Emergence & Convergence

Research Paper No. 3

July 2018

The Digitization of Biology: Understanding the New Risks and Implications for Governance

By Natasha E. Bajema, Diane DiEuliis, Charles Lutes, and Yong-Bee Lim

Dr. Natasha Bajema and Dr. Diane DiEuliis are Senior Research Fellows, Charles Lutes is the Director, and Mr. Yong-Bee Lim is an intern at the Center for the Study of Weapons of Mass Destruction at the National Defense University. Opinions, conclusions, and recommendations expressed or implied within are solely those of the authors and do not necessarily represent the views of the Defense Department or any other agency of the Federal Government.

The CRISPR-Cas9 gene editing technique has received much media attention in the life sciences over the past few years. This is due to its vast potential for transforming the field of synthetic biology, accelerating the prevention and treatment of disease, and developing new products to improve human life—CRISPR-Cas9 is also significantly cheaper, easier, and quicker to use than previous gene editing techniques.

The media hype about CRISPR-Cas9, however, has overlooked another equally profound transformation that is underway in the life sciences—the digitization of biology.[1] CRISPR-Cas9 operates like a molecular scissors and offers scientists a new tool for editing the genes of living organisms. However, the true potential of CRISPR-Cas9 depends on genomic data, i.e., accurate and digitized knowledge about gene sequences and genomes of living organisms.

In this paper, we seek to help U.S. policymakers understand the implications of the digitization of biology. To that end, we discuss the risks posed by different types of genomic data, examine the process of digitization, and outline the implications of this trend for governance in the field of synthetic biology.

The Digitization of Biology

Scientists continue to wrestle with the dual use dilemma of their field: rapid advances and development in biotechnology have tremendous medical benefits to humankind, but they can also create security, safety, or ethical issues.[2] As gene editing becomes more common-place, there is growing concern these techniques could lead to new pathways for the development of biological weapons

such as enabling the production of enhanced biological agents. CRISPR-Cas9 increases the risks that nefarious actors could use gene editing for malevolent purposes—for example, enhancing certain characteristics of existing pathogens or creating novel pathogens to cause harm. In 2016, Director of National Intelligence James Clapper went as far as including gene editing among the top WMD-related threats faced by the United States.[3]

Though a known feature of bacteria for protecting against viruses, scientists first proved the utility of CRISPR-Cas9 for modifying the genome of living organisms in 2012. The technique uses the Cas9 enzyme and an RNA molecule matching the target DNA sequence identified for editing. The RNA guides Cas9 to the correct target sequence in the genome where it cuts the DNA. After the cut, the DNA is repaired, causing the gene sequence to be disrupted or modified.

Despite being easier to use than previous techniques, scientists cannot effectively use CRISPR-Cas9 without reliable data about where to make the cuts in the sequence, how to avoid off-target effects or other unintended consequences, or even more fundamental data about what gene sequences code for what functions and how those genes are expressed in living organisms. Scientists have not yet sequenced all living organisms, and have only made partial connections between certain gene sequences and functions as expressed in organisms.

Today, researchers no longer need a physical sample of DNA to manipulate it or study it. In 2010, J. Craig Venter's team became the first scientists to create a living organism from computer data.[4] His team assembled a genome based on digitized DNA sequences, synthesized the DNA, inserted the artificial DNA into a bacterial cell, and life took over from there. The bacteria began to function, grow, and replicate.

The volume of digitized genomic data is on the rise. Over the past several years, scientists have responded to dramatic reductions in the cost of DNA sequencing and synthesis, computing power, and data storage by sequencing greater numbers of gene sequences and the genomes of living organisms and digitizing this information for storage in online databases and analysis on computers. To simplify the creation and modification of living organisms, scientists are identifying standard, interchangeable DNA sequences that code for certain functions, and are building online catalogs to make this information available. Scientists from around the world can then leverage this growing volume of genomic data to construct new genes and DNA sequences of interest, and potentially create new living organisms from scratch. Rather than acquire physical samples, researchers can now search these online catalogues for sequences of interest and analyze the data and/or have them synthesized to work with them in a lab environment.

The digitization of biology has made synthetic biology simultaneously more accessible and more powerful. Since 2004, teams of high-school and college students come together at MIT in Boston for the annual iGEM (International Genetic Engineered Machine) competition. In 2017, more than 300 teams from around the world competed to design, build, test, and measure an original biological system using standard DNA sequences and current molecular biology techniques such as CRISPR-Cas9. Meanwhile, in academic and government labs, scientists and engineers are working with sophisticated, computer-aided design and modeling tools, allowing them to rewrite and reprogram entire genomes—a growing field called bioinformatics.

As biology continues its rapid transformation into a new branch of information technology, more and more biological information is moving back and forth between the physical and digital worlds.[5] The availability, breadth, sophistication, and digitization of genomic data are growing rapidly and propelling synthetic biology forward as an important means for treating disease and innovating biologically-derived chemicals, pharmaceuticals, and biologics. As such, synthetic biology has become a vital engine of the U.S. economy.[6]

However, the digitization of biology also exacerbates the traditional national security risks associated with dangerous pathogens. It further introduces many new risks and vulnerabilities that occur at the interface between the life sciences and cyberspace. These new risks have national security implications and include issues as diverse as privacy and discrimination, loss or theft of data, unauthorized access to data, commercial sabotage, and hacking.

WHAT IS GENOMIC DATA?

With recent advances in synthetic biology, scientists now have direct access to life's genetic code and the ability to manipulate it. All living things, bacteria, viruses, plants, animals, and humans, contain genetic information that controls the way an organism grows and operates over time: deoxyribonucleic acid (DNA) or ribonucleic acid (RNA).[7] When the entirety of an organism's DNA or RNA sequence is mapped, this is referred to as its genome.

The genomes of living organisms belong to a broader category of genomic data, which consist of gene sequences, entire genomes, data that links genes to specific functions, and other types of metadata for a broad range of organisms including humans, animals, plants, and microbes. When genome data are combined with the understanding of how specific DNA sequences function, the potential result is the creation of new biological organisms and manipulation of existing organisms toward specific ends.

Genetic material *does not* determine destiny. The presence of certain genes in a genome does not always line up perfectly with what happens in an organism. Genes encode the capability and instructions for the phenotype, but their expression can be modified by the environment.

Unlike static data, which do not change over time, genomic data have dynamic features. Mutations can occur when DNA is incorrectly copied or when DNA is exposed to environmental factors such as chemicals and radiation. Exposure to hazards can lead to changes in the base sequence of DNA and explain variation in strains of pathogens and the evolution of species over time.

For the purpose of understanding the diverse range of risks of genomic data, we identify three different categories: pathogen, human, and industrial. The use and risks associated with each category of genomic data varies according to its source (specific organism), setting, and/or context. All three categories are dual-use: capable of supporting both good and nefarious uses. Each category of data is used in different ways to drive innovation and create good for humanity. At the same time, each category produces potential national security risks. In the following, we discuss each type of genomic data and how it may be used for good or for ill (See Appendix A for a more detailed breakdown of risks for each category).

Pathogen Genomic Data

In the national security realm, policymakers are most familiar with the uses and risks associated with pathogen genomic data. Pathogen genomic data refers to gene sequences and/or genomes of microorganisms such as viruses and bacteria that cause diseases in plants, animals, and people.
Scientists use pathogen genomic data to conduct infectious disease research and increase understanding of the complex interactions between pathogens and hosts that result in disease. This research also provides a foundation for the prevention and mitigation of infectious diseases as a result of natural outbreaks and/or intentional release. To date, scientists have used pathogen genomic data to create new diagnostic techniques, vaccines, and therapies.

Research conducted using pathogen genomic data to advance human health can also support nefarious intent. Already in 2002, scientists were able to recreate poliovirus from genomic data using research available online and ordering the gene sequences by mail order, a project that took over three years.[8] More recently, in 2016, scientists at the University of Alberta in Canada pieced together the genome of the horsepox virus to help develop more effective vaccines for the variola virus (smallpox), its close relative. Over the course of six months, scientists ordered DNA sequences of the virus by mail, put them together, and synthesized the virus in the lab.[9] The project cost about $100K, which is rather cheap by scientific standards.

When the scientists attempted to publish their research results, a heated controversy broke out across the scientific and policymaking communities about potential national security implications of the research at a time where gene editing techniques have become easier, cheaper and quicker to use.[10] Policymakers and security experts are concerned that with an increase in the number of actors engaging in research with pathogen genomic data, the risk of malicious use of such data also increases.

Malicious actors may be able to leverage CRISPR-Cas9 and the knowledge generated from legitimate research using pathogen genomic data to cause harm. However, there is still no evidence that such actors are developing enhanced biological agents. As gene editing technology becomes easier to use, this may change in the future.

Human Genomic Data

Unlike pathogen data, national security policymakers have only recently become aware of potential privacy and security risks arising from human genomic data. Human genomic data refers to sequences and/or entire genomes of individual people.

Scientists map and analyze human genomic data to understand how genetic differences contribute to an individual's development and determine how these differences contribute to a person's susceptibility to a variety of chronic conditions, including cardiovascular disease, Alzheimer's disease, and diabetes.

