Feature Review

Crop genome sequencing: lessons and rationales

Catherine Feuillet[1], Jan E. Leach[2], Jane Rogers[3], Patrick S. Schnable[4] and Kellye Eversole[5]

[1] Institut National de la Recherche Agronomique-Université Blaise Pascal-UMR1095-Domaine de Crouel, 63100 Clermont-Ferrand, France
[2] Department of Bioagricultural Sciences and Pest Management, Colorado State University, Fort Collins, CO 80523, USA
[3] The Genome Analysis Centre, Norwich Research Park, Colney, Norwich, NR4 7UH, UK
[4] Center for Plant Genomics, Iowa State University, Ames, IA 50011, USA
[5] Eversole Associates, Wyoming Road 5207, Bethesda, MD 20816, USA

2010 marks the 10th anniversary of the completion of the first plant genome sequence (*Arabidopsis thaliana*). Triggered by advancements in sequencing technologies, many crop genome sequences have been produced, with eight published since 2008. To date, however, only the rice (*Oryza sativa*) genome sequence has been finished to a quality level similar to that of the *Arabidopsis* sequence. This trend to produce draft genomes could affect the ability of researchers to address biological questions of speciation and recent evolution or to link sequence variation accurately to phenotypes. Here, we review the current crop genome sequencing activities, discuss how variability in sequence quality impacts utility for different studies and provide a perspective for a paradigm shift in selecting crops for sequencing in the future.

A decade of plant genome sequencing

A high-quality, reference genome sequence provides access to the relatively complete gene catalog for a species, the regulatory elements that control their function and a framework for understanding genomic variation. As such, it is a prerequisite resource for understanding fully the roles of genes in development, driving genomics-based approaches to systems biology and efficiently exploiting the natural and induced genetic diversity of an organism. Publication of the first plant genome sequence, *Arabidopsis* (*Arabidopsis thaliana*), in 2000 [1] spawned an expansion in genomics-based *Arabidopsis* research and the exploitation of annotated genes to explore orthologous genes in other plants. It also paved the way for sequencing several other model plant genomes and a few crop genomes.

Ten years after this major achievement, we examine the current situation regarding the sequencing of other plant genomes, in particular those of crops, many of which have more complex genome structures compared with model organisms and require the development of new tools and strategies for genome interpretation. How useful have the first complete sequences been and how do they compare with more recently generated draft genome sequences?

Glossary

BAC-by-BAC strategy (also known as hierarchical shotgun sequencing or a clone-by-clone shotgun strategy): based on a two-step progression. First, a physical map of the target genome (or chromosome) is established using BAC clones (typically 100–150 kb) and a set of overlapping clones representing a minimal tiling path (MTP) is then ordered along the chromosomes of the target genome. The individually mapped clones of the MTP are then subjected to shotgun sequencing. DNA from each BAC clone is fragmented randomly into smaller pieces that are either cloned into a plasmid and sequenced with Sanger sequencing technologies or directly sequenced via NGS technologies. The sequences are then aligned so that identical sequences overlap and contiguous sequences (contigs) are assembled into a finished sequence.

Homoeologous chromosomes: chromosomes that are located in different species or in different genomes in polyploid species and that originate from a common ancestral chromosome.

Next generation sequencing (NGS): sequencing technologies based on massive parallel sequencing as opposed to the Sanger sequencing technology. Most NGS technologies eliminate the need for the bacterial cloning used in Sanger sequencing and rely instead on the amplification of single isolated DNA molecules and their analysis in a massively parallel way. Hundreds of thousands, or even tens of millions, of single-stranded DNA molecules are immobilized on a solid surface,‘such as a glass slide or on beads, depending on the platform used. To date, the commercially available NGS platforms that are generally used for plant genome sequencing are Roche/454 FLX (http://454.com/products-solutions/system-features.asp), the Illumina/Solexa Genome Analyser (http://www.illumina.com/technology/sequencing_technology.ilmn) and the Applied Biosystems SOLiD^{TM} System (http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems.html).

Polyploidy: the presence of more than one chromosome set within the same nucleus. It originates from either whole-genome doubling (autopolyploidy) or by interspecific or intergeneric hybridizations followed by chromosome doubling (allopolyploidy). Polyploidy has been common throughout the evolution of plant genomes. Although most of the current genomes went through a diploidization process and only traces of ancient polyploidy events can be found today, allopolyploidy is still observed in many groups of plants, including crops such as banana (*Musa* sp.), *Brassica*, bread wheat (*Triticum aestivum*), durum wheat (*Triticum durum*), oat (*Avena sativa*), cotton (*Gossypium* sp.), canola (*Brassica napus*), coffee (*Coffea* sp.) and tobacco (*Nicotiana tabacum*).

Sanger sequencing technology (also known as dideoxy sequencing or chain termination method): used to sequence the first plant genomes following a BAC-by-BAC approach. It is based on the sequencing of DNA fragments amplified after cloning in bacterial vectors and relies on the use of dideoxynucleotides (ddNTPs) in addition to the normal nucleotides (dNTPs) found in DNA. Dideoxynucleotides are the same as nucleotides except that they contain a hydrogen group on the 3′ carbon instead of a hydroxyl group (OH). These modified nucleotides, when integrated into a sequence, prevent the addition of further nucleotides and, thus, the DNA chain is terminated. The incorporation of ddNTPs labeled with different fluorescent dyes enables the detection of all terminated sequence products through electrophoresis based on size separation.

Whole-genome shotgun (WGS) strategy: involves the assembly of sequence reads generated in a random, genome-wide fashion. In this approach, the entire genome of the target genome (chromosome) is fragmented into pieces of defined sizes that are subcloned into plasmid vectors or directly sequenced via NGS technologies. Sequence reads are generated from many subclones so as to produce highly redundant sequence coverage across the genome or chromosome. The sequences are then assembled using various computational methods to produce a consensus sequence.

Corresponding author: Feuillet, C. (catherine.feuillet@clermont.inra.fr).

1360-1385/$ – see front matter © 2010 Elsevier Ltd. All rights reserved. doi:10.1016/j.tplants.2010.10.005

Stimulated by the success of the human genome project and the prospect of DNA sequencing becoming a routine medical practice, the costs in sequencing have been reduced by over 10 000 times during the past ten years [2]. With further reductions in cost and time for DNA sequence generation projected, the rapid and inexpensive acquisition of draft sequences of large and complex organisms, such as those belonging to the Triticeae (5–17 Gb) or the pines (*Pinus* spp.; 18–40 Gb), is now achievable. Furthermore, large-scale resequencing of genomes is becoming more practical, enabling thorough analyses and cataloguing of genetic variation. Access to these data is changing the nature of experimentation designed to understand the structure, function and evolution of organisms, as well as transforming the methods and tools for crop improvement. In the rush to carry out experiments that are made feasible by technology, are researchers losing sight of the quality of the resources that are needed as references for the longer term? Will the whole-genome shotgun (WGS; see Glossary) draft genomes of today have the utility of the finished sequences obtained previously with map-based approaches and Sanger sequencing technology? Is this quality necessary for all genomes? What kind of strategies should be developed to ensure that the quality of crop genome sequences meets the long-term needs of biologists and plant breeders?

Most of the animal genome sequencing projects have been justified by their relationship to, and understanding of, the human genome. However, in the plant world, no central focus exists. This is because there are ~25 major food and feed crop species that nourish the world and seven that are becoming increasingly important for biofuel and biomaterial production (Table 1). Although the absolute range of genome size is comparable between plants and animals [3], more plant species have large and complex genome sizes and structures than do animal species [4]. Consequently, until recently, the choice of genomes for sequencing has been driven mainly by cost efficiency and complexity avoidance. Before the maize (*Zea mays*) and soybean (*Glycine max*) genome drafts were produced in 2010 [5,6], only plants with relatively small genomes (median size of 466 Mb) were selected for sequencing [7]. Given that the median genome size of the most important crops (Table 2) is 777 Mb, complexity problems cannot be avoided if crop breeding is to benefit from genomics tools. With the promises offered by next-generation sequencing (NGS) technologies, this paradigm can change and plants can be prioritized for sequencing in relation to their value to humans.

Here, we examine the different strategies that have been used so far to sequence the genomes of crop plants. We compare the characteristics of the sequences obtained in light of their applications for crop improvement, examine the use and contribution of model genome sequences, the first applications of crop genome sequences and discuss the potential offered by NGS technologies to produce new genome sequence resources.

