J Inf Process Syst, Vol.14, No.3, pp.655~665, June 2018
https://doi.org/10.3745/JIPS.01.0029

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Information Technology Infrastructure for Agriculture Genotyping Studies

Bens Pardamean*, James W. Baurley*, Anzaludin S. Perbangsa*,
Dwinita Utami**, Habib Rijzaani**, and Dani Satyawan**

Abstract

In efforts to increase its agricultural productivity, the Indonesian Center for Agricultural Biotechnology and Genetic Resources Research and Development has conducted a variety of genomic studies using high-throughput DNA genotyping and sequencing. The large quantity of data (big data) produced by these biotechnologies require high performance data management system to store, backup, and secure data. Additionally, these genetic studies are computationally demanding, requiring high performance processors and memory for data processing and analysis. Reliable network connectivity with large bandwidth to transfer data is essential as well as database applications and statistical tools that include cleaning, quality control, querying based on specific criteria, and exporting to various formats that are important for generating high yield varieties of crops and improving future agricultural strategies. This manuscript presents a reliable, secure, and scalable information technology infrastructure tailored to Indonesian agriculture genotyping studies.

Keywords

Agriculture Biotechnology, Big Data, Genotyping, Infrastructure, IT

1. Introduction

In a 2009 report, United Nation's Food Agriculture Organization estimated that by the year 2050, the world's population will reach 9.1 billion. To sustain this substantially larger population, food production must increase by 70%. Simultaneously, there is a significant growth in the utilization of crops for bio-energy and other industrial purposes. Grain production per annum would need to rise to approximately 3 billion tons from the current 2.1 billion while meat production per annum would need a rise of over 200 million tons to reach 470 million tons [1]. The rapid growth in demand for agricultural products would subsequently place added pressure on already scarce agricultural resources. The agricultural sector would be competing against urban settlement expansions for land and water supplies while facing other major challenges, including adapting to and contributing to the mitigation of climate change, preservation of natural habitats, and the maintenance of biodiversity. To respond to these demands, developing countries such as Indonesia would need to increase 80% of its agricultural production through the amplification of crop yields from existing resources while the remaining 20%

Manuscript received November 19, 2015; first revision January 13, 2017; accepted January 28, 2017.
Corresponding Author: Bens Pardamean (bpardamean@binus.edu)
* Bioinformatics & Data Science Research Center, Bina Nusantara University, Jakarta, Indonesia ({bpardamean, baurley, aperbangsa}@binus.edu)
** Indonesian Center for Agricultural Biotechnology and Genetic Resources Research and Development, Bogor, Indonesia ({dnitaw-utami, habib, d.satyawan}@litbang.pertanian.go.id)

would originate from arable land expansion.

Promoting agriculture is the key to achieving food supply security. Thus, it is essential to take four actions: increase investment in agriculture, broaden access to food, improve governance of global trade, and increase productivity while conserving natural resources. The realization of the fourth action would require the provision of an array of technological options to farmers. This includes agricultural biotechnology, which represents a broad range of technology used in agriculture with regards to genetic improvements in crop and animal husbandry, the characterization and conservation of the genetic resources, and the management of plant and animal diseases [2]. Advances in genomics and molecular biology create opportunities to improve plant produce and live stocks through the identification of new genes. These improvements would then lead to information for the optimization of product use and seed performance [3]. Through this concept, agricultural biotechnology would contribute to food supply security through improvements in yield and quality. For maximum impact, a grassroots approach is necessary, i.e., small-scale farmers are able to adopt the technology seamlessly. Therefore, improvement approaches that are simple and low cost with little risk to humans and the environment are needed [4].

Indonesia possesses one of the few mega biodiversities on Earth, inducing the country into one of the richest sources of genes of interests for crop and animal breeding purposes. Tapping into this diversity can be done through the use of high-throughput next generation sequencing (NGS) and genotyping microarray technologies. Known genetic markers are used to identify potential genes that could be valuable for crop and animal improvement in breeding programs [5]. The genotyping studies rely on information technology (IT) infrastructure to store, transmit, and analyze big data, a term used to describe a massive volume of both structured and unstructured data [6]. Moreover, big data usually refers to a collection of datasets that are immensely large and complex that it becomes difficult to process using conventional database management tools and data processing applications [7].