With a greater understanding of human genes, their expressions, and how they are regulated, researchers can apply this understanding to better define the pathways that lead to chronic illness and disease. Once these pathways are better defined, researchers can then develop targeted approaches to treating and curing diseases through personalized medicine.

While life science practitioners use human genomic data for good, the storage of and access to this personal data in online databases raises a number of important

privacy and discrimination issues. Unfortunately, a person's genomic data cannot be replaced like a hacked or stolen credit card account: there is no replacement for an individual's genomic data. Nonetheless, growing volumes of insufficiently protected human genomic data might be used in ways that we have yet to anticipate.

For example, in April 2018, the Sacramento police department arrested 72-year-old Joseph James DeAngelo, suspected to be the long-sought Golden Gate Killer and believed to be responsible for killing 12 people and raping more than 50 women in the 1970s and 1980s.[11] The Sacramento police used the GEDmatch, a free, publicly accessible online database to match DNA found at the crime scenes to profiles of the killer's distant relatives and narrowed in on a suspect.

According to their website, "GEDmatch provides DNA and genealogical analysis tools for amateur and professional researchers and genealogists."[12] After a quick user registration, individuals who purchase DNA analysis from companies such as 23andMe or AncestryDNA can enter their own DNA profile in order to find and locate other family members.[13] Users of GEDmatch agree to make their data public in the hopes of being connected to lost relatives.

When Sacramento police officers searched the database, more than 100 users matched the DNA profile of their suspected killer as a distant relative.[14] Police began contacting distant relatives of the suspected killer. After four months of developing a family tree, they honed in on Joseph James DeAngelo. To make the final confirmation, they retrieved a discarded DNA sample from his garbage and matched it to the killer's DNA.

In addition to privacy concerns, growing repositories of human genomic data present national security risks as well. For example, malicious actors may be able to exploit research on disease pathways and vulnerabilities to increase the likelihood and/or severity of chronic illnesses and infectious diseases. This is a different risk than engineering a pathogen to be more dangerous. Rather, it refers to exploiting host vulnerabilities in such a way to increase the likelihood or severity of an illness or condition. Finally, adversaries could potentially use large datasets of human genomic data to find DNA patterns that are shared by groups of people and target specific groups. Depending on the type of attack used by an adversary, genocide might become possible.

Industrial Genomic Data

Despite the growing U.S. bioeconomy, national security policymakers have not given much consideration to the more abstract and less immediate risks of genomic data used for manufacturing organisms within the field of synthetic biology. Industrial genomic data refers to sequences and/or genomes of microorganisms that are used to fuel biologically-derived economic activities. Unlike the case of pathogen genomic data, these microbes are not known to cause disease in humans and animals and therefore do not pose a direct risk related to biological weapons.

Bio-industrial companies use genomic data collections to identify gene sequences that produce viable consumer products when introduced into engineered yeast cells, other engineered cells or cell-free pathways. For example, Ginkgo Bioworks uses genomic data to link key compounds of flavorings and aromatics to their corresponding gene sequences. These gene sequences are introduced into engineered yeast cells to produce sustainable synthetic alternatives to natural fragrances and flavorings such as rose and vanilla.

Beyond fragrances and flavors, companies use genomic data to provide goods and services that are marketable and economically competitive in fields such as food security, human health, and biofuels. Specific examples include Biohydrin, a synthetic version of natural rubber and BioSteel, a synthetic version of spider silk. This approach to making products is aimed at maximizing crop yields, accelerating drug discovery, and expanding biomanufacturing into new sectors. With market growths to exceed over $10 billion by 2018, this bioindustrial economy, also known as the bioeconomy, is poised to be highly lucrative.[15]

Although industrial genomic data do not directly contribute to the increased spread of disease, they support the research and development of the field of synthetic biology, enhance the contribution of biotechnology to the U.S. economy, and thus, may be a determinant of the economic strength of states in the future. Malicious actors might be able to leverage the knowledge generated from industrial genomic data research to do economic harm.

For example, malicious actors can manipulate and/or steal industrial genomic data. Many bio-industrial companies use specialized algorithms and computer analytics to assist in their research. These specialized algorithms, which are often proprietary, are directed towards very specific tasks. For example, a company like Ginkgo Bioworks uses proprietary algorithms on gene sequences to find better ways to produce compounds and proteins in engineered yeast cells. With such proprietary algorithms and analytics, companies may find gene sequences that increase the production and enhance the stability of produced compounds and proteins.

If this proprietary data is housed on network servers and is connected to the Internet, a competitor could hack the network and tamper with such information, thus sabotaging a rival company. At a minimum, the sabotage of data would force the company to halt research and production until the problem could be corrected.

Beyond manipulating data, adversaries can exploit security vulnerabilities to steal proprietary information. This proprietary information could include both the raw genomic data, as well as the data and algorithms arising from the data analytics. Once hackers steal this proprietary information, companies face the risk of losing their market edge to new competition. In other sectors, companies like American Superconductor (AMSC) and Coca-Cola have experienced such theft with varying degrees of negative repercussions.[16]

Genomic Data: Understanding the New Risks

Many of the risks of genomic data discussed above existed prior to the digital age. In the past, for example, malicious actors could gain access to vital genetic knowledge via the publication of pathogen genomes or sequences in academic journals and use such information to cause harm. Governments and industry were not immune to the accidental release of records containing human genomic data to the wrong end-user, compromising the privacy of those individuals affected. When such information is digitized, however traditional risks are exacerbated and new risks associated with cyberspace and information technology arise.

Over the last decade, researchers have taken advantage of rapid leaps in computer processing capabilities, the decreasing cost of storage space, and the steep drop in DNA sequencing costs. To gain a greater

understanding of how different organisms function based on their genetic code, researchers are producing an unprecedented volume of sequence data. In fact, data production has doubled every seven months since 2010, generating an exponential increase in the availability and volume of genomic data across a variety of living organisms.[17]

The risks associated with genomic data increase with the digitization of biology. While making genomic data available online provides immense benefits, the open-access nature of the Internet introduces new vulnerabilities that must be addressed by policymakers. Digital information moves at the speed of light and can easily be shared across electronic devices allowing for unprecedented access to information. By using the Internet as a backbone to acquire, store, and distribute genomic data, many more individuals are gaining access to all types of genomic data at any time and any place, to use as they wish. Not bound by a country's borders, digital information cannot be controlled in the same way as physical samples. Digital information is stored on computers, servers, and networks, all of which are subject to known cyber vulnerabilities: unauthorized access, theft, manipulation, and malicious use.

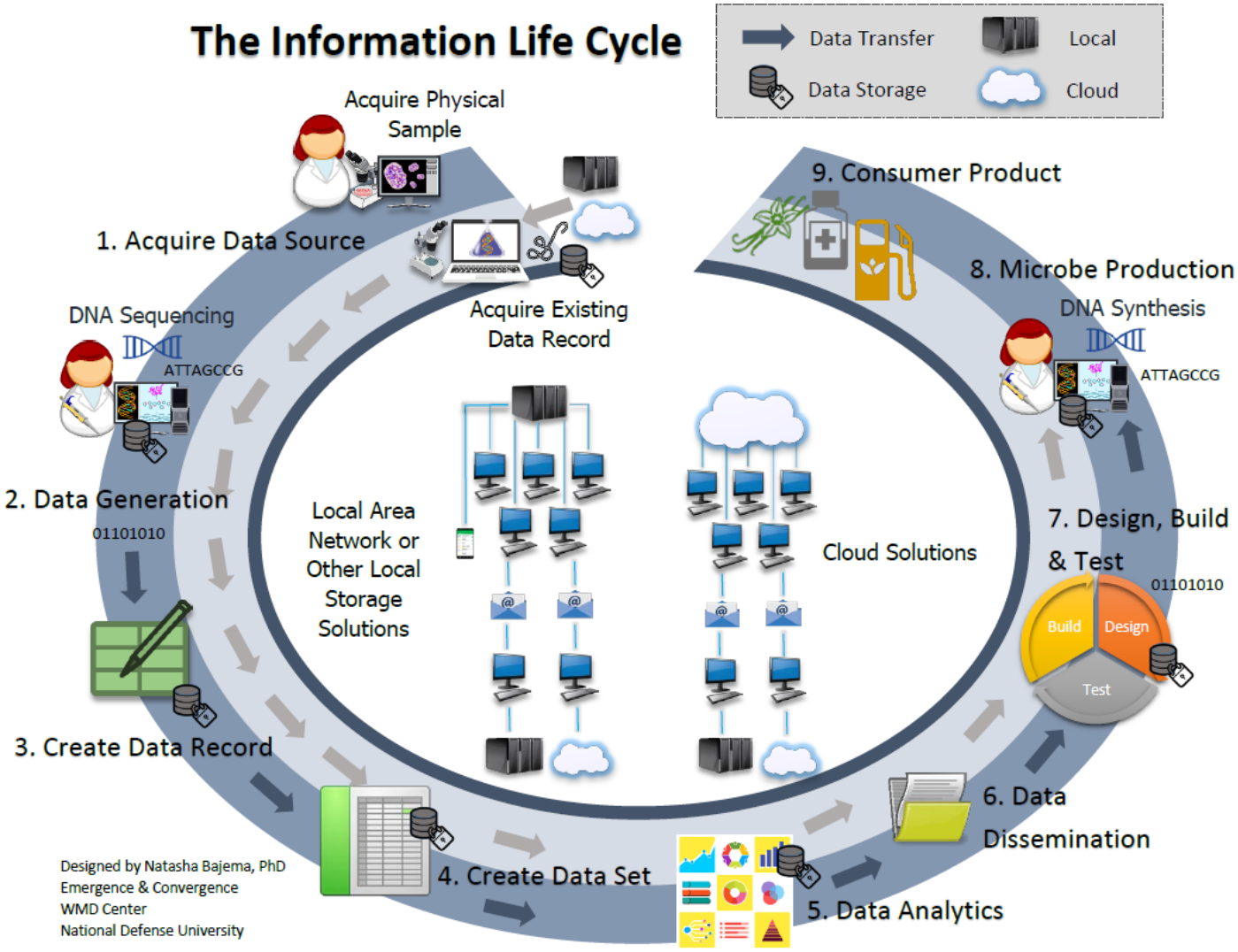
The Digitization Process: A Modern Information Life Cycle

To understand the risks of genomic data and implications for governance, we illustrate the process of digitization and discuss the different stages of the information life cycle for the life sciences—i.e., how scientists acquire, generate, and use biological information to conduct their research and/or develop biologically-derived products.

