Supporting Information

Schneeberger et al. 10.1073/pnas.1107739108

SI Materials and Methods

Sample preparation for Illumina sequencing. Plant DNA as well as single- and paired-end libraries were prepared as described (1, 2). *Arabidopsis* Biological Resource Center (ABRC) stock numbers for the *Arabidopsis thaliana* accessions L*er*-1, Bur-0, C24, and Kro-0 are CS22686, CS22679, CS22680, and CS1301, respectively. Mate-pair libraries were prepared with kits no. 1004876 and no.1005363 preRelease (Illumina) according to manufacturer's instructions, but with 5 psi pressure and 5-s duration for the first nebulization, and 35 psi/6 s for the second. Approximately 5-kb fragments were purified in the first size selection step. Data for seven flow cell lanes of the Bur-0 paired-end library were kindly provided by Illumina.

Short Read Mapping and Consensus Analysis. Short read alignment followed by a consensus analysis was used at three different stages within the assembly. First, the read partitioning was based on short read alignments against the reference sequence. Second, short reads were aligned against supercontigs for assembly correction and scaffolding. Third, short reads were aligned against the final scaffolds for per-base quality assessment and filtering.

For each such alignment-consensus analysis we used the short read analysis pipeline SHORE with GenomeMapper as alignment tool. Within the alignment we allowed for at most 10% of the positions of a read to mismatch, including 7% being involved in gaps. Repetitive alignments were removed if another alignment of the same read in combination with an alignment of the read pair was more likely to resemble the sequenced clone (paired-end correction). Base calling was performed using SHORE's quality metric for homozygous variation. For the third analysis we additionally allowed base calling in repetitive positions.

Blocks, Superblocks, and Initial Alignment. On the basis of the initial alignment we partitioned the reads according to their alignment locations. For this we defined regions with contiguous read coverage as blocks. A block ends at any region without read alignments. If such a region was spanned with discordantly mapped read pairs we expanded the block up to the next region with absence of read coverage. The rational behind this is the assumption that the discordantly mapped read pairs indicate the coherence of the blocks in the focal genome, and thus there is no need to split them. On the basis of these nonoverlapping blocks we define superblocks, by joining two or more adjacent blocks until their combined length reaches a minimum length of 12 kb. The superblocks are built up in an overlapping manner such that a minimal overlap length of 300 bp is shared between two neighboring superblocks (Fig. 1). For each noncentromeric superblock we gathered all reads mapped to comprised blocks in addition to all of the unmapped reads with a mate pair mapped to one of these blocks, called dangling reads from here on. Note reads with multiple mapping locations can be members of multiple blocks and superblocks. Each set of reads was used as input for the short read assembly tools VELVET, ABYSS, and EULER-SR.

To incorporate not only the leftover reads with a mapped mate, but all unalignable read pairs, we applied SUPERLOCAS. This short read assembly tool initially builds up one assembly graph of all unmapped reads, called leftover graph. Subsequently assembly graphs for each superblock are generated separately and linked into the leftover graph, to allow incorporation of unmapped reads as long as they have high quality overlaps with conserved regions of blocks. After the contigs of a superblock assembly have been successfully elongated, the superblock assembly graph is discarded and the reads of the next superblock will be subject to the same procedure. This presents a computationally feasible solution compared with assembling all leftover reads over and over within the assembly of each superblock. Incorporating unalignable read pairs also allows for assembly of diverged regions between blocks longer than twice the insert size of the sequenced samples. Furthermore we used VELVET to assemble all read pairs where one or both mates could not be mapped (leftover reads and dangling pairs). The resultant contigs are expected to originate in part from large insertions or highly diverged regions and can subsequently be used to bridge remaining gaps between blocks.

Running AMOScmp. We used AMOScmp (version 2.0.8) to assemble the contigs that were produced by the short read assembly tools. This program allows removing redundancy inherent in the contig assemblies. Contigs were separated by chromosome arm and assembled using the respective chromosome arm reference sequence as homology target. Within the AMOScmp script we executed all programs with default values except for casm-layout using parameter *–t* 3,500 (maximum ignorable trim length) and *make-consensus* using parameter *–o* 10 (minimum overlap bases).

