The Past, Present, and Future of Bioinformatics: An Analysis of the Field's Key Developments and Ethical Complications

Presented to the S. Daniel Abraham Honors Program

In Partial Fulfillment of the

Requirements for Completion of the Program

Stern College for Women

Yeshiva University

April 28, 2022

Tova Dorit Narrowe

Mentor: Dr. Anya Alayev, Biology

The complexity of biology is staggering. To properly study and understand the chemistry, molecular interactions, and effects of a singular protein or gene is a task that could fill several careers. The pace at which biological knowledge has developed in the last century is also astounding. One hundred years ago, scientists weren't even sure whether the genetic material of a cell was contained with its deoxyribonucleic acid (DNA) or its proteins. Now, one can reveal the nucleotide sequence of their entire genome within a few hours.

Advancements in biological knowledge have developed hand-in-hand with computer science. Experiments nowadays can not only occur *in vitro* or *in vivo*, but *in silico* as well. This term has been developed to refer to experiments that take place on a computer or in a simulation. Biologists also often now distinguish between procedures that are performed in the 'wet-lab' and the 'dry-lab.' The term 'wet-lab' refers to physical experiments that take place in the laboratory, while 'dry-lab' refers to computational experiments and/or the analysis that follows up the wet-lab work. Enormous biological databases are now publicly available through which scientists share exabytes worth of data. Computational work has become an essential part of biology, both in that it is intertwined in the processes through which we gather data, but also in how invaluable it has become for data analysis. Computation is changing the way scientific research is conducted, and, as such, biological knowledge has developed more in the last seventy years than it had in previous centuries.

The National Human Genome Research Institute defines bioinformatics to be "...a scientific subdiscipline that involves using computer technology to collect, store, analyze and disseminate biological data and information…." Essentially, it is the intersection of biology and computer science; the use of technology to answer biological questions. It is a field that is growing in prominence and popularity. Universities offer undergraduate, graduate, and doctoral degrees in

bioinformatics and many laboratories now hire individuals specifically to work as bioinformaticians. Computational skills have become a valued and essential part of the biologists' toolkit.

There are many subdisciplines to bioinformatics. Genomics, transcriptomics, and proteomics refer to the study of complete genomes, transcriptomes, and proteomes respectively. In these disciplines, vast amounts of data are evaluated in each study. The ability to study such large quantities of information provides valuable insights into molecular processes, protein interactions, and cell communications. It has allowed for the development of a whole new approach to biology; one where scientists no longer have to spend their entire career on understanding one gene or characterizing a single protein. Instead, now scientists can study systems biology, where they analyze the interactions and developments of entire cellular systems or biological communities. Another subdiscipline of bioinformatics is known as structural bioinformatics. This refers to the use of computational approximations and statistical analyses to examine and predict the structure of molecules. It can model potential interactions between proteins and has vast applications for the world of medicine and drug design. In other areas, bioinformatics focuses on the development of tools used to simplify wet-lab research or data evaluation, such as image analysis software that can be used to translate the hundreds of images generated in an experiment into readable results.

Bioinformatics as a field is crucial to the scientific future, but it is also facing many challenges as it develops. The community is currently grappling with essential questions such as how do we share data without compromising the privacy of the people involved in studies? How do we increase diversity within our research without exploiting underrepresented communities?

What technological improvements can be expected in the coming years and how will they effect the ways in which research is conducted?

This work focuses on these questions specifically as they apply to the world of genomics, but they are prevalent to all areas of bioinformatics. An analysis of some key historical developments and current prominent technologies provides insight into the directions in which bioinformatics may develop and the problems it needs to solve in the years to come.

The Beginnings of Bioinformatics

Computation was first incorporated into biology in order to help determine the amino acid sequences of proteins. In the 1950s, basic understanding of DNA was still being developed, but scientists had much stronger knowledge of the chemistry of proteins. The Edman degradation method was used to determine the order of amino acids in a protein chain. The process involved cleaving a single amino acid from the N-terminal end of the polypeptide chain and isolating it, after which it could be identified. This process could only work for polypeptides between the length of sixty to one hundred and fifty amino acids (Niall, 1973). Sequencing proteins of a larger size required that the protein first be broken down into smaller pieces, which were then sequenced individually. Then, after all the smaller sequences were determined, it was necessary to place the results into the correct order to obtain the original polypeptide sequence.

Determining this order was no simple task, and it was one that was better handled by computer programs. Margaret Dayhoff rose to the challenge of producing this software. She and Robert S. Ledley worked together to write a FORTRAN program that they called COMPROTEIN which could take the results of separate Edman runs and reassemble them into a single sequence (Gauthier et al., 2019). It identified overlaps between different fragments, also called ‘reads,’ to

identify a most likely structure (Dayhoff & Ledley, 1962). This was the first recorded occurrence of *de novo* sequencing, a term which refers to the determining the sequence only through reassembly of smaller fragments, without any known structure to reference. COMPROTEIN was used to sequence many proteins, and in 1965, Dayhoff released *Atlas of Protein Sequence and Structure,* which was the first biological sequence database. At the time of its publication, it contained sixty-five protein sequences (Gauthier et al., 2019). Additionally, Dayhoff developed the one letter amino acid code that is still in use today. She did this in order to minimize the size of the data that had to be inputted into her programs. Dayhoff came to be considered the first bioinformatician. Her contributions were so crucial that David Lipman, the former director of the National Center for Biotechnology, said that Dayhoff was "the mother and father of bioinformatics" (Gauthier et al., 2019). Margaret Dayhoff demonstrated the lengths to which computation could advance biological research, and, in that way, she laid the foundation for all of the bioinformatics work that was to come.

