F1000Reports BIOLOGY

Published: 08 December 2010

The case for resequencing studies of *Arabidopsis thaliana* accessions: mining the dark matter of natural genetic variation

Luca Santuari and Christian S Hardtke*

Address: Department of Plant Molecular Biology, University of Lausanne, Biophore Building, DBMV, CH-1015 Lausanne, Switzerland

* Corresponding author: Christian S Hardtke (christian.hardtke@unil.ch)

F1000 Biology Reports 2010, **2**:85 (doi:10.3410/B2-85)

Abstract

Ultra-high-throughput sequencing (UHTS) techniques are evolving rapidly and may soon become an affordable and routine tool for sequencing plant DNA, even in smaller plant biology labs. Here we review recent insights into intraspecific genome variation gained from UHTS, which offers a glimpse of the rather unexpected levels of structural variability among *Arabidopsis thaliana* accessions. The challenges that will need to be addressed to efficiently assemble and exploit this information are also discussed.

Introduction and context

The introduction of 'next-generation' sequencing technologies has had a tremendous impact on all areas of biology dealing with genomic information, from population genetics to comparative genomics, including the plant sciences [1,2]. This impact is expected to grow exponentially as technical advances are continuously increasing the length and quality of sequenced DNA fragments while decreasing the cost of the process. The throughput of current ultra-high-throughput sequencing (UHTS) equipment has already reached several giga base pairs per run, which means that complex multicellular organisms with an average-sized genome, such as the model plant *Arabidopsis thaliana* (Arabidopsis), can be sequenced at reasonable coverage and cost. To reliably assemble whole genome sequences from the short reads generated by UHTS, a well-defined, high-quality reference genome still remains invaluable. This is the case for Arabidopsis [3], which was one of the immediate candidate organisms for resequencing projects. Indeed, a few studies have already focused on comparative analysis of divergent Arabidopsis genomes to characterize the extent of intraspecific genome variation with respect to the Columbia-0 (Col-0) reference accession [4,5].

Major recent advances

Sufficient funding permitting, these initial studies are just the prelude to the ambitious '1001 Genomes Project', which aims to resequence divergent Arabidopsis accessions by the hundreds [6]. In a pilot study [4], Ossowski *et al.* resequenced the genome of Col-0, revealing more than 2000 homozygous single nucleotide polymorphisms (SNPs) and insertions and deletions (indels) that represent potential errors in the original annotation or spontaneous mutations (see below). Moreover, they analyzed the genome of two divergent strains and found that each of them carries more than 800,000 SNPs and 80,000 short indels. Longer indels were difficult to assess from the short reads by simple re-mapping to the reference genome. However, a *de novo* assembly approach allowed the authors to identify some larger structural variants by starting from reads that flanked regions characterized by low coverage. Using an alternative, combinatorial approach, Santuari *et al.* [5] evaluated the genome-wide abundance of large-scale deletions in four Arabidopsis strains sequenced at moderate coverage. The authors demonstrated that the intersection of signal intensities from tiling array hybridizations with UHTS read coverage accurately detects larger deletions. Hundreds of major deletions were observed, which frequently affect gene

