Journal of Genetic Engineering and Biotechnology

REVIEW Open Access

Bioinformatics approaches and applications in plant biotechnology

Yung Cheng Tan[1], Asqwin Uthaya Kumar[1,2], Ying Pei Wong[1] and Anna Pick Kiong Ling[1*]

Abstract

Background: In recent years, major advance in molecular biology and genomic technologies have led to an exponential growth in biological information. As the deluge of genomic information, there is a parallel growth in the demands of tools in the storage and management of data, and the development of software for analysis, visualization, modelling, and prediction of large data set.

Main body: Particularly in plant biotechnology, the amount of information has multiplied exponentially with a large number of databases available from many individual plant species. Efficient bioinformatics tools and methodologies are also developed to allow rapid genome sequence and the study of plant genome in the 'omics' approach. This review focuses on the various bioinformatic applications in plant biotechnology, and their advantages in improving the outcome in agriculture. The challenges or limitations faced in plant biotechnology in the aspect of bioinformatics approach that explained the low progression in plant genomics than in animal genomics are also reviewed and assessed.

Conclusion: There is a critical need for effective bioinformatic tools, which are able to provide longer reads with unbiased coverage in order to overcome the complexity of the plant's genome. The advancement in bioinformatics is not only beneficial to the field of plant biotechnology and agriculture sectors, but will also contribute enormously to the future of humanity.

Keywords: Bioinformatics, Biotic and abiotic, GWAS, NGS, Plant breeding, Plant sequencing, Plant pathogen, PRGdb sequence analysis

Background

Over the past decades, the term 'bioinformatics' has become a buzzword in all areas of research in biological science. With the continuous development and advancement in molecular biology, the explosive growth of biological information required a more organized, computerized system to collect, store, manage, and analyse the vast amount of biological data generated in the experiments from all fields [1]. Bioinformatics, as a new emerging interdisciplinary field for the past few decades, has many tools and techniques that are essential for efficient sorting and organizing of biological data into databases [1, 2]. Bioinformatics can be referred as a computer-based scientific field which applies mathematics, biology, and computer science to form into a single discipline for the analyses and interpretation of genomics and proteomics data [2, 3]. In short, the main components of bioinformatics are (a) the collection and analysis of database and (b) the development of software tools and algorithm as a tool for interpretation of biological data [2]. Bioinformatics played a crucial role in many areas of biology as its applications provide various types of data, including nucleotide and amino acid sequences, protein domains and structure as well as expression patterns from various organisms [3]. Similarly, the field of plant biotechnology has also taken advantages of bioinformatics, which

*Correspondence: anna_ling@imu.edu.my

[1] Division of Applied Biomedical Sciences and Biotechnology, School of Health Sciences, International Medical University, 126 Jalan Jalil Perkasa 19, Bukit Jalil, 57000 Kuala Lumpur, Malaysia
Full list of author information is available at the end of the article

provides full genomic information of various plant species to allow for efficient exploration into plants as biological resource to humans [1, 3, 4]. The intention of this article is to describe some of the key concepts, tools, and its applications in bioinformatics that are relevant to plant biotechnologies. The current challenges and limitations for improvement and continuous development of bioinformatics in plant science are also described.

Main text

Applications of bioinformatics in plant biotechnology

The introduction of bioinformatics and computational biology into the area of plant biology is drastically accelerating scientific invention in life science. With the aid of sequencing technology, scientists in plant biology have revealed the genetic architecture of various plant and microorganism species, such as proteome, transcriptome, metabolome, and even their metabolic pathway [1]. Sequence analysis is the most fundamental approach to obtain the whole genome sequence such as DNA, RNA, and protein sequence from an organism's genome in modern science. The sequencing of whole genome permits the determination of organization of different species and provides a starting point to understand their functionality. A complete sequence data consists of coding and non-coding regions, which can act as a necessary precursor for any functional gene that determines the unique traits possessed by organisms. The resulting sequence includes all regions such as exons, introns, regulator, and promoter, which often leads to a vastly large amount of genome information [5]. With the emergence of next-generation sequencing (NGS) and some other omics technologies used to examine plants genomics, more and more sequenced plants genome will be revealed [1, 6–8]. To deal with these vast amounts of data, the development and implementation of bioinformatics allow scientists to capture, store, and organize them in a systematic database [1, 5].

Bioinformatics databases and tools for plant biotechnology

In the field of bioinformatics, there are a variety of options of databases and tools that are available to perform analysis related to plant biotechnology. Next-generation sequencing (NGS) and bioinformatics analysis on the plant genomes over the years have generated a large amount of data. All these data are submitted to various and multiple databases that are publicly available online. Each database is unique and has its focus. For instance, CottonGen, database is solely dedicated to obtaining genomics and breeding information of any cotton species of interest [9]. The establishment of such database eases the researchers who are working on cotton genomic studies by focussing on using just one database instead of searching through other available databases. However, some databases are established and designed to cater not only to one specific species or genus, but focus on all the plant species, such as the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/) database, which as of 2021 possesses almost 21,000 plant genomes that are available for access [10]. Such a database is useful for studies that do not focus on one specific genus or species. This eases the researchers in accessing to all kinds of genomic data in one database. This section will briefly discuss some of the available plant genome databases, which are publicly accessible and not designated for one genus or species alone.

First would be the globally known and recognized database by all the researchers and biologists, which is the NCBI database. NCBI has been dedicated for gathering and analysing information about molecular biology, biochemistry, and genetics. In the NCBI database, one can download the genome information of the plant species of interest from either gene expression omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) or sequence read archive (SRA) (https://www.ncbi.nlm.nih.gov/sra) by simply stating the scientific name of the plant in the search bar and the entire genomic information of the plant can then be obtained. The GEO and SRA comprise processed or raw gene expression data or RNA sequencing of plants that are reposited in the repository. For instance, to obtain the genomics of *Rosa chinensis* (Rose plant), by inputting the name in the search bar, it will direct to the search result page where the researcher can select the most recent or suitable datasets with specific accession number. Depending on the profiling platform used in each dataset, researchers could retrieve either gene symbols, Ensemble ID, open reading frame, chromosomal location, regulatory elements, etc. The information allows researcher to further analyse the subject of study using bioinformatics tools such as gene ontology (http://geneontology.org/), Database for Annotation, Visualization and integration Discovery (DAVID) (https://david.ncifcrf.gov/), Basic Local Alignment Search Tool (BLAST) (https://blast.ncbi.nlm.nih.gov/Blast.cgi), and others that is relevant for the study.

Another database that is available for accessing plant genome database is EnsemblPlants (https://plants.ensembl.org/index.html). Unlike the NCBI database, which is not only dedicated to plant genomes, EnsemblPlants is specifically dedicated to accessing plant genomes. EnsemblPlant is part of the Ensembl project that started in 1999, where the project aimed to automatically annotate the genome and integrate the outcome of the annotation with other publicly available biological data and establish an open access archive or database online for the use of the research community [11]. Ensembl project later launched the taxonomic specific websites designated for each taxon under their project that also includes the plants. The database is a user-friendly integrative platform, where it is continuously updated with the new addition of plant species every time a plant genome is completely sequenced. Compared to the NCBI database mentioned earlier, EnsemblPlant not only provides genome sequence, gene models, and functional annotation of the plant species of interest, but also includes the polymorphic loci, population structure, genotype, linkage, and phenotype information [11, 12]. Unlike, NCBI, EnsemblPlant does also provide comparative genomics data of the plant species of interest. This indicates that the platform does not only offer genome sequence data but provide additional analytical data about the plant species of interest and help the researchers who are working on plant bioinformatics to save a lot of time by reducing the tedious work in running the analysis. Yet, the researchers could re-assess the data if necessary, depending on the stringency of their work.

