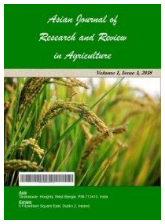

Asian Journal of Research and Review in Agriculture

3(3): 5-14, 2021; Article no.AJRRA.453

Bioinformatics Approaches toward Plant Breeding Programs

Marina A. Ibrahim[1], Marina A. Shehata[1], Nancy N. Nasry[1], Mariam S. Fayez[1], Sabah K. Bishay[1], Mariam M. Aziz[1], Nardeen R. Ratib[1], Nessma K. Ahmed[1], Shimaa M. Ali[1], E. Ismail[1], Galal A. R. El-Sherbieny[1] and Haitham M. A. Elsayed[1*]

[1]Department of Genetics, Faculty of Agriculture, Sohag University, 82786 Sohag, Egypt.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Review Article

Received 06 April 2021
Accepted 12 June 2021
Published 15 June 2021

ABSTRACT

Plants are comprised of interrelated traits, where a change in one trait may cause a change in another, or in a combination of traits. Bioinformatics tenancies were distinguished by wide accessibility of computers to different aspects of genomes. Nucleic acid sequences and information from a wide range of genomes become possible through genomics. Genomics made this information accessible to further analysis and experimentation. Therefore, development of computerized models for quantitative traits used bioinformatics techniques can reduce the time and cost in creating new plant variety, and can significantly improve breeding efficiency by constructing reliable predictive estimates and identifying selectable genotypes by greatly accelerating the progress in both fundamental plant science and applied breeding research. Moreover, it clarify the function of key genes and the interaction of responsible genes. Thus, a variety software's and web-based tools have been developed to help with these issues. So, this article highlights the functional information and tools for genome annotation, gene ontology and gene network by stating the art regarding genome assembly. In addition, we show how phenotypic data yield new trait-trait correlations by linking phenotypic data to genomic data together.

Keywords: Genome annotation; gene ontology; gene network; genome assembly; correlation.

1. INTRODUCTION

The development of crop breeding is largely constrained by the lack of theoretical level of many branches of modern genetics, the basis of which should be appropriate mathematical models. The presence of such models can significantly simplify and accelerate the solution of genotype identification problems, as well as predict the combination of quantitative traits of

Corresponding author: Email: dhaithamm@yahoo.com;

newly created varieties. So, combining traditional crop modelling with new breeding methods and genetic modelling will help accelerate the creation of new plant varieties for different conditions [1]. A large portion of these tools and techniques are related to OMICs category [2]. The OMICS tools can enhance the quality nutritional composition of food crops, increasing agricultural production for food, feed, and energy. With the use of OMICS, the consistency and predictability of plant breeding programmes have been improved, reducing the time and the expense of stress tolerant varieties [3]. One of the most important subcategories of OMICs is transcriptomics which attracts a large number of biologists especially in plant breeding area [4,5,6,7].

As with genomic developments, there are promising advances in plant phenotyping technology, such as the use of automated phenotyping machinery [8] and advanced image analyses [9]. This has resulted in unprecedented insights into plant physiology, architecture, and performance. Compared with genomic research, data output produced by established systems in plant phenotyping is still manageable [10].

According to National Center of Biotechnology Information (NCBI), bioinformatics is the field of science that merge biology, computer science and information technology into a single context [11]. The classical definition of bioinformatics is the mathematical, statistical and computing methods aimed to resolve biological problems using DNA, RNA, amino acid sequences and related information [12]. Bioinformatics has been involved in different aspects of sciences including plant breeding. From the past to the present, plant breeding has been extended through development and deployment a large number of methods and tools with respect to specific objectives [13].

Sequencing of a genome is only the first step toward understanding genome organization, gene structure and gene expression patterns. Up to date, near-complete genome sequences are available for only two model species, Arabidopsis and rice [12]. However, many genome sequencing projects have been undertaken with respect to different plant species like wheat and maize [14]. The initial post-Sanger sequencing advancement came in the form of high-throughput short-read technologies, frequently termed second-generation sequencing [15]. Analysis that combine advanced phenotyping and genomic datasets offer great potential for the discovery of novel insights, such as in GWAS [16] or genomic prediction technologies, even within the scope of a single project. Furthermore, machine learning and other data science techniques can extract novel insights from meta-analyses of multiple datasets. Selection from the point of view of modern understanding of plant genetics and the increasing complexity of experiments and instruments has significantly increased the rate of output varieties [1]. So, the use of modern approaches of bioinformatics is an urgent need for modern plant breeding to reduce labor and material costs in short time. Because, the variety model is a scientific prediction showing what combination of traits plants should have to provide a given level of productivity, sustainability and other required production qualities [17].