Running BAMBUS. All read pairs that align to two different supercontigs define a connection (bridge) between the respective supercontigs, suggesting that these two supercontigs, although not assembled together, are in local vicinity in the focal genome and have a defined order. In addition we used MUMmer/nucmer to infer links between supercontigs based on nucmer alignments to the reference sequence, as described in the BAMBUS manual. For this step we only allowed for anchor matches that are unique in the reference and postfiltered the resulting MUMmer links removing links connecting supercontigs with opposite alignment orientation or more than 10 kb inferred distance relative to the reference. Furthermore the contig order and orientation proposed by MUMmer links was not allowed to disagree with the contig layout proposed by AMOScmp in the previous step.

Any supercontig providing at least five bridges to other supercontigs is classified as essential. Next, essential supercontigs smaller than 50 bp and nonessential supercontigs smaller than 100 bp and, further, essential supercontigs with more than 1 error per 200 bp and nonessential supercontigs with more than 1 error per 1,000 bp are removed.

We ran BAMBUS (version 2.33) with default parameter setting as described in the manual, in detail we started goBambus, untangle, and printScaff. The configuration file for goBambus was set to require at least six bridges for paired-end and mate pair libraries and the preferedBridges was set to equal or larger than 10. printScaff was run with the parameter –nomerge to prevent the concatenation of the contigs with 60 bp. Instead we calculated the most likely number of positions between contigs and introduced that many N's.

By default, BAMBUS incorporates 60 N's between contigs neighbored in a scaffold to report one sequence per scaffold. The read pairs aligning to these two different contig might suggest a different distance, however. Thus, we calculated the most likely distance between the two contigs on the basis of all pairs aligning to both contigs. First, we calculated the insert size distribution for each of the sequencing libraries on the basis of unique alignments of read pairs to one single contig. This was directly translated into a probability distribution for clone lengths. On the basis of this probability distribution and the

alignment locations, we calculated the most likely number of N's. If there were reads from multiple libraries spanning the same contigs usually the most likely number of N's did not match. Thus, we prioritized the paired-end libraries over the mate pairs, as their SD was much smaller. Conflicting read pairs within a library were excluded as well.

After the first run of BAMBUS we manually checked all connections between contigs for spurious connections and removed ~10 per chromosome arm. Afterward we ran BAMBUS a second time now permitting these connections.

Comparison with a Standard Alignment-Consensus Approach. We performed a standard resequencing analysis on all four strains to analyze the difference between the reference-guided assembly and the alignment-consensus methods, comparing both the contig sizes, genome coverage, and the resultant polymorphism calls. We used the same set of reads as for the assembly and applied SHORE's resequencing pipeline using GenomeMapper as alignment tool. We allowed for 10% and 7% of the nucleotide of a read to mismatch and or to gap, respectively. Concatenating adjacent base calls (including reference, SNP, and microindel calls) generated the alignment-consensus contigs.

Running MUMmer to Generate Whole-Genome Alignment. We used the MUMmer whole-genome alignment tool to align all scaffolds of each assembly to the reference sequence. We followed the instructions for "Mapping a draft sequence to a finished sequence" (http://mummer.sourceforge.net/manual/#mappingdraft). For this we ran *nucmer* using a parameter setting favoring specificity over sensitivity ("*nucmer–mum -b 100 -g 90 -l 35 -c 80 -f–prefix= outputFolder referenceSquence assemblySequence*"), where the *–f* parameter was omitted when aligning de novo assemblies. Therefore, we only allowed for alignment anchors that were unique in both the reference and query. Further we allowed nucmer to extend alignments across poor scoring regions by maximally 100 edit distance, whereas longer diverged regions or indels larger than 50 bp always lead to an alignment break. Finally we increased nucmer's default values for minimum length of a single match and a cluster of matches and restricted the alignment to matches of the forward strand of the query.

