Time Series Analysis in DNA Sequencing for Monitoring Disease Progression and Treatment Response

Ritik Roshan  
School Of Computer Science Engineering & Technology  
*Bennett University*Greater Noida, Uttar Pradesh, India  
e23mcag0012@bennett.edu.in Yatharth Aggarwal  
School Of Computer Science Engineering & Technology  
*Bennett University*Greater Noida, Uttar Pradesh, India  
e23mcag0012@bennett.edu.in Dr. Eth e Sham  
School Of Computer Science Engineering & Technology  
*Bennett University*Greater Noida, Uttar Pradesh, India  
ethesham@bennett.edu.in

*Abstract*— Time series analysis is an effective method for comprehending dynamic biological processes, such as the progression of disease and the response to treatment. Time series data in the context of DNA sequencing provide an extensive picture of molecular changes over time, yielding important insights into underlying mechanisms of disease and the effectiveness of treatment interventions. This study examines the use of time series analysis in DNA sequencing to track the advancement of disease and the effectiveness of treatment. We go over several statistical techniques and computational strategies used to examine gene expression profiles, DNA methylation patterns, and chromatin accessibility dynamics in longitudinal genomic data. In addition, we draw attention to the difficulties and possibilities associated with combining clinical data with time series analysis to support precision medicine and personalized treatment. We demonstrate how time series analysis can be used to better understand disease processes and to untangle the temporal dynamics of biological systems. In conclusion, we go over upcoming directions and new developments in this multidisciplinary subject, highlighting how time series analysis may help advance precision medicine’s prognostics, therapeutic approaches and diagnostics.

Keywords—Time Series Analysis, DNA Sequencing, Disease Progression, Treatment Response, Gene Expression, Deep Learning, Dynamic Modeling, Longitudinal Data Analysis, RNA, Precision Medicine, Bioinformatics.

# Introduction

In biological research, time series analysis has become a crucial tool that allows for unparalleled granularity in studying dynamic events that occur over time. Time series data provide the opportunity to analyse the temporal dynamics of molecular events driving illness progression and treatment response in the context of DNA sequencing. It is crucial to comprehend these dynamic shifts in order to clarify the intricate processes behind illnesses and to create patient-specific treatment plans that work.

Biomedical research has been transformed by high-throughput sequencing technologies, which offer an unparalleled amount of genomic data. DNA sequencing allows for the investigation of dynamic changes in gene expression, epigenetic alterations, and chromatin structure across time, in addition to making it easier to identify genetic variants linked to disease. Researchers can find novel biomarkers, therapeutic targets, and predictive models by tracking the evolution of genetic signatures linked to disease states and therapeutic interventions through the analysis of longitudinal genomic data.

In this review, we concentrate on the use of time series analysis in DNA sequencing to track the advancement of disease and the effectiveness of treatment. We explore the statistical and computational techniques used to evaluate longitudinal genomic data, such as network interference, dynamic modelling, and time series clustering. To facilitate precision healthcare and tailored medicine, we also investigate the integration of time series analysis with clinical data, including patient outcomes and treatment plans.

Our goal is to demonstrate the value of time series analysis in deciphering the temporal dynamics of biological systems through an extensive literature survey and illustrative examples. Time series analysis has the potential to improve patient outcomes and advance our understanding of diseases by clarifying the complex interactions among genetic modification, disease development, and therapeutic interventions. We also explore the potential and difficulties of using time series analysis for precision medicine and suggest future lines of inquiry for this quickly developing field of study.

# Litreture review

Understanding the complex dynamics of illness progression and therapy response has grown increasingly dependent on the use of time series analysis to DNA sequencing data. Taking advantage of the abundance of genetic data that high-throughput sequencing technologies give, the emerging area provides a thorough understanding of molecular changes across time. Numerous studies have emerged in this field that have revealed the temporal subtleties of chromatin dynamics, epigenetic changes, and gene expression in a variety of illness situations.

