Positional Information Resolves Structural Variations and Uncovers an Evolutionarily Divergent Genetic Locus in Accessions of *Arabidopsis thaliana*

Alvina G. Lai[1], Matthew Denton-Giles[1], Bernd Mueller-Roeber[2,3], Jos H. M. Schippers[2,3], and Paul P. Dijkwel*[,1]

[1]Institute of Molecular BioSciences, Massey University, Private Bag 11-222, Palmerston North 4442, New Zealand

[2]Department of Molecular Biology, Institute of Biochemistry and Biology, University of Potsdam, 14476 Potsdam-Golm, Germany

[3]Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam-Golm, Germany

*Corresponding author: E-mail: p.dijkwel@massey.ac.nz.

Accepted: 17 April 2011

Abstract

Genome sequencing of closely related individuals has yielded valuable insights that link genome evolution to phenotypic variations. However, advancement in sequencing technology has also led to an escalation in the number of poor quality–drafted genomes assembled based on reference genomes that can have highly divergent or haplotypic regions. The self-fertilizing nature of *Arabidopsis thaliana* poses an advantage to sequencing projects because its genome is mostly homozygous. To determine the accuracy of an *Arabidopsis* drafted genome in less conserved regions, we performed a resequencing experiment on a ~371-kb genomic interval in the Landsberg *erecta* (L*er*-0) accession. We identified novel structural variations (SVs) between L*er*-0 and the reference accession Col-0 using a long-range polymerase chain reaction approach to generate an Illumina data set that has positional information, that is, a data set with reads that map to a known location. Positional information is important for accurate genome assembly and the resolution of SVs particularly in highly duplicated or repetitive regions. Sixty-one regions with misassembly signatures were identified from the L*er*-0 draft, suggesting the presence of novel SVs that are not represented in the draft sequence. Sixty of those were resolved by iterative mapping using our data set. Fifteen large indels (>100 bp) identified from this study were found to be located either within protein-coding regions or upstream regulatory regions, suggesting the formation of novel alleles or altered regulation of existing genes in L*er*-0. We propose future genome-sequencing experiments to follow a clone-based approach that incorporates positional information to ultimately reveal haplotype-specific differences between accessions.

Key words: haplotype, allelic variants, drafted genomes, genome partitioning, comparative genomics.

Introduction

The number of genome projects of various scales has increased substantially over the years due to a reduction in sequencing costs as technology advances (Chain et al. 2009). Many laboratories benefit from this impressive technological advancement in terms of rapid generation of high-depth sequence data. However, next-generation sequencing (NGS) platforms are compromised in their ability to generate long reads. Read length reduction is compensated by an increase in coverage where 20- to 30-fold redundancy has been reported as the acceptable criterion by most genome projects (Bentley et al. 2008; Ossowski et al. 2008). Due to the nature of short-read data sets, drafted genomes are assembled based on preexisting published sequences and the quality of the resulting data has yet been sufficiently diagnosed. This has led to the mass release of drafted genomes (Chain et al. 2009), many of whose qualities are only assessed by identifying the number of assembly gaps. Other valuable diagnostic criteria such as the number of errors and misassemblies are potentially missing and can only be revealed with fine-scale analysis.

Computational algorithms have been developed specifically to tackle short-read data sets (Butler et al. 2008; Ossowski et al. 2008; Zerbino and Birney 2008; Simpson

et al. 2009). The identification of single nucleotide polymorphisms (SNPs; Shen et al. 2010) and small insertion–deletion polymorphisms (indels; Krawitz et al. 2010) using a combination of multiple assembly algorithms that are each designed and optimized for different purposes had seemed to be the end goal of genome projects as other forms of deviations relative to the reference genome remain challenging to detect. Resolving SVs, that is, changes that are not single nucleotide variants, such as duplications, inversions, large indels, and copy number variations (CNV) (Feuk et al. 2006; Frazer et al. 2009), have been proven problematic for short-read assemblers (Snyder et al. 2010). Prior to the arrival of NGS technology, comparative genomic hybridization using oligonucleotide arrays have been extensively used as analysis tools for the discovery of submicroscopic SVs (Sebat et al. 2004; Gresham et al. 2008). Recently, several methods have been developed to detect SVs from NGS data sets (Korbel et al. 2007; Chen et al. 2009; Snyder et al. 2010). The accuracy of these techniques remains to be sufficiently tested particularly on highly complex eukaryotic genomes. An example that can potentially result in assembly error is when a tandem duplication spanning across an inversion allele may be interpreted as a de novo complex duplication if only one inversion haplotype is represented in the reference genome (Zhang et al. 2009). The lack of strategies to transverse across rearrangements and co-occurrences of SVs between chromosomal haplotypes can cause assembly gaps as sequence reads from paralogous regions are mistaken as allelic overlaps when they map to a single location (Bailey et al. 2001; Sharp et al. 2006). This problem further complicates accurate variant calling and may hamper large indel detection in such regions. Improper placement of scaffolds may also introduce nonexistence of heretical evolutionary breakages (Lewin et al. 2009).

