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

Korbinian Schneeberger[a,b,1], Stephan Ossowski[a,c,1], Felix Ott[a], Juliane D. Klein[d], Xi Wang[a], Christa Lanz[a], Lisa M. Smith[a], Jun Cao[a], Joffrey Fitz[a], Norman Warthmann[a], Stefan R. Henz[a], Daniel H. Huson[d], and Detlef Weigel[a,2]

[a]Department of Molecular Biology, Max Planck Institute for Developmental Biology, D-72076 Tübingen, Germany; [b]Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, D-50829 Cologne, Germany; [c]Genomic and Epigenomic Variation in Disease Group, Genes and Disease Program, Center for Genomic Regulation (CRG) and Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain; and [d]Center for Bioinformatics Tübingen, Eberhard Karls University, D-72076 Tübingen, Germany

Contributed by Detlef Weigel, May 16, 2011 (sent for review November 18, 2010)

We present whole-genome assemblies of four divergent *Arabidopsis thaliana* strains that complement the 125-Mb reference genome sequence released a decade ago. Using a newly developed reference-guided approach, we assembled large contigs from 9 to 42 Gb of Illumina short-read data from the Landsberg *erecta* (L*er*-1), C24, Bur-0, and Kro-0 strains, which have been sequenced as part of the 1,001 Genomes Project for this species. Using alignments against the reference sequence, we first reduced the complexity of the de novo assembly and later integrated reads without similarity to the reference sequence. As an example, half of the noncentromeric C24 genome was covered by scaffolds that are longer than 260 kb, with a maximum of 2.2 Mb. Moreover, over 96% of the reference genome was covered by the reference-guided assembly, compared with only 87% with a complete de novo assembly. Comparisons with 2 Mb of dideoxy sequence reveal that the per-base error rate of the reference-guided assemblies was below 1 in 10,000. Our assemblies provide a detailed, genomewide picture of large-scale differences between *A. thaliana* individuals, most of which are difficult to access with alignment-consensus methods only. We demonstrate their practical relevance in studying the expression differences of polymorphic genes and show how the analysis of sRNA sequencing data can lead to erroneous conclusions if aligned against the reference genome alone. Genome assemblies, raw reads, and further information are accessible through http://1001genomes.org/projects/assemblies.html.

In the past decade, there has been a growing appreciation that individuals of the same species are not only distinguished by small-scale differences such as single nucleotide polymorphisms (SNPs), but that copy number variants (CNVs) often account for an even greater difference in genetic material, both within and between closely related species. Since the advent of next generation sequencing (NGS) technologies, the main challenge in genome sequencing has shifted from data generation to reconstruction of genomes from short reads.

We have previously estimated that up to 7% of the *Arabidopsis thaliana* noncentromeric genome are highly diverged (1–3), and other plant species can be even more polymorphic (4, 5). The first NGS analysis of *A. thaliana* revealed a density of at least 1 SNP every 200 bp between random pairs of strains (3). This study followed the alignment-consensus approach and thus excluded most regions of high divergence or repetitiveness; the actual extent of genomic differences in *A. thaliana* is therefore expected to be substantially higher.

Prediction methods for structural variants (SVs) have been developed to annotate diverged regions on the basis of paired-end sequencing (6–16). Unfortunately, these predictions do not include the actual sequence of the variants, and they often miss larger rearrangements, complex changes, and small insertion/deletions (indels). In addition, regions of similar length but with dissimilar sequence will not reveal themselves in the form of paired alignments with unexpected distance or orientation to each other. To overcome these shortcomings, it has been suggested to locally assemble regions of high dissimilarity between sample and reference sequence (3, 17, 18). One way to reduce reference bias is to use multiple references as alignment target, each representing different strains or diverged regions in strains of the same species (19).

Perhaps the simplest way to bypass all problems specific to reference-based approaches is de novo assembly. This has been initially attempted with complex genomes analyzed with short NGS reads only, but the resulting contigs and scaffolds tended to be short and a substantial portion of the genome was not captured in these assemblies (20, 21). Different studies have tried to reduce the complexity by introducing reduced-representation libraries (22). Alternatively, jumping libraries with large insert sizes can greatly improve de novo assembly, but the production of such libraries is technically challenging (23).

Here we present a unique multitiered approach of de novo assembly guided by homology to a reference genome and compare it to complete de novo assembly. We demonstrate how information from the reference-guided assemblies can be used for more accurate annotation of the effects of sequence differences, as well as for improving estimates of strain-specific differences in expression of mRNAs and small RNAs.

Results

Reference-Guided Assembly. Our reference-guided assembly approach is outlined in Fig. 1. We used paired-end reads of 36–80 bp generated on the Illumina Genome Analyzer platform, with average library insert lengths from 177 to 4,700 bp (Table S1). Some of the reads had been produced previously (24, 25). Filtering and alignment of the short reads against the *A. thaliana* reference sequence were performed using GenomeMapper (19). We partitioned reads on the basis of their alignment locations and defined regions with constant coverage or adjacent regions connected by aligned mate pairs, i.e., two reads generated from the same fragment, as *blocks*. Adjacent blocks were combined into superblocks, with neighboring superblocks sharing at least one block. Each superblock contained all reads that aligned to the constituent blocks. We also included “dangling” reads where only one mate aligned to one of the constituent blocks (*SI Materials and Methods*). About 14 Mb of the reference sequence correspond to highly repetitive pericentromeric and centromeric sequences (1). Because they attract many erroneous mappings (3), we discarded superblocks overlapping with these regions.

