$

$$P(h_j = 1|\mathbf{v}) = \frac{1}{1 + e^{-\Delta E}} = \frac{1}{1 + e^{-(c_j + \sum_i w_{ij} v_i)}}$$

$$P(h_j = 1|\mathbf{v}) = \sigma \left(c_j + \sum_i w_{ij} v_i \right)$$

sigmoid function $\sigma(x) = \frac{1}{1 + e^{-x}}$



From the original papers

$$w_{total} = w_h e^{-w_h \bar{e}}$$

$$\begin{aligned} &= \ln \frac{p(e|h)}{p(e|\bar{h})} - \ln \frac{p(\bar{e}|h)}{p(\bar{e}|\bar{h})} \\ &= \ln \frac{p(e,h)[1-p(e)-p(h)+p(e,h)]}{[p(e)-p(e,h)][p(h)-p(e,h)]} \end{aligned} \quad (6)$$

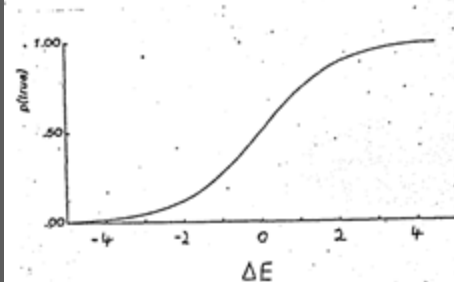
Equation 6 is symmetrical in e and h , so in solving the problem of how to make the negation of e have the correct effect on h we have also solved the second problem -- the required weights are now symmetrical. The more complicated weight in Eq. 6 does not alter the fact that the probability of a hypothesis has the form of the Boltzmann distribution for a unit with two energy states.

$$p_k = \frac{1}{(1 + e^{-\Delta E_k/T})} \quad (3)$$

where T is a parameter which acts like temperature (see fig. 1). This parallel algorithm ensures that in thermal equilibrium the relative probability of two global states is determined solely by their energy difference, and follows a Boltzmann distribution.

$$\frac{P_\alpha}{P_\beta} = e^{-(E_\alpha - E_\beta)/T} \quad (4)$$

At low temperatures there is a strong bias in favor of states with low energy, but the time required to reach equilibrium may be long. At higher temperatures the bias is not so favorable but equilibrium is reached faster.



The point of a Boltzmann Machine

- Boltzmann machine parameters are *learned* from training data. That is, the parameters are identified that maximize the probability of the training data. The word “learning” is an anthropomorphic way of saying the phrase “parameter estimation” from statistics.
 - The *Restricted Boltzmann Machine* (RBM), in which the Boltzmann Machine graph is bipartite with all the visible nodes on one side, and the hidden nodes on the other, facilitates a fast parameter estimation algorithm.
 - Fast parameter estimation is possible via *backpropagation*, discovered by Rumelhart in the 1980s, via ideas in control theory from the 1950s and 1960s.

Backpropagation and the brain

Perspective | Published: 17 April 2020

Backpropagation and the brain

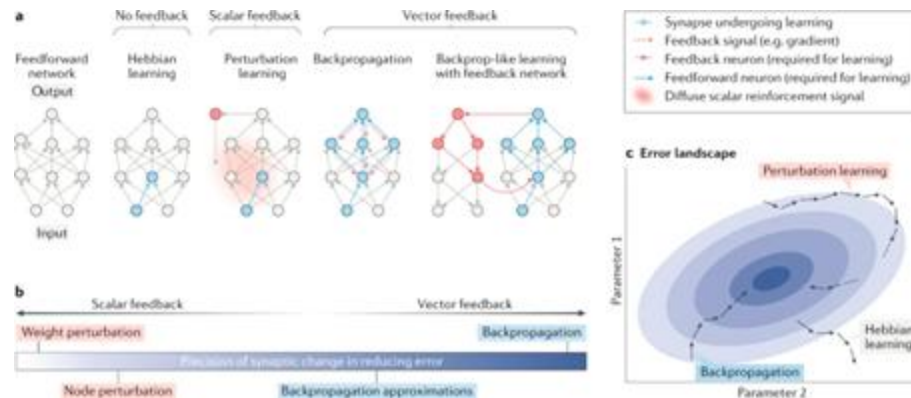
Timothy P. Lillicrap , Adam Santoro, Luke Marris, Colin J. Akerman & Geoffrey Hinton 

Nature Reviews Neuroscience 21, 335–346 (2020) | [Cite this article](#)

42k Accesses | 288 Citations | 514 Altmetric | [Metrics](#)

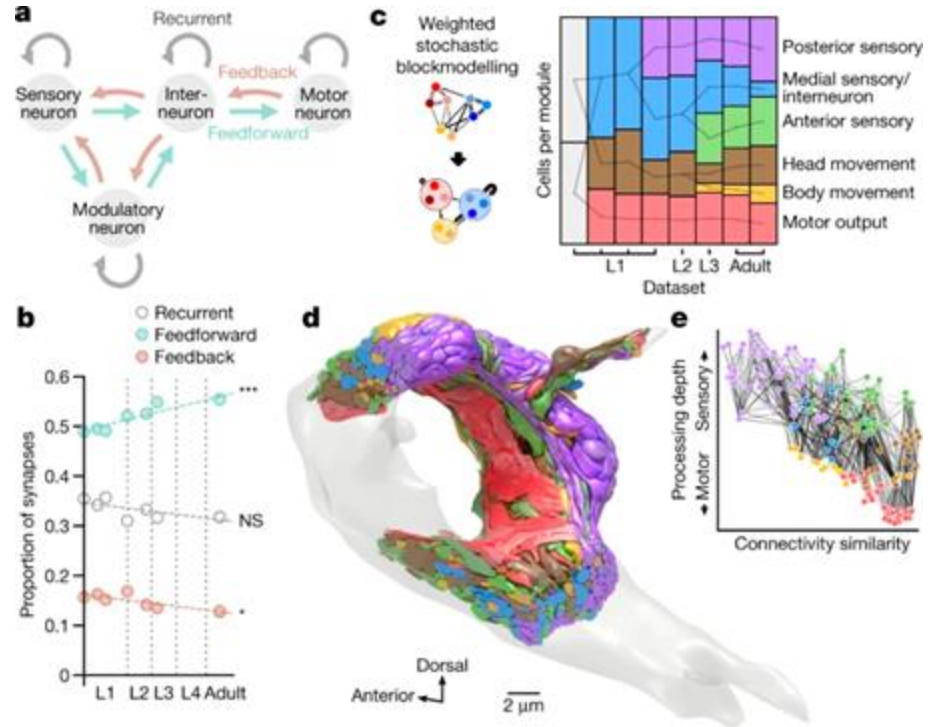
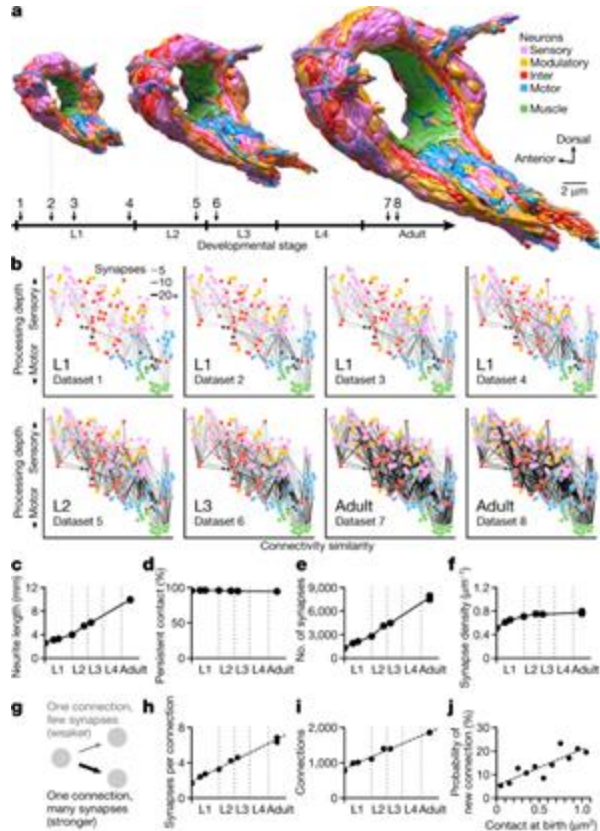
Abstract

