Sustainable Agriculture and Food Security
Series Editor: Rajeev K. Varshney

Manish K. Pandey · Alison Bentley
Haile Desmae · Manish Roorkiwal
Rajeev K. Varshney *Editors*

Frontier Technologies for Crop Improvement

Springer

Chapter 3
Bioinformatics for Plant Genetics and Breeding Research

Yogesh Dashrath Naik, Chuanzhi Zhao, Sonal Channale, Spurthi N. Nayak, Karma L. Bhutia, Ashish Gautam, Rakesh Kumar, Vidya Niranjan, Trushar M. Shah, Richard Mott, Somashekhar Punnuri, Manish K. Pandey, Xingjun Wang, Rajeev K. Varshney, and Mahendar Thudi

Abstract Global food demand is expected to increase between 55 and 70% by 2050. Plant breeders and geneticists are constantly under pressure to develop high-yielding climate-resilient varieties using novel approaches. The quest for simplifying complex traits and efforts for developing high-yielding varieties during the twenty-first century led to a paradigm shift from phenotypic-based selection to genome-based breeding. On one hand, the development and utilization of diverse genetic

Y. D. Naik · K. L. Bhutia
Dr. Rajendra Prasad Central Agricultural University (RPCAU), Pusa, Bihar, India

C. Zhao · X. Wang
Shandong Academy of Agricultural Sciences (SAAS), Jinan, Shandong, China

S. Channale
University of Southern Queensland (USQ), Toowoomba, Queensland, Australia

S. N. Nayak
University of Agricultural Sciences, Dharwad, Karnataka, India

A. Gautam · R. Kumar
Central University of Karnataka, Kalaburagi, Karnataka, India

V. Niranjan
RV College of Engineering, Bengaluru, Karnataka, India

T. M. Shah
International Institute of Tropical Agriculture (IITA), Nairobi, Kenya

R. Mott
University College London, London, UK

S. Punnuri
College of Agriculture, Family Sciences and Technology, Dr. Fort Valley State University, Fort Valley, Geogia, USA

M. K. Pandey
Center of Excellence in Genomics and Systems Biology (CEGSB), International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India

Shandong Academy of Agricultural Sciences (SAAS), Jinan, Shandong, China

M. K. Pandey et al. (eds.), *Frontier Technologies for Crop Improvement*, Sustainable Agriculture and Food Security, https://doi.org/10.1007/978-981-99-4673-0_3

resources, and advances in genomics on the other hand provided a kick start for the understanding the genetics of economically important complex traits at a faster pace. Further, the next-generation sequencing revolutionized our understanding of the genome architecture. As a result, there has been an increasing demand for statistical and bioinformatics tools to analyse and manage the enormous amount of data generated from sequencing of genomes, transcriptomes, proteome and metabolomes. In this chapter, we review the intervention of bioinformatics and computational tools for deploying the tremendous wealth of data for plant genetics and breeding research.

Keywords Bioinformatics · Next-generation sequencing · Database · Pangenome · Haplotype · Artificial intelligence

3.1 Introduction

Climate change and increasing population growth at an alarming rate poses the biggest challenges to food and nutritional security across the globe. By 2050, the global population is predicted to increase by 55 to 70%, as a result the proportion of people at risk of hunger may increase to around 8% (van Dijk et al. 2021a). With diminishing resources and limited arable land, sustainable production to cater the food and nutritional demands has been a daunting task. Plant breeders and geneticists are constantly under pressure to develop improved crop varieties that are climate-resilient and high-yielding to meet the food and nutritional demands. Low genetic diversity, prolonged breeding cycles, and limited access to high-quality seeds for cultivation have been serious obstacles to achieve greater genetic advancements (Varshney et al. 2020). Although conventional breeding programs contributed to the development of improved varieties, to achieve "zero hunger," the Sustainable Developmental Goal 2 adopted by United Nations Organization advocated the integration of modern breeding approaches in agriculture (Varshney et al. 2018).

Ever since the rediscovery of Mendelian laws, there has been a paradigm shift in understanding the phenotype-based trait genetics to the use of molecular markers, genomics, genomes and sequence-based trait dissection (Varshney et al. 2019; Thudi et al. 2023). During the last two decades, genomics and NGS (next-generation sequencing) technologies have not only revolutionized our understanding of

University of Southern Queensland (USQ), Toowoomba, Queensland, Australia

R. K. Varshney
Murdoch's Centre for Crop and Food Innovation, State Agricultural Biotechnology Centre, Food Futures Institute, Murdoch University, Murdoch, WA, Australia

M. Thudi (✉)
Dr. Rajendra Prasad Central Agricultural University (RPCAU), Pusa, Bihar, India

Shandong Academy of Agricultural Sciences (SAAS), Jinan, Shandong, China

University of Southern Queensland (USQ), Toowoomba, Queensland, Australia

molecular basis of economically important traits, but also increased the rate of adoption of modern breeding approaches to develop climate-resilient crop varieties (Thudi et al. 2020; Varshney et al. 2021a). To date, draft genomes of more than 1000 plants representing 788 species are available in public domain (Sun et al. 2021). Not only draft genomes, gold standard reference genomes to platinum standard reference genomes are available in crops like rice (Zhou et al. 2020) and also in cetacean species (Morin et al. 2020). Efforts are also underway to sequence all the known eukaryotic species through "The Earth BioGenome Project" that provides insights into the biology of life (Lewin et al. 2018). Apart from draft genomes, several germplasm lines including wild species accessions have been sequenced in several crops including pearl millet (Varshney et al. 2017a), chickpea (Thudi et al. 2016; Varshney et al. 2021b), pigeon pea (Varshney et al. 2017b), rice (Wang et al. 2018; Stein et al. 2018). Development of pangenomes and super-pangenomes are underway in many crop species (Khan et al. 2020). With the rapid availability of biological data in public domain, rate-limiting factor in genomics research has shifted from sequencing to computer analysis (Kathiresan et al. 2017). The statistical, bioinformatics tools and algorithms developed earlier are becoming obsolete and computational tools and algorithms that handle "BIG data" are gaining importance (Edwards et al. 2009; Batley and Edwards 2016).

In this chapter, we review the NGS data analysis and available databases that are developed to store and retrieve biological information produced from different omics approaches. In addition, we also discuss the computational tools and approaches that enable development of pangenome, identification of haplotypes and editing genomes. Besides highlighting the challenges, we also highlight the scope of improving the bioinformatics approaches for effective use in crop improvement.

3.2 Understanding Genetic Diversity and Trait Mapping

Genetic diversity plays a major role for gaining greater insights and simplifying complex traits. Prior to advent of molecular markers, the phenotypic plasticity in a crop species was assessed using simple experimental analyses and programmes like XLstat or SPSS (Addinsoft 2021; IBM Corp Ibm 2017). In addition, statistical packages like INDOSTAT is being used to analyse variance, D^2 statistics, canonical roots, path analysis etc. (Khetan and Ameerpet 2015). The statistical tool for agricultural research (STAR) has modules for randomization and layout of crop research experimental designs, data management, and fundamental statistical analysis, including descriptive statistics, hypothesis testing, and ANOVA of designed experiments (Gulles et al. 2014). The stability of a crop over different locations and years is one of the crucial prospects in plant breeding. Software like GGE biplot, GEA-R, STABILITYSOFT, and AMMISOFT are used to analyse Genotype × Environment (G × E) interaction studies (Yan 2001; Pacheco et al. 2015; Pour-Aboughadareh et al. 2019; Gauch and Moran 2019). Stability and performance are examined simultaneously using these tools, allowing for a comprehensive

Table 3.1 List of commonly used software packages for plant breeding

Software/program	Key features	References
XLstat, SPSS	These are used for simple experimental analyses	Addinsoft (2021); IBM Corp Ibm (2017)
R and INDOSTAT	Used for analysis of variance, covariance matrices with ANOVA and ANCOVAS, D^2 statistics with Mahalanobis, stability model analysis, Diallel analysis, Heterosis, line × tester analysis, path analysis, joint scaling test (Cavilli), North Carolina design 1, North Carolina design 3, augmented design, double cross analysis, triple cross analysis and triple test cross	Ledesma (2008); Team (2013); Khetan and Ameerpet 2015
GGE biplot, GEA-R, STABILITYSOFT, and AMMISOFT	Analyses genotype × environment analysis for stability analysis	Yan (2001); Pacheco et al. (2015); Pour-Aboughadareh et al. (2019); Gauch and Moran (2019)
Mapmaker-QTL	It can perform only simple interval mapping	Lincoln et al. (1993)
QTL cartographer	Offers options for carrying out the majority of the documented QTL mapping methods	Basten et al. (2002)
Win-QTL cartographer	It maps quantitative trait loci (QTL) in cross populations from inbred lines	Wang (2005)
PLABQTL	Its primary goal is to identify and describe QTL in populations resulting from a biparental cross by selfing or the creation of doubled haploids. A rapid multiple regression approach achieves simple and composite interval mapping	Utz and Melchinger (1996)
MapQTL	It analyses composite interval mapping, interval mapping, nonparametric mapping, automatic cofactor selection, and permutation test for interval mapping	Van Ooijen and Maliepaard (1999)
STRUCTURE	Used for determining population structure	Pritchard et al. (2000)
TASSEL	It is used for evaluation of trait associations, evolutionary patterns, LD statistics, GLM, MLM, CMLM, P3D: Genomic selection; graphical interphase, PCA, and kinship analysis	Bradbury et al. (2007); Gupta et al. (2015)

understanding of the crop's behavior across different environments and conditions. (Table 3.1).

With the availability of molecular markers, efforts were made to map the genomic regions or genes responsible for the complex traits using both linkage mapping or QTL mapping and linkage disequilibrium-based mapping or association

analysis. The most common software packages used for maapping genomic refions are Mapmaker-QTL, QTL Cartographer, Win-QTL Cartographer, PLABQTL, MapQTL are command-line software (Lincoln et al. 1993; Basten et al. 2002; Wang 2005; Utz and Melchinger 1996; Van Ooijen and Maliepaard 1999; Bradbury et al. 2007; Gupta et al. 2015. Mapmaker-QTL can only perform simple interval mapping (Lincoln et al. 1993). The most versatile QTL mapping software is QTL Cartographer. A range of software tools, including the widely used STRUCTURE, are available for determining population structure (Pritchard et al. 2000). Using this software, you can choose the number of subpopulations by using all marker data or a subset of unlinked markers from the marker collection. Alternatively, using the given marker data, principal component analysis (PCA) can be performed and the first few components used as variables to adjust for population structure. Association analysis can be done with TASSEL. Even without forming a core, one can test a population for its suitability as an association panel. Then it can be directly used for TASSEL analysis. However, some prerequisite analysis is required, like population structure, kinship analysis, and principal component analysis (PCA) (Bradbury et al. 2007). It uses marker data to calculate kinship, which helps to address family relatedness and population structure (Table 3.1) (Gupta et al. 2015).