Sanger Sequencing

Bioinformatics started in the world of proteins, but it wasn't long before scientists set their sights on the much more computationally complicated field of genomics. It was time to figure out how to determine the nucleotide sequence of genetic samples. The first widely used process for reading the DNA of a sample was called 'Sanger sequencing' (Gauthier et al., 2019). Developed in 1977 by Frederick Sanger and his colleagues, this method was often also referred to as the chain termination method (Millipore Sigma, n.d.). The process involved performing a polymerase chain reaction (PCR) using labeled dideoxyribonucleotides (ddNTPs) in order to amplify the sample. These ddNTPs lack the 3`-OH group, which prevents any other nucleotides from being integrated

into the chain after them. As the sample is amplified, the ddNTPs are incorporated, which results in fragmented copies of the DNA with different lengths. These copies are then run through gel electrophoresis in order to sort them by size (Adams, 2008; Heather & Chain, 2016; Millipore Sigma, n.d.). Then, the sequence can be determined through analysis of the gel, with the shorter reads representing nucleotides at the beginning of the sample and longer reads revealing the nucleotides at the end. Originally, when Sanger sequencing had to be done manually, an X-ray image was taken of the gel. This process would show the differences between the labels on the ddNTPs, revealing the type of nucleotide that it had bound to, and allowing the scientists to see the genetic sequence (Heather & Chain, 2016; *The Dawn of DNA Sequencing*, 2021). While Sanger and his team had to then manually look through each gel to determine the genetic sequence, this process would later be automized. When a computer analyzes a Sanger sequence, it excites the ddNTPs with a laser, analyzes the wavelength of light emitted and records it as a nucleotide (Millipore Sigma, n.d.).

Sanger used his method to sequence the first genome of an organism; specifically, that of the bacteriophage ΦX174 (Sanger et al., 1977). This process was revolutionary and set the foundation for all sequencing methods to come. However, Sanger sequencing was rather limited in what it could accomplish. Specifically, the number of base pairs that could be determined in one round of Sanger sequencing was between two hundred and five hundred (Adams, 2008). Much like when sequencing proteins, determining longer nucleic acid sequences required cutting the sample into smaller pieces and sequencing those sections individually. The results then had to be reassembled into one pattern. Determining the order of these pieces was a very complex problem, even more so than reassembling Edman reads, particularly since the DNA had to be replicated

multiple times before being cut up to ensure that no portions of the sequence were lost in the PCR process.

While Sanger and his team had to do their analysis by hand, it wasn't long before the task was given to computers. In 1979, Roger Staden published software specifically to analyze Sanger reads. The 'Staden Package,' as it came to be known, handled the computationally complex components of assembling a genetic sequence, such as identifying overlaps between the copied sequences to compile 'contigs,' or contiguous segments of DNA. The software also annotated and maintained sequence files (Gauthier et al., 2019). The Staden Package vastly improved the efficiency of Sanger sequencing and set the stage for it to be used for significant genomic projects, such as the Human Genome Project.

The Human Genome Project

The developments of Sanger sequencing and the Staden Package meant that it was now time for scientists to set their sights on a once seemingly impossible task: it was time to overcover the sequence of the human genome. Significant genetic sequencing projects were already being undertaken. For example, in 1995 R. D. Fleischmann published the first genomic sequence of a free-living organism, *Haemophilus influenzae* (Fleischmann et al., 1995). However, uncovering the sequence of the human genome was a task on a different level. To this day, the Human Genome Project is considered to be the world's largest biological research project. It was an international undertaking, with scientists from all around the globe contributing their findings. The project was formally established in 1990 when its goal of sequencing the entire human genome was announced by a committee from the U.S. National Academy of Sciences. The project was declared complete

in 2003, and its findings, including the sequence of ninety nine percent of the euchromatic genome, were formally published in 2004 (National Human Genome Research Institute, 2018a).

A key principle of the Human Genome Project was the idea that the human genome should be considered public information and that scientists everywhere should have access to it. In order to ensure this, the Bermuda Principles were established in 1998. They declared that all contributors must adhere to the following rules: "automatic release of sequence assemblies >1 kb, preferably within 24 h; immediate publication of finished annotated sequences; and making the entire sequence freely available in the public domain for both research and development in order to maximize its benefits to society." (Barranco, 2021). Crucially, these principles fought the notion that any entity could have the right to patent any part of the human genome. This was essential, because private companies were also hard at work to determine the sequence. Most notably, the company Celera Genomics wanted to complete the task before the Human Genome Project did and intended to require scientists to purchase subscriptions in order to access their data (Barranco, 2021). Eventually a deal was brokered between the company and the Human Genome Project, after which they shared data and simultaneously published their results in 2001.

The two entities approached their shared task in different ways. The Human Genome Project used the 'hierarchical shotgun sequencing' method. In this approach, the genome is divided into pieces of roughly one hundred and fifty kilobases in length, and those pieces are kept in order. Then, one piece at a time, each fragment is broken down into smaller sections for sequencing, which are then reassembled. As such, as each section was reassembled, its rough location on the chromosomes was known, since the original order of the pieces was maintained. Conversely, Celera Genomics used the 'whole genome shotgun sequencing' method. This required breaking down the entire genome into pieces roughly five hundred base pairs in length. The pieces were

then all sequenced from both ends of the genome, generating ‘mate pair’ fragments. All the pieces then had to be reassembled into one sequence (Barranco, 2021; Gauthier et al., 2019). The whole genome shotgun sequencing method was simpler than the hierarchical shotgun sequencing method in that its wet-lab processes contained less steps, but it presented a significantly more difficult computational challenge when it came to reassembling the fragments.