During learning, the brain modifies synapses to improve behaviour. In the cortex, synapses are embedded within multilayered networks, making it difficult to determine the effect of an individual synaptic modification on the behaviour of the system. The backpropagation algorithm solves this problem in deep artificial neural networks, but historically it has been viewed as biologically problematic. Nonetheless, recent developments in neuroscience and the successes of artificial neural networks have reinvigorated interest in whether backpropagation offers insights for understanding learning in the cortex. The backpropagation algorithm learns quickly by computing synaptic updates using feedback connections to deliver error signals. Although feedback connections are ubiquitous in the cortex, it is difficult to see how they could deliver the error signals required by strict formulations of backpropagation. Here we build on past and recent developments to argue that feedback connections may instead induce neural activities whose differences can be used to locally approximate these signals and hence drive effective learning in deep networks in the brain.



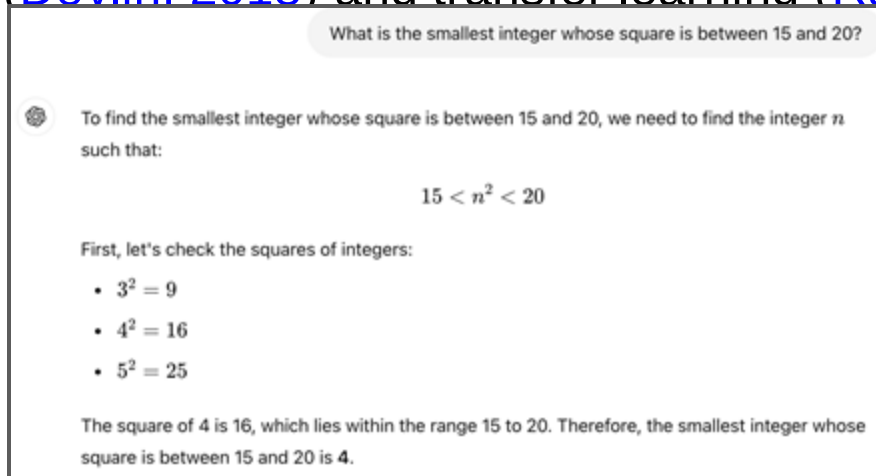
[Lillicrap et al., 2020](#)

The best understood brain: *C. elegans* brain



The AI tool of today

We have all experienced *large language models* based on transformers ([Devlin, 2018](#)) and transfer learning ([Raffel, 2019](#)) via chatGPT:



The possibilities of such AI for navigating the morass of biological data are not hard to imagine.

What is “biomedical data”?

- Sequence data is readily available due to vast amounts of publicly shared sequence reads. The sequence read archive (SRA) contains >30 petabytes of sequence.
- The protein data bank (PDB) contains more than 220,000 protein structures.
- Imaging data both from biology experiments and medical testing.
- Electronic health records.
- Literature.

Tremendous opportunity. Enormous challenges.

- Why do transformers work well for natural language?

Journalism Quarterly

DEVOTED TO RESEARCH STUDIES IN THE FIELD OF MASS COMMUNICATIONS

FALL 1953

"Cloze Procedure": A New Tool For Measuring Readability

BY WILSON L. TAYLOR*

Here is the first comprehensive statement of a research method and its theory which were introduced briefly during a workshop at the 1953 AEJ convention. Included are findings from three pilot studies and two experiments in which "cloze procedure" results are compared with those of two readability formulas.

The same principle applies to language. Given "Chickens cackle and _____ quack," almost anyone can instantly supply "ducks." If that word really is the same as the one omitted, the person scores one cloze unit for correctly closing the gap in the language pattern.

Note that the sentence pattern is a complex one made up of many sub-patterns. One must know not only the meanings (i.e., patterns of symbol-meaning relationships) and forms (patterns of letters) of all the five words, but also the meanings of given combinations of them—plus the fact that the sentence structure seems to demand a term parallel to "cackle" but associated with ducks instead of chickens. In other words, one must guess what the mutilated sentence means as a whole, then complete its pattern to fit that whole meaning.

Consider the SRA

- Individual datasets consist of massive numbers of sequence reads from complex experiments with poorly structured, and frequently missing, metadata.
- The purpose of reads is **bean counting**. The sequence content is not necessarily very informative.
- Our models of the molecular biology of the cell are nascent, and vastly under parameterized.

The origin of modern genomics: Wold and Myers, 2008

COMMENTARY

Nature Methods **5**, 19 - 21 (2008)

Published online: 19 December 2007 | doi:10.1038/nmeth1157

Sequence census methods for functional genomics

Barbara Wold¹ & Richard M Myers²

1. Barbara Wold is in the Division of Biology, California Institute of Technology, Pasadena, California 91125, USA. e-mail: bwold@caltech.edu

2. Richard M. Myers is in the Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA. e-mail: myers@shgc.stanford.edu

Next-generation sequencing technologies are beginning to facilitate genome sequencing. But in addition, new applications and new assay concepts have emerged that are vastly increasing our ability to understand genome function.

When Thomas Edison invented the phonograph, playing music was well down his list of possible uses. Something similar may now be happening in the genome world.