Aside from the abovementioned databases that are widely used for retrieving plant genome sequence, there are still other plant databases such as PlantGDB, MaizeDIG, and Phytozome that can also be considered. Table 1 lists the available database and tools that are widely applied in plant biotechnology.

Biotechnology and bioinformatics for plant breeding

Plant breeding can be defined as the changing or improvement of desired traits in plants to produce improved new crop cultivars for the benefits of humankind [8]. Jhansi and Usha [13] mentioned a few benefits brought by genetically engineered plants such as improved quality, enhanced nutritional value, and maximized yield. The revolution of life science in molecular biology and genomics has enabled the leaps forward in plant breeding by applying the knowledge and biological data obtained in genomics research on crops [6, 8, 13]. In modern agriculture, transgenic technology on plants refers to genetic modification, which is done on plants or crops by altering or introducing foreign genes into the plant, to make them useful and productive and enhance their characteristic [13, 14]. As mentioned above, the evolution of next-generation sequencing (NGS) and other sequencing technologies produces a large size of biological data which require databases to store the information. The accessibility of whole genome sequences in databases allows free association across genomes with respect to gene sequence, putative function, or genetic map position. With the aid of software, it is possible to formulate predictive hypothesis and incorporate the desired phenotypes from a complex combination into plants by looking at those genetic markers which score well and gives a higher reliability in breeding [2, 15]. Other than genome sequence information, databases which store the information of metabolites also play a crucial role in the study of interaction with proteomics and genomics to reflect the changes in phenotype and specific function of an organism [1]. Some of the most widely used metabolomics databases for plants and crops such as Metlin (http://metlin.scripps.edu), provides multiple metabolite searching and about 240,000 metabolites, nearly 72,000 high-resolution MS/MS spectra, and PlantCyc (https://plantcyc.org/), a database which stores information about biochemical pathway and their catalytic enzyme and genes from plants [1, 16]. Moreover, single-nucleotide polymorphism markers also benefit from the revolution of NGS and other sequencing technologies. By using NGS, RNA sequencing (RNA-seq) allows direct measure of mRNA profile in order to identify known single-nucleotide polymorphism (SNP) [1]. SNP is the unique allelic variation within a genome of same species, which can be used as biological markers to locate the genes associated with desired traits in plants [17, 18]. Besides, transcriptome resequencing using NGS allows rapid and inexpensive SNP discovery within a large, complex gene with highly repetitive regions of a genome such as wheat, maize, sugarcane, avocado, and black currant [17]. Figure 1 illustrates briefly the process involved in plant breeding using NGS and bioinformatics.

Rice

Ever since the first transgenic rice production in 2000, there has been a significant revolution in crop genome sequencing projects, along with the advancement in technologies, rapidly increasing the pace in genetically modified organism (GMO) [2, 13, 19]. Among all the products in rice biotechnology, one of the most widely known GM rice is golden rice. Golden rice is a variety of rice engineered by introducing the biosynthetic pathway to produce β-carotene (pro-vitamin A) into staple food in order to resolve vitamin A deficiency. The World Health Organization has classified vitamin A deficiency as public health problem as it causes half a million of children to childhood blindness [13]. Vitamin A is an essential nutrient to humans as it helps with development of vision, growth, cellular differentiation, and proliferation of immune system; insufficient intake of vitamin A may lead to childhood blindness, anaemia, and reduced immune responsiveness against infection [20]. Being the first crop genome to be sequenced, rice has become the most suitable model to initiate the development and improvement of other species in genomic aspect [21–24]. The particular reason is due to its small genome size and diploidy, which enables rice to be an excellent model for other cereal crops with

Table 1 List of bioinformatics databases and tools applied in plant biotechnology

Function	Servers	URL	Details
Genome Database	ArrayExpress	www.ebi.ac.uk/arrayexpress	Archive for functional genomics data from microarray and sequencing platforms
	BarleyGenes	https://ics.hutton.ac.uk/barleyGenes/	Gene and RNA-seq database for Barley
	Chrysanthemum Transcriptome Database	http://www.icugi.org/chrysanthemum	Database of chrysanthemum's gene and transcripts
	Cottongen	https://www.cottongen.org/	Genomics, genetics, and breeding database for cotton
	Expression Atlas EMBL-EBI	www.ebi.ac.uk/gxa/home	Open resource platform for gene and protein expression information
	Ensembl Plants	https://plants.ensembl.org/index.html	Genomic database (genome sequence, gene models, functional annotation) of more than 33 plant species
	Gene Expression Omnibus	www.ncbi.nlm.nih.gov/projects/geo	Functional genomics data repository
	MaizeDIG	https://maizedig.maizegdb.org/	Genotypic-phenotypic database for maize
	MaizeGDB	https://www.maizegdb.org/	Genetic and genomics database for maize
	MaizeMine	https://maizemine.rnet.missouri.edu/maizemine/begin.domine/begin.do	Archive to access literature, genomic, interaction and proteomic data for maize
	VvGDB	www.plantgdb.org/VvGDB/	Database for grape genome
	Phytozome	https://phytozome-next.jgi.doe.gov/	Comparative genomics portal of plants
	Plant Promoter Database	http://linux1.softberry.com/berry.phtml?topic=plantprom&group=data&subgroup=plantprom	Database of plant promoter sequences and experimentally determined transcription start site of various plant species
	PLEXdb	http://www.plantgdb.org/prj/PLEXdb/	Plant expression database with gene expression profile data sets, structural genomics, and phenotypic data
	Pomamo	https://www.gabipd.org/projects/Pomamo/	Archive of potato sequences, literature, maps, and tools
	PRGdb	http://prgdb.org/prgdb/	Archive of pathogen receptor genes of various plant species
	Rice Expression Database	http://expression.ic4r.org/	Repository of gene expression profiles of rice from RNA-seq data on tissues spanning an entire range of rice growth
	Rice expression profile Database	http://ricexpro.dna.affrc.go.jp/	Repository of gene expression profiles of rice from microarray analysis
	SolGenomics Network	https://solgenomics.net/gem/experimental_design.pl?id=2	Database and tool of genomics and genetics approach for tomato and some other plant species such as eggplant
	TAIR	https://www.arabidopsis.org/	Database of genetic and molecular biology data for *Arabidopsis thaliana*
	The Rice Annotation Project Database	https://rapdb.dna.affrc.go.jp/	Archive of rice genome sequence, structure, and function
	Tomato Functional Genomics Database	http://ted.bti.cornell.edu	Archive of tomato microarray data, metabolite and RNA-seq
Metabolomic Database	Metabolights	www.ebi.ac.uk/metabolights/	Metabolomics database (cross-species, metabolite structures, biological role)

Table 1 (continued)

