Review article

Translational genomics for plant breeding with the genome sequence explosion

Yang Jae Kang¹, Taeyoung Lee¹, Jayern Lee¹, Sangrea Shim¹, Haneul Jeong¹, Dani Satyawani¹,2, Moon Young Kim¹ and Suk-Ha Lee¹,3,*

¹Department of Plant Science Research Institute of Agriculture and Life Sciences, College of Agriculture and Life Sciences, Seoul National University, Seoul, Korea
²Indonesian Center for Agricultural Biotechnology and Genomic resources Research and Development (ICABIOGRAD-IAARD), Bogor, Indonesia
³Plant Genomics and Breeding Institute, Seoul National University, Seoul, Korea

Received 13 May 2015; revised 4 July 2015; accepted 10 July 2015.

*Correspondence (Tel: +8228804545; fax: +8228774550; email sukhalee@snu.ac.kr)

Keywords: translational genomics, next-generation sequencing, genomic resources, high-throughput genotyping, genome-assisted breeding, Ontology-driven database.

Summary

The use of next-generation sequencers and advanced genotyping technologies has propelled the field of plant genomics in model crops and plants and enhanced the discovery of hidden bridges between genotypes and phenotypes. The newly generated reference sequences of unstudied minor plants can be annotated by the knowledge of model plants via translational genomics approaches. Here, we reviewed the strategies of translational genomics and suggested perspectives on the current databases of genomic resources and the database structures of translated information on the new genome. As a draft picture of phenotypic annotation, translational genomics on newly sequenced plants will provide valuable assistance for breeders and researchers who are interested in genetic studies.

Introduction

The genetic improvement of important crops by breeding programmes has increased the yield for global food security. The breeding programmes have adopted various strategies for efficient breeding that increase total yield and decrease the whole breeding period. These strategies include traditional phenotype-based selection, marker-assisted selection and, more recently, genome-assisted breeding (GAB) (Varshney et al., 2013). Consequently, various aspects, including the elevation of resistance or tolerance against biotic and abiotic stresses and the fine-tuning of flowering time and maturity, have increased crop yields (Varshney et al., 2009). Major cereal crops, such as rice, maize and soya bean, have been intensively studied, and many important genomic resources that can be applied to the breeding programme have been published (Bolger et al., 2014). These resources facilitate both molecular breeding and genetic engineering to ensure the high yield and quality. The continuous attention on improving the yield and ingredients of major crops has highly enriched genetic and genomic resources regarding important agricultural traits (Monaco et al., 2014; Varshney et al., 2014).

In addition to the total yield, the improvement of particular ingredients is mostly focused on the trends of consumer’s favourites or diversified industrial purposes, such as medicine and cosmetics. In particular, medicinal plants have potential properties that would alleviate oxidative stresses with regard to human diseases, such as cancer, cardiovascular diseases and inflammatory diseases. These properties are found in extracts from the stems, roots, bark, leaves, fruits and seeds of plants (Krishnaiah et al., 2011). Recently, specific attention has been placed on noncrop plants in an effort to improve traits of interest for the mass production of expensive chemical compounds that would be achieved by the breeding programme (Canter et al., 2005; Ngo et al., 2013). Although the noncrop plant does not have enough genomic resources that can be applied to the breeding programme, recent technological advances have used various approaches to build up genomic resources at an unprecedented fast rate. Orphan legume crops, such as chickpeas and pigeon peas, have become crops with rich genomic resources via the intensive application of next-generation technologies in sequencing and genotyping (Varshney et al., 2013).

Leading manufacturers of the next-generation sequencer (NGS), such as Illumina and Pacific Biosciences (PacBio), have continuously launched new sequencing chemistries and platforms that decrease sequencing cost and increase read length and reliability. These sequencing platforms can produce enormous sequences from sample species and cover several folds of the genome at cheaper costs. The computational algorithms for mapping, assembly and statistics have also co-evolved with the sequencing platforms and enable single research groups to assemble fragmented sequences into a draft reference genome (Nagarajan and Pop, 2013). Moreover, population-level genotyping can be accomplished within several days at low costs using NGS, high-throughput genotyping technologies based on the automation of polymerase chain reaction (PCR) steps, or microarrays. These technologies boost the production of genome sequences and revolutionize genomic research in many important plants, thus allowing for genomics-assisted breeding.

Here, we review the current status of genome sequencing, genomics-related technologies and the genetic and genomic resources of model plants that are highly enriched with the
flooding of genome sequences and databases. Moreover, we (i) focus on the strategies of translational genomics on newly constructed draft genome sequences using the genomic resources of model plants, (ii) discuss the current needs of phenotypic classification on the genomic resources of model plants and (iii) suggest the schematic database structure of transferred genomic resources. Translational genomics-derived genome annotations can provide an estimation of the phenotypic contribution of genomic regions that would benefit the breeding of noncrop plants with limited budgets. This especially enables researchers and breeders to select informative genetic markers for traits of interest and allows them to perform marker-assisted breeding and genome selection. We expect that the integrated databases of the genetic and genomic resources of model crops and plants would propel translational genomics on newly constructed genome sequences, and the phenotype-driven database of transferred genomics resources would be essential for the breeding of noncrop plants.

**Status of genome sequencing**

Since the completion of the reference genome sequences of model plants and crops, such as Arabidopsis, rice and soya bean, the publications of reference genomes have accelerated with the emergence of NGS technology, which enables cheap, high-throughput sequencing (Figure 1a) (http://www.genome.gov/sequencingcosts/). The sequencing cost decreased rapidly with the emergence of NGS platforms in 2008. Afterwards, the sequencing cost per single base continued to decrease as the leading companies (Illumina, Roche and PacBio) improved their sequencing chemistries and platforms (Figure 1b). Currently, a ~500-Mb genome can be sequenced by a single run of Illumina HiSeq2000 for draft genome assembly. The computational algorithm has been well developed to achieve genome assembly by a single-server computer. For example, a server equipped with 32 processes of Intel(R) Xeon(R) CPUs, ~500 GB memory, and over 10 Tb storage space can cover the *de novo* assembly of a ~500-Mb plant genome using over 200X Illumina Hiseq sequence reads by the ALLPATHS-LG software (Kang et al., 2014; Gnerre et al., 2011). With these technological and analytical advances, the number of genome-related publications increased in the period of 2012–2014 (Figure 1c). Most draft genome sequences were constructed by NGS technology alone within this period (Bolger et al., 2014). The majority of published plant genomes are around ~500 Mb in size (Figure 1d). The PacBio sequencing platform has evolved rapidly in recent years and has developed much longer read lengths than the Illumina platform that can generate short reads of ~250 bp. In fact, the PacBio sequencing platform is now able to generate long reads with N50 read lengths of 10–15 Kb (Figure 1e). This would solve any assembly problems that may be caused by long repetitive sequences, which can misdirect assembly algorithms. Hence, this long-read technology will accelerate the genome projects of Giga-sized genomes and improve pre-existing reference genomes with many gap sequences.