The reasoning behind using strict alignment parameters is that relaxed alignments tend to produce false positives due to aligning regions that are not orthologous to each other. Long indels can nonetheless be accurately defined by annotating the alignment breakpoints and the distance between high-scoring segment pairs (HSPs).

Resultant scaffold to reference alignments were parsed to retrieve SNPs, insertions, and deletions without any further filtering except that ambiguous insertions featuring more then 10% N's were removed. Additionally we analyzed alignments with multiple HSPs by annotating the alignment breaks (gaps between HSPs) to distinguish between simple deletions or insertions, highly diverged regions, and spurious alignments in repetitive regions. Therefore, a deletion was defined if more than 20 bp of the reference sequence are not matched by scaffold sequence, whereas the scaffold sequence could be fully aligned to the HSPs upstream and downstream of the break. Vice versa, an insertion is defined if more than 20 bp of scaffold sequence is not matched by reference sequence. Finally we defined a highly diverged region (HDR) if more than 20 bp from both reference and scaffold could not be aligned against each other, thus the break between the HSPs represents diverged but not deleted alleles in the reference and the analyzed strain. The last category includes all spurious alignments, e.g., negative distance between alignment breakpoints (overlapping HSP alignments indicating wrongly aligned or assembled repeats), and was removed from further analysis. Table S4 shows a complete overview of all variation found in the four strains using the assembly and whole-genome alignment approach compared with variation found with the alignment-consensus approach.

Annotation of Polymorphisms. All polymorphisms overlapping exons were characterized as either major (deleterious) or minor changes. Deleterious changes encompass long indels and HDRs as well as microindels causing a frameshift or SNPs introducing or removing a stop codon. Microindels changing the length of the coding sequence by a factor of three (including multiple compensating indels in the same gene) are classified as minor changes as are any amino acid changes except for stop mutations. Genes not featuring any mutation or only synonymous SNPs are classified as conserved.

Sample Preparation for Expression Analysis. The *A. thaliana* (Bur-0 and C24) sRNA data were generated from inflorescences including flowers up to stage 14 grown at 23 °C and in a 16-h light period. Libraries were constructed as described (3), except that sRNAs were isolated from a 15% denaturing polyacrylamide gel, and RNA amplicons were reverse transcribed using the Revertaid kit (Fermentas) before PCR amplification using the Phusion polymerase (Finnzymes). For tiling array analysis, probes were synthesized from RNA extracted from inflorescences as for sRNAs and hybridized to Affymetrix GeneChip *Arabidopsis* Tiling 1.0R arrays. Triplicate biological replicates were used, and array data were analyzed to generate RMA expression summaries.

1. Ossowski S, et al. (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res* 18:2024–2033.
2. Mirouze M, et al. (2009) Selective epigenetic control of retrotransposition in *Arabidopsis*. *Nature* 461:427–430.
3. Mosher RA, et al. (2009) Uniparental expression of PolIV-dependent siRNAs in developing endosperm of *Arabidopsis*. *Nature* 460:283–286.

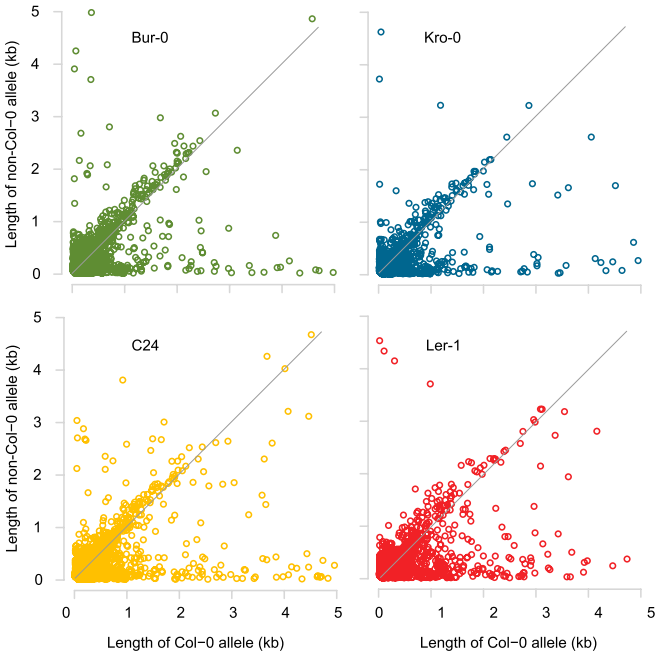

Fig. S1. Allele length comparisons of highly diverged regions (HDRs).