Function	Servers	URL	Details
Pathway Database	Reactome	www.reactome.org	Pathway interaction repository
	RiceCyc	http://pathway.gramene.org/gramene/ricecyc.shtml	Database of known and predicted biochemical pathways from rice
	PlantCyc	http://plantcyc.org	Metabolic pathway reference database from over 350 plant species
RNA Analysis Tool	qTeller	https://qteller.maizegdb.org/	Comparative RNA-seq expression tool
Chemical Compound Database	ChEBI	www.ebi.ac.uk/chebi/	Archive of molecular entities that focuses on small chemical compounds
	PubChem	http://pubchem.ncbi.nlm.nih.gov/	Archive of chemical information of various chemical compounds
Mass Spectrum Database	ReSpect for Phytochemicals	http://spectra.psc.riken.jp/menta.cgi/index	Database for phytochemicals MS spectra data and literature
	Metlin	http://metlin.scripps.edu/	Comprehensive MS/MS database
Resistance Analysis	Disease Resistance Analysis and Gene Orthology (DRAGO2)	http://prgdb.org/prgdb/drago2	Annotate resistance genes
Networking and Interaction Analysis	PathoPlant	www.pathoplant.de/	Database of plant-pathogen interaction and components related to plant pathogenesis
	AraNet	www.inetbio.org/aranet/	Probabilistic functional gene network tool for *Arabidopsis thaliana*
	PLANET	http://aranet.mpimp-golm.mpg.de/	Platform for visualization and analysis of co-function networks of photosynthetic organisms
Genomics Tool	The Bio-Analytic Resource	http://bar.utoronto.ca/	Plant bioinformatics tools resource platform (gene expression, mapping, molecular markers, and genomics)

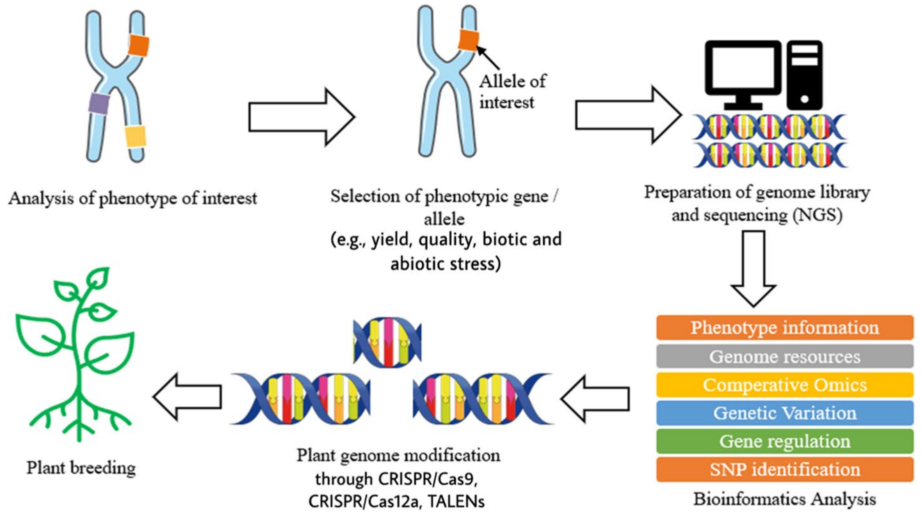
Allele of interest
Analysis of phenotype of interest
Selection of phenotypic gene / allele
(e.g., yield, quality, biotic and abiotic stress)
Preparation of genome library and sequencing (NGS)
Phenotype information
Genome resources
Comperative Omics
Genetic Variation
Gene regulation
SNP identification
Bioinformatics Analysis
Plant genome modification through CRISPR/Cas9, CRISPR/Cas12a, TALENs
Plant breeding

Fig. 1 Brief process of plant breeding involving NGS and bioinformatics

larger genomes, such as maize and wheat [21, 23]. Song et al. [22] reported the complete genome sequence of two rice subspecies, *japonica* and *indica*, in 2005 that laid a strong foundation for molecular studies and plant breeding research [22, 24]. With recent advancement in bioinformatics, it is now possible to run the sequence alignment between large and complex genome from other crop species with genomic data available from rice, by using different software or tools, in order to find out the shared conserved sequence through comparative genomics [2, 7]. Vassilev et al. stated some of the most commonly used programmes such as BLAST and FASTA format allowed rapid sequence searching in databases and give the best possible alignment to each sequence [25]. The programming algorithm calculates the alignment score to measure the proportion of homology matching residue between sequence from related species [2].

Wheat

Wheat, as the most widely grown consumed crops, together with rice and maize contributes more than 60% of the calories and protein for our daily life [26, 27]. To meet the demands of human population growth, it is necessary to achieve more understanding in wheat research and breeding in order to accelerate the production of wheat yield by 2050 [26–28]. Despite its importance, the improvement of wheat has been challenging as the researchers have to overcome the complexity of the wheat genome such as highly repetitive and large polyploid in order to get a fully sequenced reference genome [26, 29]. Advances in next-generation sequencing (NGS) platforms and other bioinformatics tools have revealed the extensive structural rearrangements and complex gene content in wheat, which revolutionized wheat genomics with the improvement of wheat yield and its adaptation to diversed environments [26, 29]. The NGS platforms allow the swift detection of DNA markers from the huge genome data in a short period of time. These NGS-based approaches have undoubtedly revolutionized the allele discovery and genotype-by-sequencing (GBS). By providing a high-quality reference genome of wheat in databases, it allows more sequence comparison between wheat and other species to find out more homologous gene. Moreover, the development of sequencing technologies in both high-throughput genotyping and read length, combining with biological databases, allow the rapid development of novel algorithm to complex wheat genome [29, 30]. For instance, genome-wide association studies (GWAS) are an approach used in genome research which allows rapid screening of raw data to select specific regions with agronomic traits [29, 31]. It allows multiple genetic variants across genome to be tested to study the genotype-phenotype association; thus, this method can be used to facilitate improvement in crop breeding via genomic selection and genetic modification [16, 29].

Maize

Maize, a globally important crop, not only has a wide variety of uses in terms of economic impact, but can also serve as genetic model species in genotype to phenotype relationship in plant genomic studies [32, 33]. Besides, due to its extremely high level of gene diversity, maize has high potential in the improvement of yield to meet the demands of population growth [33]. Despite the combination of economic and genomic impact, the progress in generating a whole genome sequence in maize has been a computational challenge due to the presence of tremendous structural variation (SV) in its genome [34]. The introduction of NGS techniques in several crops including maize allowed the rapid de novo genome sequencing and production of huge amount genomics and phenomics information [1, 35]. A better integration of data within multiple genome assemblies is much needed to study the connection between phenotype and genotype in order to achieve yield and quality improvement of maize [35]. Nowadays, some user-friendly online databases such as qTeller, MaizeDIG, and MaizeMine are designed to ease the comparison and visualization of relationships between genotypes and phenotypes [36]. MaizeGDB, a model organism database for maize, provides the access of data on genes, alleles, molecular markers, metabolic pathway information, phenotypic images with description, and more which are useful for maize research [35, 36]. MaizeMine is a data mining resource under MaizeGDB, which was designed to accelerate the genomics analysis by allowing the researchers to better script their own research data in downstream analysis [36] whereas MaizeDIG is a genotype-phenotype database which allows the users to link the association of genotype with phenotype expressed by image [35, 36]. Cho et al. [35] reported that with the accessibility via image search tool, the relationship between a gene and its phenotype features can be visualized within image. The integration and visualization of high-quality data with these tools enables quick prioritizing phenotype of interest in crops, which play a crucial role in the improvement of plant breeding.

Bioinformatics for studying stress resistance in plants

The understanding of the stress response on plants is vital for the improvement of breeding efforts in agriculture, and to predict the fate of natural plants under abiotic change especially in the current era of continuous climate change [37]. Stress response in plants can be divided into biotic and abiotic. Biotic stress mainly refers to negative influence caused by living organism such as virus, fungi, bacteria, insects, nematodes, and weeds [38] while abiotic stress refers to factors such as extreme temperature, drought, flood, salinity, and radiation which dramatically affect the crop yield [37]. NGS technologies and other potent computational tools, which allowed sequencing of whole genome and transcriptome, have led to the extensive studies of plants towards stress response on a molecular basis [1, 2, 37]. The tremendous amount of plant genome data obtained from genome sequencing allows the investigation of correlations between the molecular backbone of living organism and their adaptations towards the environment [16].

