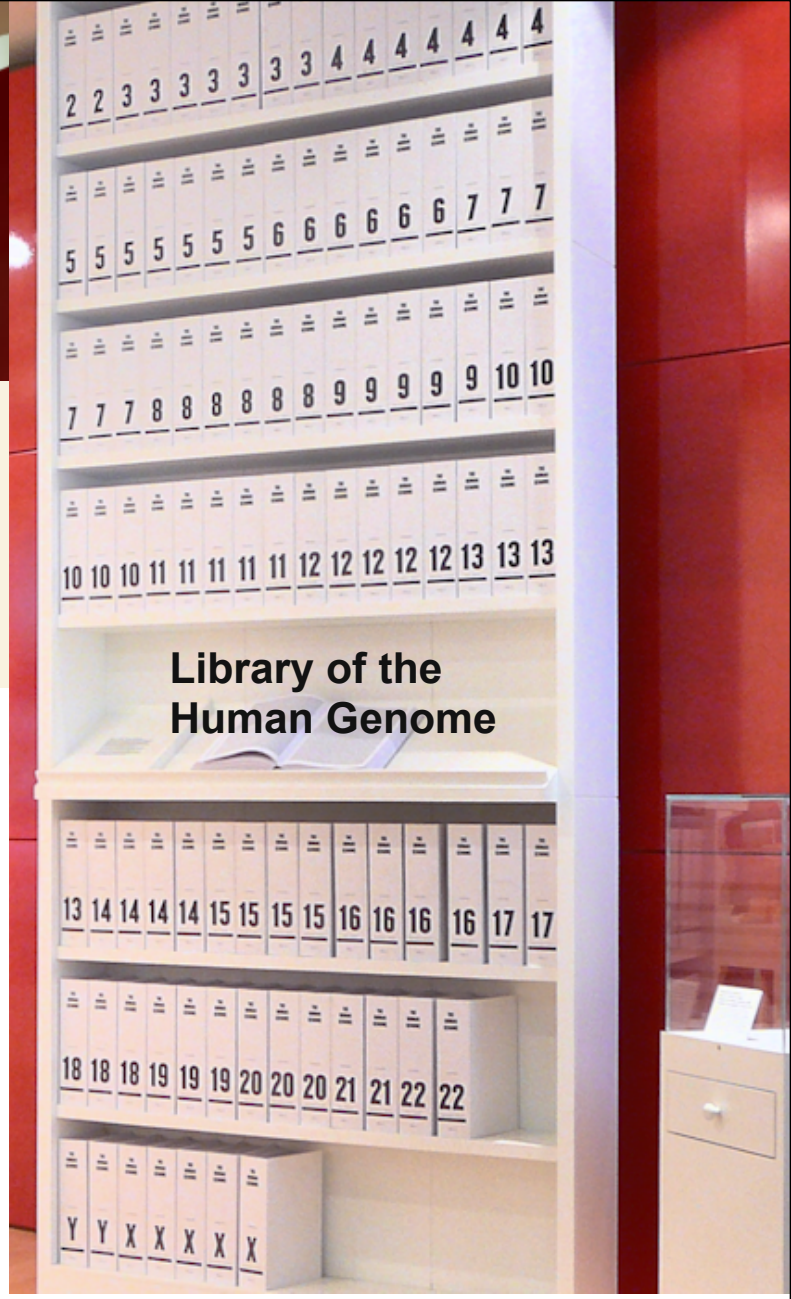
Great feats of exploration in human genome

Computational Biology
Group Presentation

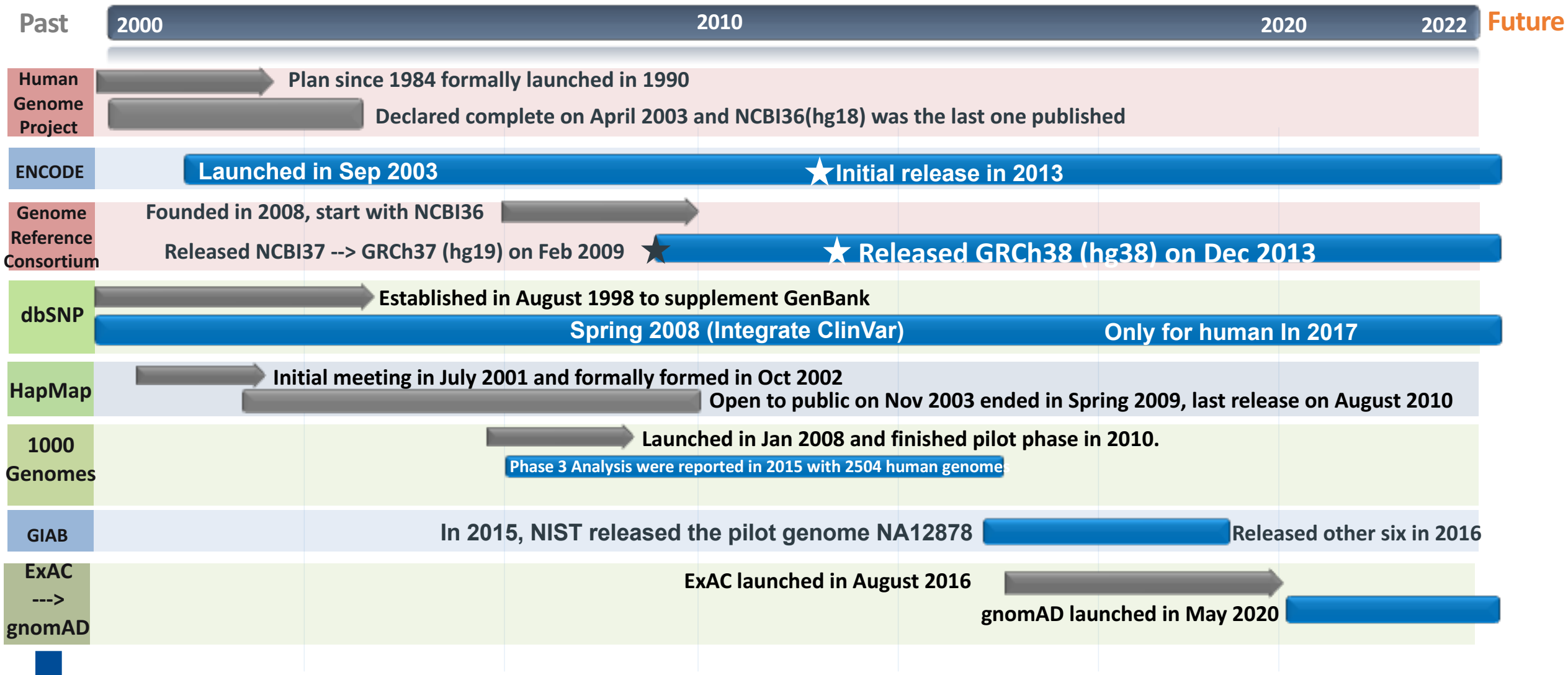
Xiang Li
Jan.13.2022



Ref: https://wellcomecollection.org/articles/WcvK4CsAANQR59Up

SANA CONFIDENTIAL – ©2019 Sana Biotechnology. All rights reserved.

