Takuji Sasaki · Motoyuki Ashikari
Editors

Rice Genomics, Genetics and Breeding

Springer

Rice Genomics, Genetics and Breeding

Takuji Sasaki • Motoyuki Ashikari
Editors

Rice Genomics, Genetics and Breeding

Springer

Editors
Takuji Sasaki
NODAI Research Institute
Tokyo University of Agriculture
Setagaya, Tokyo, Japan

Motoyuki Ashikari
Bioscience and Biotechnology Center
Nagoya University
Nagoya, Aichi, Japan

ISBN 978-981-10-7460-8 ISBN 978-981-10-7461-5 (eBook)
https://doi.org/10.1007/978-981-10-7461-5

Library of Congress Control Number: 2018930351

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Rice is a staple food for about half of the world population and has sustained mankind since the dawn of agriculture. As the mid-twenty first century approaches, the world is faced with the monumental task of feeding 9 billion people. To address this challenge, sustainable increases in cereal grain production has to be met. Increasing rice production would require significant improvements in two key aspects: (1) crop management strategies including soil, water, and pest and disease management and (2) development of cultivars with better characteristics including higher yield and improved grain quality. Both these factors would be critical in realizing sustainable rice cultivation.

This book focuses on the body of genetics, genomics, and breeding researches that have been carried out to characterize the rice plant itself towards the purposeful development of better rice cultivars. Existing rice varieties that are being grown in localities worldwide are results of breeding efforts in response to demands by farmers for cultivars that are adapted to specific regional environments and by consumers with established quality preferences. For a long time, breeding for improved cultivars relied on visual observations and selection of recombinants with favorable phenotypes. With the discovery of Mendel's law, however, it has been shown that many important phenotypes or traits are too complex to be explained by mere visual observations.

As in the case of many other organisms, advancement in molecular genetics has revolutionized our understanding of rice as a plant. The map-based, genome sequence of a standard rice variety that was released in 2004 via an international collaboration has facilitated the identification and cloning of genes and quantitative trait locus (QTL) controlling various traits. This sequence information also opened the doors to the field of genomics which aims to understand genome-wide variation in rice in terms of gene expression, metabolite profile, and hormonal level during development. These information are in turn used to elucidate genetic networks that make up the rice plant.

Combined with the fast pace of technological advancements, the mega-volume of genomic information has allowed targeted genetic manipulations to induce variations in a given allele using genome-editing technologies such as TALENS or CRISPR/Cas9 system. To this day, genome editing has found applications not only in validating gene function but also in generating new alleles that can give more favorable phenotypes.

This book is composed of 28 chapters that describe the recent progress and future perspectives in *Rice Genomics, Genetics and Breeding*. Each chapter, written by established rice researchers who are experts in their field, is a comprehensive look at the genetics and genomics machineries underlying various traits in rice that can be used to address issues on food security. This is especially intended for rice scientists, breeders, post-docs, and graduate and undergraduate students as a standard reference that can be used to device strategies towards solving the 9 billion people challenge.

The editors would like to gratefully acknowledge the respective authors for their outstanding contributions towards the realization of this wonderful book.

Special thanks go to Ms. Yoko Niimi of Nagoya University for carefully checking all the citations in each chapter and to Ms. Sowmya Ramalingam of Springer for the final editing of this book.

Tokyo, Japan — Takuji Sasaki
Aichi, Japan — Motoyuki Ashikari

Contents

Chapter 1
Genome Sequences of *Oryza* Species

Masahiko Kumagai, Tsuyoshi Tanaka, Hajime Ohyanagi, Yue-Ie C. Hsing, and Takeshi Itoh

Abstract This chapter summarizes recent data obtained from genome sequencing, annotation projects, and studies on the genome diversity of *Oryza sativa* and related *Oryza* species. *O. sativa*, commonly known as Asian rice, is the first monocot species whose complete genome sequence was deciphered based on physical mapping by an international collaborative effort. This genome, along with its accurate and comprehensive annotation, has become an indispensable foundation for crop genomics and breeding. With the development of innovative sequencing technologies, genomic studies of *O. sativa* have dramatically increased; in particular, a large number of cultivars and wild accessions have been sequenced and compared with the reference rice genome. Since de novo genome sequencing has become cost-effective, the genome of African cultivated rice, *O. glaberrima*, has also been determined. Comparative genomic studies have highlighted the independent domestication processes of different rice species, but it also turned out that Asian and African rice share a common gene set that has experienced similar artificial selection. An international project aimed at constructing reference genomes and examining the genome diversity of wild *Oryza* species is currently underway, and the genomes of some species are publicly available. This project provides a platform for investigations such as the evolution, development, polyploidization, and improvement of crops. Studies on the genomic diversity of *Oryza* species, including wild species, should provide new insights to solve the problem of growing food demands in the face of rapid climatic changes.

M. Kumagai · T. Tanaka · T. Itoh (✉)
Advanced Analysis Center, National Agriculture and Food Research Organization, Tsukuba, Ibaraki, Japan
e-mail: taitoh@affrc.go.jp

H. Ohyanagi
King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Thuwal, Kingdom of Saudi Arabia

Y.-I. C. Hsing
Academica Sinica, Institute of Plant and Microbial Biology, Taipei, Taiwan

T. Sasaki, M. Ashikari (eds.), *Rice Genomics, Genetics and Breeding*,
https://doi.org/10.1007/978-981-10-7461-5_1

Keywords Genome sequencing · Reference genome · Annotation · Next-generation sequencing technology · Resequencing · Biodiversity · Comparative genomics

1.1 Introductory Overview of Rice Genome Sequencing

The central dogma of molecular biology suggests that all the biological processes of an organism should derive from the information encoded in its genomic DNA. As the genome is considered a blueprint of cellular life forms, knowledge of the entire genome sequence should be equivalent to understanding the whole biological mechanism. In the 1980s, biologists envisaged that genome-wide sequencing would expedite molecular biological studies to a much greater extent than the piecemeal analyses of a handful of genes (Dulbecco 1986). However, the sequencing technologies available at that time were not sufficient to determine billions of nucleotides in a short time. Therefore, to achieve rapid whole-genome sequencing of higher eukaryotes, several innovative sequencing technologies have been developed in the twenty-first century. In this chapter, we provide an overview of the history of genome sequencing in *Oryza* species over the last two decades, which was strongly affected by the advent of novel sequencing platforms.

Since the completion of the genome sequencing of *Haemophilus influenzae* in 1995 (Fleischmann et al. 1995), genome-wide sequencing has played a pivotal role in current biology, although the era of genomics emerged slightly later for plants than microbes and animals. Sequencing using the Sanger method was neither massive nor fast for large genomes, and the determination of higher eukaryote genomes therefore generally took a long time. Although the genome sequence of *Arabidopsis thaliana*, the best studied model plant, was published in 2000 (Arabidopsis Genome Initiative 2000), genome-wide sequence data for crops were not available until 2002, when the genomes of *japonica* and *indica* cultivars of rice (*O. sativa* L.) were deciphered using the whole-genome shotgun method (Goff et al. 2002, Yu et al. 2002). However, these two genome sequences were extensively fragmented, reflecting the limitations of the sequencing strategy, and a much more accurate genome sequence that would meet the demands for further genomic studies was therefore anticipated. Long before the publication of the genomes obtained using the shotgun method, other efforts toward rice genome sequencing were conceived and initiated in the early 1990s in Japan (Sasaki 1998); however, these techniques were not sufficient to sequence the complete genome at that time. In 1998, the International Rice Genome Sequencing Project (IRGSP) was organized by pioneering researchers from ten countries and regions, including Japan (Matsumoto et al. 2016). This international collaborative project employed precise genome sequencing based on a physical map of P1-derived artificial chromosome (PAC)/bacterial artificial chromosome (BAC) clones to generate a high-quality genome sequence. In fact, the resultant genome sequence published in

2004 (International Rice Genome Sequencing Project 2005), which was later improved in 2013 (Kawahara et al. 2013), is accurate and still serves as the essential foundation of cereal genomics (Matsumoto et al. 2016).