3.3 Identification and Understanding Key Genes Using Multi-Omics Approaches

Interpretation of molecular complexity and variability at several levels, such as genome, transcriptome, proteome and metabolome, is necessary for comprehensive understanding of organism's entire metabolism. The data from various levels are together referred to as "multi-omics" data. Multi-omics data obtained from various approaches provide insights into the flow of biological information at various levels, can aid in figuring out the biological state of interests underlying mechanisms.

In the last decade, technological advancement in DNA sequencing (Le Nguyen et al. 2019), transcriptomics analysis via RNA-seq (Mashaki et al. 2018), SWATH-based proteomics (Zhu et al. 2020) and metabolomics via UPLC-MS and GC-MS (Balcke et al. 2012) has made a significant contribution in biological data. The first omics field to emerge is genomics that deals with study of complete genomes. Genomic studies like QTL/association mapping has been used to detect genomic regions associated with agronomically important traits (Varshney et al. 2014, 2021b; Bhatta et al. 2019; Thudi et al. 2021; Yoshida et al. 2022) and provide basic framework for other omics approaches. Additionally, differentially expressed genes under several biotic and abiotic stresses were identified using transcriptomics studies in several crop plants (Nayak et al. 2017; Channale et al. 2021; Chen et al. 2022; Pal et al. 2022). Gene expression atlas provides insights into the subsets of genes expressed during different growth stages for pigeon pea (Pazhamala et al. 2017), chickpea (Kudapa et al. 2018), groundnut (Sinha et al. 2020). The spatial transcriptomics method developed by Giacomello et al. (2017) enables

Table 3.2 Summary of widely used databases in plant genetics and breeding research

Databases	Key features	Link
AtMAD	Provide high-quality multi-omics data of *Arabidopsis thaliana*	http://www.megabionet.org/atmad
GoMapMan	Gene functional annotations in the plant sciences	http://www.gomapman.org/
HapRice	SNP-haplotype database for rice	http://qtaro.abr.affrc.go.jp/index.html
NPACT	Plant-derived natural compounds exhibiting anticancerous activity	https://webs.iiitd.edu.in/raghava/npact/faq.html
PGDD	Database for gene and genome duplication in plants	http://chibba.agtec.uga.edu/duplication/
PGDJ	DNA marker and linkage database	http://pgdbj.jp/plantdb/plantdb.html
Phytozome	Plant comparative genomics	https://phytozome-next.jgi.doe.gov/
PIECE	Plant intron exon comparison and evolution database	https://data.nal.usda.gov/dataset/piece-plant-intron-exon-comparison-and-evolution-database
Plant rDNA	Plant rDNA database	https://www.plantrdnadatabase.com/
PlantGDB	Plant genome browsers	https://www.plantgdb.org/prj/GenomeBrowser/
PlantRNA	Database for tRNAs of photosynthetic eukaryotes	http://seve.ibmp.unistra.fr/plantrna/
PlnTFDB	Plant transcription factor prediction	http://planttfdb.gao-lab.org/
PLUTO	Contains information on plant varieties	http://www.upov.int/pluto/en/
PMRD	Plant microRNA database	http://bioinformatics.cau.edu.cn/PMRD/
PTGBase	Plant tandem duplicated genes database	http://ocri-genomics.org/PTGBase/
SALAD	Comparison of proteome data among the species	https://salad.dna.affrc.go.jp/salad/en/

high-throughput and spatially resolved transcriptomics in plant tissues using a combination of histological imaging and RNA sequencing. Functional analysis of translated regions of the genome is understood using proteomics, while metabolomics serves as a diagnostic tool for assessing the plant performance under different stimuli (Villate et al. 2021). A number of repositories were developed to organise data generated from different experiments and sequencing studies. The repositories include DNA, RNA and protein sequence databases, as well as specialized databases for specific information (Lai et al. 2012; Thudi et al. 2020). Based on different types of omics data, databases can be classified into four classes: (1) genomics databases contain nucleotide sequence or genomic sequence, (2) transcriptomics databases include functional RNA sequences, (3) proteomics databases contain information related to amino acid sequence and protein structure, and (4) metabolomics databases contain information about metabolites and metabolic pathways (Table 3.2, Fig. 3.1a).

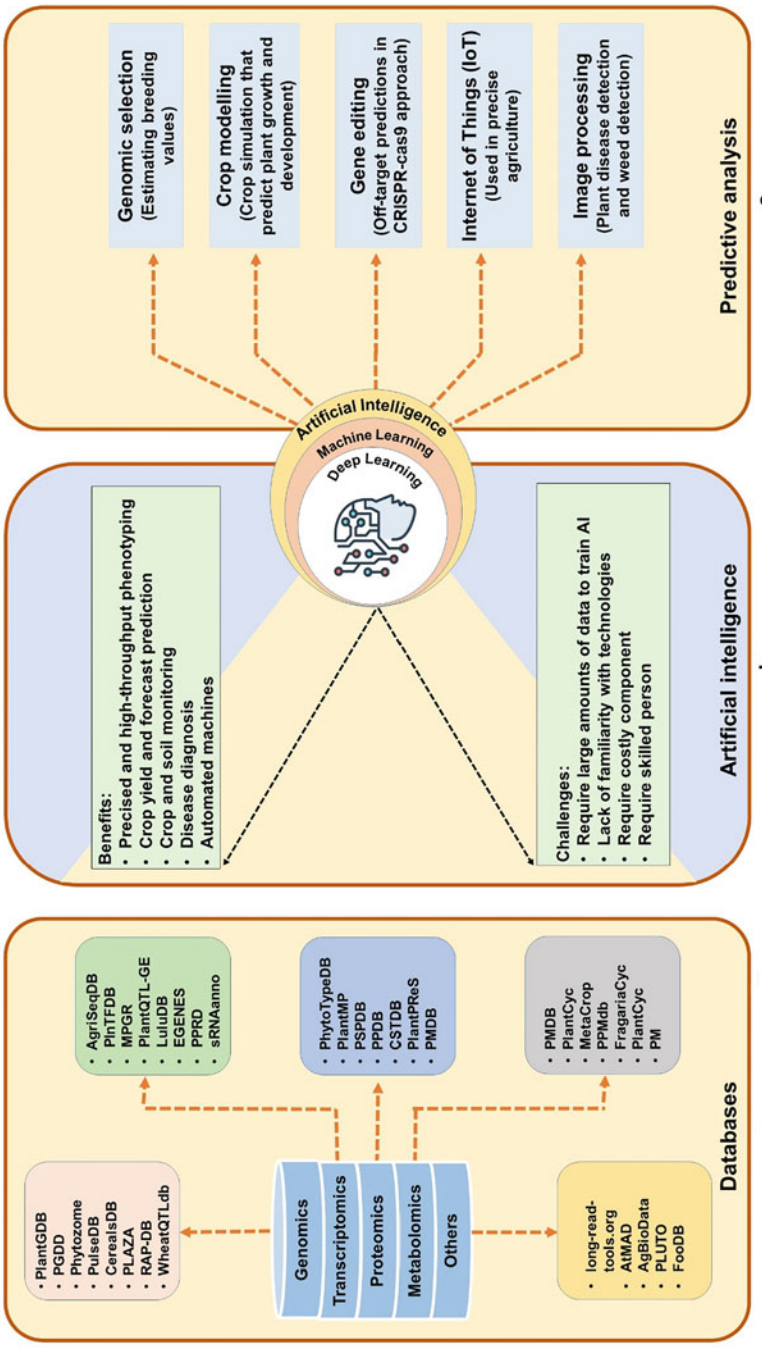

Fig. 3.1 Summary of databases and applications of artificial intelligence in agriculture: (**a**) Represent databases developed to store and retrieve the biological information produced from various omics approaches, includes Genomics, Transcriptomics, Proteomics and Metabolomics, (**b**, **c**) Showed different kinds of predictive analysis based on AI

Databases include PlnTFDB (planttfdb.gao-lab.org/) for plant transcription factor, widely used for expression analysis or functional genomics. This database allows user to get sequence information of known plant transcription factors. Phytozome (phytozome-next.jgi.doe.gov/) database provides access to the selected plant genome sequences and improved platform for comparative analysis of genomes. Breeders have access to useful tools like molecular markers that can speed crop improvement program. In case of chickpea, "CicArVarDB" database provides information of single nucleotide polymorphisms (SNP) and insertion/deletion (Indel) variations which can be utilized for advanced genetics research (Doddamani et al. 2015). Additionally, AgBioData consortium (Harper et al. 2018) works together across different agricultural-related databases to identify approaches for integrating and standardizing database operations. This collaborative effort aims to develop database products that exhibit more interoperability. The major challenge is to manage and translate the sequence information for the crop improvement.

3.4 Evolution of Sequencing Technologies and Tools

About 25 years after discovering the double helical structure of DNA, the first-generation sequencing technologies like Sanger sequencing and Maxam and Gilbert sequencing were available for sequencing both smaller and large genomes. Nevertheless, a plethora of sequencing technologies have evolved during last 15 years and there is an increased data output, read lengths, efficiencies, and applications. Second-generation sequencing technologies had improvement in sequencing throughput, required time and read length with low cost. Short-read sequencing technologies (up to 600 bp) have been widely used in genomics research as it supports wide range of statistical analysis using cost-effective pipelines (Heather and Chain 2016). However, sequencing of short reads created complications in reconstruction of larger fragment or original molecules due to the presence of homopolymers. Long-read sequencing (up to 10 kb) is a highly accurate approach that can be used to sequence traditionally challenging genomes and facilitate de novo assembly, also help in the transcript isoform identification and structural variant identifications. It helps to construct better pangenome than short-read sequencing. In case of rice, third-generation sequencing with long reads were used to construct pangenome using 105 accessions and found 604 Mb novel sequences which was not present in reference genome (Zhang et al. 2022). Specialised analytical tools that consider the properties of long-read data are needed, but the speed at which these tools are being developed can be daunting. Currently, more than 350 long-read analysis tools are available that are generally utilized in Nanopore and SMART sequencing platform (Amarasinghe et al. 2020). For choosing appropriate tool, there is a publicly available database named as "long-read-tools.org," which has a collection of long-read analysis tools and allows us to choose appropriate tools for analysis (Amarasinghe et al. 2021). In order to analyse and interpret the NGS data, there is

a need of highly qualified and competent bioinformaticians. For accurate downstream analysis of sequencing data, appropriate analysis tools are essential and it involves conversion of raw signal data to sequence data.

