Chapter 9 Bioinformatic practical applications in biotechnology, medicine, environmental and agricultural sciences

Capítulo 9 Aplicaciones prácticas de la bioinformática en biotecnología, medicina y ciencias medioambientales y agropecuarias

RAGGI, Luciana†*[1], GODOY-LOZANO, Elizabeth Ernestina[2], JIMENEZ-JACINTO, Verónica[3] and ESCOBAR-ZEPEDA, Alejandra[4]

[1] CONAHCYT - Instituto de Investigaciones Agropecuarias y Forestales, Universidad Michoacana de San Nicolás de Hidalgo
[2] Centro de Investigación Sobre Enfermedades Infecciosas, Instituto Nacional de Salud Pública, Cuernavaca, Morelos, Mexico
[3] Unidad Universitaria de Secuenciación Masiva y Bioinformática, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Morelos, Mexico.
[4] Microbial Informatics Team, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK

ID 1er Author: *Luciana, Raggi* / **ORC ID:** 0000-0001-8502-4834

ID 1er Co-author*: E. Ernestina, Godoy-Lozano* / **ORC ID:** 0000-0001-6927-9132

ID 2do Co-author: *Verónica, Jiménez-Jacinto* / **ORC ID:** 0000-0001-6742-1537

ID 3rd Co-author: *Alejandra, Escobar-Zepeda* / **ORC ID:** 0000-0003-3549-9115

DOI: 10.35429/H.2023.7.1.103.124

L. Raggui, E. Godoy, V. Jimenez and A. Escobar

* luciana.raggi@umich.mx

Á. Marroquín, L. Castillo, J. Olivares and N. Olguín Cruz. (VV. AA.) CIERMMI Women in Science T-XXI Biological, Humanities and Social Sciences. Handbooks-©ECORFAN-México, Querétaro, 2023.

Abstract

The new massive sequencing technologies of nucleic acids (DNA and RNA) have allowed a great advance in health, biology, environmental, agricultural, and biotechnology sciences. However, the gigantic amount of data (big data) obtained from each experiment requires increasingly demanding computational power and also experimented computer scientists. Therefore, the field of bioinformatics, or the use of computing applied to the understanding of biological systems, requires specialized analysts who have an understanding of both biology and computational systems.

In this chapter, we set down examples of bioinformatics around the study of a) microorganisms, b) food science, c) health studies concerning the immunological repertoire, and d) studies in agricultural sciences.

Metagenomics, Transcriptomics, Metaprofiling, Microbiomics, Omics Data Science

Resumen

Las nuevas tecnologías de secuenciación masiva de ácidos nucleicos (ADN y ARN) han permitido un gran avance en las ciencias de la salud, la biología, el medio ambiente, la agricultura y la biotecnología. Sin embargo, la gigantesca cantidad de datos (big data) que se obtiene de cada experimento requiere una potencia de cálculo cada vez más exigente o, tal vez, unos informáticos cada vez más experimentados. Por lo tanto, el campo de la bioinformática, o el uso de la informática aplicada a la comprensión de los sistemas biológicos, requiere analistas especializados que comprendan tanto la biología como los sistemas computacionales.

En este capítulo, exponemos ejemplos de bioinformática en torno al estudio de a) los microorganismos, b) la ciencia de los alimentos, c) los estudios de salud relativos al repertorio inmunológico, y d) los estudios en ciencias agropecuarias.

Metagenómica, Transcriptómica, Perfiles metagenómicos, Microbiómica, Ciencia de datos ómicos

9.1 Introduction

Computer science or informatics is a young science that has had an accelerated development, that is, in less than a human generation, more than 20 generations of computers or ways of programming have been developed. Thus, computing has quickly permeated all areas of knowledge.

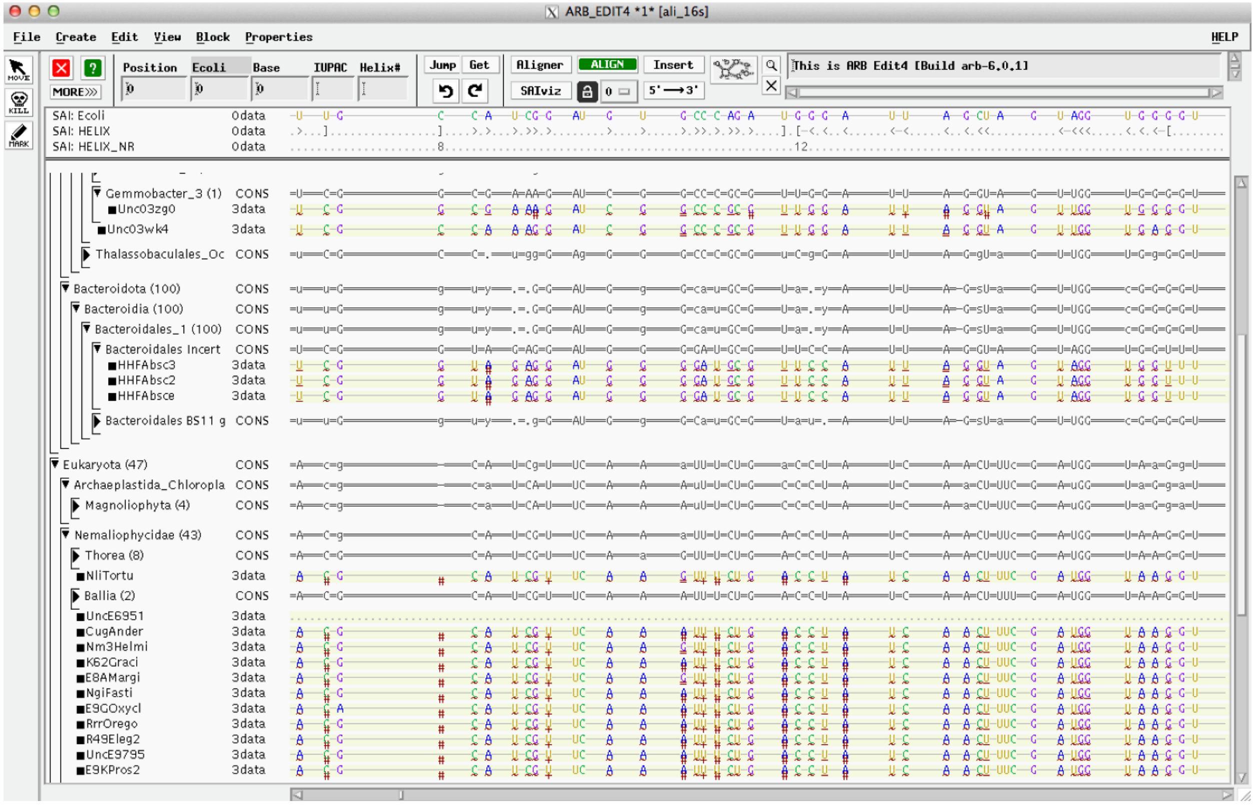
Bioinformatics is the use of computing applied to the understanding of biological issues, and it arose in the mid-1960s when computers began to reach the hands of scientists in the area of life sciences, such as Margaret O. Dayhoff, a pioneer in the area and who published the first atlas of protein sequences and their structures (Dayhoff, 1979). In this chapter we show our research, as scientists in the area of life sciences, using bioinformatics oriented to omics sciences.

With the development of molecular methods, particularly those used for the manipulation, characterization, and sequencing of DNA and RNA, it was possible to establish molecular protocols aiming to characterize biological species, particularly microbial communities associated with complex environments, and on the other hand, they are used to study the immunological repertoire of humans and other beings with a similar defense system. It was in 1988 that the term metagenomics, coined by Jo Handelsman (Handelsman, 2004), became widely used to refer to the study of environmental DNA associated with a biome.

Studies of metagenomic DNA for the scrutiny of specific metabolic functions began by preparing metagenomic libraries in plasmids and transforming them into competent strains (microorganisms that are easy to modify genetically, such as *Escherichia coli* or *Saccharomyces cerevisiae*) for the subsequent scrutiny of enzymatic functions of interest.

At the same time, sequencing using DNA chain termination inhibitors, now called Sanger sequencing (Sanger et al., 1977), was developed, which was a spark that started an explosion of sequencing of all model organisms within laboratories, and afterward all living organisms. Dendrograms derived from this sequencing, and from the alignment of these sequences (Fig. 9.1) currently show the connection of all living organisms in an evolutionary context (Fig. 9.2).

