Congressional Research Service
Informing the legislative debate since 1914

IN FOCUS

March 23, 2023

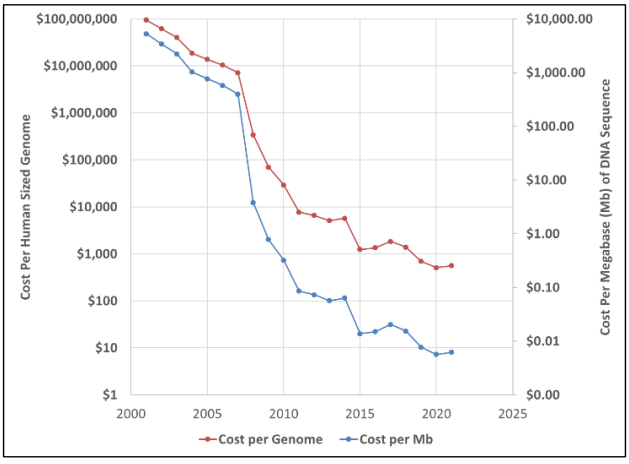
Digital Biology: Implications of Genetic Sequencing

Deoxyribonucleic acid (DNA) is the molecule that carries genetic information of an organism. This genetic code is composed of nucleotide bases (A, T, C, and G). The sequence of these bases encodes information that can, for example make a protein. A genome is the complete set of DNA in an organism. A gene sequencer reads DNA. Gene synthesis technologies can write DNA. It is this ability to both read and write DNA that researchers in the field of engineering biology use to reprogram cellular systems at the genetic level for a specific functional output. To do so, researchers require data about what gene sequences to code for, what functions those genes impact, and how those genes are expressed in living organisms.

Many emergent technologies, such as artificial intelligence (AI), require large datasets, often referred to as "big data." Theoretically, as more data becomes available, the capabilities of those technologies increase. This includes applications that require the use of genetic sequence data.

Sequencing technologies have evolved rapidly, making it possible to sequence entire genomes more efficiently and at lower cost. Sequences are collected and stored in databases, many of which are publicly funded and freely accessible, while others are privately held. The volume of genetic sequence information in databases has grown as sequencing technology has evolved. See **Figure 1** and **Figure 2**.

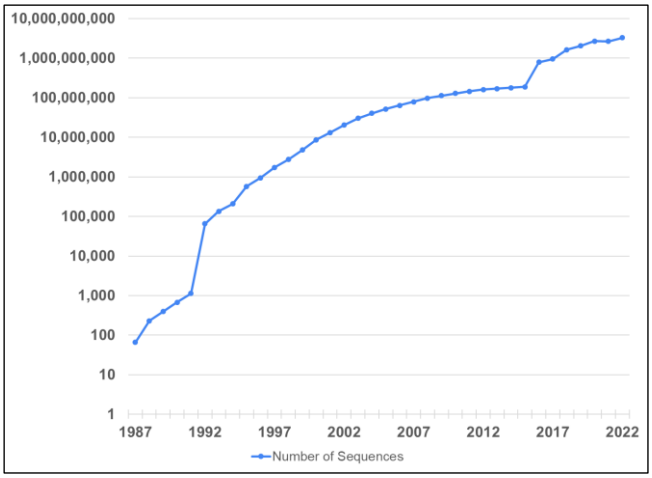
Figure 1. Cost of DNA Sequencing over Time



Source: CRS analysis of data from Kris Wetterstrand, "DNA Sequencing Costs: Data from the National Human Genome Research Institute Genome Sequencing Program," National Institutes of Health, at http://www.genome.gov/sequencingcostsdata.

Notes: A megabase (Mb) is a unit of measurement for DNA. One megabase = 1 million bases. For generating the "Cost per Genome," the assumed genome size was 3,000 Mb.

Figure 2. Growth of Sequences in the International Nucleotide Sequence Database Collaboration



Source: CRS analysis of data from the International Nucleotide Sequence Database Collaboration (INSDC).

Notes: INSDC includes sequence data from the DNA Data Bank of Japan; the European Nucleotide Archive; and GenBank, the National Institutes of Health genetic sequence database.

Sequencing Life on Earth

Private companies and public research groups produce large amounts of genetic sequence data. For example, the Broad Institute of MIT and Harvard claims to produce roughly 500 terabases (500 trillion bases) of genomic data per month. There is great potential value in the aggregate volume of genetic datasets that can be collectively mined to discover and characterize relationships among genes.

In 2018, the National Institutes of Health launched the All of Us precision medicine research program, which aims to collect clinical, lifestyle, electronic health record, and genomic data from at least 1 million people to advance the development of precision medicine. Since its launch, the program has made available about 100,000 whole genome sequences. Genomic data, along with other information, including data about the communities where participants live, is available via a cloud-based platform. All direct identifiers are removed from the data, and other privacy requirements have been put in place for researchers seeking access, in order to protect participants' privacy. This combination of data may help researchers better understand how genes can cause or influence diseases in the context of other health determinants.