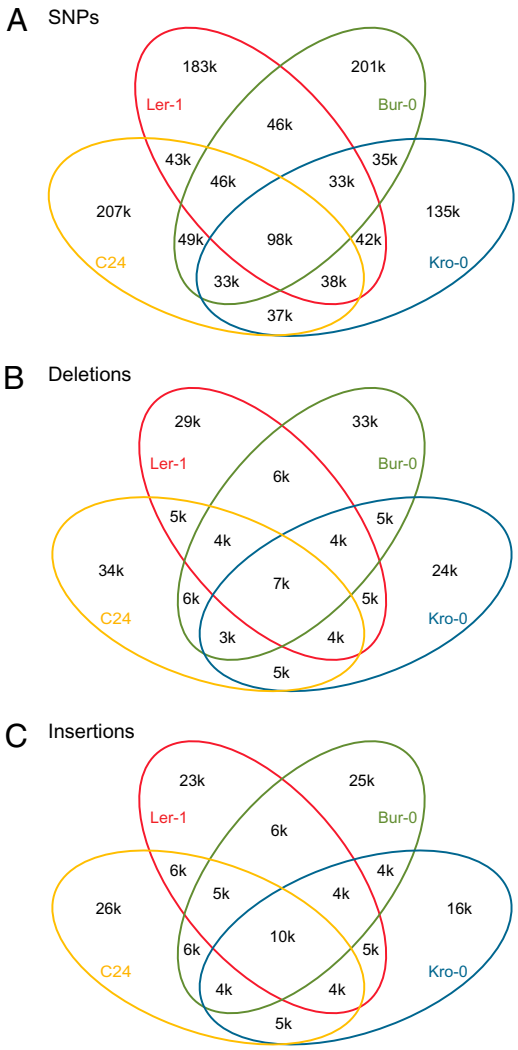

Fig. S2. Shared SNPs (*A*), deletions (*B*), and insertions (*C*) relative to Col-0 reference.