Figure 9.1 16S/18S sequences alignment in ARB software utilizing SILVA/ARB database

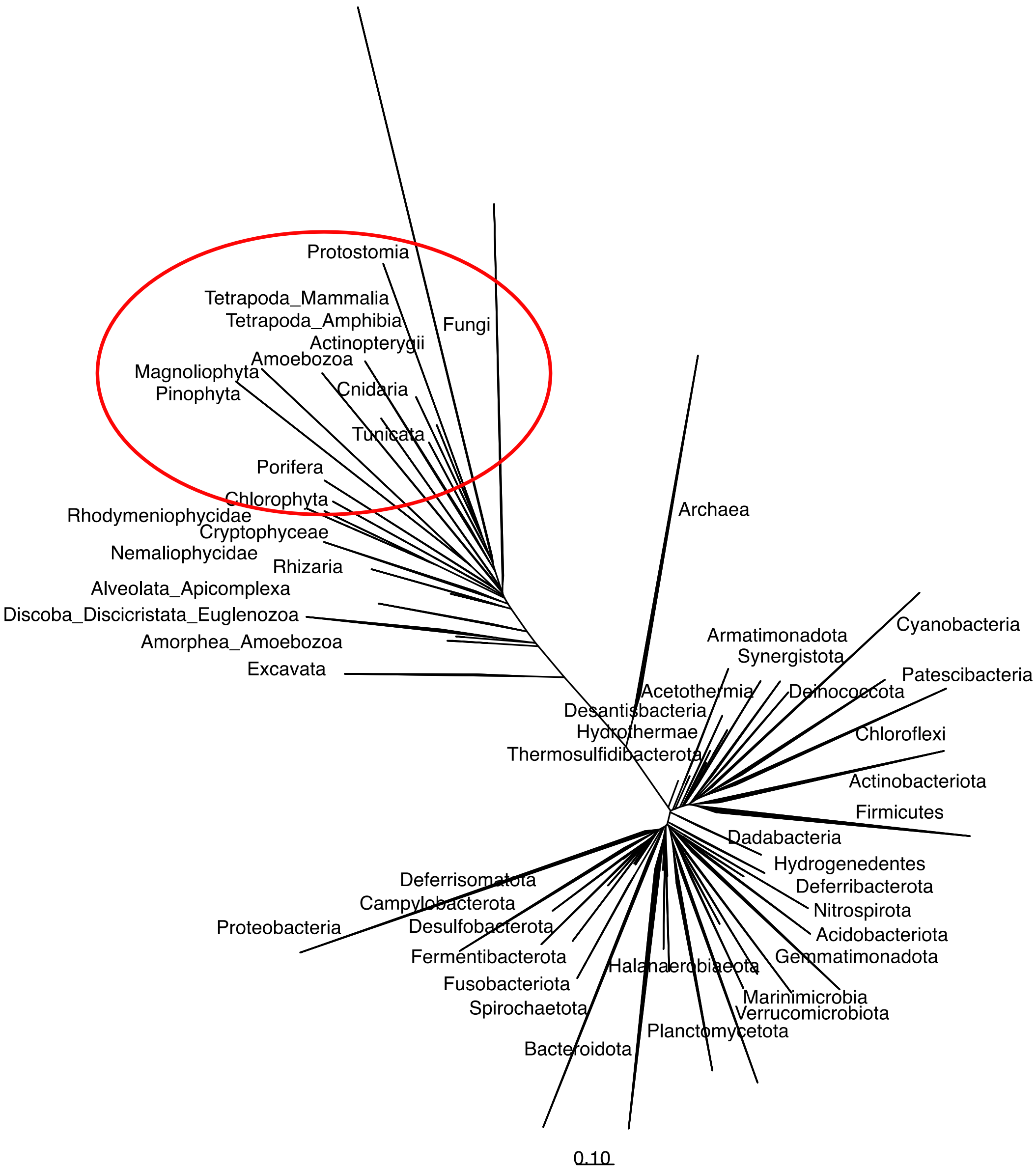


The taxonomy or classification of microorganisms is currently based on the study of ribosomal genes that Carl Woese worked on in the 70s (Woese, 1966; Woese & Fox, 1977). Phylogenetic studies are based on the alignment of the sequences of highly conserved genes, particularly those that code for ribosomal RNA (16S in prokaryotes or 18S in eukaryotes). All living organisms have these genes in their genome, in such a way that their sequence can be aligned and then a dendrogram or phylogenetic tree of all living organisms currently shows that the largest quantity of living beings (biodiversity) are microscopic, that is, they are microorganisms (Fig. 9.2).

Massive sequencing of DNA came later with new generation sequencing (NGS) technologies being 454 the first, followed by Torrent and several others until reaching the sequencing by synthesis of the Illumina® company, which is the most relevant by volume of data generated and therefore the most used in today (Table 1). New sequencing technologies are currently trying to overcome the limitation of sequencing by short pieces of DNA, Oxford Nanopore technology has generated very long reads with a polymerase-independent chemistry, however, it has still a high error rate, and the second one is the PacBio Hifi that has managed to generate long reads even with a low error rate.

Myriad computer programs can currently analyze millions of sequences efficiently and allow the design of workflows with different capacities; finally, the DNA and RNA sequences are 4-letter (GCAT nucleotides) sequences and their analysis becomes mathematically relevant with the combinatorics and statistics that a computer can particularly analyze.

Figure 9.2 Dendrogram based on the alignment of 16S (prokaryotes) and 18S (eukaryotes) sequences in which the red circle surrounds living organisms with macroscopic dimensions (macrobiota) and all the other groups that dominate in quantity belong to the microscopic world (microbiota).



9.2 Bioinformatics in the study of microorganisms

The microbiota is the large number of microorganisms associated with a certain habitat, whether it is a terrestrial or aquatic environmental ecosystem or a more complex organism, such as a fungus, plant, or animal, that makes a habitat for microorganisms. The rules of host-microorganisms interaction are not yet fully known, however, some of the niches provided by a complex organism are densely colonized by endogenous microbiota (Medzhitov, 2007). It is calculated that the density of microorganisms, for example in an intestine, can become of the order of 10^{10} per cm^3. One of the favorite sites for microbial colonization is the digestive tract or alimentary canal of both vertebrates and invertebrates. At present, there are extensive studies on the characterization of the intestinal microbiota of various hosts, which are carried out thanks to technological advances. "Microbiomics" studies, as they are called, are based on omics analysis of the microbiota: through the development of DNA, RNA, and protein sequencing techniques, and thus trying to elucidate the function of the microbiota on organisms: digestion, energy homeostasis, synthesis of vitamins, amino acids and fatty acids, and direct interaction with the immune system.

It is well known that an immune system depends on an interrelationship with microorganisms and their molecules, consequently, the paradigm of growing animals with a strengthened immune system is reinforced with the idea of continuous exposure to microorganisms, contrary to keeping them in a sterile environment with a depressed immune system; thus, experiments are going in that direction.

The microbiota is becoming a "proxy" or biomarker, in both environmental and health studies, that is, patterns and communities of healthy or altered intestinal microbiota are analyzed, contaminated or healthy soils or aquatic bodies, and the health of other several habitats is beginning to be microbiologically categorized. Microbiota will indicate the state of each environment, showing for example a possible dysbiosis in organisms, caused by stress or contamination of the system (Perry et al., 2020). In nutritional science, many studies have been carried out in various species, including humans, with prebiotic and/or probiotic supplements added to diets to observe their influence on the microbiota and host (namely the holobiont, that is made of these two components), and the health effects.

Currently, holo-studies recognize the symbiotic and therefore natural and essential association of an organism with its endogenous microbiota and its interaction with the environmental microbiota (Limborg et al., 2018). Organisms are exposed to a great diversity of pathogenic bacteria and viruses and their microbiota is seen as a defense barrier against these pathogens (Chiu et al., 2017).

9.3 Metagenomics of agricultural and aquaculture systems

At present, global climate change has an impact on agricultural systems. Microorganisms contribute to climate change, the clearest example of how microbial life contributes to atmospheric changes was the oxygenation of our atmosphere in the early days of Earth's geological history. Today, microorganisms continue to be the main players in atmospheric changes at all levels, including terrestrial, oceanic, and urban areas. From gases produced by cows' rumens to melting soils in permafrost regions, symbiotic coral systems in oceans, and carbon waste from our cities, microbial metabolism produces and absorbs gases that can affect the environment and climate (Tiedje et al., 2022).

At the level of food production systems, it is important to take into account the microorganisms of the system and observe them as holo-systems or holobiomes (Gutiérrez-Pérez et al., 2022). Integrated agro-aquaculture systems encompass the integration of aquaculture production with agriculture, where waste from one system is passed on to the other, and this recirculation of nutrients is attainable owing to the microorganisms that usually accumulate and proliferate in biofilters to leverage their metabolism. Microbiology is an indispensable component of these nutrient recirculation and water purification systems, and other environmental services (e.g., bioremediation and alternative energy), thanks to its biogeochemical characteristics, metabolic diversity, resilience capacity, rapid adaptation, and evolution. The integration of aquaculture into terrestrial systems provides an additional source of protein and high-quality fatty acids (e.g. DHA) that are essential for food security. Fish are born and perish in an aquatic environment that is densely populated with microorganisms (~10^6 bacteria and 10^9 viruses per mL of water, Whitman et al., 1998), in contrast to the terrestrial/aerial environment; and even though environmental conditions tend to be more constant in the aquatic environment, the microbial load and diversity are also high. Consequently, the immune system of aquatic organisms co-develops with this high density of microorganisms, thus establishing an interdependent relationship between the host and its microbiota (Kogut et al., 2020), which, in fact, co-regulate and co-evolve.

There are numerous microbiological occurrences in the agricultural systems that can be capitalized on, and through understanding them, we may be able to regulate and modify the systems to make them ecologically more efficient and therefore ecologically sustainable.

9.4 Metagenomics of fermented food

9.4.1 Fermented foods are cultural markers

Fermented foods hold a significant place in the cultural practices of all societies, intricately woven into their historical and societal fabric, and closely intertwined with the establishment of settled communities.

As humans evolved, the process of cultivation led to the deliberate selection of the most nutritious plant varieties, optimizing harvest yields and fostering specialization among communities in the production and processing of essential grains and vegetables, forming the cornerstone of their dietary patterns. In contemporary times, corn, wheat, and rice have emerged as the most extensively cultivated cereals globally (Erenstein et al., 2021). Just as agriculture spurred progress, the transition to sedentary lifestyles facilitated the domestication and controlled breeding of animals for human consumption. Moreover, the confluence of diverse cultural traditions through cultural exchanges and syncretism has enriched the spectrum of consumable goods by amalgamating raw materials and techniques from disparate cultural heritages.

In tandem with the advancements in food production, preservation technologies have evolved, encompassing a spectrum that spans from rudimentary methods like drying and salting to more intricate techniques such as smoking, pickling, and fermentation. These techniques, adept at bestowing the desired attributes upon the end product, have endured through the establishment of standardized manufacturing processes and the continual introduction of fresh batches of raw materials. Coincidentally, this very mechanism has inadvertently catalyzed the natural selection of microorganisms, perpetuating their prevalence and influence.