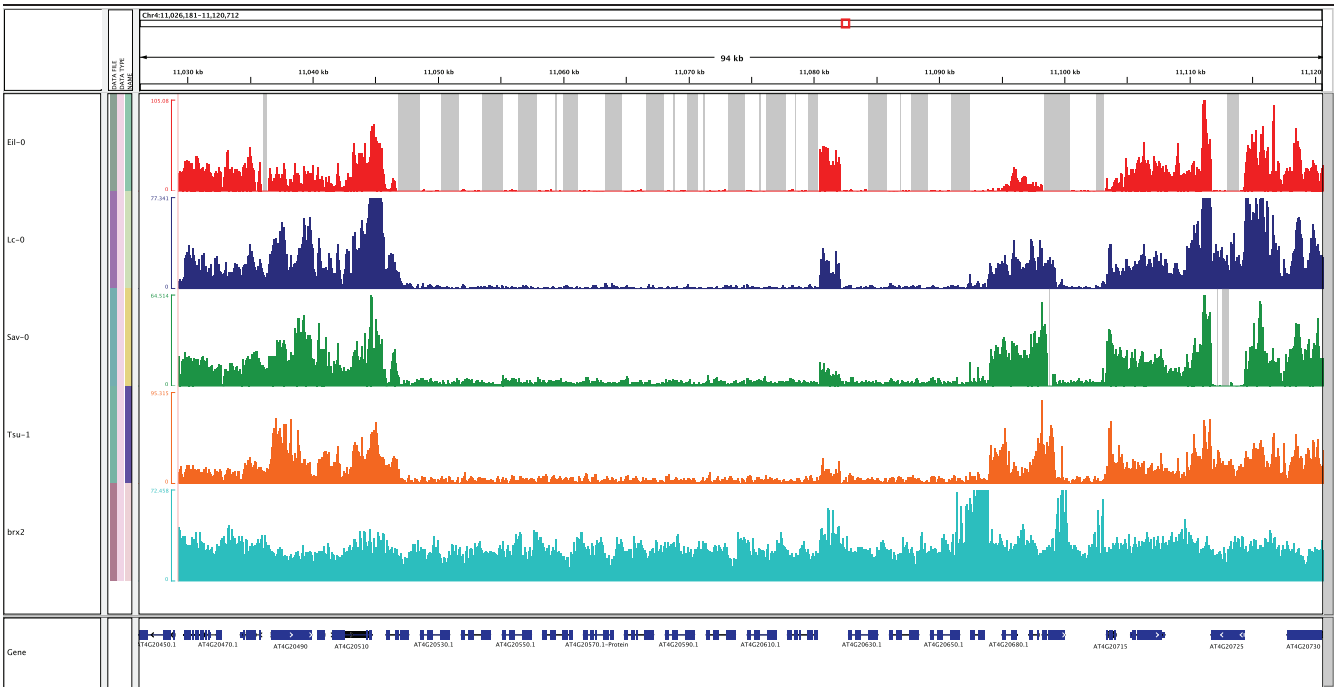
function. Among them, transposable elements were found to be overrepresented, suggesting that the majority of genomic rearrangements identified result from the activity of mobile elements. Such activity was recently also observed in real time [7]. Individual deletions were frequently observed in two or more of the accessions examined, suggesting that variation in gene content partly reflects a common history of deletion events.

In summary, the characterization of only a limited number of divergent Arabidopsis genomes has already identified an unexpected degree of structural diversity that significantly affects gene content and function [4,5,8] (Figure 1). Although the majority of those polymorphisms supposedly originated in the wild, recent studies have highlighted the dynamics of the evolutionary process over merely a few generations, as exemplified in another pioneering publication by Ossowski *et al.* [9]. In this paper, the authors used next-generation sequencing to evaluate the rate and accumulation of spontaneous mutations across several generations. They analyzed five Col-0 lines that had been maintained for 30 generations and found a mutation per site rate equaling 7×10^{-9} base pairs per generation. Thus, Arabidopsis geneticists are confronted with the finding that mutant lines and reference backgrounds might be more disparate than suspected. In practice, this should not pose a major problem as most mutations observed are essentially neutral. However, another study analyzing structural genome variation over several inbred generations subjected to stress treatment found major, stress-induced structural variation that could significantly affect phenotype [10]. Although purely array-based and thus not ultimately conclusive, if confirmed by UHTS, these findings might force us to rethink our notions of genome stability.

Future directions

As this perspective is written, it is already likely to be superseded by ongoing technological progress, which should soon permit routine identification of structural genome variants. For instance, using paired-end sequencing it is already possible to follow the movements of transposons, which are highlighted as mismappings of the two ends of a sequenced fragment on the reference genome. Beyond hardware improvements, the development of bioinformatics tools will also be a major driver. Already several tools have been developed to detect inter- and intra-chromosomal rearrangements [11,12].

