

## **A combined method of genomics and bioinformatics for plant breeding programs**

**Debaghya Biswas** 

\*Corresponding author. Assistant Professor, Department of CS & IT, Kalinga University, Raipur, India. E-mail address: ku.debaghyabiswas@kalingauniversity.ac.in

**Pooja Sharma** 

Research Scholar, Department of CS & IT, Kalinga University, Raipur, India. E-mail address: pooja.sharma@kalingauniversity.ac.in

---

### ***Abstract***

#### **Objectives**

Given the rapid growth of the global human population, it is imperative to enhance agricultural productivity to fulfill the increasing demand for crops. Enhancing crops through selective plant breeding (PB) is a sustainable method to augment both the quantity and consistency of yields without escalating the need for fertilizers and pesticides. Moreover, data generation in agriculture and biotechnology has greatly increased in recent years due to the very rapid development of high-performance technologies. Similar to advancements in genomics, there are encouraging progressions in plant phenotyping technologies, including mechanized phenotyping apparatus and sophisticated picture analysis. This has led to an unparalleled understanding of PB, structure, and function. Therefore, the aim of this study was to investigate a combined method of genomics and bioinformatics for plant breeding programs.

#### **Results**

The latest advancements in genomics and bioinformatics offer possibilities for expediting crop enhancement. Third-generation sequencing (TGS) techniques are aiding in the resolution of difficulties in plant genome assembly arising from polyploidy and the presence of repetitive regions. There is a growing availability of superior crop-referencing genomes, which significantly aids in the analysis of genetic variations and the identification of specific PB objectives within the genome. Machine learning (ML) aids in identifying genomic areas of agricultural significance

by assisting in the annotation of genomes and allowing for the efficient measurement of agronomic variables in controlled and natural environments.

### Conclusions

Crop datasets that combine the increasing amount of genotype and phenotypic data offer a valuable tool for breeders and a chance for data mining methods to discover new candidate genes related to traits. Moreover, with the increasing understanding of agricultural genetics, the techniques of genomic selections and genome engineering provide the potential for developing crops resistant to illnesses and adaptable to stress while achieving high production.

**Keywords:** Bioinformatics, Crops, Genomics, Plant Breeding

**Paper Type:** Review Paper.

**Citation:** Biswas D, Sharma P (2024) A Combined Method of Genomics and Bioinformatics for Plant Breeding Programs. *Agricultural Biotechnology Journal* 16 (3), 211-228.

---

*Agricultural Biotechnology Journal* 16 (3), 211-228. DOI: 10.22103/jab.2024.23993.1601

Received: July 31, 2024.

Received in revised form: September 25, 2024.

Accepted: September 26, 2024.

Published online: September 30, 2024.

Publisher: Faculty of Agriculture and Technology Institute of Plant Production, Shahid Bahonar University of Kerman-Iranian Biotechnology Society.



© the authors

---

### Introduction

The progress of Plant Breeding (PB) is primarily limited by the need for a sufficient theoretical foundation in various aspects of current genetics, which necessitates the development of suitable mathematical frameworks (Swarup et al. 2021). The existence of such algorithms can significantly streamline and expedite the resolution of genotype identity issues, as well as forecast the amalgamation of quantitative characteristics of newly developed varieties. Integrating conventional crop modeling with innovative PB techniques and genetic modeling will expedite the development of novel cultivars suited for diverse environmental circumstances (Srinivasa et al. 2023). Many of these instruments and processes are associated with the OMICs grouping (Brière et al. 2021). Moreover, data generation in agriculture and biotechnology has greatly increased in recent years due to the very rapid development of high-performance technologies

(Mohammadabadi et al. 2024). These data are obtained from studying biological molecules, such as metabolites, proteins, RNA, and DNA, to understand the role of these molecules in determining the structure, function, and dynamics of living systems (Pour Hamidi et al. 2017). Functional genomics is a field of research that aims to characterize the function and interaction of all the major components (DNA, RNA, proteins, and metabolites, along with their modifications) that contribute to the set of observable characteristics of a cell or individual (i.e., phenotype) (Williamson & Leiros 2020). Furthermore, in a breeding program, genetic improvement can be maximized through accurate identification of superior organisms that are selected as parents of the next generation, thereby achieving breeding goals (Mohammadabadi et al. 2024). Artificial neural networks have been proposed to alleviate this limitation of traditional regression methods and can be used to handle nonlinear and complex data, even when the data is imprecise and noisy (Pour Hamidi et al. 2017). Agricultural data can be too large and complex to handle through visual analysis or statistical correlations. This has encouraged the use of machine intelligence or artificial intelligence (Ghotbaldini et al. 2019). The utilization of OMICS technologies can optimize the nutritional value of crop residues, hence augmenting agricultural productivity for nutrition, animal feed, and energy generation (Kamusoko et al. 2021). OMICS has enhanced the reliability and foreseeability of PB programs, resulting in reduced time and cost for developing stress-tolerant crop types (Veerasamy & Fredrik 2023). Transcriptomics, a significant subcategory of OMICs, appeals to biologists, particularly those involved in PB (Kamusoko et al. 2021).