9.4.2 The role of microbial communities in fermented foods

Intricate matrices rich in concentrated nutrients are found in fermented foods, serving as substrates that provide sustenance to microorganisms either naturally present or deliberately introduced during processing.

Throughout the fermentation process, the dynamic enzymatic activities of these microorganisms orchestrate a remarkable transformation of the food matrix. This metamorphosis encompasses alterations in critical physicochemical attributes, such as pH and redox potential, alongside enhancements in microbiological attributes. This microbial-driven evolution fosters the emergence of robust strains that adeptly navigate the novel environmental milieu and adeptly withstand the presence of antimicrobial agents.

In parallel, the microbial metabolic endeavors not only reshape the fundamental sensory properties of the food but also bestow upon it a spectrum of secondary metabolites and release degradation byproducts from its elemental building blocks—carbohydrates, proteins, and lipids. This intricate interplay generates a multifaceted medley of compounds intricately woven into a complex tapestry of aroma and flavor profiles (Bamforth, 2005).

The environment serves as a pivotal wellspring of fungi and bacteria that enrich fermented food processing. The diverse array of microorganisms found in the environment defies replication within aseptic confines or in settings apart from their indigenous origins. This notion forms a cornerstone in comprehending the innate uniqueness of traditional fermented foods, a uniqueness that, in select instances, finds safeguarding under the mantle of Designation of Origin (Reinders et al., 2019).

Exemplifying this concept are several traditional non-distilled fermented beverages deeply rooted in Mexican heritage. These include maguey mead pulque, pineapple tepache, coconut palm mead tuba, red prickly pear colonche, and corn-derived tegüino and pozol, among others. While these treasures possess rich tradition, an exceptional nutritional value, and embody the artistry of local artisans, it's noteworthy that none of these legacies enjoys the shield of designation of origin.

9.4.3 Biotechnological applications of microorganisms associated with fermented foods

Traditional fermented foods have garnered substantial interest in industrial realms due to their status as a secure reservoir of microorganisms, thoughtfully culled through the annals of tradition, and suitable for human consumption. The daily incorporation of these time-honored original communities creations has been empirically linked to a wealth of health advantages for consumers (Cuamatzin-García et al., 2022).

In the quest for elucidating the mechanisms underpinning these benefits, scrutiny has turned towards identifying the presence of probiotic strains (Soemarie et al., 2021) and the extraction of prebiotic compounds, envisioning their integration as additives in diverse comestibles. This strategic action aims to foster the proliferation of beneficial human gut bacteria (Christensen et al., 2022), thereby enhancing the intestinal microbiota's desirable equilibrium.

Concurrently, probing into the microbial profiles of foods serves as a guiding compass in crafting starter cultures, steering the preparation of authentic 'type' foods, and exploring novel product frontiers within the industry. This endeavor is not merely confined to creating delectable offerings but underscores paramount considerations of safety and transformative efficiency (Hansen, 2002).

Moreover, it is a fact that the orchestration of physicochemical changes during fermentation and maturation ushers in a selective process to the microorganisms primed for a substrate battle. In this enthralling contest, the victors often bear the mantle of producing antimicrobial compounds or harbor steadfast defense mechanisms. These strategic attributes confer a distinct competitive edge. Leveraging this insight, the realms of food and biomedicine converge in their fervent exploration of harnessing nature’s arsenal, such as the remarkable bacteriocins. Embodied by nisin as the pioneering industrial example, the application of such natural preservatives beckons as an avenue of great interest (Lahiri et al., 2022).

9.5 Application of Bioinformatics in Immune Repertoire Studies

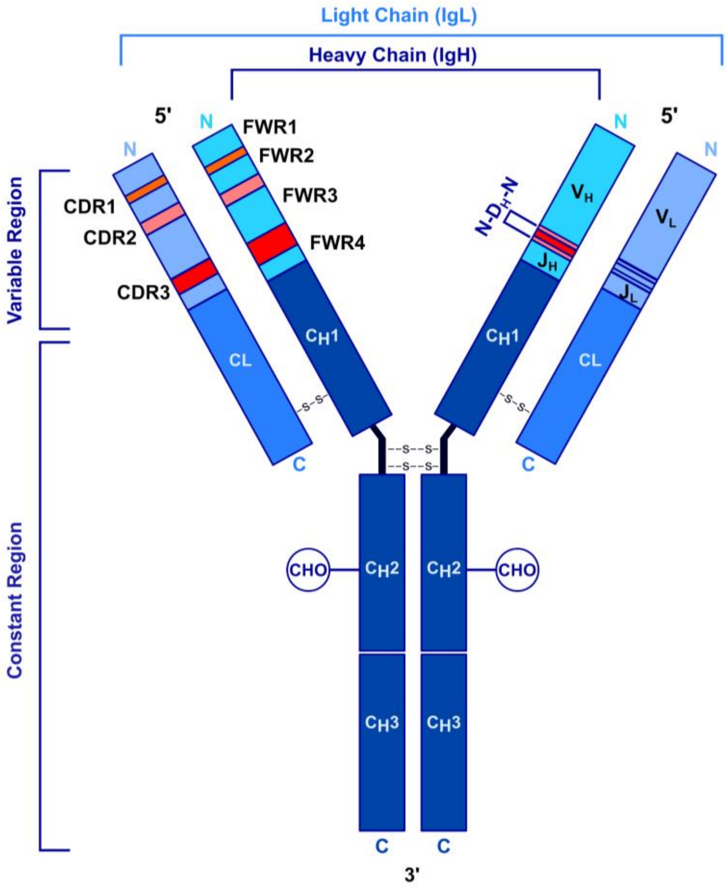
9.5.1 Introduction to the Immune Repertoire

The human immune system shows an astonishing ability to produce a diverse array of antibodies capable of recognizing an extensive range of antigenic structures. The immune repertoire encompasses all B cell receptors (BCRs or antibodies) and T cell receptors (TCRs) within an individual. This assembly of receptors and cells emerges through various processes, including, but not limited to, combinatorial diversity, somatic recombination, class switching, and somatic hypermutation. These receptors display remarkable variability, and the advent of high-throughput sequencing has revolutionized our capacity to scrutinize them in greater detail (Calis & Rosenberg, 2014).

Currently, our understanding of the human immune response to vaccinations, cancer, and viral infections relies heavily on advanced "omics" technologies. These innovative methodologies facilitate the quantification of genetic behavior, mRNA (single-cell transcriptomics), proteins (proteomics), metabolites (metabolomics), cells (mass cytometry), and epigenetic modifications (ATAC-seq). In conjunction with computational approaches, bioinformatics plays an indispensable role in the exploration of these domains. (Pulendran & Davis, 2020).

When focusing on the diversity of antigen receptors, specifically TCRs and BCRs, it is of paramount importance to thoroughly explore the mechanisms that foster such diversity. This process begins by estimating the theoretical diversity, which indicates the potential of germline segments to combine. The maximal amino acid diversity of immune repertoires is estimated to be approximately 10^{140}, calculated as $20^{110} \times 2$. This calculation considers the 20 amino acids, the 110-amino-acid variable region of immune receptors, and the two variable regions constituting each receptor (IGVL-IGVH for B lymphocytes or TCRVα-TCRVβ for T lymphocytes). However, it is important to note that this vast diversity in humans is confined by initial segments of V, D, and J genes, resulting in a conceivable diversity range between 10^{13} y 10^{18}. At any given moment, only a fraction of this potential diversity can manifest in an individual due to limitations in the number of circulating B and T cells (humanos: $10^{11\text{-}12}$) and the count of distinct clones, as defined by clonotypes, which approximates 10^{9} in humans and $10^{6\text{-}7}$ in mice (Miho et al., 2018).

Figure 9.3 BCR Structure. The soluble form of the B-cell receptor (BCR) is the immunoglobulin, consisting of two identical heavy chains (IgH) and two light chains (IgL) connected by disulfide (S-S) bonds. The variable region of the heavy chain comprises the VDJ segment, while the light chain's variable region comprises only the VJ segments. The junction of the V(D)J segments for IgH and VJ segments for IgL forms the CDR3 region. The variable region of the antibody is composed of three complementarity-determining regions (CDRs) and four framework regions (FRs) of both IgH and IgL. The constant region consists of two or three constant domains from the heavy chains, depending on the antibody's class



9.5.2 Structure of Immune Receptors

9.5.2.1 B cell receptor (BCR)

The B cell receptor, known as BCR, is composed of two identical heavy chains (IgH) and two identical light chains (IgL) that are linked by noncovalent interactions and disulfide bonds. The IgL chains consist of two types, lambda (Igλ) and kappa (Igκ), and their relative proportion varies among species, with an average ratio of 2:1 in humans (Schroeder & Cavacini, 2010) (Fig. 3).

The BCR plays a crucial role in recognizing and responding to antigens, inducing a cascade of events that leads to cell proliferation and differentiation into plasma B cells. Subsequently, plasma cells produce immunoglobulins, commonly known as antibodies, which represent the secreted form of the BCR and effectively neutralize pathogens.

Antibodies can be classified into two functional regions: the variable region, located at the amino terminus, responsible for recognizing and binding antigens, and the constant region, determining the antibody's isotype and effector functions. BCRs on lymphocyte surfaces lack effector functions due to the constant region being embedded in the cell membrane. The heavy chain isotype type determines the antibody's functional properties, and five major types are recognized: IgM, IgD, IgG, IgA, and IgE (Schroeder & Cavacini, 2010). Within the antibody's variable region, there are three complementarity-determining regions (CDRs) and four IgH and IgL Frameworks (FRs). The CDRs, mainly CDR3, exhibit the most variability and are crucial for pathogen recognition. The constant region comprises two or three constant domains from both heavy chains, and the number varies based on the antibody's class.