Overall, due to great impact of plant breeding to develop new genotypes that are genetically superior to the currently existing ones for a specific environment, it is necessary to assess the role of bioinformatics toward plant breeding science. Accordingly, the main target of this research review is to outline the current state of the art in genomics, plant phenotyping, and standardization by address a list of online application tools being used in the data analysis of different molecular biology assays to facilitate plant improvement.

2. JOINING COMPUTER SIMULATION WITH PLANT BREEDING

Because computer simulation is fast and uses few physical resources, we can easily see why simulations and plant breeding are compatible. Computer simulation can accommodate genetic models with multiple genes, multiple alleles, pleiotropic and epistatic effects. A remarkable utility of computer simulation in plant breeding is combining knowledge from quantitative genetics, molecular genetics, and plant breeding to predict the efficiency of different selection methods [18]. An analytical approach makes it difficult to account for parental selection, pedigree relationships, Linkage, and recombination, but a simulation experiment can easily accommodate all of these and more. Thus, simulation can be a valuable tool for breeders to find the most efficient path to the target cultivar using parental selection, predicting cross performance, comparing selection strategies, breeding by design, etc. [19]. The efficiency of MAS, an indirect selection method, is determined by the genetic correlation between the target traits and the markers used for selection. This correlation is determined by the strength of linkage disequilibrium (LD) between the markers and the QTL as well as the percentage of additive genetic variance explained by the markers [20]. The optimal number of markers depends on the degree of LD between the marker and QTL. LD is determined by the average number of recombination per generation in that region of the genome, the number of generations since the original mutation, and the population size [21]. If LD is large, more markers will not necessarily lead to a higher selection response. This result occurs for two possible reasons. First, when LD is large, multiple makers are in LD with each other, creating colinearity, meaning that their effects overlap; second, covariance in these markers makes estimating QTL effects inaccurate, thus decreasing the efficiency of MAS. Increasing the number of markers in a finite population could even result in spurious linkages between markers and unlinked QTLs [19]. Based on the literature, the major applications of computer simulation can be partitioned into four areas: (1) breeding method comparison: finding the best breeding scheme taking account of multiple factors; (2) gene mapping: validating the effectiveness of new mapping methods, calculating LOD score and confidence interval; (3) plant-breeding platforms: integrative simulation programs that can simulate the whole plant-breeding process; and (4) crop modeling: combining crop models, genetic architecture of traits, and environmental information to fill the gap between genotype and phenotype as revealed in Fig. 1.

Computer simulation can be used to (1) compare breeding methods, (2) compare gene mapping strategies and calculate significance threshold and confidence interval, (3) simulate gene networks and genotype-by-environment interactions, and (4) simulate crop growth using crop models to assess the influence of genes, environment, and climate change, among other things [19]. Therefore, simulation tools are important in designing and testing breeding strategies and this role becomes increasingly important as our understanding of the genetic architecture of quantitative traits improves.

3. FROM SEQUENCES TO GENOMES

Many standard pipelines and tools that can potentially assemble a reasonable quality genome, are available (Fig. 2). The primary factor determining the choice of assembly pipeline is the type of reads in the dataset, as short and long reads are generally assembled using very different approaches. In the case of short-read data from the Illumina platform, reads are typically quality controlled using for example FASTQC followed by adapter/quality trimming [22]. After read trimming, the assembly process can be performed using a variety of short-read assemblers such as ABySS, DISCOVAR (*de novo*), Velvet or SOAPdenovo [23]. Alternatively, commercial software such as the CLC assembler can be used with small computational resources and offers a graphical user interface, whereas the commercial NRGene suite enables the

