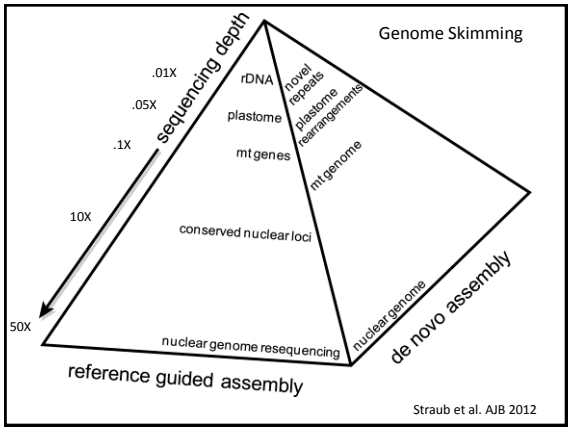
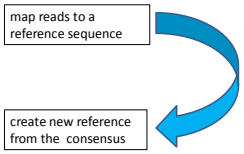
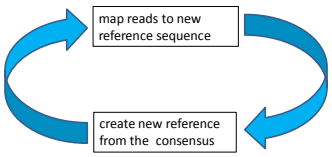


Reference Guided Assembly



Reference Guided Assembly



Repeat until no further improvement

Machine learning = A branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data. Wikipedia

Assembly Pipeline

AlignReads (Straub et al. 2011)

A five-stage automated pipeline for assembly of short reads.

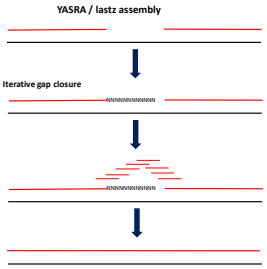
1) YASRA (Ratan 2009)

2) NUCMer (Kurtz et al. 2004)

3) Delta Filter (Kurtz et al. 2004)

4) Sumqual (Straub et al. 2011)

5) Qualtofa (Straub et al. 2011)



Assembly Pipeline

AlignReads (Straub et al. 2011)

A five-stage automated pipeline for assembly of short reads.

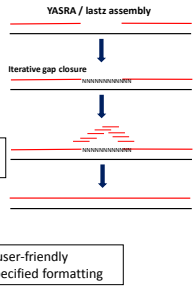
1) YASRA (Ratan 2009) — iterative assembler

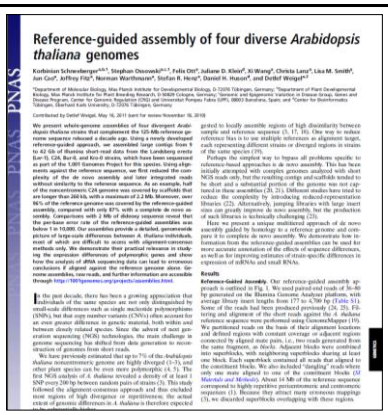
2) NUCMer (Kurtz et al. 2004) — align contigs to reference

3) Delta Filter (Kurtz et al. 2004) — align contigs to reference

4) Sumqual (Straub et al. 2011)

5) Qualtofa (Straub et al. 2011) — convert to user-friendly and user-specified formatting





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

PNAS

SUPERLOCAS Klein et al. 2011

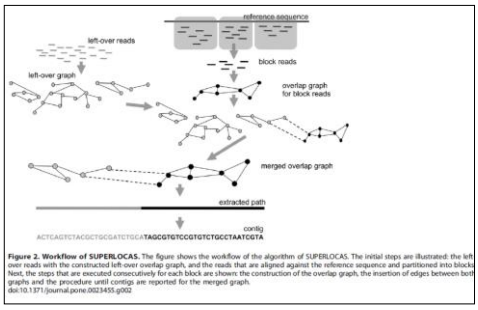


Figure 2. Workflow of SUPERLOCAS. The figure shows the workflow of the algorithm of SUPERLOCAS. The initial steps are illustrated: the left-over reads with the constructed left-over overlap graph, and the reads that are aligned against the reference sequence and partitioned into blocks. Next, the steps that are executed consecutively for each block are shown: the construction of the overlap graph, the insertion of edges between both graphs and the procedure until contigs are reported for the merged graph.
doi:10.1371/journal.pone.0023455.g002

Schneeberger et al. 2011

Table 1. Assembly statistics

	Bur-0	C24	Kro-0	Ler-1
Coverage	83.2×	75.0×	72.7×	322.4×
Libraries	2	2	2	3
N50 (intrinsic)	193	109	161	113
L50, kb	147.3	273.2	163.5	272.5
N50 (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

N50/L50 (intrinsic) using total length of the scaffolds as reference size.
N50/L50 (target) using the expected genome size as reference (105.2 Mb).

N50 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.