9.*5.2.2 T Cell Receptor (TCR)*

The T cell receptor (TCR) is composed of two chains that bear resemblance to immunoglobulins. However, a notable distinction exists: TCRs are not secreted; instead, they remain consistently associated with the cell membrane. Consequently, they traverse a segment in both chains that spans the lipid bilayer of the membrane, encompassing a small intracellular portion. These two chains are identified as TCRα and TCRβ and placed next to each other through disulfide bonds. CD3, CD4, and CD8 represent specific molecules present on the surface of T cells that stabilize both TCR-mediated interactions and intracellular communication.

9.5.3 Library Preparation and Sequencing Platforms for Repertoire Studies

The analysis of immune repertoires can involve the selection of DNA or RNA as source material. The integration of antigen receptor sequences into sequencing libraries is typically achieved through targeted PCR amplification. However, amplifying the variable region sequences of highly diverse BCRs and TCRs presents a significant challenge. Designing PCR strategies that allow unbiased and complete amplification of these exceptionally variable receptors is particularly demanding. This challenge becomes critical in experiments that necessitate high-throughput massive sequencing for quantifying receptors and lymphocyte clones, as imbalances in amplification efficiencies can introduce biased clonal frequency measurements. To tackle these difficulties, diverse amplification approaches have been utilized. These methods encompass multiplex PCR with intricate combinations of direct primers, multi-step PCR, and utilization of multiple primer mixes.

Alternatively, PCR initiated from engineered adapter sites has also been employed, particularly for RNA input. This technique involves introducing a consistent adapter sequence into the variable regions, followed by PCR amplification using a single forward primer and reverse primers that target J segments or constant regions. This innovative approach effectively mitigates amplification bias and generates a library encompassing the complete variable region (ORF), suitable for functional assessments of TCRs or antibodies. Given the unique characteristics of the amplicons to be sequenced, the selection of a suitable sequencing technology becomes crucial. The landscape of sequencing methodologies continually advances in terms of depth and precision. Consequently, depending on the sequencing technology employed, the intended application and its implementation can vary (see Table 9.1).

Table 9.1 Common Platforms Used for Sequencing the Immune Repertoire (Modify of Chaudhary & Wesemann, 2018).

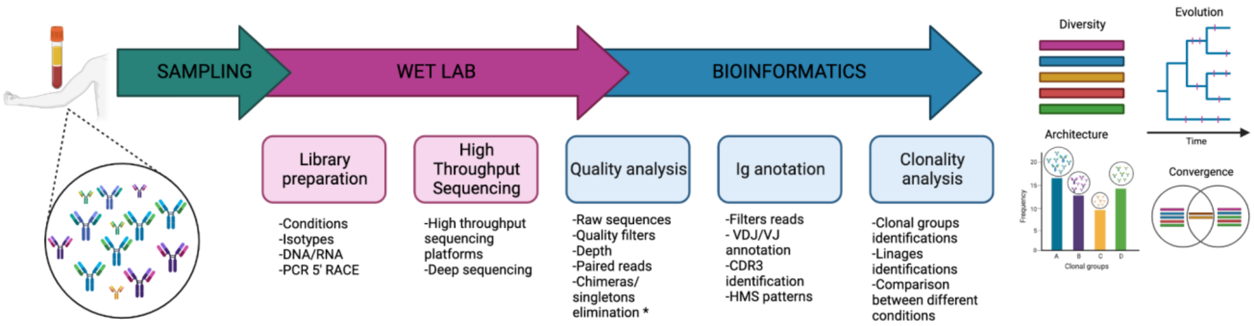
High Sequencing platform	Method	Read size	Depth sequencing	Reads per run	Covered Region	Accuracy and error rate (%)	Error
Roche's 454 GS FLX	Pyrosequencing	700 pb	450 Mb	~1 millon	FR1- Constant region	1	Indel
Illumina MiSeq	Dye terminator sequencing	300 pb x 2	13.2-15 Gb	~ 3 million	FR1- Constant region	~0.1	Substitution
Illumina HiSeq	Dye terminator sequencing	250 pb x 2	500-1000 Gb	~ 2 billion	FR1- Constant region	~0.1	Substitution
Ion torrent	Synthesis (detects H+ ions)	> 100 - 200 pb	30 Mb - 2 Gb	~60 -80 million	FR3- Constant region	~1	Indel
PacBio	Synthesis (fluorescence tag attached to phosphate chain)	860 - 1100 pb	5-10 GB	~0.01 million	Amplification of Linked Heavy and Light Chains	~13	Indel

Note: Each platform has its own strengths and limitations, and the choice of platform depends on the specific research goals and characteristics of the immunological repertoire being studied

9.5.4 Immune Repertoire Analysis

The immunological repertoire is characterized by its diversity, architecture, evolution, and convergence (Fig. 9.9.4). The wide-ranging diversity of immune repertoires is a crucial attribute that facilitates the recognition of a broad spectrum of antigens. To measure the diversity of a repertoire we can use alpha diversity descriptors such as clonotype richness and clonotype abundance. Accurate diversity assessment hinges on the accurate annotation of sequencing reads. The annotation process follows these key steps: (i) identification of the V, D, and J segments, (ii) identification of the positions that define the frameworks and CDRs, (iii) identification of nucleotides inserted and deleted in the binding region, and (iv) quantification of somatic hypermutation (in the context of antibodies). (Miho et al., 2018).

Figure 9.4 Immune repertoire analysis process. In the context of repertoire analysis, sample selection is a crucial step in experimental design. This selection must accurately reflect the study conditions and populations being compared, including any enrichment of B or T cells that may be performed. Typically, the immune repertoire is accessed through peripheral blood samples. The subsequent library preparation step involves the isolation and amplification of genetic material fragments, such as genomic DNA or mRNA (in cases where mRNA is used, an additional step is required to convert RNA to DNA). Following this, the region of interest - typically the variable region - is amplified through PCR 5'RACE. The desired sequencing type is chosen and the necessary steps to add adapters for massive sequencing and barcodes corresponding to each individual/condition are taken. Once raw sequences are obtained, a quality control process is implemented to filter out sequences with a quality Q>20, and eliminate chimeras and singletons. Subsequently, sequenced regions are annotated. VDJ segments for the heavy chain and VJ segments for the light chain are identified, CDR3 is determined, and patterns of somatic hypermutation are compared with the germ line. In clonality analysis, sequences are grouped based on their characteristics into clonal groups and lineages, and subsequently, comparisons are made between different study conditions. With these components, diversity, evolution, architecture, and convergence of the immune repertoire can be studied. Image created with BioRender.com.



The full spectrum of similarity relationships among immune receptor sequences is referred to as the similarity architecture within an immune repertoire. Consequently, the diversity of the immune repertoire, derived from frequency profiles of clonotypes, differs from the sequence similarity architecture, which is based on similarity ratios of clonotypes, regardless of their frequency. The degree of similarity among immune receptors significantly influences the breadth of antigen recognition; as receptor differences increase, the coverage of the antigen space expands. Network theories provide a framework to delve into this subject. Clonal networks are established by defining clonotypes as network nodes, and connections (edges) between clonotypes are established based on specific similarity conditions, measured by a distance, resulting in undirected Boolean networks.

To deduce ancestral evolutionary relationships among individual B cells, lineage trees are constructed from sets of sequences belonging to the same clonotype. A clonal lineage encompasses receptor sequences originating from the same recombination event. When creating lineage trees, a common preliminary step involves pooling sequences with identical V and J genes, as well as the same CDR3 length. In the realm of antibody repertoire phylogenetics, no consensus has been reached regarding the optimal method for inferring lineage evolution. Techniques such as LD, Neighbor-Joining (NJ), Maximum Parsimony (MP), Maximum Likelihood (ML), and Bayesian Inference (BEAST) are employed.

The reconstruction of phylogenetic trees for antibodies necessitates a profound understanding of the physical and temporal dynamics of somatic hypermutation, a process integral to antigen-driven antibody sequence evolution. Mutation statistics can be leveraged to infer the probability of mutation, which is unevenly distributed across the VDJ region of the antibody.

Convergence of immune repertoires refers to the exchange of identical or similar immune receptor sequences between two or more individuals. Sequence convergence can be indicated by the interchange of clonotypes (public clonotypes, full clonal sequence, or pool of clonotypes) or motifs (substrings of sequences). Numerous researchers in this field have endeavored to quantify the degree of convergence of the naive repertoire and the antigen-modified repertoire, employing a wide range of computational approaches that determine sequence similarity between individuals.

9.5.5 Application of Machine Learning in Immune Repertoires

Machine Learning (ML) has become an indispensable subset of Artificial Intelligence (AI) that has gained substantial recognition in recent times due to its application in various research domains. Notably, DeepMind's Alpha Fold tool (Jumper et al., 2021), which is founded on AA, has achieved a significant breakthrough in structural biology. The tool has the capacity to predict the three-dimensional structure of proteins with high precision, which is arguably the most noteworthy contribution of AI to the advancement of scientific knowledge. Moreover, AA has been utilized to forecast susceptibility to ailments such as cancer, disease recurrence, and life expectancy.

In the field of immunology, AA has been employed to scrutinize the response of the adaptive immune system to vaccines and infections and to recognize molecules with potential therapeutic benefits, such as monoclonal antibodies. In this sense, AA has enabled the identification of new monoclonal antibodies *in silico* based on extensive sequencing data of the variable regions repertoire (Greiff et al., 2020). Additionally, AA has facilitated the optimization of binding to the target molecule and comprehension of the biophysical properties of antigen-antibody interaction.

Recently, AA has been utilized to identify potential candidate monoclonal antibodies to treat COVID-19 based on the CDR3 sequence, as well as to identify biomarkers that correlate with disease severity (Magar et al., 2021).

