NABIC: A New Access Portal to Search, Visualize, and Share Agricultural Genomics Data

Young-Joo Seol, Tae-Ho Lee, Dong-Suk Park and Chang-Kug Kim

Genomics Division, National Academy of Agricultural Science (NAAS), Jeonju, Korea.

Libertas Academica
FREEDOM TO RESEARCH

ABSTRACT: The National Agricultural Biotechnology Information Center developed an access portal to search, visualize, and share agricultural genomics data with a focus on South Korean information and resources. The portal features an agricultural biotechnology database containing a wide range of omics data from public and proprietary sources. We collected 28.4 TB of data from 162 agricultural organisms, with 10 types of omics data comprising next-generation sequencing sequence read archive, genome, gene, nucleotide, DNA chip, expressed sequence tag, interactome, protein structure, molecular marker, and single-nucleotide polymorphism datasets. Our genomic resources contain information on five animals, seven plants, and one fungus, which is accessed through a genome browser. We also developed a data submission and analysis system as a web service, with easy-to-use functions and cutting-edge algorithms, including those for handling next-generation sequencing data.

KEYWORDS: agricultural genomics, NABIC, omics database

Introduction

The National Agricultural Biotechnology Information Center (NABIC, http://nabic.rda.go.kr) has played a leading role in coordinating genome biotechnology efforts for agricultural species in Korea since 2002. Recent technological advances in next-generation sequencing (NGS) and transcriptomics have led to an accumulation of "-omics" and functional genomic data.[1] In response, we have implemented a national policy to manage these data through the Next-Generation BioGreen 21 Program and the postgenome project, Agriculture Science Technology Information System (http://atis.rda.go.kr/).

Many genome portals that provide valuable data resources for biotechnologists are available. Some databases provide primary data and offer integrated views of different data types, allowing the user to easily perform customized queries over large datasets and compare different types of data.[2] Several integrated systems for agricultural data and resources are available, including the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/), the European Bioinformatics Institute (EBI, http://www.ebi.ac.uk/), Beijing Genomics Institute (http://www.genomics.cn/), and DNA Data Bank of Japan (http://www.ddbj.nig.ac.jp/). Resources particularly relevant for the present work include the International Nucleotide Sequence Database Collaboration (http://www.insdc.org), which provides public domain nucleotide sequence information; ExPASy (http://www.expasy.org/), which provides access to proteomics, genomics, and systems biology databases; and Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/), a database resource for investigating pathways in biological systems.

There are several plant agricultural databases. Gramene (http://www.gramene.org) is a curated online resource for comparative functional genomics in crops and model plant species. The database hosts annotated whole genomes of more than two dozen plant species and partial assemblies for almost a dozen wild rice species in the Ensembl browser.[3] MaizeGDB (http://www.maizegdb.org) is a highly curated, community-oriented database and informatics service for researchers focused on the crop plant and model organism *Zea mays* ssp. *mays*.[4] GrainGenes (http://wheat.pw.usda.gov/) is a comprehensive resource for molecular and phenotypic information on Triticeae and *Avena*, including wheat, barley, rye, and oat. The website hosts a database that includes genetic maps, genes, alleles, genetic markers, phenotypic data, quantitative trait loci studies, experimental protocols, and publications.[5]

Current databases do not specialize in comprehensive agricultural species or in South Korean resources. Here, we describe a platform for omics research of agricultural organisms that specializes in South Korean resources. This database can be utilized to identify region-specific characteristics of biological mechanisms and generate evolutionary insights, all of which can be accessed through a simple and intuitive interface.

Materials and Methods

Data collection. Agricultural biotechnology information was collected from the Rural Development Administration (http://www.rda.go.kr/), National Institute of Agricultural Sciences (http://www.naas.go.kr/), National Institute of Crop Science (http://www.nics.go.kr/), National Institute of Horticultural and Herbal Science (http://www.nihhs.go.kr/), National Institute of Animal Science (http://www.nias.go.kr/), the genetic resources project (http://www.genebank.go.kr/), seven centers affiliated with the Next-Generation BioGreen 21 Program (http://atis.rda.go.kr/), and other universities and institutes in Korea. Genomic information was collected from several collaborative and public institutes, such as NCBI, PlantGDB, and the International Rice Genome Sequencing Project (http://rgp.dna.affrc.go.jp/IRGSP/). We also integrated reference data from public database, such as Ensembl, NCBI RefSeq, and NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/).

