A brief history of the genomic era

Sean Davis, MD, PhD

March 21, 2022

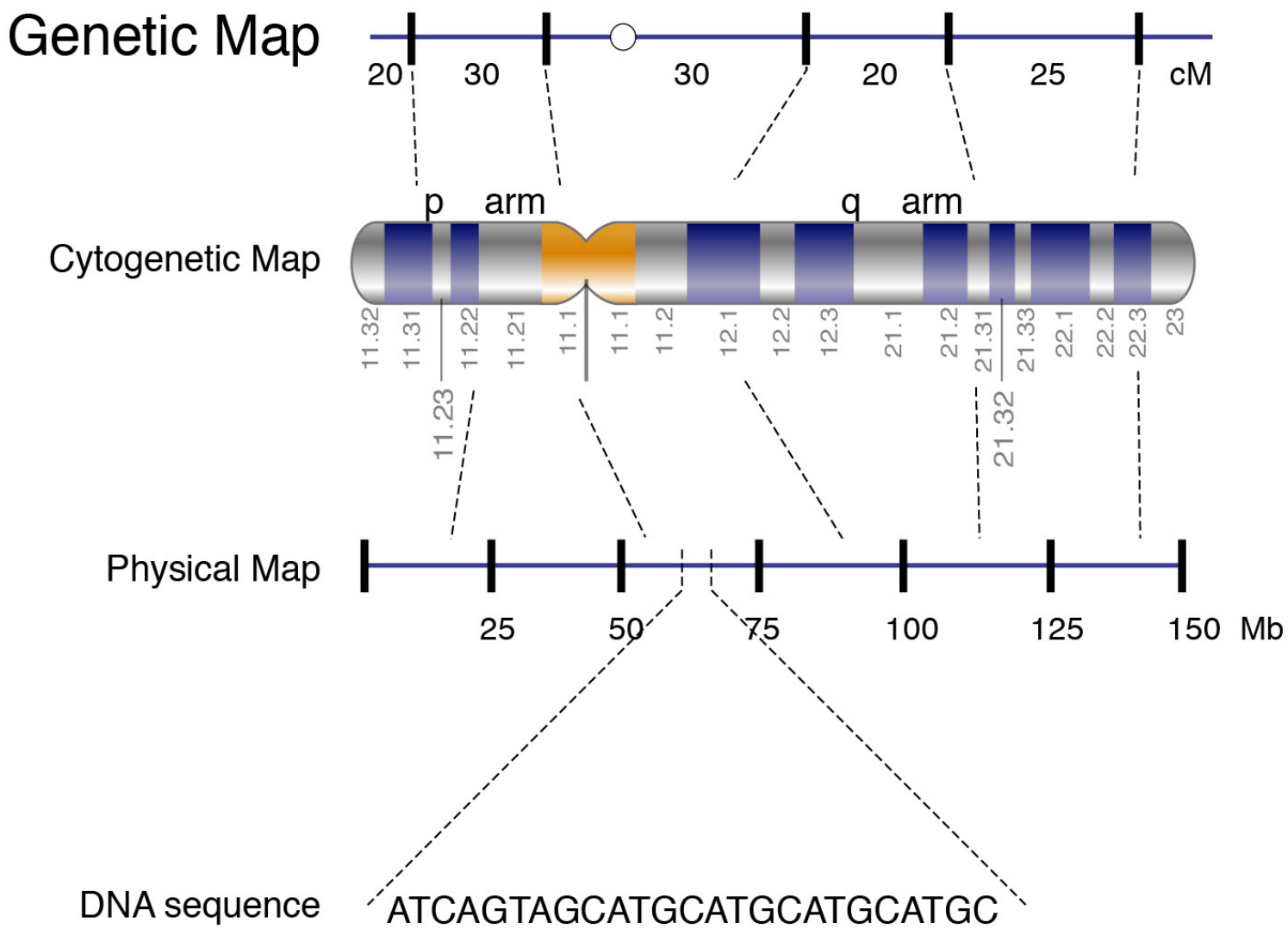
UVI BIO 361

What is a genome?

A genome is an organism's complete set of deoxyribonucleic acid (DNA), a chemical compound that contains the genetic instructions needed to develop and direct the activities of every organism.

https://www.genome.gov/human-genome-project/Completion-FAQ

What is a genome?



The Human Genome Project

- International scientific research effort to determine the complete sequence of base pairs that make up human DNA and all the genes it contains
- It was and still is the world's largest collaborative biological project.
- Planning started in 1984
- Project launched in 1990
- Declared complete in 2003

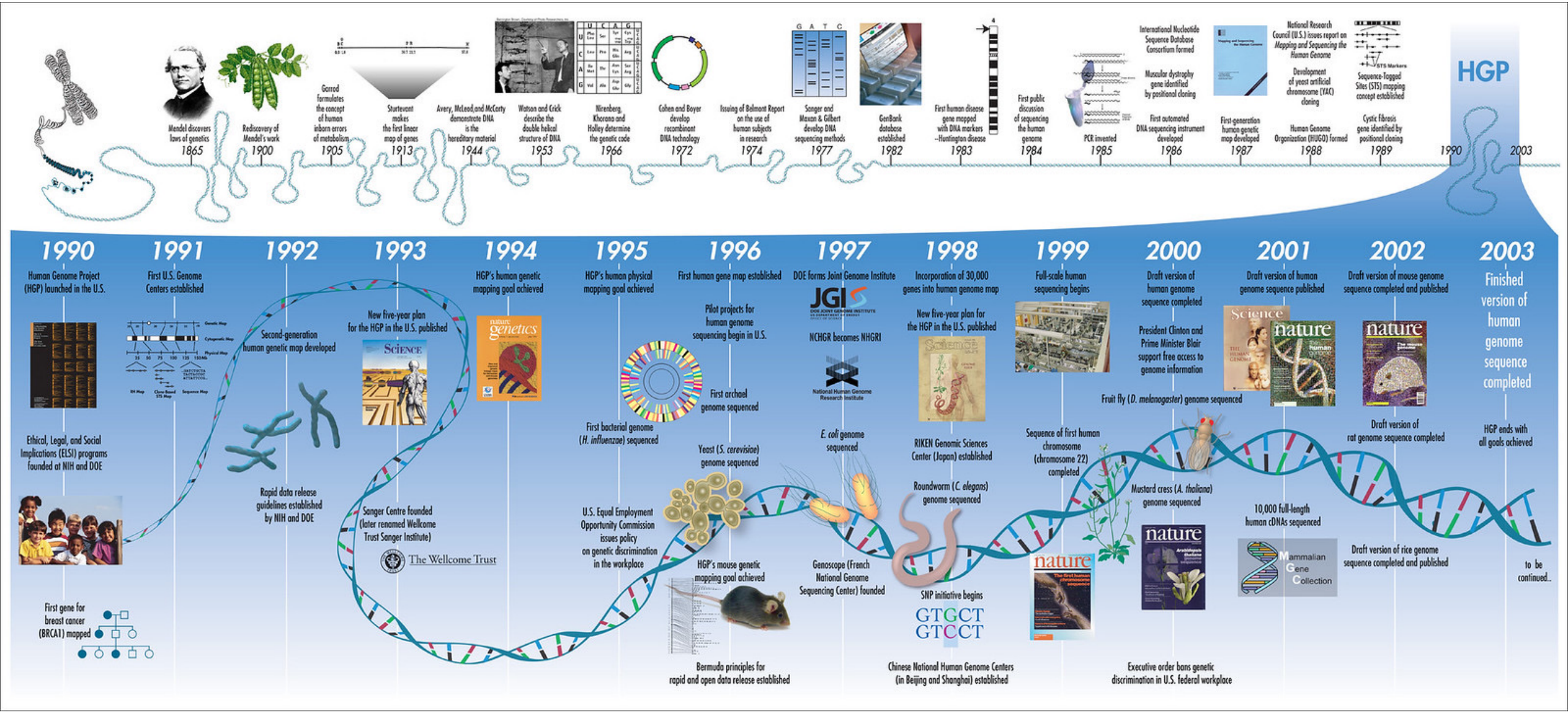
The Human Genome Project

- Original aim was to map all bases in a human haploid reference genome.
- The genome of any individual is unique.
- The reference human genome is derived from a small number of individuals and then assembling them to create a complete sequence for each chromosome.
- The ”finished” human genome is, therefore, a mosaic with contributions from several volunteers.

