2024

Creation of a Digital Storage System for Genome Sequencing Metadata

Jacquelin W. Olexa
jo162493@umconnect.umt.edu

Creation of a Digital Storage System for Genome Sequencing Metadata

By: Jackie Olexa

In Collaboration With: Drs. Jeffrey Good and Travis Wheeler

Presented to:

University of Montana Conference on Undergraduate Research, 2024

Wildlife Biology Senior Thesis Committee, 2nd May 2024

Fulfills Requirements for:

Davidson Honors College to graduate with University Scholar Distinction from the University of Montana

Department of Wildlife Biology Senior Thesis to graduate with Wildlife Biology Honors from the University of Montana

Under the Supervision of:

Dr. Jeffrey Good, Good Lab PI, Department of Ecology and Evolutionary Biology, University of Montana

Abstract

As the field of computational genomics continues to expand in both potential and application, it is now more imperative than ever to ensure that massive genetic sequencing datasets are properly stored in an accessible manner. This project sought to establish a practical, user-friendly, secure system for a genomics research lab (the Good Lab; thegoodlab.org) at the University of Montana. A MySQL database and connected web application was ruled the best configuration to maximize utility and accessibility for the lab's researchers. Building the logical framework for the database, creating the server, and sourcing data occurred over several months. The dataset ranged from experimental details of sequencing (such as experiment dates, sequencing platform, and provider) to metadata of the samples (specific biological specimen information, molecular protocols). A combination of lab notebooks and a master Excel spreadsheet yielded over 3,500 individual biological sequencing samples that spanned terabytes of archived data. These data represent 10 years of lab sequencing efforts, with numerous examples of incomplete or non-standardized documentation. Once the database was seeded with these data, efforts transitioned to user functionality and the front end. One goal became the creation of a web application that allows efficient execution of basic functions (insertions, selective deletions, updates, and queries) for individuals without a MySQL background. However, due to such an interfaces' complexity, a temporary substitute in the form of a thorough backend users' guide was designed to allow for maximum usability of the system in the immediate future. Ultimately, the fundamental goal was accomplished: a clear, organized system for sequencing data was created with a structure and function that will permit many years of continued data collection and recall in a manner befitting the importance of the data being collected. Areas for future improvement and development for the stack were also identified.

Background

Arising in the 1990s and continuing to expand into the modern age, the next technological revolution has already brought about groundbreaking advancements such as the world wide web, bionic prosthetics, and artificial intelligence. Herein, it is the intersection of two undervalued advancements- high-throughput genome sequencing and digital data storage- that provide a valuable guidebook and, perhaps, cautionary tale regarding the future of computational genomics. To better elucidate the later methodologies and results, it is essential to first provide a framework of understanding around genomics, data storage, and their intersection.

To begin, what is a genome? As most people learned in biology class, all of life is based on its deoxyribonucleic acids, or "DNA" (with some viruses possessing a ribonucleic acid, "RNA" genome). The unique arrangements and patterns of DNA are what create the structures and molecules that create life. Ultimately, a genome is the full, organized collection of DNA base pairs that an organism carries in each of its cells (Giani 2020). Genomics is the study of the "structure, function, and inheritance" of genomes and how genomes can be used to learn more about individuals, populations, and species (Griffiths 2023). But for any meaningful genomic analysis to be performed, DNA sequencing must first be conducted. DNA sequencing is the process of determining the order of nucleotides within DNA fragments and delineating the fragments by source individual (NHGRI 2023). Sequencing has changed dramatically over the last several decades, and understanding the accelerating trajectory of sequencing technology is useful to understanding the changing landscape of genomics (Hutchison III 2007).

Since 1953 when Rosalind Franklin and Maurice Wilkins first observed the three-dimensional structure of DNA (famously analyzed by Watson and Crick), the study of DNA has advanced at an impressive clip (Heather and Chain 2016). Only 23 years later, the first complete

bacteriophage genome was sequenced, followed a year later by the first DNA genome (Fiers et. al. 1976, Sanger et. al. 1977). Over the next several decades, sequencing techniques advanced dramatically, culminating with the first whole genome synthesis of a free-living organism in 1995, that of bacteria *Haemophilus influenzae* Rd. (Fleischmann et. al. 1995). Following this genomic advancement, many more bacterial and several eukaryotic genomes were synthesized (NHGRI 2022). Then, in 2001, the first draft of the highly-publicized, whole human genome (and subsequent 2004 high-quality assembly) was sequenced by the Human Genome Project (International Human Genome Sequencing Consortium 2001, 2004). The Human Genome Project was the first successful genome sequencing of a mammal, proving to be a major step in genome sequencing. The project took an estimated 3 billion dollars and over a decade to complete (Hood and Rowen 2013).

Since the late 1990s and early 2000s, the industry has experienced a radiation event rivaling its subject material; by 2015, over 30,000 distinct genomes had been sequenced (Hug et. al. 2016). The cause of this industry growth has been an on-going technological leap in sequencing through the development of "second generation" and "third generation" methods (Slatko et al 2018). One of the first groups of advancing "second generation" sequencing techniques was "Sequencing by synthesis" (SBS), a technique derived from the earlier Sanger sequencing (Slatko et al 2018). The first marketed SBS method was pyrosequencing which utilized pyrophosphate (a nucleotide incorporation byproduct) to determine DNA chain base order (Margulies et. al. 2005). However, the most prevalent SBS technique is Illumina sequencing (based on Solexa and Lynx Therapeutics processes), which uses the process of amplified DNA fragments clustering along oligonucleotide fragments to allow parallel sequencing (Brenner et al 2000, Slatko et al 2018). The 2010s and beyond saw the rise of "third

generation" sequencing, namely Single Nucleotide Real-Time sequencing by Pacific Biosciences and Nanopore sequencing by Oxford Nanopore Technologies (Giani et al 2020, Braslavsky et. al. 2003, Church et. al. 1998). These processes allow for the parallel sequencing of longer reads, creating an even more efficient sequencing approach (Slatko et al 2018). With these new technologies, projects like the Earth BioGenome Project are arising, seeking to sequence an additional 1.5 million eukaryotic species over the next 10 years, backed by billions in funding ("Roadmap: Project Plan" 2022). However, the challenge with all second and third generation sequencing is that these parallelized methods generate millions of sequencing reads, resulting in extensive data storage requirements (Slatko et al 2018).