The NABIC portal (http://nabic.rda.go.kr/) is the official national management and certification center for government-funded biotechnology research projects. We receive data submissions and perform data quality checks, storage, and management. We also issue official certifications for all data products. Gene and genome data from public databases are updated according to related in-house data. For quality control, we employed two software programs to validate our data: FastQC 0.11 (Babraham Bioinformatics, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and FastX-Toolkit (Hannon laboratory, http://hannonlab.cshl.edu/fastx_toolkit/). The data architecture was designed using several open standard protocols and dual networks (Fig. 1).

Database. We employed the BioSQL schema supporting Oracle 10 g RDBMS (http://www.oracle.com) to construct a standard database covering public and private platforms, which are derived from NCBI and exclusively in-house data. Data verification was performed by checking data for accuracy and eliminating inconsistencies after data migration. Then, data were validated using information from the project documentation. External databases are regularly synchronized with our database pipeline for the maintenance of up-to-date information. Users can search for information through an HTTP RESTful interface and web URL address, with security ensured by OAuth. The underlying platform was constructed on MySQL Enterprise 5.5, IBM General Parallel File System (GPFS), Red Hat Enterprise Linux 6.1, and RESTful architecture.

Website. We developed a web-portal system to enable searches for agricultural omics data and provide services, such as NGS assembly and genome-wide association studies (GWAS), as well as for differential expression, microbial community, and systems biology analyses (Fig. 2).

This web-based system integrates several open-source software projects and allows users to search multiple databases with a single query. A system diagram for NABIC website hardware and software is presented in Figures 3 and 4.

Results

Database for meta-omics information. The collected meta-omics database consists of 10 data types (ie, NGS Sequence Read Archive [SRA], genome, gene, nucleotide, DNA chip, expressed sequence tag [EST], interactome, protein structure, molecular marker, and single-nucleotide polymorphism [SNP]) under six schema categories (ie, European Nucleotide Archive/Sequence Read Archive [EBI/SRA], NCBI/GEO, dbEST, BioSQL, HUPO/PSI-MI, and Protein Data Bank [PDB]) for 162 agricultural organisms. The six schema categories are defined by 10 data types of collected meta-omics data.[6–10] Figure 5 shows the relationship between our local data and the models of the incorporated public databases, including the relationships between the 10 data types and six schema categories.
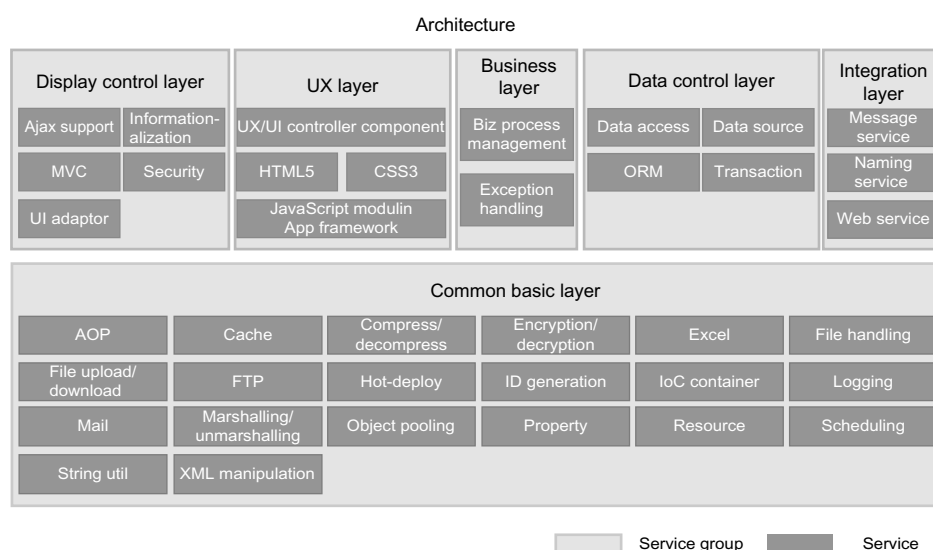
Figure 1. NABIC database system architecture contains six layers and was designed using various open standard protocols and a dual network.

Figure 2. A snap shot of the NABIC portal, which shows eight information menus for agricultural species. The website is organized into introduction and agricultural news at the top, and South Korean-native agricultural organisms listed to give more genomic information in the middle area. The omics databases and submission status are shown at the bottom.