Biotic and abiotic stress management

How the plants and crops respond towards stress environment is the key to ensure their growth and development, and to avoid the great crop yield penalty caused by harsh condition [35, 39]. Therefore, the utilization of bioinformatic tools is important to study and analyse the plant transcriptome in response to biotic and abiotic stress. Besides, the application of bioinformatics tools on plants and crops genome can benefit the agricultural community by searching the desired gene among genome from different species and elucidate their function on the crops [35]. The genome databases play a crucial role in storing and mining large and complex genome sequence from the plants. Besides data storage, some genome databases are also able to perform gene expression profiling to predict the pattern of gene expressed at the level of transcript in cell or tissues. By using in silico genomic technologies, the disease resistance gene-enzyme with their respective transcription factor, which plays a role in defence mechanism against stress, are able to be identified [40, 41]. For instance, a large-scale transcriptome sequencing of chrysanthemum plants was carried out by Xu et al. [40] to study the dehydration stress in chrysanthemum plants. An online database called Chrysanthemum Transcriptome Database (http://www.icugi.org/chrysanthemum) was developed to allow the storage and distribution of transcriptome sequence and its analysis result among research community [40]. With the aid of different protein databases, the biochemical pathway and kinase activity of chrysanthemum in response to dehydration stress are able to be predicted [40]. Xu et al. [40] also reported a total of 306 transcription factor and 228 protein kinase that are important upstream regulator in plants when encountered with various biotic and abiotic stresses.

Bioinformatics approaches to study resistance to plant pathogen

One of the challenges in modern agriculture to supply the nutrition's demand along with the world population growth is the crop loss due to disease. The study of plant pathogen plays an essential role in the study of plant diseases, including pathogen identification, disease aetiology, disease resistance, and economic impact, among others [41]. Plants protect themselves through a complex defence system against variety of pathogen, including insects, bacteria, fungi, and viruses. Plant-pathogen interaction is a multicomponent system mediated by the detection of pathogen-derived molecules in the form of protein, sugar, and polysaccharide, by pattern recognition receptor (PRRs) within the plants [42–45]. After the recognition of enemy molecules, signal transduction is carried out accordingly and plant immune systems will respond defensively through different pathways involving different genes [42]. According to Schneider et al. [46], the development of molecular plant pathology can be broadly divided into three eras, begins with the disease physiology starting from early 1900s until 1980s [46]. In the second era of molecular plant genetic studies, one or a few genes of bacterial pathogens were focused whereas the third era of plant genomic studies began in 2000 with the sequencing of genome, and the first complete genome of bacterial pathogen, *Xylella fastidiosa*, was obtained [46]. The recent advance in DNA sequence technologies allow researchers to study the immune system of plants on genomic and transcriptomics level [1, 41, 42]. Genomics has revealed the mystery and complexity and consequently the various information about phytopathogen. A clearer picture of plant-pathogen interactions in the context of transcriptomic and proteomics can be visualized through the application of different bioinformatics tools, which in turn made feasible the engineering resistance to microbial pathogen in plant [43].

PRGdb: bioinformatics web for plant pathogen resistance gene analysis

Plants have developed a wide range of defence mechanism against different pathogen and ultimately inhibit growth and spread of pathogen [47, 48]. Plant defence system is mediated by resistance (R) gene [47]. R gene plays an important role in defence mechanism. They encode for protein that recognizes specific avirulent (Avr) pathogen proteins and initiated the defence mechanism through one or more signal transduction pathway in a hypersensitive response (HR) [41, 47, 48]. However, the essential components needed for protein to exert their resistance are still unidentified [48]. With the intention to study and identify more novel R gene, high-throughput genomic experiments and plant genomic sequence are essential to explore their function and new R gene discovery [47]. In 2009, Plant Disease Resistance Gene database (PRGdb), a comprehensive bioinformatics resource across hundreds of plant species, was launched in order to facilitate the plant genome research on discovery and predict plant disease resistance gene [47, 48]. To date, PRGdb 3.0 has been released with 153 reference resistance genes and 177,072 annotated candidate pathogen receptor genes (PRGs) [49]. This database act as an important reference site and repository to all the research studies on exploration and use of plant resistance genes [48, 49].

Apart from resistance gene storage, this easily accessible platform also allows different tools that are essential for exploration and discovery of novel R gene. For instance, the DRAGO 2.0 tool, which was built to explore known and novel disease resistance gene, can be launched on any transcriptome or proteome to annotate and predict PRG from DNA or amino acid with high accuracy [49]. Besides, BLAST search tools available in PRGdb provide comparison of different sequences which allowed the determination of gene homology and expression analysis. Apart from the database, plant pathology field also benefited from whole genome sequence technologies. The new DNA sequencing technologies such as NGS and Sanger sequencing allowed the study of genomics, proteomics, metabolomics, and transcriptomics on both the host plant and the pathogen [1]. The phytopathogen genomes which have been sequenced are expected to provide valuable information on the molecular basis for infection of plant host and explore the potential novel virulence factors [1]. Figure 2 illustrates a brief process involved in producing stress-resistant plant using bioinformatics approach.

Metagenomics in plant biotechnology and Cas9 modification

The effects of environment microorganisms' community, especially soil microorganism on plants, may contribute to plant's growth and pathogenesis. Through metagenomics approaches, the soil microorganism community that contributed to plant growth may provide a great genomic insight into physiology and pathology [50–53]. In metagenomics approaches, the overall genetic materials obtained from soil are sequenced and advancing to microbial community analysis via data analytics [53–55]. The extracted genetic materials from the soil were subjected to high-throughput metagenomics analysis via various NGS approaches such as 16S rRNA sequencing, shotgun metagenomic sequencing, MiSeq sequencing [54–56] for microbial species identification, functional genomics study, and structural metagenomic analysis. A NGS produces huge genomics data for each study; thus, application of bioinformatics tools would add value in the metagenomics analysis as the target genes identified could advance into elucidation of plant growth, plant disease, soil contamination, and microbial taxonomy [52]. For example, the use of UNITE (https://unite.ut.ee/) for fungi identification [57], SILVA (https://www.arb-silva.de/) for 16S rRNA [58], and MGnify (https://www.ebi.ac.uk/metagenomics/) possesses metagenomics data of microbiome [59]. These databases allow the researchers to retrieve and analyse the relevant metagenomic sequenced data for a specific study.

