Journal of Biotechnology 261 (2017) 46–52

Contents lists available at ScienceDirect

Journal of Biotechnology

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

From plant genomes to phenotypes

Marie Bolger[a], Rainer Schwacke[a], Heidrun Gundlach[b], Thomas Schmutzer[c], Jinbo Chen[c], Daniel Arend[c], Markus Oppermann[c], Stephan Weise[c], Matthias Lange[c], Fabio Fiorani[a], Manuel Spannagl[b], Uwe Scholz[c], Klaus Mayer[b], Björn Usadel[a,*]

[a] Forschungszentrum Jülich (FZJ), Institute of Bio- and Geosciences (IBG-2) Plant Sciences, Wilhelm-Johnen-Straße, 52425 Jülich, Germany
[b] Helmholtz Zentrum München (HMGU), Plant Genome and Systems Biology (PGSB), Ingolstädter Landstraße 1, 85764, Neuherberg, Germany
[c] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Corrensstraße 3, 06466, Seeland, Germany

ARTICLE INFO

Keywords:
Plant genomes
Plant bioinformatics
Plant genome annotation
Phenotyping

ABSTRACT

Recent advances in sequencing technologies have greatly accelerated the rate of plant genome and applied breeding research. Despite this advancing trend, plant genomes continue to present numerous difficulties to the standard tools and pipelines not only for genome assembly but also gene annotation and downstream analysis.

Here we give a perspective on tools, resources and services necessary to assemble and analyze plant genomes and link them to plant phenotypes.

1. Introduction

The last decade has seen tremendous progress in the field of plant genomes which began with the model plant, *Arabidopsis thaliana*, whose genome was published in 2000 (Arabidopsis Genome, 2000). This was followed shortly with the genomes of the first crop plants, rice and poplar and since then both crop and non-crop plants from diverse clades have been sequenced and assembled (Bolger et al., 2014c) (http://www.plabipd.de/portal/sequence-timeline). Until recently, these sequencing projects were carried out by large genome sequencing consortia bringing together expertise in many different fields. The next generation sequencing era gradually enabled individual labs or small consortia to undertake whole plant genome assembly projects (Jiao et al., 2017; Pucker et al., 2016). Further improvements in long-read technologies help bridge repeat regions, a major obstacle to completing a genome, which previously required complicated mate-pair libraries (Jiao et al., 2017). Still, despite these tremendous advances, many plant genomes contain particular complexities which make genome assemblies and analysis difficult (Claros et al., 2012).

Given the fact that the “1000$ human genome” has been called a reality by Illumina, one might wonder why some plants are resilient to having their genomes assembled (Fig. 1).

One of the primary issues with plants is the highly repetitive nature of many plant genomes which presents a major inherent problem for the assembly process. A recent study has highlighted this issue by systematically comparing over 40 plants with over 60 vertebrate genomes using an unbiased kmer based approach and showed the high repeat content in plants (Jiao and Schneeberger, 2017). Though this issue has been greatly reduced by long-read technologies, these reads are still unable to span the large tandem repeat regions found in many plant genomes. This problem is further exacerbated in many cases by the sheer size of plant genomes which in extreme cases can reach to tens of gigabases. In addition, plant genomes frequently have high ploidy levels which resulted from either duplications as seen in the autopolyploid potato genome or from a combination of genomes from different species which gives rise to allopolyploid species such as wheat, Camelina and rape seed. As an example, the ca 17 GB wheat genome features three subgenomes (International Wheat Genome Sequencing, 2014), each of which contains a set of homologous genes (genes with purportedly the same ancestor from different genome complements). Last but not least, many plants are self-incompatible (Fujii et al., 2016) thereby resisting attempts to achieve a homozygous state. The genome assembly process is greatly simplified however when there is only a single allele per locus to assemble. These genomic features combine to magnify the seemingly trivial task of producing a high quality genome assembly.

Plants are extremely versatile organisms featuring many hundreds of thousands of metabolites occurring through the plant kingdom and probably more than ten thousand metabolites per species (De Luca and St Pierre, 2000; Fernie, 2007). Although many of these metabolites share common pathways, there are still a large number of genes involved overall. In some cases, divergent genes which have largely retained sequence homology need to be separately annotated due to their involvement in different pathways and thereby different metabolites.

* Corresponding author.
E-mail address: b.usadel@fz-juelich.de (B. Usadel).

http://dx.doi.org/10.1016/j.jbiotec.2017.06.003
Received 17 February 2017; Received in revised form 27 May 2017; Accepted 7 June 2017
Available online 09 June 2017