Since the public release of the IRGSP genome, next-generation sequencing (NGS) technologies have dramatically altered molecular biological studies. Genome sequencing is now rather easier and more cost-effective than gene-by-gene approaches. For example, the causative mutation of a phenotype of interest may be more rapidly observed by sequencing than through other conventional molecular genetic methods. Such high-throughput sequencing studies are particularly effective in rice because we can map newly sequenced reads to the high-quality reference genome. In addition, while current sequencing technologies are suitable for massive sequence production, these data generally contain a significant number of errors, several orders of magnitude larger than the errors generated using the Sanger method. NGS-based genome sequencing is currently widely used for resequencing, as observed in rice, for which we have an appropriate reference genome for comparison.

To construct a high-quality reference genome, yet another innovative technology is needed. The single-molecule real-time sequencing technology of Pacific Biosciences (PacBio) (Eid et al. 2009) is currently a promising method for de novo genome sequencing and is therefore being applied to crop species (Sakai et al. 2015; Du et al. 2017), but the assemblies presented thus far still contain some gaps. Another emerging platform is the MinION device, a single-molecule nanopore sequencer from Oxford Nanopore Technologies (Michael et al. 2017). MinION is highly cost-effective and presents a strong capability for sequencing in the field, leading to its widespread usage. In addition to these methods, the combination of Illumina with Hi-C, which can reutilize sequence data generated in the past, exhibits great potential (Dudchenko et al. 2017). Therefore, since low-cost/high-quality de novo genome sequencing is anticipated in the near future, multiple reference genomes of *japonica* cultivars, *indica* cultivars, and wild rice accessions will be available, and rice genomic research will be further accelerated (Fig. 1.1).

1.2 Rice Genome Sequencing Projects and the Release of the IRGSP Reference Genome

Two rice draft genomes, one for *japonica* and the other for *indica*, were made available prior to the completion of IRGSP. Beijing Genomics Institute (BGI) sequenced an *indica* cultivar, representing the major cultivar group in China and the southern part of Asia (Yu et al. 2002). They sequenced the genome of cultivar 93-11, the parental cultivar of super-hybrid rice Liang-You-Pei-Jiu (LYP9). Two other groups, Monsanto Co. and Syngenta Co., sequenced the *japonica* cultivar Nipponbare, which was the same cultivar sequenced by IRGSP (Barry 2001; Goff et al. 2002). These draft genome sequencing projects employed whole-genome shotgun sequencing,

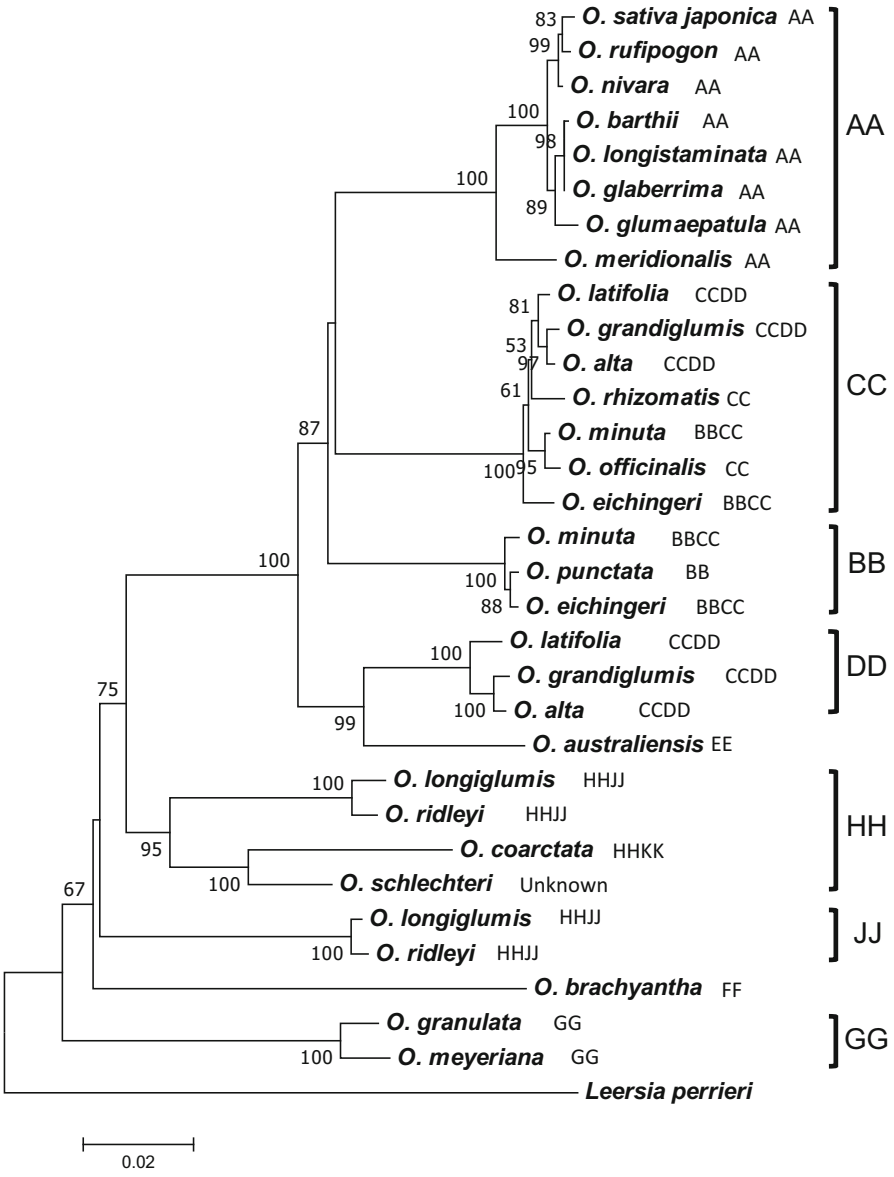

Fig. 1.1 Phylogeny of the genus *Oryza* based on the nucleotide sequences of *Adh1* (Ge et al. 1999). The tree was inferred using the neighbor-joining method, and the evolutionary distances were computed using Kimura's two-parameter method. The bootstrap test values (1000 replications) are shown next to the branches. The genome type of each accession is indicated with two or four letters. The estimated ancestral diploid genome of each allele was inferred from the tree topology and is indicated on the right side of the tree (Ge et al. 1999). This analysis was conducted using MEGA7 (Kumar et al. 2016)

in which Sanger sequencing reads were assembled without anchoring to chromosomes. Therefore, a draft genome could be rapidly constructed. However, the quality was not comparable to other reference genomes, such as those of *Arabidopsis*, yeast, *Drosophila*, and *C. elegans*. The assembled genome sequences were divided into 791 BAC contigs in *japonica* and 103,044 scaffolds in *indica*. While comparative genomics and gene analysis could be performed using these genome data, a high-quality rice reference genome sequence was desired by the rice research community as well as researchers focused on cereal crops and other plants.

