Received: 17 August 2024 | Revised: 17 October 2024 | Accepted: 31 October 2024 | Published online: 13 November 2024

REVIEW

Medinformatics
2024, Vol. 00(00) 1–18
DOI: 10.47852/bonviewMEDIN42024120

BON VIEW PUBLISHING

Machine Learning in Genomics: Applications in Whole Genome Sequencing, Whole Exome Sequencing, Single-Cell Genomics, and Spatial Transcriptomics

Saheed Adegbola Adeyanju[1],* and Taiwo Temitope Ogunjobi[2]

[1]*Department of Bioinformatics, Teesside University, UK*

[2]*Department of Biochemistry, Ladoke Akintola University of Technology, Nigeria*

Abstract: The application of machine learning (ML) to genomics has transformed the process of analyzing and interpreting large-scale, complex datasets, leading to important breakthroughs in our knowledge of biological systems. This review provides a comprehensive overview of ML applications in key genomic areas: whole genome sequencing (WGS), whole exome sequencing (WES), single-cell genomics, and spatial transcriptomics. In WGS and WES, ML techniques are employed for variant calling, genome-wide association studies, rare variant analysis, and the prediction of pathogenicity. In single-cell genomics, ML facilitates clustering, trajectory inference, and cell type identification, while in spatial transcriptomics, it aids in deciphering spatial patterns of gene expression and tissue heterogeneity. This review further explores the application of ML in related omics fields, including proteomics, transcriptomics, metagenomics, epigenomics, and microbiome research. These applications encompass protein structure prediction, functional annotation, microbial community profiling, and the analysis of epigenetic modifications. We address the challenges caused by high dimensionality, variability in the data, and the requirement for interpretable ML models when dealing with genomic data. Emerging technologies like explainable AI and federated learning are highlighted for their potential to address these challenges. Additionally, the review addresses ethical considerations, data privacy issues, and the necessity for standardized protocols in ML applications. This comprehensive examination underscores the transformative impact of ML in genomics and highlights its potential to drive future innovations in personalized medicine and biological research.

Keywords: machine learning, genomics, whole genome sequencing (WGS), whole exome sequencing (WES), spatial transcriptomics, metagenomics, epigenomics

1. Introduction

1.1. Overview of machine learning (ML) in genomics

ML has been a transformative force in genomics, revolutionizing the processing and interpretation of complex biological data. By utilizing the massive volumes of high-dimensional data produced by cutting-edge sequencing technology, ML techniques have been integrated into genomics to provide insights that were previously unattainable by conventional statistical methods. This integration improves the capacity to identify trends, forecast results, and produce theories that inform future investigations and therapeutic uses [1]. Numerous approaches are involved in the use of ML in genomics, and each one has advantages over the other when it comes to examining various kinds of genomic data. Supervised learning is widely used for applications like disease outcome prediction, mutation impact prediction, and gene expression classification. It involves training algorithms using labeled data. Among the algorithms commonly used to build predictive models that classify samples, identify biomarkers, and elucidate the biological mechanisms causing genomic variations are support vector machines (SVMs), random forests, and neural networks [2].

Conversely, unsupervised learning methods are vital for uncovering the underlying structures in genomic data in the absence of predetermined labels. Using expression profiles or other genomic markers, clustering techniques like k-means and hierarchical clustering are used to identify different groups of genes or samples. Principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) are two dimensionality reduction approaches that are useful for visualizing complex multi-dimensional data and revealing hidden patterns that guide the functional categorization of genes and pathways. The use of multi-layered neural networks, or deep learning, is a subset of ML that has improved the capabilities of genetic data analysis [3]. While recurrent neural networks and transformers are useful for managing sequential data, such as

*Corresponding author: Saheed Adegbola Adeyanju, Department of Bioinformatics, Teesside University, UK. Email: b1197921@tees.ac.uk

DNA or RNA sequences, convolutional neural networks (CNNs) are skilled at understanding spatial hierarchies in genomic data, such as chromatin interactions and sequence patterns. Deep learning techniques enable the extraction of high-level characteristics from unprocessed genomic data, improving accuracy in tasks like functional genomics, variation interpretation, and alignment of genomic sequences [4].

The use of ML in whole genome sequencing (WGS) and whole exome sequencing (WES) data has one of the biggest effects in genomics. ML techniques are used to predict the functional effects of genetic changes, identify benign and pathogenic variations, and improve the accuracy of variant calling. This feature is especially noteworthy in the context of precision medicine, where patient categorization based on genetic profiles and the identification of novel genetic variations linked to complicated diseases are made possible by ML algorithms. ML tackles the problems caused by the high levels of noise and intrinsic heterogeneity in single-cell RNA sequencing (scRNA-seq) data in the field of single-cell genomics [5].

Traditional methods in genomics, such as rule-based algorithms and statistical tools, predominantly emphasized linear models, sequence alignment, and hypothesis-driven strategies, depending on established rules and simple computations. These methods proved useful for small-scale data but encountered constraints with high-dimensional and big datasets. Contemporary ML techniques, particularly deep learning and ensemble methods, provide enhanced flexibility by autonomously discerning complicated patterns from data without explicit programming, rendering them optimal for the analysis of the extensive and intricate data produced in genomics today. However, present hurdles exist, including the integration of multi-omics data, model interpretability, and scalability of algorithms to manage exponentially expanding genomic datasets. These problems underscore the compelling need for breakthrough ML algorithms that might address these limits and generate new insights into genomics [6].

The application of ML in genomics originated in the late 20th century, as the increasing accessibility of genetic data from initiatives such as the Human Genome Project necessitated sophisticated computer methods. Early efforts concentrated on sequence analysis, including gene prediction and motif identification, employing fundamental statistical models. With the increase in processing power and data accessibility, advanced techniques such as SVMs and hidden Markov models were utilized for gene expression analysis and functional annotation. The advancement of deep learning in the 2010s signified a pivotal transformation, facilitating intricate applications including WGS, variant identification, and precision medicine. Currently, ML is crucial in genomics, facilitating developments in personalized medicine, drug discovery, and single-cell analysis, underpinned by ongoing improvements in algorithms and computational frameworks that accommodate the increasing complexity of genomic data [2].

1.2. Case studies of ML in genomics

1.2.1. WGS for the detection of disease variants

A notable case study examines the application of ML algorithms in WGS to identify harmful variations associated with genetic disorders. DeepVariant, a ML tool created by Google, utilizes deep learning to identify variations from sequencing data. Utilizing millions of genomic samples, it surpasses conventional techniques by enhancing the precision of single nucleotide polymorphism (SNP) and indel identification, which are essential for elucidating the genetic foundations of disorders like cancer, cystic fibrosis, and congenital heart disease [7].

1.2.2. Single-cell genomics and cell types classification

ML algorithms are progressively employed in scRNA-seq to categorize and forecast cell types based on gene expression profiles. An illustrative example is the utilization of Seurat, a widely used R-based software that implements dimensionality reduction and clustering methods to categorize cells into discrete kinds. Researchers can discover unusual cell types, such as cancer stem cells, that contribute to tumor heterogeneity and medication resistance by employing this method. This is essential in cancer research, as targeted medicines are designed based on the molecular and cellular properties of the tumor [8].

1.2.3. Cancer genomics and neoantigen discovery

In cancer genomics, ML has been employed to predict neoantigens tumor-specific antigens that originate from mutations facilitating individualized cancer vaccinations. One notable use is the use of deep learning to model and forecast the binding affinity of neoantigens to major histocompatibility complex molecules, a vital step in assessing whether new antigens might elicit an immune response. For example, NetMHCpan, an ML tool, has been effectively applied to predict possible neoantigens for building cancer immunotherapies, presenting a new approach in tailored treatment.

1.2.4. Genomic prediction in agriculture

ML is also transformative in agricultural genomics, particularly in genomic selection to boost crop yields and disease resistance. In this scenario, ML models like genomic best linear unbiased prediction and random forests have been applied to predict phenotypic features based on genomic data. For example, in maize (Zea mays), ML algorithms have helped forecast resistance to diseases such as Northern Corn Leaf Blight, enabling breeders to select for resistant genotypes with higher accuracy, hence enhancing crop resilience and productivity [1].

1.2.5. Epigenomics and predictive modeling of gene regulation

In epigenomics, ML models are utilized to predict regulatory elements, like as promoters and enhancers, from chromatin immunoprecipitation sequencing (ChIP-seq) and ATAC-seq data. A case in point is the use of CNNs in the DeepSEA framework, which analyzes genomic sequences and epigenomic profiles to identify functional non-coding variations and their possible impact on gene regulation. This use is crucial in understanding complicated disorders like schizophrenia and diabetes, where regulatory alterations in non-coding areas play significant roles [6].

1.3. Importance of genomics in understanding biological systems

Understanding the genetic makeup of all living things through genomics provides a comprehensive understanding of biological systems, which is crucial to comprehending them. Researchers can examine the complex relationships, roles, and interactions between genes in various biological contexts by studying genomes, the entire collection of DNAs within an organism. This in-depth exploration of an organism's genetic code provides important new understandings of physiology, development, and disease causes. Clarifying the genetic basis of complex traits and diseases is one of genomics' most significant contributions categories. Genetic variables that contribute to susceptibility and resistance to diseases can be found by studying changes in the genome, such as copy number variations (CNVs) and SNPs [9]. The identification of genetic variations associated with

complicated ailments like cancer, cardiovascular diseases, and neurological disorders has been made possible through the use of genome-wide association studies (GWAS). These findings provide opportunities for the development of tailored medicine and focused therapeutics in addition to improving our understanding of the origins of disease [10].

Furthermore, genomics makes it easier to investigate gene activity and regulation using a variety of high-throughput methods. Through the use of functional genomics techniques like gene knockdown and overexpression studies, scientists may examine how specific genes affect cellular functions and the development of entire organisms. Transcriptomics offers insights into the regulation of genes in various tissues, developmental stages, and in response to environmental stimuli by analyzing gene expression profiles. Understanding how genetic and epigenetic variables affect cellular activity and organismal features requires knowledge of this material. Genomic research not only clarifies gene function but also provides important insights into the dynamics of gene networks and interactions [11].

2. WGS

ML techniques play a significant role in processing WGS data by automating and optimizing numerous key procedures. Initially, raw WGS data require preprocessing, which includes quality control, read alignment, and error correction. ML models, such as random forests and SVMs, are typically applied to increase the accuracy of these processes by learning patterns from the data. For variant detection, deep learning models like CNNs are used to discover single nucleotide variations (SNVs), insertions, and deletions by examining the alignment data and separating real variants from sequencing errors. Feature selection approaches are utilized to minimize the dimensionality of the huge WGS datasets, enabling more efficient model training. Additionally, ensemble learning approaches can incorporate many algorithms to increase forecast accuracy, making ML a vital tool in handling the complexity and quantity of WGS data [7].

2.1. Definition and significance

WGS is a comprehensive and high-resolution sequencing method that identifies every variation of an organism's DNA (Figure 1) [12]. The whole genome, including all coding and non-coding areas, is covered in great detail by WGS, in contrast to targeted sequencing techniques that concentrate on particular regions of interest. By sequencing both nuclear and mitochondrial DNA, this method enables a comprehensive examination of genetic variants throughout the whole genome. The value of WGS is in its capacity to offer a comprehensive picture of an organism's genetic makeup, which is essential for several applications in both clinical and scientific settings. Researchers can identify a variety of genetic changes, including SNPs, insertions, deletions, and structural variants, by sequencing the entire genome. These differences may have significant effects on our comprehension of hereditary susceptibilities to illnesses, our comprehension of the genetic foundation of intricate features, and our investigation of the functional functions of various genomic areas [13].