Understanding the temporal dynamics of gene expression profiles is an important area of research. For example, research like that done by Liu et al. (2019) [1]has used time series RNA sequencing data to identify the dynamic transcriptional responses that occur during infection, providing insights into immune evasion mechanisms and host - pathogen interactions. Comparably, Zhu et al. (2020) [2]investigated the temporal development of gene expression patterns in cancer cells treated with chemotherapy, discovering new targets for treatment and indicators of prognosis. These studies highlight the values of time - resolved transcriptome analysis in understanding the molecular mechanisms underlying the development of disease and the effectiveness of therapeutic approaches.

Chromatin dynamics and epigenetic alterations are new areas of investigation for time series analysis of DNA sequencing data. In order to understand the dynamic changes in DNA methylation patterns during cellular reprogramming studies like that conducted by Farlik et al. (2016) [3]have used time series analysis. This has allowed researchers to gain insight into the epigenetic regulation of pluripotency and differentiation. Additionally, Buenrostro et al. (2018) used time series ATAC- seq data to investigate how chromatin accessibility dynamically changes in response to development cues and environmental stimuli, revealing regulatory mechanisms that underlie decisions about cellular destiny. These studies demonstrate how dynamic chromatin organization and epigenetic changes are, and how important they are in determining gene expression patterns and cellular phenotypes throughout a range of temporal scales.

 In the middle of these efforts, statistical and computational techniques have grown, providing a toolkit for the analysis of longitudinal genetic data. Ernst et al. (2021) have suggested Bayesian hierarchical models that are noteworthy for their ability to discover dynamic signalling networks and temporally coordinated gene regulatory modules from time series gene expression data. Furthermore, Hughey et al. (2019) created the DYAN framework, which makes it easier to identify the regulatory relationships underpinning cellular functions and disease phenotypes by inferring dynamic regulatory networks from time series genomic data. By advancing our understanding of the dynamic regulation of gene expression and cellular function, these computational advancements highlight the synergy between time series analysis and computational biology.

Time series analysis and clinical data integration has significant consequences for precision healthcare and personalized therapy. Personalized treatment strategies based on molecular signatures have been made possible by studies like that conducted by Almanzar et al. (2020), which combined time series gene expression data with clinical metadata to create predictive models for the course of the disease and the response to treatment in patients with inflammatory bowel disease. Similar to this, Patel et al (2021) identifies temporal patterns of disease development and treatment response in cancer patients by combining longitudinal genetic data with electronic health records. This allowed for early identification and precision oncology. By bridging the gap between genetic findings and clinical applications, time series analysis has the potential to revolutionize patient care and therapeutic tactics. These integrated approaches highlight this study of time series analysis in DNA sequencing.

In summary, time series analysis in DNA sequencing is a multidimensional method to decipher the dynamic landscape of molecular changes that drive disease development and therapy response. Researchers have uncovered temporal trends in gene expression, epigenetic alterations, and chromatin dynamics by utilizing longitudinal genomic data and computational approaches. These findings provide insights into the mechanisms underlying disease and potential treatment targets. Personalized medicine has great potential when it integrates clinical data, as this can facilitate the application of genomic insights in clinical practice and enhance patient outcomes. To advance the discipline, interdisciplinary cooperation and creative solutions are needed to address ongoing issues with data integration, computational scalability, and clinical implementation. [4]

# Objectives

This study’s main goal is to examine current computational techniques and methodologies for time series analysis of DNA sequencing data. Examining numerous statistical methods, bioinformatics tools, and computer algorithms for interpreting temporal dynamics in the analysis of longitudinal genomic data is part of this research paper.

The use of time series analysis is to comprehend the biological mechanism behind illness progression. Examining dynamic shifts in chromatin structure, epigenetic alterations, and gene expression across time in various illness situations, including cancer, autoimmune disorders, and infectious diseases, is part of this research paper.

Talking about the implications for personalized medicine of integrating time series analysis with clinical data is a key goal. This involves investigating the ways in which clinical decision-making and patient care can be enhanced by combining longitudinal genetic data with clinical information, patient outcomes, and longitudinal clinical profiles.