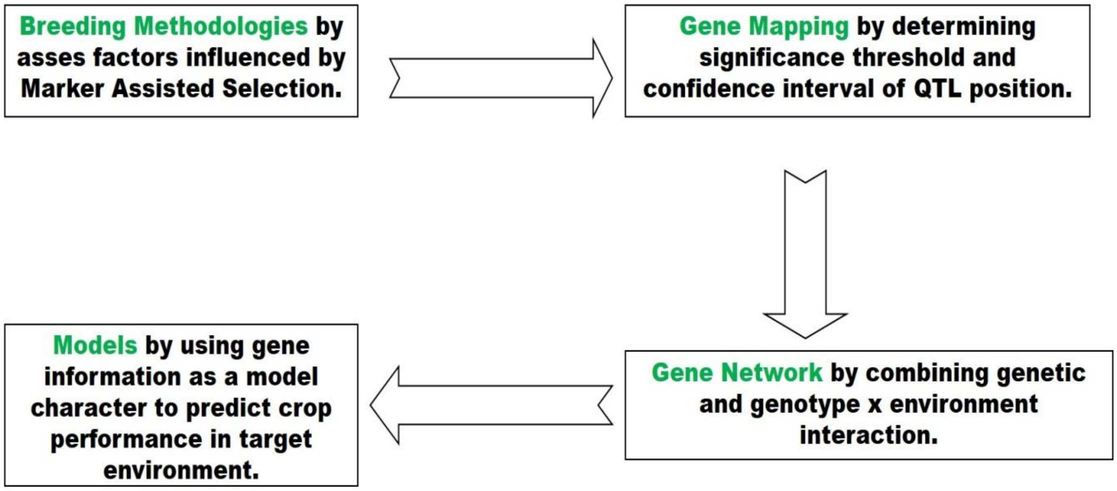
Breeding Methodologies by asses factors influenced by Marker Assisted Selection.
Gene Mapping by determining significance threshold and confidence interval of QTL position.
Gene Network by combining genetic and genotype x environment interaction.
Models by using gene information as a model character to predict crop performance in target environment.

Fig. 1. Bioinformatics relationship with plant breeding

analysis of complex genomes using short-read data [23,24,25]. However, when sufficient long-read data are available, a long-read assembly approach will generally give a better result. Short reads can be used before assembly to correct the individual long reads (Fig. 2) or after assembly to correct the contigs (Fig. 3), a process commonly referred to as ‘polishing’ the assembly. Due to costly long-read sequencing technologies prohibitively expensive, both in terms of sequencing and computation, due to the high repeat content of many plant genomes, with repetitive elements derived from a wide range of sources, including transposons and tandem gene duplications. Therefore, for pre-assembly read correction, the simplest approach is to map individual short reads onto long reads and use the short-read consensus to correct the long reads. This approach is implemented in tools such as Proovread [26] and/ or LSC [27]. Post-assembly polishing using short reads can be performed using Pilon, while Racon supports polishing with either short or long reads. Although polishing with accurate short reads can dramatically improve assembly accuracy, in practice, this often applies only to unique genome regions [23]. In addition, especially for some plant genomes, many short reads cannot be accurately mapped to one location due to transposon-derived repeats and homologous genes with a high degree of identity, making the long-read assembly errors unrecoverable by short reads [23]. Short-read sequencing technologies, in conjunction with annotated reference genomes, can be readily applied to a variety of biological questions, including detection [28] and analysis of gene expression [29], DNA methylation [30], identification of transcription factor binding sites and the detection of causal regions and mutations in mutant screens [31,32] or populations [33]. However, the importance of next-generation sequencing beyond the context of a single reference accession has long been recognized [34]. As sequencing became more accessible in terms of cost and availability, plant projects frequently sequenced multiple accessions or species in order to investigate natural diversity. This was initially applied to plant species with relatively small genomes such as rice or Arabidopsis, but has since been extended to field crops such as tomato [35].

A reduced representation of a genome is potentially the cheapest way to gain SNP and marker information in order to enable genome-wide association studies (GWAS) and genomic selection studies [36]. The key idea was to reduce the sequencing cost per sample by only sequencing corresponding parts of genomes, albeit at the cost of a more complex library preparation, using restriction enzymes to selectively cut the DNA, therefore focusing the sequencing around the restriction sites. While techniques based on mapping reads to reference genomes are well suited to GWAS and genomic selection, they are inadequate in identifying new genome variants, such as novel genes not present in the reference. In maize, it was estimated that an early genomic reference did not capture about a quarter of the low-copy gene fraction from all inbred lines [23]. Therefore, a full genome annotation will usually first rely on an automatic functional annotation based on domain analyses and sequence similarity searches. In order to provide consistency, most tools that automatically annotate genomes frequently employ formalized ontologies such as Gene Ontology (GO) or MapMan ontology [37]. There are many tools available that automatically annotate genes using ontologies such as the Mercator automated annotation tool [38], BLAST2GO [39], KEGG Automatic Annotation Sever (KAAS) [40], and TRAPID [41]. The overarching goal of these tools is the rapid automatic annotation of genes to a high standard, approaching that of manual annotation.