Timeline



National Center for Biotechnology Information (NCBI)

SANA CONFIDENTIAL – ©2019 Sana Biotechnology. All rights reserved.

Sana Biotechnology

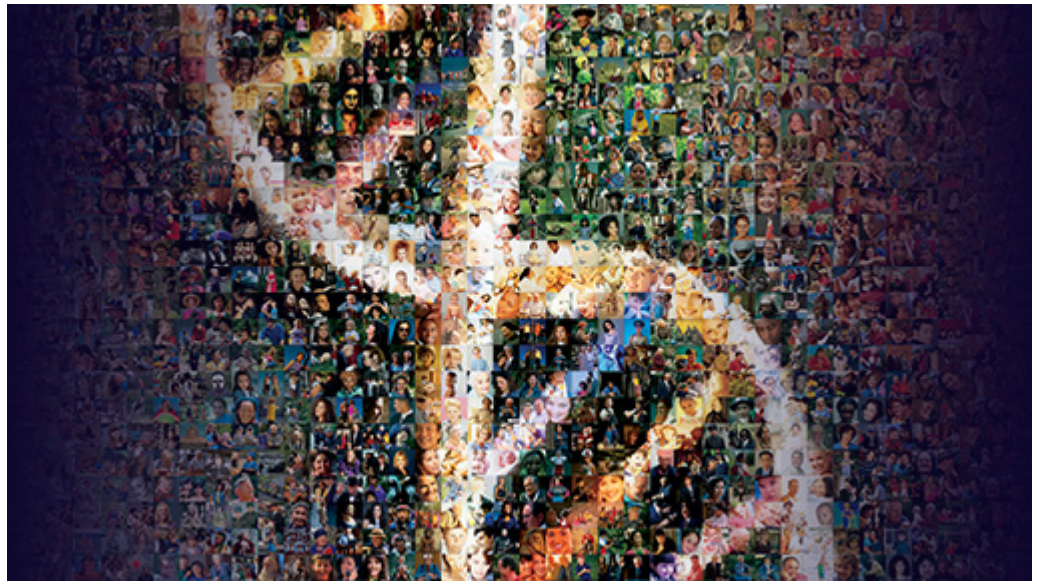
Human Genome Project (HGP)

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

The Santa Cruz Workshop—May 1985

ROBERT L. SINSHEIMER

Department of Biological Sciences, University of California, Santa Barbara, California 93106



Ref: https://www.genome.gov/human-genome-project

- 1988 NIH, lay out a plan for the human genome project.
- 1989 National Center for Human Genome Research (NCHGR) with James D. Watson as the first director. (resigned in 1992)
- 1993 Francis S. Collins appointed as NCHGR director
- 1994 Meet first goal, a comprehensive human genetic linkage map
- 1995 Complete a physical map that contains actual, physical locations of identifiable landmarks on chromosomes.
- 1997 NCHGR became NHGRI (national human genome research institute)
- 1999 Completes the pilot phase of sequencing the human genome and launch full-scale effort to sequence all 3 billion letters. In December, first reveal 33.5M letters for chr22.

"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 64. Weather map, page 29.

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 33 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 "become embedded in routine police practice" without causing any measurable difficulty for prosecutors, there was no justification for doing

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 strategy-delayed refighting of a previous generation's battle over the rights of criminal suspect. 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 deliver the decision in the case, Dickerson v. United States, No. 99-5525. The announcement meant that he was the majority opinion's author. Given

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 told 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 26.]

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.

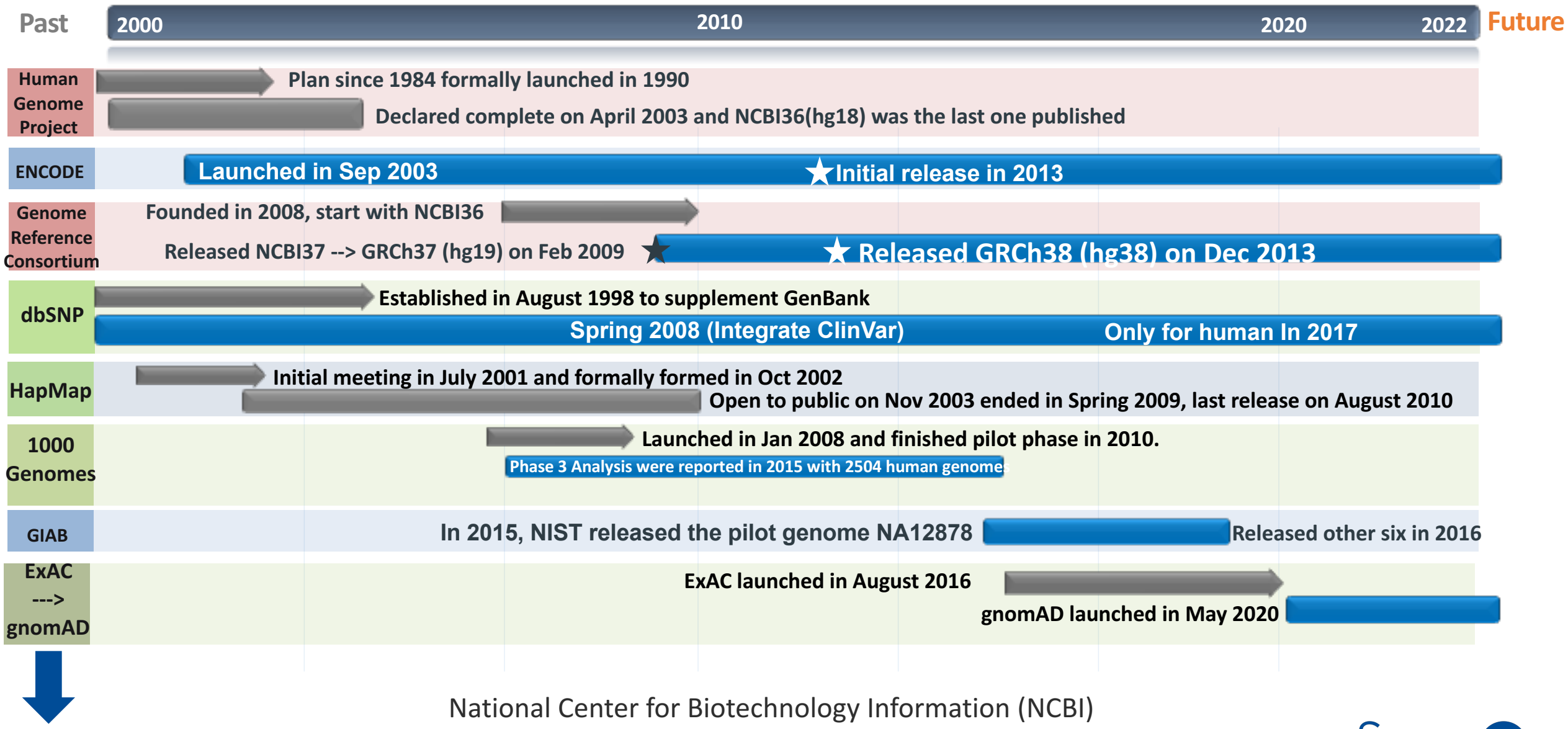
The human genome, the ancient script that has now been deciphered, consists of two sets of 23 giant DNA molecules, or chromosomes, with each set — one inherited from each parent — containing more than three billion chemical units.

