Next-generation Sequencing Promoted the Release of Reference Genomes and Discovered Genome Evolution in Cereal Crops

Yong Huang†, Haiyang Liu† and Yongzhong Xing*

National Key Lab of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, China.

*Correspondence: yzxing@mail.hzau.edu.cn
†These authors contributed equally

https://doi.org/10.21775/cimb.027.037

Abstract

In recent decades, next-generation sequencing (NGS) was developed and brought biology into a new era. Rice, maize, wheat, sorghum and barley are the most important cereal crops and feed most of the world's population. Great progress in the study of cereal genomes has been made with the help of NGS. Reference genome sequence assembly and re-sequencing have grown exponentially. Thus, evolution and comparative genomics are renewed, including origin verification, evolution tracking and so on. In this review, we briefly record the development of sequencing technology, the comparison of NGS methods and platforms and summarize the bioinformatics tools used for NGS data analysis. We describe how NGS accelerates reference genome assembly and new evolutionary findings. We finally discuss how to discover more valuable resources and improve cereal breeding in the future.

Introduction

Over the three decades since Frederick Sanger and Walter Gilbert received the Nobel Prize for Chemistry in 1980, extraordinary progress has been made in genome sequencing technologies. Particularly since the new century, genome sequencing has made breakthroughs that have greatly increased the understanding of plant genomes and biology. To date, the most widely used first-generation sequencer is a typical Applied Biosystems' (ABI) 3730XL capillary electrophoresis sequencer, which has a longer read length (more than 1000 bp) and high accuracy rate (99.999%); the cost per 1000 bases is $0.5, and a many as 600,000 bases are generated per day (Shendure and Ji, 2008). However, due to the limitations of throughput and the relatively high cost of sequencing, the first-generation sequencer is unable to meet the needs of large-scale genome sequencing, such as deep sequencing and repeated sequencing, which prompted the birth of second-generation sequencing technologies, also known as next-generation sequencing (NGS).

Theories and platforms of NGS technologies

Over the last decade, NGS technologies have formed gradually and developed maturely. Roche/454 pyrosequencer, Illumina sequencer and ABI SOLiD sequencer are the main representatives. The approaches of NGS used in these three sequencers fall into two categories: sequencing by synthesis (SBS) and sequencing by ligation (SBL) (Fig. 2.1). No matter what kind of methods are used for sequencing, a library needs to be prepared

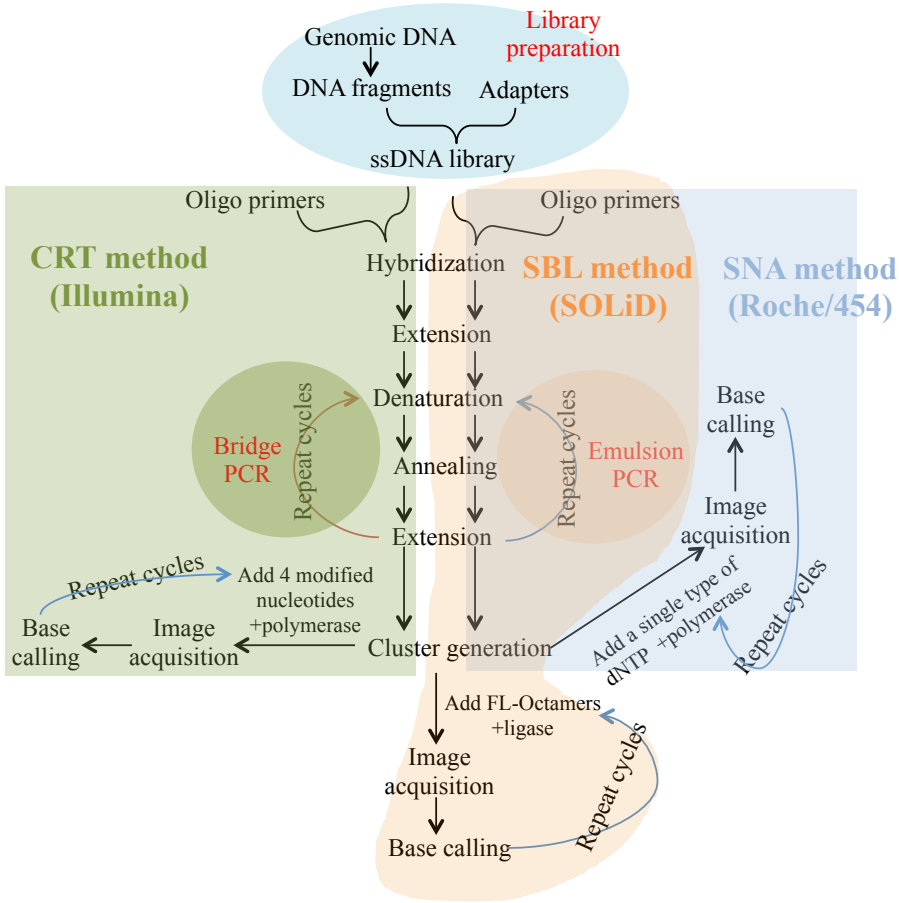

Figure 2.1 The workflow of the approaches of next-generation sequencing. Aqua green region represents library-construction process. Olive area, light blue region and light orange region represent the process of CRT, SNA and SBL methods, respectively. In the CRT method, modified nucleotide is that fluorophore-labelled, terminally blocked nucleotides. ssDNA, single-strand DNA; CRT, cyclic reversible termination; SNA, single-nucleotide addition; SBL, sequencing by ligation; FL-Octamers, fluorescence labelling octamers.

by randomly breaking the genomic DNA into small fragments and adding common adapters to the fragment ends. Subsequently, a DNA template is primed by a sequence that is complementary to the adapter sequence. In order to achieve the required signal for sequencing, different types of polymerase chain reactions (PCRs) are performed at different positions. After several rounds of amplification, respective DNA clusters are generated.

In SBS approaches, PCR and fluorescence sequencing are repeatedly performed on DNA samples filled with chips. SBS approaches are divided into cyclic reversible termination (CRT) and single-nucleotide addition (SNA). The uniqueness of the CRT approach is that the ribose 3′-OH group of the nucleotide is blocked (Fig. 2.1). During each cycle, a fragment of DNA template combined with an adapter will incorporate just one nucleotide as the blocked 3′ group prevents additional incorporates. After nucleotide incorporation, unincorporated nucleotides are washed away, and the image is obtained to determine which base was incorporated in each cluster; then, the cycle begins again. The Illumina platform, currently the most cost-effective and most widely used, is a representative of the CRT approach (Goodwin *et al.*, 2016; Mardis, 2008; Metzker, 2010). Unlike CRT approaches, SNA approaches do not require the dNTPs to be blocked because only one type of dNTP is flowed across the PicoTiterPlate (PTP) wells in each round of the sequencing reaction. SNA approaches are applied on a Roche/454 pyrosequencing platform, which is

the founder of and greatly promotes the development of NGS technologies (Rothberg and Leamon, 2008). In the CRT approaches of Illumina, bridge PCR and sequencing are carried out on the surface of the flow cell, while in the SNA approaches of Roche/454, emulsion PCR and sequencing are carried out in two different places: microbeads and PTP wells, respectively.

