# DATA SCIENCE IN PHARMA RESEARCH

ALVARO SEBASTIAN

WOMEN IN MACHINE LEARNING & DATA SCIENCE

21 FEBRUARY 2020 - POZNAN

SIXTH RESEARCHER
www.sixthresearcher.com

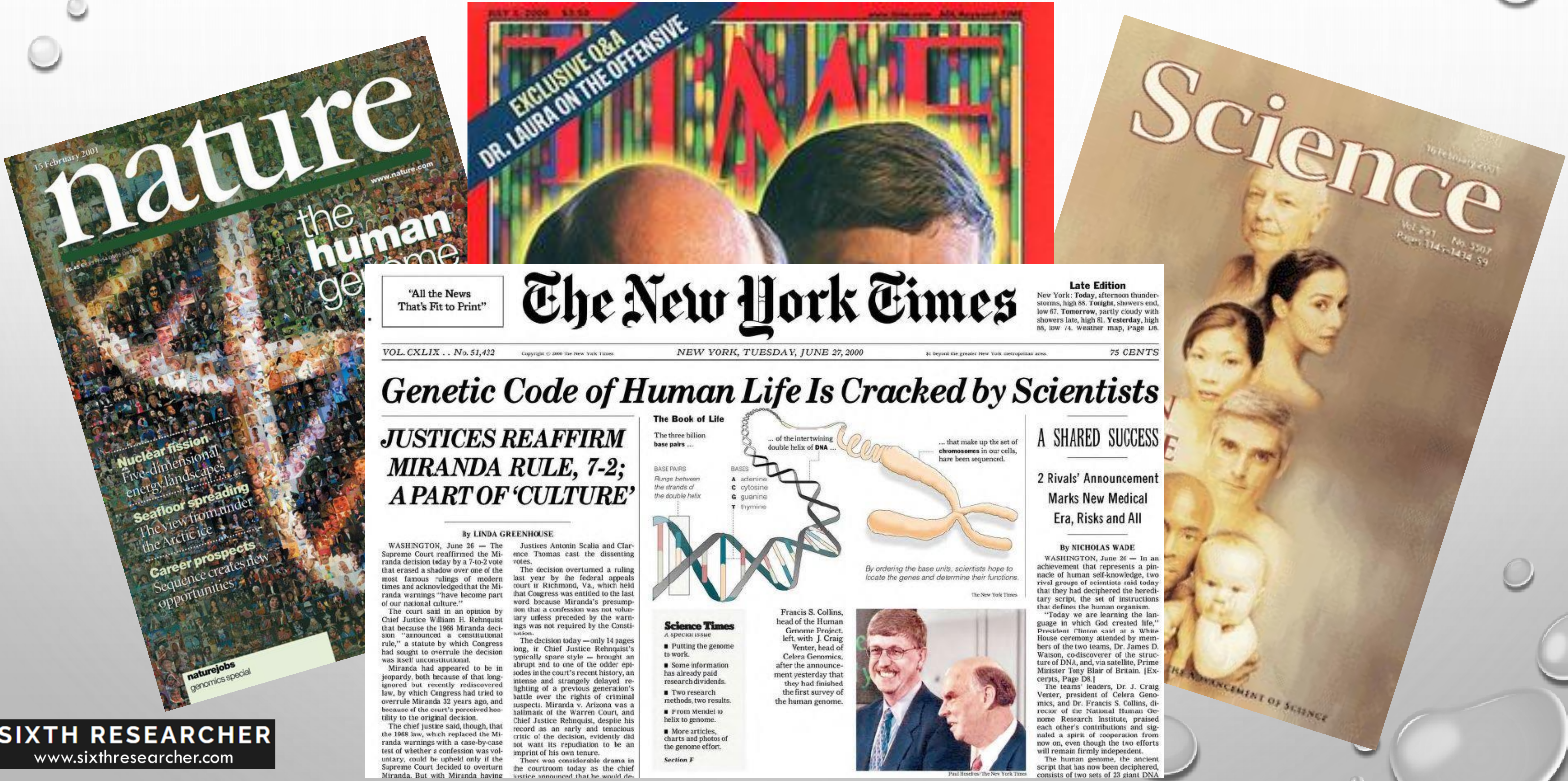
# THE ORIGINS…

# THE HUMAN GENOME

On June 26, 2000, a 'rough draft' of the genome was ann
jointly by U.S. President Bill Clinton (photo) and the British
Minister Tony Blair (via satellite).

https://en.wikipedia.org/wiki/Human_Genome_Project

# THE HUMAN GENOME



"All the News That's Fit to Print"

The New York Times

**Late Edition**

New York: **Today**, afternoon thunderstorms, high 88. **Tonight**, showers end, low 67. **Tomorrow**, partly cloudy with showers late, high 81. **Yesterday**, high 88, low 74. Weather map, Page D8.

VOL. CXLIX . . No. 51,432 — Copyright © 2000 The New York Times — NEW YORK, TUESDAY, JUNE 27, 2000 — $1 beyond the greater New York metropolitan area. — 75 CENTS

## Genetic Code of Human Life Is Cracked by Scientists

### JUSTICES REAFFIRM MIRANDA RULE, 7-2; A PART OF 'CULTURE'

By LINDA GREENHOUSE

WASHINGTON, June 26 — The Supreme Court reaffirmed the Miranda decision today by a 7-to-2 vote that erased a shadow over one of the most famous rulings of modern times and acknowledged that the Miranda warnings "have become part of our national culture."

The court said in an opinion by Chief Justice William H. Rehnquist that because the 1966 Miranda decision "announced a constitutional rule," a statute by which Congress had sought to overrule the decision was itself unconstitutional.

Miranda had appeared to be in jeopardy, both because of that long-ignored but recently rediscovered law, by which Congress had tried to overrule Miranda 32 years ago, and because of the court's perceived hostility to the original decision.

The chief justice said, though, that the 1968 law, which replaced the Miranda warnings with a case-by-case test of whether a confession was voluntary, could be upheld only if the Supreme Court decided to overturn Miranda. But with Miranda having

Justices Antonin Scalia and Clarence Thomas cast the dissenting votes.

The decision overturned a ruling last year by the federal appeals court in Richmond, Va., which held that Congress was entitled to the last word because Miranda's presumption that a confession was not voluntary unless preceded by the warnings was not required by the Constitution.

The decision today — only 14 pages long, in Chief Justice Rehnquist's typically spare style — brought an abrupt end to one of the odder episodes in the court's recent history, an intense and strangely delayed refighting of a previous generation's battle over the rights of criminal suspects. Miranda v. Arizona was a hallmark of the Warren Court, and Chief Justice Rehnquist, despite his record as an early and tenacious critic of the decision, evidently did not want its repudiation to be an imprint of his own tenure.

There was considerable drama in the courtroom today as the chief justice announced that he would de-

**The Book of Life**

The three billion **base pairs** ...

BASE PAIRS
*Rungs between the strands of the double helix*

BASES
**A** adenine
**C** cytosine
**G** guanine
**T** thymine

... of the intertwining double helix of **DNA** ...

... that make up the set of **chromosomes** in our cells, have been sequenced.

*By ordering the base units, scientists hope to locate the genes and determine their functions.*

The New York Times

**Science Times**
*A special issue*

- Putting the genome to work.
- Some information has already paid research dividends.
- Two research methods, two results.
- From Mendel to helix to genome.
- More articles, charts and photos of the genome effort.

*Section F*

Francis S. Collins, head of the Human Genome Project, left, with J. Craig Venter, head of Celera Genomics, after the announcement yesterday that they had finished the first survey of the human genome.

Paul Hosefros/The New York Times

### A SHARED SUCCESS

### 2 Rivals' Announcement Marks New Medical Era, Risks and All

By NICHOLAS WADE

WASHINGTON, June 26 — In an achievement that represents a pinnacle of human self-knowledge, two rival groups of scientists said today that they had deciphered the hereditary script, the set of instructions that defines the human organism.

"Today we are learning the language in which God created life," President Clinton said at a White House ceremony attended by members of the two teams, Dr. James D. Watson, co-discoverer of the structure of DNA, and, via satellite, Prime Minister Tony Blair of Britain. [Excerpts, Page D8.]

The teams' leaders, Dr. J. Craig Venter, president of Celera Genomics, and Dr. Francis S. Collins, director of the National Human Genome Research Institute, praised each other's contributions and signaled a spirit of cooperation from now on, even though the two efforts will remain firmly independent.

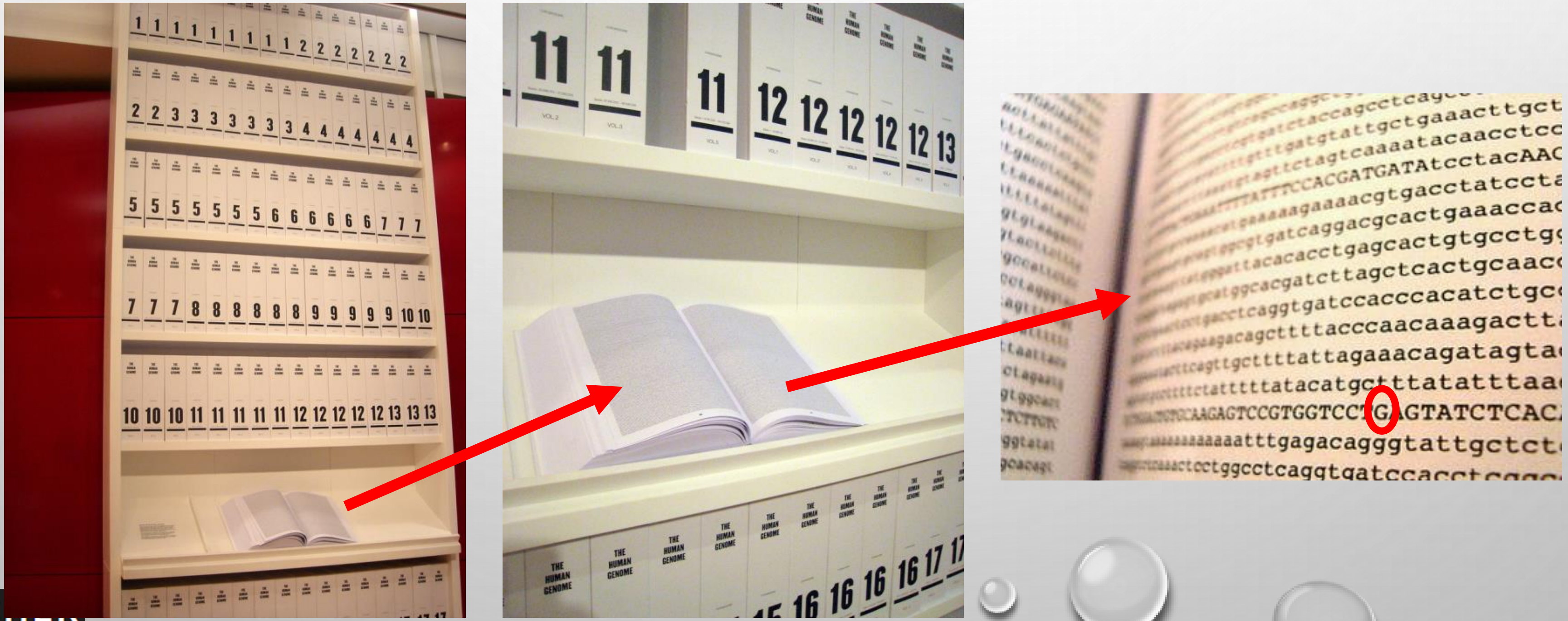
The human genome, the ancient script that has now been deciphered, consists of two sets of 23 giant DNA

# THE HUMAN GENOME

**The total length of the haploid human genome is 3.3 billion base pairs (3.3E9).**

*Don Quixote*, the Spanish novel by Miguel de Cervantes contains around 2 million of letters, so **the human genome has as many letters as 1500 copies of *Don Quixote*.**

There are a lot of letters in that stack and a lot of information that we are trying to understand. For example, **a genetic disease is like having a typo in one of those copies of *Don Quixote*.**