The Human Genome Project took thirteen years from start to finish and became the foundation for genetic research in the twenty first century. The project took so long to complete not only due to the enormity of the task of sequencing three billion base pairs, but also because of the limitations imposed by Sanger sequencing. In a single run, a Sanger sequencer could analyze ninety-six reads containing eight hundred base pairs each. Therefore, forty thousand runs were required to cover the human genome a single time, and the complete genome had to be sequenced multiple times over to account for any potential mistakes (Gauthier et al., 2019). Sanger sequencing machines were also expensive, and this was a massive, publicly funded project. At the time of its completion, the Human Genome Project had cost around 2.7 billion U.S. dollars.

The conclusion of this project didn’t mean that the entire human genome had now been uncovered. It only focused on the euchromatic regions of the genome, as those areas are less condensed and are associated with active gene expression. Important heterochromatic regions were not researched, meaning eight percent of the human genome had not been sequenced. The Telomere-to-Telomere Consortium was a public group dedicated to bridging this gap. They recently published their findings, presenting “gapless assembly for all chromosomes except Y” (Sergey et al., 2022). This involved sequencing and mapping nearly two hundred million more genes and led to the identification of nearly two thousand more gene predictions, ninety-nine of which are believed to code for proteins (Sergey et al., 2022).

The impact of the Human Genome Project was wide reaching and is significant to this day. Most obviously, the development of a reference human genome was a crucial foundation for all of human genomics. It gave scientists the background they needed to further understand cellular activity, and it continues to have critical applications for medicine, both in comprehending disease processes and for the development of treatments. Additionally, the project led to the sequencing of other organisms' genomes. Scientists were concurrently studying the genomes of bacteria and several multicellular organisms, including *Escherichia coli*, *Caenorhabditis elegans*, *Drosophila melanogaster*, mice and rats (National Human Genome Research Institute, 2018b). These are all organisms that are often used in laboratory experiments, and as such, sequencing their genomes provided a significant advantage to scientific research. The sequencing of the human genome and the development of genetic maps simplified the process of sequencing other organisms, as large portions of all genomes are homologous. These accomplishments were all significant; however, one often overlooked yet crucial impact of the Human Genome Project was how it established a new way to conduct research. It set standards for data sharing that still benefit the field and it paved the way for large consortium-based research projects (Green et al., 2015). The Human Genome Project not only broadened humanity's knowledge of genetics, it demonstrated the heights that can be achieved through collaboration.

Next-Generation Sequencing

Sanger sequencing was revolutionary but rather limited. It could only process a small number of base pairs at a time, which made sequencing a rather slow process. Multiple rounds of sequencing had to happen before genome assembly could begin. It was also expensive, and its

limitations effected the speed and cost of the Human Genome Project. It was only a matter of time until a more efficient method was introduced.

The genetic sequencing methods introduced after Sanger sequencing came to be collectively known as next-generation sequencing (NGS). Often scientists distinguish between second-generation sequencing, the methods that immediately followed Sanger sequencing, and third-generation sequencing, which came later. These methods are also often referred to as massively parallel sequencing or high-throughput sequencing (Alekseyev et al., 2018). As the names imply, the main improvement of the newer sequencing methods was the parallelization of the sequencing. Unlike Sanger sequencing, second-generation sequencing does not require the physical separation of reactions, only their spatial separation (Alekseyev et al., 2018). Millions of DNA fragments can be loaded onto beads for simultaneous amplification, drastically decreasing the amount of time scientists need to wait for the sequencing reactions to complete. NGS also uses a different process to determine the type of nucleotide at each position than Sanger sequencing did. Instead of labeling the ddNTPs that were incorporated into the synthesized sequences, second-generation sequencing recorded the amounts of pyrophosphate released as each nucleic acid was added to the chain. This was enough to identify the type of nucleotide, which was then recorded by the machine (Heather & Chain, 2016). A single bead can produce a read of four hundred to five hundred base pairs in length, and the new machines were able to analyze millions of beads at once. Therefore, second-generation sequencing could sequence a genome dramatically faster than Sanger sequencing could. While the Human Genome project took thirteen years, NGS allowed for the comparable sequencing of a human genome in two months at a way lower price (Wheeler et al., 2008).

The first company to sell second-generation sequencing machines was 454 Life Sciences, but many other companies quickly produced their own NGS technology. Most notably, Illumina released their NGS technology two years after 454 Life Sciences, causing prices to drop even lower. The Human Genome Project cost over two billion U.S. dollars. By 2017, sequencing a human genome cost less than one thousand U.S. dollars (Alekseyev et al., 2018). This reduction in price had immense implications for the field of bioinformatics. Now genetic sequencing was not limited only to scientists in the most well-funded and prestigious labs. Sequencing was no longer a luxury, but another useful research tool.

Nowadays, sequencing is so widespread that scientists face a new problem: standardization. Multiple companies produce NGS technology, and many different tools and services for analysis are available to the public. New developments are made and released rapidly, occasionally leaving older technologies obsolete within years of their development. These differences in data production and analysis could inhibit scientists from sharing their results with one another. However, this is a very difficult problem to address. It is challenging to convince scientists to invest the necessary time and resources into learning new tools and systems when they are already trained in suitable alternatives (Gauthier et al., 2019). Instead, the solution may lie within the creation of standard data analytical measures and file formats. Authorities in the field, whether they be prominent research agencies or journals, should insist that scientists whom they employ, work with, and/or publish adhere to certain data standards. Doing so will hopefully motivate leading companies to adhere to these guidelines within their products and therefore circumvent any future complications that could arise from the lack of standardization across sequencing results.