The rapid progress of sequencing technologies enables single research groups to carry out multiple genome projects and construct high-quality reference genomes. Currently, ~99 plant species have been sequenced into draft genome sequences. We expect that the larger number of reference genomes will be generated in the science community in a few years, and this number will exceed the amount of genomes that have been sequenced in the past 15 years. Although NGS-derived draft genomes are mostly incomplete with regard to assembly contiguity, number of gaps and genome coverage to the estimated genome size, these data provide abundant genomic resources for downstream analyses. Genome sequences at the scaffold level show many types of genomic resources, including simple sequence repeats (SSRs), transposable elements, protein-coding genes and their traces, and gene orders. Combining these sequences with pre-existing genetic maps will enable us to map genomic regions to well-studied quantitative trait loci (QTLs).

To enrich genomic resources, many sequencing strategies that use different libraries, such as direct DNA resequencing, genotype by sequencing (GBS), DNA sequencing and small RNA profiling, are implemented, based on the reference genome (Elshire et al., 2011; Lister et al., 2008; Oszolak and Milos, 2011). As of 1 April 2015, 1,244,872 run data from 997,587 experiments have been deposited in live status at the Sequence Read Archive (SRA), which was established in 2007 as a database for sharing NGS data among researchers. Information regarding experimental designs and raw sequencing reads can be readily downloaded to a local computer by the Aspera plug-in, which is provided by the National Center for Biotechnology Information (NCBI). The largest amounts of sequence reads from rice and Arabidopsis were deposited as models of monocot and dicot plant, respectively. Major crop species, such as *Hordeum vulgare*, *Zea mays*, *Triticum aestivum*, *Sorghum bicolor*, *Solanum lycopersicum* and *Glycine max*, were also abundantly sequenced (Figure 1f). Currently, the SRA is largely composed of sequence reads from the Illumina platform, followed by reads from the sequencing platform, GS-FLX (Figure 1g).

**Status of high-throughput genotyping**

Due to reference genomes and resequencing efforts, numerous genetic markers have been developed for applications in breeding and genetic studies (Yang et al., 2012). The genotyping efforts of developed genetic markers on a large set of germplasm and breeding populations are essential for genetic studies, as well as molecular breeding. Using NGS technology, the resequencing strategy can be applied to reveal the maximum number of variations in the interested individual against the reference genome. GBS is cost-effective and a fast strategy for developing single nucleotide polymorphisms (SNPs), constructing genetic maps and mapping QTLs (Elshire et al., 2011). Along with the rapid development of NGS technology, genotyping technologies have also evolved. By reducing the time required for manual pipetting, PCR and detection (agarose electrophoresis) processes, these technologies aim to generate large amounts of genetic data within a short time with less costs. Advanced Analytical Technologies has developed the Fragment Analyzer™ (http://aati-us.com/product/fragment-analyzer), which allows for the high-throughput genotyping of the SSR marker by automated capillary electrophoresis. This technology significantly reduces the highly laborious gel electrophoresis step and processes up to 288 samples (3 × 96-well plates) in a single run. Fluidigm® (https://www.fluidigm.com/) released the SNP Type™ assay, which uses integrated fluidic circuit (IFC) technology to automate the PCR and detection processes. IFC can automatically assemble samples and primers by a network of microfluidic channels to decrease the number of pipetting steps. Moreover, the researcher can
choose among different types of IFC plates, depending on the numbers of primers and samples. This platform can read 2304–9216 genotypes in a single run using IFC plates of 48 (primers) × 48 (samples) to 96 × 96, respectively. In addition, Douglas Scientific has attempted to fully automate genotyping steps using robotic pipetting arms and Array Tape technology (http://www.douglasscientific.com/). They recently released IntelliQube®, which integrates the automation technologies of genotyping, including pipetting, PCR and detection processes, in a single platform. This platform uses 96 or 384 channels of the CyBi FeliX Pipette for sample dispensing and 4 channels of the Dispense Jet for the primer pairs of genetic markers. A total of 24 960 genotypes can be generated by the 65 × 384 array tape in 8 h. With the LGC KASP™ primer mixture (http://www.lgcgroup.com/products/kasp-genotyping-chemistry/), which is designed for specific binding to the SNP of interest, both platforms, SNP Type™ and IntelliQube®, can synergistically facilitate genotyping. In addition to PCR-based genotyping, array-based systems, such as Illumina Infinium® HD and Affymetrix Axiom®, allow for the high-throughput genotyping of up to millions of loci per sample and are suitable for genomewide association study (GWAS). The increasing availability of predesigned genotyping arrays can allow researchers to focus on precise experimental designs and sample preparation schemes for crops.

**Genetic studies of major crops**

Crop genetics have been intensively studied for the generation of various genomic resources. The development of genetic markers and the refinement of their application in breeding populations have enabled researchers to construct genetic maps and to find approximate positions of qualitative and quantitative traits on the map. Fast construction of the reference genome has boosted genetic marker development and genotyping processes that were highly laborious and time-consuming before the advancement of technologies. Moreover, these technologies accelerate gene function analyses and genetic studies that are used to perform statistical associations between genotypes and phenotypes, such as QTL mapping and GWAS. It is necessary to combine accumulated genomic resources in the reference genome to collectively infer the contribution of genomic regions to a particular phenotype. These phenotype-associated genomic resources would be useful hints for breeders to determine the informative marker set for the genotyping of germplasms. The first step for this objective would be the construction of an integrated database of genomic resources, which would give easy accessibility to users.

Several well-known databases of genomic resources have been constructed for each plant species (Table 1). A growing number of reference genomes and genome browsers, which display...
various genomic features, are established in several databases, such as Phytozone (http://phytozome.jgi.doe.gov/pz/portal.html) and PlantGDB (http://www.plantgdb.org). These databases provide well-organized user interfaces, including keyword search, Basic Local Alignment Search Tool (BLAST) and genome browser, as well as direct links to bulk data for bioinformatics analyses. In particular, Phytozone added the web application, PhytoMine, which provides templates for retrieving genomic information. These templates provide defined genomic features, such as gene expression levels of various tissues and a list of genes in selected genomic regions. They also provide application programming interfaces (API) that can be directly included in various programming languages, such as Perl, Ruby, Java and Python. These APIs enable the bioinformatician to merge the genomic information of different species without downloading heavy genomic data sets to a local computer. The newly published genomes deposit their own databases, as well as public databases, for the agile update of sequence improvement and their genomic resources.

In addition to information regarding the reference genome, several databases deal with integrated genetic and genomic resources, such as QTL information, genetic markers, gene expression, traces of DNA methylation and small RNAs. (Table 1). Genetic markers are genomic regions that show polymorphisms, thus allowing us to distinguish between different species or cultivars and to observe the genotypes of target species that are linked to phenotypes. Currently, SSR and SNP markers are used for genetic studies and molecular breeding. SNPs provide higher genome resolution due to their frequent occurrence in the genome and are genotyped according to four discrete nucleotides. Importantly, SNPs are suitable for the high-throughput digital genotyping pipeline. In contrast, SSRs have been reported to be highly polymorphic and are more appropriate for diversity analysis (Singh et al., 2013). Because the reference genomes and resequencing efforts have evolved, it is now possible to develop a large number of SSR and SNP markers. For public access to genetic marker data, lists of published genetic markers and QTLs are available in various databases (SoyBase, Gramene, CerealDB, GrainGenes, etc.) (Table 1). Recently, the available 20 000 000 SNPs across the rice genome developed by the rice 3000 genomes project were organized into SNP-Seek database (Alexandrov et al., 2015; The 3000 rice genomes project, 2014). However, more web-based tools are needed to provide the easy retrieval of necessary data for biologists and breeders. The PhytoMine application in Phytozone provides a template titled, 'Variants Near an Annotated Gene', for easy access to known variants around the querying gene. This feature would greatly benefit researchers who are interested in finding polymorphic sites around target genes.