Information technology is holding an increasingly crucial role in the efforts to achieve sustainable agricultural development; in particular, to provide appropriate and location-specific technologies and to furnish proficient advice for farmers in a timely manner. IT is also a valuable tool for improving agricultural research and education [8]. With the increased application of IT, more opportunities can be explored and more challenges addressed, leading to higher agricultural output and lower unit cost, and improved time-critical decisions [9]. The information and communication infrastructure development will increase the availability and transparency of information and reduce agricultural trade transaction costs [10].

There are many examples of potentially effective application and initiatives of IT in agriculture. However, there is still much that remains to be done. Several important emerging trends include [11]:

1. The convergence of media and tools for communication;
2. Increased web-based storage of agricultural information;
3. Inexpensive and improved connectivity for rural communities;
4. Increased government recognition of the importance of IT use in rural development;
5. Increased tailor-made, quality agricultural information services.

The “right” choices in technology development has been the result of interplay among many factors: scientific discoveries, evolution in business image and interests, fluid trends in consumer demand, government regulation, global citizens' movements, and the emergence of novel institutions and paradigms [12].

An information technology infrastructure designed for genotyping studies in agriculture has requirements in common with an infrastructure for handling big data, including fast processors; flexible, scalable storage and backup systems of the raw data; large memory to process data; methods to retrieve relevant variables for analysis; and sufficient bandwidth to transfer data among different sites and researchers. Therefore, with this study, an information technology infrastructure was designed specifically to address these problems for agricultural genotyping. The infrastructure is a flexible and scalable framework consisting of hardware paired with a database and software application. The database and application helps reduce programming and computing overhead while accumulating samples and variables. The application is connected to a centralized database, allowing researchers concurrent access to study data and the ability to share results in real time. The application also allows for the management and analysis of genotypes, traits, and annotations. The next section provides an overview of the system specification and descriptions of the hardware, database, and application.

2. System Requirements

The specifics system requirements include:

1. The integration of genotypes, trait, heredity, stress, and climate data for a diverse set of crops planted in various locations with distinct climate conditions across multiple seasons in a relational database and the linking of these data to other national and international resources. The implementation of integration requires the system to develop protocols and a web-based application for beneficiaries to interface to these data, to build powerful queries, and to expand data collection for multiple trials and crop varieties.

2. The performance of genome-wide and genome-environment-wide (GxE) analyses of complex traits in the diversity of crop varieties using appropriate generalized linear modeling.

3. The development and utilization of statistical methodologies to assess systematically the relationships among numerous factors (genetic, ancestry, climatic, soil, and agronomic factors) within relevant abiotic biological pathways. Numerous factors are comprehensively modeled using conventional and advanced multivariate techniques.

4. The combination of evidences from independent datasets and consortia as well as the evaluation of the replication and predictive properties of promising models.

5. The development of a result that provides prioritized lists of marginal and interaction associations, and well-fitting models for breeders to evaluate for marker-assisted selection programs; evaluation and advising on high-throughput/screening technologies and the design of follow-on trials; and the integration of findings into the database and web-based application, subsequently providing access to the toolkit to qualified crops and agro-genomic researchers.

3. System Specification

3.1 Overview

The large quantity of data includes thousands of potential study samples in the form of millions of genotype and phenotype variants. The management of the sheer volume of data calls for three

components to work together: hardware, database, and application. The first component, the hardware, comprised a fast processor and large memory workspace for data processing and analysis as well as reliable network connectivity with high bandwidth for data transfer. The next component, the database, employed a high performance data management system to store and backup raw data. The last component, the application, included customized software with specific features, such as a germplasm database application and statistical genetics algorithms.

Through the development of database applications, researchers can easily obtain data on genotypes and phenotypes with certain criteria that need further analysis. The database application workflow begins with data collection from the genotyping instrument. Then the data set is entered into the database in batches. Once in the database, researchers can filter and sort data according to criteria of interest. The results can be exported to a variety of data formats in accordance with the statistical tools used for analysis. With the application of this database tool, researchers can conduct studies whose results may be used for downstream agriculture research.