The human genome project

- 1953 - Watson & Crick discover the structure of the DNA double helix
- 1975 - First complete DNA genome sequenced ([φX174](#))
- 1977 - Maxam, Gilbert and Sanger publish sequencing protocols
- 1980 - Sanger and Gilbert receive Nobel prize
- 1987 - First automated sequencing machine (ABI 370)
- 1990 - Human genome project launched
- **1995 - Venter and Smith publish the genome of Haemophilus Influenzae**
- 1998 - Celera: intent to sequence human genome by shotgun method
- 2001 - Draft of human genome(s) published
- 2003 - Human genome “finished”. Cost of project: **\$3 billion**.
- **2020 - Nebula Genomics (w/ BGI) announces \$300 human genome.**



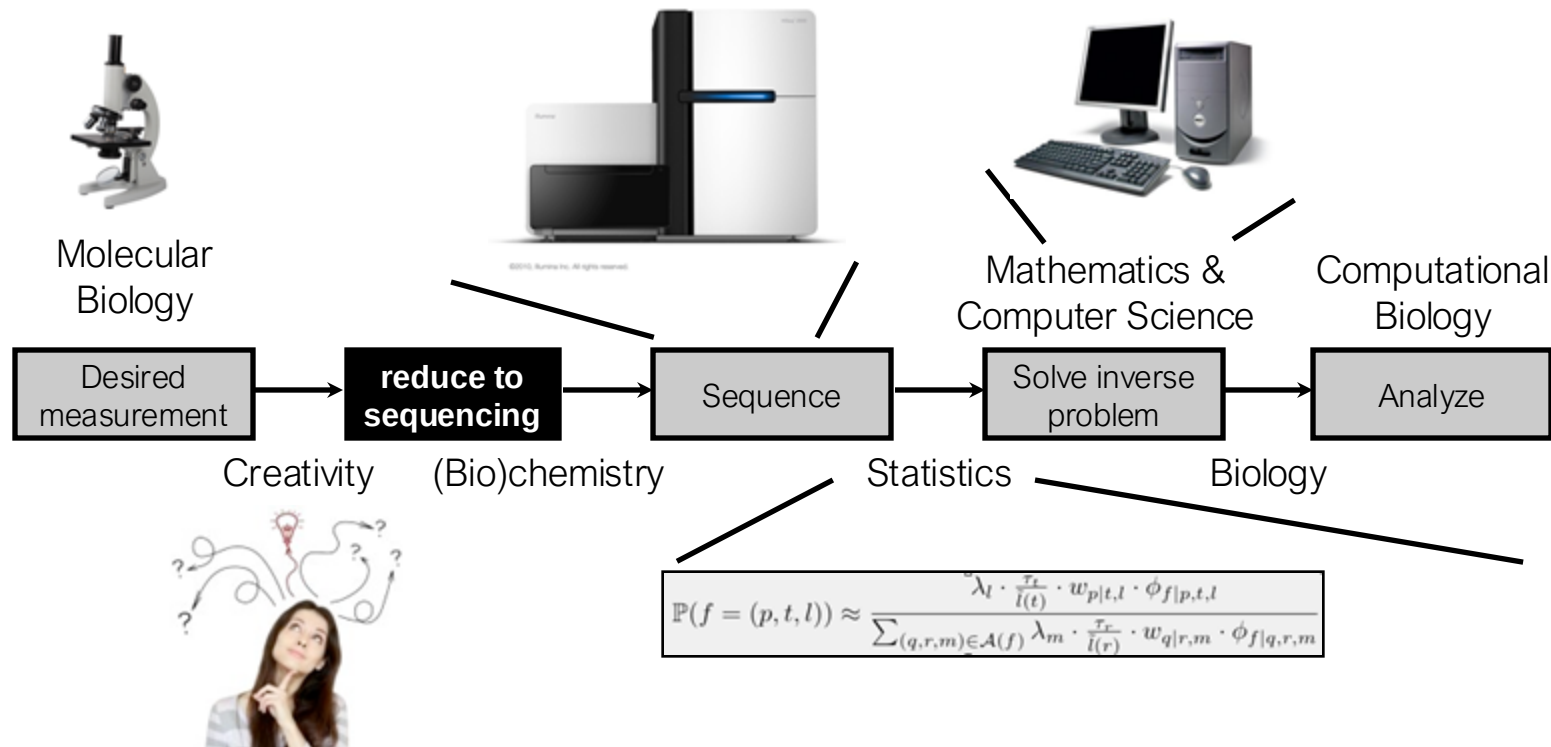
An unexpected application

- Instead of being applied solely for the purpose of determining DNA sequence, high throughput sequencers have come to be utilized as **bean counters**. Their use for this purpose is termed “sequence census”, or * Seq.
- *** Seq experiments have become the predominant application of sequencing.**
- The number of * Seq assays is now in the thousands.
- Main applications are trisomy 21 detection and RNA-Seq, an experiment for measuring *gene expression*.

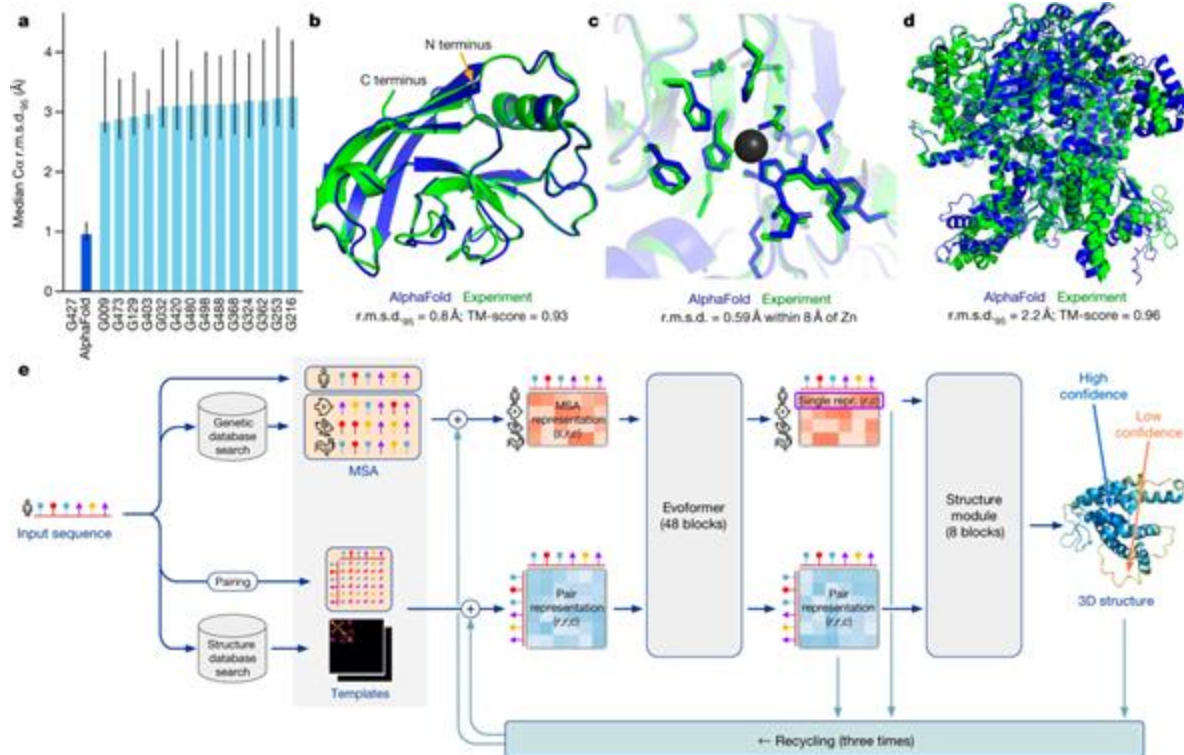


Thomas Edison

Measuring by sequencing (sequence census)

