

Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes

Katharina Schenker^{1,2,3}, Stephan Ossowski^{1,2,3}, Felix Graf¹, Julian D. Klein⁴, Xi Wang¹, Chelia Lian¹, Lisa M. Smith¹, Jan Graf¹, Jeffrey Fong¹, Harman Walthering¹, Stefan R. Heur¹, David H. Hwang¹, and Detlef Weigel¹

Abstract

We present whole-genome assemblies of four diverse *Arabidopsis thaliana* accessions that complement the 100-Mb reference genome sequence released in 2004. Using a novel iterative reference-guided approach, we assembled large contigs from 1 to 4.5 Gb of short-read data from the Arabidopsis Genome Project and the 100 reference genomes. This approach allows us to assemble the complete set of the de novo assembly and later integrate well-supported contigs into the reference sequence. For example, the 100 reference genomes were used to assemble the complete set of the de novo assembly and later integrate well-supported contigs into the reference sequence. For example, the 100 reference genomes were used to assemble the complete set of the de novo assembly and later integrate well-supported contigs into the reference sequence.

Results

Reference-guided assembly. Our reference-guided assembly approach is outlined in Fig. 1. We used paired-end reads of 36 bp that were generated from the four diverse accessions, with average library insert lengths from 177 to 436 bp (Table S1). Some of the reads had been produced previously (24, 25). Following an algorithm of our short-read assembler, the reference-guided assembly was performed using LASTZ (Stephan et al. 2009). We performed reads on the basis of their distance between contigs and their overlap with contigs. We used a greedy approach to assemble contigs with maximal overlap to adjacent regions covered by the same read pair. We then merged adjacent contigs into supercontigs, with neighboring supercontigs sharing at least one block. Each supercontig was compared to the reference contigs. We also included "clipping" reads with the same alignment as one of the reference reads (Graf et al. 2009). The algorithm concatenates supercontigs and their subcontigs into supercontigs. In addition, the actual overlap of genes identified in *A. thaliana* is thereby expected.

SUPERLOCAS Klein et al. 2011

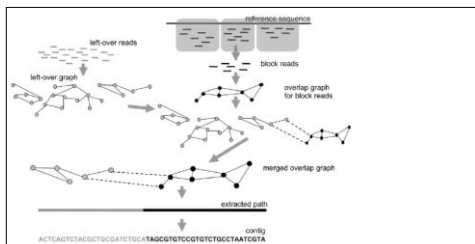


Figure 2. Workflow of SUPERLOCAS. The figure shows the workflow of the algorithm of SUPERLOCAS. The initial steps are illustrated: the left-over reads with the constructed left-over overlap graph, and the reads that are aligned against the reference sequence and partitioned into blocks. Next, the steps that are executed consecutively for each block are shown: the construction of the overlap graph, the insertion of edges between block graphs and the procedure until contigs are reported for the merged graph.

Schneeberger et al. 2011

Table 1. Assembly statistics

	Bur-0	C24	Kro-0	Ler-1
Coverage	83.2x	75.0x	72.7x	322.4x
Libraries	2	2	2	3
NSO (intrinsic)	193	109	161	113
L50, kb	147.3	273.2	163.5	272.5
NSO (target)	208	117	178	121
L50, kb	139.6	260.4	151.8	261.9
Scaffolds	2,526	2,052	2,670	1,528
Total length, Mb	101.0	101.3	99.9	100.8
Longest scaffold, Mb	1.12	2.18	1.48	1.09
Ambiguous bases, %	4.0	3.6	5.1	1.3

NSO(L50 (intrinsic)) using total length of the scaffolds as reference size.
 NSO(L50 (target)) using the expected genome size as reference (105.2 Mb).

NSO and L50, which indicate the total number and minimum length, respectively, of all scaffolds that together account for 50% of the genome.

Schneeberger et al. 2011

Table 2. Assembly validation

	Ler-1 (MN2010)	C24 (MN2010)	Bur-0 (MN2010)	Bur-0 (shotgun)
Sanger reads	1,139	1,139	1,110	955
Organelle/centromere hits	48	48	49	267
No significant hits	12	4	6	52 (30)*
Euchromatic hits	1,079	1,087	1,055	658
Identical	1,069	1,074	1,046	629
With mismatching bases	6	9	4	17
With indels in simple repeats	2	4	4	4
With indels (up to 476 bp)	2	0	1	8
Nucleotides queried, kb	580	584	563	285
No. mismatching bases	11	14	8	22

*Fifty-two reads were blasted against NCBI nonredundant database. Thirty reads did not feature alignments that were related to rDNA or human DNA.