Sanjeev Gupta · Nagasamy Nadarajan
Debjyoti Sen Gupta *Editors*

Legumes in the Omic Era

Springer

Legumes in the Omic Era

Sanjeev Gupta • Nagasamy Nadarajan
Debjyoti Sen Gupta
Editors

Legumes in the Omic Era

Springer

Editors
Sanjeev Gupta
Indian Institute of Pulses Research
Kanpur, UP, India

Nagasamy Nadarajan
Indian Institute of Pulses Research
Kanpur, UP, India

Debjyoti Sen Gupta
Department of Plant Sciences
North Dakota State University
Fargo, ND, USA

ISBN 978-1-4614-8369-4 ISBN 978-1-4614-8370-0 (eBook)
DOI 10.1007/978-1-4614-8370-0
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013951148

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Enormous amount of biological information are available today, particularly after the completion of whole genome sequencing project in legumes like *Medicago truncatula*, Soybean, Pigeonpea and Chickpea. Large-scale sequencing projects on two other legumes, *Lotus japonicus* and *Vigna radiata*, are also near completion. The information generated from genome sequencing calls for producing a complete functional interpretation of whole genome. This demand coincides with another technological development in plant biology called "OMICS" revolution. The technologies such as transcriptomics, proteomics and metabolomics are being developed in major legume species with the aim to analyze molecular data on a genome-wide scale. These developments are now becoming major landmarks in understanding legume biology in a precise manner.

The present book is an excellent review of the recent advances in the grain legumes' genomics research and applications. In this book efforts have been made to gather and present available recent information for individual grain legume species in a logical order. Genomic resources, structural and functional genomics, progress towards whole genome sequencing and use of genome sequence information in crop improvement are major aspects which are described in detail for each grain legume species in respective chapters. More focus is given to showcase the potential and practical use of genomic tools and resources available today in these species for crop improvement. Information is also shared on the advances in bioinformatics tools and techniques in grain legumes research. The genomic tools' used in revealing legume genome evolution are also discussed in detail. Legume biofortification research and importance of genomic tools in nutritional improvement of grain legumes are presented briefly.

This book contains 15 chapters authored by scientists/researchers who are actively involved in analyzing and improving particular legume genome. Their contribution is enormous in presenting up-to-date information on the subject. Some figures included by the authors in the respective chapter were published elsewhere previously. The necessary permission has been obtained by the authors to use them again for their chapters. We record our acknowledgements to all such publishers and authors for their generosity and goodwill. There are many people around the globe

who were there always during the entire developmental period of this book influencing positively to make this project feasible for us: Dr. Shiv Kumar, ICARDA, Rabat, Morocco, Dr. Rajeev Varshney, ICRISAT, Patancheru, India, Dr. Clare J Coyne, USDA/ARS, WSU, Pullman, USA, and Dr. Dil Thavarajah, NDSU, Fargo, USA.

We editors are highly thankful to Dr. S. Ayyappan, Director General, ICAR, and Secretary, DARE, Ministry of Agriculture, Govt. of India for being the source of guidance and encouragement to take up this project. Thanks are also due to Dr. S. K. Datta (DDG, Crop Science, ICAR) and Dr. B. B. Singh (ADG, Oilseeds and Pulses, ICAR) for continuous support and encouragement. We thank our parent organization, Indian Council of Agricultural Research (ICAR) for supporting the scientific endeavor like "Legumes in the Omic Era".

We thank our families for being patient and supportive in this long journey, without their moral support it would not be possible. We thank Misses Hannah Smith and Melissa Higgs (Associate Editors, Springer Science+Business Media, New York) and Mr. Michael D. Sova (Developmental Editor) for enormous support to make the publication a reality. Mr. Ashutosh Kumar, Senior Research Fellow, Crop Improvement Division of Indian Institute of Pulses Research, Kanpur, is highly appreciated for typographical help.

Kanpur, UP, India — Sanjeev Gupta
Kanpur, UP, India — Nagasamy Nadarajan
Fargo, ND, USA — Debjyoti Sen Gupta

Contents

Contributors

Tuba K. Anjum, M.Sc. Crop Improvement Division, Indian Institute of Pulses Research, Kanpur, Uttar Pradesh, India

Francisco J.L. Aragao, Ph.D. Embrapa Recursos Genéticos e Biotecnologia, Norte Brasília, DF, Brazil

Michael Baum, Ph.D. Biodiversity and Integrated Gene Management (BIGM) Program, Center for Agricultural Research in the Dry Areas (ICARDA), Aleppo, Syria

Abhishek Bohra, Ph.D. Crop Improvement Division, Indian Institute of Pulses Research, Kanpur, Uttar Pradesh, India

P. Dhanasekar, M.Sc. (Ag.) Nuclear Agriculture and Biotechnology Division, Bhabha Atomic Research Centre, Mumbai, Maharashtra, India

Rebecca Dickstein, Ph.D. Department of Biological Sciences, University of North Texas, Denton, TX, USA

Pooran M. Gaur, Ph.D. Grain Legume Genetics Program, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, Andhra Pradesh, India

Debjyoti Sen Gupta, M.Sc. (Ag.) Cereal Science Graduate Program, Department of Plant Sciences, North Dakota State University, Fargo, ND, USA

Sanjay Gupta, Ph.D. Crop Improvement Section, Directorate of Soybean Research, Khandwa Road, Indore, MP, India

Sanjeev Gupta, Ph.D. Indian Institute of Pulses Research, Kanpur, Uttar Pradesh, India

Aladdin Hamwieh, Ph.D. Biodiversity and Integrated Gene Management (BIGM) Program, International Center for Agricultural Research in the Dry Areas (ICARDA), Aleppo, Syria

S.M. Husain, Ph.D. Crop Improvement Section, Directorate of Soybean Research, Indore, MP, India

Arvind M. Kayastha, Ph.D. Faculty of Sciences, School of Biotechnology, Banaras Hindu University, Varanasi, Uttar Pradesh, India

Eva Konečná, Ing. Department of Plant Science, Mendel University in Brno, Brno, Czech Republic

Jitendra Kumar, Ph.D. Crop Improvement Division, Indian Institute of Pulses Research, Kanpur, Uttar Pradesh, India

Shiv Kumar, Ph.D. Biodiversity and Integrated Gene Management (BIGM) Program, International Center for Agricultural Research in the Dry Areas, Rabat, Morocco

Giriraj Kumawat, Ph.D. Crop Improvement Section, Directorate of Soybean Research, Indore, MP, India

Marc Libault, Ph.D. Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK, USA

Nupur Malviya, Ph.D. Crop Improvement Division, Institute of Pulses Research, Kanpur, Uttar Pradesh, India

Alagu Manickavelu, Ph.D. Biodiversity and Integrated Gene Management (BIGM) Program, International Center for Agricultural Research in the Dry Areas, Rabat, Morocco

Pooja Bhatnagar Mathur, Ph.D. International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, Andhra Pradesh, India

Phillip E. McClean, Ph.D. Department of Plant Sciences, North Dakota State University, Fargo, ND, USA