A Pearl and a Hodgepodge: Human DNA

Biotechnology

SANA CONFIDENTIAL – ©2019 Sana Biotechnology. All rights reserved.

Timeline



National Center for Biotechnology Information (NCBI)

SANA CONFIDENTIAL – ©2019 Sana Biotechnology. All rights reserved.

Sana Biotechnology

Genome Reference Consortium (GRC)

GRC (Blog http://genomeref.blogspot.com/)

- Started in 2008 with the last assembly NCBI36(hg18) from Human genome project described https://pubmed.ncbi.nlm.nih.gov/15496913/

- GRCh37(hg19) In Feb 2009, GRC released a more accurate of the human genome,
  - left more than 300 gaps

- GRCh38(hg38) In Dec 2013, the first major assembly update since 2009, (http://genomeref.blogspot.com/2013/12/)
  - while160 such gaps remained
  - Latest minor released GRCh38.p13 on March 2019.

Following update:

- May 2020, the GRC reported 79 "unresolved" gaps, accounting for as much as 5% of the human genome, months later the application of new long-range sequencing techniques led to the first telomere-to-telomere assembly.

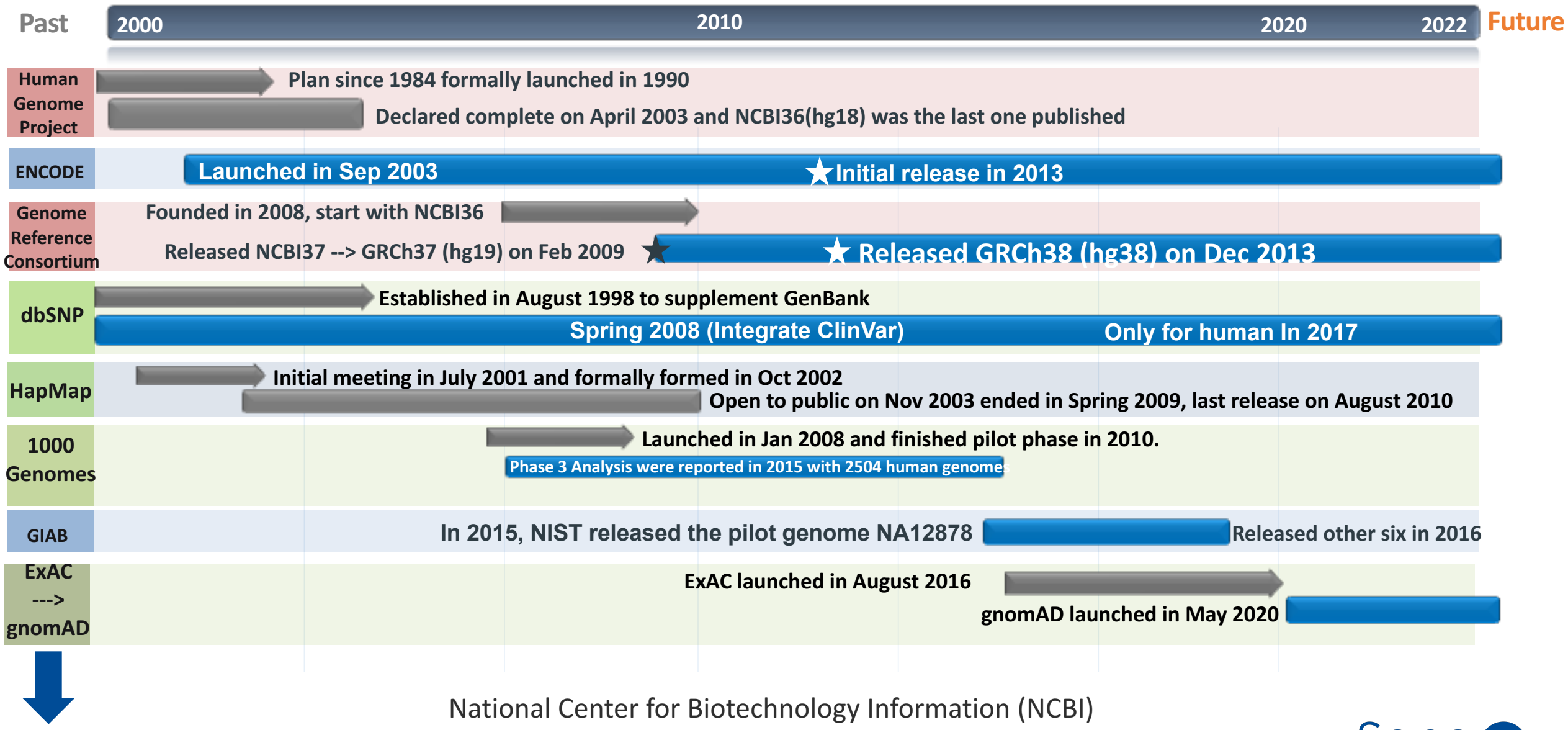
- In 2021 it was reported that the Telomere-to-Telomere (T2T) consortium had filled in all of the gaps. Thus there came into existence a complete human genome with almost no gaps, but it still had five gaps in ribosomal DNA.
  - In May, those last five rDNA gaps were found and published as version 1.1. This sequence does not contain chrY.

GRC Genome Reference Consortium

SANA CONFIDENTIAL – ©2019 Sana Biotechnology. All rights reserved.

Sana Biotechnology

Timeline



National Center for Biotechnology Information (NCBI)

SANA CONFIDENTIAL – ©2019 Sana Biotechnology. All rights reserved.