Antibody therapeutics have become highly effective biotherapeutics, securing four of the top ten therapeutics in terms of sales in 2021. Furthermore, antibody-based biotherapeutics, comprising antibody-drug conjugates and bispecific antibodies, have exhibited potential as therapeutic modalities. Traditionally, experimental approaches, such as phage or yeast display, and animal immunization have propelled the discovery and development of antibodies. However, these methods are time-consuming and labor-intensive and have various limitations, including challenges in specifying antibody binding sites (epitopes) and manufacturing antibodies at scale. Despite the various reported strategies to optimize the experimental workflow, significant challenges remain. In recent years, computational and AI-based methodologies have gained importance at various stages of the antibody development workflow. This is analogous to small molecule drug development, where computational methods have made significant progress. Specifically, the prediction of interactions between drugs and therapeutic targets has considerably benefited from the marked improvement in the performance of computational methods.

9.6 Methods

There is an abundance of techniques and computer software currently available, and the challenge in sequence analysis is selecting the most suitable one. The databases selected, particularly for sequence annotation (i.e., aligning new sequences with existing ones from repositories), will always serve as a crucial element in the analysis (Escobar-Zepeda et al. 2018). Among the most meticulously curated collections and repositories of ribosomal subunits (including 16S, 18S, 23S, and 28S rRNA gene sequences) is the SILVA-ARB database, which also offers a local sequence manipulation application known as ARB (Quast et al., 2013; Yilmaz et al., 2014). Both the database and the program were utilized in this study to construct the 16S and 18S dendrogram (Fig. 2).

With the advent of RNAseq high throughput sequencing technologies (transcriptomics), the power of genomics data analysis and the global transcription of an organism's collection of mRNAs have increased. However, such analyses require new capabilities in computing infrastructure and large data handling skills on the part of the data analyst.

A typical workflow in RNAseq data analysis commences with the extraction of genetic material from the messenger RNAs and their transport to a massive sequencing center. This center generates text files in the fastq format, much like the case of DNA analysis. The fastq format often employs four lines to describe each sequence, with the first line containing the name of the sequence, the second line containing the actual nucleotide sequence (ATCG), the third line starting with a "+" symbol, which may repeat the sequence's name or display the "+" symbol, and the fourth line containing the sequence's qualities in PREHD 33 format.

When a reference genome is present, the next step is to align the fastq files against the reference genome using software designed for this purpose. Over 100 software are available for this step, but the most commonly used or referenced in articles are bowtie (Langmead & Salzberg, 2012) and BWA (Li & Durbin, 2009), with their selection closely linked to the genetic material used and the size of the sequences in the fastq files. An analysis of the advantages and disadvantages of five highly popular aligners is available at this site: https://www.ecseq.com/support/benchmark.

In the absence of a reference genome, the recommended approach is to perform a *de novo* assembly of the genome or transcript using specialized software. Trinity is a popular software for this purpose, and a highly detailed execution protocol can be found in the protocol published in nature protocols (Haas et al., 2013).

In the third step, a table of transcript abundances is generated whether one has aligned against a known reference genome or with the transcript assembly. This table consists of a transcript in each line and a column for each library that has been sequenced. It is crucial to sequence a basal or "Wild type" condition and a condition against which one wants to contrast, be it an experimental condition (due to a specific environmental condition, a drug, a disease, a tissue, a strain), etc., a mutant or variant. Moreover, it is imperative to consider biological replicates. Multiple publications have shown that the greater the number of replicates, the greater the understanding of the experiment under study (Schurch et al., 2016; Lamarre et al., 2018). Although the methodologies for generating replicates can be expensive or complicated, it is recommended that there should never be less than 3 replicates per condition. It is also vital to avoid the "batch" effect, which is generated when samples of the same condition are prepared at different times, by different people, or in different situations. The greatest variability occurs when the genetic material is extracted, so it is imperative to reduce the variables that could generate greater dispersion (e.g. time, environmental conditions, temperature, sex, size, etc.).

The fourth step involves the differential expression analysis (DEA). There are numerous R-language software packages available for this step, but proficiency in the language is required. Fortunately, we have developed a website for this last step. The DEA analysis begins with the abundance table described in step 3 and defines the value of LFC (logfoldChange or value of change) that is considered necessary, a relevant p-value, and a CPM (counts per million). These three values are set by default in LFC>=1, pvalue<=0.5, and CMP=1, which are common.

Integrative Analysis of Differential Expression for Multiple Experiments (IDEAMEX) (Jiménez-Jacinto et al., 2019) performs differential expression analysis using four Bioconductor packages: DESeq2, EdgeR, NOIseq, and Limma. It also includes a module for integrating results that reports the coincidences in the sets of differentially expressed genes using Venn Diagrams, plain text lists, and heatmaps. All of this is presented in an easy and user-friendly visualization environment, enabling analysis to be carried out without any prior experience in handling the R language.

The condition of samples used in the study of immune repertoires holds paramount importance for ensuring the validity and relevance of research findings. These samples need to accurately reflect the specific conditions under investigation, such as infections, vaccinations, or autoimmune responses, to provide insights applicable to real-world scenarios. Comparable sample conditions are essential for accurate comparisons between populations or experimental conditions, preventing biases and confounding variables. Using representative samples also avoids introducing artifacts, supports data reproducibility, and aids in the development of clinically relevant therapies, vaccines, and diagnostics. Moreover, maintaining consistent sample conditions before and after enrichment processes is crucial to attribute observed changes to the experimental manipulations rather than to technical variations. In essence, the accuracy of antibody repertoire analysis hinges on the fidelity of sample conditions, influencing the reliability, applicability, and clinical significance of the research outcomes. To access a sample of the immune repertoire for study, a multi-step process is undertaken. Initial sample collection involves obtaining biological material rich in immune cells, often from peripheral blood or other relevant tissues. The isolated immune cells are then processed to extract their genetic material, which encodes the antibodies of interest. This genetic material is transformed into sequencing libraries through fragmentation, amplification, and tagging steps. High-throughput sequencing technologies are employed to generate vast amounts of sequence data.

Within the field of immune repertoires, it is important to consider certain limitations, including the depth and quality of sequencing. In cases where sequencing depth is limited or quality is poor, segment allocation is restricted, thereby limiting the characterization of sample diversity. Moreover, while some studies have focused solely on heavy chain clonotypes, few have examined full clonotypes. Finally, the databases used in such studies often fail to reflect the variability of alleles across populations.

Given the vast amount of data generated by High-throughput sequencing technologies in the study of immune repertoires, the application of bioinformatics tools is critical. Currently, a variety of software programs exist that provide the necessary tools for such studies (as outlined in Table 2). Notably, the majority of these software have been developed with a focus on the antibody repertoire, given its high degree of variability.

Table 9.2 Specialized Software for Immunological Repertoire Analysis

Software	Description	Repertoire type	Reference
ImmunediveRsity	Tool based mainly on R language for comprehensive data analysis of the B cell repertoire. Performs clonal and lineage grouping.	BCR	(Cortina-Ceballos et al., 2015)
IMGT/High V-Quest	Web tool that allows rapid identification of the germ line (allele assignment), determination of the structure of TCR and BCR.	BCR, TCR	(Li et al., 2013)
VDJFasta	Tool that uses Hidden Markov Models to determine all CDRs, performs frequency analysis.	BCR	(Glanville et al., 2009)
ImmuneDB	Tool that identifies genes, determines clones, builds lineages and provides information such as selection pressure and mutation analysis.	BCR	(Rosenfeld et al., 2017)
immunarch	An R package designed to analyze TCR and BCR repertoires, designed primarily for medical scientists and bioinformaticians.	BCR, TCR	(Samokhina et al., 2022)
Immcantation	Tool that provides an end-to-end analytical environment for high-throughput AIRR-seq data sets. From raw reads, Python and R packages are provided for preprocessing, population structure determination, and repertoire analysis.	BCR	(Gupta et al., 2015)
IgBLAST	Tool identifies germline gene matches, assesses rearrangements, and delineates IG V domain regions, supporting both nucleotide and protein sequences with parallel database searches for comprehensive insights.	BCR, TCR	(Ye et al., 2013)
IGGalaxy	A web application using the Galaxy GUI, and can be used on a single computer and on a server.	BCR	(Moorhouse et al., 2014)
SONAR	Tool is specifically designed to analyze the development of antibody lineages over time.	BCR	(Schramm et al., 2016)

Bioinformatics analysis follows, encompassing quality control, alignment, and annotation of the sequences to identify key regions like antibody variable domains. Clonality analysis groups similar sequences to reveal immune cell population dynamics, while comparative analysis provides insights across different conditions. Several steps can be identified for the analysis of the immune repertoire. The initial step involves data pre-processing, whereby the primary objective is to rectify sequencing errors and eliminate noise from the raw sequences. This step bears much resemblance to procedures in other types of studies. Subsequently, germline annotation of the crude sequences is carried out, which is deemed one of the most crucial steps as it involves the inference of the correct germline alleles that recombined to produce each TCR/BCR/antibody. Following this, clonal assignment, which is mostly accomplished by CDR3 sequence homology at either the amino acid or nucleotide level, is conducted. In essence, it can be stated that a sequence that originates from the same V and J segment and has the same CDR3 can be classified as belonging to the same clonal group. Within the clonal group, different lineages that give shape to that clonal group can be grouped at the complete sequence level. Lastly, the characteristics of the repertoire can be described in terms of diversity, architecture, evolution, and convergence in the various study conditions.

9.7 Results and Discussion

9.7.1 Application of metagenomics of microorganisms in environmental sciences

The consideration of microbial contributions to carbon fluxes to and from the atmosphere is imperative in all climate change models. The microbial realm has the potential to become a crucial accomplice in endeavors to mitigate the outcomes of human greenhouse gas emissions, as it may be feasible to encourage alterations in microbial activities in various environments to consume more and generate fewer gases that contribute to global warming. To address intricate issues, more interdisciplinary research is required to probe into the relationships between microorganisms, climate change, and human well-being. Microorganisms' adaptation to a warming world may directly affect human well-being through modified patterns of host-microbe interactions, microbial biogeography, and altered terrestrial, aquatic, and urban microbiology. Consequently, omics studies of environmental microbiology play a pivotal role in continuing to comprehend the microbial world and its vast diversity that surpasses that of the macroscopic world (Fig. 2).