Figure 1. Example of a major common deletion affecting multiple genes

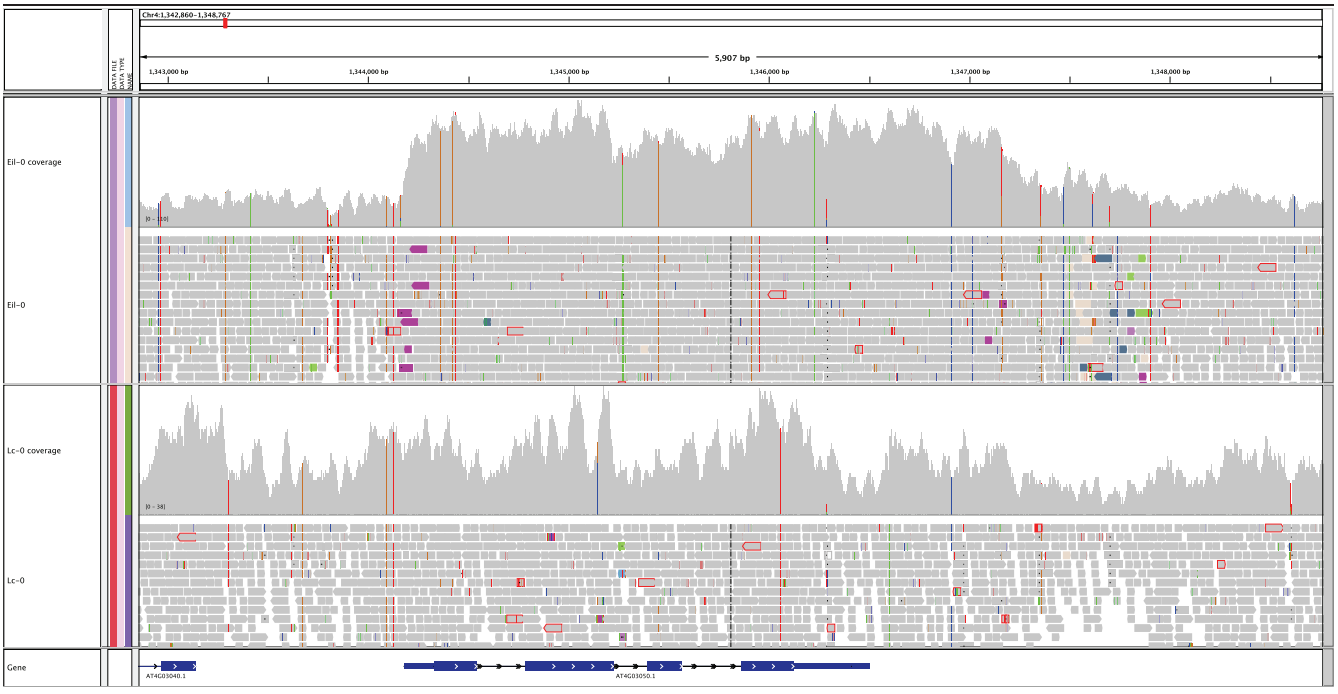


A region on Arabidopsis chromosome 4 of approximately 68 kilo base pairs (kb) contains a series of receptor-like protein kinase-related proteins in Col-0, but shows a clear lack of read coverage in the Eilenberg-0 (Eil-0) strain (top), as depicted by the grey bars, and significant decrease in coverage for the other strains in the diagram. A closer look reveals that a major number of reads mapped to this region are probably mis-mapped and thus misleading due to very low mapping quality. Overall, this picture indicates a series of large scale deletions comprising at least 15 genes that are present in the Col-0 strain, represented by the *brx*-2 mutant obtained from a Col-0 background, but missing in all four other strains. Lc-0, Loch Ness-0; Sav-0, Slavice-0, Tsu-1, Tsushima-1.