Goals of the Human Genome Project

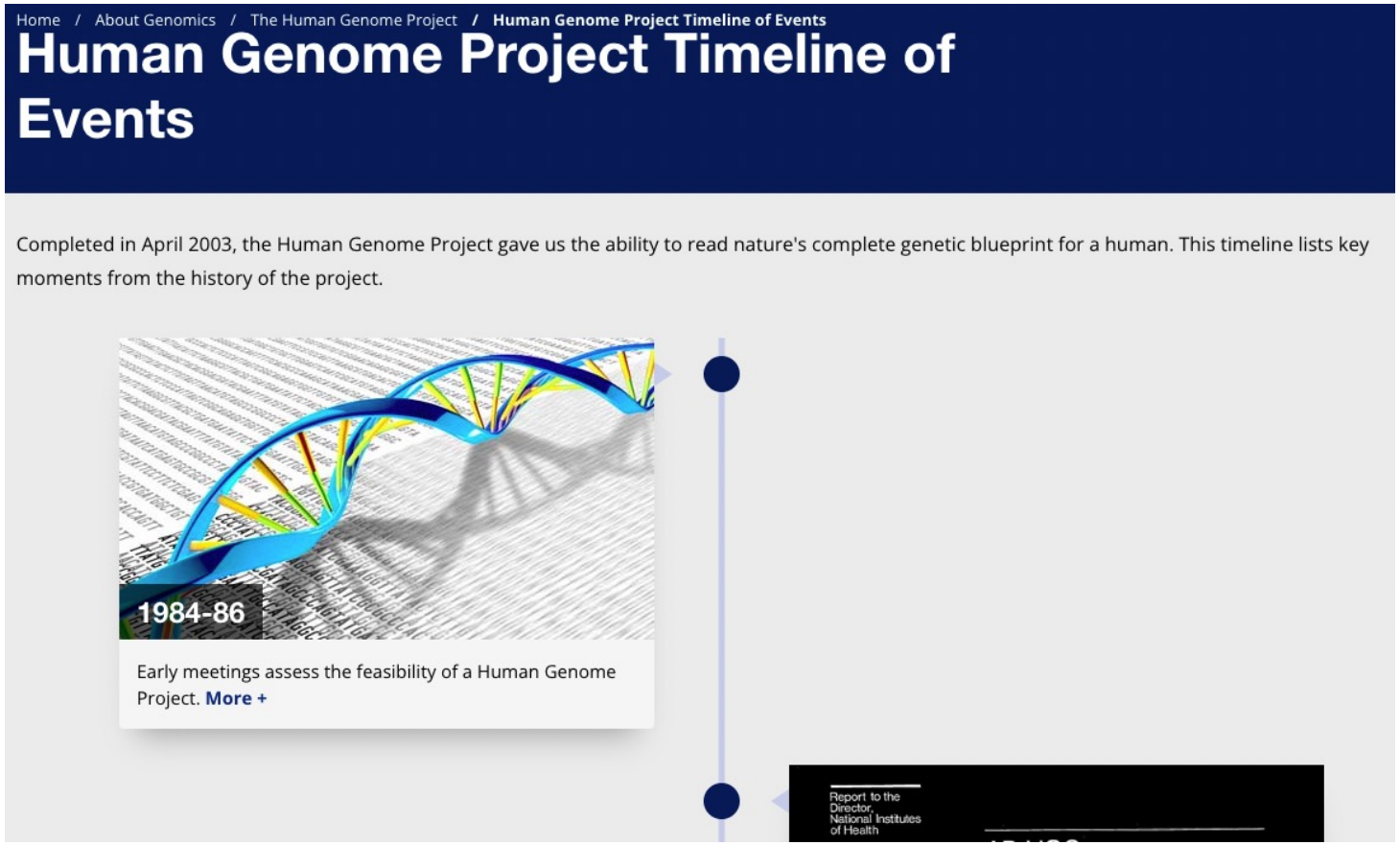
- To determine one complete DNA reference genome that includes all 3 billion bases
- To identify all ~22,000 genes in human DNA, both functionally and physically
- To build efficient approaches for analyzing the data
- To store all data in databases for access and use
- To allow public and private sector access and use of the information and technologies arising from the project
- To address ethical, legal, and social issues.