One of the primary advantages of WGS is its capacity to uncover rare and novel genetic variants that may be missed by targeted sequencing approaches. This comprehensive coverage is especially helpful when researching complex genetic disorders, as the risk of the condition is influenced by several genetic factors. For example, WGS has helped uncover novel genes and pathways involved in disease processes, clarified the genetic basis of uncommon and inherited illnesses, and identified earlier unidentified disease-associated variations [14]. WGS is also an effective method for investigating the structure and function of the genome. A comprehensive investigation of genomic characteristics, including gene structure, regulatory elements, and non-coding RNA sequences, is made possible by it. Understanding transcriptional networks, gene regulation, and the functional effects of chromosomal changes all depend on this knowledge. Furthermore, WGS makes it easier to investigate epigenetic changes and how they affect gene expression, offering insights into the dynamic interplay between genetic and epigenetic variables. WGS advances our knowledge of genetic diversity and population structure in the context of population genomics and public health [15].

2.2. ML applications in WGS data analysis

The ability of ML to analyze complicated and high-dimensional genetic information has greatly increased the interpretation of WGS data. These computational methods are applied to variation discovery, functional annotation, and illness association, among other WGS-related difficulties. Accurately identifying and classifying genomic variations is one of the main uses of ML in WGS data analysis. To increase the accuracy of variant calling, algorithms like deep learning models and SVMs are utilized to differentiate between actual variations and noise and sequencing artifacts. By using extensive datasets for training, these models can identify patterns and characteristics that point to real genetic differences, which improves the accuracy of the variant identification procedure [16].

By predicting their possible functional consequences, ML approaches can make it easier to interpret variant effects. Neural network- and ensemble-based tools examine the connection between genetic variations and phenotypic outcomes, offering insights into the potential roles of particular mutations in disease. In therapeutic situations, where it facilitates the discovery of pathogenic variations and helps determine their significance for patient diagnosis and therapy, this predictive ability is especially useful. ML is also used to integrate WGS data with other omics data, such as transcriptomics and proteomics, to give a deeper understanding of gene function and regulation [15]. ML methods can anticipate the downstream effects of genetic variants on cellular processes and clarify the biological pathways they affect by merging genomic information with gene expression profiles and protein interaction data. The ability to find new insights into disease mechanisms and possible therapeutic targets is improved by this integrative approach [17].

2.2.1. Variant calling and annotation

Using ML approaches has substantially helped variant calling, a critical step in WGS data interpretation. The process involves identifying genetic variants, such as SNPs and structural alterations, that differ from a reference genome. Two ML techniques that increase the accuracy of variant detection are SVMs and deep neural networks. These techniques learn from large datasets and discern patterns amid sequencing noise and artifacts that signal the presence of genuine genetic variants. Several issues with WGS data processing are addressed by the ML integration in variant calling. Conventional techniques frequently have trouble telling the difference between sequencing errors and true variations, which can result in missed or false positives. By training on annotated variant datasets, ML algorithms overcome these drawbacks and become more adept at distinguishing between real genomic changes and sequencing artifacts [18].

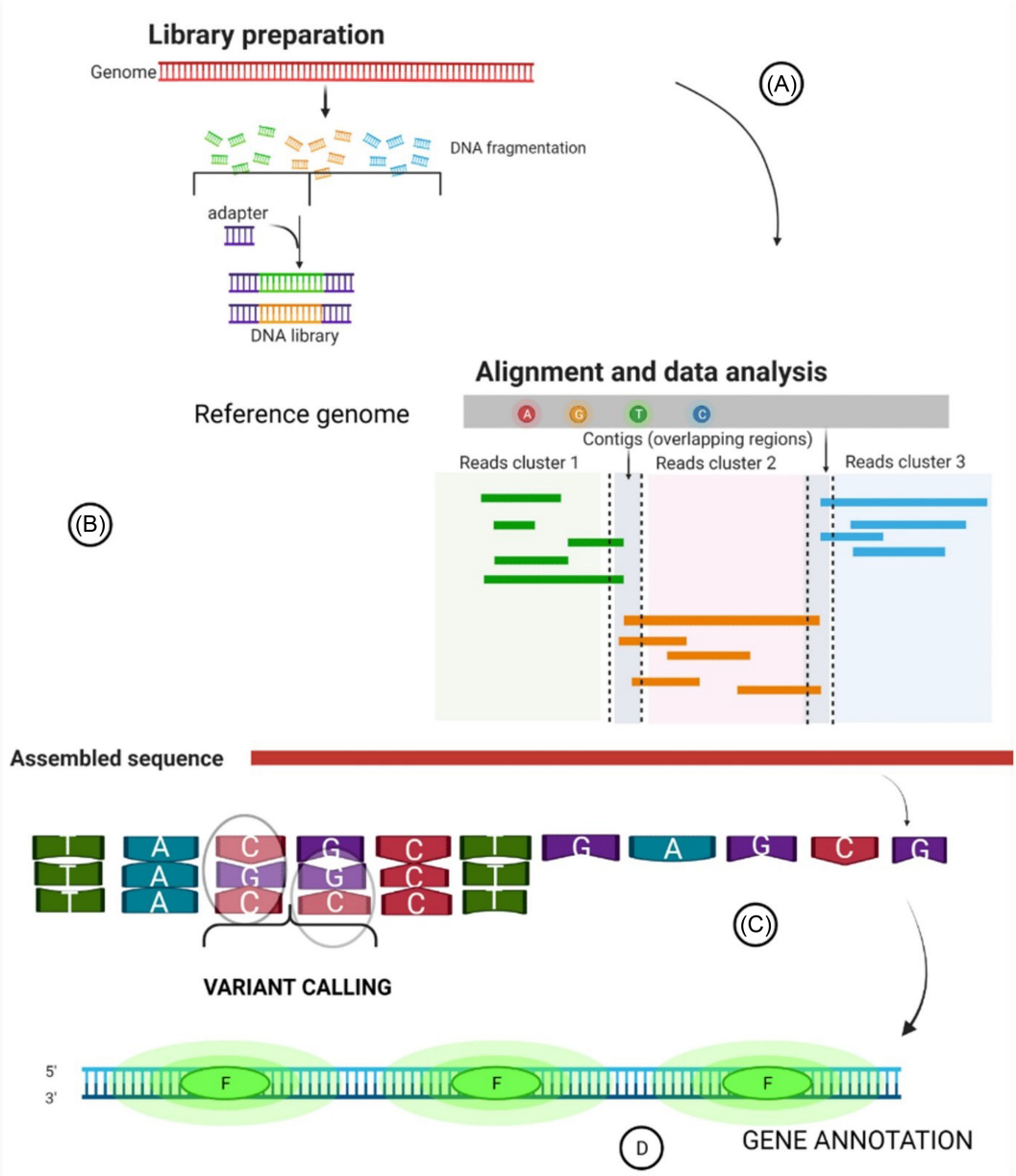

Figure 1. Diagram of a WGS process. (A) Illustrates the library preparation process. (B) Discusses the mapping of nucleotide sequences to genomes of reference. (C) Displays the variant calling and detection (SNP). (D) Gene annotation: identification of functional elements across the genome. The letters represent functional proteins.

Beyond improving the accuracy of variant detection, ML approaches facilitate the analysis of complex variant types, including structural variants and CNVs. These types of genetic alterations often present challenges due to their complexity and size, making traditional methods less effective. ML models, particularly those utilizing deep learning architectures, can analyze patterns and relationships within large genomic datasets to detect and characterize these complex variants more efficiently [5]. Some genomic tools/algorithms based on ML architecture for variant calling and annotations are listed below in (Table 1)

2.2.2. GWAS

ML applications have greatly increased the efficacy of GWAS, which have become a critical method for discovering genetic variations linked to complex traits and disorders. To find correlations between genetic variations and phenotypic variables across sizable populations, GWAS generally entails scanning the complete genome. Regression models, clustering algorithms, and ensemble approaches are a few examples of machine learning techniques that have been used more and more to increase the accuracy and power of these investigations. By enhancing the detection of connections between genetic variations and characteristics through sophisticated data analysis techniques, ML improves GWAS [19]. Algorithms that can handle high-dimensional data and reveal complex patterns that typical statistical methods can miss include random forests and gradient-boosting machines. These machine learning algorithms are capable of processing enormous volumes of phenotypic and genomic data, and they can decipher intricate connections between trait outcomes and genetic variants. This skill enables the

Table 1. Genomic tools for variant calling and annotations using machine learning

Tool/Algorithm	Machine learning method	Primary function	Key features
DeepVariant	Convolutional Neural Networks (CNN)	Variant Calling	High accuracy, trained on diverse datasets
GATK HaplotypeCaller	Random Forest	Variant Calling	Detects SNPs and indels, widely used in genomics pipelines
ANNOVAR	Decision Trees	Variant Annotation	Supports multiple annotation databases, user-friendly interface
SnpEff	Rule-based Machine Learning	Variant Annotation	Fast processing, customizable annotations
DeepSNV	Deep Learning	Detection of Rare Variants	Highly sensitive, suitable for low-frequency variants
VEP (Variant Effect Predictor)	Ensemble Learning	Variant Annotation	Extensive database support, functional impact prediction

Table 2. Genomic tools for GWAS using machine learning

Tool/Algorithm	Machine learning method	Primary function	Key features
PRSice-2	Lasso Regression	Polygenic Risk Score Calculation	Fast computation, automated clumping
GEMMA	Bayesian Regression	GWAS, Mixed Model Analysis	Handles relatedness, population structure
SnpEff	Decision Trees	Variant Annotation	Fast, supports multiple genomes
PLINK	Support Vector Machines	GWAS, Data Management	Efficient data handling, filtering options
DeepVariant	Deep Learning (CNN)	Variant Calling	High accuracy, trained on diverse datasets
PrediXcan	Elastic Net Regression	Transcriptome Prediction	Integrates GWAS and expression data

identification of novel loci that may contribute to phenotypic variation or disease susceptibility, as well as the detection of modest genetic connections [20].

Furthermore, the integration of GWAS results with other genomic data types, like gene expression profiles and epigenomic markers, is made easier by ML approaches (Figure 1) [12]. Through the utilization of integrative models, scientists may evaluate the impact of detected genetic variations on gene expression and regulatory networks, thereby offering a more profound understanding of the biological processes that underlie observed correlations. Through the connection of genetic variations to functional outcomes and possible treatment targets, this integrated method improves the interpretation of GWAS results. Moreover, the issue of multiple testing in GWAS is addressed by ML models, as the large number of statistical tests conducted may result in exaggerated false-positive rates [21]. Table 2 provides a concise overview of some tools and algorithms used in GWAS that incorporate ML, highlighting their ML methods, primary functions, and key features.

3. WES

ML techniques are used to process WES data by improving various phases of data analysis. In the preprocessing phase, methods like SVMs and random forests help filter out sequencing noise and enhance read alignment to the reference genome [16]. For variant calling, which focuses on coding regions, deep learning models such as CNNs are employed to detect SNVs and small insertions or deletions with high accuracy by examining the aligned reads. Additionally, ML can be employed in feature extraction and prioritization to find clinically relevant variants, decreasing the computational burden by focusing solely on coding exons. Ensemble approaches, combining the outputs of several models, can further refine variant categorization, enhancing the detection of rare variations in WES datasets. These methods boost both the efficiency and precision of WES data analysis [22].

3.1. Definition and significance

WES is a targeted sequencing method that concentrates on the exomes, or coding sections of the genome, which make up between 1% and 2% of the human genome. These regions comprise the great majority of known genetic variants that impact protein function, while making only a small portion of the genome (Figure 2) [16]. This makes WES an effective method for discovering mutations linked to a wide range of disorders. Focusing on exomes, WES enables scientists and medical professionals to thoroughly examine gene sequences that code for proteins which are thought to contain the majority of mutations that cause disease. The value of WES is in its capacity to offer an in-depth examination of genetic variants that may cause illness [16]. Researchers can identify pathogenic variations that may go undetected by traditional diagnostic techniques by sequencing the exomes of affected individuals. This allows for accurate genetic diagnosis and informs suitable treatment therapies [20].