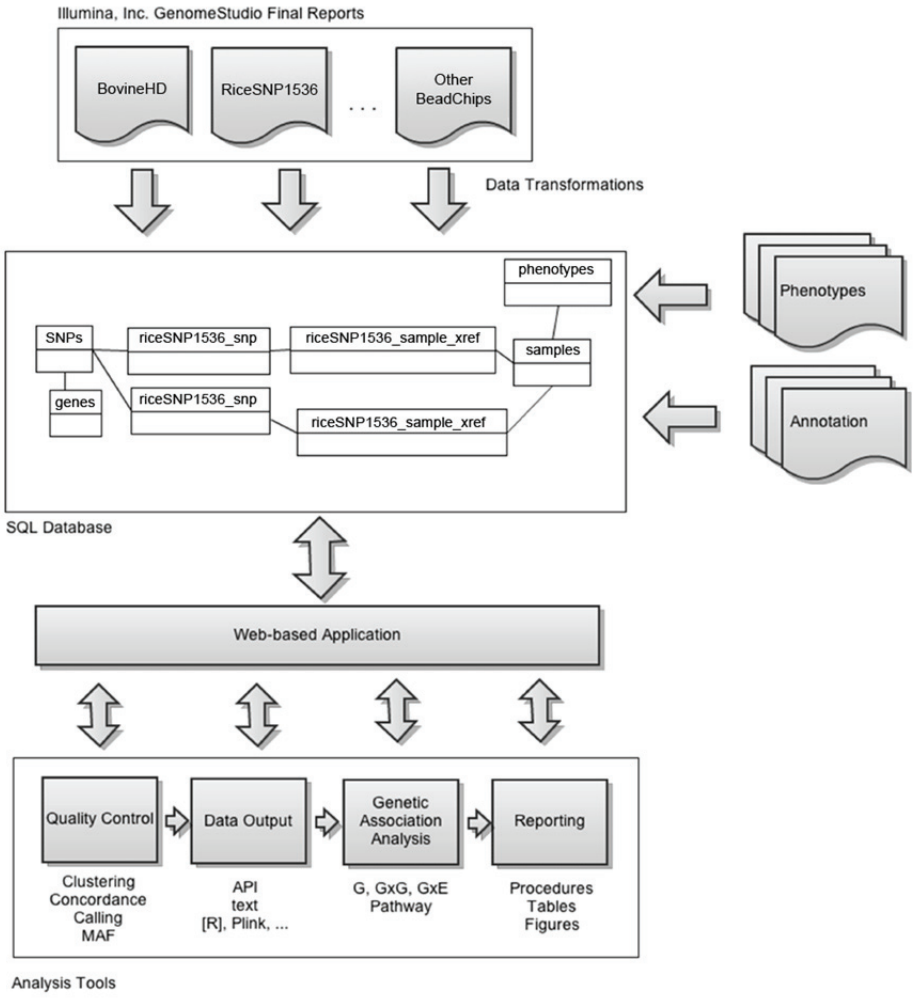

Fig. 1. System overview. From Baurley et al., International Journal of Bio-Science and Bio-Technology, vol. 5, no. 6, pp. 33-42, 2013 [13].

The Indonesian Center for Agricultural Biotechnology and Genetic Resources Research and Development (ICABIOGRAD) has recently deployed the database application to store genetic, trait, and annotation data, as seen in the System Overview shown in Fig. 1. Genetic data are the result of the process of sequencing and genotyping in the laboratory using biotechnology instruments while phenotypic data are obtained from laboratory measurements and field observations [13]. Prior to database input, each data is formatted, validated, and verified by scripts. For each species, the data are organized using schemas in the database. Each schema consists of SNP (single-nucleotide polymorphism), Sample, Final Report, and Phenotype tables. Registered users can access the database online.

With these database applications, users can perform: quality control, queries based on criteria of interest, data management, and data export to various formats for further analysis with statistical tools. Descriptive statistics are calculated for each trait, utilizing the statistics tool R by grouping them based on location and year. For continuous variables, descriptive statistics are shown as mean, median, standard deviation, and range. Furthermore, t-test or analysis of variance (ANOVA) is conducted to analyze continuous variables [14].

