Indian J. Plant Genet. Resour. 35(3): 200–203 (2022)
DOI 10.5958/0976-1926.2022.00069.9

Role of Bioinformatics in the Development of Plant Genetic Resources

Tanwy Dasmandal, Dwijesh Chandra Mishra* and Anil Rai
ICAR-Indian Agricultural Statistics Research Institute, Library Avenue, Pusa Campus, New Delhi-110012, India

Bioinformatics plays a significant role in the development of many fields of biological science and plant genetic resources (PGR) is one of them. With the advent of high throughput sequencing technology, bioinformatics continues to make considerable progress in biology by providing scientists with access to the genomic information. There are many areas of plant genetic resources such as development of core set, trait associated gene discovery, genetic diversity analysis, Genome Wide Association Studies (GWAS), phylogenetic and evolutionary analysis, database development and its management etc. where bioinformatics plays important roles. Main role of bioinformatics is to provide computational algorithm and software tools to accelerate the research of PGR. Bioinformatics is a new paradigm in the genomic research, which provides PGR research a great thrust. However, there are many areas such as pan genomics, multi locus GWAS, Genomic Selection with epistasis effects where bioinformatics can play better role.

Key Words: Bioinformatics, High-throughput Technology, PGR, GWAS, Core Set, Phylogenetic Analysis

Introduction

Plant genetic resources can be defined as all materials that are available for improvement of a cultivated plant species (Becker, 1993). A more effective use of plant genetic diversity is required to address the issues of development, food security, and poverty alleviation, according to the FAO's Global Plan of Action for the Conservation and Sustainable Utilization of Plant Genetic Resources (FAO, 1996b). For this purpose, extensive *ex situ* and *in situ* PGR conservation must be ensured, as well as evaluation of conserved accessions and their use by plant breeders needs to be facilitated. Through genetic advancement and the promotion of less common, neglected, or underutilized crop species, the goal should be to promote interspecific variety in agriculture as well as intraspecific variation within a crop. Modern technologies have paved the path towards fulfilment of these aims through the advanced sequencing technologies and the efficient amalgamation of information technologies with biological science or 'Bioinformatics'. With the generation of huge amount of biological data, bioinformatics has become almost an indispensable part for imparting meaning to the raw data through their analysis and visualization and also for easy maintenance and retrieval of the biological data. In the field of plant genetic resources, bioinformatics has found its role in several areas like in developing germplasm core collection, gene discovery, genomic characterization, database creation and management and so on.

Role in Developing Germplasm Core Collection

Capturing haplotypes: Single nucleotide polymorphism (SNP) data sets from across the genome have enormous potential to enhance *ex situ* conservation. However, there are two issues that have been noticed regarding their use in the production of core collections. Firstly, due to the huge number of SNPs, it may be challenging to assemble the collections that will maximize diversity. To address this issue bioinformatics plays its role in developing computer program (like M+) for identifying optimized core collections from arbitrarily large genotypic data sets. MSTRAT is a popular program for producing genetically variable core collections.

Secondly, it is uncertain how the genome's natural partitioning into linked regions or haplotype blocks would affect the diversification and collection optimization. Large samples are necessary to determine haplotype block structure, and the process is methodologically complicated. However, bioinformatics can be used for simulating the basic structure of haplotype blocks using program like HAPLOTYPISTA.

**Author for Correspondence: Email-dwij.mishra@gmail.com*

Genomics-based gene discovery: One of the main goals of germplasm research is gene discovery. In germplasm research using genomics, there are four approaches to finding new genes: map-based, association-based, allele mining-based, and comparative genomics-based. Among these methods bioinformatics has been extensively used for association based gene discovery using germplasm collection through GWAS analysis, and for comparative genomics based gene discovery using orthologous gene strategy where gene function and sequence have already been determined in a model (or other) species.

Graphical Tools for Germplasm Analysis

Bioinformatics has also been extensively used in developing graphical tools for germplasm analysis like GENE-MINE software. These tools have found applications in various fields of PGR studies such as-

a. a geographical tool that could show the origin of accessions and the distribution of genetic diversity;
b. a haplotype tool showing genotype information for accessions;
c. tool for generating graphs that might display phylogenetic trees or networks (graphs containing closed loops, which can be used to represent genetic exchange between organisms) as well as pedigrees;
d. tool that produces scatter plots of genetic marker distribution on relevant linkage maps, such as principal component analysis and diversity distances between pairs of accessions

Extraction of Functional Genetic Diversity from Heterogeneous Germplasm Collections for Crop Improvement

Large collection sizes, uneven trait characterization, and unpredictable distribution of allelic diversity across diverse accessions restrict efficient use of genetic variation in plant germplasm collections. Conventional and precision breeding might be streamlined by distributing compact subsets of the whole collection that include the largest amount of allelic variation at functional loci of interest.

