Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Plant Biology

Short read sequencing in studies of natural variation and adaptation

Ben Hunter*, Kevin M Wright* and Kirsten Bomblies

Today's high throughput sequencing approaches, coupled with equally revolutionary advances in bioinformatics, allow us to describe and analyze genomes in unprecedented detail. Short Read Sequencing (SRS) approaches have been especially instrumental in bringing genomic analysis to a wide range of questions and species in plant biology. We can now connect genotypes and phenotypes with greater efficiency, and investigate the molecular basis of natural variation and adaptation in a genomic framework. New and creative applications of SRS and other genomic approaches are not only reshaping how we study natural variation, but also our overall understanding of gene and genome evolution. Here we discuss examples of the application of SRS technologies to the characterization of genetic diversity, genome evolution and adaptation in plants.

Address
Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA

Corresponding author: Bomblies, Kirsten (kbomblies@oeb.harvard.edu)
* These authors contributed equally.

Current Opinion in Plant Biology 2013, **16**:85–91

This review comes from a themed issue on **Growth and development**

Edited by **Michael Scanlon** and **Marja Timmermans**

For a complete overview see the Issue and the Editorial

Available online 20th November 2012

http://dx.doi.org/10.1016/j.pbi.2012.10.003

Introduction

The ongoing revolution in genomics is unlikely to have escaped the notice of any biologist working today [1,2]. Our unprecedented power to sequence and analyze entire genomes, coupled with improved analytical tools and declining costs have opened the floodgates for exceptionally detailed studies of variation and adaptation. Large-scale efforts in model systems have enabled rigorous analysis of forces shaping diversity in natural populations [3–5], while extensions to non-model species broaden the applications to novel questions in plant diversity and evolution [6].

Some researchers expressed initial fears that creativity might be stifled by the relegation of genomics to large wealthy facilities focused on 'idea-free' data generation [7], but this is not the reality of genomics. Improvements in technology and analysis have rendered genomic approaches more cost-effective and user-friendly, so even with moderate funding these tools are accessible to small labs or new investigators. This democratization of genomics is bearing fruit, as we will discuss below, focusing in particular on the application of Short Read Sequencing (SRS) technology to questions in natural variation and adaptation.