The unique feature of SBL approaches is that they replace the polymerization reaction during sequencing by synthesis with a ligation reaction (Fig. 2.1). ABI SOLiD sequencing technology is a typical SBL method. With SOLiD, each sequencing cycle introduces a degenerate population of fluorophore-labelled octamers composed of known nucleotides in the first and second positions. After ligation and imaging, the octamer is chemically cleaved between positions 5 and 6, removing the fluorescent label and leaving a free end for the next cycle of ligation. Single sequencing includes five rounds of sequencing reactions and ultimately can complete all positions of the nucleotide sequence; each position of the base is queried twice (Metzker, 2010).

In order to better explore DNA sequence information, a new generation of sequencing methods has been developed as single-molecule sequencing technology, also known as third-generation sequencing technology. It performs faster data reading and has great application potential. The read length of third-generation sequencing technology can reach several or even dozens of kilobases, thereby allowing the resolution of highly complex genomes with many long repetitive elements, copy number variations (CNVs) and structural variations. The most popular of the third-generation sequencing platforms is the single-molecule real-time (SMRT) sequencing method used by Pacific Biosciences (PacBio). In SMRT approaches, the polymerase is fixed at the bottom of the well and allows the DNA template to go through the zero-mode waveguides (ZMW) where sequencing can take place rather than polymerase binding and travelling along the DNA template as in NGS methods. All four dNTPs are labelled and available for incorporation. When the polymerase incorporates a fluorophore-labelled nucleotide, the emitted light is recorded by a camera, and the corresponding base sequence information then is transformed by the computer analyser (Goodwin *et al.*, 2016).

Comparison of NGS platforms

As new technologies emerge, existing problems are exacerbated or new problems arise. Due to double detection, the original sequencing accuracy of SOLiD is as high as 99.94%, and the accuracy of 15× coverage is even up to 99.999%. Therefore, SOLiD should provide the highest sequencing accuracy in the current NGS technology. Despite its accuracy, the SOLiD platform still has a number of limitations, such as displaying AT bias (Rieber *et al.*, 2013). The maximum read length of the Illumina, 454, SOLiD and PacBio platforms is 600 bp, 1000 bp, 75 bp and ~20,000 bp, respectively (Table 2.1). The very short read length may be the greatest limitation of the SOLiD platform preventing its widespread adoption. In NGS technologies, due to dependence on the template amplification phase prior to sequencing, the assembly of a genome with a high GC content is particularly limited (Aird *et al.*, 2011; Dohm *et al.*, 2008; Niu *et al.*, 2010). The unbiased and longer read length of SMRT sequencing markedly improved genome assembly with a high GC content through gap filling and repeat resolution (Shin *et al.*, 2013).

Although providing huge amounts of data, NGS platforms have higher error rates (~0.1 to 15%) and generally shorter (35 to 700 bp) read lengths than those of traditional Sanger sequencing platforms (Liu *et al.*, 2012). Compared with the huge genome, short read length makes sequence assembly extremely difficult. Although having vast quantities of sequence data and a depth of sequencing coverage reaching dozens or even hundreds of times, many users are still unable to complete the genome sequence assembly. Although third-generation sequencing technology overcomes the read length limitation of the NGS approaches, there remains a series of problems, such as greater cost, lower throughput and higher error rate of ~14% with insertion/deletion (InDel) errors (Carneiro *et al.*, 2012). Therefore, third-generation sequencing is always not used independently. Combinations of first-, second- and third-generation sequencing technologies play to their respective advantages and complement each other, leading to a high-quality genome sequence. For example, the high-quality genomes of ZS97 and MH63 were sequenced using a bacterial artificial chromosome (BAC)-by-BAC approach and PacBio SMRT technology, supplemented with

Table 2.1 Property summary of NGS platforms

Platform	MRL (bp)	Throughput	Run time	RPFC	Data quality	PE	ER (%)
Illumina Miniseq	2×150	0.6–7.5 Gb	4–24 hours	25 million	≥80% of base above Q30 at 2×150 bp	Substitution	<1
Illumina MiSeq	2×300	0.3–15 Gb	5–55 hours	25 million	≥80% of base above Q30 at 2×150 bp	Substitution	<0.1
Illumina Nextseq500	2×150	20–39 Gb	15–26 hours	130 million	≥75% of base above Q30 at 2×150 bp	Substitution	<1
	2×150	30–120 Gb	12–30 hours	400 million	≥75% of base above Q30 at 2×150 bp	Substitution	<1
Illumina HiSeq2500	2×250	10–300 Gb	7–60 hours	300 million	≥80% of base above Q30 at 2×150 bp	Substitution	0.1
	2×125	50–1000 Gb	<1–6 days	2 billion	≥80% of base above Q30 at 2×100 bp	Substitution	0.1
Illumina HiSeq3000	2×150	125–750 Gb	<1–3.5 days	2.5 billion	≥75% of base above Q30 at 2×150 bp	Substitution	0.1
Illumina HiSeq4000	2×150	125–1500 Gb	<1–3.5 days	2.5 billion	≥75% of base above Q30 at 2×150 bp	Substitution	0.1
Illumina HiSeq X	2×150	900–1800 Gb	<3 days	3 billion	≥75% of base above Q30 at 2×150 bp	Substitution	0.1
SOLiD5500xl	1×75	80–320 Gb	6–10 days	1.4 billion	~99.99%	A-T bias	>0.01
454 GS FLX XL+	1000	700 Mb	23 hours	1 million	Q20 read length of 400 bp	Indel	1
454 GS FLX XLR70	600	450 Mb	10 hours	1 million	Q20 read length of 400 bp	Indel	1
PacBio RS II	~20,000	500 Mb–1 Gb	4 hours	55,000	Q20-Q24 for bases without indels	Indel	~14

ER, error ratio; MRL, maximum read length (bp); PE, primary errors; RPFC, reads per flow cell.
Q20 and Q30 mean the probability that the base was miscalled is 1% and 0.1%, respectively.