Sequencing data analysis includes raw read quality control, sequence alignment, variant calling, genome assembly, genome annotation and other advanced analysis. Numerous bioinformatics tools have been developed and used in sequence analysis (Table 3.3). It is essential to evaluate the raw sequence data to ensure the quality for any subsequent analysis. It can give a broad overview of read counts and lengths, coverage reads, contaminating sequences and sequence duplication level. In the first stage, adapter sequences and low-quality sequences are separated from whole genome sequencing data through a quality assessment process. FastQC is the well-known bioinformatics tool for calculating quality control of sequencing reads (Andrews 2010). More recently, fastp tool is also utilized in quality control, base correction and filtering of sequencing reads. The fastp tool is two to five times faster than previous approach (Chen et al. 2018) and ensures the read quality as well as adapter trimming.

The second step is to align the sequences with reference genome, that is, read/sequence alignment. In the case of non-availability of reference genome, de novo genome assembly method is used to generate the contigs by aligning the overlapping regions together. This step is the most crucial and important in the entire workflow. The sequence reads are precisely and quickly aligned to the appropriate places of the reference genome using a variety of tools and algorithms. Many tools have been developed for sequence alignment; the popular aligners include BWA (Li and Durbin 2009), Bowtie2 (Langmead and Salzberg 2012), CUSHAW3 (Liu et al. 2014), MOSAIK (Lee et al. 2014), and Novoalign (http://novocraft.com/). MOSAIK is the mapping tool currently available that can align reads produced by all the major sequencing technologies. Minimap2 is a flexible pairwise nucleotide sequence aligner and mapper. It can be used with short reads, assembly contigs, long noisy genomic and RNA-seq reads (Li 2018). The lra tool requires less time and memory for alignment as compared to Minimap2 (Ren and Chaisson, 2021). The recently developed kngMap (*k*-mer neighbourhood graph-based mapper) tool is specifically designed to align long noisy reads to a reference genome (Wei et al. 2022).

The third step is variant calling. The variations in the output sequences compared to the reference sequence are called as variants. The presence of SNPs, INDEL, presence/absence variations (PAVs), copy number variations and haplotypes blocks are detected using variant calling tools. Tools used for variant calling includes SAM tools (Li et al. 2009), Genome Analysis Tool Kit Haplotype Caller (GATK-HC) (McKenna et al. 2010), Freebayes (Garrison and Marth 2012), SNPSVM (O'Fallon et al. 2013), varScan (Koboldt et al. 2013), DeepVariant (Poplin et al. 2018), Torrent Variant Caller (TVC) (Life Technologies, Rockville, MD), etc. Numerous automated workflows have been developed to streamline the variant calling process. These workflows integrate various aligners and variant calling tools with other upstream and downstream tools to provide an end-to-end solution (Kanzi et al. 2020). Tools available like ToTem and Appreci8 (Tom et al. 2018; Sandmann et al. 2018) are completely automated variant calling pipelines. ToTem is becoming

Table 3.3 Bioinformatics tools used for NGS data analysis

Approach	Tool	Key feature	Link
Quality check	FastQC	Quality control checks on raw sequence data coming from high-throughput sequencing pipelines	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
	fastp	It can perform quality control, adapter trimming, quality filtering	https://github.com/OpenGene/fastp
Sequence alignment	BWA	Mapping low-divergent sequences against a large reference genome	http://bio-bwa.sourceforge.net/
	Bowtie2	Bowtie2 is an ultrafast and memory efficient tool for aligning sequencing reads	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
	CUSHAW3	Mapping with high computational efficiency	http://cushaw3.sourceforge.net/homepage.htm#latest
	kngMap	Align long reads to a reference sequence	https://github.com/zhang134/kngMap
	MOSAIK	Mapping second- and third-generation sequencing reads	https://github.com/wanpinglee/MOSAIK
	Novoalign	Mapping of short reads onto a reference genome from different NGS platforms	http://www.novocraft.com/products/novoalign/
	SOAP3-dp	SOAP3 is the first short-read alignment tool that leverages the multiprocessors in a graphic processing unit (GPU) to achieve a drastic improvement in speed. SOAP3 is the first short-read alignment tool that leverages The multiprocessors in a graphic processing unit (GPU) to achieve a drastic improvement in speed. SOAP3 is the first short-read alignment tool that leverages the multiprocessors in a graphic processing unit (GPU) to achieve a drastic improvement in speed. Consider alignment with Indels in addition to mismatches.	http://soap.genomics.org.cn/
	MAQ	Builds assembly by mapping short reads to reference sequences	http://maq.sourceforge.net/
	Minimap2	It is accurate and efficient for long noisy genomic and RNA sequences	https://github.com/lh3/minimap2
Variant calling	GATK	Set of bioinformatics tools for analysing high-throughput sequencing and variant call format data	https://software.broadinstitute.org/gatk/
	Freebayes	It is a haplotype-based variant detector and is a great tool for calling variants from a population	https://github.com/ekg/freebayes
	DeepVariant	It is an analysis pipeline that uses a deep neural network to call genetic variants	https://github.com/google/deepvariant
	Platypus	It is a haplotype-based variant caller	

(continued)

Table 3.3 (continued)

Approach	Tool	Key feature	Link
			http://www.well.ox.ac.uk/platypus
	VarScan	An open source tool for variant detection that is compatible with several short-read aligners	http://dkoboldt.github.io/varscan/
	ToTem	Primary role is to automatically generate, execute and benchmark different variant calling pipeline settings	https://totem.software/
	Appreci8	That combines and filters the variant calling results of eight different tools	https://hub.docker.com/r/wwuimi/appreci8/
Data visualization	IGV	It is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets	https://igv.org/
	VISTA	Based on global alignment strategies and a curve-based visualization technique and it also used for comparative analysis	https://genome.lbl.gov/vista/index.shtml
	R software	Gosling: It is a grammar for scalable and interactive genomics data visualizations	http://gosling-lang.org/

a popular tool because it has automated pipeline optimization and efficient analysis management. Appreci8 gives an accurate variant calling as it uses eight different tools to perform the same task that filters and combines the outputs for appropriate calling. Final step is data visualization; there are various tools available for visualization depending on the experiments and the research objectives. One of the popular choices of visualization tool for reference genomes is integrated genome viewer (Thorvaldsdottir et al. 2012). VISTA is also visualization tool which can be used for comparing difference between two genomic sequences. To aid the biologists with no or little knowledge of using perl/python languages, desktop solutions for a wide range of genomic analysis needs, including transcriptomics, variant calling, epigenomics, metagenomics, comparative genomics, are available like Qiagen CLC Genomics Workbench, geWorkbench, Partek Genomics Suite, JMP Genomics, DNA Baser-NextGen Sequence Workbench, etc.

During NGS analysis, numerous intermediate analysis and result files are generated that require large storage. It is difficult to interpret these complicated NGS data files in terms of converting data into knowledge for important traits, especially for aggregated vast volumes of variants or heterogeneous sequencing data require a high-performance computational resource. The NGS data after analysis could be effectively interpreted using machine learning-based techniques.

3.5 Approaches for Development of Genome and Pangenome Assemblies

The wild relatives have a large genetic diversity and ability to survive under various biotic and abiotic stresses. Crop domestication and evolution have significantly decreased the genetic diversity in cultivated species, which has led to the loss of key loci that govern crucial traits. The traditional crop improvement approaches include selection of superior traits from either cultivated varieties or the wild relatives and utilizing them in the breeding programs (Dempewolf et al. 2017). During the process of selection, the crops became more susceptible to different stresses due to impact of climate change and evolution of pathogens and pests. To address these limitations, it is necessary to utilize crop wild relatives, which are known to have genes for several biotic/abiotic stress tolerance traits that have been lost during domestication or breeding procedures. As a result of advancement in sequencing technologies, reference genome sequences for a number of crops have been accessible, serving as the foundation for efforts to boost crop improvement programme (Varshney et al. 2017a, 2017b). In addition to cultivated crop genome, de novo assembled genomes of a number of wild relatives have also been made available. In addition, the idea of pangenomes is being adopted more widely due to the growing recognition that a single reference genome cannot capture the diversity contained within a species.

Pangenome is the collection of genes or DNA sequence in a species to provide useful sources for functional genomics, evolutionary studies that can be used for crop improvement. Pangenomic studies have been conducted in various model and crop plants including *Arabidopsis*, stiff brome, wheat, cabbage, tomato, soybean, rice, rapeseed, barley, chickpea and sorghum (Hurgobin et al. 2018; Gao et al. 2019; Jayakodi et al. 2020; Barchi et al. 2021; Ruperao et al. 2021; Varshney et al. 2021b; Jha et al. 2022) (Table 3.4). Genome assembly is the process of arranging nucleotides in the proper order. Sequence read lengths are currently far shorter than most of genomes or even most of the genes; therefore, it is important to assemble reads and construct genome or pangenome. In plants or other eukaryotic organisms, genes are found in the same physical place on the chromosome, but the frequency of copies and repeating sequences can vary, making assembly more difficult. Pangenomes have been constructed via de novo, iterative, and graph-based assembly techniques. The de novo assembly is straightforward and simplest approach for development of pangenome. This approach includes assembly using overlapping regions and does not require reference genome. It requires high depth sequencing of all the targeted accessions, then creates unique de novo assemblies for each accession. The comparison of the resulting individual assemblies identifies conserved and variable genomic regions across the genomes. Advancement in long-read sequencing technologies and complementary strategies like creation of Hi-C and BioNano maps make it possible to obtain high-quality plant genomes at the chromosomal level (Miga 2020). Comparative analysis is used to identify all types of variations and characterized genes found in core and dispensable regions (Mahmoud et al. 2019).