Fred J. Muehlbauer, Ph.D. Grain Legume Genetics Physiology Research Unit, USDA, WSU, Pullman, WA, USA

Nagasamy Nadarajan, Ph.D. Indian Institute of Pulses Research, Kanpur, Uttar Pradesh, India

Juan M. Osorno, Ph.D. Department of Plant Sciences, North Dakota State University, Fargo, ND, USA

Aditya Pratap, Ph.D. Crop Improvement Division, Indian Institute of Pulses Research, Kanpur, Uttar Pradesh, India

Neha Rajan, M.Sc. Krishi Vigyan Kendra, Ramakrishna Mission Ashram, Ranchi, Jharkhand, India

S.V. Ramesh, Ph.D. Crop Improvement Section, Directorate of Soybean Research, Indore, MP, India

Milind B. Ratnaparkhe, Ph.D. Crop Improvement Section, Directorate of Soybean Research, Indore, MP, India

Gopesh C. Saha, Ph.D. Department of Crop and Soil Sciences, Washington State University, Pullman, WA, USA

C.V. Sameerkumar, Ph.D. Pigeonpea Breeding, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, Andhra Pradesh, India

K.B. Saxena, Ph.D. Pigeonpea Breeding, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, Andhra Pradesh, India

Rachit K. Saxena, Ph.D. Center of Excellence in Genomics (CEG), International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, Andhra Pradesh, India

Tilak Raj Sharma, Ph.D. Department of Agricultural Biotechnology, CSK Himachal Pradesh Agricultural University, Palampur, Himachal Pradesh, India

A.K. Singh, Ph.D. Department of Genetics and Plant Breeding, Institute of Agricultural Sciences, Banaras Hindu University, Varanasi, Uttar Pradesh, India

B.D. Singh, Ph.D. Faculty of Sciences, School of Biotechnology, Banaras Hindu University, Varanasi, Uttar Pradesh, India

B.B. Singh, Ph.D. Department of Soil and Crop Sciences, Texas A&M University, College Station, TX, USA

Mritunjay Singh, M.Sc. Crop Improvement Division, Indian Institute of Pulses Research, Kanpur, Uttar Pradesh, India

Vinay Kumar Singh, M.Sc. Faculty of Sciences, Centre for Bioinformatics, School of Biotechnology, Banaras Hindu University, Varanasi, Uttar Pradesh, India

Petr Smýkal, Ph.D. Department of Botany, Palacky University in Olomouc, Olomouc, Czech Republic

J. Souframanien, Ph.D. Nuclear Agriculture and Biotechnology Division, Bhabha Atomic Research Centre, Mumbai, Maharashtra, India

Srinivasan Samineni, Ph.D. Grain Legume Genetics Program, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, Andhra Pradesh, India

Ekta Srivastava, M.Sc. Crop Improvement Division, Indian Institute of Pulses Research, Kanpur, Uttar Pradesh, India

Dil Thavarajah, Ph.D. School of Food Systems, North Dakota State University, Fargo, ND, USA

Pushparajah Thavarajah, Ph.D. School of Food Systems, North Dakota State University, Fargo, ND, USA

Mahendar Thudi, Ph.D. Grain Legume Genetics Program, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, Andhra Pradesh, India

Michael P. Timko, Ph.D. Department of Biology, University of Virginia, Charlottesville, VA, USA

Rakhi Tomar, M.Sc. Crop Improvement Division, Indian Institute of Pulses Research, Kanpur, Uttar Pradesh, India

Rajeev K. Varshney, Ph.D. Center of Excellence in Genomics (CEG), International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, Andhra Pradesh, India

Chapter 1
Legumes in Omic Era: Retrospects and Prospects

Sanjeev Gupta, Nagasamy Nadarajan, and Debjyoti Sen Gupta

Abstract Legumes are important for nutritional security of humans and livestocks and ecological sustainability of agricultural production systems of the world. The adaptability and productivity of legumes are limited by major biotic and abiotic stresses. Therefore, there is a crucial need to increase tolerance against various stresses, which is a major challenge in legume improvement programs for enhancing yield. Breeding methods complemented adequately by genomics approaches could lead to simpler and more effective gene-based approach for legume improvement. This requires adequate genomic resources and information for each legume species of economic importance. Major developments made in recent past, like genome sequencing, the "omic" research and bioinformatics have provided scope for utilization of genomic resources for legume improvement. A good progress has been made in genome sequencing of some legumes and this will increase even more due to novel sequencing technologies called next generation sequencing. Since the release of genome sequences of *Lotus japonicus*, *Glycine max*, *Medicago truncatula*, *Cajanus cajan* and *Cicer arietinum*, a number of comprehensive tools such as bioinformatics tools for sequence assembly and functional annotation, microarray platforms for high-throughput gene expression, transformation systems, and large cDNA and gDNA libraries have been developed for important legumes. These tools need to be integrated to understand genome structure and function of legumes. More comprehensive approaches, including quantitative and qualitative analyses of gene expression

S. Gupta, Ph.D. (✉)
Indian Institute of Pulses Research,
Kanpur, Uttar Pradesh 208024, India
e-mail: saniipr@rediffmail.com

N. Nadarajan, Ph.D.
Indian Institute of Pulses Research,
Kanpur, Uttar Pradesh, India

D.S. Gupta, M.Sc. (Ag.)
Cereal Science Graduate Program, Department of Plant Sciences, North Dakota State University, Fargo, ND, USA

S. Gupta et al. (eds.), *Legumes in the Omic Era*, DOI 10.1007/978-1-4614-8370-0_1,

products are further necessary at the transcriptomic, proteomic, and metabolomic levels for better understanding the functioning of genomes and their gene, including their regulatory networks by combining at computational approaches with translational genomics. The progress in omics research will considerably contribute to better understanding of the molecular and genetic basis of yield and tolerance to biotic and abiotic stress and accelerate molecular and transgenic breeding of legumes of economic importance.

Keywords Legumes • Genome sequence • Transcriptomics • Proteomics • **Metabolomics**

Introduction

Legumes are the third largest family of higher plants with more than 20,000 species having major impact on agriculture, human and livestock nutrition and environment. These are second only to grasses in agricultural importance (Doyle 2001). Major grain legumes like common beans (*Phaseolus* spp.), pea (*Pisum sativum* L.), chickpea (*Cicer arietinum* L.), broad bean (*Vicia faba* L.), pigeonpea (*Cajanus cajan* L.), cowpea (*Vigna unguiculata* L.), and lentil (*Lens esculentum* L.) together constitute 33 % of dietary protein needs of human (Vance et al. 2000). Moreover, grain legumes, predominantly soybean (*Glycine max* L.) and peanut (*Arachis hypogeae* L.), are also a major source for vegetable oil, providing more than 35 % of the world's processed vegetable oil. Among fodder legumes, medics (*Medicago* spp. L.), clovers (*Trifolium* spp.), Vetches (*Vicia* spp.) and Stylos (*Stylosenthes* spp.) assume major importance for animal productivity and nitrogen economy in most parts of the world.