With the history of genome sequencing established, the very reasonable question is raised as to why genome sequencing matters. In response to these questions, numerous articles have been published regarding the value of genome sequencing. Perhaps most importantly to the populous, since the Human Genome Project's publication and due to sequencing advancements, human genome sequencing has declined to a cost of less than $1000 per individual, meaning that the applications of genome sequencing in medical treatments are becoming increasingly economically possible (Wetterstrand 2023). Among the areas highlighted for potential genome sequencing use include prenatal, newborn, and adult disease screening, targeted tumor therapy, and disease monitoring (McCormick and Calzone 2016). Such applications may allow for significantly improved preventive care, as well as more effective treatments of various diseases including many types of cancer (McCormick and Calzone 2016). Likewise, genome sequencing significantly contributed to the rapid and effective response to the SARS-CoV-2 (Covid-19) pandemic, as sequencing helped determine probable origins, diversity, and evolutionary trajectory and ultimately contributed significantly to the development of the Covid-19 vaccine

(Saravanan et. al. 2022). In parallel, genomics can tell researchers a great deal about the natural, non-human world as well. Access to a species' genomic code allows researchers greater insights into evolutionary histories, species diversity, hybridization, inbreeding potential, and a myriad of other valuable measurements useful to conservation efforts (Theissinger et. al. 2023). In fact, the ability to understand species is critical to defining populations and regions of conservation concern, as well as understanding how historical and current management practices may be impacting species (Cook et. al. 2023). Such efforts are exceptionally valuable in the face of growing climate concerns and limited funding to the field. Clearly, whether anthropocentric or not, genomics is a powerful tool that has the potential to improve lives and reshape humanity's understanding of the natural world.

However, there is a steep cost to genome sequencing. This cost comes from several sources: time investment, monetary investment, and data storage. Given the scope of this specific project, the remainder of the discussion will focus on non-human sequencing, as that is the type sequencing data presented herein. Firstly, preparing samples for sequencing can be a time-intensive process. From collecting samples from wild, captive, or museum specimens, to extracting the DNA, to cleaning the DNA of contaminants, segment repair and amplification, labeling samples with unique DNA "barcodes", pooling samples, shipping the samples to the sequencing facility, and then the formal sequencing itself, a complete preparation and sequencing can take several weeks' worth of time. While several of these steps can be done on many samples simultaneously, the investment of time into this process is not insignificant. Likewise, the financial cost is comprised of many sub-elements from paying staff to prepare the sequences, to the special kits needed to prepare the samples, and the sequencing cost itself. While the kits may only run about $30 to $50 for any given sample (in the instance of Illumina preparations),

consider that an ideal sample size is usually over 50 individuals and may stretch far higher (Thermo Fischer Scientific 2024). Likewise, while sequencing has reached a record low, down from $1000 per Megabase (one million base pairs) in 2004 to under $0.01 per Megabase in 2022, if the sample is eukaryotic this can still mean sequencing costs of upwards of $1000 per individual for high coverage methods (Wetterstrand 2023). Finally, the data storage cost for genomics can be substantial. While raw sequencing run FASTQ files may only take 1-10 Gigabytes per individual, projects often require tens or hundreds of individuals to gain key insights into variation between individuals and groups. Put into perspective, a 256 Gigabyte laptop (a relatively standard capacity in 2024) could contain hundreds of raw sequencing files. But potentially thousands of samples may be generated for various projects and comparisons. Additionally, some file types including genome assemblies can be terabytes in size, greatly exacerbating the issue of data storage for genomics. Ultimately, decades of sequencing and the necessities of many, much larger files for analysis means that much larger storage systems are required to handle the data complexity required for genomics. Clearly, not only is further monetary cost required to obtain means of storing such data, but storing tens or hundreds of genome sequences is simply challenging to achieve.

Given the immense cost and complexity of genome sequencing, it becomes readily apparent that ensuring sequencing data is never lost is vitally important. Importantly though, this goes beyond the raw sequencing data and encompasses the much broader collection of metadata. The metadata of a sequence is all the information surrounding its origins, preparations, functional sequencing, and later uses. The loss of any element of this metadata means the loss of valuable information regarding its' sample. For instance, if information on the source individual is lost, understanding sex-linked traits may be impossible. If how the sample was prepared is

lost, then it calls into question the quality of the sequence. The loss of any of this information may require resequencing or the removal of that sequence from study, resulting in unnecessary expenditure either way. As such, a system to consistently track and securely store genome sequencing metadata.

