Genome Technology

Status and Future Directions

Presentation for BIOE291, Fall 2023

Dr. James Van Deventer

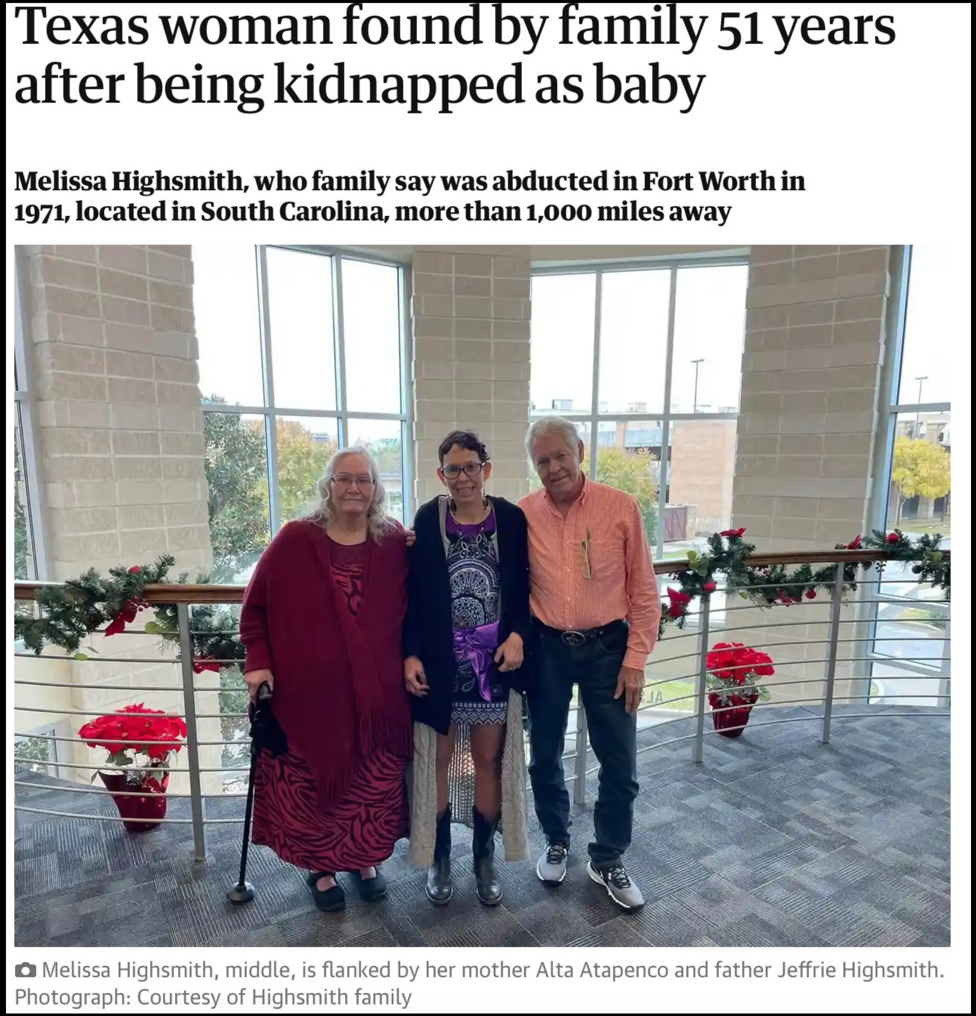
by

Adelaide Rhodes, Ph.D.

Senior Bioinformatics Scientist

Tufts University

Genomes Affect Our Everyday Lives

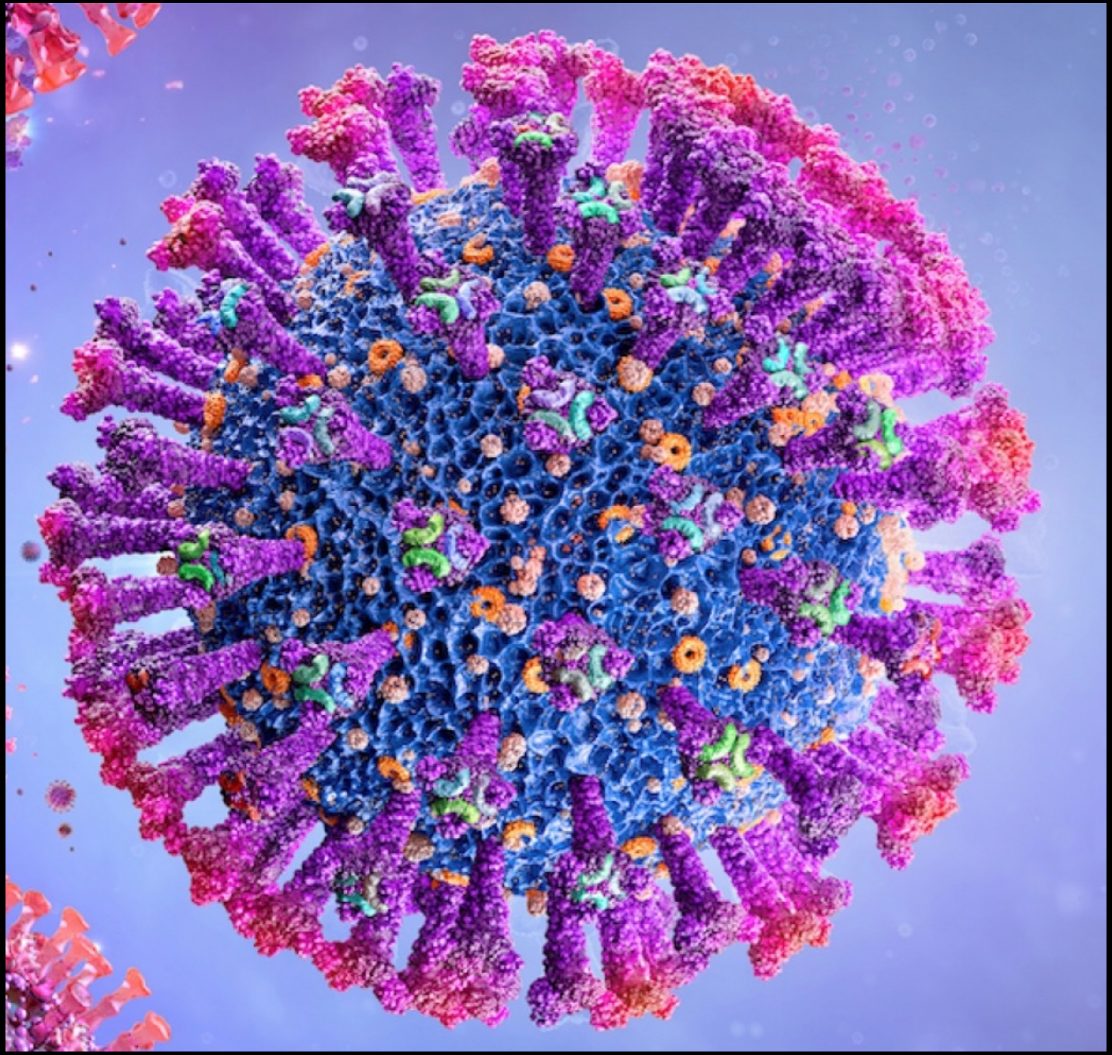


Texas woman found by family 51 years after being kidnapped as baby

Melissa Highsmith, who family say was abducted in Fort Worth in 1971, located in South Carolina, more than 1,000 miles away

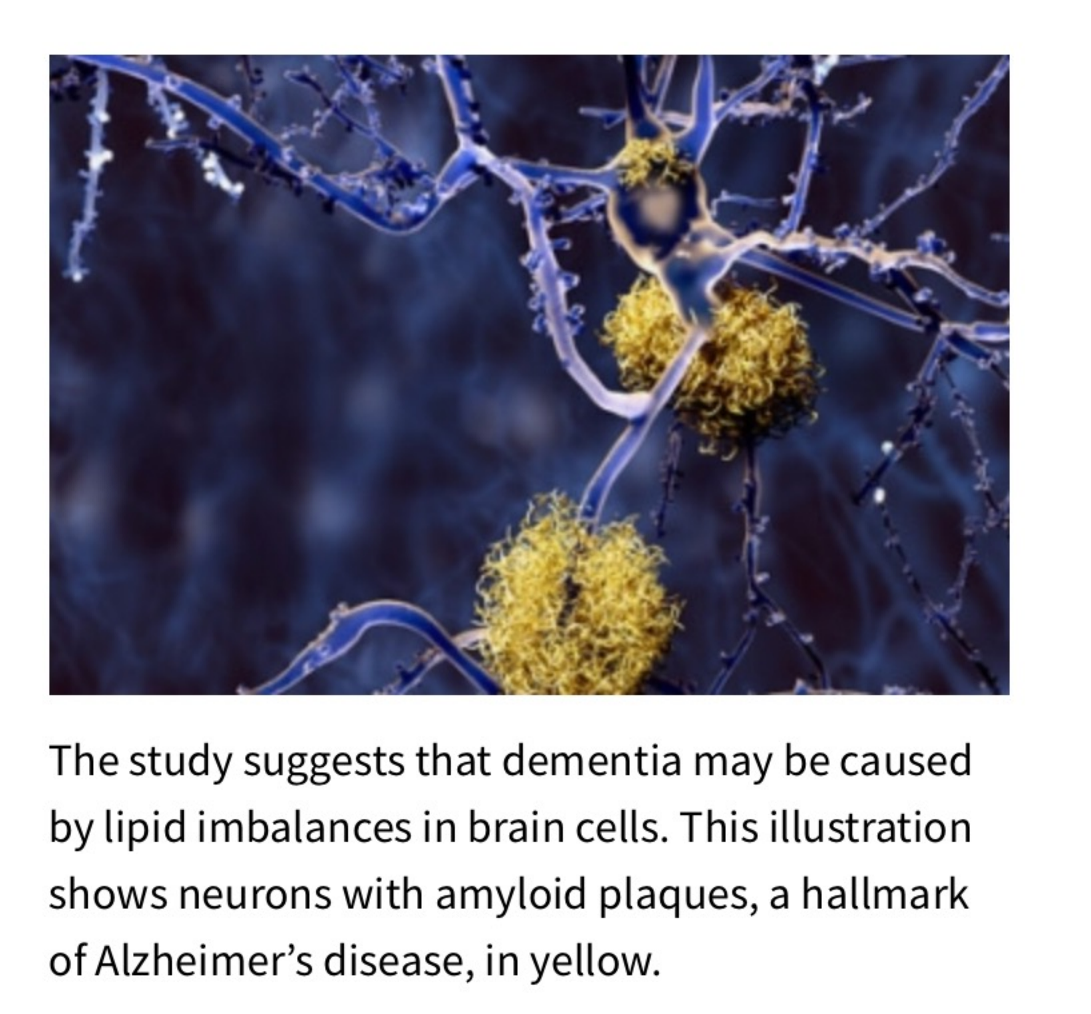
Melissa Highsmith, middle, is flanked by her mother Alta Atapenco and father Jeffrie Highsmith. Photograph: Courtesy of Highsmith family

https://www.theguardian.com/us-news/2022/nov/28/texas-woman-melissa-highsmith-found-south-carolina



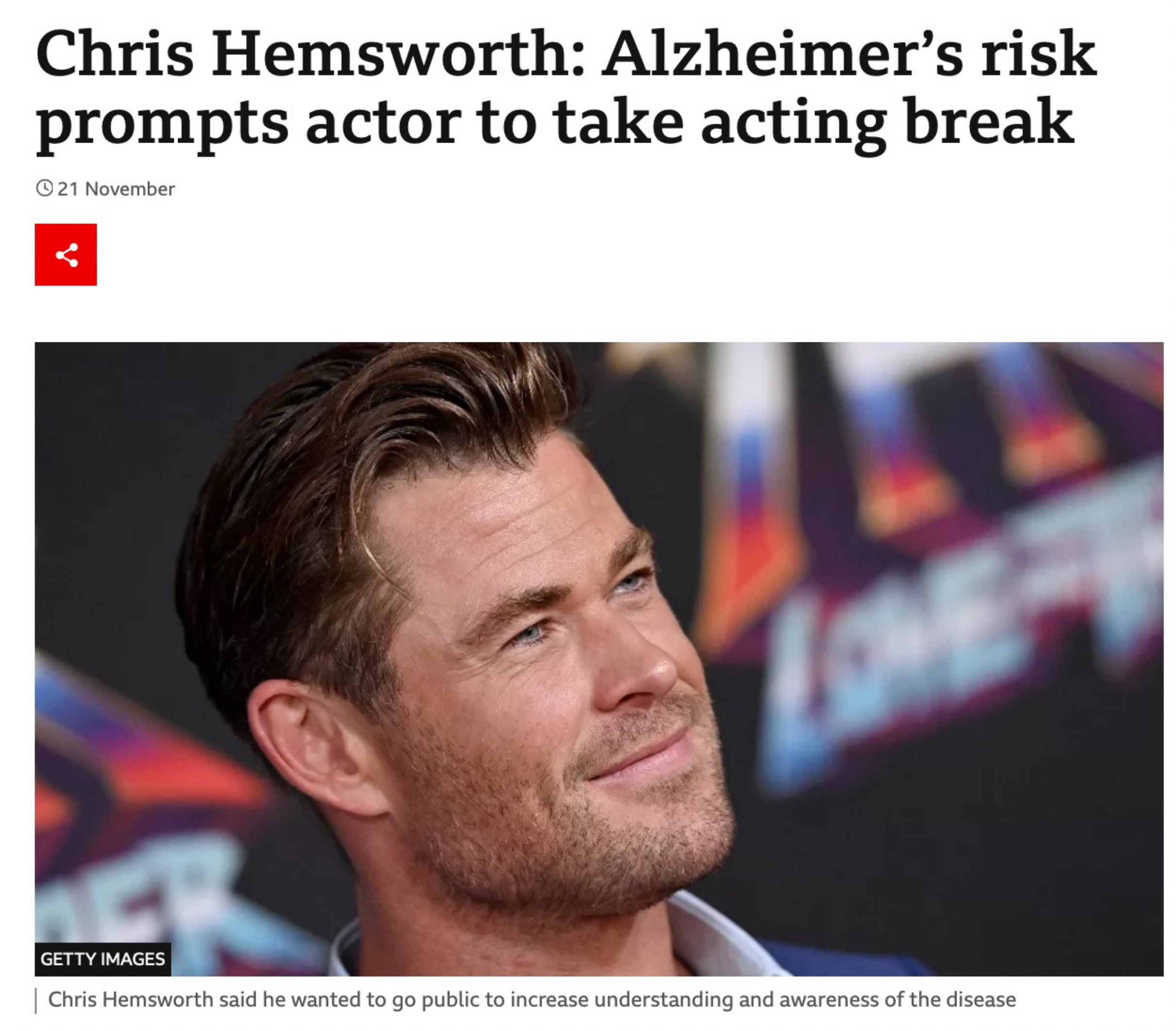
https://weillcornell.org/news/the-covid-19-delta-variant-here’s-what-we-know-so-far