The Earth Microbiome Project (EMP) is a global research project to sequence global microbial life funded by public and private entities. Its goal was to sequence 200,000 samples from different biomes to produce a global Gene

Atlas. The project is currently at capacity and not accepting new samples until additional funding is identified.

Estimates suggest less than 0.1% of the identified plant and animal species have been sequenced. The Earth Biogenome Project, an international network of public, private and nonprofit institutions, is attempting to sequence, catalogue, and characterize all known animal, plant, and fungal species within 10 years. Launched in 2020, the project suggests that accomplishing this goal could have numerous scientific and societal impacts. These include better understanding of evolutionary relationships among organisms; better understanding of ecosystem composition and functions; the discovery of new species; the study of the role of climate change on biodiversity; better understanding and management of future pandemics; and identification of genetic variations for improving agriculture and developing new biomaterials.

White House Initiative

In March 2023, the White House Office of Science and Technology Policy released *Bold Goals for U.S. Biotechnology and Biomanufacturing*. Among other issues associated with genetic sequencing, it announced a goal of sequencing 1 million microbial species' genomes within five years and stated, "Storing and analyzing huge amounts of genome and phenotype data will require innovations in computing, including artificial intelligence."

Societal Concerns

The declining cost of sequencing has expanded the collection of genetic data, including by testing companies that give consumers access to their own genetic information. Sequencing and related capabilities have raised concerns over who is collecting the data, where it is being stored, what it can be used for (e.g., forensics), and who "owns" the data. For example, when an individual submits a sample to a genetic testing company, depending on the user agreement, the genetic data can be accessed by or sold to other users. Concerns over the publication and access to other types of genetic sequences (e.g., viruses) have raised additional biosafety and biosecurity concerns.

National Security Concerns

The Intelligence *Community's 2023 Annual Threat Assessment* stated that the fields of AI and biotechnology are "being developed and are proliferating faster than companies and governments can shape norms, protect privacy, and prevent dangerous outcomes." The report identified genomic sequence data as a particular area of interest, pointing toward efforts by countries, universities, and private companies that have created, or are creating, centralized databases to collect, store, process, and analyze genetic data. The report further identified China's efforts to collect U.S. health and genomic data through its acquisitions of and investments in U.S. companies, as well as through cyberattacks. This analysis followed a 2021 assessment by the National Counterintelligence and Security Center suggesting China understands that the collection and analysis of large genomic data sets from diverse populations can help foster new medical discoveries and cures with substantial commercial value and can advance its AI and precision medicine industries.

On March 2, 2023, the Bureau of Industry and Security (BIS) in the Department of Commerce amended the Export Administration Regulations (EAR) by adding BGI Research, BGI Tech Solutions Co., Ltd., and Forensic Genomics International to the Entity List pursuant to § 744.11 of the EAR. BIS states that the addition of these entities is based on information indicating their collection and analysis of genetic data poses a significant risk of contributing to monitoring and surveillance by the government of China, which has been utilized in the repression of ethnic minorities in China. BIS also indicates that the actions of these entities concerning the collection and analysis of genetic data present a significant risk of diversion to China's military programs.

International Governance

Digital sequence information (DSI) has been the focus of recent debate in multiple international forums. Debate has focused on how publication of and access to genetic sequences may affect international access and benefit-sharing (ABS) agreements around genetic material. While the United States is not a party to all of the agreements discussed below, outcomes of these negotiations may affect the strategic competitiveness of U.S. researchers and companies.

DSI issues have been raised in the context of the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization. Parties to the protocol have been negotiating whether DSI falls within its scope, and, if so, whether current ABS mechanisms are sufficient or a new mechanism is needed. While the United States has not ratified this protocol, it participates in the discussions. A decision adopted in December 2022 establishes a process to develop and operationalize a multilateral mechanism for ABS associated with DSI.

In March 2023, a draft agreement under United Nations Convention on the Law of the Sea protocol would establish a series of ABS requirements for DSI related to marine genetic resources collected in international waters. Although the United States has not ratified this protocol, it participates in the discussions. One draft requirement stipulates that covered DSI be entered into publicly accessible repositories and databases that are maintained nationally or internationally. The agreement, if adopted, also would establish a multilateral benefit-sharing mechanism, including monetary payments derived from any utilization of covered DSI.

The Plant Treaty, ratified by the United States in 2016, aims to guarantee food security through the conservation, exchange, and sustainable use of plant genetic resources. Parties are discussing how DSI may affect its ABS mechanisms, but no formal decisions have been made.

Considerations for Congress

Policymakers may consider how current federal efforts related to research, collection, use, and retention of genomic sequence data impact U.S. competiveness and national security concerns. Another issue for Congress may be whether the federal government should facilitate or regulate access to certain genomic data or regulate certain uses.

Todd Kuiken, Analyst in Science and Technology Policy

Disclaimer