Human Genome Project Timeline

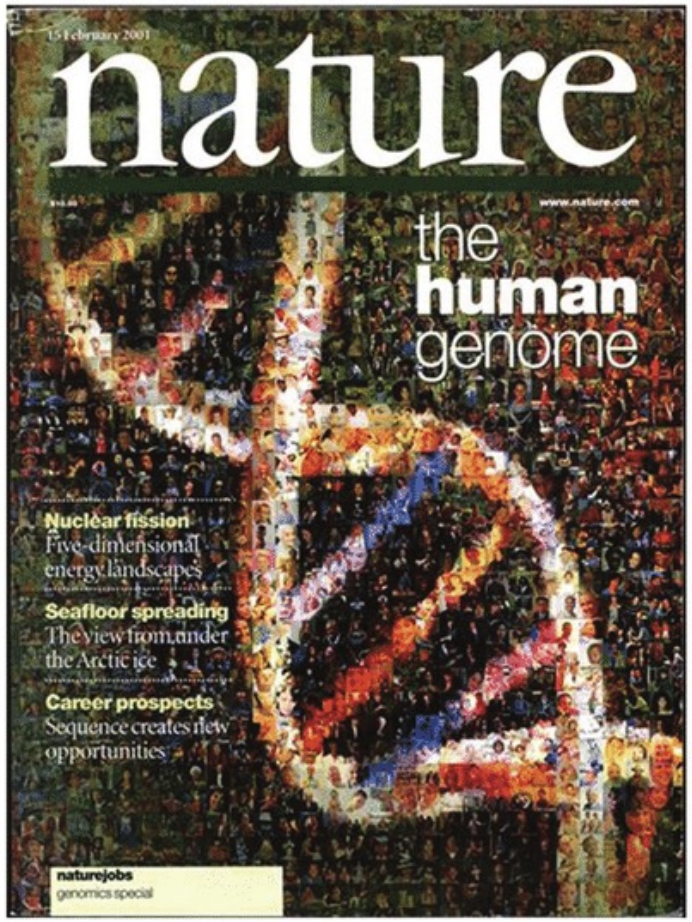


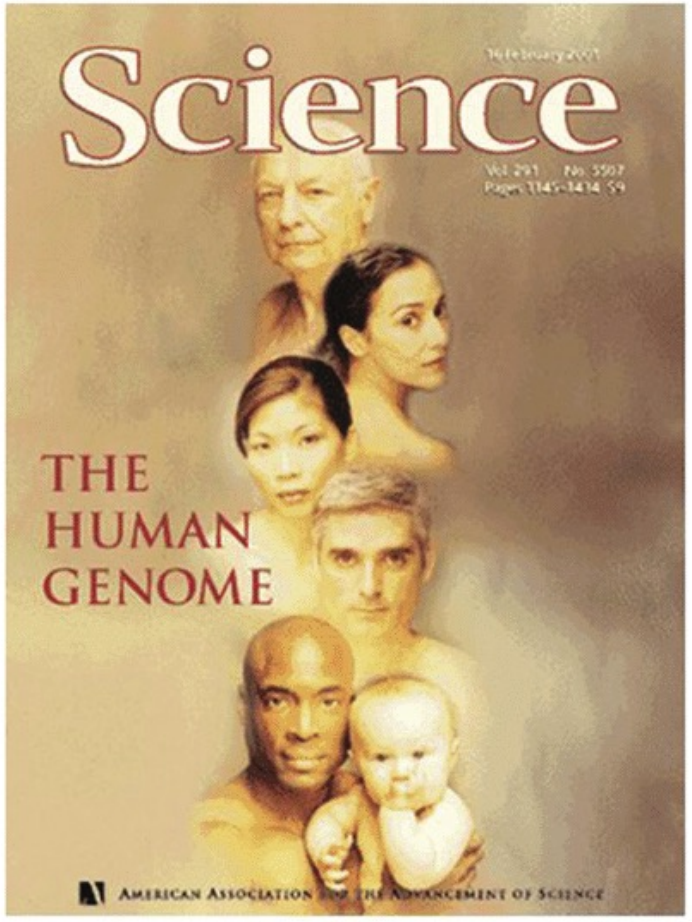
Interactive timeline

- https://www.genome.gov/human-genome-project/Timeline-of-Events

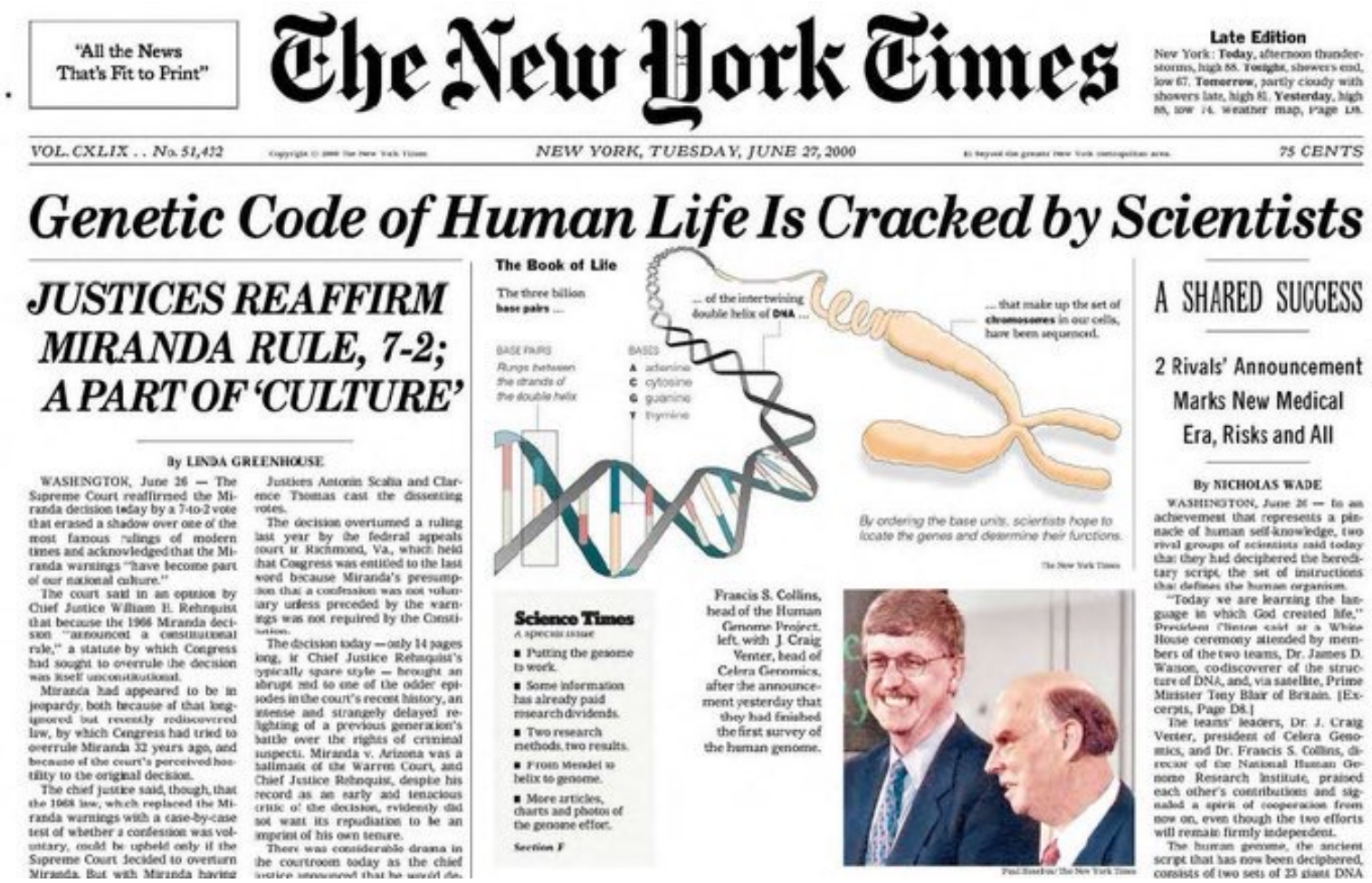


HGP 2001: Draft completed





Scientists finish 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**, shower's 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 NEW YORK, TUESDAY, JUNE 27, 2000 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 relighting 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 ... of the intertwining double helix of DNA ... that make up the set of chromosomes in our cells, have been sequenced.

BASE PAIRS
Rungs between the strands of the double helix

BASES
A adenine
C cytosine
G guanine
T thymine

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

US DNA day: April 25

108TH CONGRESS
1ST SESSION

S. CON. RES. 10

Designating April 2003 as "Human Genome Month" and April 25 as "DNA Day".

IN THE SENATE OF THE UNITED STATES

FEBRUARY 27, 2003

Mr. GREGG (for himself, Mr. KENNEDY, Ms. SNOWE, and Mr. DASCHLE) submitted the following concurrent resolution; which was considered and agreed to

---

CONCURRENT RESOLUTION

Designating April 2003 as "Human Genome Month" and April 25 as "DNA Day".

Whereas April 25, 2003, will mark the 50th anniversary of the description of the double-helix structure of DNA by James D. Watson and Francis H.C. Crick, considered by many to be one of the most significant scientific discoveries of the 20th Century;

The New York Times

MATTER

Scientists Finish the Human Genome at Last

The complete genome uncovered more than 100 new genes that are probably functional, and many new variants that may be linked to diseases.

By Carl Zimmer

Published July 23, 2021 Updated July 26, 2021

Sign up for Science Times Get stories that capture the wonders of nature, the cosmos and the human body. Get it sent to your inbox.

Two decades after the draft sequence of the human genome was unveiled to great fanfare, a team of 99 scientists has finally deciphered the entire thing. They have filled in vast gaps and corrected a long list of errors in previous versions, giving us a new view of our DNA.

https://www.nytimes.com/2021/07/23/science/human-genome-complete.html

Human Genome Sequencing completed in 2021!

The complete sequence of a human genome

Sergey Nurk[1,*], Sergey Koren[1,*], Arang Rhie[1,*], Mikko Rautiainen[1,*], Andrey V. Bzikadze[2], Alla Mikheenko[3], Mitchell R. Vollger[4], Nicolas Altemose[5], Lev Uralsky[6,7], Ariel Gershman[8], Sergey Aganezov[9], Savannah J. Hoyt[10], Mark Diekhans[11], Glennis A. Logsdon[4], Michael Alonge[9], Stylianos E. Antonarakis[12], Matthew Borchers[13], Gerard G. Bouffard[14], Shelise Y. Brooks[14], Gina V. Caldas[15], Haoyu Cheng[16,17], Chen-Shan Chin[18], William Chow[19], Leonardo G. de Lima[13], Philip C. Dishuck[4], Richard Durbin[21], Tatiana Dvorkina[3], Ian T. Fiddes[22], Giulio Formenti[23,24], Robert S. Fulton[25], Arkarachai Fungtammasan[18], Erik Garrison[11,26], Patrick G.S. Grady[10], Tina A. Graves-Lindsay[27], Ira M. Hall[28], Nancy F. Hansen[29], Gabrielle A. Hartley[10], Marina Haukness[11], Kerstin Howe[19], Michael W. Hunkapiller[30], Chirag Jain[1,31], Miten Jain[11], Erich D. Jarvis[23,24], Peter Kerpedjiev[32], Melanie Kirsche[9], Mikhail Kolmogorov[33], Jonas Korlach[30], Milinn Kremitzki[27], Heng Li[16,17], Valerie V. Maduro[34], Tobias Marschall[35], Ann M. McCartney[1], Jennifer McDaniel[36], Danny E. Miller[4,37], James C. Mullikin[14,29], Eugene W. Myers[38], Nathan D. Olson[36], Benedict Paten[11], Paul Peluso[30], Pavel A. Pevzner[33], David Porubsky[4], Tamara Potapova[13], Evgeny I. Rogaev[6,7,39,40], Jeffrey A. Rosenfeld[41], Steven L. Salzberg[9,42], Valerie A. Schneider[43], Fritz J. Sedlazeck[44], Kishwar Shafin[11], Colin J. Shew[20], Alaina Shumate[42], Yumi Sims[19], Arian F. A. Smit[45], Daniela C. Soto[20], Ivan Sović[30,46], Jessica M. Storer[45], Aaron Streets[5,47], Beth A. Sullivan[48], Françoise Thibaud-Nissen[43], James Torrance[19], Justin Wagner[36], Brian P. Walenz[1], Aaron Wenger[30], Jonathan M. D. Wood[19], Chunlin Xiao[43], Stephanie M. Yan[49], Alice C. Young[14], Samantha Zarate[9], Urvashi Surti[50], Rajiv C. McCoy[49], Megan Y. Dennis[20], Ivan A. Alexandrov[3,7,51], Jennifer L. Gerton[13], Rachel J. O'Neill[10], Winston Timp[8,42], Justin M. Zook[36], Michael C. Schatz[9,49], Evan E. Eichler[4,24,†], Karen H. Miga[11,†], Adam M. Phillippy[1,†]

Completed Human Genome: Improvements from current reference

Table 1. Comparison of GRCh38 and T2T-CHM13 human genome assemblies.