State of the art of plant genome sequencing

(R)evolution of sequencing strategies and technologies

At the beginning of October 2010, 25 genome sequences were publicly available for 19 plant species (Table 2) and 15 projects were underway or not publicly available (Table 3). Of the 19 plant species with available genome sequences, 13 are crops (Table 2), most of which have been sequenced since 2005 and eight have been published since 2008. Sequencing these genomes was driven by different goals: (i) as a model for plant biology (e.g. *A. thaliana*) or crop genomics [e.g. *Brachypodium distachyon*, *Medicago truncatula*, poplar (*Populus trichocarpa*) and rice]; (ii) as an essential crop for food, animal feed, or energy [e.g. rice, maize, soybean, sorghum (*Sorghum bicolor*) and grapevine (*Vitis vinifera*)]; (iii) as a component of the tree of life for understanding evolution through comparative genomics (e.g. *Amborella trichopoda)*; and (iv) to provide efficient genome-wide molecular tools to support crop improvement through breeding programs. Technical feasibility, financial opportunities and goals have influenced the selection of genomes that have been sequenced, the strategies used and the level of completion (from standard draft to finished sequence) [8].

The first sequenced plant genomes, those of *A. thaliana* and rice, were both produced using state-of-the-art technology for the late 1990 s: sequencing of overlapping bacterial artificial chromosome (BAC) clones selected from a physical map. Sanger dideoxynucleotide chemistry was used to generate a draft assembly of each clone from the sequences of random subclones and to carry out directed sequencing to close gaps and resolve errors or ambiguities. Rice and *Arabidopsis* are the only plant species that have finished genome sequences available to date (Table 2) and they were obtained after several years of international effort involving hundreds of people and, for *A. thaliana*, at an estimated cost of US$70 million. The demonstration that WGS sequencing could be used to sequence bacteria [9] and rapidly generate draft sequences of *Drosophila* [10], human [11] and mouse genomes [12], without requiring prior construction of a physical map, stimulated the widespread adoption of this approach for multiple plant genomes (Table 2) and other organisms [13]. With the exception of maize, all other recently published plant genome sequences were generated using a WGS approach to generate paired-end reads from cloned DNA fragments of multiple sizes (plasmid, fosmid and BAC) to support the final assembly (Table 2). Regardless of the strategy selected, until 2009, plant genomes were assembled from 600–800-base Sanger sequence reads that facilitated the assembly of repeat-rich genomes.

Since 2005, new platforms have emerged (the so-called 'NGS' technologies) that offer improvements in throughput and cost efficiency by using massively parallel sequencing systems [14,15]. Thus, the first plant genomes sequenced primarily with NGS technologies are beginning to be reported. For example, 95% of the reads used to assemble the cucumber (*Cucumis sativus*) genome sequence were short reads (50 bases) sequenced on the Illumina GA platform [16]; whereas, for the cassava (*Manihot esculenta*) genome, 61 million Roche 454 GS-FLX Titanium reads (average length 400 bases) were combined with onefold genome coverage in Sanger reads, providing a first assembly of 416 Mb (55% of the total genome) (http://www.phytozome.net/cassava). Both genome sequences are highly fragmented, but 95% coverage of the genes is assumed based on comparisons with cDNA databases.

Table 1. Major crops for food, feed and non-food uses

	Crop name	Amount produced[a]	Order/Family[b]	Genome size (Mb)	Sequencing status
Food crops					
Cereals	Maize[c]	791.7	Poales/Poaceae	2500	Yes
	Rice (paddy)	659.5	Poales/Poaceae	389	Yes
	Wheat	605.9	Poales/Poaceae	17 000	Project
	Barley	133.4	Poales/Poaceae	5000	Project
	Sorghum	63.3	Poales/Poaceae	736	Yes
	Millet[c]	33.9	Poales/Poaceae	515	Project
	Oats	24.8	Poales/Poaceae	13 000	No
	Rye	14.7	Poales/Poaceae	8000	No
Root crops	Potatoes	309.3	Solanales/Solanaceae	840	Project
	Cassava[c]	214.5	Malpighiales/Euphorbiaceae	770	project
	Sweet potatoes	107.6	Solanales/Convovulaceae	1500	No
Vegetables and melon	Tomatoes	129.9	Solanales/Solanaceae	950	Project
	Watermelons/melons	97.4	Cucurbitales/Cucurbitaceae	480	Project
	Cabbages and other brassicas	68.9	Brassicales/Brassicaeae	530	Project
	Onions, dry	66.0	Asparagales/Alliaceae	16 000	No
	Cucumbers and gherkins	44.2	Cucurbitales/Cucurbitaceae	367	Yes
	Eggplants (aubergines)	32.2	Solanales/Solanaceae	1100	No
	Carrots and turnips	27.2	Apiales/Apiaceae	473	No
	Lettuce and chicory	23.3	Asterales/Asteraceae	2500	No
	Pumpkins, squash and gourds	21.0	Cucurbitales/Cucrbitaceae	500	No
	Cauliflowers and broccoli	17.7	Brassicales/Brassicaeae	880	No
	Garlic	15.8	Asparagales/Alliaceae	11 100	No
	Spinach	14.0	Caryophyllales/Amaranthaceae	990	No
Oil crops	Soybeans[c]	220.5	Fabales/Fabaceae	1115	Yes
	Oil palm[c]	192.6	Arecales/Arecaceae	1800	Yes
	Seed cotton	73.6	Malvales/Malvaceae	2250	No
	Coconuts	61.5	Arecales/Arecaceae	2150	No
	Rapeseed[c]	50.6	Brassicales/Brassicaeae	600	Yes/project
	Groundnuts, with shell	37.1	Diverse order/families	800	No
	Sunflower seed[c]	26.8	Asterales/Asteraceae	3000	Project
	Olives	17.4	Lamiales/Oleaceae	1500	No
Fruits	Bananas/plantains	119.8	Zingiberales/Musaceae	600	Project
	Oranges/clementines/lemons	104.3	Sapindales/Rutaceae	367	Project
	Grapes	67.2	Vitales/Vitaceae	500	Yes
	Apples	66.0	Rosales/Rosaceae	750	Yes
	Mangoes, mangosteens and guavas	33.4	Diverse order/families	400	No
	Pineapples	20.9	Poales/Bromeliaceae	440	No
	Pears	20.6	Rosales/Rosaceae	500	No
	Peaches and nectarines	17.4	Rosales/Rosaceae	250	Yes
	Papayas	7.2	Brassicales/Caricaceae	372	Yes
Non-food crops					
Energy crops[c]	Poplar	NA	Malpighiales/Salicaceae	485	Yes
	Eucalyptus	NA	Myrtales/Myrtaceae	600	Project
	Jatropha	NA	Malpighiales/Euphorbiaceae	400	Yes
	Switchgrass	NA	Poales/Poaceae	480	Project
	Castor bean	NA	Malpighiales/Euphorbiaceae	400	Yes
	Miscanthus	NA	Poales/Poaceae	3300	Project
	Sugarcane	NA	Poales/Poaceae	2300	Project

Abbreviation: NA, not available.

[a] Crops are sorted according to their relative importance in terms of production in 2007 (million tons yr^{-1}; source: http://faostat.fao.org/).

[b] The order and family names are from the Angiosperm Phylogeny website (http://www.mobot.org/MOBOT/research/APWeb/welcome.html).

[c] Species used as both food and energy crops.

Specific considerations and challenges of plant genome sequencing

Plant genomes have several features that present challenges for the elucidation of their sequences, with size and complexity representing the most significant difficulties. The average representative size of a plant genome is ~6 Gb [4], which is an order of magnitude larger than the average size of genomes sequenced so far (Table 2). Two factors underpin the large variation in plant genome size: the amount of repetitive and transposable element DNA [from 10% in *Arabidopsis* [1] to >80% in wheat (*Triticum aestivum*)] and the level of ploidy (from diploid to octaploid and higher).