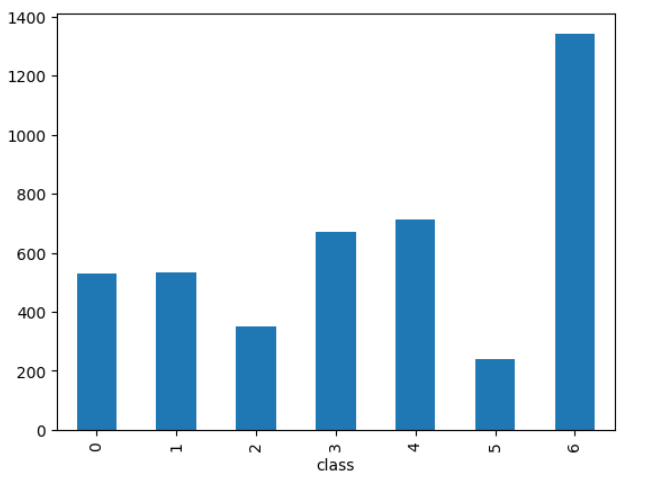
The goal of the project is to use time series analysis of DNA sequencing data to look at the dynamics of therapy response, explaining mechanisms and relapse, and characterizing temporal changes in molecular profiles after therapeutic treatments.

Finally, the research looks for future possibilities and obstacles in the field of time series analysis in DNA sequencing. In order to overcome these obstacles and move the field closer to precision medicine and personalized therapies, this entails talking about technological constraints, computational bottlenecks, and methodological gaps as well as putting up creative solutions and interdisciplinary partnerships.

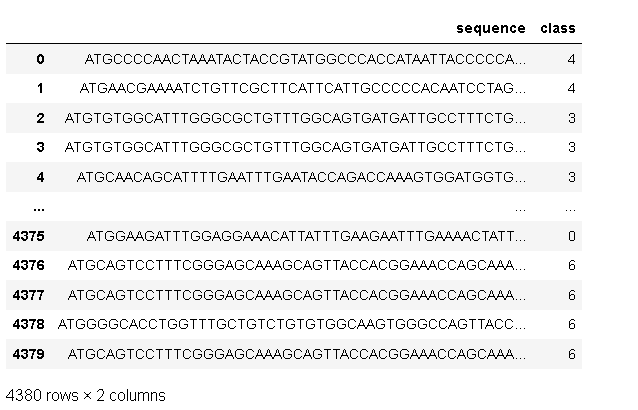
# Research Methodology

**Data Collection -**

The entire genetic sequence of viruses, including SARS, MERS, dengue, hepatitis and influenza, was retrieved from “The National Centre for Biotechnology Information (NCBI)” a public nucleotide sequence database.[5] The DNA sequence data is in FASTA format and contains metadata. The sequence is between 1 to 4800 nucleoids long. The class distribution for each class with the number of samples displays sample DNA sequences from the collection that include a virus’s entire genomic sequence, the length of the sequence, and the class labels.



**Figure 1. Shows the distribution of each class and sample sequence in a dataset**

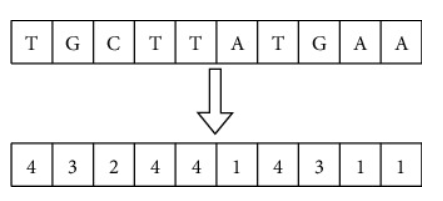


**Figure 2. Shows the sample dataset with genomic sequences of Human DNA & its length**

The dataset makes it very evident that there is a problem with an imbalanced dataset. To address this issue, SMOTE (Synthetic Minority Oversampling Technique) is used. There are insufficient numbers of MERS and dengue DNA sequence in our collection.[5] The SMOTE method is used to create synthetic samples that closely resemble the majority class for the minority classes, such as MERS and dengue. SMOTE looks for the k nearest minority class neighbours of a randomly selected minority class instance. Then, to create the synthetic instance, a line segment is supplied by joining a and b with the feature space, randomly selecting one of the k nearest neighbours, b. The two selected cases, a and b. Are combined in a convex manner to construct the synthetic instances. For the minority classes, this process can be used to create an artificial instance.