Table 3.4 Summary of important tools in various plant genetics and genomics approaches

Approach	Tool	Key feature	Link
Pangenome	EUPAN	Large-scale eukaryotic pangenome analyses and detection of gene PAVs at a relatively low sequencing depth	https://cgm.sjtu.edu.cn/eupan/index.html
	GET_HOMOLOGUES-EST	Highly customized and automated pipeline especially designed for people with non-bioinformatics background	https://github.com/eead-csic-compbio/get_homologues/releases
	PAN2HGENE	Computational tool that allows identification of gene products missing from the original genome sequence	https://sourceforge.net/projects/pan2hgene-software/
	Panakeia	Providing a detailed view of the pangenome structure which can efficiently be utilised for discovery, or further in-depth analysis	https://github.com/BioSina/Panakeia
	Pantools	A versatile tool for mapping the metagenomic and genomic reads in both prokaryotes and eukaryotes	https://git.wur.nl/bioinformatics/pantools
	PanViz	An interactive visualization tool to compare the individual genomes to the pangenome	https://github.com/thomasp85/PanViz/blob/master/package.json
	PATO	It performs common tasks of pangenome analysis and also integrates all the necessary functions for the complete analysis with high speed	https://github.com/irycisBioinfo/PATO
	PGAP	Perform pangenome profiling, gene cluster analysis, species evolution analysis, gene enrichment, and genetic variation analysis	https://sourceforge.net/projects/pgap/
	PGAP-X	Analyse pangenome profile curve, gene distribution analysis, genomic region variations, and comparative analysis of genome structure	http://pgapx.ybzhao.com/
	ppsPCP	Detect presence/absence variations and assembled comprehensive pangenome	http://cbi.hzau.edu.cn/ppsPCP/
	RPAN	Rich source for rice genomic research and breeding	https://cgm.sjtu.edu.cn/3kricedb/
Haplotype	Falcon phase	Groups long-read contigs into two separate haplotypes based on hi-C data	https://github.com/phasegenomics/FALCON-Phase

(continued)

Table 3.4 (continued)

Approach	Tool	Key feature	Link
	Hap10	Novel algorithm for haplotype assembly of polyploid genomes using linked reads	https://sourceforge.net/projects/sdhap/
	HapCut2	Robust and accurate haplotype assembly for diverse sequencing technologies	https://github.com/vibansal/HapCUT2
	HaploConduct	Package designed for reconstruction of individual haplotypes	https://github.com/HaploConduct/HaploConduct
	HaplotypeTools	Analysing hybrid or recombinant diploid or polyploid genomes and identifying parental ancestry for sub-genomic regions	https://github.com/rhysf/HaplotypeTools
	HAPLOVIEW	Analysis and visualization of LD and haplotype maps	https://www.broadinstitute.org/haploview/haploview
	HAPPE	Facilitates informative displays wherein data in plots are easy to read and access	https://github.com/fengcong3/HAPPE
	HapTree	Provide polyploid haplotype assembly tool based on a statistical framework.	http://cb.csail.mit.edu/cb/haptree/
	Hifiasm	Fast haplotype-resolved de novo assembler for PacBio HiFi reads	https://github.com/chhylp123/hifiasm
	SDip	Graph-based approach to haplotype-aware assembly	https://github.com/shilpagarg/sdip
	WhatsHap	Reconstruct the haplotypes and then write out the input VCF augmented with phasing information	https://whatshap.readthedocs.io/en/latest/
k-mer	BFCounter	Program for counting *k*-mers in DNA sequence data	http://pritch.bsd.uchicago.edu/bfcounter.html
	iMOKA	Utilized fast and accurate feature reduction step	https://github.com/RitchieLabIGH/iMOKA
	KAT	Multi-purpose software toolkit for reference-free quality control (QC) of WGS reads and de novo genome assemblies	https://github.com/TGAC/KAT
	KITSUNE	Identifying optimal *k*-mer length for alignment free phylogenomic analysis	https://github.com/natapol/kitsune
	KmerGO	Identify group-specific sequences using *k*-mers	

(continued)

Table 3.4 (continued)

Approach	Tool	Key feature	Link
			https://github.com/ChnMasterOG/KmerGO
Genome editing	CHOPCHOP	Web-based tool to select target sites for CRISPR/Cas9- or TALEN-directed mutagenesis	https://chopchop.cbu.uib.no/
	CLD	Suitable for the design of libraries using modified CRISPR enzymes and targeting non-coding regions	https://github.com/boutroslab/cld
	CRISPETa	Design optimal pairs of sgRNAs for deletion of desired genomic regions	http://crispeta.crg.eu/
	CRISPOR	Finds guide RNAs in an input sequence and ranks them according to different scores	http://crispor.tefor.net/
	CRISPR-FOCUS	Web-based platform to search and prioritize sgRNAs for CRISPR screen experiments	http://cistrome.org/crispr-focus/
	CROPSR	Highly effective and efficient to design gRNA in crop plants	https://github.com/H2muller/CROPSR
	E-CRISP	Computational tool to design and evaluate guide RNAs for use with CRISPR/Cas9	http://www.e-crisp.org/E-CRISP/

Several bioinformatics tools have been developed for assembling the prokaryotic pangenome and having the ability to handle less complex genomic content (Khan et al. 2020). For constructing eukaryotic pangenomes, some tools have been developed (Table 3.4) that include EUPAN (Hu et al. 2017), GET_HOMOLOGUES (Contreras-Moreira and Vinuesa 2013), PanTools (Sheikhizadeh et al. 2016), etc. One of the first attempts to examine eukaryotic pangenomes was EUPAN, which supported genome assembly, identification of core and dispensable gene databases using read coverage, and gene annotation of the pan-genomic dataset. GET_HOMOLOGUES can be used in eukaryotic pangenome development and it is written in Perl and R language platform. Additionally, Panconda tool (Warren et al. 2017) is used to compare whole genome multiple sequence and representing relations between sequence as graph and it is the initial step for the de Bruijn graph which can be used for pangenome construction. PanTools is also used to construct and visualize pangenome, the representation of pangenome depending on the de Bruijn graphs. PAN2HGENE (Silva de Oliveira et al. 2021) recently developed computational tools for pangenome analysis, which can do automated comparison analysis for both full and draft genomes and identifies gene that are missing from the original genome sequence.

3.6 Bioinformatics Tools Used in *K*-Mer Analysis

The importance of supporting sequencing technologies has been highlighted by our growing understanding of biological information and its implications for the vast volume of DNA data. Counting *k*-mers is an essential component for many bioinformatics techniques, such as nucleotides assembly, metagenomic sequencing and sequencing error correction (Melsted and Pritchard 2011). A *k*-mer is unique sub-sequence of nucleotide sequence. The distribution of statistically significant *k*-mers in a genomes and other regulatory subregions has been described in a number of recent studies (Hashim and Abdullah 2015; Cserhati et al. 2018). It has also been also employed in comparative studies (Cserhati et al. 2019), and major advantages of alignment-free approaches based on *k*-mer are their speed and ability to remove biases. Most of the association mapping studies has been done using SNPs. However, this approach has some limitations (Rahman et al. 2018). A *k*-mer-based analysis is alternative method to address some limitations of SNP-based analysis.

At its most basic, *k*-mer count analysis simply considers two parameters: the length of the *k*-mer and whether the orientation of the DNA strand is known. *k* is normally selected to be at least 20 and frequently falls between 20 and 31. Too small *k* will give redundant count information because the probability that a *k*-mer is unique to a genome is reduced. However, as *k* increases the probability that a *k*-mer contains an error increases. There are a number of bioinformatics tools developed to analyse the *k*-mer and further utilization of *k*-mers. BFCounter is a program that is used for counting *k*-mers in DNA sequence data (Melsted and Pritchard 2011) (Table 3.4). KAT (*k*-mer Analysis Toolkit) is a multipurpose tool for reference-free quality control and de novo assembly (Mapleson et al. 2017). iMOKA (interactive multi-objective *k*-mer analysis) is bioinformatical tool/software that enables comprehensive analysis of large collections of sequencing data based on *k*-mer. It uses efficient and effective steps that combines Naive Bayes classifier augmented by an adaptive entropy as well as graph-based filter to reduce search time (Lorenzi et al. 2020). KmerGO software is utilized to identify group-specific nucleotide sequences between two different groups. Furthermore, it is also used to check association between nucleotide sequence and quantitative traits (Wang et al. 2020). KITSUNE is a tool to identify the empirically optimal *k*-mer length for phylogenetic analysis and provides alternative alignment tool for comparative studies (Pornputtapong et al. 2020).

3.7 Artificial Intelligence

Artificial intelligence (AI) is the simulation of human intelligence processes by computer systems and it holds marvellous promise for better utilization of the available dataset to appropriate prediction and better understanding of genetic complexity (Fig. 3.1a, b). The three cognitive skills that make up AI encoding are

learning (acquiring data and then developing algorithms to transform it into usable information), reasoning (selecting the appropriate algorithm to arrive at a desired result), and self-correction (constantly adjusting designed algorithms to ensure that they deliver the most accurate results) (Gharaei et al. 2019). Breeders have access to an ever-growing suite of high-throughput sensors and imaging techniques for a wide range of traits and situations in the field. In addition, novel genomic assays are constantly being developed that can reveal missing heritability (Harfouche et al. 2019). Nowadays, a major challenge in the advancement of technologies is the management and utilization of big data. The utilization of data with AI technologies can accelerate the breeding program to increase productivity and development of climate-resilient crop by phenotyping, efficient and effective diagnosis of disease and precise selection of individual for breeding (Fig. 3.1c). AI can also help breeders to quickly determine which plants grow the quickest in a specific climate, which genes support plant growth and adaptation, produce the best gene combination for a given location and choosing traits that increase yield and fend off the effects of a changing climate.

One of the important elements in AI is machine learning (ML), which helps to use data more efficiently and that uses statistical and mathematical approaches for appropriate predictions (Ayed and Hanana 2021). The ML has ability of ML to distinguish between various types of genomic regions, for instance, distinguishing active genes and pseudogenes, using feature like DNA methylation (Sartor et al. 2019). Additionally, ML was utilised to foresee the locations of DNA crossover (Demirci et al. 2018). Single-cell RNA sequencing is fascinating the new area in which ML is essential (Speranza et al. 2021; van Dijk et al. 2021b). This method makes it possible to examine cellular development and responses to environmental stimuli in diverse tissues. Digital plant phenotyping has been an active study area to accelerate plant science studies. Different imaging systems can be used to study the various macroscopic levels, for example, real-time stomata phenotyping using microscopic observation (Toda et al. 2021). Numerous sensors have been employed to accurate phenotyping, and it includes spectral sensor, lidar/laser sensor, fluorescence sensor, ultrasonic sensor and thermography (Qiu et al. 2018).

AI systems currently in use neural networks (NNs) and extreme gradient boosting (XGboost), both of which are popular machine learning models employed for a variety of tasks including regression and classification (Chen and Guestrin 2016). Deep learning techniques are based on neural networks, sometimes referred to as artificial/ simulated neural networks, which are a subset of machine learning. Leveraging AI in agriculture shows impressive results in image-based disease identification using deep learning model. It uses publicly available image datasets for disease identification (Mohanty et al. 2016). However, the supervised branch of machine learning includes the tree-based method known as XGboost. In maize, different models were used to predict yield using AI and found better results using XGBoost (Nyeki et al. 2019). These AI systems internal working and decision-making procedures are mysterious. It is possible to see the results, but it is not clear why a particular choice was picked. As a result, the introduction of new explainable AI algorithms that not only have a prediction model but also gives the appropriate

reasons for choice is needed. It is the first stage in the development of next-generation AI (Harfouche et al. 2019).