- GAS: Genome annotation server
- NAS[1]: NGS anlysis server
- OGS*: Open grid system

- NAS[2]: Network attached stirage
- WAS: Weg application server
- LDAP: Lightweight directory access protocol
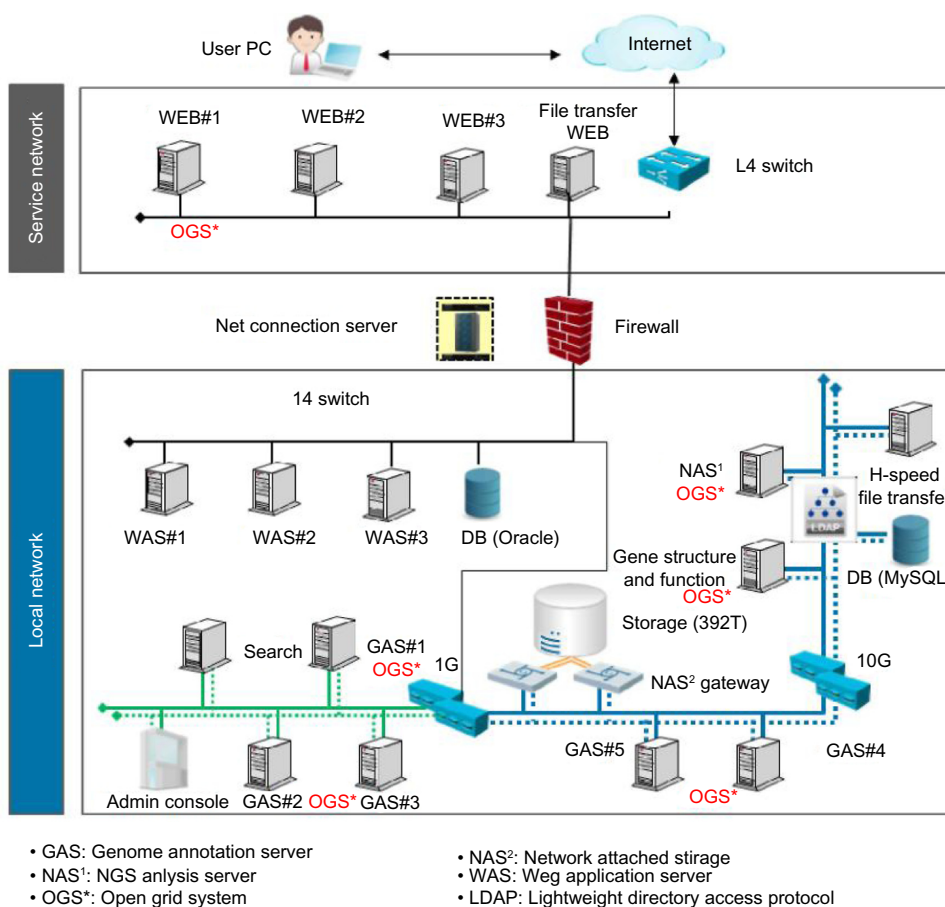
Figure 3. System diagram for hardware architecture of the NABIC database. The NABIC system consists of NGS analysis, genome structure/function analysis server, and high-speed file transfer, with underlying GAS server, NGS analysis server (NAS[1]), OGS, WAS, and network-attached storage (NAS[2]).
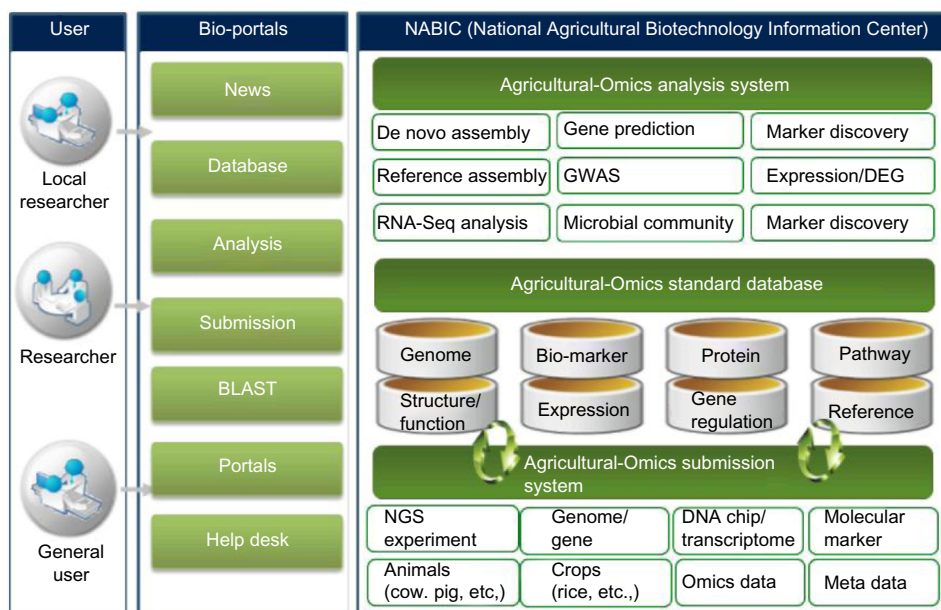
Figure 4. System diagram for software architecture of the NABIC database, which was classified using various analysis functions.