Sana Biotechnology

HapMap (Haplotype Map)

- PHASE I
  - One common SNP was genotyped every 5,000
  - 1 million SNPs typed
- PHASE II
  - 3.5 million SNPs typed
- PHASE III
  - 11 global ancestry groups have been assembled
  - Genotyped 1.6 million common single nucleotide

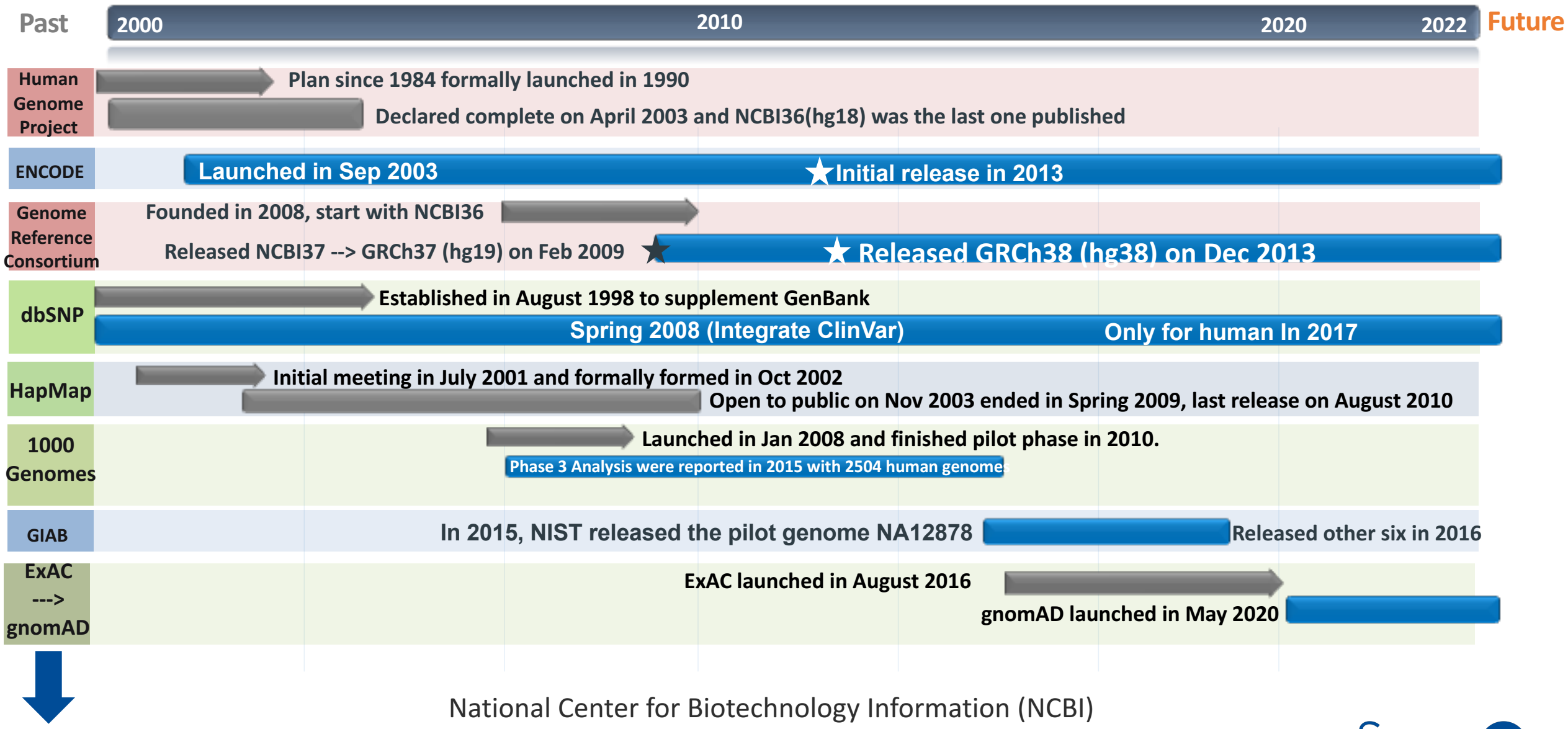
- Research:
  - Modern humans arose in Africa and replaced other human species across the globe.
  - NA12878 picked up by NIST as HG001

- Legacy Hapmap site was finally taken down by NCBI on June 2016 for a security flaws.
- Since 1000 genomes (1KG) has established itself as a research standard for population genetics and genomics.

Ref: https://en.wikipedia.org/wiki/International_HapMap_Project
https://www.genome.gov/10001688/international-hapmap-project

SANA CONFIDENTIAL – ©2019 Sana Biotechnology. All rights reserved.

Timeline



National Center for Biotechnology Information (NCBI)

SANA CONFIDENTIAL – ©2019 Sana Biotechnology. All rights reserved.

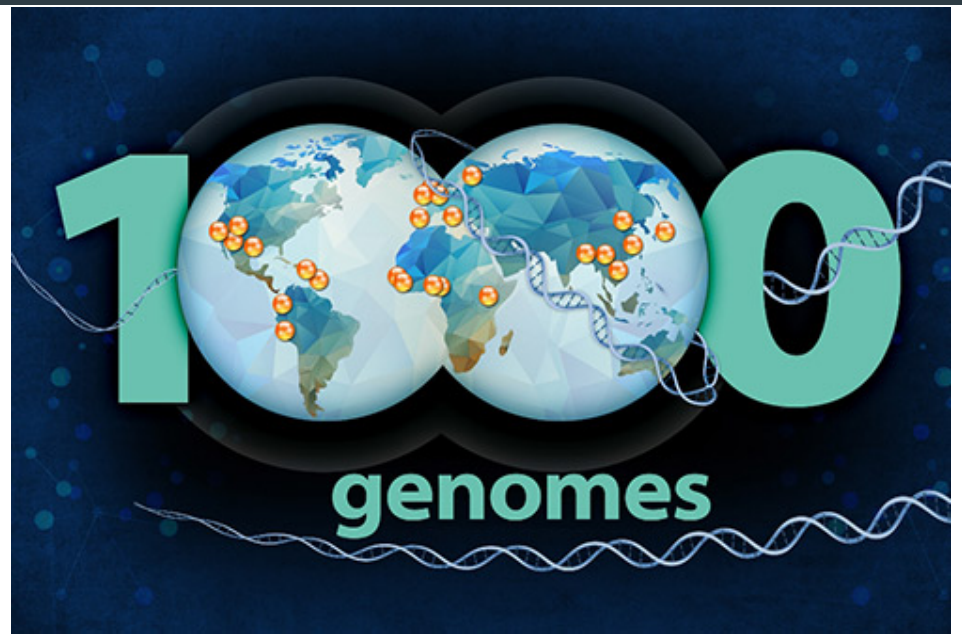
Sana Biotechnology

1000 Genomes Project

- Goal:
  - Find common genetic variants with frequencies of at least 1% in the populations studied.

- Design of the Project
  - The Project was planned in September 2007. The project was conducted in four stages: A pilot phase, phases one and three produced data and phase two concentrating on technical development.