Illumina whole-genome shotgun (WGS) data (Zhang *et al.*, 2016).

Bioinformatic tools for analysing NGS data

The emergence of NGS technology makes it possible to re-sequence entire genomes more efficiently and economically and in greater depth than ever before. When the reference genome is available, NGS data can be assembled into a new genome by alignment. However, the assembly of *de novo* sequencing is a greater challenge without a reference genome. Compared with traditional sequencing technology, the NGS raw data are more prodigious because of short sequences with a higher error rate. Therefore, further approaches for short-read assembly need to be proposed and implemented (Imelfort and Edwards, 2009). In addition, the identification of DNA sequence polymorphism within a genome is of utmost importance for dissecting genetic variation. The process of polymorphic screening includes the assembly and alignment of raw data, single nucleotide polymorphism (SNP) discovery and haplotype analysis.

Assembly and alignment of raw data

Different NGS platforms have different types of errors in sequencing. The Illumina, 454 and SOLiD platforms primarily produce errors of substitutions, InDel and A-T bias, respectively. The per-base quality score (Q) measures the error possibility of the base in the read. When Q is higher, the base error probability is smaller. The formula is $Q = -10\log P(\text{error})$. For example, Q20 and Q30 represent the probabilities of base error of 1% and 0.1%, respectively. In general, Q30 is the evaluation criteria for the Illumina platform. Illumina guarantees that the Illumina HiSeqX, the most widespread used high-throughput sequencing platform, has

a greater than 75% base quality above Q30 at 2×150 bp.

Correct sequence alignment is the key to polymorphism detection. Alignment errors may result in SNP and genotype detection errors. Therefore, alignment accuracy is vital for variation detection. Separating the true difference between point mutations and InDels requires improved alignment algorithms. There is a major issue in efficiently aligning short reads to a reference genome and handling ambiguity or lack of accuracy in this alignment. Most alignment algorithms for NGS data are based on either 'Hashing' or 'Burrows–Wheeler transform' (BWT). 'Hashing' includes the software MAQ, SOAP, Novoalign and Stampy. MAQ is used to map and assemble short reads. It can also report SNPs and InDels using a simple assembly visualizer (Li *et al.*, 2008a). SOAP efficiently aligns gapped and ungapped short reads onto reference sequences (Li *et al.*, 2008b). Stampy uses a hybrid mapping algorithm and a detailed statistical model to achieve both speed and sensitivity. It is particularly suitable for the reads that include sequence variation. Both Novoalign and Stampy show good performance for longer paired-end reads (Lunter and Goodson, 2011). BWT-based aligners, for example, Bowtie (Langmead *et al.*, 2009), SOAP2 (Li *et al.*, 2009) and BWA24 (Li and Durbin, 2009), which are less sensitive than the best hash-based mappers, are fast and particularly useful for aligning repetitive reads. The greater is the difference between the sequencing and reference genomes, the more difficult is the alignment; thus, the alignment can be solved by analysing the longer reads and pair-end reads.

SNP discovery and haplotype analysis

SNP is DNA sequence variation that occurs when a single nucleotide of the genome sequence is altered. SNP is the most abundant form of genetic variation existing between any diverse genotypes. Many high-throughput technologies, such as the Illumina platform, have been developed to efficiently genotype SNPs. The main aims underlying SNP calling is to obtain polymorphic loci or different base sites with reference sequences. With low-coverage sequencing data, it is difficult to detect SNPs and assemble the haplotypes with low non-reference allele frequency, as there is often considerable uncertainty associated with the results. A very high coverage is needed because certain regions may not be covered by the reads. According to the quality value of Q20, analyses filter and retain high-confidence bases. The bioinformatics tools that are used for SNP discovery and haplotype analysis include BreakDancer, VarScan, MAQ, INTERSNP, and Atlas-SNP, as well as HapHunt, HapCUT, BEAGLE 3.0, and Hap-seqX, respectively. BreakDancer can sensitively and accurately detect InDels ranging from 10 base pairs to 1 megabase pair, which are difficult to detect via a single conventional approach (Chen *et al.*, 2009). VarScan can detect SNPs and indels with high sensitivity and specificity in both Roche/454 sequencing and Illumina sequencing (Koboldt *et al.*, 2009). MAQ is accurate and efficient, rapidly aligns short reads to the reference genome, and discovers variants, including SNPs and short indels (Li *et al.*, 2008a). INTERSNP examines each position in the genome; a SNP is called whenever two samples differ at the same position (Herold *et al.*, 2009). Atlas-SNP is used for SNP and InDel discovery from genome resequencing using NGS technologies (Wheeler *et al.*, 2008). HapHunt uses K-means clustering to solve the haplotype-phasing problem and consists of identifying all haplotypes in an individual or population (Page *et al.*, 2014). HapCUT, which is based on computing max-cuts, is an efficient and accurate algorithm for the haplotype assembly problem (Bansal and Bafna, 2008). BEAGLE 3.0, which uses a haplotype frequency model, produces most likely haplotypes and sampled haplotypes for each individual (Browning and Browning, 2009). Hap-seqX can save all the intermediate results to memory, which not only solves the memory issue but also significantly improves the running time, and predicts haplotypes from whole-genome sequencing data (He and Eskin, 2013).

The pathways to assemble a reference genome in cereal crops

The reference genome sequence is the foundation of functional genomics and comparative genomics. With a high-quality reference genome, it is easy to develop millions of polymorphism markers and construct a super high-density genetic linkage map in crops, which can be used to discover more favourable genes for crop breeding. Since first-generation

sequencing technology was invented in 1973, scientists have attempted to sequence the genomes of microbes, plants, animals and humans. There are two main strategies to sequence the whole genome. One is WGS, which requires constructing a sequencing library using whole-genome DNA and sequencing enough folds to cover the whole genome. The other is clone-to-clone strategy, which requires constructing a library; selecting the minimum tiling path clones, which cover the whole genome sequence; and then sequencing the clones directly or sequencing in shotgun.