The NABIC database has three major characteristics: (1) updates are periodically performed for data from public databases and our local genome data, (2) all data records are linked with corresponding URL addresses, enabling users to access the database through a RESTful interface for building scalable web services,[11] and (3) users can perform specific queries using the Application Programming Interface and can search and perform analysis using keywords and the Basic Local Alignment Search Tool (BLAST) program. The database contains 2,938,025 records totaling 28.4 TB (Table 1) and consists of 10 data types in four categories (ie, biosequence, transcriptome, proteome, and variation).

Genome research. The NABIC database consists of data for 13 species of selected animals (cow, dog, pig, chicken, and goat), plants (Chinese cabbage, rice, grape, soybean, radish, maize, and chrysanthemum), and a fungus (*Flammulina velutipes*) and provides a genome browser embedding GBrowser V0.4.0 from the public NCBI and PlantGDB databases.[12]

This browser includes a genome description and detailed information on genes for each chromosome. Users can visualize a particular chromosomal region by selecting a tab at the top of the browser to further annotate functional units, and all these regions can be saved, shared, and compared with the user's data. We also provide an SNP marker database, genetic map, and Bacterial Artificial Chromosome (BAC) sequencing list, similar to those provided in the *Brassica rapa* project[13] and the International Rice Genome Sequencing Program.[14] We continuously update genome research data from agriculture-related organisms after an internal review process (Supplementary Fig. 1).

Data submission system. To enable effective sharing and review of data from government-funded research projects, we developed a submission system for agricultural omics data as part of the NABIC portal. We categorized omics data into one of the 10 metadata types: NGS SRA, genome, gene, nucleotide, DNA chip, EST, interactome, protein structure, molecular
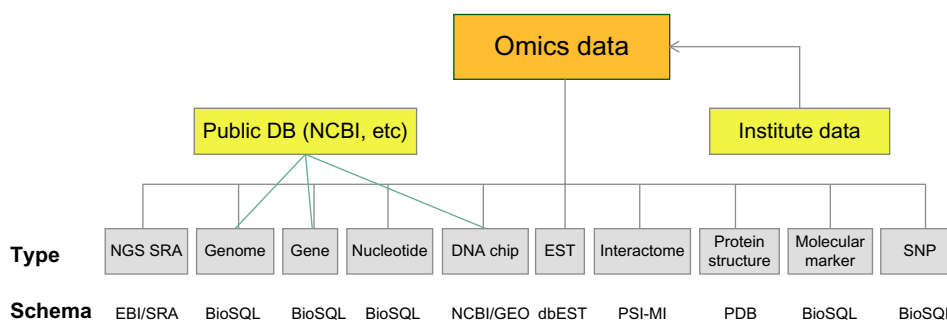
Figure 5. Structure of agricultural omics database in the NABIC portal. The NABIC database contains several types of omics data derived from public databases and in-house data. The data schema primarily uses BioSQL schema and schema from SRA, GEO, PSI-MI, and dbEST.

**Table 1.** Data category and statistics of agricultural omics database.

CATEGORY	TYPE	SCHEMA	NUMBER OF ENTRIES	SIZE (GB)
Biosequence	NGS SRA	EBI/SRA	2,208	29,081.6
	Genome	BioSQL	474,580	5.8
	Nucleotide	BioSQL	4,987	2.2
	Gene	BioSQL	163,673	1.1
Transcriptome	DNA chip	NCBI/GEO	41	4.9
	EST	dbEST	2,230,407	2.0
Proteome	Interactome	PSI-MI	1	0.0
	Protein structure	PDB	13	0.1
Variation	Molecular marker	BioSQL	7,805	10.0
	SNP	BioSQL	54,310	0.0
Total			2,938,025	28.4TB

marker, or SNP. All submitted data are standardized and integrated into the omics database. The submission menu provides the quality management activities using multivalidation steps. Finally, data collected from users are stored in the NABIC. Our system allows individual researchers to use a high-throughput protocol (InnoEX solution, http://www.innorix.com/) during data submission and retrieval, which is faster than common FTP. Figure 6 shows the flow chart for the data submission process.

