Adv Biochem Eng Biotechnol
DOI: 10.1007/10_2017_56

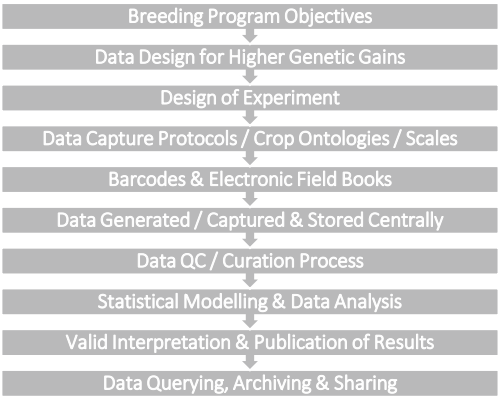
Current Status and Future Prospects of Next-Generation Data Management and Analytical Decision Support Tools for Enhancing Genetic Gains in Crops

Abhishek Rathore, Vikas K. Singh, Sarita K. Pandey, Chukka Srinivasa Rao, Vivek Thakur, Manish K. Pandey, V. Anil Kumar, and Roma Rani Das

Abstract Agricultural disciplines are becoming data intensive and the agricultural research data generation technologies are becoming sophisticated and high throughput. On the one hand, high-throughput genotyping is generating petabytes of data; on the other hand, high-throughput phenotyping platforms are also generating data of similar magnitude. Under modern integrated crop breeding, scientists are working together by integrating genomic and phenomic data sets of huge data volumes on a routine basis. To manage such huge research data sets and use them appropriately in decision making, Data Management Analysis & Decision Support Tools (DMASTs) are a prerequisite. DMASTs are required for a range of operations including generating the correct breeding experiments, maintaining pedigrees, managing phenotypic data, storing and retrieving high-throughput genotypic data, performing analytics, including trial analysis, spatial adjustments, identifications of MTAs, predicting Genomic Breeding Values (GEBVs), and various selection indices. DMASTs are also a prerequisite for understanding trait dynamics, gene action, interactions, biology, GxE, and various other factors contributing to crop improvement programs by integrating data generated from various science streams. These tools have simplified scientists' lives and empowered them in terms of data storage, data retrieval, data analytics, data visualization, and sharing with other researchers and collaborators. This chapter focuses on availability, uses, and gaps in present-day DMASTs.

A. Rathore (✉), V. K. Singh, S. K. Pandey, C. S. Rao, V. Thakur, M. K. Pandey, V. Anil Kumar, and R. R. Das
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India
e-mail: a.rathore@cgiar.org

Graphical Abstract



Keywords Analytical Decision Support Tool, Data management, Genetic gains, Plant breeding

Contents

1 Introduction
2 DMAST for Phenotypic Evaluation of Datasets
3 DMAST for Molecular Maker Datasets Including Genomics Data
4 DMAST for Metabolomics and Proteomics Data
5 DMAST for Molecular Breeding
6 Integrated Pipelines for Plant Breeding Data Management
7 DMASTs for Data Sharing and Visualization
8 Breeder Requirements for Enhancing Genetic Gains
8.1 Pipeline to Understand the Association Between Phenotype and Genotype
8.2 High-Throughput and Precision Phenotyping
8.3 New Web-Based Interface with Better Organization
8.4 Trait Ontology Inference as Part of the Data Management Pipeline
8.5 Better Support from Plant Genomics
8.6 Better Support for Data Analysis and Investments
8.7 Integration from "Omics" Information
9 Conclusion
References

1 Introduction

Good quality research experiments, precise data, appropriate data analysis, and data-driven decision making make up the backbone of modern agricultural research and integrated breeding. Integrated breeding exploits high-throughput phenomics and genomics, and has opened the floodgates of data pouring into crop specialists of all disciplines. Many international consortia and research centers are engaged in plant

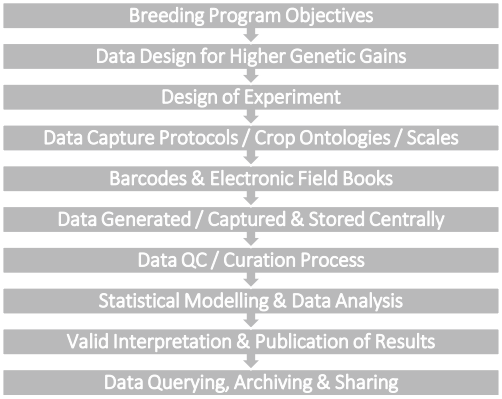


Fig. 1 Data management pipeling of a successful breeding program. Traditionally, data management has been interpreted as storing research data in online repositories and share. However, data management has a much wider definition, which starts with designing the study and concludes with appropriate analysis and making the data publicly available through repositories

research and are generating huge amounts of sequencing/genotyping and phenotyping data. These data sets require the appropriate capacities for processing, analysis, management, and storage. Often it becomes very difficult to analyze these data sets and convert them into information through conventional data-management tools and analysis strategies. It has also been observed that researchers very frequently limit the definition of data management and interpret it in terms of mere physical data storage and access. However, the scientific data management scope is much wider and includes a complete life cycle. Figure 1, explains a typical data-management workflow in a breeding pipeline.