The read length provided by the current versions of most UHTS platforms coupled with different insert size libraries will ultimately make it feasible to move from analysis that is focused on read mapping onto a reference genome towards a reference-free, *de novo* assembly approach. This is critical to overcome the intrinsic limitations dictated by relying on a single reference sequence. The collection of high-quality scaffolds and contigs from divergent Arabidopsis accessions could be used to define genomic regions that may have accumulated a degree of divergence that would prevent their accurate elucidation by classical mapping approaches (Figure 2). In particular, in the case of duplicated or partially conserved regions, the assembly itself is limited by the highly repetitive content of these sequences. The integration of data from the reference mapping with assembled scaffolds will help to reconstruct specific regions that would otherwise be hard to decipher using either mapping or assembly alone. *De novo* assembly algorithms that take into account the read mapping position on the reference sequence are already being developed, such as LOCAS [13] or the recent Columbus algorithm implemented in Velvet [14,15]. Notably, these tools can also assemble genomes from low coverage data, further decreasing costs. Another important challenge in analyzing these data is how to make them accessible and useful to researchers. This will probably be driven by increasingly advanced and intuitive genome browsers, such as the recent version of Ensembl Plants or GenomeMapper [16].

Beyond the utility of UHTS in every day lab approaches, such as mutant mapping [17], one might ask: why should we generate these data? Clearly, one of the greatest promises lies in their integration into genome-wide association studies [18], which would enable us to move from assessing the qualitative and quantitative effects of a single locus towards identifying and evaluating the systemic effects of multiple genes involved in a trait of interest [18-20]. The genome sequences of specific parental accessions would also greatly accelerate standard quantitative genetics approaches, such as quantitative trait locus analysis of recombinant inbred lines. Maybe most importantly, sequencing hundreds of strains, as proposed by the '1001 Genomes Project', will not only indicate which genes are divergent, missing, or not functional with respect to the Col-0 reference sequence in a given

Figure 2. Ambiguity in resolving loci that include duplicated genes and deletions



The *AOP3* gene, involved in glucosinolate biosynthesis and thus pathogen defense, appears to be duplicated in an exact copy in the Eilenberg-0 (Eil-0) accession, as indicated by homozygous single nucelotide polymorphisms depicted with colored vertical lines. Breakpoints are genomic regions where both ends of a fragment sequenced with the paired-end library are mapped on the reference genome at a distance that is significantly different from the insert size of the library, suggesting structural rearrangements. Here, the reads supporting breakpoints are shown in different colors, with each color representing the chromosome where the other end of the fragment is located. Following these reads, it is possible to reconstruct where the second copy of *AOP3* is located, which appears to be 10 kilo base pairs downstream of its original locus, immediately downstream of the *AOP2* locus in Col-0. Lc-0, Loch Ness-0.

accession, but will also lead to the discovery of genes that are present in the worldwide Arabidopsis population (but absent from Col-0). Without this 'dark matter' of the Arabidopsis genome, defining the full gene complement of the species and gaining a complete understanding of the ecological-evolutionary and developmental history of this plant cannot be attained.

Abbreviations

bp, base pair; Col-0, Columbia-0; Eil-0, Eilenberg-0; indel, insertion-deletion; SNP, single nucleotide polymorphism; UHTS, ultra-high-throughput sequencing.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The UHTS projects in our lab are supported by the SystemsX 'Plant Growth in a Changing Environment' network.

References

1. Gilad Y, Pritchard JK, Thornton K: **Characterizing natural variation using next-generation sequencing technologies.** *Trends Genet* 2009, **25:**463-71.
2. Lister R, Gregory BD, Ecker JR: **Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond.** *Curr Opin Plant Biol* 2009, **12:**107-18.
3. Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.** *Nature* 2000, **408:**796-815.
4. Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D: **Sequencing of natural strains of Arabidopsis thaliana with short reads.** *Genome Res* 2008, **18:**2024-33.