Many important agricultural traits, such as flowering time, maturity and yield, are quantitative traits. These traits are usually related to multiple loci, thus making it complicated to implement functional analyses of single genes. Therefore, statistical approaches, including QTL mapping and GWAS, are used to dissect the responsible genomic regions for quantitative traits by high-density genetic markers. The approximate physical locations of QTLs are highly informative, because they can be regarded as the statistical phenotype annotation of reference genomes. For PCR-based genotyping, the physical location of QTLs can be deduced from the sequence information of correlated genetic markers, such as the sequences of primer pairs and PCR clones. For NGS-based genotyping, the physical location would be directly retrieved from read-mapping results. For the array-based genotyping platform, the probe information usually contains the genomic locations. Among all model crops, soya bean and rice

| Table 1 List of databases that deposited various genetic and genomic resources |
|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| **DB name** | **Contents** | **Plant species** | **URL** | **Reference** |
| PLEXdb | Gene expression | Arabidopsis, Barley, Brachypodium, Citrus, Cotton, Grape, Maize, Medicago, Poplar, Rice, Soya bean, Sugarcane, Tomato, Wheat | http://www.plexdb.org | Dash et al. (2012) |
| RiceXPro | Gene expression | Rice | http://ricepro.dna.affrc.go.jp/ | Sato et al. (2013) |
| CerealsDB | Genetic markers | Wheat | http://www.cerealsdb.uk.net/cerealgenomics/ | Wilkinson et al. (2012) |
| SoyKb | Genetic markers, genomic resources | Soya bean | http://soyk.org/ | Joshi et al. (2014) |
| PGDBj | Genetic markers, QTLs, genomic resources | 80 plant species | http://pgdbj.jp/ | Asamizu et al. (2014) |
| SoyBase | Genetic markers, QTLs, genomic resources | Soya bean | http://soybase.org/ | Grant et al. (2010) |
| SNP-Seek | SNP | Rice | http://www.oryzasnp.org/iorc-portal/ | Alexandrov et al. (2015) |
| Phytozone | Genome | 48 plant genomes | http://phytozone.jgi.doe.gov/pz/portal.html# | Goodstein et al. (2012) |
| PlantGDB | Genome | 27 plant genomes | http://www.plantgdb.org/ | Duvick et al. (2008) |
| Gramene | Genome, Genetic markers, QTLs | 39 plant genomes | http://www.gramene.org/ | Monaco et al. (2014) |
| GrainGenes | Genome, Genetic markers, QTLs, genomic resources | Wheat, barley | http://wheat.pw.usda.gov/GG3/ | Carollo et al. (2005) |
| ASRP | Small RNA | Arabidopsis | http://asrp.danforthcenter.org/ | Backman et al. (2008) |
| CSRDB | Small RNA | Maize, Rice | http://sundarilab.ucdavis.edu/srnmas/ | Johnson et al. (2007) |
have been intensively studied, and their QTLs have been located on the genetic map. SoyBase and Gramene provide large amounts of QTL information for soya bean and rice, respectively (Table 1). In SoyBase, 3075 QTLs of soya bean are listed. In Gramene, 8216 rice QTLs are categorized into 236 essential traits, which are assigned to nine trait categories, including ‘Abiotic stress’, ‘Anatomy’, ‘Biochemical’, ‘Biotic stress’, ‘Development’, ‘Quality’, ‘Sterility or fertility’, ‘Vigor’ and ‘Yield’ (Figure 2a, b).

GWAS can be also used to infer statistical associations between variations in genomic regions and phenotypes of populations, such as landraces, breeding population and the nested association mapping population (Rafalski, 2010). The advancement of NGS and various genotyping technologies makes it feasible to perform association analyses in larger populations with more dense markers increasing the genome resolution to detect precise linkage disequilibrium blocks and recombination breakpoints. So far, the GWAS has been implemented for several important crops (Table 2). These analyses usually result in several significant loci or markers. Zhou et al. (2015) performed large-scale resequencing-based GWAS on 302 soya beans, including wild, landrace and cultivated soya bean accessions. They found various domestication signatures, such as 230 selective sweeps and 162 selected copy number variants. With domestication-related phenotypic traits, they found several novel associations between 13 SNPs to oil content, plant height and pubescence form. Numerous loci in crop species have also been revealed by GWAS (Table 2). It is necessary to construct the GWAS database to consolidate phenotype-associated genomic regions for the utilization of functional gene studies and crop breeding. The results of GWAS are highly dependent on the structure of the testing population, the phenotypic variation of individuals and the number of markers used; hence, this additional information should also be added to the database.

Analysing the expression of genes in various organs at each growth stage would be important for understanding the contribution of genes to particular traits. NGS-based RNA-sequencing and microarray-based technologies have allowed for genome-wide gene expression analyses of sampled tissues. These enabled researchers to observe gene expression changes by controlled factors, such as outside treatments (biotic or abiotic stresses) or genotype variations between wild-type and mutant species. Differentially expressed genes can be good candidates for further functional analyses of the phenotypes of interest. Even if candidate genes are not verified by functional analyses, they can be still important clues for understanding the roles of genomic regions. Furthermore, functional gene networks can be inferred from shared gene expression patterns from various samples or experiments of crop species (Lee et al., 2010). For the consolidation of expression data, the Plant Expression Database was constructed for 14 plant species and provides a gene expression atlas with detailed experimental designs and tested germplasm information, as well as visualization tools. Soybean Knowledge Base (SoyKb) and RiceXPro also contain gene expression profiles for soya bean and rice, respectively (Table 1).

On the other hand, small RNA and DNA methylation traces in a genome have been detected by NGS technologies, based on the variations in library preparation methods (McGinn and Czech, 2014; Urich et al., 2015). Small RNAs, such as microRNA and

![Figure 2](image-url)

**Figure 2** Genomic resources that have been deposited in databases. (a) The numbers of quantitative trait loci (QTLs) of rice and soya bean, which are categorized by the nine plant ontologies. (b) QTL and known gene distribution in the rice genome; the red dots represent functionally identified genes. (c) The numbers of QTL and genetic markers of various crop species that have been deposited in the PGDBj database.
small-interfering RNA, play a role in gene regulation by the RNA interference (RNAi) pathway in the cytoplasm. In the nucleus, the RNAi pathway also affects gene regulation with DNA methyltransferases (Castel and Martienssen, 2013). DNA methylation seems to have varying functions in gene regulation, according to the sequence contexts of methylation (Jones, 2012), and is widespread in plant genomes, which have been reported in Arabidopsis and soybean (Lister et al., 2008; Schmitz et al., 2013). Therefore, the polymorphisms of these genomic traces should be considered in genetic studies to unravel the causalities of subtle phenotypic variations. For example, in soybean, Schmitz et al. (2013) found inherited DNA methylation traces and the evidence for local methylQTL from B3 recombinant inbred lines by BSseq approach that uses bisulphite conversion of unmethylated genomic DNA. The potential effect of small RNA and DNA methylation on gene regulation and inheritance enables us to regard them as important genomic resources for crop breeding. Several public RNA-related databases have been constructed, including the Arabidopsis Small RNA Project Database, the Cereal Small RNAs Database and SoyKb (Table 1).