IRGSP published the whole-genome sequence of a *japonica* cultivar in 2004 (IRGSP 2005). The genome was sequenced using map-based sequencing and clone-by-clone Sanger sequencing. The BACs, PACs, and fosmid clones in the physical map were sequentially sequenced, independently assembled, and used for reconstruction of the genome assembly. Genomic libraries from Nipponbare, a temperate *japonica* cultivar that is widely used for experiments, were employed to establish the physical map. Sequence gaps were cautiously resolved by selecting gap-bridge clones and PCR fragments and through the direct sequencing of BACs. Each clone was sequenced via shotgun sequencing with tenfold coverage. The quality of the obtained assembly was expected to exceed the 99.99% accuracy standard (less than one error in 10,000 bases). The finished genome provided direct evidence of a rice genome size of 389 Mb, which is three times larger than the *Arabidopsis thaliana* genome. The completed IRGSP genome sequence was 370 Mb in length, representing 95% coverage of the rice genome. However, comparative analysis against the two previously published draft genome sequences showed that the coverage of these draft genomes compared with the IRGSP genome was 69% and 78% in *indica* and *japonica*, respectively. Ab initio gene finding predicted a total of 37,544 non-transposable-element-related protein-coding sequences, and 2859 rice genes were not previously observed in the *Arabidopsis* genome. The completeness of the IRGSP genome enabled an analysis of centromeres in the untraversed genomic region. Deciphering these centromere sequences was a remarkable effort. The entire centromere sequences of chromosomes 4 and 8 were determined, showing sizes of 59 and 69 kb, respectively, based on clustered CentO repeats (Nagaki et al. 2004; Wu et al. 2004; Zhang et al. 2004). *Oryza sativa* became the first eukaryotic species with a complex genome structure whose complete centromere sequence was analyzed. The high-quality map-based genome sequence of Nipponbare remains the only monocot genome and serves as the role model for genome sequencing projects in other cereal crops with large genomes and complex chromosome contents (Matsumoto et al. 2016).

1.3 Genome Annotations and the Release of the Revised Genome Assembly, IRGSP-1.0

Genome annotation is absolutely essential for utilizing genome information in biological studies. Prior to completing the IRGSP genome sequencing project, The Institute for Genomic Research (TIGR), a member of the IRGSP, initiated a gene annotation project, currently known as the Rice Genome Annotation Project (RGAP), and has successively released the results to the research community (Yuan et al. 2003). This annotation database is now maintained by Michigan State University (MSU, http://rice.plantbiology.msu.edu/cgi-bin/gbrowse/rice/). IRGSP members launched the official genome annotation project, the Rice Annotation Project (RAP), an initiative of the National Institute of Agrobiological Sciences (NIAS) in Japan, and the data are released from RAP-DB (http://rapdb.dna.affrc.go.jp, Ohyanagi et al. 2006). RAP-DB is currently maintained by National Agriculture and Food Research Organization. RAP adopted evidence-based annotation; therefore, representative genes are associated with experimental results, such as full-length rice cDNA sequences, rice ESTs (expressed sequence tags), and mRNA/protein sequences from closely related species. In the case of ESTs, several ab initio gene prediction programs are combined to generate intact gene structures. In addition, RAP has continued improving the reference genome assembly and is currently manually curating gene structures and functional descriptions based on the literature. In 2013, the most recent genome assembly, Os-Nipponbare-Reference-IRGSP-1.0 (IRGSP-1.0), was published as a joint effort of the RAP and RGAP annotation projects (Kawahara et al. 2013). This unified genome represents the revised version of these projects, based on the use of NGS technology and the minimal tiling path of clones with optical mapping. Resequencing data from two Nipponbare individuals obtained from the Illumina Genome Analyzer II/IIx platform and Roche 454 GS FLX was used for correcting single nucleotide and short insertion/deletion errors (Kawahara et al. 2013). After mapping the NGS reads to the reference genome, a total of 4886 single-nucleotide sequencing errors and five insertions/deletions were detected in the whole Nipponbare genome. This result validated the quality of the original assembly, showing an average error rate of 0.15 per 10,000 nucleotides (Kawahara et al. 2013). The size of revised Nipponbare genome was estimated to be 384.2–386.5 Mbp based on revised assembly and gap size estimation by fluorescence in situ hybridization and estimated length of the rDNA regions. After the release of IRGSP-1.0, the two annotations, RAP and RGAP, can be compared directly on the new assembly.

1.4 De Novo Assembly of the *indica* Genome via NGS

Because the map-based genome sequence of IRGSP was obtained using the *japonica* cultivar and because the draft genome of *indica* was far from completion, determining the genome sequence of *indica* remained one of the major goals. According to studies on genetic diversity and molecular phylogeny, significant genetic diversity of *O. sativa* species has been observed. For example, *japonica* and *indica* originated from different ancestral populations of the wild rice species *O. rufipogon* (Cheng et al. 2003; Londo et al. 2006). Therefore, there are many structural differences between *japonica* and *indica*, including structural variations at the chromosomal segment level. Moreover, high within-population genetic diversity has been observed in *indica*, and at least two varietal groups (*indica* and *aus*) were defined based on classical observation of physical and physiological traits, which were subsequently confirmed based on modern genetics and genomic data (Glaszmann 1987; Garris et al. 2005; McNally et al. 2009). Hence, researchers demanded genome sequences for *indica* varietal cultivars. Thus far, several de novo genome assembly studies in *indica* varieties, including the *aus* group, have been performed using NGS technologies (Sakai et al. 2014; Schatz et al. 2014; Du et al. 2017). Although the strategy for these de novo assembly studies primarily involves whole-genome shotgun sequencing, similar to past draft genome sequencing projects, computational algorithms for assembling the genome sequence enabling researchers to construct a higher-quality genome have been developed and applied under some NGS methods, such as those involving mate pair libraries and ultra-long read sequencing (e.g., the PacBio sequencing platform, whose read length N50 (a statistic defined as the shortest sequence length for the top 50% of sequences) can be 10 kb and more). In particular, long sequence reads generated using a third-generation sequencer, such as the PacBio platform, are powerful for de novo genome assembly in species with large complex genomes (Sakai et al. 2015). This technique has frequently been used in recent de novo genome assembly studies and employed for *indica* genome assembly (Du et al. 2017). Regarding the comparative quality of the obtained genomes, the first-draft genomes generated using Sanger sequencing exhibited N50 sizes of 6.69 kb for contigs (which is the minimum unit of genome assembly) and 11.76 kb for scaffolds (Yu et al. 2002), while the most recently reported *indica* de novo assembly using PacBio reads showed N50 sizes ranging from several hundreds of kb to 1.1 Mb (which varies based on the assembly software used) for contigs and 2.48 Mb scaffolds (Du et al. 2017).