Arabidopsis thaliana, a flowering plant from the Brassicaceae family, is one of the best studied plant species due to its tractability and the number of research tools available. The self-compatible nature of *Arabidopsis* has allowed each accession or lineage to evolve independently yielding diverse populations that display a multitude of phenotypic variations (Koornneef et al. 2004). Several groups have embarked on the 1001 *A. thaliana* genome project (Weigel and Mott 2009) dedicated to generate genome sequences from numerous accessions of this species. Comparative genomics have frequently been used as a tool to study evolution by natural selection (Feuillet and Keller 2002; Nishiyama et al. 2003; Bowman et al. 2007; Koonin 2009). By comparing two or more genomes, one can infer how natural selection acts in different lineages in driving sequence evolution in genes and nongenic regions and how these changes relate to phenotypic evolution and adaptation (Ellegren 2008). Investigating patterns of divergence around known functional elements could yield insights on the effect that different forces, for example, purifying selection and genetic hitchhiking (Cai et al. 2009), have on genetic polymorphisms (Altshuler et al. 2010).

It has been reported that approximately one quarter of the *A. thaliana* reference genome involves regions that are highly divergent with the presence of rare alleles in at least one accession (Zeller et al. 2008). Genomic SVs underlie phenotypic differences between *A. thaliana* accessions (Fransz et al. 2000; Meyers et al. 2005; Alonso-Blanco et al. 2009). SVs are predominantly multigenic or even multiloci and may not be represented in the reference accession. The role of SVs in chromosomal speciation has been shown in several models (White 1978), an example being the suppressed-recombination model where a genetic barrier is formed between populations. Substitutions linked to these rearranged chromosomes cannot be exchanged, thereby promoting genomic incompatibilities and hence speciation (Rieseberg et al. 1999; Perry et al. 2008; Bikard et al. 2009; Marques-Bonet et al. 2009; Alcázar et al. 2010). Complex SVs also promote genome instability by long-distance nonallelic homologous recombination leading to further CNV (Johnson et al. 2006). Orthologous regions enriched with ancestral segmental duplications may serve as hot spots for constant genomic turnover, and recurrent CNV genesis happens as a result of evolutionarily shared duplications occurring across and within species (Perry et al. 2008).

The stream of drafted genomes released has far outnumbered the small group of high-quality genomes (Chain et al. 2009). Downstream comparative genomics heavily depends on the fidelity of these drafts. A poor quality draft is therefore prone to misinterpretations (Choi et al. 2008; Meader et al. 2010). Here, we performed a fine-scale assessment of the Landsberg *erecta* (L*er*-0) drafted genome at a selected polymorphic locus. We identified and resolved novel SVs in a contiguous L*er*-0 locus using high-coverage Illumina reads that were generated from an experimental method that incorporates positional information. This work not only highlights the importance of rectifying errors on drafted genomes before they are used in downstream applications but also provides an unprecedented view on genomic divergence in an inbred species. We propose future genome projects to proceed in a manner that incorporates positional information in order to improve genome assembly and to reveal large deviations from reference genomes.

Materials and Methods

Genomic DNA Extraction

Arabidopsis thaliana seed stocks for the L*er*-0 accession were obtained from the Nottingham Arabidopsis Stock Centre (ID: NW20). High-quality genomic DNA suited for long-range (LR)–polymerase chain reaction (PCR) amplification was extracted from 21-day-old frozen leaf material according to the modified method of van der Biezen (van

der Biezen et al. 1996). Four grams of leaf tissue was ground in liquid nitrogen and vortexed in 25 ml chilled extraction buffer (0.35 M sorbitol, 0.1 M Tris–HCl, 5 mM ethylenediaminetetraacetic acid [EDTA], pH 7.5, 20 mM $Na_2S_2O_5$). The crude extract was centrifuged at 14,000 revolutions per minute (rpm) for 1 h at 4 °C, and the supernatant was discarded. A 1.25 ml of extraction buffer, 1.75 ml nucleus lysis buffer (0.2 M Tris–HCl, 50 mM EDTA, 2 M NaCl, 2% hexadecyl-trimethyl-ammonium bromide pH 7.5), and 0.6 ml of 5% sarkosyl were used to dissolve the pellet. The mixture was subsequently incubated for 1 h at 65 °C. Chloroform/isoamylalcohol (24:1 v/v) extraction was performed by adding 7.5 ml of the solvent mixture to the tube, followed by centrifugation at 14,000 rpm for 15 min. Clear supernatant was transferred to a clean tube, and DNA was precipitated with an equal volume of chilled isopropanol and incubated on ice for 20 min before centrifugation at 14,000 rpm for 15 min. The isopropanol was decanted, and the pellet was washed with 70% ethanol and air dried for 20 min. The pellet was dissolved in 500 μl Tris–ethylenediaminetetraacetic acid (TE) buffer containing 10 μl of 10 mg/ml RNaseA. Genomic DNA was stored at 4 °C to prevent multiple freeze-thaw sessions that might hamper LR-PCR amplifications.