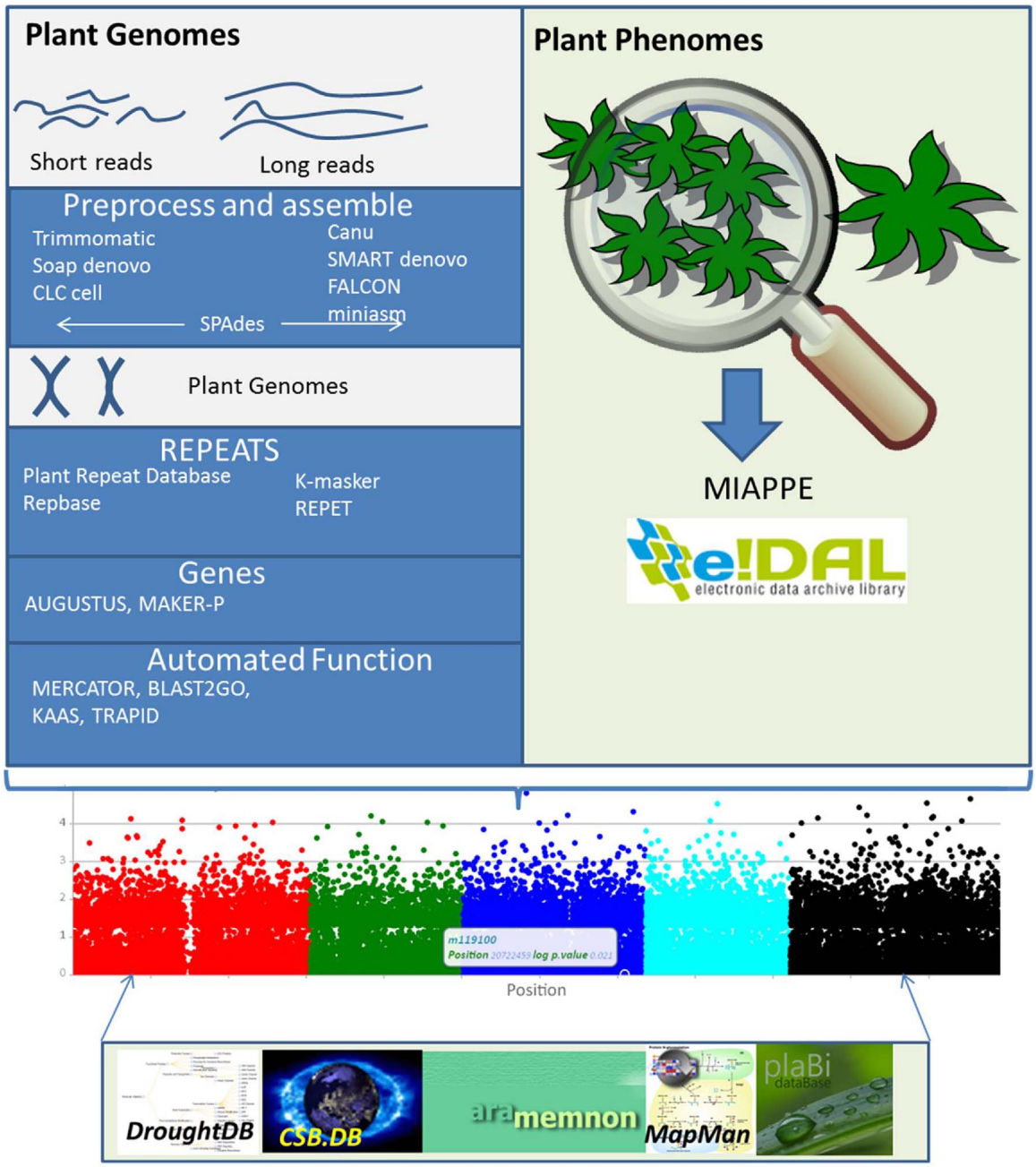

Fig. 1. From genomes and phenomes to candidate genes: Fig. 1 shows which tools one can typically use to assemble and annotate plant genomes and what standards are being developed for plant phenotyping.
The analysis of genomes and phenotypes leads to candidate regions which can be further delineated.

Furthermore, plant sciences feature several communities with specific needs and interests. As an example, a researcher working on sugar beet (Dohm et al., 2014) or carrots (Iorizzo et al., 2016) will be more interested in below ground organs and their development than a researcher working on tomatoes or barley. However, a common ground is that the plants are (also) looked at from a breeding perspective in order to improve plant yield and/or resilience. This process is greatly expedited by the development of statistical analysis and model based analysis of plant genetic (genomic) and phenotypic data (Hammer et al., 2006) as well as the maturation and development of plant phenotyping technology (Fiorani and Schurr, 2013).

Within this review we attempt to shed light on the analysis of plant genomes, describe current problems as well as how plant genomes can be best leveraged in conjunction with high throughput phenotyping to accelerate selective plant breeding.

We provide a detailed list of tools which can be used in the process of genome assembly, annotation and linking it to phenotypic plant data.

2. *De novo* genome assembly

Plant *de-novo* genome assembly is notoriously difficult (Claros et al., 2012), mainly due to the problems mentioned above. This has prompted the development of tools which can cope with these difficulties, some of which also serve the wider scientific community. A notable example of this is the raw data preprocessing tools 'Trimmomatic' http://www.usadellab.org/cms/?page=trimmomatic (Bolger et al., 2014b) which was developed due to the necessity for a highly efficient adapter trimmer during a plant genome sequencing project and has since been widely adopted by the whole scientific community due to its flexibility, speed and efficiency.