Reference genomes in the Sanger sequence era

The first reference genome sequence in crops came from rice, the model monocot, in 2002 (Goff *et al.*, 2002; Yu *et al.*, 2002). After 6-fold WGS with Sanger sequencing, Beijing Genomics Institute (BGI) and Syngenta completed the draft genome sequence of *indica* cultivar 93–11 and *japonica* cultivar Nipponbare separately in 2 years. Both draft genomes coved more than 90% per cent of the genome (Goff *et al.*, 2002; Yu *et al.*, 2002). However, for scientists, a draft map for model plants was not enough. In 2005, the IRGSP (International Rice Genome Sequencing Project) released the map-based genome sequence of Nipponbare (including the Syngenta sequencing data), which took seven years and coved 95% of the genome in the first version. When comparing the draft and map-based genome sequences, nearly 80% match well (International Rice Genome Sequencing, 2005). The rice reference genome has greatly promoted rice molecular genetics and biology. The achievement of the rice reference genome inspired other crop consortiums to form genome-sequencing programs. In 2009, the maize genome B73 was completed by BAC-to-BAC in Sanger and covered 89% of the whole 2.3 Gb genome. The maize genome is abundant with transposons, which occupy more than 85% of whole genome (Schnable *et al.*, 2009). In the same year, sorghum reference genome was assembled with WGS data and BAC pair-end sequences (Paterson *et al.*, 2009). In 2010, the soybean Williams 82 genome was finished by 8-fold WGS in Sanger and covered 80% of the whole 1115 Mb (Schmutz *et al.*, 2010). In the Sanger sequencing era, the WGS was timesaving but had difficulty covering the whole genome; the genome coverage was not as high as that from clone-by-clone, especially for complex genomes. Clone-by-clone is standard at first, but it's more time consuming and, more importantly, requires longer scaffolds in complex genome assembly.

Reference genomes assembled from next-generation sequencing

As mentioned above, second-generation sequencing includes massively parallel sequencing. The length and quantity are not as good as those of Sanger sequencing, but second-generation sequencing is cheaper and higher throughput; thus, genome sequencing has entered a new stage. The human 1000 Genomes Project, 1000 Plants Genome Project and the 3000 Rice Genome Project, among others, have been initiated (1000 Genomes Project Consortium, 2015; Li *et al.*, 2014; Matasci *et al.*, 2014). Crop reference genomes have been made breakthroughs, and several cereal crops have had reference genome sequences released in recent decades (Table 2.2).

It is unknown whether there is a significant difference in reference sequence quality between NGS and Sanger sequencing. Huang *et al.* (2009) reported that with 68× WGS in Illumina and 4× WGS in Sanger, the assembly genome coved 73% of the cucumber whole genome, which demonstrated that massively parallel sequencing was powerful for genome sequencing. Schatz *et al.* (2014) reported *de novo* assemblies of three strains of rice with 110-fold WGS in Illumina HiSeq 2000. When comparing the *de novo* assembled Nipponbare sequence with the IRGSP genome sequence, 91.2% was covered and 99.94% was identical, on average. The error rate of sequencing and assembly is at most 0.06%. That means that regarding accuracy, the *de novo* assembly genome was nearly as high quality as the reference genome from clone-to-clone assembly (Schatz *et al.*, 2014). Then, some complex genome underwent *de novo* assembly of the reference genome with high-fold WGS in Illumina HiSeq 2000.

Using wheat-A genome progenitor *Triticum urartu* var. G1812 as an example, 100-fold WGS in Illumina HiSeq 2000 obtained 79% coverage of the whole 4.94 Gb (Ling *et al.*, 2013). Compared with the Illumina HiSeq platform, the Roche/454 platform has longer reads of 600 to 1000 bp, which is a similar length as that of Sanger sequencing.

Table 2.2 Reference genome sequence obtained by NGS in cereal crops

Species	Chromosome no.	Genome size	Coverage	Gene no.	TEs (%)	Strategies	Reference
Oryza sativa ssp. *japonica* cv. Nipponbare	$2n=2x=24$	389 Mb	82%	39,083	–	WGS, 110×	Schatz *et al*., 2014
Oryza sativa ssp. *indica* cv. IR64	$2n=2x=24$	389 Mb	81%	37,758	–	WGS, 110×	
Oryza sativa ssp. *Indica* cv. DJ123	$2n=2x=24$	389 Mb	83%	37,812	–	WGS, 110×	
Oryza sativa ssp. *indica* cv. Zhenshan 97	$2n=2x=24$	384 Mb	91%	54,831	41	Clone-by-clone, and WGS, 200×	Zhang *et al*., 2016
Oryza sativa ssp. *indica* cv. Minghui 63	$2n=2x=24$	386 Mb	93%	57,174	42		
Triticum aestivum var. Chinese spring	$2n=6x=42$	17 Gb	61%	124,201	77	WGS	IWGSC, 2014
Sorghum bicolor var. BTx623	$2n=2x=20$	730 Mb	96%	34,496	55	WGS, 8.5× +clone libraries	Paterson *et al*., 2009
Setaria italica var. Zhang gu	$2n=2x=18$	490 Mb	86%	38,801	46	WGS, 82×	Zhang *et al*., 2012
Hordeum vulgare cv. Morex	$2n=2x=14$	5.1 Gb	90%	24,287	84	Clone-by-clone and WGS, 50×	International Barley Genome Sequencing *et al*., 2012
Hordeum vulgare var. Nudum	$2n=2x=14$	4.48 Gb	89%	36,151	81	WGS, 178×	Zeng *et al*., 2015

Tes, transposon elements; WGS, whole-genome shotgun; IWGSC, International Wheat Genome Sequencing Consortium.

Would the Roche/454 platform perform better than Illumina with shorter reads? Chalhoub *et al.* first assembled the allopolyploid *Brassica napus* oilseed genome using Illumina HiSeq data and some Roche/454 and Sanger data (Chalhoub *et al.*, 2014). According to Diguistini *et al.*'s study in fungus, Roche/454 has more advantages than Illumina, and Illumina, Roche/454 and Sanger combined together are powerful (Diguistini *et al.*, 2009). With 94-fold WGS in Roche/454, the wheat D-genome progenitor *Aegilops tauschii* var. AL8/78 assembly genome covered 97% of the 4.36 Gb genome (Zeng *et al.*, 2015). Even with the complex hexaploid bread wheat *Triticum aestivum* var. Chinese spring, a 17 Gb draft genome assembly was obtained by WGS on the Roche/454 platform based on chromosome isolation methods, covering 61% of the total genome (Consortium, 2014). As a special advantage of long reads and high throughput, Roche/454 combined with other platforms, such as Illumina HiSeq 2000 and Sanger sequencing, is a powerful strategy.

Zhang *et al.* completed five *Oryza* AA genome assemblies with 88–95% coverage using Roche/454 and Illumina HiSeq 2000 (Zhang *et al.*, 2014). In addition, after sequencing BAC clones by Roche/454 and assembly with 50-fold WGS sequence data, 90% of the 5.1 Gb *Hordeum vulgare* cv. Morex genome was obtained (International Barley Genome Sequencing *et al.*, 2012). Combined with the third-generation sequencing of long read sequences by 20 kb, it is possible to assemble a more complex genome and obtain a more accurate genome. The two cases of cereal crop genome assembly with third-generation sequencing are Zhenshan 97 and Minghui 63. With 110× Minimum tiling paths (MTPs) of the BAC clone sequence, the assembled genome with 200× WGS Illumina sequence data covered 90.6% and 93.2%, respectively (Zhang *et al.*, 2016).

