Chromosome-level assembly of *Arabidopsis thaliana* Ler reveals the extent of translocation and inversion polymorphisms

Shorter title (for mobile devices): *De novo* assembly of *A. thaliana* Ler

Luis Zapata\(^{1,2,†}\), Jia Ding\(^{3,†}\), Eva-Maria Willing\(^4\), Benjamin Hartwig\(^4\), Daniela Bezdan\(^{1,2}\), Wen-Biao Jiao\(^4\), Vipul Patel\(^4\), Geo Velikkakam James\(^4\), Maarten Koornneef\(^3,5\), Stephan Ossowski\(^{1,2}\), Korbinian Schneeberger\(^4\)

Author affiliations:

\(^1\) Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, 08003 Barcelona, Spain. \(^2\) Universitat Pompeu Fabra (UPF), Plaça de la Mercè 10-12, 08002 Barcelona, Spain. \(^3\) Department of Plant Breeding and Genetics, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Cologne, Germany. \(^4\) Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Cologne, Germany. \(^5\) Laboratory of Genetics, Wageningen University, NL-6708 PE, Wageningen, Netherlands.

\(^†\) These authors contributed equally.

Present addresses:

Agilent Technologies, Cartagenia, Technologialaan 3, 3001 Leuven, Belgium (JD);

KWS SAAT SE, RD-DM, Grimsehlstr. 31, 37574 Einbeck, Germany (VP);

Rijk Zwaan R&D Fijnaart, Rijk Zwaan, Fijnaart, Netherlands (GVJ);
Abstract

Resequencing or reference-based assemblies reveal large parts of the small-scale sequence variation. However, they typically miss to separate such local variation into co-linear and re-arranged variation, as they usually do not recover the complement of large-scale rearrangements including transpositions and inversions. Besides the availability of hundreds of genomes of diverse Arabidopsis thaliana accessions, there is so far only one full-length assembled genome, the reference sequence. We have assembled 117 Mb of the A. thaliana Ler genome into five chromosome-equivalent sequences using a combination of short Illumina reads, long PacBio reads and linkage information. Whole-genome comparison against the reference sequence revealed 564 transpositions and 47 inversions comprising around 3.6 Mb, in addition to 4.1 Mb of non-reference sequence mostly originating from duplications. Though rearranged regions are not different in local divergence from co-linear regions, they are drastically depleted for meiotic recombination in heterozygotes. Using a 1.2 Mb inversion as example, we show that such rearrangement-mediated reduction of meiotic recombination can lead to genetically isolated haplotypes in the world-wide population of A. thaliana. Moreover, we found 105 single copy genes, which were only present in the reference sequence or the Ler assembly and 334 single-copy orthologs, which showed an additional copy in only one of the genomes. This work gives first insights into the degree and type of variation, which will be revealed once full-length assemblies will replace resequencing or other reference dependent methods.
**Significance Statement**

Despite the widespread reports on deciphering the sequences of all kinds of genomes, most of these reconstructed genomes rely on a comparison of short DNA sequencing reads to a reference sequence, rather than being independently reconstructed. This limits the insights on genomic differences to local, mostly small-scale variation, as large re-arrangements are likely overlooked by current methods.

We have *de novo* assembled the genome of a common strain of *Arabidopsis thaliana* Landsberg erecta and revealed hundreds of re-arranged regions. Some of these differences suppress meiotic recombination, impacting on the haplotypes of a worldwide population of *A. thaliana*. In addition to sequence changes, this first comparison of an independent, chromosome-level assembled *A. thaliana* genome revealed hundreds of unknown, accession-specific genes.
Introduction

Landsberg *erecta* is presumably the second most used strain of *Arabidopsis thaliana* after the reference accession Columbia (Col-0). It is broadly known as Ler-0, which is an abbreviation for its accession code La-1 and a mutation in the *ERECTA* gene. In 1957, George Rédei, at the University of Missouri-Columbia, irradiated La-1 samples, which were provided by Friedrich Laibach and were collected in Landsberg an der Warthe now called Gorzów Wielkopolski, Poland, where Ler-0 related genotypes are still present (1). Some seeds of the original batch were irradiated with X-rays, resulting among others in the isolation of the *erecta* (*er*) mutant (2)(3). Will Feenstra received this mutant from George Rédei in 1959 as he was interested in its erect growth habit and introduced it as the standard strain in the Department of Genetics at Wageningen University. There he started a mutant induction program that was later continued by Jaap van der Veen and Maarten Koornneef. Mutants from this program as well as parental lines of recombinant inbred lines were mainly distributed as Ler-0 lines to other laboratories reflecting the increasing interest in *A. thaliana*. Some descendants of Ler-0 were later renamed to Ler-1 and Ler-2 to identify genotypes used in different laboratories, but most likely all derived from the original mutant isolated by George Rédei and we will collectively refer to them as ‘Ler’.

First comparative analyses of the Ler genome included cytogenetic studies using pachytene cells and in situ hybridization (4)(5), suggesting a large inversion on the short arm of chromosome 4 as well as differences between 5S rDNA clusters as compared to the genome of Col-0 (4). The first large-scale analysis of the Ler genome sequence was published together with the Col-0 reference sequence in 2000 (6). Within 92 Mb of random shotgun dideoxy sequencing reads 25,274 SNPs and 14,570 indels were identified. Though this was a severe underestimation
(7)(8)(9)(10)(11), the authors already observed that many of the large indels contained entire active genes, half of which were found at different loci in the genome of Ler, whereas others were entirely absent (6).

The advent of next generation sequencing greatly expanded the knowledge on natural genetic diversity in *A. thaliana* (reviewed in (12) and (13)). However, genome-wide studies on gene absence/presence polymorphisms were not repeated, as short read analyses focused on small-scale changes only. To resolve large variation, reference-guided assemblies (8)(9)(14) and structural variation identifying tools (15) were introduced. However, such methods mostly reveal local differences, which do not include the complement of large-scale rearrangements including inversions or transpositions.

So far two *de novo* assemblies of Ler have been published. The first was based on Illumina short read data and resulted in an assembly with an N50 of 198 kb showing similar performance as a reference-guided assembly (8). The second was based on a set of previously released Pacific Bioscience’s single-molecule, real-time sequencing (PacBio) data (16) and assembled the genome into 38 contigs with an N50 of 11.2 Mb (17). This drastic improvement outlines the potential of long read sequencing technologies such as PacBio sequencing (18) to overcome the limitations of short read methods as long reads can span many of the repetitive regions, which are presumably the most common reason for assembly breaks in short read assemblies. Alternatively, low-fold long read sequencing could be combined with cheaper short read sequencing, either by using the short reads for error correcting the long reads (19) or for integrating long read information into short read assemblies or vice versa (20).