3.8 Identification of Superior Haplotype for Crop Improvement

Second-generation molecular markers have been successfully used in plant breeding for development of improved varieties and also utilised in genome mapping, but gives low resolution of QTLs (Zargar et al. 2015). Advancement in the NGS technologies provide sequence-based markers (SNPs) having wide coverage with high density (Gouda et al. 2021), and have wide applications in plant breeding. These markers help to increase the resolution of genome mapping and the accuracy of genomic selection (Yadav et al. 2019). However, identified SNPs have some limitations which includes bi-allelic nature, difficult to identify rare alleles, less polymorphic, linkage drag problem and giving false positive results (Voss-Fels and Snowdon 2016; Bhat et al. 2021). In this context, the haplotype-based approaches are a successful strategy to get over SNPs limitations and boost the resolution of genomic regions (Qian et al. 2017). Haplotype is combination of nucleotide or markers that inherit together from polymorphic sites in the same or different chromosome having strong linkage disequilibrium between them (Bhat et al. 2021). Number of studies have demonstrated that a haplotype-based association study can find variants that would not be detected by a typical SNP-based investigation (Zakharov et al. 2013). Additionally, a recent study also identified several important genes, that can be utilized as important molecular markers for the purpose of genetic manipulation to design and develop robust and resistant crop cultivars (Pal et al. 2022).

The detection of haplotypes and their use in genetic investigations is significantly impacted by the availability of high-throughput sequencing technologies. Second-generation sequencing technologies generate 150 base pairs short reads. Therefore, the haplotypes identification is difficult and requires powerful statistical tools (Delaneau et al. 2019). On the other hand, third-generation sequencing technologies, such as Oxford Nanopore and Pacific Biosciences, generates long reads from which the haplotypes can be constructed directly (Maestri et al. 2020). The haplotype mining can be used to dissect complex traits by using approaches like haplotype-based breeding, haplotype-GWAS, haplotype-assisted genomic selection (Table 3.4).

Haplotype identification, characterization and visualization are important for utilization of haplotype for crop improvement. Many tools have been developed to estimate and visualize haplotypes. Haplotype identification/estimation also called as "phasing," is a process of estimation or construction of the haplotype sequences from genotypic data and it is utilized for understanding sequence-specific variation. Haplotype-based GWAS analysis is complicated as compared to SNP-based analysis

to identify the associations, because it involves three major steps: phasing/haplotype estimation, block determination and statistical analysis. Estimation of haplotypes required pooled information of all individuals present in sample. Number of unrelated individuals is an important factor that can influence the estimation of haplotypes, and more individuals can give better results. However, related individuals can be phased by considering haplotypes shared by members of families which are descended from one another (Browning and Browning 2011). Numerous phasing techniques that enable the construction of haplotypes from long-read sequencing data have recently been established, such as reference-based phasing, de novo genome assembly and strain-resolved metagenome assembly (Garg 2021; Bhat et al. 2021). Choice of appropriate phasing, block determination algorithms and their interaction are important factors that can influence accuracy of phasing the haplotype blocks (Bkhetan et al. 2019). Various haplotype analysis approach combined with different computational tools such as DESMAN, Falcon phase, HapCut2, HapTree, Hifiasm, MetaMaps, POLYTE, SDip, and WhatsHap are extensively reviewed by Garg (2021). The combination of different analysis approaches and computational tools with long-reads sequencing technologies has allowed us to fully utilise the potential of these sequencing methodologies for haplotype construction. SNPViz v2.0 (Zeng et al. 2020) is a web-based tool that enhances the identification of large-scale haplotype blocks. HaplotypeTools (Farrer 2021) is tool to phase variant, based on detecting the reads overlapping ≥ 2 heterozygous positions and then extent of the reads; it is also a powerful tool for analysing hybrid and polyploid genomic regions. Recently, python coded tool HAPPE (Feng et al. 2022) was developed to construct and visualize the haplotypes easily (Table 3.4). Additionally, Practical Haplotype Graph is a powerful tool for storage, retrieval and imputation of haplotypes that can be used for genomic studies (Bradbury et al. 2022).

3.9 Genome Editing

CRISPR/Cas9 is the potent genetic modification technique that is a great example of genome editing technologies. This technology is proved to be extremely effective tool not only in the field of basic science but also in the plant breeding. The development of genome editing technologies (ZFN, TALEN, CRISPR/Cas9, etc.) drawn a lot of attention, because they eliminate the restrictions of traditional breeding approaches (Matres et al. 2021). These methods enable precise and effective targeted genome modifications, greatly shortening the time needed to obtain plants with desired traits for the development of new crop varieties. Sequence-specific nucleases and small guide RNA are the key components of CRISPR-based gene editing approach to generate precise modification. The CRISPR/Cas system is still evolving, but there are two significant obstacles: off-target effects and on-target efficiency (Xu et al. 2015; Zhang et al. 2015; Liu et al. 2020). To overcome these issues, optimizing small guide RNA by effective computer methods assist in silico gRNA design that plays an important role (Doench

et al. 2016; Hassan et al. 2021). One of the key factors affecting gRNA effectiveness is the nucleotide content of a target sequence. The PAM (Protospacer Adjacent Motif) sequence and its nearby nucleotide is significantly important for the better efficiency (Liu et al. 2020). Guanines are favoured at first and second nucleotide position before the PAM sequence while thymines are not preferred within four nucleotides upstream/downstream of PAM sequence. Furthermore, sequences upstream of PAMs have no discernible influence, although sequences downstream can affect gRNA efficiency (Doench et al. 2014). At cleavage site, cytosine is preferred and GC content at downstream of the PAM sequence that increases high efficiency to gRNA. Numerous efficiency prediction models are available built using this important information. Various tools have been developed based on these models to design gRNA either by alignment-based, hypothesis-driven and/or learning-based models (Konstantakos et al. 2022). Hypothesis-driven and learning model-based tools perform better than alignment-based models. Several tools have been developed to predict gRNA with high target efficiency includes E-CRISP, CHOPCHOP, CRISPR-FOCUS, PROTOSPACER, CLD, CRISPOR, and CRISPETa (Table 3.4). WheatCRISPR is a web-based bioinformatics tool which is generally used for constructing target-specific gRNA in wheat (Cram et al. 2019). Additionally, CROPSR is the first open source bioinformatics tool to help design genome-wide guide RNA for CRISPR-based genome editing with high speed that reduces the challenges of complex crop genome (Paul et al. 2022).

3.10 Major Challenges in Bioinformatics

NGS technologies have made genomic revolution by generating enormous amount of data quickly and affordably. The use of bioinformatics in life science research is becoming more and more essential at the moment. Data analysis is frequently the main bottleneck because of the exponential growth in amount and complexity of life science data over the past two decades. Handling, analysing, and storing information has become a new barrier for biologists. Efficient data processing is necessary and there are many algorithms available for these specific tasks. To increase efficiency and accuracy, it needs combination of tools and enough resources for smooth operation. Another challenge for the biologists is to learn the languages like python, Perl or R for efficient handling of the data and lack of training in the field by the expert bioinformatician who knows biological problems and associated complexities. Genome assembly has gained more and more attention as advance sequencing technology are developed. Despite the abundance of genome assembly tools available, de novo genome assembly using next-generation reads still faces four significant obstacles: sequencing errors, sequencing bias, topological complexity of repetitive regions and huge computational resource consumption (Liao et al. 2019). The accuracy of results can have a big impact on downstream analysis of sequencing data. False positives and inaccurate findings may result from the errors during data processing. On the other side, poorly chosen approaches or tools may

produce false negatives, which would result in the loss of genuine variants. Therefore, finding a suitable balance between accuracy of results and sensitivity is thus another big problem for data analysis. The application of ML in plant research is also an important issue. Traditionally, statistical techniques have been used to predict genotype-phenotype relationships. These techniques have been very effective and successful throughout the past century. Decision-making for researchers and practitioners typically involves the use of confidence measures and model interpretation. Further, data-driven flexibility of ML offers a range of advantages over stringent statistical approaches that make it a powerful tool for solving complex problems and extracting valuable insights from diverse and dynamic datasets.

3.11 Future Prospective and Conclusions

Bioinformatics has been emerging and cross-cutting different fields of agricultural sciences for enhancing our understanding of the complex mechanism underlying different traits in different crop plants in crop improvement (Fig. 3.2). A paradigm shift in the field of life sciences has been brought by NGS and has transformed genomics research. In addition to being crucial for fundamental genomic and molecular biology research, bioinformatics also has a significant influence on many fields of agricultural and medical sciences. Suitable computational tools and the right resources are essential for identifying biological information that adds value and offers novel insights into biological systems. The rise in omics-based research needs education in the relevant technologies and bioinformatics in order to correctly translate experimental and computational efforts. AI-based solutions are help to increase efficiency and regulate a number of factors, including crop yield, soil profile, crop irrigation, weeding, and crop monitoring (Bhardwaj et al. 2022). The possibility of using AI in agriculture will increase as the field of AI matures and more trained algorithms are added. Recently, the development of genetic algorithm-based Internet of precision agricultural things (IopaT) and becoming famous in rural areas to solve the real-time problems. Genetic algorithmic system is developed to predict water requirement (Roy and De 2020). This kind of system will also help in decision-making in agriculture, like crop patterns and water management at particular place (Xu et al. 2022a). Future applications of AI/ML in plant research include predicting which regions of the genome should be modified to produce a particular phenotype and providing the best possible local growing conditions by monitoring crop performance in vivo in the greenhouse or on the field. We are still very early in the genomics era, and undoubtedly, a long way from accomplishing the ambitious objective. In fact, efforts are still required for in-depth and appropriate analyses of genome, transcriptome, and metagenome data to identify link between organization and functionality. Moreover, chemical genomics approaches aid in the comprehension of overcoming stress conditions and improving crop yield and productivity (Pa et al. 2022; Adhinarayanreddy et al. 2022). Utilizing integrated multi-omics data, big data technology, and artificial intelligence proposed the new term called