Genomes Can Offer Hope & Worry



The study suggests that dementia may be caused by lipid imbalances in brain cells. This illustration shows neurons with amyloid plaques, a hallmark of Alzheimer's disease, in yellow.

https://www.nia.nih.gov/news/study-reveals-how-apoe4-gene-may-increase-risk-dementia



Chris Hemsworth: Alzheimer's risk prompts actor to take acting break

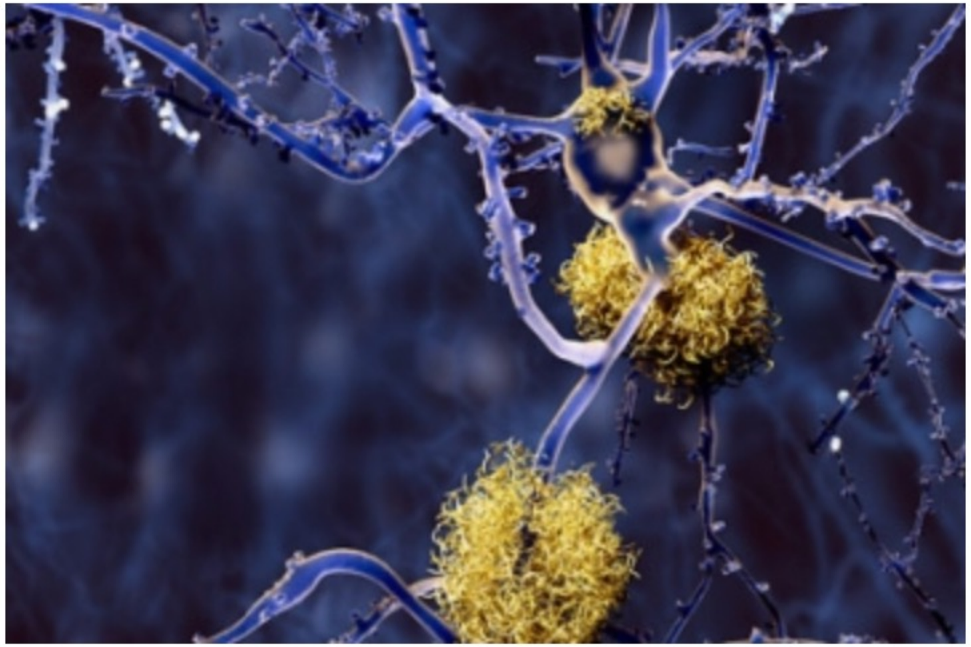
21 November

GETTY IMAGES

Chris Hemsworth said he wanted to go public to increase understanding and awareness of the disease

https://www.bbc.com/news/entertainment-arts-63668310

Genomes Can Offer Hope & Worry



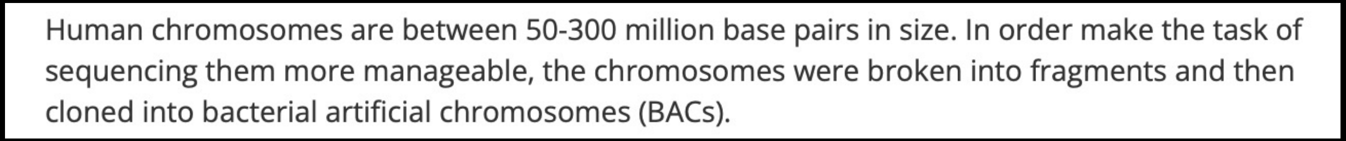
The study suggests that dementia may be caused by lipid imbalances in brain cells. This illustration shows neurons with amyloid plaques, a hallmark of Alzheimer's disease, in yellow.

One of the most significant genetic risk factors is a form of the *apolipoprotein E* gene called *APOE4*. About 25% of people carry one copy of *APOE4*, and 2 to 3% carry two copies. *APOE4* is the strongest risk factor gene for Alzheimer's disease,

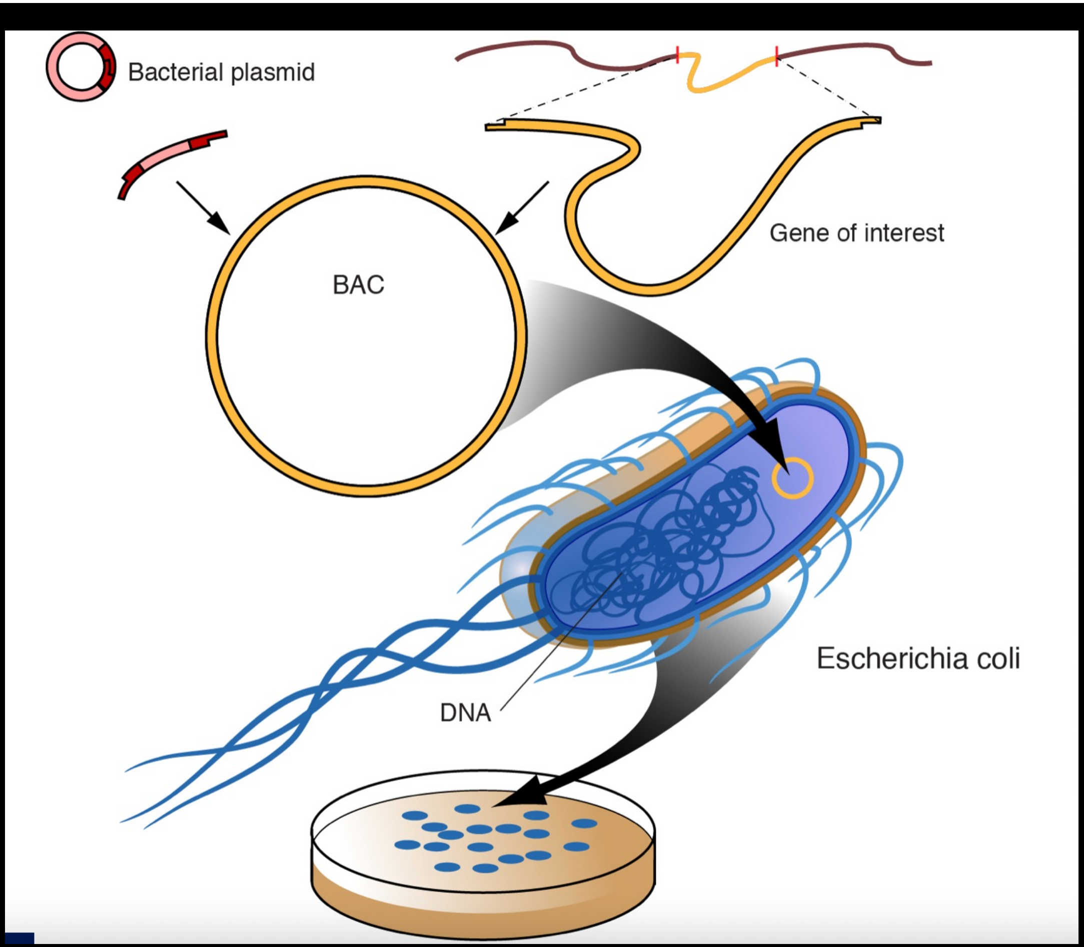
although inheriting *APOE4* does not mean a person will definitely develop the disease.

https://www.nia.nih.gov/news/study-reveals-how-apoe4-gene-may-increase-risk-dementia

Human Genome Project, 1990 - ???



Human chromosomes are between 50-300 million base pairs in size. In order make the task of sequencing them more manageable, the chromosomes were broken into fragments and then cloned into bacterial artificial chromosomes (BACs).



https://www.genome.gov/genetics-glossary/Bacterial-Artificial-Chromosome

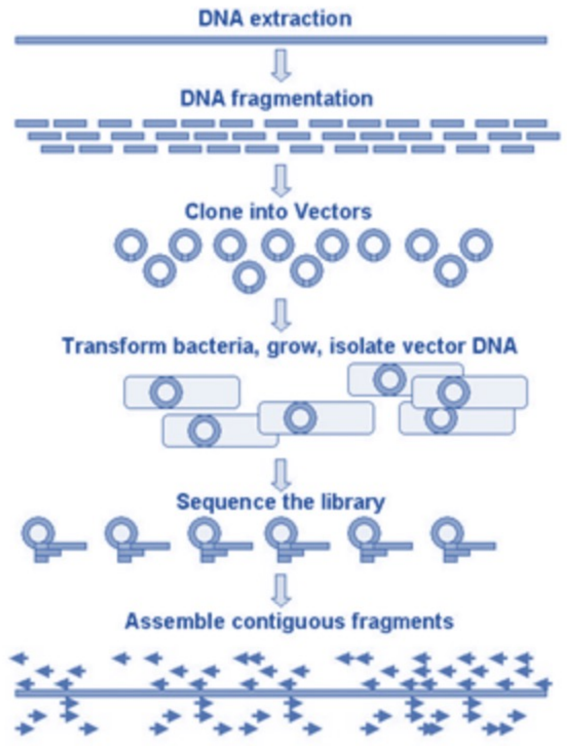
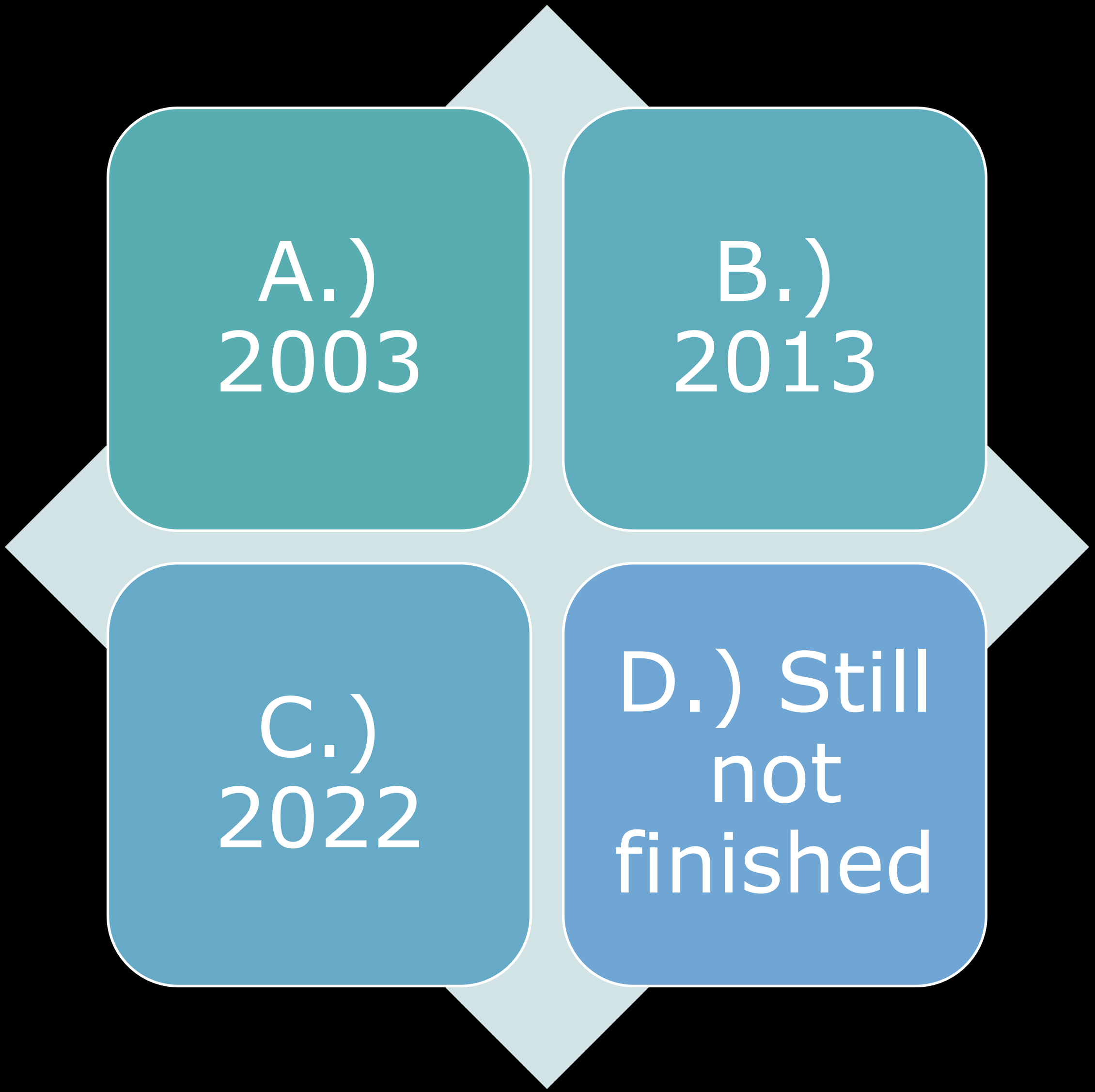

DNA extraction
DNA fragmentation
Clone into Vectors
Transform bacteria, grow, isolate vector DNA
Sequence the library
Assemble contiguous fragments

Diagram of DNA sequencing method used in the Human Genome Project.
Photo Credit: [Abizar Lakdawalla].

https://www.stressmarq.com/june-26-2000-dna-sequence-released-by-human-genome-project/?v=7516fd43adaa

What year was the Human Genome Completed?