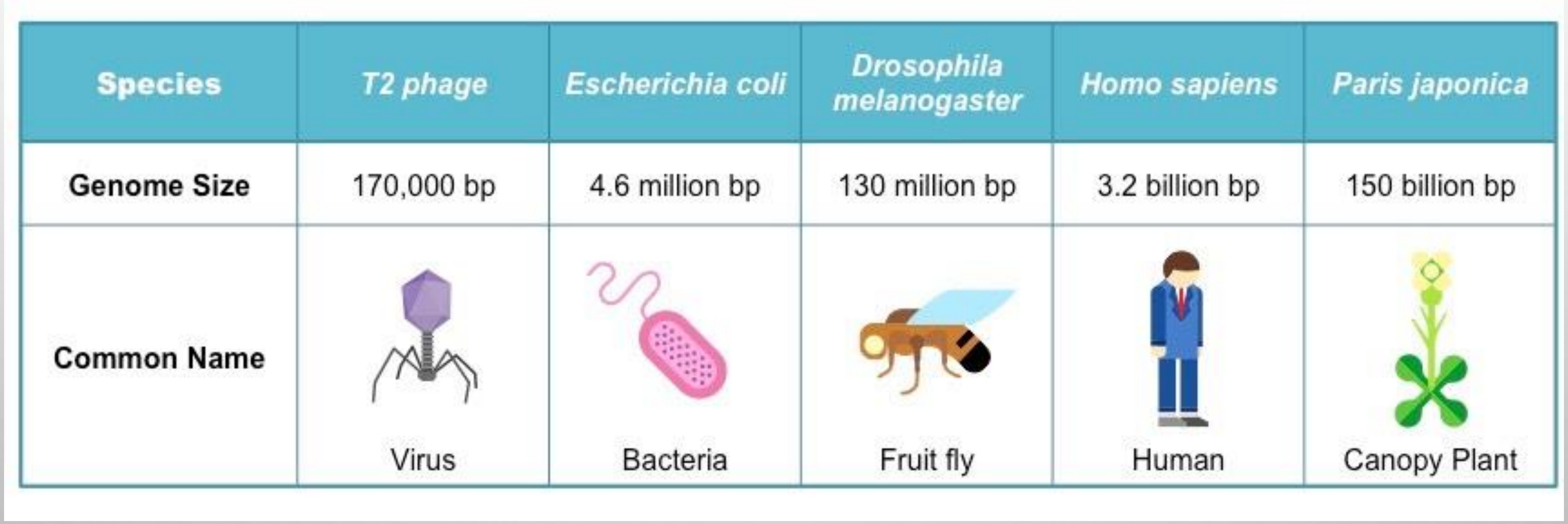


# THE HUMAN GENOME

**The human genome occupies around 750 Megabytes that is about 1 CD of space.**

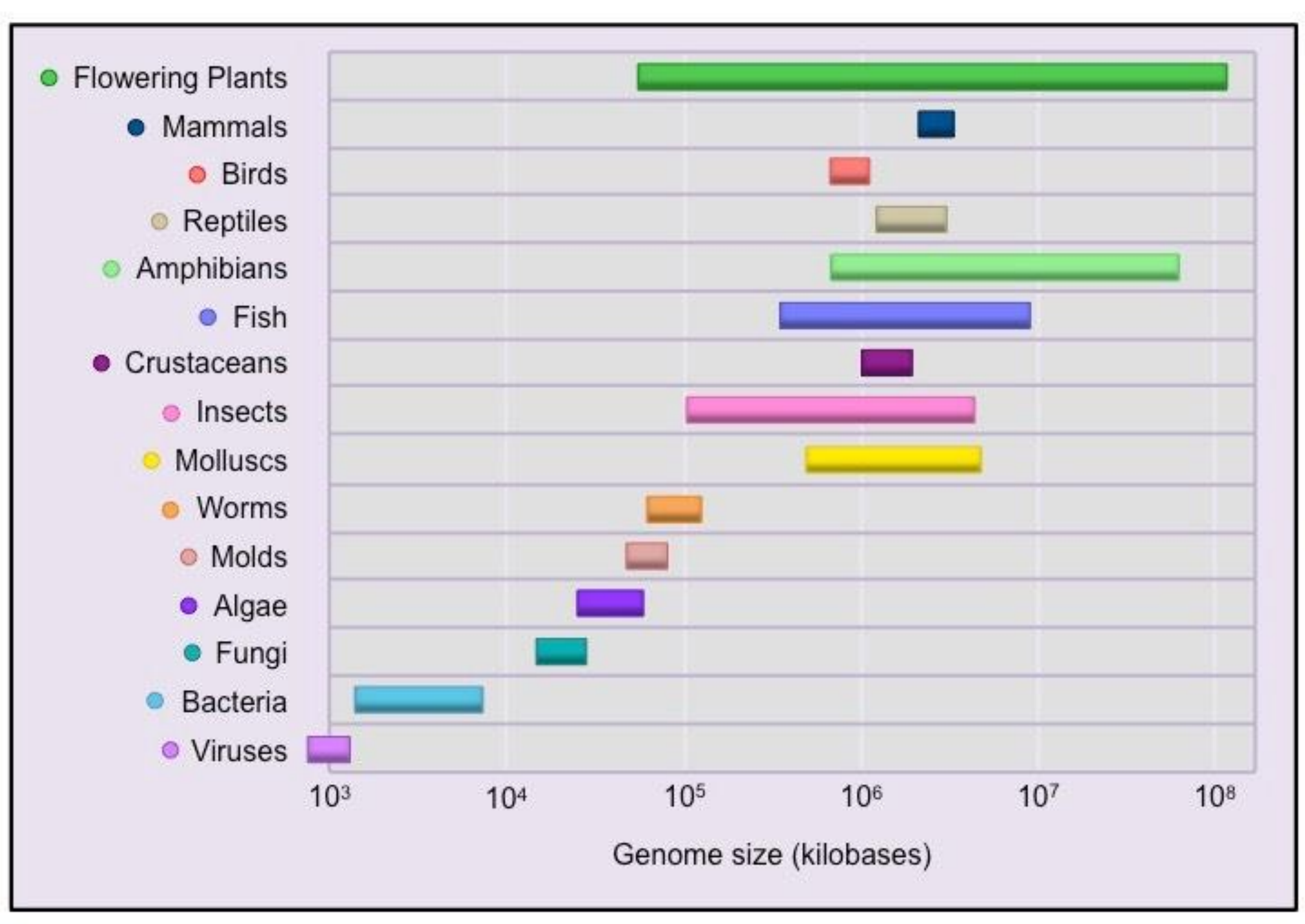
3×10^9 base pairs/haploid genome x 2 bits/base pairs x 1 byte/8 bits = 0.75E9 bytes

**That is nothing! The Canopy Plant Genome is 50 times bigger!**



| Species | *T2 phage* | *Escherichia coli* | *Drosophila melanogaster* | *Homo sapiens* | *Paris japonica* |
|---|---|---|---|---|---|
| **Genome Size** | 170,000 bp | 4.6 million bp | 130 million bp | 3.2 billion bp | 150 billion bp |
| **Common Name** | Virus | Bacteria | Fruit fly | Human | Canopy Plant |