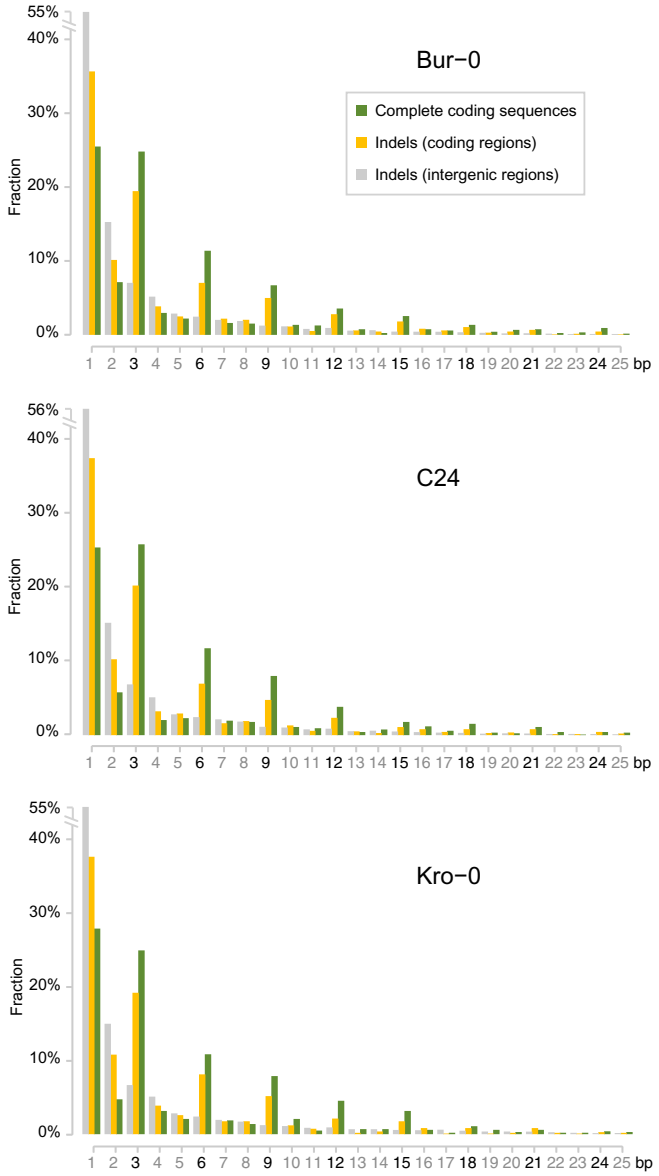

Fig. S3. Length variation in coding sequences, relative to Col-0 reference.

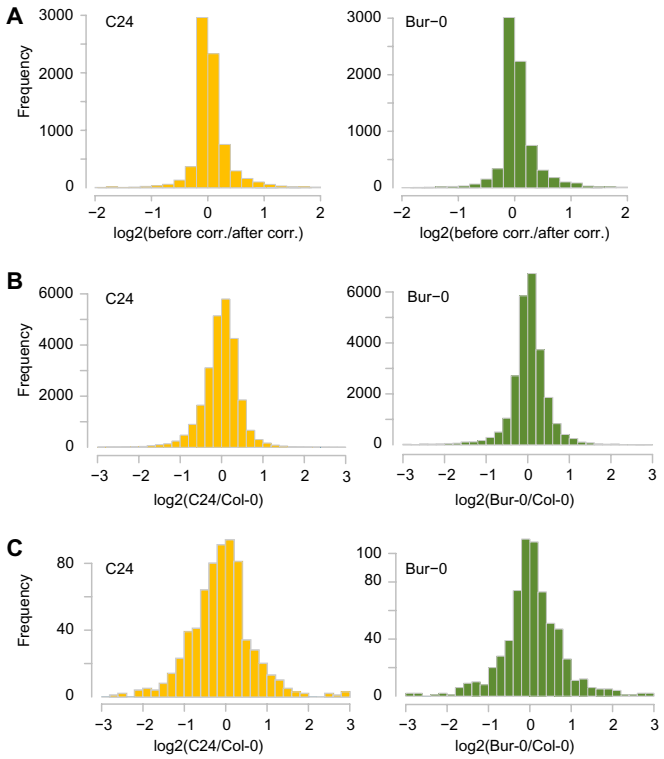

Fig. S4. Tiling array expression analysis. (*A*) Effect of probe correction on expression estimates for 7,056 genes for which half or more of all probes were removed. Note that the distribution is skewed toward the estimates being higher after correction. (*B*) Expression of conserved genes (at least 97.5% of exonic nucleotides conserved between Col-0, Bur-0, and C24). (*C*) Expression of polymorphic genes (at least 2.5% of exonic nucleotides differ between Col-0, Bur-0, and C24).

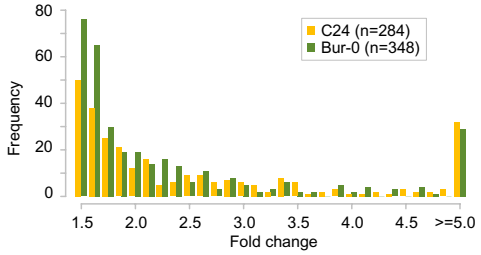

Fig. S5. sRNA expression analysis. Increase in expression estimates for distinct sRNA loci resulting from incorporating information from genome assemblies.