Author contributions: K.S., S.O., D.H.H., and D.W. designed research; C.L., L.M.S., J.C., and N.W. performed research; K.S., S.O., F.O., J.D.K., X.W., J.F., and S.R.H. analyzed data; and K.S., S.O., and D.W. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]K.S. and S.O. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: weigel@weigelworld.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1107739108/-/DCSupplemental.

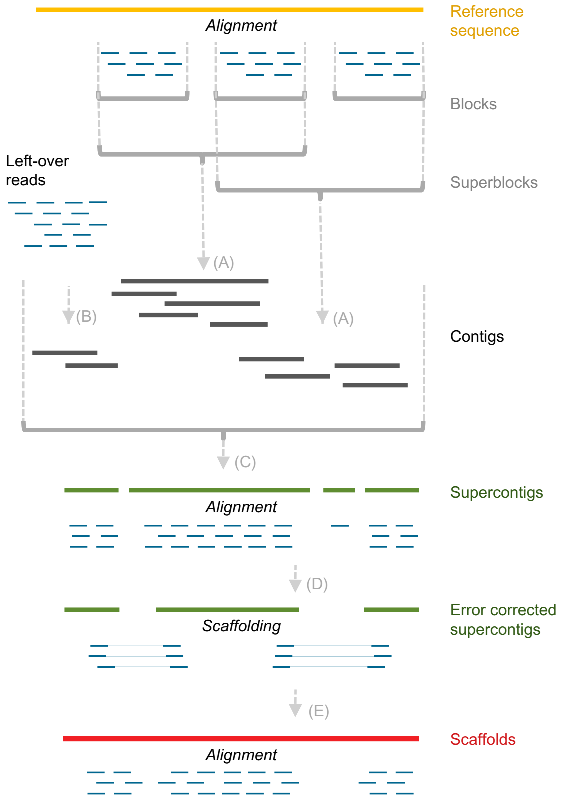

Fig. 1. Illustration of reference-guided assembly. Reads and their alignments are shown in blue. Regions of constant coverage were defined as blocks. Adjacent blocks were combined into superblocks until they reached a minimal length of 12 kb. Superblocks were defined in an overlapping fashion, such that blocks could belong to several superblocks. All reads of a superblock were assembled with reads that had not been aligned. Resulting contigs (dark blue) were merged into a nonredundant set of supercontigs (green). Short read alignments against the supercontigs allowed for error correction and scaffolding. Short read alignments against the scaffolds (red) enabled a final quality assessment and filtering.

Reads corresponding to each superblock were assembled separately using the de Bruijn graph-based tools ABYSS (26), VELVET (27), and EULER-SR (28) (Fig. 1*A* arrows). We executed each tool eight times per read set using eight different kmer sizes. The local assemblies were combined with an assembly of all leftover reads using SUPERLOCAS (*SI Materials and Methods*). Finally, we assembled all unmapped reads (including pairs with a dangling read) de novo with VELVET, to separately reconstruct long stretches of nonreference sequence (Fig. 1*B* arrow).

As single reads can contribute to different assemblies our workflow introduces redundancy into the contigs. As this redundancy generates overlaps between the contigs we used the homology guided Sanger assembler AMOScmp (29) to merge all contigs of each chromosome arm into a set of nonredundant supercontigs (Fig. 1*C* arrow).

To validate the supercontigs, we aligned all original reads against these. Consistent differences between supercontigs and reads were taken as indications of misassemblies. With this information we corrected or, in the most extreme cases, removed supercontigs (Fig. 1*D* arrow). Read pairs with ends that aligned to different supercontigs were used for scaffolding with BAMBUS (30) (Fig. 1*E* arrow).

BAMBUS scaffolds were used as alignment target for a third and final round of read mapping and consensus analysis. We developed a simple metric to assign per-base quality values according to the base qualities of the consensus analysis. Scaffolds shorter than 500 high-quality base pairs were discarded, and low-quality positions were masked. Additionally we ran a more stringent base masking to produce a high-quality, although less comprehensive assembly. Both types of assemblies can be downloaded at http://1001genomes.org/projects/assemblies.html.

Assembly Statistics. We used a mix of single-end, paired-end, and mate-pair libraries for the genomes of the four *A. thaliana* strains Landsberg *erecta* (L*er*-1), C24, Burren (Bur-0), and Krotzenburg (Kro-0) (Table S1). Total coverage was greater than 70× for all and greater than 320× for L*er*-1. Common metrics to assess assembly quality are N50 and L50, which indicate the total number and minimum length, respectively, of all scaffolds that together account for 50% of the genome. After exclusion of centromeric regions, we had targeted sequences that correspond to around 105 Mb of the reference sequence. Based on this sequence, N50 and L50 for our assemblies ranged from 117 to 208, and 140 to 262 kb, respectively, with the longest scaffold in each assembly being between 1.1 and 2.2 Mb (Table 1). For comparison, contigs derived from concatenating consecutively called positions of an alignment consensus yielded N50 and L50 values of 6,147 and 4.1 kb, respectively, for L*er*-1 (Table S2). The cumulative lengths of all scaffolds were about 5% shorter than the target of 105 Mb. This was mainly because of failure to assemble repetitive sequences, as indicated by the fact that the nonpericentromeric portions of the reference sequence not covered by assembly contigs had a 36–41% repeat content, compared with an average of 8% in noncentromeric regions, as assessed with 36-mers (3).