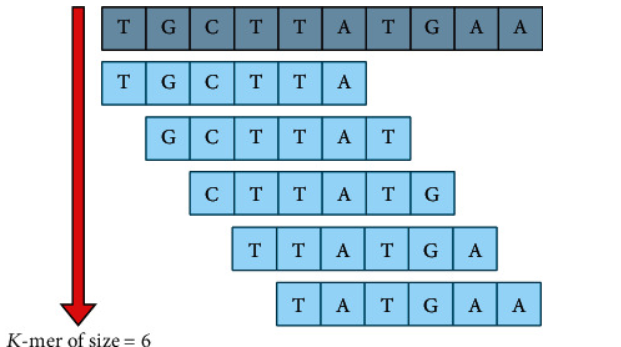
**Data Preprocessing -**

For the majority of machine learning and deep learning algorithms that use numerical rather than categorical data, preprocessing the data is the most important stage. The DNA dataset’s genomic sequence is classified. [6]The categorical data can be converted to numerical using a variety of methods. The method of translating the nucleotide’s category data into numerical form is known as the encoding approach. The DNA sequence encoded in this work using label encoding method affects the categorization accuracy. [7]Label encoding assigns an index value to each nucleotide in the DNA sequence, such as A-1, C-2, G-3 and T-4. Using Label Binarizer, the complete DNA sequence is transformed into a range of numbers ().



**Figure 3. Sequence data encoding using Label Binarize**

In the K-Mer encoding process, the raw DNA sequence is turned into an English-like statement by creating K0mers. In every DNA sequence is converted into K-Mer of size m, and all of the K-Mers produced are concatenated to create a sentence. It is now possible to classify DNA sequences using the natural language processing techniques. This work converts the K-Mer text into a dense feature matrix using the word embedding layer.[7]



**Figure 4. Sequence data encoding using K-mer technique**

## Classification Models-

  In this research paper, DNA sequence categorization is done using three different classification models: - CNN, CNN-LSTM, and CNN-Bidirectional LSTM.

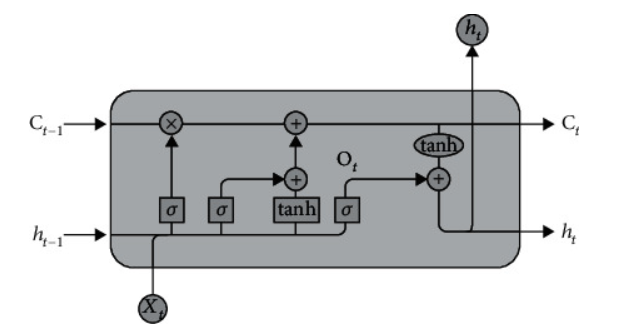
The DNA sequence is encrypted using the label encoding and K-mer techniques, which maintain the positional information of every nucleotide in the sequence. The data from the first two methods are embedded using the embedding layers. The CNN layer is employed as the step for extracting features, and it provides the input for the classification stages of LSTM and bidirectional LSTM.