In addition to cereal crops, many other crops have also been recently sequenced (URL: https://genomevolution.org/wiki/index.php/Sequenced_plant_genomes). The past decade was the genome

sequencing explosion decade, as massively parallel sequencing became popular. A combination of cheaper and high-throughput NGS with increasing computational capabilities would create an age in which sequencing could flourish.

Challenges for cereals with complex genomes

In addition to sequencing technology, other technology also accelerates genome assembly. During scaffold assembly, a physical map is useful. The fingerprint of the BAC end is mostly used in the past, but Hi-C, Optical Nano-mapping and nano-channel array technologies were established independently of BAC clone libraries, which are time-consuming to construct (Hastie *et al.*, 2013; Kaplan and Dekker, 2013; Korbel and Lee, 2013; Lam *et al.*, 2012). In spite of the advanced sequencing methods and mapping technology, some crops with complex genome structures still did not achieve a high-quality reference genome, such as sugar cane, which is challenged by high polyploidy and aneuploidy and a large size of 10 Gb. Sequenced wheat also has a draft genome sequence, which only covered 61% of the genome, still a long way from the 'reference genome'. In the future, to quickly obtain higher quality reference genomes for such crops, especially for genomes with complex structures, there are several pathways for improvement: (1) longer and high-fidelity sequence technology – a longer read will significantly improve the genome assembly and the highly repetitive region; and (2) new technology to sample chromosome-specific DNA, such as isolating every chromosome specifically. With high-quality and less complex samples, it will be easier to sequence and assemble genomes.

The genome evolution and germplasm resources in cereal crops

Maize, rice and wheat provide the most important foods in the world, and they all are cereals. However, it is not clear why and how they evolved into completely different crops. The reference genome provides us the chance to answer this evolution question. Evolution has always been presented with a phylogenetic tree. However, most phylogenetic trees are built on a single gene or several genes and are most likely incorrect (Degnan and Rosenberg, 2006). A genome sequence comparison would provide a more accurate phylogenetic tree. When building phylogenetic trees with oxtail millet (*Setaria italica*) and other *Poaceae* species, all the *Panicoideae* millet, sorghum and maize genome sequences cluster together, and millet appears to have split from sorghum and maize approximately 27 Mya. When looking at phylogenetic trees in detail, it was found that a whole genome duplication (WGD) in *Poaceae* occurred before the split millet (Jiao *et al.*, 2014; Zhang *et al.*, 2012).

Wild rice has different types of genomes that exhibit different characters in genome size and plant architecture (Table 2.3); even wild rice in the same AA genome type has diversified to adapt to different environments (Ge *et al.*, 1999; Jacquemin *et al.*, 2013; Zhang *et al.*, 2014). Domestication from wild species to cultivars sometimes is complex due to the interference from artificial selection. Scientists have debated the mechanism of rice domestication for many years before NGS. When building phylogenetic trees with a few different domestication genes, two completely opposite hypotheses are generated. One that *indica* and *japonica* cultivars originate from the same progenitor, while the other hypothesis is that these cultivars have independent origins from two different progenitors (Sang and Ge, 2007). Both hypotheses are supported by casual domestication related genes; using the no shattering trait as an example, *shattering 4* mutated by a functional SNP was found in all cultivar subpopulations, while the dominant *SH4* existed only in wild rice (Li *et al.*, 2006; Lin *et al.*, 2007). The domestication of *sh4* indicated that all the cultivars originate from a single progenitor. However, the result was opposite in *qSH1*, in which the shattering alleles were found mainly in *indica* subpopulations, indicating that *indica* and *japonica* originate from two different progenitors (Konishi *et al.*, 2006). Finally, this problem was solved by a Chinese group who re-sequenced 446 wild rice and 1083 cultivars. Then, a phylogenetic tree based on whole-genome sequences indicated that the wild rice was divided into three subgroups (Or-I, Or-II, Or-III), the ancient *japonica* subgroup was first domesticated from Or-III, and the ancient *japonica* subgroup evolved the *aromatic*, *temperate japonica* and *tropical japonica* subspecies. Then, the *indica* subgroup generated from the cross of ancient *japonica* and Or-I subgroup wild rice eventually resulted in the

Table 2.3 Reference genomes of rice wild relatives revealed by NGS

Species	Genome type	Genome size	Coverage	Gene no.	TE (%)	Strategies	Reference
Oryza rufipogon var. W1943	AA	–	406 Mb	–	–	WGS, 100×	Huang *et al*., 2012
Oryza brachyantha var. IRGC101232	FF	297 Mb	88%	32,038	29	WGS 104×, BAC clones	Chen *et al*., 2013
Oryza glaberrima cv. CG14	AA	358 Mb	88%	33,164	34	Clone-by-clone and WGS	Wang *et al*., 2014
Oryza glumaepatula	AA	366 Mb	91%	41,605	30	WGS, 86×	Zhang *et al*., 2014
Oryza meridionalis	AA	388 Mb	88%	39,106	30	WGS, 60×	
Oryza glaberrima	AA	370 Mb	93%	41,476	29	WGS, 56×	
Oryza barthii	AA	376 Mb	89%	42,283	30	WGS, 51×	
Oryza nivara	AA	395 Mb	95%	41,490	28	WGS, 73×	
Oryza barthii	AA	–	308 Mb	–	–	WGS, 110×	http://ensembl.gramene.org/species.html
Oryza longistaminata	AA	–	326 Mb	–	–	WGS	
Oryza glumaepatula	AA	–	373 Mb	–	–	WGS	
Oryza meridionalis	AA	–	336 Mb	–	–	WGS	
Oryza nivara	AA	–	338 Mb	–	–	WGS	
Oryza punctata	BB	–	394 Mb	–	–	WGS	

TE, transposon element; WGS, whole-genome shotgun.

indica and *aus* subspecies (Huang *et al*., 2012). A similar method was used to analyse the origin of African rice, cotton, wheat, fleshy fruit, etc. (Meyer *et al*., 2016).