Most of the grain legumes are attributed for low yield. The adaptability and productivity of legumes are limited by major biotic and abiotic stresses, including fungal and viral diseases, insect pests, drought, heat, frost, chilling, water-logging, salinity and mineral toxicities (Dita et al. 2006). Fusarium wilt, blights and viral diseases inflicts severe losses. Similarly, pod borers and sucking pests also causes serious problems in crop management. Foliar and root diseases in forage legumes also constitute major production constraints. Hence, there is a crucial need to increase the abiotic and biotic stress tolerance in legumes, which is a major challenge in crop improvement programs for enhancing yield. Although conventional plant breeding and enhanced management strategies have addressed several constraints that limit crop productivity or quality, there are situations where the existing genetic resources lack the requisite traits. Breeding methods complemented adequately by genomics approaches could lead to simpler and more effective gene-based approach for legume improvement. This requires adequate genomic resources and information for each legume species of economic importance.

In the beginning, the progress in legume biology has been made by the development of model systems to investigate the genetics of important traits. Like *Arabidopsis thaliana*, the two legumes, *Lotus japonicus* and *Medicago truncatula*,

due to their small and diploid genomes, autogamous nature, short generation times, and prolific seed production were emerged out as model legume plant systems (Cook 1999; Handberg and Stougaard 1992). Since then, powerful genetic and genomic tools have been developed that include genome sequencing (Kato et al. 2003), isolation of expressed sequence tags (ESTs) (Asamizu et al. 2004; Kulikova et al. 2001), and establishment of genetic and physical maps for each model species (Pedrosa et al. 2002; Thoquet et al. 2002). The increasing wealth of genetic and genomic data and the high degree of synteny between legume genomes (Kalo et al. 2004; Stracke et al. 2004) make these two species valuable models for the molecular genetic study of various traits related to increased productivity of legumes.

Recent advances in plant genomics have moved beyond model systems to various plant species of economic importance. Since the release of genome sequences of *Arabidopsis* and rice in the past (Goff et al. 2002; Yu et al. 2002; Lin et al. 1999; Mayer et al. 1999) and of *Lotus japonicus*, *Glycine max*, *Medicago truncatula*, *Cajanus cajan* and *Cicer arietinum* recently (Sato et al. 2008; Schmutz et al. 2010; Young et al. 2011; Singh et al. 2012; Varshney et al. 2010, 2011), a number of comprehensive tools such as bioinformatics tools for sequence assembly and functional annotation, microarray platforms for high-throughput gene expression, transformation systems, and large cDNA and gDNA libraries have been developed for a range of species, including the important legumes. Now a major challenge is to integrate these various tools to better understand genome structure and function.

While sequence information is invaluable and a necessary starting point, it is insufficient to answer questions concerning gene function, regulatory networks, and the biochemical pathways activated in response to biotic and abiotic stresses. To address these questions, more comprehensive approaches, including quantitative and qualitative analyses of gene expression products are necessary at the transcriptomic, proteomic, and metabolomic levels. These developments will provide opportunities for better understanding the functioning of genomes and their gene, including their regulatory networks by combining at computational approaches with translational genomics.

Retrospects

Three major developments made in recent past, like genome sequencing, the OMICS research and bioinformatics have revolutionized the plant biology (Weckwerth 2011a, b). A good progress has been made in genome sequencing of some legumes and this will increase even more due to novel sequencing technologies called next generation sequencing (Weckwerth 2011a, b; Ideker et al. 2001). Based on sequence information, genome assembly is developed. After the assembly of a full genome, functional annotation is established for each sequence assembly. Predicted genes are searched for homology against databases of characterized genes and proteins. It is obvious that this initial functional annotation is not capable of producing a complete functional interpretation of the whole genome and a prediction of the molecular phenotype (Weckwerth 2011a, b). Consequently, the molecular

phenotype needs to be measured for the functional interpretation of the genotype. This requires the support of "omics" research. A modest beginning has been made in this area for legumes. Recent progress indicates that appropriate bioinformatics platform has also been developed for utilization of sequence and "omic" information for improvement of legumes.

Genome Sequencing

The nuclear genomes of legumes vary greatly in size, from 370 Mbp in *Lablab niger* to as large as 13,000 Mbp of *Vicia faba*. Efforts have been made to generate complete sequence information of some important legumes. The whole genome sequencing of three legumes viz., *Medicago*, *Lotus* and *Glycine* has been completed till date (Young et al. 2011; Sato et al. 2008; Schmutz et al. 2010). The draft genome sequences of *Cajanus* and *Cicer* are already available (Singh et al. 2012; Varshney et al. 2010, 2011). Genome sequencing in some other legumes like mungbean, peas, alfalfa, peanut, cowpea and common bean are at various stages of progress, with the later being expected to be completed shortly. With the advent of next generation sequencing, the task of whole genome sequencing of these crops can be efficiently completed. The next-generation sequencing not only is a dramatic advance over capillary-based sequencing but also meets significant challenges in assembly and sequence accuracy due to short read lengths, method-specific sequencing errors, and the absence of physical clones. However, the promise of much lower sequencing cost with the now proven concept of next-generation expressed sequence tag sequencing which will allow assessment of plant genomes at least at the functional level (Ohtsu et al. 2007).

"OMICS" Research

"Omics" research involves functional genomics, transcriptome profiling, proteomics, and metabolomics for analyzing molecular data of living systems on a genome scale (Somerville and Dangl 2000; Weckwerth 2003; Ideker et al. 2001). This provides genome-scale molecular data in combination with a genomic template. The ultimate goal is to derive a model of metabolism that is driven by genome data and predicts the phenotype (Weckwerth 2011a, b). Transcriptomics, proteomics and metabolomics data can be exploited for gene prediction and functional gene annotation in fully sequenced organisms (Castellana et al. 2008; Wienkoop et al. 2010). Major studies in plant model systems such as *Arabidopsis thaliana* and *Chlamydomonas reinhardtii* have demonstrated the applications of proteo- and metaproteogenomics (Castellana et al. 2008; Wienkoop et al. 2010). The progress in omics research will considerably contribute to better understanding of the molecular and genetic basis of yield and tolerance to biotic and abiotic stress that has been an important bottleneck for legume molecular and transgenic breeding.

Functional Genomics

Since the 1990s, genomics has been the most active research field in biological science generating a huge amount of information, while structural genomics has emerged at the methodological level to understand gene expression and function. A broad range of genomic resources has been developed to accelerate legume improvement. These include expression sequence tag (EST) database, genome sequences (whole or partial), physical maps, molecular maps, DNA chips and bacterial artificial chromosome (BAC) or similar genomic libraries. BAC libraries have been widely used in different aspects of genome research. The value of large insert libraries has long been recognized in genome analysis. BAC libraries are important genomic resources that have been used for (1) physical mapping of genomes, (2) cost effective molecular marker development of microsatellite markers (Shultz et al. 2007), (3) map-based cloning of genes or QTL for important agronomic traits, (4) evolutionary study of multigene families, (5) karyotyping genomes through BAC-FISH, and (6) whole genome sequencing. BAC libraries have been constructed for many species and are usually developed by cloning size-fractionated DNA fragments partially digested with restriction enzymes. In near future, the BACs of legumes should have potential applications in comparative genomics and functional genomics as well owing to the macro- and microsynteny widespread within legumes. Among the grain legumes, soybean has been more intensively studied and according to the legume information system data, over 1.3 million ESTs were developed from different cDNA libraries, which is the largest in number among the individual grain legume ESTs. The availability of a large number of EST and BAC sequences facilitated the discovery of new SNP and SSR markers toward the construction of high-resolution genetic maps of various legumes. With the availability of whole genome sequence information, large numbers of ESTs are identified for biotic and abiotic response in two other legumes, chickpea and pigeonpea. However, progress made in this area is also satisfactory in cowpea and groundnut.