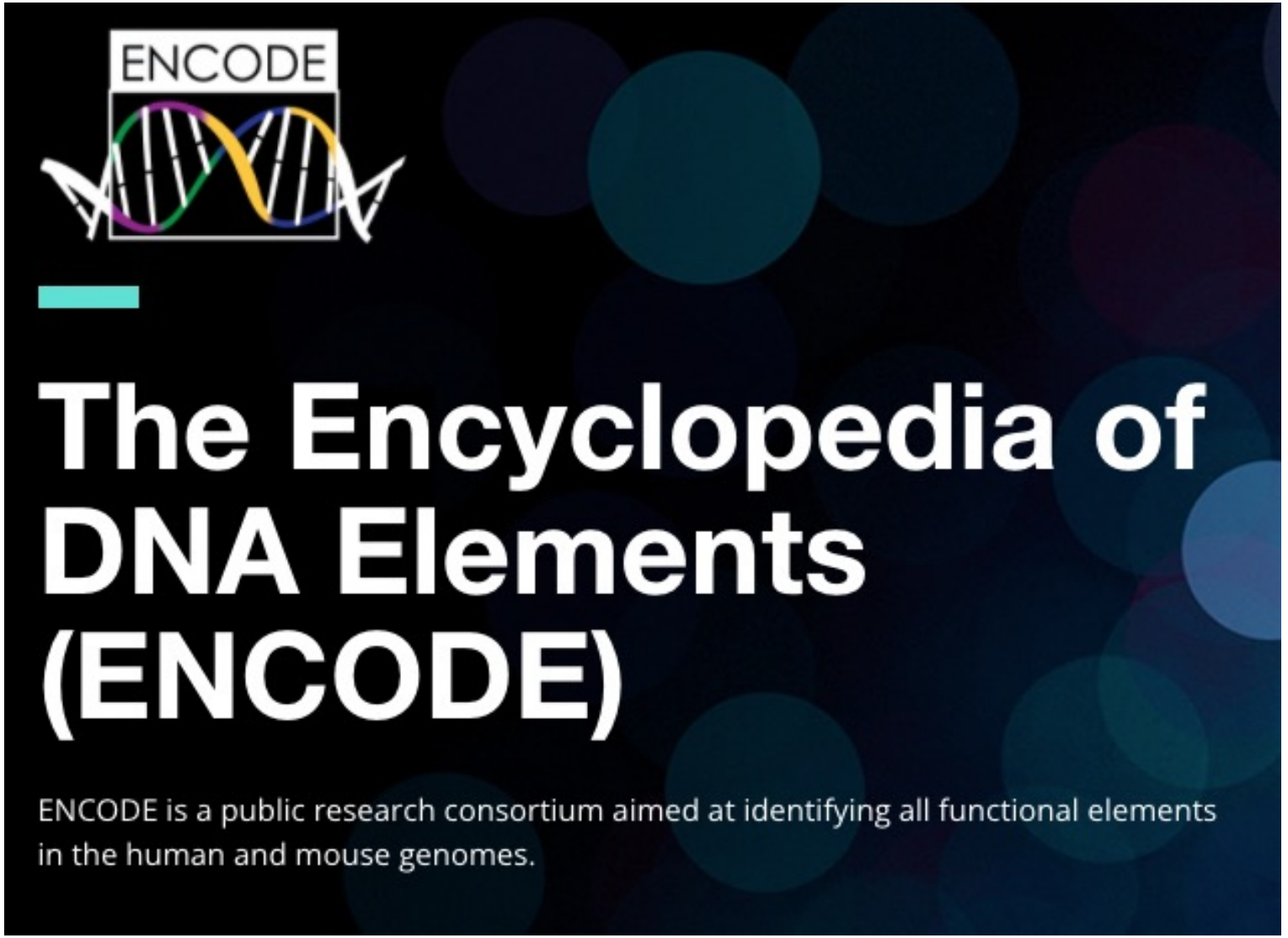
Summary	GRCh38p13	CHM13v1.1	±%
Assembled bases (Gbp)	2.92	3.05	+4.5%
Unplaced bases (Mbp)	11.42	0	-100.0%
Gap bases (Mbp)	120.31	0	-100.0%
# Contigs	949	24	-97.5%
Ctg NG50 (Mbp)	56.41	154.26	+173.5%
# Issues	230	46	-80.0%
Issues (Mbp)	230.43	8.18	-96.5%
Gene Annotation			
# Genes	60,090	63,494	+5.7%
protein coding	19,890	19,969	+0.4%
# Exclusive genes	263	3,604	
protein coding	63	140	
# Transcripts	228,597	233,615	+2.2%
protein coding	84,277	86,245	+2.3%
# Exclusive transcripts	1,708	6,693	
protein coding	829	2,780	
Segmental duplications (SDs)			
% SDs	5.00%	6.61%	
SD bases (Mbp)	151.71	201.93	+33.1%
# SDs	24097	41528	+72.3%
RepeatMasker			
% Repeats	50.03%	53.94%	
Repeat bases (Mbp)	1,516.37	1,647.81	+8.7%
LINE	626.33	631.64	+0.8%
SINE	386.48	390.27	+1.0%
LTR	267.52	269.91	+0.9%
Satellite	76.51	150.42	+96.6%
DNA	108.53	109.35	+0.8%
Simple repeat	36.5	77.69	+112.9%
Low complexity	6.16	6.44	+4.6%
Retroposon	4.51	4.65	+3.3%
rRNA	0.21	1.71	+730.4%

Genome Sequence

TACTCAAGCAATGGCCCCATTTCCCTTGGGGAATCCATCTCTCTCGCAGGCTTAGTCCCAGAGCTTCAGG
TGGGGCTGCCCACAGAGCTCCTCAGTCTAAGCCAAGTGGTGTGTCATAGTCCCCTGGCCCCATTAATGGA
TTCTGGGATAGACATGAGGACCAAGCCAGGTGGGATGAGTGAGTGTGGCTTCTGGAGGAAGTGGGGACAC
AGGACAGCATTCTTTCCTGCTGGACCTGACCCTGTGTCATGTCACCTTGCTACCACGAGAGCATGGCCTG
TCTGGGAATGCAGCCAGACCCAAAGAAGCAAACTGACATGGAAGGAAAGCAAAACCAGGCCCTGAGGACA
TCATTTTAGCCCTTACTCCGAAGGCTGCTCTACTGATTGGTTAATTTTTGCTTAGCTTGGTCTGGGGAGT
TCTGACAGGCGTGCCACCAATTCTTACCGATTTCTCTCCACTCTAGACCCTGAGAAGCCCACGCGGTTCA
TGCTAGCAATTAACAATCAATCTCGCCCTATGTGTTCCCATTCCAGCCTCTAGGACACAGTGGCAGCCAC
ATAATTGGTATCTCTTAAGGTCCAGCACGAGGTGGAGCACATGGTGGAGAGACAGATGCAGTGACCTGGA
ACCCAGGAGTGAGGGAGCCAGGACTCAGGCCCAAGGCTCCTGAGAGGCATCTGGCCCTCCCTGCGCTGTG
CCAGCAGCTTGGAGAACCCACACTCAATGAACGCAGCACTCCACTACCCAGGAAATGCCTTCCTGCCCTC
TCCTCATCCCATCCCTGGGCAGGGGACATGCAACTGTCTACAAGGTGCCAAGTACCAGGACAGGAAAGGA
AAGACGCCAAAAATCCAGCGCTGCCCTCAGAGAAGGGCAACCACGCAGTCCCCATCTTGGCAAGGAAACA
CAATTTCCGAGGGAATGGTTTTGGCCTCCATTCTAAGTGCTGGACATGGGGTGGCCATAATCTGGAGCTG
ATGGCTCTTAAAGACCTGCATCCTCTTCCCTAGGTGTCCCTCGGGCACATTTAGCACAAAGATAAGCACA
AAAGGTGCATCCAGCACTTTGTTACTATTGGTGGCAGGTTTATGAATGGCAACCAAAGGCAGTGTACGGG
TCAAGATTATCAACAGGGAAGAGATAGCATTTCCTGAAGGCTTCCTAGGTGCCAGGCACTGTTCCATTCC
TTTGCATGTTTTGATTAATTTAATATTTAAAATAATTCTACCAGGAAGCTACCATTATTACCACAACTTC
ACAAATGAGAACACCGAGGCTTAGAGGGGTTGGGTTGCCCAAGGTTACAGAGGAAGAAAACAGGGGAGCT
GGATCTGAGCCAAGGCATCAACTCCAAGGTAACCCCTCAGTCACTTCACTGTGTGTCCCCTGGTTACTGG
GACATTCTTGACAAACTCGGGGCAAGCCGGTGAGTCAGTGGGGGAGGACTTTCAGGAAGAGGTGGGTTCC
CAGTTGGTGACAGAAGAGGAGGCTGCAAAGTGAAGGAGCAGGGGCTCCAGGTCTGGCGACAACCAGGGAA
GGGACAGGGCAGGGATGGCTTGGACCACGAGAGGCACCTGAGTCAGGCAGTCACATACTTCCCACTGGGG
TCTACCATGTGAGGCATGGTGTGGGATCCTGGGAAGGAGACCAAGCCTCATTTCAGTTTGCTTATGGCCA
AAGACAGGACCTGTGTACCCGACAACCCCTGGGACCTTTACCAAAAAAAGAGCAAACACCATTCACTCAC
TCATGTTAGATAAACACTGAGTGAAGTCACTGGAGCCCAAGGACTGTGCGAGGTCAGCACTGCCAATACA
AGAAGCTGCAGCCCTCCAGCTCGCCTCCCTCAATGGCCACTCCGTGCTCCAGCCATGCTGGCTTCCTTTT
AGGTCCTCCACCTCCAGGCTGTAGTTCATGTGCTTCTTTCTGGAATGTTCTTCCCAACCTACCCACTCAA
CCCTCAGACTTTACCATAAATGTCATTTCCTCACGTCTGCCTTCCCTGACCTGAGACCAAGCCAGGCTTC
CCATGACGAGCCTCACAGTACCCCATCTCCCCTGAACAGATGCAGTAATAACCTACATAACCCGGGGCCA
TGATCTATGGCTTTGAATCCTGGCTCTGTCACTAGGCCAGGTCTCTCAGCCCTTCTGTGCCTCAGTTTCC
TCATCTATAAAATGAGATGACGGCAGTGCCTGCTCATGAAGTGTGAGTTAATGCACTCAAATCAATGGTT
GTGCACGGTTTATATGAATATTAGTGATTACAAAATATTATCAATAGACCTTGTCACAACTGTTATTGAA
GAACTAATCATCTATTGCTTATTTAGGTCTTTCTCTCCTGCCAGAATGTGCGCTCCAGGTGGAGAGGTAT
GTTGCCTTATCCGTGGCTGGATATATAGAGATTCCCACACTGCCTTGCACACGAGCACTGCTGGGTAAAT
ATTTGTTGGCTGCAGGAAAACGTGAAGGAATAGGCCCTCCAATGGGAGGAAAAGCATGAGTTGTGAGAGC

