Biotechnology Advances 32 (2014) 122–136

Contents lists available at ScienceDirect

Biotechnology Advances

journal homepage: www.elsevier.com/locate/biotechadv

Research review paper

Advances in plant chromosome genomics

Jaroslav Doležel *, Jan Vrána, Petr Cápal, Marie Kubaláková, Veronika Burešová, Hana Šimková

Centre of the Region Haná for Biotechnological and Agricultural Research, Institute of Experimental Botany, Šlechtitelů 31, Olomouc CZ-78371, Czech Republic

ARTICLE INFO

Available online 7 January 2014

Keywords:
BAC library
Chromosome sorting
Cytogenetics
DNA arrays
DNA marker development
Flow cytometry
Genome assembly
Mitotic metaphase chromosomes
Next generation sequencing
Physical mapping

ABSTRACT

Next generation sequencing (NGS) is revolutionizing genomics and is providing novel insights into genome organization, evolution and function. The number of plant genomes targeted for sequencing is rising. For the moment, however, the acquisition of full genome sequences in large genome species remains difficult, largely because the short reads produced by NGS platforms are inadequate to cope with repeat-rich DNA, which forms a large part of these genomes. The problem of sequence redundancy is compounded in polyploids, which dominate the plant kingdom. An approach to overcoming some of these difficulties is to reduce the full nuclear genome to its individual chromosomes using flow-sorting. The DNA acquired in this way has proven to be suitable for many applications, including PCR-based physical mapping, *in situ* hybridization, forming DNA arrays, the development of DNA markers, the construction of BAC libraries and positional cloning. Coupling chromosome sorting with NGS offers opportunities for the study of genome organization at the single chromosomal level, for comparative analyses between related species and for the validation of whole genome assemblies. Apart from the primary aim of reducing the complexity of the template, taking a chromosome-based approach enables independent teams to work in parallel, each tasked with the analysis of a different chromosome(s). Given that the number of plant species tractable for chromosome sorting is increasing, the likelihood is that chromosome genomics – the marriage of cytology and genomics – will make a significant contribution to the field of plant genetics.

Contents

* Corresponding author. Tel.: +420 585 238 703.
E-mail addresses: dolezel@ueb.cas.cz (J. Doležel), vrana@ueb.cas.cz (J. Vrána), capal@ueb.cas.cz (P. Cápal), kubalakova@ueb.cas.cz (M. Kubaláková), buresova@ueb.cas.cz (V. Burešová), simkovah@ueb.cas.cz (H. Šimková).

0734-9750
http://dx.doi.org/10.1016/j.biotechadv.2013.12.011

1. Sequencing of plant genomes

The last decade has seen a major leap in our understanding of plant genome structure, function and evolutionary dynamics. The main driver of this advance has been the elaboration of next generation sequencing (NGS) platforms, which allow for the parallel acquisition of huge numbers of reads, representing hundreds of billions of nucleotides; in concert, advances in bioinformatics have been necessary to enable this ocean of DNA sequence to be analyzed. The first plant genome to be fully sequenced was that of *Arabidopsis thaliana*, chosen for its small genome of ~150 Mb; although this represented a logistical challenge in the context of 1990s sequencing technology, it would no longer do so, given the capacity of modern instruments, which can generate up to 60 Gb of sequence per run. The *A. thaliana* genome was acquired using a clone-by-clone (CBC) strategy (The Arabidopsis Genome Initiative, 2000). The minimum set of clones to be sequenced, termed the "minimum tiling path" (MTP), is elaborated from the physical map, which is constructed on the basis of overlapping large-insert DNA clones. The second plant species to be sequenced was rice, using a similar strategy (Matsumoto et al., 2005). Apart from its importance as a crop species, rice was selected also because of its relatively small genome size (~400 Mb). The acquisition of these two whole genome sequences marked a new departure for plant genetics, allowing, for the first time, a holistic view of the entire genome. Since the beginning of the present century, the pace of sequencing has accelerated, so that by 2010, a number of important plant species had been sequenced.