The final endpoint of a genome assembly is ordering and orienting the assembled sequences to form chromosomal pseudomolecules. This can be guided by marker sequences from an independently determined genetic map. Alignment of these marker sequences against the assembly allows the approximate chromosomal position and potentially orientation of each scaffold to be determined. This last step is often not reached, as it is either not needed for the planned analyses or high resolution genetic maps are not available. However, in the context of combining genotypes with phenotypes, the exact chromosomal position of genes is essential for their correlation with known QTLs. For plants it was notably applied to the 5 Gb barley and 12 Gb wild emmer genome [24,42] and has allowed chromosome-scale assemblies without a genetic map for example for raspberry [43]. In summary, using multiple-reference genomes, it is possible to find new genes or new regulatory *cis* elements. This would not be possible with only one reference. Especially in the case of regulatory elements, line-specific transposon insertions bringing their own regulatory elements might play an important role [44].

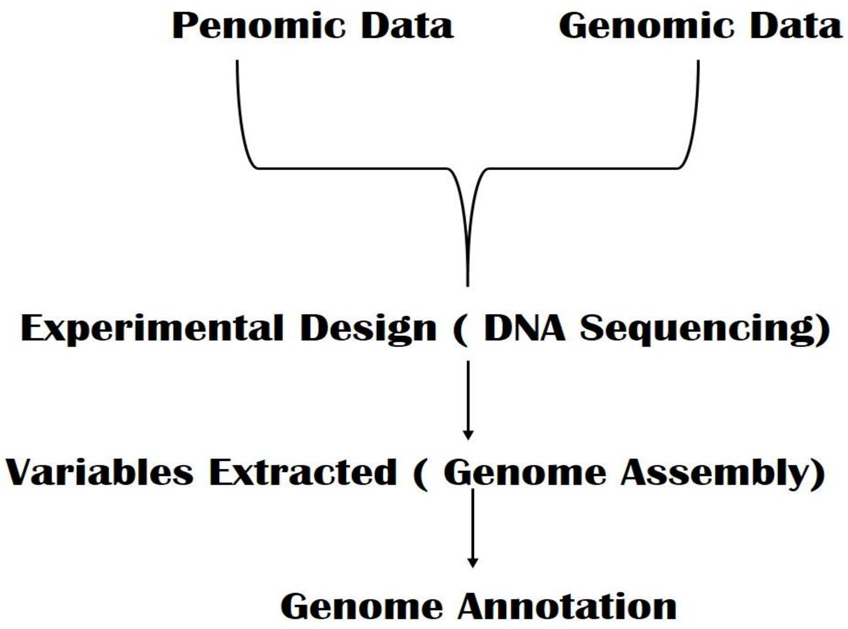
Penomic Data
Genomic Data
Experimental Design (DNA Sequencing)
Variables Extracted (Genome Assembly)
Genome Annotation

Fig. 2. Flow chart of genomic and phenomic data for new genome analysis

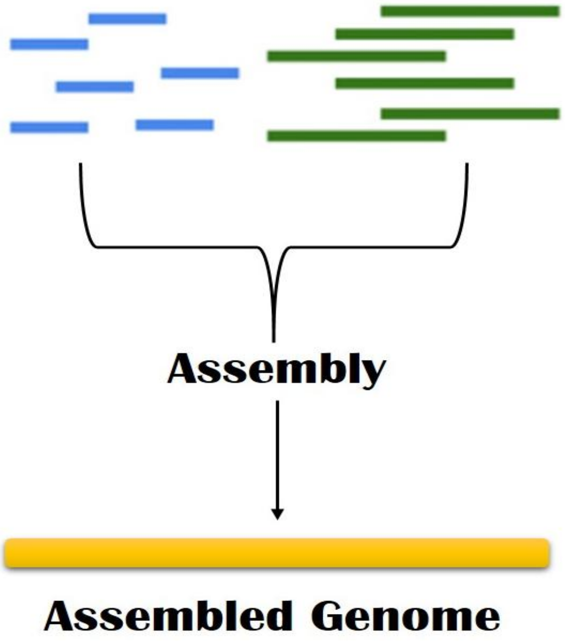

Fig. 3. Ideal Approach for genome sequencing

4. EXAMPLES TOOLS FOR FUNCTIONAL INFORMATION ANALYSIS

Computer programs are mainly designed to clarify the inheritance of quantitative traits using various genetic and mathematical methods [45]. When developing technologies for optimizing the construction of models of varieties, it is necessary to have information that comprehensively reflects the interaction of the biological system (variety) with different soil and climatic conditions. Therefore, before proceeding to the correction of expert models for the construction of varieties, we need data on the basis of which it would be possible to develop a complex system of the variety model, which allows the introduction of certain input data to solve specific breeding problems [45]. Modeling in biology is still a very serious research activity related to hypotheses and tests, as there is not enough information about the context of the problem to give a clear guide on how to "uniquely" determine the likely solution to various possibilities [46]. The creation of an interactive information and research database will provide a systematic knowledge of plant genomics, gene interactions, the manifestation of the genes in the form of phenotypic features and the impact on these processes of environmental factors.