The functionally identified genes are valuable resources for translational genomics and crop breeding. The functionally annotated genes can be mapped to known pathways and can enable the researchers to deduce whether the target genes participate in the traits of interest. These genes are systematically collected in the UniProt database (Consortium, 2012). In Swiss-Prot of the UniProt database, functionally identified genes have been gathered from the literature and curator-evaluated computational analyses. In flowering plants (Magnoliophyta), 32 880 genes have been functionally identified (Table 3). Among them, the function of 13 882 and 3334 genes has been revealed in Arabidopsis and rice, respectively (Figure 2b).

The shared sequence context and QTL annotation may implicate the possible contribution of genomic content to the quantitative trait. For the comparative analysis of QTL among different crop species, it is necessary to integrate published QTLs

| Species           | Trait                        | Number of markers or loci associated with traits | Reference                  |
|-------------------|------------------------------|------------------------------------------------|----------------------------|
| Oryza sativa      | Grain-filling rate           | 31                                             | Liu et al. (2015)           |
|                   | Aluminium tolerance          | 48                                             | Famoso et al. (2011)        |
|                   | Yield and others             | 52                                             | Begum et al. (2015)         |
|                   | Chlorophyll content          | 46                                             | Wang et al. (2015)          |
|                   | Salinity tolerance           | 64                                             | Kumar et al. (2015)         |
|                   | Blast disease                | 30                                             | Wang et al. (2014)          |
|                   | Ozone tolerance              | 16                                             | Ueda et al. (2015)          |
|                   | Yield and others             | 141                                            | Yang et al. (2014)          |
|                   | Agronomic traits             | 80                                             | Huang et al. (2010)         |
| Zea mays          | Leaf architecture            | ~300                                           | Tian et al. (2011)          |
|                   | Resistance to head smut      | 18                                             | Wang et al. (2012)          |
|                   | Fusarium ear rot disease     | 7                                              | Zila et al. (2014)          |
|                   | Resistance to the Mediterranean corn borer | 25 | Samayoa et al. (2015) |
|                   | Carotenoid biofortification  | 40                                             | Suwarno et al. (2015)       |
|                   | Root development             | 28                                             | Abdel-Ghani et al. (2015)   |
|                   | Seedling root development    | 268                                            | Pace et al. (2015)          |
|                   | Developmental timing of vegetative phase | 13 | Foerster et al. (2015) |
|                   | Seed protein and oil content | 65                                             | hhwang et al. (2014)        |
|                   | Iron deficiency chlorosis    | 33                                             | Mamidi et al. (2014)        |
|                   | Carbon isotope ratio         | 39                                             | Dhanapal et al. (2015)      |
|                   | Seed size and shape          | 41                                             | Hu et al. (2013)            |
|                   | Yield and yield components   | 29                                             | Hao et al. (2012)           |
|                   | Domestication and improvement| 13                                             | Zhou et al. (2015)          |

| Scientific name | Common name | Number of identified genes |
|-----------------|-------------|----------------------------|
| Arabidopsis thaliana | Mouse-eared cress | 13 822                     |
| Oryza sativa subsp. indica | Rice | 3334                       |
| Zea mays | Maize | 752                       |
| Oryza sativa subsp. indica | Rice | 704                       |
| Nicotiana tabacum | Tobacco | 465                       |
| Solanum lycopersicum | Tomato | 428                       |
| Solanum tuberosum | Potato | 401                       |
| Glycine max | Soybean | 390                       |
| Pisum sativum | Garden pea | 387                       |
| Triticum aestivum | Wheat | 369                       |
| Hordeum vulgare | Barley | 349                       |
| Spinacia oleracea | Spinach | 285                       |
| Daucus carota | Wild carrot | 185                       |
| Sorghum bicolor | Sorghum | 172                       |
| Vitis vinifera | Grape | 167                       |
| Phaseolus vulgaris | Kidney bean | 159                       |
| Brassica rapa | Rape | 150                       |
| Gossypium hirsutum | Upland cotton | 150                       |
| Helianthus annuus | Sunflower | 142                       |
| Populus trichocarpa | Western balsam | 128                       |

| List of GWAS publications of three major crops including O. sativa, Z. mays and G. max |
|----------------------------------|---------------------------------|
| Species                          | Trait                           | Number of markers or loci associated with traits |
| Oryza sativa                     | Grain-filling rate              | 31                                             |
|                                  | Aluminium tolerance             | 48                                             |
|                                  | Yield and others                | 52                                             |
|                                  | Chlorophyll content             | 46                                             |
|                                  | Salinity tolerance              | 64                                             |
|                                  | Blast disease                   | 30                                             |
|                                  | Ozone tolerance                 | 16                                             |
|                                  | Yield and others                | 141                                            |
|                                  | Agronomic traits                | 80                                             |
| Zea mays                         | Leaf architecture               | ~300                                           |
|                                  | Resistance to head smut         | 18                                             |
|                                  | Fusarium ear rot disease        | 7                                              |
|                                  | Resistance to the Mediterranean corn borer | 25 | |
|                                  | Carotenoid biofortification     | 40                                             |
|                                  | Root development                | 28                                             |
|                                  | Seedling root development       | 268                                            |
|                                  | Developmental timing of vegetative phase | 13 | |
|                                  | Seed protein and oil content    | 65                                             |
|                                  | Iron deficiency chlorosis       | 33                                             |
|                                  | Carbon isotope ratio            | 39                                             |
|                                  | Seed size and shape             | 41                                             |
|                                  | Yield and yield components      | 29                                             |
|                                  | Domestication and improvement   | 13                                             |
of the crop species into a single database. For this purpose, a Japanese research group published the database, PGDBj (Asamizu et al., 2014) (http://pgdbj.jp/), which combines the genetic markers and QTL data from various plant species. The primer sequences are available for 98 333 genetic markers of 40 plant species, and 3144 QTLs of 27 plant species are listed (figure 2c).

Because biological data points are growing at a rapid rate, the traditional format of databases (e.g. keyword search, long lists with alphabetical index, many tables in one page) is insufficient for handling massive data and is also lacking readability to the users. For single species data, genome browsers (e.g. Gbrowse, Jbrowser and UCSC Genome Browser) have been favoured for the visualization of large data, as layers of various features can be easily added (Donlin, 2009; Skinner et al., 2009). For comparative genomics, CoGe (https://genomesevolution.org/coge/) was designed to provide useful web-based bioinformatics tools, as well as solutions for visualization. CoGe focuses on the conservation of gene order (synteny) among plant species and displays dot plots and protein alignments (Lyons et al., 2008).

These databases are commonly focused on providing information regarding genes or genomic regions. However, crop breeders usually query agriculturally important phenotypes rather than certain genes or genomic regions because their goals are to improve a specific target trait. For the breeder’s purpose, we suggest using phenotype-based databases to classify genetic and genomic resources according to the plant ontology (PO) system (Jaiswal et al., 2005) (http://www.plantontology.org/). Phenotypes can be represented by functionally identified genes, QTLs and trait-associated genomic regions from GWAS. Moreover, this knowledge can be transferred to newly constructed genomic sequences using the translational genomics approach, which takes advantage of the conserved features in plant genome sequences.