The digitization of biology refers to the translation of biological information for use in the digital world. This dynamic process can occur in either direction to facilitate scientific research (from physical to digital) or bio-industrial production (from digital to physical), but the primary purpose is to allow biological information to be used on computing platforms, stored in databases, analyzed using software, shared online, and sent by email.

Conversion from physical DNA of living organisms to digital information involves the sequencing of genes and/or genomes and conversion of the base pairs of As, Cs, Gs and Ts into ones and zeros that can be read by computers. Gene synthesis or the writing of DNA entails the reverse process of translating digital information stored on computers into a physical DNA sample and possibly into a living organism. Scientists use computer models to modify existing living organisms or design new ones. Once the desired organism has been created, the computer generates the DNA sequences which can then be synthesized and used to create a living organism from scratch in a laboratory.

The information life cycle refers to the process throughout which genomic data is produced, used, and transferred between the physical and digital worlds. As scientists conduct their research, they use computers and the Internet to efficiently generate, store, distribute, and analyze data at each stage of their study. As a result, genomic data is exposed to the many risks and vulnerabilities inherent in all digital information stored on a specific computer, local area network or cloud services. The illustration on the next page shows the different stages of the information life cycle, which are then discussed in greater detail.



At each stage in the information life cycle, genomic data is exposed to the many risks and vulnerabilities inherent in all digital information. The above illustration shows two pathways for the information life cycle. In the first pathway (dark gray), the researcher acquires a physical sample as the first step. In the second, the researcher acquires a data record for a genome or sequence of interest as the first step. The two pathways merge as digital information in the creation of data sets when all biological information becomes digitized regardless of its original source (physical or digital).

The illustration shows the many touch points along the research and bio-industrial production pathways where data is transferred and stored. These touch points correspond to potential vulnerabilities for the unintentional release, unauthorized access, hacking, theft, human error, and sabotage of genomic data.

Acquiring Samples (1)

The information life cycle begins when scientists collect gene sequences of interest to support their research. In general, scientists obtain samples in two ways: existing data records of gene sequences and/or genomes (digital) and physical samples of DNA to be collected and then sequenced (converted to digital). In contrast, existing data records are typically stored in an online database or available from a database on local computer networks.

Scientists typically start their research by searching a genome database such as the U.S. National Library of Medicine's National Center for Biotechnology Information (NCBI).[18] The NCBI's genomic database holds the gene sequences of over 33,000 organisms and is accessible to scientists and the public alike. Scientists can download custom datasets from the online database by using text queries or entering a set of unique identifiers.

In the event that a digital record for a gene sequence is not available in NCBI, scientists have other options. They may reach out to other scientists and researchers in the field. For example, academics may have sequences of interest stored as digital records on their own computers that can be transferred electronically. As members of a collaborative community, scientists often share their data sets. However, larger data collections would require shipping a hard disk by mail or using peer-to-peer file sharing technologies.[19]

Scientists may also approach private companies with proprietary genome databases and pay money to obtain electronic samples. If no digital records are available, scientists may decide to obtain physical samples for sequencing.

In general, scientists have three ways of obtaining physical samples of DNA. First, scientists can go out into the field and collect the samples from living organisms for sequencing. Second, they can obtain physical samples from other researchers. For example, scientists from the U.S. who are interested in variations in a virus endemic to Africa could reach out to researchers in the region to see if they are able to transfer blood samples of patients. Third, scientists can acquire physical samples from a bio-resource center like the American Type Culture Collection, a repository of microorganisms, cell lines, and other bio-related materials.[20]

Sequencing the Samples (2)

Unlike digital records, physical samples have to be sequenced to obtain their genomic data (converted into A's, T's, C's and G's and then 0's and 1's). As gene sequencers have become cheaper and more accurate over the past decade, scientists are increasingly sequencing a majority of physical samples in-house.

For developing countries and other laboratories without gene sequencing equipment, physical samples are typically sent to gene sequencing companies. While much cheaper than in the past, this method is not ideal for two reasons. First, sending physical samples overseas requires permits because they are subject to export control laws if shipped overseas. In addition, once put into the mail, the scientists no longer have complete control over the physical samples, which may introduce the potential for inaccurate results in their research.

Recording Sequence Data and Metadata (3)

Scientists create individual data records for all genomic data originating from sequenced physical samples. Each record contains the raw gene sequence as well as descriptive

data about the sample, including the organism's name, the strain, where the sample was collected, and any other important attributes. This descriptive data about the sequence is called metadata.

Metadata is important because it summarizes basic information about the sample. By providing this basic information, metadata can help scientists find and work with data easier if stored in a large dataset. For example, scientists may be able to explain variations in the gene sequence based on the origin of the physical sample.

Data Storage (4)

To develop countermeasures, vaccines and therapies, scientists need to collect and analyze large datasets of genomic data. These large, analyzed datasets direct scientists towards specific targets and mechanisms that might be leveraged to treat and prevent initial infection or the spread of a disease.

Once several samples have been sequenced, the collection of individual records becomes a dataset. This dataset is stored on a specific computer, the local area network and/or on the Internet. If this local area network is not connected to the Internet, access to the data is limited to certain individuals.

If local storage is not desirable or impossible due to capacity issues, scientists may turn to commercial bioinformatics platforms or cloud services such as those provided by Amazon, Microsoft, and Google to store and manage their datasets. In addition to providing storage services, bioinformatics platforms often offer access to advanced analytic tools.[21] Cloud services allow users to control access to the dataset and engage in collaboration and information-sharing with other scientists.

Scientists may also choose to submit their data records to online databases like NCBI, which operates its own cloud storage and allows public access to data through a user account. While this option limits the scientist's ability to control access to the data, it provides valuable information for the scientific community at large.

Data Analysis (5)

Once the data is stored, scientists can analyze the dataset using various software tools. Scientists apply algorithms and analytic techniques to make sense of the massive amount of information and produce findings from their research.

Data scientists at bioinformatics companies have developed software tools designed to mine raw data, perform biological modeling on the data, and then generate complex computational analysis of biological metadata, systems, and pathways. By sharing tutorials on how to address 'big data' issues and offering up code to help others achieve 'big data' tasks or functions, data scientists are empowering more individuals to engage in big data analysis.

Many of the same software tools are widely used across academia and industry and built from open-source codes. These bioinformatics techniques are used to develop biofuels, precision medicine, and market-competitive bio-manufactured products.

Data Dissemination (6)

After completing complex data analyses of genomic data, researchers may publish their results in journal articles and conference presentations as a means to share information on new developments in the field. More recently, blog posts and posts on personal websites have also been ways to disseminate ideas and generate discussion.

Product Development (7,8 & 9)

To create or engineer biological organisms, it is essential to understand the specific functions (phenotypic data) coded in a vast array of DNA sequences of different organisms. Bio-industrial companies often work in the reverse process when compared to scientific researchers. Unlike a researcher building a dataset for the purpose of conducting research on a disease, bio-industrial companies leverage existing genomic data collections to develop high-quality products derived from living organisms.

Bio-industrial companies apply the engineering principle of "design build test" to the life sciences to create these products:

- **Design:** A bio-production company uses industrial genomic data to identify gene sequences of interest and select a prototypical function.
- **Build:** The developers introduce the sequences and create a first instantiation of the designed prototype.
- **Test:** Researchers and developers then test the built system for the purity of the product and feasibility and scalability of the production process.

As a company works to identify a suitable microbe for producing the desired product, they work back and forth through several stages in the information life cycle to recreate and synthesize the microbes from their genetic sequences to test them for their viability. Once a microbe is selected, these companies synthesize the organisms to produce the consumer product.

To streamline the production process, industry stakeholders are particularly interested in eliminating the need for actual cells to produce compounds. This method, referred to as a cell-free pathway, enjoys several advantages over the existing fermentation pathway. Companies save money since less equipment is required for cell-free pathways than fermentation pathways. In addition, cell-free pathways produce the product directly. This approach, which relies more heavily on genomic data, is far more efficient than the fermentation pathway, where the custom organism produces the desired product indirectly: a byproduct of its natural biological functions.