Assembly Validation with 2 Mb of Dideoxy Data. To assess the quality and error rate of the assemblies, we used 955 random Sanger reads of the Bur-0 genome (*SI Materials and Methods*; the "shotgun" set), and a published set of 3,388 fragments of the C24, L*er*-1, and Bur-0 genomes produced by targeted Sanger resequencing of mostly unique, genic fragments (31) (the "MN2010 set"). Sanger reads were aligned against the respective assembly and the pericentromeric portions of the reference (32) (Table 2). About 4% of MN2010 and 28% of shotgun fragments aligned to organelles or pericentromeric regions. Our assemblies agreed very well with the remaining MN2010 data. At least 98.8% aligned uniquely and without any mismatch, and only 0.4–0.8% had mismatches. An additional 0.2–0.4% revealed short indel errors, all of which were associated with low sequence complexity including simple repeats. Only three MN2010 reads revealed long indel errors not associated with simple repeats, of up to 476 bp. The total per-base error measured across all MN2010 alignments (excluding the three with long indel errors) was less than 1 in 40,000 bp.

Table 1. **Assembly statistics**

	Bur-0	C24	Kro-0	L*er*-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).

Table 2. Assembly validation

	L*er*-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.

The per-base error estimate with the shotgun set for Bur-0 was higher, but still less than 1 in 10,000 bp. Eight reads out of 658 revealed long indel errors. This was not unexpected, as the shotgun set was randomly sampled from the genome and included more intergenic and repetitive sequences, which are more difficult to assemble. In addition, the shotgun reads had not been subjected to similarly extensive manual curation as the MN2010 set and were thus likely to contain more errors themselves.

We compared all reads of the shotgun set without significant BLAST hit (E value $< e^{-10}$) against National Center for Biotechnology Information (NCBI)'s nonredundant database (32). Twenty-one of 52 reads corresponded to rDNA, and one was the result of contamination with human DNA. The remaining reads, or 4.4% of all reads excluding organelles, centromeres, and contamination, present an upper boundary for the "unassembled space." This is in agreement with the total scaffold length of 96.2% of the target (Tables 1 and 2), and less than what had been estimated to be inaccessible using alignment-consensus analysis (3).

Sequence Assemblies Capture Large-Scale Variations. To determine the extent of large-scale sequence differences captured in the assemblies, we performed whole-genome alignments against the reference genome with MUMmer (33), using parameters that favored correctly placed alignments over sensitivity. The portion of the reference genome that could not be aligned against our assemblies was as low as 3.7%, whereas in the best case for the alignment-consensus approach, at least 10.3% of the reference could not be aligned (Table S3). In aligned regions, we annotated SNPs, indels, and highly diverged regions (HDRs) that are anchored within the whole-genome alignment by flanking sequences (*SI Materials and Methods*).

There was good concordance between SNPs and microindels (1–3 bp) predicted on the basis of either the whole-genome alignments or by the alignment-consensus approach (Table S3). The assemblies, however, revealed more small-scale changes: On average, 12% more SNPs, 29% more microdeletions, and 23% more microinsertions.

We also analyzed the length distributions of apparent deletions and insertions relative to the reference and HDRs (Table 3 for L*er*-1, Table S4 for the other strains). Over 1.7 Mb of reference sequence was missing from the L*er*-1 assembly, with the majority in deletions over 2 kb. As expected, deleted regions were significantly enriched for transposable elements (63.5%, compared with 13.7% of all noncentromeric positions). To assess the potential origin of novel, nonreference sequences, we selected 36 L*er*-1 regions that were at least 500 bp long and at least 10 times longer than the reference allele. Of these, 14 sequences shared similarity with *Arabidopsis lyrata* (34) over at least half of their lengths, indicating that the reference genome lacks sequences present in the last common ancestor of *A. thaliana* and *A. lyrata*.

Even though they were too divergent to be aligned directly, the lengths of HDR alleles were strongly correlated (Fig. S1), with an overrepresentation of HDRs with a longer reference allele. This might again be due to the difficulty of assembling long

Table 3. Variants of different lengths in L*er*-1

	Deletions		Insertions		HDRs > ~30 bp*	
Variant length (bp)	*n*	Length (bp)†	*n*	Length (bp)†	*n*	Length (bp)†
1	35,370	35,370	34,261	34,261		
2	9,861	19,722	10,060	20,120		
3–4	8,305	28,221	7,963	27,148		
5–8	5,816	36,809	5,677	35,766		
9–16	3,757	43,673	3,505	40,435		
17–32	1,824	41,552	1,238	27,800	66	1,752
33–64	663	30,310	579	26,413	165	8,133
65–128	296	26,190	340	29,810	379	35,178
129–256	219	40,825	127	21,676	406	76,128
257–512	204	74,045	63	22,600	359	129,491
513–1,024	240	176,491	20	12,823	217	155,935
1,025–2,048	160	223,702	2	3,376	138	192,553
>2,048	208	996,542	4	16,129	99	538,179

*Length in reference genome.

†Cumulative length of all variants of the class in that row.

insertions, also reflected in the smaller amount of sequence found only in one of the four new genomes, compared with reference-only sequences. Finally, regions with reverse complementary alignments revealed eighteen inversions (Table S5).

Comparison with Complete de Novo Assembly. Some of the most impressive de novo assemblies of short-read data have been produced with ALLPATHS-LG in combination with sequencing libraries that had large insert sizes (23). Because ALLPATHS-LG requires overlapping paired-end reads, we generated an additional 64.5 million 101-bp paired-end reads from one of the L*er*-1 libraries, which we knew to have an average clone length of 178 bp. We combined these new data with 8.7 million 40-bp mate-pair reads.