Since metagenomics analysis provides the greater output on plant-microbe interaction, the genes that are responsible for plant immunity may play a crucial role in protecting against disease-causing microorganism [60, 61]. With the emergence of Clustered Regularly Interspaced Short Palindrome Repeats (CRISPR) gene editing technique, Cas9 modification could produce a better plant trait and disease-resistant plant [62, 63]. The CRISPR/Cas9 system is employed in studying the

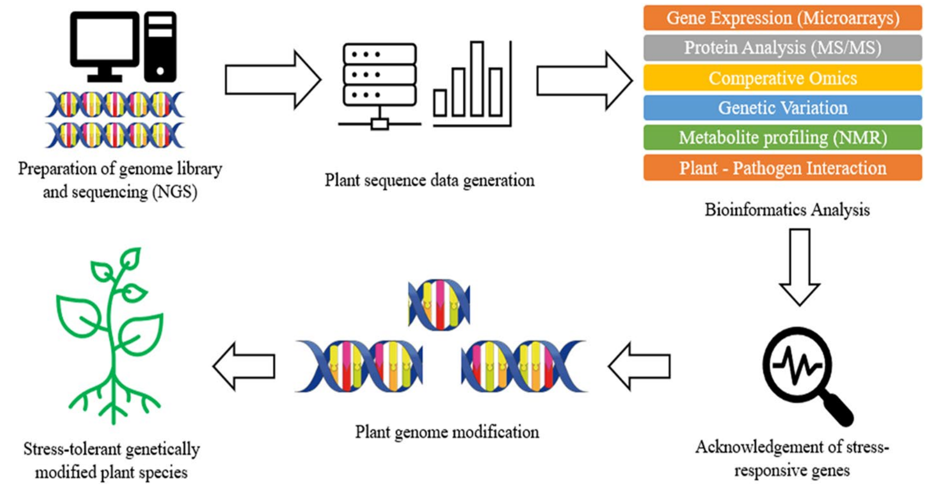
Preparation of genome library and sequencing (NGS)
Plant sequence data generation
Gene Expression (Microarrays)
Protein Analysis (MS/MS)
Comperative Omics
Genetic Variation
Metabolite profiling (NMR)
Plant - Pathogen Interaction
Bioinformatics Analysis
Acknowledgement of stress-responsive genes
Plant genome modification
Stress-tolerant genetically modified plant species

Fig. 2 Brief process involved in producing stress-resistant plant using bioinformatics approach

functional genomics in plants in relation to plant-microbe interaction. CRISPR/Cas9 system facilitated the gene editing by creating a mutant through double-stranded break forming a targeted gene mutation and followed by genome repair [63–65]. The CRISPR/Cas9 modification on *OsSWEET14* genes protects the Super Basmati Rice from bacterial blight causes by *Xanthomonas oryzae pv. oryzae* [66]. Gene editing to knockout *OsMPK5* and *OsERF922* genes in rice protects against *Magnaporthe grisea* and *Magnaporthe oryzae*, respectively [67–69]. Besides that, Cas9 modification on Cs*WRKY22* and *TcNPR3* increased host defence immunity through regulating salicylic acid in *Citrus sinensis* and *Theobroma cacao*, respectively [70, 71]. Thus, CRISPR/Cas9 modification could be one of important science advancements to validate the metagenomics analysis on plant-microbe interaction.

Current challenges of bioinformatics applications in plant biotechnology

Despite the beneficial prospect of the bioinformatics applied in plant biotechnology, there are many challenges and limitations must be addressed in order to fully utilize their potentials [1]. Along with the rapid growth in plant genome data mining and database development, there are a few challenges faced by bioinformaticians and scientists which can be divided into number of areas as mentioned in the subsections below.

Bioinformatic data management and organization and synchronize update resources

Since the introduction of the next-generation sequencing (NGS), which is commercially available in 2004, enormous amount of data has been generated in plant genome research. Thousands of Gb of plants sequences are deposited in various public databases monthly [1, 72, 73]. Moreover, the constantly sequenced and re-sequenced of the plant genome has developed a vast amount of new genome sequence in all public databases. The increase in sequenced plant genome driven by technological improvement has led to a problem that arises along with the storage and update of a large amount of data [72, 74]. The update process should occur in all the comparative databases, not just solely individual genome database [72]. With this, the synchronized update of genome data resources among different plant genomic platform is able to provide a strong, updated, reliable database community that all the plant researchers can rely on [72].

Complexity of plant genetic content

Other than the tremendous amount of genome sequence generated, the complexity of the plant genetic content is also a challenging issue faced by plant research community. Even though the arrival of next-generation sequencing technologies has allowed the rapid DNA sequencing for non-model or orphan plant species, the sequencing pace for plants is far from that of animal and microorganism [74]. The main factor which contributes to this situation is because sometimes the plant genome can be nearly hundred times larger than the currently sequenced animal and microorganism genome [73]. Needless to say, some of the plant genome even can have polyploidy, a duplication of an entire genome, which is estimated to occur in 80% of the plant species [73, 75]. According to Schatz et al., the genome assembly in the case of large size plant genome with abundance of repetitive sequence can be metaphorically described as build-up of a large puzzle consisting of blue sky separated by nearly indistinguishable white clouds of small gene [73]. The particular reason for this is mainly because the sequence length in NGS is relatively shorter than in Sanger sequencing and required dedicated assembly algorithm [74]. Therefore, most plant genomes sequenced by NGS can only be used for establishing gene catalogues, interpreting the repeat content, glimpsing evolutionary mechanism, and performing on comparative genomics in early study [74].

Advance in sequencing technologies

There are two basic approaches to genome assembly, i.e. comparative genome assembly and de novo genome assembly [75]. It is important to distinguish between these two different approaches. Comparative is a reference-guided method which use a genome or transcriptome, or both, for guidance, whereas de novo assembly refers to reconstruction of a genome from organisms that have not been sequenced before [74, 75]. Table 2 compares some of the available assembly and NGS technology available for genome sequencing. However, these two approaches are not completely exclusive due to a lack of bioinformatic tools designed to cope with the unique and challenging features of plant genomes [74, 75]. One of the biggest challenges in the development of bioinformatic software is the algorithm development [76]. As is known, all the programmes or software in bioinformatic are very computationally intensive. As most of the assemblies available now solely rely on single assembly, a development in better algorithm in terms of resource requirement is essential for combining different assemblers by using a different underlying algorithm in order to give a more credible final assembly [74, 76].

Database accessibility

To date, there are about 374,000 known plant species in the world [77]. The first full plant genome sequencing was completed on A*rabidopsis thaliana* through Sanger

Table 2 Comparison between next-generation sequencing technologies

Method	Illumina	Pacific Bio	Nanopore	Pyrosequencing (454)	SOLiD
Read length per run	50–300 base pair	10–25 kilo base pair	500–2.3 mega base pair	Approximate 800 base pair	50 base pair
Time taken per run	1 to 10 days	Up to 30 h	1 min–72 h	24 h	1 to 2 weeks
Cost	$148 per Gb	$2000 Gb	$60–80 per sample	$7000 per sample	$15,000 per 100 Gb
Accuracy	98%	99.9%	98.9–99.6%	99.9%	99.9%
Advantages	Cost-effective, high-yield sequence reads	Fast, long read lengths	Real-time analysis, long read lengths	Fast, long read lengths	High accuracy
Disadvantages	Instrument cost, high maintenance of instrument, read length	Low high throughput	Error prone	Homopolymer error	Long run time, low read length

sequencing methods in 2000 [78]. Although introduction of molecular biology decades ago may have facilitated the species identification, obtaining the full plant genomic data remains challenging due to the genome complexity. The development of NGS platform may foster the plant genome sequencing, yet there are limited sequenced datasets reposited to the database. To date, there are only 29 plant genome databases accessible in PlantGDB genome browser allowing researchers to retrieve the information about gene structure, matched GSS contigs, similar protein, spliced alignments EST, etc. Besides, the PlaD database (http://systbio.cau.edu.cn/plad/index.php) that focuses on the microarray data of the plants developed by China Agricultural University comprises transcriptomic database for plant defence against pathogen. However, it is limited to *Arabidopsis*, rice, maize, and wheat [79]. The Plant Omics Data Center (http://plantomics.mind.meiji.ac.jp/podc/) is another publicly available web-based plant database featuring omics data for co-expressed profile, regulatory network, and plant ontology information [80]. Although curated omics datasets could be retrieved from PODC, information are restricted for certain plants and crops such as *Arabidopsis*, tobacco, earthmoss, barrelclover, soybean, potato, rice, tomato, grape, maize, and sorghum. Furthermore, all these publicly available databases require constant updating with new released data or resequencing data so that the researcher could obtain the most updated version of genome datasets for their research.