Genome Assembly

Whether one uses Sanger sequencing or NGS, computational help is necessary to obtain the final genetic sequence from the wet-lab results. The individual nucleotides are not determined sequentially from an intact genome. Rather, accurate sequencing requires that the genome be replicated multiple times and then broken down into smaller fragments. Even after sequencing has been completed in the wet-lab, significant computational work has to be performed before the genome can be understood.

The computational challenge of genome assembly involves aligning the different DNA reads so that they form a plausible sequence. Reads are also often referred to as 'k-mers,' where the k represents the length of the sequence. A round of sequencing returns a set of k-mers that ideally would consist of all k length subsequences of the original sample. The software now has to identify the order of these k-mers. This is no trivial task, as sequencing often returns thousands or millions of reads.

This reassembly is accomplished through the use of graph theory. One specific type of graph, called a de Bruijn graph, is particularly useful. A de Bruijn graph is formed by analyzing the 'prefixes' and 'suffixes' of a set of k-mers. A k-mer's prefix is a (k-1)-mer that is formed by removing its final nucleotide from the sequence, while its suffix is formed by removing its first nucleotide. When one k-mer's suffix equals a different k-mer's prefix, those two k-mers are said to have an overlap and it is possible that they follow one another in the final sequence. For example, the 5-mers "AAGCA" and "AGCAT" contain an overlap because they share the 4-mer "AGCA." Combining these two 5-mers would produce the 6-mer "AAGCAT." A de Bruijn graph displays all overlaps within a set of k-mers. Each node in the de Bruijn graph represents a prefix/suffix contained within the set of k-mers. The in-edges for the node represent all the k-mers

that have that pattern as a suffix, while the out-edges represent all the k-mers that have it as a prefix. In order to reconstruct a genome from a de Bruijn graph, one has to find a path through the graph that visits every edge exactly once. This is called the Eulerian Path (Compeau & Pevzner, 2014).

It is possible to determine the Eulerian Path through a de Bruijn graph in linear time. First, it must be determined whether such a path exists. Euler's Theorem states that a graph contains an Eulerian Cycle if and only if it is strongly connected and every node has an equal amount of in-edges and out-edges. It contains an Eulerian Path if only two nodes don't have an equal amount of in and out-edges, meaning that an Eulerian Cycle could be formed through the addition of a single edge (Compeau & Pevzner, 2014). After determining that the graph is Eulerian, the algorithm starts at the node that has more out-edges than in-edges. The algorithm chooses an edge to follow, adds the k-mer it represents to the final sequence, and continues this process until it reaches the node with more in-edges than out-edges. If none of the nodes that were visited have any unvisited edges left, then the path has been found. If they do, then starting from that node, the same algorithm is applied until it has arrived back at that node. The newly generated cycle is then incorporated into the original path. The same process continues until every edge in the graph has been used exactly once.

It is possible that a de Bruijn graph would have more than one Eulerian Path through it. If two separate cycles start from the same node, how is the scientist to know which one comes first in the genome? One solution is to use longer reads. A (k+1)-mer could reveal which path should be taken first. When longer k-mers are produced from sequencing, the de Bruijn graph becomes less tangled and the reassembled genome is easier to derive. Another solution would be to produce paired reads during sequencing. Paired reads are two k-mers which are known to be connected by

a certain number of unknown base pairs. Then, when determining the order of the cycles, the scientists considers which option would keep the paired reads the appropriate distance apart.

There are many other factors that complicate genome assembly. One such complication comes from the fact that genetic information can be lost during sequencing. Sequencing does not always have perfect coverage, meaning it does not generate all possible k-mers of the sample sequence. Researchers can address this by dividing their reads into smaller sizes. There may not be perfect coverage in the 100-mers that were produced in sequencing, but, if those 100-mers are broken down into 20-mers, then perhaps the whole sequence will be covered. This process is called read breaking and is used by many modern assembly programs (Compeau & Pevzner, 2014). However, read breaking is not always enough to generate k-mers that cover the whole genome, and scientists often have to settle for compiling contigs that represent larger portions of a genome rather than the entire genomic sequence.

Genome assembly also struggles to deal with tandem repeats within genetics samples. The genetic material is replicated many times before sequencing, so, when repeated reads are observed, it is difficult to identify whether the k-mer is in fact repeated within the genome. This is unfortunate, as these repeats are common in genetics, with at least a third of all human protein sequences containing tandem repeats (Tørresen et al., 2019). Additionally, sequencers often misidentify nucleotides and introduce erroneous reads into the k-mer set. When these errors are incorporated into the de Bruijn graph alongside the correct reads, it causes “bubbles” to be observed in the graph. The assembler is supposed to identify and resolve bubbles, but they do not always remove the incorrect read, and errors are included in the final genome sequence. Inexact repeats in the DNA can also introduce bubbles into the graph. An inexact repeat is when there is a segment within a genetic sequence that differs from another portion by only one nucleotide. To

the assembly program, this can appear like an erroneous read. Occasionally, the final genomic sequence that is produced is shorter than the actual sequence was because inexact repeats are removed by the assembler (Compeau & Pevzner, 2014).

One final complication that must be addressed in genome assembly is the fact that the reads produced in sequencing come from both strands of the DNA. Both strands are replicated and sequenced simultaneously, so it isn't possible for the assembler to know which k-mers belong to the same strands. Therefore, before assembly, the reverse compliment of each read is added to the k-mer set. Ideally, this would produce a de Bruijn graph with two connected components which have identical Eulerian Paths, with each component corresponding to one DNA strand. However, many genomes contain subsequences that are the reverse compliment of one another, and therefore the de Bruijn graphs of the two strands end up being interconnected. Clearly, genome assembly is a tricky and essential part of genetic sequencing and one that, while widely in use, still requires refinement.