3.2 Hardware

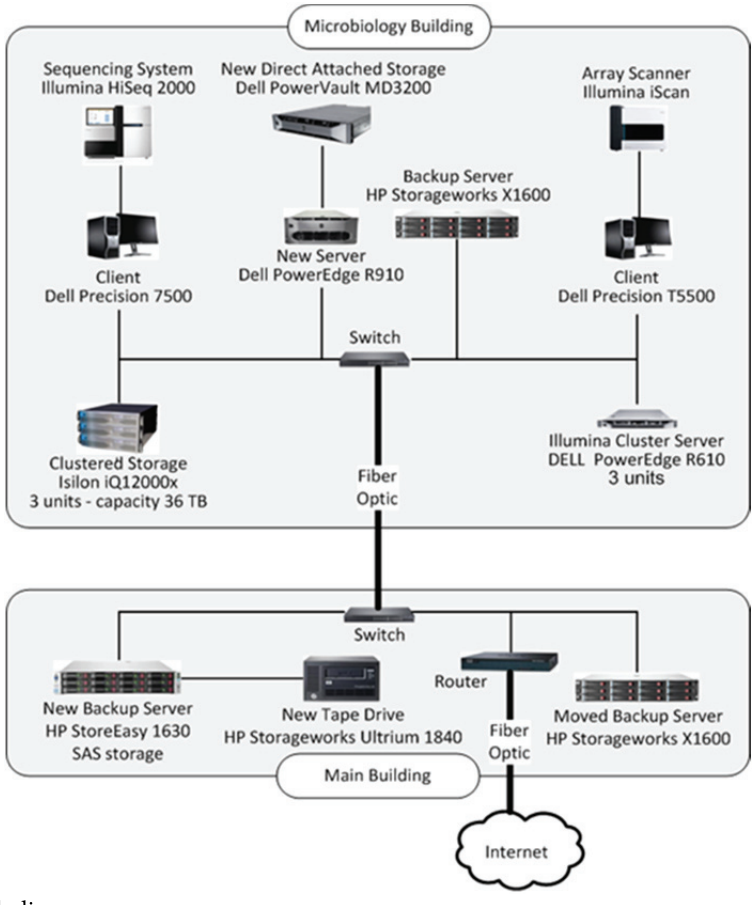

Fig. 2. Network diagram.

Fig. 2 illustrates the hardware and network configurations that were implemented. Illumina HiSeq 2000 Sequencing System and iScan Array Scanner with GenomeStudio software installed, produced, and stored raw data temporarily in the Dell Precision 7500 and Dell Precision T5500, respectively. On a regular schedule, the Dell PowerEdge R610 with Illumina Cluster Manager Technology automatically distributed and backed-up the data to three units of EMC Isilon IQ Storage Server 2000, each with a capacity of 12 TB (RAID 1). The backup process can also be conducted manually.

The data stored on the Isilon Storage were remotely synchronized to the backup servers located in the Microbiology Building and in the Main Building. Then the backup data in the Main Building were compressed to tar.gz and stored as tape media. Once the compressed file was successfully recorded into the tape, it can then be removed from the Isilon Storage. Since the two backup servers were synchronized, deleting the data manually on each server was not necessary.

Dell PowerEdge R910 was specifically used to conduct data analysis and was connected to the cluster to enable data exchange. Dell PowerVault MD3200 was a server for storing data analysis results from the Dell PowerEdge R910. HP Storageworks Ultrium 1840 Tape Drive was also used to backup data into the HP StoreEasy 1630 Storage media.

Gigabit unmanaged switches were employed to connect the rack server with system users and to accelerate data transfer to the storage system. Fiber optic cables were used as the network backbone to connect the Microbiology Building and the Main Building as well as between the Main Building and the Internet or VPN.