The theoretical basis for translational genomics

Advancements in sequencing and genotyping technologies have guided the abundant genomes and genomic resources of model crops. ‘Translational genomics’ is performed to transfer the rich genomic knowledge of model plants to newly constructed reference genomes of orphan crops or minor plants. The transferred information would serve as valuable hints for breeding and functional biology, along with various objectives for the species. Translational genomics aims to take advantage of the perspectives of comparative genomics, systemic genomics and evolutionary genomics.

Comparative genomics focuses on the similarities and differences among plant species with regard to DNA sequence, protein sequence and gene orders, as well as epigenetic signatures. The conservation of these features suggests the possibility of common phenotype contribution, and this information can be used for the functional annotation of the newly sequenced genome. More information would be transferred from model crops to phylogenetically closer species than distantly related plant species, as the annotation highly depends on pairwise similarity and conservation among neighbouring species. The legume plants have an advantage in translational genomics, due to the presence of multiple model species, including soya bean, Medicago truncatula and Lotus japonicus (Cannon et al., 2009).

Once the highly confident gene model in a new reference genome is extracted from the results of RNA-sequencing experiments in various tissues, the protein sequences of the gene model can be functionally annotated by sequence homology and domain architecture. The protein homology against well-known genes can be calculated using the BLAST algorithm against protein databases of the phylogenetically close model plant or well-classified databases such as UniProt (Consortium, 2012). The domain architecture, which is the functional region within a protein sequence, can be detected by the Hidden Markov model (HMM) algorithm. This algorithm recruits sequence conservation among species. The protein family database contains the HMM profiles of various protein domains that are built based on the emission and transition probabilities from the alignment of known domain sequences (Finn et al., 2014). For the collective annotation of domain architecture and classification of querying proteins, the InterProScan software has been chosen (Hunter et al., 2012).

For a deeper understanding of gene functions that result from different types of gene interactions, gene networks are modelled based on the probabilities of gene co-expression and protein–protein interactions. In addition to various accumulated experimental data, RNA-sequencing data that produce a massive amount of co-expression data in the public have been shown to significantly increase the prediction power. AraNet, a genome-wide gene network for Arabidopsis thaliana, was constructed to predict reliable gene annotation towards plant traits (Lee et al., 2010). In addition, the network modelling was performed to understand the Arabidopsis immune signal network against the bacterial pathogen, Pseudomonas syringae, from mRNA profiles of 22 Arabidopsis immune mutants (Sato et al., 2010). The inferred gene network model could accurately predict regulatory relationships, including previously reported cases. With these general or conditional gene network models, we can further understand the roles of genes in model plants. Hence, based on the homology with genes in model plants, gene networks can also be transferred to new genomes as functional clues for genes.

Furthermore, the conservation of gene order, which is termed as synteny or gene co-linearity, can be used for genome annotation (Paterson et al., 2010). Even though the chromosomes are shuffled among neighbouring species after speciation, the conservation of gene order within each chromosome has been detected in various genome projects, suggesting that the neighbouring plant species are diverged from a common ancestor. The syntonic relationships of Medicago truncatula, Brassica rapa and Oryza sativa to the Glycine max genome showed gene order conservation even between monocots and dicots, although the extent of these relationships decreased as phylogenetic distances increased (Figure 3a). If we postulate that the traits of interest exist in a common ancestor, we can assume that the conserved gene orders would contribute to the common traits of the descendant species. Based on this assumption, the phenotypic annotations around the syntonic regions of model plants can be transferred to the corresponding genomic regions of neighbouring species. This scheme would be useful for transferring QTLs that are represented by multiple genetic markers linked to the responsible genomic regions of the traits of interest. Supporting this assumption, we showed cases in which QTL conservation was accompanied with synteny relationship between species (Figure 3b). The disease resistance and flowering time QTLs were conserved along with syntonic regions, even between dicot and monocot species. It is hard to generalize that transferred QTLs are strong candidates for the traits; however, the presence of multiple QTL annotations from various species would be highly informative for breeding purposes. For QTL
annotations in the adzuki bean, of which genetic studies are lacking, 2010 QTL-associated SSR marker positions in the soya bean genome were transferred using the gene order conservation of 569 orthologous synteny blocks (Kang et al., 2015).

Similarly, genetic markers that are significantly associated with traits of interest via GWAS can also be transferred to the new genome by synteny relationship. In addition, GWAS can reveal domestication traces in the genome. The history of human selection activities on crops from unconscious selection to modern breeding practice has accumulated traces in genome sequences. The features of altered allele frequency and selective sweep, as well as selected copy number variants, signify the domestication traces. GWASs using cultivated and wild accessions identified domestication-related genomic regions (Chung et al., 2014; Zhou et al., 2015). Even though the domestication processes vary according to ecological, evolutionary, cultural and technological factors, the domestication-related phenotypes commonly include low-seed shattering, large fruit or seed size, changes in branching and stature, change in reproductive strategy and changes in secondary metabolites. These are collectively called the ‘domestication syndrome’ (Meyer et al., 2012). Therefore, the translation of domestication traces to the new genome would be also beneficial for domestication-related phenotypes.

Evolutionary signatures, including gene or genome duplication, can be used to understand the genomic features of new genomes. This approach uses evolutionary features in the genome and is especially useful for newly sequenced species that are distantly related to the model plants. The gene balance model explains the fate of duplicated genes (Freeling, 2009). As plants experience multiple rounds of large-scale duplication and frequent small-scale duplication, their gene copy numbers and the dosage of gene products increase. It would be expected that the core genes in the major pathway would disturb plant homeostasis if their dosages were abnormally increased. However, plants can actively decrease the core gene dosage into proper levels by gene loss processes, such as fractionation (Freeling, 2009). On the other hand, the copy number of downstream genes or surveillance genes tends to be retained even after duplication events. As a result, these genes remain as high-copy genes, as they do not disturb gene connectivity in major pathways even when their dosages are increased.

Figure 3 Synteny relationships among plant species. (a) The synteny blocks of soya bean to Medicago truncatula, Brassica rapa and Oryza (O.) sativa; the x-axis and y-axis represent the genomic location and Ks values of each synteny block, respectively. (b) The detailed view of microsynteny between species that share disease resistance and flowering time QTL.
stress-response genes are, reportedly, highly enriched in tandemly arrayed gene families with high copy numbers (Hanada et al., 2008). Furthermore, genes that recognize signals (surveillance gene) from abiotic and biotic stresses at the surface of cells have been shown to elicit defensive pathways, such as the jasmonic and salicylic pathway, to finally turn on genes that produce the secondary metabolites (downstream genes) to cope with stresses. Therefore, the high-copy-number genes of the tandem array in a genome could be the primary genomic regions for crop breeding to improve stress resistance (Kang, 2015). The colocalization of the tandemly duplicated genes with disease-related QTLs in rice and soya bean supports the role of tandemly duplicated genes (Figure 4a). We visualized cases that show the colocalization of tandemly duplicated genes with disease-related QTLs in rice and soya bean. In chromosome 1 of rice, C547, which is a genetic marker that is associated with 10 disease-related QTLs, is localized proximally to diverse tandemly duplicated genes. In soya bean, satt244 and satt191 markers, which are associated with disease-related QTLs, are also surrounded by tandem duplicated gene clusters (Figure 4b).