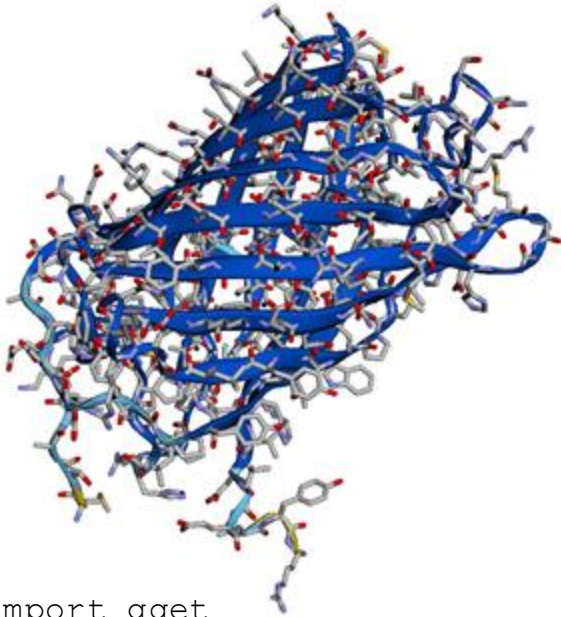


A contrast with protein sequence and structure database that enabled successful use of AI for protein folding (AlphaFold)

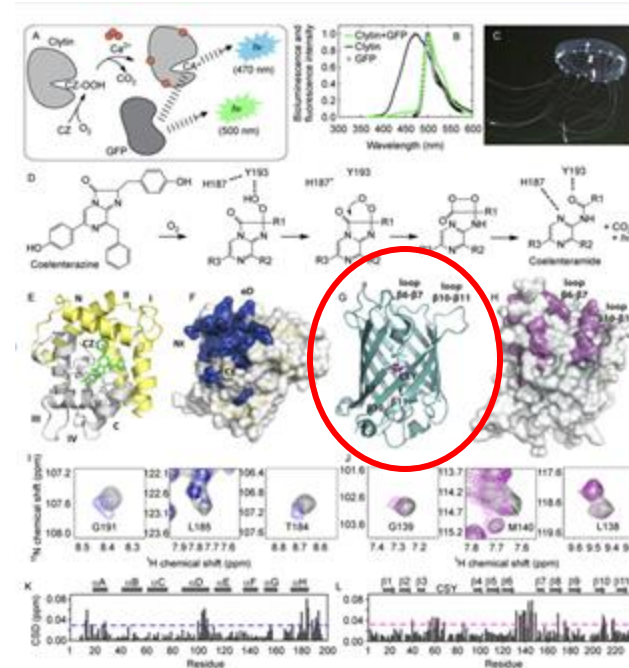


[Jumper et al., 2021](#)

Beyond genomics

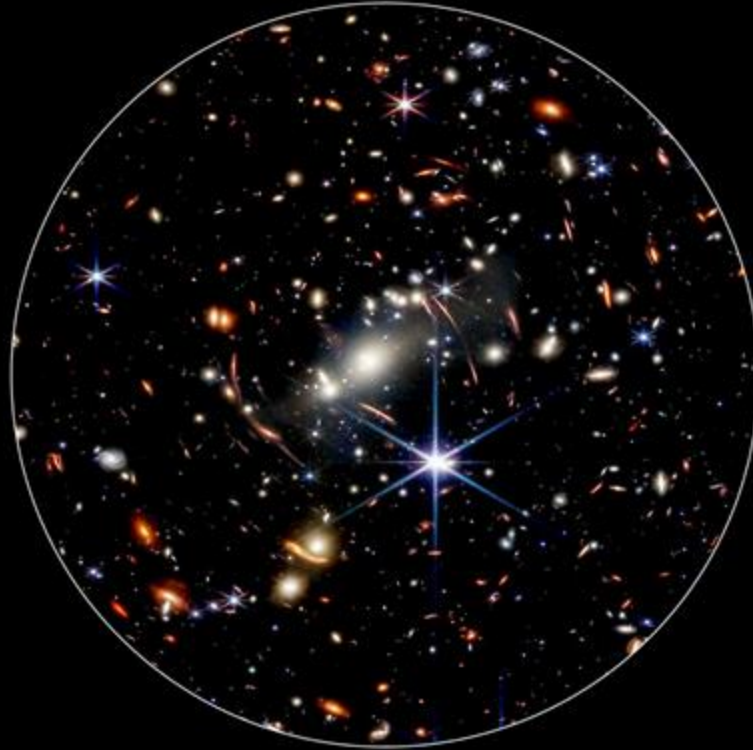


```
import gget
gget.setup("alphafold")
gget.alphafold("<sequence>")
https://github.com/pachterlab/gget
```



[Titushin et al., 2011](#)