- Publications
  - Pilot Analysis
    - A map of human genome variation from population-scale sequencing *Nature (28 October 2010)*
  - Phase 1 Analysis
    - An integrated map of genetic variation from 1,092 human genomes *Nature (01 November 2012)*
  - Phase 3 Analysis
    - A global reference for human genetic variation *Nature (01 October 2015)*
    - An integrated map of structural variation in 2,504 human genomes *Nature (01 October 2015)*



Ref: https://www.internationalgenome.org/1000-genomes-summary

Summary

- Analyse 2,504 genomes from 26 populations.
- ~88 million variants (84.7M SNPs, 3.6M indels, 60K structural variants)

  (includes >99% of SNP variants with a frequency of >1% for a variety of ancestries)

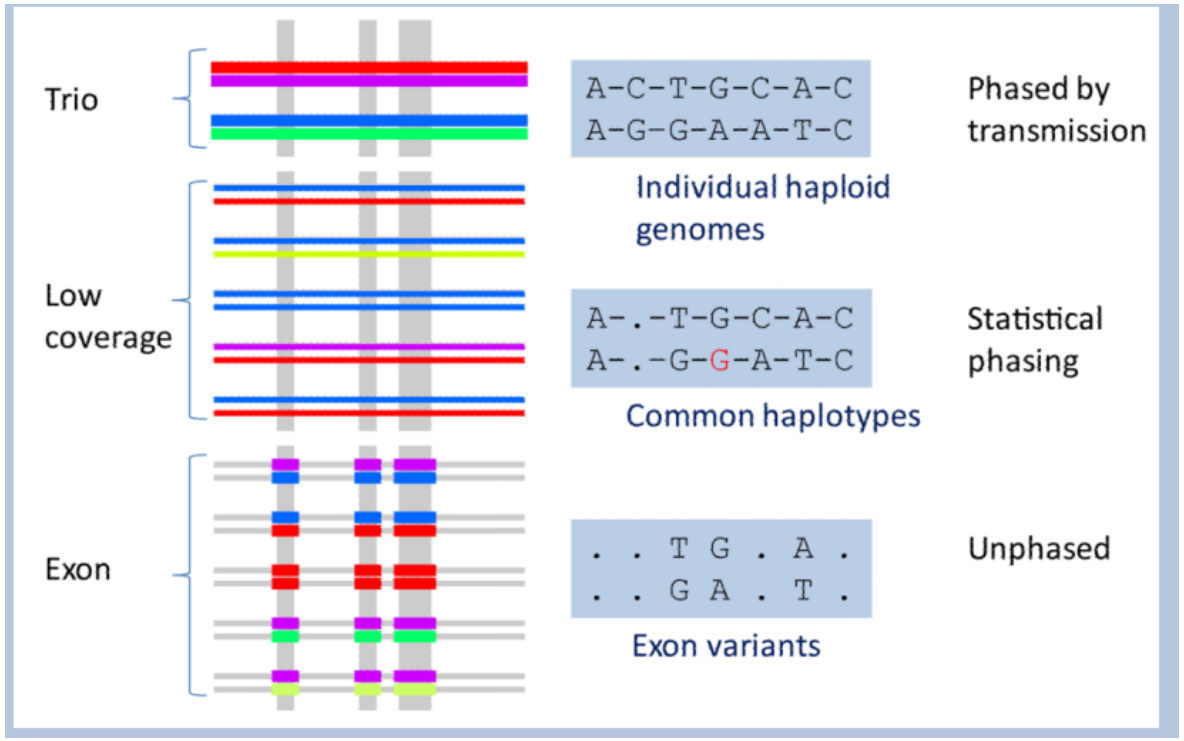
SANA CONFIDENTIAL – ©2019 Sana Biotechnology. All rights reserved.

https://confluence.corp.sana.com/display/~li.xiang/The+1000+Genomes+Project

1000 Genomes Project

Launched in Jan 2008 and finished pilot phase in 2010.

2010, the project finished its pilot phase, which was described in detail in a publication https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3042601/
2012, the sequencing of 1092 genomes was announced https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3498066/
2015, two publications marked the completion of the project and opportunities for future research

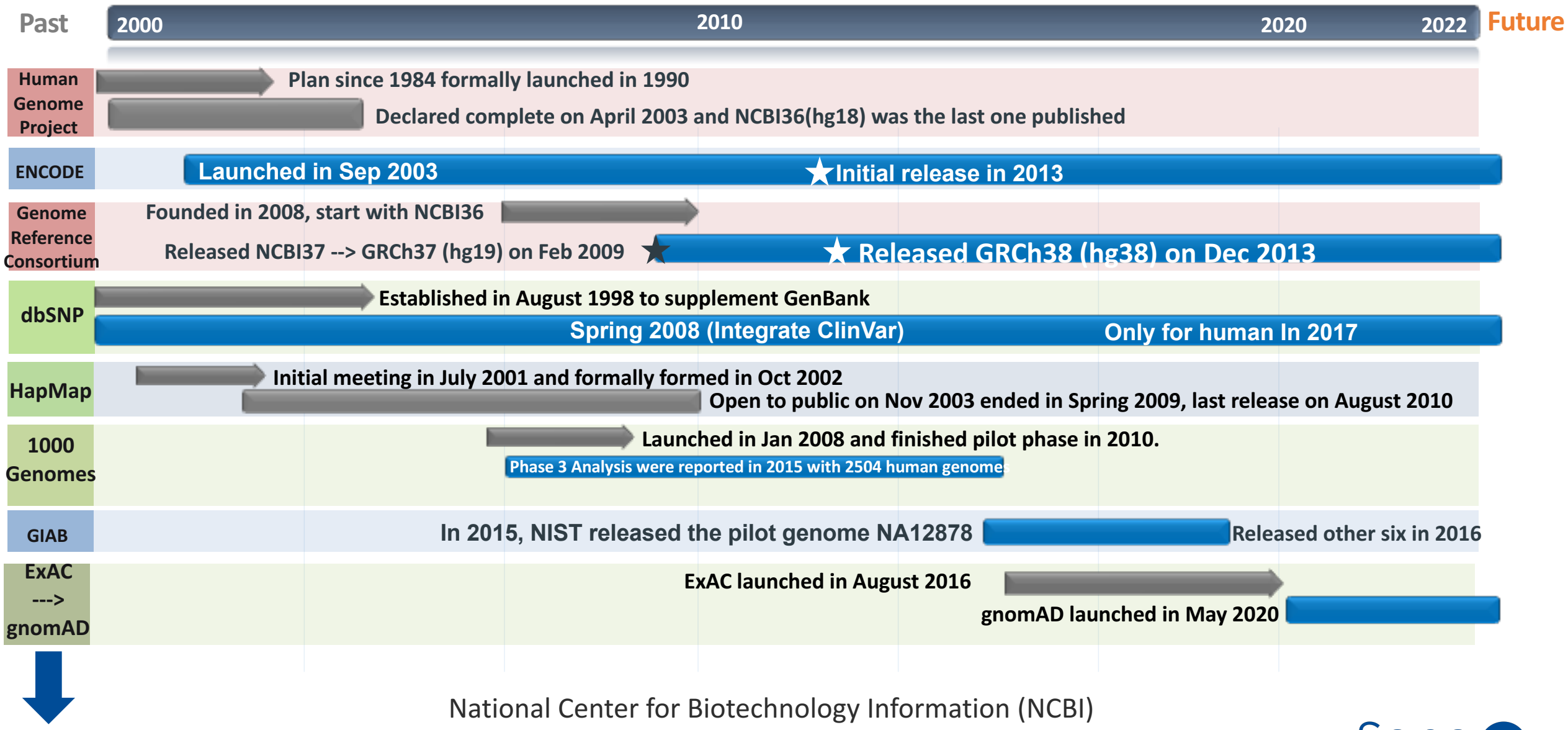


Three projects:

1. Trio project: whole genome shotgun sequencing at high coverage
2. Low coverage: whole genome sequencing of 179 individuals from four populations,
3. High coverage sequencing of two mother-father-child trios, and exon targeted sequencing of 697 individuals from seven populations.
4. Exon Project: targeted capture of the exons from nearly 1000 randomly selected genes(1.4Mb)

SANA CONFIDENTIAL – ©2019 Sana Biotechnology. All rights reserved.