9.7.2 Our understanding of fermented foods through metagenomic studies

The investigation of microbial communities has been enhanced by the emergence of new techniques for the molecular characterization of microorganisms. The employment of laboratory isolation and culture methods renders the restoration of the original microbial community unattainable, as community members often survive solely as a consortium, which cannot be isolated. Additionally, the enumeration of species richness present in the sample is skewed due to the inability of only certain species to grow under laboratory culture conditions, leading to the serious issue of pathogenic species not being detected (Fakruddin et al., 2013).

The characterization of microbiotas through the reconstruction of the taxonomic profile by high throughput massive sequencing of metagenomic DNA has highlighted that cultivable organisms are not necessarily the dominant ones (Escobar-Zepeda et al., 2016). These studies, in combination with other omics, have also demonstrated the complexity of the microbiota of fermented foods and the interactions between the populations of fungi, yeasts, and bacteria (Saak et al., 2023). Therefore, the development of starter cultures and simulation of processing conditions for large-scale production must be replaced with efforts to preserve traditional processing techniques.

The implementation of high-performance sequencing technologies in the field of food has provided evidence of the safety of fermented foods produced by traditional methods. The abundance of beneficial bacteria and a low proportion of opportunistic pathogens and antibiotic resistance genes, as compared to non-fermented foods, serve as indicators of high microbiological quality, especially in fermented foods of dairy origin (Xu et al., 2022).

Similarly, the clarification of the functional profile of the microbial communities associated with fermented foods has demonstrated the role of specific species in the degradation of compounds and the production of bacteriocins and other antimicrobials (Yasir et al., 2022; You et al., 2022).

9.7.3 Application of the study of immune repertoires in health

The examination of immune repertoires has been employed to investigate various aspects of health, disease, and vaccination conditions. The primary objective of such studies is to explore a particular signature of the repertoire under specific circumstances, while also providing insight into the immunological processes that take place.

The investigation of the human B cell response to dengue virus (DENV) infection is crucial in understanding serotype-specific protection and cross-reactive subneutralizing response. While serotype-specific protection is advantageous and represents the main objective of vaccination, cross-reactive subneutralizing response has been linked to the development of severe disease, occurring in a small but significant fraction of DENV secondary infections. Primary and secondary infections are associated with the production of polyreactive and cross-reactive IgG antibodies. Studies of the antibody repertoire in DENV indicate that during the acute phase of the disease, there is an increase in the diversity of IgG B cells, and changes in the relative use of the IGHV1-2, IGHV1-18, and IGHV1-69 segments are observed (Godoy-Lozano et al., 2016). Convergent patterns in the antibody repertoire that are specific to DENV have been identified and serve to define prevalent and specific indicators of DENV infection. These immune signatures have the potential to be useful in the development of protein- or nucleic acid-based diagnostic tools designed to detect acute dengue, as well as to evaluate and monitor DENV exposure in endemic communities (Parameswaran et al., 2013).

A primary goal in developing vaccines against swiftly mutating viruses like influenza or HIV is to generate antibodies with the capacity to neutralize a broad spectrum of variants. Nevertheless, it is important to acknowledge that B cell precursors, capable of evolving into broadly neutralizing antibodies (bNAbs), are usually rare in the immune system.

In the specific case of HIV-1 vaccine development, emphasis has been placed on activating B cell receptors to generate naïve antibodies or bNAb precursors, followed by expansion and maturation of intermediate B cell lineages, which ultimately results in the production of bNAbs with high affinity. The conserved regions of the HIV-1 coat glycoprotein trimer, commonly referred to as Env, are the intended targets of bNAbs (Spencer et al., 2021). When present during viral exposure, these antibodies can block infection. The potential therapeutic application of bNAbs is very promising, and efforts are currently underway to facilitate their development for broad clinical use. Studies of immunological repertoires have described these antibodies (Caskey, 2020).

Recently with the COVID-19 pandemic, investigating the modulation of the immune repertoire in this disease has become one of the priorities of the scientific community. Through investigating the antibody repertoire in infected patients, particular patterns of the use of segments were found. These segments include IGHV3-30, IGHV3-53, IGHV3-23, and IGHV3-9 of the heavy chain, and IGKV1-39, IGKV1-33, IGLV3-21, IGLV3-25, and IGLV6-57 of the light chain (Nielsen et al., 2020; Robbiani et al., 2020; Zost et al., 2020). The low percentage of somatic hypermutation at the beginning of the pandemic suggests a primary infection. However, with the advance of the pandemic, highly neutralizing antibodies have been described with a higher somatic hypermutation.

9.7.4 Application of transcriptomics in agronomy

At the National Center for Disciplinary Research in Animal Health and Safety, National Institute of Forestry, Agriculture, and Livestock Research (Mexico), mechanisms of resistance to ivermectin of the parasite *Haemonchus contortus*, which affects small ruminants, were investigated to develop new control and diagnostic strategies. Samples of the nematodes were obtained, and total RNA was extracted and purified with chloroform and isopropanol, then precipitated with 75% ethanol. The RNA concentration was estimated at 3 μg using spectrophotometry. RNA purity and integrity were evaluated through 1% agarose gel electrophoresis, stained with ethidium bromide, and assessed by fluorometry with a Bioanalyzer 2100 following the manufacturer's instructions (Agilent, Santa Clara, CA, EE. UU.). This process aimed to achieve an RNA Integrity Number (RIN) ≥ 6, which was deemed sufficient for the assay

The *de novo* assembly was conducted using all RNAseq sequencing and samples of the two *H. contortus* strains, IVMs, and IVMr. The nucleotide sequence data is available in GenBank, under BioProject PRJNA877658. The bioinformatics analysis was carried out using the computational cluster of the Massive Sequencing and Bioinformatics University Unit at the Institute of Biotechnology/UNAM. First, the quality of the sequences was analyzed using the FastQC v0.11.8 software, which indicated no presence of adapters and confirmed that the qualities were on average above one Q30. The sequences were then aligned against the reference assembly genome GCA_000469685 using the Smalt v0.74 software, but the alignment percentage was very low (56.11%). Consequently, the transcript was assembled *de novo* using Trinity software v 3.0. Subsequently, the RSEM software was utilized to generate the table of abundances. The data was then analyzed with four methods from the EdgeR Bioconductor package, NOISeq, limma, and DESeq2, and the results provided by DESeq2 were the most consistent with the other methods. Finally, the annotation was conducted with the Trinotate software (Griffith et al., 2015).

The novel information presented in this study indicates the presence of significant genetic diversity in various populations of *H. contortus*, which could be a result of several factors, including anthelmintic drug pressure, the prevalence of nematodes in their environments, and geographic regions. This diversity may also be a consequence of different factors related to host-parasite interactions. The species *H. contortus* has evolved genetic properties that enable them to withstand anthelmintic drugs and evade the host's immune response through specific up/down-regulated genes.

9.8 Conclusions

The limitations of the repertoire sequencing, metagenomic, and transcriptomic analyses are numerous, yet it is crucial to consider sequencing depth, sequencing quality, the use of population-specific databases, and appropriate computational methods for sample type. These biases constrain the accurate characterization of immunological and microbial diversity.

For many scientists, data analysis is a challenging task, as most available tools are implemented in a UNIX-based environment and require programming languages like R, Python, or Perl, which are intended for students with at least a basic understanding of these languages and coding skills.

However, alternative tools on the web such as Galaxy (The Galaxy Community et al., 2022), and servers like MGnify (Richardson et al., 2023) provide both public and private repositories of data analysis that are accessible upon request. Additionally, these resources offer a collection of Jupyter Notebooks that are easily comprehensible and navigable (http://notebooks.mgnify.org).

Differential expression analysis and RNA-Seq have become popular and useful methods to evaluate gene expression changes in any organism. The IDEAMEX web server was developed to address the above-mentioned difficulties (Jiménez-Jacinto et al., 2019). IDEAMEX requires a raw headcount table for as many replicates and conditions as required, enabling the user to select which conditions are compared. The whole process consists of three main steps: (i) Data analysis, which provides a preliminary quality control analysis based on the data distribution per sample using various types of graphs; (ii) Differential expression, which performs differential expression analysis with or without error for a batch effect and generates reports for each method, using the Bioconductor, NOISeq, limma-Voom, DESeq2, and edgeR packages; (iii) Integration of results, which reports the integrated results using different graphical outputs, including correlograms, heatmaps, Venn diagrams, and gene lists in text files. IDEAMEX provides easy interaction during the analysis process, error tracing, and debugging by generating output log files. The server is currently accessible on http://www.uusmb.unam.mx/ideamex/, where documentation and example input files are provided, and can help researchers with no prior bioinformatics background perform differential expression analysis of RNAseq data easily.

Bioinformatics has brought about a paradigm shift across diverse domains, wielding transformative influence in biotechnology, medicine, environmental science, and agriculture. In biotechnology, its progress is evident in the accelerated design of genetically modified organisms, fostering innovations in enzymes, biofuels, and biomaterials with far-reaching implications for sustainable practices. Meanwhile, in the medical area, bioinformatics stands as a guiding tool, shaping personalized healthcare through in-depth genomics, nutrigenomics, pharmacogenomics, clinical data analysis, and immune repertoire analysis. By pinpointing disease biomarkers, it facilitates early diagnosis and tailored treatments, and its role in drug and vaccine discovery expedites the possibilities to improve the treatment of diseases.