Type your Guess in the Chat



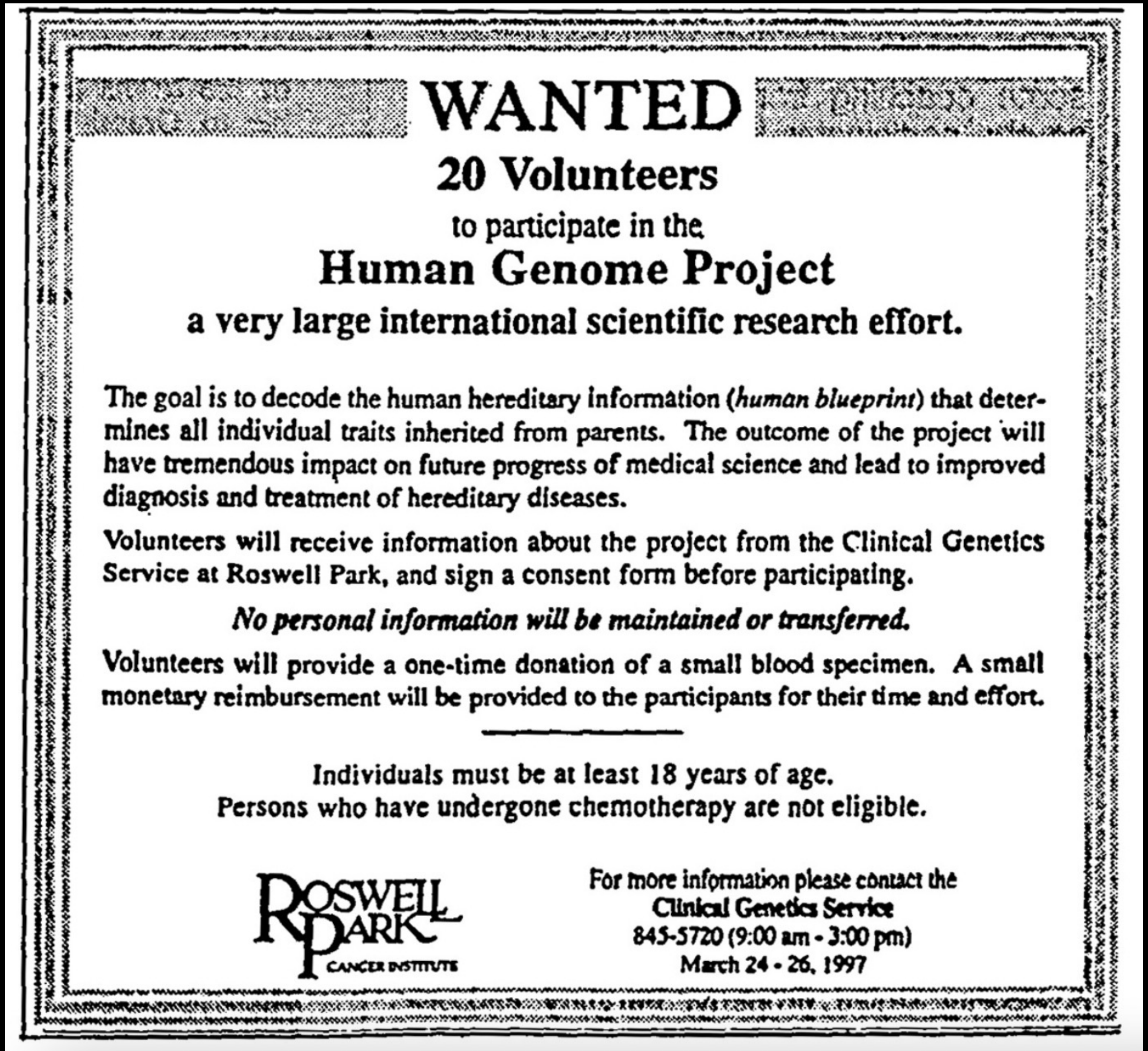
A.) 2003

B.) 2013

C.) 2022

D.) Still not finished

How many people's DNA were used for the initial Human Genome Project?



WANTED

20 Volunteers

to participate in the

Human Genome Project

a very large international scientific research effort.

The goal is to decode the human hereditary information (*human blueprint*) that determines all individual traits inherited from parents. The outcome of the project will have tremendous impact on future progress of medical science and lead to improved diagnosis and treatment of hereditary diseases.

Volunteers will receive information about the project from the Clinical Genetics Service at Roswell Park, and sign a consent form before participating.

No personal information will be maintained or transferred.

Volunteers will provide a one-time donation of a small blood specimen. A small monetary reimbursement will be provided to the participants for their time and effort.

Individuals must be at least 18 years of age.
Persons who have undergone chemotherapy are not eligible.

ROSWELL PARK
CANCER INSTITUTE

For more information please contact the
Clinical Genetics Service
845-5720 (9:00 am - 3:00 pm)
March 24 - 26, 1997

A.) 4
B.) 12
C.) 20
D.) Unknown

https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project

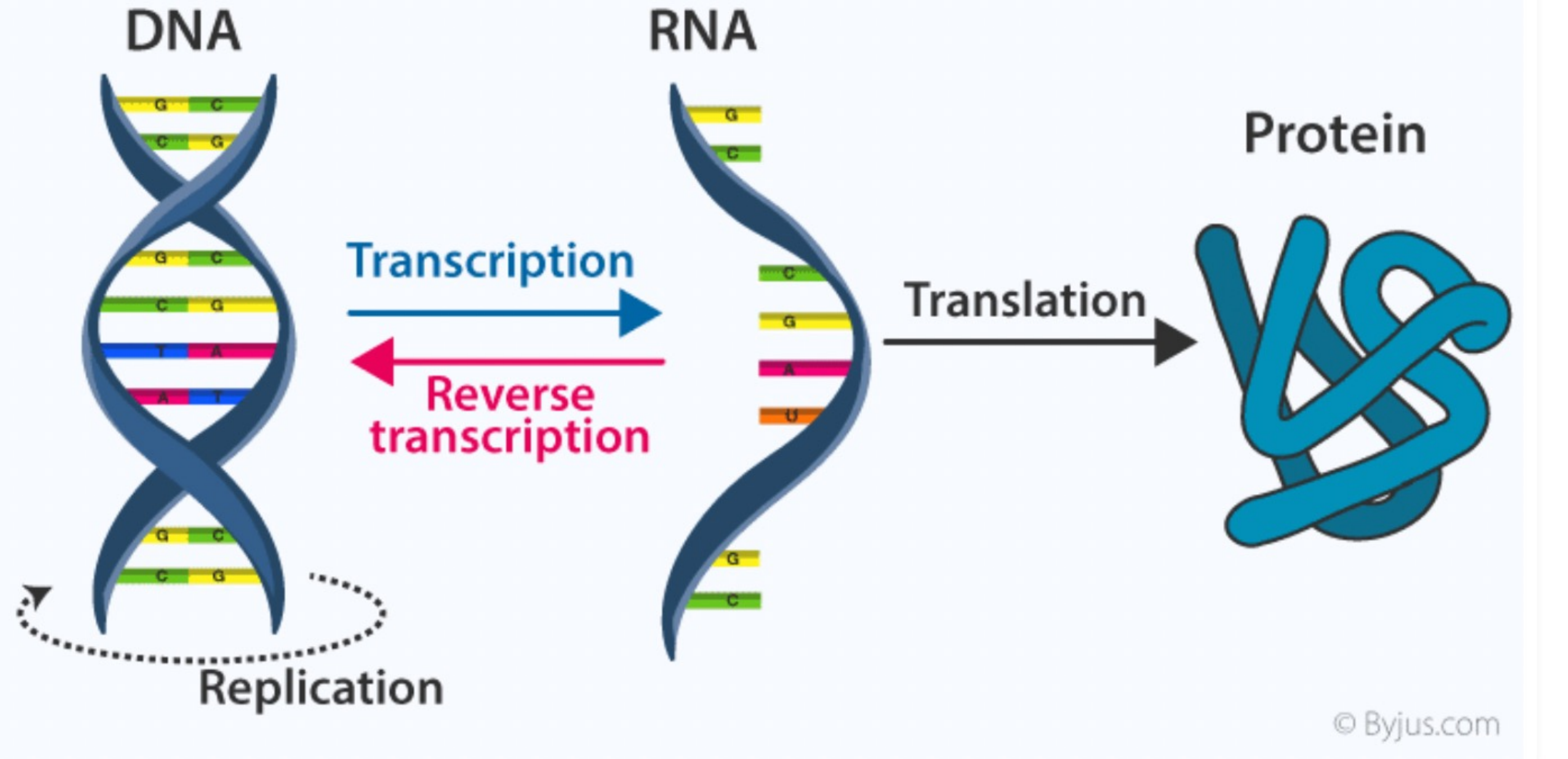
Human Genome Project – Who?

- Multiple people whose identities were intentionally made anonymous to protect their privacy.
- Volunteers provided informed consent to give their blood.
- Most donors were from Buffalo, New York:
    - 93% from 11 donors
    - 70% from one donor.
- Two male and two female donors were randomly selected from a pool of 20 volunteers. The identity of the final 4 donors remains unknown even to them.

What does this approach imply about the assumptions of the Researchers on the HGP?

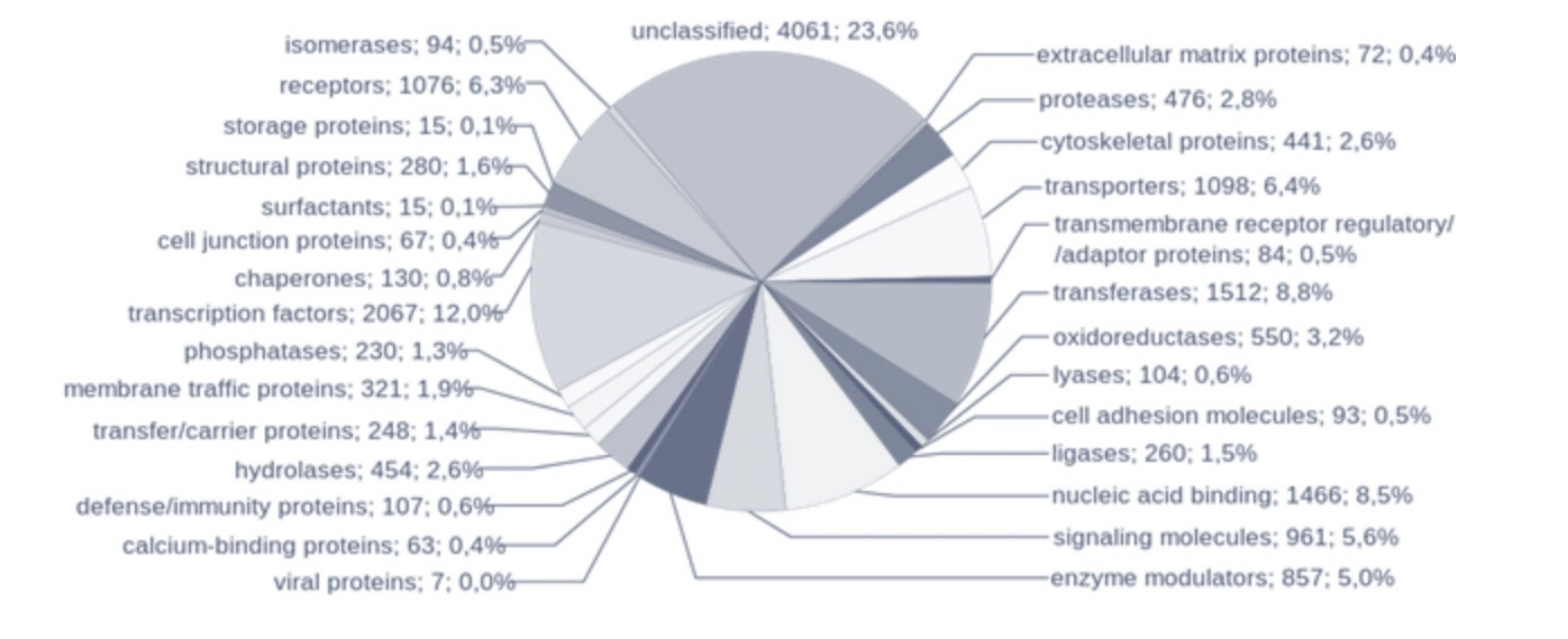
https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project

The Central Dogma



2003 – Gold Standard of Human Genome

- 20,000 Genes
  - Only 1.5% code for proteins – 1800 diseases from mutation identified
  - 98.5% of the genome is transcribed into
    - functional non-coding RNA strands
    - origins of replication, centromeres, and telomeres
- Only 1.5%? That seems low, right?