3.3 Database

The database was designed with batch data input and generated outputs in various formats for analytical tools such as R, TASSEL, PLINK, and SAS. Inputs to the system included genetic data from DNA genotyping or sequencing machines, and phenotypic/trait data that were measured in the greenhouse and the field. The rice germplasm database collected 1536-SNP and 384-SNP entries linked to genotype and phenotype data, which can be grouped by age and yield. The 384-SNP used International Rice Research Institute (IRRI)'s identification to cross-breeds Indica-Indica, Japonica-Japonica, and Indica-Japonica.

The data produced by array scanning software could be exported to a text file then entered into the database based on the entity-relationship diagram, as shown in Fig. 3. The SNP Map table with 12 variables contained array data such as position, polymorphic nucleotides, and other attributes. The DNA data samples were inserted into the Sample Map table with 8 variables. Phenotype data or traits were inserted into the Phenotype table with 17 variables, and connected to the Sample Map table with one-to-many relationships. The Final Report table stored genotypes data and provided call rate information for quality measure with 39 variables.

PostgreSQL was chosen as the database management system to store and retrieve study data because of its secure, scalable, and open source features. There are four datasets managed by the system. One bovine dataset and three rice datasets namely 1536, 384 ICABIOGRAD, and 384 IRRI. These were used for benchmarking purposes.

The bovine data set used the Illumina BovineSNP50 and BovineHD arrays, both of which contain evenly spaced SNPs spanning the entire bovine genome. The BovineSNP50 contains 54,609 SNPs with an average spacing of 49.4 kilobases (kb) and the BovineHD contains SNPs with spacing <3 kb. The

1536 rice dataset contains data for the purpose of genome-wide association analysis (GWAS). The 384 ICABIOGRAD rice dataset contains information that is used by ICABIOGRAD for characterization activities on Indonesian rice germplasm varieties based on its DNA profile. The final rice dataset, the 384 IRRI contains information of 384-SNP markers used by IRRI for a subset of activities that involve genotyping samples of Indonesian rice germplasm for benchmarking.

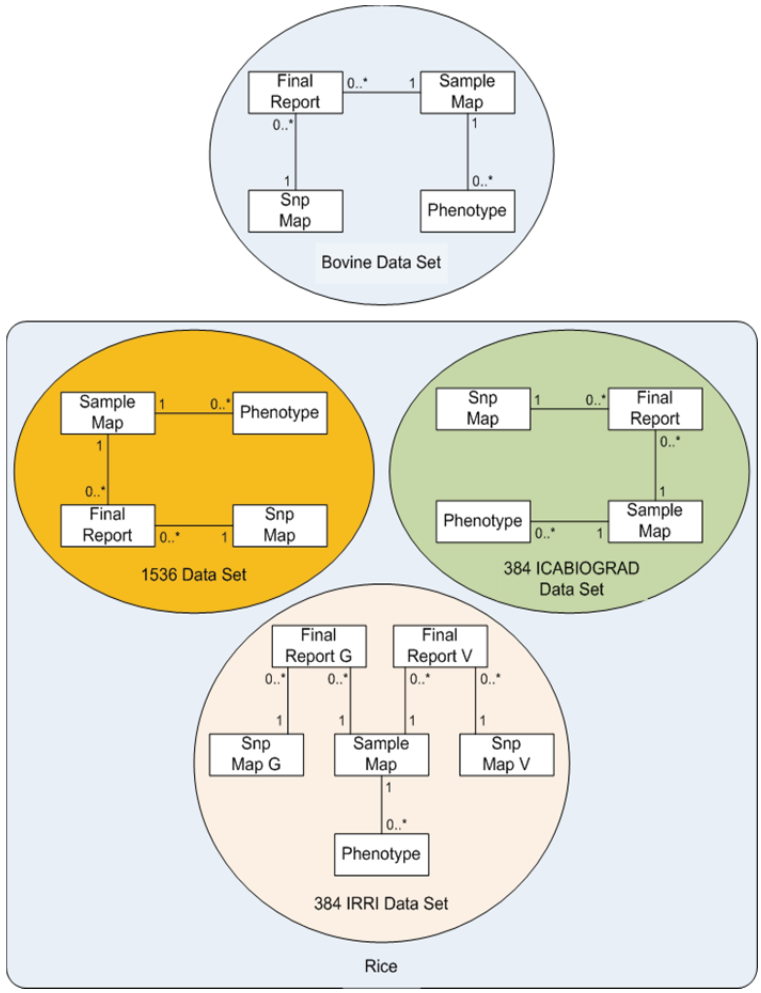

Fig. 3. Entity-relationship diagram.