Nanopore Sequencing

The computational complexity and limitations of genetic assembly prompt the obvious question: is it possible to sequence the genome in such a way that reassembly is minimized or even not required at all? The answer has come in the form of a new type of sequencing, often referred to as nanopore sequencing. Both Sanger sequencing and NGS are forms of 'sequencing-by-synthesis,' meaning that copies of the DNA must be generated in order to determine its sequence. Conversely, nanopore sequencing can examine genetic material as it is, without the need for synthesis. Nanopore sequencing is accomplished by passing intact DNA strands through a single pore in a membrane. The ion current flow is measured as the sample passes through the pore.

Changes in the current can be used to identify the type of nucleotide that has just been passed through. As such, the nucleic acid sequence can be produced in order, with no need to reassemble fragmented reads. Each read can also be significantly longer than any reads obtained from sequencing-by-synthesis (Jain et al., 2016; Lu et al., 2016).

Pacific Biosciences (PacBio) was the first company to successfully release nanopore sequencing technology. Their devices are capable of sequencing roughly ten kilobases in a single read, although they do have a higher error rate than sequencing-by-synthesis methods. Sequencing-by-synthesis has an error rate of less than two percent, while comparatively, at the time of its release, PacBio's synthesis technology had an error rate of ten to fifteen percent (Lu et al., 2016). However, since the PacBio technology can generate reads very quickly and efficiently, it is very simple to run samples multiple times. Doing so allows scientists to build a consensus sequence based on their multiple results; this process is often called circular consensus sequencing (Lu et al., 2016).

PacBio wasn't the only company to see the benefits of producing newer sequencing technologies. In 2014, Oxford Nanopore Technologies (ONT) produced their own nanopore sequencing device, called MinION. It is a portable device capable of being plugged into a standard computer using a USB port. It is approximately ten by three by two centimeters in size and weighs roughly ninety grams (Jain et al., 2016; Lu et al., 2016). The MinION device allows for both single stranded and double stranded sequencing, which are referred to as one-dimensional and two-dimensional sequencing respectively. Two-dimensional sequencing usually results in reads of higher accuracy. Samples require minor preparation before being added to the MinION device. Specifically, adapters are added to both ends of the DNA sample, which then attach to the membrane and guide the sample to the pore. If two-dimensional sequencing is being performed,

then a hairpin adapter is also required to covalently attach the two strands so that they can both be sequenced. Each device contains five hundred and twelve pores and can sequence reads of fifty kilobases at ninety two percent accuracy (Jain et al., 2016). The computer to which the MinION is plugged in most be running the specialized MinKNOW software, which is responsible for data acquisition, processing and real time analysis (Lu et al., 2016).

MinION is ideal to be used for widespread genetic sequencing for many reasons. Its portability, simple sample preparation process, and basic hardware requirements make it easily accessible to any hospital, university, lab, or even private individuals who are interested in genetic sequencing. The simplicity of genetic sequencing that MinION provides makes it ideal for multiple health applications. For example, MinION can be used on prenatal samples to detect aneuploidy in a fetus in less than four hours (Wei & Williams, 2016). It can also be a vastly important tool for understanding and managing diseases. In fact, the MinION device has been utilized in many recent disease outbreaks, including the Ebola outbreak in 2016 and the SARS-CoV-2 pandemic. When the Ebola virus was spreading and causing mass deaths in West Africa, MinION was used on-site to study the virus. Sequencing could be accomplished in under an hour once the samples were loaded onto MinION devices, meaning that there was less than a day between sample collection and result production (Quick et al., 2016). This allowed scientists to study the evolution of the virus in real time, as well as its response to potential vaccines and treatments. The ability to study a virus on-site not only decreases the time necessary to get results, it also reduces the likelihood of mishandling of samples between collection and testing sites that can cause inaccurate results. These same benefits made MinION a useful tool that was utilized in the recent coronavirus pandemic as well (Bull et al., 2020).

Nanopore sequencing, and specifically the MinION device, will continue to push the field of bioinformatics forward. Accessibility and simplicity of sequencing technology means that more and more researchers and scientists are able to perform their own sequencing. However, nanopore sequencing has not yet been perfected, and there are many problems that still need to be addressed. For example, nanopore sequencing, while mostly accurate, is still relatively inaccurate when compared to NGS. This inaccuracy can currently be mostly overcome through repetitive sequencing but improving the initial accuracy of nanopore sequencing would eliminate extra steps and further simplify the process. Nanopore sequencing may be capable of sequencing very long reads, but developing sequencing technology for even longer segments of DNA would be extremely useful. If an entire chromosome could be sequenced in a single read, or even through only a couple of reads, it could eliminate the need for complex genome reassembly. This would vastly simplify the software necessary for genetic sequencing and reduce the likelihood of reassembly related errors. Genetic sequencing will continue to be the basis of bioinformatics, and, as such, further improvements in sequencing technology will advance scientific research immeasurably.