Owing to their repetitive nature, transposable elements complicate genome assembly, particularly when short-read technologies are used [17]. The two largest and most repetitive sequenced plant genomes to date [of maize (2.5 Gb) and soybean (1.1 Gb)], are incomplete, partly because of the difficulty and cost associated with sequencing and assembling the large amount (>60%) of repetitive and transposable element DNA. The strategy chosen by the Maize Genome Sequencing Consortium to circumvent this

Table 2. Overview of higher plant genome sequencing projects for which sequences are publicly available

Species and genotype	Genome size (Mb)	Ploidy	Sequencing strategy	Coverage	Refs
Finished genome sequences[a]					
Arabidopsis thaliana (thale cress, cv. Columbia)	125	Diploid	BAC-by-BAC	~15×	[1]
Oryza sativa ssp. *japonica* (rice, cv. Nipponbare)	389	Diploid	BAC-by-BAC	10×	[38]
Improved high-quality draft genome sequences					
Sorghum bicolor (sorghum, cv. BTx623)	770	Diploid	WGS	8.5×	[72]
Vitis vinifera (grapevine, Pinot noir cv. PN40024)	487	Dihaploid	WGS	8.4×	[22]
Brachypodium distachyon (purple false brome, line Bd21)	300	Diploid	WGS	8×	[86]
Glycine max (soybean, cv. Williams 82)	1100	Diploid	WGS	8×	[6]
Populus trichocarpa (black cottonwood, poplar, cv. Nisqually-1)	485	Diploid	WGS	7.5×	[25]
Zea mays ssp. *mays* (maize, cv. B73) (gene space)	2600	Diploid	BAC-by-BAC	6×	[5] http://maizesequence.org/
Cucumis sativus (cucumber, IL 9930)	367	Diploid	WGS	72.2× (NGS)	[16]
High-quality draft genome sequences					
Oryza sativa ssp. *japonica* (rice, cv. Nipponbare)	433	Diploid	WGS	6×	[87]
Oryza sativa ssp. *indica* (rice, cv. 93-11)	466	Diploid	WGS	6.3	[40]
Oryza sativa ssp. *japonica* (rice, cv. Nipponbare)	399	Diploid	BAC-by-BAC	5×	[39]
Vitis vinifera (grapevine, Pinot noir cv. ENTAV 115)	505	Diploid	WGS and SBS	6.5/4.2×	[23]
Carrica papaya (transgenic papaya, 'Scv.unUp')	372	Diploid	WGS	3×	[88]
Prunus persica (Peach, cv. Lovell)	220	Dihaploid	WGS	7.7×	http://www.rosaceae.org/peach/genome
Malus x domestica Borkh (Apple, cv. Golden delicious)	742	Dihaploid	WGS	16.9x	[89]
Standard draft genome sequences					
Medicago truncatula (barrel medic, cv. Jemalong A17)	500	Diploid	BAC-by-BAC	ND[c]	http://www.medicago.org/
Zea mays (popcorn, cv. Palomero Toluqueno)	2100	Diploid	WGS	3.2×	[90]
Lotus japonicus (trefoil, cv. Miyakojima MG-20)	472	Diploid	BAC-by-BAC	8.4×	http://www.kazusa.or.jp/lotus/
Mimulus guttatus (common monkeyflower)	430	Diploid	WGS	7×	http://www.phytozome.net/mimulus
Ricinus communis (castor bean, cv. Hale)	400	Diploid	WGS	4×	[91]
Solanum tuberosum (potato, DM1-3-516 R44)	840	Dihaploid	WGS	70× (NGS)	http://www.potatogenome.net
Manihot esculenta (cassava, cv AM560-2) (gene space 416 Mb)	770	Amphiploid	WGS	ND	http://www.phytozome.net/cassava
Solanum lycopersicum (common tomato, cv esculentum x pennellii)	950	Diploid	BAC-by-BAC		http://sgn.cornell.edu/about/tomato_project_ overview.pl
			WGS	22 x	http://mips.helmholtz-muenchen.de/plant/tomato/index.jsp
Arabidopsis lyrata (rock cress)	230	Diploid	WGS	ND	http://www.phytozome.net/alyrata

Abbreviations: ND, no data; SBS, sequencing by synthesis.

[a]The sequences have been classified according to the criteria proposed by Chain *et al.* [8].

problem was to focus on a BAC-by-BAC sequencing strategy. The genome was sequenced using a minimum tiling path of 16 848 genetically and physically anchored BACs that were shotgun sequenced (four- to sixfold coverage) and assembled [5]. Only the non-repetitive, low-copy regions of BACs were improved using automated and manual sequencing approaches. Anchoring of the maize reference genome to a high-resolution gene map [18] enabled the positioning of most BAC contigs along chromosomes, but, as a consequence of the sequencing strategy, the order and orientation of many contigs within a given BAC remain unknown. Approximately 7% of the genome (160 Mb) is missing from the draft B73 RefGen_v1 sequence and, because the BAC-based physical map is missing ~6% of the genes, it is unlikely that the filtered gene set is complete. Approximately 85% (950 Mb) of the soybean genome was assembled from WGS sequence data [6]. All but 20 of the 397 sequence scaffolds that were assembled into 20 chromosome-level pseudomolecules have been oriented unambiguously on the chromosomes by genetic mapping; whereas ~18 Mb of mostly repetitive scaffold sequences that contain ~450 predicted genes remain unanchored.

Polyploidy, a general feature of plant genomes [19], adds to the difficulties in sequencing and assembling some of the largest genomes. Although it is not a problem for most plant species, given that the polyploidy nature is ancient and dates back to >10 million years, it remains crucial for many important crops that are true polyploids; for example, bananas (*Musa* spp.), potato (*Solanum tuberosum*), cotton (*Gossypium hirsutum*), wheat and sugarcane (*Saccharum* ssp.). The redundancy created by the presence of two or more sets of genes within a nucleus can affect the accuracy of genome sequence assembly and the need to differentiate between homoeologs could influence the ultimate utility of the sequence. Consequently, to date, no polyploid plant species has been sequenced. Rapeseed (*Brassica napus*, 2n = 4x = 38, AACC genome) and wheat (*Triticum aestivum*, 2n = 6x = 42, AABBDD genome) are major crops and both represent typical examples of recent allopolyploids. The rapeseed genome was recently sequenced by a consortium driven by a private company (Table 3) but the sequence is not yet publicly available. The described strategy was to combine sequences from the *Brassica rapa* (AA genome) and *Brassica oleracea* (CC

Table 3. Overview of higher plant genome sequencing projects that are either underway or not yet publicly available

Species and genotype	Genome size (Mb)	Ploidy	Sequencing strategy	Refs
Aquilegia formosa (Western columbine)	400	Diploid	WGS	http://www.jgi.doe.gov/genome-projects/
Brassica oleracea	600	Diploid	WGS	http://www.genomesonline.org
Brassica rapa ssp. *pekinensis* (Chinese cabbage, cv. Chifu 401-42)	529	Diploid	BAC-by-BAC	http://www.brassica.info/resource/sequencing.php
Brassica rapa ssp. *pekinensis* (Chinese cabbage, cv. Chifu 401-42)	492	Diploid	WGS	http://www.intl-pag.org/18/abstracts/W14_PAGXVIII_104.html
Brassica rapa (B3)	530	Diploid	WGS	http://www.jgi.doe.gov/genome-projects/pages/projects.jsf?kingdom5Plant
Brassica napus (rapeseed) (*oleracea*/*rapa*)	1100	Tetraploid	WGS	http://www.bayercropscience.com/bcsweb/cropprotection.nsf/id/EN_20091009?open&l5EN&ccm5500020
Capsella rubella (pink shepherds purse)	250	Diploid	WGS	http://www.jgi.doe.gov/genome-projects/
Citrus sinensis (sweet orange, cv. Ridge pineapple)	382	Diploid	WGS	http://www.citrusgenome.ucr.edu/
Elaeis guineensis (oil palm, cv. *tenera* × *dura*)	1800	Diploid	BAC pools and WGS	http://www.syntheticgenomics.com/media/press/52108.html
Eucalyptus grandis (BRASUZ1)	600	Diploid	WGS	http://www.jgi.doe.gov/genome-projects/
Gossypium raimondii (cotton)	880	Diploid	WGS	http://www.jgi.doe.gov/genome-projects/
Jatropha curcas (synthetic genomics)	400	Diploid	Unpublished	http://www.syntheticgenomics.com/media/press/52009.html
Nicotiana tabacum (tobacco, cv. Hicks broadleaf)	4500	Tetraploid	Methyl filtration	http://www.tobaccogenome.org/
Setaria italica (foxtail millet, Yugu1)	515	Diploid	WGS	http://www.jgi.doe.gov/genome-projects/
Solanum tuberosum (potato, RH89-039-16)	840	Diploid	BAC-by-BAC	http://www.potatogenome.net [92]
Vigna unguiculata (cowpea)	620	Diploid	Methyl filtration	[93]
Zea mays (maize, cv. Mo17)	2100	Diploid	WGS	http://www.maizegdb.org/sequencing_project.php

genome) with *B. napus* (http://www.bayercropscience.com/bcsweb/cropprotection.nsf/id/8AAD8BA3537C1DDB-C125764A002DE77C). In the absence of publicly available sequence and community curation, it is impossible to assess the completeness and quality of the sequence or to discuss the advantages and disadvantages of using such an approach. It would be interesting to compare it with the results obtained by the public Multinational Brassica Genome Project, which has started multiple efforts to sequence the Brassica A, C and AC genomes (http://www.brassica.info/resource/sequencing.php). For wheat, the International Wheat Genome Sequencing Consortium (http://www.wheatgenome.org) has established a road map for sequencing the hexaploid bread wheat (*T. aestivum*) genome rather than the individual, ancestral diploid genomes because: (i) it is the species grown on 95% of the wheat-growing areas; and (ii) the ABD genome does not correspond structurally and functionally to the sum of the ancestral A (*Triticum urartu*), B (unknown species that are likely to be related to *Aegilops speltoides*) and D (*Aegilops tauschii*) genomes. To overcome the difficulties associated with sequencing the hexaploid wheat genome, the consortium is following a chromosome-based strategy [20] to construct physical BAC clone maps and subsequently to sequence each of the 21 individual chromosomes. The first physical map of the largest wheat chromosome, 3B (1 Gb), was produced recently [21] and its sequencing using NGS technologies is underway (http://urgi.versailles.inra.fr/index.php/urgi/Projects/3BSeq).