3.4 Application

The application was developed using an iterative development process with frequent prototypes and feedback from agriculture researchers. Several important software design properties were built into the application including accessibility, security, usability, and scalability. The application may be accessed from web browser, meaning no specialized hardware or software is needed on-site to use the application. It can run on any operating system that supports a web-browser, including mobile devices. The specific system design and technology stack used in the application is described in Table 1.

Table 1. System design and technology stack [13]

Feature	Description
Security	Authentication, Authorization
System design	Web application
Web server	Apache 2.2.9
Server script	PHP 5.2.6
Client script	Javascript 1.7, HMTL 5, CSS 3
Database	PostgreSQL 9.2
Server operating system	Ubuntu 12.04.1 LTS
Client operating system	All
Statistical support	R 2.14

The database application users are divided into two groups, administrator and researcher. Administrator has the authority to determine who may use the application and limits users access rights. Researchers with access rights can manage, maintain quality, and export the data into a format that compatible with statistical software for further analysis. Use case diagram in Fig. 4 describes the key features contained in the database application.

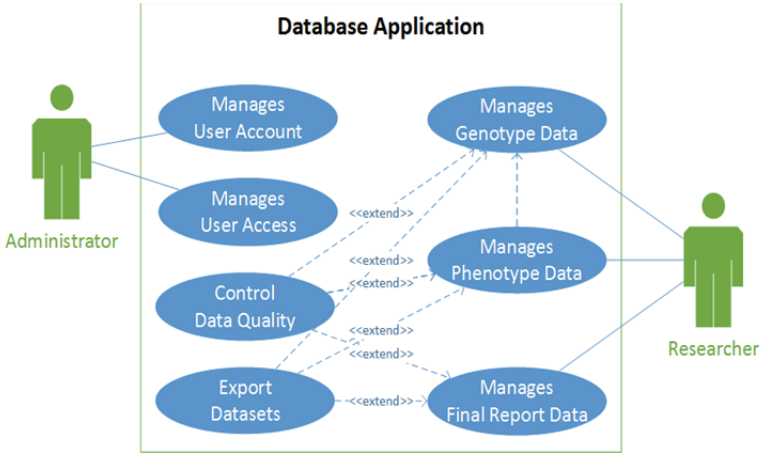

Fig. 4. Use case diagram.

To access the system, a user must have an account and his/her activities are limited by the access permission assigned to him/her. For instance, a user may be allowed access to data for only a particular species or a user can be granted access to view limited amount of data. Authentication and authorization is managed within the application for easy administration.

A user, once logged into the system, can select from amongst available species and studies. Once a species is selected, the genotype data are displayed and the user can begin filtering the results. A subset of columns in the final report is displayed in the web application (Fig. 5). The user can then filter the genotyping results by samples, SNPs, and common quality control variables. The entries of the "No", "SNP Name" and "Sample ID" columns are links to detailed information about the genotype, SNP, and sample respectively. Filtered data can be exported to various formats (e.g., delimited text, Excel) for

external analysis.

While the data imported from the array scanner cannot be modified in the user interface, the user may add or replace genotype data by importing new reports. Phenotype data linked to a DNA sample can be added or modified directly in the application as needed.