Despite the unprecedented contiguity of this long read assembly, both earlier studies focused on methodological aspects and did not perform any whole-genome comparisons of the Ler genome or gene annotations and as a consequence comprehensive reports on large-scale re-arrangements and non-reference genes are
still sparse. We have generated an advanced de novo assembly of Ler consisting of 117 Mb arranged into five sequences representing the five chromosomes, which is based on a combination of short read assembly, long read-based gap closure and scaffolding based on genetic maps. This chromosome-scale assembly and its comparison to the reference assembly revealed features, which are typically not analyzed within next generation sequencing assemblies, including the location of a polymorphic rDNA cluster and centromeric repeats as well as the exact make-up of all large re-arrangements including a 1.2 Mb inversion on the short arm of chromosome 4. This inversion suppresses meiotic recombination in Ler and Col-0 hybrids, and we show that this suppression introduced genetically isolated inversion haplotypes into the worldwide population of A. thaliana. De novo gene annotation revealed hundreds of copy number polymorphisms as well as novel genes that are entirely absent in one or the other genome. Finally, we report on variation in different Ler genomes suggesting that some Ler lines feature unexpected footprints of an additional mutagenesis event.

Results

Karyotype-resolving assembly of A. thaliana Ler

We deeply sequenced the genome of Ler using Illumina libraries of different insert sizes (SI Appendix, Table S1). Combined with recently generated data (8), this yielded an initial assembly with an N50 of 7.5 Mb using ALLPATHS-LG (21) and SSPACE (22). For further improvement we generated long read data from ten PacBio SMRTcells. However, during the course of this project Pacific Biosciences released sequencing reads of Ler with higher quality (16), which we used for gap-closure (20) and scaffolding (23) to generate an assembly with an N50 of 12.8 Mb consisting of 65 scaffolds longer than 50 kb. Following the POPSEQ approach (24) we anchored 31 of these scaffolds to two public genetic maps (25)(26), which
allowed us to generate five chromosome-representing sequences using stretches of
Ns as indication of assembly breaks (Table 1, SI Appendix, Fig. S1 and
Supplementary Methods). As some of the short scaffolds were too small and did not
overlap with enough markers, we introduced additional seven scaffolds with a
combined length of 1.4 Mb into the reconstructed chromosomes based on homology
information from Col-0. The final assembly consisted of five pseudo-molecules and
25 unplaced scaffolds (including scaffolds representing the organelle genomes) with
a combined length of 118.3 Mb including less than 1 Mb of ambiguous bases.

Within the assembly, we found sequence similarity to telomeric repeats at five
out of eight chromosome ends without heterochromatic nuclear organizing region
(NOR) (i.e. shorter arms of chromosome 2 and 4), and centromeric repeat
sequences and rDNA clusters within most of the peri-centromeric regions (Fig. 1).
Telomeric repeats were also found as interstitial repeats near or in the peri-
centromeric regions, which is similar to their location in the Col-0 genome (27).
Interestingly, we found a few rDNA copies in the upper arm of chromosome 3
revealing the location of a (not fully assembled) rDNA cluster, which is in agreement
with earlier findings of a Ler-specific rDNA locus in this region (4). To test the
assembly quality further we analyzed NB-LRR gene loci, which include regions that
are known to be structurally diverse between accessions and are challenging to
assemble due to high levels of local repeats (28). For 154 of the 159 genes we could
identify the orthologous regions in the Ler assembly. Only ten (6%) of these regions
included ambiguous sequence, which would indicate failures in sequence assembly.
In contrast, 49 (31%) of them revealed differences larger than 100 bp corroborating
their strong divergence.

Finally, we used Ler wild-type RNA-seq reads from two public data sets
(9)(29) and homology to Col-0 protein sequences to annotate 27,170 protein-coding
genes in the assembly of Ler (as compared to 27,416 protein-coding genes in the
reference annotation) (SI Appendix, Material and Methods).
The complement, divergence and impact of non-allelic re-arrangements

In contrast to resequencing, full-length assemblies facilitate direct identification of large-scale genomic rearrangements, including transpositions and inversions. These types of higher order differences constitute a second layer of genetic variation, as local differences (e.g. SNPs) can be found in co-linear as well as in rearranged regions (Fig. 2A). However, as resequencing studies typically cannot distinguish between re-arranged and non-rearranged regions this distinction was so far not possible (30).

A whole-genome alignment of the Ler and Col-0 genome assemblies (31) revealed 512 co-linear (allelic) and 611 re-arranged (non-allelic) regions comprising around 107.6 and 3.6 Mb (Fig. 2B, SI Appendix, Material and Methods). Among the non-allelic regions, we identified 47 inversions and 383 and 181 inter- and intra-chromosomal transpositions (Dataset S1). Most of the transpositions resided in pericentromeric regions, whereas inversions were also found in chromosome arms (Fig. 2C). Nearly 40% of the transposed sequences overlapped with transposable elements (TEs), however, only a minor fraction of the inversion sequences were related to TEs. The inversion breakpoints overlapped with seven and ten genes in the Ler and Col-0 assemblies, which did not have syntenic orthologs in the other assembly, suggesting that these genes have been deleted by the inversion events.

The by far largest sequence difference between the two assemblies was an inversion of 1.2 Mb on chromosome 4, which we confirmed by PCR (SI Appendix, Fig. S2 and S3). Moreover, 7.9 Mb of the Col-0 and 4.1 Mb of the Ler assembly were not aligned to any homologous region of the other genome at all. This sequence space was separated into 713 and 535 regions, respectively (Fig. 2B). Even though not aligned in a strict one-to-one whole-genome alignment, large portions of these unaligned regions showed extensive similarity to regions in the other genome...
suggesting that most of these regions originated from duplication events (Fig. 2A and B) (SI Appendix, Materials and Methods).

To compare sequence divergence in allelic and non-allelic regions, we searched for local differences including SNPs, small and large indels as well as highly divergent regions (HDRs) (Fig. 2A) (SI Appendix, Materials and Methods, Dataset S2). Around 4.5 Mb (4.2%) of the 107.6 Mb of allelic sequence were polymorphic with the majority of this variation organized in long indels and HDRs (Fig. 2D). Around 39% of the long indels had flanking (short or tandem) repeat sequences indicating that a large proportion of the indel mutations were introduced by homology-dependent events, whereas only 16% of them highly overlapped with TEs. Even though re-arranged (non-allelic) sequence harbored less large local differences, presumably due to their short length (average lengths: transpositions, 4 kb; inversions, 9 kb (without the 1.2 Mb inversion); allelic regions, 209 kb), the average pairwise difference in the alignments of non-allelic regions was still similar compared to the differences in allelic regions (Fig. 2D and E).