The strategy for integration of the translated knowledge into a database

We reviewed the basis of translational genomics that transfer the genomic resources from model plants to the newly constructed genome. Protein or genomic regions can be annotated with phenotypic clues by sequence homology or gene order conservation. The gene or genome evolution traces can also be another signature for inferring stress-response genes. It is necessary to integrate transferable genomic resources into a single database with their respective genomic locations. However, a few hurdles

Figure 4  Tandem duplication and its effect on disease resistance in the rice and soya bean genomes. (a) The distribution of tandemly duplicated genes (upper) and disease-related QTLs (bottom) in the genome of *O. sativa* and *Glycine max*. (b) The detailed view of tandemly duplicated genes and their proximal disease-related QTLs; the line-linked colour boxes represent tandemly duplicated genes, and the inverted red triangle represents the genetic marker position that is associated with disease resistance.
need to be overcome for this to occur. First, the current GWAS results of plants can only be accessed via laborious literature searches, and it is necessary to integrate the GWAS results of multiple model plants into the fixed data format for easy data accessibility. There are consolidation examples of various GWAS results into a single database in human genomics (Johnson and O’Donnell, 2009; Li et al., 2012). The NHGRI-EBI GWAS Catalog (http://www.ebi.ac.uk/gwas/) provides the most recent GWAS results of human traits, the search and visualization tools, as well as bulk downloads (Welter et al., 2014). Second, it is essential for the QTL data to be organized in an informative manner. The Gramene format for the rice QTL is a good example, as it provides the essential fields, including the QTL ID and name, the trait ontology and name and the chromosomal location.

Once the fundamental databases of the genomic resources of model plants are established, we propose that the genomic resources would be tagged based on reasonable phenotypic categories, such as PO, to allow for the ontology-driven search to occur. The ontology-tagged information would be transferred to the new genome to enable researchers and breeders to select informative genetic markers for their traits of interest. It would be cost-effective to determine the highly informative marker set for crop breeding, even though current technologies provide high-throughput genotyping. We schematically designed the PO-based database of a new genome that is annotated by the genomic resources of model plants (Figure 5a). This database starts with the PO to list the related genomic regions. If the user selects a PO-related genomic region, the genome browser performs the visualization of the selected genomic regions to display synteny-based QTL and GWAS results, gene annotations and gene networks (Figure 5b). The GWASdb for human genomics adopts the ontology-driven database structure that maps GWAS results based on a phenotype query, such as Human phenotype Ontology and Disease Ontology. Furthermore, the GWASdb visualizes genomic regions that are related to the querying ontology with the respective P-value (Li et al., 2012).

As indirect evidence, transferred genomic resources on the new genome would help researchers and breeders to select specific genetic markers for genetic studies and breeding. The current Welter’s of GAB includes genomic selection (GS), marker-assisted recurrent selection and advanced backcross breeding, which have all been intensively reviewed (Varshney et al., 2013). Among these strategies, the GS approach estimates breeding values using an inferred statistical model from genomewide genotypes and phenotypes of the training population to determine the potential superior parental lines for the improvement of traits of interest. Dense genomewide markers are needed for GS, and Spindel et al. (2015) suggested a saturation point of one marker per 0.2 cM. The prior knowledge in the genomic regions via translational genomics might allow variations in the genetic marker effects on the phenotypes for model training based on Bayesian variable selection models (Su et al., 2014). Moreover, if the budget is limited, we can select genomic regions that are enriched with the transferred data as informative markers to train a model and to estimate breeding values.

Conclusion

The rapid advancement of NGS and high-throughput genotyping technology opened the bio-data era, and massive amounts of...
biological data have been produced at an unprecedented rate. The reference genome sequences of various crops, model plants and minor plants are constructed on the strength of technological and analytical progress. Along with numerous reference genomes, genetic and genomic resources have also been enriched by genomewide analyses using types of resequencing and genotyping approaches to reveal hidden bridges between genomic variations and diverse phenotypes in plant species. To allow for precise breeding and quick breeding procedures, this enriched genetic knowledge needs to be translated into the crop breeding field. Therefore, translational genomics can be applied to enable minor plants to overcome the scarcity of cumulative genomic resources.

For the newly constructed genome sequences of orphan crops or minor plants, the database can be constructed by the translational genomics approach. This would lead to a draft picture for further genetics on their own genetic data that can be quickly enriched by sequencing and genotyping technologies. We suggest that genetic studies should be accumulated into a database with PO-based categorization and their respective genomic positions of the genetic markers. If their genetic data are highly accumulated to the database following the PO-driven database format, the database would be another model database for the translational genomics, which would calibrate the effectiveness of the transferable genomic resources, and also for the comparative genomics to reveal the highly conserved gene family not only in the protein sequences but also in the contribution for the phenotype among plant species.

Acknowledgements
This research was supported by a grant from the Next-Generation BioGreen 21 Program (Code No. PJ011026), Rural Development Administration, Republic of Korea.

References
Abdel-Ghani, A.H., Kumar, B., Pace, J., Jansen, C., Gonzalez-Portilla, P.J., Reyes-Matanomos, J., San Martin, J.P., Lee, M. and Lubberstedt, T. (2015) Association analysis of genes involved in maize (Zea mays L.) root development with seedling and agronomic traits under contrasting nitrogen levels. Plant Mol. Biol. 88, 133–147.

Alexandrov, N., Tai, S., Wang, W., Mansueto, L., Palis, K., Fuentes, R.R., Ulat, V.I., Chebotarow, D., Zhang, G., Li, Z., Mauleon, R., Hamilton, R.S. and V.J., Chebotarov, D., Zhang, G., Li, Z., Mauleon, R., Hamilton, R.S. and V.J., Chebotarov, D., Zhang, G., Li, Z., Mauleon, R., Hamilton, R.S. and V.J., Chebotarov, D., Zhang, G., Li, Z., Mauleon, R., Hamilton, R.S. and v. J. (2015) SNP-Seek database of SNPs derived from 3000 rice genomes. Nucleic Acids Res. 43, D1023–D1027.

Asamizu, E., Ichihara, H., Nakaya, A., Nakamura, Y., Hirakawa, H., Ishi, T., Tamura, T., Fukami-Kobayashi, K., Nakajima, Y. and Tabata, S. (2014) Plant Genome Database Japan (PGDB): a portal website for the integration of plant genome-related databases. Plant Cell Physiol. 55, e68.

Backman, T.W., Sullivan, C.M., Cumbie, J.S., Miller, Z.A., Chapman, E.J., Fahlgren, N., Givan, S.A., Carrington, J.C. and Karess, K.D. (2008) Update of ARSP: the Arabidopsis Small RNA Project database. Nucleic Acids Res. 36, D982–D985.

Begum, H., Spinдель, J.E., Lalusiun, A., Borromeo, T., Gregorio, G., Hernandez, I., Virk, P., Collard, B. and McCouch, S.R. (2015) Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (Oryza sativa). PLoS ONE, 10, e0119873.