Different from our reference-guided assembly, the complete de novo assembly contained both noncentromeric and centromeric sequences, with a total length of 112.6 Mb in 1,705 scaffolds, compared with the 119 Mb of mostly noncentromeric reference sequence. Half of the assembly was contained in 102 scaffolds (N50) of minimum lenth 198 kb (L50). The N50 and L50 values were better than those of the reference-guided assembly, but a whole-genome alignment to the reference revealed that the scaffolds covered only 92 Mb of the noncentromeric regions. Thus, whereas only 3.7% of the noncentromeric reference genome sequence was absent from the reference-guided assembly, the complete de novo assembly lacked 12.6% of noncentromeric sequences.

Shared Polymorphisms and Their Effect on Genes. When comparing only four individuals, a large fraction of polymorphisms is expected to be found in just a single strain, and many polymorphisms that segregate at low or intermediate frequency will be missed (1–3, 31) (Fig. S2). Of 27,929 genes (excluding transposable elements and pseudogenes), over 95% could be at least partially detected in our assemblies. Over half, 55%, had at least one nonsynonymous change (Table 4). In each accession, over 3% of the genes with completely aligned sequences featured large disruptions of their coding sequence (Table 4). Partial alignments indicated that between 4.3 and 5.5% of all genes were interrupted by an HDR.

In humans, indels in coding regions occur preferentially in multiples of 3 bp, which avoids frameshifts (35). In our assemblies, 1-bp deletions were the most prevalent group, but there were distinct peaks at multiples of 3 bp, which were not seen in intergenic sequences. When considering all indels in a gene, the total variation in coding sequence length showed more pronounced peaks at multiples of 3 bp, indicating that additional indels could restore the ORF (Fig. 2 and Fig. S3). The pairwise alignments of all assembled genes can be accessed through our Web tool (http://1001genomes.org/projects/assemblies.html).

Correcting Expression Estimates for Protein-Coding Genes. Although RNA-seq is starting to eclipse microarray-based investigations of genomewide expression profile, both suffer from reliance on reference sequences, and lack of sequence conservation between individuals easily confounds expression estimates (36). We therefore investigated whether our genome assemblies would improve the interpretation of tiling array data (37) for Bur-0 and C24.

About 90% (27,607) of genes had poylmorphic probes. After probe removal, 8% (2,432) of genes could no longer be considered, because fewer than three probes had been retained. By excluding polymorphic probes, average estimates of expression levels increased slightly and were changed for many loci, especially for genes where half or more of the probes targeted polymorphic sequences (Fig. S4*A*). The variance in expression estimates for conserved genes, i.e., genes with less than 2.5% of exonic positions differing between Col-0, Bur-0, and C24, was also substantially lower than for polymorphic genes, even though the average estimates were the same (Fig. S4 *B* and *C*).

Correcting Expression Estimates for Small RNA Loci. Loci that spawn populations of small RNAs (sRNAs) are more difficult to annotate than mRNA producing loci, because they are defined by a collection of molecules. Because sRNAs are short, even small-scale differences between the focal accession and the reference will greatly affect the number of correctly mapped sRNAs.

We sequenced sRNA libraries from C24 and Bur-0 inflorescences (38). We defined sRNA loci by consecutive and overlapping alignments of sequence reads from an sRNA library and used the normalized number of reads in such segments to estimate expression of the entire locus. For Bur-0 and C24, 6.5 + 5.6 and 5.8 + 6.8 million reads from two replicate sRNA libraries, respectively, were aligned against the reference genome with one mismatch. Unaligned reads were further aligned against the Bur-0 or C24 assemblies, again allowing for one mismatch. The second step increased the alignable reads by 7%.

On the basis of the reference alignments, we defined 30,787 segments with continuous coverage of at least 10 reads from each replicate for Bur-0 and 28,174 segments for C24. Taking not only

Table 4. Functional annotation of polymorphisms in 27,929 noncentromeric genes

	Bur-0	C24	Kro-0	L*er*-1
Accessible genes	26,842	26,823	26,673	26,727
Fully aligned	23,220	23,262	23,448	23,770
Conserved	7,986	7,918	10,354	8,897
Minor change	14,320	14,438	12,306	14,007
Nonsynonymous	14,224	14,350	12,237	13,904
Deletion*	379	398	311	380
Insertion*	315	342	305	378
Major change	914	906	788	866
Deletion*	342	317	291	311
Insertion*	338	325	283	319
Stop	300	336	271	314
Stop “reversion”	99	92	73	83
Partially aligned	3,622	3,561	3,225	2,957
HDR in genes	1,369	1,540	1,212	1,461
HDR in exons	374	422	314	365

*Minor-effect indels have a length that is a multiple of 3 bp.