Usually databases and analytical tools are required for efficient utilization, retrieval, analysis, and decision making at each step of the genetic gain-enhancement process. Considering this need, experts in the area of bioinformatics, biometrics, and statistical genomics, in collaboration with plant geneticists, have developed many software tools and protocols for analyzing the data. Nevertheless, there is still tremendous scope for developing more efficient and user-friendly analysis and decision making to speed up the process of achieving higher genetic gain in crop plants.

The need for informatic intervention is required at all the steps of integrated breeding, such as selection of the appropriate experimental design, determining the size of the population, modern ways of data collection, use of modern databases, BLUE/BLUP with spatial adjustments being made during phenotypic selections, enabling sample tracking for DNA sample collections, genetic map construction, population structure, identification of marker-trait association, background and foreground selection, combining favorable alleles for complex traits using marker-assisted recurrent selection (MARS), and estimating genomics estimated breeding values (GEBVs) in genomic selection (GS). There are various open-source and proprietary tools available that cater to each step discussed above. Selection of these tools varies according to the hardware requirements, operating system, the degree of computer

skills required, user-friendliness, statistical models and algorithms used for analysis, corroboration of input data, and visualization of output results.

Development of integrated pipelines combining various useful software in one place also played a major role in the efficient integrated breeding program. Several such analytical pipelines are available that combine the analysis of phenotypic and genotypic data, such as running mixed models for statistical analysis, construction of linkage maps, mapping of quantitative traits, GWAS, and GS, etc. Nevertheless, some important DMASTs have been developed not only for helping with integrated breeding, but are also helpful in managing a larger set of data management for a future breeding program. We discuss here several informatics tools, data sharing and visualization platforms, their comparative usefulness experienced by researchers, and their ease of use, popularity, and prospects of improvement with regard to current technological needs and statistical methods.

We present several DMASTs in different sections, which include (a) DMASTs for phenotypic evaluation of datasets, (b) DMASTs for molecular maker datasets, (c) DMASTs for metabolomics and proteomics data, (d) DMASTs for molecular breeding, (e) integrated pipelines for plant breeding data management, and (f) DMASTs for data sharing and visualization. The last section is on breeder requirements for enhancing genetic gain.

2 DMAST for Phenotypic Evaluation of Datasets

High-throughput phenotyping generates a large volume of different types of data including nominal, categorical, ordinal, and ratio types of data sets. To capture variation and make good interpretations out of generated datasets, data should be subjected to appropriate statistical techniques. Good analysis will only be possible if the study was designed keeping the hypothesis in mind and data were also subjected to appropriate quality checks. The data must be cleaned, curated, and well summarized before final analysis and interpretation of results.

There is a range of statistical analysis available for the analysis of agricultural experiment data. A description of all of these is not possible in a single chapter and also out of the scope of this book, but we recommend appropriate selection of random and fixed factors and the use of mixed models, possibly with spatial adjustments. It is worth mentioning that analysis should be performed by standard and well-known statistical packages. Several commercial, open-source, and free software systems for statistical analysis are available. Often the commercial software is expensive, whereas the majority of the freeware has limited functions or is sometimes difficult to use. Among the commercially available software, ASREML (https://www.vsni.co.uk), Genstat (https://www.vsni.co.uk/software/genstat/), MINITAB, Statistical Package for Social Sciences (SPSS; http://www.spss.co.in/), Statistical Analysis System (SAS; https://www.sas.com), Statistica (https://software.dell.com/products/statistica/), and STATA (www.stata.com/) are very common, relatively easy to use, and can perform most data analyses and visualization for making breeding decisions.

There are plenty of free and open-source tools that are also available for performing statistical data analysis. R (https://www.r-project.org/) and Python (https://www.python.org/) are two such environments for performing sound statistical computing and visualization. These languages have become increasingly popular due to their versatility and availability of sound programming environments similar to many commercially available environments. R has a wide community support and has a core component that implements many classical and modern statistical methods. One can build on top of core functionality, develop their own code, and pack in R packages to perform customize analysis. The benefit of community support is that many analyses that are not available as part of core functions are also available to end users. R can also be interfaced with other programming languages and GUI development tools, such as Galaxy, Java, and Tk/Tcl. PBTools and CropStat (http://bbi.irri.org/products) are such free applications for plant breeders. Under open-source statistical software, R has emerged as the leader and most important software for analyzing data from all agricultural disciplines.