Another form of redundancy, heterozygosity, is widespread among flowering plants and the commercial varieties of many economically important crops are highly heterozygous. This has led many in the genome sequencing community to select homozygous derivatives for sequencing, rather than commercially important heterozygous varieties, following the rationale that an assembled genome sequence will then form a template onto which the variation observed in heterozygotes can be mapped. This was the rationale used by a public–private French–Italian Consortium to select the haplodiploid grape cultivar Pinot noir cv. PN40024 [22] for sequencing over the heterozygous cv ENTAV 115 chosen for sequencing by an Italian public–private initiative [23,24].

The black cottonwood tree (*Populus trichocarpa*) is the only other example of a highly heterozygous plant species that has been sequenced to date. It was selected by the US Department of Energy Joint Genome Institute (JGI) as a model for sequencing forest species because of its small genome size (~485 Mb). A single genotype, Nisqually-1, was sequenced to 7.5× coverage using a WGS strategy (Table 2) [25]. Although a BAC-based physical map was used to guide the sequence assembly, only the euchromatic region of the genome could be assembled, leaving a substantial portion (~75 Mb, 15%) of the WGS reads unassembled in the first draft [25,26]. Although these projects provided insights into the methods for assembling heterozygous genomes [24], their relative incompleteness illustrates the difficulty of producing a high-quality reference genome sequence for a heterozygous cultivar with current sequencing technologies.

What quality of sequence is required for different studies?

There are lively discussions among plant scientists and within species-specific research communities about the value of having finished genome sequences as opposed to draft sequences. This relates specifically to the question of what can be gained from a more in-depth, time-consuming and costly effort to generate high-quality complete sequences? There are also widespread debates about the

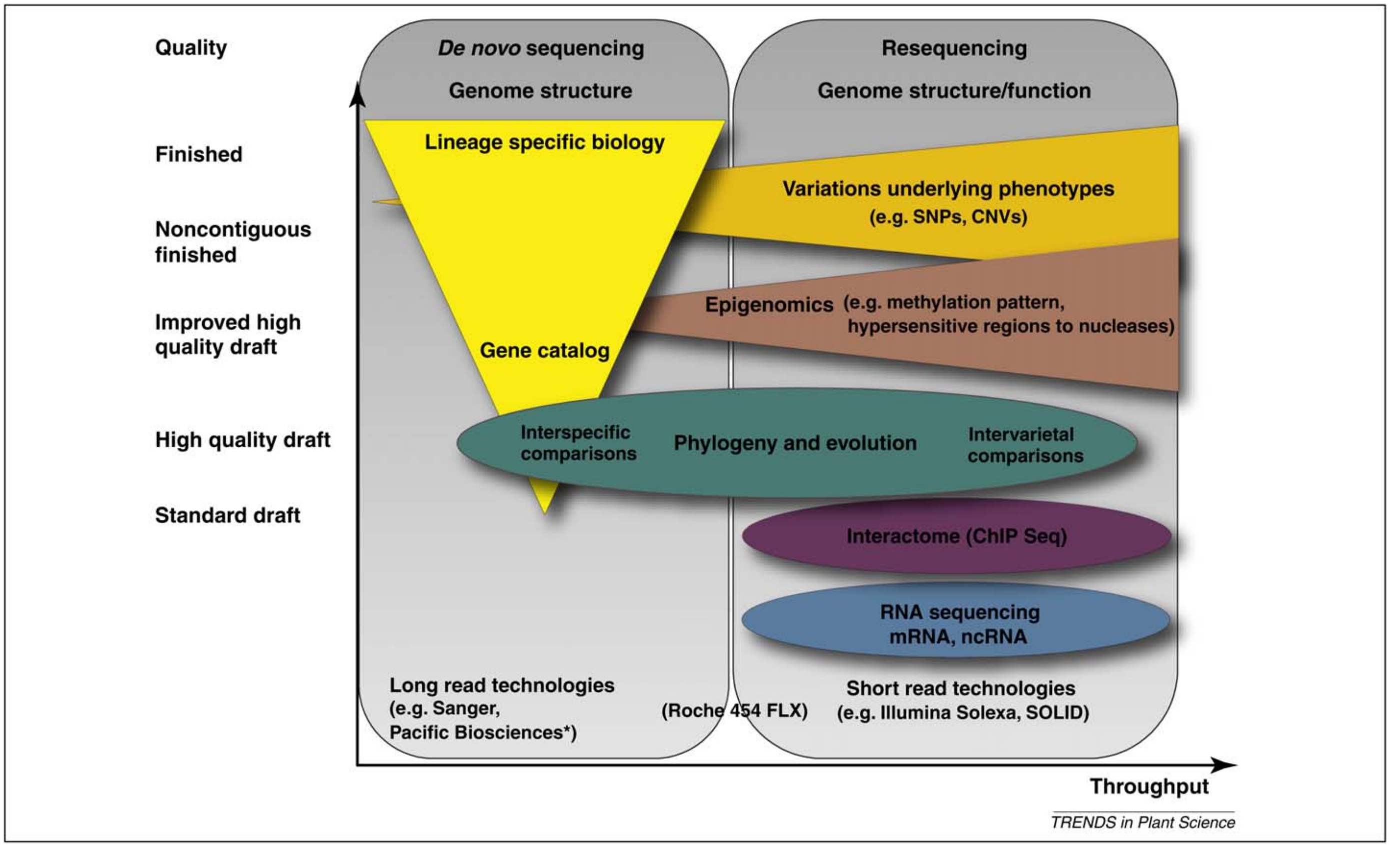

Figure 1. Different applications for genome sequencing in relationship to sequence completion, quality and throughput of different sequencing technologies available in 2010 (Sanger, Roche 454, Illumina Solexa GA and ABI SOLID) and announced but not yet used for plant genome sequencing (e.g. Pacific Biosciences). Abbreviations: ChIP Seq, chromatin immunoprecipitation sequencing; mRNA, messenger RNA; nc RNA, non-coding RNA.

levels of biological information that can be accessed from different qualities of sequence and the information that will be useful for research studies and applications (Figure 1). Low coverage or standard draft genome sequences are suitable for establishing a relatively comprehensive catalog of the gene and repeat content. Furthermore, they provide insights into evolutionary mechanisms and synteny through comparative genomics (Figure 1) and phylogenetic shadowing, as demonstrated in mammals [13,27] and proposed in plants [28]. However, as demonstrated with animal projects, it is increasingly clear that inferences about lineage-specific features can only be achieved with high-quality complete sequences (Figure 1). Many of the animal genomes that were sequenced originally to a depth of twofold added additional sequence because the low-coverage sequences were inadequate for investigators interested in the biology of the target species [29]. One of the key limitations with low coverage or draft sequences is the difficulty in capturing accurately the information embedded in the repetitive fraction of the genome, even though it is now apparent that the repetitive sequences and transposable elements have key roles in evolutionary changes and regulatory innovation. This is of crucial concern for plant genomes that are primarily composed of repetitive elements.

A short coming of draft sequences is readily apparent when analyses requiring complete features need to be undertaken. For example, it is difficult to distinguish genes from pseudogenes with such incomplete information and recent segmental duplications generally collapse into the draft assemblies. Recently, Church *et al.* [30] demonstrated that the finished mouse genome sequence provided additional information about the rodent-specific biology and the delineation of ancestral biological functions that are shared with humans, thereby also providing additional understanding of human biology. By comparing the WGS approach and the finished clone-based sequence, the authors discovered that ~5% of the genome (267 Mb) sequence was misassembled or missing in the WGS, including complete regions that corresponded to recent segmental duplications harboring lineage-specific gene families and interspersed repeats. Supported by similar results from the dog and macaque rhesus genome projects, the authors concluded that WGS assemblies will always poorly reflect lineage-specific biology. Species-specific architectural features that, in some cases, could be linked to important functions and/or represented in variable copy number, often prove to be the most difficult regions to sequence to high quality and are frequently misrepresented or missing when comparative genome analyses are based upon genome sequences of varying qualities [31]. As genome sequences are used increasingly for biological research, it is becoming clear that high-quality sequence data are essential for the full annotation and functional analysis of genomes.