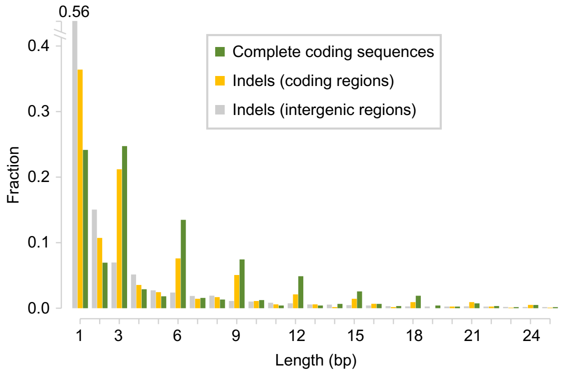

Fig. 2. Frequency of indel lengths and variation of coding sequence lengths in L*er*-1. Indels with lengths that are a multiple of three are enriched in coding regions (yellow), but not in noncoding regions (gray). This is even more apparent when comparing total length differences between orthologous coding sequences between Col-0 and Ler-1 (green). This trend can only be explained by complex changes in coding sequences that together restore the frame use. See Fig. S3 for other accessions.

the reads that aligned against the reference, but also those that aligned against the respective assembly into account, significantly changed the expression estimates of 348 segments (1.1%) for Bur-0 and 284 (1.0%) for C24 (Fig. S5). In addition, 1,283 (4.0%) of Bur-0 segments and 1,184 (4.0%) of C24 segments, could only be revealed by alignments against the nonreference genome. Finally, 579 (1.9%) of Bur-0 segments, and 556 (2.0%) of C24 segments defined by reference alignments alone were merged with neighboring segments by adding the alignments to the strain assemblies.

Discussion

Since the release of the *A. thaliana* reference genome sequence 10 y ago (39), no other whole-genome assemblies have been reported for this species. The reference genome was generated for some $70 million with a bacterial artificial chromosome (BAC)-by-BAC strategy using dideoxy sequencing (40). Differently from the first human genomes, a single individual, from the laboratory strain Col-0, was assembled.

The reference was essential for efforts to record sequence variation between natural strains of *A. thaliana* (41). However, simple alignment-based methods do not provide the complete picture to mine and exploit specieswide sequence diversity. Complete de novo assemblies would be the obvious alternative, but one of the best assembly methods available (23) resulted in significant underrepresentation of the targeted reference genome. Reasons for this might be the complexity of the *A. thaliana* genome, the length and quality of NGS reads as well as the limited insert sizes of our libraries compared with the human genome assemblies reported by Gnerre and colleagues (23). Previous work with plant genomes that achieved similar results to ours includes that of Huang and colleagues (42), who reported a whole-genome assembly of Illumina short read data for the 367-Mb cucumber genome. Whereas their L50 values were not too dissimilar from ours, they also found a larger fraction of the genome being completely missing from their assembly. Moreover, without independent validation of the cucumber results, a direct comparison with our work is difficult.

Notably, there was a limit to improving assembly statistics with additional short read data. We do not know whether this reflects an inability of the assembly tools to exploit more than about 70× coverage or whether this is an intrinsic property of read lengths and library insert sizes used and genomic repeat content. Assemblies also did not improve significantly with longer reads. For example, the C24 and L*er*-1 assemblies had almost the same N50 and L50, despite C24 being sequenced mostly with 40-bp reads, and L*er*-1 with 80-bp reads.

An obvious next application of reference-guided assemblies would be to improve draft genome sequences, through an iterative process, in which the initial draft genome is used as a reference for an assembly of the same strain. Two limitations are the current insert lengths of standard sequencing libraries, which do not span all classes of repetitive elements, as well as lack of automated software.

Our assemblies have also shed light on the functional consequences of sequence variants. For example, we have shown that a substantial fraction of 1- and 2-bp indels in coding regions are compensated by nearby indels that restore the coding frame (Fig. 2). Current genomewide association studies generally do not consider the nature of a variant, because not all variants are analyzed or even known. With better information, it will be possible to annotate the predicted effect of the combination of sequence variants in an allele, and subsequently base genomewide association studies on classes of alleles with reduced or increased activity, rather than ignoring such information.

In addition to comparing genome sequences, there is great interest in studying individual patterns of DNA methylation, chromatin modifications, and RNA expression. We have already demonstrated how our assemblies improve mRNA and sRNA expression studies, and we expect a similar impact on DNA methylation analyses.

In summary, our reference-guided assembly approach greatly reduces the bias introduced when next generation sequencing reads are only aligned against a reference genome. The availability of several whole-genome assemblies should thus improve the identification of variants in the 1,001 Genomes projects that is underway for *A. thaliana*, by exploiting all known variants as targets for mapping of short reads (19).

Materials and Methods

Combing the Contigs with AMOScmp. The block assemblies computed by VELVET, ABYSS, and Euler-SR, the SUPERLOCAS run and the VELVET assembly of leftover reads introduced high levels of redundancy into the combined set of contigs. This was evident at five levels: (*i*) different runs of different tools assembling identical sets of input reads; (*ii*) reads used twice if their respective block was allocated to overlapping superblocks; (*iii*) unmapped reads contributing to multiple superblock assemblies; (*iv*) reads reused due to repetitive alignments to multiple blocks; and (*v*) assembly of leftover and dangling pairs also included in the block assemblies. To assemble the contigs and purge the redundancy, we used AMOScmp, which applies an alignment-layout-consensus approach using alignments against the reference to guide the overlap calculation of contigs (29). To reduce complexity and hardware requirements we ran AMOScmp on each chromosome arm separately (*SI Materials and Methods*).

Correction of Misassemblies. We aligned all reads against the set of supercontigs using GenomeMapper and performed consensus analysis with SHORE. Differences between aligned reads and reference sequence revealed misassemblies in the supercontigs. Any supercontig shorter than 100 bp, featuring only low read coverage as well as supercontigs with mostly repetitive read alignments were removed. All remaining supercontigs were split at any region where variant predictions indicated misassemblies including uncovered regions, local clusters of differences, and regions with mate pairs that did not align in the expected order and orientation (*SI Materials and Methods*).