In addition to leveraging the vast supply of genomic data, bio-industrial companies are increasingly turning to automation and information technologies to enhance their operations. Companies aspire to make their production processes more efficient by using advanced robotics to monitor fermentation conditions and run repetitive activities. Automation would free up human resources for other intellectual work, such as research and testing. Though increasingly automated, in most cases today, laboratory technicians still operate the controls on-site.

As remote automation becomes less unwieldy, these companies may use computer networks to allow people to monitor the production cycle remotely. These networks may be local, closed networks: a group of computers and associated devices that share a common communications line that is not connected to the Internet. However, these networks may also be connected to the Internet so employees can more easily monitor and control the production cycle remotely.

Information Life Cycle: Understanding the New Risks

Throughout the information life cycle, both researchers and bio-industrial companies store and use genomic data on computers, local area networks, and/or cloud services and transfer such data between users over email or peer-to-peer sharing technologies. As such,

genomic data are exposed at many touch points throughout their use to the risks and vulnerabilities common to cyberspace such as hacking, data theft, sabotage, and unauthorized access. In most cases, only minimal encryption or other cybersecurity safeguards are used to secure genomic data at these touch points in the information life cycle.

Implications for Governance

The digitization of the life sciences, the rise of accessible gene editing tools, and the growing volume of genomic data available online introduces a host of diverse risks and vulnerabilities related to cyberspace that have yet to be addressed by policymakers concerned about biosecurity. These risks and vulnerabilities are exacerbated by a general lack of awareness among scientists and researchers and the absence of effective governance measures for protecting genomic data in the first place (See Appendix B for a comprehensive review of existing domestic and international governance).

The U.S. government has primarily focused its biosecurity efforts on restricting access to physical samples of high-risk agents found on the Select Agent List. Even before the digitization of biology, this framework had limitations because pathogens exist in nature and are endemic to specific regions. Today, the availability of pathogen genomic data on the Internet could enable malicious actors to produce dangerous pathogens using gene sequencing and synthesis technologies without having to hunt for physical samples in nature or undergo a series of checks and investigations. Moreover, these actors could use digital genomic data to produce novel pathogens that are not on the Select Agent List. Such list-based approaches will have limited success in a world where malicious actors can order gene sequences from private companies.

Due to these risks, the gene synthesis industry has largely adopted voluntary screening of gene sequence orders. However, the decreasing cost of DNA synthesis will make screening costs even less financially attractive and put pressure on companies focused on the bottom line.

The governance structure for human genomic data and industrial genomic data is weak to non-existent. Despite protections against discrimination and violations of privacy related to health information, human genomic data are not even considered personally identifiable information under current U.S. law. Likewise, there are few protections for industrial genomic data except those afforded by intellectual property law, which are challenged by the difficulties of controlling digital information.

Furthermore, governance of genomic data should not be effectively viewed as simply a subset of cybersecurity. Protection of genomic data requires an understanding of how bio-scientists use and could potentially misuse such information. To be sure, good cyber hygiene is important, but not sufficient for protection against the misuse of genomic data.

Developing effective governance to simultaneously manage the risks and promote the opportunities of the life sciences is a difficult undertaking. To address emergent genomic data issues, policymakers must strike a balance between two factors: the perceived risks of genomic data and the incentives to share and use genomic data to foster innovation.

At minimum, U.S. policymakers should begin promoting awareness among scientists, researchers, and bio-industrial companies of the risks and vulnerabilities that occur as a result of the digitization of biology—at the

interface of the life sciences and cyberspace. The culture of responsibility in the life sciences for biosafety and biosecurity should include extending the stewardship of genomic data into the digital realm. Working with stakeholders, U.S. policymakers might consider facilitating the development of new standards of practice among scientists and bio-industrial companies to better protect all types of digitized genomic data.

Moreover, U.S. policymakers should explore the adoption of advanced encryption algorithms used by banks in the financial sector as a means for protecting digitized genomic data.[22] This will require striking a balance between the need for security and the ethos of the scientific community toward openness, sharing, and collaboration. Although advanced cybersecurity tools can mitigate some of the risks, such security measures come at the expense of efficiency, remote controllability, and ease of use, possibly impeding innovation through the use of genomic data. U.S. policymakers should strike a balance between safeguarding genomic data and making genomic data useable and accessible.

About the Authors:

Dr. Natasha Bajema is a Senior Research Fellow at National Defense University, the principal investigator for Emergence and Convergence, and Course Director for an elective entitled Through the Film-maker's Lens: Contemporary Issues in Combating Weapons of Mass Destruction.

Dr. Diane DiEuliis is a Senior Research fellow at National Defense University. Her research areas focus on emerging biological technologies, biodefense, and preparedness for biothreats. Dr. DiEuliis also studies issues related to dual use research, disaster recovery research, and behavioral, cognitive, and social science as it relates to important aspects of deterrence and preparedness.

Mr. Charles D. Lutes is the Director of the Center for the Study of Weapons of Mass Destruction (WMD) at the National Defense University (NDU) in Washington, D.C. Prior to joining the Center full time in 2013, Mr. Lutes served in the Office of the Undersecretary of Defense for Policy, where he was a senior advisor for Countering WMD and Acting Principal Director for Nuclear and Missile Defense Policy. Mr. Lutes culminated his 28-year Air Force career as Director for Counterproliferation and Nonproliferation on the National Security Council Staff under Presidents George W. Bush and Barack Obama.

Mr. Yong-Bee Lim is an intern at the WMD Center, National Defense University. He is a PhD candidate in George Mason University's Biodefense program and writing his dissertation on the biosecurity implications of the DIY community. Mr. Lim is a 2018 fellow of the Emerging Leaders in Biosecurity Initiative (ELBI) at Johns Hopkins Center for Health Security.

The authors are indebted to the following individuals for their reviews and insightful comments on earlier drafts of this paper: Dr. Tom Slezak, Mr. John Caves, and Dr. Seth Carus

Emergence & Convergence Study

In its multi-year study entitled *Emergence and Convergence*, the WMD Center is exploring the risks, opportunities, and governance challenges for countering WMD introduced by a diverse range of emerging technologies. The WMD Center identified advanced robotics as one of several emerging technologies for deeper assessment. Toward this end, the WMD Center has developed an exploratory framework for first identifying the emerging technologies that will have greatest impact on the WMD space for state and non-state actors and then for evaluating the nature of that impact on the current tools and approaches for countering WMD.

The Emergence and Convergence study is supported by several offices within the Office of Secretary of Defense (OSD) and receives its primary funding from the CWMD Systems Program Office within the Office of the Assistant Secretary of Defense for Nuclear, Chemical and Biological Defense Programs/Threat Reduction and Arms Control (NCB/TRAC/CWMD Systems). The support by Mr. Jim Stokes, Director, CWMD Systems Program, has been critical to the project's success.

APPENDIX A: THE RISKS OF GENOMIC DATA

This appendix discusses in further detail the risks and vulnerabilities associated with each type of genomic data (See Table 1 on the next page for a complete overview of the risks for each category of genomic data). Although there is some overlap in solution sets, we grouped the risks into four subtypes: capability risks, data risks, cybersecurity risks, and societal risks of privacy and discrimination. Each subtype represents a distinct problem set requiring specific solutions.

- **Capability Risks**: In conjunction with gene editing techniques, both pathogen and human genomic data can be used to ease the acquisition and development of biological weapons. These enhanced capabilities include:
  - Obtaining electronic genomic data to do harm;
  - Using genomic data to engineer new pathogens;
  - Using genomic data to recreate extinct, high-impact pathogens;
  - Using genomic data to modify low-risk pathogens to become high-impact pathogens;
  - Using genomic data to increase the likelihood of disease;
  - Using genomic data to enhance targeting of the recipient; and
  - Using genomic data to enhance the deleterious effects of pathogens.
- **Data Risks:** Scientists produce and store more data online every year. As data accumulates, collections of data face the same fundamental issues of all big data ventures:
  - Creating inaccurate data through machine or human error;
  - Finding ways to consistently catch and fix inaccurate data; and
  - Prioritizing data storage as storage space diminishes.
- **Cybersecurity Risks**: All three types of genomic data are available directly on the Internet or devices connected to the Internet, exposing them to the many vulnerabilities that exist in an Internet-connected world. The level of security to protect these data collections from misuse varies. Compared to certain pathogen and human genomic data which exist in open databases, biomanufacturers tend to restrict their data as proprietary and secure industrial genomic data and their analytics to remain globally competitive. These cybersecurity risks include:
  - Transferring data securely to the correct end users;
  - Accessing proprietary or high-risk information without authorization;
  - Editing data deliberately to be incorrect;
  - Stealing proprietary or high-risk data; and
  - Stealing proprietary tools to analyze datasets.
- **Societal Risks of Privacy and Discrimination:** The intrinsic link between an individual person and their human genomic data creates ethical issues that do not exist for the other categories of genomic data identified in this paper. Societal risks include:
  - Releasing human genomic data unintentionally;
  - Releasing human genomic data intentionally; and
  - Engaging in discriminatory practices based on human genomic data.