A comparison of genome sequences told us not only about the domestication but also the improvement process. With 5.4× coverage on average, several teams from China re-sequenced six elite maize inbred lines that were the parents of the most productive hybrids in China and identified the complementation of presence/absence and other deletion variation contributing to heterosis together, which means these variations were selected during modern breeding (Lai *et al*., 2010). Furthermore, they increased the population to 278 temperate maize inbred lines from different stages of breeding history and found that after domestication, there was some introgression from wild relatives to increase the diversity of cultivars, and the favourable alleles of key loci for agronomy traits were selected for past breeding. They also noted that some relative fraction of rare alleles would be helpful in future breeding programmes (Jiao *et al*., 2012). In a word, NGS also helped to identify variations, favourable alleles and breeding processes during past artificial selection.

At the same time, to capture all the QTLs from whole germplasm resources, only studying the natural variations that exist in reference genome sequence is not enough, as some of the cultivar- or subpopulation-specific DNA sequences or genes would be missed. The pan-genome containing a core genome and variable genome has been assembled in many crops using NGS, and much more accurate QTLs were identified in the variable genome when studied in association with the pan-genome. In a study by Yao *et al*. in rice, when 840 metabolic traits associated with the pan-genome from 1483 cultivars were compared with the reference genome, 23.5% of the traits had higher association signals and 41.6% had associated SNP locations concordant with associated variable genome sequences (Schatz *et al*., 2014; Yao *et al*., 2015).

Crops always suffer from artificial selection. Empirical selection could enhance cultivars' yield potential and stress tolerance. However, it is not enough to directly or efficiently improve a target trait. If more genes were identified in nature, they could be used for marker-assisted selection for breeding. Therefore, mining favourable alleles from genetic resources becomes more important. Genome-wide association studies (GWAS) are a

robust way to establish the relationship between genome variation and phenotype. Hundreds of QTLs/genes have been detected based on natural variation in cereals (Bai *et al.*, 2016; Han *et al.*, 2016; Huang *et al.*, 2010; Magwa *et al.*, 2016; Tian *et al.*, 2011). It is believed that more QTLs will be discovered from nature by GWAS for the future breeding programmes. Therefore, NGS is encouraged to mine the favourable alleles recently mutated in cultivars. From what has been discussed above, to trace all the natural variation and artificial selection and to more sufficiently explore and apply the potential of germplasm resources in the era of NGS, we should make the following efforts.

Build higher quality genome sequences. Genome sequences are the foundation of the study of genome evolution. With a pan-genome, we could gather complete information for a species, mainly in association studies now but later in other genome study areas, such as comparative genomics (Yao *et al.*, 2015).

Develop more accurate sequencing technologies. Now in population genetic and evolution studies, most studies pay more attention to SNPs but little attention to InDels due to the low confidence of InDels with NGS (Huang *et al.*, 2012; Huang *et al.*, 2010). We could cover both SNP and InDel variations if the InDel information could be obtained as easily and accurately as SNPs.

Increase the germplasm resources, such as adding more wild species and neighbour species. More germplasm resources would provide more natural variations to explore, study and apply (Zhang *et al.*, 2014).

Identify genetic interaction within populations. The natural variation was contributed not only by loci but also by loci–loci interactions and loci–environment interactions. It has been studied in rice yield heterosis by sequencing an immortalized F2 population and a sufficiently large F2 population (Huang *et al.*, 2016; Zhou *et al.*, 2012). Genetic improvement would be pushed forwards by using favourable alleles from wild relatives and the latest mutations in cultivars identified by NGS.

Acknowledgements

This work was partially supported by natural science foundation of China (91535301) and Agriculture Public Welfare Scientific Research Project of China (201303008).

References

1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74. http://dx.doi.org/10.1038/nature15393

Aird, D., Ross, M.G., Chen, W.S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. *12*, R18. http://dx.doi.org/10.1186/gb-2011-12-2-r18

Bai, X., Zhao, H., Huang, Y., Xie, W., Han, Z., Zhang, B., Guo, Z., Yang, L., Dong, H., Xue, W., *et al.* (2016). Genome-wide association analysis reveals different genetic control in panicle architecture between and rice. Plant Genome *9*, 02. http://dx.doi.org/10.3835/plantgenome2015.11.0115

Bansal, V., and Bafna, V. (2008). HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. Bioinformatics *24*, i153–9. http://dx.doi.org/10.1093/bioinformatics/btn298

Browning, B.L., and Browning, S.R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. *84*, 210–223. http://dx.doi.org/10.1016/j.ajhg.2009.01.005

Carneiro, M.O., Russ, C., Ross, M.G., Gabriel, S.B., Nusbaum, C., and DePristo, M.A. (2012). Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics *13*, 375. http://dx.doi.org/10.1186/1471-2164-13-375

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I.A., Tang, H., Wang, X., Chiquet, J., Belcram, H., Tong, C., Samans, B., *et al.* (2014). Plant genetics. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science *345*, 950–953. http://dx.doi.org/10.1126/science.1253435

Chen, J., Huang, Q., Gao, D., Wang, J., Lang, Y., Liu, T., Li, B., Bai, Z., Luis Goicoechea, J., Liang, C., *et al.* (2013). Whole-genome sequencing of Oryza brachyantha reveals mechanisms underlying *Oryza* genome evolution. Nat. Commun. *4*, 1595. http://dx.doi.org/10.1038/ncomms2596

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., *et al.* (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods. *6*, 677–681. http://dx.doi.org/10.1038/nmeth.1363

Consortium, I.W.G.S. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science *345*, 1251788. http://dx.doi.org/10.1126/science.1251788

Degnan, J.H., and Rosenberg, N.A. (2006). Discordance of species trees with their most likely gene trees. PLOS Genet. *2*, e68. http://dx.doi.org/10.1371/journal.pgen.0020068

Diguistini, S., Liao, N.Y., Platt, D., Robertson, G., Seidel, M., Chan, S.K., Docking, T.R., Birol, I., Holt, R.A., Hirst, M., *et al.* (2009). De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. Genome Biol. *10*, R94. http://dx.doi.org/10.1186/gb-2009-10-9-r94

Dohm, J.C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets

from high-throughput DNA sequencing. Nucleic Acids Res. *36*, e105. http://dx.doi.org/10.1093/nar/gkn425

Ge, S., Sang, T., Lu, B.R., and Hong, D.Y. (1999). Phylogeny of rice genomes with emphasis on origins of allotetraploid species. Proc. Natl. Acad. Sci. U.S.A. *96*, 14400–14405.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., 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. http://dx.doi.org/10.1126/science.1068275

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. *17*, 333–351. http://dx.doi.org/10.1038/nrg.2016.49

Han, Z., Zhang, B., Zhao, H., Ayaad, M., and Xing, Y. (2016). Genome-wide association studies reveal that diverse heading date genes respond to short and long day lengths between Indica and Japonica Rice. Front. Plant Sci. *7*, 1270. http://dx.doi.org/10.3389/fpls.2016.01270