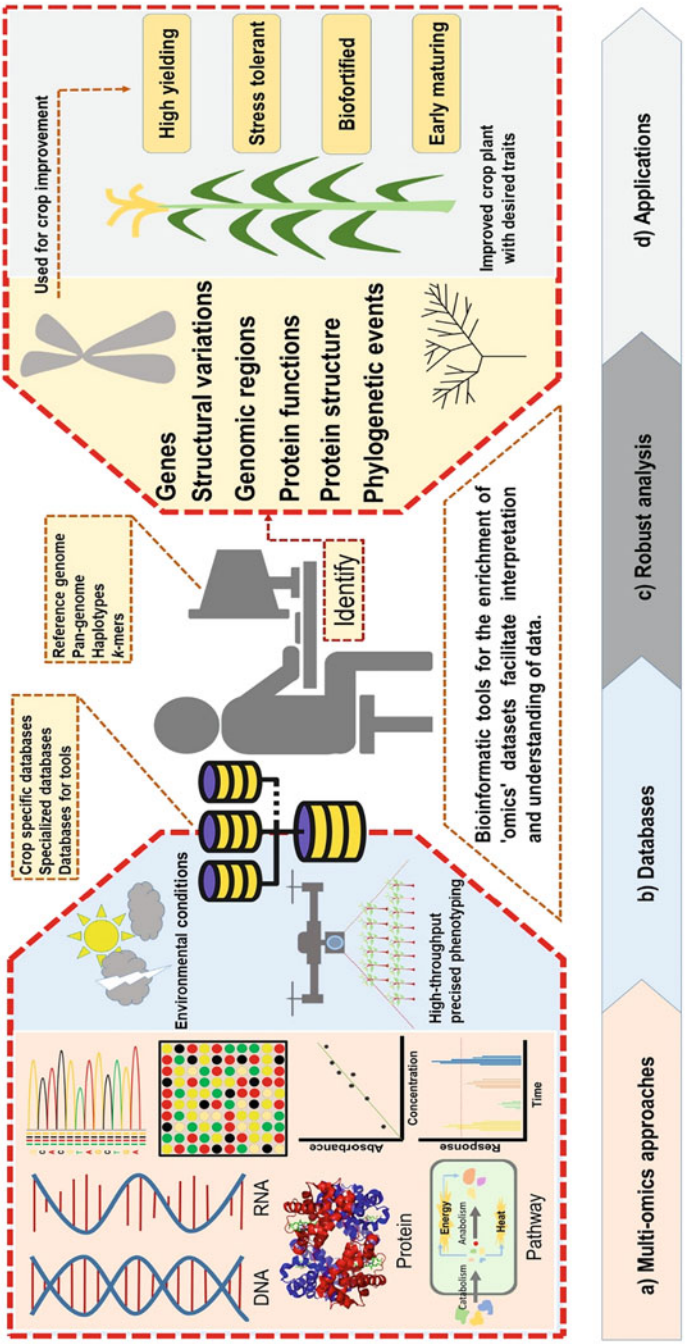

Fig. 3.2 Role of bioinformatics in genetics and plant breeding research for developing climate-resilient crops and sustainable food production. (**a**) Generation of biological data from various omics approaches as well as phenotyping data from multiple environments; (**b**) Storage and processing of different omics data generated. (**c**) Robust analysis of the raw data and transforming to useful information using bioinformatical tools for appropriate interpretation; (**d**) Application of bioinformatics in agricultural research

integrated genomic-enviromic prediction (Xu et al. 2022b), as an extension of genomic prediction will provide accelerating breeding programs. With the use of big data, AI and robust bioinformatical analysis, plant breeding in the future will become increasingly smart. The establishment of integrative plant breeding platforms and open-source breeding initiatives can help translate smart breeding efforts into genetic gains.

References

Addinsoft (2021) XLSTAT statistical and data analysis solution. New York, USA

Adhinarayanreddy V, Vijayaraghavareddy P, Vargheese A, Sujitha DA, Uttarkar A, Niranjan V, Anuradha CV, Sheshshayee MS, Vemanna R (2022) A simple and rapid oxidative stress screening method of small molecules for functional studies of transcription factor. Rice Sci 2022:3

Amarasinghe SL, Ritchie ME, Gouil Q (2021) Long-read-tools. Org: an interactive catalogue of analysis methods for long-read sequencing data. GigaScience 10(2):1–7

Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q (2020) Opportunities and challenges in long-read sequencing data analysis. Genome Biol 21(1):1–16

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data

Ayed BR, Hanana M (2021) Artificial intelligence to improve the food and agriculture sector. J Food Qual 2021:1–7

Balcke GU, Handrick V, Bergau N, Fichtner M, Henning A, Stellmach H, Tissier A, Hause B, Frolov A (2012) An UPLC-MS/MS method for highly sensitive high-throughput analysis of phytohormones in plant tissues. Plant Methods 8(1):1–11

Barchi L, Rabanus-Wallace MT, Prohens J, Toppino L, Padmarasu S, Portis E, Rotino GL, Stein N, Lanteri S, Giuliano G (2021) Improved genome assembly and pan-genome provide key insights into eggplant domestication and breeding. Plant J 107(2):579–596

Basten CJ, Weir BS, Zeng ZB (2002) QTL cartographer, version 1.17. Department of Statistics, North Carolina State University, Raleigh, NC

Batley J, Edwards D (2016) The application of genomics and bioinformatics to accelerate crop improvement in a changing climate. Curr Opin Plant Biol 30(2):78–81

Bhardwaj A, Kishore S, Pandey DK (2022) Artificial Intelligence in Biological Sciences. Life 12: 1430

Bhat JA, Yu D, Bohra A, Ganie SA, Varshney RK (2021) Features and applications of haplotypes in crop breeding. Communications Biology 4(1):1–12

Bhatta M, Morgounov A, Belamkar V, Wegulo SN, Dababat AA, Erginbas-Orakci G, Bouhssini ME, Gautam P, Poland J, Akci N, Demir L (2019) Genome-wide association study for multiple biotic stress resistance in synthetic hexaploid wheat. Int J Mol Sci 20(15):3667

Bkhetan AZ, Zobel J, Kowalczyk A, Verspoor K, Goudey B (2019) Exploring effective approaches for haplotype block phasing. BMC Bioinform 20(1):1–14

Bradbury PJ, Casstevens T, Jensen SE, Johnson LC, Miller ZR, Monier B, Romay MC, Song B, Buckler ES (2022) The practical haplotype graph, a platform for storing and using pangenomes for imputation. Bioinform 38(15):3698–3702

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinform 23(19):2633–2635

Browning SR, Browning BL (2011) Haplotype phasing: existing methods and new developments. Nat Rev Genet 12(10):703–714

Channale S, Kalavikatte D, Thompson JP, Kudapa H, Bajaj P, Varshney RK, Zwart RS, Thudi M (2021) Transcriptome analysis reveals key genes associated with root-lesion nematode *Pratylenchus thornei* resistance in chickpea. Sci Rep 11(1):1–11

Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 785–794

Chen C, Shang X, Sun M, Tang S, Khan A, Zhang D, Yan H, Jiang Y, Yu F, Wu Y, Xie Q (2022) Comparative transcriptome analysis of two sweet sorghum genotypes with different salt tolerance abilities to reveal the mechanism of salt tolerance. Int J Mol Sci 23(4):2272

Chen S, Zhou Y, Chen Y, Gu J (2018) Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinform 34(17):884–890

Contreras-Moreira B, Vinuesa P (2013) GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol 79(24):7696–7701

Cram D, Kulkarni M, Buchwaldt M, Rajagopalan N, Bhowmik P, Rozwadowski K, Parkin IA, Sharpe AG, Kagale S (2019) WheatCRISPR: a web-based guide RNA design tool for CRISPR/Cas9-mediated genome editing in wheat. BMC Plant Biol 19(1):1–8

Cserhati M, Xiao P, Guda C (2019) K-mer-based motif analysis in insect species across anopheles, drosophila, and Glossina genera and its application to species classification. Computational and mathematical methods in medicine 1–16

Cserhati MF, Mooter ME, Peterson L, Wicks B, Xiao P, Pauley M, Guda C (2018) Motifome comparison between modern human. Neanderthal and Denisovan BMC Genomics 19(1):1–9

Delaneau O, Zagury JF, Robinson MR, Marchini JL, Dermitzakis ET (2019) Accurate, scalable and integrative haplotype estimation. Nat Commun 10(3):1–10

Demirci S, Peters SA, de Ridder D, van Dijk AD (2018) DNA sequence and shape are predictive for meiotic crossovers throughout the plant kingdom. Plant J 95(4):686–699

Dempewolf H, Baute G, Anderson J, Kilian B, Smith C, Guarino L (2017) Past and future use of wild relatives in crop breeding. Crop Sci 57(3):1070–1082

Doddamani D, Khan AW, Katta MA, Agarwal G, Thudi M, Ruperao P, Edwards D, Varshney RK (2015) CicArVarDB: SNP and InDel database for advancing genetics research and breeding applications in chickpea. Database 2015:1–7

Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard R, Virgin HW (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol 34(2):184–191

Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, Sullender M, Ebert BL, Xavier RJ, Root DE (2014) Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. Nat Biotechnol 32(12):1262–1267

Edwards D, Stajich J, Hansen D (eds) (2009) Bioinformatics: tools and applications. Springer, New York

Farrer RA (2021) HaplotypeTools: a toolkit for accurately identifying recombination and recombinant genotypes. BMC Bioinform 22(1):1–15

Feng C, Wang X, Wu S, Ning W, Song B, Yan J, Cheng S (2022) HAPPE: a tool for population haplotype analysis and visualization in editable excel tables. Front Plant Sci 13:1–7

Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, Burzynski-Chang EA, Fish TL, Stromberg KA, Sacks GL, Thannhauser TW (2019) The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat Genet 51(6):1044–1051

Garg S (2021) Computational methods for chromosome-scale haplotype reconstruction. Genome Biol 22(1):1–24

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. arXiv 1207:3907

Gauch HG, Moran DR (2019) AMMISOFT for AMMI analysis with best practices. BioRxiv 538454

Gharaei A, Karimi M, Shekarabi SAH (2019) An integrated multi-product, multi-buyer supply chain under penalty, green, and quality control polices and a vendor managed inventory with