Timeline



National Center for Biotechnology Information (NCBI)

SANA CONFIDENTIAL – ©2019 Sana Biotechnology. All rights reserved.

Sana Biotechnology

Genome in a Bottle (GIAB)

$664 for 10ug (~1.6M cells)

A publicly available study on human genome Hosted by NIST

Goal:

- Develop the technical infrastructure (reference standards, reference methods, and reference data) to enable translation of whole human genome sequencing to clinical practice
- Improve sequencing technologies, SNP, indel, and structural variant calling.

Genome	Coriell cell line ID	NIST ID	NIST RM #	NCBI BioSample	PGP ID
CEPH Mother/Daughter	GM12878	HG001	RM8398	SAMN03492678	Not PGP
AJ Son	GM24385	HG002	RM8391* (son)/RM8392* (trio)	SAMN03283347	huAA53E0
AJ Father	GM24149	HG003	RM8392* (trio)	SAMN03283345	hu6E4515
AJ Mother	GM24143	HG004	RM8392* (trio)	SAMN03283346	hu8E87A9
Chinese Son	GM24631	HG005	RM8393*	SAMN03283350	hu91BD69
Chinese Father	GM24694	N/A†	N/A†	SAMN03283348	huCA017E
Chinese Mother	GM24695	N/A†	N/A†	SAMN03283349	hu38168C

- **Data (12 tech sources)**
  - BioNano Genomics, Complete Genomics paired-end and LFR, Ion Proton exome, Oxford Nanopore, Pacific Biosciences, SOLiD, 10X Genomics GemCode WGS, and Illumina exome and WGS paired-end, mate-pair, and synthetic long reads.
- [Best Practices for Benchmarking Germline Small Variant Calls in Human Genomes]

National Institute of Standards and Technology

Ref: https://www.nature.com/articles/sdata201625

SANA CONFIDENTIAL – ©2019 Sana Biotechnology. All rights reserved.
https://confluence.corp.sana.com/pages/viewpage.action?pageId=43386483

Data & Methods in GIAB

Table 1 Summary of data available for GIAB samples.

From: Extensive sequencing of seven human genomes to characterize benchmark reference materials

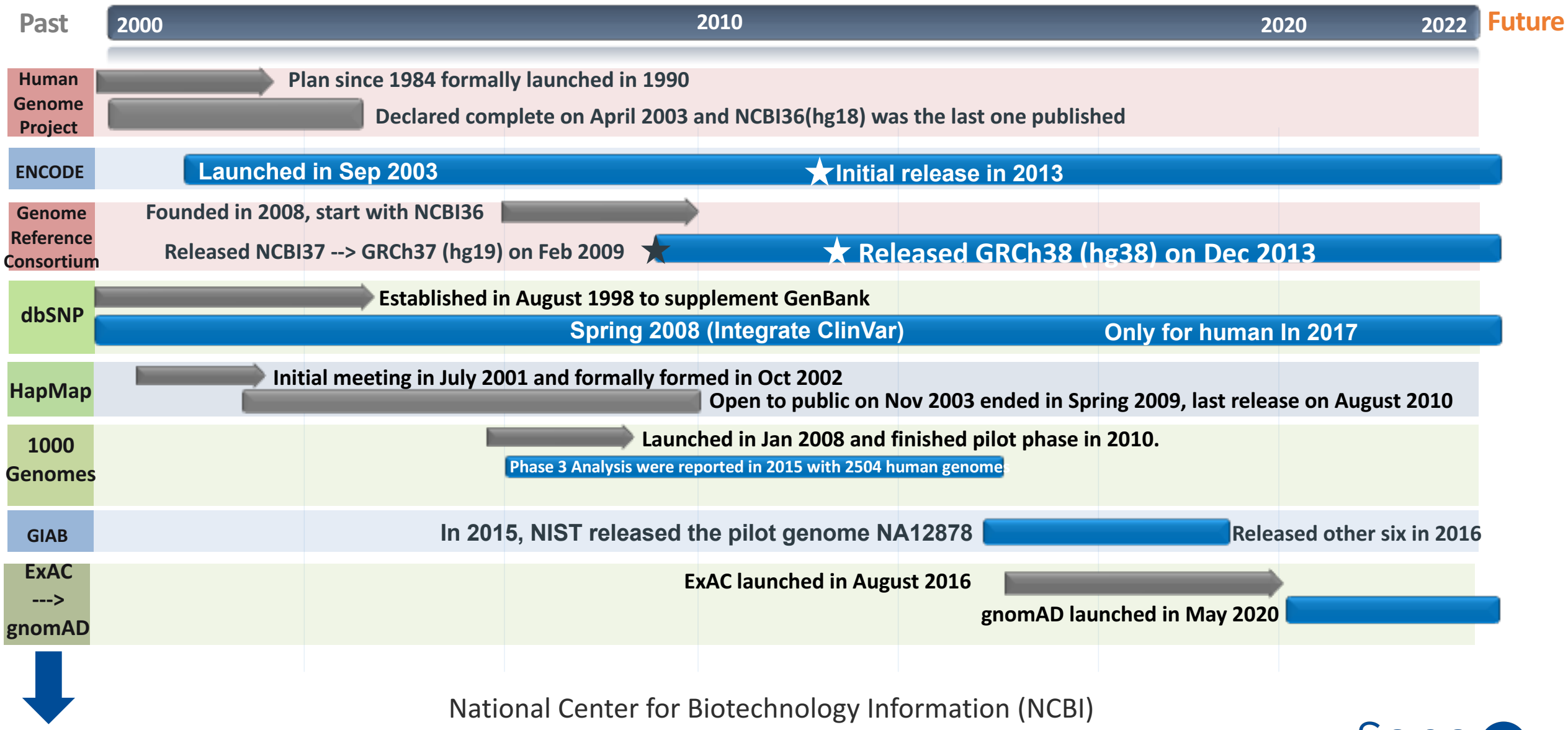
Lab	Technology	Library Type	Date Produced	Mean Paired end distance	Mean read length	Mean coverage per individual for AJ trio	Mean coverage per individual for Chinese trio
NIST	Illumina HiSeq 2500 Rapid SBS	WGS paired end PCR-free	Jul to Sep 2014	565 bp	2×250 (Chinese son), 2×148 (others)	296.83x	300x (son) 100x/parent
NIST/ Stanford	Illumina HiSeq 2500 v3/v4	WGS mate-pair	Dec 2014 to Mar 2015	6000 to 7000 bp	100	13x to 14x	14x to 17x
NIST/ Illumina	Illumina HiSeq 2500 v4	Synthetic Long Reads	Mar to Apr 2015	N/A	2×125	20x to 30x	20x to 30x
Oslo University Hospital	Illumina HiSeq 2500 v4	WES paired end	Oct 2015	202 bp	2×125	192x to 225x	181x (Son)
10X Genomics	10X GemCode WGS	WGS Linked-Reads	June to July 2015	200 to 220 bp	2×98	25x (son) 24x (mother) 22x (father)	None
Complete Genomics	Complete Genomics	WGS paired end	Sep 2014	388	26	101x	98x
Complete Genomics	Complete Genomics	WGS LFR	Dec 2014	278	26	100x	None
Thermo Fisher	Ion Proton	WES	Jun 2014	N/A	190	1020x	1036x
NIST	SOLiD 5500×l Wildfire	WGS single end	Jan to Aug 2015	N/A	55 to 57	79x (Son)	62x (Son)
BioNano Genomics	BioNano Genomics	WG optical mapping	Feb to Dec 2014	N/A	251 kb (Chinese son), 195 kb (AJ Son), 213 kb (AJ Mother), 246 kb (AJ Father)	92x (Mother), 87x (Father), 112x (Son)	57x (Son)
NIST/Mt. Sinai	PacBio P6-C4 (90%) & P5-C3 (10%)	WGS single end	Oct 2014 to Mar 2015	N/A	10–11 kb N50	69x son; 30–32x/parent	None
Weill Cornell	Oxford Nanopore	2D reads	May 2015	N/A	5.8 kb	0.017x (son)	None