Python on other hand is also gaining popularity, but it seems it will take some more time to gain a strong place in the academic world. The advantages of Python are an easy learning curve and good graphics and visualization capabilities. Python has been used widely in web development and hence the development of online web-based analytical applications is a clear advantage with Python.

3 DMAST for Molecular Maker Datasets Including Genomics Data

To analyze genetic diversity with a moderate level of markers and genotypes, NTSYSpc (numerical taxonomy and multivariate analysis system) [1] is one of the most widely used programs, as is evident from the citation index. MEGA7 (Molecular Evolutionary Genetic Analysis) is another highly cited and widely used program, and was originally developed in 1993 [2]. This software can estimate the evolutionary distance or the phylogenetic tree calculation of genetic distance, using DNA or protein sequences data. This is a flexible and easy-to-use genetic data analysis system, and it can import unlimited sizes of datasets from various programs. DARwin (http://darwin.cirad.fr/) is a freely available software package developed for diversity and phylogenetic analysis by evolutionary dissimilarities. DAMBE (data analysis for molecular biology and evolution) is a phylogenetic analysis software first released in 2001 and recently updated as DAMBE5, with many new functions [3]. PAUP (phylogenetic analysis using parsimony) (http://paup.csit.fsu.edu/) is another widely-used program for inferring and interpreting evolutionary trees. It includes analysis of parsimony, distance matrix, invariance and maximum likelihood methods, and other statistical analysis.

To analyze population genetics, GENEPOP is a widely used population genetics software [4]. This software can estimate the number of tests (null allele estimates,

exact tests, Markov chain probabilities, and test statistics), multi-locus F-statistics, microsatellite allele sizes, RST, and rST, etc. Arlequin [5] is highly cited software for analyzing population genetics, and this software can handle a large number of datasets including molecular variance in the population regarding AMOVA, which is the unique feature of this software. Power Marker [6] is a program designed for SSR or SNP marker data for population genetic analysis, with a user-friendly graphic interface. DnaSP v5 (DNA Sequence Polymorphism) [7] is another software package for the analysis of nucleotide polymorphism from aligned DNA sequence data. SMOGD (Software for the Measurement of Genetic Diversity) [8] is a web-based application for the calculation of advanced proposed genetic diversity indices G′ST and Dest. GenAlEx 6.5 [9] is a Microsoft Excel-based software, and it offers a wide range of population genetic analysis options for the full spectrum of genetic markers.

Genome-mapping methods such as the construction of a genetic/linkage map and a physical map make up one of the basic steps involved in the identification of genes/QTLs for the trait of interest in the target environment. Mapping involves some steps such as determining recombination fractions, using a mapping function (Haldane or Kosambi), testing for appropriate linkages (LOD scores), grouping and ordering of markers into linkage groups, and bridging different genetic maps to develop a consensus map. Within the toolkit available for this work, MAPMAKER [10] open-source software released during the 1980s led the way towards computational strategies in the construction of the genetic map. This software uses a multipoint likelihood objective function [11] by combining the EM algorithm and the Hidden Markov Model (HMM) method, which significantly lowers the computational time when large datasets are used for analysis. The MAPMAKER software is still quite popular among geneticists, as the paper from Lander et al. [10] shows, with more than 6,000 citations in http://scholar.google.co.in/. However, due to the command prompt interface, it is not very user-friendly, and good quality graphic representation cannot be generated using it. JoinMap [12] is a widely used software as it has several positive features such as the user-friendly MS-Windows interface, ability to integrate maps from different mapping populations, continuous development, and professional support. JoinMap utilizes maximum likelihood and regression mapping algorithms for marker order strategies. For a better graphic representation, MapChart [13] and cMap [14] are also quite popular tools. For handling large numbers of marker datasets, special software packages have recently been developed, such as MadMapper [15] and MSTmap [16], for making a high-density genetic map.

Once the genetic map is made, the next step is to identify marker-trait associations by QTL analysis. As most agronomically important traits/phenotypes are polygenic in nature, many statistical and genetic models have been developed. MapMaker/QTL [17] was one of the most widely used open-source software during the 1990s and utilizes interval mapping (IM). As it is a command prompt-based interface and does not handle complex statistical models such as multiple interval mapping (MIM) and composite interval mapping (CIM), it is currently not much in use. QTL Cartographer [18] and QGene [19] do both MIM and CIM analysis. Software IciMapping [20] has a better QTL analysis model called inclusive composite interval mapping (ICIM). Recently, a new software called QTLnetwork [21] is becoming quite popular among geneticists as

it can analyze all types of genetic models such as additive QTLs, additive and epistatic QTLs, and QTL × environment interactions [22].