Scaffolding. Read pairs with reads aligned to two different supercontigs defined a connection (bridge) between the respective supercontigs. Bridges suggested that two supercontigs were in local vicinity and had a defined order in the focal genome. We also used homology of supercontigs to the reference sequence to infer additional connections, as described in the BAMBUS (30) manual (*SI Materials and Methods*). As mate pair libraries suffer from high rates of potential clonal events we did not consider mate pair reads that aligned to the exact positions as others of the same library.

After running BAMBUS with the set of filtered bridges and connections based on homology as input, the final scaffolding graph was plotted, manually evaluated, and suspicious connections were removed (*SI Materials and Methods*). By default, BAMBUS connects contigs within scaffolds using a fixed number of 60 N's. Instead, we predicted the most likely distance between connected contigs on the basis of the alignment locations of read pairs mapped to two connected contigs, and introduced the corresponding number of N's into the scaffolds.

Base Quality Assessment and Masking. To assign a final per-base quality value, we aligned reads against the respective assembly and used SHORE's resequencing pipeline for consensus analysis (*SI Materials and Methods*). On the basis of SHORE's positionwise quality values q_{ref} (reference) and q_{var} (variation) we assigned a per-base quality q_{ass} to each residue. If there was mere support for the reference allele, q_{ass} was set to q_{ref}. If only a nonreference allele was supported, q_{ass} was set to 0. If there was evidence for two alleles, q_{ass} was assigned the maximum of 0 and the difference of q_{ref} and q_{var}. Every base that was assigned a quality value of less than 10 was masked. Scaffolds with less than 500 unmasked bases were discarded. N's at the beginning of scaffolds were removed. For a more stringent, but less comprehensive assembly, we masked all bases with a quality of less than 15. Additionally we masked all unmasked regions that were shorter than 100 bp.

ALLPATHS-LG de Novo Assembly. ALLPATHS-LG version allpathslg-35762 (23) was applied. Default parameters were used, except that USE_LONG_JUMPS was set to "false," as long jumping libraries with size larger than 20 kb were not included in the analysis.

ACKNOWLEDGMENTS. We thank all developers of the assembly tools used in this study. We thank Illumina for sequencing seven paired-end lanes of Bur-0. This work was supported by European Community Framework Programme 7 (FP7) Marie Curie Fellowship PIEF-GA-2008-221553 and a European Molecular Biology Organization (EMBO) long-term fellowship (to L.M.S.), Bundesministerium für Bildung und Forschung (BMBF) Genomanalyse im biologischen System Pflanze - Genome-wide Natural Diversity Exploitation in the Arabidopsis Population (GABI-GNADE) (D.H.H. and D.W.), European Community Framework Programme 6 (FP6) Integrated Programme (IP) Silencing RNAs: Organizers and Coordinators of Complexity in Eukaryotic Organisms (SIROCCO) (Contract LSHG-CT-2006-037900), Transnational Plant Alliance for Novel Technologies - Towards Implementing the Knowledge-based Bio-economy in Europe (PLANT-KBBE) project Transcriptional Networks and Their Evolution in the Brassicaceae (TRANSNET) (BMBF), a Gottfried Wilhelm Leibniz Award of the Deutsche Forschungsgemeinschaft (DFG), and the Max Planck Society (D.W.).

1. Clark RM, et al. (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* 317:338–342.
2. Zeller G, et al. (2008) Detecting polymorphic regions in *Arabidopsis thaliana* with resequencing microarrays. *Genome Res* 18:918–929.
3. Ossowski S, et al. (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res* 18:2024–2033.
4. Springer NM, et al. (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet* 5: e1000734.
5. Gore MA, et al. (2009) A first-generation haplotype map of maize. *Science* 326: 1115–1117.
6. Quinlan AR, et al. (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res* 20:623–635.
7. Lam HY, et al. (2010) Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol* 28:47–55.
8. Hormozdiari F, et al. (2010) Next-generation VariationHunter: Combinatorial algorithms for transposon insertion discovery. *Bioinformatics* 26:i350–i357.
9. Lee S, Hormozdiari F, Alkan C, Brudno M (2009) MoDIL: Detecting small indels from clone-end sequencing with mixtures of distributions. *Nat Methods* 6:473–474.
10. Korbel JO, et al. (2009) PEMer: A computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol* 10:R23.
11. Chiang DY, et al. (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* 6:99–103.
12. Campbell PJ, et al. (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* 40: 722–729.
13. Lee S, Cheran E, Brudno M (2008) A robust framework for detecting structural variations in a genome. *Bioinformatics* 24:i59–i67.
14. Wang K, et al. (2007) PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 17:1665–1674.
15. Korbel JO, et al. (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318:420–426.
16. Tuzun E, et al. (2005) Fine-scale structural variation of the human genome. *Nat Genet* 37:727–732.
17. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18:1851–1858.
18. Bentley DR, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59.
19. Schneeberger K, et al. (2009) Simultaneous alignment of short reads against multiple genomes. *Genome Biol* 10:R98.
20. Dalloul RA, et al. (2010) Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): Genome assembly and analysis. *PLoS Biol* 8:e1000475.
21. Li R, et al. (2010) The sequence and de novo assembly of the giant panda genome. *Nature* 463:311–317.
22. Young AL, et al. (2010) A new strategy for genome assembly using short sequence reads and reduced representation libraries. *Genome Res* 20:249–256.
23. Gnerre S, et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA* 108:1513–1518.
24. Mirouze M, et al. (2009) Selective epigenetic control of retrotransposition in *Arabidopsis*. *Nature* 461:427–430.
25. Laitinen RA, Schneeberger K, Jelly NS, Ossowski S, Weigel D (2010) Identification of a spontaneous frame shift mutation in a nonreference *Arabidopsis* accession using whole genome sequencing. *Plant Physiol* 153:652–654.
26. Simpson JT, et al. (2009) ABySS: A parallel assembler for short read sequence data. *Genome Res* 19:1117–1123.
27. Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18:821–829.
28. Chaisson MJ, Pevzner PA (2008) Short read fragment assembly of bacterial genomes. *Genome Res* 18:324–330.
29. Pop M, Phillippy A, Delcher AL, Salzberg SL (2004) Comparative genome assembly. *Brief Bioinform* 5:237–248.
30. Pop M, Kosack DS, Salzberg SL (2004) Hierarchical scaffolding with Bambus. *Genome Res* 14:149–159.
31. Nordborg M, et al. (2005) The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol* 3:e196.
32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.
33. Kurtz S, et al. (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5:R12.
34. Hu TT, et al. (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat Genet* 43:476–481.
35. Pelak K, et al. (2010) The characterization of twenty sequenced human genomes. *PLoS Genet* 6:e1001111.
36. Plantegenet S, et al. (2009) Comprehensive analysis of *Arabidopsis* expression level polymorphisms with simple inheritance. *Mol Syst Biol* 5:242.
37. Laubinger S, et al. (2008) At-TAX: A whole genome tiling array resource for developmental expression analysis and transcript identification in *Arabidopsis thaliana*. *Genome Biol* 9:R112.
38. Mosher RA, et al. (2009) Uniparental expression of PolIV-dependent siRNAs in developing endosperm of *Arabidopsis*. *Nature* 460:283–286.
39. Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815.
40. Theologis A (2001) Goodbye to 'one by one' genetics. *Genome Biol* 2:2004.1–2004.9.
41. Rounsley SD, Last RL (2010) Shotguns and SNPs: How fast and cheap sequencing is revolutionizing plant biology. *Plant J* 61:922–927.
42. Huang S, et al. (2009) The genome of the cucumber, Cucumis sativus L. *Nat Genet* 41: 1275–1281.