Table 1: The Risks of Genomic Data

	Pathogen	Human	Industrial
Capability Risks			
Obtaining genomic data to do harm	☑	☑	☑
Using genomic data to engineer new pathogens	☑	☑	☑
Using genomic data to recreate extinct, high-impact pathogens	☑	☒	☒
Using genomic data to modify low-risk pathogens to become high-impact	☑	☑	☑
Using genomic data to increase the likelihood of disease	☑	☑	☒
Using genomic data to enhance targeting of the recipient	☑	☑	☒
Using genomic data to enhance pathogens	☑	☑	☑
Data Risks			
Creating inaccurate data through machine or human error	☑	☑	☑
Finding ways to consistently catch and fix inaccurate data	☑	☑	☑
Prioritizing data storage as storage space diminishes (data loss)	☑	☑	☑
Cybersecurity Risks			
Transferring data securely to the correct end users	☑	☑	☑
Accessing proprietary or high-risk information without authorization	☑	☑	☑
Editing data deliberately to be incorrect	☑	☑	☑
Stealing proprietary or high-risk data	☑	☑	☑
Stealing proprietary tools to analyze datasets	☑	☑	☑
Societal Risks of Privacy and Discrimination			
Releasing human genomic data unintentionally	☒	☑	☒
Releasing human genomic data intentionally	☒	☑	☒
Engaging in discriminatory practices	☒	☑	☒

☑ Applies to this type of genomic data

☒ Doesn't apply to this type of genomic data

APPENDIX B: RELEVANT GOVERNANCE

In this appendix, we review the current U.S. national and international governance that apply to the three types of genomic data (pathogen, human, and industrial) and highlight the limitations of each governance measure.

Pathogen Genomic Data

As illustrated in Table 1, pathogen genomic data pose a range of different risks from enhancing the capabilities of malicious actors (e.g., enabling them to engineer new pathogens), to unintentional errors (human error), to deliberate sabotage of datasets (commercial sabotage). The U.S. government has primarily focused its biosecurity efforts on restricting access to physical samples of high-risk agents.

At the domestic level, the U.S. governs the risks associated with pathogen genomic data through the Federal Select Agent Program, the Department of Health and Human (HHS) Screening Guidance Framework, and the Dual-Use Research of Concern (DURC) policy. In the following, we outline each of these policies and the risks they address and highlight relevant gaps pertaining to pathogen genomic data.

The Federal Select Agent Program

The Federal Select Agent Program (FSAP) creates a framework for the oversight of possession, use, and transfer of physical samples of select agents. The framework focuses on restricting access to physical agents and toxins through three mechanisms. First, HHS and the U.S. Department of Agriculture are responsible for maintaining, implementing, and enforcing the Select Agent List—a list of high-risk pathogen agents with the potential to pose a severe threat to public health and safety. Select agents require stringent measures to ensure the agents are used and researched properly and are not misappropriated for malicious purposes.

Once the select agent list is developed, access to agents on the list is restricted by conducting security risk assessments of end users and maintaining a national database of where select agents are located. Finally, the framework enables law enforcement officials to conduct investigations in cases of non-compliance.

This framework addresses the risks of malicious actors gaining access to and misusing physical pathogen samples, potentially to cause harm. It also discussed several shortcomings for mitigating the risks posed by digital information.

Even before the digitization of biology, the framework suffered from a significant problem: select agents are not confined to research laboratories and pathogen repositories like the American Type Culture Collection. Pathogens, including select agents, exist in nature and are endemic to specific regions. In the past, the Soviet Union used highly-trained scientists to acquire virulent strains of the pathogens that cause anthrax, rabbit fever, and plague from outbreaks in nature. Even non-state actors, have attempted to harvest infectious agent from the soil near areas of known contamination.

The digitization of biology exacerbates the weaknesses of the framework. Some pathogen genomic data is now openly available in databases and published in journal articles on the Internet. Historically, this data was limited to a select audience, such as security experts and scientists with access to hard copies of papers or academic conferences. The availability of pathogen genomic data on the Internet could enable actors to gain access to a dangerous pathogen. Recent advances in

gene sequencing and synthesis may enable individuals to use digital records to produce high-risk pathogens without having to hunt for physical samples in nature or undergo a series of checks and investigations.

Using genomic data online, individuals can place orders to a number of companies for gene sequences of interest. For example, scientists achieved the chemical synthesis of poliovirus from scratch in 2002 and published their results. More recently, Canadian researchers synthesized the horsepox virus, related to the smallpox virus, from gene sequences ordered through the mail. In both of these cases, though, scientists conducted experiments to better understand the disease and potential treatment.

HHS Screening Framework Guidance

The Department of Health and Human Services (HHS) developed its Screening Framework Guidance to address advances in gene sequencing and synthesis over the past decade.

The HHS Screening Framework Guidance addresses the risk of individuals gaining access to high-risk pathogens from gene synthesis companies. The framework suggests that companies that produce and sell synthetic gene sequences to customers should be responsible for both customer screening and sequence screening to prevent malicious individuals from gaining access to high-risk gene sequences.

As their first responsibility, synthesis companies should understand the profiles of their current and potential customers. The guidance recommends that companies develop customer screening mechanisms to determine the legitimacy of customers who order gene sequences, such as confirming a customer's identity, identifying potential red flags, and ensuring that customers conform to U.S. trade restrictions and export control regulations. As their second responsibility, synthesis companies should screen orders for potentially dangerous sequences associated with agents on the select agent list. Known as "sequences of concern", the guidance suggests that companies should engage in additional follow-up procedures if fulfillment of the order would allow individuals to gain access to dangerous sequences. Such follow-up procedures could include verifying the legitimacy of the customer, the principal user, and/or the end-use of the ordered sequence.

U.S. policymakers considered feedback from synthesis companies in the development of this guidance. As such, it avoids overregulation often feared by the private sector and complements business practices. Adherence to the guidance remains voluntary.

Although the guidance acknowledges the risks of genomic data, it has several limitations for reducing risk, especially in the future. First, as already mentioned, the guidance is voluntary, and companies may choose whether or not to comply. As the barriers to entry decrease and a wider array of companies enter the synthesis market, the monopoly of existing synthesis firms that follow these screening standards may erode. Second, as the cost of gene synthesis continues to drop, the relative cost of screening sequence orders will begin to impact the bottom line and potentially become a disincentive for voluntary compliance. Third, the guidance is based on the select agents list. With the advent of synthetic biology, harmful entities could be created which lie outside the list. Finally, advances in new technologies such as the DNA printer, which integrates and automates the DNA synthesis and assembly process into a desktop device, will enable individuals to produce their own gene sequences.

Dual-Use Research of Concern

Advances in the life sciences are a double-edged sword: this issue is known as the dual-use dilemma. Scientific research can be used both for beneficial and malicious purposes. In 2012, the U.S. government issued a policy for biological dual-use research of concern (DURC), where DURC is "life sciences research that, based on current understanding, can be reasonably anticipated to provide knowledge, information, products, or technologies that could be directly misapplied to pose a significant threat with broad potential consequences to public health and safety, agricultural crops and other plants, animals, the environment, materiel, or national security."[23]

The DURC policy was created to address the risk of dual-use research with pathogen genomic data, including the engineering of new pathogens, recreating old pathogens, modifying low-risk pathogens, increasing the likelihood of disease, enhancing targeting, and increasing the severity of disease. The policy addresses the dual use dilemma by providing a framework for researchers to define DURC projects, determine how DURC projects should be assessed and funded, and impose repercussions for violations. DURC projects apply to a subset of 15 select agents from the FSAP or one of seven experiment categories of concern:

- Enhancing the harmful consequences of an agent or toxin;
- Disrupting the immunity or the effectiveness of an immunization against the agent or toxin without clinical or agricultural justification;
- Conferring resistance to clinically or agriculturally useful prophylactic or therapeutic interventions against agents or toxins;
- Facilitating a toxin's or agent's ability to evade detection methods;
- Increasing the stability, transmissibility, or the ability to disseminate the agent or toxin;
- Altering the host range of the agent or toxin.[24]

DURC projects are assessed by institutional biosafety committees (IBC's), which evaluate the projects for their risks and benefits as well as the strength and thoroughness of the project's risk mitigation plan.

The majority of DURC projects are funded by the federal government through agencies such as the National Institutes of Health (NIH) and the Centers for Disease Control (CDC). Policymakers can enforce the policy on DURC by denying funding to researchers if they violate the DURC policy.

The DURC policy has enhanced national security by providing additional controls on what research should and should not be funded by the government. However, the policy is bound in its scope by the Select Agent list, the utility of which is eroded by advancements in synthetic biology that make possible the design of novel pathogens.

In addition to domestic governance, the U.S. government participates in international regimes that restrict the behavior of states actors related to pathogen genomic data: the Biological Weapons Convention and the Australia Group. Moreover, gene synthesis companies participate in the International Gene Synthesis Consortium.

Biological Weapons Convention

To address the risks associated with the deliberate misuse of life sciences knowledge, technology, and materials, the international community worked together to negotiate the Biological and Toxin Weapons Convention (BWC). This convention is a legally-binding treaty that prohibits the development, production, acquisition, or retention of

biological agents and toxins of both types and quantities that have no justification for prophylactic, protective, or peaceful purposes.