Bögler, M.E., Weissbaehr, B., Scholz, U., Stein, N., Usadel, B. and MAYER, K.F. (2014) Plant genome sequencing - applications for crop improvement. Curr. Opin. Biotechnol. 26, 31–37.

Cannon, S.B., May, G.D. and Jackson, S.A. (2009) Three sequenced legume genomes and many crop species: rich opportunities for translational genomics. Plant Physiol. 151, 970–977.

Carter, P.H., Thomas, H. and Ernst, E. (2005) Bringing medicinal plants into cultivation: opportunities and challenges for biotechnology. Trends Biotechnol. 23, 180–185.

Carolo, V., Matthews, D.E., Lazo, G.R., Blake, T.K., Hummel, D.D., Lui, N., Hane, D.L. and Anderson, O.D. (2005) GrainGenes 2.0: an improved resource for the small-grains community. Plant Physiol. 139, 643–651.

Castel, S.E. and Martienssen, R.A. (2013) RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. Nat. Rev. Genet. 14, 100–112.

Chung, W.H., Jeong, N., Kim, J., Lee, W.K., Lee, Y.G., Lee, S.H., Yoon, W., Kim, J.H., Choi, I.Y., Choi, H.K., Moon, J.K., Kim, N. and Jeong, S.C. (2014) Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes. DNA Res. 21, 153–167.

Consortium, U. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res. 40, D71–D75.

Dash, S., Van Hempt, J., Hong, L., Wise, R.P. and Dickerson, J.A. (2012) PLEXdb: gene expression resources for plants and plant pathogens. Nucleic Acids Res. 40, D1194–D1201.

Dhanapal, A.P., Ray, J.D., Singh, S.K., Hoyos-Villegas, V., Smith, J.R., Purcell, L.C., Andy King, C., Cregan, P.B., Song, O. and Fritschi, F.B. (2015) Genome-wide association study (GWAS) of carbon isotope ratio (delta13C) in diverse soybean [Glycine max (L.) Merr.] genotypes. Theor. Appl. Genet. 128, 73–91.

Donlin, M.J. (2009) Using the Generic Genome Browser (GBrowse). Curr. Protoc. Bioinformatics, Chapter 9, Unit 9.9.

Duivick, J., Fu, A., Muppirla, U., Sabharwal, M., Wilkerson, M.D., Lawrence, C.J., Lushbough, C. and Brendel, V. (2008) PlantGDB: a resource for comparative plant genomics. Nucleic Acids Res. 36, D959–D965.

Elshir, R.J., Gläubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6, e19379.

Famoso, A.N., Zhao, K., Clark, R.T., Tung, C.W., Wright, M.H., Bastamante, C., Kochian, L.V. and McCouch, S.R. (2011) Genetic architecture of aluminum tolerance in rice (Oryza sativa) determined through genome-wide association analysis and QTL mapping. PLoS Genet. 7, e1002221.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L., Tate, J. and Punta, M. (2014) Pfam: the protein families database. Nucleic Acids Res. 42, D222–D230.

Foerster, J.M., Beissinger, T., de Leon, N. and Kaepler, S. (2015) Large effect QTL explain natural phenotypic variation for the developmental timing of vegetative phase change in maize (Zea mays L.). Theor. Appl. Genet. 128, 529–538.

Freeling, M. (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu. Rev. Plant Biol. 60, 433–453.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Sheap, T.P., Sykes, S., Berlin, A.M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Grinius, A., Nusbaum, C., Lande, E.S. and Jaffe, D.B. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. USA, 108, 1513–1518.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazio, J., Mitros, T., Dirks, W., Heilsten, U., Putnam, N. and Rokhsar, D.S. (2012) Phytorez: a comparative platform for green plant genomics. Nucleic Acids Res. 40, D1178–D1186.

Grant, D., Nelson, R.T., Cannon, S.B. and Shoemaker, R.C. (2010) SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res. 38, D843–D846.

Hanada, K., Zou, C., Lehti-Shiu, M.D., Shinozaki, K. and Shiu, S.H. (2008) Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol. 148, 993–1005.

Hao, D., Cheng, H., Yin, Z., Cui, S., Zhang, D., Wang, H. and Yu, D. (2012) Identification of single nucleotide polymorphisms and haplotypes associated with yield and yield components in soybean (Glycine max) landraces across multiple environments. Theor. Appl. Genet. 124, 447–458.

Hu, Z., Zhang, H., Kan, G., Ma, D., Zhang, D., Shi, G., Hong, D., Zhang, G. and Yu, D. (2013) Determination of the genetic architecture of seed size and shape via linkage and association analysis in soybean (Glycine max L. Merr.). Genetica, 141, 247–254.
Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C., Zhu, C., Lu, T., Zhang, Z., Li, M., Fan, D., Guo, Y., Wang, A., Wang, L., Deng, L., Li, W., Lu, Y., Weng, Q., Liu, K., Huang, K., Zhou, T., Jing, Y., Li, W., Lin, Z., Buckler, E.S., Qian, Q., Zhang, Q.F., Li, J. and Han, B. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. 42, 961–967.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T.K., Bateman, A., Bernard, T., Binnis, D., Bork, P., Burge, S., de Castro, E., Coggill, P., Corbett, M., Das, U., Daugherty, L., Duqueinne, L., Finn, R.D., Fraser, M., Gough, J., Hatt, D., Hulo, N., Kahn, D., Kelly, E., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McArthur, A., McDowall, D., McMinn, A., Met, M., Mutowo-Muellenet, P., Mulder, N., Natalie, D., Oregno, C., Pesseat, S., Punta, M., Quinn, A.F., Rivoire, C., Sangrador-Vegas, A., Selengut, J.D., Sigrist, C.J., Scheremetjew, M., Tate, J., Thimmakanathanarat, M., Thomas, P.D., Wu, C.H., Yeats, C. and Yong, S.Y. (2012) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res. 40, D306–D312.

Hwang, E.Y., Song, Q., Jia, G., Specht, J.E., Hyten, D.L., Costa, J. and Cregan, P.B. (2014) A genome-wide association study of seed protein and oil content in soybean. BMC Genom. 15, 1.

Jaiswal, P., Anravhal, S., Ilic, K., Kellogg, E.A., McCouch, S., Pujara, A., Reiser, L., Rhee, S.Y., Sachs, M.M., Schaeffer, M., Stein, L., Stevens, F., Vincent, L., Ware, D. and Zhai, F. (2006) Plant ontology (PO): a controlled vocabulary of plant structures and growth stages. Comp. Funct. Genomics, 6, 388–397.

Johnson, A.D. and O’Donnell, C.J. (2009) An open access database of genome-wide association results. BMC Med. Genet. 10, 6.

Johnson, C., Bowman, L., Adai, A.T., Vance, V. and Sundaresan, V. (2007) CSRDB: a small RNA integrated database and browser resource for cereals. Nucleic Acids Res. 35, D829–D833.

Jones, P.A. (2012) Functions of DNA methylation: islands, start sites, gene transcription, epigenetic memory and disease. Trends Genet. 28, 105–113.