Beyond its use in diagnosis, WES is essential to genetic research since it makes it easier to find new genes linked to disease. This holds particular significance for complex illnesses that could entail various hereditary and environmental components. By analyzing exomes across large datasets, WES uncovers common and uncommon variants that increase the risk of disease and sheds light on the genetic architecture of complex traits. Finding these variations may help us comprehend the mechanisms underlying the disease and may also point to new treatment options. Additionally, WES can be applied to population genomics to investigate genetic diversity and the

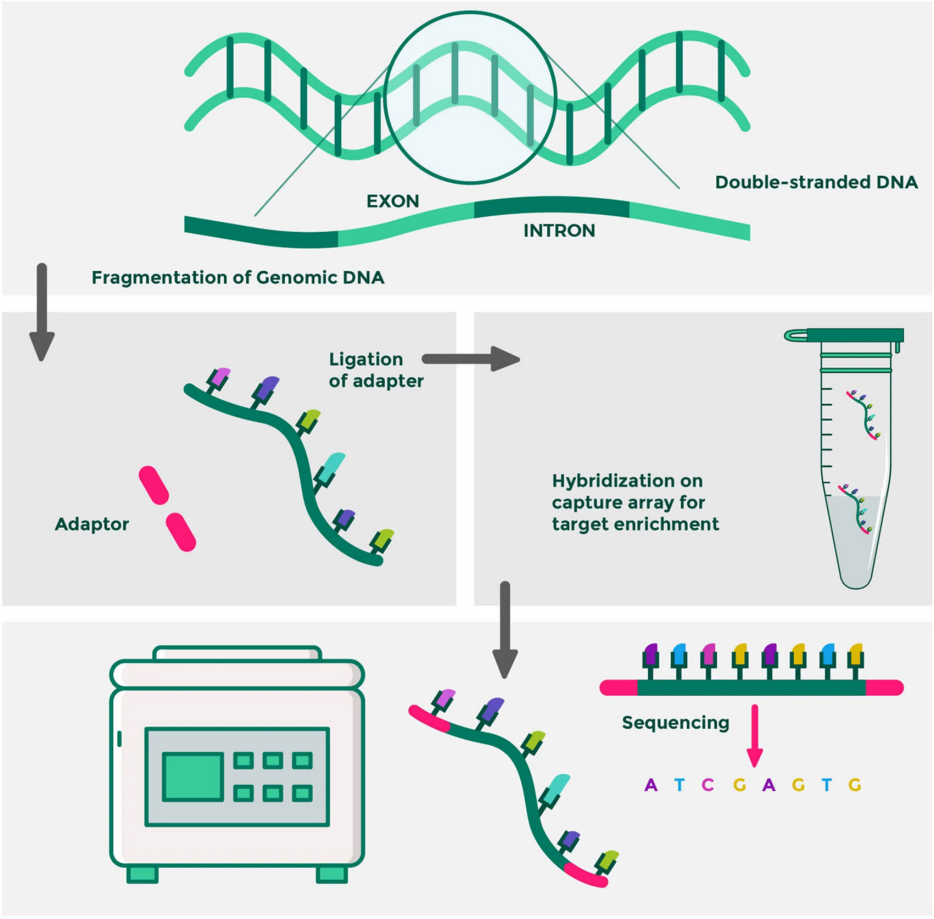

Figure 2. Whole exome sequencing: The steps involved
Note: Firstly, genomic DNA is isolated from the sample. Subsequently, the DNA is fragmented, and these fragments are hybridized to capture probes that specifically target the exonic regions. The captured exons undergo enrichment and amplification via PCR techniques. The enriched library undergoes sequencing through high-throughput sequencing platforms. Ultimately, the generated sequence data are aligned with a reference genome, and computational techniques are utilized to identify alterations within the exonic regions, facilitating the identification of mutations linked to disease.

evolutionary background of human communities, providing information on the distribution of genetic variations among various populations [16]. The affordability of WES in comparison to WGS is another important benefit. Even though WGS scans both coding and non-coding sections of the genome, processing and analyzing the massive quantity of data it produces can be expensive. Because WES concentrates on the exomes, less data are produced, which makes it a more feasible choice for large-scale research and standard clinical diagnoses. Since the majority of clinically significant mutations are discovered in the coding regions that WES targets, this cost-effectiveness does not come at the risk of losing important information. Therefore, WES achieves a compromise between accessibility and detailed genetic analysis, leading to its widespread use as a tool in both clinical and scientific settings [22].

3.2. ML applications in WES data analysis

The processing of (WES) data has been transformed by ML algorithms, which have improved our capacity to understand and apply the enormous amounts of genetic data gathered. The majority of known disease-causing mutations occur in the protein-coding sections of the genome, which is the focus of WES. Nonetheless, there are several difficulties in identifying, interpreting, and translating these genetic variations into therapeutic practice. ML provides strong tools to overcome these issues and progress the science of genomics because of its ability to process and evaluate complicated, high-dimensional data. Variant calling is one of the main uses of ML in WES data processing [5]. Finding genetic variations between an individual's exome and a reference genome, such as SNPs and insertions or deletions (indels), is known as variant calling. A few ML techniques that have been applied to increase the precision and effectiveness of this process are SVMs, random forests, and neural networks. These algorithms can discriminate between actual genetic variants and sequencing errors since they were trained on labeled datasets with known variants and artifacts. ML models reduce false positives and false negatives by improving the sensitivity and specificity of variation detection through learning from these examples [23].

Apart from variants calling, ML is essential for the functional annotation of variants found using WES. Anticipating the possible effects of genetic variations on protein function and illness risk is known as functional annotation. This attempt is difficult because of the great variety of possible modifications and the subtle nature of their effects. ML models, particularly those that employ deep learning techniques, can include a range of datasets, such as

sequence conservation scores, gene expression data, and protein structural features, to predict the functional implications of alterations. By spotting minute patterns and connections that conventional bioinformatics techniques can miss, these models can provide more precise variant pathogenicity predictions [23]. Prioritizing variations for additional research is another function that ML makes easier. Prioritizing the variations most likely to be clinically significant is crucial, as most WES analyses identify a large number of variants. Based on anticipated pathogenicity, allele frequency, and correlation with established illness characteristics, machine learning models can rank variations [24].

3.2.1. Identification of disease-causing mutations

To diagnose and comprehend genetic illnesses, one of the most important aspects of medical genetics and genomics is the detection of mutations that cause disease. This technique entails identifying genetic variants that result in aberrant regulation or function of proteins, which can induce or predispose people to diseases. In this investigation, WES is a valuable tool since it concentrates on sequencing the genome's protein-coding regions, where the majority of known disease-causing mutations are found [16]. Bioinformatics and computational techniques are used to evaluate the possible influence of genetic variations on gene function after they have been identified by sequencing. Evaluating the variations' functional effects, this involves assessing them in databases of known pathogenic and benign mutations. Predictive algorithms are also used in this process. Variables like the biochemical properties of modified amino acids, the conservation of evolutionary traits, and the anticipated impact on the structure and function of proteins are often considered by these algorithms. ML algorithms have enhanced this procedure by integrating multiple data types and increasing the accuracy of pathogenicity prediction [11].

Finding the mutations that cause a disease is very important in clinical settings because it helps with diagnostic and treatment choices. Finding the causing mutation for uncommon genetic illnesses can lead to a conclusive diagnosis, saving patients and their families from a diagnostic maze. Additionally, it makes it possible to predict the course of a disease and possible therapeutic responses, which helps to inform individualized treatment plans. Furthermore, genetic therapy can be made easier by detecting mutations in patients, enabling at-risk family members to receive testing and information about their genetic status. Through the identification of novel genes and pathways involved in pathogenesis, the discovery of disease-causing mutations in research broadens our understanding of the genetic basis of disorders [25]. New therapeutic targets and strategies may result from this understanding. Moreover, the detection of these mutations aids in clarifying the mechanisms underlying complex disorders, which frequently involve the interplay of several genetic and environmental variables. Therefore, finding the mutations that cause a disease is crucial for improving public health outcomes, promoting scientific research, and providing care for specific patients [26].

3.2.2. Predictive models for phenotypic traits

In genetics and genomics, predictive models for phenotypic traits are a crucial tool that allow the assessment of individual features based on genetic and environmental factors. These models use statistical and computational methods to examine huge datasets that include phenotypic data and genetic variants, which are frequently obtained from whole genome or WES. Finding genetic markers linked to particular qualities and predicting how these traits will manifest in individuals based on their genetic profile are the main objectives. There are various crucial phases involved in building predictive models for phenotypic features. Research has shown that the identification of mutations that cause disease broadens our knowledge of the genetic basis of diseases by identifying new genes and pathways involved in pathogenesis. The accuracy and reliability of these predictions are influenced by factors such as the sample size, the genetic architecture of the trait, and the presence of gene-environment interactions [27].

Predictive models for phenotypic features have been developed and applied with great improvement attributed to ML techniques. The intricacy and large dimensionality of genetic data are handled by techniques like deep learning algorithms, SVMs, and regularized regression. To produce more thorough and precise predictions, these models might include a wide range of variables, such as genetic, epigenetic, and environmental influences. These models are resilient and generalizable to various populations thanks to the application of cross-validation and other validation techniques. There are several uses for predictive models of phenotypic features in agriculture, evolutionary biology, and customized medicine [28]. These models have the potential to predict an individual's susceptibility to specific diseases, which opens the door to early intervention and tailored therapy in personalized medicine. They support breeding efforts in agriculture by forecasting desirable features in animals and crops. Predictive models aid in the understanding of how genetic diversity influences phenotypic variation and adaptation in the field of evolutionary biology. Ultimately, with major ramifications for both science and society, the creation and improvement of predictive models for phenotypic qualities represent a crucial nexus of computational biology, bioinformatics, and genetics [29].

4. Single-Cell Genomics

ML techniques are used to process WES data by improving various phases of data analysis. In the preprocessing phase, methods like SVMs and random forests help filter out sequencing noise and enhance read alignment to the reference genome. For variant calling, which focuses on coding regions, deep learning models such as CNNs are employed to detect single SNVs and small insertions or deletions with high accuracy by examining the aligned reads. Additionally, ML can be employed in feature extraction and prioritization to find clinically relevant variants, decreasing the computational burden by focusing solely on coding exons. Ensemble approaches, combining the outputs of several models, can further refine variant categorization, enhancing the detection of rare variations in WES datasets. These methods boost both the efficiency and precision of WES data analysis [8].

4.1. Definition and significance

Within molecular biology, single-cell genomics is a cutting-edge area that focuses on the analysis of the genome, transcriptome, or epigenome at the level of individual cells (Figure 3) [8]. Researchers can capture the molecular heterogeneity and dynamic processes taking place within individual cells using single-cell genomics, in contrast to typical bulk sequencing methods that evaluate averaged signals from a population of cells. Understanding complex biological systems requires this ability since it offers insights into cellular diversity, development, and disease states that cannot be obtained via bulk analysis [8]. Single-cell genomics is important because it may reveal the distinct transcriptome and genetic profiles of individual cells. Even while all cells in multicellular organisms have the same genetic makeup, they develop into diverse types with unique roles [8].