Table S1. Read statistics

	Bur-0	C24	Kro-0	Ler-1
		Single end		
Reads	142,532,346	27,033,381	4,443,603	10,076,255
Mb	5,118.6	1,113.2	183.8	550.0
Coverage	42.7x	9.3x	1.5x	4.6x
		Paired end (library 1)		
Pairs	55,811,985	89,737,786	91,624,757	189,763,954
Avg. insert size	187	185	177	178
SD	24	27	17	23
Mb	4,094.9	7,210.9	8,124.6	26,774.8
Coverage	34.1x	60.1x	67.7x	223.1x
		Paired end (library 2)		
Pairs	—	—	—	84,223,339
Avg. insert size	—	—	—	458
SD				45
Mb	—	—	—	10,803.5
Coverage	—	—	—	90.0x
		Mate pair*		
Pairs	9,676,627	9,319,898	5,900,939	4,169,512
Avg. insert size	3795	4617	4700	3711
SD	508	920	571	477
Mb	770.2	671.0	424.9	564.0
Coverage	6.4x	5.6x	3.5x	4.7x

*Including potential clonal events.

Table S2. Comparison of alignment-consensus and assembly-derived contigs

	Bur-0			C24			Kro-0			Ler-1		
	CA	AS	sAS	CA	AS	sAS	CA	AS	sAS	CA*	AS*	sAS*
N50 (intrinsic)	6,563	193	185	6,154	109	105	6,831	161	154	4,405	113	108
L50, kb	3.7	147.3	147.1	4.0	273.2	273.7	3.6	163.5	167.3	5.7 kb	272.5	270.8
N50 (target)	7,788	208	216	7,265	117	119	8,011	178	181	5,016	121	126
L50, kb	3.3	139.7	135.0	3.5	260.4	251.2	3.2	151.8	145.6	5.2	261.9	246.5
Scaffolds	145,683	2,526	2,143	138,438	2,052	1,740	160,535	2,670	2,408	104,403	1,528	1,261
Total length, Mb	96.7	101.0	96.5	96.8	101.3	98.1	97.3	99.9	96.7	98.6	100.8	96.3
Longest scaffold	59 kb	1.12 Mb	1.12 Mb	64 kb	2.18 Mb	2.18 Mb	51 kb	1.48 Mb	1.48 Mb	88 kb	1.09 Mb	1.09 Mb
Ambiguous bases, %	0.0	4.03	8.30	0.0	3.60	6.81	0.0	5.10	8.12	0.0%	1.3%	8.53%

CA, consensus-alignment approach; AS, assembly; sAS, stringently masked assembly.

Table S3. Comparison of accessibility, SNPs, deletions, and insertions obtained by assembly and alignment-consensus approach, respectively

	Accessibility (Mb, %)	SNPs	Microdeletion	Microinsertion
		Assembly		
Bur-0	101.0 (96)	541,713	52,429	49,421
C24	101.3 (96.3)	552,177	53,157	50,596
Kro-0	99.9 (94.9)	451,928	43,847	40,659
Ler-1	100.8 (95.8)	530,081	50,230	49,025
		Consensus Q25		
Bur-0	93.9 (89.2)	487,550	37,231	38,136
C24	94.1 (89.4)	484,757	37,340	37,035
Kro-0	94.4 (89.7)	391,301	32,203	31,271
Ler-1	93.7 (89.1)	478,925	47,902	47,731
		Overlap		
Bur-0	N/A	440,254	31,815	30,553
C24	N/A	439,990	32,457	31,002
Kro-0	N/A	355,170	27,159	26,005
Ler-1	N/A	426,107	36,247	35,658