Conclusion

The application of bioinformatics in plant biotechnology represents a fundamental shift in the way scientists study living organisms. Bioinformatics play a significant role in the development of agriculture sector as it helps to study the stress resistance and plant pathogen, which are critical in advancing crop breeding [75]. NGS and other sequencing technologies will make more plant genome data accessible in all public databases and enable the identification of genomic variants and prediction of protein structure and function [75, 76]. Moreover, GWAS, which allows the identification of loci and allelic variation related to valuable traits, eased the crop modification and improvement [74]. In brief, the advance in bioinformatics application in plant biotechnology enables researchers to achieve fundamental and systematic understanding of economically important plant. However, despite all these exciting achievement by the application of bioinformatic on plant biotechnology, it is still a long way from automated full genome sequencing and assembly at a low cost [76]. There is a critical need for effective bioinformatic tools which are able to provide longer reads with unbiased coverage in order to overcome the complexity of the plant's genome. To achieve this, an enhanced algorithm development is essential to enable data mining and analysis, comparison, and so on. Therefore, bioinformaticians and experts with mathematical and programming skills will play an important role in bringing fresh approaches and knowledge into bioinformatics, not only for the advancement in plant biotechnology and agriculture sector, but the future of humanity as well.

Abbreviations

GWAS: Genome-wide association studies; NGS: Next-generation sequencing; PRGdb: Plant Disease Resistance Gene database; RNA-seq: RNA sequencing; SNP: Single-nucleotide polymorphism.

Acknowledgements

The authors wish to thank Prof. Hoe I. Ling of Columbia University (New York, USA) for his editorial input and proofread the manuscript.

Authors' contributions

YCT designed the content and was a major contributor in writing the manuscript. AUK and YPW edited the manuscript. APKL designed and edited the manuscript. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Division of Applied Biomedical Sciences and Biotechnology, School of Health Sciences, International Medical University, 126 Jalan Jalil Perkasa 19, Bukit Jalil, 57000 Kuala Lumpur, Malaysia. [2]School of Biosciences and Biotechnology, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Malaysia.

Received: 30 November 2021 Accepted: 5 July 2022
Published online: 15 July 2022

References

1. Gomez-Casati DF, Busi MV, Barchiesi J, Peralta DA, Hedin N, Bhadauria V (2018) Applications of bioinformatics to plant biotechnology. Curr Issues Mol Biol 27:89–104. https://doi.org/10.21775/cimb.027.089
2. Zhang SY, Liu SL (2013) Bioinformatics. In: Maloy S, Hughes K (eds) Brenner's Encyclopedia of Genetics, 2nd edn. Academic Press, London. https://doi.org/10.1016/B978-0-12-374984-0.00155-8
3. Tiwari A, Singh P, Kumawat S (2020) Applications of bioinformatics in plant breeding system. Int J Curr Microbial App Sci. 11:2825–2831
4. Rhee SY, Dickerson J, Xu D (2006) Bioinformatics and its applications in plant biology. Annu Rev Plant Biol 57:335–360. https://doi.org/10.1146/annurev.arplant.56.032604.144103
5. Normand EA, Van den Veyyer IB (2019) Next-generation sequencing for gene panels and clinical exomes. In: Leung PCK, Qiao J (eds) Human Reproductive and Prenatal Genetics, 1st edn. Academic Press, London. https://doi.org/10.1016/B978-0-12-813570-9.00025-5
6. Blätke MA, Szymanski JJ, Gladilin E, Scholz U, Beier S (2021) Editorial: advances in applied bioinformatics in crops. Front Plant Sci 12:640394. https://doi.org/10.3389/fpls.2021.640394
7. Kushwaha UKS, Deo I, Jaiswal JP, Prasad B (2017) Role of bioinformatics in crop improvement. Glob J Sci Front Res D Agric Vet 17(1):13–23
8. Caligari PDS, Brown J (2017) Plant Breeding, Practice. In: Thomas B, Murray BG, Murphy DJ (eds) Encyclopedia of Applied Plant Sciences, 2nd edn. Academic Press, London. https://doi.org/10.1016/B978-0-12-394807-6.00195-7
9. Yu J, Jung S, Cheng CH, Lee T, Zheng P, Buble K et al (2021) CottonGen: the community database for cotton genomics, genetics, and breeding research. Plants. 10(12):2805. https://doi.org/10.3390/plants10122805
10. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC et al (2022) Database resources of the national center for biotechnology information. Nucleic Acids Res 50(D1):D20–D26. https://doi.org/10.1093/nar/gkab1112
11. Howe KL, Contreras-Moreira B, De Silva N, Maslen G, Akanni W, Allen J et al (2019) Ensembl Genomes 2020 – enabling non-vertebrate genomic research. Nucleic Acids Res 48(D1):D689–D695. https://doi.org/10.1093/nar/gkz890
12. Bolser D, Staines DM, Pritchard E, Kersey P (2016) Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomics data. In: Edwards D (ed) Plant Bioinformatics. Methods in Molecular Biology, vol 1374. Humana Press. https://doi.org/10.1007/978-1-4939-3167-5_6
13. Jhansi Rani S, Usha R (2013) Transgenic plants: Types, benefits, public concerns and future. J Pharm Res 6(8):879–883. https://doi.org/10.1016/j.jopr.2013.08.008
14. Barragán-Ocaña A, Reyes-Ruiz G, Olmos-Peña S, Gómez-Viquez H (2019) Transgenic crops: trends and dynamics in the world and in Latin America. Transgenic Res 28(3-4):391–399. https://doi.org/10.1007/s11248-019-00123-8
15. Platten JD, Cobb JN, Zantua RE (2019) Criteria for evaluating molecular markers: Comprehensive quality metrics to improve marker-assisted selection. PLoS One 14(1):e0210529. https://doi.org/10.1371/journal.pone.0210529
16. Filho HA, Machicao J, Bruno OM (2018) A hierarchical model of metabolic machinery based on the kcore decomposition of plant metabolic networks. PLoS One 13(5):e0195843. https://doi.org/10.1371/journal.pone.0195843
17. Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla S (2012) SNP markers and their impact on plant breeding. Int J Plant Genomics 728398:1–11. https://doi.org/10.1155/2012/728398
18. Hoskins RA, Phan AC, Naeemuddin M, Mapa FA, Ruddy DA, Ryan JJ et al (2001) Single nucleotide polymorphism markers for genetics mapping in *Drosophila melanogaster*. Genome Res 11(6):1100–1113. https://doi.org/10.1101/gr.gr-1780r
19. Edwards D, Batley J (2010) Plant genome sequencing: applications for crop improvement. Plant Biotechnol J 8(1):2–9. https://doi.org/10.1111/j.1467-7652.2009.00459.x
20. Tang G, Qin J, Dolnikowski GG, Russell RM, Grusak MA (2009) Golden Rice is an effective source of vitamin A. Am J Clin Nutr 89(6):1776–1783. https://doi.org/10.3945/ajcn.2008.27119
21. Yu J, Hu S, Wang J, Wong GKS, Li S, Liu B et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *Indica*). Science. 296(5565):79–92. https://doi.org/10.1126/science.1068037
22. Song S, Tian D, Zhang Z, Hu S, Yu J (2018) Rice genomics: over the past two decades and into the future. Genomics Proteomics Bioinformatics 16(6):397–404. https://doi.org/10.1016/j.gpb.2019.01.001
23. Jackson SA (2016) Rice: The First Crop Genome. Rice. 9(14). https://doi.org/10.1186/s12284-016-0087-4
24. Jain R, Jenkins J, Shu S, Chern M, Martin JA, Copetti D et al (2019) Genome sequence of the model rice variety KitaakeX. BMC Genomics 20(905). https://doi.org/10.1186/s12864-019-6262-4
25. Vassilev D, Leunissen J, Atanassov A, Nenov A, Dimov G (2005) Application of bioinformatics in plant breeding. Biotechnol Biotechnol Equip 19(sup3):139–152. https://doi.org/10.1080/13102818.2005.10817293
26. Walkowiak S, Gao L, Monat C, Haberer G, Kassa MT, Brinton J et al (2020) Multiple wheat genomes reveal global variation in modern breeding. Nature. 588(7837):277–283. https://doi.org/10.1038/s41586-020-2961-x
27. Appels R, Eversole K, Stein N, Feuillet C, Keller B, Rogers J et al (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science. 361(6403). https://doi.org/10.1126/science.aar7191
28. Gill BS, Appels R, Borta-Oberholster AM, Buell CR, Bennetzen JL, Chalhoub B et al (2004) A workshop report on wheat genome sequencing: International Genome Research on Wheat Consortium. Genetics. 168(2):1087–1096. https://doi.org/10.1534/genetics.104.034769
29. Babu P, Baranwal DK, Harikrishna PD, Bharti H, Joshi P et al (2020) Application of genomics tools in wheat breeding to attain durable rust resistance. Front Plant Sci 11:567147. https://doi.org/10.3389/fpls.2020.567147
30. Guan J, Garcia DF, Zhou Y, Appels R, Li A, Mao L (2020) The battle to sequence the bread wheat genome: a tale of the three kingdoms. Genomics Proteomics Bioinformatics 18(3):221–229. https://doi.org/10.1016/j.gpb.2019.09.005
31. Bolser D, Staines DM, Pritchard E, Kersey P (2016) Ensembl plants: integrating tools for visualizing, mining and analyzing plant genomics data. Methods Mol Biol 1374:115–140. https://doi.org/10.1007/978-1-4939-3167-5_6
32. Haberer G, Young S, Bharti AK, Gundlach H, Raymond C, Fuks G et al (2005) Structure and architecture of the maize genome. Plant Physiol 139(4):1612–1624. https://doi.org/10.1104/pp.105.068718
33. Li C, Song W, Luo Y, Gao S, Zhang R, Shi Z et al (2019) The HuangZaoSi maize genome provides insights into genomic variation and improvement history of maize. Mol Plant 12(3):402–409. https://doi.org/10.1016/j.molp.2019.02.009
34. Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, Wang T et al (2015) High-resolution genetic mapping of maize pan-genome sequence anchors. Nat Commun 6:6914. https://doi.org/10.1038/ncomms7914
35. Cho KT, Portwood JL, Gardiner JM, Harper LC, Lawrence-Dill CJ, Friedberg I et al (2019) MaizeDIG: maize database of images and genomes. Front Plant Sci 10:1050. https://doi.org/10.3389/fpls.2019.01050