Similar to advancements in genomics, there are encouraging progressions in plant phenotyping technologies, including mechanized phenotyping apparatus and sophisticated picture analysis. This has led to an unparalleled understanding of PB, structure, and function. When comparing genomic studies to Plant Phenotyping (PP) (Bhattacharya et al. 2021), the data output generated by established methods in PP is still easily handled (Sayers et al. 2023). As per the National Center of Biotechnology Information (NCBI), bioinformatics is a scientific discipline that integrates biology, computer science, and technological expertise into a unified framework (Sayers et al. 2023). Bioinformatics is a field that uses mathematical, empirical, and computational techniques to address biological challenges by utilizing DNA, RNA, protein acid patterns, and related data (Cao et al. 2020). Bioinformatics has played a role in various scientific disciplines, including PB. PB has expanded by utilizing different strategies and equipment to achieve specific goals. The sequencing of a genome is merely the initial phase in comprehending genome organization, gene framework, and gene expression trends (Wang et al. 2021). There are just two model organisms, Arabidopsis and grain, for which the research has genome sequences that are both recent and almost complete. Numerous genome sequencing initiatives have been conducted for various plants, such as wheat and corn. The first significant improvement after

Sanger sequencing was the development of short-read technology, sometimes called second-generation sequencing (Sayers et al. 2023).

Machine Learning (ML) and other data science approaches can uncover new and valuable information by analyzing various datasets (Wan et al. 2022). The present knowledge about plant genetics and the rising sophistication of tests and tools have greatly accelerated the production of different plant types (Gladkov & Gladkova 2021). Utilizing contemporary bioinformatics methodologies is imperative for modern PB to expedite the reduction of labor and supplies expenses within a short timeframe (Chen et al. 2023). The variety framework is a scientific prediction that determines a crop's specific features to achieve a desired efficiency level, sustainability, and other necessary production attributes (Suruliandi et al. 2021). In machine learning, feature learning refers to the automated procedure by which a model maximizes its performance on a particular task by identifying and optimizing important patterns, structures, or traits (termed "features") from raw data. It is crucial because it allows machines to learn the most important traits automatically, rather than having to be manually engineered. This dramatically improves prediction accuracy and efficiency. The significant influence of PB in creating genetically better genotypes for particular environments necessitates an evaluation of bioinformatics's contribution to the PB research field (Chen et al. 2023). The primary objective of this study is to provide an overview of the current advancements in genomics, PP, and regulation. This will be achieved by examining a range of online software tools that are utilized to analyze the results of various molecular biology tests to enhance plant modification.

With the advent of Third-Generation Sequencing (TGS), the practice of sequencing and repetition genomes has become commonplace in various fields of PB (Xiao & Zhou 2020). TGS has many drawbacks, including intrinsic biases and uncertain alignment of repeated components. As a result of confirmation bias, it's possible to misunderstand crops or choose to disregard data that challenges the preconceptions. This may lead to issues with genome identification in the case that the approach misses a problem or makes a wrong decision. These shortcomings result in draft genome assembly and hinder the investigation of concealed indels and structural variations (Wan et al. 2022). TGS, such as Pacific Biosciences (PacBio) single-molecule sequencing and Oxford Nanopore Technology (ONT) sequencing (Mikheenko et al. 2022), has facilitated the fabrication of longer reads and improved the accuracy and continuity of genome assembly (Suruliandi et al. 2021). The integration of long-read sequencing, mapping methods, and chromosomal conformation capturing has enabled smaller facilities and non-model cultivars to achieve remarkably congruent chromosome-level crop genomic assembly (Dessy et al. 2023; Veerasamy & Fredrik 2023). Emerging optical mapping technologies, like BioNano Genomics, can

efficiently mark lengthy DNA molecules exceeding 250 kb (Bhattacharya et al. 2021). This facilitates the identification of structural variations and the creation of a well-constructed framework at a relatively cheap expense. A great example is Bionano Genomics, which builds optical maps of DNA molecules using third-generation sequencing (TGS) technology. Among the many applications for Bionano's optical maps are: Enhancing and verifying sequencing assembly, Find gene rearrangements between others, structural variations (SVs) in plant and animal genomes, and SVs in both plant and animal genomes. The model details are illustrated in Table 1.

**Table 1. Literature Survey on Previous Models**

Ref	Model/Method/Algorithm	Inference
(Wan et al. 2022)	Third-Generation Sequencing (TGS)	Reduced intrinsic biases and uncertain component alignment.
(Suruliandi et al. 2021)	Single-molecule sequencing and Oxford Nanopore Technology (ONT)	Improved accuracy and continuity.
(Veerasingam & Fredrik 2023)	integration of long-read sequencing, mapping methods	Achieved congruent chromosome-level crop genomic assembly.
(Bhattacharya et al. 2021)	Optical mapping technologies	Efficiently marked lengthy DNA molecules.