**CNN -** For the majority of classification problems, CNN is a popular deep-learning method that produces state-of-the-art results. CNN may yield strong accuracy on text data in addition to performing well on image classification. For the most part, CNN is used to automatically extract features from the input dataset, with machine learning models, however, features must be chosen by the user. For image and video data, 2D CNN and 3D CNN are utilized, where 1D CNN is employed for text classification. Nucleotide A,C,G, and T are the letters that make up the DNA sequence. The DNA sequence is transformed into numerical values by either one hot encoding or label encoding as CNN can only process numerical input.[8] A sequence of convolutional layers of the CNN architecture are used to extract features from the input dataset. Following each convolutional layer is a max pooling layer, which reduces the size of the retrieved features. Function extraction in the convolutional layer is significantly influenced by the kernel’s size. The number of filters and kernel size are the model’s hyperparameters. The proposed CNN model’s whole architecture is summarized. The embedded layer, which has dimension eight, is the first layer. Using the frequency with which a word occurs near other words, this layer transforms the words into a vector space model. To determine the embedding for each phrase in the training dataset, the embedding layer employs random weights. RLU is used as an activation function for feature extraction in a kernel of size(2x2), and two convolutional layers with filters of 128 and 64 are added to the model. [8]The addition of a max pooling layer with a size of 2x2 reduces the dimensions of the feature map. Lastly, the flatter layer is used to transform the feature maps into a single-column vector. A dense layer comprising neurons 128 and 64 receives the output. The classification layer, which can work effectively for the multiclass problem, uses the SoftMax layer, N is the number of units. Every unit has a complete connection to the previous layer and used Equation to calculate the probability of every class on N.

is the weight matrix connecting the mth unit to the previous layer, x is the last layer’s output, and is the mth unit’s bias.

**LSTM -** Sequence prediction and classification use long short-term memory (LSTM), a recurrent neural network (RNN) that can learn long-term dependencies in a sequence. It consists of a set of cells, or memory blocks, with three gates in each: forget, output, and input. In this situation, the LSTM will selectively recall and forget items. The general design of the LSTM model is shown below the figure. It can detect the lengthy sequence and learn new things. The current state is calculated using the equation.

Where is the input state,is the current state, and is the previous state.

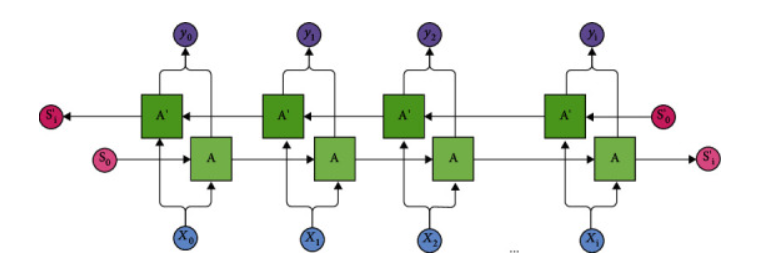


**Figure 5. Architecture of the LSTM model.**

In the LSTM, the forget gate is responsible for eliminating any data from the cell state. The forget gate transmits a value of 0, indicating that the data should be deleted from the memory cell, when detailed information is no longer valid for the sequence categorization. Two inputs are used by this gate: (input from the present state) and (input from the past state). The input will be added with ias after being multiplied by a weight. The sigmoid function is then used. The output of the sigmoid function is a number between 0 and 1. The LSTM’s input gate is in charge of adding each pertinent value to the cell state. It involves two activation functions. The values that contribute to the cell state are first controlled by the sigmoid function.[8]

The second function is called tanh and it returns values between -1 and +1, which represent all possible values that could be applied to the cell state. The sigmoid activation function and the tan h activation function are applied to the cell state by the output gate of an LSTM to determine what value can be in the output. In our model, the convolutional layers are followed by an LSTM layer with 100 memory units, which predicts that classification labels. The LSTM layer receives the features that the convolutional layer extracted as an input for classification. CNN and LSTM are merged into a hybrid model in various NLP tasks in order to increase classification accuracy. It’s astonishing how well this hybrid model has performed at classifying texts. To lessen the overfitting issue, the LSTM model incorporated dropout layers and regularization strategies.

**CNN-Bidirectional LSTM -** For classifying DNA sequences, a bidirectional hybrid model consisting of LSTM and CNN is employed. The model classifies using bidirectional LSTM and extracts features using CNN. With regard to learning the dependencies in the forward sequence and the dependencies in the backward sequence, the bidirectional LSTM comprises two RNN.



**Figure 6. Architecture of bidirectional LSTM**

# Result and Discussion

Training for Bidirectional LSTM-RNN and GRU Model:

As mentioned before, about 80% of the whole dataset was used to train the model. The training dataset consists of two components, Train X and Train Y, which together make up approximately 80% of the dataset. All nucleotide sequences and their corresponding labels. The TensorFlow library computes accuracy metrics, which show how frequently predictions and labels match, as well as the loss between true and predicted labels.