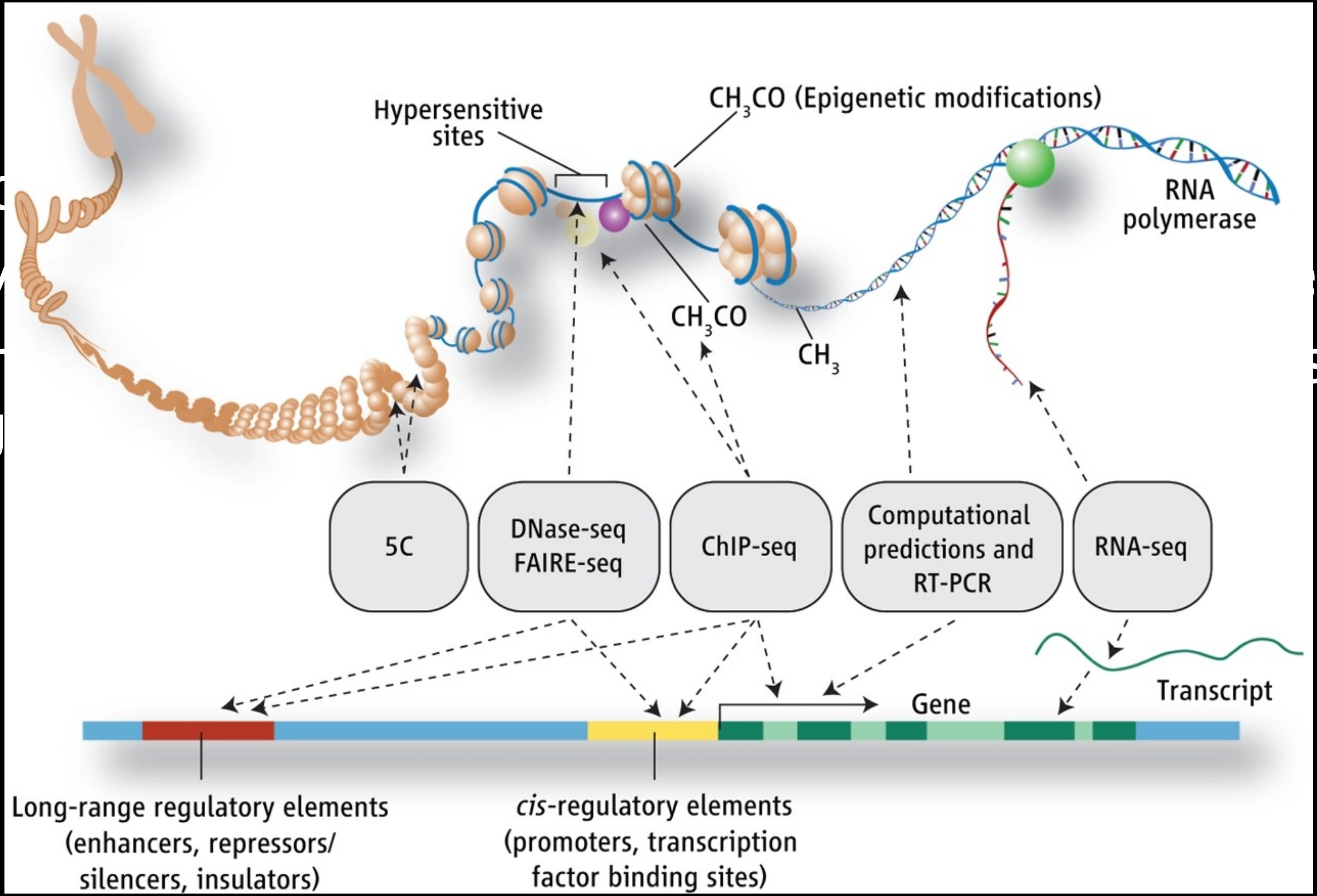
“ENCODE Project Writes Eulogy for Junk DNA”

- ”Junk DNA” refers to regions of the genome that do not produce functional proteins.
- However, not all functional activity is in the proteins.
- The Encyclopedia of DNA Elements describes these non-coding functions.

https://www.science.org/doi/10.1126/science.337.6099.1159

“ENCODE Project Writes Eulogy for Junk DNA”

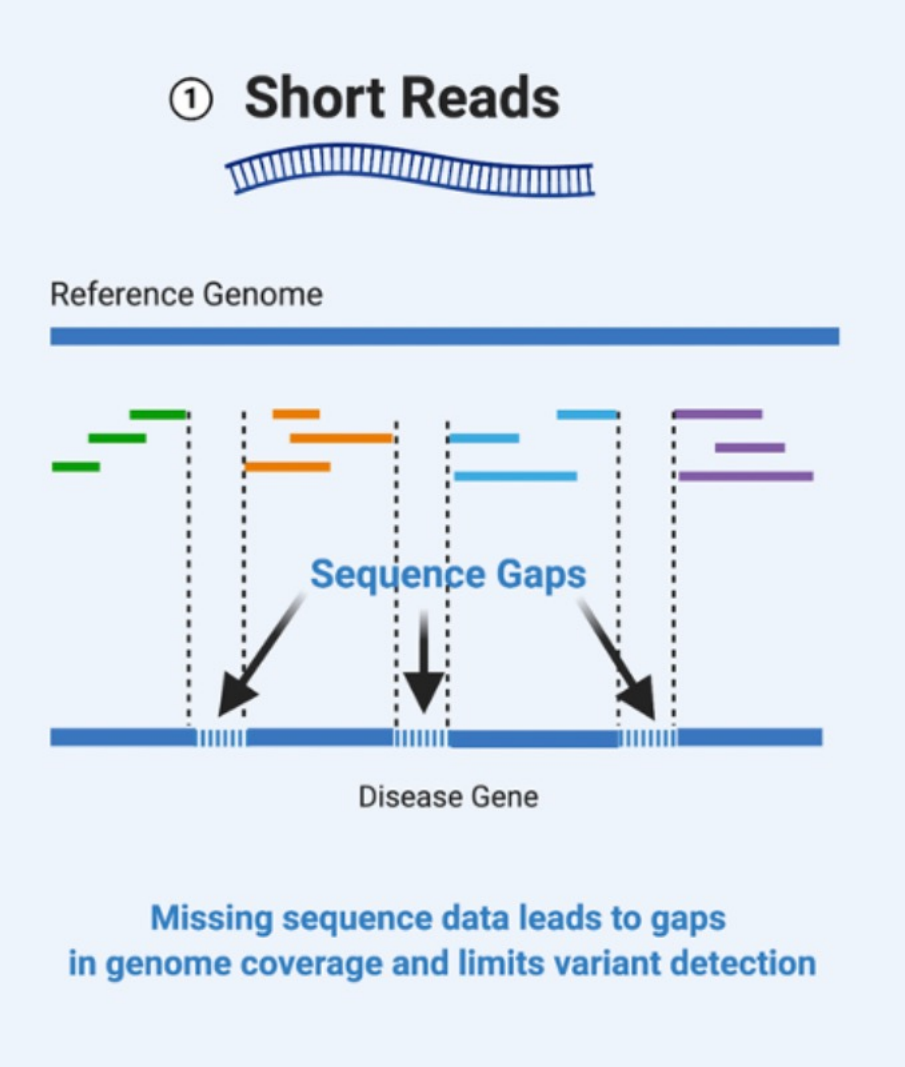
- ”Junk do not produc
- Howev eins
- The En se non-coding



Hypersensitive sites
CH$_3$CO (Epigenetic modifications)
RNA polymerase
CH$_3$CO
CH$_3$
5C
DNase-seq
FAIRE-seq
ChIP-seq
Computational predictions and RT-PCR
RNA-seq
Transcript
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)

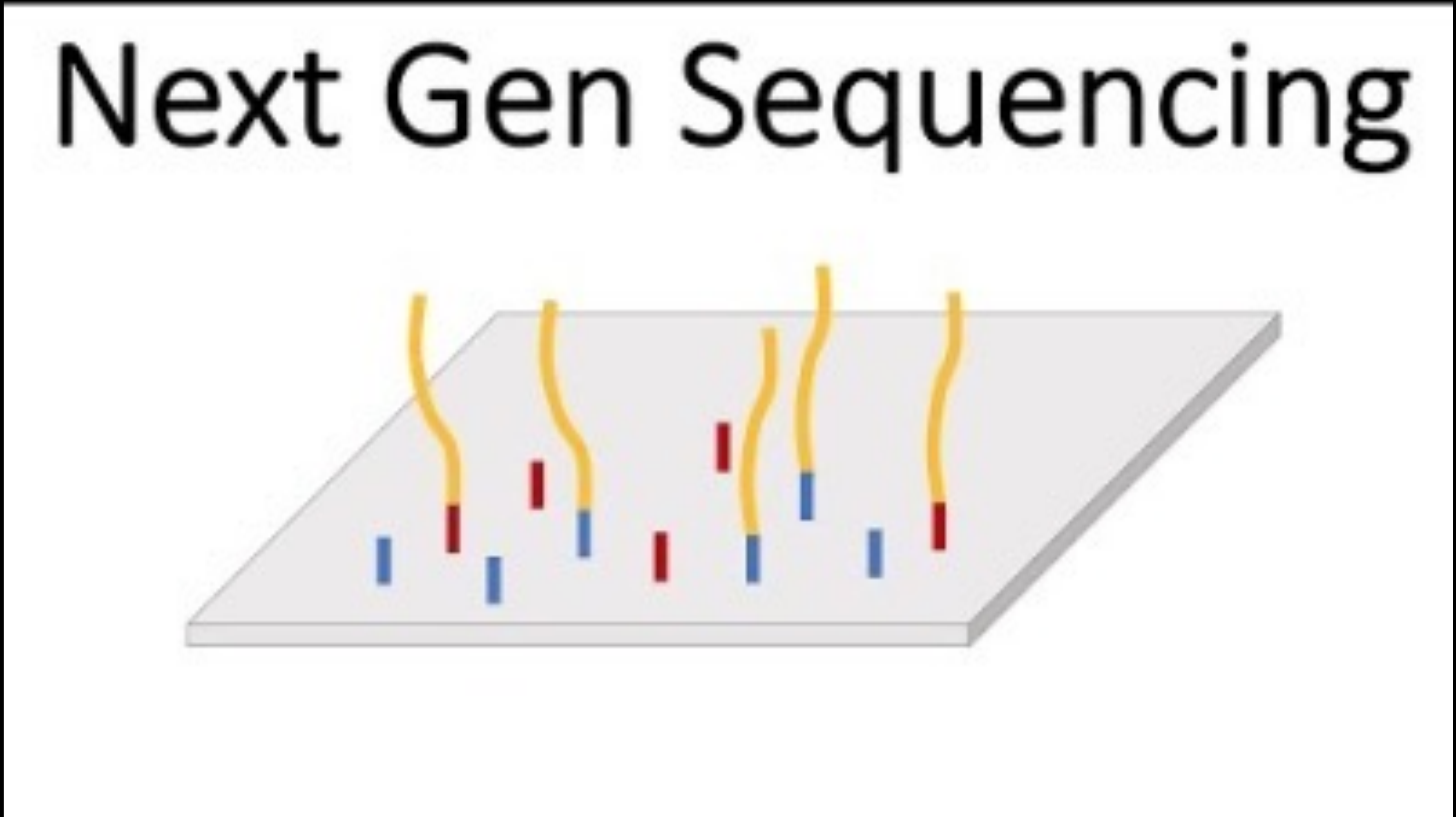
Waiting for the Technology to Catch Up

- The development of sequencing technologies that acted directly on the DNA sample generated millions of short reads at one time.
- The bottleneck at this juncture became algorithms to reassemble the sequence from these shorter sequences.
- Think of chopping up an encyclopedia into 250-letter chunks and then trying to put the books back together in the correct order.


① Short Reads
Reference Genome
Sequence Gaps
Disease Gene
Missing sequence data leads to gaps
in genome coverage and limits variant detection

https://www.hudsonalpha.org/hudsonalpha-researchers-use-highly-accurate-long-read-sequencing-technology-to-help-diagnose-rare-disease/

Advent of Next-Gen Sequencing (~2000)