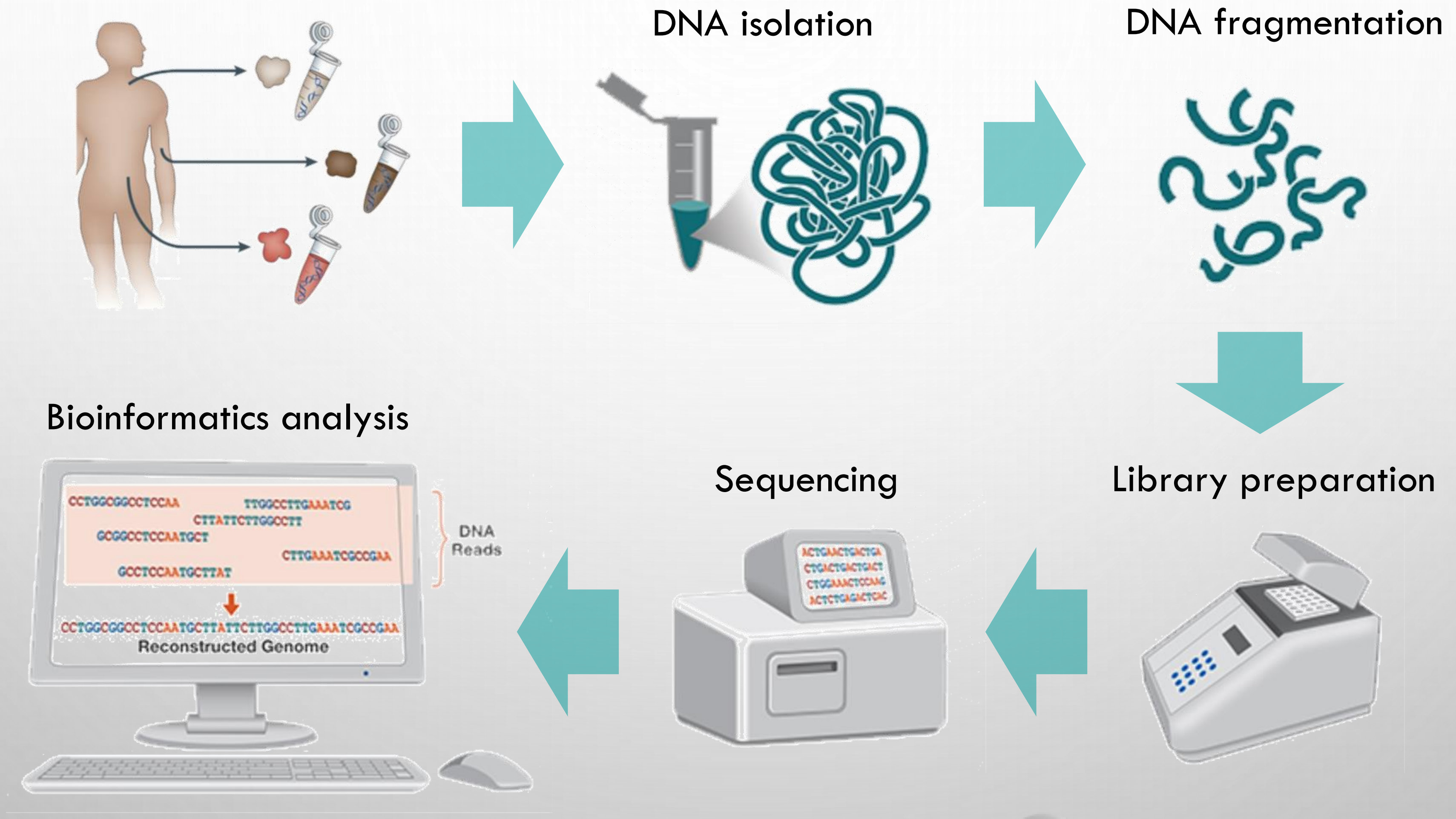
# THE HUMAN GENOME



SIXTH RESEARCHE
www.sixthresearcher.com

# THE METHODS…

# GENOME SEQUENCING



DNA isolation

DNA fragmentation

Library preparation

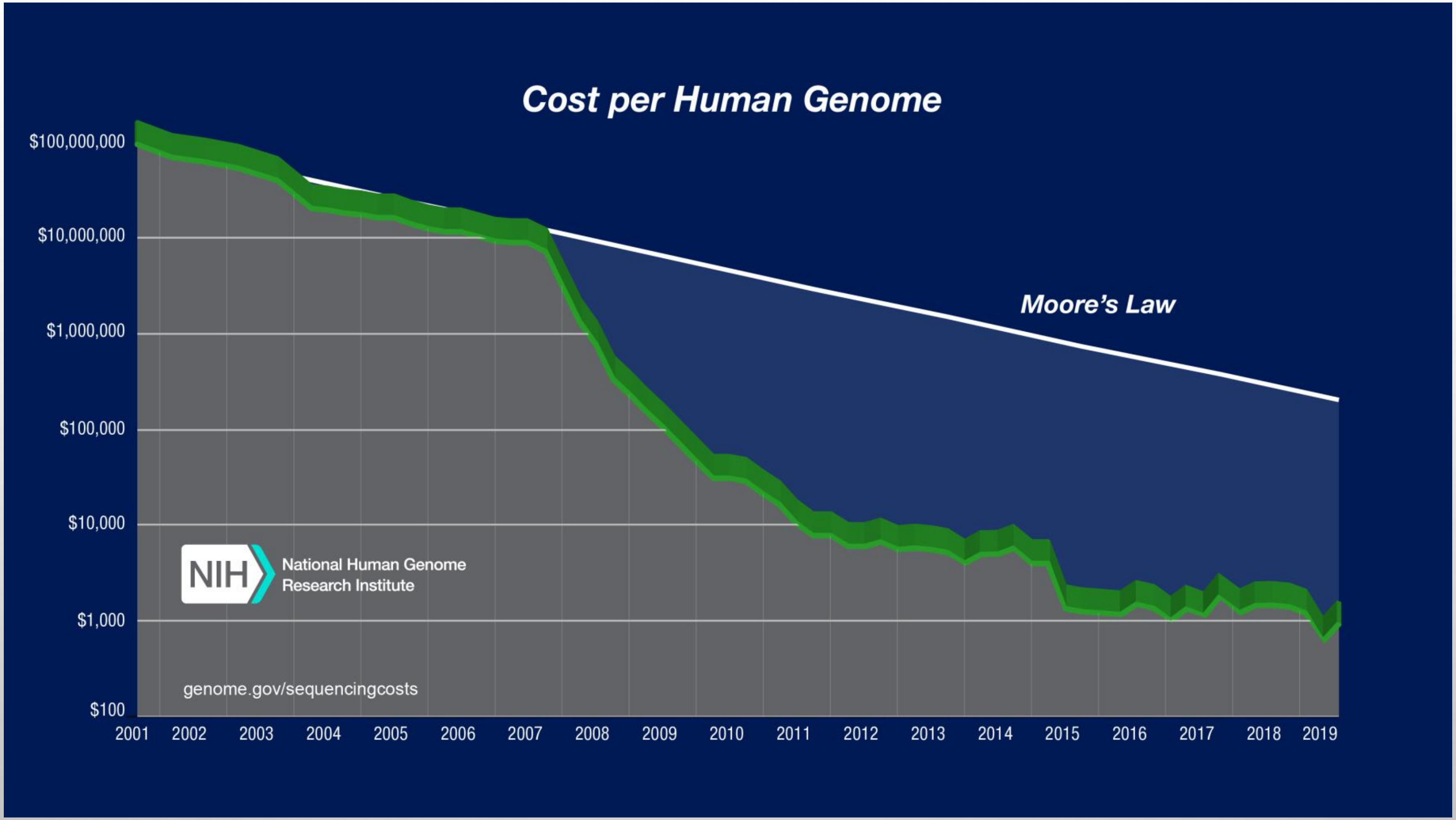
Sequencing

Bioinformatics analysis

# GENOME SEQUENCING

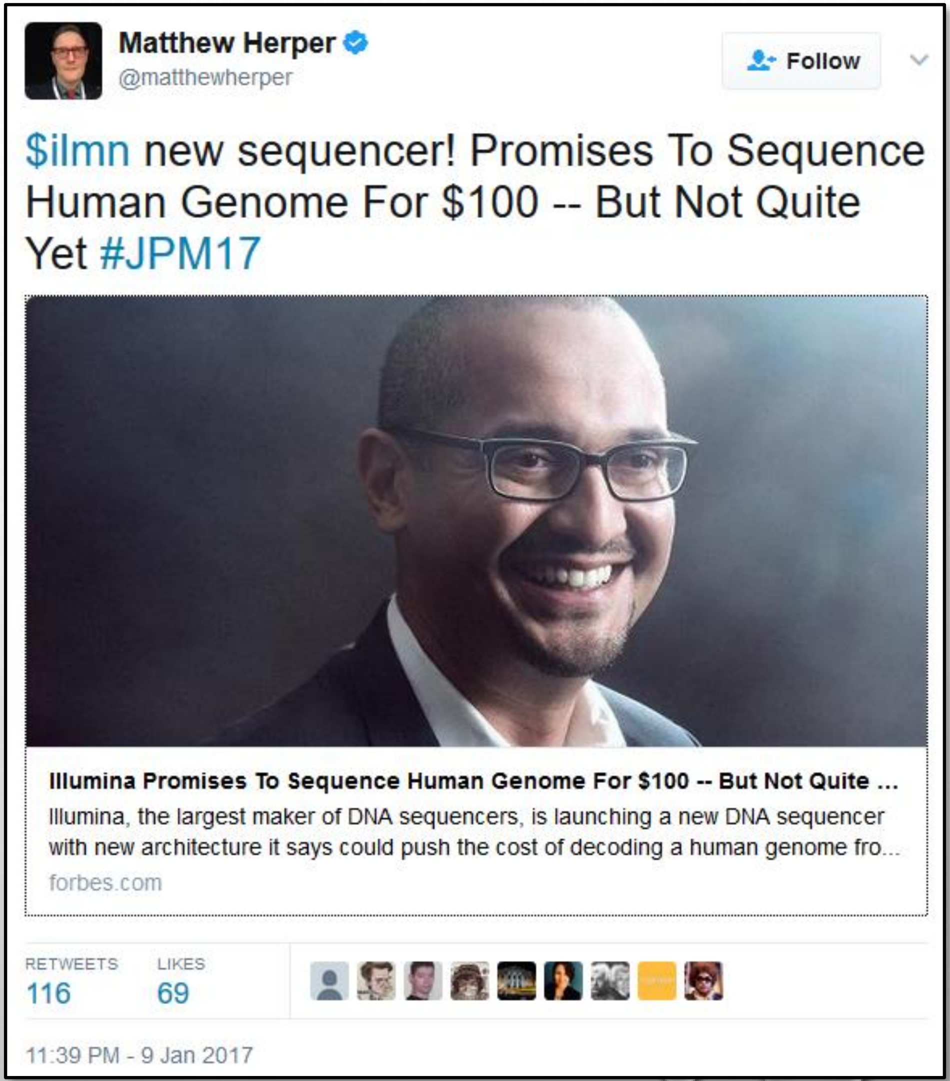
# THE PRICE…

# HUMAN GENOME COST



https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

# HUMAN GENOME COST



Matthew Herper
@matthewherper

$ilmn new sequencer! Promises To Sequence Human Genome For $100 -- But Not Quite Yet #JPM17

**Illumina Promises To Sequence Human Genome For $100 -- But Not Quite ...**
Illumina, the largest maker of DNA sequencers, is launching a new DNA sequencer with new architecture it says could push the cost of decoding a human genome fro...
forbes.com

RETWEETS 116
LIKES 69

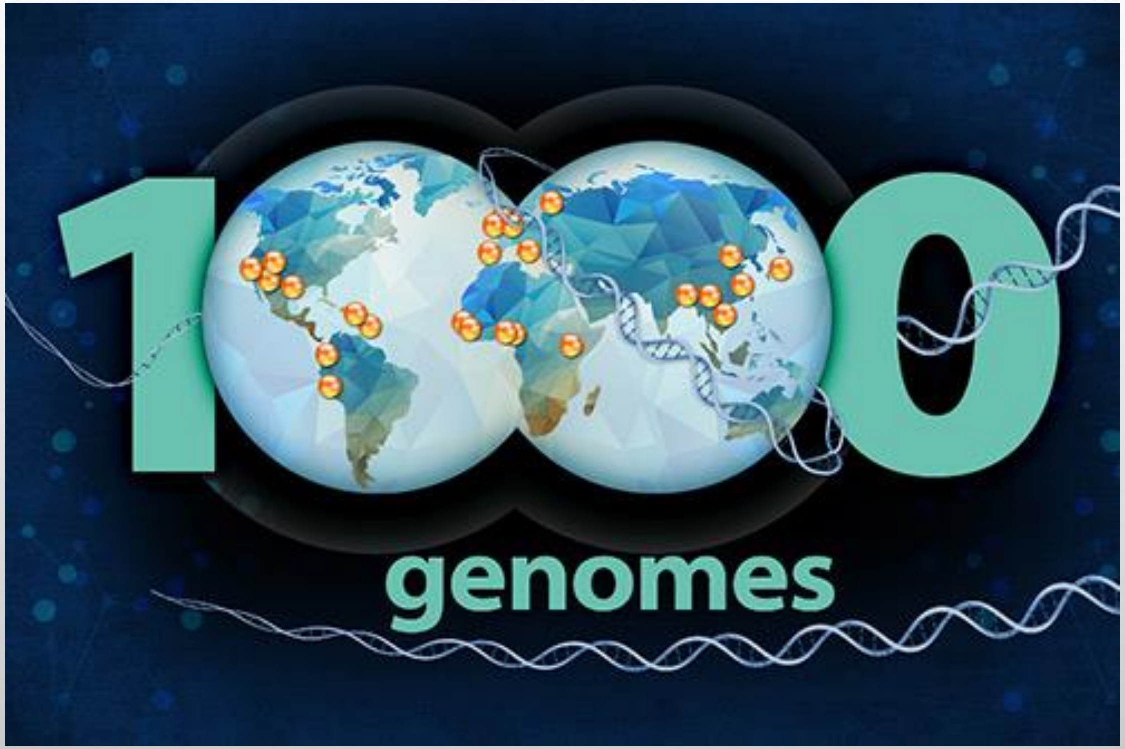
11:39 PM - 9 Jan 2017

# MORE GENOMES…

# THE 1000 GENOMES PROJECT

The 1000 Genomes Project, launched in January 2008, consisted in sequencing the genomes of at least one thousand anonymous participants from a number of different ethnic groups.

In 2012, the sequencing of 1092 genomes was announced in a Nature publication



https://en.wikipedia.org/wiki/1000_Genomes_Project

SIXTH RESEARCHER
www.sixthresearcher.com

# THE 100,000 GENOMES PROJECT

The project was established by the UK government to sequence 100,000 genomes from NHS patients affected by a rare disease, or cancer.

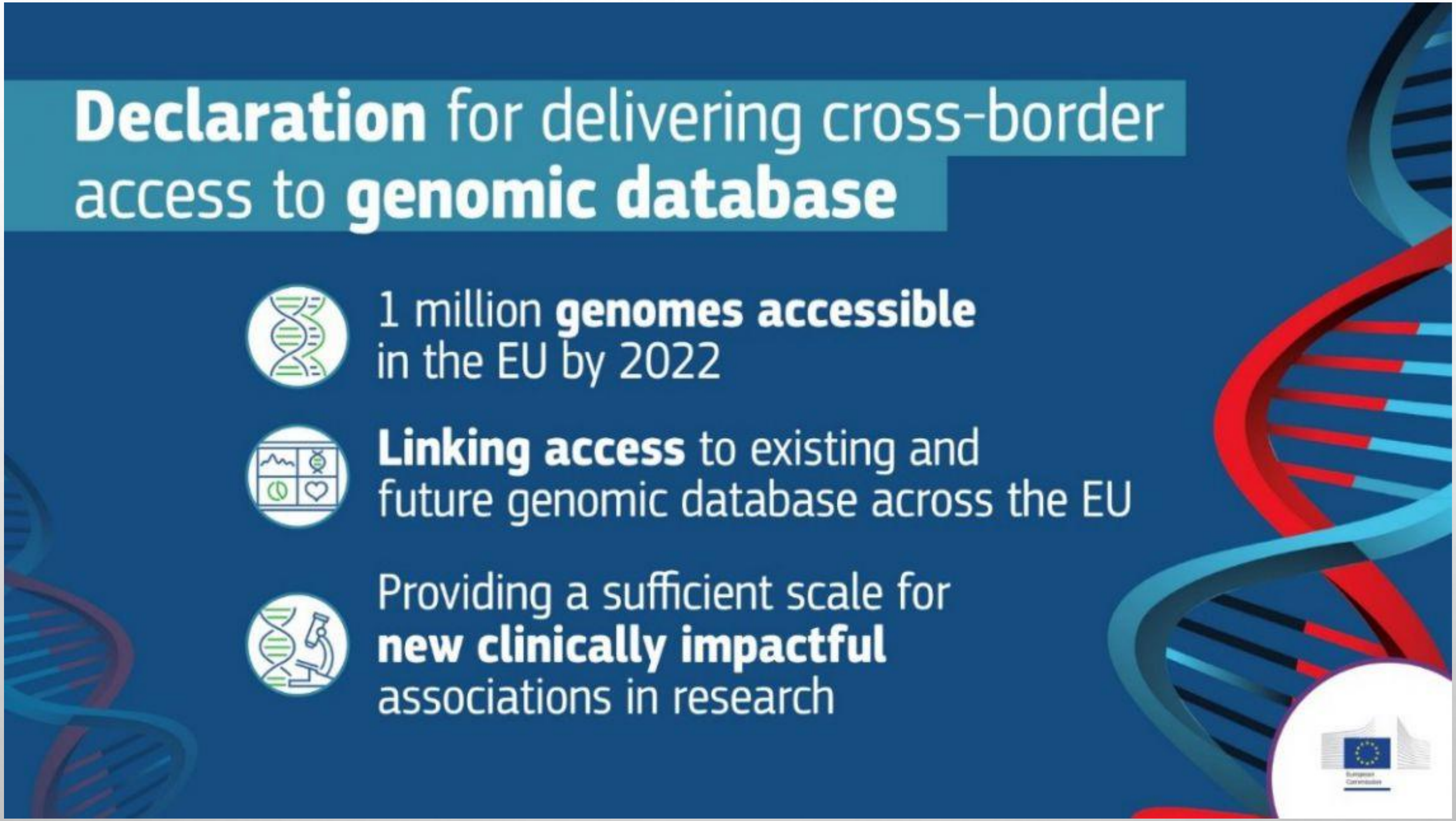
Recruitment of participants to the 100,000 Genomes Project was completed in 2018, with the 100,000th sequence achieved in December 2018.