Testing Results for the Bidirectional LSTM-RNN and GRU Model:

To test the model, twenty percent of the entire dataset was used. Twenty percent of the dataset is made up of Tests X and Y. To evaluate the model, Twenty percent of the testing data is used. Twenty percent of the test data, along with the X and Y vectors including all nucleotide sequences, window sizes, and labels, are contained in Table 3. The model’s capacity to abstract from diverse knowledge sources is tested using test data. Using the test data, the model attains an accuracy rate of 96.1 percent. The highest performance we achieved was 50 cycles off 70 epochs when we tested 20:80 epochs with a cycle rate ranging from 10 to 70. An input sequence of a randomly selected C. parvum entire genome was used both before and after testing the model to predict the exons from the trained model. An intron length-based filter was incorporated into the model to increase the precision of exon and intron prediction. This improved the precision of the model.

The model will be given the labels 0, 1, and 2 when it is executed. This illustrates how the counter indicates that 0 = G.T., 1 equals A.G., and 2 equals no-site inside a specific window of sequences. If a window contains more counts for each label than a predetermined amount, then this is the case. The predicted exons and introns of the proposed model. These forecasts are contrasted with C. parvum annotated genome [36]. The length of the C. Parvum ATCC mRNA annotated CDS sequence. Using short-read sequencing data to precisely delineate UTR boundaries is difficult due to high gene density and short gene distances. In the absence of reference annotation or the genome, RNA-Seq transcriptome assembly leads to a high chimerism rate. Consequently, we used reference annotation to construct the assembly process and optimized the parameters to reduce the number of transcripts that were erroneously fused. Spliced exons Favor longer introns because of the length restrictions imposed by the two basic splicing methods. DNA portions known as exons are where a gene’s amino acid sequence is found. In both plants and animals, the majority of gene sequences are broken up by one or more DNA segments known as introns. If a different exon was selected, it was undoubtedly among the several possible exons found further down the line.

In addition, exons from adjacent genes were identified as corresponding to fourteen of the OGA’s Coding regions in the revised annotation. Our research led us to discover 451 previously unidentified introns and 511 extra exons. There are two to ten exons in genes with introns, with an average of 2.6 exons per gene. This indicates that the model is evaluated with around 96.1 accuracy using annotated C. parvum benchmark data. When compared to the state-of-the-art, the suggested strategy performs better in Table 5 when it comes to predicting splice positions. On this test, our unique bidirectional LSTM-RNN and GRU outperform earlier DL algorithms. We covered how bidirectional LSTM-RNN and GRU allow neural networks to store both forward and backward information from the data states of hidden sequences. The machine can now learn far more efficiently as a result. Data provided in a sequential fashion is best handled by an RNN model [28]. By Increasing the number of interactions in RNN, only the LSTM is able to boost accuracy and overcome the problem of vanishing gradients. The network does not get better when we operate and change the state of our computer, as Table 5 illustrates. It is known as a vanishing gradient phenomenon. The bidirectional LSTM-RNN and GRU were selected for this research project because they can used in both directions.[9]

# Conclusion

Personalized sequencing has emerged, providing previously unattainable insights into individual genetic variants and their consequences for health and disease. This marks a tremendous development in the study and medical management of disorders. Personalized sequencing has gained momentum despite obstacles and has proven useful in medical applications, especially in the diagnosis and treatment of diseases. By using customized sequencing, we can shed light on the underlying genetic pathways of both uncommon and well-studied diseases, including cancer, and begin to unravel their complicated aetiologies. Its effectiveness in uncommon Mendelian illness is especially noteworthy, as the discovery of single genetic variants can clarify disease symptoms and direct clinical treatment by enabling precise diagnosis.