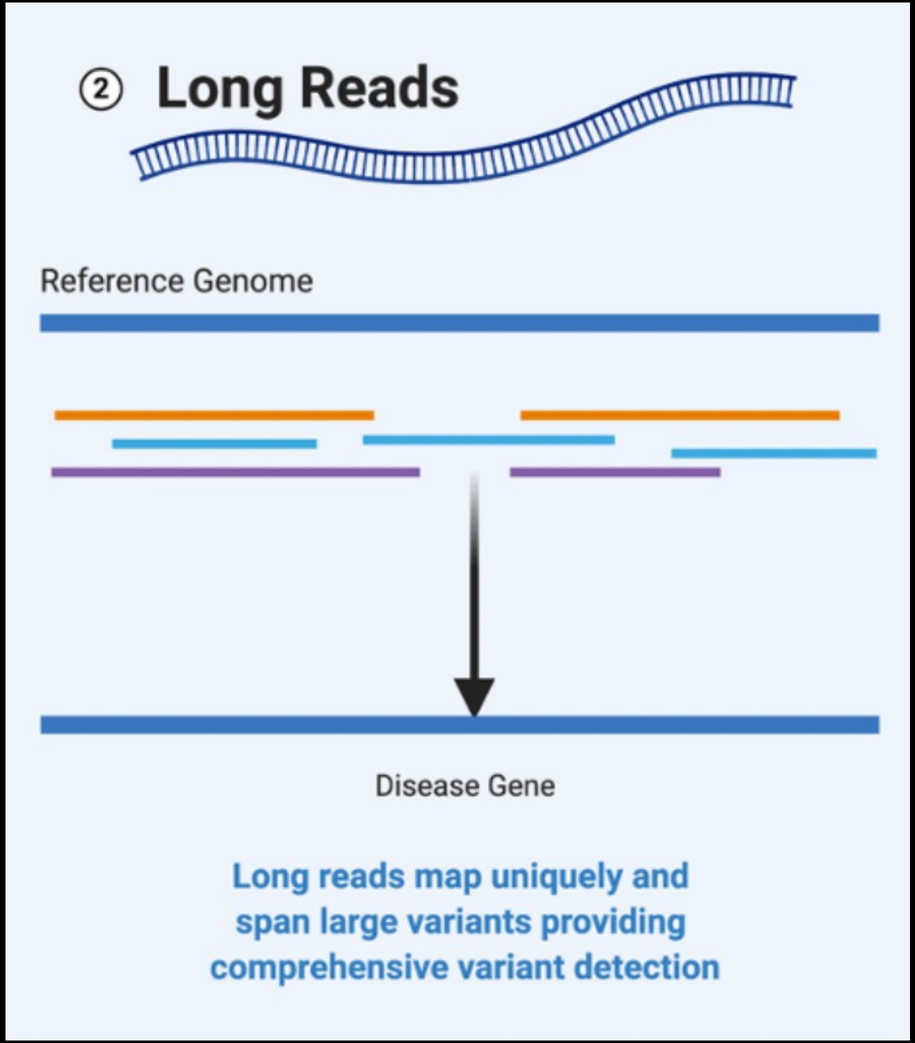
Next Gen Sequencing

https://www.youtube.com/watch?v=CZeN-IgjYCo

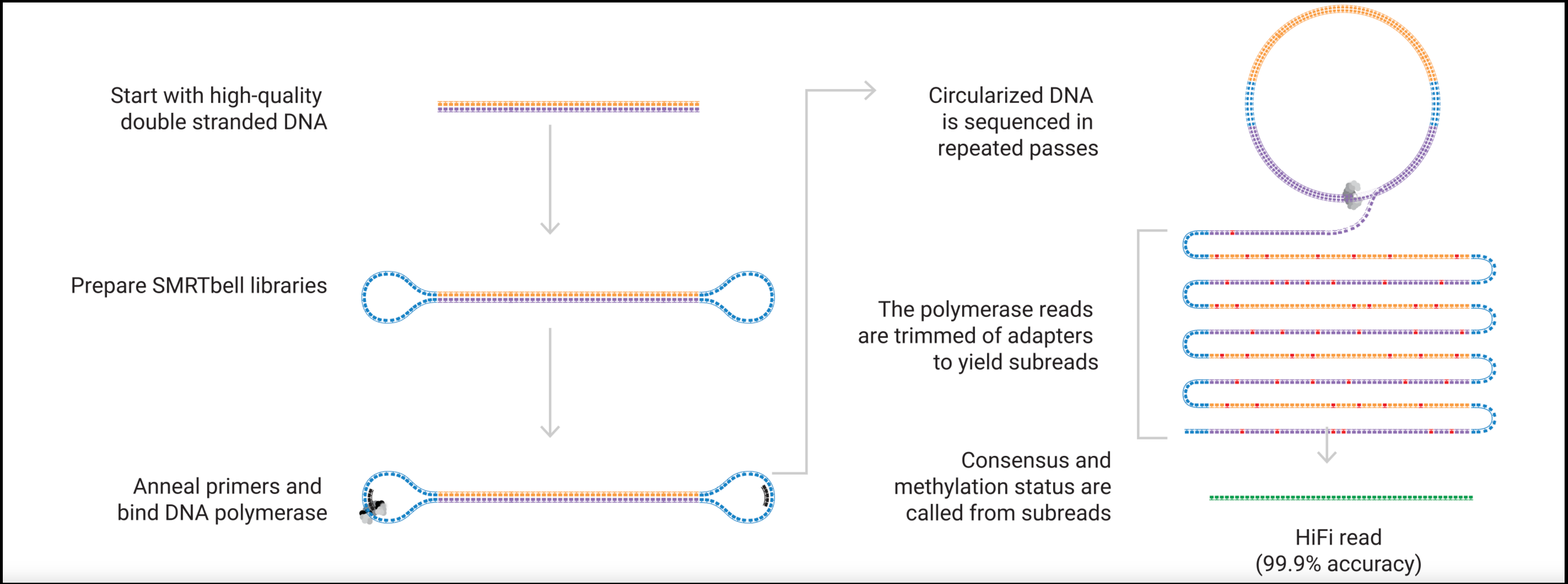
Next Next Gen Sequencing – Long Reads

- Around 2015, PacBio introduced single-molecule real-time (SMRT) sequencing to generate highly accurate (99.8%) long high-fidelity (HiFi) reads
- For the first time, phased sequencing allowed for determination of the contributions from either parent to a child's genome, improving haplotype determination for individual genomes.

https://www.nature.com/articles/s41587-019-0217-9


② Long Reads
Reference Genome
Disease Gene
Long reads map uniquely and span large variants providing comprehensive variant detection

https://www.hudsonalpha.org/hudsonalpha-researchers-use-highly-accurate-long-read-sequencing-technology-to-help-diagnose-rare-disease/



Start with high-quality double stranded DNA

Prepare SMRTbell libraries

Anneal primers and bind DNA polymerase

Circularized DNA is sequenced in repeated passes

The polymerase reads are trimmed of adapters to yield subreads

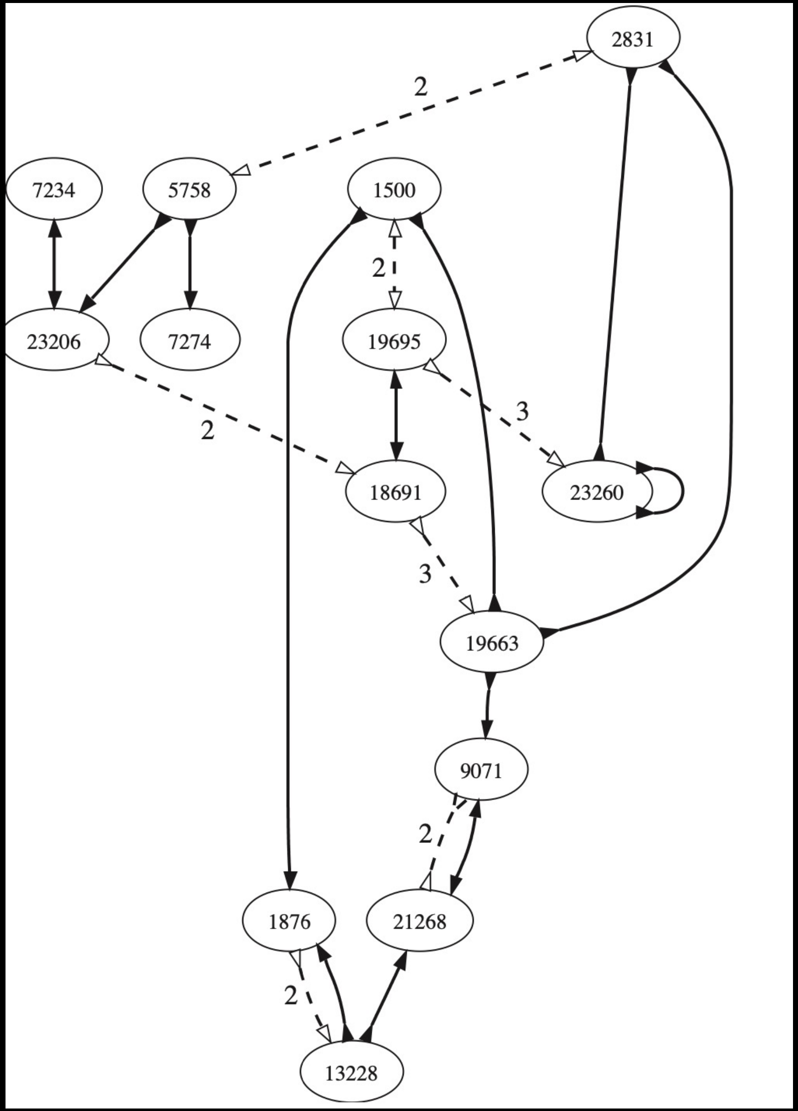
Consensus and methylation status are called from subreads

HiFi read (99.9% accuracy)

https://www.pacb.com/technology/hifi-sequencing/

Bioinformatics had to Catch Up as well

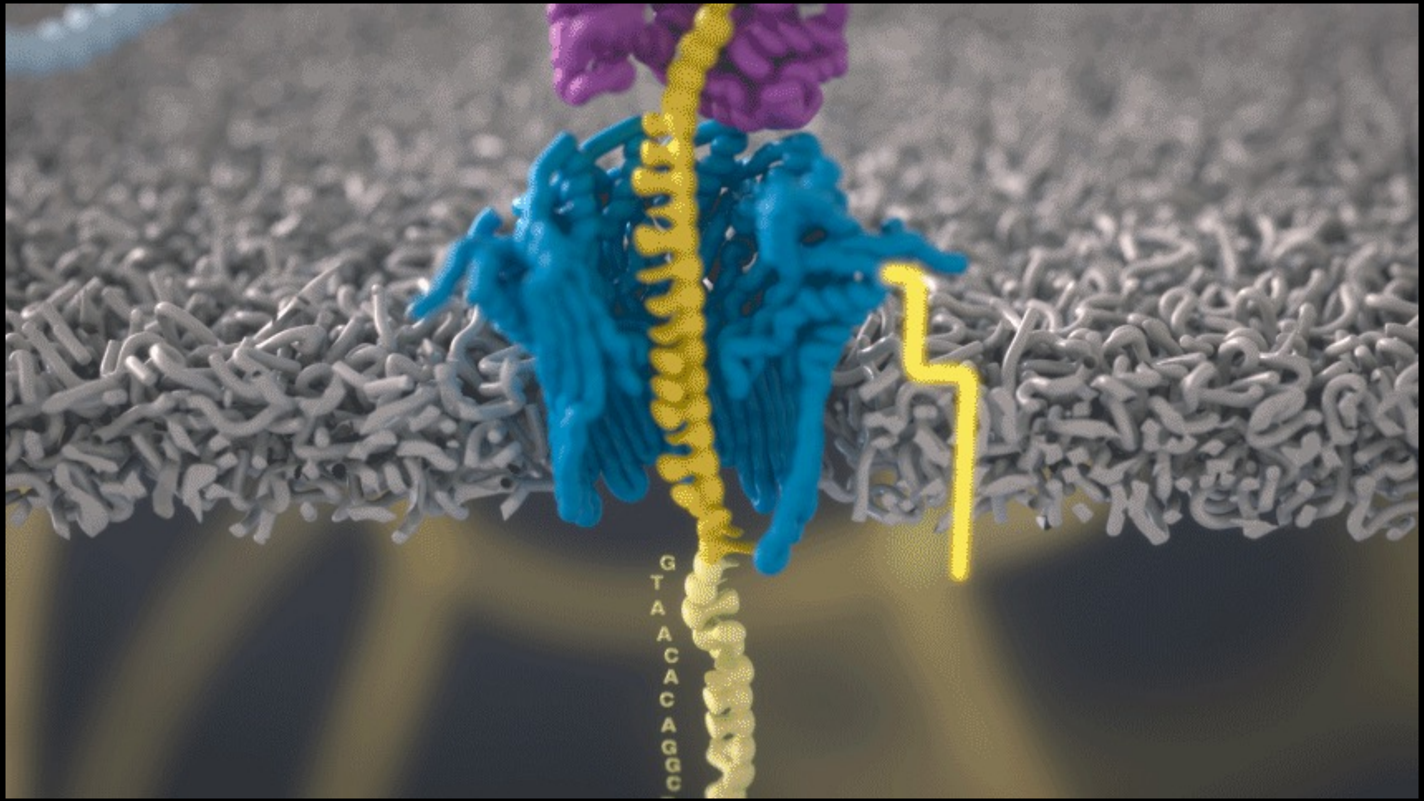
- Traditional assemblers built for short reads were based on algorithms that were optimized for that approach.
- String graph algorithms were more useful once HiFi long reads became more available.



E. W. Myers, The fragment assembly string graph. *Bioinformatics***21**, ii79–ii85 (2005).

Next Next Gen Sequencing – Oxford Nanopore Technology

- Around the same time, Oxford nanopore introduced a disruptive, electronic, single-molecule sensing system



Each bit of DNA passing through the "nanopores" in the flowscell is a charged molecule. The software allows a user to actually reverse the voltage on an individual molecule, which has the effect of ejecting it out of the nanopore. Each base (CGTA) has a distinctive squiggly line decoded by the sequencer.

https://stackoverflow.blog/2021/12/24/sequencing-your-dna-with-a-usb-dongle-and-open-source-code/

https://nanoporetech.com/how-it-works

Oxford Nanopore Technology – Sequencing in Space and Antarctica

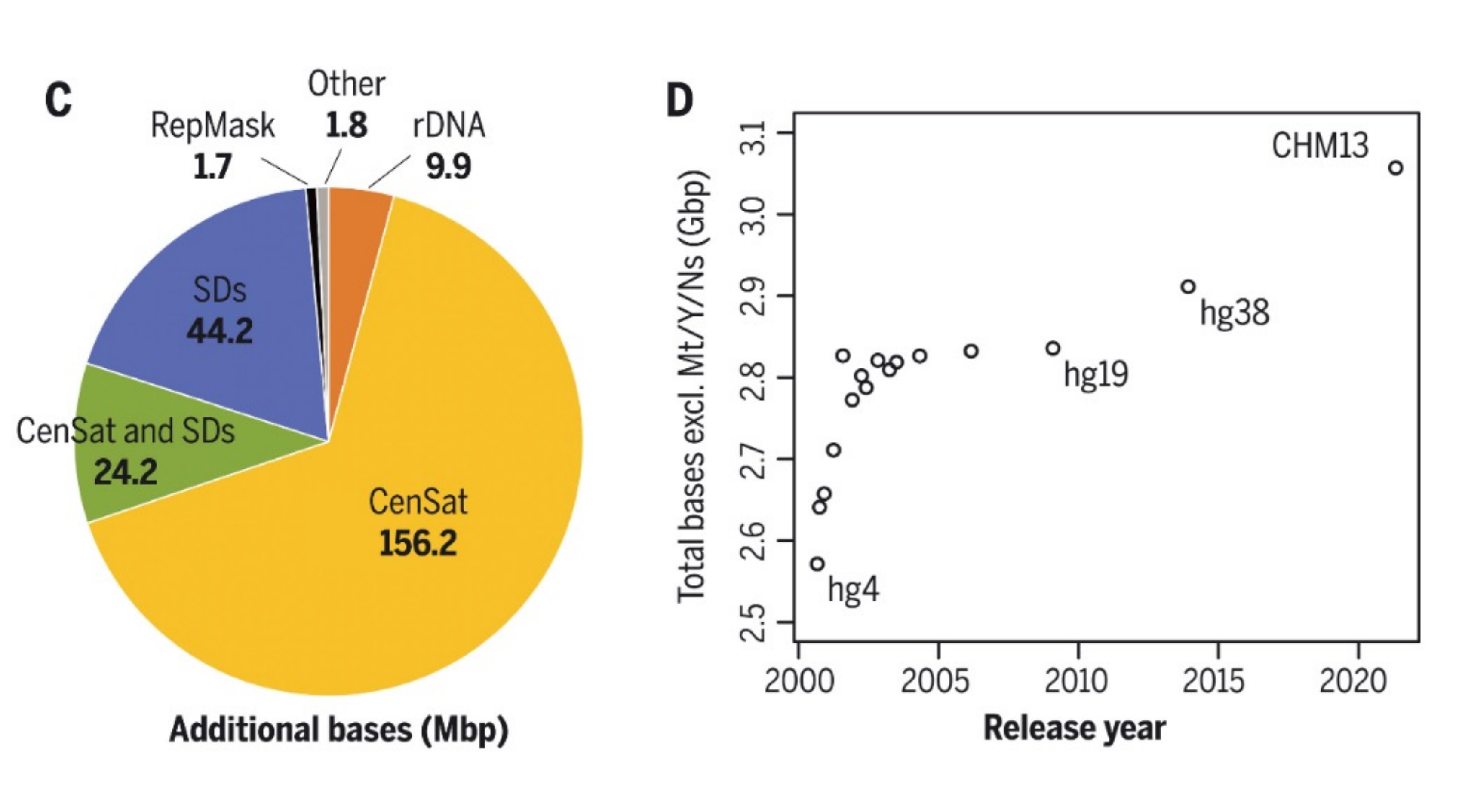
NASA Astronaut Kate Rubins sequenced DNA in space for the first time ever for the Biomolecule Sequencer investigation, using the MinION sequencing device.