   F1000 Factor 7
   Evaluated by Bernd Weisshaar 10 Nov 2008, Julin Maloof 10 Nov 2008,
5. Santuari L, Pradervand S, Amiguet-Vercher AM, Thomas J, Dorcey E, Harshman K, Xenarios I, Juenger TE, Hardtke CS: **Substantial deletion overlap among divergent Arabidopsis genomes revealed by intersection of short reads and tiling arrays.** *Genome Biol* 2010, **11:**R4.

   F1000 Factor 6
   Evaluated by Michael Lassner 31 Mar 2010
6. Weigel D, Mott R: **The 1001 genomes project for Arabidopsis thaliana.** *Genome Biol* 2009, **10:**107.
7. Mirouze M, Reinders J, Bucher E, Nishimura T, Schneeberger K, Ossowski S, Cao J, Weigel D, Paszkowski J, Mathieu O: **Selective epigenetic control of retrotransposition in Arabidopsis.** *Nature* 2009, **461:**427-30.
8. Plantegenet S, Weber J, Goldstein DR, Zeller G, Nussbaumer C, Thomas J, Weigel D, Harshman K, Hardtke CS: **Comprehensive analysis of Arabidopsis expression level polymorphisms with simple inheritance.** *Mol Syst Biol* 2009, **5:**242.
9. Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M: **The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana.** *Science* 2010, **327:**92-4.

   F1000 Factor 6
   Evaluated by Motoaki Seki 27 Jan 2010
10. DeBolt S: **Copy number variation shapes genome diversity in Arabidopsis over immediate family generational scales.** *Genome Biol Evol* 2010, **2:**441-53.

    F1000 Factor 6
    Evaluated by Jiming Jiang 29 Jul 2010
11. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER: **BreakDancer: an algorithm for high-resolution mapping of genomic structural variation.** *Nat Methods* 2009, **6:**677-81.
12. Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoix-ne P, Nicolas A, Delattre O, Barillot E: **SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data.** *Bioinformatics* 2010, **26:**1895-6.
13. **LOCAS low-coverage short-read assembler.** [http://www-ab.informatik.uni-tuebingen.de/software/locas]
14. Zerbino DR: **Using the Velvet de novo Assembler for Short-Read Sequencing Technologies.** *Curr Protoc Bioinformatics* 2010, **31**(Suppl):11.5.1-12.
15. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18:**821-9.
16. Schneeberger K, Hagmann J, Ossowski S, Warthmann N, Gesing S, Kohlbacher O, Weigel D: **Simultaneous alignment of short reads against multiple genomes.** *Genome Biol* 2009, **10:**R98.
17. Laitinen RA, Schneeberger K, Jelly NS, Ossowski S, Weigel D: **Identification of a spontaneous frame shift mutation in a nonreference Arabidopsis accession using whole genome sequencing.** *Plant Physiol* 2010, **153:**652-4.
18. Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, Jiang R, Muliyati NW, Zhang X, Amer MA, Baxter I, Brachi B, Chory J, Dean C, Debieu M, de Meaux J, Ecker JR, Faure N, Kniskern JM, Jones JD, Michael T, Nemri A, Roux F, Salt DE, Tang C, Todesco M, *et al.*: **Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines.** *Nature* 2010, **465:**627-31.

    F1000 Factor 10
    Evaluated by Bernd Weisshaar 09 Jun 2010
19. Nemri A, Atwell S, Tarone AM, Huang YS, Zhao K, Studholme DJ, Nordborg M, Jones JD: **Genome-wide survey of Arabidopsis natural variation in downy mildew resistance using combined association and linkage mapping.** *Proc Natl Acad Sci U S A* 2010, **107:**10302-7.
20. Todesco M, Balasubramanian S, Hu TT, Traw MB, Horton M, Epple P, Kuhns C, Sureshkumar S, Schwartz C, Lanz C, Laitinen RA, Huang Y, Chory J, Lipka V, Borevitz JO, Dangl JL, Bergelson J, Nordborg M, Weigel D: **Natural allelic variation underlying a major fitness trade-off in Arabidopsis thaliana.** *Nature* 2010, **465:**632-6.