Diversity of Genome-Wide Association Studies

Advances in genetic sequencing technology have made the analysis of an individual's genome a commonplace procedure. Now scientists often perform genome-wide association studies. This is when an individual's entire genome is sequenced and analyzed to correlate their phenotypic traits with their genetics. They are a powerful research tool that most crucially provides critical insight into the genetic basis for disease. Doctors use genome-wide association studies to identify single nucleotide polymorphisms (SNPs) within patients that could be the cause of their

medical conditions. Entire databases are dedicated to the results of these studies and are referenced to determine whether the identified SNPs within a patient have ever been correlated with the observed phenotype. Additionally, they can help with preventative medical care. If a doctor observes that a patient has a mutation that is associated with a disease that the patient has not yet developed, they can recommend lifestyle changes and preventative treatments that will hopefully inhibit the disease from ever presenting. Genome-wide association studies are also used to prompt research into potential treatments and drug development.

Despite their vitality to the medical world, genome-wide association study databases have one critical weakness: they lack diversity. The vast majority of individuals who have participated in genome-wide association studies are of European decent. This has been a recognized problem in the field for many years, yet diversity has not improved. In 2009, ninety six percent of all genomes included in genome-wide association study databases came from people of European descent. Seven years later, that amount had only changed to eighty percent. This change mainly came from increased studies of people of Asian descent. The amount of data from African, Hispanic, Latin American, and indigenous peoples had barely changed (Popejoy & Fullerton, 2016). As of 2019, seventy eight percent of people represented in these studies were of white ancestry (Peterson et al., 2019).

The lack of diversity in this data has enormous consequences. Firstly, it contributes to and exacerbates already existing discrepancies in medical care between white people and people of color. One important purpose of performing widespread genome-wide association studies is to determine the diseases for which a patient is likely to be at risk. They can reveal which genetic markers doctors should be screening for. Without adequate quantities of data, significant conclusions cannot be drawn, and doctors remain unaware for which conditions their patients are

likely to have a genetic predisposition. People of color receive insufficient medical care due to shortage of relevant data and that is an injustice that needs to be corrected. Expanding the diversity in genomic research will also prompt studies into new treatment methods and could lead to new medicines and potential cures. It could also cause scientists to reexamine any previous incorrect conclusions that were drawn because the only data that was studied came from populations of European descent.

There are also nonmedical benefits to expanding the diversity of genetic datasets. Contained within the human genome are insights into not only biology, but human history as well. Evolution has left a record of our history in our DNA. Identifying shared traits between populations that are currently geographically separated can reveal ancient migration patterns. Preserved mutations can indicate details of the lifestyle of prehistoric populations, such as diet and disease exposures (Bentley et al., 2017). Diversity in genome-wide association studies will help fill in the blanks of our collective history and better our understanding of how the world as we know it came to be.

So how can this discrepancy be addressed? The National Institute of Health has sponsored multiple initiatives with the goal of increasing the diversity of genetic data. For example, the *All of Us* project is encouraging people to volunteer to contribute their genomes to be available for use in research, with the goal of acquiring one million samples from a largely diverse group. The NIH also funds the Human Pangenome Reference Program, which is aiming to replace the singular human reference genome that was generated by the Human Genome Project with a set of reference genomes that represent different communities. These initiatives might be a good start, but there are other ways to encourage diverse research. For example, studies that will focus on underrepresented populations can be given priority for grants and funding (Popejoy & Fullerton,

2016). Scientific journals also have the ability to potentially implement change. The desire to be published drives the academic research world, and thus journals hold a lot of power. Journals implementing diversity requirements or incentives could significantly improve the diversity of genetic datasets. Additionally, increased diversity among people who work in bioinformatics must be encouraged. People from low-income or underrepresented ethnicities will provide new perspectives and attention to issues that currently aren't being addressed. As the workforce becomes more diverse, so will the data.

It is important to avoid exploiting underrepresented communities in an attempt to increase the diversity in databases. Many minority populations have a history of being exploited by researchers and medical professionals in the name of pushing science forward. All efforts to solicit genetic samples must focus on how the individuals who are donating will benefit from their contributions. When possible, research should be conducted locally. Encouraging this may require funding new facilities and training programs around the world, like, for example, the Human Health and Hereditary in Africa (H3Africa) Initiative (Bentley et al., 2017). Decisions concerning these research initiatives, such as the scope of sample collection and the rights of researchers involved, should be left up to ethics committees comprised of members of the effected communities. Ethically increasing diversity will be a difficult task, but ultimately a doable and essential one.

Data Sharing and Privacy

Another one of the biggest issues currently facing the field of bioinformatics has to do with data privacy and sharing. Now that genetic sequencing can be accomplished relatively quickly and cheaply, it has become a standard and often-performed procedure. Doctors can have their

patients' genomes sequenced to try to identify the genetic basis for their conditions, and the rise of companies like 23andMe allow anyone to have their genes analyzed in order to learn more about their family history. The quantity of human genetic information available nowadays is staggering. And, as with any academic field, the question now being asked is how much of their data should researchers have to share?

The answer is simpler when discussing non-human genetic data. There, it is clear that the field as a whole benefits from the open sharing of information. First of all, sharing data saves researchers' time. When data is publicly shared, it prevents scientists from having to replicate experiments that have already been successful, just so they can have access to results that already exist. Open sharing of data also allows scientists to formulate more precise hypotheses that can drive their research further faster. Most prominent scientific journals recognize this and make data availability a requirement for publication. For example, the Editorial Policies of the journal *Nature* states that "A condition of publication in a *Nature Portfolio* journal is that authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications." (Nature, n.d.). This enables other scientists to replicate and verify their results and contributes to the amount of publicly available data.