Credits: NASA

https://www.frontiersin.org/articles/10.3389/fnano.2021.628861/full

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5362188/

Other Discoveries – Repeat Regions and Haplotypes, Regions that were not describable using existing sequencing technology



In six papers in Science from 2021-2022 the Telomere-to-Telomere (T2T) Consortium—named for the chromosomes' end caps—fills in all but five of the hundreds of remaining problem spots, leaving just 10 million bases and the Y chromosome only roughly known.

The T2T consortium recently announced in a tweet it had deposited a correct sequence assembly of the missing Y.

https://www.science.org/content/article/most-complete-human-genome-yet-reveals-previously-indecipherable-dna

https://www.science.org/doi/10.1126/science.abj6987

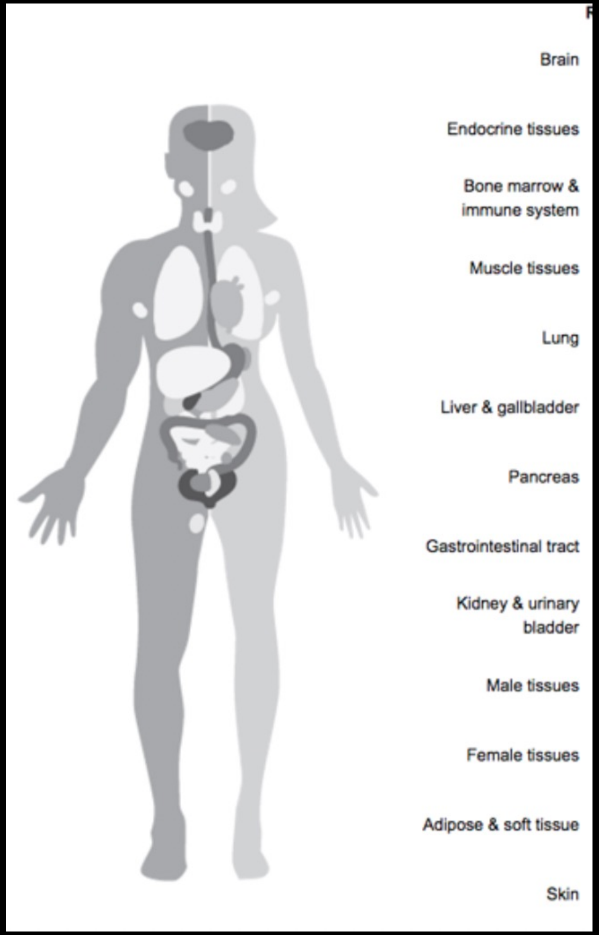
Is the Human Genome Finished?

- In 2022, the Telomere-to-Telomere (T2T) consortium finished the first truly complete sequence of a human genome.
- The resulting CHM13 reference assembly uncovered approximately 200 Mbp of new genomic sequence, comprising the centromeric satellite arrays, recent segmental duplications, and the short arms of all five acrocentric chromosomes.
- Combined, these regions contain thousands of new gene predictions and enable millions of new variant calls.
- The completed human genome will enable a new era of comparative genomics with a focus on segmental duplications and complex structural variation, ultimately providing a more complete link between genotype and phenotype.
- The next developmental stage will generate diploid human genomes as well as high quality genomes for non-model organisms.

https://www.science.org/doi/10.1126/science.abj6987

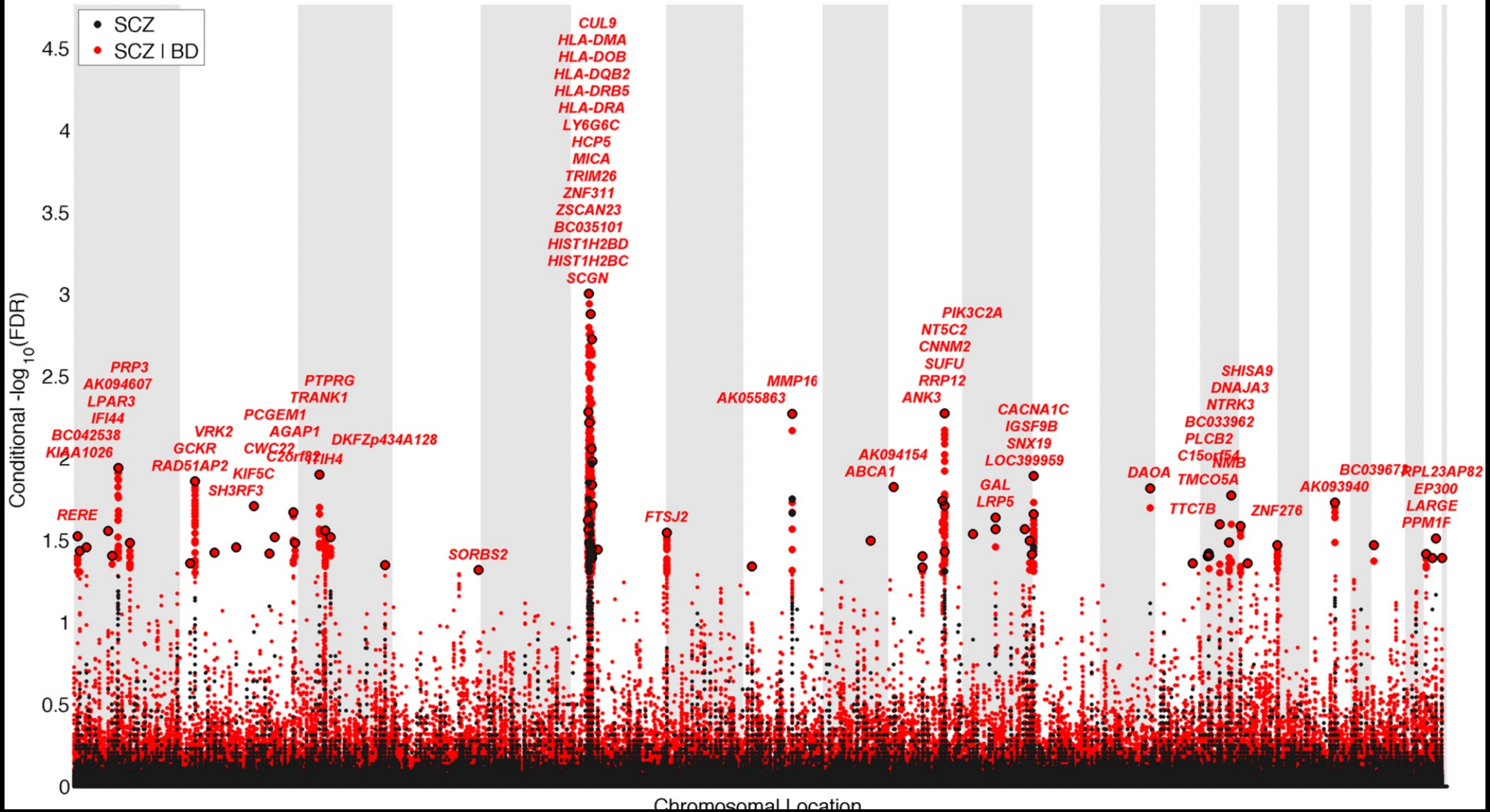
We have a complete genome now, so what?

- In 2005 **HapMap** started to characterize all the single-nucleotide polymorphisms (SNPs) across individuals and populations. SNPs are areas in the genome where one nucleotide has been altered.
- The **Cancer Genome Atlas** wants to map all the genetic abnormalities found in various types of cancer. These efforts will pinpoint the specific areas where our genomes differ as well as wehere tumor genomes differ from normal tissue samples.
- **Tissue Atlas** classifies the differential expression of protein coding regions among various tissues.
- These approaches are part of a new field called <u>“Precision Medicine”</u> that allows for targeted treatments for individual patients.



https://www.atlasantibodies.com/resources/human-protein-atlas/tissue-atlas/

GWAS – Genome Wide Association Studies



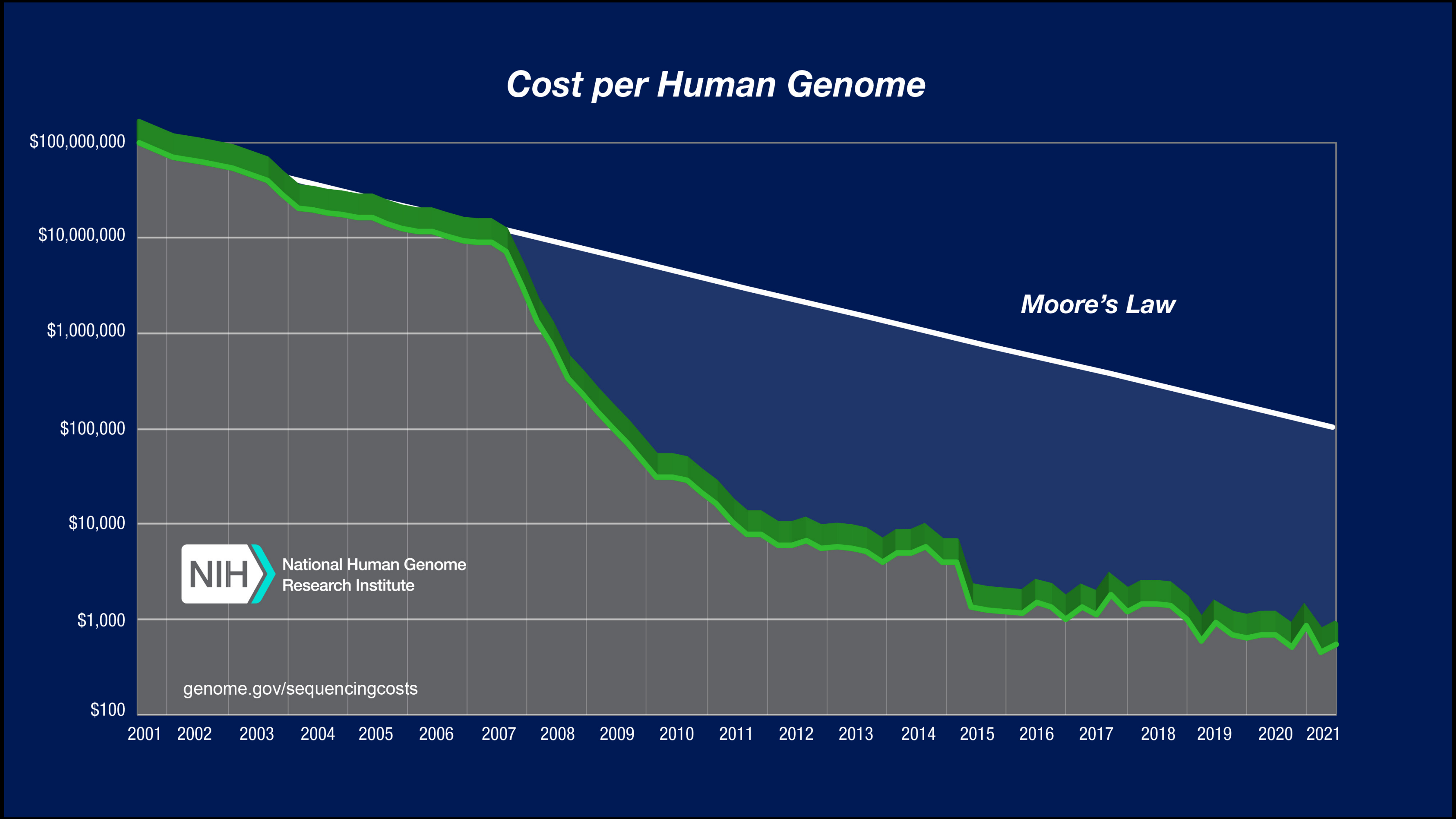
The schizophrenia (SCZ) GWAS summary statistics results were obtained from the PGC Schizophrenia Work Group [13], which consisted of 9,394 cases with schizophrenia or schizoaffective disorder and 12,462 controls (52% screened) from a total of 17 samples from 11 countries.