To avoid recombination between non-allelic regions, meiotic recombination requires pairing of homologous sequence between allelic regions (11) implying that re-arranged regions should be suppressed for meiotic recombination in heterozygotes (Fig. 2F). To test this, we overlapped the precise location of 362 crossover (CO) events from Col-0/Ler hybrids previously collected from different studies (32) with allelic and non-allelic regions (SI Appendix, Materials and Methods).

Only five of the CO events did not reside in co-linear, allelic regions (expected: 35 CO events), of which four COs were located in non-aligned regions, and only one CO was mapped to a transposition (Fig. 2G). This highly significant underrepresentation of COs in non-allelic regions evidences that re-arranged regions are in fact suppressed for meiotic recombination (p-value = 1.96e-06, Chi-squared test).
**The effects of inversions on natural haplotypes**

Inverted regions were most drastically suppressed for meiotic recombination and in consequence genetic exchange between the two alleles of an inversion is expected to be minimized (33). Despite this strong impact, no population-wide analysis of inversions in Arabidopsis and only few inversions have been reported so far (5)(8)(34).

Two inversions between Col-0 and Ler were larger than 100 kb including a 170 kb inversion on chromosome 3 and the 1.2 Mb inversion on chromosome 4. The latter inversion was already described by Fransz *et al* (5) and was found to be associated with a polymorphic, heterochromatic knob on the short arm of the Col-0 chromosome, presumably due to inverting parts derived from the peri-centromere (5)(35). This inversion could also be observed between Col-0 and the closely related plant *Arabidopsis lyrata* further corroborating that the Col-0 allele is the derived form (36). Overlapping the regions of the inversion with meiotic recombination frequency data (26) showed locally reduced recombination rates co-occurring with both of these two large inversions (Fig. 3A, *SI Appendix*, Fig. S4). The effect of the chromosome-4-inversion extended its recombination suppression into the heterochromatic peri-centromere and was in agreement with early reports of reduced recombination specific to hybrids of these accessions (26)(37).

Selection acting on one of the alleles of an inverted region can have dramatic effect on its allele frequency and haplotype divergence across the entire inverted regions. To estimate the impact of these large inversions on the population of *A. thaliana*, we genotyped a worldwide selection of 409 accessions using public whole-genome sequencing data (15)(30)(38) (*SI Appendix*, Materials and Methods). For this, we simultaneously aligned the short reads against the Col-0 and Ler reference sequences and calculated the ratio of alignments to either of the inversion breakpoints to assign the respective allele (*SI Appendix*, Fig. S5). Surprisingly, only
sequencing reads of Ler matched the Ler inversion breakpoints of the 170 kb inversion on chromosome 3, while the reads of all other accessions matched the Col-0 breakpoints suggesting that this inversion was either specific to the La-1 accession or was introduced during the mutagenesis leading to the Ler genotype. In contrast, genotyping for the 1.2 Mb inversion revealed 26 accessions as carriers of the Col-0 allele and 383 accessions as carriers of the Ler allele (Fig. 3B). Most accessions could be characterized at the distal inversion breakpoint, however some accessions showed an additional re-arrangement on the proximal breakpoint complicating short read alignments. Despite this, none of the accessions showed contradicting genotypes at the two breakpoints.

The relatedness of the accessions was estimated using 20,408 SNPs from within the inversion and revealed a perfectly separated sub-clade, which matched the assignment Col-0-like inversion allele during the breakpoint genotyping (SI Appendix, Materials and Methods) (Fig. 3C). This suggested that suppressed recombination and genetic exchange separated the population in two distinct groups, in particular as this separation was not mirrored by geographic isolation (Fig. 3D). However, this separation was not absolute, across the 20,408 sites we found 13 (0.06%) sites with shared variation between both groups (while 20,306 sites were polymorphic only within the Ler-like inversion accessions and 89 sites were polymorphic only within the Col-0-like accessions). Though this very low level of shared variation still could indicate the existence of rare (double) recombination events between two inversion alleles, such patterns might also be explained by gene conversions (11), or accumulation of false positive SNP prediction within the set of shared SNPs.

The minor Col-0 inversion allele mostly co-occurred together with the Ler-like alleles across its entire distribution range in Central and Northern Europe and even in recently invaded North America (39). While the Ler-like genotypes showed genetic diversity within the inversion region, which was not different from the rest of the genome, the diversity among the Col-0-like genotypes was greatly reduced within the
inversion (Fig. 3E, see SI Appendix, Fig. S6) and extended across the low recombining peri-centromere. The drastic reduction in diversity most likely reflects a bottleneck event as a result of the inversion generating the derived Col-0 allele, which was maintained by the suppression of recombination in inversion heterozygotes.

Together this suggests that the inversion was introduced by a single event, and in consequence, population differentiation between the two groups was not uniform across the chromosome. Wright's fixation index ($F_{ST}$) was elevated in the inversion and again the effect was extended into the pericentromeric region (Fig. 3F, see SI Appendix, Fig. S7).

Finally, we assessed patterns of genic selection in the inversion by calculating $K_a/K_s$ to explain the relatively high allele frequency of the inversion allele. One of the genes in the inversions, RLP46, a LRR-receptor domain-containing gene with function in defense response was among the four genes with the highest $K_a/K_s$ values across the entire genome. However, the numbers of non-synonymous and synonymous changes were not high enough to prove a significant difference from neutral evolution, preventing a final conclusion whether selection of a gene in the inversion helped to retain the inverted allele or if the inversion simply drifted into the population.

**Genic copy number variation between two A. thaliana lines**

To assess the amount of genes that are specific to Ler or Col-0, we first calculated groups of orthologous genes between the two genomes (SI Appendix, Materials and Methods). This revealed initial sets of 212 and 240 single-copy genes that appeared specific to either of the genomes. These gene sets were then further filtered to exclude the possibility that their accession-specific occurrence is due to failures in the assemblies, annotations, or assignment of orthologs (SI Appendix, Materials and Methods). This conservative approach led to 40 unique genes specific
for Ler and 63 genes specific for Col-0. Using the genome of the close relative A. lyrata (36)(40) as out-group, we found that the majority of these polymorphic genes (77%) evolved via deletions within the genome where they are absent, rather than by a spontaneous appearance in the genome they are present in (Fig. 4A). In fact, the number of genes is presumably underestimated as the A. lyrata assembly might not be complete and might lack some genes leading to a false categorization of genes.