Recently, using high-quality sequence data, Knowles and McLysaght [32] provided the first evidence in the human genome of *de novo* genes (orphan genes with no

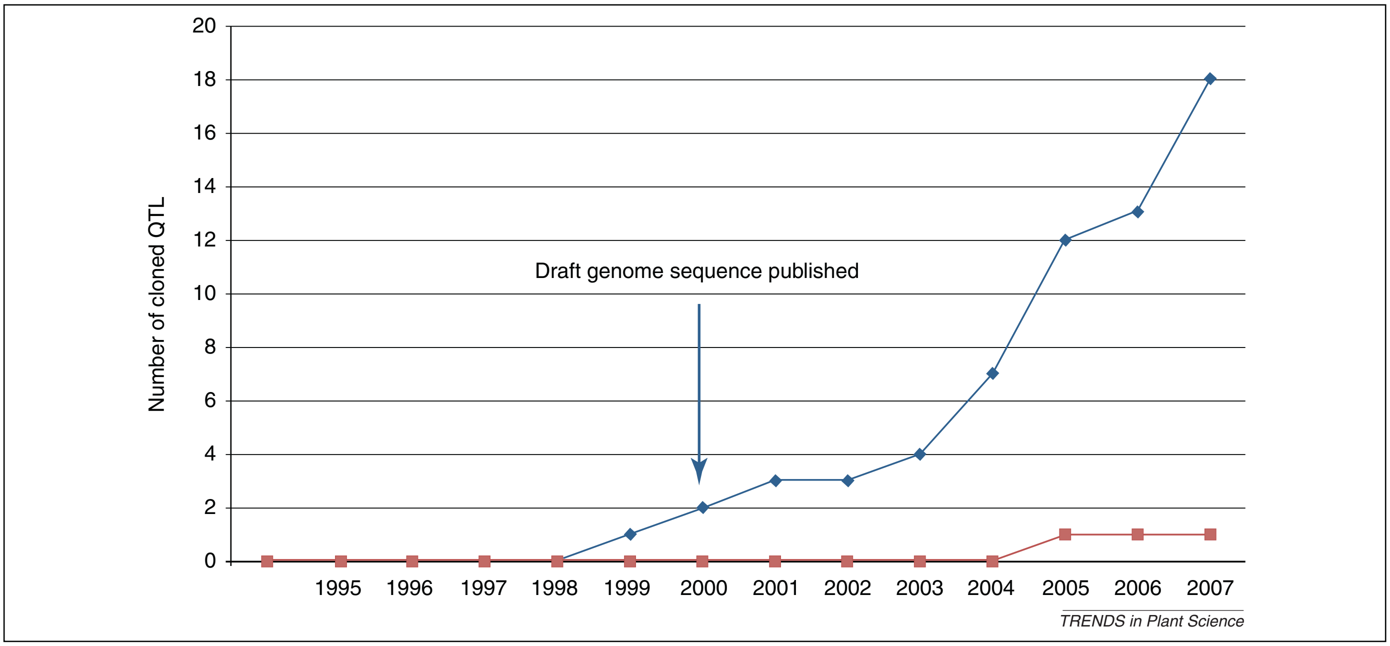

Figure 2. Number of QTL cloned in rice (blue) and wheat (red) since 1995. The blue arrow indicates the year in which the rice genome sequence became available and spurred the number of cloned genes and QTL (published source: NCBI and [42,61]). The Y axis represents the number of cloned QTL.

protein-coding homologs in other species). Furthermore, data emerging from the ENCODE and modENCODE consortia (that are exploring regulatory features and functional regulation of human and model organism genomes, respectively) illustrate the importance of high-quality, contiguous reference sequences to predict and interpret regulatory signatures accurately [33,34]. Finally, projects for the systematic generation of conditional knock-outs of every mouse gene (http://www.knockoutmouse.org/) are dependent upon accurately annotated gene structures in finished sequence data.

With the wide-ranging technologies and strategies in use today, how can one ensure that the products of genome sequencing projects meet the needs of researchers using them for downstream applications? From the beginning, the *Arabidopsis* genome initiative adopted the position that a ‘complete’ sequence was required and that this would be achieved when each of the ten chromosome arms is represented in a single contig. Each contig should end in telomeric or nuclear repeats at one end and centromeric satellite repeats at the other [35]. However, the first projects did not include random sequencing datasets and only relied on BAC-by-BAC sequencing. Completeness was often overestimated, as indicated by the comparison of WGS sequences with the BAC-by-BAC finished genome sequence of *A. thaliana* in 2005 [36]. This underlies the importance of developing combined strategies of WGS and BAC-by-BAC to increase the probability of capturing the maximum information [13]. The rice genome sequencing efforts have also provided insights into how different strategies can contribute to the completeness and quality of a finished genome sequence. The International Rice Genome Sequencing Project (IRGSP) initiated the sequencing of *Oryza sativa* Nipponbare (ssp. *japonica*) using a hierarchical clone-by-clone strategy in 1998 (summarized in [37]). The finished sequence covered 95% of the 389-Mb rice genome and was published in 2005 [38]. While the IRGSP effort was underway, three other groups, Monsanto [39], the Beijing Genome Institute (BGI) [40] and Syngenta [41], produced draft sequences of rice. Monsanto produced a draft of ~259 Mb of Nipponbare using a BAC sequencing approach, whereas the BGI and Syngenta used whole-genome strategies to generate the sequences of two genomes, *indica* line 93–11 and *japonica* Nipponbare, respectively. The WGS assemblies provided an overview of genome structure, but this approach resulted in considerable misassembly because of the difficulties of correctly positioning repeats in the genome assembly. However, information from the draft sequences was integrated ultimately into the IRGSP effort and contributed to the highly accurate, ‘gold standard’ map-based sequence published by the IRGSP [38]. Moreover, the BAC-by-BAC approach provided the ability to link directly the sequence to the maps that carry genetic and phenotypic information, thereby enabling efficient isolation of agronomically relevant genes. Indeed, the release of the high-quality rice genome sequence performed by the IRGSP has accelerated cloning of, for example, quantitative trait loci (QTLs) for flowering time, disease resistance, plant architecture and abiotic stress (salt and submergence) [42], as illustrated in Figure 2.

To date, complete and accurate sequence assemblies have only been achieved through a costly, time-consuming ‘finishing effort’ that involves additional directed sequencing of existing subclones or from purified BAC DNA or from amplified PCR products. With the exponential increase in whole-genome sequencing projects triggered by NGS technologies in which clones are no longer used for sequence generation, this ‘finishing effort’ is generally avoided, resulting in an increasing number of draft sequences [8]. The closure of bacterial genome sequences and other small genomes from WGS assemblies has been undertaken by combinatorial PCR from genomic DNA, but this approach would be costly and time consuming for large genomes, such as those of many crops, and is likely to be only moderately successful if the genome is repetitive. Directed sequencing of regions is possible if clone end sequences are incorporated into genome assemblies to provide scaffolding information for contiguation, enabling the clones to be sequenced using NGS or Sanger methods and the data incorporated into subsequent genome assemblies. In the longer term, the issues of long-range assembly and high accuracy might become easier to address with the development of technologies (so-called 'third-generation' sequencing technology) that are able to generate long sequence reads at a cost that enables high levels of redundancy to be achieved.

To assess what can be done with a particular genome sequence, it is becoming important to be able to compare and evaluate the quality of the genome sequence assemblies. Recently, a group of internationally recognized sequencing centers proposed six quality levels to describe genome sequences: (i) standard draft; (ii) high-quality draft; (iii) improved high-quality draft; (iv) annotation-directed improvement; (v) noncontiguous finished; and (vi) finished [8]. We applied these criteria to the current plant genome sequences (Table 2) and propose the systematic use of these standards in each of the published genome sequences, a policy that should be reinforced by journals and funding agencies. The value of higher quality genome sequences should be emphasized and strategies developed for cost-effective sequence improvement based on the use of NGS technologies.

Applications of the plant genomes sequenced to date: breeding and scientific perspectives

Translational biology: from models to crops

Global research on *A. thaliana* triggered by the sequencing of its genome ten years ago revolutionized understanding of plant biology by unraveling basic mechanisms in plant development, tolerance to abiotic and biotic stresses and adaptation. Given that many of these basic pathways are common to all plants, *Arabidopsis* genes can be used either directly in heterologous systems or as candidate genes for identifying orthologs in crops. This is particularly true for basic developmental processes and abiotic stress tolerance. The most successful translations of a gene from *A. thaliana* to improve a trait in crops were achieved with genes involved in abiotic stress tolerance and, primarily with transcription factors, because of their central role in controlling cellular processes. For example, CRT binding factors (CBFs) from *Arabidopsis* were expressed in tomato (*Solanum lycopersicum*), rapeseed, strawberry (*Fragaria* ssp.), rice and wheat, providing evidence of improved freezing, salt and drought tolerance [43,44]. Generally, such translational biology is more complex and inefficient for disease resistance, partly because of the two resistance mechanisms [pathogen-associated molecular pattern-triggered immunity (PTI) and effector-triggered immunity (ETI)] that are superimposed on each other in plants. ETI corresponds to classic race-specific disease resistance that has evolved rapidly and more specifically in each plant species, thereby making it a difficult subject for direct transfer from model to crops. By contrast, the PTI translation affords more opportunities, as was demonstrated recently with the successful engineering of broad-spectrum disease resistance in tomato and tobacco (*Nicotiana tabacum*) by the expression of an *Arabidopsis* elongation factor Tu receptor (EFR) [45].