https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003455#s4

Sequencing is no longer the challenge

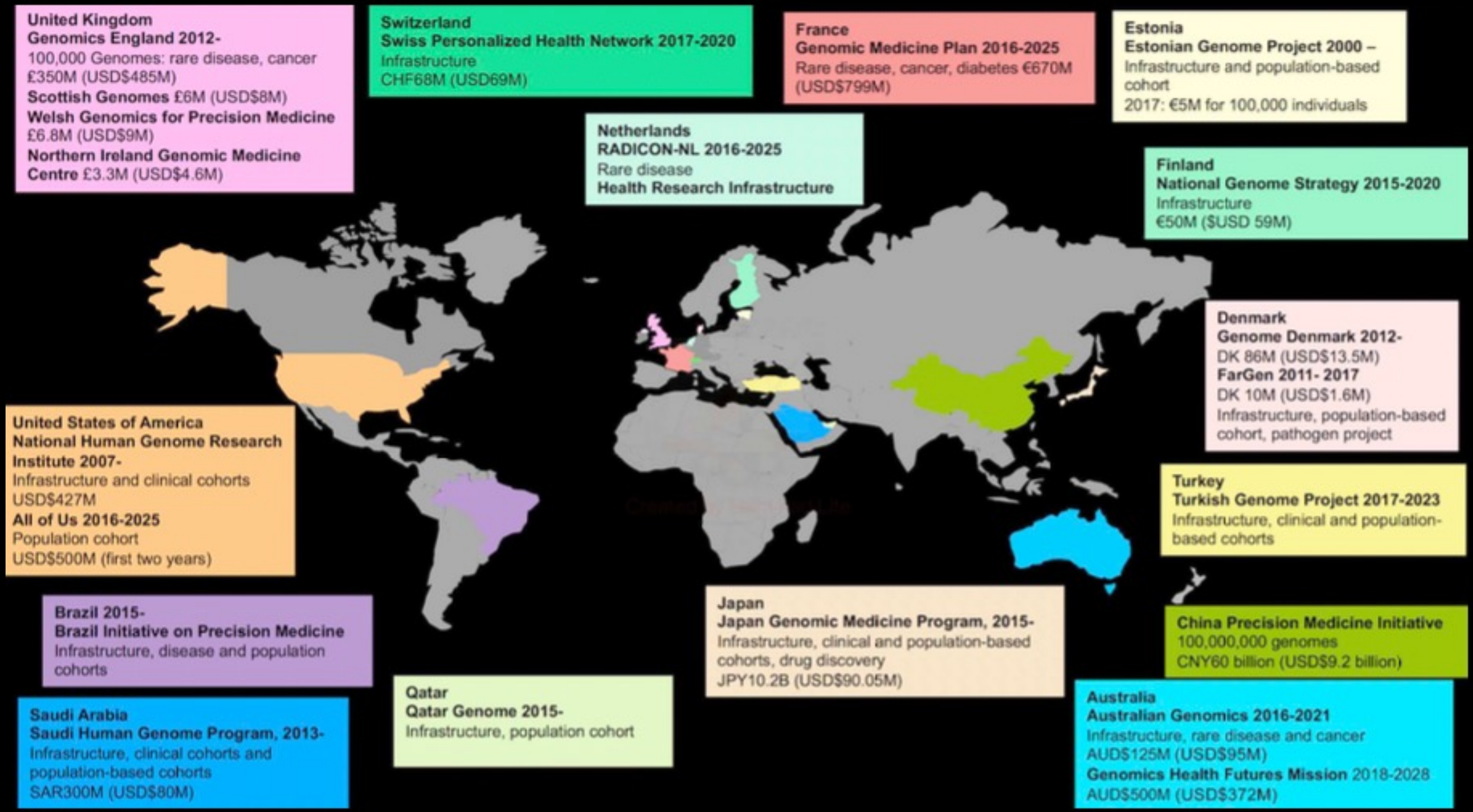
What do you think the next challenge will be?

The Human Genome Project only Sequenced One Representative Genome



Recently, the development of personal genome sequencers for $100 has been announced

https://www.darkdaily.com/2022/07/01/california-based-genomics-startup-secures-600-million-in-funding-to-deliver-100-whole-human-genome-with-its-new-high-throughput-low-cost-sequencing-platform/



United Kingdom
Genomics England 2012-
100,000 Genomes: rare disease, cancer
£350M (USD$485M)
Scottish Genomes £6M (USD$8M)
Welsh Genomics for Precision Medicine
£6.8M (USD$9M)
Northern Ireland Genomic Medicine Centre £3.3M (USD$4.6M)

Switzerland
Swiss Personalized Health Network 2017-2020
Infrastructure
CHF68M (USD$69M)

France
Genomic Medicine Plan 2016-2025
Rare disease, cancer, diabetes €670M (USD$799M)

Estonia
Estonian Genome Project 2000 –
Infrastructure and population-based cohort
2017: €5M for 100,000 individuals

Netherlands
RADICON-NL 2016-2025
Rare disease
Health Research Infrastructure

Finland
National Genome Strategy 2015-2020
Infrastructure
€50M ($USD 59M)

Denmark
Genome Denmark 2012-
DK 86M (USD$13.5M)
FarGen 2011- 2017
DK 10M (USD$1.6M)
Infrastructure, population-based cohort, pathogen project

United States of America
National Human Genome Research Institute 2007-
Infrastructure and clinical cohorts
USD$427M
All of Us 2016-2025
Population cohort
USD$500M (first two years)

Turkey
Turkish Genome Project 2017-2023
Infrastructure, clinical and population-based cohorts

Brazil 2015-
Brazil Initiative on Precision Medicine
Infrastructure, disease and population cohorts

Japan
Japan Genomic Medicine Program, 2015-
Infrastructure, clinical and population-based cohorts, drug discovery
JPY10.2B (USD$90.05M)

China Precision Medicine Initiative
100,000,000 genomes
CNY60 billion (USD$9.2 billion)

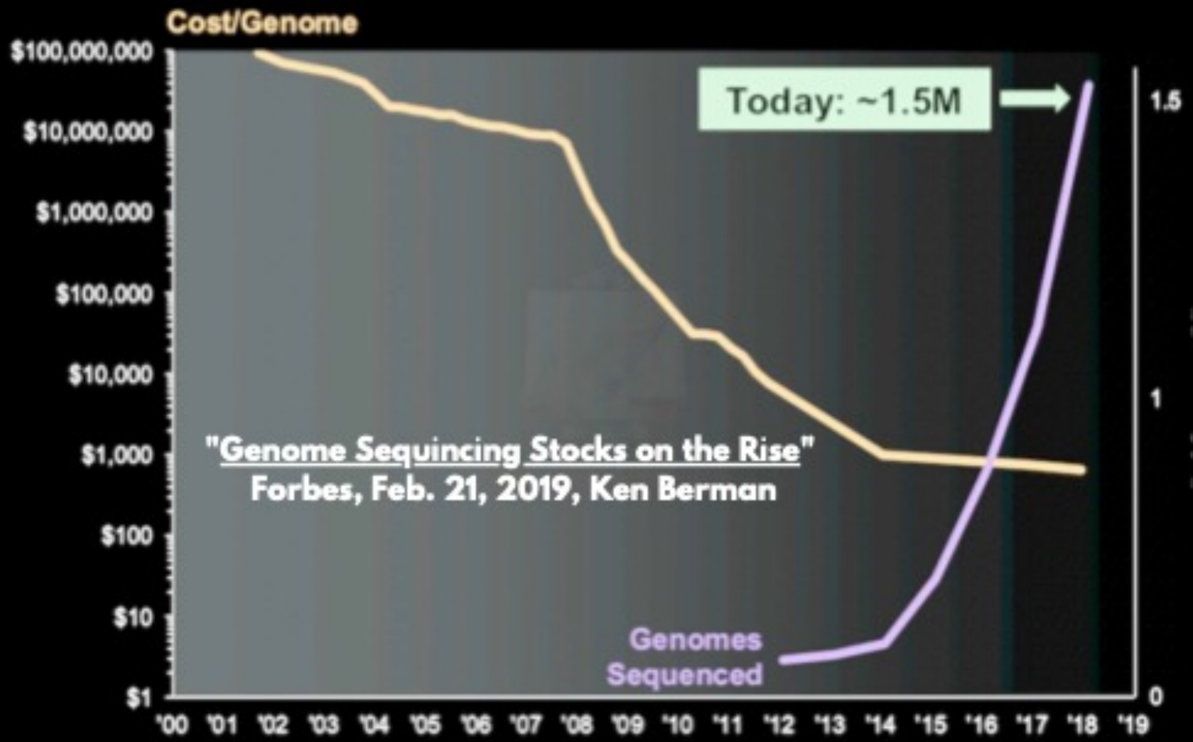
Saudi Arabia
Saudi Human Genome Program, 2013-
Infrastructure, clinical cohorts and population-based cohorts
SAR300M (USD$80M)

Qatar
Qatar Genome 2015-
Infrastructure, population cohort

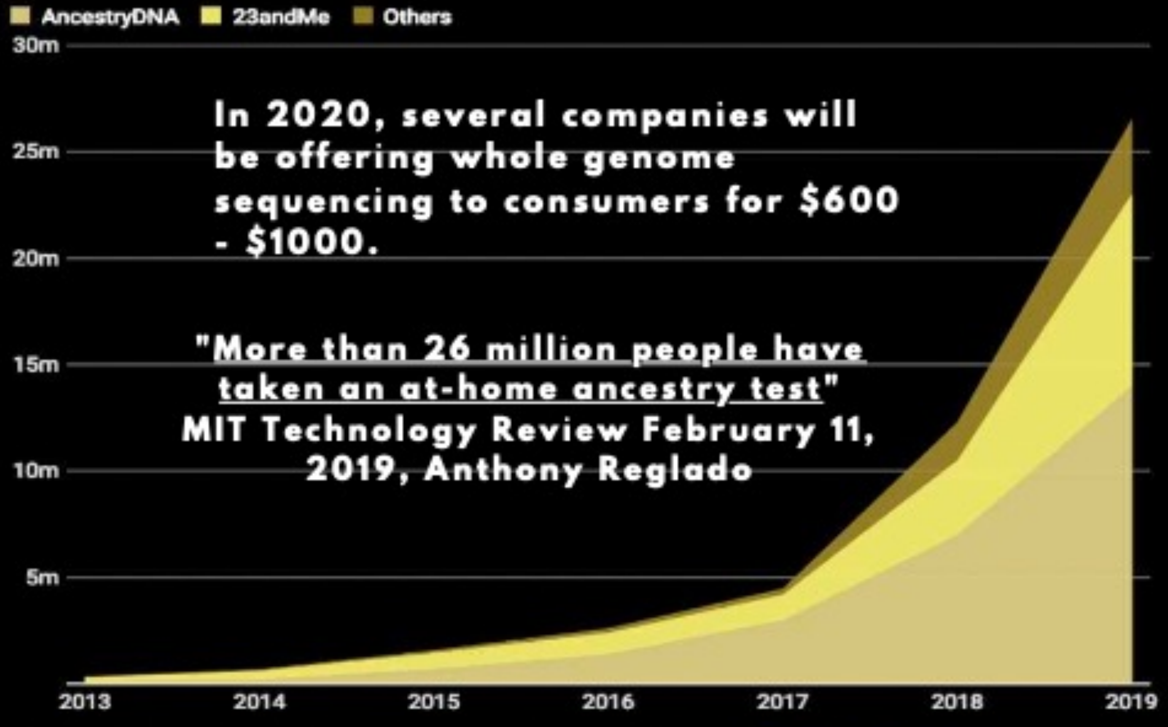
Australia
Australian Genomics 2016-2021
Infrastructure, rare disease and cancer
AUD$125M (USD$95M)
Genomics Health Futures Mission 2018-2028
AUD$500M (USD$372M)

STARK ET AL. INTEGRATING GENOMICS INTO HEALTHCARE: A GLOBAL RESPONSIBILITY.
THE AMERICAN JOURNAL OF HUMAN GENETICS 104.1 (2019): 13-20.