In theory, this treaty covers the risks posed by pathogen genomic data and would address any digital transfers that aid in the development of biological weapons. However, the BWC suffers from several major weaknesses.

First, the BWC does not target specific activities. Proponents of the BWC argue that the dual-use nature of the life sciences make it unfeasible to ban specific activities: the same activities that can be used for producing medical countermeasures and treatments can also be used to create biological weapons. However, without a ban on specific activities, state parties that adhere to the BWC must determine violations by determining the *intent* of actors.

Second, the BWC lacks effective verification procedures to ensure that member states are meeting their obligations. Many would dispute the possibility for viable verification measures given the unique challenges posed by the life sciences. While the BWC provides for a mechanism to review its status and future direction every five years, the lack of verification measures hinders its effectiveness overall.

Third, as with every multilateral treaty, the BWC applies to the activities of non-state actors only indirectly. Member states are obligated to mitigate risks of individuals gaining access to dangerous microbes to cause harm, but the language is vague, and the treaty doesn't propose or require specific measures.

Finally, the review process for the BWC, which takes place every five years, is unable keep up with current advances in technology and the rate of information dissemination. Recent proposals to ensure the relevance of the BWC were not approved in the most recent Review Conference.

Australia Group

Created in response to Iraq's use of chemical weapons during the Iran-Iraq war, the Australia Group (AG) coordinates and implements export controls on materials, equipment and technology related to chemical and biological weapons.

Composed of 42 member states, the AG is an informal and voluntary arrangement with several objectives:

- Encourage member and non-member states to create national control laws and procedures in their respective countries;
- Create Common Control Lists (CCLs), which harmonize export control lists of precursors, equipment, agents, and organisms between states;
- Provide guidance to industry stakeholders on how to detect potential proliferation transactions;
- Facilitate information-sharing between member states;
- Conduct outreach to non-member states through expert meetings, consultations, and sharing information like CCLs.

The AG suffers from two major problems. First, the recommendations of the AG are not legally binding. This has created a patchwork landscape of governance where certain states have robust national control laws and procedures, while others have none. Malicious actors can exploit these gaps to obtain the desired material, equipment, and agents.

The AG also suffers from the same flaw as other governance mechanisms like the FSAP and the DURC policy: it relies upon pre-determined lists of pathogens and specific equipment. As technology continues to advance rapidly and becomes increasingly accessible on the

Internet, the relevance of export control lists and export controls erodes quickly. Rather than depend on gaining access to physical samples, malicious actors may instead order genomic data online, synthesize the microbe, and then grow the samples in a lab.

International Gene Synthesis Consortium

In 2009, a coalition of gene synthesis companies, called the International Gene Synthesis Consortium (IGSC), formed to address security issues related to gene sequencing and synthesis.[25] Without a framework for baseline oversight of the customers or the gene sequences, gene synthesis companies acknowledged the risk of inadvertently supplying the materials for biological weapons.

The IGSC developed a common protocol to screen both gene sequence orders and the customers who place them. The group also created its own set of compulsory guidelines called the Harmonized Screening Protocol to reduce the risk of gene sequences being misused.

IGSC companies agree to verify new customers (and even new contacts within an existing account) through a series of restricted party screening protocols. In addition, IGSC companies agree to limit their sales to legitimate companies, universities, and institutes. Companies agreed not to fulfill orders made by private persons.

IGSC companies also pledged to comply with the HHS Screening Framework Guidance. Gene sequences orders are screened against an internal IGSC database containing high-risk gene sequences drawn from domestic and international biosecurity lists, including the Australia Group's CCLs, and the joint USDA/HHS Select Agent List. If a gene sequence order matches a high-risk organism, IGSC companies agreed to follow-up with customers to document their intended use of the gene sequence and ensure proper import/export permits.

Currently, 80% of the commercial gene synthesis market are IGSC members. In addition, the academic and industry communities have offered their support to the IGSC by only purchasing sequences from IGSC members. While the IGSC helps address the misuse of synthetic gene sequences, membership remains voluntary and there are no repercussions for failing due diligence except expulsion from the group.

Human Genomic Data

The governance structure for human genomic data is weak both domestically and internationally. This is particularly troubling since human genomic data poses risks from all four subtypes risk highlighted in Table 1.

In the absence of international governance, the U.S. governs the risks associated with human genomic data through the Health Insurance Portability and Accountability Act (HIPAA) and the Genetic Information Nondiscrimination Act (GINA).

Health Insurance Portability and Accountability Act of 1996 (HIPAA)

Until 1996, there was no broadly-accepted framework for protecting health information of individuals across the health care industry. Rather, individual companies in the health care industry set their own standards. This resulted in a patchwork of policies and requirements that varied in their effectiveness.

In recent decades, the health care industry has transitioned from paper-based to electronic information systems and relies upon computers and the Internet to pay claims, answer eligibility questions, provide health

information, and other administrative and clinically-based functions.

To address the changing technological and health care industry landscapes, U.S. policymakers created the Health Insurance Portability and Accountability Act of 1996 (HIPAA), which requires the Secretary of HHS to develop regulations to protect both the privacy and security of certain health information.

To achieve its legislated obligations, HHS published two rules: the Privacy Rule and the Security Rule. The Privacy Rule establishes national standards for protecting certain health information, including individuals' medical records and other personal health information. It requires health plans, health care clearinghouses, and health care providers to protect the privacy of personal health information with appropriate safeguards. In addition, this rule sets limits and conditions on the uses and disclosures of personal health information without patient authorization. Finally, the rule gives patients certain rights over their health information, including the right to examine and obtain a copy of their health records.

The Security Rule established a national set of security standards for protecting certain health information that is stored or transferred in electronic form. The Security Rule puts the principles of the Privacy Rule into action by addressing the technical and non-technical safeguards that the health care industry must put in place to secure 18 unique identifiers of individuals collectively known as the protected health information (PHI).

While the HIPAA regulation provides privacy for individuals, the regulation does not currently consider genomic data as PHI. In the age of big data, it is becoming more difficult for researchers to keep their human subjects anonymous. This is particularly true when the genomic data of subjects is cross-referenced with other data sources, including genealogical, and geographic data.

However, HIPAA only applies to information that is handled and located in the U.S. If health data is siphoned out through international partnerships or collaborations, HIPAA regulations no longer apply, and privacy requirements are no longer enforceable. This is a big concern for the United States as countries like China create partnerships with U.S.-based institutions and siphon health records overseas.

Genetic Information Nondiscrimination Act (GINA)

GINA, adopted in 2008, protects individuals from genetic discrimination in health insurance and employment, and is enforced by the Department of Labor, the Department of Treasury and the Department of Health and Human Services. The statute defines genetic information as any information about an individual's genetic tests, genetic tests of family members, family history, or any requests for genetic testing. A genetic test is defined as an analysis of DNA, RNA, chromosomes, proteins or metabolites that detect genotypes, mutations, or chromosomal changes. Routine tests such as complete blood counts, cholesterol tests and liver-function tests are not protected.

GINA makes it illegal for health insurance companies to use an individual's genetic information to determine eligibility or health premiums, contributions or coverage. For example, health insurance companies are not allowed to consider family history or a genetic test rest as pre-existing condition or require individuals to undergo genetic testing. Similarly, GINA prohibits prospective employers from requesting or using genetic information to make decisions about hiring, firing, promotion,

salary or benefits. Moreover, employers must not treat an employee differently based on his or her genetic information.

Like HIPAA, GINA focuses primarily on issues of discrimination using human genomic data for health insurance and employment and does little to address the broader risks of genomic data. GINA, for example, does not apply to life insurance, disability insurance or long-term care insurance. The degree of protection varies from state to state, but GINA sets the minimum standard to be met in all states. Moreover, the legislation applies solely to the United States and does not address discrimination of U.S. citizens overseas.

Industrial Genomic Data

Like human genomic data, the governance structure for mitigating the risks of industrial genomic data at the international level is nearly non-existent. That said, the Coordinated Framework for the Regulation of Biotechnology and intellectual property laws offer some relevant measures for protecting citizens from the effects of products made with biological processes and balancing intellectual property rights with research and innovation.

<u>Coordinated Framework for the Regulation of Biotechnology</u>

Genetic engineering and manipulation capabilities rapidly increased in the 1970s and 1980s, leading to greater numbers of consumer products produced with biotechnology including food, medicine and pesticides. In 1986, the White House Office of Science and Technology Policy released the Coordinated Framework for the Regulation of Biotechnology to outline a comprehensive Federal regulatory policy to ensure the safety of biotechnology products.

The Coordinated Framework led to the proper allocation and coordination of oversight responsibilities for the safety of biotechnology products among three federal agencies: U.S. Environmental Protection Agency (EPA), the Food and Drug Administration (FDA), and the U.S. Department of Agriculture (USDA). Over the past 30 years, these oversight agencies have developed regulations and guidance documents to ensure the safety of biotechnology products. Some of the principles that guide these regulations for all three agencies are as follows:

- Biotechnology products have applications in many areas.
- Products are regulated based on their specific uses. Therefore, all products with the same use are subject to the same types of oversight.
- Each agency uses its existing authorities and regulations to ensure that biotechnology products are both safe and used for their intended purpose(s).
- Risk of a biotechnology product is determined by the characteristics of the product, the environment it will function in, and the application of the product.
- The risk of a biotechnology product should not be determined by the process used to make the product, but the use(s) of the biotechnology product.