The question of if data should be publicly accessible becomes much more complicated when discussing human genetic data. In that case, the desire to push science forward has to be balanced with the rights of the people from whom the data came. A person's genetic sequence is highly valuable and identifiable information. It can describe a person's family history and ancestry as well as their current medical status. It can even provide information about their current lifestyle. As such, this data must be well protected. If someone has their DNA sequenced by recommendation of their doctor to shed light on their condition, that data should of course not be

entered into a database for public use and studies. In fact, the Health Insurance Portability and Accountability Act (HIPAA) and the Genetic Information Nondiscrimination Act (GINA) ensure that genetic information is included in information that a doctor is not allowed to disclose (Clayton et al., 2019).

It is more difficult to protect genetic information when it is being used for research. Anyone's genome can be used to identify them, so it is impossible to protect the privacy of research participants without modifying the data. Attempts have been made to protect this data in the past, but simply removing identifying features from the sequences isn't enough. In 2013, Melissa Gymrek and her colleagues published a paper describing how you could use the supposably protected genomes to identify individuals. They realized that short tandem repeats (STRs) on the Y chromosome are specific enough to identify a family. They used a publicly available genealogy database to find a mostly likely surname based on the STRs, which they then combined with the state of residence and the age connected with the sample. This was enough data to identify the individual (Gymrek et al., 2013). Clearly the genetic information hadn't been protected enough. This study was enough to prompt the National Institute of Health, the maintainers of the database used, to add extra protections to who could access their databases (National Human Genome Research Institute, 2021).

Furthermore, Direct-to-Consumer (DTC) genetic tests are widely in use. These are private companies, such as 23andMe or Ancestry, who offer genetic sequencing and analysis to the public. They provide different services, such as identifying which genetic diseases a person may be a carrier for or finding distant relatives. There is some federal regulation of the quality of such tests; for example, the FDA reviews some medical DTC tests to investigate the accuracy of their results (U.S. Food and Drug Administration, 2019). However, there is a lack of oversight when it comes

to what these companies are allowed to do with the data. There are no federal laws preventing DTC companies from selling the data they obtain to other parties (National Human Genome Research Institute, 2021). There are also no federal laws against surreptitious DNA testing. Surreptitious DNA testing is when samples of someone's DNA are acquired and sequenced without their knowledge. Some individual states have regulations against surreptitious DNA testing that vary in their specificity and severity (National Human Genome Research Institute, 2021; Strand, 2016).

Moving forward, in order to protect the privacy of everyone, both people who consent to join studies and those who haven't, decisive steps have to be taken. Federal legislation is required to prevent the exploitation of people's DNA. There needs to be laws that criminalize surreptitious DNA testing and prevent DTC companies from sharing data with third parties without the informed consent of the customer. In research settings, there have to be procedures and protections in place that remind scientists that they are not the owners of this genetic data. Anyone who has their genome sequenced should have the sole rights to that data and the ability to evaluate requests from researchers who wish to use their genome in a study. Researchers should be required to fully explain the intent of their studies and the methods that will be used, as well as informing the people from whom they are requesting data where that information could end up. This will be difficult to manage, as replicability is key in validating scientific research and therefore requires access to the same resources that the original scientist used. Perhaps part of agreeing to allow a researcher to use your genome would be allowing for it to be used by other researchers for the sake of validating results. Participants would have to be notified when their information is being passed along and be given the right to sue any parties who do not respect their wishes. Such regulation might be

difficult to implement and oversee, but doing so would ensure the ethicality of future research, thereby enhancing the field of bioinformatics.

Conclusion

Bioinformatics may not have existed a century ago, but today it is a thriving field at the forefront of science. Advancements in genomics have revolutionized our understanding of humanity and the world around us. Next-generation sequencing and nanopore sequencing technologies have made deciphering genetic content a commonplace procedure, prompting many ethical debates alongside the progression of knowledge. Issues like the lack of diversity and concerns for privacy expand beyond only genomics; they effect all realms of bioinformatics. The recommendations listed in this work should not be limited to the context of genomics in which they were discussed. They are applicable to all of bioinformatics and many other scientific fields as well.

Technological advancements are always exciting and prompt scientific research to reach beyond its previous limitations, but they must always be considered hand-in-hand with the ethical questions they raise. Researchers do not only have a responsibility to further scientific knowledge. First and foremost, they have a duty to their fellow human beings and are obligated to place their wellbeing above research. We delve into attempting to understand the world around us with the ultimate goal of improving people's lives. Therefore, taking decisive action to address the issues that have arisen is imperative.

Additionally, it is important to acknowledge that addressing ethical dilemmas within the scientific world is a task that will never be fully complete. No solution will ever be perfect, and it is essential that the consequences of all further actions continue to be considered and addressed.

Technology will continue to advance, and it is important that our ethics develop at the same pace. Only in this way will bioinformatics fulfill its potential and help usher humanity into a better future.

References

Adams, J. U. (2008). DNA Sequencing Technologies. *Nature Education*.

Alekseyev, Y. O., Fazeli, R., Yang, S., Basran, R., Maher, T., Miller, N. S., & Remick, D. (2018). A Next-Generation Sequencing Primer-How Does It Work and What Can It Do? *Academic Pathology*, *5*, 2374289518766521–2374289518766521. https://doi.org/10.1177/2374289518766521

Barranco, C. (2021). *The Human Genome Project*. Nature Milestones: Genomic Sequencing. https://www.nature.com/articles/d42859-020-00101-9

Bentley, A. R., Callier, S., & Rotimi, C. N. (2017). Diversity and inclusion in genomic research: why the uneven progress? *Journal of Community Genetics*, *8*(4), 255–266. https://doi.org/10.1007/s12687-017-0316-6

Bull, R. A., Adikari, T. N., Ferguson, J. M., Hammond, J. M., Stevanovski, I., Beukers, A. G., Naing, Z., Yeang, M., Verich, A., Gamaarachchi, H., Kim, K. W., Luciani, F., Stelzer-Braid, S., Eden, J.-S., Rawlinson, W. D., van Hal, S. J., & Deveson, I. W. (2020). Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nature Communications*, *11*(1), 6272. https://doi.org/10.1038/s41467-020-20075-6

Clayton, E. W., Evans, B. J., Hazel, J. W., & Rothstein, M. A. (2019). The law of genetic privacy: applications, implications, and limitations. *Journal of Law and the Biosciences*, *6*(1), 1–36. https://doi.org/10.1093/jlb/lsz007

Compeau, P., & Pevzner, P. (2014). *Bioinformatics Algorithms: An Active Learning Approach*. Active Learning Publishers.