What does it all mean?

2003: ENCODE



https://www.genome.gov/Funded-Programs-Projects/ENCODE-Project-ENCyclopedia-Of-DNA-Elements

Encyclopedia of DNA Elements: ENCODE

Vol 447|14 June 2007|doi:10.1038/nature05874 nature

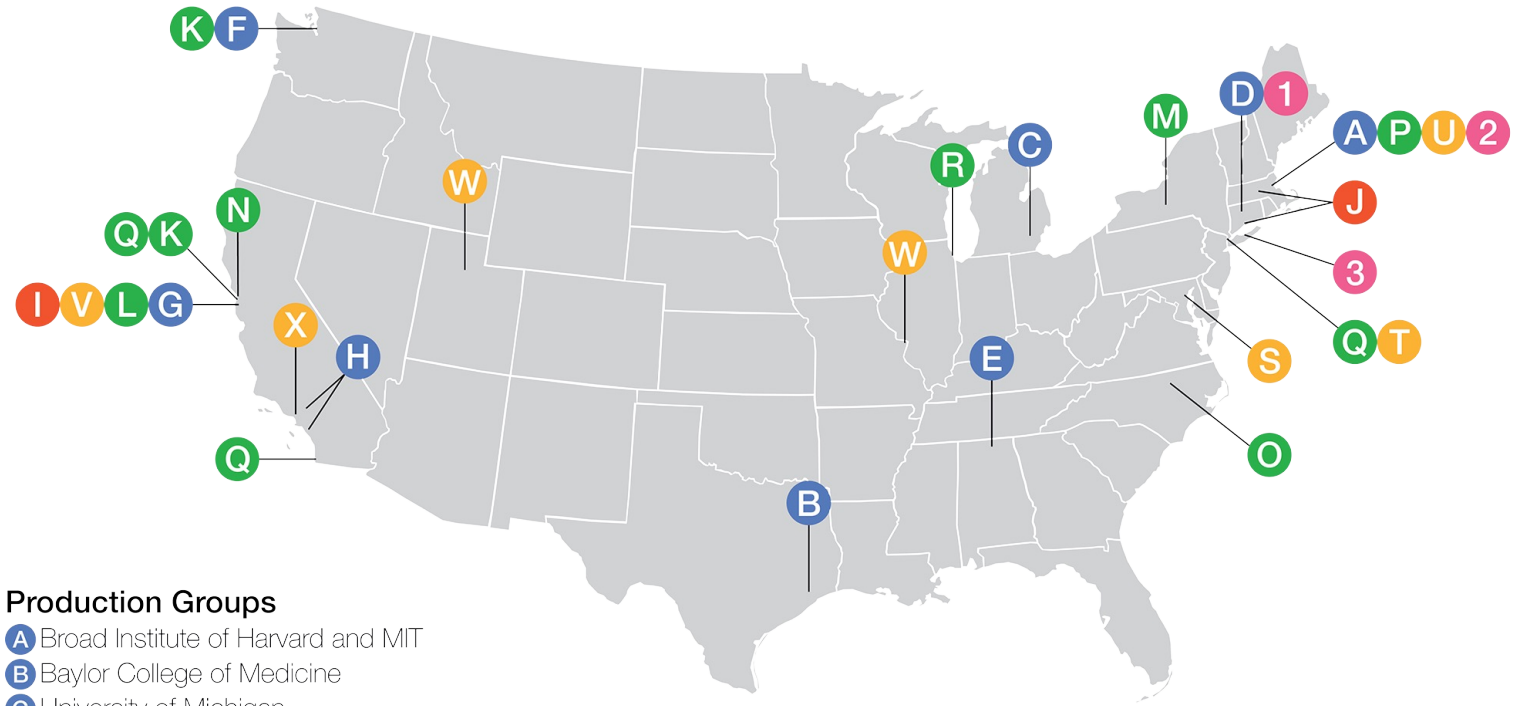
ARTICLES

Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project

The ENCODE Project Consortium*

We report the generation and analysis of functional data from multiple, diverse experiments performed on a targeted 1% of the human genome as part of the pilot phase of the ENCODE Project. These data have been further integrated and augmented by a number of evolutionary and computational analyses. Together, our results advance the collective knowledge about human genome function in several major areas. First, our studies provide convincing evidence that the genome is pervasively transcribed, such that the majority of its bases can be found in primary transcripts, including non-protein-coding transcripts, and those that extensively overlap one another. Second, systematic examination of transcriptional regulation has yielded new understanding about transcription start sites, including their relationship to specific regulatory sequences and features of chromatin accessibility and histone modification. Third, a more sophisticated view of chromatin structure has emerged, including its inter-relationship with DNA replication and transcriptional regulation. Finally, integration of these new sources of information, in particular with respect to mammalian evolution based on inter- and intra-species sequence comparisons, has yielded new mechanistic and evolutionary insights concerning the functional landscape of the human genome. Together, these studies are defining a path for pursuit of a more comprehensive characterization of human genome function.

ENCODE Centers



Production Groups

- A Broad Institute of Harvard and MIT
- B Baylor College of Medicine
- C University of Michigan
- D The Jackson Laboratory
- E HudsonAlpha Institute for Biotechnology, University of Alabama in Huntsville
- F Altius Institute for Biomedical Sciences
- G Stanford University
- H California Institute of Technology, University of California, Irvine

Data Coordination Center

- I Stanford University

Data Analysis Center

- J University of Massachusetts Medical School; Yale University

Characterization Centers

- K University of California, San Francisco; University of Washington
- L Stanford University
- M Cornell University
- N Lawrence Berkeley National Laboratory
- O Duke University
- P Broad Institute of Harvard and MIT
- Q University of California, San Francisco; University of California, San Diego; Ludwig Institute for Cancer Research
- R University of Chicago