1.5 *O. sativa* Genome Resequencing Project

Reflecting the dramatic reduction in sequencing costs after the emergence of NGS technologies, many genome sequencing studies of cultivated rice have been performed in this decade. With the improvement of sequencing platforms and chemistry, accompanied by reduced costs, we have obtained an increasing amount of individual whole-genome data. The scale of genome sequencing studies has gradually expanded. Reference mapping-based studies of *O. sativa* for detecting single-nucleotide variants (SNVs) of each variety were initiated on a small scale, involving one to a few samples, which then increased to dozens and hundreds of samples, finally reaching thousands of genomes (Yamamoto et al. 2010; Huang et al. 2010; Xu et al. 2012, The 3000 rice genomes project 2014, Yano et al. 2016). These investigations were primarily aimed at understanding the relationship between genotype and phenotype based on genome-wide association studies (GWAS) and provided information such as variety-specific genetic polymorphisms, within- and between-population genetic diversity, and insights into the history of Asian rice domestication. The population genomic analyses of several samples enabled the detection of footprints of artificial selection in past domestication and breeding efforts. A large-scale genome resequencing study involving more than 1000 accessions of *O. sativa* in China and its wild ancestor *O. rufipogon* revealed that many causal genes for domestication-related phenotypes, such as grain shattering, grain size, plant architecture, and grain color, were located in candidate regions for selective sweeps (Huang et al. 2012). This successful work indicated the usefulness of determining whole-genome polymorphisms for large-scale sample collection and the detection of selective sweeps using population genetic statistics to identify candidate gene alleles related to beneficial traits. The most recent large-scale sequencing project for *O. sativa* is the 3000 Rice Genomes Project. This ongoing project has resequenced a core collection of 3000 rice germplasm accessions, including both *japonica* and *indica* cultivars, selected from resources of the International Rice Research Institute (IRRI) and the Chinese Academy of Agricultural Sciences (CAAS), comprising accessions from 89 countries distributed in Southeast Asia (33.9%), South Asia (25.6%), and China (17.6%). Each genome of 3000 accessions contained sequences with 14X genome coverage on average, indicating that this amount of data provided an adequate depth for the detection of reliable SNVs, with 17TB of data being obtained using the Illumina platform in total. Based on reference mapping to IRGSP-1.0, approximately 18.9 M single-nucleotide polymorphisms (SNPs) were identified (The 3000 rice genomes project 2014). These data will serve as a fundamental resource for the discovery of novel alleles for important phenotypes that are useful for rice improvement and adaptation to changing environments.

1.6 Domestication History of *O. sativa* and Contribution of Genomic Studies

The domestication history of crop species has attracted much attention from a variety of research fields. Understanding the domestication process is a subject of much interest and will contribute to the management of next-generation agriculture in the coming era, with ongoing, rapid environmental change. Recent efforts in genomic studies have shed light on the history of Asian rice domestication. The mystery of the origin and domestication process of Asian cultivated rice has been argued for decades. The main issues were the origins of *japonica* and *indica* and whether these varietal groups were independently domesticated. Previous genetic studies employing *O. sativa* and a diverse panel of *O. rufipogon* indicated that *japonica* and *indica* showed close affinity to different *O. rufipogon* populations and suggested multiple origins of *O. sativa*, with *japonica* originating in China and *indica* originating in South/Southeast Asia (Cheng et al. 2003; Londo et al. 2006; Rakshit et al. 2007), and their divergence time predated the onset of domestication (Ma and Bennetzen 2004; Vitte et al. 2004; Zhu and Ge 2005). This hypothesis presented a good fit for archaeological data demonstrating the existence of old rice culture ruins in the Yangtze River basin in China and the Ganges River basin in India (Fuller 2006). However, the story is more complicated, as some population genetic studies employing a larger number of loci have suggested nonindependent domestication of *japonica* and *indica* (Gao and Innan 2008; Molina et al. 2011). Recent studies based on whole-genome data clarified the detailed process of Asian rice domestication, showing that *japonica* and *indica* exhibit divergent genomic backgrounds, coming from different wild rice populations, and that gene introgression of domestication-related genes from *japonica* to *indica* has occurred (Huang et al. 2012; Yang et al. 2012). Furthermore, two *indica* rice varieties (*indica* and *aus*) also have different origins (Civáň et al. 2015; Choi et al. 2017). Phylogeographical analysis of worldwide wild rice panels comprising more than 400 accessions, together with cultivated rice, demonstrated that wild rice collected from the middle of the Pearl River region in southern China showed the closest genetic affinity to cultivated rice in domestication-related gene regions (Huang et al. 2012).

1.7 Sequencing of *O. glaberrima*, African Cultivated Rice

In addition to Asian rice, African cultivated rice, *O. glaberrima*, was independently domesticated from the wild rice species *Oryza barthii* in West Africa approximately 3000 years ago (Linares 2002). *O. glaberrima* is well adapted to cultivation conditions in Africa but presents a lower yield potential than that of *O. sativa* (Linares 2002). An early *O. glaberrima* genome sequencing effort resulted in a partial genome-wide sequence of this species (Sakai et al. 2011). To more precisely

determine the genome, a BAC library was prepared to construct a minimum tilling path (MTP) (Wang et al. 2014). The MTP hybrid BAC pools were then subjected to sequencing using Roche/454 technology, with a sequence coverage of 31X. The reference sequences for 12 chromosomes, with the total assembly size of 316 Mb, indicated that *O. glaberrima* was domesticated in a single region along the Niger River (Wang et al. 2014). Comparative genomic analysis highlighted the independent selection of a common set of genes during two geographically distinct domestication processes (Wang et al. 2014). Regarding non-shattering, for example, early domestication-related genes such as *qsh1* (Konishi et al. 2006) and *sh4* (Li et al. 2006) exhibited totally different haplotypes in *O. sativa* and *O. glaberrima* (Wang et al. 2014). Additionally, the *heading date 1* (*Hd1*) gene, related to photoperiod sensitivity, was deleted from *O. glaberrima* (Sanyal et al. 2010; Wang et al. 2014), while single-nucleotide polymorphisms (SNPs) or small insertions/deletions (indels) were responsible for the loss of function of *Hd1* in *O. sativa* (Yano et al. 2000; Takahashi et al. 2009). The accession sequences as well as the assembly size and percentages of repetitive sequences in these African cultivated and wild rice species are listed in Table 1.1.

1.8 Wild *Oryza* Species

The genetic resource center of IRRI maintains a collection of more than 4000 accessions of wild *Oryza* species (Sanchez et al. 2013) comprising 22 relatives found in a wide range of habitats, including areas of Asia, Australia, Africa, and South and Central America, representing 15 million years of evolutionary diversification. This collection includes 14 diploids and 8 polyploids, with genome sizes ranging from approximately 300 Mb to 1.2 Gb. The chromosome number of these accessions is 24 or 48, representing the AA, BB, CC, BBCC, CCDD, EE, FF, GG, HHJJ, and HHKK genome types (Table 1.2, Sanchez et al. 2013). These wild species exhibit significant diversity in terms of morphological traits, plant height, tillering number and stature, flowering behavior, growth habits and panicle, and leaf and seed characteristics. In addition, these species adapt to different habitats and are resistant to a range of biotic and abiotic stresses.