Single-copy gene loss is an extreme type of modification of the gene content of a genome, as no additional copy can replace the function of the absent gene. In contrast, small gene families with multiple copies are expected to be at least partially functionally redundant and therefore more variable for copy number. In fact, 330 unique one-to-one orthologous gene pairs showed one additional copy in either the Ler or Col-0 genome (again after strict filtering for artifacts) (SI Appendix, Materials and Methods). In 151 of the cases the additional gene was found in the genome of Col-0, whereas in 179 of the cases the additional copy was found in Ler. In contrast to accession-specific genes, copy number variation did not primarily evolve through deletion events; instead copy loss events were underrepresented (36%) as compared to acquiring new copies by novel duplication (64%) (again, gain and loss events were distinguished using A. lyrata as an out-group) (Fig. 4B). These additional genes resided in 134 duplicated regions in Col and in 148 duplicated regions in Ler, implying that some of the duplicated genes were introduced by the same duplication event (Fig. 4C). Interestingly, while the loss of additional copies was independent of the location, the gain of additional copies was clearly enriched for copies residing close to the original gene (Fig. 4D) suggesting that novel genes primarily evolved by local duplication events. This was further supported by differences in the similarity between new copies and the original genes. Gained copies were much more similar to each other as compared to copies, where the other genome lost one copy, corroborating that gene gains predominantly result from recent events (Fig. 4E).
Performing a GO enrichment analysis, we found that polymorphic single-copy genes are enriched for signaling and signal transduction related pathways, whereas genes with copy number changes were enriched for defense-related categories, protein polymerization and translation.

**Polymorphisms between diverse Ler lines**

Earlier comparisons of different Ler lines have revealed a polymorphic premature stop codon in the HUA2 gene (41). This allele, hua2-5, was specific only to some but not all Ler lines and could not be identified in any of 29 other A. thaliana accessions (41) and still has not been identified in any of the hundreds of other accessions released by the 1001 Genomes Project (www.1001genomes.org). This suggested that the hua2-5 mutation occurred after the original La-1 mutagenesis, and because it was not found in some flowering time mutants isolated in the 1960’s but present in mutants isolated in the early 1970’s at the Genetics department in Wageningen this mutation most likely occurred in that laboratory (41).

We have re-analyzed six different data sets, all of which have been labeled as whole-genome sequencing data of Ler (8)(9)(10)(11) (SI Appendix, Table S2). We aligned the reads of all sets against the Ler reference sequence and after stringent filtering we identified only a marginal amount of structural variations between these lines, including five deletions and two duplication events using read copy and read pair analysis. However, we also identified an unexpected amount of 723 single-nucleotide variations including the hua2-5 mutation, which was present in two of the six lines. Surprisingly most of the other polymorphic sites were not specific to a single genome as expected for inbred strains distributed to different labs, but were associated to the hua2-5 mutation. Nearly all of these mutations were C=>T or G=>A mutations and were found in large clusters throughout the entire genome (Fig. 5). As such mutations and their clustering are characteristic for EMS mutations, it is likely
that the *hua2-5* mutation (and all other associated mutations) resulted from an EMS mutagenesis.

**Discussion**

We have used a combination of short Illumina reads, long PacBio reads and genetic maps to generate a chromosome-level assembly of *A. thaliana* Ler. Despite the unprecedented contiguity that we achieved by combining different data types, integration was not straightforward. In contrast, *de novo* assemblies based only on long reads are easier to perform and start to become a powerful alternative (17)(42), although they do not assemble entire chromosomes yet. Independent of the assembly type, the amount of re-arranged sequence that was revealed by comparing two assemblies underlines the general importance of chromosome-level assemblies, in particular as rearrangements have been recognized as confounding factors in genome-wide screens (11)(43) and are essential to fully understand the segregation and evolution of natural haplotypes.

Recent meiotic recombination estimates suggested that high levels of sequence divergence themselves are not inhibitory for meiotic recombination (26)(44)(45), which is in agreement with a positive correlation of ancestral recombination frequency and regions with high sequence divergence (46). However, this refers to local sequence divergence without considering re-arrangements like transpositions or inversions. In fact, by analyzing crossover events, which were observed in Col-0/Ler hybrids, in respect to their occurrence in co-linear and re-arranged regions we could show that non-allelic regions have significantly reduced levels of meiotic recombination, despite no obvious difference in local sequence divergence.

Such cross-specific suppression of recombination was described for individual regions before including the 1.2 Mb inversion on the upper arm chromosome 4 (5)(26)(37), which was found between Col-0 and Ler as well as Ws
and Ler. Similar suppression patterns were found in a 2.2 Mb region on upper arm of chromosome 3 within the offspring of Bay-0 and Sha (47) (and were later also observed in crosses between Col-0 and Sha (45)(48)(49)), a 2-3 Mb region on the upper arm of chromosome 5 specific to RRS7, and 0.2 Mb region on the lower arm of chromosome 1 specific to Bay-0 (49). Recently, local suppression of recombination on the lower arm chromosome 4 between a cross of Col-0 and Ws-2 led to the identification of an 1.8 Mb inversion, which perfectly matched the region without any crossover events (34).

Using an 1.2 Mb inversion on the upper arm of chromosome 4 as example we studied the effects of such inversion-mediated suppression of meiotic recombination on the global population of *A. thaliana*. Suppression of genetic exchange between these two inversion alleles introduced a new genetically isolated haplotype into the global population. Although accessions carrying the Col-0-like allele showed a clear concentration in Germany and Sweden, it was surprising that these genotypes were rather widely distributed. Other examples of widely distributed polymorphisms include deletions in FRIGIDA (50), which could be associated with different selective pressures. We could not identify similar signals for any of the alleles in the derived form of the inversion, which nevertheless does not exclude a possible selective advantage of this inversion allele, but might point to a more complex scenario of selection.

In addition to re-arrangements, an even larger portion of the genome was not assigned to any orthologous region in the whole-genome alignment. Many of these regions resulted from duplication events, which had a drastic impact on the gene content of the genomes. Gene absence/presence polymorphisms have been connected to phenotypic variation (1) and genetic incompatibilities (51)(52)(53)(54) before, but the genome-wide extend of hundreds of polymorphic genes in both genomes was surprisingly high even though we only looked at low-copy differences. The functional classes, which were enriched among the duplicated genes, included
defense-related genes, which can have a selective advantage when duplicated as they can evolve race specific resistances exemplified by the RPP1 locus where Ler has 70 kb additional sequence including several strain specific genes (1)(55).