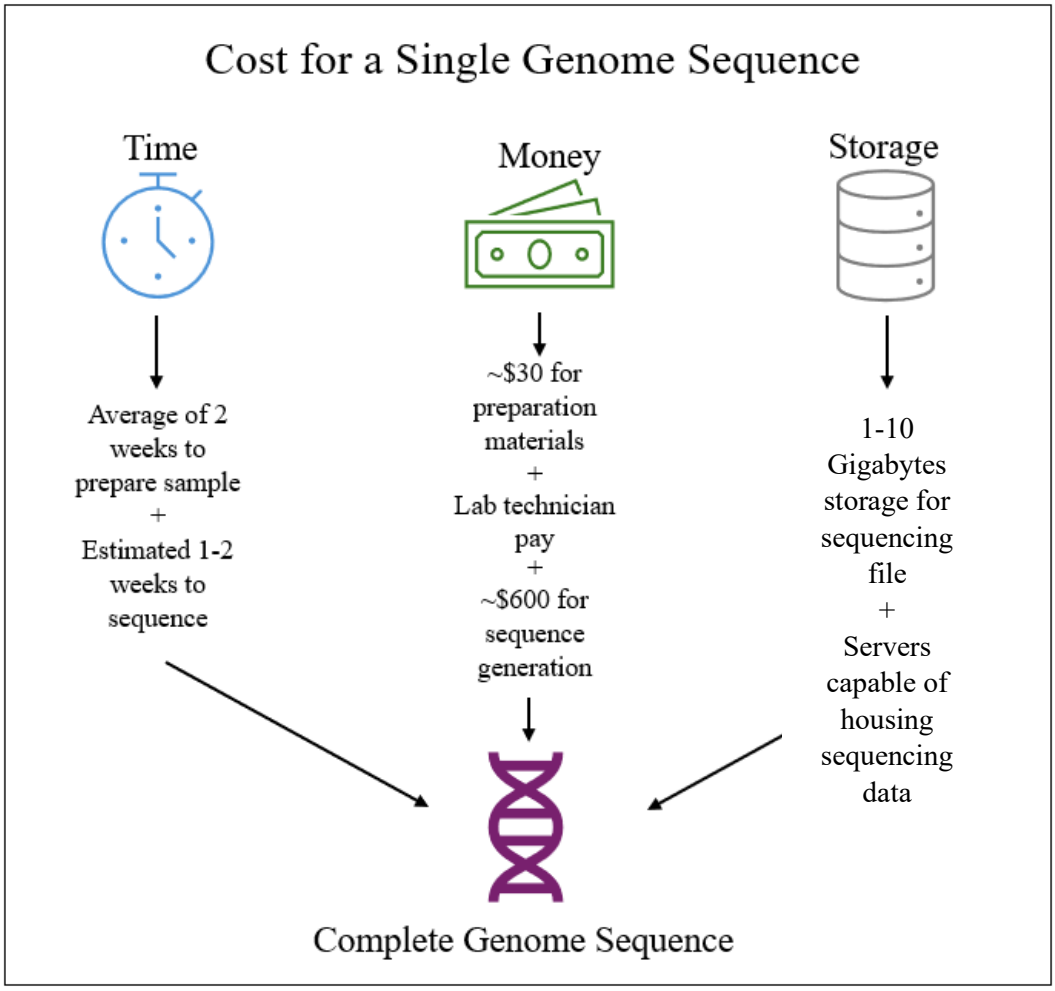


Figure 1. Visual representation of the complex costs associated with creating a complete genome sequence. Major cost areas are time, money, and storage.

Unfortunately, with the bulk of genomics labs populated with biologists, geneticists, ecologists, and conservationists, individuals with the skill set necessary for implementing and maintaining systems for storage of genome metadata are uncommon. With some labs having

done genomics for upwards of two decades, such storage problems are becoming increasingly concerning. Sifting through files of information of potentially thousands of individuals, often with erroneous or insufficient content, has led many organizations to start considering options in terms of storage. While professionally-developed digital organizational systems have begun to emerge in response to this growing problem, many are developed by individuals removed from genomics or with such a broad user base in mind that useful functionalities are overlooked or misapplied. As such, many labs have yet to find a system that works well for their specific, genomic needs. This project outlines the necessary considerations and methodology needed to create one possible solution to the complex problem of genome metadata storage.

Introduction

One genomics lab struggling to find a permanent metadata storage solution was the University of Montana's Good Lab, helmed by Dr. Jeffrey Good. Founded in 2010, the lab set quickly to researching mice and other mammals, focusing on genome evolution, speciation, and genetic causes of phenotypic variation. Throughout that time, Good Lab was using genome sequencing. This extensive use resulted in an expansive collection of genome sequences and metadata. A robust, redundant system of raw data storage with synchronized cloud-backups has ensured that all data are secured prior to depositing into public databases at publication. However, the full scope of these datasets, projects, and metadata were not effectively and centrally organized. For many years, Good Lab relied on a stand-alone Microsoft Excel spreadsheet to organize the metadata. The earliest recorded instance of the spreadsheet's usage was summer 2011, suggesting that the spreadsheet should have contained over a decade's worth of sequencing efforts.

However, the last recorded additions to the sheet occurred in 2020, coinciding with the global Covid-19 pandemic and a significant turnover in the Good Lab staff. A combination of these events led to lax enforcement of data archiving protocols, resulting in the datasheet's lack of use. The subject of this spreadsheet arose in spring of 2023 in a Good Lab weekly meeting, as concerns flared about proper data storage following a handful of private server issues. It quickly became apparent that not only were years of sequencing metadata missing from the repository, but years of misuse had occurred with inconsistent formatting or data missing altogether. The lab unanimously decided that a new, permanent, user-friendly system was needed to ensure that over a decade's worth of genome sequencing remained usable to avoid incurring costs for resequencing.

But therein lay a secondary decision, on what type of system should be used. As previously stated, expensive organizational systems do exist for metadata storage; but oftentimes, these systems fail to consider the complexity of genome metadata. Additionally, having a system that allows extremely complex searches across all elements of metadata was also deemed important by the lab, as it allows pin-point searches for specific clusters of sequences that may not otherwise be clustered together. The need for greater metadata complexity and organization lends itself exceptionally well to a database.