1.9 The International Collaborative for Wild *Oryza* Species Genome Sequencing

An ambitious comparative genomic program entitled the "*Oryza* Map Alignment Project" (OMAP) was established in the early 2000s (Wing et al. 2005). The long-term objective of this work was to generate a genome-level closed experimental system for the genus *Oryza* as a research platform to study evolution, development,

Table 1.1 Current rice reference genomes

Species	Genome type	Accessions	Origin	Assembly size (Mb)	Repeat (%)	References
O. sativa (*japonica*)	AA	Nipponbare	Asia	389	38.9	IRGSP (2005) and Sakai and Itoh (2010)
O. sativa (*indica*)	AA	93-11	Asia	375	42.8	Yu et al. (2002)
O. sativa (*indica*)	AA	Zhenshan 97	Asia	384	41.3	Zhang et al. (2016)
O. sativa (*indica*)	AA	Minghui 63	Asia	386	41.6	Zhang et al. (2016)
O. sativa (*indica*)	AA	Shuhui498	Asia	390	42.1	Du et al. (2017)
O. nivara	AA	W0603	Asia	338	36.4	I-OMAP (2018)
O. rufipogon	AA	W1943	Asia	338	42.1	I-OMAP (2018)
O. glaberrima	AA	IRGC: 96717,	Africa	285	39.3	Wang et al. (2014)
O. barthii	AA	IRGC:105608	Africa	308	38.3	I-OMAP (2018)
O. glumaepatula	AA	GEN1233	America	373	31.4	I-OMAP (2018)
O. meridionalis	AA	W2112	Australia	336	27.2	I-OMAP (2018)
O. punctata	BB	IRGC:105690	Africa	394	49.6	I-OMAP (2018)
O. brachyantha	FF	IRGC:101232	Africa	261	28.7	Chen et al. (2013)
Leersia perrieri	–	IRGC:105164	Africa	267	26.7	I-OMAP (2018)

genome organization, polyploidy, domestication, gene regulatory networks, and crop improvement (Wing et al. 2005). In this project, these researchers constructed and aligned BAC/STS (sequence tagged site)-based physical maps of 11 wild rice species, including 6 diploid genomes (AA, BB, CC, EE, FF, and GG) and 4 tetraploid genomes (BBCC, CCDD, HHJJ, and HHKK) (Wing et al. 2005, Ammiraju et al. 2006, 2010), followed by BAC-end sequencing (Kim et al. 2008). With these resources, the research community revealed gene conversion (Jacquemin et al. 2013), duplication, de novo origination, movement, loss (Zhao et al. 2015), and transposable element (TE) dynamics (Jacquemin et al. 2014). In 2007, this project was transformed into the International *Oryza* Map Alignment Project (I-OMAP) with the aim of generating RefSeq and transcriptome datasets for eight AA genome species and one BB genome species (Jacquemin et al. 2013). These datasets included two Asian AA genome species, *O. rufipogon* and *O. nivara*; one

Table 1.2 Species in the genus *Oryza*

Species	Chromosome number	Genome	Geographical distribution
Oryza sativa complex			
O. sativa	24	AA	All over the world
O. rufipogon	24	AA	Asia, Oceania
O. nivara (also known as annual ecotype of *O. rufipogon*)	24	AA	Asia, Oceania
O. glaberrima	24	AA	West Africa
O. barthii	24	AA	Africa
O. longistaminata	24	AA	Africa
O. meridionalis	24	AA	Australia
O. glumaepatula	24	AA	Central and South America
O. officinalis complex			
O. officinalis	24	CC	Asia
O. minuta	48	BBCC	Philippines
O. rhizomatis	24	CC	Sri Lanka
O. eichingeri	24	CC	Africa, Sri Lanka
O. punctata	24, 48	BB, BBCC	Africa
O. latifolia	48	CCDD	Central and South America
O. alta	48	CCDD	Central and South America
O. grandiglumis	48	CCDD	South America
O. australiensis	24	EE	Australia
O. ridleyi complex			
O. ridleyi	48	HHJJ	Asia, New Guinea
O. longiglumis	48	HHJJ	New Guinea
O. granulata complex			
O. granulata	24	GG	Asia
O. meyeriana	24	GG	Asia
Others			
O. brachyantha	24	FF	Africa
O. schlechteri	48	Unknown	New Guinea
O. coarctata (also known as *Porteresia coarctata*)	48	HHKK	Coastal region of India, Pakistan, and Bangladesh

Australian AA genome species, *O. meridionalis*; one American AA genome species, *O. glumaepatula*; and one African AA genome species, *O. barthii*, the species from which *O. glaberrima* was domesticated. In addition, one African BB genome species, *O. punctata*, and the diploid *Leersia perrieri* as an outgroup species were also added (I-OMAP Consortium 2018). Shotgun sequences with various insert size libraries at a minimum of depth of 100X coverage were generated using Illumina technology for assembly. For the two African species *O. barthii* and *O. punctata*, additional sequence coverage of 10–20X was obtained using Roche/454 technology. Final super-scaffolds were manually constructed using paired BAC-end

sequences and alignment to the Nipponbare sequences as guide information. The total lengths of the resulting genome sequences, with 12 chromosomes in each species, ranged from 267 to 394 Mb (I-OMAP Consortium 2018). The accessions are listed in Table 1.1 with the assembly sizes. Using all 11 species, including the two major Asian cultivars, the African cultivated species, and all wild species mentioned above, several conclusions were reached. Phylogenic analysis indicated that the "crown" age of the AA clade is approximately 2.5 million years, with a rapid diversification rate of ~0.50 net new species/MYR (I-OMAP Consortium 2018). Detailed sequence analysis also showed that extensive introgression has occurred in *Oryza* species, particularly between South American *O. glumaepatula* and the African AA species (I-OMAP Consortium 2018). The turnover rate of LTR retrotransposons within the AA genome lineage was one to two orders of magnitude faster than those estimated for flies and mammals, respectively (I-OMAP Consortium 2018). In addition, thousands of candidate disease resistance genes were discovered in heterogeneous gene pairs organized in a head-to-head fashion, supporting the integrated decoy model for disease resistance (I-OMAP Consortium 2018). The assembled genome sequences and the annotated open reading frame (ORF) amino acid sequences have all been available to the community at Ensembl Plants (http://plants.ensembl.org/index.html) since 2015 and have been extensively employed. For example, using the sequence information for two Asian wild rice species (*O. rufipogon* and *O. nivara*) from I-OMAP, Choi et al. conducted comparative analyses with two *indica* rice varieties (IR64 and 93-11), two *aus* rice varieties (Kasalath and DJ123), and the *japonica* variety Nipponbare (Choi et al. 2017) and concluded that domestication occurred only once, with multiple origins; thus, each domesticated rice subpopulation, including *japonica*, *indica*, and *aus*, arose separately from *O. rufipogon* and/or *O. nivara* progenitors. Furthermore, when Baldrich et al. (2016) analyzed polycistronic miRNAs in the cultivated and wild rice using this resource, they discovered new rice polycistronic miRNAs and suggested that most polycistronic and candidate polycistronic miRNAs showed a pattern of conservation in the genomes of rice species with an AA genome (Baldrich et al. 2016).

1.10 Distantly Related Wild *Oryza* Species

Oryza species that are only distantly related to cultivated rice species in terms of evolution exhibit more diverged morphologies and a wider variety of resistant phenotypes related to biotic and abiotic stresses than closely related wild species (Jena 2010; Nonomura et al. 2010). In addition, previous studies have shown genomic diversity with respect to genome size, genomic contents, and genomic polyploidy (Vaughan et al. 2003; Buell et al. 2005; Wing et al. 2005). These studies highlighted distantly related *Oryza* species as invaluable genetic and genomic resources for exploring and exploiting the hidden molecular mechanisms of agronomically important traits in breeding science and have made a good evolutionary

case for investigating the complex polyploidy problem in basic biology. In this NGS era, efforts to meet the challenge of deciphering distantly related wild *Oryza* genome sequences at a reasonable quality have been initiated. The achievements of several completed and ongoing projects thus far will be reviewed in the following sections.