Even in an organism with so many resequenced genomes as A. thaliana, a single chromosome-level genome assembly enabled us to analyze a second layer of genetic variation, which so far could not be considered in most short read based genome-wide studies. These regions showed a drastic impact on natural haplotypes and introduced great variability into the gene space, giving a first glimpse of the degree of natural variation, which can be revealed once more chromosome-level de novo genome assemblies will become available.


**Materials and Methods**

Material and Methods are described in full length in the supplementary information (SI Appendix).

**Data availability**

Raw sequencing data was deposited at NCBI SRA; 180bp fragment library, SRX1567556; 8kb jumping library, SRX1567557; 20kb jumping library, SRX1567558, raw reads with unknown insert size, SRX1567559. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession LUHQ00000000. The version described in this paper is version LUHQ01000000.

**Acknowledgements**

We thank Erik Wijnker for help with interpretation of meiotic recombination suppression in non-allelic regions, Petra Tänzler for help with DNA extractions, Avraham Levy and Shay Shilo for sharing their recombination data and the 1001 Genomes Project for releasing their data allowing us to search for natural variants of the *hua2-5* allele. We acknowledge support of the Spanish Ministry of Economy and Competitiveness, ‘Centro de Excelencia Severo Ochoa 2013-2017’, SEV-2012-0208. Luis Zapata has been supported by the International PhD scholarship program of La Caixa at CRG.

**Authors’ contribution**

Conceived and designed the project: KS, SO, MK. Performed the experiments: BH, DB. Genome assembly: LZ, JD, VP, SO. Genome annotation: LZ, GVJ. Genome and gene comparisons: LZ, JD. Population genetics: JD, EMW, WBJ. Wrote the paper: KS, LZ with the help of all other authors.
References


11. Wijnker E, et al. (2013) The genomic landscape of meiotic crossovers and


modulating salicylic acid pathway activation. *Proc Natl Acad Sci U S A*


**Figure 1.** Chromosome-level sequence assembly reveals the karyotype and the arrangement of structural features of the *A. thaliana Ler* genome. Ideogram of the Ler genome showing idealized chromosomes at pachytene and chromosomal marks, which were revealed with cytogenetics including heterochromatic clusters (dark grey), clusters of centromeric repeats (red) and rDNA (blue) (4)(5) (figure is a redrawing of (56)). On the right is an illustration of the assembly including five chromosomal sequences. The colored bars next to the chromosomes indicate the location of sequence similarity to telomeric repeat sequence (green), centromeric repeat (red), rDNA (blue) as well as major gaps in the sequence (dark grey). The blue star marks a Ler-specific rDNA cluster, which was earlier identified by cytogenetics (4), and was also found in the assembly.
Figure 2. Higher order sequence variation. (A) Schematic of local (top) and higher order (bottom) sequence variation as revealed by a whole-genome alignment. Local sequence divergence does not only include small-scale variation like SNPs and small indels, but also structural variation like large indels and highly divergent regions (HDRs). Higher order variation includes transpositions and inversions, which do not reside in the orthologous regions in the other genome. Both, co-linear (allelic) and rearranged (non-allelic) regions can harbor local variation. (B) Amount of aligned and non-aligned regions in a non-redundant whole-genome alignment of Col-0 and Ler. Aligned regions can be separated into co-linear (grey) and rearranged regions (inversions and transpositions, red). Non-aligned regions, typically residing in the breaks between allelic and non-allelic regions, are shown for Col-0 and Ler.
separately, including the amount of putatively duplicated regions. (C) Location of
transpositions and inversions. (D) Genomic space involved in different types of local
sequence variation, separately shown for allelic and non-allelic regions. (E)
Sequence divergence in allelic and non-allelic alignments. (F) Schematic examples
for the consequences of meiotic recombination (crossing over (CO)) events in
transposed (top) and inverted (bottom) regions. Chromosome arm exchange in non-
allelic regions can lead to extreme chromosomal re-arrangements. (G) Distribution of
the location of 362 CO events in respect to their occurrence in allelic (grey), non-
aligned (green) and non-allelic (red) regions in contrast to the genomic fractions of
these regions; complete genome on top, considering only chromosome arms below.
Figure 3. Impact of large-scale inversions on meiotic recombination and haplotype diversity in a worldwide collection of *A. thaliana* accessions. (A) Male meiotic recombination frequencies across chromosome 3 and 4 contrasted with the location of the two large-scale inversions (dark grey boxes) and the pericentromeric regions (light grey boxes) (recombination data generated by Giraut et al (26)). Recombination frequency was measured between markers with an average distance of 316 kb. Both inversions co-occur with locally reduced recombination frequencies. The interval harboring the inversion on chromosome 3, however, showed residual recombination activity, which does not imply recombination in the inverted region but might arise from recombination in 111 kb of non-inverted sequence in this interval. (B) The names of 409 accessions colored by the inferred chromosome-4-inversion allele (blue: Ler allele, red: Col-0 allele) as assessed on the left and right breakpoints of the inversion. The accessions were ordered after their
occurrence in the haplotype clustering shown in (C). (C) Haplotype clustering based on 9,198 SNPs located within the chromosome-4-inversion revealing two distinct clusters, which perfectly matched the two chromosome-4-inversion alleles. (D) Distribution of the accession origins in central Europe, colored by their respective chromosome-4-inversion alleles. (E) Haplotype diversity within the accessions carrying a Col-0-like (red) or Ler-like allele (blue) of the chromosome-4-inversion. (F) Population differentiation ($F_{st}$) between these two groups of accessions. Inversion and peri-centromere shown with dark and light grey boxes.

Figure 4. Gene absence/presence polymorphisms between Col-0 and Ler. (A) Amount of single-copy, polymorphic genes in Col-0 and Ler. The genes were separated by the presence (left) or absence (right) of an ortholog in the related species A. lyrata. (B) Amount of single-copy genes with one additional copy in Col-0 and Ler. Cases were separated by the presence of one or two orthologs in the genome of A. lyrata like in (A). (C) Dot plot (57) of an example of a local duplication event coping multiple genes in the genome of Ler. Identically colored arrows indicate similarity between underlying gene loci. (D) Amount of local or dispersed gene copies separately shown for copy loss or copy gain events as defined in (B). (E) Sequence identity between gene copies separately shown for copy loss or copy gain events as defined in (B).
Figure 5. Single nucleotide polymorphisms between six Ler genomes from different laboratories. Location and type of SNPs distinguishing six genomes published as the genome of Ler. Genome-wide visualization revealed large blocks of C=>T and G=>A mutations specific to two Ler lines.
SI Appendix, Material and Methods

Chromosome-level assembly of *Arabidopsis thaliana* Ler reveals the extent of translocation and inversion polymorphisms

Luis Zapata†, Jia Ding†, Eva-Maria Willing, Benjamin Hartwig, Daniela Bezdan, Vipul Patel, Geo Velikkakam James, Maarten Koornneef, Stephan Ossowski, Korbinian Schneeberger

Next generation DNA sequencing

Preparation of the 180 bp Illumina paired-end library and whole-genome DNA sequencing on an Illumina HiSeq 2000 machine has been performed at the Center for Genomic Regulation (Barcelona, Spain). Long-range jumping libraries, with 8 and 20 kb average insert sizes, were generated by Eurofins Genomics (Ebersberg, Germany) and sequenced at the Center for Genomic Regulation. Additional sequencing data was acquired from a previous study (1), including another 180bp paired-end and one 3kb mate-pair library sequenced on an Illumina GAIIx at Max-Planck Institute Tübingen, Germany. An overview of fragment and read length as well as read numbers for all sequencing libraries used in this study can be found in SI Appendix, Table S1.