To understand the germplasm, STRUCTURE (http://pritchardlab.stanford.edu/structure.html) is the most extensively used software to detect population genetic structure. This program generates clusters caused by admixture between populations [23]. EIGENSOFT and Bayesian Analysis of Population Structure (BAPS) are the two another widely used statistical packages for detection and correction of population stratification in GWAS analysis [24, 25]. A detailed list of other available software packages for linkage disequilibrium analysis can be found in the following link: http://www.genes.org.uk/software/LD-software.shtml. Trait Analysis by aSSociation, Evolution, and Linkage (TASSEL) is the most common and highly cited software for performing marker-trait association analysis in GWAS studies in plants [26]. PLINK is another highly cited open-source whole genome association analysis toolset, which performs a range of basic, large-scale analyses in a computationally efficient manner [27].

4 DMAST for Metabolomics and Proteomics Data

In addition to genomics, proteomics and metabolomics hold a great perspective for serving as pillars for crop improvement. Complex and multi-omics studies have increased in the recent past, which integrate genomics data with metabolomics, epigenetics, and proteomics data. It is anticipated that in the future metabolomics will emerge as a significant part of crop improvement programs for achieving complex breeding objectives. Therefore, metabolomics techniques will be integrated with other "omics" technologies in order to identify and understand biochemical mechanisms and their consequences [28]. The few commonly used software programs available are BioCyc (http://biocyc.org), iPath (http://pathways.embl.de), KaPPA-View (http://kpv.kazusa.or.jp/en/), KEGG (http://www.genome.jp/kegg/pathway.html), MapMan (http://mapman.gabipd.org/web/guest/mapman), MetabolomeExpress (https://www.metabolome-express.org/), MetaboAnalyst (http://www.metaboanalyst.ca/faces/home.xhtml), Metscape (http://metscape.ncibi.org), MGV (http://www.microarray-analysis.org/mayday), Paintomics (http://www.paintomics.org), Pathos (http://motif.gla.ac.uk/Pathos/), Pathvisio (http://www.pathvisio.org/), PRIMe (http://prime.psc.riken.jp/), ProMetra(http://www.cebitec.uni-bielefeld.de/groups/brf/software/prometra_info/), Reactome(http://www.reactome.org), VANTED (http://vanted.ipk-gatersleben.de), and MetPA(http://metpa.metabolomics.ca). Most computational tools available are largely intended for metabolite identification. However, in order to gain some biological insight, it is necessary to have an integrated tool that can perform metabolite identification, functional analysis, detection of associated compounds, and metabolic modeling [28]. At the same time, there has been a swift addition of proteomics data due to advances in proteomics technologies such as high-throughput experimental

platforms [29]. There are a number of data repositories as well as data analysis and visualization tools available for proteomics [30–32].

PRoteomicsIDEntifications database (PRIDE; http://www.ebi.ac.uk/pride) is a comprehensive database of protein and peptide identifications; MSDA (https://msda.unistra.fr/) is a proteomics suite for detailed Mass Spectrometry Data Analysis; COMPASS (https://github.com/dbaileychess/Compass) is a suite of pre- and post-search proteomics software tools for OMSSA; PICR (http://www.ebi.ac.uk/Tools/picr/) or CRONOS [33] are web-based algorithms that associate names of the protein with their corresponding gene names; Gene Ontology terms (http://www.geneontology.org) are used to connect the protein identifier with its associated Gene. Obtained MS/MS spectra are interpreted with Mascot (http://www.matrixscience.com) and SEQUEST (https://omictools.com/sequest-tool) algorithms. Some functional databases such as the "Uniprot knowledge base" (www.uniprot.org/help/uniprotkb) and Ensembl (www.ensembl.org/) are being widely used in the field of proteomics along with other detailed pathway databases like KEGG (www.genome.jp/kegg/pathway.html), Reactome (http://www.reactome.org), and Ingenuity Pathway Knowledge Base (https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/). In addition to comprehensive resources, precise databases have been established for signal transduction processes, such as PANTHER (http://pantherdb.org/about.jsp). Information on protein interactions in complexes are deposited in interaction databases such as BioGRID (https://thebiogrid.org/) and IntAct (http://www.ebi.ac.uk/intact). Further, STRING (https://string-db.org/) and Cytoscape (www.cytoscape.org/) are graphic tools for visualizing and analyzing biological pathways. EnrichNet (www.enrichnet.org/) serves as a web-based platform, integrating pathway and interaction analysis in several databases (KeGG, Gene Ontology, Reactome, Wiki, and NCI pathways (http://www.wikipathways.org/index.php/WikiPathways). A few other programs like Pfam (http://pfam.xfam.org/), Interpro (https://www.ebi.ac.uk/interpro/), SMART (http://smart.embl-heidelberg.de/), and DAVID (https://david.ncifcrf.gov/) are among the commonly used software programs.

Therefore, in the future there is a need for an integrated tool to support the analysis and interpretation of multi-omics data generated from different fields consisting of large populations. Development and deployment of the DMASTs for metabolomics and proteomics in accordance with the necessity of breeding programs will help to achieve breeding targets efficiently and rapidly.