There have been several examples in which candidate genes underpinning basic traits have been identified in crops by screening with sequences of functional or structural orthologs for closely related or model species, such as *Arabidopsis*. For example, Nelson *et al.* [46] obtained drought tolerance after transforming maize with a maize ortholog (*ZmNF-YB2*) of the *Arabidopsis* transcription factor *AtNF-YB1* that was identified through a screen for drought tolerance in *Arabidopsis*. However, even if orthologs can be identified, the candidate gene approach suffers from limitations because one-way translational biology cannot explain all the differences in species. *In silico* comparisons of genome sequences offer advantages of speed and cost for gene identification. This was exemplified when, as soon as the rice genome was completed, comparative analyses with the *Arabidopsis* genome accelerated discoveries in rice biology [47]. Thus, knowledge of biological processes provided by model plants can be applied most efficiently when one also has access to the crop genome sequence.

In addition, direct comparisons of a crop genome sequence with other genomes, model or other crops, are useful for identifying species-specific differences that can underlie essential traits and can be as important, if not more so, than the conserved elements. For example, map-based cloning of flowering-time genes in rice was instrumental in determining that *Arabidopsis* and rice share common regulatory pathways but functional analyses demonstrated differential regulation that results from reverse function of a key central regulator [48]. These findings have refined understanding of what triggers the differences between short-day and long-day plants. The cloning and characterization of mode of action for the vernalization genes *Vrn1*, *Vrn2* and *Vrn3* in wheat [49–51] offer similar examples of the power of comparative genomics.

Finally, model organisms, although useful for understanding basic biological mechanisms, cannot reflect all the specificities and complexities of adaptation mechanisms of the other species. Genotype × environment interactions are important for plants and their regulation from one species to another could be different from that of model species. Moreover, crop species result from decades of domestication and selection for specific traits related to food and feed end usage (e.g. shattering, bread making or oil quality) that have specifically affected genes and networks. Furthermore, several economically important plant species are polyploid, thereby resulting in more complex regulation between homoeologous genes and possibly obscuring the orthologous relationships between models and polyploid crop genomes. Thus, to understand complex adaptive and agronomically relevant traits in the most important food, feed and fiber crops, the full breadth of genomic resources must be developed for those crops (Table 1).

Applications of sequences from crops to crop improvement

As the first crop genome sequenced, rice provides an excellent opportunity to illustrate the impact on plant biology and breeding of having access to a finished genome sequence for a species of major socio-economic importance. Using existing genetic and genomic resources and tools (mutants, genetic populations and transformation techniques [52,53]), rice researchers were rapidly able to integrate and apply genome sequence information to understand rice genome structure and evolution as well as to discover and mine genes, including those underlying complex traits of agricultural importance (reviewed in [42,54–56]). Members of large gene families [such as transcription factors, peptide transporters, kinases, nucleotide binding leucine-rich repeats (NB-LRRs), microRNAs and germins] have been discovered through genome-wide surveys, enabling their cellular functions to be dissected and their roles in plant growth and development to be elucidated (e.g. 151 members of the rice NAC transcription family [57,58]). In another example, genome sequence-enabled identification and positional cloning of genes responsible for traits selected during domestication, including the seed-shattering trait, led to the identification of molecular changes selected during domestication [55,59,60].

Perhaps the most anticipated outcome of the rice genome sequence was the promise of high-throughput development of molecular markers to assist genetic analysis, gene discovery and breeding programs [61]. Indeed, rice is now rich in tools for mapping and breeding, including high-density simple sequence repeats (SSRs) (~51 SSR Mb^{-1}), comprehensive single nucleotide polymorphisms (SNPs), insertion–deletion polymorphisms (IDPs) and custom-designed (candidate gene) markers for marker-assisted breeding. One of the first outcomes of genome sequence-based marker development has been the increase in the map-based cloning of QTL underlying agronomically important traits in rice (Figure 2) [42]. Projects that generally take eight to ten years in the absence of a genome sequence (e.g. in wheat; Figure 2) can now be completed in only a couple of years and require less effort in rice. A second outcome of having access to the rice genome sequence has been the development of a comprehensive collection of SNPs. Early on, a genome-wide rice DNA polymorphism survey of the genomes of Nipponbare and 93–11 revealed 1 703 176 SNPs, with approximately one SNP every 268 bp [62]. Subsequent genome-wide SNP discovery projects used various strategies and the advent of NGS to resequence diverse rice varieties and compare them to the reference genomes [63,64]. Analysis of shared SNPs among 20 diverse land races and varieties revealed chromosomal segments that were introgressed from one varietal group (e.g. *japonica*) into another (e.g. *indica*), providing snapshots of the long and extensive breeding history of rice [63]. Some introgressions correlated with genomic regions responsible for traits transferred between varietal groups, including regions containing semi-dwarfing and salt-tolerance traits, whereas others represented candidates for additional events of potential significance for breeding. The application of this comprehensive SNP data for high-resolution genotyping and for understanding relationships among rice varieties was later validated using a subset of verified SNPs to genotype 395 diverse *O. sativa* accessions [65].

Sequence-enabled allele mining, the exploration and exploitation of naturally occurring allelic variation at candidate genes controlling key agronomic traits, is an important pre-breeding tool for rice improvement programs [56,66]. The approach can help trace the evolution of alleles, identify new useful haplotypes and guide the development of allele-specific markers for use in marker-assisted selection. This has proven useful in the identification and functional validation of candidate genes for complex traits governing disease resistance [54,67–69]. For example, Carillo *et al.* [67] compared sequences of alleles of an oxalate oxidase gene family that contributes to a disease-resistance QTL and discovered an indel in the promoter of one gene family member from the QTL donor that was lacking in susceptible varieties. Association and functional analyses of the oxalate oxidase gene demonstrated that presence of the indel was associated with QTL effectiveness and that the polymorphism created by the indel is a useful marker for the QTL. In another example, Fukuoka *et al.* [68] used sequence-based markers from the blast disease resistance gene *pi21* region to identify recombinants between *pi21* and another gene located 37 kb apart that confers poor eating quality. With this information, the authors were able, for the first time in decades, to select for rice varieties that combined durable resistance and good eating quality [61]. The above examples are but a few of the many describing progress built on the foundation of a high-quality, well-annotated reference rice genome. Fortunately, the rice research community has received resources to continue improving the annotation [70,71], which will support systematic studies of rice gene function.

Following the completion of the rice genome sequence and the application of genomic tools, efforts have been undertaken to exploit the recently released draft sequences of other crop species, such as sorghum, maize and grapevine. As a first step towards application of the sequence into breeding programs, the sorghum genome sequence [72] was coupled with dense molecular marker maps that were previously impossible to link because of the lack of common markers across populations. This enabled, for the first time, the integration of major effect genes into a single map, thereby providing a foundation for breeders to link these easily recognizable landmarks to QTL studies and improve breeding programs [73].

SNP discovery by resequencing whole-genome or sub-genome representations is often one of the first uses of a reference genome sequence (Figure 1). For example, with knowledge from the B73 genome sequence features, Gore *et al.* [74] targeted the gene fraction of the maize genome for resequencing in the founder inbred lines of the Nested Associated Mapping (NAM) population [75]. Two datasets comprising 3.3 million SNPs were used to produce a first haplotype map (‘HapMap’) and to analyze the distribution of recombination and diversity along the maize chromosomes. This Maize HapMap and comparative genome hybridization (CGH) experiments enabled the identification of >100 low-diversity regions that are possibly associated with the domestication and geographic differentiation of maize. In grapevine, several initiatives are underway to resequence different cultivars for SNP discovery (e.g.

http://urgi.versailles.inra.fr/index.php/urgi/Projects/GrapeReSeq) to support the development of high-throughput genotyping platforms [76]. Finally, there is increasing interest from breeders to explore the causal effects of other types of polymorphism that are related to structural variations, such as copy number variation (CNV), IDPs and presence–absent variants (PAVs). Using the maize genome sequence, Springer *et al.* [77] revealed extensive structural variation, including hundreds of CNVs and thousands of PAVs, among maize lines. Many of the PAVs contain intact, expressed, single-copy genes that are present in one haplotype but absent from another, including hundreds of expressed genes potentially involved in heterosis [77]. Thus, access to the maize genome sequence enabled a fundamental phenomenon of plant biology to be unveiled for the first time, thereby leading the way to identifying the underlying genes or small RNAs with the prospect of improving heterosis in maize and other crops.