The 100,000 Genomes Project

Genomics England & Partners

https://en.wikipedia.org/wiki/100,000_Genomes_Project

# EUROPEAN 1+ MILLION GENOMES



**Declaration** for delivering cross-border access to **genomic database**

1 million **genomes accessible** in the EU by 2022

**Linking access** to existing and future genomic database across the EU

Providing a sufficient scale for **new clinically impactful** associations in research

https://ec.europa.eu/digital-single-market/en/european-1-million-genomes-initiative

# THE PURPOSE…

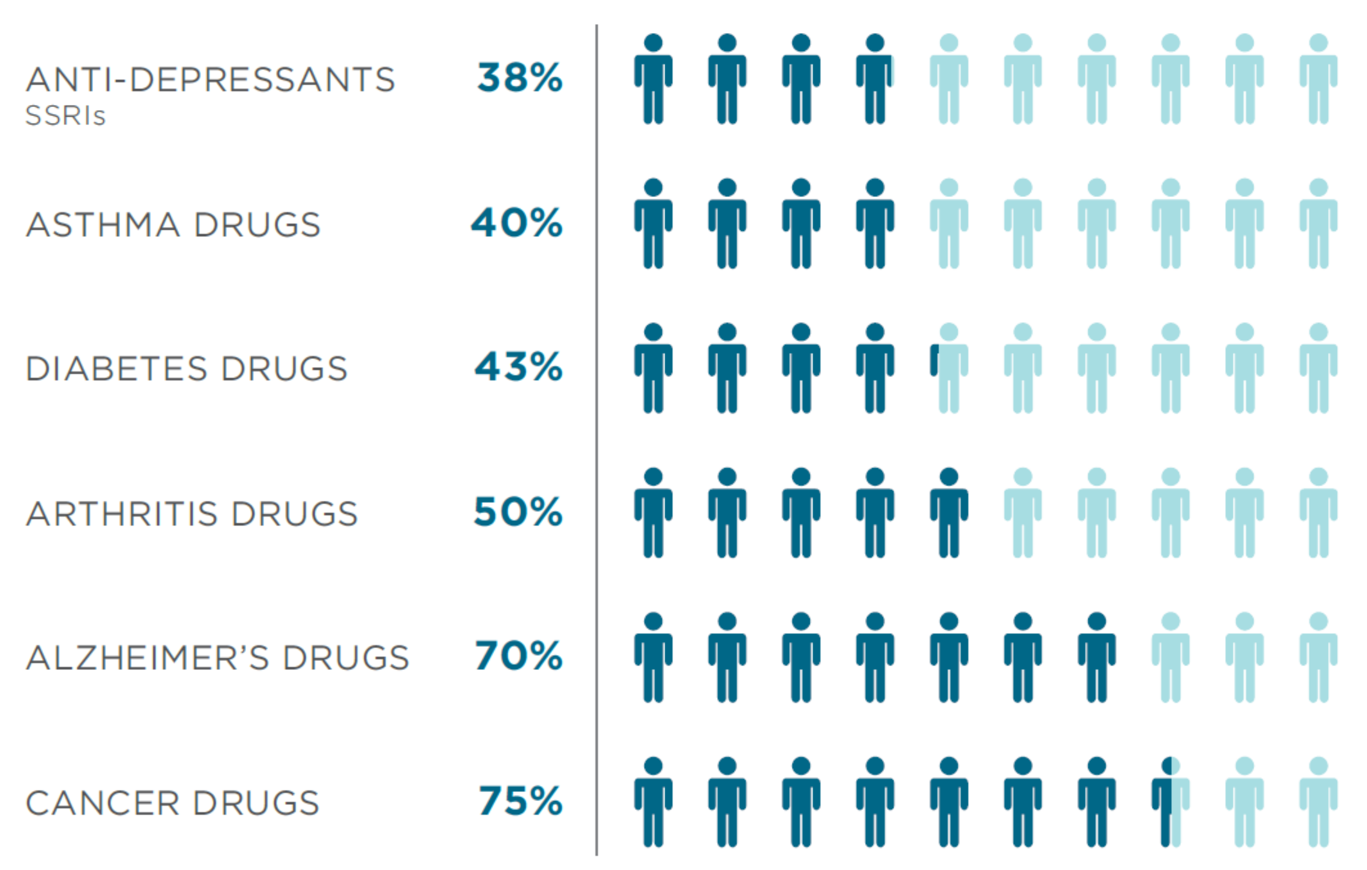
# PERSONALIZED MEDICINE

Genome sequencing can reveal **alterations in DNA that influence diseases** ranging from cystic fibrosis to cancer.

**Personalized medicine** takes advantage of the results from these techniques to design the most appropriate therapy for each patient.

# PERSONALIZED MEDICINE

Percentage of the patient population for which a particular drug in a class is ineffective, on average:



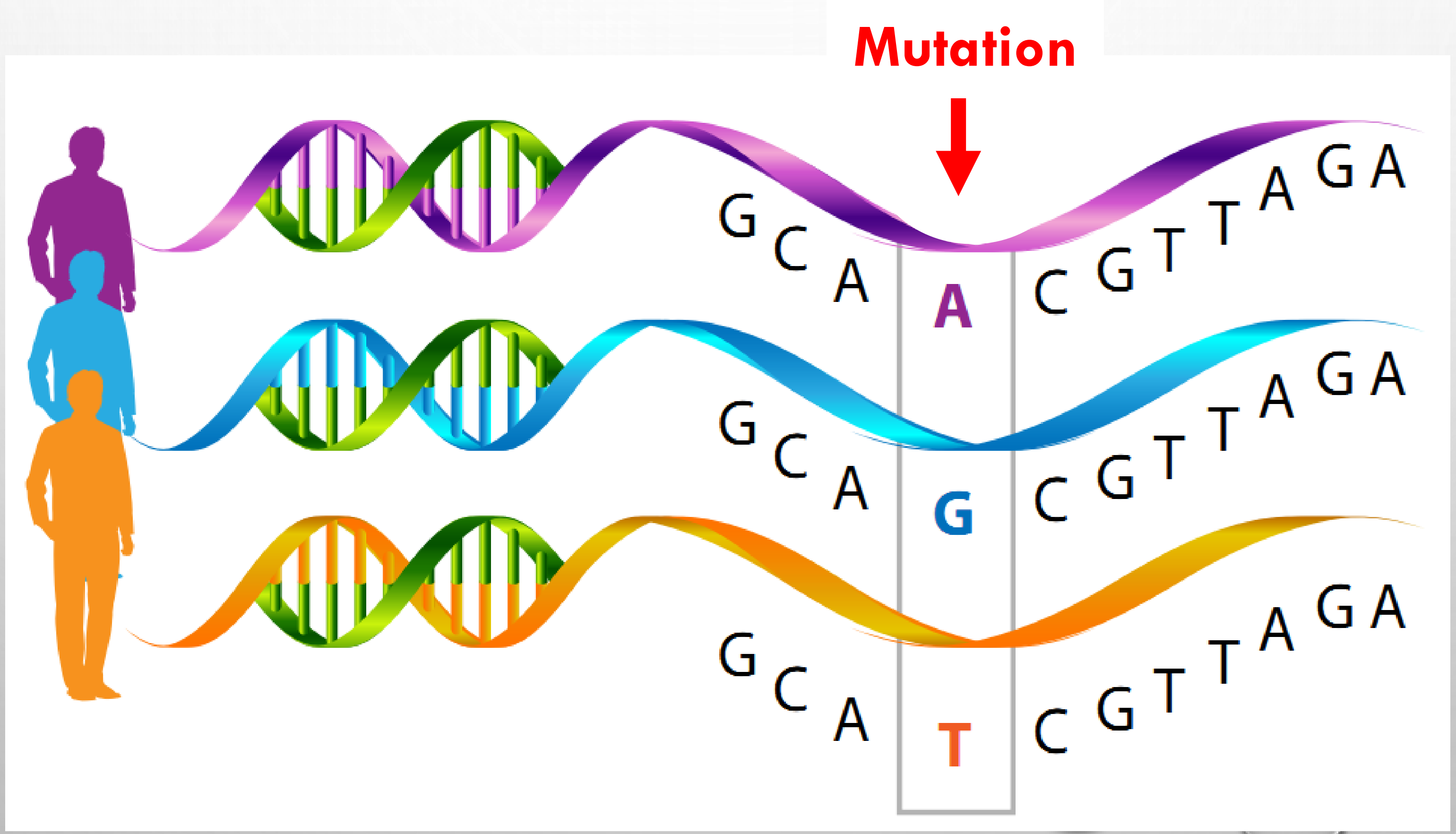
Brian B. Spear, Margo Heath-Chiozzi, Jeffrey Huff, "Clinical Trends in Molecular Medicine," Volume 7, Issue 5, 1 May 2001, pages 201-204.