Although the framework has been reviewed several times over its history, it continues to neglect the topic of genomic data. Under the current framework, biotechnology products are assessed by the uses of the product rather than the process of creating the product (when genomic data comes into play).

In addition, old rules and regulations require adjustments to address the influx of new players such as do-it-yourself biology (DIYBio) community laboratories, at-home and direct-to-consumer biotechnology developers, and crowdfunded biotechnology ventures. These new players blur existing legal distinctions between product sponsors, product

developers, and manufacturers. While the framework has held up well in the past, it fails to account for how technologies *converge:* interact with other disciplines and technologies.

Intellectual Property Laws

Even before the entire human genome was mapped, many stakeholders understood the enormous economic potential of genomic data. To tap into this potential, private and public companies filed patent requests for human genes and gene sequences. These gene patents allowed companies to have exclusive rights to a specific sequence of DNA. The holder of the patent could dictate how the gene could be used in both commercial and noncommercial settings for 20 years from the date of the patent. Therefore, patent holders could potentially leverage intellectual property laws to address cybersecurity issues such as unauthorized access, data tampering, as well as the theft of data and analytics.

To license a gene patent, companies submit a patent application to the U.S. Patent and Trademark Office. A patent officer examines the application to ensure that the requested gene sequence fulfills the standards for patentability. Based on the assessment, a patent license is issued to a successful patent applicant. Patent holders can use their patents to exclude others from making, using, or selling an invention. In the microbial genomic data setting, this allows companies to create proprietary therapeutics. Furthermore, these patents provided companies with legal recourse in the event of a violation.

APPENDIX C: NDU'S GENOMIC DATA WORKSHOP

The Center for the Study of Weapons of Mass Destruction at National Defense University (NDU) launched a multi-year study in 2016 to examine the impact of emerging and converging technologies on national security, the threat posed by WMD, and efforts to counter WMD.[26] As part of our analysis of the risks and opportunities of synthetic biology, we identified bioinformatics and genomic data, i.e., the digitization of biology, as a critical issue for further exploration.[27] Particularly in the WMD context, the convergence of synthetic biology tools with reliable, accessible genomic data could enable actors to pursue the development and use of biological weapons.

While policymakers have discussed privacy considerations at length, they have paid less attention to the biosecurity risks associated with genomic data. To address this gap, the WMD Center held a **Deep Dive Workshop entitled "The Age of Genomic Data" on 10 August 2017** to explore the impact of genomic data on the risks posed by synthetic biology. Workshop participants from government, academia, and industry considered how the increasing volume of genetic information interacts with the growing ability of a broad set of actors to tinker with DNA. In addition, workshop participants examined the ways in which bad actors might leverage genomic data as well as the informatics used to store, access, and manipulate genomic data. In the concluding session, workshop participants discussed the availability of governance tools for responding to this risk and the existing governance gaps for biosecurity.

The workshop was hosted under NDU's policy of non-attribution, in which remarks are not attributed to speakers or participants without their express permission. This appendix provides a brief summary of the proceedings and finding.

Mr. Chuck Lutes, Director of the WMD Center, opened the workshop by framing the discussion on genomic data and bioinformatics as part of WMD Center's broader approach for exploring the impact of emerging technologies. The remainder of the deep dive consisted of three panels: 1) Types of Pathogen Data; 2) Data Tools and Data Use and; 3) Data Governance. Each panel featured subject matter experts on the risks, opportunities, and governance challenges of genomic data.

Dr. Diane DiEuliis, Senior Research Fellow at the WMD Center, moderated the first panel on the types of genomic data. Dr. Tom Slezak, Lawrence Livermore National Laboratory, provided an overview of the current state of bioinformatics for pathogen data. Dr. Patrick Boyle, Ginkgo Bioworks, offered an industry perspective on the use of genomic data for producing biological products. Mr. Steve Mason, FBI WMD Directorate, highlighted the risks associated with big data on personal information including genomic and lifestyle data.

Mr. Charles Lutes, Director of the WMD Center, moderated the second panel on data tools and data use. Dr. MJ Rosovitz, National Biodefense Analysis and Countermeasures Center, discussed the utility of genomic data for next generation detection technologies. Dr. Corey Hudson, Sandia National Laboratory, provided an overview of cybersecurity vulnerabilities associated with the use, storage and transport of genomic data. Dr. Christina Ting, Sandia National Laboratory, spoke about encryption and the myth of anonymity with regards to genomic data. Dr. Anup Singh, Sandia National Laboratories, discussed nanotechnology, precision medicine and dual use issues.

The workshop concluded with a panel moderated by Dr. Kavita Berger, Gryphon Scientific, on data governance. Dr. Sarah Carter, Scientific Policy Consulting, LLC, provided an overview of relevant governance measures and assessed their effectiveness for addressing genomic data issues. Dr. Eleanor Celeste, National Institute for Standards and Technology, discussed policies that govern ethical considerations related to genomic data, health insurance, employment and other areas where discrimination might occur.

Throughout the deep dive, workshop participants considered the interactions between two factors: increasing volumes of genomic data and the growing ability of a broader set of actors to manipulate DNA. In addition to assessing the interactions between these two factors, workshop participants also discussed what governance tools are currently available for responding to this risk. Below are the key observations and insights from each part of the workshop:

Pathogen Genomic Data:

- **When used together, gene editing technologies and genomic data information can lower the bar for engineering new pathogens, as well engineering pathogens with customized capabilities.**
- **As researchers use nanotechnology and other disciplines to improve how drugs and treatments are delivered, the potential for the misuse of pathogen genomic data is likely to increase.** By using advances in therapy and precision medicine applications, malicious actors may be able to achieve greater targeting, penetration, or delivery of a harmful bioagent to a target.
- **The rapid growth of pathogen genomic datasets has complicated our current approach to pathogen detection technologies**. The increasing complexity of genomic data not only has implications for classical taxonomies, it can also undermine detection. As the data complexity grows, current detection technologies face increasing difficulty in distinguishing between the signal (a detectable pathogen of interest) from the noise (environmental microbes, false positives, and other impediments). This further complicates the detection of unknown threats.

Human Genomic Data

- **As understanding of the human genome advances, it may lower the bar to creation of "precision maladies", or engineered agents that target individuals or populations.**
- **Although privacy and biosecurity are distinct issues which require different solutions, both exist on a continuum of risk.** A breach of privacy could eventually lead to a security concern depending on the context.
- **To understand genomic data in the human context, other human health data is needed.** Data is much broader than genomics data. Other vital data is frequently contained within a person's health record, in self reporting, and blood samples. Other unstructured clinical data are captured through health and activity monitoring devices (e.g., Fitbits or other "wearables" that monitor physical attributes).
- **Researchers, corporations, and states are increasingly monitoring and sharing genomic data through Internet applications and cloud-based systems.** Interested parties can access this human genomic data, as well as other personally-identifiable information (PII) from any geographic location through such systems. To access this information, the user only

requires an Internet connection and a user account.

Industrial Genomic Data

- **There is no common understanding of biosecurity and biosafety risks in the growing bioeconomy.** A growing number of small startups and larger companies are pursuing synthetic biological manufacturing. These companies not only have their own incentives, but also vary widely in their use of genomic data, datasets, techniques for analyzing data, and their bioinformatics capabilities.
- **Many components of the bioindustrial manufacturing process can generate points of biosecurity risk.** Bio-manufacturing is a convergence of biology (in the use of genomic data for biological systems design), engineering, bioinformatics, automation and complex computation.
- **Minimal encryption or other safeguards are used at these risk points in the information life cycle.** While cybersecurity tools can be applied to mitigate some of the risks, such security comes at the expense of efficiency, remote controllability, and ease of use.
- **Corporate espionage takes many forms.** Beyond stealing proprietary data to gain competitive economic advantage, espionage can involve tampering with a company's data. By introducing error into a company's genomic data, production may be slowed, biosafety or biosecurity could be compromised, or research may be sabotaged.

Data Tools and Data Use

- **The relationship between genotype and phenotype is still poorly understood.** Researchers still struggle to understand how genetic information is expressed. Researchers are using advancements in bioinformatics and high performance computing to address this issue and, over time, hope to produce a comprehensive understanding of the complex relationships between genotype and phenotype.
- **Genomic data is large and fragmented. Therefore, researchers need to develop complex tools and analyses to understand genomic data better.** Actors, regardless of intent, cannot achieve specific goals by merely leveraging large volumes of raw data. Complex analyses are needed to find patterns in the raw data. Unfortunately, researchers run into challenges with analytics, including the underestimation of the variation in genomes from humans, bacteria, and viruses.
- **The ability to perform analysis on genomic metadata** (the descriptive data about single genomes, including information on the organism, isolate information, host information, sequence information, phenotype information, and other types of information) **relies on sophisticated computational bioinformatics and hardware**. Not all actors have this capability at this time.
- **Quantity does not always equal quality (at this time) for genomic data.** With increased accuracy and reduce costs, researchers are using gene sequencing more and more. However, researchers continue to experience errors when they sequence genes. These errors lead to incomplete and/or erroneous information associated with genomic data in publicly available resources such as the GeneBank. It is expected that better annotation will correct many of these errors in the future.
- **Data management is still challenging.** The amount of data generated, e.g., raw data and analyzed data, are becoming larger. Genomic data users must find ways to either accommodate the increasing input of data, or to narrow areas of interest to

minimize the amount of data that needs to be stored.