consignment stock agreement: the outer approximation with equality relaxation and augmented penalty algorithm. Appl Math Model 69:223–254
Giacomello S, Salmen F, Terebieniec BK, Vickovic S, Navarro JF, Alexeyenko A, Reimegard J, McKee LS, Mannapperuma C, Bulone V, Stahl PL (2017) Spatially resolved transcriptome profiling in model plant species. Nat Plants 3(6):1–11
Gouda AC, Warburton ML, Djedatin GL, Kpeki SB, Wambugu PW, Gnikoua K, Ndjiondjop MN (2021) Development and validation of diagnostic SNP markers for quality control genotyping in a collection of four rice (*Oryza*) species. Sci Rep 11(1):1–11
Gupta AK, Zhang X, Andrews JG (2015) Potential throughput in 3D ultradense cellular networks. In 49th Asilomar conference on signals, systems and computers,1026–1030. IEEE
Gulles AA, Bartolome VI, Morantte RI, Nora LA, Relente CE, Talay DT, Caneda AA, Ye G (2014) Randomization and analysis of data using STAR [Statistical Tool for Agricultural Research]. Philippine J Crop Sci 39:137
Harfouche AL, Jacobson DA, Kainer D, Romero JC, Harfouche AH, Mugnozza GS, Moshelion M, Tuskan GA, Keurentjes JJ, Altman A (2019) Accelerating climate resilient plant breeding by applying next-generation artificial intelligence. Trends Biotechnol 37(11):1217–1235
Harper L, Campbell J, Cannon EK, Jung S, Poelchau M, Walls R, Andorf C, Arnaud E, Berardini TZ, Birkett C, Cannon S et al (2018) AgBioData consortium recommendations for sustainable genomics and genetics databases for agriculture. Database 2018:1–32
Hashim EK, Abdullah R (2015) Rare k-mer DNA: identification of sequence motifs and prediction of CpG Island and promoter. J Theor Biol 387:88–100
Hassan MM, Chowdhury AK, Islam T (2021) In silico analysis of gRNA secondary structure to predict its efficacy for plant genome editing. In: Islam, Molla (eds) CRISPR-Cas methods, New York, NY, pp 15–22
Heather JM, Chain B (2016) The sequence of sequencers: the history of sequencing DNA. Genomics 107(1):1–8
Hu Z, Sun C, Lu KC, Chu X, Zhao Y, Lu J, Shi J, Wei C (2017) EUPAN enables pan-genome studies of a large number of eukaryotic genomes. Bioinform 33(15):2408–2409
Hurgobin B, Golicz AA, Bayer PE, Chan CKK, Tirnaz S, Dolatabadian A, Schiessl SV, Samans B, Montenegro JD, Parkin IA, Pires JC (2018) Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. Plant Biotechnol J 16(7): 1265–1274
IBM Corp Ibm, SPSS (2017) Statistics for windows, version 25.0. IBM Corp, Armonk, NY
Jayakodi M, Padmarasu S, Haberer G, Bonthala VS, Gundlach H, Monat C, Lux T, Kamal N, Lang D, Himmelbach A, Ens J (2020) The barley pan-genome reveals the hidden legacy of mutation breeding. Nat 588(7837):284–289
Jha UC, Nayyar H, von Wettberg EJ, Naik YD, Thudi M, Siddique KH (2022) Legume Pangenome: status and scope for crop improvement. Plan Theory 22:3041
Kanzi AM, San JE, Chimukangara B, Wilkinson E, Fish M, Ramsuran V, De Oliveira T (2020) Next generation sequencing and bioinformatics analysis of family genetic inheritance. Front Genet 11:e544162
Kathiresan N, Temanni R, Almabrazi H, Syed N, Jithesh PV, Al-Ali R (2017) Accelerating next generation sequencing data analysis with system level optimizations. Sci Rep 7(1):1–11
Khan AW, Garg V, Roorkiwal M, Golicz AA, Edwards D, Varshney RK (2020) Super-pangenome by integrating the wild side of a species for accelerated crop improvement. Trends Plant Sci 25(2):148–158
Khetan M, Ameerpet M (2015) Indostat package for data analysis. Windostat version 9.3 from indostat services, Hyderabad
Koboldt DC, Larson DE, Wilson RK (2013) Using VarScan 2 for germline variant calling and somatic mutation detection. Curr Protoc Bioinform 44(1):15–14
Konstantakos V, Nentidis A, Krithara A, Paliouras G (2022) CRISPR–Cas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning. Nucleic Acids Res 50(7):3616–3637

Kudapa H, Garg V, Chitikineni A, Varshney RK (2018) The RNA-Seq-based high resolution gene expression atlas of chickpea (*Cicer arietinum* L.) reveals dynamic spatio-temporal changes associated with growth and development. Plant Cell Environ 41(9):2209–2225

Lai K, Lorenc MT, Edwards D (2012) Genomic databases for crop improvement. Agron 2(1):62–73

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. Nat Methods 9(4): 357–359

Le Nguyen K, Grondin A, Courtois B, Gantet P (2019) Next-generation sequencing accelerates crop gene discovery. Trends Plant Sci 24(3):263–274

Ledesma R (2008) Software de análisis de correspondencias múltiples: una revisión comparativa. Metodología de encuestas 10(1):59–75

Lee WP, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT (2014) MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. PLoS One 9(3): e90581

Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards SV, Forest F, Gilbert MTP, Goldstein MM (2018) Earth BioGenome project: sequencing life for the future of life. Proc Natl Acad Sci 115(17):4325–4333

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinform 34(18):3094–3100

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinform 25(14):1754–1760

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAMtools. Bioinform 25(16):2078–2079

Liao X, Li M, Zou Y, Wu FX, Wang J (2019) Current challenges and solutions of *de novo* assembly. Quantitat Biol 7(2):90–109

Lincoln SE, Daly MJ, Lander ES (1993) Constructing genetic linkage maps with MAPMAKER/EXP Version 3.0: a tutorial and reference manual. A whitehead institute for biomedical research technical report, 3

Liu G, Zhang Y, Zhang T (2020) Computational approaches for effective CRISPR guide RNA design and evaluation. Comput Struct Biotechnol J 18(2):35–44

Liu Y, Popp B, Schmidt B (2014) CUSHAW3: sensitive and accurate base-space and color-space short-read alignment with hybrid seeding. PLoS One 9(1):e86869

Lorenzi C, Barriere S, Villemin JP, Dejardin Bretones L, Mancheron A, Ritchie W (2020) iMOKA: k-mer based software to analyze large collections of sequencing data. Genome Biol 21(1):1–19

Maestri S, Maturo MG, Cosentino E, Marcolungo L, Iadarola B, Fortunati E, Rossato M, Delledonne M (2020) A long-read sequencing approach for direct haplotype phasing in clinical settings. Int J Mol Sci 21(23):9177

Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ (2019) Structural variant calling: the long and the short of it. Genome Bology 20(1):1–14

Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ (2017) KAT: a k-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinform 33(4):574–576

Mashaki MK, Garg V, Nasrollahnezhad Ghomi AA, Kudapa H, Chitikineni A, Zaynali Nezhad K, Yamchi A, Soltanloo H, Varshney RK, Thudi M (2018) RNA-Seq analysis revealed genes associated with drought stress response in kabuli chickpea (*Cicer arietinum* L.). PLoS One 13(6):e0199774

Matres JM, Hilscher J, Datta A, Armario-Nájera V, Baysal C, He W, Huang X, Zhu C, Valizadeh-Kamran R, Trijatmiko KR, Capell T (2021) Genome editing in cereal crops: an overview. Transgenic Res 30(4):461–498

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9):1297–1303

Melsted P, Pritchard JK (2011) Efficient counting of k-mers in DNA sequences using a bloom filter. BMC Bioinform 12(1):1–7

Miga KH (2020) Centromere studies in the era of 'telomere-to-telomere'genomics. Exp Cell Res 394(2):e112127

Mohanty SP, Hughes DP, Salathe M (2016) Using deep learning for image-based plant disease detection. Front Plant Sci 7:1419

Morin PA, Alexander A, Blaxter M, Caballero S, Fedrigo O, Fontaine MC, Foote AD, Kuraku S, Maloney B, Mccarthy M, Mcgowen M (2020) Building genomic infrastructure: sequencing platinum-standard reference-quality genomes of all cetacean species. Mar Mamm Sci 36:1356–1366

Nayak SN, Agarwal G, Pandey MK, Sudini HK, Jayale AS, Purohit S, Desai A, Wan L, Guo B, Liao B, Varshney RK (2017) *Aspergillus flavus* infection triggered immune responses and host-pathogen cross-talks in groundnut during in-vitro seed colonization. Sci Rep 7(1):1–14

Nyeki AE, Kerepesi C, Daroczy BZ, Benczúr A, Milics G, Kovacs AJ, Nemenyi M (2019) Maize yield prediction based on artificial intelligence using spatio-temporal data precision agriculture '19, eds: John V Stafford, 1011–1017

O'Fallon BD, Wooderchak-Donahue W, Crockett DK (2013) A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data. Bioinform 29(11):1361–1366

Pa V, Vijayaraghavareddy P, Uttarkar A, Dawane A, KC B, Niranjan V, MS S, CV A, Makarla U, Vemanna RS (2022) Novel small molecules targeting bZIP23 TF improve stomatal conductance and photosynthesis under mild drought stress by regulating ABA. FEBS J 289(19):6058–6077

Pacheco A, Vargas M, Alvarado G, Rodríguez F, Crossa J, Burgueño J (2015) GEA-R (genotype x environment analysis with R for windows) version 4.1. hdl 11529(10203):16

Pal G, Bakade R, Deshpande S, Sureshkumar V, Patil SS, Dawane A, Agarwal S, Niranjan V, Prasanna MK, Vemanna RS (2022) Transcriptomic responses under combined bacterial blight and drought stress in rice reveal potential genes to improve multi-stress tolerance. BMC Plant Biol 22(1):1–20

Paul MH, Istanto DD, Heldenbrand J, Hudson ME (2022) CROPSR: an automated platform for complex genome-wide CRISPR gRNA design and validation. BMC Bioinform 23(1):1–19

Pazhamala LT, Purohit S, Saxena RK, Garg V, Krishnamurthy L, Verdier J, Varshney RK (2017) Gene expression atlas of pigeonpea and its application to gain insights into genes associated with pollen fertility implicated in seed formation. J Exp Bot 68(8):2037–2054

Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, Gross SS (2018) A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol 36(10):983–987

Pornputtapong N, Acheampong DA, Patumcharoenpol P, Jenjaroenpun P, Wongsurawat T, Jun SR, Yongkiettrakul S, Chokesajjawatee N, Nookaew I (2020) KITSUNE: a tool for identifying empirically optimal k-mer length for alignment-free phylogenomic analysis. Front Bioeng Biotechnol 23(8):556413

Pour-Aboughadareh A, Yousefian M, Moradkhani H, Poczai P, Siddique KH (2019) STABILITYSOFT: a new online program to calculate parametric and non-parametric stability statistics for crop traits. Appl Plant Sci 7(1):e01211

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Qian L, Hickey LT, Stahl A, Werner CR, Hayes B, Snowdon RJ, Voss-Fels KP (2017) Exploring and harnessing haplotype diversity to improve yield stability in crops. Front Plant Sci 8(1):1–11

Qiu R, Wei S, Zhang M, Li H, Sun H, Liu G, Li M (2018) Sensors for measuring plant phenotyping: a review. International Journal of Agricultural and Biological Engineering 11(2):1–17

Rahman A, Hallgrimsdottir I, Eisen M, Pachter L (2018) Association mapping from sequencing reads using k-mers. elife 13(7):e32920

Ren J, Chaisson MJ (2021) lra: a long read aligner for sequences and contigs. PLoS Comput. Biol 17(6):e1009078

Roy SK, De D (2020) Genetic algorithm based internet of precision agricultural things (IopaT) for agriculture 4.0. Internet of Things 18:100201

Ruperao P, Thirunavukkarasu N, Gandham P, Selvanayagam S, Govindaraj M, Nebie B, Manyasa E, Gupta R, Das RR, Odeny DA, Gandhi H (2021) Sorghum pan-genome explores the functional utility for genomic-assisted breeding to accelerate the genetic gain. Front Plant Sci 12(1):963–980

Sandmann S, Karimi M, de Graaf AO, Rohde C, Gollner S, Varghese J, Ernsting J, Walldin G, van der Reijden BA, Müller-Tidow C, Malcovati L (2018) appreci8: a pipeline for precise variant calling integrating 8 tools. Bioinform 34(24):4205–4212