In general, three bioinformatic approaches come into play to extract functional genetic diversity.

First, in a "candidate gene" approach, subsets that maximized haplotypic diversity are assembled.

Secondly, to find regulatory loci and assembled subsets representing genome-wide regulatory gene diversity, a general source of phenotypic variation, one can do a keyword search against the Gene Ontology.

Thirdly, machine-learning approach can be developed to rank semantic similarity between Gene Ontology term definitions and the textual content of scientific publications on crop adaptation to stress, a complex breeding objective.

Role in Germplasm Characterization

Trait Mapping using GWAS

For germplasm managers, fundamental researchers, and plant breeders, the capacity to precisely detect and analyze genome-wide genetic variation or specific molecular variations through generations of individuals offers a potent tool. Thanks to the development of NGS, GWAS is presently a useful method to investigate allelic variation in a wider context for extensive phenotypic diversity and improved resolution of QTL mapping. Using GWAS, many research projects have been done to investigate the association between genetic variation and valuable plant traits. At the current stage, several bioinformatics approaches have been introduced as GWAS acceleration tools. Following are some examples: Heap which is a SNPs detection tool for NGS data with special reference to GWAS and BioGPU, a high-performance computing tool for GWAS. But, focusing on one main SNP that correlates with a specific phenotype as normal GWAS output may miss the key genetic variants with particular environment response in the context of complex traits. For this issue, bioinformatics approach is again a current solution. Generalize multifactor dimensionality reduction (GMDR) algorithm on a computing system with graphics processing units (GPUs) is one in some available methods at the moment that can screen potential candidate variants and then use the mixed liner model to detect the epistatic and gene-environment interactions.

Characterization on Climate Change Adaptation

Studying the transcriptome of populations using bioinformatic approaches growing along an environmental gradient may also reveal changes in gene expression of the same set of genes, thus potentially shedding light on the genetics of adaptation. Genomic information emanating from studies on climate change adaptation will support decision-making on what genetic resources to conserve in a gene bank in future. Genomics will also facilitate identification of novel alleles emerging because of climate change adaptation. These novel alleles, which

give plants the adaptive capacity, should be prioritized for conservation, as they are important in developing climate resilient crops.

Phylogenomics and Evolutionary Analysis

There are two important goals in phylogenomic research to accomplish. First is to discover the evolutionary patterns among plant species using nuclear genomic information. Second is to derive new hypothesis for the unknown function of plant genes associated to major divergence events in the evolution of plant species [95]. Bioinformatics have been extensively explored to develop methods and tools for performing plant phylogenomics like ExaML (Exascale Maximum Likelihood), MEGA, PHYLIP and so on.

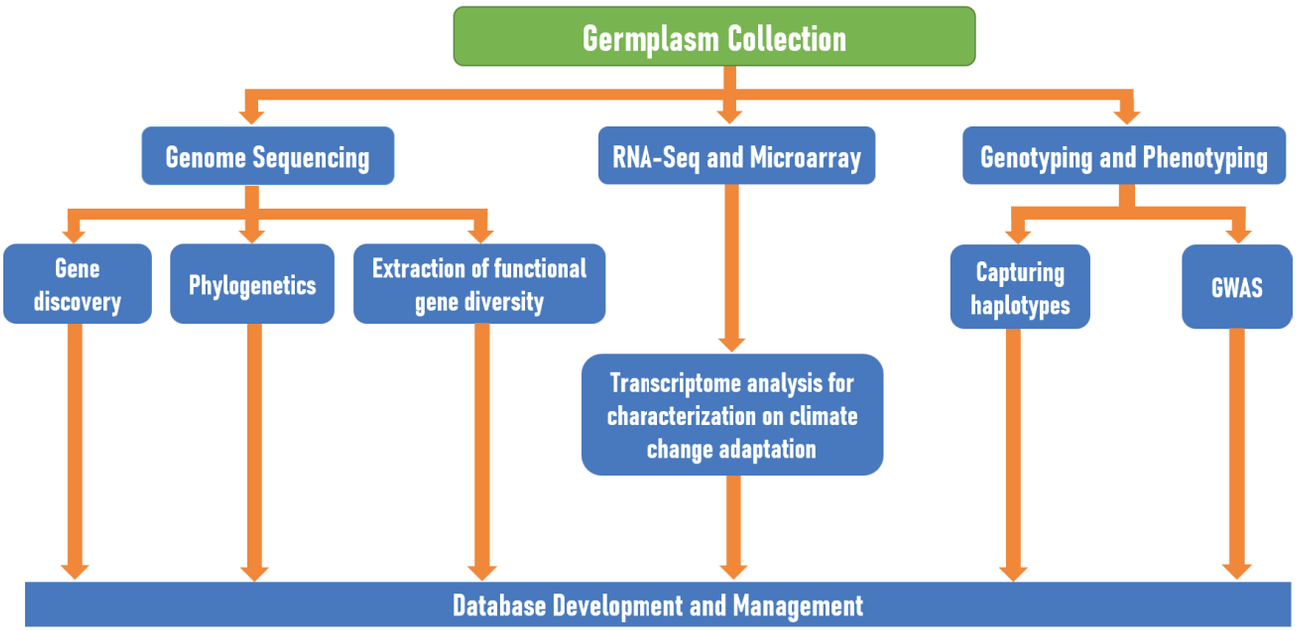
Role in Database Development and Management