# CANCER…

# TYPE OF MUTATIONS

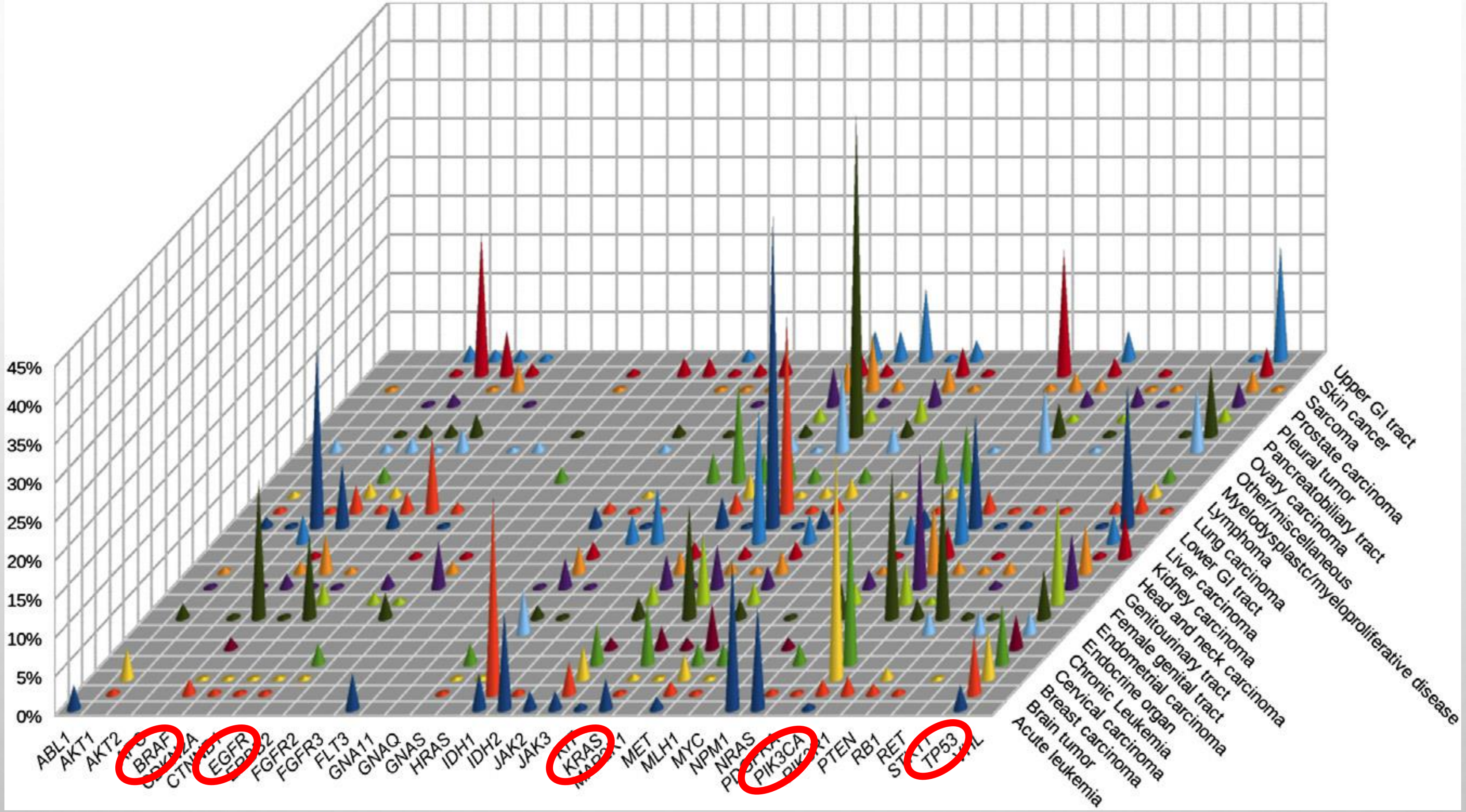
A **germline mutation** is a constitutional mutation that and is transmitted to offspring via the germ cells. **is inherited, present in all the body cells.**

A **somatic mutation** is not inherited from a parent, **is spontaneously generated during life,** and also not passed to offspring.



# MUTATIONS AND CANCER

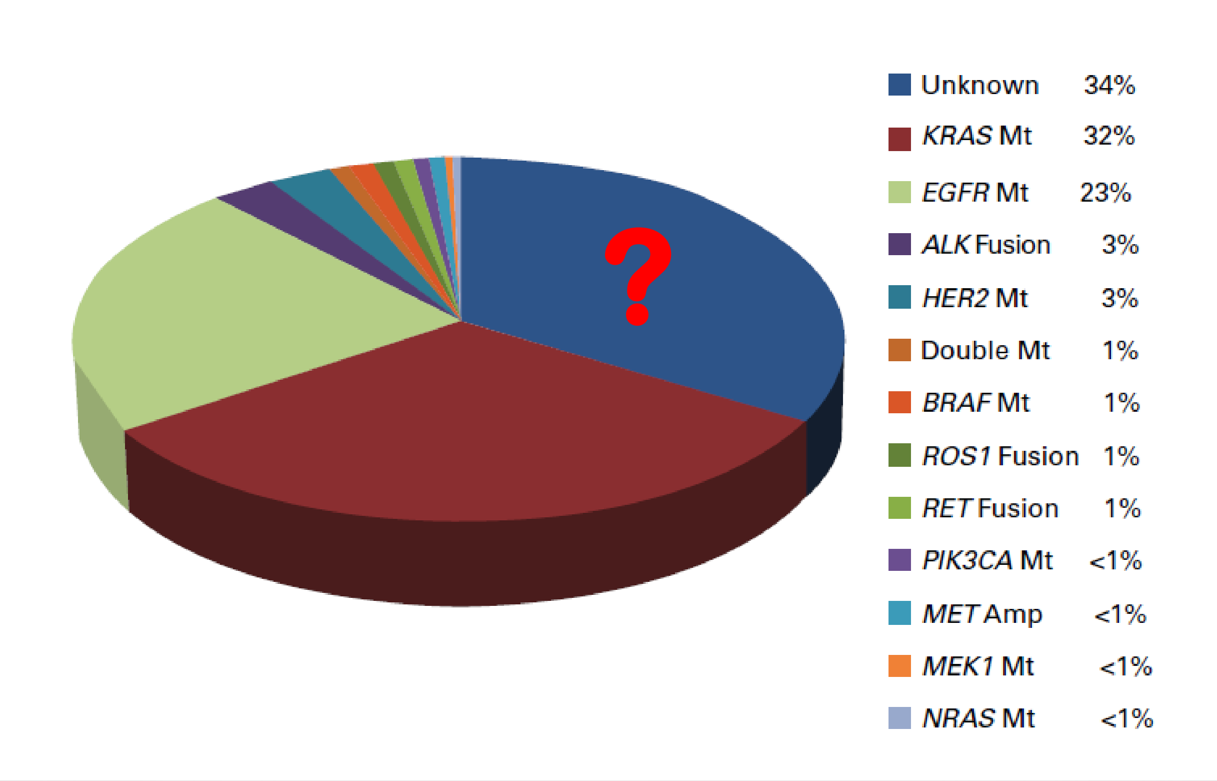
Genomic Landscape of 5000 Human Cancers:



MacConaill, L. E., Garcia, E., Shivdasani, P., Ducar, M., Adusumilli, R., Breneiser, M., … Lindeman, N. I. (2014). Prospective Enterprise-Level Molecular Genotyping of a Cohort of Cancer Patients. *The Journal of Molecular Diagnostics*, *16*(6), 660–672.

# MUTATIONS AND CANCER

Still we do not know many of the cancer driven genomic alterations:



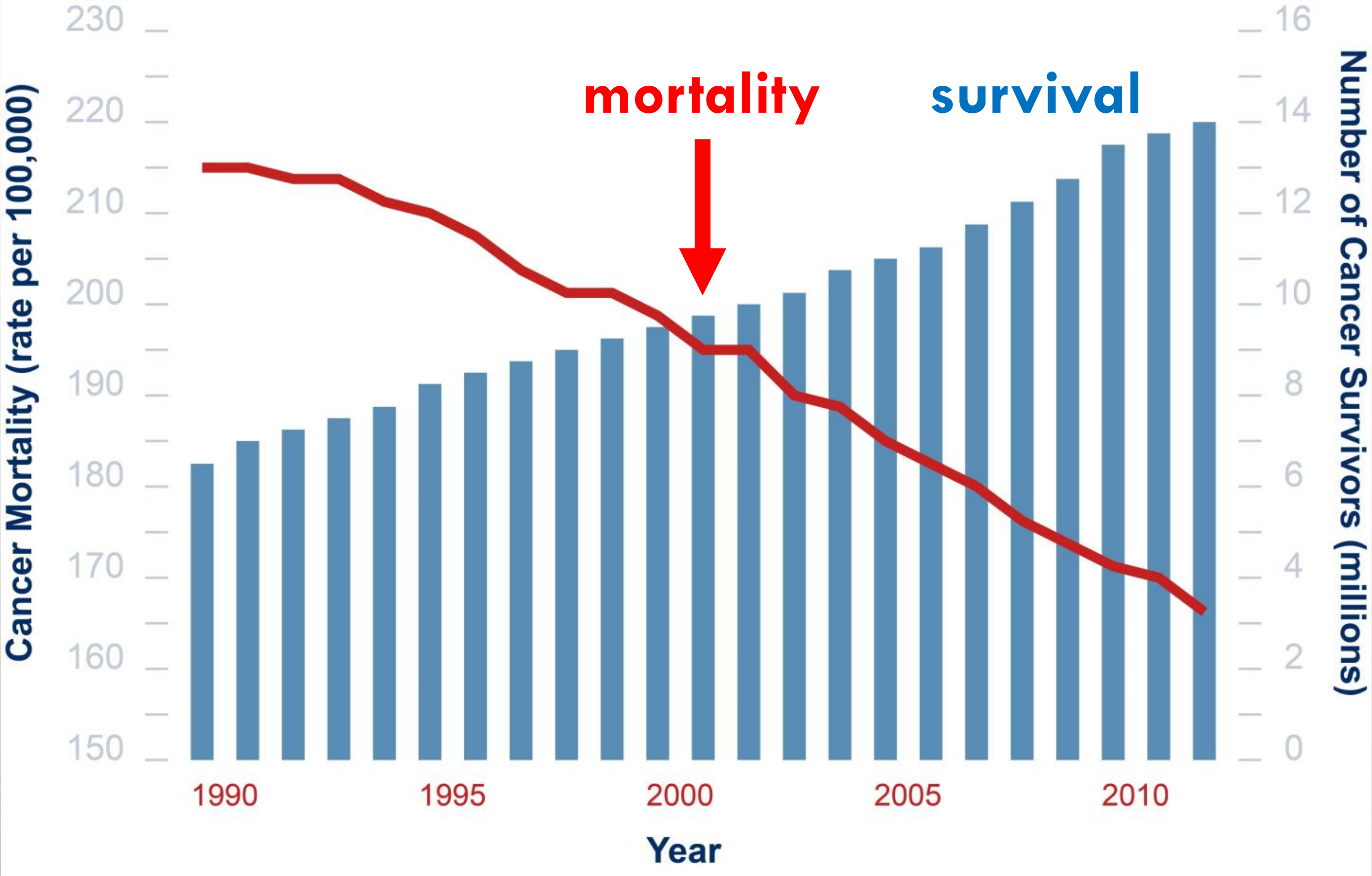
Representative pie charts from molecular diagnostic testing of NSCLC using a combination of assays at Memorial Sloan Kettering Cancer Center (MSKCC). Sanger sequencing, IHC, FISH, multiplex hotspot mutational testing, and multiplex sizing assays were used as part of a diagnostic algorithm for lung adenocarcinomas.

Naidoo, J., & Drilon, A. (2014). Molecular Diagnostic Testing in Non-Small Cell Lung Cancer.
The American Journal of Hematology/Oncology, 10(4)(september), 4–11.

# THE RESULTS…

# CANCER THERAPIES

Ongoing improvements in cancer treatments, survivorship up, mortality down:



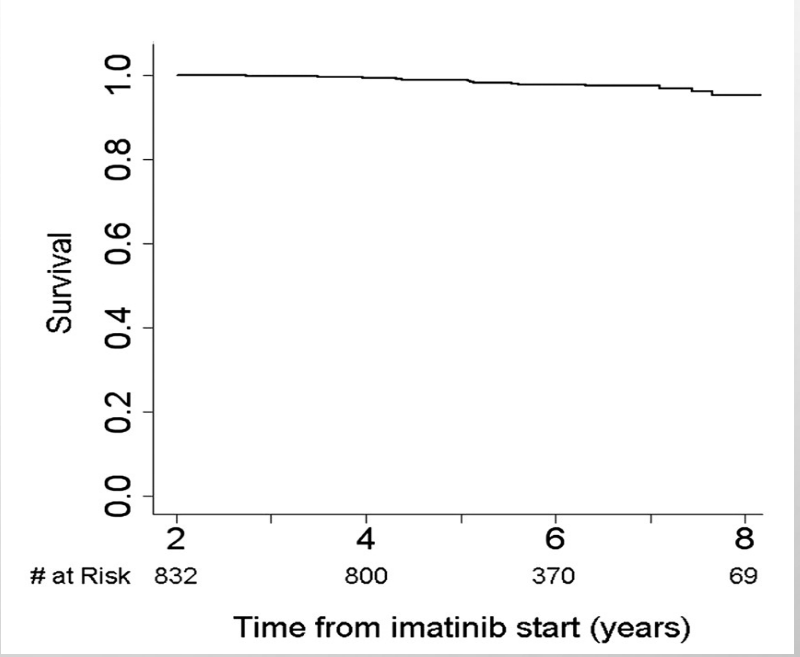
Sources: US Mortality Files, National Center for Health Statistics, CDC. DeSantis C, Chunchieh L, Mariotto AB, et al. (2014). Cancer Treatment and Survivorship Statistics, 2014. CA: A Cancer Journal for Clinicians.