Joshi, T., Fitzpatrick, M.R., Chen, S., Liu, Y., Zhang, H., Endacott, R.Z., Jones, P.A. (2012) Functions of DNA methylation: islands, start sites, gene transcription, epigenetic memory and disease. Trends Genet. 28, 105–113.

Kang, Y.J. (2015) Small-scale duplication as a genomic signature for crop domestication: an historical review and quantitative analysis of 203 global food crops. New Phytol. 196, 29–48.

Mamidi, S., Lee, R.K., Goos, J.R. and McClean, P.E. (2014) Genome-wide association studies identifies seven major regions responsible for iron deficiency chlorosis in soybean (Glycine max). PLoS ONE, 9, e107469.

Meyers, R.S., DuVal, A.E. and Jensen, H.R. (2012) Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. New Phytol. 196, 29–48.

Monaco, M.K., Stein, J., Nathani, S., Wei, S., Dharmawardhana, P., Kumar, S., Amraasinghe, V., Youens-Clark, K., Thomason, J., Prece, J., Parnenak, S., Olson, A., Jiao, Y., Lu, Z., Bolser, D., Kerhomou, A., Stelns, D., Walts, B., Wu, G., D’Eustachio, P., Haw, R., Croft, D., Kersey, P.J., Stein, L., Jaiswal, P. and Ware, D. (2014) Gramene 2013: comparative plant genome resources. Nucleic Acids Res. 42, D1193–D1199.

Nagarajan, N. and Pop, M. (2013) Sequence assembly demystified. Nat. Rev. Genet. 14, 157–167.

Ngo, L.T., Okogon, J.I. and Folk, W.R. (2013) 21st century natural product research and drug development and traditional medicines. Nat. Prod. Rep. 30, 584–592.

Ozsolak, F. and Milos, P.M. (2011) RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet. 12, 87–98.

 Pace, J., Gardner, C., Romay, C., Ganapathybrumanian, B. and Lubberstedt, T. (2015) Genome-wide association analysis of seedling root development in maize (Zea mays L.). BMC Genom. 16, 47.

Paterson, A.H., Freeling, M., Tang, H.B. and Wang, X.Y. (2010) Insights from the comparison of plant genome sequences. Annu. Rev. Plant Biol. 61, 349–372.

Rafalski, J.A. (2010) Association genetics in crop improvement. Curr. Opin. Plant Biol. 13, 174–180.

Samayoa, L., Malvar, R., Olukolu, B.A., Holland, J.B. and Buttron, A. (2015) Genome-wide association study reveals a set of genes associated with resistance to the Mediterranean corn borer (Sesamia nonagrioides L.) in a maize diversity panel. BMC Plant Biol. 15, 35.

Sato, M., Tsuda, K., Wang, L., Collier, J., Watanabe, Y., Glazerbrook, J. and Katagiri, F. (2010) Network modeling reveals prevalent negative regulatory relationships between signaling sectors in Arabidopsis immune signaling. PLoS Pathog. 6, e1001011.

Sato, Y., Takehisa, H., Kamatsuki, K., Minami, H., Namiki, N., Ikawa, H., Ohyanagi, H., Sugimoto, K., Antonio, B.A. and Nagamura, Y. (2013) RiceXPro version 3.0: expanding the informatics resource for rice transcriptome. Nucleic Acids Res. 41, D1206–D1213.

Schmitz, R.J., He, Y., Valdes-Lopez, O., Khan, S.M., Jost, T., Urich, M.A., Nery, J.R., Diers, B., Xu, D., Stacey, G. and Ecker, J.R. (2013) Epigenome-wide inheritance of cytokine methylations variants in a recombinant inbred population. Genome Res. 23, 1663–1674.

Singh, N., Choudhury, D.R., Singh, A.K., Kumar, S., Srinivasan, K., Tyagi, R.K., Singh, N.K. and Singh, R. (2013) Comparison of SSR and SNP markers in estimation of genetic diversity and population structure of Indian rice varieties. PLoS ONE 8, e84136.

Skinner, M.E., Uzlov, A.V., Stein, L.D., Mungall, C.J. and Holmes, I.H. (2009) JBrowse: a next-generation genome browser. Genome Res. 19, 1630–1638.

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redona, E., Atlin, G., Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redona, E., Atlin, G., Jannink, J.L. and McCouch, S.R. (2015) Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS Genet. 11, e1004982.

Su, G., Christensen, O.F., Janss, L. and Lund, M.S. (2014) Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. J. Dairy Sci. 97, 6547–6559.

Suwamw, W.B., Pixley, K.V., Palacios-Rojas, N., Kaeppler, S.M. and Babu, R. (2015) Genome-wide association analysis reveals new targets for carotenoid biofortification in maize. Theor. Appl. Genet. 128, 851–864.

The 3000 Rice Genomes Project (2014) The 3,000 rice genomes project. GigaScience, 3, 7.
Tian, F., Bradbury, P.J., Brown, P.J., Hung, H., Sun, Q., Flint-Garcia, S., Rocheford, T.R., McMullen, M.D., Holland, J.B. and Buckler, E.S. (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat. Genet. 43, 159–162.

Ueda, Y., Frimpong, F., Qi, Y., Matthias, E., Wu, L., Holler, S., Kraska, T. and Frei, M. (2015) Genetic dissection of ozone tolerance in rice (Oryza sativa L.) by a genome-wide association study. J. Exp. Bot. 66, 293–306.

Urich, M.A., Nery, J.R., Lister, R., Schmitz, R.J. and Ecker, J.R. (2015) MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. Nat. Protoc. 10, 475–483.

Varshney, R.K., Hossinginton, D.A., Nayak, S.N. and Graner, A. (2009) Molecular plant breeding: methodology and achievements. Methods Mol. Biol. 513, 283–304.

Varshney, R.K., Mohan, S.M., Gaur, P.M., Gangarao, N.V., Pandey, M.K., Bohra, A., Savargaonkar, S.L., Chitikineni, A., Kimurto, P.K., Janila, P., Saxena, K.B., Fikre, A., Sharma, M., Rathore, A., Pratap, A., Tripathi, S., Datta, S., Chaturvedi, S.K., Mallikarjuna, N., Anuradha, G., Babbar, A., Choudhary, A.K., Mhase, M.B., Bharadwaj, C., Mannur, D.M., Harer, P.N., Guo, B., Li, X., Nadarajan, N. and Gowda, C.L. (2013) Achievements and prospects of genomics-assisted breeding in three legume crops of the semi-arid tropics. Biotechnol. Adv. 31, 1120–1134.

Varshney, R.K., Kudapa, H., Pazhamalai, L., Chitikineni, A., Thudi, M., Bohra, A., Gaur, P.M., Janila, P., Fikre, A., Kimurto, P. and Ellis, N. (2014) Translational genomics in agriculture: some examples in grain legumes. Crit. Rev. Plant Sci. 34, 169–194.

Wang, M., Yan, J., Zhao, J., Song, W., Zhang, X., Xiao, Y. and Zheng, Y. (2012) Genome-wide association study (GWAS) of resistance to head smut in maize. Plant Sci. 196, 125–131.

Wang, C., Yang, Y., Yuan, X., Xu, Q., Feng, Y., Yu, H., Wang, Y. and Wei, X. (2014) Genome-wide association study of blast resistance in indica rice. BMC Plant Biol. 14, 311.