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.

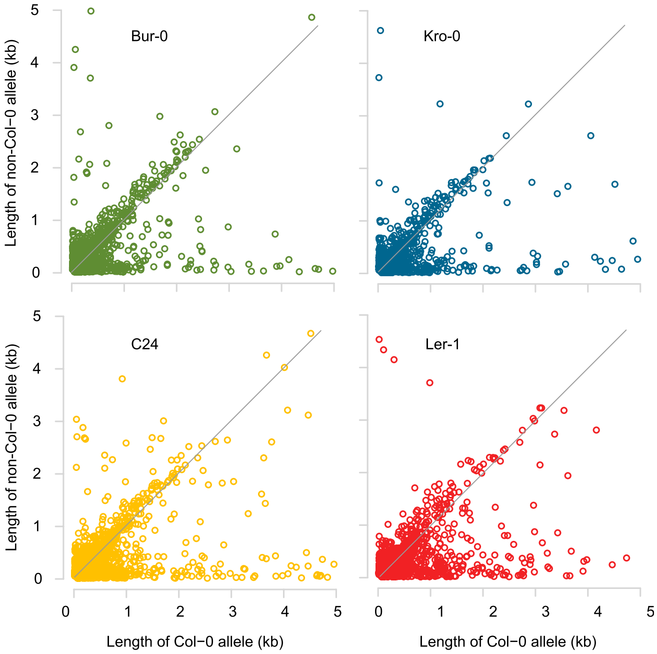

Fig. S1. Allele length comparisons of highly diverged regions (HDRs).

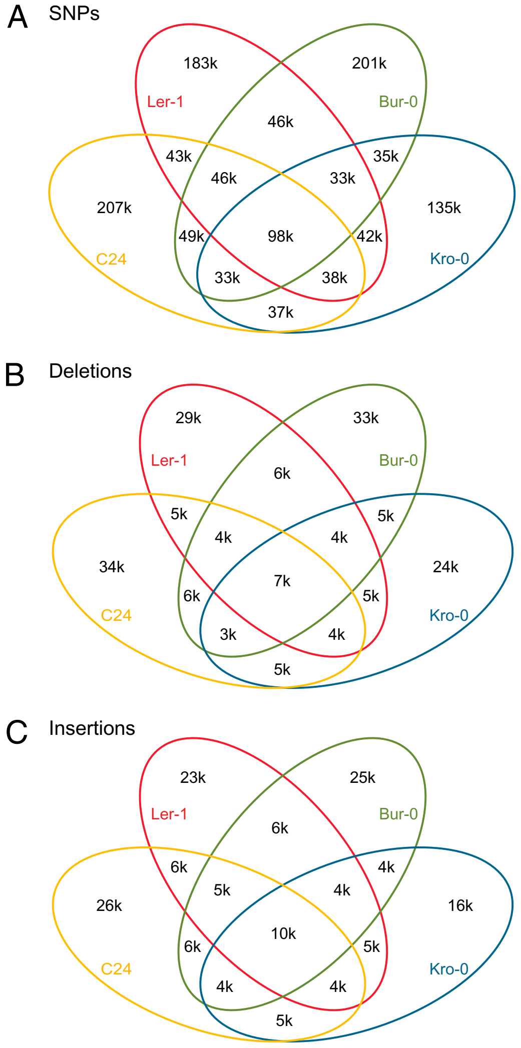

Fig. S2. Shared SNPs (*A*), deletions (*B*), and insertions (*C*) relative to Col-0 reference.