To efficiently store and manage data, the use of state-of-the-art database management back-end infrastructure is indispensable. In this context well aligned binary data storage, Relational Database Management System (RDBMS), homogeneous designed data schemata, customized data import and user interfaces must be combined into an all-in-one back-end. Besides these back-end services, efficient data-sharing and data-access is a further challenging task for scientific data engineering. Data re-use and exchange is a key component to gain scientific knowledge in natural sciences [9] and particularly in plant science.

As for example, there are several database systems designed to manage this information for crop improvement; such as the integrated breeding platform (https://www.integratedbreeding.net/) and the Triticeae toolbox (https://triticeaetoolbox.org). GnpIS-Asso is another such example which is a generic database for managing and exploiting plant genetic association studies.

Future Prospects

- Pangenomics is the new age concept which needs to be explored for germplasm characterization. The pangenome refers to a collection of genomic sequence found in the entire species or population rather than in a single individual. So, instead of sequencing and analysing some selected germplasm, every germplasm in the collection is needed to be analysed and explored.
- Genomic selection is an area of bioinformatics which has been merely exploited in crop science. Genomic selection is an advance form of marker assisted selection which has the potential to reduce the breeding cycle and thus increasing the genetic gain in crops. This technique can extensively be used for crop improvement programs.
- Till date GWAS is limited to single locus study which can explain only a little part of the genetic variance associated with the phenotype. Thus, GWAS studies needs to be extended to multi locus studies which



Role of Bioinformatics in PGR Studies

can capture majority of the genetic variances and can also detect the epistatic interactions as well.

- With the changing time, the demand of agriculture is constantly changing. Unlike previous demands of increasing yields and other quantitative traits in crops, recent era focuses more on qualitative enrichments of crops like nutrient content, aroma etc. And to keep pace with these changing demands bioinformatics has to be accepted as an indispensable part of crop improvement. Bioinformatics analysis of qualitative traits need to be carried out on a large scale on all major as well as minor crop species.
- Role of Bioinformatics in PGR has been largely applied in areas of genomics so far with very little attention paid to other areas of omics like transcriptomics, proteomics and metabolomics which can give better insights of crop genetics and breeding.

References

Jia, J., Li, H., Zhang, X., Li, Z., & Qiu, L. (2017). Genomics-based plant germplasm research (GPGR). *Crop J.* **5(2):** 166-174.

Reeves, P.A., Tetreault, H.M., & Richards, C.M. (2020). Bioinformatic extraction of functional genetic diversity from heterogeneous germplasm collections for crop improvement. *Agronomy*, **10**(4): 593.

Davenport, G., Ellis, N., Ambrose, M., & Dicks, J. (2004). Using bioinformatics to analyse germplasm collections. *Euphytica*, **137**(1): 39-54.

Reeves, P.A., & Richards, C.M. (2017). Capturing haplotypes in germplasm core collections using bioinformatics. *Genet. Resour. Crop Evol*, **64**(8): 1821-1828.

Ong, Q., Nguyen, P., Phuong Thao, N., & Le, L. (2016). Bioinformatics approach in plant genomic research. *Current genomics*, **17**(4): 368-378.