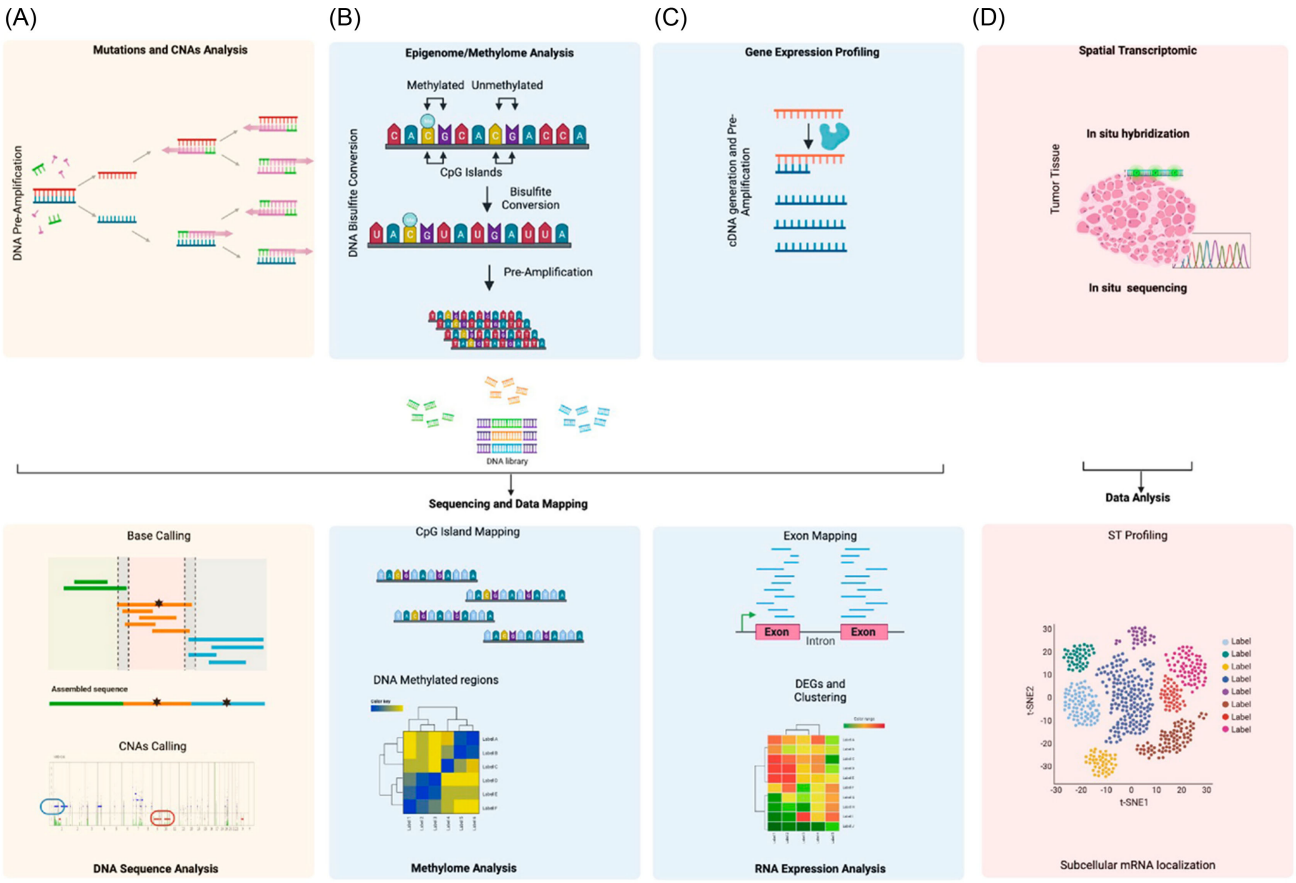

Figure 3. The principle of single-cell sequencing and analysis

Note: (A–C) Isolated single cells undergo breakdown and are subsequently analyzed for mutations and CNAs: the amplified DNA is employed for sequencing and library preparation. The sequences are aligned with a specific DNA reference to detect mutations and copy number alterations. Analysis of the epigenome and methylome involves treating DNA with bisulfite, which converts unmethylated cytosines into uracil. Thymine undergoes conversion from uracil residues during a pre-amplification process. After that, libraries are organized and prepared. Methylation levels are determined by mapping the obtained reads with a particular DNA reference; (C) Gene expression profiling includes cDNA reverse transcription of RNA, pre-amplified, and sequencing and library preparation. After the mapping of the acquired reads in accordance with exome references, DEG clustering will be carried out; (D) Figure illustrating the principle of spatial transcriptomics. Both sequencing and in situ hybridization can be used to determine mRNA expression. The acquired data make it possible to map a single cell's expression profile to a particular coordinate.

Single-cell genomics has significant implications for comprehending the cellular basis of illness in the context of disease. Significant cellular heterogeneity is present in a wide range of diseases, including cancer, neurological illnesses, and immune-related problems. Tumors, for instance, are made up of many subpopulations of cancer cells that have unique gene expression profiles and genetic abnormalities that can affect how well a patient responds to treatment and how quickly the illness progresses. These subpopulations can be recognized and described through the use of single-cell genomics, which facilitates the creation of tailored treatment plans and targeted medications. This technique aids in the understanding of immune evasion and treatment resistance mechanisms by offering a comprehensive perspective of the tumor microenvironment, encompassing interactions between immune cells and cancer cells [30]. Single-cell genomics is based on technology that includes isolating individual cells, amplifying their genetic information, and sequencing the DNA or RNA that is produced. Techniques such as scRNA-seq and single-cell DNA sequencing are widely used to study gene expression and genetic changes, respectively. The ability to characterize epigenetic alterations at the single-cell level has also been made possible by recent developments, offering a thorough understanding of gene regulation [31].

4.2. ML applications in single-cell genomics

In single-cell genomics, the study of an individual cell's genome, transcriptome, or epigenome, ML has become an essential tool for data analysis and interpretation. Large and complex datasets generated by this field represent the variety of cellular populations, rendering typical analytical techniques inadequate for deriving biologically significant conclusions. ML is essential for tackling these problems and expanding our knowledge of cellular biology because of its ability to handle high-dimensional data, recognize patterns, and make predictions. The detection and categorization of cell types is one of the main uses of ML in single-cell genomics. RNA-seq data obtained from scRNA-seq may reveal the gene expression patterns of thousands of distinct cells [7]. However, advanced computational techniques are needed to identify different cell types and subtypes within this

data. ML algorithms, such as clustering approaches (k-means, hierarchical clustering) and dimensionality reduction techniques (PCA, t-SNE), are used to group cells with comparable expression patterns. These techniques aid in the differentiation of well-known cell types and the discovery of unusual or uncommon cell populations, all of which are important for comprehending tissue formation and heterogeneity [32].

Furthermore, combining single-cell data from several sources is a common use of ML models. Multimodal data, including chromatin accessibility, protein levels, and RNA expression, are frequently used in single-cell genomics. Biological insights are more accurate and have higher resolution when these many data various types are integrated. To gain a more thorough understanding of cellular states and functions, deep learning techniques such as autoencoders and neural networks, for example, can combine multi-omics data to identify correlations between various cellular properties. Finding the regulatory networks and pathways that govern gene expression and cell fate decisions is made easier with the help of this integration [33]. ML applications in single-cell genomics have significant implications for disease context, including biomarker identification and disease mechanism understanding. ML can recognize unique biological patterns linked to particular diseases, including cancer, autoimmune diseases, or neurodegenerative disorders, by examining single-cell data from diseased tissues [34].

4.2.1. Cell type identification

A crucial component of single-cell genomics is cell type identification, which makes it possible to characterize the diversity of cells within tissues in great detail. Through the examination of gene expression profiles at the individual cell level, scientists differentiate between different cell types and comprehend their distinct molecular signatures. Understanding developmental phases, disease processes, and cellular heterogeneity all depend on this process. In real terms, scRNA-seq data with high dimensions must be analyzed to identify the cell type. Computational techniques are used to group cells according to similarities in their gene expression profiles. Using techniques like k-means clustering, hierarchical clustering, and more sophisticated algorithms like graph-based clustering, cells are classified into groupings that suggest potential cell types. PCA and t-SNE are two commonly used methods for reducing the dimensionality of data and facilitating visualization [35].

After clusters are found, they are annotated by cross-referencing them with reference cell type atlases or recognized marker gene sets. Annotation can be done manually or automatically using algorithms that estimate cell types based on expression profiles using pre-existing datasets. Precise identification of cell types is crucial for comprehending typical biology as well as for locating uncommon or unique cell populations that might be important in the context of disease. Cell type identification continues to offer important insights into the intricacy of cellular systems and their implications for health and illness as single-cell technologies progress [36].

4.2.2. Trajectory inference

A crucial computational method in single-cell genomics, trajectory inference seeks to clarify the dynamic mechanisms driving cellular differentiation and development. Reconstructing the series of cellular states or transitions that give rise to discrete cell types or stages is the aim of this method. Trajectory inference offers insights into how cells go through distinct stages of differentiation, respond to stimuli, or proceed through diseases by examining high-dimensional gene expression data from individual cells. Finding cells with similar gene expression profiles is the first step in the trajectory inference process. These cells are then arranged into a continuous trajectory or developmental route [37]. This reconstruction is achieved using various computational methods that model the progression of cellular states as a trajectory through a high-dimensional gene expression space. Techniques such as PCA and t-SNE are used to reduce the complexity of the data and visualize the trajectories in a lower-dimensional space. Algorithms such as Monocle, Slingshot, and Pseudotime are specifically designed to infer the ordering of cells along these trajectories, allowing researchers to trace the path from precursor states to differentiated cell types [30].

In the research of diseases and developmental biology, trajectory inference is particularly valuable. It aids in the understanding of cell differentiation in developmental biology by highlighting key stages and transitions that take place during development. Understanding the regulatory mechanisms that dictate cell fate decisions and the function of particular genes or signaling pathways in these processes requires knowledge of this information. Trajectory inference can be used to understand how disease states change over time, including how immune cells react to infections or therapies or how cancer cells develop malignant characteristics. Trajectory inference also makes it possible to identify important regulatory genes or factors that mediate changes in cellular states [38].

5. Spatial Transcriptomics

5.1. Definition and significance

Through the use of cutting-edge technology called “spatial transcriptomics”, researchers can examine the spatial arrangement of transcriptomes within intact tissues by fusing gene expression profiling with spatial information (Figure 3(D)) [8]. Through the preservation of the spatial context of gene expression, a critical component in comprehending the structure and operation of complex tissues, this method overcomes the drawbacks of conventional bulk RNA-seq and scRNA-seq. Spatial transcriptomics facilitates a better understanding of the spatial variability and organization of cellular processes by offering a map of the locations of particular transcripts within tissue sections. Spatial transcriptomics is crucial because it can provide light on the in vivo interactions between cells and their surroundings [39]. Spatial information is lost when tissues are homogenized or divided into individual cells in traditional gene expression research techniques. This is especially problematic in tissues, like the brain, where different regions have different cellular compositions and functions, or in tumors, where the interaction between cancer cells and the surrounding stroma affects the course of the disease and the patient’s response to treatment. Spatial transcriptomics preserves this spatial context, enabling researchers to link gene expression patterns directly to their anatomical locations [40].

Spatially barcoded slides or arrays are used in the technology that enables spatial transcriptomics to extract mRNA from tissue sections. Every area on the array is marked with a distinct barcode and relates to a particular place in the tissue. The spatial barcode enables researchers to track the gene expression data back to its original position in the tissue once mRNA from the tissue is extracted and read. A high-resolution map of the transcriptome activity across the tissue segment is the outcome of this. The simultaneous acquisition of morphological data, which can be combined with transcriptome data for thorough analysis, has been made possible by technological advancements in imaging, further

improving spatial transcriptomics [41]. One of the primary areas in which spatial transcriptomics is applied is developmental biology, where researchers can learn more about the mechanisms of tissue patterning and organogenesis by examining the spatial distribution of gene expression in developing tissues [42].

5.2. ML applications in spatial transcriptomics

ML is essential for analyzing and interpreting spatial transcriptomics data. This allows researchers to extract valuable information from this technique's complex, high-dimensional datasets. By combining spatial information from tissue sections with gene expression data, spatial transcriptomics offers a full picture of the molecular environment inside a biological sample. To find patterns, recognize different cell types, deduce spatial organization, and comprehend cellular interactions, the enormous volume of data generated demands sophisticated computational techniques. These activities are perfectly suited for ML algorithms because of their prowess in handling enormous datasets and identifying complex patterns [43]. The recognition and categorization of spatial domains inside tissues is one of the key uses of ML in spatial transcriptomics. Regions of tissue with unique gene expression profiles are referred to as spatial domains; these regions are frequently associated with particular cell types or functional domains. Spots or regions are often grouped based on the similarities in their gene expression using clustering algorithms like k-means and hierarchical clustering, as well as unsupervised learning approaches like PCA and t-SNE. These clusters can help map tissue architecture and identify regions with distinct molecular signatures by revealing the spatial organization of various cell populations [41].

Additionally, the integration of spatial transcriptomics data with other modalities, including histology images or other omics datasets, is made easier by ML. One use of deep learning models is the alignment and integration of spatial transcriptomics data with corresponding histology pictures, specifically using CNNs. The incorporation of this feature improves the spatial resolution and offers a more comprehensive framework for analyzing gene expression patterns concerning tissue shape [44]. Also, transcriptome data can be combined with proteome or metabolomic data through multi-omics integration utilizing ML, providing a more thorough understanding of cellular relationships and functions [45].