1.11 Sequencing of *O. punctata* (BB) Genome

O. punctata, which belongs to the *Oryza officinalis* complex (Table 1.2), is a wild rice species distributed in Africa. The *Oryza officinalis* complex comprises diploid and tetraploid species with BB, CC, DD, and EE genomes. The DD genome has not yet been observed in a diploid state and has only been found in CCDD tetraploids. The species in the *O. officinalis* complex are geographically widely distributed in Asia (and the northern part of Australia), Africa, and South America. *O. punctata* accessions are categorized into two subtypes according to their genome types: diploids (BB) and tetraploids (BBCC), which grow in separate habitats. Their traits and genome sequences are of immediate interest, particularly to breeding scientists, since these species demonstrate resistance to multiple biotic and abiotic stresses (Jena 2010; Sanchez et al. 2013). The diploid BB *punctata* has $2n = 24$ chromosomes, and the nuclear genome size has been estimated as nearly equal to that of *O. sativa* (~400 Mb, Table 1.1) in flow cytometry experiments. An effort to decipher a diploid *punctata* (BB) accession genome has been made by Arizona Genomics Institute (AGI) under the activities of I-OMAP. A BAC-pooled WGS method using Illumina technology was undertaken, and the scaffolds were finally aligned to chromosomal coordinates according to the BAC-based physical map. As expected, the total assembly size was ~400 Mb, with a similar proportion of repetitive contents in the total genome to cultivated rice. This *O. punctata* genome sequencing has not yet been reported but in preparation, although the pre-publication assembly and baseline genome annotations have been released under the guidelines of the Fort Lauderdale Agreement (https://www.genome.gov/10506537/) in Ensembl Plants (http://plants.ensembl.org). From an evolutionary point of view, the availability of both BB and CC diploid genomes (see below) will be an epochal advance in the characterization of ancient genome-scale evolutionary events.

1.12 Sequencing of *O. officinalis* (CC) Genome

O. officinalis (Table 1.2) is a wild rice species that grows in various environments in South and Southeast Asia. Due to the wider habitats and putative core components of BBCC and CCDD tetraploid species, significant roles of *O. officinalis* in the

evolutionary history of rice have been speculated, whereas the relationship with other tetraploid species is not yet known. From a breeding perspective, similar to *O. punctata*, *O. officinalis* shows resistance traits related to multiple plant diseases (Jena 2010; Sanchez et al. 2013) and has been utilized as a genetic resource for introgression into cultivated rice species (Huang et al. 2001; Sanchez et al. 2013). However, there are limited genomic resources publicly available for this species. Consistent with other diploid species in the *Oryza* genus, this species comprises 2n = 24 chromosomes, but the nuclear genome size has been estimated as more than 600 Mb, which is one and a half times as large as those of AA species based on flow cytometry experiments (Uozu et al. 1997; Miyabayashi et al. 2007).

To serve as a quality genomic reference resource in breeding science to explore the biology of ploidy, the National Institute of Genetics of Japan has been promoting a nuclear genome sequencing project of an accession of *O. officinalis* under the collaborative activities of I-OMAP. A hybrid WGS approach using both Illumina and PacBio sequencing technologies has been adopted, and the resultant scaffolds have been aligned based on chromosomal coordinates according to the BAC-based physical map provided by AGI/I-OMAP. Baseline genome annotation and comparative genomic analysis have been conducted. A manuscript reporting these results is currently under preparation. The genomic reference of *O. officinalis* will be critical for providing a foundation for exploring the ploidy biology of BBCC, CC, CCDD, and unknown DD diploid species in the *O. officinalis* complex.

1.13 Sequencing of *O. brachyantha*, the Smallest *Oryza* Genome

O. brachyantha is a wild rice species distributed in tropical Africa and is located in the basal lineage in the phylogeny of the genus *Oryza* (Zou et al. 2008). The genome type of this species is FF, and the genome size is the smallest in the genus *Oryza*. This species exhibits many biotic- or abiotic-resistant traits, such as broad-spectrum resistance to rice bacterial leaf blight (Ram et al. 2010). Through a whole-genome shotgun sequencing approach using Illumina technology and BAC-end sequences generated via Sanger technology, sequence coverage of 104X was generated for *O. brachyantha* (Chen et al. 2013). The assembled sequence blocks were subsequently anchored to chromosomes using a cytogenetic approach, generating 12 chromosome sequences with a total of 261 Mb in size (Chen et al. 2013). The low activity of long terminal repeat (LTR) retrotransposons and high frequency of internal deletions of ancient long terminal repeat elements in the *O. brachyantha* genome led to its compact genome. Approximately 32,000 protein-coding genes were annotated in its genome, and only 70% of these genes are located in collinear positions compared with the *O. sativa* genome. These

nonlinear genes were enriched at pericentromeric or heterochromatic knobs compared with euchromatic regions in the *O. sativa* genome, resulting in a reduced level of gene collinearity in recombination-inert regions (Chen et al. 2013).

1.14 Perspectives

The rice reference genome has been contributing to the studies of a wide range of areas including plant physiology, molecular genetics, and breeding. It is expected that rice genomics will further extend to research on rice genetic diversity and genotype-phenotype interaction, such as quantitative trait loci analysis and GWAS for traits of agronomical importance. Since there should be a large number of unrecognized useful genes in *O. sativa* landraces as well as wild *Oryza* species, which were possibly lost in the modern cultivars including Nipponbare, comparative genomics approaches will discover such genes that can help create novel elite cultivars with beneficial traits.

It is interesting that known domestication-related traits of cereal crops were generally endowed through loss-of-function-type mutations or gene losses. Therefore, to detect large-scale indels, inversions, translocations, and present/absent variations of chromosomal segments, reference class complete genomes of a lot of accessions including wild species are envisaged. For this purpose, recently emerging ultra-long read sequencing technologies will be a key factor for future rice genomics. Moreover, development of computer algorithms to cope with such a gigantic dataset is an urgent issue. As novel experimental methods such as genome editing is quickly spreading, the next-generation breeding will be based on the combination of genome information analysis and such experimental technologies.

Acknowledgments This work was supported by a grant from the Ministry of Agriculture, Forestry, and Fisheries of Japan (Genomics-based Technology for Agricultural Improvement, IVG2001).

References

Ammiraju JSS, Luo M, Goicoechea JL et al (2006) The *Oryza* bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus *Oryza*. Genome Res 16:140–147. https://doi.org/10.1101/gr.3766306

Ammiraju JSS, Song X, Luo M et al (2010) The *Oryza* BAC resource: a genus-wide and genome scale tool for exploring rice genome evolution and leveraging useful genetic diversity from wild relatives. Breed Sci 60:536–543

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815. https://doi.org/10.1038/35048692

Baldrich P, Hsing Y-IC, San Segundo B (2016) Genome-wide analysis of polycistronic microRNAs in cultivated and wild rice. Genome Biol Evol 8:1104–1114. https://doi.org/10.1093/gbe/evw062

Barry GF (2001) The use of the Monsanto draft rice genome sequence in research. Plant Physiol 125:1164–1165

Buell CR, Yuan Q, Ouyang S et al (2005) Sequence, annotation, and analysis of synteny between rice chromosome 3 and diverged grass species. Genome Res 15:1284–1291. https://doi.org/10.1101/gr.3869505

Chen J, Huang Q, Gao D et al (2013) Whole-genome sequencing of *Oryza brachyantha* reveals mechanisms underlying *Oryza* genome evolution. Nat Commun 4:1595. https://doi.org/10.1038/ncomms2596

Cheng C, Motohashi R, Tsuchimoto S et al (2003) Polyphyletic origin of cultivated rice: based on the interspersion pattern of SINEs. Mol Biol Evol 20:67–75. https://doi.org/10.1093/molbev/msg004

Choi JY, Platts AE, Fuller DQ et al (2017) The rice paradox: multiple origins but single domestication in Asian rice. Mol Biol Evol 34:969–979. https://doi.org/10.1093/molbev/msx049