Whole genome assembly of Ler

Data generation and quality filtering

The two libraries of 180bp fragment size were merged for further processing.

To improve raw sequence quality we computed the overlap of the paired reads using
FLASH (2) and kept only overlapping read pairs with less than 2 mismatches per 10 bp in the overlapping region. Jumping libraries were filtered for chimeras and fragments containing the spacer sequence using a custom algorithm adapted from (3). Finally, we filtered the reads of each library with SGA (functions: preprocess, correct) (4) and tagDust (5).

Assignment of mate-pair reads with unidentifiable barcode

The two long jumping libraries (8kb, 20kb) were sequenced in the same Illumina sequencing lane. For unknown reasons, the fraction of read pairs with unidentifiable barcodes comprised around one third of the total sequence data (SRR3157034). We used alignments against the reference sequence to assign those reads to the mate pair libraries. Pairs with a distance between 5 kb and 12 kb were assigned to the 8 kb library and pairs aligning with a distance between 15 kb and 35 kb, to the 25 kb library. Thereby, we obtained 3,731,986 and 609,046 additional reads for 8 kb and 20 kb library, respectively.

De novo assembly with ALLPATHS-LG and SSPACE

We ran ALLPATH-LG (r40912, standard parameters) (6, 7) using all data from all preprocessed libraries. We ran SSPACE-shortread (8) (SSPACE-SR, parameters: -x 0 -m 32 -o 20 -t 0 -k 3 -a 0.3 -n 15 -p 1 -v 0) using the 3 kb, 8 kb and 20 kb libraries to connect scaffolds generated by ALLPATHS-LG. We manually inspected the proposed connections and only merged scaffolds larger than 100kb.

Assessing assembly statistics in respect to sequencing depth

To interrogate if additional sequencing depth for the lowly sequenced jumping libraries would have improved the assembly, we performed additional runs of ALLPATHS-LG using subsets of the data. Scaffold N50 of an assembly generated without the 8 kb and 20 kb insert data was 260 kb. Adding half of the data of the 8 kb library increased N50 to 1.1 Mb, but did not improve much more when adding the
entire read set. The same trend, however at an N50 of 4.1 Mb, was observed comparing the use of 50% or 100% of data from the 20 kb library. When combining reads from both libraries, already 10% of the data from both libraries were sufficient to achieve optimal results, evidencing that despite the overall low depth-of-coverage with long-range mate-pairs, sufficient sequencing data was available for scaffolding (SI Appendix, Figure S1).

Integration of PacBio data using PBJelly and SSPACE-LongRead

We integrated publicly available 17x coverage long-read PacBio Ler data (9) into the Illumina assembly. For this, we ran PBJelly (10) (v15.2.20, standard parameters) to close gaps and SSPACE-LongRead (11) to connect scaffolds (-i 90 -t 24 -l 2 -k 1) (SI Appendix, Table S3). These procedures reduced the total number of scaffolds from 3,147 to 2,804, and reduced the number of ambiguous bases from 8.1 Mb to 7 Mb. We also tested the impact of using a subset of PacBio reads for both algorithms independently (SI Appendix, Table S3).

Integration of genetic maps

For incorporation of genetic information, we used two public genetic maps, broadly following the POPSEQ approach (12). The first map was calculated from genotyping data of 100 RILs derived from a cross between Col-0 and Ler by 25 bp-oligo-array hybridization of marker regions (13). The second map was based on Col-0/Ler F1 hybrids, which were backcrossed to Col-0 using the F1 as the male. Of the resultant population, 1,505 plants were genotyped (14). Overall, we could assign 676 and 386 markers of the first and second map to a location on the scaffolds. This allowed us to split five long scaffolds, where contig connections were not in agreement with the genetic information and to connect 31 scaffolds into five pseudo-molecules with a cumulative length of 116.9 Mb. Additional integration of seven scaffolds based on synteny assumptions with a combined length of 1.4 Mb led to the final assembly of five chromosome-representing scaffolds. We connected the
scaffolds by inserting a stretch of 100 Ns to represent the un-assembled space.

Finally, we used a public PacBio-only assembly (https://github.com/PacificBiosciences/DevNet/wiki/Arabidopsis-P5C3) to correct 3,483,293 ambiguous bases (Ns) in the assembly. We manually ensured that the newly inserted bases corresponded to the correct region by confirming a perfect match of 1 kb flanking to the gap.

**De novo gene annotation**

We ran AUGUSTUS (v2.7) to perform a *de novo* gene annotation on the final assembly. In order to construct gene models, AUGUSTUS was run with hint files from different sources, including repetitive elements identified by RepeatMasker (http://www.genome.washington.edu/uwgc/analysistools/repeatmask.htm), alignments generated from public RNA-seq data (15, 16) using Tophat2 (17), alignments generated from *de novo* assembled transcripts using Trinity (18) with the same RNA-seq data and alignments of the TAIR10 reference protein sequences. Finally, we used transposable elements identified by RepeatMasker to identify genes containing TE sequences and differentiate them from the set of protein coding genes.

**Identification of large re-arrangements and local differences**

**Whole-genome alignment**

The whole genome alignment was performed using MUMmer (19). The output of *nucmer* (maxmatch, c=100, b=500, l=50) was filtered by *delta-filter* using minimum alignment length of 50, minimum identity of 95% and maximum gap size of 100. To visualize the whole-genome alignment, we used *mummerplot* after filtering alignments shorter than 10 kb (*SI Appendix*, Figure S2).
Identification of allelic and non-allelic regions (higher order variation)

First, we classified the homology blocks generated by MUMmer’s `show-diff` function into allelic (co-linear) and non-allelic (re-arranged) blocks. For this, we searched for the longest syntenic path through all homology blocks, in a way that all homology blocks in this path are not re-arranged to each other. This defined the allelic regions and thereby the syntenic backbone of the chromosomes. Homology blocks, which were not part of these allelic regions, were defined as transpositions or inversions depending on their location and orientation to their neighboring blocks. Finally, neighboring blocks, which were assigned to the same allelic or non-allelic region, were joined into one co-linear or re-arranged region, respectively.