Transcriptomics

Unlike genome analysis, transcriptome analysis offers a full profile of gene function information under various conditions, and it differs with dissimilar environments, cell types, developmental stages, and cell states (Moe et al. 2011). It provides a powerful tool for differential gene expression, mutant splicing, SSR or SNP analysis, and functional genetics studies. The typical analysis of the dynamic transcriptome is usually performed with microarray technology and is one of the pioneering genome-scale, hypothesis-free screening methods. Several large-scale studies have revealed differential gene expression under different conditions and almost every gene in *Arabidopsis thaliana* is already characterized based on RNA sequencing data under specific conditions (www.arabidopsis.org). Nowadays, NGS provides an alternative technology for RNA sequencing (Brautigam and Gowik 2010; Wang et al. 2010). However, this technology is still in development and very expensive

because several fold genome coverage has to be measured to obtain statistically significant data. With development of new technologies it is expected that transcriptomic analysis will be performed extensively in major legumes in near future.

Since trancriptomics in legumes started its beginning, it is desirable that gene expression studies performed in *Arabidopsis*, and resulting knowledge from such studies be used in legumes through comparative genomics. For example, Ishitani et al. (2004) selected 100–200 genes from the *Arabidopsis* database and showed that at least three DREB-like genes, thought to be key transcriptional regulators of drought and/or cold tolerance, in common bean. Similarly, in Arabidopsis, analysis of the transcriptome changes occurring during cold, drought, and salt stress in a survey of 7,000 genes showed a shared response for a majority of cold and drought stress regulated genes (Seki et al. 2002). The Arabidopsis model is likely to be very different from legumes in terms of responses to stress in relation to grain filling, nitrogen utilization, fixation, and transport, root architecture, and interactions, all physiological processes that are fundamentally different in legumes. Hence, the usefulness of developing a legume model has become increasingly relevant in recent years. In legumes, the gene expression patterns following biotic stresses have been more extensively studied than those following abiotic stresses. With respect to abiotic stress, gene expression analyses have been mainly based on studies with cloned genes (Singh et al. 2004). Significant progress is being made at the genetic and genomic levels using the model legume *M. truncatula* through macro- and microarray analysis, reverse genetics, genome sequencing, and other high-throughput techniques (Thompson et al. 2005; Oldroyd 2005). The analysis of almost 200,000 ESTs of *M. truncatula*, isolated from many different libraries constructed from diverse stages and treatments, was facilitated by searchable databases such as MtDB2 (Lamblin et al. 2003) and the TIGR Gene Index (http://www.tigr.org). Recent reports have also shown that transcriptomic tools are a good option for legume breeding to abiotic and biotic stresses.

Proteomics

Proteins act directly on biochemical processes, and thus must be closer to the phenotype, compared to DNA-based markers. The studies on proteins expressed by the genome of a cell, tissue, or organism at a specific time (proteome) is necessary to understand the biological function of a cell or an organism. Although research on plant responses to stress on the DNA or RNA level provided an important insight into stress tolerance, the proteomics approach is very important in evaluating stress responses since the mRNA levels may not always correlate with protein accumulation (Gygi et al. 1999). In addition, many proteins are modified by posttranslational modifications such as phosphorylation, glucosylation, and ubiquitinylation, which significantly influence protein functions. Proteomics, understood as protein biochemistry on an unprecedented and high-throughput scale, is becoming a promising and active approach in this postgenomic period. However, its application to plants is rather limited compared to other biological systems (Jorrin et al. 2006), although

good technical progress has been achieved in the separation of proteins and their identification by mass spectrometry. Studies have evaluated changes in protein levels in plant tissues in response to stresses (Canovas et al. 2004; Kim et al. 2003). However, these studies have mainly focused on nonlegume species such as Arabidopsis and rice (Canovas et al. 2004) and some legumes recently (Jorrin et al. 2006). As a result, only a handful of studies have been carried out in legumes, although in the next few years there should be a significant increase in the number of legume species. So far, pea has been more intensively studied, with the analysis of induced protein expression in roots in response to salt (Kav et al. 2004) and to cadmium stress (Repetto et al. 2003). Recently, *M. truncatula* has been the subject of several proteomic studies that represent the most extensive proteomic description to date and provide a reference map for future comparative proteomics and functional genomics studies of biotic and abiotic stress responses on legumes (Lei et al. 2005).

Metabolomics

Metabolomics provide the most direct tools for the quantitative measurement of the metabolism in an organism. Transcriptomic and proteomic data are important in deciphering a complex biological process, but they are still insufficient since most biological processes are ultimately mediated by cell metabolites. Thus, the complete understanding of both gene function and molecular events controlling complex plant processes requires analysis of transcriptome, the proteome, and the metabolome in an integrative manner (Dixon 2001). Metabolite profiling and metabolic fingerprinting are two different approaches in metabolomics that can be used for a large range of applications, including phenotyping of genetically modified plants, substantial equivalence testing, determining gene function, and monitoring responses to biotic and abiotic stresses. Recently, a promising platforms for metabolomics has been developed with the combination of two-dimensional gas chromatography and fast acquisition rate mass spectrometry (Scherling et al. 2010). Due to their specific technology, both technologies provide a complementary view of the metabolome such as amino acids, sugars, organic acids, free fatty acids, etc. However, most of the metabolomics platforms still need further method validation and quality checks. This is an essential requirement to guarantee meaningful biological applications. Furthermore, databases, experimental standards and data exchangeability between labs is an urgent issue for further developments in metabolomics (Weckwerth 2011a, b; Sansone et al. 2007).

In legumes, the metabolomic approach has been used in *M. truncatula* to determine the responses to various stimuli (Bell et al. 2001). Although, large-scale comprehensive metabolomic studies are difficult, a number of targeted analyses have been performed to assess the involvement of subsets of metabolites in various stresses. Although the preliminary results from combining metabolic approaches with transgenics indicates the potential of increasing intrinsic stress resistance levels in legume crops and strengthens the potential role of "omic" research in crop improvement (He and Dixon 2000; Wu and Van Etten 2004), it must be emphasized

that most metabolic pathways are interconnected in highly complex networks. Thus, modulating one metabolic pathway may have negative impacts on another, leading to concomitant deleterious traits in the modified crop. Large-scale metabolic analyses are therefore necessary to observe the metabolic networks important for plant growth and development under a range of environmental conditions.