5 DMAST for Molecular Breeding

Once the genomic region has been identified through QTL analysis, these regions are then introgressed/pyramided into elite cultivars through the marker-assisted backcrossing (MABC) approach. To quickly introgress the targeted genomic regions, strategies such as foreground selection, recombination selection, and recovery of recurrent parent genome (RPG) through background selection are utilized.

Several visualization tools have been developed in the past such as GGT (graphical genotype) [34] and Flapjack [35], and are currently being used alone or as part of pipelines such as iMAS (http://www.icrisat.org/bt-biomatrics-imas.htm) and ISMU (Integrated SNP Mining and Utilization) [36]. The Marker-assisted Back-crossing Tool (MABT) is another JAVA-based decision-making software program that enables users to calculate the percentage of recovery of the recurrent parent at each generation (https://www.integratedbreeding.net/ib-tools/breeding-decision/marker-assisted-back-crossing-tool). To implement MARS in the breeding program through accelerated genetic gain by assembling favorable alleles issued from diverse parents, OptiMAS [37] has been developed with the following interactive graphical interface: (a) to trace parental alleles throughout generations, (b) to select the best plants based on estimated molecular scores, and (c) for an efficient inter-mating strategy to recombine positive alleles in a single genetic background. Genomic selection (GS) is a new molecular breeding approach using whole-genome profiling with a large number of markers and offers many advantages involved with improving the rate of genetic gain in crop breeding programs. solGS [38] and ISMU2 are two programs available for the calculation of GEBVs for the selection of individuals.

6 Integrated Pipelines for Plant Breeding Data Management

Data management plays a major role in creating a basis for sound scientific decision making, increased efficiency of resource use, and ultimately leads to enhanced research quality and reliability [39]. Data management software is not just a database but signifies appropriate experimental design, analysis, interpretation, archiving, and sharing of data. One of the biggest challenges for effective data management in public plant breeding is a lack of access to public data management systems to track samples, manage and analyze breeding data, and support breeding decisions. To overcome this hindrance, a few commercial software programs have been developed that offer breeding management systems; however, these come with an additional cost to the research organizations. Intensive crop improvement data demands a single integrated platform that can be used for data management, data mining, analysis, and sharing.

Many attempts have been made from both public and private sectors to provide advanced systems for data management. However, some of the systems have multiple features while others have specific applications [40]. Most importantly, the DMASTs need to evolve with the pace of volume and type of data generated in fast-evolving genetic and breeding methodologies. For this reason, currently no single data-management tool can be used for all the applications. Nevertheless, the scientific community is now well aware of such a need and soon there will be a few initiatives to work in this direction, for example, the development of the International Crop Information System (ICIS) (www.icis.cgiar.org) by the CGIAIR and partners, a database system for the management and integration of global information on crop improvement and genetic resources for any crop [41].

To efficiently manage the regular movement of data from lab to the breeder and to integrate information from genotyping and phenotyping, comprehensive crop-improvement data-management tools are required. To deal with the constraints in present-day data management, the Integrated Breeding Platform (IBP) (http://www.integratedbreeding.net), established by the CGIAR's Generation Challenge Program (GCP) and partners, offers a web-based frontline platform of technology and services for managing both traditional and modern breeding activities. From phenotyping to complex genotyping, it provides information, analytical tools, and related services to conduct modern breeding research. The Breeding Management System (BMS) of the IBP is an interconnected application specifically designed for managing breeding activities through all phases of research using various types of data management, statistical analysis, and decision support tools. Presently, the BMS is the only publicly available data management solution that supports various crops and has inbuilt international crop ontologies. The BMS is actively used by many CGIAR institutes including ICRISAT, CIAT, and IITA, with many more institutes adopting it. ICRISAT is one of the first centers to adopt it on an institutional scale and to implement it on a cloud. The BMS has an advantage of hosting several crops on one installation.

Breeding4Rice (B4R), is a breeding information management system at IRRI that provides an integrated, user-friendly information management system, developed using modern web technologies, and is deployed to a cloud infrastructure. The system is being extended to various other crops and will soon be available for maize and wheat. CassavaBase (https://www.cassavabase.org/) is an integrated information management system for breeding programs that deals with phenotyping, low-density marker, pedigree management, and selection decision support. Katmandoo (http://www.katmandoo.org/) is a data management system of biosciences primarily developed to be used by breeders and researchers in breeding programs. It is mainly focused on providing single tools for dealing with both phenotypic and genotypic data.