Inferring cell-cell connections and communication networks within the geographical context is another important use of ML. ML algorithms can uncover signaling pathways that are active in particular places or anticipate possible connections between distinct cell types by examining gene co-expression patterns and spatial closeness. Classifiers such as random forests or SVMs, for instance, can be trained to identify connections based on the ligand and receptor expression levels in nearby cells. Understanding how cells affect one another's behavior and function—especially when it comes to the tumor microenvironment or immune response—is made possible by the knowledge provided by this material [43].

6. Proteomics, Transcriptomics, Metagenomics Epigenomics, Microbiome Research in ML

Proteomics, transcriptomics, metagenomics, epigenomics, and microbiome research represent critical omics fields where ML has become indispensable for handling the vast and complex datasets generated (Figure 4) [45]. In proteomics, ML algorithms are employed to predict protein-protein interactions (PPI) and functional annotations, enabling deeper insights into cellular functions and disease mechanisms. Transcriptomics leverages ML to analyze differential gene expression and construct gene regulatory networks (GRNs), facilitating the understanding of gene activity and its regulation under various conditions. In metagenomics, ML aids in profiling microbial communities and predicting their functional potential, which is vital for understanding microbial dynamics in diverse environments. Epigenomics benefits from ML in the analysis of methylation patterns and histone modifications, offering crucial information about gene regulation and its role in diseases such as cancer [45].

6.1. Overview of each omics field

The large-scale study of proteins, which are essential macromolecules with a range of structural and functional roles in cells, is known as proteomics. This study focuses on understanding the structure, function, and connections of the

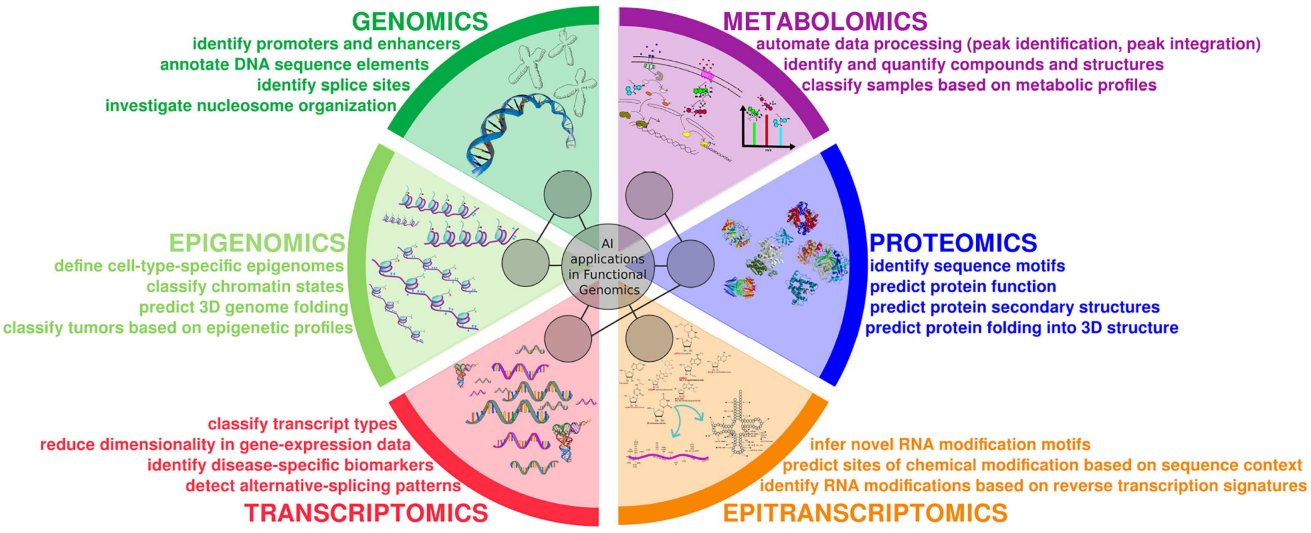

Figure 4. ML application in functional genomics

entire proteome—the collection of proteins expressed by a genome, cell, tissue, or organism [46]. Proteomics includes identifying and quantifying proteins, characterizing their post-translational modifications, and investigating their interactions and localization within the context of cells. Commonly employed techniques for protein analysis include mass spectrometry and protein microarrays. Understanding cellular mechanisms, signaling pathways, the molecular underpinnings of illnesses, and the identification of potential biomarkers and therapeutic targets are all made possible with the help of proteomics [47]. Study of transcriptomics focuses on the transcriptome, which is the entire collection of RNA transcripts generated by the genome in a particular cell or under particular conditions. All forms of RNA, such as transfer RNA, messenger RNA, ribosomal RNA (rRNA), and non-coding RNA, are included in this field. A snapshot of the active genes and their expression levels at a specific time is provided by transcriptomic studies. High-throughput technologies are used to measure RNA levels and investigate the dynamics of gene expression, such as RNA-seq and microarrays. Understanding gene control, cellular differentiation, and responses to environmental stimuli is made possible by transcriptomics. Additionally, it is essential in determining the patterns of gene expression linked to specific disorders, which facilitates the identification of therapeutic targets and diagnostic markers [47].

By thoroughly examining genetic material extracted from environmental samples, a process known as metagenomics makes it possible to investigate microbial populations without the requirement for culture. In this field, the diversity, dynamics, and functions of microbial populations are studied by looking at the collective genomes of microorganisms in a variety of settings, including soil, water, and the human body. By identifying and describing microbial species, their genes, and metabolic pathways using sequencing technology, metagenomic techniques shed light on the functions that these organisms play in ecosystems and host health. Metagenomics has proven to be a valuable tool in the research of the human microbiome, or the community of bacteria that live in the human body, and how it affects immunity, physiology, and illness [48]. The study of epigenetic modifications, or heritable variations in gene expression without changes to the DNA sequence, is included in the field of epigenome research. These alterations control chromatin structure and gene expression together. They include DNA methylation, histone changes, and non-coding RNA-associated gene silencing. The goal of epigenomics is to map these changes throughout the genome to comprehend how they affect gene activity and support cellular identity and function. Epigenetic modifications are important for proper cellular differentiation and development. They also have a major impact on several diseases, such as autoimmune diseases, cancer, and neurological disorders [47].

6.1.1. Proteomics: PPI prediction, functional annotation

Understanding cellular functions and molecular pathways depends on crucial components of proteomics, the large-scale study of proteins, such as functional annotation and PPI prediction. PPIs are necessary for almost all biological activities because proteins often work in complexes rather than alone. Anticipating these interactions contributes to our understanding of how proteins work together to perform tasks including metabolic pathways, signal transduction, and structural organization. Empirical data on these interactions can be obtained using a variety of experimental methods, albeit their applicability and context-specificity may be limited. Examples of these methods include yeast two-hybrid screening, co-immunoprecipitation, and affinity purification followed by mass spectrometry [49]. Predicting PPIs involves substantial use of computer tools to supplement experimental data. By applying techniques like sequence homology, motif discovery, and structural modeling, these methods make use of protein sequence and structural data. The accuracy of PPI predictions has greatly improved due to ML and deep learning algorithms, which analyze vast datasets of known interactions and find patterns that point to binding interfaces and interaction motifs. These prediction models provide a more thorough knowledge of protein interactions under various biological settings by integrating a variety of biological data, such as transcriptomic, functional, and genomic data [50].

Another vital aspect of proteomics is functional annotation, which attempts to assign biological roles to proteins according to their interactions, structures, and sequences. Proteins are categorized using characteristics like conserved domains, motifs, and established interactions into functional groupings, such as enzymes, receptors, or structural components. Bioinformatics techniques are frequently used in conjunction with experimental methods, like mutagenesis and biochemical experiments, to anticipate the functionalities of proteins that lack experimental characterization. These methods offer direct insights into the functioning of proteins. Protein functions can be systematically annotated with the help of databases such as Gene Ontology and InterPro, which provide structured vocabularies and classification schemes. Accurate functional annotation requires a combination of computational predictions and experimental validation [51]. Homology-based techniques are extensively used to predict the function of proteins by comparing their sequences to those of well-characterized proteins. By assuming that identical sequences frequently suggest similar functions, these techniques enable researchers to deduce the purpose of proteins that have not yet been fully identified. Furthermore, by employing knowledge about interacting partners to forecast activities, network-based methods take into account the larger context of protein interactions. This is predicated on the idea that proteins that interact are probably engaged in similar biological routes or processes [52].

Proteomics is essential for determining the molecular causes of disorders as well as for comprehending normal cellular functions. Diseases including cancer, neurological conditions, and cardiovascular conditions might result from abnormalities in PPIs or protein functions. One way to find oncogenic drivers and possible treatment targets in cancer is to identify disordered protein networks. Analyzing protein interactions in neurodegenerative illnesses can also reveal pathways that contribute to the course of the disease and provide information about possible treatments. Proteomics technology and computational methods hold great potential to improve our comprehension of the intricate networks and roles of proteins as they develop further [53]. Proteomics' integration with other omics technologies, including transcriptomics, metabolomics, and genomes, expands our understanding by offering a systems biology viewpoint that is essential for thorough biological and clinical insights. This comprehensive method is becoming more and more important in precision medicine, since knowledge of the proteome may guide individualized treatment plans that are catered to each patient's unique molecular profile [53].

6.1.2. Transcriptomics: Differential expression analysis, GRNs

Transcriptomics, a comprehensive study of RNA transcripts generated by the genome in particular conditions, offers vital information about the regulation and expression of genes. A key method in transcriptomics is differential expression analysis, which contrasts RNA levels at various biological states or

intervals. The analysis reveals genes whose expression levels fluctuate dramatically, providing insight into the molecular mechanisms behind physiological functions or pathological conditions. Researchers can discover biological pathways, find possible illness biomarkers, and learn how organisms react to treatments or environmental changes by measuring these differences [53]. High-throughput sequencing technologies, such as RNA-seq, are commonly employed in the process of differential expression analysis. The quantitative assessment of RNA abundance provided by RNA-seq makes it possible to identify both known and unknown transcripts. Following sequencing, the sequences are aligned to a reference genome, and expression levels are measured using bioinformatics methods. The significance of expression variations is then evaluated using statistical approaches that take sample size and variability into account. With the use of this thorough method, scientists can accurately and specifically identify alterations in the transcriptome, giving them a comprehensive picture of gene activity under various circumstances [30].

Another key aspect of transcriptomics is the study of GRNs, which concentrate on the intricate relationships that control gene expression. Genes, transcription factors, and other regulatory components make up GRNs, which regulate the timing and degree of gene expression. These regulatory linkages, which can be direct —like transcription factor binding—or indirect—like intermediary signaling pathways—must be identified to construct GRNs. GRNs offer a framework for comprehending how particular cellular responses are regulated by genes and how these networks are impacted by illnesses [54]. Integrating several data sources, including transcriptomic, ChIP-seq, and epigenomic data, is frequently necessary for the creation of GRNs. From these data, computational approaches like network inference methods and ML are used to forecast regulatory connections. These models can identify important regulators that regulate big gene sets, like master transcription factors.

Understanding the molecular underpinnings of development, differentiation, and illness requires an understanding of GRNs. GRNs, for instance, can show how tumor suppressors and oncogenes interact with other genes to promote the growth of tumors in cancer research. Researchers can find possible treatment targets and biomarkers by mapping these networks. Additionally, because different cell types or states may exhibit diverse regulatory networks, GRNs offer insights into cellular heterogeneity. This information is especially crucial for fields such as developmental biology and regenerative medicine, where tissue engineering and therapy tactics can be informed by knowledge of the regulatory environment [55]. Transcriptomics provides an effective instrument for investigating the functional landscape of the genome through differential expression analysis and GRN assembly. The comprehension of basic biological processes is aided by these methods, which also have useful applications in biotechnology, medicine, and agriculture [56].