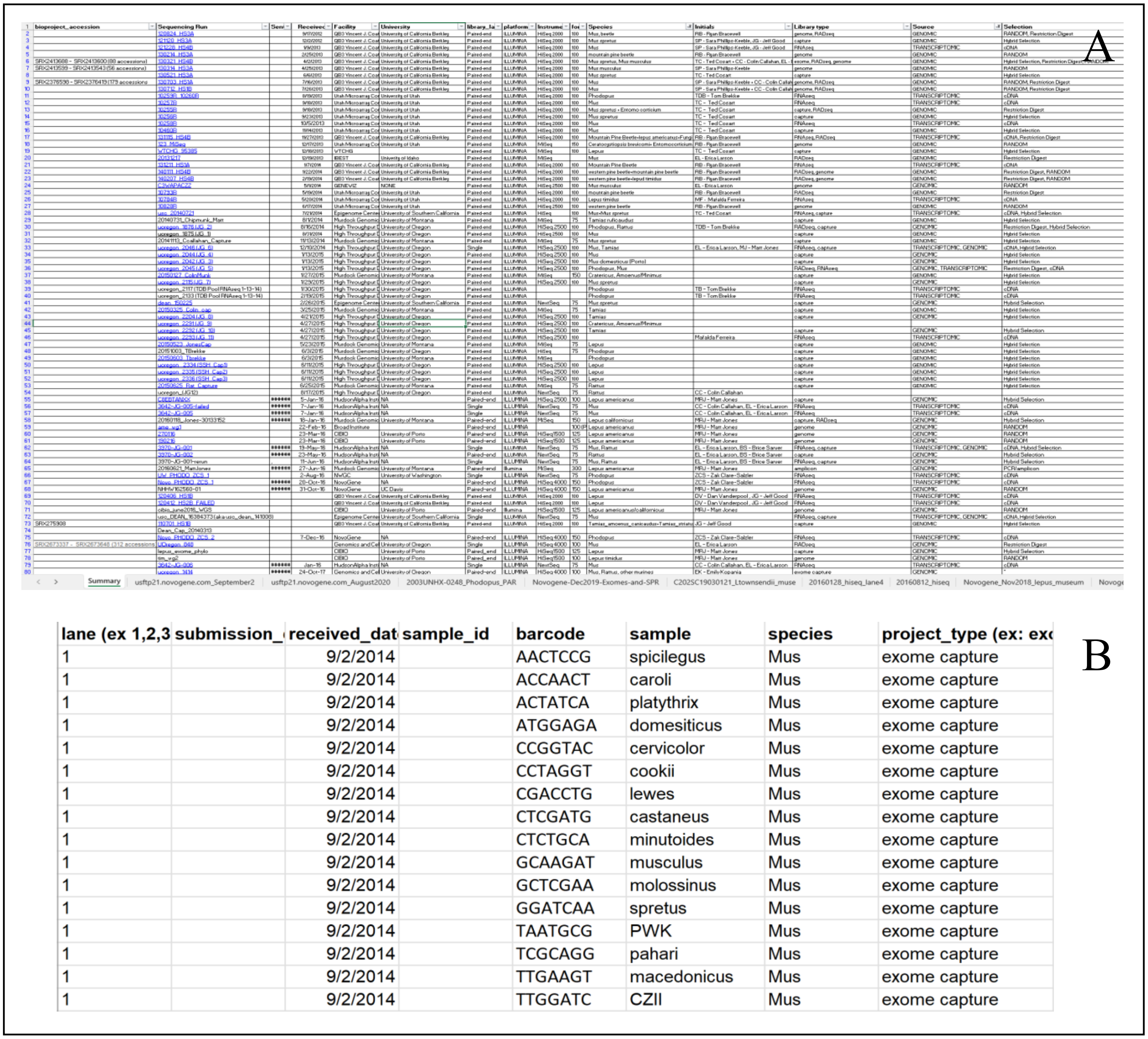
Ultimately, given my background in both genomics and computer science, we identified this project as an ideal focus for my Senior Thesis. By keeping the database's construction internal, greater attention could be paid to Good Lab's specific needs and goals, generating the best possible product to serve the lab. Thus, a year-long project to create the ideal Good Lab Genome Sequencing Database began in early summer of 2023.

Methods

The implementation of a database has become a relatively standardized procedure over the last several decades, with small but significant modifications made each step to create custom systems for projects ranging from simple systems like an employee catalogue to highly complex systems like that behind platforms such as Amazon. This project closely followed the standard steps of database design, such as those outlined in the textbook Fundamentals of Database Systems (Elmasri and Navathe 2016).

Like with projects spanning from review papers to experimental studies, the first step to database design was the data collection. The data was sourced from the outdated Good Lab spreadsheet, lab sample preparation notebooks, conversations with current lab members about valued metadata elements, and reviewing existing literature. Of these sources, the most data was sourced from the spreadsheet, with 300 Kilobytes of data to organize. The underlying structure of the spreadsheet was one central page housing the submissions of sequencing preps as groups done at the same time (Figure 2A) and a linked page for each submission housing the metadata for each sequence within that submission (Figure 2B). Notebooks also took a substantial amount of time to process, with an estimated 1,000 pages of data to sift through and organize by preparation methods and source. Ultimately, with the initial data survey complete, the process of categorizing data into metadata elements began. Within the spreadsheet, data was already largely categorized by metadata type, such as submission title, sequencing facility, and sequence sample source. This facilitated an efficient development of the general metadata categorization.

The goal with this categorization is to visualize what one complete instance of data will contain or, in other words, all the domains that data should be added to every time someone add new data. The importance of this “data abstraction” is that it enforces a uniform structure for



lane (ex 1,2,3	submission_	received_dat	sample_id	barcode	sample	species	project_type (ex: exc
1		9/2/2014		AACTCCG	spicilegus	Mus	exome capture
1		9/2/2014		ACCAACT	caroli	Mus	exome capture
1		9/2/2014		ACTATCA	platythrix	Mus	exome capture
1		9/2/2014		ATGGAGA	domesiticus	Mus	exome capture
1		9/2/2014		CCGGTAC	cervicolor	Mus	exome capture
1		9/2/2014		CCTAGGT	cookii	Mus	exome capture
1		9/2/2014		CGACCTG	lewes	Mus	exome capture
1		9/2/2014		CTCGATG	castaneus	Mus	exome capture
1		9/2/2014		CTCTGCA	minutoides	Mus	exome capture
1		9/2/2014		GCAAGAT	musculus	Mus	exome capture
1		9/2/2014		GCTCGAA	molossinus	Mus	exome capture
1		9/2/2014		GGATCAA	spretus	Mus	exome capture
1		9/2/2014		TAATGCG	PWK	Mus	exome capture
1		9/2/2014		TCGCAGG	pahari	Mus	exome capture
1		9/2/2014		TTGAAGT	macedonicus	Mus	exome capture
1		9/2/2014		TTGGATC	CZII	Mus	exome capture

Figure 2. Pages of the Good Lab metadata spreadsheet that served as the basis for much of the database's original structure. (A) The central page of the original Good Lab genome metadata spreadsheet. Contains details about each of the 97 submissions made between 2011 and 2020. (B) A representative sequencing data page which represents the metadata for each sequencing sample that occurred within the same broader submission.

data, versus leaving the types of data added up to user discretion. With this complete set of data, an initial structure for a MySQL relational database was designed (Figure 3A). This model is dubbed the Entity-Relationship (ER) Model, as it consolidates domains of data into broader

tables (or “entities”) that contain domains (now referred to as “attributes”) that are important to that entity. Entities are then connected to each other, via “relationships”. For instance, the entity “Submission” is the abstract table wherein all attributes about the broad sequencing submission (such as date submitted and the unique name of the submission), which is then connected or “related” to the specific sequence’s “Sample” entity that may be found within that submission, via a relationship “Contained”. This relational structure is what fundamentally allows the degree of “querying” (or searching) through the data in the database.