Small-scale genomic variation

Small-scale polymorphisms like SNPs can be captured in the alignments of the homology blocks and were extracted using `show-snps`. Depending on the type of the homology block, local variation was classified as allelic or non-allelic variation.

Classification of local structural variation

Next, we classified the gaps in between the homology blocks of the same regions, excluding gaps residing in between an allelic and a non-allelic homology block. Gaps between allelic regions describe local events and we distinguished between long indels mediated by repeats or not, tandem CNVs and highly diverged regions (HDR). Indels without flanking repeats feature no overlap between the two flanking alignments and can be identified as only one of the two genomes shows a gap between these flanking alignments. In contrast, repeat-mediated events feature an overlapping alignment in one genome whereas the other genome shows a gap in between flanking repeats. Tandem CNVs feature overlapping alignments in both genomes, having a varying size in only one of them. Highly diverged regions (HDRs) have a gap in both genomes, displaying high nucleotide diversity between the two gaps. All cases are illustrated in Figure 2A.
**Analysis of non-aligned regions**

Finally, we defined all remaining homology block breaks (those that reside between allelic and non-allelic regions) as non-aligned regions. We extracted the sequences of these non-aligned regions from both genome assemblies and aligned them to the respective other genome using blastn. Regions with a significant (e-value cutoff 1e-04) hit to regions, which themselves are involved in the whole-genome alignment, where defined as duplicated regions.

**Inversion analysis**

**Genotyping of inversion breakpoints**

To characterize the worldwide population of *A. thaliana* with regard to inversion alleles, we used three publically available read data sets (20-22). Of a total of 477 accessions, 409 were uniquely assignable to their accession code and geographic location and showed a sufficient sequencing quality. Short read alignments were performed against the reference sequence and the newly assembled Ler genome using GenomeMapper (23). For each of the inversion breakpoints, we calculated the ratio of alignments to the Col-0 like breakpoint sequence divided through the sum of reads aligned to the breakpoints in both genomes. This generated two clearly separated classes of ratios (histograms of ratios shown in SI Appendix, Fig. S5), which allowed us to characterize each of the accessions for inversion alleles using ratios between 0.0-0.2 and 0.8-1.0 as confident assignments of the Ler and Col-0 haplotypes, respectively. For control we included whole-genome sequencing data of Col-0 and Ler, which were assigned to their respective inversion alleles.

**CO Analysis**

A recent study (26) generated and collected information on the precise location of crossing over events of Col-0 x Ler hybrids. A total of 737 CO sites were
recorded. Of those, 363 were assigned to regions shorter than 1.5 kb, which were further used for overlap analysis with re-arranged regions.

Population differentiation, haplotype diversity, and genic selection patterns

Pairwise $F_{st}$ was calculated between two groups of accessions (with different chromosome 4 inversions alleles) in a sliding window of 250 kb with a 100 kb step size. Haplotype diversity was calculated with all SNPs with a minor allele frequency of at least 5% and complete genotypic information across all accessions. Haplotypes were defined as clusters of identical genotypes in a 10 kb window (according to (20)). Mean haplotype diversity was calculated in a sliding window of 250 kb with a 100 kb step size. $K_a/K_s$ values of all 1-to-1 orthologs were estimated as described above for $K_s$ values.

Identification of homologous, duplicated and novel genes

Calculation of orthologous groups

We ran Inparanoid (v4.0) (27) to identify clusters of orthologs (COGs) between Ler and Col-0. In addition we ran Scipio (v1.4.1) (28) to align the predicted Ler protein sequences to the Col-0 genome and vice versa. Inparanoid predicted 23,475 COGs and 23,255 of them referred to 1-to-1 orthologs between both accessions. The number of proteins that were not assigned to a COG by Inparanoid was surprisingly high, considering that we compared protein annotations from two strains of the same species. We therefore converted Inparanoid COGs and Scipio gene-to-gene pairings into an undirected graph representing gene relationships. Based on this graph we classified all genes into categories such as 1-to-1 orthologs, duplicates and novel genes.

Duplicated genes were then further classified as “dispersed” or “local” depending on their genomic positions. Duplicated genes within syntenic regions (based on Scipio protein-DNA alignment coordinates) were manually inspected and
labeled as “local” if the genes were in close proximity (less than 10 genes or 100kb apart), otherwise they were labeled as “dispersed”. Protein sequences that could not be aligned to the respective other genome were considered as potentially novel genes.

Filtering for accession-specific genes

Potentially novel genes were realigned to all scaffolds (including discarded scaffolds) using tblastn in order to identify cases where a gene was absent due to incomplete assembly of its sequence. In addition, we compared the novel genes to predicted structural variants and transposable elements. Moreover, we checked the expression of the putative novel genes in Ler to reduce potential false positive gene annotations. We excluded genes that apparently do not have an annotated homolog in the syntenic region, but have a gene structure with a functional ORF. These genes may correspond to wrong (missing or false) annotations in one of the genomes.

Functional analysis

Functional enrichment analysis of the accession-specific and the duplicated genes was performed using the TAIR10 gene identifier and the Gene Ontology consortium website (http://geneontology.org/page/go-enrichment-analysis).

Analysis of NBS-LRR genes

We aligned the 5kb up- and downstream sequence of 159 NBS-LRR genes against the Ler assembly. We selected the best alignment and considered only those that aligned at least 30% of their sequence at a minimum of 75% sequence identity. If the gene was not found, we used blat to align the coding sequence of the gene and manually inspected the position in the assembly. We then overlapped these regions with structural variant prediction and checked for ambiguous bases.
Figure S1. Assembly performance of different subsets of long-range insert size libraries. Scaffold N50 increased with more data, however, already 50% of the data of the 8 and the 20 kb libraries were generating close-to-optimal results. Interestingly, using small fractions of both libraries together outperformed the results gained when using all of the data of a single library.
Figure S2. Dot-plot of the whole-genome alignment of Col-0 and Ler. The 1.2 Mb inversion on chromosome 4 can be easily spotted in a broad visualization of the whole-genome alignment generated with MUMmer. (Each dot refers to a broad homologous region. Dots in red for forward alignments, blue dots for reverse alignments.)
Figure S3. PCR sequencing of the chromosome-4-inversion breakpoints in the genomes of Col-0 and Ler. The primer sequences to validate the breakpoints in Col-0 and Ler are shown in the figure. Primers were generated to span the breakpoints for more than 200 bp on each side, using Primer3 plus (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi). Bands 1 to 4 were cut from the gel, the DNA was purified using the Qiagen gel extraction kit protocol and was Sanger-sequenced using the specific input PCR-primer and confirmed the PCR products. All primers were pooled for the negative control.
Figure S4. Local impact of large inversions on female meiotic recombination rate. Green line, cM/Mb values across the chromosomes; grey boxes, location of inversions.