Data analysis system. Researchers can analyze 10 different types of omics data from in-house sources and user uploads using a wide range of analytical approaches. We focused on developing web-based systems for the analysis of NGS, GWAS, systems biology, gene expression, and microbial community data. All these followed the *Korean e-Gov Standard Framework*, with user interfaces for

tracking the progress of analyses, displaying results, and downloading data.

Genome analysis. We organized a variety of open-source tools for NGS data analysis in four categories: genome assembly, RNA sequencing (RNA-seq), gene prediction, and variant discovery. In the genome assembly pipeline, we provide de novo assembly for species that do not have available reference genomes. We adopted several algorithms for this purpose, including Velvet, SOAP de novo, and CLC de novo assembler. FASTA, SFF, and Illumina's FASTQ are supported file formats for de novo assembly. For example, the CLC de novo assembler (CLC bio, a QIAGEN Company) offers comprehensive support for a variety of data formats, including both short and long reads, and mixing of paired reads, such as those with different insert sizes and orientation (Supplementary Fig. 2). For assembly with reference data, we
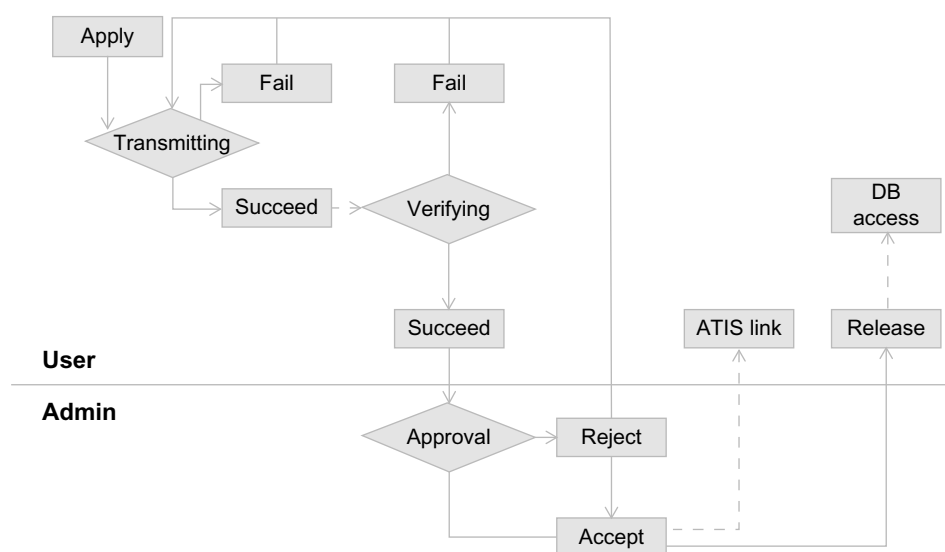
Figure 6. Process of data submission system in the NABIC. The NABIC submission system provides requirements for conducting quality management activities for all data collection. Data submitted to NABIC are only available to the public after sufficient validation steps. Finally, data collected from users are stored in the NABIC.

implemented algorithms, including Bowtie, BWA, MAQ, LASTZ, and CLC assembler. Supplementary Figure 3 shows reference assembly processing using the Bowtie2 and SAMtools methods.

For the RNA-seq analysis pipeline with NGS data, we used TopHat to align RNA-seq reads derived from a de novo or reference assembler, then Cufflinks to assemble and estimate the relative abundance of transcripts. In addition, Cuffmerge (for the Cufflinks assemblies), Cuffcompare (for the comparison of multiple experiments), and Cuffdiff (for the identification of significant changes in transcript quantity and structure) are available in the RNA-seq analysis pipeline (Fig. 7).

We implemented a gene prediction tool for three organisms: rice, human, and *Arabidopsis*. FASTA format is supported as an input data type. We use FGENESH, AUGUSTUS, and GlimmerHMM tools for accurate and comprehensive prediction. For the discovery of variants, we make SAMtools available to identify primary SNPs, and SpliCQ to report splicing events using NGS data (Supplementary Table 1).