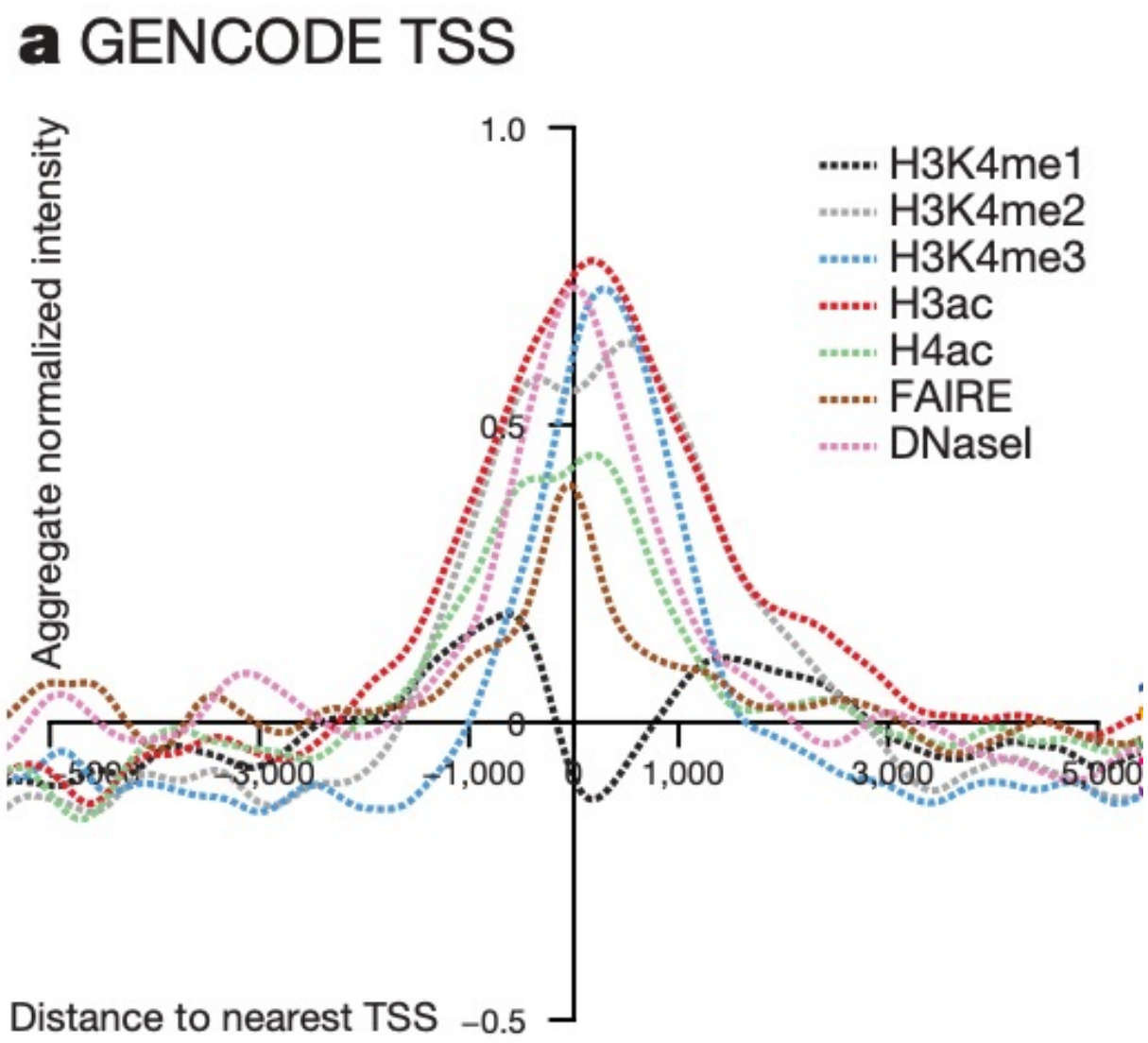
Computational Analysis Groups

- S Johns Hopkins University
- T Memorial Sloan Kettering Cancer Center
- U Harvard University; Brigham and Women's Hospital
- V Stanford University
- W Washington University; University of Utah
- X University of California, Los Angeles

Affiliated Groups

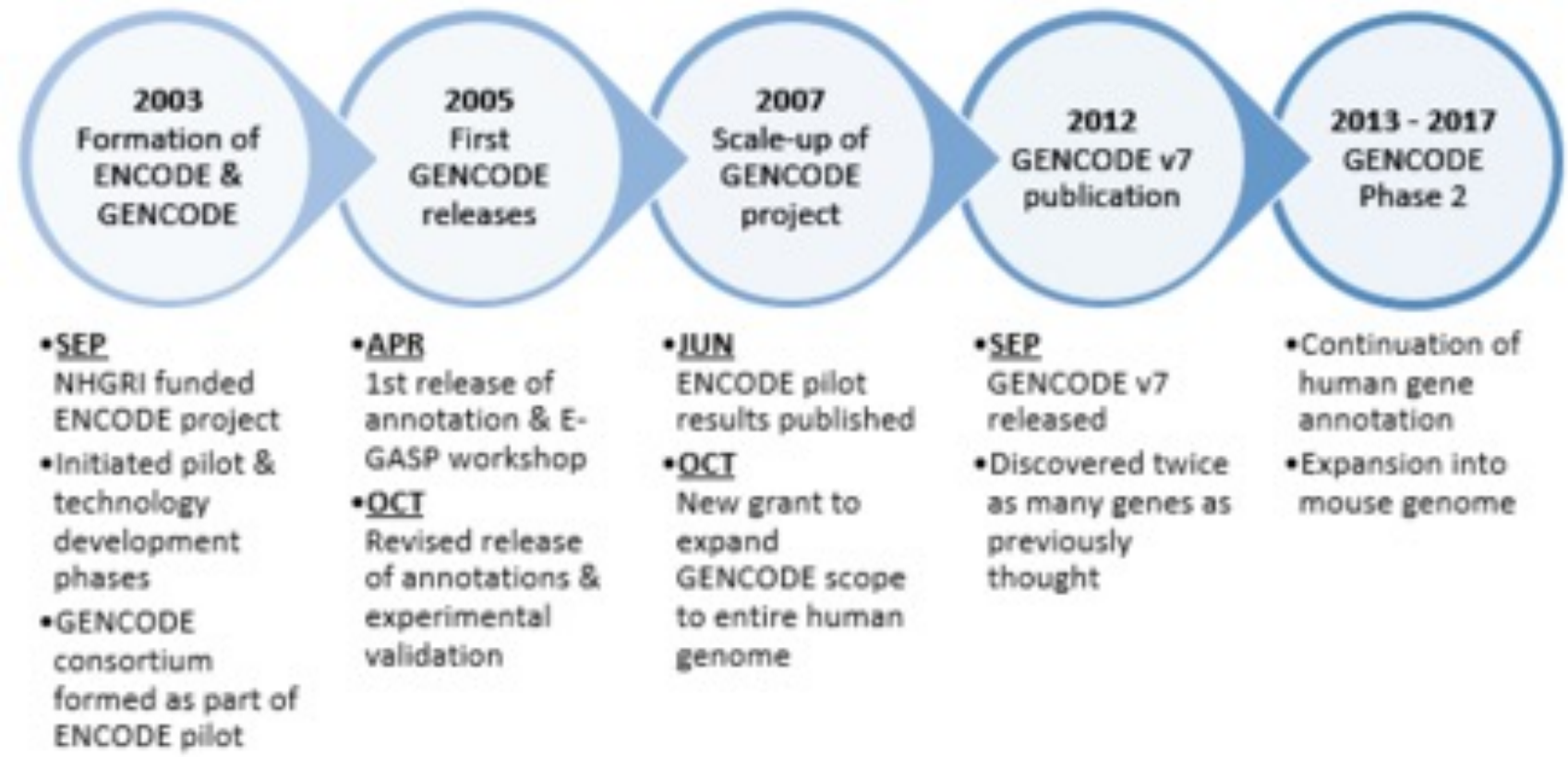
- 1 University of Connecticut Health Center
- 2 Dana-Farber Cancer Institute
- 3 Cold Spring Harbor Laboratory