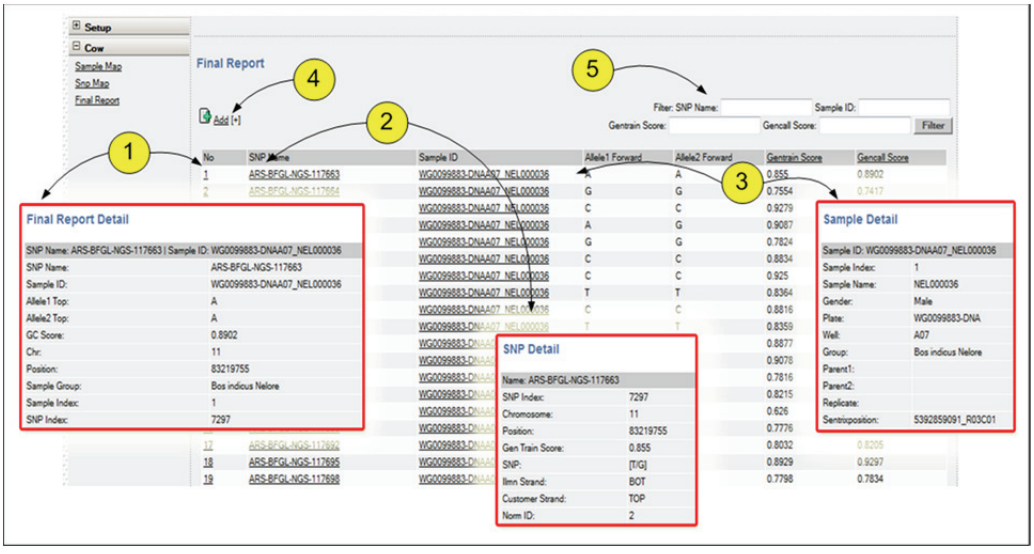

Fig. 5. Application user interface. From Baurley et al., International Journal of Bio-Science and Bio-Technology, vol. 5, no. 6, pp. 33-42, 2013 [13].

3.5 Analysis Tools

Various popular statistical programming environments can be used to analyze the genotype data on plant and animal diversity, such as R, PLINK, TASSEL, and SAS. For example, R has features for generating box-plot graphs and performing descriptive statistics that can be used to control the data quality based on certain variables. The need for a comprehensive statistical tool as well as user accessibility will continue to facilitate the integrated analysis of genotype data sets [15].

4. Conclusion

Developing the information technology infrastructure for Indonesian agriculture genotyping studies is a challenging task due to big data management. This infrastructure has been prepared based upon demands to deal with a large quantity of data produced every time a genome is typed. Data and processing are distributed across multiple processing units to provide data much faster. The database is maintained with PostgreSQL allowing ongoing analysis of the data. The web application also contributed to the management of big data by caching and compressing the requested data. The careful assessment of requirements to store, transmit, or analyze this massive volume of data has led to the implementation of scalable hardware, database, and web application to support agriculture genotyping studies in Indonesia.

Acknowledgement

This research was funded by the Indonesian Center for Agricultural Biotechnology and Genetic Resources Research and Development, Indonesian Agency for Agricultural Research and Development, Ministry of Agriculture of Indonesia.

References

[1] Food and Agriculture Organization, "FAO's Director-General on How to feed the world in 2050," *Population and Development Review*, vol. 35, no. 4, pp. 837-839, 2009.

[2] J. Ruane and A. Sonnino, "Agricultural biotechnologies in developing countries and their possible contribution to food security," *Journal of Biotechnology*, vol. 156, no. 4, pp. 356-363, 2011.

[3] J. K. Bull, A. W. Davis, and P. W. Skroch, "How smart-IT systems are revolutionizing agriculture," *The Bridge*, vol. 41, no. 3, pp. 14-21, 2011.

[4] D. Gahakwa, T. Asiimwe, N. Senkensha, J. Kajuga, P. Rukundo, E. Munganyinka, and J. Kahia, "Biotechnology for improving food security in Rwanda," *Rwanda Journal, Series E: Agricultural Sciences*, vol. 28, pp. 95-106, 2012.

[5] Ministry of National Development Planning/National Development Planning Agency, *Biodiversity Action Plan for Indonesia*. Jakarta, Indonesia: Ministry of National Development Planning/National Development Planning Agency, 1993.

[6] P. Kumar and K. Pandey, "Big data and distributed data mining: an example of future networks," *International Journal of Advance Research and Innovation*, vol. 1, no. 2, pp. 36-39, 2013.

[7] T. White, *Hadoop: The Definitive Guide*. Sebastopol, CA: O'Reilly, 2012.