GWAS. GWAS are a popular approach for identifying genomic variation underlying valuable traits in agricultural organisms, such as crops and livestock. We integrated a range of bio-agricultural data, including markers, traits, Quantitative Trait Locus (QTLs), and linkage information, and exploited the PLINK v1.07 toolset to enable comprehensive GWAS in a web environment. The steps in this analysis include (1) SNP array quality control by analysis of minor allele frequencies, Hardy–Weinberg equilibrium tests, and calling rates and (2) linkage disequilibrium tests using uploaded genotype data, or association tests with corrections for multiple tests, to reveal traits of interest in phenotype data (Supplementary Fig. 4).

Microbial community. Agricultural environments contain microorganisms that can significantly affect the ecology of crops and livestock. Generally, the microbial environment is studied at the population level. For microbial population analysis, our system uses the open-source platform mothur v1.32.1 and supports a broad range of analyses and data formats, such as SFF and FASTA. Users can upload their own samples or retrieve data from our database and assign taxonomic units by sequence alignment and clustering and identify operational taxonomic units using various approaches. In addition to performing taxonomic- and operational taxonomic unit-based analysis, our pipeline allows data preprocessing, including denoising and chimera removal, to ensure data quality. Comparisons among different microbial communities are available through the generation of Venn diagrams and parsimony-based structural similarity analysis (Supplementary Fig. 5).

Marker analysis. We developed a molecular marker analysis pipeline to help breeding researchers. Before executing a run, users import reads and assign parameters for mapping, SNP-calling, and primer design. Users can check the results of marker analysis according to job title, data type, analysis status, reference genome, or read name. Detailed information on the analysis is provided with an organized report, and several filters are available to identify SNPs within a particular chromosome or restriction enzyme. All results are prepared for download in Excel format (Supplementary Fig. 6).
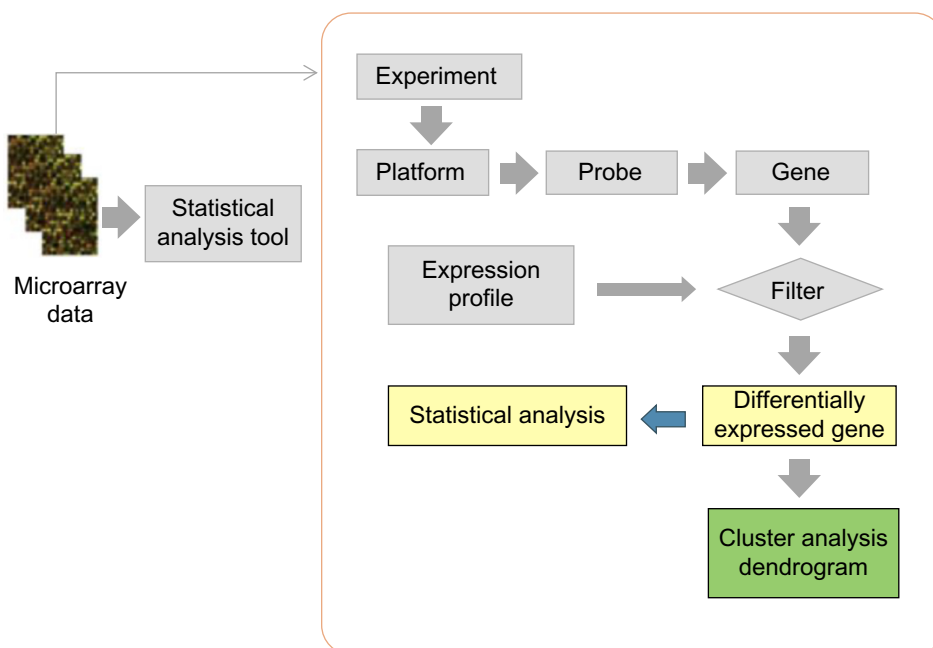
Differential expression analysis. Microarrays and RNA-seq are increasingly used for gene expression profiling. Analysis of these data is a major challenge, and development of statistical and computational methods is essential for drawing meaningful conclusions from these large datasets. We used two methods for detecting differentially expressed genes, the *t*-test and Wilcoxon rank-sum test, which are performed using edgeR 3.12 (http://bioinf.wehi.edu.au/edgeR/) in a Bioconductor package.[15] The user can choose corrections for multiple tests, including the False Discovery Rate (FDR) and Bonferroni methods. Input files are in matrix format with intensity values, and the results of an analysis can be reused. This analysis menu has six essential features, namely: (1) introduction and help sections, (2) sample information, (3) information on a series of samples for a specific experiment, (4) information on expressed genes, (5) Differential Expression Analysis (DEG) analysis and gene expression profiles, and (6) display of results with hierarchical clustering and dendrograms (Fig. 8).