Characteristics of these data for the AJ and Chinese trios are described here, and in more detail in this manuscript.

SANA CONFIDENTIAL – ©2019 Sana Biotechnology. All rights reserved.

Sana Biotechnology

Timeline

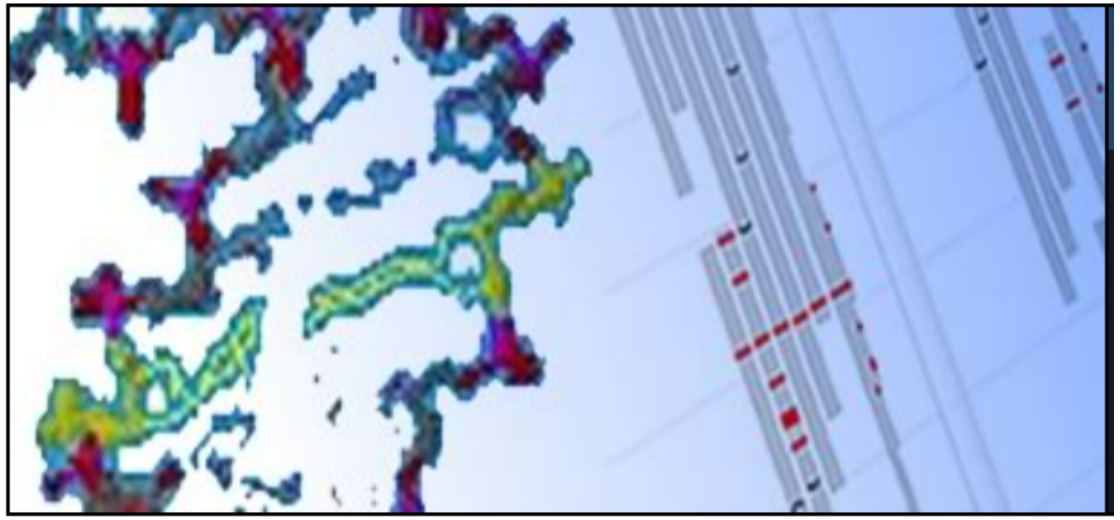


National Center for Biotechnology Information (NCBI)

SANA CONFIDENTIAL – ©2019 Sana Biotechnology. All rights reserved.