The reach of bioinformatics extends to environmental science, where its analytical power might uncover intricate ecological interactions. Metagenomics and transcriptomics, enabled by bioinformatics, shed light on microbial communities in ecosystems, enabling comprehension of their roles in nutrient cycling and ecosystem stability. This knowledge, in turn, informs strategies for pollution control, bioremediation, and sustainable resource management. In agriculture, bioinformatics has heralded a new era by contributing to the development of genetically enhanced crop varieties, displaying augmented yields, disease resistance, and improved nutrition. Additionally, precision agricultural techniques, harnessed through bioinformatics, optimize resource allocation, fostering resource-efficient practices and bolstering global food security.

As technology and knowledge continue to evolve, the interdisciplinary partnership between life sciences and computational analysis is poised to reshape these fields even further, propelling us into a future characterized by sustainable solutions, personalized interventions, and profound scientific insights.

References

Bamforth, C. (2005). *Alimentos, ferementación y microorganismos* (Acribia). https://www.editorialacribia.com/libro/alimentos-fermentacion-y-microorganismos_53995/

Calis, J. J. A., & Rosenberg, B. R. (2014). Characterizing immune repertoires by high throughput sequencing: Strategies and applications. *Trends in Immunology*, *35*(12), 581–590. https://doi.org/10.1016/j.it.2014.09.004

Caskey, M. (2020). Broadly neutralizing antibodies for the treatment and prevention of HIV infection. *Current Opinion in HIV and AIDS*, *15*(1), 49–55. https://doi.org/10.1097/COH.0000000000000600

Chaudhary, N., & Wesemann, D. R. (2018). Analyzing Immunoglobulin Repertoires. *Frontiers in Immunology*, *9*, 462. https://doi.org/10.3389/fimmu.2018.00462

Chiu, L., Bazin, T., Truchetet, M.-E., Schaeverbeke, T., Delhaes, L., & Pradeu, T. (2017). Protective Microbiota: From Localized to Long-Reaching Co-Immunity. *Frontiers in Immunology*, *8*, 1678. https://doi.org/10.3389/fimmu.2017.01678

Christensen, C. M., Kok, C. R., Auchtung, J. M., & Hutkins, R. (2022). Prebiotics enhance persistence of fermented-food associated bacteria in in vitro cultivated fecal microbial communities. *Frontiers in Microbiology*, *13*, 908506. https://doi.org/10.3389/fmicb.2022.908506

Cortina-Ceballos, B., Godoy-Lozano, E. E., Sámano-Sánchez, H., Aguilar-Salgado, A., Velasco-Herrera, M. D. C., Vargas-Chávez, C., Velázquez-Ramírez, D., Romero, G., Moreno, J., Téllez-Sosa, J., & Martínez-Barnetche, J. (2015). Reconstructing and mining the B cell repertoire with ImmunediveRsity. *MAbs*, *7*(3), 516–524. https://doi.org/10.1080/19420862.2015.1026502

Cuamatzin-García, L., Rodríguez-Rugarcía, P., El-Kassis, E. G., Galicia, G., Meza-Jiménez, M. D. L., Baños-Lara, Ma. D. R., Zaragoza-Maldonado, D. S., & Pérez-Armendáriz, B. (2022). Traditional Fermented Foods and Beverages from around the World and Their Health Benefits. *Microorganisms*, *10*(6), 1151. https://doi.org/10.3390/microorganisms10061151

Dayhoff, M. O., & National Biomedical Research Foundation (Eds.). (1979). *Atlas of protein sequence and structure. Vol. 5, Suppl. 3* (Vol. 5). National Biomedical Research Foundation.

Erenstein, O., Chamberlin, J., & Sonder, K. (2021). Estimating the global number and distribution of maize and wheat farms. *Global Food Security*, *30*, 100558. https://doi.org/10.1016/j.gfs.2021.100558

Escobar-Zepeda, A., Godoy-Lozano, E. E., Raggi, L., Segovia, L., Merino, E., Gutiérrez-Rios, R. M., Juarez, K., Licea-Navarro, A. F., Pardo-Lopez, L., & Sanchez-Flores, A. (2018). Analysis of sequencing strategies and tools for taxonomic annotation: Defining standards for progressive metagenomics. *Scientific Reports*, *8*. https://doi.org/10.1038/s41598-018-30515-5

Escobar-Zepeda, A., Sanchez-Flores, A., & Quirasco Baruch, M. (2016). Metagenomic analysis of a Mexican ripened cheese reveals a unique complex microbiota. *Food Microbiology*, *57*, 116–127. https://doi.org/10.1016/j.fm.2016.02.004

Fakruddin, Md., Mannan, K. S. B., & Andrews, S. (2013). Viable but Nonculturable Bacteria: Food Safety and Public Health Perspective. *ISRN Microbiology*, *2013*, 1–6. https://doi.org/10.1155/2013/703813

Glanville, J., Zhai, W., Berka, J., Telman, D., Huerta, G., Mehta, G. R., Ni, I., Mei, L., Sundar, P. D., Day, G. M. R., Cox, D., Rajpal, A., & Pons, J. (2009). Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proceedings of the National Academy of Sciences*, *106*(48), 20216–20221. https://doi.org/10.1073/pnas.0909775106

Godoy-Lozano, E. E., Téllez-Sosa, J., Sánchez-González, G., Sámano-Sánchez, H., Aguilar-Salgado, A., Salinas-Rodríguez, A., Cortina-Ceballos, B., Vivanco-Cid, H., Hernández-Flores, K., Pfaff, J. M., Kahle, K. M., Doranz, B. J., Gómez-Barreto, R. E., Valdovinos-Torres, H., López-Martínez, I., Rodriguez, M. H., & Martínez-Barnetche, J. (2016). Lower IgG somatic hypermutation rates during acute dengue virus infection is compatible with a germinal center-independent B cell response. *Genome Medicine*, *8*, 23. https://doi.org/10.1186/s13073-016-0276-1

Greiff, V., Yaari, G., & Cowell, L. G. (2020). Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. *Current Opinion in Systems Biology*, *24*, 109–119. https://doi.org/10.1016/j.coisb.2020.10.010

Griffith, M., Walker, J. R., Spies, N. C., Ainscough, B. J., & Griffith, O. L. (2015). Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. *PLOS Computational Biology*, *11*(8), e1004393. https://doi.org/10.1371/journal.pcbi.1004393

Gupta, N. T., Vander Heiden, J. A., Uduman, M., Gadala-Maria, D., Yaari, G., & Kleinstein, S. H. (2015). Change-O: A toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics*, *31*(20), 3356–3358. https://doi.org/10.1093/bioinformatics/btv359

Gutiérrez-Pérez, E. D., Vázquez-Juárez, R., Magallón-Barajas, F. J., Martínez-Mercado, M. Á., Escobar-Zepeda, A., & Magallón-Servín, P. (2022). How a holobiome perspective could promote intensification, biosecurity and eco-efficiency in the shrimp aquaculture industry. *Frontiers in Marine Science*, *9*, 975042. https://doi.org/10.3389/fmars.2022.975042

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews*, *68*(4), 669–685. https://doi.org/10.1128/MMBR.68.4.669-685.2004

Hansen, E. B. (2002). Commercial bacterial starter cultures for fermented foods of the future. *International Journal of Food Microbiology, 78*(1–2), 119–131. https://doi.org/10.1016/S0168-1605(02)00238-6

Jiménez-Jacinto, V., Sanchez-Flores, A., & Vega-Alvarado, L. (2019). Integrative Differential Expression Analysis for Multiple EXperiments (IDEAMEX): A Web Server Tool for Integrated RNA-Seq Data Analysis. *Frontiers in Genetics, 10*, 279. https://doi.org/10.3389/fgene.2019.00279

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kogut, M. H., Lee, A., & Santin, E. (2020). Microbiome and pathogen interaction with the immune system. *Poultry Science, 99*(4), 1906–1913. https://doi.org/10.1016/j.psj.2019.12.011

Lahiri, D., Nag, M., Dutta, B., Sarkar, T., Pati, S., Basu, D., Abdul Kari, Z., Wei, L. S., Smaoui, S., Wen Goh, K., & Ray, R. R. (2022). Bacteriocin: A natural approach for food safety and food security. *Frontiers in Bioengineering and Biotechnology, 10*, 1005918. https://doi.org/10.3389/fbioe.2022.1005918

Lamarre, S., Frasse, P., Zouine, M., Labourdette, D., Sainderichin, E., Hu, G., Le Berre-Anton, V., Bouzayen, M., & Maza, E. (2018). Optimization of an RNA-Seq Differential Gene Expression Analysis Depending on Biological Replicate Number and Library Size. *Frontiers in Plant Science, 9*, 108. https://doi.org/10.3389/fpls.2018.00108

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England), 25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, S., Lefranc, M.-P., Miles, J. J., Alamyar, E., Giudicelli, V., Duroux, P., Freeman, J. D., Corbin, V. D. A., Scheerlinck, J.-P., Frohman, M. A., Cameron, P. U., Plebanski, M., Loveland, B., Burrows, S. R., Papenfuss, A. T., & Gowans, E. J. (2013). IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nature Communications, 4*(1), 2333. https://doi.org/10.1038/ncomms3333

Limborg, M. T., Alberdi, A., Kodama, M., Roggenbuck, M., Kristiansen, K., & Gilbert, M. T. P. (2018). Applied Hologenomics: Feasibility and Potential in Aquaculture. *Trends in Biotechnology, 36*(3), 252–264. https://doi.org/10.1016/j.tibtech.2017.12.006

Magar, R., Yadav, P., & Barati Farimani, A. (2021). Potential neutralizing antibodies discovered for novel corona virus using machine learning. *Scientific Reports, 11*(1), 5261. https://doi.org/10.1038/s41598-021-84637-4