Another advancement in mapping technology is the chromosomal conformational capturing sequence, which relies on the natural physical joining of DNA segments (Williamson & Leiros 2020; Xiao & Zhou 2020). Combining Hi-C data and visual mapping enables enhanced chromosomal phasing and scaffolding (Camgözlü & Kutlu 2023; Surendar et al. 2024). Hi-C allows for the selective isolation of chimeric DNA ligation junctions followed by deep sequencing by including a biotin-labeled nucleotide at the ligation junction. The ability to identify chromatin connections on an unprecedented scale is made feasible by Hi-C's interoperability with next generation sequencing tools. Xiao et al. (2022) successfully constructed the wheat genome, characterized by high repetition and polyploidy, by integrating brief reads, visual mapping, and chromatin interface mapping information. The assembly achieved an N50 value of 1.9 Mb. TGS offers a greater level of sequencing contiguity, which can enhance genomics-based PB methods such as trait identification.

One of the most effective uses of TGS in PB is to create better and more complete crop genomes (Bohra et al. 2020). When choosing a sequencing method for a project to assemble the genome of crops, it is crucial to consider factors such as genome dimensions, fertility, levels of repeated material, and accessible financial resources. TGS is often not cost-effective for repeated sequencing investigations of crop species (Camgözlü & Kutlu 2023; Dessy et al. 2023). With the decreasing cost of TGS, these approaches expand to a wider spectrum of genomic study beyond genome construction. In investigations requiring barcoding of more than a certain number of samples, TGS may prove to be a more economical option than other approaches. After the sample sizes reached 61 for Flongle, 183 for MinION, and 356 for PacBio, TGS became more cost-effective than Sanger sequencing for barcoding.

### **Genomics and Bioinformatics for Plant Breeding**

The objective of PB is to generate novel plant cultivars. This enduring endeavor commences with fundamental research and frequently spans numerous years, requiring a substantial monetary commitment. Genomics-assisted PB is a highly efficient and cost-effective method that has been extensively utilized. The phenotyping platforms, genetic databases, and high-throughput genotyping technologies are used by GAB to accomplish these aims. Among the many outcomes of GAB are: Tapaswini rice: This variety benefits from a mix of genes that make it resistant to both biotic and abiotic tolerant genes. The Pusa Chickpea 10216 is a hybrid that was created by crossing the Pusa 372 with genes that make the plant more resistant to drought. Genomics aids in comprehending the structure and operation of biological structures. It can monitor the molecular alterations during growth in various circumstances, such as shifts in plant physiological science, pathogen influence, or environmental adjustments. Genomic specimens can be taken from the same or distinct individuals within a species or from various species altogether.

Comparison genomics enables examining specific characteristics in closely related plants by preserving genetic sequences between species possessing smaller, more manageable genomes and those possessing larger and more intricate genomes. They have been employed to investigate genetic variation and detect advantageous alleles for many aesthetic and resistance features, such as plant structure and inflorescence attributes, waterlogging acceptance, aphid opposition, and adaptation to drought. Su et al. (2024) converted a significant Single Nucleotide Polymorphism (SNP) associated with waterlogging resistance in Chrysanthemum (CS) into a Dynamic Cut Amplification Polymorphism Sequences (d-CAPS) marker (Williamson & Leiros 2020; Su et al. 2024). The accuracy of this marker was determined to be 71.3% and was validated in 50 different cultivars or ancestors. With this accuracy rating, the prediction will work better on different

cultivars. Agrawal et al. (2020) proposed two d-CAPS markers linked to the CS blooming stage and head dimensions. The d-CAPS indicators show promise for use in the genetic improvement of CS. These strategies will offer innovative and potent resources for future CS cultivation.

ML enables programs to analyze data by acquiring trends through iterative processes. ML offers significant benefits over other analytical methods when dealing with extensive, varied, and unstructured information, such as those produced by optical imaging or sequence. ML enables PB to analyze PP and extract valuable insights from various datasets, including identifying correlations between DNA sequencing and characteristics. Plant breeders may benefit significantly from machine learning (ML), a subfield of AI. It is capable of optimising in vitro breeding techniques, predicting crossovers and regions of natural selection, identifying QTLs, assessing genetic architecture, classifying genotypes, modeling attributes, phenotypic prediction, and complicated relationship interpretation. Additionally, ML can improve in vitro breeding procedures, model plant properties, analyze complicated relationships, and categorize genotypes of plants using morphological and genetic markers. Additionally, it may predict crossovers, identify areas of natural selection where mutations have been almost fixed, and differentiate between active genes and pseudogenes, among other genomic regions. Moreover, the epigenome comprising different mechanisms e.g. DNA methylation, remodeling, histone tail modifications, chromatin microRNAs and long non-coding RNAs, interact with environmental factors like nutrition, pathogens, climate to influence the expression profile of genes and the emergence of specific phenotypes (Barazandeh et al., 2016). Multi-level interactions between the genome, epigenome and environmental factors might occur (Amiri Roudbar et al., 2020). Furthermore, numerous lines of evidence suggest the influence of epigenome variation on health and production (Alavi et al., 2022). The expression of eukaryotic genes is temporarily and multidimensionally controlled. Only a relatively small set of the entire genome is expressed in each type of tissue, and the expression of genes depends on the stage of development (Heidarpour et al. 2011; Khabiri et al. 2023). Therefore, gene expression in eukaryotes is specific to each tissue (Safaei et al. 2022). Also, the amount of gene products that are made in the same tissue as well as in other tissues that make up that product, regulates the expression of that gene. One of the basic activities is the study of genes and proteins related to traits and their study at the cellular or chromosomal level (Bordbar et al., 2022).