The first successes in sequencing rice and, subsequently, the first large and complex crop genomes have paved the way for sequencing other important crops. Several initiatives are now underway to tackle the wheat, barley (*Hordeum vulgare*), millet (*Setaria italica*), banana (*Musa acuminata*), clementine (*Citrus clementina*), tomato (*Solanum lycopersicum)*, sunflower (*Helianthus annuus*), eucalyptus (*Eucalyptus grandis*) and sugarcane genomes (Table 1) using NGS technologies. These research communities are eager to gain access to the tools and resources that have proven useful in addressing fundamental scientific questions and in supporting breeding for the first crop genomes sequenced.

New sequencing technologies: a revolution that will hold its promises?

Excellent recent reviews describe in detail the methods and potential applications of the first generation of DNA sequencing technologies to succeed Sanger dideoxy sequencing [14,15,78,79] and we will not detail those here. Recently, Varshney *et al.* [80] also reviewed the potential applications of NGS for genomics-assisted breeding in crops with and without a reference genome sequence. Among those, NGS offers the possibility of cost-efficient transcriptome and interactome profiling as well as in-depth analyses of epigenomics modifications (Figure 1). Here, we focus our discussion on the perspectives offered for obtaining *de novo* reference genome sequences. Although new sequencing technologies have expanded the possibility of obtaining such sequences, current read lengths (400 bp for Roche 454/FLX, 150 bp for Illumina Solexa GAIIe and 75 bp for ABI SOLID systems) limit the capacity to assemble sequences into long stretches representing the chromosomes [81]. Consequently, the trend is to generate draft as opposed to finished genome sequences [8] and, as stated above, this will limit the range of applications for the resulting genome sequences. Initially, the tendency was to develop a strategy in which a reference genome sequence is produced and then other cultivars are resequenced to identify the genetic variation associated with specific cultivars. However, comparative analyses between maize inbred lines have demonstrated that a single genome sequence does not necessarily reflect the entire genic complement of a species, leading to the concept of pan-genomes with a core genome shared between individuals and a variety-specific dispensable genome [77,82]. This promoted the idea that *de novo* sequencing of several individuals would be preferable to resequencing for the purpose of capturing the diversity present in a species, in particular for species with large genomes having a high proportion of transposable elements. For this, *de novo* sequencing has to be cost efficient and needs to provide high-quality sequence assemblies. NGS technologies are cost efficient, but the quality of the sequence assemblies that can be achieved currently is still generally lower than can be achieved with a ‘classic’ Sanger sequencing approach. To date, the shorter read length provided by some of the highest throughput sequencing technologies limit their accuracy in assembling highly repetitive genomes, which are common among crop genomes. Paired-end reads of >700 bp delivered by the Sanger sequencing platforms provide a template for sequence assembly that cover many smaller repeats and, because they can be generated from cloned fragments ranging in size from two to 150 kb, they can span the long repeats (>8 kb) that exist in many genomes. Methods to capture more structural information with the NGS technologies are being developed, but have proved difficult to implement efficiently. Thus, strategies are still needed to ensure that NGS technologies deliver *de novo* genome sequences that are equivalent or higher in quality to those produced with Sanger methodologies. Long read sequencing technology developments, such as the single molecule real-time DNA sequencing [83], hold promise to deliver the read length and/or structural information required for assembling complex genomes and regions. However, it is unclear whether this will deliver high-quality sequences and whether this can be achieved within the near future.

While waiting for these achievements, combinations of strategies are being explored. For example, in a combination of ‘old’ and ‘new’ technologies, a high-quality draft sequence was obtained for an entire *Oryza barthii* chromosome arm by sequencing multiple pools of BAC clones that represented a minimum tiling path using a combination of 454 FLX Pair-end reads and Titanium technologies [84]. Following a similar strategy combined with Illumina Solexa reads of sorted chromosomes, sequencing of the first wheat chromosome (3B) is underway (http://urgi.versailles.inra.fr/index.php/urgi/Projects/3BSeq).

Conclusion

A decade ago, technological limitations forced the plant biology community to select a few species as models. Although this strategy was necessary at the time, the subsequent continued focus on developing resources and high-quality genome sequences for only a few model plant species has limited the impact of translational biology on crop improvement. The new information coming out of human, animal, microbial and plant genome sequencing projects, coupled with rapid technological developments demand that researchers change how they approach crop genome sequencing projects. It is no longer necessary to consider a draft crop genome sequence as good enough simply because it delivers some of the tools required for short-term breeding objectives. Given the importance of crops and the immense challenges that agriculture faces over the next decade [85], the plant community should be inspired by the more productive vision developed for animal systems, where models are not promoted as replacements for human research, but rather as guides.

As a consequence of continued improvements in sequencing technologies, methods and bioinformatic capabilities, sequencing goals need no longer be limited because of high cost or complexity avoidance. The target should not be unimproved drafts that are 'good enough for the moment'; rather, researchers should imagine what will be possible with continued technological advancements and aim to provide access to all the high-quality crop genome information that biologists and breeders need. For large and complex genomes, this will require strategies that profit from the new sequencing technologies as soon as they become reasonably affordable while maintaining the possibility to add quality until the high-quality reference sequence is achieved. This implies that funding does not end when a first draft of a genome is released and that it is accepted by funders that a genome sequencing project can be deployed in several phases. The resulting crop genome sequencing revolution is likely to provide a paradigm shift in the approach to plant biology and crop breeding. Thus, the prospects for meeting future global food, fiber and fuel needs through crop improvement, even in the face of global climate change, are within reach if research and funding communities embrace the challenges of crop genomes sequencing.

Acknowledgements

Research in C.F.'s laboratory (Genetics, Diversity and Ecophysiology of Cereals, INRA Clermont-Ferrand) is supported by the ANR-GPLA06001G SMART and ANR- 09-GENM-025 grants from the Agence Nationale de la Recherche and France Agrimer and the FP7-212019 grant from the European Community's 7th Framework Program (FP7/ 2007-2013). Activities of K.E. are supported by the International Wheat Genome Sequencing Consortium (http://www.wheatgenome.org). J.R. is supported by the Biotechnology and Biological Sciences Research Council.

References

1 AGI (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815
2 Collins, F. (2010) Has the revolution arrived? *Nature* 464, 674–675
3 Gregory, T.R. (2005) The C-value enigma in plants and animals: a review of parallels and an appeal for partnership. *Ann. Bot.* 95, 133–146
4 Gregory, T.R. *et al.* (2007) Eukaryotic genome size databases. *Nucleic Acids Res.* 35, D332–D338
5 Schnable, P.S. *et al.* (2009) The B73 maize genome: complexity, diversity and dynamics. *Science* 326, 1112–1115
6 Schmutz, J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183
7 Eversole, K. *et al.* (2009) Wheat and barley genome sequencing. In *Genetics and Genomics of the Triticeae* (Feuillet, C. and Muehlbauer, G.J., eds), pp. 713–742, Springer
8 Chain, P.S. *et al.* (2009) Genomics. Genome project standards in a new era of sequencing. *Science* 326, 236–237
9 Fleischmann, R.D. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512
10 Myers, E.W. *et al.* (2000) A whole-genome assembly of *Drosophila*. *Science* 287, 2196–2204
11 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
12 Consortium, M.G.S. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562
13 Green, E.D. (2001) Strategies for the systematic sequencing of complex genomes. *Nat. Rev. Genet.* 2, 573–583
14 Metzker, M.L. (2009) Sequencing technologies – the next generation. *Nat. Rev. Genet.* 11, 31–46
15 Mardis, E.R. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* 9, 387–402
16 Huang, S. *et al.* (2009) The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* 41, 1275–1281
17 Schatz, M.C. *et al.* (2010) Assembly of large genomes using second-generation sequencing. *Genome Res.* 20, 1165–1173
18 Wei, F. *et al.* (2009) The physical and genetic framework of the maize B73 genome. *PLoS Genet.* 5, e1000715
19 Hancock, J.F. (2004) *Plant Evolution and the Origin of Crop Species*, CABI Publishing
20 Dolezel, J. *et al.* (2007) Chromosome-based genomics in the cereals. *Chromosome Res.* 15, 51–66
21 Paux, E. *et al.* (2008) A physical map of the 1-gigabase bread wheat chromosome 3B. *Science* 322, 101–104
22 Jaillon, O. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467
23 Velasco, R. *et al.* (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS One* 2, e1326
24 Zharkikh, A. *et al.* (2008) Sequencing and assembly of highly heterozygous genome of *Vitis vinifera* L. cv Pinot Noir: Problems and solutions. *J. Biotechnol.* 136, 38–43
25 Tuskan, G.A. *et al.* (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313, 1596–1604
26 Kelleher, C.T. *et al.* (2007) A physical map of the highly heterozygous *Populus* genome: integration with the genome sequence and genetic map and analysis of haplotype variation. *Plant J.* 50, 1063–1078
27 Blakesley, R.W. *et al.* (2004) An intermediate grade of finished genomic sequence suitable for comparative analysis. *Genome Res.* 14, 2235–2244
28 Paterson, A.H. (2006) Leafing through the genomes of our major crop plants: strategies for capturing unique information. *Nat. Rev. Genet.* 7, 174–184
29 Green, P. (2007) 2x genomes: does depth matter? *Genome Res.* 17, 1547–1549
30 Church, D.M. *et al.* (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol.* 7, e1000112
31 Blakesley, R. *et al.* (2010) Effort required to finish shotgun-generated genome sequences differs significantly among vertebrates. *BMC Genomics* 11, 21
32 Knowles, D.G. and McLysaght, A. (2009) Recent *de novo* origin of human protein-coding genes. *Genome Res.* 19, 1752–1759
33 Birney, E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816
34 Celniker, S.E. *et al.* (2009) Unlocking the secrets of the genome. *Nature* 459, 927–930
35 Martienssen, R.A. (2000) Weeding out the genes: the *Arabidopsis* genome project. *Funct. Integr. Genomics* 1, 2–11
36 Liu, R. and Bennetzen, J.L. (2008) Enchilada redux: how complete is your genome sequence? *New Phytol.* 179, 249–250
37 Matsumoto, T. *et al.* (2008) Development in rice genome research. based on accurate genome sequence. *Int. J. Plant Genomics* 2008, 348621
38 International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* 436, 793-800.
39 Barry, G.F. (2001) The use of the Monsanto draft rice genome sequence in research. *Plant Physiol.* 125, 1164–1165
40 Yu, J. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp indica). *Science* 296, 79–92
41 Goff, S.A. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp japonica). *Science* 296, 92–100
42 Yamamoto, T. *et al.* (2009) Towards the understanding of complex traits in rice: substantially or superficially? *DNA Res.* 16, 141–154
43 Century, K. *et al.* (2008) Regulating the regulators: the future prospects for transcription-factor-based agricultural biotechnology products. *Plant Physiol.* 147, 20–29
44 Zhang, J.Z. *et al.* (2004) From laboratory to field. Using information from *Arabidopsis* to engineer salt, cold and drought tolerance in crops. *Plant Physiol.* 135, 615–621