Bioinformatics Platform

"Omic" era in the twenty-first century gives us opportunities to understand the legume genome at sequence-structural-functional levels. The rapid development of various genomic tools and techniques including large scale analysis of genome organization, gene expression, protein-protein interaction etc. are generating enormous amount of data which need to be analyzed and interpreted to develop a biologically meaningful concepts. The genome sequencing projects on different legumes generated the wealth of sequence data. These data need to be properly analysed to enable prediction of the potential functional elements, genes and transcription factors. Bioinformatics tools and databases help us in the analysis as well as understanding of the various features of the sequenced genome (Kushwaha et al. 2011; Dutt et al. 2010; Kumari et al. 2010). The availability of different biological databases related to legumes provides valuable information resource for research and analysis. Illustrated Legume Genetic Resources Database (www.gene.affrc.go.jp), LegumeTFDB (www.legumetfdb.psc.riken.jp) Bioinformatics resources for legume researchers (www.legumes.org), Chickpea Transcriptome Database (http://59.163.192.90:8080/ctdb/) are some of the important bioinformatics platforms providing important resources for legumes. These experimental datasets give us opportunities to understand the functional and biological roles of unknown genes/proteins from different legumes. Most of the assembler tools and packages were also developed e.g., short oligonucleotide analysis package and *de novo* assembly tools were developed by Beijing Genomics Institute (BGI). Several bioinformatics tools are available for annotation, genome sequence alignment, *de novo* assembly, sequence alignments and RNA sequence analysis. The basic level of genome annotation can be done using Basic Local Alignment Search Tool to find out similarities and differences. However, nowadays more and more additional information is added to the annotation platform. Bioinformatics tools developed for many non-legume species provides useful platform for legumes also.

Applications in Crop Improvement

All these technological platforms described above enable the genome-wide molecular analysis of different genotypes. This integrated high throughput analysis of metabolites, proteins and transcripts allow the definition of biochemical

phenotypes and their relationship to the corresponding genotype (Weckwerth 2008). The integration of metabolite and protein profiling has already been demonstrated to significantly improve pattern recognition and the selection and interpretation of multiple physiological and biomarkers for plant systems and different plant genotypes under different environmental conditions such as day-night rhythms or cold stress (Morgenthal et al. 2005; Wienkoop et al. 2008). Integration of metabolite and transcript data was also demonstrated to reveal the relationship between mRNA expression and dynamics of secondary metabolism (Tohge et al. 2005). The exploitation of these technologies in QTL-based marker-assisted breeding approaches (Fernie and Schauer 2009; Collard and Mackill 2008) is an obvious development. Most of the studies are focused on DNA markers. In recent studies the successful application of these technologies was also demonstrated for proteomics and metabolomics. Such efforts need to be accelerated for legumes. Marker-assisted selection could accelerate this process for the identification of useful traits in the early years of the selection process. It is anticipated that new technologies such as genomics, proteomics and metabolomics will yield such marker systems, however, these technologies have hardly reached the stage of application for breeding. A multitude of diagnostic marker assays will therefore be required for marker-assisted selection (Gebhardt et al. 2006). However, the robustness of these markers must be analyzed with higher statistical power from a higher number of samples. Both data sets—the metabolomics data and the proteomics data—show a good cultivar discrimination, however, the sample pattern can be interpreted differently depending on the characteristics of the different cultivars. Thus, the metabolite data carry different information to the protein data. Integration of these data leads consequently to optimized pattern recognition processes and improved interpretation of the molecular data with respect to the molecular phenotype which was indeed observed in several previous studies (Morgenthal et al. 2005; Wienkoop et al. 2008).

Prospects

Legumes are important for nutritional security and ecological sustainability. Exploitation of natural variation, population dynamics and a better understanding of the genotype-phenotype relationship is very crucial for improving productivity and stress adaptation of legumes. The development of genomic research during last decade may lead to a refinement of classical and molecular breeding approaches for legume improvement. A quite progress has been made for genome-scale investigation of some legumes to understand adaptation mechanisms and to provide fundamental knowledge for genetic variation, also for trait selection and genome/marker-assisted breeding approaches. Genomic resources provide the breeders a platform for rapid realization of resistance breeding objectives. A crucial pre-requisite for the deployment of markers to support stress tolerance or resistance breeding is the development of a genetic map, followed by identification of gene-based or gene linked markers to be used in marker assisted selection (MAS). Basic requirement of

availability of genomic resources for successful application of molecular markers in most of the legumes are now in place. Sequencing efforts have already made their strides in complete genome sequencing of few legumes like *M. truncatula*, *L. japonicus* and *Glycine max*, and draft genome sequencing of *Cicer arietinum* and *Cajanus cajan* genomes and the genome sequencing projects of some other legumes. This will generate large scale SNPs, SSRs and intron length polymorphic markers, which can help to saturate the linkage maps. Current genetic linkage maps of most of the legumes display an inadequate level of marker density. To improve the utility of such maps, it will be required to further saturate the map with additional markers.

Large-scale analysis by using different omics technologies are providing extensive data sets that will help identify potential candidate genes for enhanced productivity and stress adaptations in important legume crops. Identification of these candidate genes may allow their direct application in crop improvement through marker assisted breeding. However, in most cases, the roles of these candidate genes remain unknown and it will be important to carry out functional studies as a preliminary step toward their use in genetic improvement. The traditional pursuit of a gene starting with a phenotype (forward genetics) has paved the way for the opposite situation where the gene sequences are known but not their functions. The challenge is to decipher the function of thousands of genes identified by genome projects where reverse genetics methodologies will be the key tools. The ability to knockout genes or suppress their expression are powerful tools to determine the function of a gene. This can be done by antisense RNA suppression, targeted gene replacement, insertional mutagenesis, gene silencing through RNAi, and targeted induced local lesion in genome (TILLING) approaches.

Successful application of omics to legume improvement requires knowledge of stress response at molecular level, which includes gene expression to protein or metabolite and its phenotypic effects. Availability of genome sequence of legumes has a potential to facilitate positional cloning and other approaches and their applications for legumes research. A genome-wide expression profiling with next-generation sequencing approaches could circumvent the various problems in studying the legume genome. Compared to analysis of the transcriptome, analysis of the plant proteome and metabolome in response to abiotic and biotic stresses is still limited to *M. truncatula* and protein reference maps of soybean to stress responses are now available. More recently, few proteomics studies are available on chickpea and groundnut and they have to be extensively carried out for other legumes. Moreover, the recent progress in the mass-scale profiling of the genome, transcriptome, proteome, and metabolome offers the possibility of investigating the concerted response of thousands of genes to biotic and abiotic stresses. The mapping of abiotic stress QTL in legume is still at an early stage and gene pyramiding has not been applied yet. Nevertheless, with the establishment of the model legumes, *M. truncatula* and *L. japonicus*, there is now applicable information on legumes. Among the grain legumes, soybean has been more intensively studied, and the availability of more numbers of ESTs and genome sequences will facilitate mapping of major QTL in other legumes. Rapid progress in legume improvement will be possible with identification of candidate genes for desired traits in legumes. It is now possible

to target almost all legume crops with a variety of omics approaches for genetic improvement. All these efforts will lead to enhanced crop productivity of legume and ensure progress towards attaining nutritional security and ecological sustainability.