SIXTH RESEARCHER
www.sixthresearcher.com

# CANCER THERAPIES

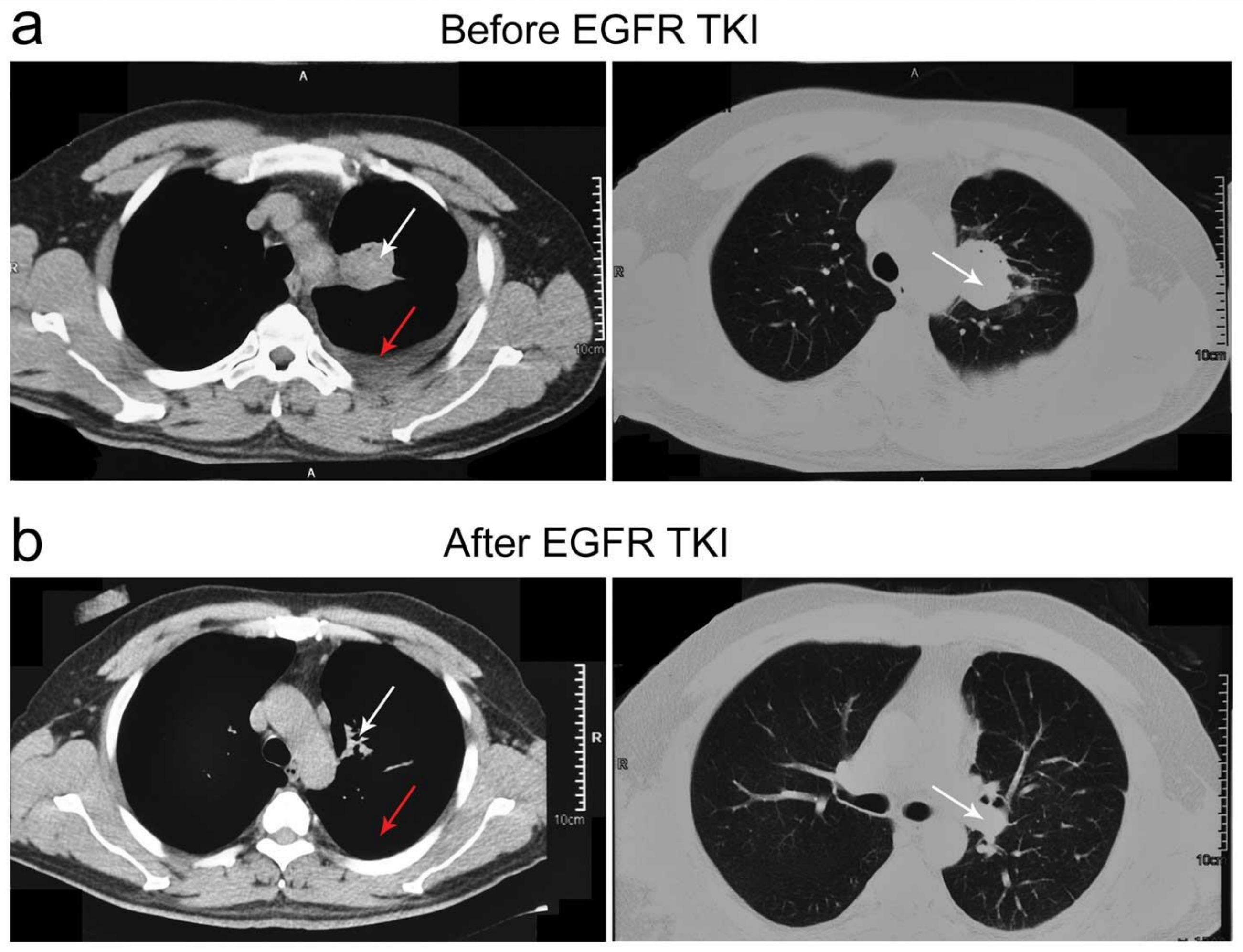
**Imatinib opened the new era of Cancer Targeted Therapies.** A simple pill putting an end to treatments with serious side effects that had limited success in prolonging life beyond the first year of diagnosis.

A 2011 study concluded that Chronic Myeloid Leukemia (CML) patients whose disease is in remission after 2 years of imatinib treatment have the same life expectancy as those who never had this disease.



Gambacorti-Passerini, C., Antolini, L., Mahon, F.-X., Guilhot, F., Deininger, M., Fava, C., … Kim, D.-W. (2011). Multicenter independent assessment of outcomes in chronic myeloid leukemia patients treated with imatinib. Journal of the National Cancer Institute, 103(7), 553–61.

# CANCER THERAPIES



Chu, H., Zhong, C., Xue, G., Liang, X., Wang, J., Liu, Y., … Bi, J. (2013). Direct sequencing and amplification refractory mutation system for epidermal growth factor receptor mutations in patients with non-small cell lung cancer. Oncology Reports, 30(5), 2311–2315.

# THE CODE…

# CODE EXAMPLES

**Google Colaboratory Notebooks:**

- Exploring human mutations related with cancer:

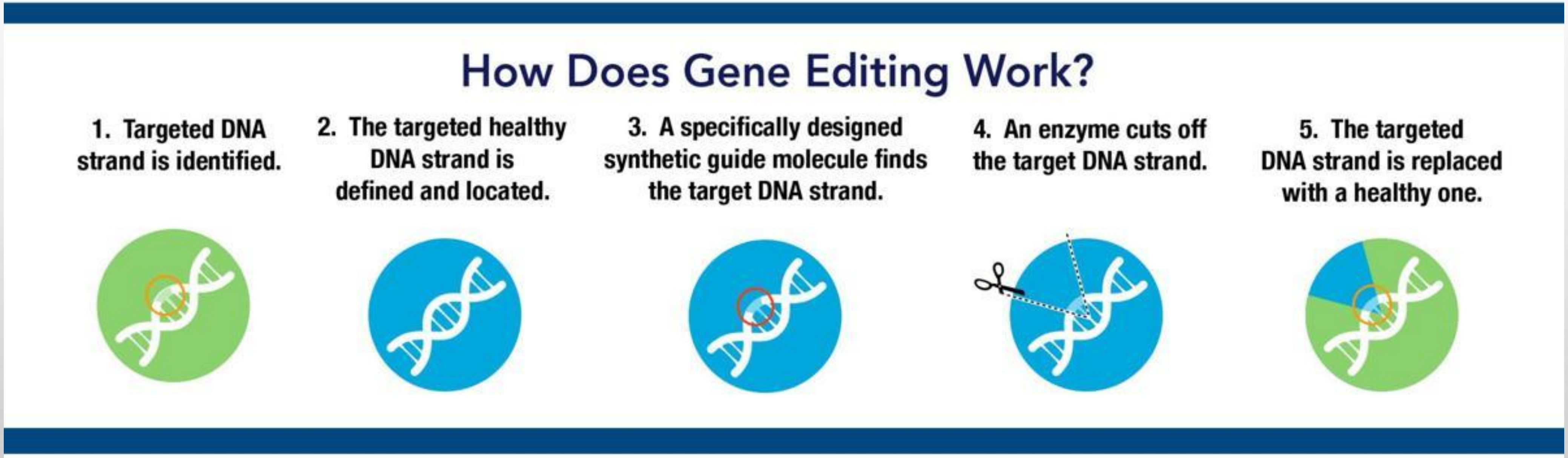
  https://colab.research.google.com/drive/1xOkGnrLVPiqwj1BfcMgfKdVilUOES5gd

- Looking for EGFR gene mutations at NGS data from lung cancer patients:

  https://colab.research.google.com/drive/1jffxhQoswPEW5-JMMk_y6HFbBL0dLjqD

# THE FUTURE…

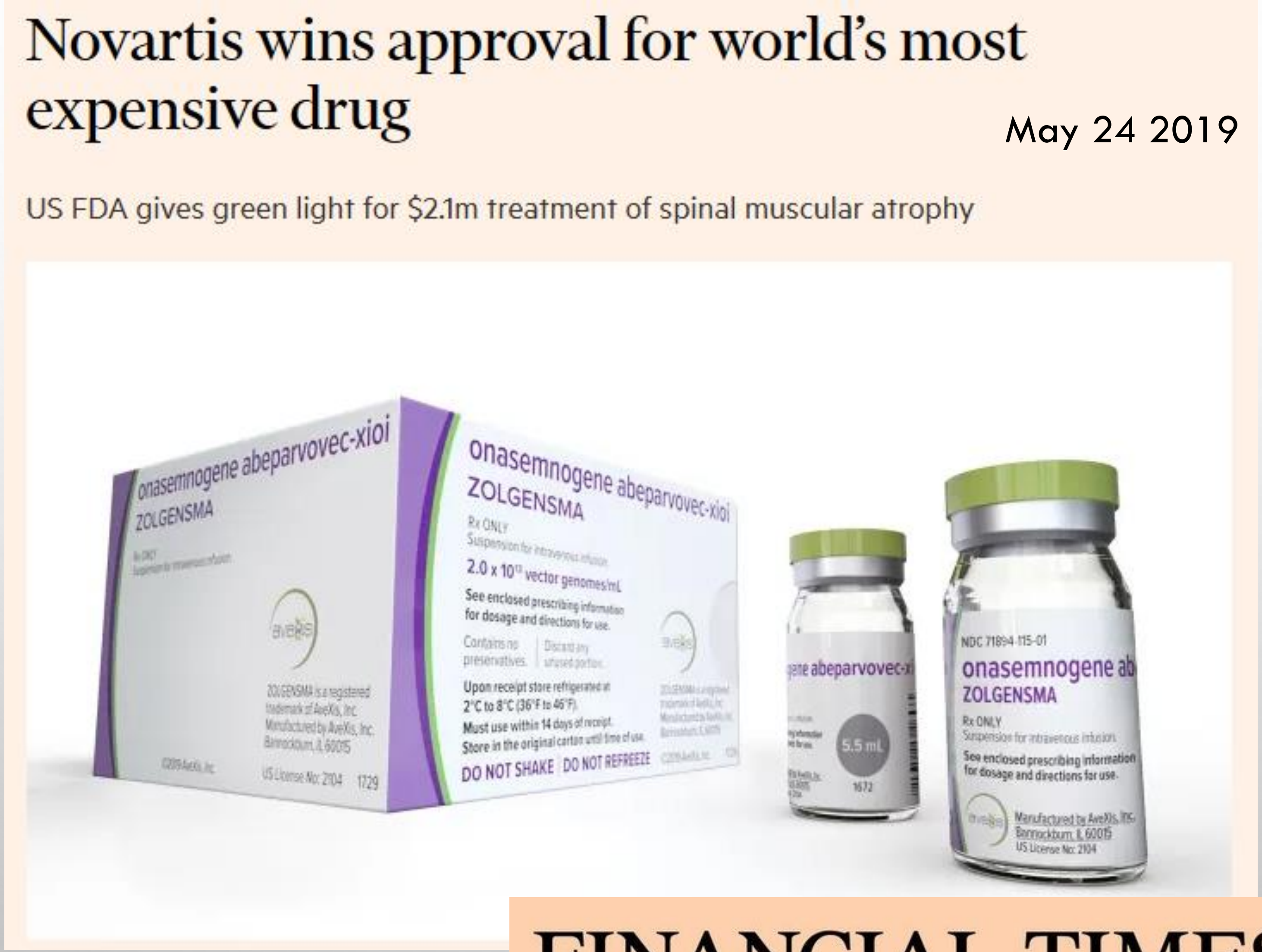
# GENE EDITING



## How Does Gene Editing Work?

1. Targeted DNA strand is identified.
2. The targeted healthy DNA strand is defined and located.
3. A specifically designed synthetic guide molecule finds the target DNA strand.
4. An enzyme cuts off the target DNA strand.
5. The targeted DNA strand is replaced with a healthy one.