However, the ER model is extremely convoluted and overlooks key elements such as the fact that each sequencing facility will always belong to one and only one university or commercial lab, or that a project may have lots of publications- which the ER model does not account for in its structure. Thus, the ER model underwent “normalization” to ensure that each data table properly accounts for elements such as a single element of entity A may have many possible relationships with entity B, an attribute in entity C may be directly correlated with entity D, or several other possibilities. The goal of normalization is to fulfill the cardinal rule of coding: Do Not Repeat Yourself. If there are ways to minimize the number of times the same information must be entered without diminishing the structural framework of the database, it is best practice to take that action. Once normalization is completed, a much more legible “relational schema” arose which consisted of entities as tables, containing their attributes, but with additional tables and “foreign keys” (those highlighted in blue) that mimic the role of the relationships in the ER model (Figure 3B).

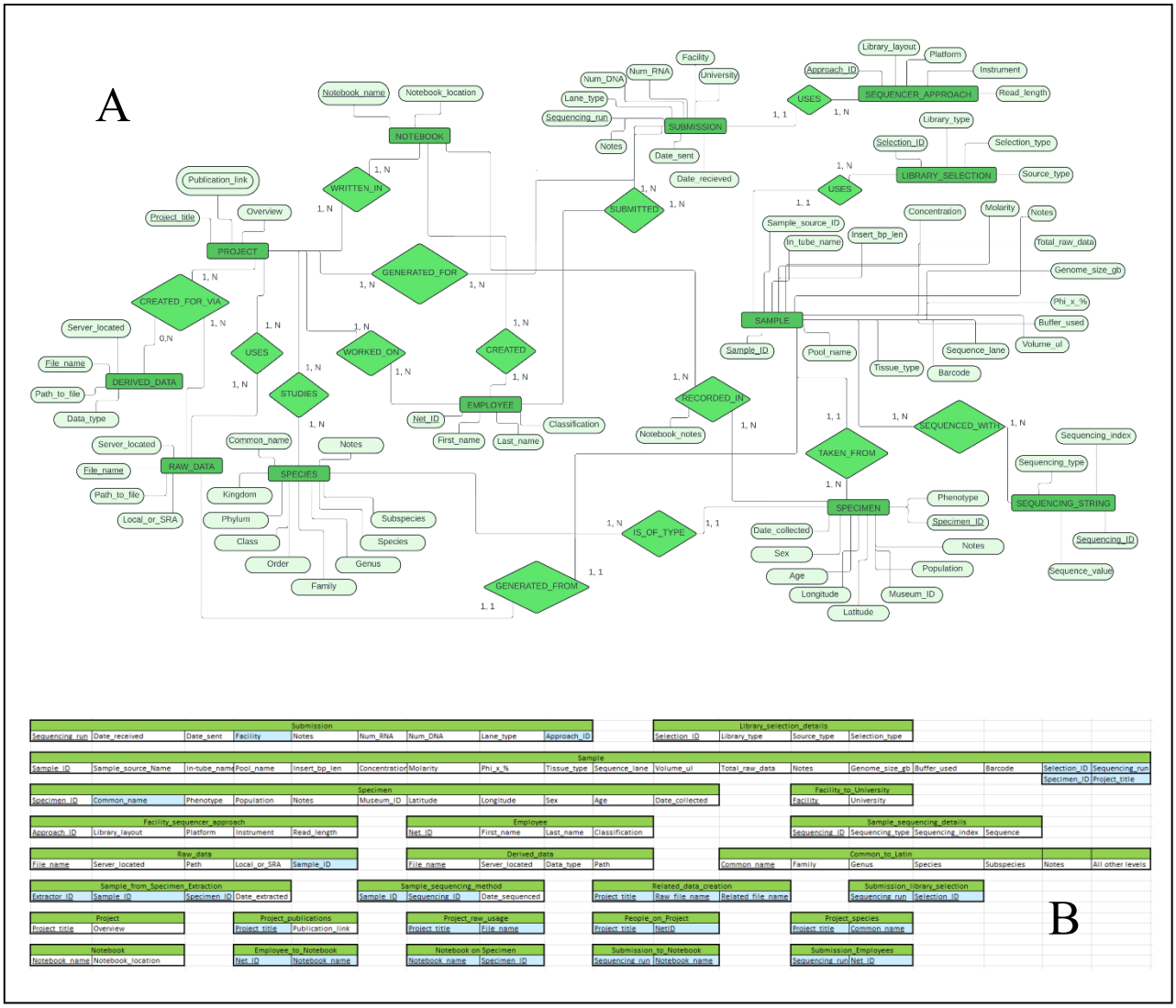


Figure 3. The starting and end points of the visual constructure of the Good Lab Genome Sequencing Database. (A) The complete, Entity Relationship Model of the Good Lab Genome Sequencing Database. Rectangles represent entities, diamonds relationships, and ovals attributes. (B) The normalized database schema. Each green label is a data table, and sub-elements are attributes. Blue cells are foreign keys that reference other tables.

With this fleshed-out relational schema, the database had to be implemented, as the schema was still little more than a complex drawing. To do so, a server to host the data was necessary. While initially a Microsoft Azure database server was initialized, a combination of cost and unneeded features resulted in a transition to an Akamai Linode server with built-in

MySQL functionality. Given that database data, even in large quantities, takes up little storage space, the specifications of the Linode were a “Nanode” with 1 Gigabyte of RAM, 1 CPU core, and 25 Gigabytes of storage. MySQL was initialized at version 8.0.36, the most recent publicly available version. While many relational database programming languages exist, MySQL was chosen based on personal experience with the language, its wide-spread use, and relatively straightforward structure.

To initialize the database itself, the design schema was translated into its equivalent MySQL database notation. Attention was paid to the data types of each attribute in each table, ensuring that there was as limited room for human error as possible; attributes were set to integers, dates, decimals, set-length text, et cetera based on what the attribute required. Certain attributes of high importance such as essential preparation details, sequence specimen identification, and submission dates were made mandatory using the “NOT NULL” command, ensuring that new additions to the database do not overlook the most important database elements. The file containing the complete database table framework was then added to the Linode server and run to initialize the database and its structure.

Following the formal creation of the database, extensive time was spent on compiling the known genome metadata into the form necessary for uploading to the database. Given the wide spread of data sources and inconsistencies in data forms, to complete the full collection, organization, and uploading of data took several months of daily work. However, it was deemed a necessary use of time, as the database is only so useful as its contents.