References

Asamizu E, Nakamura Y, Sato S, Tabata S (2004) Characteristics of the *Lotus japonicus* gene repertoire deduced from large-scale expressed sequence tag (EST) analysis. Plant Mol Biol 54:405–414

Bell CJ, Dixon RA, Farmer AD, Flores R, Inman J, Gonzales RA, Harrison MJ, Paiva NL, Scott AD, Weller JW, May GD (2001) The *Medicago* Genome Initiative: a model legume database. Nucleic Acids Res 29:114–117

Brautigam A, Gowik U (2010) What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. Plant Biol (Stuttg) 12:831–841

Canovas F, Dumas-Gaudot E, Recorbet G, Jorrin J, Mock HP, Rossignol M (2004) Plant proteome analysis. Proteomics 4:285–298

Castellana NE, Payne SH, Shen ZX, Stanke M et al (2008) Discovery and revision of Arabidopsis genes by proteogenomics. Proc Natl Acad Sci U S A 105:21034–21038

Collard BC, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos Trans R Soc Lond B Biol Sci 363:557–572

Cook DR (1999) *Medicago truncatula*: a model in the making! Commentary. Curr Opin Plant Biol 2:301–304

Dita MA, Rispail N, Prats E, Rubiales D, Singh KB (2006) Biotechnology approaches to overcome biotic and abiotic stress constraints in legumes. Euphytica 147:1–24. doi:10.1007/s10681-006-6156-9

Dixon RA (2001) Natural products and plant disease resistance. Nature 411:843–847

Doyle JJ (2001) Leguminosae. In: Brenner S, Miller JH (eds) Encyclopedia of Genetics. Academic, San Diego, pp 1081–1085

Dutt S, Singh VK, Marla SS, Kumar A (2010) *In silico* analysis of sequential, structural and functional diversity of wheat cystatins and its implication in plant defense. Genomics Proteomics Bioinformatics 8(1):42–56, PMID: 20451161

Fernie AR, Schauer N (2009) Metabolomics-assisted breeding: a viable option for crop improvement? Trends Genet 25(1):39–48

Gebhardt C, Bellin D, Henselewski H, Lehmann W et al (2006) Marker-assisted combination of major genes for pathogen resistance in potato. Theor Appl Genet 112:1458–1464

Goff SA, Ricke D, Lan TH, Presting G, Wang RL, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H et al (2002) A draft sequence of the rice genome (*Oryza sativa* L ssp *japonica*). Science 296:92–100

Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19:1720–1730

Handberg K, Stougaard J (1992) *Lotus japonicus*: an autogamous, diploid legume species for classical and molecular genetics. Plant J 2:487–496

He XZ, Dixon RA (2000) Genetic manipulation of isoflavone 7-Omethyltransferase enhances biosynthesis of 40-O-methylated isoflavonoid phytoalexins and disease resistance in alfalfa. Plant Cell 12:1689–1702

Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet 2:343–372

Ishitani M, Rao I, Wenzl P, Beebe S, Tohme J (2004) Integration of genomics approach with traditional breeding towards improving abiotic stress adaptation: drought and aluminum toxicity as case studies. Field Crop Res 90:35–45

Jorrin JV, Rubiales D, Dumas-Gaudot E, Recorbet G, Maldonado A, Castillejo MA, Curto M (2006) Proteomics: a promising approaches to study biotic interaction in legumes: a review. Euphytica 147:37–47

Kalo P, Seres A, Taylor SA, Jakab J, Kevei Z, Kereszt A, Endre G, Ellis THN, Kiss GB (2004) Comparative mapping between *Medicago sativa* and *Pisum sativum*. Mol Genet Genomics 272L:235–246

Kato T, Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S (2003) Structural analysis of a Lotus japonicus genome V sequence features and mapping of sixty-four TAC clones which cover the 64 Mb regions of the genome. DNA Res 10:277–285

Kav NNV, Srivastava S, Goonewardene L, Blade SF (2004) Proteome-level changes in the roots of *Pisum sativum* in response to salinity. Ann Appl Biol 145:217–230

Kim ST, Cho KS, Yu S, Kim SG, Hong JC, Han CD, Bae DW, Nam MH, Kang KY (2003) Proteomic analysis of differentially expressed proteins induced by rice blast fungus and elicitor in suspension cultured rice cells. Proteomics 3:2368–2378

Kulikova O, Gualtieri G, Geurts R, Kim DJ, Cook D, Huguet T, de Jong JH, Franszn PF, Bisseling T (2001) Integration of the FISH pachytene and genetic maps of *Medicago truncatula*. Plant J 27:49–58

Kumari A, Singh VK, Fitter J, Polen T, Kayastha AM (2010) Alpha-amylase from germinating soybean (*Glycine max*) seeds-purification, characterization and sequential similarity of conserved and catalytic amino acid residues. Phytochemistry 71(14–15):1657–1666, PMID: 20655076

Kushwaha H, Gupta S, Singh VK, Rastogi S, Yadav D (2011) Genome wide identification of D of transcription factor gene family in sorghum and its comparative phylogenetic analysis with rice and *Arabidopsis*. Mol Biol Rep 38(8):5037–5053, PMID: 21161392

Lamblin AFJ, Crow JA, Johnson JE et al (2003) MtDB: a database for personalized data mining of the model legume *Medicago truncatula* transcriptome. Nucleic Acids Res 31:196–201

Lei Z, Elmer AM, Watson BS, Dixon RA, Mendes PJ, Sumner LW (2005) A two-dimensional electrophoresis proteomic reference map and systematic identification of 1367 protein from a cell suspension culture of the model legume *Medicago truncatula*. Mol Cell Proteomics 4:1812–1825

Lin XY, Kaul SS, Rounsley S, Shea TP, Benito ML, Town CD, Fujii CY, Mason T, Bowman CL, Barnstead M et al (1999) Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature 402:761

Mayer K, Schuller C, Wambutt R, Murphy G, Volckaert G, Pohl T, Dusterhoft A, Stiekema W, Entian KD, Terryn N et al (1999) Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. Nature 402:769

Moe KT, Chung JW, Cho Y, Moon JK, Ku JH, Jung JK, Lee J, Park YJ (2011) Sequence information on simple sequence repeats and single nucleotide polymorphisms through transcriptome analysis of mungbean. J Integr Plant Biol 53:67–73

Morgenthal K, Wienkoop S, Scholz M, Selbig J, Weckwerth W (2005) Correlative GC–TOF–MS based metabolite profiling and LC–MS based protein profiling reveal time-related systemic regulation of metabolite–protein networks and improve pattern recognition for multiple biomarker selection. Metabolomics 1:109–121

Ohtsu K, Smith MB, Emrich SJ, Borsuk LA, Zhou R, Chen T, Zhang X, Timmermans MCP, Beck J, Buckner B, Buckner DJ, Nettleton D, Scanlon MJ, Schnable PS (2007) Global gene expression analysis of the shoot apical meristem of maize (*Zea mays* L). Plant J 52:391–404