MASSIVE AMOUNTS OF DATA



"<u>Genome Sequincing Stocks on the Rise</u>"
Forbes, Feb. 21, 2019, Ken Berman

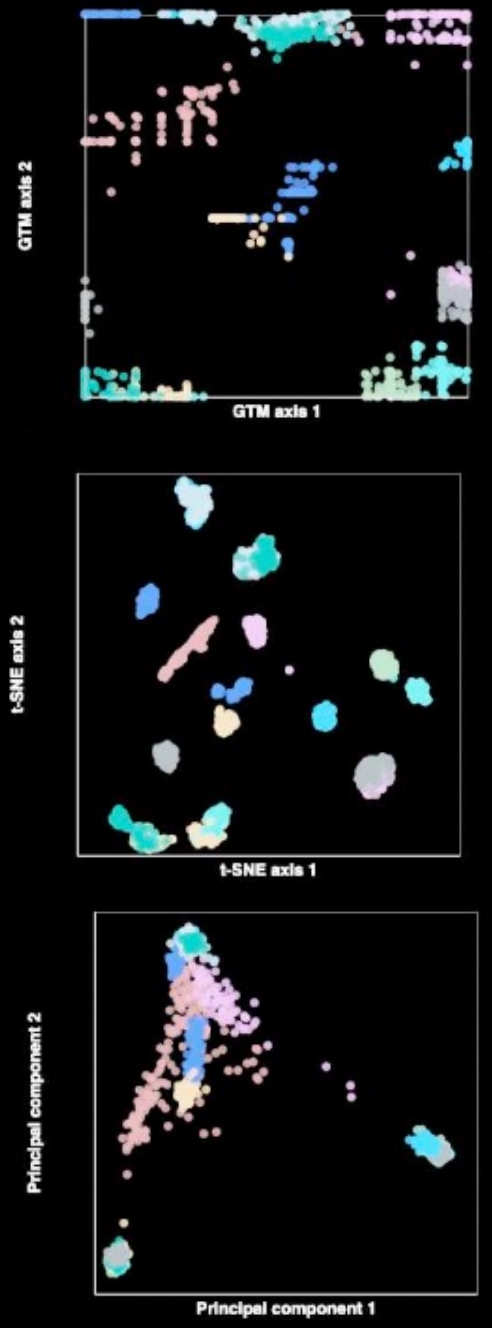


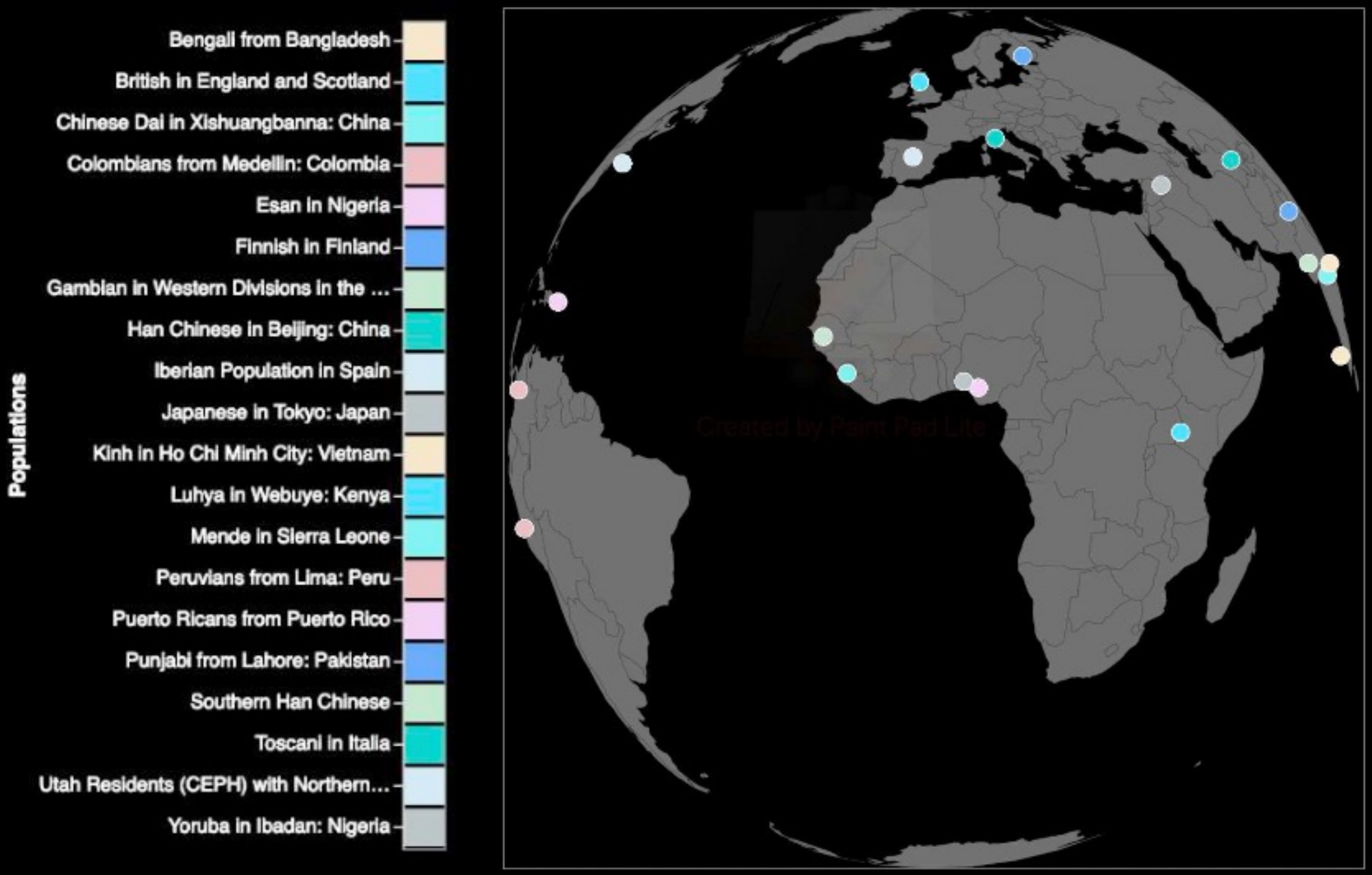
In 2020, several companies will be offering whole genome sequencing to consumers for $600 - $1000.

"<u>More than 26 million people have taken an at-home ancestry test</u>"
MIT Technology Review February 11, 2019, Anthony Reglado

Chart: MIT Technology Review • Source: Company reports, Leah Larkin, ISOGG • Created with Datawrapper

- EACH GENOME GENERATES 100 GB DATA FOR DOWNSTREAM ANALYSIS,
- THIS REQUIRES STORAGE BEYOND WHAT THE TYPICAL HPC ACCOMMODATES
- 100 MILLION GENOMES X 100 GIGABYTES OF DATA = 10 EXABYTES OF DATA




Populations
Bengali from Bangladesh
British in England and Scotland
Chinese Dai in Xishuangbanna: China
Colombians from Medellin: Colombia
Esan in Nigeria
Finnish in Finland
Gambian in Western Divisions in the ...
Han Chinese in Beijing: China
Iberian Population in Spain
Japanese in Tokyo: Japan
Kinh in Ho Chi Minh City: Vietnam
Luhya in Webuye: Kenya
Mende in Sierra Leone
Peruvians from Lima: Peru
Puerto Ricans from Puerto Rico
Punjabi from Lahore: Pakistan
Southern Han Chinese
Toscani in Italia
Utah Residents (CEPH) with Northern...
Yoruba in Ibadan: Nigeria

GASPAR, H.A., BREEN, G. PROBABILISTIC ANCESTRY MAPS: A
METHOD TO ASSESS AND VISUALIZE POPULATION SUBSTRUCTURES IN GENETICS.
BMC BIOINFORMATICS 20, 116 (2019).
HTTPS://DOI.ORG/10.1186/S12859-019-2680-1

PERSONAL GENOMICS FOR HEALTHCARE WILL EXCEED CLOUD DEMAND FOR CLINICAL AND ACADEMIC GENOMICS RESEARCH

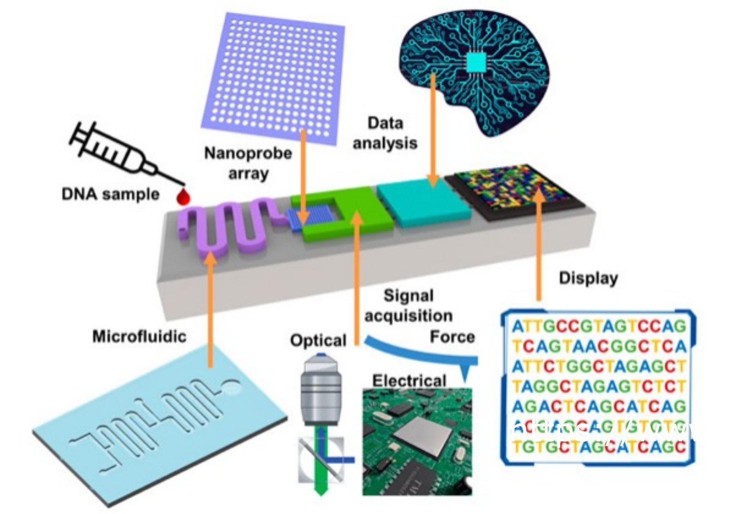
BY 2024
$340 BILLION
DOLLARS/YEAR WILL BE SPENT ON CLOUD COMPUTING

THE U.S. SPENDS
35%
OF OVERALL WORLD FUNDING ON GENOMICS RESEARCH

BY 2025
60 MILLION
WILL HAVE THEIR GENOME SEQUENCED IN A HEALTHCARE CONTEXT

Sources: https://www.marketwatch.com/press-release/global-cloud-computing-market-size-2019-industry-trends-share-statistics-worldwide-overview-key-players-analysis-research-by-types-services-regional-outlook-and-forecasts-till-2024-2019-11-13
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2576262/,
https://www.cell.com/ajhg/fulltext/S0002-9297(18)30422-1

The Promise of New Technologies



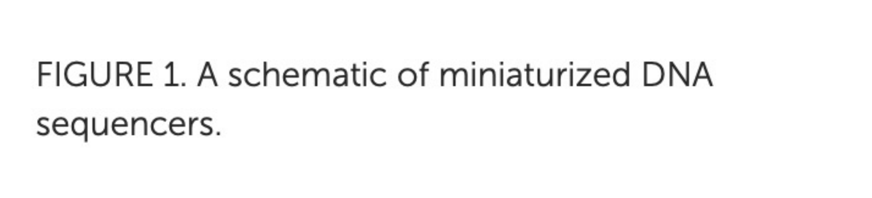
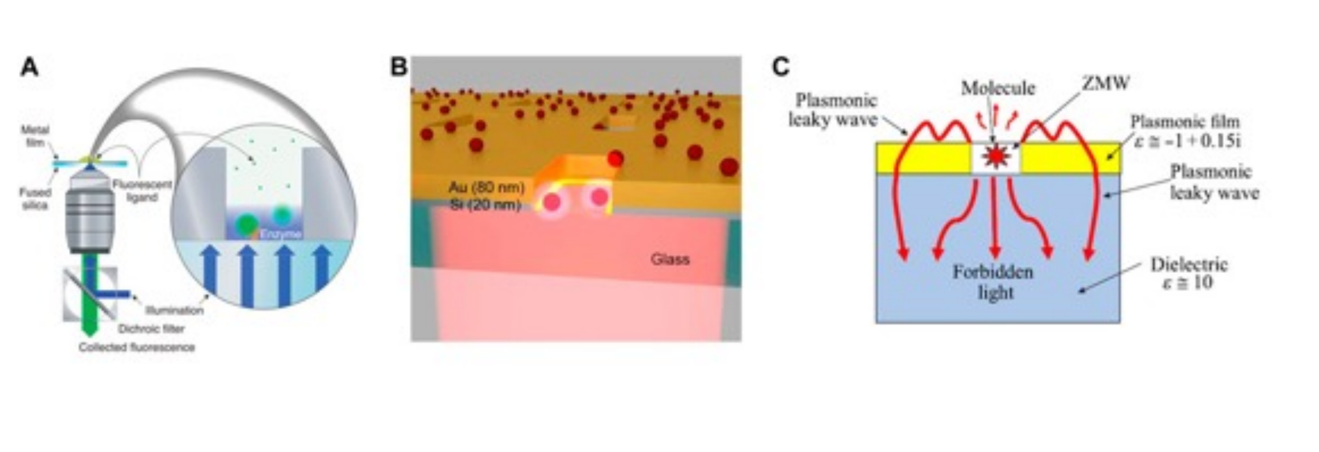


FIGURE 1. A schematic of miniaturized DNA sequencers.

ontiersin.org/articles/10.3389/fnano.2021.



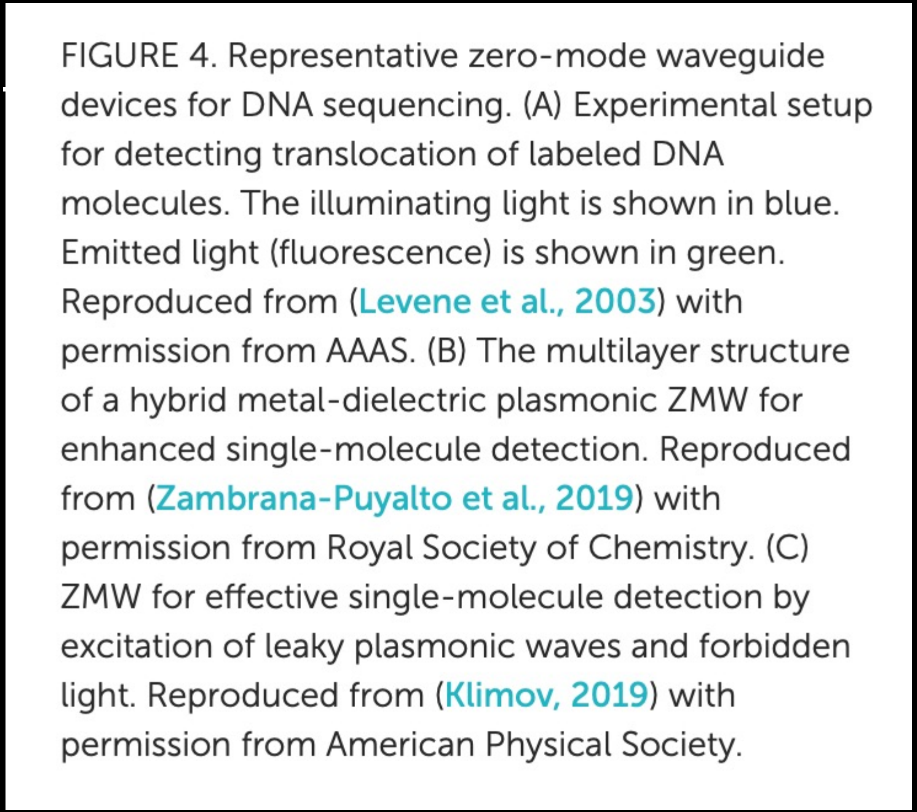


FIGURE 4. Representative zero-mode waveguide devices for DNA sequencing. (A) Experimental setup for detecting translocation of labeled DNA molecules. The illuminating light is shown in blue. Emitted light (fluorescence) is shown in green. Reproduced from (Levene et al., 2003) with permission from AAAS. (B) The multilayer structure of a hybrid metal-dielectric plasmonic ZMW for enhanced single-molecule detection. Reproduced from (Zambrana-Puyalto et al., 2019) with permission from Royal Society of Chemistry. (C) ZMW for effective single-molecule detection by excitation of leaky plasmonic waves and forbidden light. Reproduced from (Klimov, 2019) with permission from American Physical Society.

https://www.frontiersin.org/articles/10.3389/fnano.2021.628861/full

Discussion