[8] S. R. Zahedi and S. M. Zahedi, "Role of information and communication technologies in modern agriculture," *International Journal of Agriculture and Crop Sciences*, vol. 4, no. 23, pp. 1725-1728, 2012.

[9] K. C. Ting, T. Abdelzaher, A. Alleyne, and L. Rodriguez, "Information technology and agriculture: global challenges and opportunities," *The Bridge on Agriculture and Information Technology*, vol. 41, no. 3, pp. 6-13, 2011.

[10] S. Bojnec and I. Ferto, "Information and communication infrastructure development and agro-food trade," *Agricultural Economics*, vol. 57, no. 2, pp. 64-70, 2011.

[11] J. Stienen, W. Bruinsma, and F. Neuman, "How ICT can make a difference in agricultural livelihoods" in *Commonwealth Ministers Reference Book 2007*. London: Henley Media Group in conjunction with the Commonwealth Secretariat, 2007, pp. 2-4.

[12] P. J. Vergragt, "How technology could contribute to a sustainable world," The Tellus Institute, Boston, MA, 2006.

[13] J. W. Baurley, A. S. Perbangsa, A. Subagyo, and B. Pardamean, "A web application and database for agriculture genetic diversity and association studies," *International Journal of Bio-Science and Bio-Technology*, vol. 5, no. 6, pp. 33-42, 2013.

[14] J. W. Baurley, B. Pardamean, A. S. Perbangsa, D. W. Utami, H. Rijzaani, and D. Satyawan, "A bioinformatics workflow for genetic association studies of traits in Indonesian rice," in *Information and Communication Technology*. Heidelberg: Springer, 2014, pp. 356-364, 2014.

[15] J. D. Robin, A. T. Ludlow, R. LaRanger, W. E. Wright, and J. W. Shay, "Comparison of DNA quantification methods for Next Generation Sequencing," *Scientific Reports*, vol. 6, article no. 24067, 2016.

Bens Pardamean https://orcid.org/0000-0002-7404-9005

He earned a doctoral degree in informative research from the University of Southern California in 2007, as well as a master's degree in computer education and a bachelor's degree in computer science from California State University, Los Angeles in 1988 and 1994, respectively. He currently holds a dual-appointment as the Director of Bioinformatics & Data Science Research Center and as an Associate Professor of Computer Science at the University of Bina Nusantara (BINUS) in Jakarta, Indonesia.

James W. Baurley https://orcid.org/0000-0003-4116-7723

He earned a bachelor's degree in computer science from Clemson University, as well as a master's degree in biostatistics and a Ph.D. in statistical genetics and genetic epidemiology from the University of Southern California. He currently is an adjunct faculty and bioinformatics research consultant at BINUS University in Jakarta, Indonesia.

Anzaludin S. Perbangsa https://orcid.org/0000-0002-0343-8951

He earned a bachelor's degree in computer science as well as a master's degree in information system from BINUS University. Currently, he is a researcher at Bioinformatics & Data Science Research Center and a faculty member at School of Information System, BINUS University in Jakarta, Indonesia.

Dwinita Utami https://orcid.org/0000-0002-5873-0570

She received a bachelor's degree in biology from Gajah Mada University in 1992, a master's degree in biotechnology in 2000 and a doctoral degree in agronomy in 2005 from Bogor Agricultural University. She is currently a researcher at Indonesian Center for Agricultural Biotechnology and Genetic Resources Research and Development, Bogor, Indonesia.

Habib Rijzaani https://orcid.org/0000-0001-5981-1869

He received a bachelor's degree in biotechnology from University of Queensland, Australia in 1997 and a master's degree in biotechnology in 2000 from Bandung Institute of Technology. He is currently a researcher at Indonesian Center for Agricultural Biotechnology and Genetic Resources Research and Development, Bogor, Indonesia.

Dani Satyawan https://orcid.org/0000-0002-8166-7709

He received a bachelor's degree in biotechnology from University of Queensland, Australia in 1997 and a master's degree in biotechnology in 2000 from Bandung Institute of Technology. Currently, he is a researcher at Indonesian Center for Agricultural Biotechnology and Genetic Resources Research and Development, Bogor, Indonesia.