Oldroyd G (2005) Sequencing the model legume *Medicago truncatula*. Grain Legumes 41:23

Pedrosa A, Sandal N, Stougaard J, Schweizer D, Bachmair A (2002) Chromosomal map of the model legume *Lotus japonicus*. Genetics 161:1661–1672

Repetto O, Bestel-Corre G, Dumas-Gaudot E, Berta G, Gianinazzi-Pearson V, Gianinazzi S (2003) Targeted proteomics to identify cadmium-induced protein modifications in *Glomus mosseae* inoculated pea roots. New Phytol 157:555–567

Sansone SA, Fan T, Goodacre R, Griffin JL et al (2007) The metabolomics standards initiative. Nat Biotechnol 25:846–848

Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K et al (2008) Genome structure of the legume, *Lotus japonicus*. DNA Res 15:227–239

Scherling C, Roscher C, Giavalisco P, Schulze ED, Weckwerth W (2010) Metabolomics unravel contrasting effects of biodiversity on the performance of individual plant species. PLoS One 5:e12569

Schmutz J, Cannon S, Schlueter J, Ma J, Mitros T, Nelson W, Hyten D, Song Q, Thelen J, Cheng J, Xu D, Hellsten U, May G, Yu Y, Sakurai T, Umezawa T, Bhattacharyya M, Sandhu D et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183

Seki M, Narusaka M, Ishida J, Nanjo T, Fujita M, Oono Y, Kamiya A, Nakajima M, Enju A, Akurai T et al (2002) Monitoring the expression profiles of 7000 *Arabidopsis* genes under drought, cold, and high salinity stresses using a full-length cDNA microarray. Plant J 31:279–292

Shultz L, Kazi S, Bashir R, Afzal JA, Lightfoot DA (2007) The development of BAC-end sequence-based microsatellite markers and placement in the physical and genetic maps of soybean. Theor Appl Genet 114:1081–1090

Singh BN, Mishra RN, Agarwal PK, Goswami M, Nair S, Sopory SK, Reddy MK (2004) A pea chloroplast translation elongation factor that is regulated by abiotic factors. Biochem Biophys Res Commun 320:523–530

Singh NK, Gupta DK, Jayaswal PK, Mahato AK, Dutta S, Singh S, Bhutani S et al (2012) The first draft of the pigeonpea genome sequence. J Plant Biochem Biotechnol 21(1):98–112

Somerville C, Dangl J (2000) Genomics. Plant biology in 2010. Science 290:2077–2078

Stracke S, Sato S, Sandal N, Koyama M, Kaneko T, Tabata S, Parniske M (2004) Exploitation of colinear relationships between the genomes of *Lotus japonicus*, *Pisum sativum* and *Arabidopsis thaliana*, for positional cloning of a legume symbiosis gene. Theor Appl Genet 108:442–449

Thompson R, Ratet P, Kuster H (2005) Identification of gene functions by applying TILLING and insertional mutagenesis strategies on microarray-based expression data. Grain Legumes 41:20–22

Thoquet P, Gherardi M, Journet EP, Kereszt A, Ane JM, Prosperi JM, Huguet T (2002) The molecular genetic linkage map of the model legume *Medicago truncatula*: an essential tool for comparative legume genomics and the isolation of agronomically important genes. BMC Plant Biol 2:1

Tohge T, Nishiyama Y, Hirai MY, Yano M et al (2005) Functional genomics by integrated analysis of metabolome and transcriptome of *Arabidopsis* plants over-expressing an MYB transcription factor. Plant J 42:218–235

Vance CP, Graham PH, Allan DL (2000) Biological nitrogen fixation, phosphorus: a critical future need. In: Pedrosa FO, Hungria M, Yates MG, Newton WE (eds) Nitrogen fixation: from molecules to crop productivity. Kluwer, Dordrecht, The Netherlands, pp 506–514

Varshney RK, Glaszmann JC, Leung H, Ribaut JM (2010) More genomic resources for less-studied crops. Trends Biotechnol 28:452–460

Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MTA, Azam S et al (2011) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nat Biotechnol 30:83–89

Wang L, Li P, Brutnell TP (2010) Exploring plant transcriptomes using ultra high-throughput sequencing. Brief Funct Genomics 9:118–128

Weckwerth W (2003) Metabolomics in systems biology. Annu Rev Plant Biol 54:69–89

Weckwerth W (2008) Integration of metabolomics and proteomics in molecular plant physiology—coping with the complexity by data-dimensionality reduction. Physiol Plant 132:176–189

Weckwerth W (2011a) Green systems biology—from single genomes, proteomes and metabolomes to ecosystems research and biotechnology. J Proteomics 75:285–305

Weckwerth W (2011b) Unpredictability of metabolism—the key role of metabolomics science in combination with next-generation genome sequencing. Anal Bioanal Chem 400:1967–1978

Wienkoop S, Morgenthal K, Wolschin F, Scholz M et al (2008) Integration of metabolomic and proteomic phenotypes: analysis of data covariance dissects starch and RFO metabolism from low and high temperature compensation response in Arabidopsis thaliana. Mol Cell Proteomics 7:1725–1736

Wienkoop S, Weiss J, May P, Kempa S et al (2010) Targeted proteomics for Chlamydomonas reinhardtii combined with rapid subcellular protein fractionation, metabolomics and metabolic flux analyses. Mol Biosyst 6:1018–1031

Wu QD, Van Etten HD (2004) Introduction of plant and fungal genes into pea (*Pisum sativum* L) hairy roots reduces their ability to produce pisatin and affects their response to a fungal pathogen. Mol Plant Microbe Interact 17:798–804

Young ND, Debellé F, Oldroyd GE, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KF, Gouzy J, Schoof H, Van de Peer Y, Proost S, Cook DR, Meyers BC, Spannagl M, Cheung F, De Mita S, Krishnakumar V et al (2011) The *Medicago* genome provides insight into the evolution of rhizobial symbioses. Nature 480:520–524

Yu J, Hu SN, Wang J, Wong GKS, Li SG, Liu B, Deng YJ, Dai L, Zhou Y, Zhang XQ et al (2002) A draft sequence of the rice genome (*Oryza sativa* L ssp *indica*). Science 296:79–92

Chapter 2
Advances in Functional Genomics in Legumes

Marc Libault and Rebecca Dickstein

Abstract Functional genomics encompasses RNA transcription, protein expression and metabolomics as well as forward and reverse genetics. In recent years several resources like transcriptomes, proteomes, metabolomes, regulatory elements and mutant libraries using TILLING methods have been developed for legumes. These provide powerful molecular resources to identify the legume genes playing essential roles in plant resistance to biotic and abiotic stresses, regulating protein and oil accumulation in seeds and controlling beneficial plant-microbe symbiosis. Functional genomic studies on model legumes as well as on legumes of economic interest have already provided valuable information for enhancing legume productivity and holds promise for the future.