Classical breeding depends largely on the phenotypic selection and the experience of the breeder, so the efficiency of breeding is low and the forecasts are inaccurate. Along with the rapid development of molecular biology and biotechnology, there is a large amount of biological data for genetic studies of the most important plant traits, which, in turn, allows for genotypic selection in the breeding process. However, gene information is still not effectively used in plant breeding, due to the lack of appropriate tools and technologies. The simulation approach can use extensive and varied genetic information, predict cross-performance and compare different breeding methods. Thus, it is possible to identify the most effective crossing technologies and effective breeding strategies. Modeling the breeding process allows to determine the complex genetic model consisting of multiple alleles and genes, effects of pleiotropy and epistasis, the interaction with the environment and represents a useful tool for breeders to effectively use a wide range of genetic data and other available information [47].

The new generation phenotyping generates significantly more data than previously and requires new data management systems, data access and storage, increased use of ontologies to facilitate data integration, and new statistical tools for enhancing experimental design and extracting biologically meaningful signal from environmental and experimental noise. To ensure relevance, the implementation of effective and informative experiments full also requires familiarity with the various resources germoplasm, structures of the populations and the target populations of environment.

With the growing needs of the population in quality agricultural products, the growing of the diversity of varieties, the identification of species and varieties of DNA markers becomes an increasingly important task. Molecular markers and developed on the basis of their molecular passports varieties of Kazakhstan breeding will acquire a single value in seed production to control the varietal purity during the multiplication of varieties.

Improved DNA sequencing and computational technologies allow us to track plant genetic variability at a level unthinkable a few years ago. This radically improves plant breeding and makes the process much cheaper and faster. This allows scientists to work with very complex properties and makes it possible to use the valuable genetic diversity of wild relatives of crops. The arrival of affordable, high-throughput DNA sequencing, combined with improved bioinformatics and statistical analysis, provides significant advances in molecular plant breeding [48]. Multi-disciplinary programs for the breeding of major crops allow genomic variations of DNA sequences to be investigated and associated with the inheritance of very complex traits controlled by many genes [49].

Assessment and focus on the network of molecular interactions are the most attractive aspects of molecular biology in the post genomic era unraveling the complex roles of genes their products and the cellular environments in different biological processes [50]. Therefore, interrelationships of genes and their products (especially protein product) can be investigated through a complex network of interactions. With respect to biological systems, genes are interacted each other [51,52] and the function of a gene depends on its cellular context, and the activity of a cell is determined by which genes are being expressed and by interactions with others [53].

The scope of the genes function is infinite in different genomes. Many tools are available for interpreting of gene expression studied which are:

4.1 FlexArray

Is a Windows software package that has been created to simplify data analysis of expression micro arrays. This tool is suited for normalization algorithms, statistical tests and other complex data-processing tasks. It is also an exploration tool for analysis methods and algorithms [54]. This software can be found athttp://genomequebec.mcgill.ca/FlexArray.

4.2 BioConductor

Is based primarily on the R programming language although it is friendly with different programming languages. Also, is open development software project being used in the analysis and comprehension of genomics data, especially microarray data. The Bioconductor q-value package was also used for p value correction [55]. For more details about the Bioconductor follow http://www.bioconductor.org/ [56].

4.3 Maanova

Is suited for the analysis of both small- and large-scale microarray experiments, which is implemented in Matlab is an add-on package for the statistical language R for Analysis Of Variance. This software is friendly to run on any platform that supports these packages. This package provides a complete work flow for different aspects of microarray data analysis i.e. data-quality checks (visualization and transformation), ANOVA model fitting (both fixed and mixed effects models), statistical tests (F and Fs statistics), p-value (using sampling and residual shuffling permutation approach) and summarize the results in tables and graphics including volcano plot and bootstrapping-based tree cluster.

4.4 Gene Ontology (GO)

This tool reported gene expression study with different wheat cultivars [57]. Such information are a description of the molecular function, biological process, and cellular component of gene products. To reach GO data follow (http://www.geneontology.org/).