Civáň P, Craig H, Cox CJ, Brown TA (2015) Three geographically separate domestications of Asian rice. Nat Plants 1:15164. https://doi.org/10.1038/nplants.2015.164

Du H, Yu Y, Ma Y et al (2017) Sequencing and de novo assembly of a near complete *indica* rice genome. Nat Commun 8:15324. https://doi.org/10.1038/ncomms15324

Dudchenko O, Batra SS, Omer AD et al (2017) *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science 356:92–95. https://doi.org/10.1126/science.aal3327

Dulbecco R (1986) A turning point in cancer research: sequencing the human genome. Science 231:1055–1056

Eid J, Fehr A, Gray J et al (2009) Real-time DNA sequencing from single polymerase molecules. Science 323:133–138. https://doi.org/10.1126/science.1162986

Fleischmann RD, Adams MD, White O et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Fuller DQ (2006) Agricultural origins and frontiers in South Asia: a working synthesis. J World Prehistory 20:1–86. https://doi.org/10.1007/s10963-006-9006-8

Gao L-Z, Innan H (2008) Nonindependent domestication of the two rice subspecies, *Oryza sativa* ssp. *indica* and ssp. *japonica*, demonstrated by multilocus microsatellites. Genetics 179:965–976. https://doi.org/10.1534/genetics.106.068072

Garris AJ, Tai TH, Coburn J et al (2005) Genetic structure and diversity in *Oryza sativa* L. Genetics 169:1631–1638. https://doi.org/10.1534/genetics.104.035642

Ge S, Sang T, Lu B, Horton M (1999) Phylogeny of rice genomes with emphasis on origins of allotetraploid species. Proc Natl Acad Sci U S A 96:14400–14405

Glaszmann JC (1987) Isozymes and classification of Asian rice varieties. Theor Appl Genet 74:21–30. https://doi.org/10.1007/BF00290078

Goff SA, Ricke D, Lan T-H et al (2002) A draft sequence of the rice genome (*Oryza sativa L.* ssp. *japonica*). Science 296:92–100. https://doi.org/10.1126/science.1068275

Huang Z, He G, Shu L et al (2001) Identification and mapping of two brown planthopper resistance genes in rice. TAG Theor Appl Genet 102:929–934. https://doi.org/10.1007/s001220000455

Huang X, Wei X, Sang T et al (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42:961–967. https://doi.org/10.1038/ng.695

Huang X, Kurata N, Wei X et al (2012) A map of rice genome variation reveals the origin of cultivated rice. Nature 490:497–501. https://doi.org/10.1038/nature11532

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800. https://doi.org/10.1038/nature03895

Jacquemin J, Bhatia D, Singh K, Wing RA (2013) The international *Oryza* map alignment project: development of a genus-wide comparative genomics platform to help solve the 9 billion-people question. Curr Opin Plant Biol 16:147–156. https://doi.org/10.1016/j.pbi.2013.02.014

Jacquemin J, Ammiraju JSS, Haberer G et al (2014) Fifteen million years of evolution in the *Oryza* genus shows extensive gene family expansion. Mol Plant 7:642–656. https://doi.org/10.1093/mp/sst149

Jena KK (2010) The species of the genus *Oryza* and transfer of useful genes from wild species into cultivated rice, *O. sativa*. Breed Sci 60:518–523. https://doi.org/10.1270/jsbbs.60.518

Kawahara Y, de la Bastide M, Hamilton JP et al (2013) Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice (N Y) 6:4. https://doi.org/10.1186/1939-8433-6-4

Kim H, Hurwitz B, Yu Y et al (2008) Construction, alignment and analysis of twelve framework physical maps that represent the ten genome types of the genus *Oryza*. Genome Biol 9:R45. https://doi.org/10.1186/gb-2008-9-2-r45

Konishi S, Izawa T, Lin SY et al (2006) An SNP caused loss of seed shattering during rice domestication. Science 312:1392–1396. https://doi.org/10.1126/science.1126410

Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870–1874. https://doi.org/10.1093/molbev/msw054

Li C, Zhou A, Sang T (2006) Rice domestication by reducing shattering. Science 311:1936–1939. https://doi.org/10.1126/science.1123604

Linares OF (2002) African rice (*Oryza glaberrima*): history and future potential. Proc Natl Acad Sci U S A 99:16360–16365. https://doi.org/10.1073/pnas.252604599

Londo JP, Chiang Y-C, Hung K-H et al (2006) Phylogeography of Asian wild rice, *Oryza rufipogon,* reveals multiple independent domestications of cultivated rice, *Oryza sativa*. Proc Natl Acad Sci U S A 103:9578–9583. https://doi.org/10.1073/pnas.0603152103

Ma J, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci U S A 101:12404–12410. https://doi.org/10.1073/pnas.0403715101

Matsumoto T, Wu J, Itoh T et al (2016) The Nipponbare genome and the next-generation of rice genomics research in Japan. Rice (N Y) 9:33. https://doi.org/10.1186/s12284-016-0107-4

McNally KL, Childs KL, Bohnert R et al (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. Proc Natl Acad Sci U S A 106:12273–12278. https://doi.org/10.1073/pnas.0900992106

Michael TP, Jupe F, Bemm F, et al (2017) High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. bioRxiv:1–18. https://doi.org/10.1101/149997

Miyabayashi T, Nonomura K-I, Morishima H, Kurata N (2007) Genome size of twenty wild species of *Oryza* determined by flow cytometric and chromosome analyses. Breed Sci 57:73–78. https://doi.org/10.1270/jsbbs.57.73

Molina J, Sikora M, Garud N et al (2011) Molecular evidence for a single evolutionary origin of domesticated rice. Proc Natl Acad Sci U S A 108:8351–8356. https://doi.org/10.1073/pnas.1104686108

Nagaki K, Cheng Z, Ouyang S et al (2004) Sequencing of a rice centromere uncovers active genes. Nat Genet 36:138–145. https://doi.org/10.1038/ng1289

Nonomura K-I, Morishima H, Miyabayashi T et al (2010) The wild *Oryza* collection in National BioResource Project (NBRP) of Japan: history, biodiversity and utility. Breed Sci 60:502–508. https://doi.org/10.1270/jsbbs.60.502

Ohyanagi H, Tanaka T, Sakai H et al (2006) The Rice Annotation Project Database (RAP-DB): hub for *Oryza* sativa ssp. japonica genome information. Nucleic Acids Res 34:D741–D744. https://doi.org/10.1093/nar/gkj094

Rakshit S, Rakshit A, Matsumura H et al (2007) Large-scale DNA polymorphism study of *Oryza sativa* and *O. rufipogon* reveals the origin and divergence of Asian rice. Theor Appl Genet 114:731–743. https://doi.org/10.1007/s00122-006-0473-1

Ram T, Laha GS, Gautam SK et al (2010) Identification of new gene introgressed from *Oryza brachyantha* with broad-spectrum resistance to bacterial blight of rice in India. Rice Genet Newsl 25:57–58

Sakai H, Itoh T (2010) Massive gene losses in Asian cultivated rice unveiled by comparative genome analysis. BMC Genomics 11:121. doi: 10.1186/1471-2164-11-121

Sakai H, Ikawa H, Tanaka T et al (2011) Distinct evolutionary patterns of *Oryza glaberrima* deciphered by genome sequencing and comparative analysis. Plant J 66:796–805. https://doi.org/10.1111/j.1365-313X.2011.04539.x

Sakai H, Kanamori H, Arai-Kichise Y et al (2014) Construction of pseudomolecule sequences of the *aus* rice cultivar Kasalath for comparative genomics of Asian cultivated rice. DNA Res 21:397–405. https://doi.org/10.1093/dnares/dsu006