To conclude: a vision for AI in biology



Viruses by the numbers

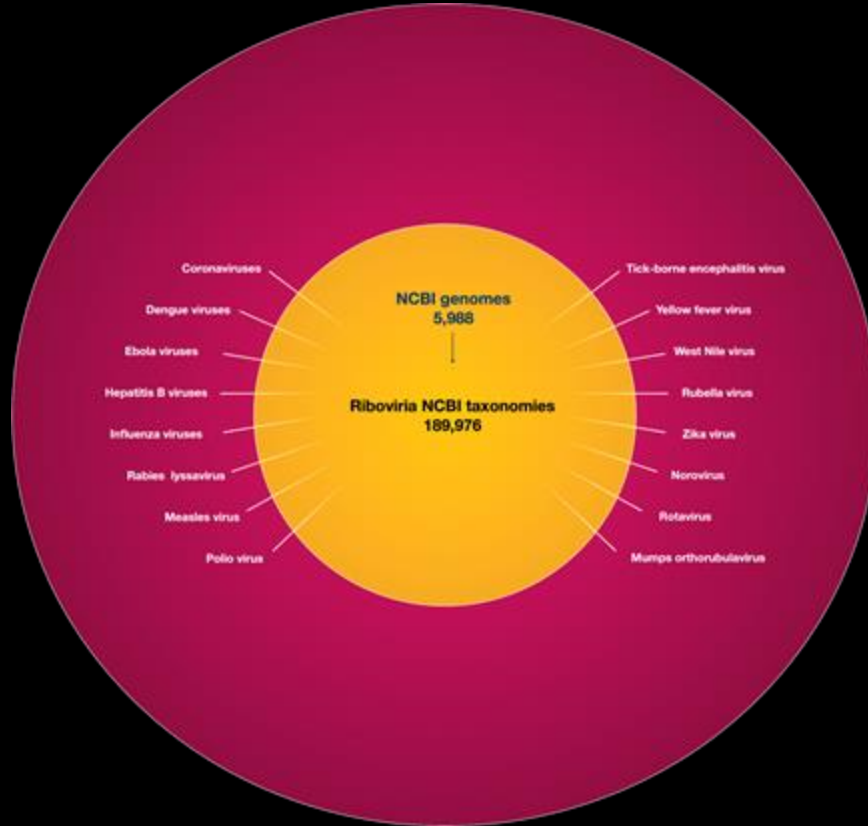


Stars in the known universe
 10^{22} - 10^{24}

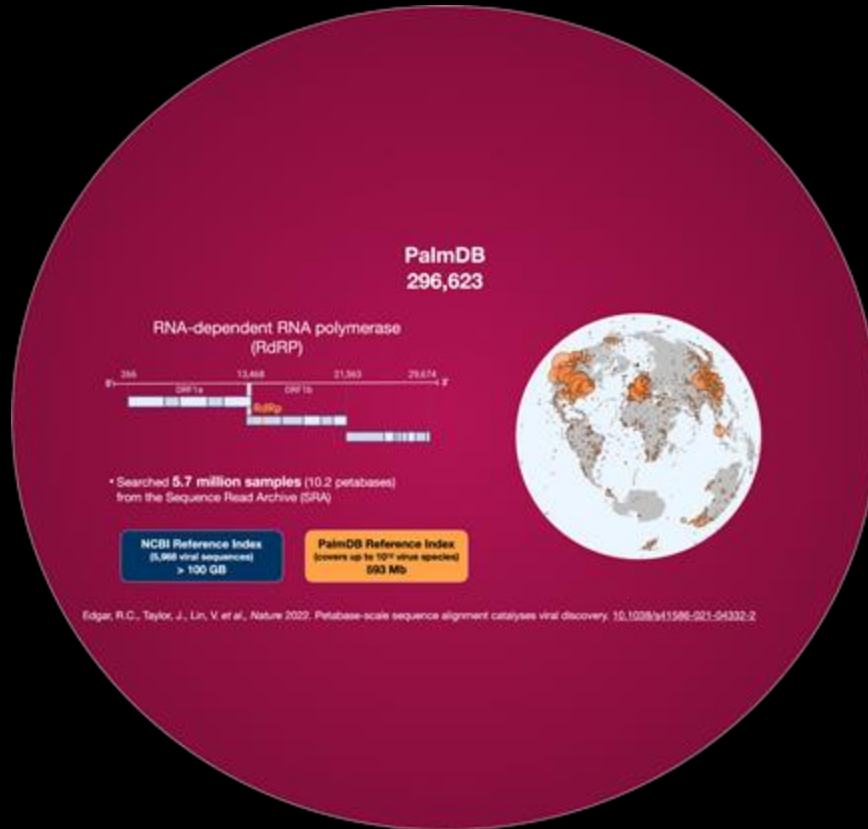
Viruses by the numbers



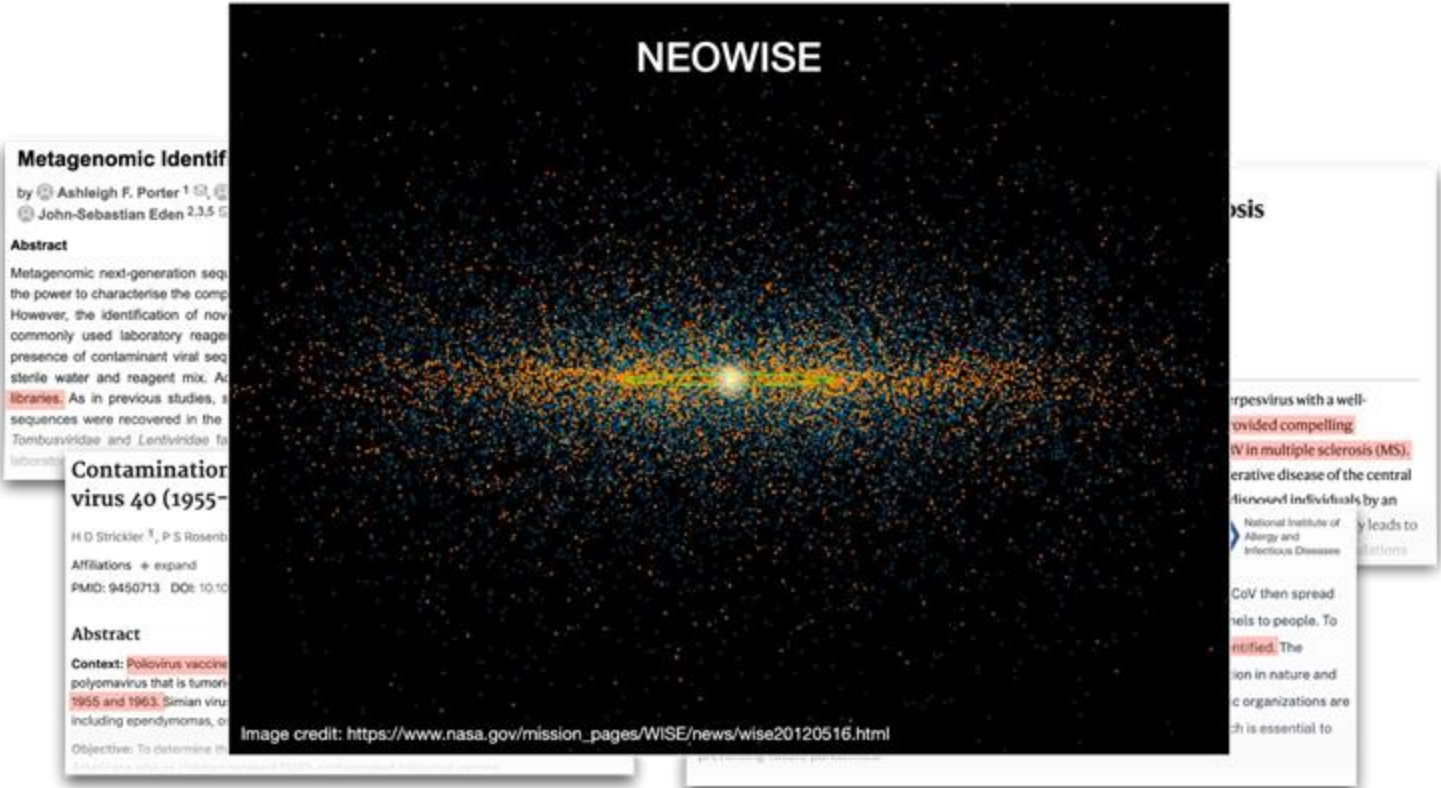
Viruses by the numbers



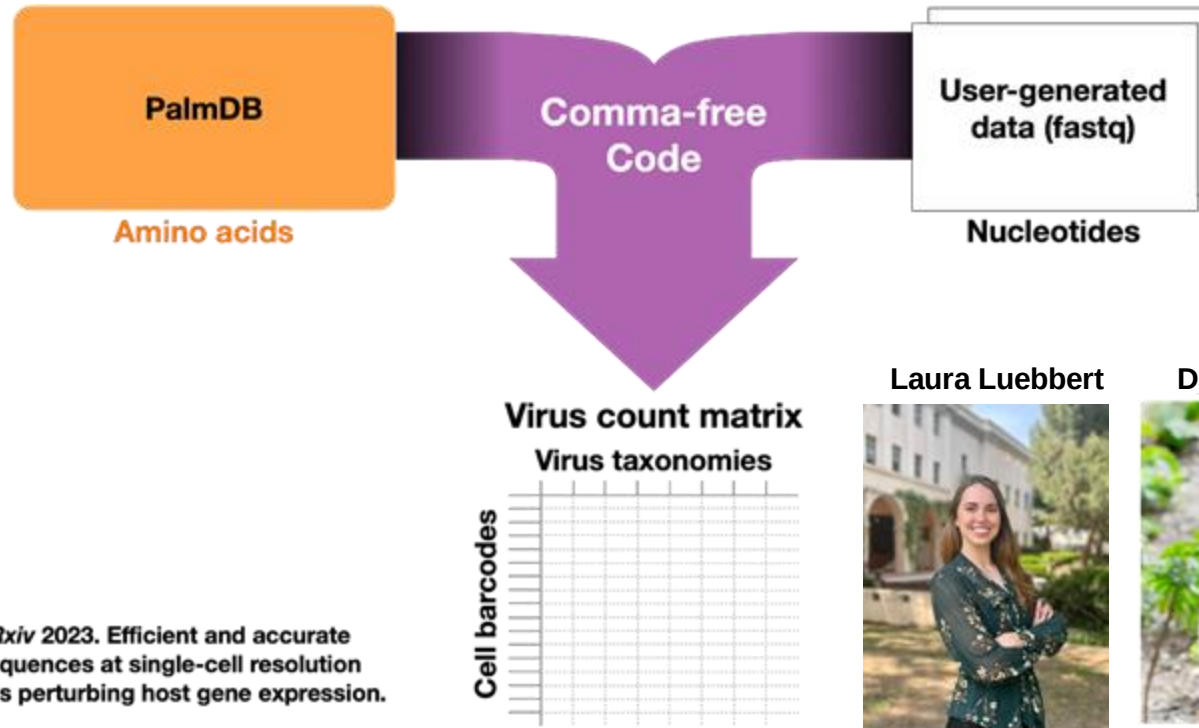
Viruses by the numbers



Imagining an early detection program for viruses



Translated search with kallisto



Luebbert et al., *bioRxiv* 2023. Efficient and accurate detection of viral sequences at single-cell resolution reveals novel viruses perturbing host gene expression. [2023.12.11.571168](https://doi.org/10.1101/2023.12.11.571168)

Laura Luebbert



Delaney Sullivan



Command line

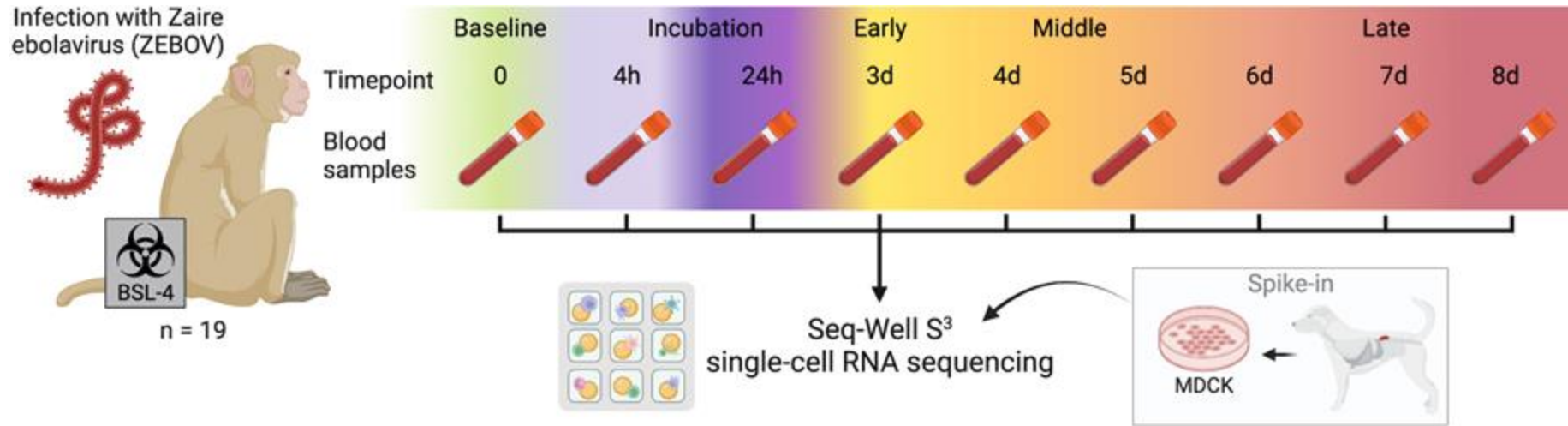
```