6.1.3. Metagenomics: Microbial community profiling, functional potential prediction

A thorough characterization of microbial communities is made possible by the analysis of DNA extracted directly from environmental samples; a process known as metagenomics. Researchers may investigate the variety and makeup of microorganisms in many environments, such as soil, the ocean, and the human body, thanks to this field. By sequencing the whole genomes of all the microorganisms in a sample, microbial community profiling eliminates the need to culture each individual species. Through the identification of various species' existence and relative abundance, this method offers a picture of the microbial landscape. While shotgun metagenomic sequencing provides a more comprehensive perspective by incorporating viruses and fungi, techniques like 16S rRNA gene sequencing are frequently utilized for bacterial and archaeal profiling [57]. Microbial community profiling frequently yields large, complicated data sets that require analysis using sophisticated bioinformatics methods. By allocating sequences to recognized species or operational taxonomic units following sequence similarity, these methods aid in taxonomic classification [57].

An additional key aspect of metagenomics is functional potential prediction, which focuses on the genes and metabolic pathways that are present in a microbial community. To predict the functional traits of the community, such as energy generation, nutrient cycling, or pathogenicity, the metagenomic data must be analyzed. The presence of particular genes or gene clusters that encode enzymes and other proteins involved in metabolic processes is used to infer the functional potential. By mapping these genes to well-known metabolic pathways, programs like Kyoto Encyclopedia of Genes and Genomes (KEGG) and MetaCyc shed light on the functional roles that bacteria play in their particular surroundings [58]. Understanding ecosystem activities and biogeochemical cycles is made possible by the capacity to anticipate functional potential from metagenomic data. Metagenomics, for instance, can show how microbial communities influence the global climate processes by contributing to the cycling of carbon and nitrogen in marine environments. Similar effects on plant health and soil fertility can be observed in soil ecosystems due to the functional potential of bacteria. Within the framework of human health, metagenomic analyses of the gut microbiota have revealed pathways related to vitamin production, immune system modulation, and food component digestion [59].

Furthermore, metagenomics allows researchers to identify genes associated with virulence factors or antibiotic resistance, offering crucial insights for public health and medical procedures. One of the biggest obstacles to treating infections is the proliferation of genes for antibiotic resistance in both pathogenic and non-pathogenic microorganisms. Antibiotic resistance can be monitored and controlled with the help of metagenomic analysis, which can trace the distribution and prevalence of these resistance genes in diverse settings [60]. All things considered, metagenomics provides an effective method for examining the variety, dynamics, and functional potential of microbial communities. Researchers can obtain a comprehensive understanding of microbial ecosystems and learn about the roles that microorganisms play in environmental processes, illness, and health by combining taxonomic characterization with functional potential prediction [61].

6.1.4. Epigenomics: Methylation pattern analysis and histone modification prediction

Heterologous modification prediction and methylation pattern analysis are fundamental to the field of epigenomics, which is the study of heritable variations in gene function without alterations to the DNA sequence. The importance of developing patterns for the prediction of epigenetic modifications is illustrated as a flowchart in (Figure 5) [62]. DNA methylation is a key epigenetic mechanism in which a methyl group is added to the cytosine residues in DNA, typically at CpG dinucleotides. This modification can decrease DNA accessibility to the transcriptional machinery, which in turn can suppress gene expression. It does this by altering the binding affinities of transcription factors or drawing chromatin-compacting proteins. Normal cellular differentiation and development depend on methylation patterns,

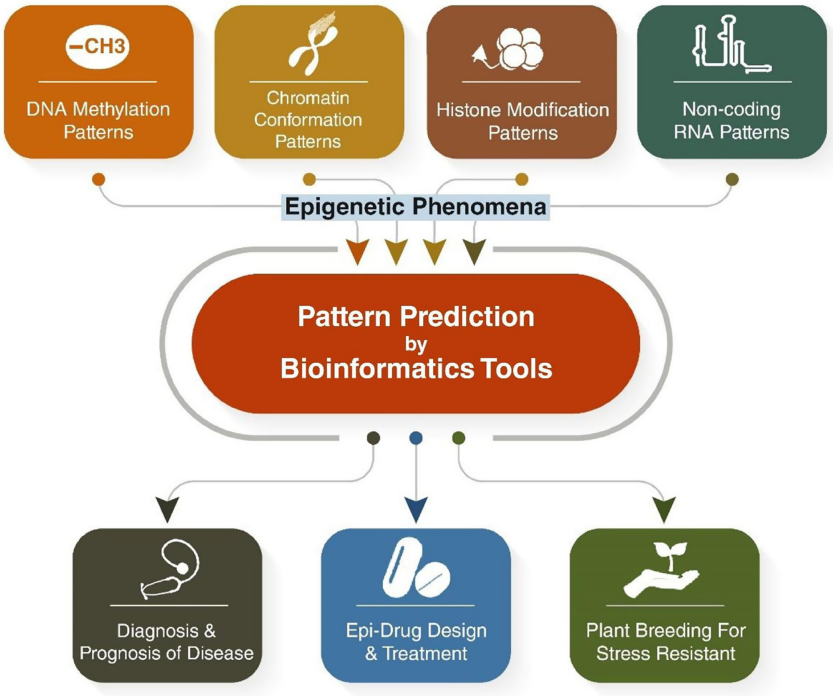

Figure 5. Importance and application of predicting epigenetic patterns
Note: DNA methylation, histone post-translational changes, chromatin conformation, and non-coding RNA regulation are all instances of epigenetic inheritance. Epi-drugs, crop improvement against stresses, and illness detection and therapy all use patterns that forecast these four modifications.

and abnormalities in these patterns are frequently linked to illnesses like cancer [63]. One essential component of epigenomic research is the analysis of DNA methylation patterns. High-resolution maps of the methylation sites throughout the genome are produced by methods like bisulfite sequencing. Comprehending these patterns is especially crucial for cancer research, as abnormal methylation may function as a target for therapy as well as a biomarker for diagnosis and prognosis [64].

Yet another significant component of the epigenome is histone modifications, which are chemical alterations to the histone proteins that encircle DNA. Depending on their kind and position, these modifications—which include acetylation, methylation, phosphorylation, and ubiquitination—can either activate or repress transcription. For instance, acetylation of histone tails is typically linked to an open chromatin conformation and active transcription, but depending on the particular lysine residue altered, certain methylation marks can either inhibit or activate transcription. Understanding chromatin dynamics and gene regulation requires the prediction and analysis of histone alterations [65]. Histone modification mapping throughout the entire genome is made possible by technologies like ChIP-seq. This technique includes precipitating DNA-protein complexes unique to changed histones using antibodies, which are subsequently sequenced to determine the chromosomal sites of these modifications. The collected information shows areas of active or repressed transcription and offers a thorough picture of the chromatin landscape. In order to forecast histone changes and understand their functional ramifications, computational techniques—such as ML algorithms—are utilized. This aids in the clarification of the intricate regulatory networks that control gene expression [66].

The field of epigenomic studies holds significant implications for comprehending disease, differentiation, and development. The dynamic alterations in methylation and histone modifications play a crucial role in embryonic biology by directing cell fate decisions and defining cellular identity. To ensure that differentiated cells preserve their identity over cell divisions, these epigenetic marks are also crucial in preserving cellular memory. The epigenome is frequently disturbed in disease states, including cancer, which results in alterations in gene expression that fuel malignancy. Tumorigenesis can be facilitated, for example, by mutations in the enzymes that write, read, or remove histone marks. These mutations can cause aberrant gene silence or activation [67]. A more comprehensive understanding of the epigenetic regulation of the genome is possible through the combination of data on methylation and histone modifications. To create epigenetic therapeutics and comprehend how the epigenome affects health and illness, a holistic perspective is essential [68].

6.1.5. Microbiome research

Research on the various microbial communities that live in a variety of settings, such as the human body, soil, oceans, and more, is greatly aided by studies on microbiomes. The process of classifying these microbial communities entails recognizing and cataloging the many microbial species that exist within a particular environment. The main methods for doing this are high-throughput sequencing technologies, like whole metagenome sequencing, which offers a more comprehensive perspective of all microbial life, including viruses, fungi, and protists, and 16S rRNA gene sequencing for bacteria and archaea. Through the examination of these microbes' genetic material, scientists can ascertain the species makeup, variety, and relative abundance within a given microbial community [69]. Understanding the dynamics and structure of microbial communities is just as important to their classification as identifying their species. This

involves researching the ways in which microbial populations fluctuate over time and in reaction to different stimuli, including sickness, nutrition, and environmental circumstances. For example, changes in the microbial composition of the human gut microbiome have been associated with a number of health issues, such as diabetes, obesity, inflammatory bowel disease, and mental health issues. Comprehending these changes can offer valuable perspectives on the origins of various illnesses and direct the creation of diagnostic and treatment approaches [70].

Beyond species identification, functional annotation of microbial communities makes predictions about the metabolic and functional capacities of the microbiome. This procedure entails assigning recognized functions, including enzymatic or metabolic pathways, to genes found in the microbial genome. It is common practice to assign functions to genes based on their sequences using programs like MetaCyc and KEGG. Functional annotation is important because it establishes a connection between the existence of particular microorganisms and their possible functions in the environment or host organism, such as the cycling of nutrients, the breakdown of contaminants, or the synthesis of bioactive substances [71]. Microbial communities' functional capacity can have a significant impact on human and environmental health. Microbes are crucial to the decomposition of organic matter, carbon sequestration, and nutrient cycling in environmental microbiomes, such as those found in soil and marine environments. The resilience and stability of ecosystems depend on these mechanisms. The digestion of dietary fibers, the production of vital vitamins, and immune system modulation are just a few of the functions of the gut microbiota about human health. A healthy microbial balance is crucial because dysbiosis, or an imbalance in the content and function of the microbiome, has been linked to several disorders [72].

Additionally, research on the microbiome aids in the discovery of biomarkers for the diagnosis and prognosis of disease. Certain functional features or microbiological profiles can act as markers of disease states or therapy responses. As an illustration, specific bacterial signatures have been linked to colorectal cancer and may operate as non-invasive biomarkers for early identification. Furthermore, knowledge of the functional activities of bacteria can guide the creation of innovative therapies that attempt to modify the microbiome beneficially, such as probiotics and prebiotics. These treatments have the power to improve or repair microbiological processes that support well-being and ward off illness [73].

7. Challenges and Future Directions

The integration and harmonization of diverse datasets present significant challenges in the rapidly developing fields of genomics and personalized medicine. Genomic data are frequently generated from different platforms and technologies, each with unique formats, resolutions, and quality metrics. Comprehensive analyses that can uncover new insights into biological processes and disease mechanisms depend on the integration of these heterogeneous data sources; however, differences in data types, such as sequencing depth, coverage, and annotation standards, can complicate this integration. Moreover, harmonizing datasets from different populations or study designs requires careful consideration to prevent biases and guarantee that the results are generalizable [74]. Another major obstacle to applying these technologies to genomic data is the interpretability of ML models. Even though deep learning models in particular have demonstrated considerable promise in locating intricate patterns across huge datasets, ML algorithms frequently serve as "black boxes", providing little information about the process by which particular predictions are generated. In therapeutic contexts, where knowing the reasoning behind a model's predictions is essential for fostering trust and helping patients make educated decisions, this lack of interpretability might be problematic. For instance, to predict the risk of a disease using genomic data, physicians and patients must both be aware of the specific genetic variants that are influencing the risk and how [75].

Given how sensitive and private genetic data is, ethical and privacy concerns should be taken very seriously when gathering, storing, and analyzing it. Strict privacy regulations are necessary since genetic data may be misused for discriminatory purposes, such as insurance or job discrimination. Moreover, questions of consent and people's power to regulate how their genetic information is used are brought up by the sharing of genomic data for study. It is essential to put in place data governance systems that provide informed consent, safeguard personal information, and permit restricted access to data. Also, developments in methods like safe multi-party computation and differential privacy are being investigated to allow genomic data analysis without jeopardizing personal privacy. These methods can assist in striking a compromise between the ethical requirement to preserve participant anonymity and the necessity for data access in research [76].