ENCODE Example

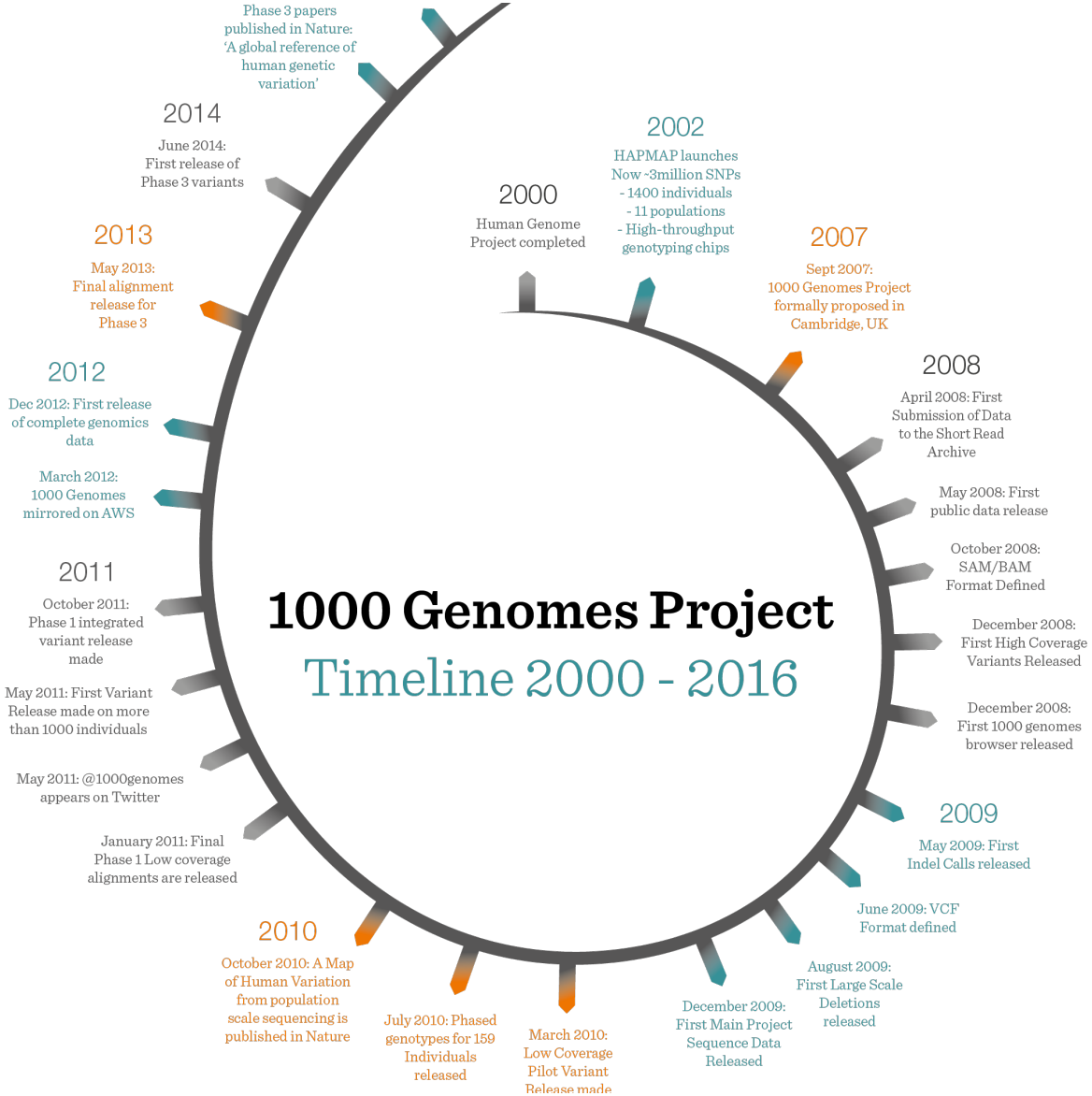


GENCODE

The goal of the GENCODE project is to identify and classify all gene features in the human and mouse genomes with high accuracy based on biological evidence, and to release these annotations for the benefit of biomedical research and genome interpretation.



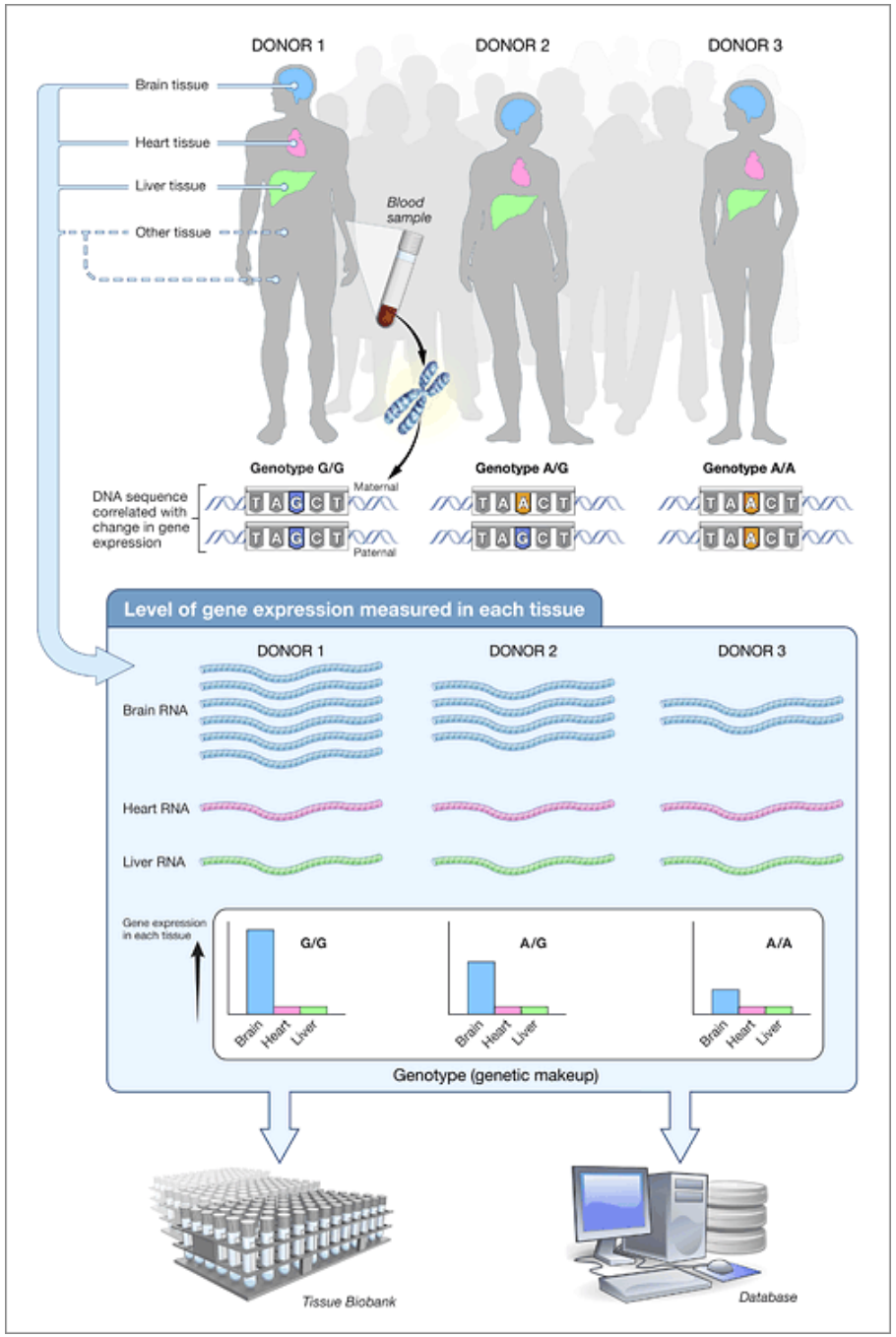
The 1000 Genomes Project



https://www.internationalgenome.org/

GTEx

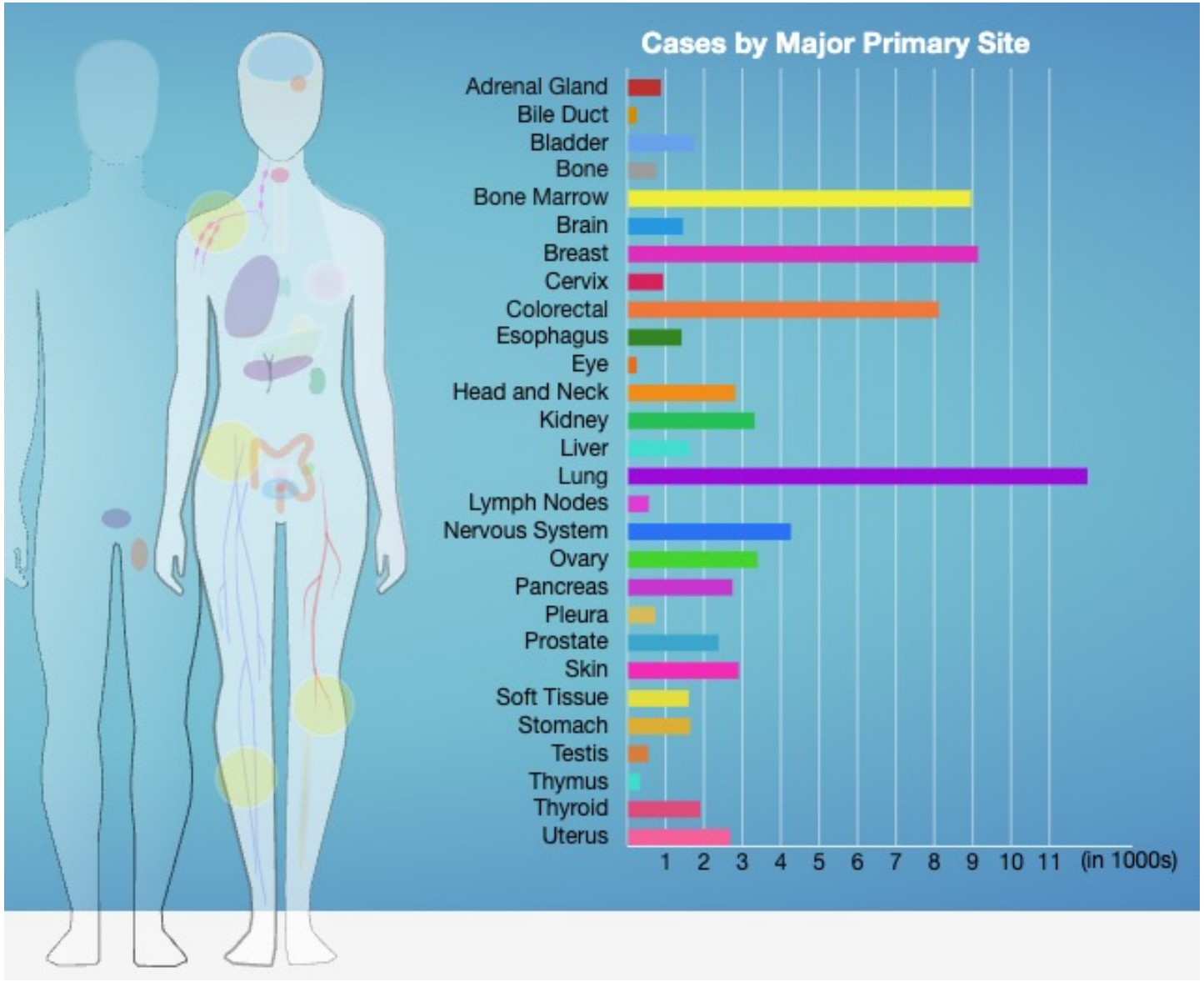
The Genotype-Tissue Expression (GTEx) project aims to provide to the scientific community a resource with which to study human gene expression and regulation and its relationship to genetic variation.