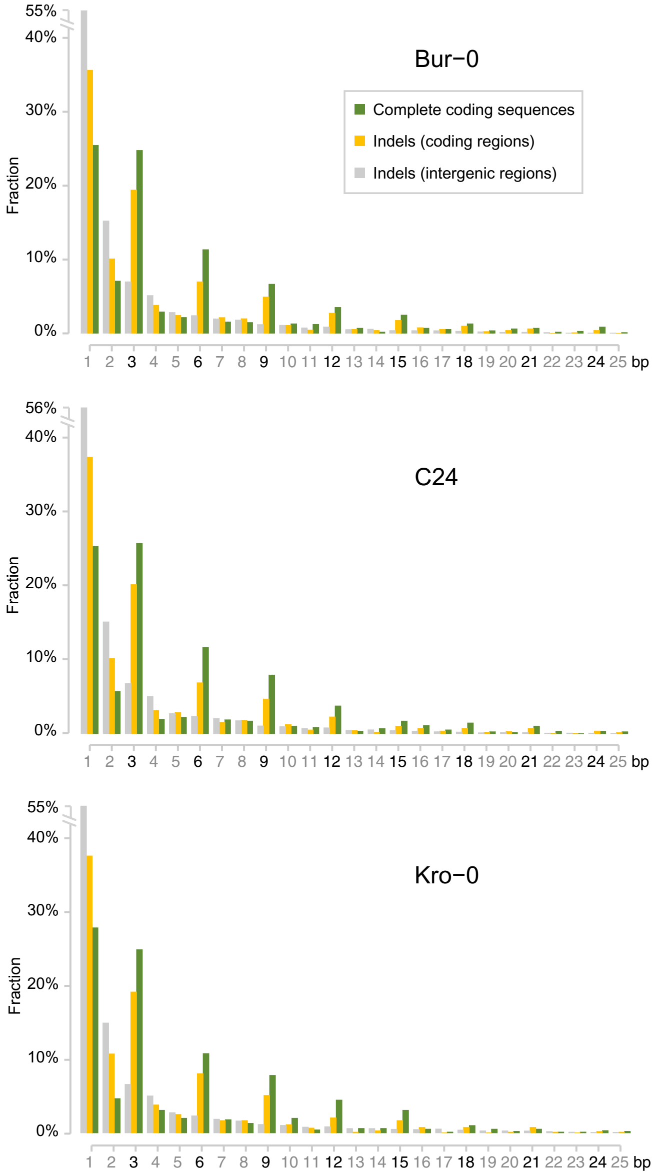

Fig. S3. Length variation in coding sequences, relative to Col-0 reference.

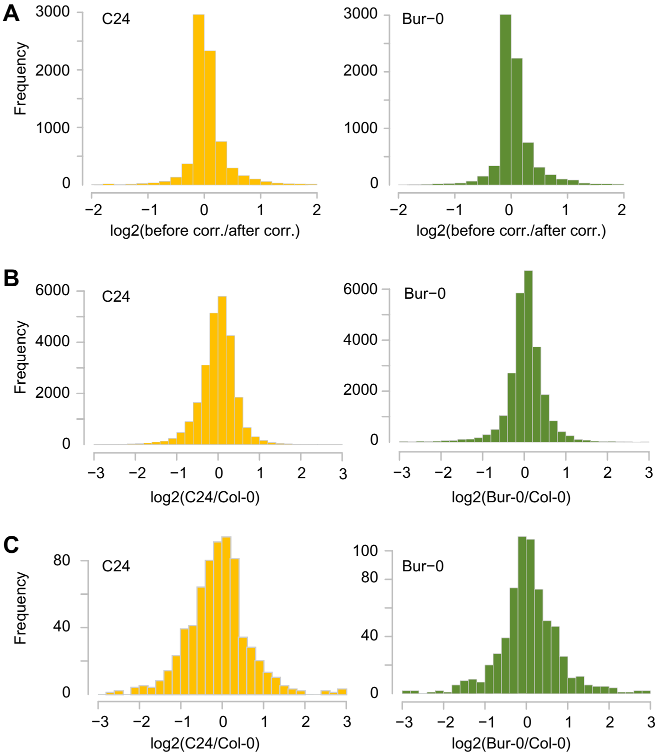

Fig. S4. Tiling array expression analysis. (*A*) Effect of probe correction on expression estimates for 7,056 genes for which half or more of all probes were removed. Note that the distribution is skewed toward the estimates being higher after correction. (*B*) Expression of conserved genes (at least 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).

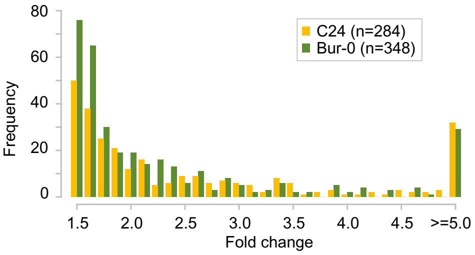

Fig. S5. sRNA expression analysis. Increase in expression estimates for distinct sRNA loci resulting from incorporating information from genome assemblies.

Table S1. Read statistics

	Bur-0	C24	Kro-0	L*er*-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			L*er*-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
L*er*-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
L*er*-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
L*er*-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
L*er*-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	—
L*er*-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	—
L*er*-1	2	10,153,298	10,153,423	125	1,010	53,559	53,434	125	—
L*er*-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	—
L*er*-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	—
L*er*-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.