Table S4. Variants of different lengths relative to reference

	Bur-0					
	Deletions		Insertions		HDRs	
Variant length (bp)	*n*	Length (bp)	*n*	Length (bp)	*n*	Length (bp)
1	36,694	36,694	34,573	34,573		
2	10,423	20,846	10,135	20,270		
3–4	8,858	30,120	7,859	26,723		
5–8	6,354	40,067	5,438	34,134		
9–16	4,334	50,323	3,274	37,834		
17–32	1,827	40,533	1,166	26,539	70	1,756
33–64	762	34,564	481	22,113	180	8,727
65–128	350	30,748	291	25,665	358	33,830
129–256	241	44,852	85	14,463	352	64,532
257–512	210	75,280	50	17,652	282	101,696
513–1,024	234	174,908	13	7,542	199	139,709
1,025–2,048	163	228,097	1	2,038	106	152,209
>2,048	189	1,087,692	3	31,965	64	327,035
	Kro-0					
	Deletions		Insertions		HDRs	
Variant length (bp)	*n*	Length (bp)	*n*	Length (bp)	*n*	Length (bp)
1	31,032	31,032	28,571	28,571		
2	8,592	17,184	8,031	16,062		
3–4	7,082	24,105	6,677	22,651		
5–8	5,141	32,314	4,583	28,824		
9–16	3,713	43,569	2,789	32,032		
17–32	1,971	43,522	946	21,576	54	1,397
33–64	554	25,448	479	22,331	154	7,480
65–128	279	23,974	236	20,369	310	28,789
129–256	215	40,143	92	15,307	254	47,401
257–512	174	63,710	21	6,990	232	83,324
513–1,024	188	139,313	6	3,319	145	105,935
1,025–2,048	112	157,612	5	5,851	80	112,926
>2,048	162	949,727	3	111,514	60	326,617
	C24					
	Deletions		Insertions		HDRs	
Variant length (bp)	*n*	Length (bp)	*n*	Length (bp)	*n*	Length (bp)
1	37,595	37,595	35,206	35,206		
2	10,355	20,710	10,457	20,914		
3–4	8,714	29,649	8,346	28,451		
5–8	6,367	40,117	5,633	35,522		
9–16	4,225	49,338	3,496	40,555		
17–32	1,851	41,941	1,216	27,618	71	1,815
33–64	720	32,754	601	27,530	176	8,547
65–128	401	34,927	306	27,125	396	36,878
129–256	251	46,516	153	25,993	415	76,388
257–512	209	76,249	64	22,170	337	119,336
513–1,024	248	180,813	20	12,599	258	187,392
1,025–2,048	186	262,125	4	5,884	124	173,796
>2,048	185	911,483	10	186,588	119	804,913

For HDRs we listed the length of the Col-0 allele (see main text for definition of HDR). Whole-genome alignment was adjusted to align through regions of high divergence being shorter than 20 bp.

Table S5. Inversions

Accession	Chr.	Begin	End	Length	Scaffold	Begin	End	Length	Affected gene
Ler-1	1	1,771,291	1,771,586	295	17	42,411	42,116	295	AT1G05870
Bur-0	1	16,614,139	16,614,251	112	729	11,212	11,100	112	—
C24	1	20,333,803	20,334,503	700	763	193,870	193,170	700	—
Ler-1	1	27,858,219	27,858,368	149	600	412,916	412,768	148	—
C24	1	27,858,227	27,858,368	141	921	156,825	156,684	141	—
Kro-0	1	27,858,237	27,858,368	131	1,166	52,004	51,873	131	—
Ler-1	2	10,153,298	10,153,423	125	1,010	53,559	53,434	125	—
Ler-1	2	17,617,485	17,617,698	213	1,081	54,963	54,756	207	AT2G42270
C24	3	3,365,964	3,366,120	156	1,937	581,455	581,299	156	—
Bur-0	3	6,620,232	6,620,327	95	2,268	243,613	243,518	95	—
Ler-1	3	8,081,491	8,081,607	116	1,234	249,654	249,538	116	—
C24	3	15,381,184	15,381,316	132	2,368	21,136	21,004	132	—
Bur-0	3	15,778,040	15,778,871	831	2,769	17,065	16,239	826	—
Kro-0	3	15,942,027	15,942,108	81	2,872	60,388	60,307	81	—
Ler-1	3	18,124,079	18,125,014	935	1,603	9,335	8,403	932	AT3G48840
Kro-0	3	18,124,430	18,125,014	584	3,028	9,709	9,125	584	AT3G48840
Bur-0	4	5,212,688	5,212,878	190	3,319	33,720	33,530	190	—
C24	4	7,555,219	7,555,851	632	3,086	7,637	7,005	632	—

Shaded rows highlight inversions that have been identified in more than one strain.