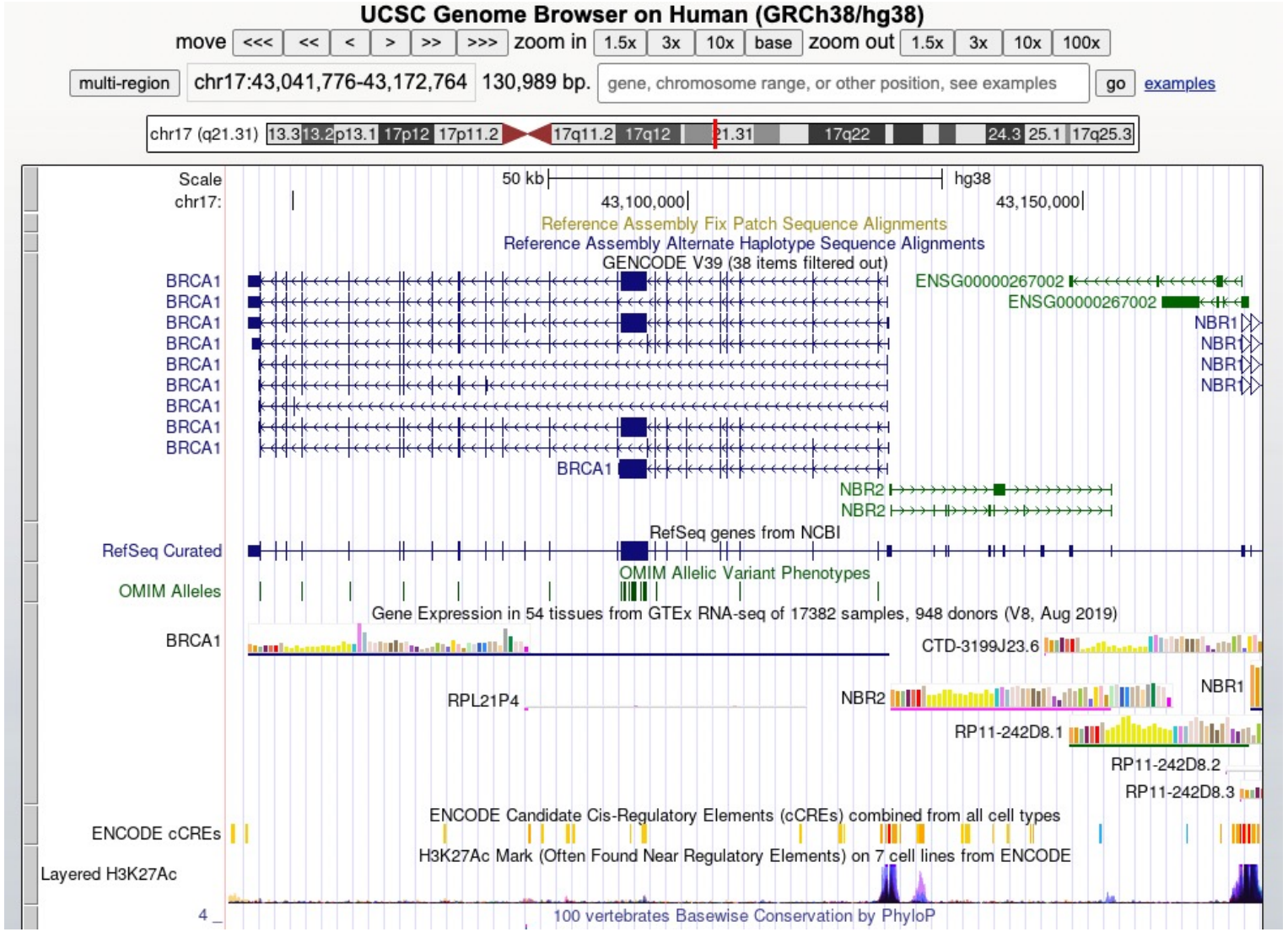
The Cancer Genome Atlas

- The Cancer Genome Atlas (TCGA), a landmark cancer genomics program, molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. This joint effort between NCI and the National Human Genome Research Institute began in 2006, bringing together researchers from diverse disciplines and multiple institutions.
- Over the next dozen years, TCGA generated over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data. The data, which has already led to improvements in our ability to diagnose, treat, and prevent cancer, will remain publicly available for anyone in the research community to use.

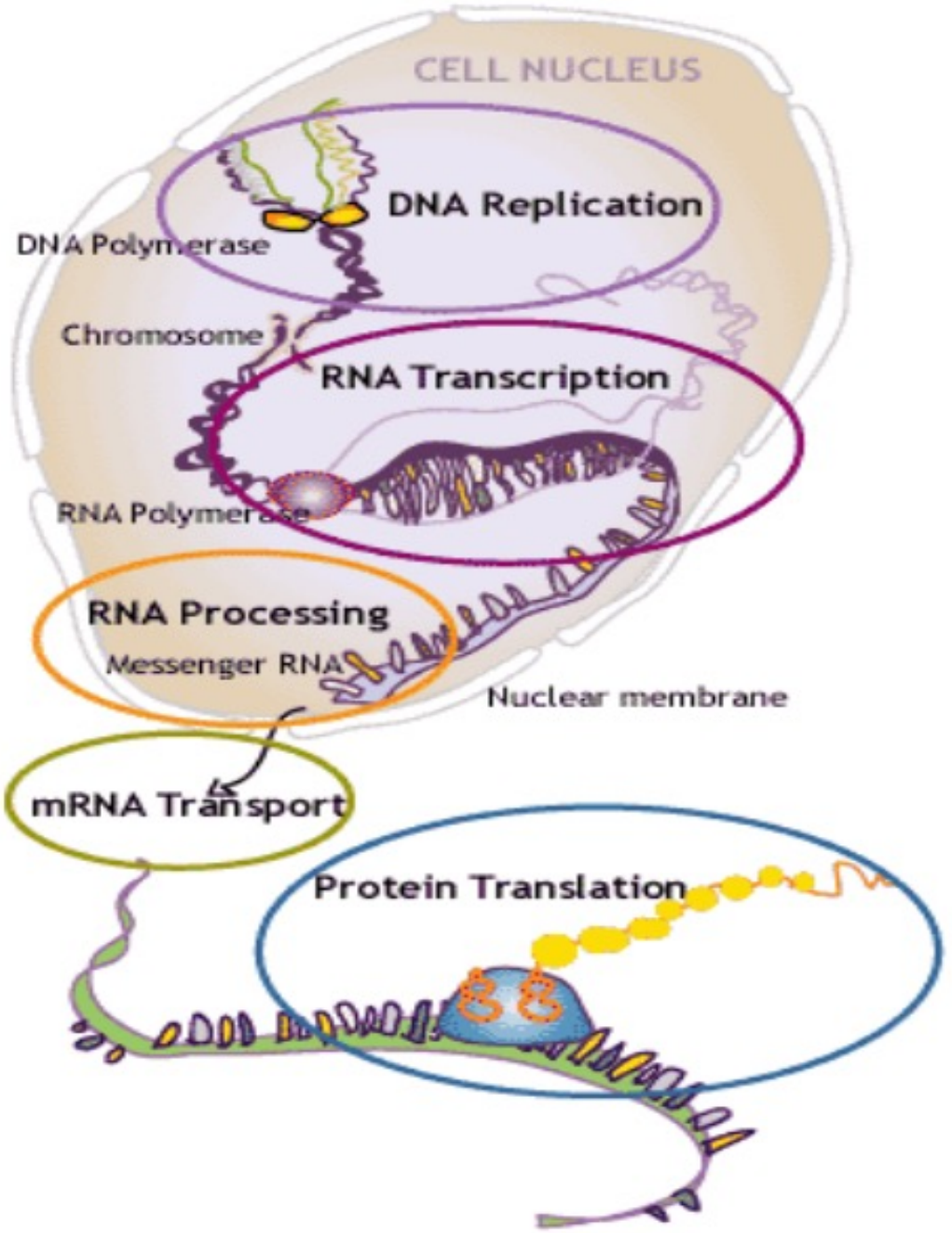
Cancer Genomic Data



Views of the Genome



The Central Dogma



Questions

Sean Davis
seandavi@gmail.com

ELSI Concerns