4.5 Gramene

Is an integrated web resource for visualizing and comparing the data of plant genomes, which is varied and included genomes, protein structures, EST sequencing, genetic and physical mapping, QTL localization, interpretation of biological pathway, functional analysis, Gene Ontology. For further details refer to http://www.gramene.org/ [58].

4.6 AGRIS

The Arabidopsis Gene Regulatory Information Server (AGRIS) provides a comprehensive resource for gene regulatory studies in *Arabidopsis thaliana* as a genetically model plant. Providing information from Arabidopsis promoters, *cis*-regulatory elements, TFs, and their interactions into regulatory networks.

The aggregation and analysis of large volumes of long-term data (Big Data) will allow to model both the variety as a whole and its individual indicators in their interaction with environmental factors, which will allow to identify the most significant features/characteristics in relation to specific soil-climatic zones. In addition, it is possible to find suitable parent pairs, which are most likely to appear in the simulated lines. Information retrieval (analytical) system allows to speed up the work of a biologist breeder, automating the process of identifying the relationship between the phenotype, genotype and the environment. With this system, the breeder can analyze hundreds and thousands of lines in a matter of minutes, which increases the speed of data processing and, accordingly, productivity.

These software’s has been used in different plant species for microarray data analysis [12,59].

5. CONCLUSION

As has been shown above, both genomic and phenomic datasets are becoming more and more mature and cost efficient. At this time, it is the model plant Arabidopsis, rather than crop plants, that contains the most extensive datasets and that may enable ontology-driven phenotype prediction. This is largely due to a number of points: (i) the availability of the machine-readable ontology termenriched phenotypic datasets for well defined genes; (ii) the largest wealth of functional data for gene annotation, which is related to the former point; (iii) the use of standardized populations from the 1001 genome consortium, facilitating abstraction at the phenotypic level; and (iv) standardization, driven for example by TAIR. Also, for genetic and genomic studies, it is necessary to note the importance of accurate phenotyping.

As we learn more about the extremely diverse conditions and climate variability faced by farmers, scientists can find the most suitable varieties/plants faster and more easily using simulation modeling. Modeling of varieties is very useful for understanding what a plant needs to achieve higher yields in a given environment [60]. The total set of the revealed morphobiological signs of plant is offered as scientifically proved model of a variety. The proposed optimal parameters of the variety model will help to improve the efficiency of breeding of economically valuable genotypes and targeted selection for adaptability to the conditions of the region for the creation of new high-yield varieties of agricultural plants.

The use of information technologies for the creation of science-based models of varieties will not only reduce the time and cost of breeding research, but also improve the efficiency of the breeding process due to the quality of interpretation of the results of experiments, reliability and reliability of the findings.
The use of modern technologies (mathematical modeling, bioinformatics, Big Data analysis, drone, satellite and aerial photography, electronic field maps, GPS tracking and geolocation programs and others) will significantly simplify, accelerate and improve the fficiency of breeding processes, reduce the cost of creating a new variety, and also to get rid of routine operations due to the quality of interpretation of the results of experiments, reliability, simplicity and reliability of research results.