**Plant phenotyping (PP):** PP involves quantifying plants' functional or structural characteristics, ranging from within cells to the organism scale. This process is complicated for conducting research on associations and enhancing crop development. Due to the rapid advancement of genomics studies and sequencing methods, there is a growing need for plant characteristics to aid in comprehending genetic data. Traditional PP methods are frequently a

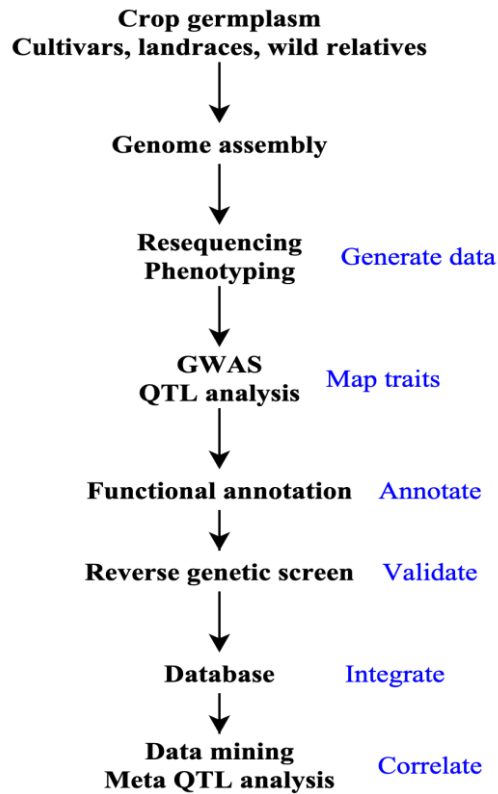
hindrance due to their subjective nature, susceptibility to errors, the requirement for significant labor and time, and their limitation on the number of characteristics, plants, and conditions assessed. The development of high-throughput photography and automatic detectors, along with advancements in ML, enables the implementation of PP. This approach overcomes the limitations of conventional human-based phenotyping by facilitating the quick creation of phenotypical characteristics and traits across huge populations. High-throughput phenotyping consists of four primary components: image or sensor-based identification, categorizing phenotypical information, quantifying features, and forecasting using specialized models or methods. Despite the practical application of ML in agricultural genomics and PP, some difficulties still need to be addressed. ML modeling necessitates extensive datasets to use in training and constructing the model. A limited training set might result in scientifically inconsequential and erroneous predictions. Obtaining an extensive dataset can be expensive and time-consuming, especially for crop metrics, which can only be conducted once each growth cycle. ML-based phenotyping is restricted to specific universities and businesses. To make ML-based phenotyping possible on future farms, decreasing the expenses associated with acquiring and operating is necessary. By analyzing plant data, machine learning (ML) may help with crop management and plant breeding. Plant phenotyping, which involves evaluating characteristics including growth, development, and resistance, may be improved by ML. Machine learning has several applications in plant phenotyping, such as:

- The application of ML allows for the development of automated phenotyping systems, which are capable of collecting and analyzing massive amounts of plant data. These systems can examine plant data in controlled conditions or in the field using robots, sensors, and imaging technologies.
- Machine learning (ML) is a powerful tool for studying gene activity and its impact on plant phenotypes in response to environmental changes.
- When it comes to developing plants to be more stress-tolerant, ML may be a huge assistance.
- Management of crops: ML is useful for managing crops since it can forecast yields, estimate biomass, and track growth phases.
- Plant health: ML can detect nutritional status and water stress, which helps in assessing plant health.
- Early disease identification in plants is possible with the use of ML.

**ML in plant genomics:** ML is utilized in various genomics studies, including genome assembling, the iterative deduction of gene regulatory systems, and identifying authentic Single



Nucleotide Polymorphisms (SNPs) in polyploid crops. Torniainen et al. (2020) provide a comprehensive list of ML techniques and corresponding open-source R modules applicable for analyzing plant science data (Torniainen et al. 2020). ML can enhance the process of assembling polyploid genomes with intricate redundancy in their genetic structure. The method of fully assembling and annotating a genome is essential for monitoring variation in genetics in an animal species and comprehending the purpose and arrangement of plant genes. This is important in crop trait development (Figure 1). Non-ML-based assembling techniques that utilize a linear method to assemble repeated sequence areas need help dealing with highly duplicated genomes. An ML technique was employed to address this constraint, identify flaws in the assembly process, and provide a superior assembly method for wheat. Portcullis's RNA-sequence mapping program utilizes ML to distinguish between genuine and synthetic splicing sites. This application of ML has been essential in annotating the bread grain genomes.