Comprehensive answers to these problems are becoming increasingly important as the science of genomics expands. International collaboration and standardization efforts will assist the integration and harmonization of heterogeneous datasets, guaranteeing that data from various sources may be easily merged and evaluated. Within the field of ML, creating more comprehensible models and techniques for elucidating intricate algorithms will augment the therapeutic usefulness of these instruments, permitting more transparent and practical insights [77].

8. Conclusion

The study of genomics has experienced a revolution with the introduction of ML, which has until recently unseen prospects for understanding the intricate biological mechanisms that underlie both health and illness. ML approaches have made it possible to analyze large amounts of genomic data efficiently, as this review has shown. This has resulted in substantial breakthroughs in fields like variant calling, disease mutation identification, and microbial community profiling. In addition to improving our knowledge of genetic variants and how they relate to diseases, these technologies have opened the door for more individualized medical care. The transformational promise of these technologies is best demonstrated by their capacity to uncover novel disease biomarkers, predict phenotypic features from genomic data, and clarify GRNs. Despite these developments, there are still several difficulties, namely data integration, model interpretability, and ethical issues. Integration of datasets and consistent, dependable studies are significantly hampered by the variety of genomic data sources and the absence of standards. The creation of strong bioinformatics tools and universal data standards that can harmonize various data types is necessary to address these issues. Moreover, integrating computational predictions into clinical practice requires ML models to be interpretable. To promote confidence and acceptance in clinical settings, it is critical to create techniques that can clarify how ML algorithms arrive at their predictions as these algorithms get more complex. Genomic research also raises ethical and privacy concerns, especially when it comes to handling and distributing genetically sensitive data. Preserving personal privacy and avoiding improper use of data is

critical. Establishing thorough ethical standards and data governance frameworks that guarantee informed consent, openness, and data protection is essential as the sector develops. These steps will assist in striking a compromise between the ethical requirement to protect people's genetic information and the necessity for data accessibility in research.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support this work are available upon reasonable request to the corresponding author.

Author Contribution Statement

Saheed Adegbola Adeyanju: Conceptualization, Methodology, Validation, Data curation, Writing – original draft, Writing – review & editing, Supervision. **Taiwo Temitope Ogunjobi:** Methodology, Investigation, Writing – original draft, Writing – review & editing.

References

[1] Wu, J., & Zhao, Y. (2019). Machine learning technology in the application of genome analysis: A systematic review. *Gene*, *705*, 149–156. https://doi.org/10.1016/j.gene.2019.04.062

[2] Reel, P. S., Reel, S., Pearson, E., Trucco, E., & Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology Advances*, *49*, 107739. https://doi.org/10.1016/j.biotechadv.2021.107739

[3] Wang, J., Zou, Q., & Lin, C. (2022). A comparison of deep learning-based pre-processing and clustering approaches for single-cell RNA sequencing data. *Briefings in Bioinformatics*, *23*(1), bbab345. https://doi.org/10.1093/bib/bbab345

[4] Sarumathi, S., Ranjetha, P., Saraswathy, C., & Gayathri, S. (2022). Deep neural network algorithms and their role: A systematic review. *International Research Journal of Computer Science*, *9*(2), 16–25. https://doi.org/10.26562/irjcs.2022.v0902.004

[5] Vadapalli, S., Abdelhalim, H., Zeeshan, S., & Ahmed, Z. (2022). Artificial intelligence and machine learning approaches using gene expression and variant data for personalized medicine. *Briefings in Bioinformatics*, *23*(5), bbac191. https://doi.org/10.1093/bib/bbac191

[6] Xu, H., Cong, F., & Hwang, T. H. (2021). Machine learning and artificial intelligence–driven spatial analysis of the tumor immune microenvironment in pathology slides. *European Urology Focus*, *7*(4), 706–709. https://doi.org/10.1016/j.euf.2021.07.006

[7] Petegrosso, R., Li, Z., & Kuang, R. (2020). Machine learning and statistical methods for clustering single-cell RNA-sequencing data. *Briefings in Bioinformatics*, *21*(4), 1209–1223. https://doi.org/10.1093/bib/bbz063

[8] Massimino, M., Martorana, F., Stella, S., Vitale, S. R., Tomarchio, C., Manzella, L., & Vigneri, P. (2023). Single-cell analysis in the *omics* era: Technologies and applications in cancer. *Genes*, *14*(7), 1330. https://doi.org/10.3390/genes14071330

[9] Li, B., & Ritchie, M. D. (2021). From GWAS to gene: Transcriptome-wide association studies and other methods to functionally understand GWAS discoveries. *Frontiers in Genetics*, *12*, 713230. https://doi.org/10.3389/fgene.2021.713230

[10] Yang, W., Zhang, T., Song, X., Dong, G., Xu, L., & Jiang, F. (2022). SNP-target genes interaction perturbing the cancer risk in the post-GWAS. *Cancers*, *14*(22), 5636. https://doi.org/10.3390/cancers14225636

[11] Joshi, M., Kapopoulou, A., & Laurent, S. (2021). Impact of genetic variation in gene regulatory sequences: A population genomics perspective. *Frontiers in Genetics*, *12*, 660899. https://doi.org/10.3389/fgene.2021.660899

[12] Akoniyon, O. P., Adewumi, T. S., Maharaj, L., Oyegoke, O. O., Roux, A., Adeleke, M. A., . . . , & Okpeku, M. (2022). Whole genome sequencing contributions and challenges in disease reduction focused on malaria. *Biology*, *11*(4), 587. https://doi.org/10.3390/biology11040587

[13] Mintzer, V., Moran-Gilad, J., & Simon-Tuval, T. (2019). Operational models and criteria for incorporating microbial whole genome sequencing in hospital microbiology – A systematic literature review. *Clinical Microbiology and Infection*, *25*(9), 1086–1095. https://doi.org/10.1016/j.cmi.2019.04.019

[14] Ahmed, Z., Renart, E. G., & Zeeshan, S. (2021). Genomics pipelines to investigate susceptibility in whole genome and exome sequenced data for variant discovery, annotation, prediction and genotyping. *PeerJ*, *9*, e11724. https://doi.org/10.7717/peerj.11724

[15] Katsaouni, N., Tashkandi, A., Wiese, L., & Schulz, M. H. (2021). Machine learning based disease prediction from genotype data. *Biological Chemistry*, *402*(8), 871–885. https://doi.org/10.1515/hsz-2021-0109

[16] Kuksa, P. P., Greenfest-Allen, E., Cifello, J., Ionita, M., Wang, H., Nicaretta, H., . . . , & Leung, Y. Y. (2022). Scalable approaches for functional analyses of whole-genome sequencing non-coding variants. *Human Molecular Genetics*, *31*(R1), R62–R72. https://doi.org/10.1093/hmg/ddac191

[17] Brendel, M., Su, C., Bai, Z., Zhang, H., Elemento, O., & Wang, F. (2022). Application of deep learning on single-cell RNA sequencing data analysis: A review. *Genomics, Proteomics and Bioinformatics*, *20*(5), 814–835. https://doi.org/10.1016/j.gpb.2022.11.011

[18] Sun, S., Dong, B., & Zou, Q. (2021). Revisiting genome-wide association studies from statistical modelling to machine learning. *Briefings in Bioinformatics*, *22*(4), bbaa263. https://doi.org/10.1093/bib/bbaa263

[19] Nicora, G., Vitali, F., Dagliati, A., Geifman, N., & Bellazzi, R. (2020). Integrated multi-omics analyses in oncology: A review of machine learning methods and tools. *Frontiers in Oncology*, *10*, 1030. https://doi.org/10.3389/fonc.2020.01030

[20] Aggarwal, S. (2021). Role of whole exome sequencing for unidentified genetic syndromes. *Current Opinion in Obstetrics and Gynecology*, *33*(2), 112–122. https://doi.org/10.1097/GCO.0000000000000688

[21] Alatrany, A. S., Hussain, A. J., Mustafina, J., & Al-Jumeily, D. (2022). Machine learning approaches and applications in genome wide association study for Alzheimer's disease: A

systematic review. *IEEE Access*, *10*, 62831–62847. https://doi.org/10.1109/ACCESS.2022.3182543

[22] Bartha, Á., & Györffy, B. (2019). Comprehensive outline of whole exome sequencing data analysis tools available in clinical oncology. *Cancers*, *11*(11), 1725. https://doi.org/10.3390/cancers11111725

[23] Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1), 281. https://doi.org/10.1186/s12911-019-1004-8

[24] Roman-Naranjo, P., Parra-Perez, A. M., & Lopez-Escamez, J. A. (2023). A systematic review on machine learning approaches in the diagnosis of rare genetic diseases. *medRxiv, 2023-01*. https://doi.org/10.1101/2023.01.30.23285203

[25] Lima, Z. S., Ghadamzadeh, M., Arashloo, F. T., Amjad, G., Ebadi, M. R., & Younesi, L. (2019). Recent advances of therapeutic targets based on the molecular signature in breast cancer: Genetic mutations and implications for current treatment paradigms. *Journal of Hematology & Oncology*, *12*, 38. https://doi.org/10.1186/s13045-019-0725-6

[26] Elbracht, M., Mackay, D., Begemann, M., Kagan, K. O., & Eggermann, T. (2020). Disturbed genomic imprinting and its relevance for human reproduction: Causes and clinical consequences. *Human Reproduction Update*, *26*(2), 197–213. https://doi.org/10.1093/humupd/dmz045

[27] Tong, H., & Nikoloski, Z. (2021). Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. *Journal of Plant Physiology*, *257*, 153354. https://doi.org/10.1016/j.jplph.2020.153354

[28] Borkenhagen, L. K., Allen, M. W., & Runstadler, J. A. (2021). Influenza virus genotype to phenotype predictions through machine learning: A systematic review: Computational prediction of influenza phenotype. *Emerging Microbes & Infections*, *10*(1), 1896–1907. https://doi.org/10.1080/22221751.2021.1978824

[29] Nichol, D., Robertson-Tessi, M., Anderson, A. R. A., & Jeavons, P. (2019). Model genotype–phenotype mappings and the algorithmic structure of evolution. *Journal of the Royal Society Interface*, *16*(160), 20190332. https://doi.org/10.1098/rsif.2019.0332

[30] Li, L., Xiong, F., Wang, Y., Zhang, S., Gong, Z., Li, X., . . . , & Guo, C. (2021). What are the applications of single-cell RNA sequencing in cancer research: A systematic review. *Journal of Experimental & Clinical Cancer Research*, *40*(1), 163. https://doi.org/10.1186/s13046-021-01955-1

[31] Chen, G., Ning, B., & Shi, T. (2019). Single-cell RNA-seq technologies and related computational data analysis. *Frontiers in Genetics*, *10*, 317. https://doi.org/10.3389/fgene.2019.00317

[32] Liu, P., Liu, S., Fang, Y., Xue, X., Zou, J., Tseng, G., & Konnikova, L. (2020). Recent advances in computer-assisted algorithms for cell subtype identification of cytometry data. *Frontiers in Cell and Developmental Biology*, *8*, 234. https://doi.org/10.3389/fcell.2020.00234

[33] Raimundo, F., Meng-Papaxanthos, L., Vallot, C., & Vert, J. P. (2021). Machine learning for single-cell genomics data analysis. *Current Opinion in Systems Biology*, *26*, 64–71. https://doi.org/10.1016/j.coisb.2021.04.006

[34] Priya, P., Aneesh, B., & Harikrishnan, K. (2021). Genomics as a potential tool to unravel the rhizosphere microbiome interactions on plant health. *Journal of Microbiological Methods*, *185*, 106215. https://doi.org/10.1016/j.mimet.2021.106215

[35] Erfanian, N., Heydari, A. A., Iañez, P., Derakhshani, A., Ghasemigol, M., Farahpour, M., . . . , & Sahebkar, A. (2023). Deep learning applications in single-cell omics data analysis. *bioRxiv*. https://doi.org/10.1101/2021.11.26.470166

[36] Hia, N. T., & Ahmed, S. (2022). Automatic cell type annotation using supervised classification: A systematic literature review. *Systematic Literature Review and Meta-Analysis Journal*, *3*(3), 99–108. https://doi.org/10.54480/slrm.v3i3.45