ACKNOWLEDGEMENTS

The authors wish to acknowledge Genetics Dept., Faculty of Agric., Sohag Univ., Sohag, Egypt conducting the present work. The authors would also like to thank reviewers for their valuable comments and criticisms.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Urazaliev KR. Bioinformation Technologies in Plant Breeding. UDC. 2019;57:51-76; 57.02:001.57.
2. Barh D, Zambare V, Azevedo V. Omics: Applications in Biomedical, Agricultural, and Environmental Sciences. 2013;CRC Press.
3. Van Emon JM. The Omics Revolution in Agricultural Research. J Agr Food Chem. 2016;64(1):36-44.
4. Usadel B, Fernie AR. The Plant Transcriptome from Integrating Observations to Models. Front Plant Sci. 2013;4:48.
5. Barh D, Khan MS, Davies E. Plant Omics: The Omics of Plant Science. Springer; 2015.
6. Gürel F, Öztürk NZ, Uçarlı C. Transcriptomic Responses of Barley (*Hordeum vulgare* L.) to Drought and Salinity. In Plant Omics: Trends and Applications; 2016.
7. Hakeem KR, Tombuloğlu H, Tombuloğlu G. Plant Omics: Trends and Applications; 2016.
8. Fiorani F, Schurr U. Future Scenarios for Plant Phenotyping. Annu. Rev. Plant Biol. 2013;64:267–291.
9. Pound MP, Atkinson JA, Townsend AJ et al. Deep Machine Learning Provides State of The Art Performance in Image-Based Plant Phenotyping. Gigascience. 2017;6(10):1–10.
10. Coppens F, Wujts N, Inze D, Dhont S. Unlocking the Potential of Plant Phenotyping Data through Integration and Data Driven Approaches. Curr. Opin. Syst. Biol. 2017;4:58–63.
11. Thampi SM. Introduction to Bioinformatics. arXiv preprint arXiv:0911. 2009; 4230.
12. Shariatipour Nikwan, Bahram Heidari. Application of Bioinformatics in Plant Breeding Programmes. BAOJ Bioinfo. 2017;1(2):1-8.
13. Al-Khayri JM, Jain SM, Johnson DV. Advances in Plant Breeding Strategies: Breeding, Biotechnology and Molecular Tools. Springer International Publishing; 2015.
14. Skuse GR, Du C. Bioinformatics Tools for Plant Genomics. Int J Plant Genomics; 2008.
15. Anthony M. Bolger, Hendrik Poorter, Kathryn Dumschott, Marie E. Bolger, Daniel Arend, Sonia Osorio, Heidrun Gundlach, Klaus F.X. Mayer, Matthias Lange, Uwe Scholz, Bjorn Usadel. Computational Aspects Underlying Genome to Phenome Analysis in Plants. The Plant Journal. 2019;97:182–198.
16. Millet EJ, Welcker C, Kruijer W et al. Genome-Wide Analysis of Yield in Europe: Allelic Effects Vary with Drought and Heat Scenarios. Plant Physiol. 2016;172:749–764.
17. Malchikov PN, Vyushkov AA, Myasnikova MG. Formation of Models of Durum Wheat Varieties for the Middle Volga Region. Samara: Samar. Scientific center of RAS; 2009.
18. Cooper M, Podlich DW, Luo L. Modeling QTL Effects and MAS in Plant Breeding. In Genomics-Assisted Crop Improvement" (R. K. Varshney and R. Tuberosa, Eds.).

2007;57–95. Springer, Dordrecht, the Netherlands.

19. Li Xin, Chengsong Zhu, Jiankang Wang, Jianming Yu. Computer Simulation in Plant Breeding. Advances in Agronomy. 2012; 116(3):219-264.
20. Johnson R. Marker-Assisted Selection. Plant Breed. Rev. 2004;24(1):293–309.
21. Mackay TFC. The Genetic Architecture of Quantitative Traits. Annu. Rev. Genet. 2001;35:303–339.
22. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–2120.
23. Bolger ME, Arsova B, Usadel B. Plant genome and transcriptome annotations: from misconceptions to simple solutions. Brief. Bioinform. 2019; 19(3):437–449.
24. Avni R, Nave M, Barad O et al. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science. 2017;357(6346): 93–97.
25. Luo MC, Gu YQ, Puiu D et al. Genome sequence of the progenitor of the wheat D genome Aegilops tauschii. Nature. 2017; 551:498–502.
26. Hackl T, Hedrich R, Schultz J, Forster F. Proovread: large scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics. 2014; 30:3004–3011.
27. Au KF, Underwood JG, Lee JG, Wong WH. Improving Pac-Bio long read accuracy by short read alignment. PLoS ONE. 2012; 7:e46679.
28. Zhang R, Calixto CPG, Marquez Y et al. A high quality Arabidopsis transcriptome for accurate transcript-level analysis of alternative splicing. Nucleic Acids Res. 2017;45(9):5061–5073.
29. Ezer D, Jung JH, Lan H et al. The evening complex coordinates environmental and endogenous signals in Arabidopsis. Nat. Plants. 2017;3:17087.
30. Zhong S, Fei Z, Chen YR et al. Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. Nat. Biotechnol. 2013;31:154–159.
31. James GV, Patel V, Nordstrom KJ, Klasen JR, Salome PA, Weigel D, Schneeberger K. User guide for mapping-by-sequencing in Arabidopsis. Genome Biol. 2013; 14(6):R61.
32. Klap C, Yeshayahou E, Bolger AM, Arazi T, Gupta SK, Shabtai S, Usadel B, Salts Y, Barg R. Tomato facultative parthenocarpy results from SlAGAMOUS-LIKE 6 loss of function. Plant Biotechnol. J. 2017; 15(5):634–647.
33. Thoen MP, Davila Olivas NH, Kloth KJ et al. Genetic architecture of plant stress resistance: multi-trait genome-wide association mapping. New Phytol. 2017; 213(3):1346–1362.
34. Varshney RK, Nayak SN, May GD, Jackson SA. Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol. 2009;27:522–530.
35. Lin T, Zhu G, Zhang J et al. Genomic analyses provide insights into the history of tomato breeding. Nat. Genet. 2014; 46:1220–12266.
36. Bhat JA, Ali S, Salgotra RK et al. Genomic selection in the era of next generation sequencing for complex traits in plant breeding. Front. Genet. 2016;7:221.
37. Jaiswal P, Usadel B. Plant pathway databases. Methods Mol. Biol. 2016; 1374:71–87.
38. Lohse M, Nagel A, Herter T et al. Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. Plant, Cell Environ. 2014; 37:1250–1258.
39. Conesa A, Gotz S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int. J. Plant Genomics. 2008;619832.
40. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: An automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007;35:W182–W185.
41. Van Bel M, Proost S, Van Neste C, Deforce D, Van de Peer Y, Vandepoele K. TRAPID: an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes. Genome Biol. 2013;14:R134.
42. Mascher M, Gundlach H, Himmelbach A, et al. A chromosome conformation capture ordered sequence of the barley genome. Nature. 2017;544:427–433.
43. Van Buren R, Wai CM, Colle M et al. A near complete, chromosome-scale assembly of the black raspberry (*Rubus occidentalis*) genome. Gigascience. 2018; 7(8).