**Figure 1. Steps in genomics-assisted plant breeding**

SNPs are the most common type of genetic differences found in plant genomes. There are still obstacles to overcome when it comes to identifying SNPs in plants with several sets of chromosomes. Barbosa et al. (2021) created an ML analysis program that utilizes neural networks

and tree-bagging methods to locate and remove inaccurate positive SNPs (Barbosa et al. 2021; Su et al. 2024). Their demonstration showed that the technique can detect SNP variability and accurately select real SNPs with a precision of over 98% in generated SNP variation information from peanuts and strawberries. Neuronal networks have been used to identify SNPs in error-prone lengthy readings. This task is challenging in specimens with two or more sets of chromosomes because errors must be differentiated from actual heterozygous variations. Among the many possible future uses of machine learning in genomics are the creation of more precise pharmaceutical drugs customized to individual patients or populations, the enhancement of soil quality and agricultural yields by farmers, and the advancement of genetic screening tools for kids.

### **Bioinformatics tools**

**Biological repository:** Biological repositories are repositories of consistent information that are stored standardized and efficiently. These databases encompass a wide range of biological disciplines. A simple database could consist of a single file with several records, including the same information set. Repositories consist of hardware and software components used to manage data. The primary goal of a registry is to organize data into a collection of structured entries to facilitate efficient data access. Every record, known as an entry, should contain a range of fields that store specific data elements. To get an individual entry from the database, the consumer can specify a specific value to be searched for within a particular area and anticipate the computer to return the complete data item. This technique is referred to as formulating a query. There are two types of databases: the primary and secondary databases. The key distinction between main and secondary databases is that the primary one stores the initial data while the second analyses and expands upon it.

Primarily, primary databases are used to store first-hand information, such as findings, concepts, and theories. Publicly accessible databases, often known as primary databases, save information for future use. The European Molecular Biology Laboratory, GenBank, and the DNA Database of Japan are all examples of main databases. Data from main databases is used to construct additional layers of information in secondary databases. Due to their employment of both computer methods and human examination, secondary databases are often referred to as curated databases. Multiple resources, such as other databases, scholarly articles, and restricted vocabularies, may be mined for data via secondary databases.

**Software and tools:** For finding and grouping collections of nucleotide and protein sequences, two programs are available: DIAMOND and MMseqs2. The MMseqs2 suite of

programs is multi-purpose, functioning as both a desktop program and a local web server. The purpose of DIAMOND is to analyze massive volumes of sequencing data. Bioinformatics employs a variety of software, ranging from essential instructional tools to more intricate graphical applications and independent internet services offered by bioinformatics corporations or public organizations. The Basic Local Alignment Sequencing Tool (BLAST) is widely used for sequencing identification. Three phases are used by BLAST. The process begins by "chopping" the DNA or protein sequence into smaller "words" of 10–12 nucleotides or 3–4 amino acids. Secondly, it searches the whole database for exact matches using these short terms. Third, when a match is detected, it attempts to further align the letters by comparing them in sequence. It checks for a satisfactory match with every new set of letters. A higher score is given for a good match and a lower one for an inadequate one. A precomputed score table for each pair of nucleotides or amino acids is included into the BLAST algorithm. The method still holds the record as the most efficient way to identify particular sequences in extensive datasets and allows for quickly labeling new sequences. BLAST is widely recognized as the primary tool for detecting sequence parallels in extensive records. There are other alternatives for assembling sequencing records, with the choice depending on factors such as hardware accessibility, database magnitude, format, organization, and the genetic makeup of the species.

**Gene silencing methods:** While insertion ablation is an effective technique for creating loss of function mutations, it does have limits when it comes to redundant genes and fatal mutants. To overcome these restrictions, techniques for disrupting biological processes have been devised and utilized for the functional study of plant genes. RNA interference is a widely used technique for silencing genetics through the targeted destruction of RNA molecules with a sequence matching the introduced double-stranded RNA. This process is known as post-transcriptional gene suppression. In recent years, selective RNA interference systems have been created to meet the need for targeted silencing. These systems utilize either a chemical-inducible recombine mechanism or an inducer derived from heat shock-inducible genetics. A genetic approach called CRISPR interference (CRISPRi) may be used to silence genes in cells. It does this by directing a repressor to the transcription start site of a specific gene. A single guide RNA (sgRNA) and a catalytically inactive Cas9 protein (dCas9) are used in the CRISPRi method. When the Cas9-sgRNA complex attaches to certain DNA sequences, it prevents RNA polymerase from extending transcripts.

Research into gene function, genetic connections, and loss-of-function experiments may all be facilitated using CRISPRi. One research, for example, employed CRISPRi to make mice develop osteopetrosis and defective lymph nodes by simulating gene loss.

Depending on the cell type and the growing circumstances, CRISPRi may or may not be effective. It is possible to quantify CRISPRi's impact using assays like as flow cytometry, qRT-PCR, or genome-wide NET-seq.

### Plant database

**Plant comparative genomics and repositories:** With the recent addition of nucleotide sequences for farming species, such as crops and cattle, the research can now compare genomes between different organisms. This enables the research to identify the critical genes responsible for specific physical traits. Integrating genomic material obtained from other closely related species, such as extensive collections of DNAs and information from whole-genome sequencing initiatives, is expected to enhance information regarding gene function exchange across biological models and practical applications. This could expedite the process of understanding the molecular composition of cellular systems that are linked to significant characteristics. Many internet databases for plant genetics and suitable analytical tools have emerged.