# 1. Install kb-python (optional: install gget to fetch the host genome and/or transcriptome)
pip install kb-python gget

# 2. Create reference index
# Optional: masking of host genomic and/or transcriptomic sequences using the D-list
# Single-thread runtime: 1.5 h; Max RAM: 4.4 GB; Size of generated index: 593 MB
# Without D-list: Single-thread runtime: 3.5 min; Max RAM: 3.9 GB; Size of generated index: 592 MB
kb ref \
  --aa \
  --d-list $(gget ref --ftp -w dna homo_sapiens) \
  -i index.idx --workflow custom \
  palmdb_rdrp_seqs.fa

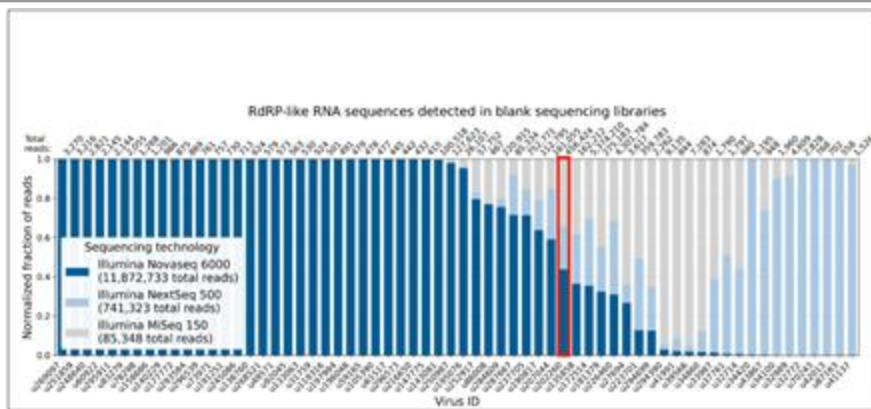
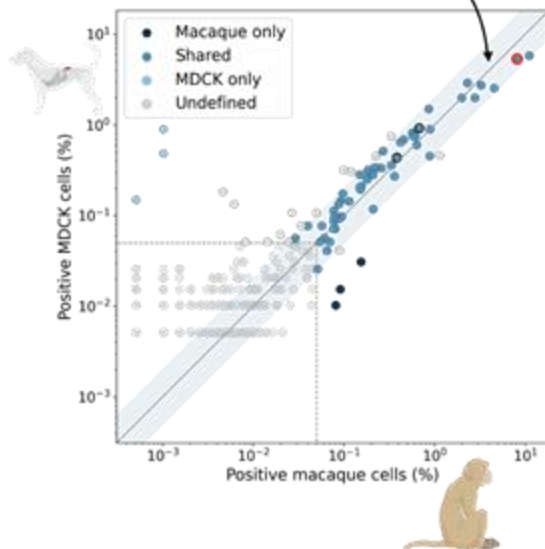
# 3. Align sequencing reads and generate count matrix
# Single-thread runtime: 1.5 min per 1 million sequences; Max RAM: 2.1 GB
kb count \
  --aa \
  -i index.idx -g palmdb_clustered_t2g.txt \
  --parity single \
  -x default \
  user_data.fastq.gz
```