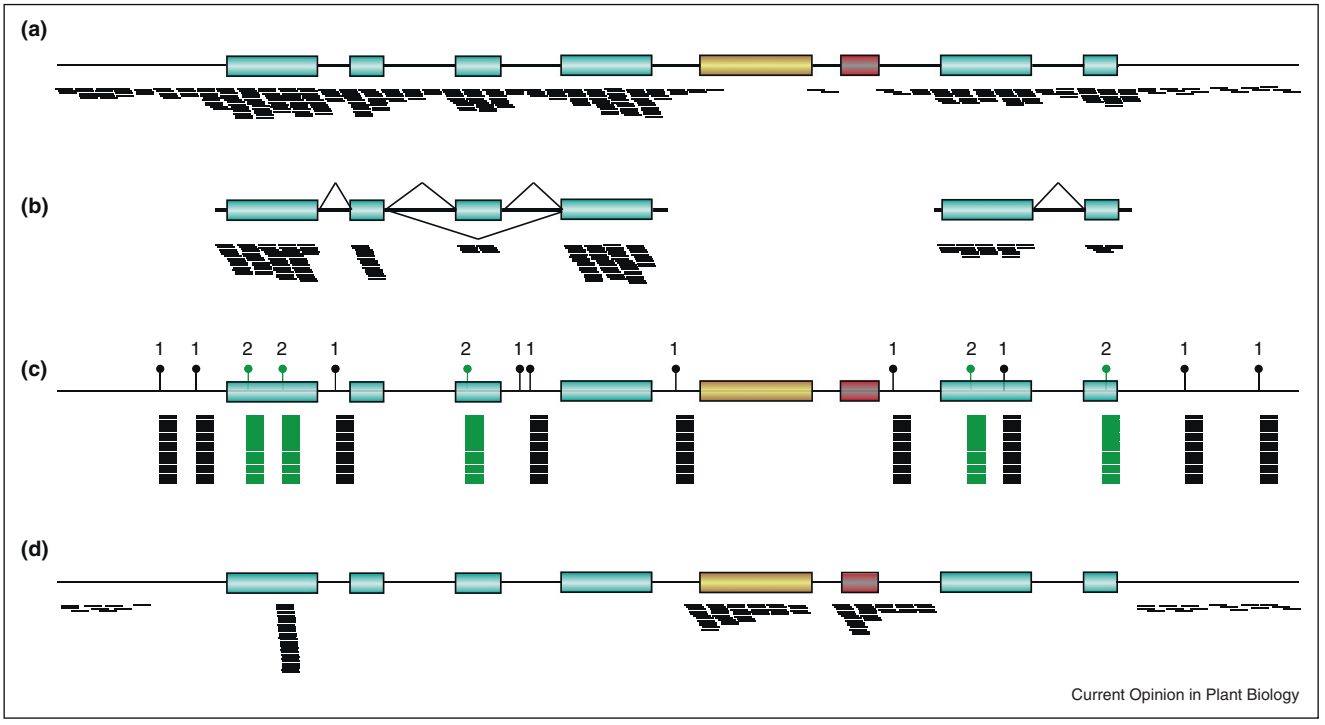
Describing genomic variation

Genome sequencing has progressed so rapidly that only twelve years after publication of the *Arabidopsis thaliana* and rice genomes [8,9] we now have comparative genomic databases with fully sequenced genomes of representatives from every major group in the plant kingdom. These are publicly available at sites such as Phytzome (www.phytzome.net), the Joint Genome Institutes (www.jgi.doe.gov) which has a collection of 31 fully sequenced plant and algae genomes, and family specific databases such as Sol Genomics Network ([10]; solgenomics.net) and grass genomic network ([11]; www.gramene.org). Such databases are much more than repositories and provide an increasingly useful framework for contextualizing and guiding sequencing efforts in new species. SRS technologies in particular (e.g. Illumina and 454) provide individual researchers the ability to generate full genome coverage of virtually any organism of interest at a relatively low cost (Figure 1; [12,13]).

The 1001 Arabidopsis genomes project (signal.salk.edu/atg1001) provides an excellent example in plants of how SRS data have been effectively used to describe genomic variation in great detail [14–18]. This project has described genomic variation in nucleotide sequence, gene expression and gene copy number [16•,18,19•], and provides an excellent model for future population genomic studies in other species. In addition to describing natural variation, SRS has also been used to estimate that base substitutions accumulate about one per generation in lab-grown *A. thaliana* [20]. Datasets such as these for core model systems facilitate comparative genomic analysis by providing assembly scaffolds and starting points for gene annotation. For example, sequencing of additional Brassicaeae such as *A. lyrata*, *Thellungiella salsuginea* and *Brassica rapa*, is facilitating analyses of genome evolution using *A. thaliana* as a reference point [21–23].

What about when there is no prior genomic information available for a particular taxon? Detailed projects like the 1001 *Arabidopsis* genome project are very valuable, but

Figure 1



Representation of different genomic regions with distinct short read approaches. The line indicates a genomic region the same for all panels with exons shown as blue boxes, and two transposons in red and gold. **(a)** Whole genome sequencing and assembly will give reads throughout the genome, but transposon or other repetitive regions will generally not assemble well. In reference-guided assemblies, exons in general have higher read coverage than introns and intergenic regions due to differences in polymorphism. **(b)** Transcriptome sequencing will give short read coverage only in regions represented in mature mRNAs. Differences in coverage can be used to estimate expression level differences, as well as alternative splicing when there expression differences among exons are observed. This will be further bolstered by the existence of reads that span exon boundaries. **(c)** RADseq is based on restriction enzyme digestion followed by ligation of adapters to the overhangs. This allows sequencing of targeted regions to high coverage. Usually this is done on many individuals in parallel to identify polymorphisms. The numbers 1 and 2 indicate restriction sites with different GC contents. In general restriction enzymes that have AT-rich recognition sites will be more likely to query intergenic regions (here #1) while GC-rich sites will primarily query gene regions (here #2). **(d)** Small RNA sequencing can identify regions that are predicted to be silenced by small RNAs (here the transposons and to a lesser degree the intergenic regions) as well as microRNA sites as here in the leftmost exon.

such an in-depth study is time consuming and costly, making it not currently feasible in most systems. Happily SRS-based techniques are available that allow us to diversify beyond the model systems. Assembling genomes *de novo* from SRS data is challenging, but available tools have greatly improved (reviewed in [24]). Furthermore, full genome assembly is not always needed. One technique which is rapidly gaining a foothold for its ease and utility in generating genetic markers is restriction site-associated DNA sequencing (RADseq; [25–27]). This method isolates small, somewhat randomly dispersed fragments of the genome that can be repeatedly isolated across individuals (Figure 1). For example, RADseq has been successfully employed in plants to describe genetic variation within domesticated strains and wild relatives of the commercially important crops Maize [28] and Grape [29]. Alternatively, genetic polymorphisms can also be identified by sequencing just the transcriptome, which greatly simplifies *de novo* assembly and provides additional information on gene expression (Figure 1; [30]). This has been used successfully to identify genetic markers in Chickpea for example [31].

In addition to describing polymorphisms from genomic DNA, SRS can be employed to characterize other aspects of variation such as epigenetic marks and small RNA profiles. Small RNAs (smRNAs) generally function as negative regulators of gene expression through a variety of mechanisms, including epigenetic modifications and mRNA degradation [32]. Natural variation for smRNAs has been well described in *A. thaliana*, *A. lyrata* [33–35], *Medicago* and multiple *Solanaceae* species [36–38]. This work uncovered an unknown mechanism of disease resistance gene regulation by smRNAs that is seemingly absent in the *Arabidopsis* genus [36–38]. This demonstrates the utility of SRS technology-facilitated studies in a diversity of organisms to better understand the origins and maintenance of variation.