# GENE EDITING



## Novartis wins approval for world's most expensive drug

US FDA gives green light for $2.1m treatment of spinal muscular atrophy

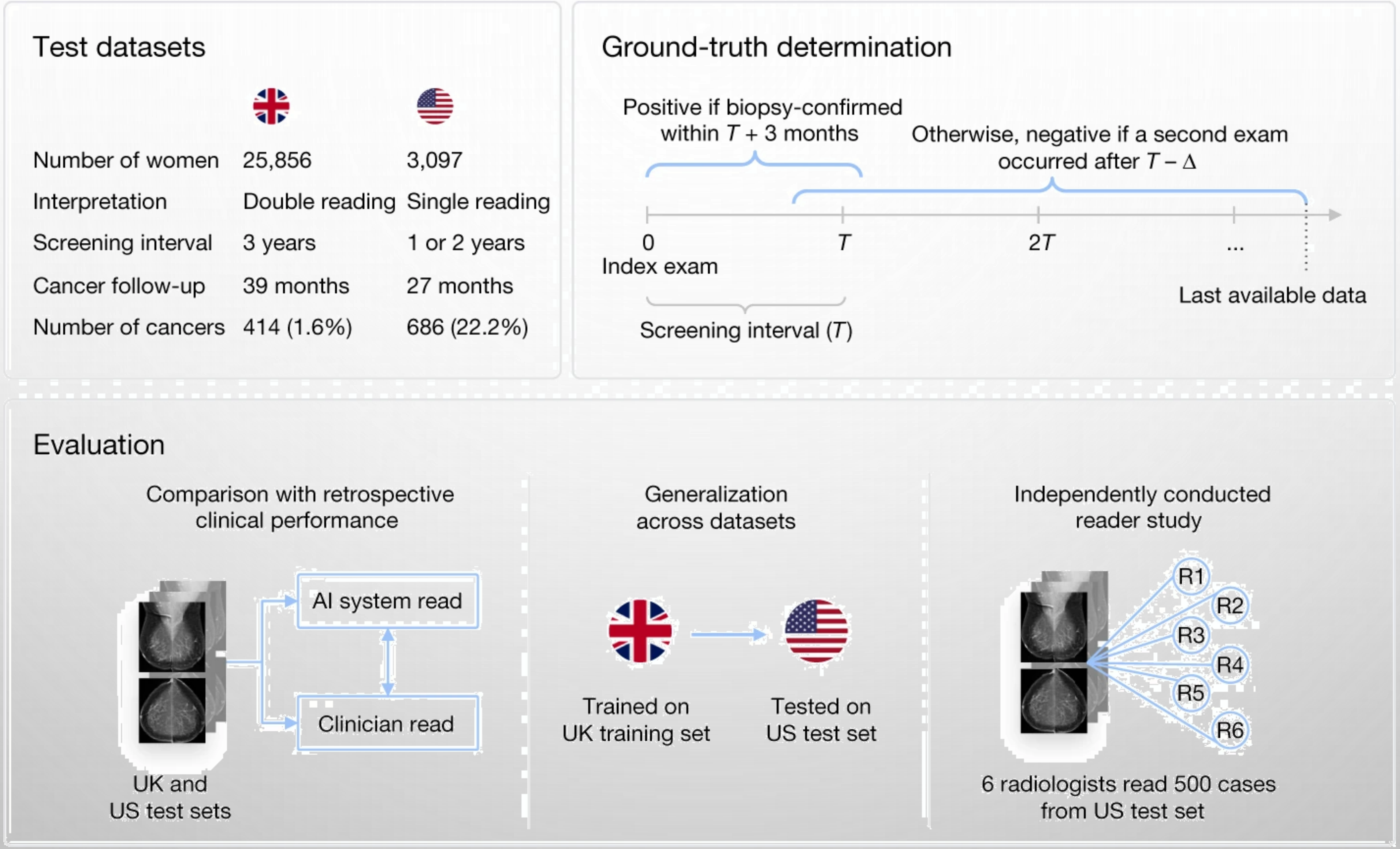
onasemnogene abeparvovec-xioi
ZOLGENSMA
Rx ONLY
Suspension for intravenous infusion
2.0 x 10¹³ vector genomes/mL
See enclosed prescribing information for dosage and directions for use.
Contains no preservatives.
Discard any unused portion.
Upon receipt store refrigerated at 2°C to 8°C (36°F to 46°F).
Must use within 14 days of receipt.
Store in the original carton until time of use.
DO NOT SHAKE
DO NOT REFREEZE
avexis
ZOLGENSMA is a registered trademark of AveXis, Inc.
Manufactured by AveXis, Inc. Bannockburn, IL 60015
US License No: 2104
5.5 mL
NDC 71894-115-01
Manufactured by AveXis, Inc. Bannockburn, IL 60015

May 24 2019

FINANCIAL TIMES

# ARTIFICIAL INTELLIGENCE…

# AI IN CANCER DETECTION



## Test datasets

| | 🇬🇧 | 🇺🇸 |
|---|---|---|
| Number of women | 25,856 | 3,097 |
| Interpretation | Double reading | Single reading |
| Screening interval | 3 years | 1 or 2 years |
| Cancer follow-up | 39 months | 27 months |
| Number of cancers | 414 (1.6%) | 686 (22.2%) |

## Ground-truth determination

Positive if biopsy-confirmed within $T + 3$ months

Otherwise, negative if a second exam occurred after $T - \Delta$

## Evaluation

Comparison with retrospective clinical performance

Generalization across datasets

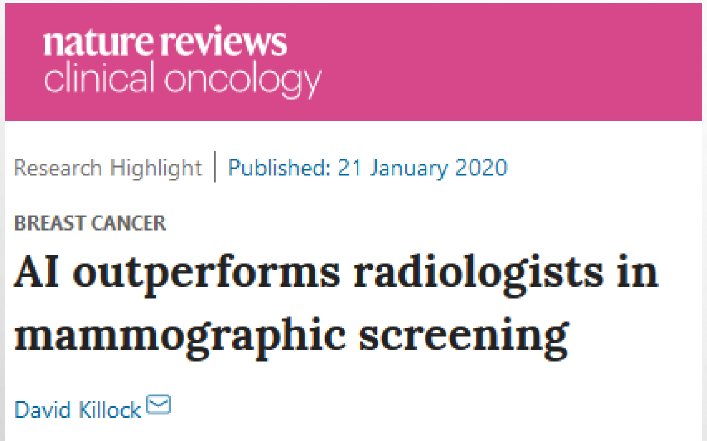
Independently conducted reader study

6 radiologists read 500 cases from US test set

McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020)

# AI IN CANCER DETECTION

An AI system that is capable of surpassing human experts in breast cancer prediction. It provides a reduction of 5.7% and 1.2% (USA and UK) in false positives and 9.4% and 2.7% in false negatives.



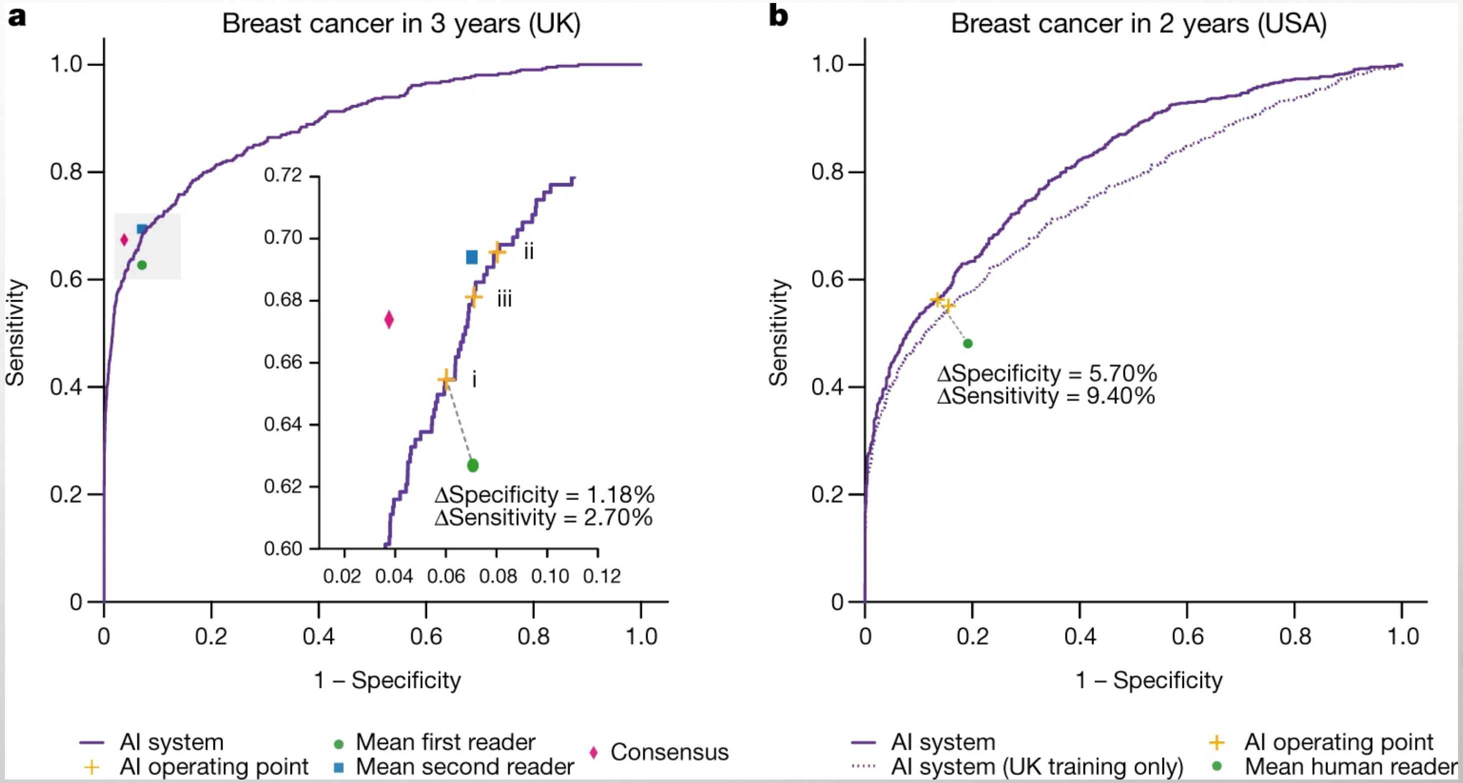
nature reviews
clinical oncology

Research Highlight | Published: 21 January 2020

BREAST CANCER

AI outperforms radiologists in mammographic screening

David Killock ✉

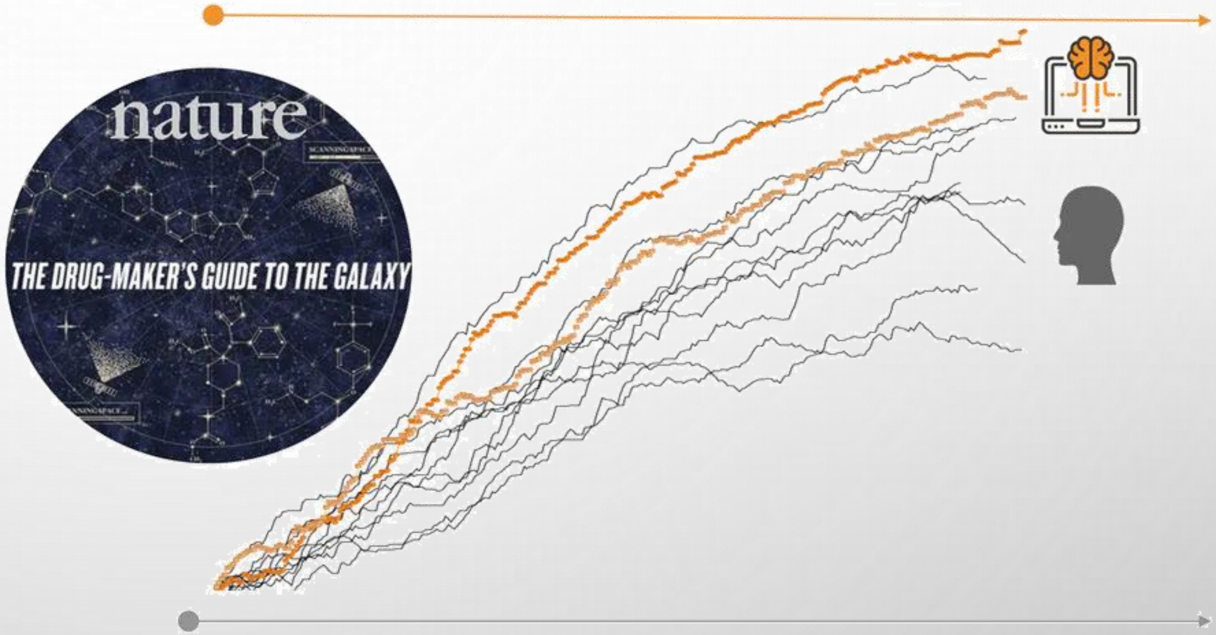


McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. *Nature 577*, 89–94 (2020)

# FIRST AI DESIGNED DRUG

In 2016, the pharmaceutical firm Sunovion gave a group of seasoned employees an unusual assignment. At the firm's headquarters in Marlborough, Massachusetts, the chemists were all asked to play a game to see who could discover the best leads for new drugs.