**Portal data resources in plants:** TAIR is a prevalent and comprehensive data resource in plant science. It serves as a crucial portal for Arabidopsis study (Xiao & Zhou 2020). The Salk Institute Genetic Analysis Laboratories is a valuable information repository that consolidates many data sets of significant omics findings primarily related to Arabidopsis. The Arabidopsis Genome Dictionary details several genomic assets developed for the Arabidopsis study (Mikheenko et al. 2022). These portal sites provide portals for extensive information about omics and bio services. These websites contain interconnected data collections that link each identified gene with its corresponding elements, such as full-length DNA replicas, gene mutations, gene expression trends, and genome homologous sequences. Internet sites employ genome websites such as G-Explore to view a gene with annotations in conjunction with its genome sequence and related data (Williamson & Leiros 2020). Gramene is a widely used portal that serves as both a comprehensive source of data on rice and a platform for advancing the study of plant genome comparison. Gramene provides comprehensive genome-focused data, encompassing gene commentary, genetic markers, and a mostly Gramineae species-oriented database (Su et al. 2024). With the advent of DNA sequencing initiatives, there has been a rise in the development of portal sites that facilitate information sharing and integrating related materials for different species. The genome network is a gateway for Solanaceae genomic assets, including essential data on the plant's genomic sequence research (Agrawal et al. 2020).

The field of agriculture encounters significant obstacles in utilizing the large amount of genetic data, which comes from many sources and is presented in different formats, to enhance

crop development. To overcome these issues, it is necessary to utilize innovative PB techniques and bioinformatics technologies to convert genomic information into improvements in crop productivity and safety. Bioinformatics is currently exerting a significant influence on the advancement of the farming industry, agro-based enterprises, the usage of agricultural by-products, and the improved management of the natural world. Genomics, which involves sequencing the genomes of hypothetical plants and plant pathogens, has advanced quickly and created numerous possibilities for enhancing the genetics of agricultural plants. Scholars might employ Meta qualitative trait loci analysis, genomic broad associations, and genetic screening to expedite the identification of robust gene trait connections. Genome editing is a valuable method for quickly introducing advantageous mutations into high-quality plant varieties. Genomic selection improves the effectiveness of selection without the need to understand the specific genetic factors involved.

**Conclusions:** The field of agriculture encounters significant obstacles in effectively utilizing the large volume of genetic data, which comes from many sources and is presented in different formats, to enhance crop development. To address these difficulties, it is necessary to employ innovative PB techniques and bioinformatics technologies to effectively convert genomic data into improved agricultural productivity and stability. Scientists can employ meta-studies and genomic screens to expedite the identification of solid connections between genes and traits. Genome editing is valuable for quickly introducing advantageous mutations into high-quality crops. However, genomic selection improves the effectiveness of selection without the need to understand the specific genetic factors involved. ML algorithms can utilize high-throughput characterization and genomic information to enhance the automation of certain aspects of the gene identification process, such as DNA annotation and picture interpretation, which are known to be particularly difficult. By combining innovative technology and approaches, emerging PB can attain the necessary rate of plant enhancement to guarantee adequate nutrition. Addressing many species concurrently may be in the future of ML. ML approaches have a wide range of potential applications in fields including comparative genomics and information transfers from model plants to target crops.

**Acknowledgement:** The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

**Conflict of Interest:** There is no conflict of interest.

## References


- Agrawal N, Tripathi R, Jain M (2020) Molecular marker tools for breeding program in crops. *Plant Ecophysiology and Adaptation under Climate Change: Mechanisms and Perspectives II: Mechanisms of Adaptation and Stress Amelioration*, Springer Singapore, pp. 567-582.
- Alavi M, Mozafari MR, Ghaemi S et al. (2022) Interaction of Epigallocatechin Gallate and Quercetin with Spike Glycoprotein (S-Glycoprotein) of SARS-CoV-2: In Silico Study. *Biomedicines* 10(12), e3074.
- Amiri Roudbar M, Mohammadabadi MR, Ayatollahi Mehrgardi A, et al. (2020) Integration of single nucleotide variants and whole-genome DNA methylation profiles for classification of rheumatoid arthritis cases from controls. *Heredity* 124(5), 658-674.
- Barazandeh A, Mohammadabadi MR, Ghaderi-Zefrehei M, Nezamabadi-Pour H (2016) Genome-wide analysis of CpG islands in some livestock genomes and their relationship with genomic features. *Czech J Anim Sci* 61(11), 487-495.
- Barbosa IDP, da Silva MJ, da Costa WG, de Castro Sant'Anna I et al. (2021) Genome-enabled prediction through machine learning methods considering different levels of trait complexity. *Crop Sci* 61(3), 1890-1902.
- Bhattacharya S, Barseghyan H, Délot EC, Vilain E (2021) nanotatoR: a tool for enhanced annotation of genomic structural variants. *BMC Genomics* 22, 1-16.
- Bohra A, Chand Jha U, Godwin ID, Kumar Varshney R (2020) Genomic interventions for sustainable agriculture. *Plant Biotechnol J* 18(12), 2388-2405.
- Bordbar F, Mohammadabadi M, Jensen J et al. (2022) Identification of candidate genes regulating carcass depth and hind leg circumference in simmental beef cattle using Illumina Bovine Beadchip and next-generation sequencing. *Animals* 12(9), e1103.
- Brière G, Darbo É, Thébault P, Uricaru R (2021) Consensus clustering applied to multi-omics disease subtyping. *BMC Bioinformatics* 22, 1-29.
- Camgözlü Y, Kutlu Y (2023) Leaf Image Classification Based on Pre-trained Convolutional Neural Network Models. *Natural and Engineering Sciences* 8(3), 214-232.
- Cao Y, Geddes TA, Yang JYH, Yang P (2020) Ensemble deep learning in bioinformatics. *Nat Mach Intell* 2(9), 500-508.
- Chen C, Wu Y, Li J, Wang X et al. (2023) TBtools-II: A “one for all, all for one” bioinformatics platform for biological big-data mining. *Mol Plant* 16(11), 1733-1742.
- Dessy A, Ratna D, Leni S, Yadi Heryadi D et al. (2023) Using Distance Measure to Perform Optimal Mapping with the K-Medoids Method on Medicinal Plants, Aromatics, and Spices Export. *J Wirel Mob Netw Ubiquitous Comput Dependable Appl* 14(3), 103-111.