The microbiome of macaques infected with Zaire ebolavirus



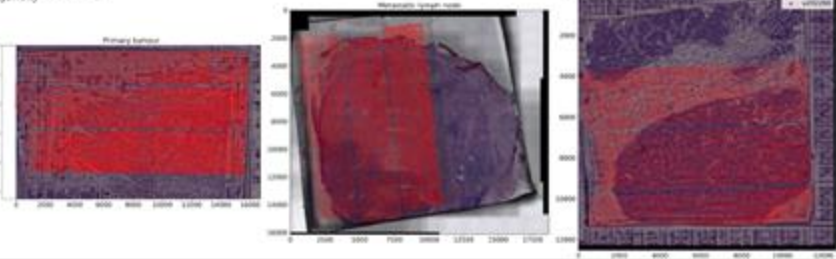
Detecting contaminants

Contamination of laboratory reagents,
cell-free RNA contamination,
widespread latent infections

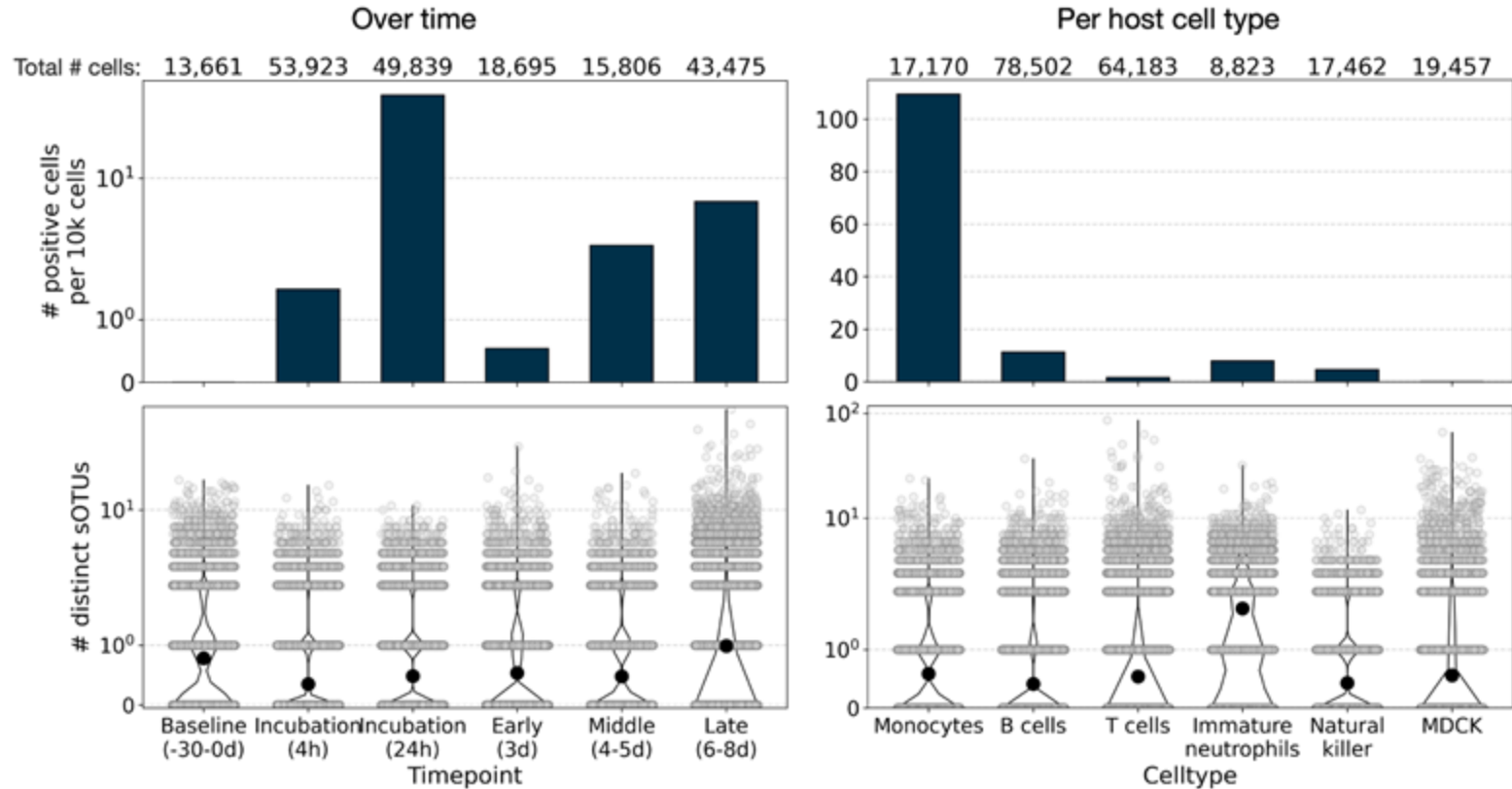


Open-ST: High-resolution spatial transcriptomics in 3D

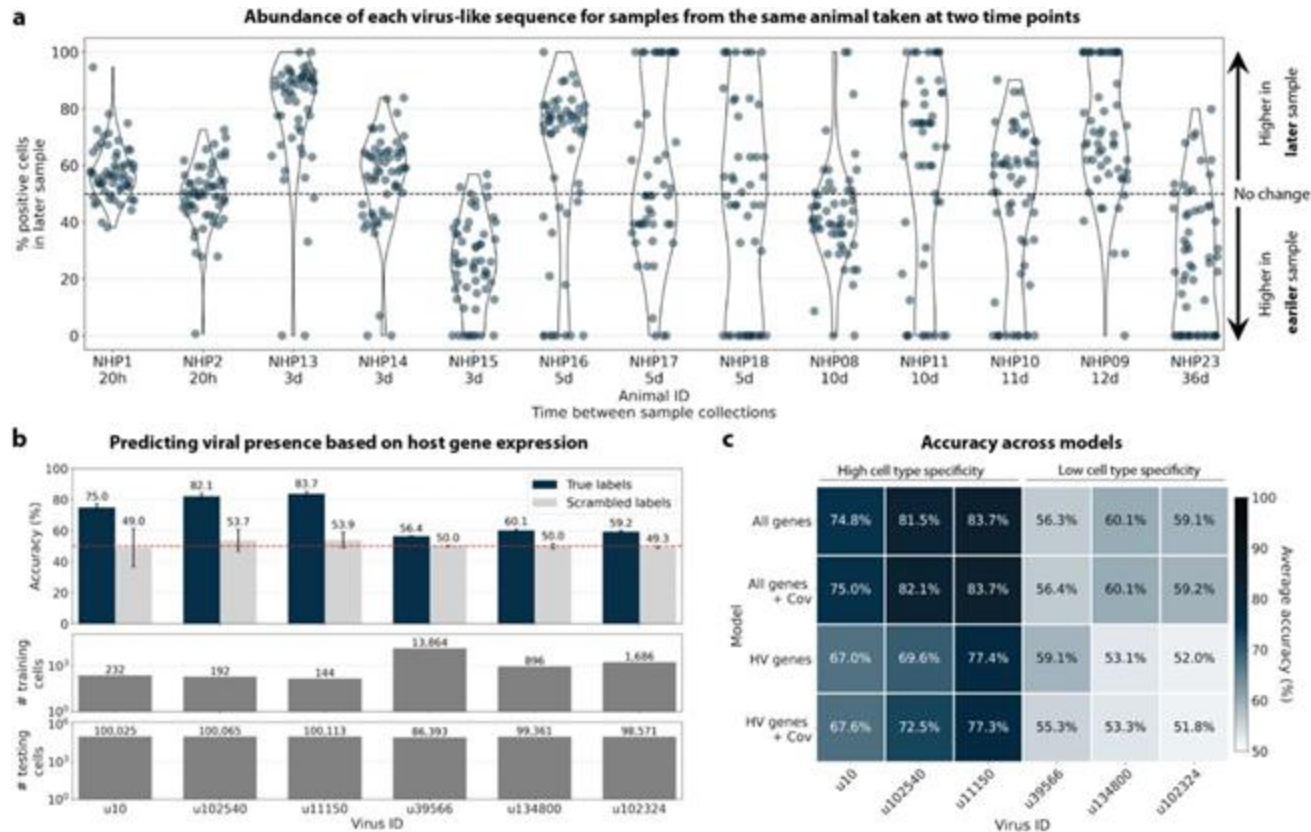
Marie Schott^{1,2}, Daniel León-Perfán^{1,2}, Elena Splendiani^{1,2,3}, Leon Strenger⁴, Jan Robin Licha⁵, Tancredi Massimo Pentimalli², Simon