45 Lacombe, S. *et al.* (2010) Interfamily transfer of a plant pattern-recognition receptor confers broad-spectrum bacterial resistance. *Nat. Biotechnol.* 28, 365–369

46 Nelson, D.E. *et al.* (2007) Plant nuclear factor Y (NF-Y) B subunits confer drought tolerance and lead to improved corn yields on water-limited acres. *Proc. Natl. Acad. Sci. U. S. A.* 104, 16450–16455

47 Rensink, W.A. and Buell, C.R. (2004) *Arabidopsis* to rice. Applying knowledge from a weed to enhance our understanding of a crop species. *Plant Physiol.* 135, 622–629

48 Hayama, R. *et al.* (2003) Adaptation of photoperiodic control pathways produces short-day flowering in rice. *Nature* 422, 719–722

49 Yan, L. *et al.* (2006) The wheat and barley vernalization gene VRN3 is an orthologue of FT. *Proc. Natl. Acad. Sci. U. S. A.* 103, 19581–19586

50 Yan, L. *et al.* (2003) Positional cloning of the wheat vernalization gene VRN1. *Proc. Natl. Acad. Sci. U. S. A.* 100, 6263–6268

51 Yan, L.L. *et al.* (2004) The wheat VRN2 gene is a flowering repressor down-regulated by vernalization. *Science* 303, 1640–1644

52 Jung, K.H. *et al.* (2008) Towards a better bowl of rice: assigning function to tens of thousands of rice genes. *Nat. Rev. Genet.* 9, 91–101

53 Krishnan, A. *et al.* (2009) Mutant resources in rice for functional genomics of the grasses. *Plant Physiol.* 149, 165–170

54 Davidson, R.M. *et al.* (2009) Germins: a diverse protein family important for crop improvement. *Plant Sci.* 177, 499–510

55 Izawa, T. *et al.* (2009) DNA changes tell us about rice domestication. *Curr. Opin. Plant Biol.* 12, 185–192

56 Negrão, S. *et al.* (2008) Integration of genomic tools to assist breeding in the *japonica* subspecies of rice. *Mol. Breed.* 22, 159–168

57 Fang, Y. *et al.* (2008) Systematic sequence analysis and identification of tissue-specific or stress-responsive genes of NAC transcription factor family in rice. *Mol. Genet. Genomics* 280, 547–563

58 Nuruzzaman, M. *et al.* (2010) Genome-wide analysis of NAC transcription factor family in rice. *Gene* 465, 30–44

59 Konishi, S. *et al.* (2006) An SNP caused loss of seed shattering during rice domestication. *Science* 312, 1392–1396

60 Shomura, A. *et al.* (2008) Deletion in a gene associated with grain size increased yields during rice domestication. *Nat. Genet.* 40, 1023–1028

61 Fukuoka, S. *et al.* (2010) Integration of genomics into rice breeding. *Rice* 3, 131–137

62 Shen, Y-J. *et al.* (2004) Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. *Plant Physiol.* 135, 1198–1205

63 McNally, K.L. *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

64 Yamamoto, T. *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

65 Zhao, K. *et al.* (2010) Genomic diversity and introgression in *O. sativa* reveal the impact of domestication and breeding on the rice genome. *PLoS One* 5, e10780

66 Kumar, G.R. *et al.* (2010) Allele mining in crops: prospects and potentials. *Biotechnol. Adv.* 28, 451–461

67 Carrillo, M. *et al.* (2009) Phylogenomic relationships of rice oxalate oxidases to the cupin superfamily and their association with disease resistance QTL. *Rice* 2, 67–79

68 Fukuoka, S. *et al.* (2009) Loss of function of a proline-containing protein confers durable disease resistance in rice. *Science* 325, 998–1001

69 Manosalva, P.M. *et al.* (2009) A germin-like protein gene family functions as a complex quantitative trait locus conferring broad-spectrum disease resistance in rice. *Plant Physiol.* 149, 286–296

70 Ouyang, S. *et al.* (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.* 35, D883–D887

71 Tanaka, T. *et al.* (2008) The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res.* 36, D1028–D1033

72 Paterson, A.H. *et al.* (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457, 551–556

73 Mace, E.S. and Jordan, D.R. (2010) Location of major effect genes in sorghum (*Sorghum bicolor* (L.) Moench). *Theor. Appl. Genet.* 121, 1339–1356

74 Gore, M.A. *et al.* (2009) A first-generation haplotype map of maize. *Science* 326, 1115–1117

75 McMullen, M.D. *et al.* (2009) Genetic properties of the maize nested association mapping population. *Science* 325, 737–740

76 Martinez-Zapater, J.M. *et al.* (2010) Grapevine genetics after the genome sequence: challenges and limitations. *Aust. J. Grape Wine Res.* 16, 33–46

77 Springer, N.M. *et al.* (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.* 5, e1000734

78 Gupta, P.K. (2008) Single-molecule DNA sequencing technologies for future genomics research. *Trends Biotechnol.* 26, 602–611

79 Mardis, E.R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24, 133–141

80 Varshney, R.K. *et al.* (2010) Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol.* 27, 522–530

81 Venter, J.C. (2010) Multiple personal genomes await. *Nature* 464, 676–677

82 Morgante, M. *et al.* (2007) Transposable elements and the plant pan-genomes. *Curr. Opin. Plant Biol.* 10, 149–155

83 Eid, J. *et al.* (2008) Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138

84 Rounsley, S. *et al.* (2009) De novo next generation sequencing of plant genomes. *Rice* 2, 35–43

85 Godfray, H.C.J. *et al.* (2010) Food security: the challenge of feeding 9 billion people. *Science* 327, 812–818

86 IBI (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* 463, 763–768

87 Goff, S.A. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296, 92–100

88 Ming, R. *et al.* (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452, 991–996

89 Velasco, R. *et al.* (2010) The genome of the domesticated apple (*Malus x domestica* Borkh.). *Nat. Genet.* 42, 833–839

90 Vielle-Calzada, J.P. *et al.* (2009) The Palomero genome suggests metal effects on domestication. *Science* 326, 1078

91 Chan, A.P. *et al.* (2010) Draft genome sequence of the oilseed species *Ricinus communis*. *Nature Biotech.* 28, 951–956

92 Visser, R. *et al.* (2009) Sequencing the potato genome: outline and first results to come from the elucidation of the sequence of the world's third most important food crop. *Am. J. Potato Res.* 86, 417–429

93 Timko, M. *et al.* (2008) Sequencing and analysis of the gene-rich space of cowpea. *BMC Genomics* 9, 103–123