Medzhitov, R. (2007). Recognition of microorganisms and activation of the immune response. *Nature, 449*(7164), 819–826. https://doi.org/10.1038/nature06246

Miho, E., Yermanos, A., Weber, C. R., Berger, C. T., Reddy, S. T., & Greiff, V. (2018). Computational Strategies for Dissecting the High-Dimensional Complexity of Adaptive Immune Repertoires. *Frontiers in Immunology, 9*, 224. https://doi.org/10.3389/fimmu.2018.00224

Moorhouse, M. J., Van Zessen, D., IJspeert, H., Hiltemann, S., Horsman, S., Van Der Spek, P. J., Van Der Burg, M., & Stubbs, A. P. (2014). ImmunoGlobulin galaxy (IGGalaxy) for simple determination and quantitation of immunoglobulin heavy chain rearrangements from NGS. *BMC Immunology, 15*(1), 59. https://doi.org/10.1186/s12865-014-0059-7

Nielsen, S. C. A., Yang, F., Jackson, K. J. L., Hoh, R. A., Röltgen, K., Jean, G. H., Stevens, B. A., Lee, J.-Y., Rustagi, A., Rogers, A. J., Powell, A. E., Hunter, M., Najeeb, J., Otrelo-Cardoso, A. R., Yost, K. E., Daniel, B., Nadeau, K. C., Chang, H. Y., Satpathy, A. T., … Boyd, S. D. (2020). Human B Cell Clonal Expansion and Convergent Antibody Responses to SARS-CoV-2. *Cell Host & Microbe, 28*(4), 516-525.e5. https://doi.org/10.1016/j.chom.2020.09.002

Parameswaran, P., Liu, Y., Roskin, K. M., Jackson, K. K. L., Dixit, V. P., Lee, J.-Y., Artiles, K. L., Zompi, S., Vargas, M. J., Simen, B. B., Hanczaruk, B., McGowan, K. R., Tariq, M. A., Pourmand, N., Koller, D., Balmaseda, A., Boyd, S. D., Harris, E., & Fire, A. Z. (2013). Convergent Antibody Signatures in Human Dengue. *Cell Host & Microbe, 13*(6), 691–700. https://doi.org/10.1016/j.chom.2013.05.008

Perry, W. B., Lindsay, E., Payne, C. J., Brodie, C., & Kazlauskaite, R. (2020). The role of the gut microbiome in sustainable teleost aquaculture. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1926), 20200184. https://doi.org/10.1098/rspb.2020.0184

Pulendran, B., & Davis, M. M. (2020). The science and medicine of human immunology. *Science*, *369*(6511), eaay4014. https://doi.org/10.1126/science.aay4014

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research, 41*(D1), D590–D596. https://doi.org/10.1093/nar/gks1219

Reinders, M. J., Banovic, M., & Guerrero, L. (2019). Introduction. En *Innovations in Traditional Foods* (pp. 1–26). Elsevier. https://doi.org/10.1016/B978-0-12-814887-7.00001-0

Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, M. L., Burdett, T., Burgin, J., Caballero-Pérez, J., Cochrane, G., Colwell, L. J., Curtis, T., Escobar-Zepeda, A., Gurbich, T. A., Kale, V., Korobeynikov, A., Raj, S., Rogers, A. B., Sakharova, E., Sanchez, S., … Finn, R. D. (2023). MGnify: The microbiome sequence data analysis resource in 2023. *Nucleic Acids Research, 51*(D1), D753–D759. https://doi.org/10.1093/nar/gkac1080

Robbiani, D. F., Gaebler, C., Muecksch, F., Lorenzi, J. C. C., Wang, Z., Cho, A., Agudelo, M., Barnes, C. O., Gazumyan, A., Finkin, S., Hägglöf, T., Oliveira, T. Y., Viant, C., Hurley, A., Hoffmann, H.-H., Millard, K. G., Kost, R. G., Cipolla, M., Gordon, K., … Nussenzweig, M. C. (2020). Convergent antibody responses to SARS-CoV-2 in convalescent individuals. *Nature, 584*(7821), 437–442. https://doi.org/10.1038/s41586-020-2456-9

Rosenfeld, A. M., Meng, W., Luning Prak, E. T., & Hershberg, U. (2017). ImmuneDB: A system for the analysis and exploration of high-throughput adaptive immune receptor sequencing data. *Bioinformatics*, *33*(2), 292–293. https://doi.org/10.1093/bioinformatics/btw593

Saak, C. C., Pierce, E. C., Dinh, C. B., Portik, D., Hall, R., Ashby, M., & Dutton, R. J. (2023). Longitudinal, Multi-Platform Metagenomics Yields a High-Quality Genomic Catalog and Guides an *In Vitro* Model for Cheese Communities. *MSystems*, *8*(1), e00701-22. https://doi.org/10.1128/msystems.00701-22

Samokhina, M., Popov, A., Ivan-Immunomind, Nazarov, V. I., Immunarch.Bot, Rumynskiy, E., Gracecodeadventures, Tsvvas, & Zarodniuk, M. (2022). *immunomind/immunarch: Immunarch 0.9.0* (0.9.0) [Software]. Zenodo. https://doi.org/10.5281/ZENODO.3367200

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

Schramm, C. A., Sheng, Z., Zhang, Z., Mascola, J. R., Kwong, P. D., & Shapiro, L. (2016). SONAR: A High-Throughput Pipeline for Inferring Antibody Ontogenies from Longitudinal Sequencing of B Cell Transcripts. *Frontiers in Immunology*, *7*. https://doi.org/10.3389/fimmu.2016.00372

Schroeder, H. W., & Cavacini, L. (2010). Structure and function of immunoglobulins. *Journal of Allergy and Clinical Immunology*, *125*(2), S41–S52. https://doi.org/10.1016/j.jaci.2009.09.046

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., Blaxter, M., & Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA (New York, N.Y.)*, *22*(6), 839–851. https://doi.org/10.1261/rna.053959.115

Soemarie, Y., Milanda, T., & Barliana, M. (2021). Fermented foods as probiotics: A review. *Journal of Advanced Pharmaceutical Technology & Research*, *12*(4), 335. https://doi.org/10.4103/japtr.japtr_116_21

Spencer, D. A., Shapiro, M. B., Haigwood, N. L., & Hessell, A. J. (2021). Advancing HIV Broadly Neutralizing Antibodies: From Discovery to the Clinic. *Frontiers in Public Health*, *9*, 690017. https://doi.org/10.3389/fpubh.2021.690017

The Galaxy Community, Afgan, E., Nekrutenko, A., Grüning, B. A., Blankenberg, D., Goecks, J., Schatz, M. C., Ostrovsky, A. E., Mahmoud, A., Lonie, A. J., Syme, A., Fouilloux, A., Bretaudeau, A., Nekrutenko, A., Kumar, A., Eschenlauer, A. C., DeSanto, A. D., Guerler, A., Serrano-Solano, B., … Briggs, P. J. (2022). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research*, *50*(W1), W345–W351. https://doi.org/10.1093/nar/gkac247

Tiedje, J. M., Bruns, M. A., Casadevall, A., Criddle, C. S., Eloe-Fadrosh, E., Karl, D. M., Nguyen, N. K., & Zhou, J. (2022). Microbes and Climate Change: A Research Prospectus for the Future. *mBio*, *13*(3), e00800-22. https://doi.org/10.1128/mbio.00800-22

Whitman, W. B., Coleman, D. C., & Wiebe, W. J. (1998). Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences*, *95*(12), 6578–6583. https://doi.org/10.1073/pnas.95.12.6578

Woese, C. R., Dugre, D. H., Saxinger, W. C., & Dugre, S. A. (1966). The molecular basis for the genetic code. *Proceedings of the National Academy of Sciences of the United States of America*, *55*(4), 966–974. https://www.pnas.org/doi/abs/10.1073/pnas.55.4.966

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, *74*(11), 5088–5090. https://doi.org/10.1073/pnas.74.11.5088

Xu, M., Su, S., Zhang, Z., Jiang, S., Zhang, J., Xu, Y., & Hu, X. (2022). Two sides of the same coin: Meta-analysis uncovered the potential benefits and risks of traditional fermented foods at a large geographical scale. *Frontiers in Microbiology*, *13*, 1045096. https://doi.org/10.3389/fmicb.2022.1045096

Yasir, M., Al-Zahrani, I. A., Bibi, F., Abd El Ghany, M., & Azhar, E. I. (2022). New insights of bacterial communities in fermented vegetables from shotgun metagenomics and identification of antibiotic resistance genes and probiotic bacteria. *Food Research International*, *157*, 111190. https://doi.org/10.1016/j.foodres.2022.111190

Ye, J., Ma, N., Madden, T. L., & Ostell, J. M. (2013). IgBLAST: An immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research*, *41*(W1), W34–W40. https://doi.org/10.1093/nar/gkt382

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., & Glöckner, F. O. (2014). The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Research*, *42*(D1), D643–D648. https://doi.org/10.1093/nar/gkt1209

You, L., Yang, C., Jin, H., Kwok, L.-Y., Sun, Z., & Zhang, H. (2022). Metagenomic features of traditional fermented milk products. *LWT*, *155*, 112945. https://doi.org/10.1016/j.lwt.2021.112945

Zost, S. J., Gilchuk, P., Chen, R. E., Case, J. B., Reidy, J. X., Trivette, A., Nargi, R. S., Sutton, R. E., Suryadevara, N., Chen, E. C., Binshtein, E., Shrihari, S., Ostrowski, M., Chu, H. Y., Didier, J. E., MacRenaris, K. W., Jones, T., Day, S., Myers, L., … Crowe, J. E. (2020). Rapid isolation and profiling of a diverse panel of human monoclonal antibodies targeting the SARS-CoV-2 spike protein. *Nature Medicine*, *26*(9), 1422–1427. https://doi.org/10.1038/s41591-020-0998-x.