 Furthermore, individual tumour DNA sequencing has become crucial in directing treatment decisions in the field of cancer, where tailored sequencing has improved our diagnostics capabilities. Personalized sequencing allows physicians to customize the therapeutic approach to each patient’s distinct genomic profile, maximizing therapy success while minimizing side effects. This is accomplished by finding certain genetic abnormalities within tumours. Additionally, germline sequencing has yielded practical insights that have informed personalized pharmacologic interventions for cancer treatment that are tailored to both patient tolerance and therapeutic effectiveness. These insights have guided lifestyle modification for individuals who are at risk of diseases such as diabetes.

Personalized sequencing has great potential for use in preventive healthcare, even though its application to clinically healthy patients’ needs more confirmation in bigger cohorts. It is becoming more and more possible to deploy personalized sequencing on a broad scale as technology develops and interpretation accuracy rises. A paradigm change of this kind has the power to completely change the healthcare industry by moving it from a reactive model that concentrates on treating diseases that have already manifested to a proactive model that emphasizes early intervention and individualized prevention. Personalized sequencing essentially signals the dawn of a new era in precision medicine, one in which medical therapies are customized to each patient’s unique genetic profile. This will ultimately result in better patient outcomes and a fundamental change in the way healthcare is provided.

# Future Scope

 Personalized sequencing has the potential to completely transform healthcare in the near future thanks to technological breakthroughs and broad clinical practice integration. Long-read and single-cell sequencing will provide deeper insights into genomic architecture and cellular heterogeneity, while anticipated advancements in the sequencing technologies will improve accessibility and affordability. Personalized sequencing will be essential for disease diagnosis, prognosis, treatment selection, and monitoring as it becomes a standard procedure in clinical care. Additionally, its entry into preventive medicine will facilitate the early detection of disease risks and customized interventions, propelling the transition to proactive approaches in healthcare. Comprehensive sequencing programs and precision public health strategies will augment our comprehension of human genetic variation and disease vulnerability, and strong ethical, legal, and societal guidelines will guarantee the conscientious application of genetic information. All things considered, personalized sequencing has enormous potential to revolutionize healthcare by facilitating accurate customized interventions and enhancing population health outcomes.

# References

[1] C. X. Liu, S. K. Guo, F. Nan, Y. F. Xu, L. Yang, and L. L. Chen, “RNA circles with minimized immunogenicity as potent PKR inhibitors,” Mol Cell, vol. 82, no. 2, pp. 420-434.e6, Jan. 2022, doi: 10.1016/j.molcel.2021.11.019.

[2] H. Zhu et al., “Gene expression profiling of type 2 diabetes mellitus by bioinformatics analysis,” Comput Math Methods Med, vol. 2020, 2020, doi: 10.1155/2020/9602016.

[3] S. Rulands et al., “Genome-Scale Oscillations in DNA Methylation during Exit from Pluripotency,” Cell Syst, vol. 7, no. 1, pp. 63-76.e12, Jul. 2018, doi: 10.1016/j.cels.2018.06.012.

[4] P. J. Canatalay and O. N. Ucan, “A Bidirectional LSTM-RNN and GRU Method to Exon Prediction Using Splice-Site Mapping,” Applied Sciences (Switzerland), vol. 12, no. 9, May 2022, doi: 10.3390/app12094390.

[5] “lim18a”.

[6] Y. Wang et al., “Circulating tumor DNA analyses predict progressive disease and indicate trastuzumab-resistant mechanism in advanced gastric cancer,” EBioMedicine, vol. 43, pp. 261–269, May 2019, doi: 10.1016/j.ebiom.2019.04.003.

[7] A. M. Alaa and M. Van Der Schaar, “Attentive State-Space Modeling of Disease Progression.”

[8] M. Byrska-Bishop, U. S. Evani, X. Zhao, and A. O. Basile, “High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios”, doi: 10.1101/2021.02.06.430068.

[9] “Enhancing Viral DNA Sequence Classification Using Hybrid Deep Learning Models and Genetic Algorithm Optimization.” [Online]. Available: https://ssrn.com/abstract=4692259