With the completion of data uploads to the database server, the lab could theoretically begin to use the database. However, given the Good Lab’s focus on programming languages designed for genomic analysis, only Dr. Good and one postdoc had database programing

language like MySQL. As such, a “Backend User’s Manual” was developed, walking the lab members through each step from remotely connecting to the database server, to how to upload and add data, to how to query for a wide variety of needs based on what the lab had indicated as priorities. However, due to the amount of time devoted to database initialization, the formal meeting to walk lab members through this manual has yet to occur, however it will ideally happen prior to the end of the spring 2024 semester or shortly thereafter. Additional work is also planned for a user-interface through a hosted website to further insulate the database while giving lab members with low coding comfort the ability to add data, query the database, and update information more efficiently.

Results

After almost one year of backend (database) development of the Good Lab Genome Sequencing Database, it is a fully functional system. All data available over the last year has been “seeded” into the database, allowing over a decade’s worth of sequencing metadata to be available to present and future members of Dr. Jeffery Good’s lab. The formal definition for the database’s functionality is, ironically, CRUD. Based on user restrictions put into place at their addition to the database server, users will be able to create (or add) new sequencing metadata to the database, read (or query) the database for any combination of data requirements, update data as new information is uncovered, or (under very limited conditions) delete information from the tables.

As reported earlier, the quantity and form of metadata added was extensive. Ultimately, there were 97 batch sequencing submissions from Good Lab to various sequencing facilities. The first data was added in 2011 and the last was dated from 2020, likely due to the disruptions that occurred in Good Lab at that time. Across the submissions, 3,403 unique samples were

sequenced meaning that, on average, roughly one sample was sequenced every day from 2011 through 2020. Interestingly, while the largest submission consisted of 312 biological samples while several other submissions contained only 1 sample. With a mean of 40.5 samples per submission and a median of 29.5 samples per submission, while there were some extremely large

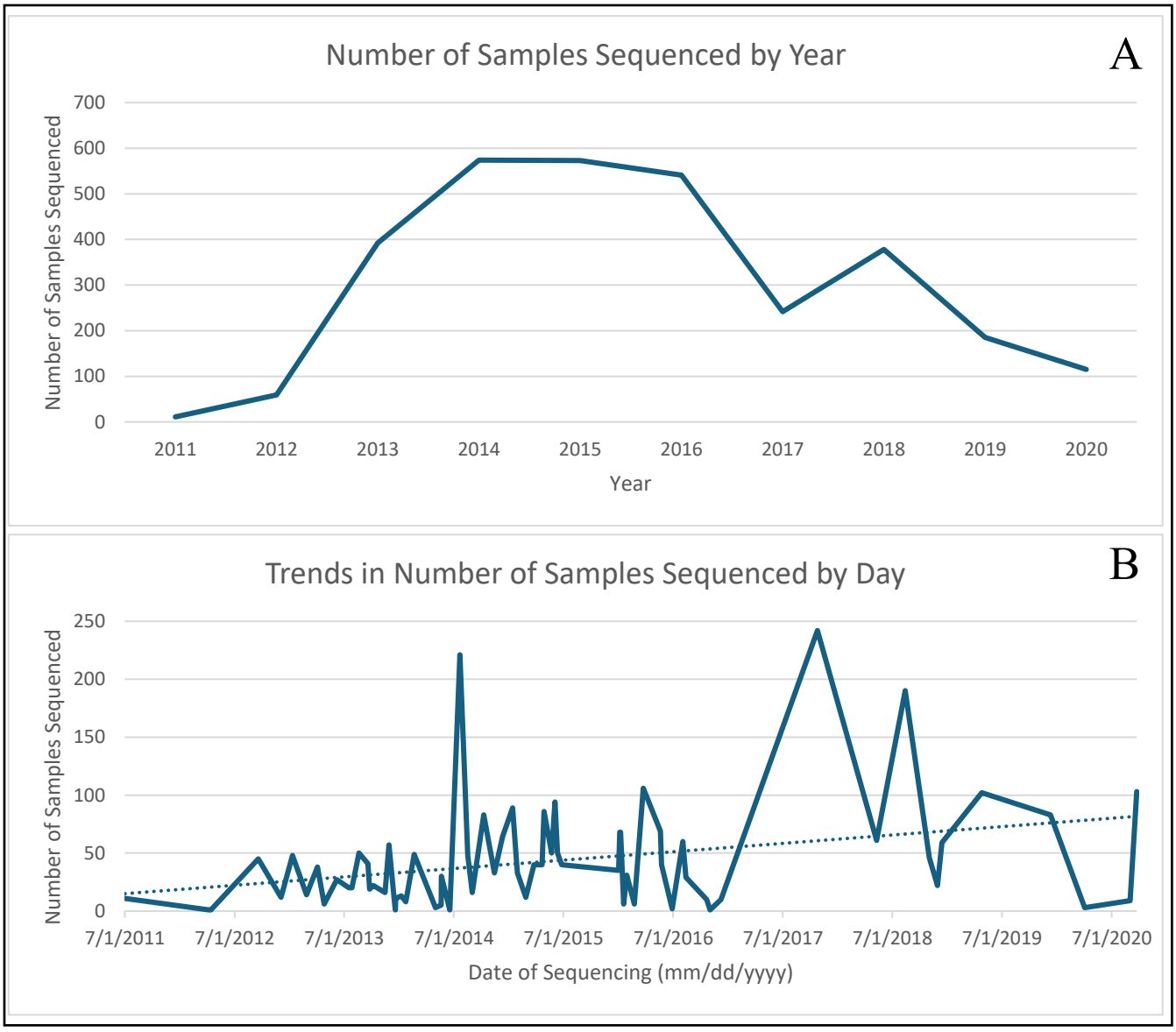


Figure 4. Trends in Good Lab genome sequencing volume from 2011 through 2020 based on database data. (A) Number of genome sequences generated collectively over each year. An evident "peak" in annual sequencing occurred in the mid-2010s, but importantly later data may be absent due to lack of reporting. (B) Daily sequencing efforts from 2011 to 2020. Trends appear to show in increase over time in the number of sequences done on days any sequencing was done, but also a decline in the frequency of sequencing days.

submissions, Good Lab tended towards smaller groups and more frequent submissions. Interestingly, despite an observed decrease in samples sequenced per year since 2014 (Figure 4A), the trend in the number of samples sequenced per event (represented in day) has increased over time (Figure 4B). However, it is important to consider that the declines in sequencing may be the result of declines in user reports, not actual Good Lab sequencing trends.