Schallenberg¹, Jonathan Allen¹, Sarah Samat Tagliaferro¹, Anastasiya Bohengagen¹, Sebastian Ehrig¹, Stefano Abbati^{1,4}, Steffen Donnerich¹, Maximiliano Pagani^{1,6}, Elisabetta Ferretti¹, Giuseppe Macino^{1,5}, Nikos Karaiskos^{1,6}, Nikolaus Rajewsky^{1,2,3,4,5,6}



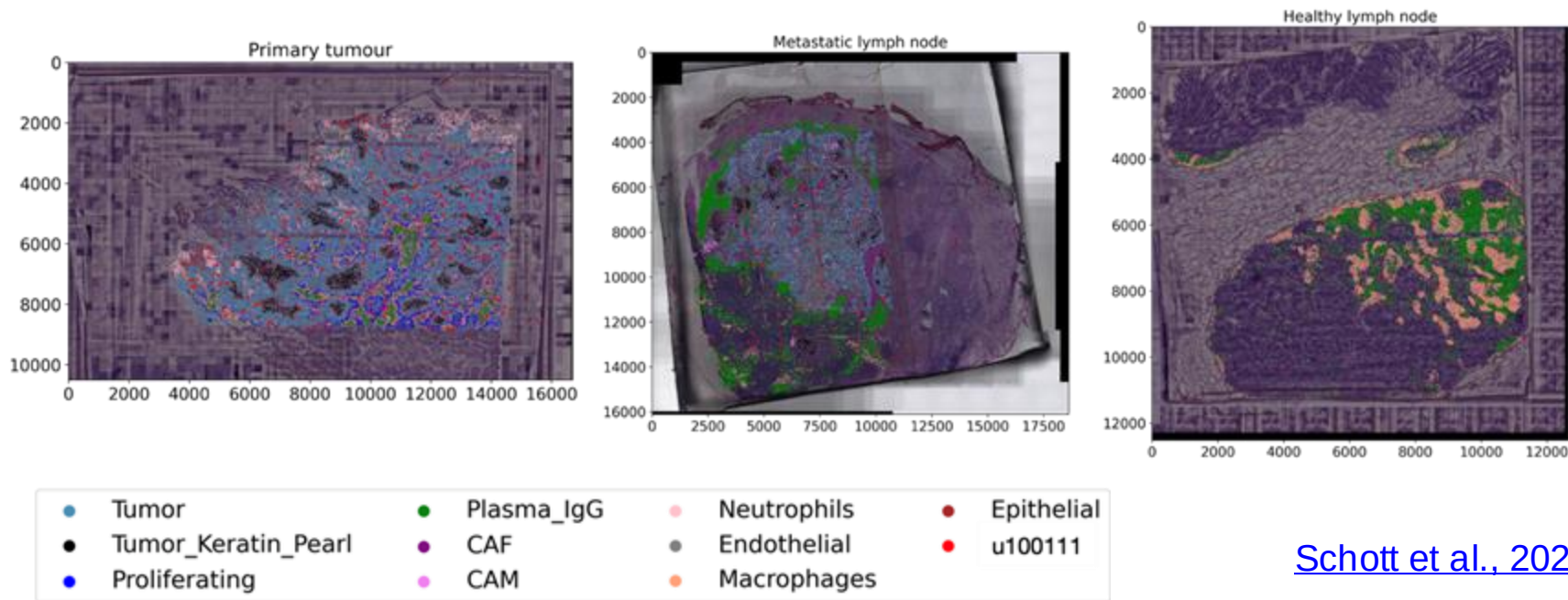
Zaire ebola virus presence in time and cell type



Prediction



Combining image analysis and sequence analysis for studying cancer



[Schott et al., 2024](#)

Contributors to the ideas and slides in this talk