A gradual shift in sequencing strategy, moving away from the CBC approach to a whole genome shotgun (WGS) one was already underway during the first phase of plant genome sequencing. The shotgun method was used for acquiring the genome sequences of poplar (Tuskan et al., 2006), grapevine (Jaillon et al., 2007) and sorghum (Paterson et al., 2009). The 2.5 Gb maize genome was published in 2009, but exceptionally relied on the CBC approach (Schnable et al., 2009). Since 2010, NGS technologies have become routine, and have greatly driven down both the price and effort required of genome sequencing. In this second phase of plant genome sequencing, already some 40 plant species have been sequenced, and the expectation is that not only reference genome sequences will be acquired for most of the economically and scientifically important plant species, but that the scale of re-sequencing will grow by orders of magnitude (The million plant and animal genomes project, 2013). Unlike *de novo* sequencing, which requires the assembly of the genome from short reads, re-sequencing is technically simpler, as the reads can be referenced to an available complete genome sequence. The quality of re-sequenced genomes is therefore determined by the quality of the reference genome sequence; the fuller the coverage of the reference sequence, the more correctly the re-sequenced contigs will be ordered. The feasibility of sequencing many individuals from the same species offers opportunities for population genetics analysis and genotype-based breeding (Long et al., 2013).

High quality reference genome sequences are particularly important for the analysis of the functional organization of DNA. The function of the nuclear genome cannot be understood without an understanding of its various components, as exemplified by the human genome ENCODE project (Gerstein et al., 2012). An unfortunate consequence of the widespread use of NGS shotgun sequencing is a drop in assembly quality, so that the highest quality genome sequences remain those of *A. thaliana*, rice and maize, which were acquired by the CBC method (Feuillet et al., 2011; Shangguan et al., 2013). Assembly is particularly problematical for large genome species such as Norway spruce (1C: ~20 Gb), where only some 25% of the genome was assemblable into scaffolds longer than 10 Kb (Nystedt et al., 2013); such issues can arise in smaller genomes too, for example in chickpea (1C: ~0.9 Gb), where the genome sequence presently comprises over 180,000 scaffolds (Jain et al., 2013). Of course, it is not always necessary to generate a gold standard sequence, since for some applications a rough genome draft is sufficient for the purpose. The difficulty arises when such draft genome assemblies are presented as reference sequences (Sierro et al., 2013). In some cases, projects relying on incomplete genome sequences may fail, and there are examples where funding proposals aimed at the acquisition of a high quality reference sequence have been declined as the donors believed that the work had already been done.

The power of NGS lies in its capacity to generate a huge volume of reads, but its weakness is that these reads are rather short. Plant genomes are populated by many families of repetitive DNA elements (Schmidt and Heslop-Harrison, 1998), and these can be impossible to resolve when only short reads are available. The problem of sequence redundancy is compounded in polyploids, which dominate the plant kingdom. Genome assembly from shotgun reads may not be straightforward even in compact genomes having a small content of repetitive DNA. A good example is the bladderwort *Utricularia gibba*, with a genome size of just 77 Mb, of which only 3% is repetitive; nevertheless an attempt at shotgun sequencing resulted in a set of >3800 sequence contigs arranged in over 1200 scaffolds (Ibarra-Laclette et al., 2013). Technical improvements in read length and/or the algorithms used for sequence assembly should in time, however, enable reference genome sequences to be produced by NGS shotgun methods (Roberts et al., 2013). NGS shotgun sequencing may be at present be of limited utility in acquiring gold standard reference sequences (Marx, 2013), but the technology is very powerful for simpler templates such as bacterial artificial chromosomes (BACs), which form the backbone of many physical maps (Feuillet et al., 2011). Incomplete sequence assembly is then limited to at most 100 Kb, the genomic location of which is known. BAC clones are commonly sequenced in pools to reduce cost (Sato et al., 2011; Steuernagel et al., 2009), and this requires a bar-coding strategy to attribute the resulting contigs to their specific BAC. The sequence redundancy typical of large and particularly of polyploid genomes, makes the construction of a physical map based on BAC clones difficult (Meyers et al., 2004; Paux et al., 2008); it is a task which would be greatly simplified if the template complexity could be reduced.

2. Reducing the complexity of the sequencing template

As both the CBC and the NGS shotgun sequencing strategies are compromised by sequence redundancy, any reduction in template complexity would be helpful. Breaking down the genome into its individual chromosomes represents an attractive option, especially for polyploid genomes, as this would abolish the problem of redundancy due to the presence of homoeologs (Fig. 1). Flow-sorting has been developed to achieve exactly this result, and this review outlines its potential for plant genome analysis and sequencing. Methods designed to simplify the assembly of shotgun sequence reads and to construct ready-to-sequence clone-based physical maps are described. Chromosome sorting is not, of course, the sole option available for reducing template complexity prior to DNA sequencing. The selection of DNA based on either its renaturation kinetics ("Cot filtration") (Peterson et al., 2002)