36. Portwood JL, Woodhouse MR, Cannon EK, Gardiner JM, Harper LC, Schaeffer ML et al (2018) MaizeGDB 2018: the maize multi-genome genetics and genomics database. Nucleic Acids Res 47(D1):D1146–D1154. https://doi.org/10.1093/nar/gky1046
37. Ambrosino L, Colantuono C, Diretto G, Fiore A, Chiusano ML (2020) Bioinformatics resources for plant abiotic stress responses: state of the art and opportunities in the fast evolving -omics era. Plants. 9(5):591. https://doi.org/10.3390/plants9050591
38. Singla J, Krattinger SG (2016) Biotic stress resistance genes in wheat. Reference Module in Food Science. https://doi.org/10.1016/B978-0-08-100596-5.00229-8
39. Costa MCD, Farrant JM (2019) Plant resistance to abiotic stresses. Plants (Basel) 8(12):553. https://doi.org/10.3390/plants8120553
40. Xu Y, Gao S, Yang Y, Huang M, Cheng L, Wei Q et al (2013) Transcriptome sequencing and whole genome expression profiling of chrysanthemum under dehydration stress. BMC Genomics 14:662. https://doi.org/10.1186/1471-2164-14-662
41. Nishad R, Ahmed T, Rahman VJ, Kareem A (2020) Modulation of plant defense system in response to microbial interactions. Front Microbiol 11:1298. https://doi.org/10.3389/fmicb.2020.01298
42. Andersen EJ, Ali S, Byamukama E, Yen Y, Nepal MP (2018) Disease resistance mechanisms in plants. Genes (Basel) 9(7):339. https://doi.org/10.3390/genes9070339
43. Dong OX, Ronald PC (2019) Genetic engineering for disease resistance in plants: recent progress and future perspectives. Plant Physiol 180(1):26–38. https://doi.org/10.1104/pp.18.01224
44. Abdulkhair WM, Alghuthaymi MA (2016) Plant pathogens. In: Rigobelo EC (ed) Plant Growth, 1st edn. InTechOpen. https://doi.org/10.5772/65325 Available from: https://www.intechopen.com/chapters/52387
45. Gupta R, Lee SE, Agrawal GK, Rakwal R, Sangryeol P, Wang Y et al (2015) Understanding the plant-pathogen interactions in the context of proteomics-generated apoplastic proteins inventory. Front Plant Sci 6:352. https://doi.org/10.3389/fpls.2015.00352
46. Schneider DJ, Collmer A (2010) Studying plant-pathogen interactions in the genomics era: beyond Molecular Koch's postulates to systems biology. Annu Rev Phytopathol 48:457–479. https://doi.org/10.1146/annurev-phyto-073009-114411
47. Sanseverino W, Hermoso A, D'Alessandro R, Vlasova A, Andolfo G, Frusciante L et al (2013) PRGdb 2.0: towards a community-based database model for the analysis of R-genes in plants. Nucleic Acids Res 41(Database Issue):D1167–D1171. https://doi.org/10.1093/nar/gks1183
48. Sanseverino W, Roma G, Simone MD, Faino L, Melito S, Stupka E et al (2010) PRGdb: a bioinformatics platform for plant resistance gene analysis. Nucleic Acids Res 38(Database Issue):D814–D821. https://doi.org/10.1093/nar/gkp978
49. Osuna-Cruz CM, Paytuvi-Gallart A, Donato AD, Sundesha V, Andolfo G, Cigliano RA et al (2018) PRGdb 3.0: a comprehensive platform for prediction and analysis of plant disease resistance genes. Nucleic Acids Res 46(D1):D1197–D1201. https://doi.org/10.1093/nar/gkx1119
50. Hily JM, Demanèche S, Poulicard N, Tannières M, Djennane S, Beuve M et al (2018) Metagenomic-based impact study of transgenic grapevine rootstock on its associated virome and soil bacteriome. Plant Biotechnol J 16(1):208–220. https://doi.org/10.1111/pbi.12761
51. Fadiji AE, Babalola OO (2020) Metagenomics methods for the study of plant-associated microbial communities: a review. J Microbiol Methods 70:105860. https://doi.org/10.1016/j.mimet.2020.105860
52. Piombo E, Abdelfattah A, Droby S, Wisniewski M, Spadaro D, Schena L (2021) Metagenomics approaches for the detection and surveillance of emerging and recurrent plant pathogens. Microorganisms. 9(1):188. https://doi.org/10.3390/microorganisms9010188
53. Chaudhary P, Khati P, Chaudhary A, Maithani D, Kumar G, Sharma A (2021) Cultivable and metagenomic approach to study the combined impact of nanogypsum and *Pseudomonas taiwanensis* on maize plant health and its rhizospheric microbiome. PLoS One 16(4):e0250574. https://doi.org/10.1371/journal.pone.0250574
54. Chukwuneme CF, Ayangbenro AS, Babalola OO (2021) Metagenomic analyses of plant growth-promoting and carbon-cycling genes in maize rhizosphere soils with distinct land-use and management histories. Genes (Basel) 12(9):1431. https://doi.org/10.3390/genes12091431
55. Zhao J, Ma J, Yang Y, Yu H, Zhang S, Chen F (2021) Response of soil microbial community to vegetation reconstruction modes in mining areas of the Loess Plateau, China. Front Microbiol 12:714967. https://doi.org/10.3389/fmicb.2021.714967
56. Babalola OO, Fadiji AE, Ayangbenro AS (2020) Shotgun metagenomic data of root endophytic microbiome of maize (*Zea mays* L.). Data Brief 31(105893). https://doi.org/10.1016/j.dib.2020.105893
57. Nilsson RH, Larsson KH, Taylor AFS, Bengtsson-Palme J, Jeppesen TS, Schigel D et al (2019) The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucleic Acids Res 47(D1):D259–D264. https://doi.org/10.1093/nar/gky1022