[37] Panina, Y., Karagiannis, P., Kurtz, A., Stacey, G. N., & Fujibuchi, W. (2020). Human Cell Atlas and cell-type authentication for regenerative medicine. *Experimental & Molecular Medicine*, *52*(9), 1443–1451. https://doi.org/10.1038/s12276-020-0421-1

[38] Paik, D. T., Cho, S., Tian, L., Chang, H. Y., & Wu, J. C. (2020). Single-cell RNA sequencing in cardiovascular development, disease and medicine. *Nature Reviews Cardiology*, *17*(8), 457–473. https://doi.org/10.1038/s41569-020-0359-y

[39] Emu, I. J., & Ahmed, S. (2022). Trajectory inference in single cell data: A systematic literature review. *Systematic Literature Review and Meta-Analysis Journal*, *3*(3), 109–116. https://doi.org/10.54480/slrm.v3i3.46

[40] Razzouk, S. (2024). Single-cell sequencing, spatial transcriptome ad periodontitis: Rethink pathogenesis and classification. *Oral Diseases*, *30*(5), 2771–2783. https://doi.org/10.1111/odi.14761

[41] Hu, B., Sajid, M., Lv, R., Liu, L., & Sun, C. (2022). A review of spatial profiling technologies for characterizing the tumor microenvironment in immuno-oncology. *Frontiers in Immunology*, *13*, 996721. https://doi.org/10.3389/fimmu.2022.996721

[42] Ferreira, R. M., Gisch, D. L., & Eadon, M. T. (2022). Spatial transcriptomics and the kidney. *Current Opinion in Nephrology and Hypertension*, *31*(3), 244–250. https://doi.org/10.1097/mnh.0000000000000781

[43] Talukder, A., Barham, C., Li, X., & Hu, H. (2021). Interpretation of deep learning in genomics and epigenomics. *Briefings in Bioinformatics*, *22*(3), bbaa177. https://doi.org/10.1093/bib/bbaa177

[44] Jiang, Y., Xie, J., Tan, X., Ye, N., & Nguyen, Q. (2023). Generalization of deep learning models for predicting spatial gene expression profiles using histology images: A breast cancer case study. *bioRxiv, 2023-09*. https://doi.org/10.1101/2023.09.20.558624

[45] Caudai, C., Galizia, A., Geraci, F., Le Pera, L., Morea, V., Salerno, E., . . . , & Colombo, T. (2021). AI applications in functional genomics. *Computational and Structural Biotechnology Journal*, *19*, 5762–5790. https://doi.org/10.1016/j.csbj.2021.10.009

[46] Liu, K., & Zhang, X. (2022). PiTLiD: Identification of plant disease from leaf images based on convolutional neural network. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *20*(2), 1278–1288. https://doi.org/10.1109/TCBB.2022.3195291

[47] Kumaraswamy, A., Leng, K. R. W., Westbrook, T. C., Yates, J. A., Zhao, S. G., Evans, C. P., . . . , & Alumkal, J. J. (2021). Recent advances in epigenetic biomarkers and epigenetic targeting in prostate cancer. *European Urology*, *80*(1), 71–81. https://doi.org/10.1016/j.eururo.2021.03.005

[48] Binnie, A., Tsang, J. L. Y., Hu, P., Carrasqueiro, G., Castelo-Branco, P., & Dos Santos, C. C. (2020). Epigenetics of sepsis.

Critical Care Medicine, *48*(5), 745–756. https://doi.org/10.1097/ccm.0000000000004247

[49] Vespero, F. H. (2023). Epigenetic regulation of gene expression in the development of neurodegenerative diseases: A narrative review. *International Journal of Medical Science and Clinical Research Studies*, *3*(10), 2385–2393. https://doi.org/10.47191/ijmscrs/v3-i10-52

[50] Fan, Y., & Pedersen, O. (2021). Gut microbiota in human metabolic health and disease. *Nature Reviews Microbiology*, *19*(1), 55–71. https://doi.org/10.1038/s41579-020-0433-9

[51] Wen, B., & Zhang, B. (2020). Computational proteomics: Focus on deep learning. *Proteomics*, *20*(21–22), 2000258. https://doi.org/10.1002/pmic.202000258

[52] Vu, T. T. D., & Jung, J. (2021). Protein function prediction with gene ontology: From traditional to deep learning models. *PeerJ*, *9*, e12019. https://doi.org/10.7717/peerj.12019

[53] Babu, N., Bhat, M. Y., John, A. E., & Chatterjee, A. (2021). The role of proteomics in the multiplexed analysis of gene alterations in human cancer. *Expert Review of Proteomics*, *18*(9), 737–756. https://doi.org/10.1080/14789450.2021.1984884

[54] Buccitelli, C., & Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics*, *21*(10), 630–644. https://doi.org/10.1038/s41576-020-0258-4

[55] Kulkarni, S. R., & Vandepoele, K. (2020). Inference of plant gene regulatory networks using data-driven methods: A practical overview. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, *1863*(6), 194447. https://doi.org/10.1016/j.bbagrm.2019.194447

[56] Haque, S., Ahmad, J. S., Clark, N. M., Williams, C. M., & Sozzani, R. (2019). Computational prediction of gene regulatory networks in plant growth and development. *Current Opinion in Plant Biology*, *47*, 96–105. https://doi.org/10.1016/j.pbi.2018.10.005

[57] Davis, B. C., Brown, C., Gupta, S., Calarco, J., Liguori, K., Milligan, E., . . . , & Keenum, I. (2023). Recommendations for the use of metagenomics for routine monitoring of antibiotic resistance in wastewater and impacted aquatic environments. *Critical Reviews in Environmental Science and Technology*, *53*(19), 1731–1756. https://doi.org/10.1080/10643389.2023.2181620

[58] Álvarez-Mercado, A. I., Navarro-Oliveros, M., Robles-Sánchez, C., Plaza-Díaz, J., Sáez-Lara, M. J., Muñoz-Quezada, S., . . . , & Abadía-Molina, F. (2019). Microbial population changes and their relationship with human health and disease. *Microorganisms*, *7*(3), 68. https://doi.org/10.3390/microorganisms7030068

[59] Azeem, M., Soundari, P. G., Ali, A., Tahir, M. I., Imran, M., Bashir, S., . . . , & Zhang, Z. (2022). Soil metaphenomics: A step forward in metagenomics. *Archives of Agronomy and Soil Science*, *68*(12), 1645–1663. https://doi.org/10.1080/03650340.2021.1921743

[60] Saito, M. A., Bertrand, E. M., Duffy, M. E., Gaylord, D. A., Held, N. A., Hervey IV, W. J., . . . , & Walsh, D. A. (2019). Progress and challenges in ocean metaproteomics and proposed best practices for data sharing. *Journal of Proteome Research*, *18*(4), 1461–1476. https://doi.org/10.1021/acs.jproteome.8b00761

[61] Sekyere, J. O., & Faife, S. L. (2021). Pathogens, virulence and resistance genes surveillance with metagenomics can pre-empt dissemination and escalation of untreatable infections: A systematic review and meta-analyses. *bioRxiv, 2021-06*. https://doi.org/10.1101/2021.06.30.450418

[62] Chenarani, N., Emamjomeh, A., Allahverdi, A., Mirmostafa, S., Afsharinia, M. H., & Zahiri, J. (2021). Bioinformatic tools for DNA methylation and histone modification: A survey. *Genomics*, *113*(3), 1098–1113. https://doi.org/10.1016/j.ygeno.2021.03.004

[63] Datta, S., Rajnish, K. N., Samuel, M. S., Pugazlendhi, A., & Selvarajan, E. (2020). Metagenomic applications in microbial diversity, bioremediation, pollution monitoring, enzyme and drug discovery. A review. *Environmental Chemistry Letters*, *18*, 1229–1241. https://doi.org/10.1007/s10311-020-01010-z

[64] Raut, J. R., Guan, Z., Schrotz-King, P., & Brenner, H. (2019). Whole-blood DNA methylation markers for risk stratification in colorectal cancer screening: A systematic review. *Cancers*, *11*(7), 912. https://doi.org/10.3390/cancers11070912

[65] Fellows, R., & Varga-Weisz, P. (2020). Chromatin dynamics and histone modifications in intestinal microbiota-host crosstalk. *Molecular Metabolism*, *38*, 100925. https://doi.org/10.1016/j.molmet.2019.12.005

[66] Liu, H., Li, P., Wei, Z., Zhang, C., Xia, M., Du, Q., . . . , & Yang, X. P. (2019). Regulation of T cell differentiation and function by epigenetic modification enzymes. In B. Gaudillière, F. Sallusto & K. Yamamoto (Eds.), *Seminars in immunopathology* (Vol. 41, pp. 315–326). https://doi.org/10.1007/s00281-019-00731-w

[67] Dobre, E. G., Constantin, C., Costache, M., & Neagu, M. (2021). Interrogating epigenome toward personalized approach in cutaneous melanoma. *Journal of Personalized Medicine*, *11*(9), 901. https://doi.org/10.3390/jpm11090901

[68] Sahafnejad, Z., Ramazi, S., & Allahverdi, A. (2023). An update of epigenetic drugs for the treatment of cancers and brain diseases: A comprehensive review. *Genes*, *14*(4), 873. https://doi.org/10.3390/genes14040873

[69] Estrela, S., Sánchez, Á., & Rebolleda-Gómez, M. (2021). Multi-replicated enrichment communities as a model system in microbial ecology. *Frontiers in Microbiology*, *12*, 657467. https://doi.org/10.3389/fmicb.2021.657467

[70] Fassarella, M., Blaak, E. E., Penders, J., Nauta, A., Smidt, H., & Zoetendal, E. G. (2021). Gut microbiome stability and resilience: Elucidating the response to perturbations in order to modulate gut health. *Gut*, *70*(3), 595–605. https://doi.org/10.1136/gutjnl-2020-321747

[71] Esvap, E., & Ulgen, K. O. (2021). Advances in genome-scale metabolic modeling toward microbial community analysis of the human microbiome. *ACS Synthetic Biology*, *10*(9), 2121–2137. https://doi.org/10.1021/acssynbio.1c00140

[72] van Bruggen, A. H. C., Goss, E. M., Havelaar, A., van Diepeningen, A. D., Finckh, M. R., & Morris Jr, J. G. (2019). One health-cycling of diverse microbial communities as a connecting force for soil, plant, animal, human and ecosystem health. *Science of the Total Environment*, *664*, 927–937. https://doi.org/10.1016/j.scitotenv.2019.02.091

[73] Garcia Gonzalez, J., & Hernandez, F. J. (2022). Nuclease activity: An exploitable biomarker in bacterial infections. *Expert Review of Molecular Diagnostics*, *22*(3), 265–294. https://doi.org/10.1080/14737159.2022.2049249

[74] Chen, T., & Tyagi, S. (2020). Integrative computational epigenomics to build data-driven gene regulation hypotheses. *GigaScience*, *9*(6), giaa064. https://doi.org/10.1093/gigascience/giaa064

[75] Galuzio, P. P., & Cherif, A. (2022). Recent advances and future perspectives in the use of machine learning and mathematical models in nephrology. *Advances in Chronic Kidney Disease*, *29*(5), 472–479. https://doi.org/10.1053/j.ackd.2022.07.002

[76] Shade, J., Coon, H., & Docherty, A. R. (2019). Ethical implications of using biobanks and population databases for genetic suicide research. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *180*(8), 601–608. https://doi.org/10.1002/ajmg.b.32718

[77] Kumar, G., Basri, S., Imam, A. A., Khowaja, S. A., Capretz, L. F., & Balogun, A. O. (2021). Data harmonization for heterogeneous datasets: A systematic literature review. *Applied Sciences*, *11*(17), 8275. https://doi.org/10.3390/app11178275

How to Cite: Adeyanju, S. A., & Ogunjobi, T. T. (2024). Machine Learning in Genomics: Applications in Whole Genome Sequencing, Whole Exome Sequencing, Single-Cell Genomics, and Spatial Transcriptomics. *Medinformatics*. https://doi.org/10.47852/bonviewMEDIN42024120