Sakai H, Naito K, Ogiso-Tanaka E et al (2015) The power of single molecule real-time sequencing technology in the de novo assembly of a eukaryotic genome. Sci Rep 5:16780. https://doi.org/10.1038/srep16780

Sanchez PL, Wing RA, Brar DS (2013) The wild relative of rice: genomes and genomics. In: Zhang Q, Wing RA (eds) Genetics and genomics of rice. Springer, New York, pp 9–25

Sanyal A, Ammiraju JS, Lu F et al (2010) Orthologous comparisons of the *Hd1* region across genera reveal *Hd1* gene lability within diploid *Oryza* species and disruptions to microsynteny in sorghum. Mol Biol Evol 27:2487–2506. https://doi.org/10.1093/molbev/msq133

Sasaki T (1998) The rice genome project in Japan. Proc Natl Acad Sci U S A 95:2027–2028

Schatz MC, Maron LG, Stein JC et al (2014) Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of *aus* and *indica*. Genome Biol 15:506. https://doi.org/10.1186/PREACCEPT-2784872521277375

Takahashi Y, Teshima KM, Yokoi S et al (2009) Variations in *Hd1* proteins, *Hd3a* promoters, and *Ehd1* expression levels contribute to diversity of flowering time in cultivated rice. Proc Natl Acad Sci U S A 106:4555–4560. https://doi.org/10.1073/pnas.0812092106

The 3000 rice genomes project (2014) The 3,000 rice genomes project. Gigascience 3:7. https://doi.org/10.1186/2047-217X-3-7

The International *Oryza* Map Alignment Consortium (2018) Sequence of 11 rice-related species unveils conservation, turnover and innovation across the genus *Oryza*. Nature Genetics. https://doi.org/10.1038/s41588-018-0040-0

Uozu S, Ikehashi H, Ohmido N et al (1997) Repetitive sequences: cause for variation in genome size and chromosome morphology in the genus *Oryza*. Plant Mol Biol 35:791–799. https://doi.org/10.1023/A:1005823124989

Vaughan DA, Morishima H, Kadowaki K (2003) Diversity in the *Oryza* genus. Curr Opin Plant Biol 6:139–146. https://doi.org/10.1016/S1369-5266(03)00009-8

Vitte C, Ishii T, Lamy F et al (2004) Genomic paleontology provides evidence for two distinct origins of Asian rice (*Oryza sativa* L.) Mol Gen Genomics 272:504–511. https://doi.org/10.1007/s00438-004-1069-6

Wang M, Yu Y, Haberer G et al (2014) The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. Nat Genet 46:982–988. https://doi.org/10.1038/ng.3044

Wing RA, Ammiraju JSS, Luo M et al (2005) The *Oryza* map alignment project: the golden path to unlocking the genetic potential of wild rice species. Plant Mol Biol 59:53–62. https://doi.org/10.1007/s11103-004-6237-x

Wu J, Yamagata H, Hayashi-Tsugane M et al (2004) Composition and structure of the centromeric region of rice chromosome 8. Plant Cell 16:967–976. https://doi.org/10.1105/tpc.019273.1

Xu X, Liu X, Ge S et al (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol 30:105–111. https://doi.org/10.1038/nbt.2050

Yamamoto T, Nagasaki H, Yonemaru J et al (2010) Fine definition of the pedigree haplotypes of closely related rice cultivars by means of genome-wide discovery of single-nucleotide polymorphisms. BMC Genomics 11:267. https://doi.org/10.1186/1471-2164-11-267

Yang C, Kawahara Y, Mizuno H et al (2012) Independent domestication of Asian rice followed by gene flow from *japonica* to *indica*. Mol Biol Evol 29:1–9. https://doi.org/10.1093/molbev/msr315

Yano M, Katayose Y, Ashikari M et al (2000) *Hd1*, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the Arabidopsis flowering time gene *CONSTANS*. Plant Cell 12:2473–2484

Yano K, Yamamoto E, Aya K et al (2016) Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat Genet 48:927–934. https://doi.org/10.1038/ng.3596

Yu J, Hu S, Wang J et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296:79–92. https://doi.org/10.1126/science.1068037

Yuan Q, Ouyang S, Liu J et al (2003) The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. Nucleic Acids Res 31:229–233. https://doi.org/10.1093/nar/gkg059

Zhang Y, Huang Y, Zhang L et al (2004) Structural features of the rice chromosome 4 centromere. Nucleic Acids Res 32:2023–2030. https://doi.org/10.1093/nar/gkh521

Zhang J, Chen LL, Xing F et al (2016) Extensive sequence divergence between the reference genomes of two elite *indica* rice varieties Zhenshan 97 and Minghui 63. Proc Nat Acad Sci 113:E5163–E5171. https://doi.org/10.1073/pnas.1611012113

Zhao Y, Tang L, Li Z et al (2015) Identification and analysis of unitary loss of long-established protein-coding genes in Poaceae shows evidences for biased gene loss and putatively functional transcription of relics. BMC Evol Biol 15:66. https://doi.org/10.1186/s12862-015-0345-x

Zhu Q, Ge S (2005) Phylogenetic relationships among A-genome species of the genus *Oryza* revealed by intron sequences of four nuclear genes. New Phytol 167:249–265. https://doi.org/10.1111/j.1469-8137.2005.01406.x

Zou X, Zhang F, Zhang J et al (2008) Analysis of 142 genes resolves the rapid diversification of the rice genus. Genome Biol. https://doi.org/10.1186/gb-2008-9-3-r49

Chapter 2
Small RNAs in Rice: Molecular Species and Their Functions

Yutaka Sato, Misuzu Nosaka-Takahashi, Toshiya Suzuki, and Sae Shimizu-Sato

Abstract Small RNAs are major components of gene regulatory pathways conserved among eukaryotes. In basic and applied sciences, RNA interference (RNAi) and artificial microRNAs (amiRNAs) are often used to modulate gene expression. The molecular mechanisms of RNAi are mainly studied in nematode or insect cells as models. Functional analyses of endogenous small RNAs, including studies of rice as a model, have greatly contributed to our understanding of plant biology. In plants, small RNA-based gene regulation has unique characteristics not found in animals, and many small RNAs regulate biological phenomena specific to plants. Recently, small RNA profiling using next-generation sequencers became possible, and various small RNA species were identified in plants including rice; their functional analyses are underway. This chapter summarizes the components of small RNA pathways, the molecular species of small RNAs, and the unique function of small RNAs in rice. It also considers the functions of small RNAs in relation to agriculturally important traits.

Keywords *Oryza sativa* · Rice · Small RNA · siRNA · miRNA · DICER · AGO

2.1 Introduction

Plant endogenous small RNA species (20–30 nucleotides (nt)) are classified mostly into microRNAs (miRNAs) and small interfering RNAs (siRNAs) on the basis of their precursor structures and processing (Table 2.1). miRNAs are produced from single-stranded RNAs with a stem–loop structure, whereas siRNAs are produced from double-stranded RNAs (dsRNAs) with nearly perfect complementarity. On

Y. Sato (✉) · M. Nosaka-Takahashi · T. Suzuki · S. Shimizu-Sato
National Institute of Genetics, Mishima, Shizuoka, Japan
e-mail: yusato@nig.ac.jp

T. Sasaki, M. Ashikari (eds.), *Rice Genomics, Genetics and Breeding*,
https://doi.org/10.1007/978-981-10-7461-5_2