Available:https://doi.org/10.1093/gigascience/giy094.

44. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: From conflicts to benefits. Nat. Rev. Genet. 2017;18(2):71–86.
45. Grebennikova IG, Aleynikov AF, Stepochkin PI, Building A. Model of the Varieties of Spring Triticale On The Basis Of Modern Information Technologies. Computing Technologies. 2016;21(1):53-64.
46. Cui ML et al. Quantitative Control of Organ Shape by Combinatorial Gene Activity. PLoS Biol. 2010;8(11): Available:http://dx.doi.org/10.1371/journal.pbio.1000538.=
47. Wang J, Wolfgang HP. Simulation Modeling in Plant Breeding: Principles and Applications. Agric. Sci. China. 2007;6(8):908-921.
48. Mcpherson JD. A defining decade in DNA sequencing a revolution in DNA sequencing technology has enabled new insights from thousands of genomes sequenced across taxa. Nat. Methods. 2014;11(10):10.1038/nmeth.3106.
49. Bassi FM et al. Breeding schemes for the implementation of genomic selection in wheat (*Triticum spp.*). Plant Sci. Elsevier. 2016;242:23-36.
50. Page GP, Coulibaly I. Bioinformatics Tools for Inferring Functional Information from Plant Microarray Data: Tools for the First Steps. Int J Plant Genomics; 2008.
51. Arnone MI, Davidson EH. The hardwiring of development: organization and function of genomic regulatory systems. Development. 1997;124(10): 1851-1864.
52. Miklos GL, Rubin GM. The role of the genome project in determining gene function: insights from model organisms. Cell. 1996; 86(4): 521-529.
53. Barabasi AL, Oltvai ZN. Network biology: Understanding the cell’s functional organization. Nat Rev Genet. 2004;5(2):101-113.
54. Blazejczyk M, Miron M, Nadon R. FlexArray: A statistical data analysis software for gene expression microarrays. Genome Quebec; 2007.
55. Garcia-Seco D, Chiapello M, Bracale M, Pesce C, Bagnaresi P, et al. Transcriptome and proteome analysis reveal new insight into proximal and distal responses of wheat to foliar infection by *Xanthomonas translucens*. Sci Rep. 2017;7.
56. Drăghici S. Statistics and data analysis for microarrays using R and bioconductor. CRC Press; 2011.
57. Poersch-Bortolon LB, Pereira JF, Nhani Junior A, Gonzáles HH, Torres GA et al. Gene expression analysis reveals important pathways for drought response in leaves and roots of a wheat cultivar adapted to rainfed cropping in the Cerrado biome. Genet Mol Biol. 2016;39(4):629-645.
58. Monaco MK, Stein J, Naithani S, Wei S. Dharmawardhana P et al. Gramene 2013: comparative plant genomics resources. Nucleic Acids Res. 2014;42(D1):D1193-1199.
59. Windram O, Madhou P, McHattie S, Hill C, Hickman R et al. Arabidopsis defense against Botrytis cinerea: chronology and regulation deciphered by high-resolution temporal transcriptomic analysis. The Plant Cell Online. 2012;24(9):3530-3557.
60. Vadez V. Crop simulation models: predicting the future of pulses. 2016;100-102.