Hastie, A., Lam, E., Chan, T., Requa, M., Austin, T.A.M., Trintchouk, F., Saghbini, M., Lai, Y., Mak, A., and Kwok, P. Structural Variation Discovery by De Novo Genome mapping of the human genome at the single molecule level using nanochannel linearization. http://bionanogenomics.com/wp-content/uploads/2013/10/ESHG2013_poster.pdf

He, D., and Eskin, E. (2013). Hap-seqX: expedite algorithm for haplotype phasing with imputation using sequence data. Gene *518*, 2–6. http://dx.doi.org/10.1016/j.gene.2012.11.093

Herold, C., Steffens, M., Brockschmidt, F.F., Baur, M.P., and Becker, T. (2009). INTERSNP: genome-wide interaction analysis guided by a priori information. Bioinformatics *25*, 3275–3281. http://dx.doi.org/10.1093/bioinformatics/btp596

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., Lucas, W.J., Wang, X., Xie, B., Ni, P., *et al.* (2009). The genome of the cucumber, Cucumis sativus L. Nat. Genet. *41*, 1275–1281. http://dx.doi.org/10.1038/ng.475

Huang, X., Kurata, N., Wei, X., Wang, Z.X., Wang, A., Zhao, Q., Zhao, Y., Liu, K., Lu, H., Li, W., *et al.* (2012). A map of rice genome variation reveals the origin of cultivated rice. Nature *490*, 497–501. http://dx.doi.org/10.1038/nature11532

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C., Zhu, C., Lu, T., Zhang, Z., *et al.* (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. *42*, 961–967. http://dx.doi.org/10.1038/ng.695

Huang, X., Yang, S., Gong, J., Zhao, Q., Feng, Q., Zhan, Q., Zhao, Y., Li, W., Cheng, B., Xia, J., *et al.* (2016). Genomic architecture of heterosis for yield traits in rice. Nature *537*, 629–633. http://dx.doi.org/10.1038/nature19760

Imelfort, M., and Edwards, D. (2009). De novo sequencing of plant genomes using second-generation technologies. Briefings Bioinf. *10*, 609–618. http://dx.doi.org/10.1093/bib/bbp039

Mayer, K.F., Waugh, R., Brown, J.W., Schulman, A., Langridge, P., Platzer, M., Fincher, G.B., Muehlbauer, G.J., Sato, K., Close, T.J., *et al.* (2012). A physical, genetic and functional sequence assembly of the barley genome. Nature *491*, 711–716. http://dx.doi.org/10.1038/nature11543

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

Jacquemin, J., Bhatia, D., Singh, K., and Wing, R.A. (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. http://dx.doi.org/10.1016/j.pbi.2013.02.014

Jiao, Y., Li, J., Tang, H., and Paterson, A.H. (2014). Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. Plant Cell *26*, 2792–2802. http://dx.doi.org/10.1105/tpc.114.127597

Jiao, Y., Zhao, H., Ren, L., Song, W., Zeng, B., Guo, J., Wang, B., Liu, Z., Chen, J., Li, W., *et al.* (2012). Genome-wide genetic changes during modern breeding of maize. Nat. Genet. *44*, 812–815. http://dx.doi.org/10.1038/ng.2312

Kaplan, N., and Dekker, J. (2013). High-throughput genome scaffolding from in vivo DNA interaction frequency. Nat. Biotechnol. *31*, 1143–1147. http://dx.doi.org/10.1038/nbt.2768

Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K., and Ding, L. (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics *25*, 2283–2285. http://dx.doi.org/10.1093/bioinformatics/btp373

Konishi, S., Izawa, T., Lin, S.Y., Ebana, K., Fukuta, Y., Sasaki, T., and Yano, M. (2006). An SNP caused loss of seed shattering during rice domestication. Science *312*, 1392–1396. http://dx.doi.org/10.1126/science.1126410

Korbel, J.O., and Lee, C. (2013). Genome assembly and haplotyping with Hi-C. Nat. Biotechnol. *31*, 1099–1101. http://dx.doi.org/10.1038/nbt.2764

Lai, J., Li, R., Xu, X., Jin, W., Xu, M., Zhao, H., Xiang, Z., Song, W., Ying, K., Zhang, M., *et al.* (2010). Genome-wide patterns of genetic variation among elite maize inbred lines. Nat. Genet. *42*, 1027–1030. http://dx.doi.org/10.1038/ng.684

Lam, E.T., Hastie, A., Lin, C., Ehrlich, D., Das, S.K., Austin, M.D., Deshpande, P., Cao, H., Nagarajan, N., Xiao, M., *et al.* (2012). Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat. Biotechnol. *30*, 771–776. http://dx.doi.org/10.1038/nbt.2303

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25. http://dx.doi.org/10.1186/gb-2009-10-3-r25

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

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. http://dx.doi.org/10.1093/bioinformatics/btp324

Li, H., Ruan, J., and Durbin, R. (2008a). Mapping short DNA sequencing reads and calling variants using

mapping quality scores. Genome Res. *18*, 1851–1858. http://dx.doi.org/10.1101/gr.078212.108

Li, J.Y., Wang, J., and Zeigler, R.S. (2014). The 3,000 rice genomes project: new opportunities and challenges for future rice research. GigaScience *3*, 8. http://dx.doi.org/10.1186/2047-217X-3-8

Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008b). SOAP: short oligonucleotide alignment program. Bioinformatics *24*, 713–714. http://dx.doi.org/10.1093/bioinformatics/btn025

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., and Wang, J. (2009). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics *25*, 1966–1967. http://dx.doi.org/10.1093/bioinformatics/btp336

Lin, Z., Griffith, M.E., Li, X., Zhu, Z., Tan, L., Fu, Y., Zhang, W., Wang, X., Xie, D., and Sun, C. (2007). Origin of seed shattering in rice (Oryza sativa L.). Planta *226*, 11–20. http://dx.doi.org/10.1007/s00425-006-0460-4

Ling, H.Q., Zhao, S., Liu, D., Wang, J., Sun, H., Zhang, C., Fan, H., Li, D., Dong, L., Tao, Y., *et al.* (2013). Draft genome of the wheat A-genome progenitor Triticum urartu. Nature *496*, 87–90. http://dx.doi.org/10.1038/nature11997

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. J. Biomed. Biotechnol. *2012*, 251364. http://dx.doi.org/10.1155/2012/251364

Lunter, G., and Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. *21*, 936–939. http://dx.doi.org/10.1101/gr.111120.110