News for agricultural biotechnology. As a web-based information portal in the field of agricultural biotechnology, our system provides up-to-date daily agricultural information, including general news, research news, and patent updates. The portal also includes rich site summary feeds, categorized and indexed literature, and advanced search options.

Portals (other Databases [DBs]). We offer three additional resource databases through the portals section: allergen, Ds-tagging rice, and wheat glutenin proteomics. The allergen database provides information on allergen characterization, which includes allergen structure and epitopes. A total of 2,939 allergens are registered and organized into 13 categories
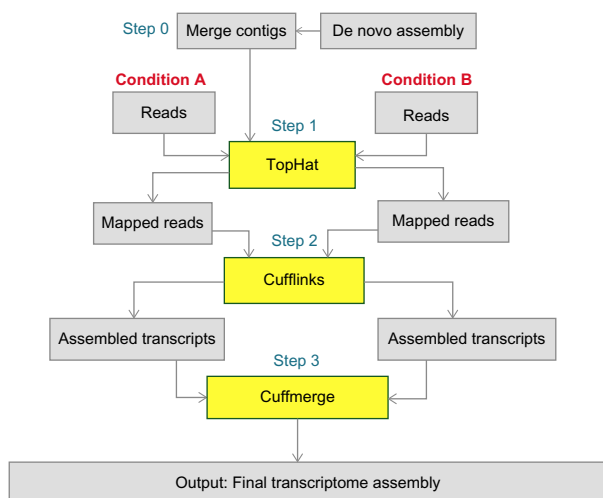
Figure 7. Pipeline of RNA-seq analysis using NGS data. NABIC RNA-seq analysis pipeline uses TopHat to align RNA-seq reads, Cufflinks to assemble and estimate the relative transcript abundance, and Cuffmerge for Cufflinks assemblies.

Figure 8. Pipeline of differential expression analysis. NABIC provides information on sample, experiment, and differentially expressed genes by performing hierarchical clustering and generating a dendrogram.

for animals, microbes, and plants. This system enables users to search for allergens and provides three computing methods for predicting allergenicity. The Ds-tagging rice database provides comprehensive information about mutant phenotypes and insertion-site sequence information for Ds-tagging lines that have been generated using japonica rice (*Oryza sativa* cv. *Dongjin*). The wheat glutenin database provides qualitative and quantitative expression levels of two glutenin proteins, such as High Molecular Weight Glutenin Subunit (HMW-GS) and Low Molecular Weight Glutenin Subunit (LMW-GS), using two-dimensional gel electrophoresis and Liquid Chromatography-tandem Mass Spectrometry (LC-MS/MS) analyses of 30 Korean wheat cultivars.

Discussion

The NABIC portal was established in 2002 for the purpose of analyzing agricultural genomes and providing related services to professional genomic research institutes and societies. Our continually updated omics-based database provides information through a user-friendly web interface that allows users to search for genetic resources and analyze large genome information datasets. This article described the development of the NABIC portal website and database, which integrates genome information of major agricultural organisms, omics data, agricultural news, and data from previously developed databases (Supplementary Table 2).

The NABIC database is a unique resource that facilitates international and in-house multiomics agricultural research and discovery. The database is a comprehensive agri-information portal and an easy-to-use analysis pipeline that easily processes the copious amounts of in-house data.

Our new database utilizes an informatics approach to agricultural biotechnology and can be extended to breeding studies for new crop cultivars. We believe that the NABIC database is a valuable resource for research on specific characteristics of South Korean agri-organisms and evolutionary history. The database includes a number of unique in-house resources, and a considerable amount of data deposited in the NABIC portal is not available in other public databases, such as NCBI and EBI. The NABIC portal will be upgraded to improve the availability of agriculture-related genomic data.

Our goal was to provide revolutionary technologies that deliver genomic information quickly and inexpensively. In 2011, we constructed a system for NGS technologies to analyze massive sequencing datasets. The system provides a range of information and tools for genomic analysis, including de novo assembly, reference assembly, RNA-seq, GWAS, microbial community analysis, and differential expression analysis. We serve the livestock genomics research community with genome data repositories for animal breeding, which is primarily aimed at understanding biological mechanisms related to traits of economic value.[16] For livestock genomics, we provide analysis tools and methods that enable researchers to optimally utilize available resources and effectively share, combine, manage, and analyze data from animal genomics/genetics studies.