Keywords Legumes • Functional genomics • Genome • Epigenome • Transcriptome • Proteome • Metabolome • Abiotic stresses • Biotic stresses

Introduction

Functional genomics contributes to molecular breeding by identifying the expressed genes, proteins and metabolites associated with specific traits. It also identifies genes associated with water use efficiency (Kang et al. 2011) including those associated with stomatal opening and closure, nitrogen use efficiency, genes that respond to high temperature stress, those associated with flowering and seed set, those that are activated in response to pathogen attack and multiple other traits that are critical

M. Libault, Ph.D. (✉)
Department of Microbiology and Plant Biology, University of Oklahoma,
770 Van Vleet Oval, Room 32, Norman, OK 73019, USA
e-mail: libaultm@ou.edu

R. Dickstein, Ph.D.
Department of Biological Sciences, University of North Texas, Denton, TX, USA

S. Gupta et al. (eds.), *Legumes in the Omic Era*, DOI 10.1007/978-1-4614-8370-0_2,

for crop improvement. In legumes, functional genomics is helping to identify the genes regulating legume development, yield, their resistance to biotic and abiotic stresses, and other factors affecting their economic value.

Functional genomics in legumes had its start in the late 1990s with the first publication of expressed sequence tags (ESTs) from root hairs and nodules in the model legumes, *Medicago truncatula* and *Lotus japonicus* (Covitz et al. 1998; Szczyglowski et al. 1997), followed by EST analyses in several tissues (Endo et al. 2000; Asamizu et al. 2000; Gyorgyey et al. 2000). Other EST analyses soon followed in *Medicago*, *Lotus*, *Glycine max* and other legumes, with tissues collected from various organs and at different developmental times in model legumes as well as crop legumes (Fedorova et al. 2002; Journet et al. 2002; Shoemaker et al. 2002; Sawbridge et al. 2003; Schroeyers et al. 2004). These ESTs were organized into tentative consensus (TC) sequences and were originally housed at TIGR (Lee et al. 2005); now the TCs are available at DFCI (http://compbio.dfci.harvard.edu/tgi).

Recently, functional genomic studies have been accelerated with the emergence of high-throughput technologies. These technologies have been applied to the characterization of physiological and molecular changes occurring in plants responding to environmental stresses, as well as between organs, tissues or even single cell types. It is assumed that the relative abundance of transcripts, proteins or metabolites would provide an important indication of their role in plant development or in the plant response to a stress.

High-throughput DNA sequencing technologies have been used to develop genetic tools and resources used daily by functional genomicists such as the generation of drafts of legumes genome sequences (e.g. (Schmutz et al. 2010; Young et al. 2011)), the identification of genetic markers such as SNPs (Single Nucleotide Polymorphisms (Hyten et al. 2010; Cortes et al. 2011; Han et al. 2011; Muchero et al. 2011; Shu et al. 2011; Varala et al. 2011; Xu et al. 2011)), the establishment of the epigenome (i.e., genomic DNA methylome profiles and the mapping of histone post-translational chemical modification in the genomic DNA; (Schmitz and Zhang 2011)) and the deep characterization of legume transcriptomes (Benedito et al. 2008; Hogslund et al. 2009; He et al. 2009; Libault et al. 2010a; Severin et al. 2010). Proteomic and metabolomics capabilities have also increased in terms of sensitivity to now allow researchers the ability to identify thousands of proteins and metabolites from smaller and smaller plant tissue sample sizes (e.g. (Watson et al. 2003; Farag et al. 2008; Brechenmacher et al. 2009)).

Development of Resources for Functional Genomics in Legumes

Transcriptomic Resources

Genes that are differentially expressed genes in response to a stress or across organs, tissues or cell types are candidates to have a role in the adaptation of the plant to the stress, in organ development or in functionality. Hence, the establishment of the

transcriptional patterns of genes is a valuable starting point to identify genes controlling a biological process. The quantification of the expression level of an organisms' set of genes, which is reflected by their mRNA abundance, leads to the establishment of the transcriptome of single cell types, tissues and organs.

Arrays have been extensively developed and used during the past years to study the transcriptome of model legumes (Kouchi et al. 2004; Kuster et al. 2004; Suganuma et al. 2004; Vodkin et al. 2004; Lohar et al. 2006; Jones et al. 2008; You et al. 2011; Takanashi et al. 2012; Zahaf et al. 2012). These arrays, which represent a collection of transcript-specific oligonucleotides, allow the discrimination and the quantification of the abundance of each transcript in an organism. In addition to being useful for same species comparisons, the arrays may be used across species, although caution must be taken. For example, taking advantage of the close evolutionary relationship between *G. max, Vigna unguiculata* and *Phaseolus vulgaris*, the soybean arrays were also used with success to quantify the expression of *P. vulgaris* genes, a non-model legume plant (Das et al. 2010; Yang et al. 2010). In addition to these "oligonucleotide" arrays, the Affymetrix Company has developed and commercialized arrays to characterize the expression pattern of *G. max* (Valdes-Lopez et al. 2011) and *M. truncatula* genes (Mitra et al. 2004a). *M. truncatula* arrays have been used to profile alfalfa genes (Kang et al. 2011). An Affymetrix array was used to generate the *M. truncatula* Gene Expression Atlas (MtGEA) (Benedito et al. 2008). This atlas groups a large number of transcriptomic analyses (Benedito et al. 2008; Naoumkina et al. 2007, 2008; Imin et al. 2008; Holmes et al. 2008; Ruffel et al. 2008; Uppalapati et al. 2009; Gomez et al. 2009; Pang et al. 2008) and is hosted by a webserver, allowing the analysis and manipulation of the transcriptomic data sets (http://mtgea.noble.org/v2/; (He et al. 2009)). Overall, the quality of the arrays and their coverage is highly dependent on the quality of the cDNA libraries used to design the oligonucleotides on the array. For example, due to the complexity of the soybean genome; i.e., its recent duplication 13 Mya, as well as the use of incomplete cDNA libraries, the first generation of the soybean Affymetrix array did not provide an optimal set of oligonucleotides, leading to a limited coverage of the soybean transcriptome as well as a lack of specificity of some of the oligonucleotides for specific transcripts (Libault et al. 2010b).

High-throughput sequencing recently became a reference technology to characterize legume transcriptomes. Various platforms exist, allowing the generation of different number of reads; i.e., sequencing products of different lengths (Fig. 2.1). Hence, biologists are now frequently using high-throughput sequencing technology to characterize legume transcriptomes. For example, the use of this technology led to the establishment of the soybean transcriptome atlases (Libault et al. 2010a; Severin et al. 2010). Coupled with the development of bioinformatics tools, these transcriptomic resources can be easily accessed from two different bioinformatics platforms: the Soyseq platform hosted on Soybase (http://soybase.org/soyseq/) and Soykb (http://soykb.org/; (Joshi et al. 2012)). The drop of the this technology's cost as well as its higher sensitivity and accuracy in measuring transcript abundance now allow scientists to use it to characterize gene expression patterns in legumes. In model legumes, use of high-throughput sequencing enables transcript abundance measurements of genes missing from microarray platforms. Also, since this