Beyond when and how much Good Lab is sequencing, what the lab is sequencing and where the specimens sequenced were from is also interestingly elucidated by the databases' efficiency. Evidently, Good Lab has sequenced roughly 135 distinct species collected across 15 U.S. states, 10 countries, and 3 continents. However, despite the high number of animals observed, most sequences came from only a handful of species. The most prevalent species (in terms of number of sub-classifications and sequencing individuals) was mice (spp. *Mus*), followed by hares (spp. *Lepus*), chipmunks (spp. *Tamias*), hamsters (spp. *Phodopus*), and pine

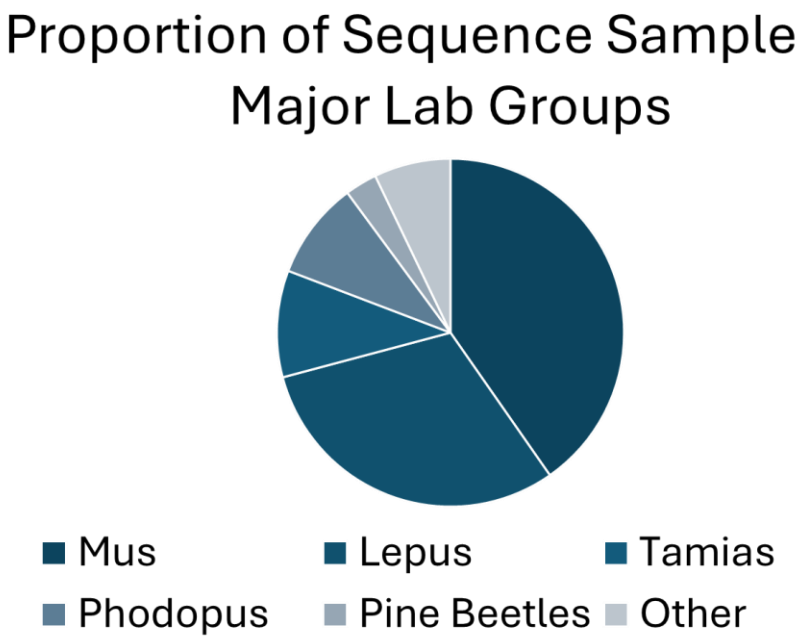


Figure 5. Proportional representation of species sequenced by Good Lab from 2011-2020. Trends follow logical patterns based on species usage in current and historic research. Proportion of pine beetle dates to mid-2010s studies into the growing epidemic of pine beetles in Montana.

beetles (Figure 5). The remaining portion of samples appear to largely originate from single sequencing events used to create complex phylogenetic trees to understand the evolutionary histories of mice, hares, chipmunks, and hamsters.

The completed database is sufficient results to have made this project worthwhile, but the interesting trends within the fundamental mechanics that make Good Lab tick are fascinating as well. However, much remains to be done to make the system as useful in perpetuity as possible.

Discussion and Future Suggestions

Despite database design having a streamlined methodology to follow, there are many pathways that can be taken along that methodical path can and do lead to both interesting and challenging results. Throughout this project, the structure of the database explored several different functional paths. Additionally, some paths are still being explored, despite the core projects' completion. These future paths include data recovery, universalizability, and, most pressingly, a front-end user interface (put simply, a website).

The first explored path was the use of a relational database design via MySQL, rather than a document database. Document databases store all contents from (in this case) a sequencing submission in a single file, while relational databases create more fragmented clusters of closely related data. In certain projects, document databases are exceptionally useful. However, the connectivity between different sequencing instances was considered exceptionally important by the members of the lab. This led to the logical preference for a relational database rather than a document database.

Furthermore, there was a period of about a month wherein an alternative form of data storage was being considered on the suggestion of Dr. Travis Wheeler. The alternative form

consisted of a text file for each submission, mirroring the structure of a document database. A complex GUI (graphical user interface) would then be used to query keywords within each document. Ultimately, while such a process would create an easier initial uploading process for users, it allows for a significant amount of flexibility in terms of documentation, which was to be avoided. Likewise, the quality and complexity of queries would have been limited significantly. It was for these reasons that a relational database built on MySQL was returned to and used for the remainder of the project.

Moving forward, it cannot be ignored that almost 4 years of sequencing data is still absent from this database due to lack of systemic backlogs and records. Certainly, that metadata does exist, but presently it is not centralized. One consideration is that in the upcoming "onboarding" to the database, individuals will be required to bring any truant data and will use that as a means of practicing using the database to not only ensure users can manipulate the database, but simultaneously adding necessary and currently absent data.

Another path that poses a future challenge and reward is that of disseminating the "white label" version of this database. The "white label" version of software is the version without any customized elements, which can be bought and sold to other organizations to minimize repetitive invention of software. As mentioned in the background review, while systems do exist for genome metadata storage, many lack widespread applicability or are extremely costly. Simultaneously, many organizations are searching for systems that fulfil their data storage requirements. The generalized structure of this genome metadata database lends itself well to applications beyond Good Lab. In fact, talks have already begun with several prospective clients on how to implement the Good Lab Genome Sequencing Database schema into several other

labs at the University of Montana. Such work is ongoing and is slowed by the need to ensure that the generalized version fulfills consumers' needs.

Finally, and most significantly, the next steps into advancing the quality of this database are being explored. While a backend system works well for those with a moderate level of comfort with database coding, Good Lab has many members who do not have this level of confidence. As such, a secondary, more user-friendly interface would prove beneficial towards ensuring the longevity of this database. By creating a website linked with several of Good Lab's other web properties, users would be able to add, update, and query the database remotely with less coding. Users would simply select the program type they are interested in using, type in a few key words as prompted, and the system will perform the commands internally. Ideally, this element of the project would have been completed within the year since the project's start, but a strong desire to ensure the database's contents and structure were as sound as possible, as well as detours through other possible approaches delayed the project significantly.

However, despite the mutli-faceted future efforts necessary to take the Good Lab Genome Sequencing Database to the next level of usability and value, the database itself has already begun to prove useful. The ability to filter through the data has allowed rapid retrieval of valuable data needed for projects, presentations, and reviews. With any luck, these efforts are only the beginning of the database's usefulness, and it will provide Good Lab with a structurally strong, accessible, efficient means of sorting through an ever-growing mountain of genome sequencing metadata.