- Ghotbaldini H, Mohammadabadi MR, Nezamabadi-pour H et al. (2019) Predicting breeding value of body weight at 6-month age using Artificial Neural Networks in Kermani sheep breed. *Acta Scientiarum Anim Sci* 41, e45282.
- Gladkov EA, Gladkova OV (2021) Plants and Maximum Permissible Concentrations of Heavy Metals in Soil. *Arch Tech Sci* 2(25), 77–82.
- Heidarpour F, Mohammadabadi MR, Zaidul ISM et al. (2011) Use of prebiotics in oral delivery of bioactive compounds: a nanotechnology perspective. *Pharmazie* 66(5), 319-324.
- Kamusoko R, Jingura RM, Parawira W, Chikwambi Z (2021) Strategies for valorization of crop residues into biofuels and other value-added products. *Biofuels Biofpr* 15(6), 1950-1964.
- Khabiri A, Toroghi R, Mohammadabadi M, Tabatabaeizadeh SE (2023) Introduction of a Newcastle disease virus challenge strain (sub-genotype VII. 1.1) isolated in Iran. *Vet Res Forum* 14(4), e221.
- Mikheenko A, Prjibelski AD, Joglekar A, Tilgner HU (2022) Sequencing of individual barcoded cDNAs using Pacific Biosciences and Oxford Nanopore Technologies reveals platform-specific error patterns. *Genome Res* 32(4), 726-737.
- Mohammadabadi M, Kheyrodin H, Afanasenko V et al. (2024) The role of artificial intelligence in genomics. *Agric Biotechnol J* 16 (2), 195-279.
- Pour Hamidi S, Mohammadabadi MR, Asadi Foozi M, Nezamabadi-pour H (2017) Prediction of breeding values for the milk production trait in Iranian Holstein cows applying artificial neural networks. *J Livestock Sci Technol* 5 (2), 53-61.
- Safaei SMH, Dadpasand M, Mohammadabadi M et al. (2022) An *Origanum majorana* Leaf Diet Influences Myogenin Gene Expression, Performance, and Carcass Characteristics in Lambs. *Animals* 13(1), e14.
- Sayers EW, Bolton EE, Brister JR, Canese K et al. (2023) Database resources of the National Center for Biotechnology Information in 2023. *Nucleic Acids Res* 51(D1), D29-D38.
- Srinivasa Rao M, Praveen Kumar S, Srinivasa Rao K (2023) Classification of Medical Plants Based on Hybridization of Machine Learning Algorithms. *Indian J Inform Source Serv* 13(2), 14-21.
- Su J, Li C, Ou X, Wen S et al. (2024) CmWAT6. 1, mined by high-density genetic map-based QTL mapping, enhances waterlogging tolerance in the chrysanthemums. *Environ Exp Bot* 219, e105612.
- Surendar A, Saravanakumar Veerappan, Sindhu S, Arvinth N (2024) A Bibliometric Study of Publication- Citations in a Range of Journal Articles. *Indian J Inform Source Serv* 14(2), 97-103.

- Suruliandi A, Mariammal G, Raja SP (2021) Crop prediction based on soil and environmental characteristics using feature selection techniques. *Math Comp Model Dyn* 27(1), 117-140.
- Swarup S, Cargill EJ, Crosby K, Flagel L et al. (2021) Genetic diversity is indispensable for plant breeding to improve crops. *Crop Sci* 61(2), 839-852.
- Torniainen J, Afara I.O, Prakash M, Sarin JK et al. (2020) Open-source Python module for automated preprocessing of near-infrared spectroscopic data. *Anal Chim Acta* 1108, 1-9.
- Veerasamy K, Fredrik ET (2023) Intelligent Farming based on Uncertainty Expert System with Butterfly Optimization Algorithm for Crop Recommendation. *J Internet Serv Inf Secur* 13(3), 158-169.
- Wan YK, Hendra C, Pratanwanich PN, Göke J (2022) Beyond sequencing: machine learning algorithms extract biology hidden in Nanopore signal data. *Trends Genet* 38(3), 246-257.
- Wang Y, Zhao Y, Bolla A, Wang Y et al. (2021) Nanopore sequencing technology, bioinformatics, and applications. *Nat Biotechnol* 39(11), 1348-1365.
- Williamson A, Leiros HKS (2020) Structural insight into DNA joining: from conserved mechanisms to diverse scaffolds. *Nucleic Acids Res* 48(15), 8225-8242.
- Xiao J, Liu B, Yao Y, Guo Z et al. (2022) Wheat genomic study for genetic improvement of traits in China. *Sci China Life Sci* 65(9), 1718-1775.
- Xiao T, Zhou W (2020) The third generation sequencing: the advanced approach to genetic diseases. *Transl Pediatr* 9(2), e163.