In addition to the above free and open-source databases, there are several commercial software solutions that are also available for handling the breeding data pipeline. As all systems are at the same stage of development, no clear-cut comparisons of these software programs are available. However, the authors of this chapter have experience in using a couple of them, and one major drawback that we observed is that once the user stops paying the annual renewal fee, there is no way one can even log in to the system and work with their past experiments. The first tool in this line is PRISM, a plant-breeding software solution (http://www.teamcssi.com/index.html) for plant researchers and agronomists. It provides user-friendly tools to manage breeding data. PRISM has been used by various public and private breeding institutions and is known for its flexible architecture. Another popular data pipeline is the Phenome One platform (http://phenome-networks.com/solutions/for-plant-breeders/). This platform supports all stages of the breeding process for field crops, horticulture crops, and ornamental plants. It is a web-based and user-friendly system, and also supports data analytics and integrated mobile application. Similarly, AGROBASE Generation II (http://www.agronomix.com) is a Windows-based agronomy software system. The CORE System

of AGROBASE Generation II offers data management and analytical tools for crop improvement. Progeny (http://www.progeno.net/software) is a Ghent University spin-off company that aims to empower plant breeders by providing access to breeding and selection methods. Several other platforms include Progen software (http://www.progeno.net/), which permits plant breeders to improve selection efficiency by incorporating phenotyping and genotyping data in the decision process. E-Brida (http://www.agripartner.nl/en-us/products/plantbreedingsoftware.aspx) is a breeding information system with several options for data recording and analysis. GeneFlow (http://www.geneflowinc.com) is a software program that provides a comprehensive tool for integrating pedigree, phenotype, and genotype data.

7 DMASTs for Data Sharing and Visualization

Research data are extremely valuable assets and resources, and good management of research data is essential for research excellence. It is essential to facilitate data sharing and ensure the sustainability and accessibility of data in the long term, and thus, their re-use for future science. This permits new and innovative research to be built on existing information, which is especially true for cases where public investment in research is to be realized. With well-organized and accurate research data we can get high quality research outputs and scientific discoveries based on evidence, while using less resources. With good data management practices and proper planning, researchers can benefit greatly, especially in saving cost and time.

Currently, many funding agencies ask for consideration of open data and data sharing for all research projects they fund, and impose research data requirements that focus on how data will be preserved and shared for public use after the project is completed. Scientific data have very important value beyond their use for the original research. Data sharing and visualization encourages scientific enquiry and debate, and promotes innovation, which may lead to new collaborations between data users and data authors, enhances the impact and visibility of research, can provide a direct credit to the researcher as a research output, and promotes the research that created the data and its outcomes. A critical part of making data findable, accessible, interoperable, and reusable with long-lasting usability is to ensure that it can be interpreted and understood by any user even in the future.

Several open-source tools are available for effective and efficient data sharing with different capacities. Data sharing helps in the reuse of existing data for new studies, which can result in innovations and new opportunities. There are many open-source data management tools available that can be used at an institute or project level. Dataverse (https://dataverse.harvard.edu/) is a research data storage and sharing platform developed by Harvard University, Cambridge, MA, USA, which is freely downloadable and can establish its own institutional open data repository. This platform is well integrated with R software modules and Geospatial map generation. Several CGIAR institutions have implementations of Dataverse and are using it as their primary data-sharing software. Dataverse is highly configurable

and can be queried through well-defined APIs. CKAN (https://ckan.org) is also an open-source data portal and data management solution that provides a streamlined way to make data discoverable and presentable with a rich collection of metadata, making it a valuable and easily searchable data catalog. Researchspace (https://www.researchspace.com) is a research management tool for Principal Investigators (PIs) and research team members of specific groups, to observe and manage lab workflows, capture, archive, organize, publish, and share the data. e!DAL (https://edal.ipk-gatersleben.de/) is a lightweight software framework for publishing and sharing research data, the main features being: version tracking, metadata management, information retrieval, an embedded HTTP(S) server for public data access, access to a network file system, and a scalable storage backend. DSpace (http://www.dspace.org) is the software of choice for academic, non-profit, and commercial organizations building open digital repositories. DSpace preserves an open access format for all types of digital content. Usually open access repositories are used for publishing digital content with more focus on long-term storage, access, and preservation. Fedora (http://fedorarepository.org) is a robust, modular, open-source repository system for the management and dissemination of digital content. It is especially suited for digital libraries and archives, for both access and preservation.

8 Breeder Requirements for Enhancing Genetic Gains

Enhancing genetic gains for crop improvement demanded several automated, integrated, straightforward, and easy to use pipelines. Based on several reports and publications, we have listed a few of the essential requirements from the breeders' perspective, which includes: (a) a pipeline to understand associations between phenotype and genotype, (b) high-throughput precision phenotyping, (c) a new web-based interface with better organization, (d) a trait ontology function inference as part of the data management pipeline, (e) better support from plant genomics, (f) better support for data analysis, and (g) integration from "omics" information. Based on the above requirements, the breeding pipeline should have a seamless interconnected analytical solution for different applications in crop improvement.