Acknowledgements

This work was supported by funding through grants from the National Institute of Health (R01-HD094787 and R01 HL159061) and the summer 2023 NSF EPSCoR CREWS project. Without this financial support, this project would not have been able to occupy as much of my time as I ended up devoting to it over the last year.

Dr. Jeffrey Good and Dr. Travis Wheeler provided invaluable insight from the genomics and software perspectives respectively. Their suggestions, modifications, and all-around support encouraged this project to reach heights and complexities that would have only been aspirational otherwise.

Words cannot describe the amount of appreciation I have for Good Lab's continual support, ideas, and understanding as this project ebbed and flowed. Many times, it was their excitement about this project that pushed me forward and encouraged me to explore the topic from new angles.

And finally, boundless thanks are owed to my friends, partner, and family for keeping me fed, watered, and exercised when I was too buried in servers and spreadsheets to recognize the need to breathe and take breaks. If not for their efforts, this project may never have reached its true potential, as a program is only as functional as its creator.

Thank you all from the bottom of my heart.

Works Cited

Braslavsky, Ido et. al. “Sequence information can be obtained from single DNA molecules.” *PNAS,* vol. 100, no. 7, 21 Mar. 2003, pp. 3960-3964. https://doi.org/10.1073/pnas.0230489100.

Brenner, S., et al. “Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays.” *Nature Biotechnology*, vol. 18, no. 6, June 2000, pp. 630–634. https://doi.org/10.1038/76469

Church, G., et. al. “Characterization of individual polymer molecules based on monomer-interface interactions.” US patent 5,795,782, 1998.

Cook, Carly N., Kent H Redford, and Mark W Schwartz. “Species conservation in the era of genomic science”. *BioScience,* vol. 73, no. 12, 16 Nov. 2023, pp. 885–890, https://doi.org/10.1093/biosci/biad098.

Elmasri, Ramez A., and Shamkant B. Navathe. *Fundamentals of Database Systems*. 7th ed., Pearson, 2016.

Fiers, W., et al. “Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene.” *Nature*, vol. 260, 8 Apr. 1976, pp. 500–507. https://doi.org/10.1038/260500a0

Fleischmann, Robert D., et al. “Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd.” *Science*, vol. 269, no. 5223, 28 Jul. 1995, pp. 496-512, DOI:10.1126/science.7542800

Giani, Alice M., et. al. “Long walk to genomics: History and current approaches to genome sequencing and assembly”. *Computational and Structural Biotechnology Journal,* vol. 18, 2020, pp. 9-19. https://doi.org/10.1016/j.csbj.2019.11.002.

Griffiths, Anthony J.F. "genomics". *Encyclopedia Britannica*, 22 Nov. 2023, https://www.britannica.com/science/genomics. Accessed 30 April 2024.

Heather, James M., and Benjamin Chain. “The sequence of sequencers: The history of sequencing DNA.” *Genomics*, vol. 107, no. 1, Jan. 2016, pp. 1–8, doi: 10.1016/j.ygeno.2015.11.003.

Hood, Leroy, and Lee Rowen. “The Human Genome Project: big science transforms biology and medicine.” *Genome Med*, vol. **5**, no. 79, 13 Sept. 2013. https://doi.org/10.1186/gm483

Hug, L., et al. “A new view of the tree of life”. *Nat Microbiol*ogy, vol. 1, no. 16048, 11 Apr. 2016, https://doi.org/10.1038/nmicrobiol.2016.48

Hutchison III, Clyde A. "DNA sequencing: bench to bedside and beyond." *Nucleic Acids Res*, vol 35, no. 18, 12 Sept. 2007, pp. 6227-37. doi: 10.1093/nar/gkm688.

International Human Genome Sequencing Consortium. "Finishing the euchromatic sequence of the human genome." *Nature*, vol. 431, 21 Oct. 2004, pp. 931–945. https://doi.org/10.1038/nature03001

International Human Genome Sequencing Consortium. "Initial sequencing and analysis of the human genome." *Nature*, vol. 409, 1 Feb. 2001, pp. 860–921. https://doi.org/10.1038/35057062

Margulies, M., et. al. "Genome sequencing in microfabricated high-density picolitre reactors." *Nature*, vol. 437, 31 July 2005, pp. 376–380. https://doi.org/10.1038/nature03959

McCormick, Kathleen A. and Kathleen A. Calzone. "The impact of genomics on health outcomes, quality, and safety." *Nursing Management*, vol. 47, no. 4, Apr. 2016, pp. 23-26, doi: 10.1097/01.NUMA.0000481844.50047.ee.

NHGRI. "DNA Sequencing Fact Sheet." *National Human Genome Research Institute*, National Institute of Health, 27 June 2023, https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Fact-Sheet.

NHGRI. "NHGRI History and Timeline of Events." *National Human Genome Research Institute*, National Institute of Health, 16 Aug. 2022, https://www.genome.gov/about-nhgri/Brief-History-Timeline.

"Roadmap: Project Plan." *Earth BioGenome Project*, Earth BioGenome Project, 2022, www.earthbiogenome.org/roadmap.

Sanger, F., et. al. "Nucleotide sequence of bacteriophage φX174 DNA." *Nature*, vol. 265, 24 Feb. 1977, pp. 687–695. https://doi.org/10.1038/265687a0

Saravanan, K.A. et. al. "Role of genomics in combating COVID-19 pandemic." *Gene*, vol 823, 20 May 2022, doi: 10.1016/j.gene.2022.146387.

Slatko Barton E, et. al. "Overview of Next-Generation Sequencing Technologies." *Current Protocols in Molecular Biology*, vol. 122, no. 1, 16 Apr. 2018. doi: 10.1002/cpmb.59. PMID: 29851291; PMCID: PMC6020069.

Theissinger, Kathrin, et. al. "How genomics can help biodiversity conservation". *Trends in Genetics*, vol. 39, no. 7, Jul. 2023, pp. 545-559, https://doi.org/10.1016/j.tig.2023.01.005.

Thermo Fisher Scientific. "DNA Sequencing." *Thermo Fischer Scientific*, Thermo Fischer Scientific Inc., 2024, https://www.thermofisher.com/search/browse/category/us/en/90220052.

Wetterstrand, Kris A. “DNA Sequencing Costs: Data.” *National Human Genome Research Institute*, National Institute of Health, 16 May 2023, https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.