Magwa, R.A., Zhao, H., and Xing, Y. (2016). Genome-wide association mapping revealed a diverse genetic basis of seed dormancy across subpopulations in rice (Oryza sativa L.). BMC Genet. *17*, 28. http://dx.doi.org/10.1186/s12863-016-0340-2

Mardis, E.R. (2008). Next-generation DNA sequencing methods. Annu. Rev. Genomics. Hum. Genet. *9*, 387–402. http://dx.doi.org/10.1146/annurev.genom.9.081307.164359

Matasci, N., Hung, L.H., Yan, Z., Carpenter, E.J., Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Ayyampalayam, S., Barker, M., *et al.* (2014). Data access for the 1,000 Plants (1KP) project. GigaScience *3*, 17. http://dx.doi.org/10.1186/2047-217X-3-17

Metzker, M.L. (2010). Sequencing technologies – the next generation. Nat. Rev. Genet. *11*, 31–46. http://dx.doi.org/10.1038/nrg2626

Meyer, R.S., Choi, J.Y., Sanches, M., Plessis, A., Flowers, J.M., Amas, J., Dorph, K., Barretto, A., Gross, B., Fuller, D.Q., *et al.* (2016). Domestication history and geographical adaptation inferred from a SNP map of African rice. Nat. Genet. *48*, 1083–1088. http://dx.doi.org/10.1038/ng.3633

Niu, B., Fu, L., Sun, S., and Li, W. (2010). Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC Bioinf. *11*, 187. http://dx.doi.org/10.1186/1471-2105-11-187

Page, J.T., Liechty, Z.S., Huynh, M.D., and Udall, J.A. (2014). BamBam: genome sequence analysis tools for biologists. BMC Res. Notes. *7*, 829. http://dx.doi.org/10.1186/1756-0500-7-829

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., *et al.* (2009). The Sorghum bicolor genome and the diversification of grasses. Nature *457*, 551–556. http://dx.doi.org/10.1038/nature07723

Rieber, N., Zapatka, M., Lasitschka, B., Jones, D., Northcott, P., Hutter, B., Jäger, N., Kool, M., Taylor, M., Lichter, P., *et al.* (2013). Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. PLOS ONE *8*, e66621. http://dx.doi.org/10.1371/journal.pone.0066621

Rothberg, J.M., and Leamon, J.H. (2008). The development and impact of 454 sequencing. Nat. Biotechnol. *26*, 1117–1124. http://dx.doi.org/10.1038/nbt1485

Sang, T., and Ge, S. (2007). The puzzle of rice domestication. J. Integr. Plant Biol. *49*, 760–768. http://dx.doi.org/10.1111/j.1744-7909.2007.00510.x

Schatz, M.C., Maron, L.G., Stein, J.C., Hernandez Wences, A., Gurtowski, J., Biggers, E., Lee, H., Kramer, M., Antoniou, E., Ghiban, E., *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. http://dx.doi.org/10.1186/PREACCEPT-2784872521277375

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., *et al.* (2010). Genome sequence of the palaeopolyploid soybean. Nature *463*, 178–183. http://dx.doi.org/10.1038/nature08670

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., *et al.* (2009). The B73 maize genome: complexity, diversity, and dynamics. Science *326*, 1112–1115. http://dx.doi.org/10.1126/science.1178534

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nat. Biotechnol. *26*, 1135–1145. http://dx.doi.org/10.1038/nbt1486

Shin, S.C., Ahn, D.H., Kim, S.J., Lee, H., Oh, T.J., Lee, J.E., and Park, H. (2013). Advantages of single-molecule real-time sequencing in high-GC content genomes. PLOS ONE *8*, e68824. http://dx.doi.org/10.1371/journal.pone.0068824

Tian, F., Bradbury, P.J., Brown, P.J., Hung, H., Sun, Q., Flint-Garcia, S., Rocheford, T.R., McMullen, M.D., Holland, J.B., and Buckler, E.S. (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat. Genet. *43*, 159–162. http://dx.doi.org/10.1038/ng.746

Wang, M., Yu, Y., Haberer, G., Marri, P.R., Fan, C., Goicoechea, J.L., Zuccolo, A., Song, X., Kudrna, D., Ammiraju, J.S., *et al.* (2014). The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. Nat. Genet. *46*, 982–988. http://dx.doi.org/10.1038/ng.3044

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T., *et al.* (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature *452*, 872–876. http://dx.doi.org/10.1038/nature06884

Yao, W., Li, G., Zhao, H., Wang, G., Lian, X., and Xie, W. (2015). Exploring the rice dispensable genome using a metagenome-like assembly strategy. Genome Biol. *16*, 187. http://dx.doi.org/10.1186/s13059-015-0757-3

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., *et al.* (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science *296*, 79–92. http://dx.doi.org/10.1126/science.1068037

Zeng, X., Long, H., Wang, Z., Zhao, S., Tang, Y., Huang, Z., Wang, Y., Xu, Q., Mao, L., Deng, G., *et al.* (2015). The draft genome of Tibetan hulless barley reveals adaptive patterns to the high stressful Tibetan Plateau. Proc. Natl. Acad. Sci. U.S.A. *112*, 1095–1100. http://dx.doi.org/10.1073/pnas.1423628112

Zhang, G., Liu, X., Quan, Z., Cheng, S., Xu, X., Pan, S., Xie, M., Zeng, P., Yue, Z., Wang, W., *et al.* (2012). Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. Nat. Biotechnol. *30*, 549–554. http://dx.doi.org/10.1038/nbt.2195

Zhang, J., Chen, L.L., Xing, F., Kudrna, D.A., Yao, W., Copetti, D., Mu, T., Li, W., Song, J.M., Xie, W., *et al.* (2016). Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. Proc. Natl. Acad. Sci. U.S.A. *113*, E5163–5171. http://dx.doi.org/10.1073/pnas.1611012113

Zhang, Q.J., Zhu, T., Xia, E.H., Shi, C., Liu, Y.L., Zhang, Y., Liu, Y., Jiang, W.K., Zhao, Y.J., Mao, S.Y., *et al.* (2014). Rapid diversification of five Oryza AA genomes associated with rice adaptation. Proc. Natl. Acad. Sci. U.S.A. *111*, E4954–62. http://dx.doi.org/10.1073/pnas.1418307111

Zhou, G., Chen, Y., Yao, W., Zhang, C., Xie, W., Hua, J., Xing, Y., Xiao, J., and Zhang, Q. (2012). Genetic composition of yield heterosis in an elite rice hybrid. Proc. Natl. Acad. Sci. U.S.A. *109*, 15847–15852. http://dx.doi.org/10.1073/pnas.1214141109