Figure S5. Alternative haplotype short read mapping approach for chromosome-4-inversion genotyping. Short reads were simultaneously aligned to the reference sequences of Col-0 and Ler (illustrated on the left). Reads aligning to the regions referring to the left and right breakpoint of the inversion in both of the genomes were counted. The number of reads aligned to the Col-0 breakpoints was divided through the sum of reads aligned against both breakpoints in both genomes. A ratio of 1 describes the situation that all reads align to the Col-0 haplotype, whereas a ratio of 0 would describe the opposite. Left and right breakpoints were assigned independently. Ratios between 0.0-0.2 and 0.8-1.0 were considered a confident assignment of the Ler and Col-0 haplotypes, respectively. On the right, the histograms show the distribution of ratios for genotyping 410 accessions of the left
and right breakpoint of the inversion on chromosome 3 (upper panels) and chromosome 4 (lower panels).

Figure S6. Genome-wide haplotype diversity within the accessions carrying the Col-0 (red) or Ler (blue) allele of the chromosome-4-inversion.
Figure S7. Genome-wide population differentiation ($F_{st}$) values between the two groups of accessions carrying either a Col-0-like or a Ler-like chromosome-4-inversion allele.
**SI Appendix, Supplementary Tables**

**Table S1. Sequencing libraries and data generated for the assembly.**

<table>
<thead>
<tr>
<th>Library type</th>
<th>Technology</th>
<th>Fragment length</th>
<th>Avg. read length</th>
<th>Read number</th>
<th>Filtered read number</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paired-end</td>
<td>Illumina</td>
<td>180</td>
<td>2x100</td>
<td>305,644,718</td>
<td>185,572,308</td>
<td>This study</td>
</tr>
<tr>
<td>Paired-end</td>
<td>Illumina</td>
<td>180</td>
<td>2x100</td>
<td>58,699,780</td>
<td>42,480,892</td>
<td>(1)</td>
</tr>
<tr>
<td>Long jumping</td>
<td>Illumina</td>
<td>8,000</td>
<td>2x100</td>
<td>51,332,776</td>
<td>38,024,342</td>
<td>This study</td>
</tr>
<tr>
<td>Long jumping</td>
<td>Illumina</td>
<td>20,000</td>
<td>2x100</td>
<td>61,030,552</td>
<td>12,060,434</td>
<td>This study</td>
</tr>
<tr>
<td>Mate pair</td>
<td>Illumina</td>
<td>3000</td>
<td>2x80</td>
<td>8,339,024</td>
<td>7,511,918</td>
<td>(1)</td>
</tr>
<tr>
<td>Single molecule</td>
<td>PacBio</td>
<td>*</td>
<td>9 kb</td>
<td>212,997</td>
<td>212,997</td>
<td>(29)</td>
</tr>
</tbody>
</table>

* PacBio sequencing data included polished reads, but no information on raw fragment length.

**Table S2. List of samples published as Ler genomes.**

<table>
<thead>
<tr>
<th>Name used here (not origin of sample)</th>
<th>Name used in original work</th>
<th>Stock center accession</th>
<th>SRA accession</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ler qrt1 (Shanghai)</td>
<td>Ler qrt1</td>
<td></td>
<td>SRP007172</td>
<td>(30)</td>
</tr>
<tr>
<td>Ler (Tübingen)</td>
<td>Ler-1</td>
<td>CS22686 (ABRC)</td>
<td>1001genomes.org</td>
<td>(1)</td>
</tr>
<tr>
<td>Ler (Cambridge)</td>
<td>Ler-0</td>
<td>CS20 (ABRC)</td>
<td>ERP000565</td>
<td>(31)</td>
</tr>
<tr>
<td>Ler qrt1 (NASC)</td>
<td>Ler qrt1</td>
<td>N8050 (NASC)</td>
<td>ERP003793</td>
<td>(32)</td>
</tr>
<tr>
<td>Ler (Cologne)</td>
<td>Ler</td>
<td></td>
<td></td>
<td>This study</td>
</tr>
<tr>
<td>Ler (Wageningen)</td>
<td>Ler wt</td>
<td></td>
<td>ERP003793</td>
<td>(32)</td>
</tr>
<tr>
<td>Ler (PacBio)</td>
<td>Ler-0</td>
<td>WT-04-19-02</td>
<td>SRX533607</td>
<td>(29)</td>
</tr>
</tbody>
</table>
Table S3. Assembly quality metrics after PacBio read data integration using PBJelly and SSPACE and three different subsets of the data.

<table>
<thead>
<tr>
<th>Amount of data</th>
<th>10%</th>
<th>50%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PBJelly</td>
<td>PBJelly</td>
<td>PBJelly</td>
</tr>
<tr>
<td># Scaffolds</td>
<td>3,147</td>
<td>3,082</td>
<td>3,074</td>
</tr>
<tr>
<td>Longest scaffold</td>
<td>15,161,445</td>
<td>15,214,508</td>
<td>15,227,620</td>
</tr>
<tr>
<td>N50 scaffold</td>
<td>7,499,974</td>
<td>7,510,181</td>
<td>7,519,546</td>
</tr>
<tr>
<td>#Ambiguous bases</td>
<td>8,178,580</td>
<td>6,586,655</td>
<td>5,924,618</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SSPACE-LR</th>
<th>SSPACE-LR</th>
<th>SSPACE-LR</th>
</tr>
</thead>
<tbody>
<tr>
<td># Scaffolds</td>
<td>2,827</td>
<td>2,805</td>
<td>2,804</td>
</tr>
<tr>
<td>Longest scaffold</td>
<td>15,161,445</td>
<td>15,214,508</td>
<td>29,446,335</td>
</tr>
<tr>
<td>N50 scaffold</td>
<td>7,508,294</td>
<td>7,537,228</td>
<td>12,809,207</td>
</tr>
<tr>
<td>#Ambiguous bases</td>
<td>8,976,222</td>
<td>7,585,348</td>
<td>7,064,230</td>
</tr>
</tbody>
</table>
Supporting Material References