## یک روش ترکیبی از ژنومیک و بیوانفورماتیک برای برنامه‌های بهنژادی گیاهی

دبرگیا بیسواس 

\*نویسنده مسئول: استادیار، گروه علوم کامپیوتر و فناوری اطلاعات، دانشگاه کالینگا، رایپور، هند. آدرس پست الکترونیکی:

ku.debaghyabiswas@kalingauniversity.ac.in

پوجا شارما 

پژوهشگر، گروه علوم کامپیوتر و فناوری اطلاعات، دانشگاه کالینگا، رایپور، هند. آدرس پست الکترونیکی:

pooja.sharma@kalingauniversity.ac.in

تاریخ دریافت: ۱۴۰۳/۰۵/۱۰ تاریخ دریافت فایل اصلاح شده نهایی: ۱۴۰۳/۰۷/۰۴ تاریخ پذیرش: ۱۴۰۳/۰۷/۰۵

### چکیده

**هدف:** با توجه به رشد سریع جمعیت جهانی بشر، تقویت بهره‌وری کشاورزی برای تحقق تقاضای روزافزون برای محصولات زراعی ضروری است. تقویت محصولات زراعی از طریق بهنژادی گیاهان (PB) روشی پایدار برای تقویت کمیت و قوام بازده بدون افزایش نیاز به کود و سموم دفع آفات است. علاوه بر این، تولید داده‌ها در کشاورزی و بیوتکنولوژی در سال‌های اخیر به دلیل توسعه بسیار سریع فناوری‌های با کارایی بالا بسیار افزایش یافته است. مشابه پیشرفت‌های ژنومیک، پیشرفت‌های دلگرم‌کننده‌ای در فناوری‌های فنوتیپ‌سازی گیاهی، از جمله دستگاه فنوتیپ مکانیزه و تجزیه و تحلیل تصویر پیچیده وجود دارد. این امر منجر به درک بی نظیری از PB، ساختار و عملکرد شده است. لذا، هدف این مطالعه، بررسی یک روش ترکیبی از ژنومیک و بیوانفورماتیک برای برنامه‌های بهنژادی گیاهی بود.

**نتایج:** آخرین پیشرفت‌ها در ژنومیک و بیوانفورماتیک امکان تسریع در تقویت محصول را فراهم می‌کند. تکنیک‌های توالی‌یابی نسل سوم (TGS) در حل مشکلات در مونتاژ ژنوم گیاهی ناشی از پلی‌پلوئیدی و وجود مناطق تکراری کمک می‌کنند. در دسترس بودن در حال رشد ژنوم‌های مرجع محصولات برتر، به طور قابل توجهی در تجزیه و تحلیل تغییرات ژنتیکی و شناسایی اهداف PB خاص در ژنوم کمک می‌کند. یادگیری ماشین (ML) با کمک در حاشیه نویسی ژنوم و امکان سنجش کارآمد متغیرهای زراعی در محیط‌های کنترل شده و طبیعی، به شناسایی مناطق ژنومی با اهمیت کشاورزی کمک می‌کند.

**نتیجه گیری:** مجموعه داده‌های زراعی که مقدار فزاینده داده‌های ژنوتیپی و فنوتیپی را ترکیب می‌کنند، ابزاری ارزشمند برای پرورش دهندگان و فرصتی برای روش‌های داده کاوی برای کشف ژن‌های جدید نامزد مربوط به صفات ارائه می‌دهند. همچنین با درک روزافزون ژنتیک کشاورزی، تکنیک‌های انتخاب ژنومی و مهندسی ژنوم پتانسیل تولید محصولات زراعی مقاوم در برابر بیماری‌ها را فراهم می‌کنند و این محصولات علاوه بر تولید بالا، با استرس نیز سازگار هستند.

**واژه‌های کلیدی:** بهنژادی گیاهی، بیوانفورماتیک، ژنومیک، محصولات زراعی

نوع مقاله: مروری.

**استناد:** دبرغیا بیسواس، پوجا شارما (۱۴۰۳) یک روش ترکیبی از ژنومیک و بیوانفورماتیک برای برنامه‌های بهنژادی گیاهی. مجله

بیوتکنولوژی کشاورزی، ۱۶(۳)، ۲۱۱-۲۲۸.

Publisher: Faculty of Agriculture and Technology Institute of Plant

Production, Shahid Bahonar University of Kerman-Iranian

Biotechnology Society.



© the authors