58. Quast C, Pruesse E, Yilmaz P et al (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41(Database issue):D590–D596. https://doi.org/10.1093/nar/gks1219
59. Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G et al (2020) MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res 48(D1):D570–D578. https://doi.org/10.1093/nar/gkz1035
60. Musidlak O, Buchwald W, Nawrot R (2014) Plant defense responses against viral and bacterial pathogen infections. Focus on RNA-binding proteins (RBPs). Herba Polonica 60:60–73. https://doi.org/10.1515/hepo-2015-0005
61. Silva MS, Arraes FBM, Campos MDA, Grossi-de-Sa M, Fernandez D, Cândido EDS et al (2018) Review: potential biotechnological assets related to plant immunity modulation applicable in engineering disease-resistant crops. Plant Sci 270:72–84. https://doi.org/10.1016/j.plantsci.2018.02.013
62. Feng Z, Zhang B, Ding W, Liu X, Yang DL, Wei P et al (2013) Efficient genome editing in plants using a CRISPR/Cas system. Cell Res 23(10):1229–1232. https://doi.org/10.1038/cr.2013.114
63. Wada N, Ueta R, Osakabe Y, Osakabe K (2020) Precision genome editing in plants: state-of-the-art in CRISPR/Cas9-based genome engineering. BMC Plant Biol 20:234. https://doi.org/10.1186/s12870-020-02385-5
64. Nekrasov V, Staskawicz B, Weigel D, Jones JD, Kamoun S (2013) Targeted mutagenesis in the model plant Nicotiana benthamiana using Cas9 RNA-guided endonuclease. Nat Biotechnol 31(8):691–693. https://doi.org/10.1038/nbt.2655
65. Langner T, Kamoun S, Belhaj K (2018) CRISPR crops: plant genome editing toward disease resistance. Annu Rev Phytopathol 56:479–512. https://doi.org/10.1146/annurev-phyto-080417-050158
66. Zafar K, Khan MZ, Amin I, Mukhtar Z, Yasmin S, Arif M et al (2020) Precise CRISPR-Cas9 mediated genome editing in super basmati rice for resistance against bacterial blight by targeting the major susceptibility gene. Front Plant Sci 11:575. https://doi.org/10.3389/fpls.2020.00575
67. Xie K, Yang Y (2013) RNA-guided genome editing in plants using a CRISPR-Cas system. Mol Plant 6(6):1975–1983. https://doi.org/10.1093/mp/sst119
68. Wang F, Wang C, Liu P, Lei C, Hao W, Gao Y et al (2016) Enhanced rice blast resistance by CRISPR/Cas9-targeted mutagenesis of the ERF transcription factor gene OsERF922. PLoS One 11(4):e0154027. https://doi.org/10.1371/journal.pone.0154027
69. Oliva R, Ji C, Atienza-Grande G, Huguet-Tapia JC, Perez-Quintero A, Li T et al (2019) Broad-spectrum resistance to bacterial blight in rice using genome editing. Nat Biotechnol 37(11):1344–1350. https://doi.org/10.1038/s41587-019-0267-z
70. Wang L, Chen S, Peng A, Xie Z, He Y, Zou X (2019) CRISPR/CAS9 -mediated editing of CsWRKY22 reduces susceptibility to *Xanthomonas citri* subsp. citri in Wanjincheng orange (*Citrus sinensis* (L.) Osbeck). Plant Biotechnol Rep 13(5):501–510. https://doi.org/10.1007/s11816-019-00556-x
71. Fister AS, Landherr L, Maximova SN, Guiltinan MJ (2018) Transient expression of CRISPR/Cas9 machinery targeting TcNPR3 Enhances defense response in theobroma cacao. Front Plant Sci 9:268. https://doi.org/10.3389/fpls.2018.00268
72. Ong Q, Nguyen P, Thao NP, Le L (2016) Bioinformatics approach in plant genomic research. Curr Genomics 17(4):368–378. https://doi.org/10.2174/1389202917666160331202956
73. Schatz MC, Witkowski J, McCombie WR (2012) Current challenges in *de novo* plant genome sequencing and assembly. Genome Biol 13(4):243. https://doi.org/10.1186/gb-2012-13-4-243
74. Claros MG, Bautista R, Guerrero-Fernández D, Benzerki H, Seoane P, Fernández-Pozo N (2012) Why assembling plant genome sequences is so challenging. Biology (Basel) 1(2):439–459. https://doi.org/10.3390/biology1020439
75. Kyriakidou M, Tai HH, Anglin NL, Ellis D, Strömvik MV (2018) Current strategies of polyploid plant genome sequence assembly. Front Plant Sci 9:1660. https://doi.org/10.3389/fpls.2018.01660

76. Mathur M (2018) Bioinformatics challenges: a review. Int J Adv Sci Res 3(6):29–33
77. Fazan L, Song YG, Kozlowski G (2020) The woody planet: from past triumph to manmade decline. Plants (Basel) 9(11):1593. https://doi.org/10.3390/plants9111593
78. Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 408(6814):796–815. https://doi.org/10.1038/35048692
79. Qi H, Jiang Z, Zhang K, Yang S, He F, Zhang Z (2018) PlaD: a transcriptomics database for plant defense responses to pathogens, providing new insights into plant immune system. Genomics Proteomics Bioinformatics 16(4):283–293. https://doi.org/10.1016/j.gpb.2018.08.002
80. Ohyanagi H, Takano T, Terashima S, Kobayashi M, Kanno M, Morimoto K et al (2015) Plant Omics Data Center: an integrated web repository for interspecies gene expression networks with NLP-based curation. Plant Cell Physiol 56(1):e9. https://doi.org/10.1093/pcp/pcu188

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.