Of the 11 players, 10 struggled through the task for several hours. But one breezed through in milliseconds… because it was an algorithm.



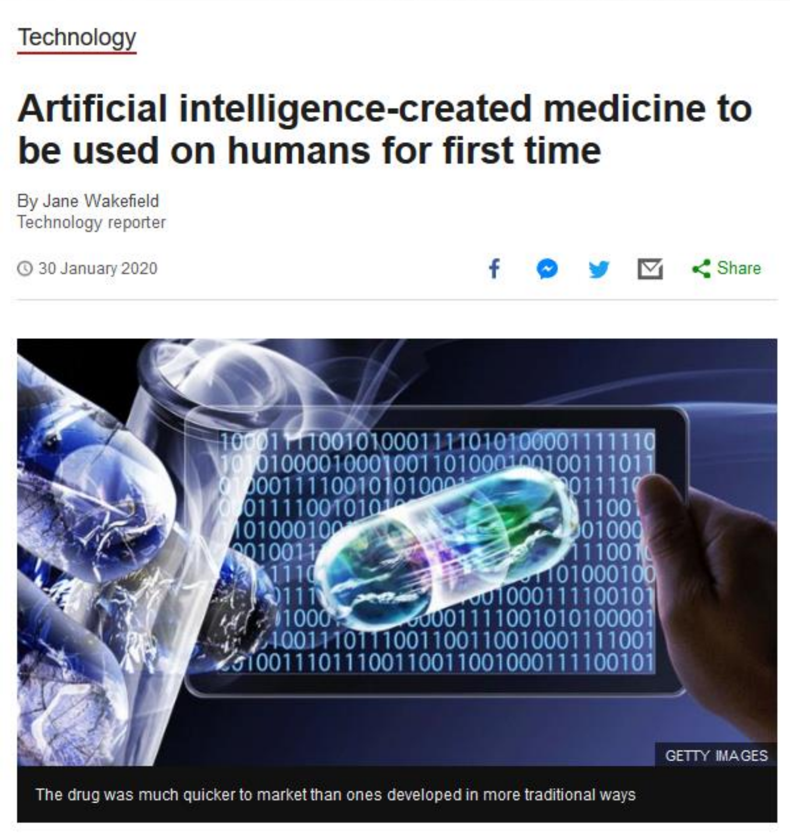
The Drug Makers Guide to the Galaxy: How machine learning and big data are helping chemists search the vast chemical universe for better medicines. Nature 26 SEP, 2017

# FIRST AI DESIGNED DRUG

A drug molecule invented entirely by artificial intelligence is set to enter human clinical trials for the first time, marking a critical milestone for the role of machine learning in medicine.

Four times faster than a typical Drug Discovery process, this AI designed drug to treat patients with obsessive-compulsive disorder will enter its clinical trials 12 months after reseach started.

The new compound was developed by Oxford-based AI start-up Exscientia in collaboration with the Japanese pharmaceutical firm Sumitomo Dainippon Pharma.



Technology

Artificial intelligence-created medicine to be used on humans for first time

By Jane Wakefield
Technology reporter

30 January 2020

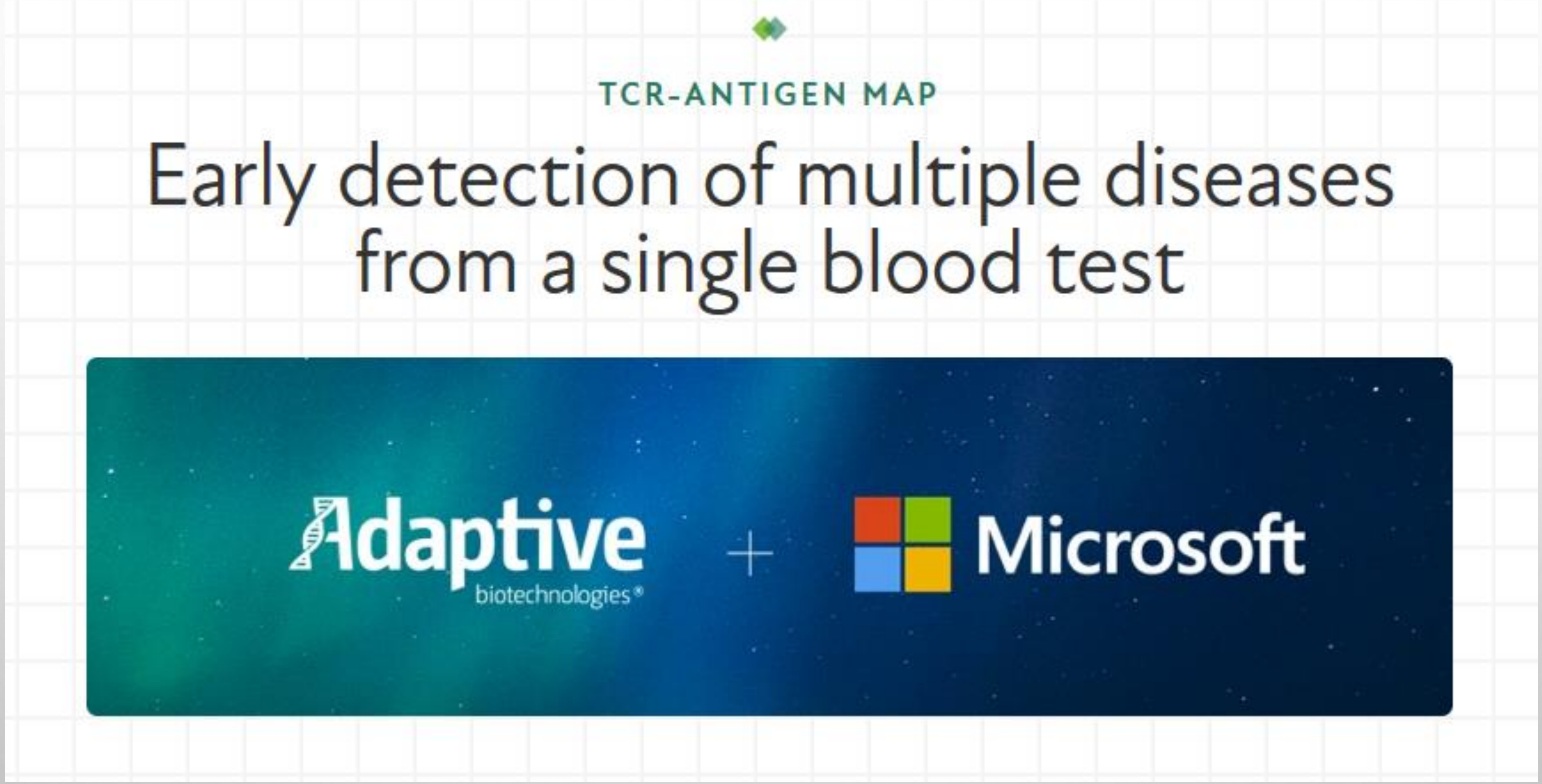
The drug was much quicker to market than ones developed in more traditional ways

BBC NEWS

https://www.bbc.com/news/technology-51315462

# WILL BE AI THE NEW THERANOS?

# WILL BE AI THE NEW THERANOS?

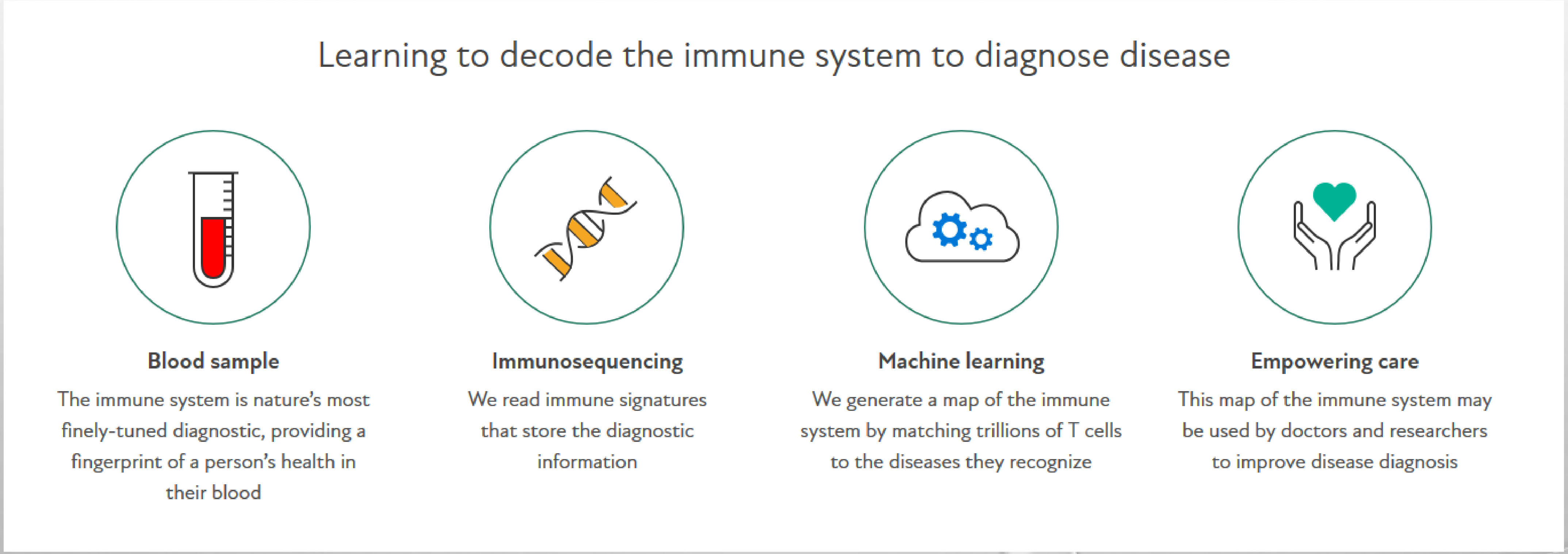


TCR-ANTIGEN MAP

Early detection of multiple diseases from a single blood test

# WILL BE AI THE NEW THERANOS?

Microsoft Healthcare NExT initiative has partnered with Adaptive to map and decode the human immune system, nature’s most finely tuned diagnostic. Together we are using immunosequencing, proprietary computational modeling, and machine learning to map T-cell receptor (TCR) sequences to the antigens they bind. Using this data, we aim to translate the natural diagnostic capability of the immune system into the clinic.



## Learning to decode the immune system to diagnose disease

**Blood sample**

The immune system is nature’s most finely-tuned diagnostic, providing a fingerprint of a person’s health in their blood

**Immunosequencing**

We read immune signatures that store the diagnostic information

**Machine learning**

We generate a map of the immune system by matching trillions of T cells to the diseases they recognize

**Empowering care**

This map of the immune system may be used by doctors and researchers to improve disease diagnosis

# THANK YOU
# DZIEKUJE
# GRACIAS

SIXTHRESEARCHER@GMAIL.COM