Sartor RC, Noshay J, Springer NM, Briggs SP (2019) Identification of the expressome by machine learning on omics data. Proc Natl Acad Sci 116(36):18119–18125

Sheikhizadeh S, Schranz ME, Akdel M, de Ridder D, Smit S (2016) PanTools: representation, storage and exploration of pangenomic data. Bioinform 32(17):487–493

Silva de Oliveira M, Thyeska Castro Alves J, Henrique Caracciolo Gomes de Sa P, Veras AADO (2021) PAN2HGENE–tool for comparative analysis and identifying new gene products. PLoS One 16(5):e0252414

Sinha P, Bajaj P, Pazhamala LT, Nayak SN, Pandey MK, Chitikineni A, Huai D, Khan AW, Desai A, Jiang H, Zhuang W (2020) *Arachis hypogaea* gene expression atlas for fastigiata subspecies of cultivated groundnut to accelerate functional and translational genomics applications. Plant Biotechnol J 18(11):2187–2200

Speranza E, Williamson BN, Feldmann F, Sturdevant GL, Pérez-Pérez L, Meade-White K, Smith BJ, Lovaglio J, Martens C, Munster VJ, Okumura A (2021) Single-cell RNA sequencing reveals SARS-CoV-2 infection dynamics in lungs of African green monkeys. Sci Transl Med 13(578): e8146

Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang C, Chougule K, Gao D, Iwata A, Goicoechea JL, Wei S (2018) Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. Nat Genet 50(2): 285–296

Sun Y, Shang L, Zhu QH, Fan L, Guo L (2021) Twenty years of plant genome sequencing: achievements and challenges. Trends Plant Sci 27(4:391–401

Team RC (2013) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. http://www. R-project. org/

Thorvaldsdottir H, Robinson JT, Mesirov JP (2012) Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Briefings in Bioinformatics 14 (2):178–192

Thudi M, Chen Y, Pang J, Kalavikatte D, Bajaj P, Roorkiwal M, Chitikineni A, Ryan MH, Lambers H, Siddique KH, Varshney RK (2021) Novel genes and genetic loci associated with root morphological traits, phosphorus-acquisition efficiency and phosphorus-use efficiency in chickpea. Front Plant Sci 1001

Thudi M, Khan AW, Kumar V, Gaur PM, Katta K, Garg V, Roorkiwal M, Samineni S, Varshney RK (2016) Whole genome re-sequencing reveals genome-wide variations among parental lines of 16 mapping populations in chickpea (*Cicer arietinum* L.). BMC Plant Biol 16(1):53–64

Thudi M, Palakurthi R, Schnable JC, Chitikineni A, Dreisigacker S, Mace E, Srivastava RK, Satyavathi CT, Odeny D, Tiwari VK, Lam HM (2020) Genomic resources in plant breeding for sustainable agriculture. J Plant Physiol 257(1):e153351

Thudi M, Samineni S, Li W, Boer MP, Roorkiwal M, Yang Z, Ladejobi F, Zheng C, Chitikineni A, Nayak S, He Z, Valluri V, Bajaj P, Khan AW, Gaur PM, van Eeuwijk F, Mott R, Xin L, Varshney RK (2023) Whole genome resequencing and phenotyping of MAGIC population for high resolution mapping of drought tolerance in chickpea. Plant Genome 30:e20333. https://doi.org/10.1002/tpg2.20333

Toda Y, Tameshige T, Tomiyama M, Kinoshita T, Shimizu KK (2021) An affordable image-analysis platform to accelerate stomatal phenotyping during microscopic observation. Front Plant Sci 12:715309

Tom N, Tom O, Malcikova J, Pavlova S, Kubesova B, Rausch T, Kolarik M, Benes V, Bystry V, Pospisilova S (2018) ToTem: a tool for variant calling pipeline optimization. BMC Bioinform 19(1):1–9

Utz HF, Melchinger AE (1996) PLABQTL: a program for composite interval mapping of QTL. J Quant Trait Loci 2(1):1–5

van Dijk ADJ, Kootstra G, Kruijer W, de Ridder D (2021b) Machine learning in plant science and plant breeding. iScience 24(1):101890

van Dijk M, Morley T, Rau ML, Saghai Y (2021a) A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050. Nat Food 2(7):494–501

Van Ooijen JW, Maliepaard CA (1999) MapQTL: version 3.0: Software for the calculation of QTL positions on genetic maps

Varshney RK, Bohra A, Yu J, Graner A, Zhang Q, Sorrells ME (2021a) Designing future crops: genomics-assisted breeding comes of age. Trends Plant Sci 26(6):631–649

Varshney RK, Roorkiwal M, Sun S, Bajaj P, Chitikineni A, Thudi M, Singh NP, Du X, Upadhyaya HD, Khan AW, Wang Y (2021b) A chickpea genetic variation map based on the sequencing of 3,366 genomes. Nat 599(7886):622–627

Varshney RK, Saxena RK, Upadhyaya HD, Khan AW, Yu Y, Kim C, Rathore A, Kim D, Kim J, An S, Kumar V (2017b) Whole-genome resequencing of 292 pigeonpea accessions identifies genomic regions associated with domestication and agronomic traits. Nat Genet 49(7): 1082–1088

Varshney RK, Shi C, Thudi M, Mariac C, Wallace J, Qi P, Zhang H, Zhao Y, Wang X, Rathore A, Srivastava RK (2017a) Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. Nat Biotechnol 35(10):969–976

Varshney RK, Sinha P, Singh VK, Kumar A, Zhang Q, Bennetzen JL (2020) 5Gs for crop genetic improvement. Curr Opin Plant Biol 56:190–196

Varshney RK, Thudi M, Nayak SN, Gaur PM, Kashiwagi J, Krishnamurthy L, Jaganathan D, Koppolu J, Bohra A, Tripathi S, Rathore A (2014) Genetic dissection of drought tolerance in chickpea (*Cicer arietinum* L.). Theor Appl Genet 127(2):445–462

Varshney RK, Thudi M, Pandey MK, Tardieu F, Ojiewo C, Vadez V, Whitbread AM, Siddique KH, Nguyen HT, Carberry PS, Bergvinson D (2018) Accelerating genetic gains in legumes for the development of prosperous smallholder agriculture: integrating genomics, phenotyping, systems modelling and agronomy. J Exp Bot 69(13):3293–3312

Varshney RK, Pandey MK, Bohra A, Singh VK, Thudi M, Saxena RK (2019) Toward the sequence-based breeding in legumes in the post-genome sequencing era. Theoretical and Applied Genetics 132(3):797–816

Villate A, San Nicolas M, Gallastegi M, Aulas PA, Olivares M, Usobiaga A, Etxebarria N, Aizpurua-Olaizola O (2021) Metabolomics as a prediction tool for plants performance under environmental stress. Plant Sci 303:110789

Voss-Fels K, Snowdon RJ (2016) Understanding and utilizing crop genome diversity via high-resolution genotyping. Plant Biotechnol J 14(4):1086–1094

Wang SCJB (2005) Windows QTL cartographer 2.5. http://statgen.Ncsu.Edu/qtlcart/WQTLCart.Htm

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, Mansueto L (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557(7703):43–49

Wang Y, Chen Q, Deng C, Zheng Y, Sun F (2020) KmerGO: a tool to identify group-specific sequences with k-mers. Front Microbiol 11:2067

Warren AS, Davis JJ, Wattam AR, Machi D, Setubal JC, Heath LS (2017) Panaconda: application of pan-synteny graph models to genome content analysis. bioRxiv 2:1–15

Wei ZG, Fan XG, Zhang H, Zhang XD, Liu F, Qian Y, Zhang SW (2022) kngMap: sensitive and fast mapping algorithm for noisy long reads based on the *k*-mer neighborhood graph. Front Genet 13:890651

Xu H, Xiao T, Chen CH, Li W, Meyer CA, Wu Q, Wu D, Cong L, Zhang F, Liu JS, Brown M (2015) Sequence determinants of improved CRISPR sgRNA design. Genome Res 25(8): 1147–1157

Xu J, Gu B, Tian G (2022a) Review of agricultural IoT technology. Artificial Intelligence in Agriculture 6:10–22

Xu Y, Zhang X, Li H, Zheng H, Zhang J, Olsen MS, Varshney RK, Prasanna BM, Qian Q (2022b) Smart breeding driven by big data, artificial intelligence and integrated genomic-enviromic prediction. Mol Plant:1–32

Yadav S, Sandhu N, Singh VK, Catolos M, Kumar A (2019) Genotyping-by-sequencing based QTL mapping for rice grain yield under reproductive stage drought stress tolerance. Sci Rep 9(1):1–12

Yan W (2001) GGE biplot-a windows application for graphical analysis of multienvironment trial data and other types of two-way data. Agron J 93(5):1111–1118

Yoshida H, Hirano K, Yano K, Wang F, Mori M, Kawamura M, Koketsu E, Hattori M, Ordonio RL, Huang P, Yamamoto E (2022) Genome-wide association study identifies a gene responsible for temperature-dependent rice germination. Nat Commun 13(1):1–13

Zakharov S, Wong TY, Aung T, Vithana EN, Khor CC, Salim A, Thalamuthu A (2013) Combined genotype and haplotype tests for region-based association studies. BMC Genomics 14(1):1–12

Zargar SM, Raatz B, Sonah H, Bhat JA, Dar ZA, Agrawal GK, Rakwal R (2015) Recent advances in molecular marker techniques: insight into QTL mapping, GWAS and genomic selection in plants. J Crop Sci Biotechnol 18(5):293–308

Zeng S, Skrabisova M, Lyu Z, Chan YO, Bilyeu K, Joshi T (2020) SNPViz v20: a web-based tool for enhanced haplotype analysis using large scale resequencing datasets and discovery of phenotypes causative gene using allelic variations. In: In 2020 IEEE international conference on bioinformatics and biomedicine, pp 1408–1415

Zhang F, Xue H, Dong X, Li M, Zheng X, Li Z, Xu J, Wang W, Wei C (2022) Long-read sequencing of 111 rice genomes reveals significantly larger pan-genomes. Genome Res 32(5): 853–863

Zhang XH, Tee LY, Wang XG, Huang QS, Yang SH (2015) Off-target effects in CRISPR/Cas9-mediated genome engineering. Molecular Therapy-Nucleic Acids 4:e264

Zhou Y, Chebotarov D, Kudrna D, Llaca V, Lee S, Rajasekar S, Mohammed N, Al-Bader N, Sobel-Sorenson C, Parakkal P, Arbelaez LJ (2020) A platinum standard pan-genome resource that represents the population structure of Asian rice. Scientific Data 7(1):1–11

Zhu FY, Song YC, Zhang KL, Chen X, Chen MX (2020) Quantifying plant dynamic proteomes by SWATH-based mass spectrometry. Trends Plant Sci 25(11):1171–1172