8.1 Pipeline to Understand the Association Between Phenotype and Genotype

The central challenge of modern data management tools are weak genomics to phenomics links. This also highlights the need for careful pipeline development and advocates for the inclusion of a robust and straightforward platform that can correlate between phenomics and genomics data seamlessly. For example, several CG centers (3,000 rice accessions from IRRI, Philippines, and 3,000 chickpea accessions from ICRISAT, Hyderabad) have generated a huge amount of genotyping/re-sequencing data. Multi-location phenotyping

data of such lines will provide meaningful results to the breeders if simple-to-use pipelines are available for understanding the association between phenotype and genotype. There exist pipelines that do part of this job and do not cycle through start to end. The current need is to bring efficiency to these tools and to link them to each other in order to undertake the huge phenotypic and genotypic datasets generated in breeding programs.

8.2 High-Throughput and Precision Phenotyping

The emphasis on high-throughput and precision phenotyping represents a significant change for breeders engaged in variety development who have traditionally favored simplicity, speed, and flexibility over sensitivity, precision, and accuracy. This is because historically the advantages of the latter could not be translated into an economically relevant genetic gain in a breeding context, and this is why easy, fast, and efficient phenotyping-capturing tools are the need at present. For example, PHENOME, Field Book, 1KK, and Coordinate are recent high-throughput phenotyping, software programs/Android apps that allow researchers to accumulate, categorize, and manage a large volume of phenotypic data using Android smartphones with barcode scanners or a Personal Digital Assistant (PDA) with a built-in barcode scanner. The collected data in the smart device could be easily transferred for data analysis in any operating system through the appropriate DMAST.

8.3 New Web-Based Interface with Better Organization

Many of the DMASTs or data management tools are stand-alone and can only be utilized through better infrastructure and with high IT skill manpower. Therefore, in the near future cloud-based, simple-to-use tools are required for breeders, which could be utilized on simple PCs. An advantage of such a web-based system will be that such tools can be used from any place or PC through a simple login with a user ID and password. The other major advantage of such a system is that the huge submitted phenotypic/genotypic datasets will be safer than those saved on standalone PCs.

8.4 Trait Ontology Inference as Part of the Data Management Pipeline

Trait ontology function should be an integral part of the data-management pipeline. This will be useful for the selection of diverse lines, for making new crosses, or for the development of new combinations of hybrids. This feature will be helpful for understanding the contributions of diverse parents in breeding lines, through their performance.

8.5 Better Support from Plant Genomics

Another important requirement from the breeders' perspective is better support from plant genomics scientists in the identification of trait-associated markers for complex traits, the selection of which is difficult in field conditions. Additionally, the development of a purity kit is important, not only for the purity of parental lines and hybrids but also for high-yielding varieties, so that the seed purity of the lines/varieties/hybrids can be tested in less time. Better GS prediction models with high prediction accuracy will also be useful for breeders for enhancing genetic gains through genomics interventions.

8.6 Better Support for Data Analysis and Investments

Meaningful and timely data analysis is the critical component of breeders' success in enhancing genetic gain. Most of the breeding trials and genotype-to-phenotype correlation requires specific DMASTs, and it is sometimes difficult for the breeders to use these tools in their breeding programs with limited infrastructure. Therefore, breeders require professional data analysis for analyzing complex datasets with specifically required tools. The information provided by such analysis of these huge datasets will be useful for making critical decisions in breeding programs. There is a need for strengthening investment in data analysis in breeding programs.

8.7 Integration from "Omics" Information

Besides genomics and phenomics, multiple studies have been conducted in other "Omics" fields in many crop plants. These "Omics" studies include transcriptomics, epigenomics, proteomics, and metabolomics. They will develop a better understanding of traits and generate meaningful information that can be used during plant selection in the field. Such integration of this information with DMASTs will increase the precision of decision making in plant selection.

9 Conclusion

This chapter discusses the status and future prospects of next-generation data management and analytical and decision support tools for crop improvement. We have presented a critical appraisal of different DMASTs and data management tools along with integrated pipelines. We have also presented the breeders' future requirements for enhancing genetic gains in terms of new required tools and easy-to-use

pipelines. We believe that the availability of GUI-based platforms with appropriated DMASTs will help breeders to make the best use of these tools in their breeding programs. Development and deployment of the right DMASTs at the right time will usher the crop improvement programs into a modernized knowledge-based crop improvement era towards sustainable crop production.