Dayhoff, M. O., & Ledley, R. S. (1962). Comprotein: A Computer Program to Aid Primary Protein Structure Determination. *Proceedings of the December 4-6, 1962, Fall Joint Computer Conference*, 262–274. https://doi.org/10.1145/1461518.1461546

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., & Merrick, J. M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (New York, N.Y.)*, *269*(5223), 496–512. https://doi.org/10.1126/science.7542800

Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2019). A brief history of bioinformatics. *Briefings in Bioinformatics*, *20*(6), 1981–1996. https://doi.org/10.1093/bib/bby063

Green, E. D., Watson, J. D., & Collins, F. S. (2015). Human Genome Project: Twenty-five years of big biology. *Nature*, *526*(7571), 29–31. https://doi.org/10.1038/526029a

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*(1), 239. https://doi.org/10.1186/s13059-016-1103-0

Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics*, *14*(5), 265–279. https://doi.org/https://doi.org/10.1016/j.gpb.2016.05.004

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., & Erlich, Y. (2013). Identifying Personal Genomes by Surname Inference. *Science*, *339*(6117), 321–324. https://doi.org/10.1126/science.1229566

Millipore Sigma. (n.d.). *Sanger Sequencing Steps & Method*. Retrieved April 23, 2022, from https://www.sigmaaldrich.com/US/en/technical-documents/protocol/genomics/sequencing/sanger-sequencing

National Human Genome Research Institute. (2018a, October 28). *What is the Human Genome Project?* https://www.genome.gov/human-genome-project/What

National Human Genome Research Institute. (2018b, December 12). *Human Genome Project Results*. https://www.genome.gov/human-genome-project/results

National Human Genome Research Institute. (2021, April 27). *Privacy in Genomics.* https://www.genome.gov/about-genomics/policy-issues/Privacy

Nature. (n.d.). *Reporting standards and availability of data, materials, code and protocols.* Retrieved April 23, 2022, from https://www.nature.com/nature/editorial-policies/reporting-standards

Niall, H. D. (1973). Automated edman degradation: The protein sequenator. In *Methods in Enzymology* (Vol. 27, pp. 942–1010). Academic Press. https://doi.org/https://doi.org/10.1016/S0076-6879(73)27039-8

Peterson, R. E., Kuchenbaecker, K., Walters, R. K., Chen, C.-Y., Popejoy, A. B., Periyasamy, S., Lam, M., Iyegbe, C., Strawbridge, R. J., Brick, L., Carey, C. E., Martin, A. R., Meyers, J. L., Su, J., Chen, J., Edwards, A. C., Kalungi, A., Koen, N., Majara, L., … Duncan, L. E. (2019). Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell*, *179*(3), 589–603. https://doi.org/https://doi.org/10.1016/j.cell.2019.08.051

Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, *538*(7624), 161–164. https://doi.org/10.1038/538161a

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., Bore, J. A., Koundouno, R., Dudas, G., Mikhail, A., Ouédraogo, N., Afrough, B., Bah, A., Baum, J. H. J., Becker-Ziaja, B., Boettcher, J. P., Cabeza-Cabrerizo, M., Camino-Sánchez, Á., Carter, L. L., … Carroll, M. W. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature*, *530*(7589), 228–232. https://doi.org/10.1038/nature16996

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Sergey, N., Sergey, K., Arang, R., Mikko, R., V, B. A., Alla, M., R, V. M., Nicolas, A., Lev, U., Ariel, G., Sergey, A., J, H. S., Mark, D., A, L. G., Michael, A., E, A. S., Matthew, B., G, B. G., Y, B. S., … M, P. A. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53. https://doi.org/10.1126/science.abj6987

Strand, N. K. (2016). Shedding Privacy Along with our Genetic Material: What Constitutes Adequate Legal Protection against Surreptitious Genetic Testing? *AMA Journal of Ethics*.

The dawn of DNA sequencing. (2021, July 21). https://www.yourgenome.org/stories/the-dawn-of-dna-sequencing

Tørresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., Gruca, A., Grynberg, M., Kajava, A. v, Promponas, V. J., Anisimova, M., Jakobsen, K. S., & Linke, D. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research*, *47*(21), 10994–11006. https://doi.org/10.1093/nar/gkz841

U.S. Food and Drug Administration. (2019, December 20). *Direct-to-Consumer Tests*. https://www.fda.gov/medical-devices/in-vitro-diagnostics/direct-consumer-tests

Wei, S., & Williams, Z. (2016). Rapid Short-Read Sequencing and Aneuploidy Detection Using MinION Nanopore Technology. *Genetics, 202*(1), 37–44. https://doi.org/10.1534/genetics.115.182311

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G. T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X., Liu, Y., … Rothberg, J. M. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, *452*(7189), 872–876. https://doi.org/10.1038/nature06884