Sana Biotechnology

dbSNP/dbVar

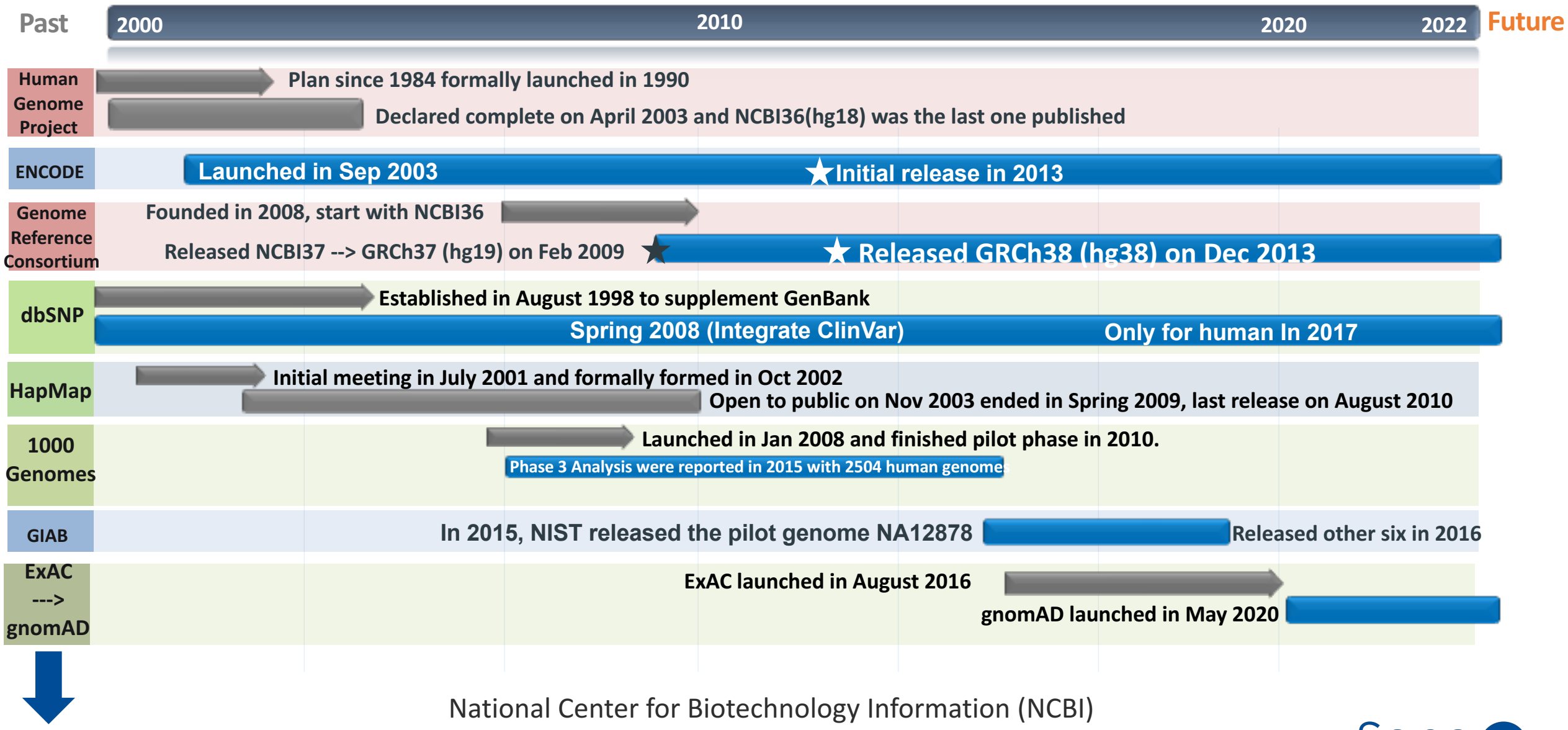


- Goal:
  - A database of small scale nucleotide variants
- Design of the Project
  - NCBI + NHGRI
  - dbSNP is an authoritative and trusted central repository for human variations.
- Data
  - Single-base nucleotide variation (SNV), short (=< 50bp) deletion/insertion polymorphisms, and other classes of small genetic variations. (dbVar for Variants > 50 bp)
  - > 675 million variants (2 billion submission)
  - ALFA Project Release 2 with 900M variants (https://www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/ALFA_20201027095038/)
  - The basis of scientific variant resources such as ClinVar, OMIM, SNPedia, the Variant Annotation Integrator at UCSC, and the Variant Effect Predictor at EBI

SANA CONFIDENTIAL – ©2019 Sana Biotechnology. All rights reserved.

Sana Biotechnology

Timeline



National Center for Biotechnology Information (NCBI)

SANA CONFIDENTIAL – ©2019 Sana Biotechnology. All rights reserved.

Sana Biotechnology

The Genome Aggregation Database (gnomAD)

Goal: for the wider scientific community

1. Improving power for disease gene discovery
2. Estimating genetic disease frequencies
3. Exploring the biological effect of genetic variation

Highlight:
Largest variant data from any species so far, incorporating exome or genome sequence data.

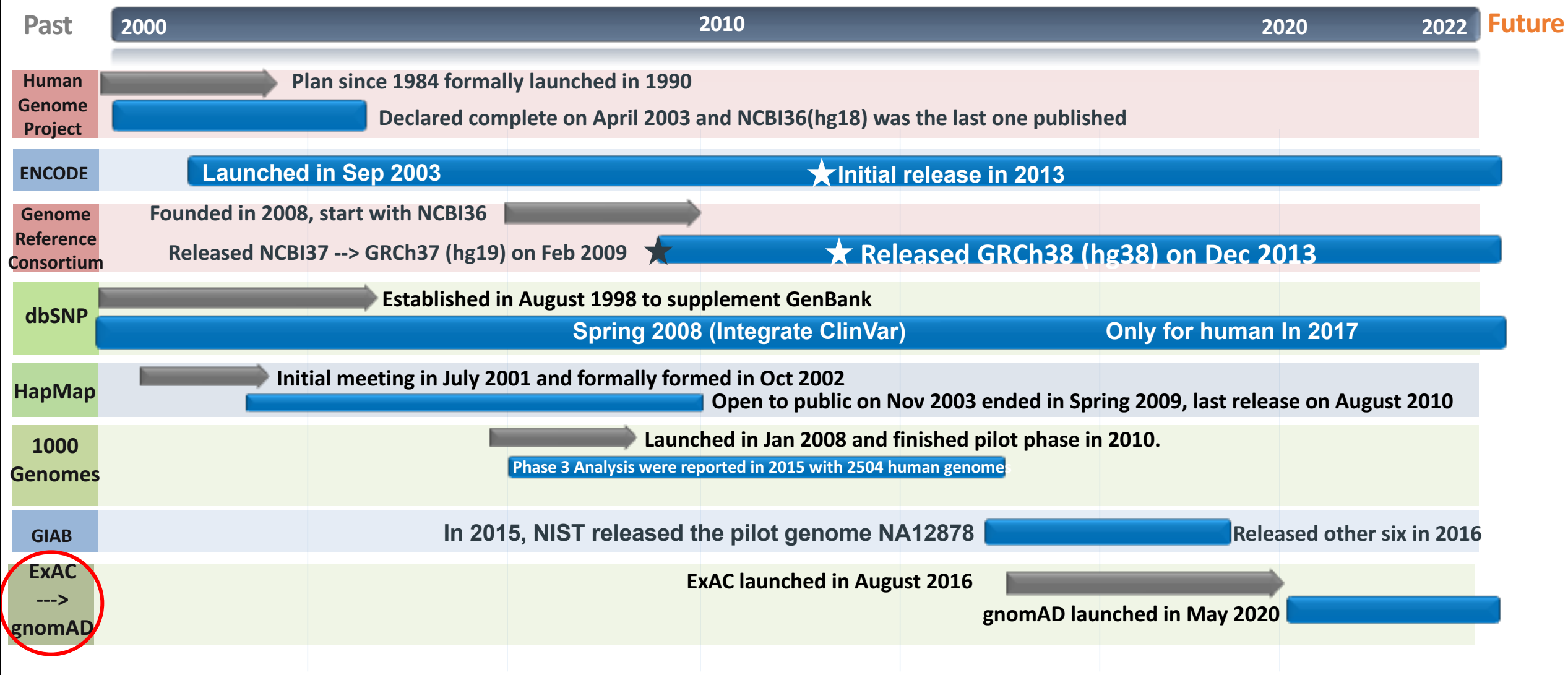
1. >3 petabytes(1000 TBs) raw data.
2. >35 TBs of variant data.
3. >140,000 humans (76,156 genomes)
4. >270 million variants.
5. >443,769 predicted loss-of-function genetic variants

- **Uniformly processed, joint variant calling basing a standardized BWA-Picard-GATK pipeline (Protocol)**
- Data analysis using Hail (https://hail.is/ & https://github.com/hail-is/hail)

https://confluence.corp.sana.com/display/~li.xiang/Introduction+of+gnomAD

SANA CONFIDENTIAL – ©2019 Sana Biotechnology. All rights reserved.

Summary



National Center for Biotechnology Information (NCBI)

SANA CONFIDENTIAL – ©2019 Sana Biotechnology. All rights reserved.

Sana Biotechnology

gnomAD

https://confluence.corp.sana.com/display/~li.xiang/Introduction+of+gnomAD

SANA CONFIDENTIAL – ©2019 Sana Biotechnology. All rights reserved.

Sana Biotechnology