References

1. Rohlf FJ (1992) NTSYS-pc: numerical taxonomy and multivariate analysis system. Appl Biostat, ISBN 9780925031181
2. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725–2729
3. Xia X (2013) DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. Mol Biol Evol 30:1720–1728
4. Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J Hered 86:248–249
5. Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinforma 1:47–50
6. Liu K, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21(9):2128–2129
7. Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25(11):1451–1452
8. Crawford NG (2010) SMOGD: software for the measurement of genetic diversity. Mol Ecol Resour 10(3):556–557
9. Peakall PE, Smouse R (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics 28:2537–2539
10. Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1:174–181
11. Lander E, Green P (1987) Construction of multilocus genetic maps in humans. Proc Natl Acad Sci U S A 84:2363–2367
12. Stam P (1993) Construction of integrated genetic linkage maps by means of a new computer package: Join Map. Plant J 3:739–744
13. Voorrips RE (2002) MapChart: software for the graphical presentation of linkage maps and QTLs. J Hered 93:77–78
14. Fang Z, Polacco M, Chen S, Schroeder S, Hancock D, Sanchez H, Coe E (2003) cMap: the comparative genetic map viewer. Bioinformatics 19:416–417
15. Kozik A, Michelmore R (2006) MadMapper and CheckMatrix-python scripts to infer orders of genetic markers and for visualization and validation of genetic maps and haplotypes. In: Proceedings of the plant and animal genome XIV conference, San Diego. Abstract P957/CP013-http://www.intl-pag.org/14/abstracts/PAG14_C013.html
16. Wu Y, Bhat PR, Close TJ, Lonardi S (2008) Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. PLoS Genet 4:e1000212
17. Lander ES, Bostein DR (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage map. Genetics 121:185–189
18. Basten CJ et al (1994) Zmap-a QTL cartographer. In: Smith C, Gavora JS, Benkel B, Chesnais J, Fairfull W, Gibson JP, Kennedy BW, Burnside EB (eds) Proceedings of the 5th World Congress on genetics applied to livestock production: computing strategies and software, vol 22. The organizing committee, 5th World Congress on genetics applied to livestock production, Guelph, pp 65–66

19. Nelson JC (1997) QGENE: software for marker-based genomic analysis and breeding. Mol Breed 3:239–245
20. Li H, Ye G, Wang J (2007) A modified algorithm for the improvement of composite interval mapping. Genetics 175:361–374
21. Yang J, Hu C, Hu H, Yu R, Xia Z, Ye X, Zhu J (2008) QTLNetwork: mapping and visualizing genetic architecture of complex traits in experimental populations. Bioinformatics 24:721–723
22. Su C, Qiu X, Ji Z (2013) Study of strategies for selecting quantitative trait locus mapping procedures by computer simulation. Mol Breed 31:947–956
23. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959
24. Corander J, Marttinen P, Sirén J, Tang J (2008) Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinform 9(1):539
25. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909
26. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635
27. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575
28. Chagoyen M, Pazos F (2013) Tools for the functional interpretation of metabolomic experiments. Brief Bioinform 14:737–744
29. MacBeath G (2002) Protein microarrays and proteomics. Nat Genet 32:526–532
30. Falkner JA, Ulintz PJ, Andrews PC (2006) A code and data archival and dissemination tool for the proteomics community. Am Biotechnol Lab 24(5):28
31. Vizcaíno AJ, Côté RG, Csordas A, Dianes JA, Fabregat A, Foster JM, Griss J, Alpi E, Birim M, Contell J et al (2013) The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res 41:D1063–D1069
32. Wein SP, Côté RG, Dumousseau M, Reisinger F, Hermjakob H, Vizcaíno AJ (2012) Improvements in the protein identifier cross-reference service. Nucleic Acids Res 40:276–280
33. Waegele B, Dunger-Kaltenbach I, Fobo G, Montrone C, Mewes HW, Ruepp A (2009) CRONOS: the cross-reference navigation server. Bioinformatics 25:141–143
34. vanBerloo R (2008) GGT 2.0: versatile software for visualization and analysis of genetic data. J Hered 99:232–236
35. Milne I, Shaw P, Stephen G, Bayer M, Cardle L, Thomas WT, Flavell AJ, Marshall D (2010) Flapjack-graphical genotype visualization. Bioinformatics 26:3133–3134
36. Azam S, Rathore A, Shah TM, Telluri M, Amindala B, Ruperao P et al (2014) An integrated SNP mining and utilization (ISMU) pipeline for next generation sequencing data. PLoS One 9: e101754
37. Valente F, Gauthier F, Bardol N, Blanc G, Joets J, Charcosset A, Moreau L (2013) OptiMAS: a decision support tool for marker-assisted assembly of diverse alleles. J Hered 104:586–590
38. Tecle IY, Edwards JD, Menda N, Egesi C, Rabbi IY, Kulakow P, Mueller LA (2014) solGS: a web-based tool for genomic selection. BMC Bioinform 15(1):398
39. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, Manoff M, Frame M (2011) Data sharing by scientists: practices and perceptions. PLoS One 6:e21101
40. Xu Y (2010) Molecular plant breeding. CAB International, Nosworthy Way
41. McLaren CG, Bruskiewich RM, Portugal AM, Cosico AB (2005) The international rice information system. A platform for meta-analysis of rice crop data. Plant Physiol 139:637–642