To develop the NABIC portal system, we focused on integrating in-house and public genomic data for a number of agriculturally important organisms. The genomic data include genome, nucleotide, DNA chip, and transcriptome. This information is easily accessible through the upper-right search bar and omics database, and genome data from major agricultural resources

are available via GBrowser.[17] The portal also provides BLAST search capability for the in-house database, current agricultural research news, and a brief introduction to omics data analysis and database submission protocol. All resources are available to the public through the NABIC portal website (http://nabic.rda. go.kr). For the present time, analysis tools, and data submission are only available to domestically registered users.

The NABIC portal has contributed to the development of informatics approaches for agricultural biotechnology to support breeding programs for new crop cultivars. We will continue to help agricultural researchers by providing a continually updated genome-based database and bioinformatics tools to solve complex biological problems.

Conclusion

We have developed the NABIC portal, which is an updated collection of agricultural omics data derived from RDA-supported research institutes. We implemented an online resource that allows users to search, view, and download genome and genetic data. The NABIC portal provides wide availability of NABIC resources. Our goal is to develop and encourage the adoption of novel informatics approaches in agricultural biotechnology and to support both molecular and conventional breeding programs in the development of new cultivars of crops and livestock.

Author Contributions

Developed and wrote the code for the NABIC portal: C-KK, Y-JS. Composed the manuscript: C-KK, Y-JS. Advised on the design and features of NABIC portal, provided overall scientific and technical guidance, and assisted with manuscript creation: T-HL, D-SP. All the authors contributed to writing and improving the manuscript, and all the authors have read and approved the final version.

Supplementary Materials

Supplementary Table 1. Algorithm sets used in the NGS analysis pipeline.

Supplementary Table 2. Functional characteristics of NABIC system by development year.

Supplementary Figure 1. Snap shot of genome research.

Supplementary Figure 2. De novo assembly process using Velvet, SOAP de novo, and CLC de novo.

Supplementary Figure 3. Reference assembly process using Bowtie2 and SAMtools.

Supplementary Figure 4. Genome-wide association study process using the PLINK program (http://pngu.mgh. harvard.edu/~purcell/plink/).

Supplementary Figure 5. Microbial community analysis using the mothur program (http://www.mothur.org/).

Supplementary Figure 6. Snap shot of the marker search menu in the NABIC portal.

REFERENCES

1. Metzker ML. Sequencing technologies – the next generation. *Nat Rev Genet*. 2010; 11(1):31–46.
2. Kodama Y, Shumway M, Leinonen R. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res*. 2012;40(D1):D54–6.
3. Monaco MK, Stein J, Naithani S, et al. Gramene 2013: comparative plant genomics resources. *Nucleic Acids Res*. 2014;42(Database issue):D1193–9.
4. Andorf CM, Cannon EK, Portwood JL, et al. MaizeGDB update: new tools, data and interface for the maize model organism database. *Nucleic Acids Res*. 2016; 44(D1):D1195–201.
5. O'Sullivan H. GrainGenes. *Methods Mol Biol*. 2007;406:301–14.
6. Boguski MS, Lowe TM, Tolstoshev CM. dbEST – database for "expressed sequence tags". *Nat Genet*. 1993;4(4):332–3.
7. Whitfield EJ, Pruess M, Apweiler R. Bioinformatics database infrastructure for biotechnology research. *J Biotechnol*. 2006;124(4):629–39.
8. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets – update. *Nucleic Acids Res*. 2013;41(D1):D991–5.
9. Orchard S. Data standardization and sharing – the work of the HUPO-PSI. *Biochim Biophys Acta*. 2014;1844(1):82–7.
10. Rose PW, Prlić A, Bi C, et al. The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res*. 2015; 43(D1):D345–56.
11. Rauf I, Porres I. Designing level 3 behavioral RESTful web service interfaces. *ACM SIGAPP Appl Comput Rev*. 2011;11(3):19–31.
12. Duvick J, Fu A, Muppirala U, et al. PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res*. 2008;36(Database issue):D959–65.
13. Wang X, Wang H, Wang J, et al; Brassica rapa Genome Sequencing Project Consortium. The genome of the mesopolyploid crop species *Brassica rapa*. *Nat Genet*. 2011;43(10):1035–9.
14. Project IRGS. The map-based sequence of the rice genome. *Nature*. 2005;436(7052): 793–800.
15. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139–40.
16. Wu XL, Beissinger TM, Bauck S, et al. A primer on high-throughput computing for genomic selection. *Front Genet*. 2011;2:4.
17. Donlin MJ. Using the generic genome browser (GBrowse). *Curr Protoc Bioinformatics*. 2009;Chapter 9:Unit99.