- **Data efficiency is also challenging**. Given the volume of genomic data produced, researchers are trying to find ways to make the process faster and more efficient. Some participants offered a solution to "take computing to the data vs. taking data to computing." Rather than moving massive amounts of data to the software, the efficiency of the process could be increased by taking the significantly smaller software package used to process the data directly into the database.
- **Data tampering is a problem.** The amount of sequenced information greatly exceeds the amount of data that can be carefully processed and annotated. This factor exacerbates fears that malicious actors can exploit the open nature of genomic databases to tamper with genomic data.
- **The available suite of cybersecurity tools has not been rigorously applied to safeguard genomic data.** There needs to be a balance between safeguarding genomic data and making genomic data useable and accessible. Participants noted that taking a more comprehensive approach to developing these cybersecurity tools could provide some immediate benefit in mitigating risk. However, participants warned that such measures could also impede innovation through the use of genomic data.

Governance Challenges

- **Protection of genomic data is based on an outdated system that focuses on controlling pathogens.** Currently, policymakers and security experts remain focused on limiting access to pathogens through programs like the Select Agent Program. Workshop participants noted that focusing exclusively on pathogens will achieve limited success in a world where malicious actors can order gene sequences through DNA sequencing companies.
- **Interim or long-term solutions are needed for gene synthesis screening**. While the gene synthesis industry has largely adopted voluntary screening of gene sequence orders, the decreasing cost of DNA synthesis will make stable voluntary screening costs even less financially attractive to the gene synthesis industry. For the gene synthesis industry, decreasing DNA synthesis costs translate to even smaller profit margins. Screening costs, which have stabilized over the years, but their cost relative to gene sequencing will put pressure on screening incentives at companies focused on the bottom line.
- **Human genomic data should be considered personally identifiable information (PII).** Participants noted that genomic data can be used to identify individuals. Since regulations on PII are meant to protect the identity of individuals, participants noted that current regulations should recognize this gap.
- **International laws or norms for the sharing of genomic data, or protection from its misuse, do not currently exist.** Participants noted that an international understanding of the promise and perils of genomic data is essential to promote good uses of genomic data while mitigating its malicious uses. Furthermore, participants highlighted other industries, such as the financial industry, that could be used to model norms for the genomic data space.

[1] See also Randall S. Murch et al, "Cyberbiosecurity: An Emerging New Discipline to Help Safeguard the Bioeconomy," *Frontiers in Bioengineering and Biotechnology,* Vol. 6, No. 39 (Apr 2018): 1-6. Available at https://www.frontiersin.org/articles/10.3389/fbioe.2018.00039/full

[2] Michael J. Selgelid, "Governance of Dual-Use Research: An Ethical Dilemma" *Bulletin of the World Health Organization*, Vol. 87 (2009): 720 - 723. Available at http://www.who.int/bulletin/volumes/87/9/08-051383/en/

[3] James Clapper, "Statement for the Record: Worldwide Threat Assessment of the US Intelligence Community," Testimony delivered to the Senate Armed Services Committee, 9 February 2016. Available at https://www.dni.gov/files/documents/SASC_Unclassified_2016_ATA_SFR_FINAL.pdf

[4] Laurie Garrett, "Biology's Brave New World," *Foreign Affairs* (Nov/Dec 2013). Available at https://www.foreignaffairs.com/articles/2013-10-15/biologys-brave-new-world

[5] Marc Goodman, *Future Crimes: Inside the Digital Underground and the Battle for Our Connected World,* New York: Anchor Books, 2015, 423.

[6] Rob Carlson, "Estimating the biotech sector's contribution to the US economy," *Nature* Vol. 34 (2016): 247-255. Available at https://www.nature.com/articles/nbt.3491

[7] Certain viruses use RNA.

[8] Andrew Pollack, "Traces of Terror: The Science; Scientists Create a Live Polio Virus," *New York Times,* 12 July 2002. Available at https://www.nytimes.com/2002/07/12/us/traces-of-terror-the-science-scientists-create-a-live-polio-virus.html

[9] Kai Kupferschmidt, "How Canadian researchers reconstituted an extinct poxvirus for $100,000 using mail-order DNA," *Science Magazine,* 6 July 2017. Available at http://www.sciencemag.org/news/2017/07/how-canadian-researchers-reconstituted-extinct-poxvirus-100000-using-mail-order-dna

[10] See for example Diane DiEuliis and Gigi Gronvall, "A Holistic Assessment of the Risks and Benefits of the Synthesis of Horsepox Virus," *mSphere* Vol. 3, No. 2 (Mar 2018), Available at http://msphere.asm.org/content/3/2/e00074-18; Diane DiEuliis, Kavita Berger, and Gigi Gronvall, "Biosecurity Implications for the Synthesis of Horsepox, an Orthopoxvirus," *Health Security,* Vol. 15, No. 6 (Nov/Dec 2017): 629-637; and Gregory Koblentz, "The De Novo Synthesis of Horsepox Virus: Implications for Biosecurity and Recommendations for Preventing the Reemergence of Smallpox," *Health Security* Vol. 15, No. 6 (Dec 2017): 620-628.

[11] Keith Allen, Jason Hanna, and Cheri Mossburg, "Police used free genealogy database to track Golden State Killer suspect, investigator says," *CNN.com,* 27 April 2018. Available at https://www.cnn.com/2018/04/26/us/golden-state-killer-dna-report/index.html

[12] GEDmatch Tools for DNA and Genealogy Research, https://www.gedmatch.com/login1.php

[13] Brittany Martin, "DNA Submitted for Genealogy Research Led Police to the Golden State Killer Suspect," *Los Angeles Magazine,* 27 April 2018, Available at http://www.lamag.com/citythinkblog/dna-golden-state-killer/

[14] Cyrus Farivar, "GEDmatch, a tiny DNA analysis firm, was key for Golden State Killer case," *ArsTechnica,* 27 April 2018, Available at https://arstechnica.com/tech-policy/2018/04/gedmatch-a-tiny-dna-analysis-firm-was-key-for-golden-state-killer-case/

[15] Gigi Kwik Gronvall, "US Competitiveness in Synthetic Biology," *Health Security,* Vol. 13, No. 6 (Dec 2015): 378–389. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4685481/

[16] Brenda Goodman, "Agents Arrest 3 in Plot to Sell Coca-Cola Secrets to PepsiCo," *New York Times,* 6 July 2006, Available at https://www.nytimes.com/2006/07/06/business/06coke.html; Erin Ailworth, "Chinese Firm Found Guilty of Stealing Wind Technology From U.S. Supplier," *Wall Street Journal,* 24 January 2018, Available at https://www.wsj.com/articles/chinese-firm-found-guilty-of-stealing-wind-technology-from-u-s-supplier-1516829326

[17] Stephens, Zachary D., Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, and Gene E. Robinson. "Big data: astronomical or genomical?" *PLoS biology* 13, no. 7 (2015): e1002195.

[18] National Center for Biotechnology Information (NCBI) https://www.ncbi.nlm.nih.gov

[19] Fabricio F. Costa, "Big Data in Genomics: Challenges and Solutions," *G.I.T Laboratory Journal,* Vol. 11/12 (2012): 2-4. Available at https://pdfs.semanticscholar.org/0e38/7bd00952c8450defcdbcdebf5c946c20f54.pdf

[20] Recently, scientists may also look into obtaining samples from the black market where market entrepreneurs purchase physical samples that cannot be used for their original purpose.

[21] See for example Appistry at www.appistry.com;

Genome International Corporation at www.genome.com.

[22] See Fabricio Costa, 4.

[23] *United States Government Policy for Oversight of Life Sciences Dual Use Research of Concern,* n.d. (March 29, 2012), 1, Available at https://www.phe.gov/s3/Documents/life-sci-dual-use.pdf

[24] Ibid, 2

[25] Diane DiEuliis, Sarah R. Carter, and Gigi Kwik Gronvall. "Options for Synthetic DNA Order Screening, Revisited." *mSphere* Vol. 2, No. 4 (2017), Available at http://msphere.asm.org/content/2/4/e00319-17

[26] Natasha E. Bajema and Diane DiEuliis, *Peril and Promise: Emerging Technologies and WMD,* Washington DC: NDU Press, May 2017, Available at http://wmdcenter.ndu.edu/Publications/Publication-View/Article/1181150/peril-and-promise-emerging-technologies-and-wmd/

[27] Diane DiEuliis, Chuck D. Lutes, and James Giordano, "Biodata Risks and Synthetic Biology: A Critical Juncture," *Journal of Bioterrorism and Biodefense,* Vol. 9, Issue 1 (April 2018), Available at http://wmdcenter.ndu.edu/Media/News/Article/1484193/biodata-risks-and-synthetic-biology-a-critical-juncture/