LR-PCR Amplification and Illumina Sequencing

Primers for LR-PCR were designed using Primer3 (Rozen and Skaletsky 2000) to amplify overlapping genomic fragments of 647–13,702 bp, spanning an ~371-kb contiguous locus in L*er*-0 (supplementary table 2A, Supplementary Material online). LR-PCR amplifications (milliQ water: 75.6 μl; 10× buffer: 10 μl; deoxyribonucleotide triphosphate [2.5 mM]: 8 μl; forward primer [10 μM]: 2 μl; reverse primer [10 μM]: 2 μl; high-fidelity Takara ExTaq enzyme [5 units/μl]: 0.4 μl; DNA template [90 ng/μl]: 2 μl for 100 μl reaction) were performed using an autosegment extension program (3 min 94 °C/30 s 94 °C, 30 s 62 °C, 5–10 min 68 °C, 30× cycles/5 min 68 °C), increasing the extension time for 15 s each cycle after 14 cycles in the PalmCycler. PCR products were separated on 0.8% (for fragments larger than 10 kb) or 1.0% (for fragments smaller than 10 kb) 1× Tris-acetate-EDTA gel for amplicon size confirmation, followed by purification using the QIAquick PCR Purification Kit. Concentration of each purified PCR product was quantified. L*er*-0 amplicons were pooled in equal molarity to yield DNA in the concentration of 5 μg/50 μl TE. Sequencing was performed on the GAII to generate a 75-bp single-read data set.

Pipeline Analysis and Read Trimming

Illumina Pipeline version 1.6 was used for pipeline analysis. Off-Line Basecaller programs, Firecrest and Bustard, were used for image analysis and base calling, respectively. Approximately 87.9% of clusters passed filtering. The GERALD module in CASAVA 1.6 was used to combine tile-based .qseq files into a single .txt file. File conversion from .qseq to .fastq was done using SSAKE (Warren et al. 2007) qseq2fastq.pl script. Reads were trimmed according to a Phred score of 20 using the TQSfastq.py script. SSAKE was further utilized to generate de novo contigs under the following parameters—*m*: 15 (minimum number of overlapping bases with the seed during overhang consensus build up) and *x*: 15 (minimum overlap between contigs to merge adjacent contigs in a scaffold).

Detection of Misassemblies and Variant Identification

Trimmed reads were assembled to either Col-0 reference or L*er*-0 draft sequence using Geneious assembler (Drummond et al. 2010) by allowing 4–6 mismatches and 5–50 bp gaps to account for indels. Misassemblies were identified by detecting aberrant assembly signatures in Geneious. Two hundred bases at the left and right flanks of the ambiguous regions were extracted and used as references for targeted iterative read mapping described in the Results section. A SHORE consensus analysis (Ossowski et al. 2008) was performed to obtain GC content and errors in read positions. SNPs were identified using the Find Variations/SNPs option in Geneious by setting the minimum coverage parameter to 100 and minimum variant frequency parameter to 0.8. Locus alignments of Col-0, L*er*-0 draft, and L*er*-0 revised sequences were generated using the progressiveMauve aligner (Darling et al. 2010).

Validation of SVs by Sanger Sequencing

Several resolved indels were randomly chosen for validation by Sanger sequencing. Genomic DNA was PCR amplified using primers designed by Primer3. Both L*er*-0 and Col-0 alleles were amplified, and size differences were visualized on an agarose gel. The L*er*-0 allele was subjected to dideoxy sequencing by the ABI-Sanger instrument followed by alignment of the sequence trace to the iteratively resolved indel for validation purposes.

Data Deposition

L*er*-0_chromosome_3_locus.fasta (GenBank: HQ698308).

Results

LR-PCR Amplification of a Polymorphic L*er*-0 Genomic Interval

An ~371-kb genomic interval on chromosome 3 (Col-0 position: 16653794–17025087) that spans six Col-0 bacterial artificial chromosomes (BACs), that is, F18N11, F9K21, T6D9, F16L2, F12M12, and F18L15, was selected for the study of the prevalence of SVs between a reference (Col-0) and a nonreference (L*er*-0) *Arabidopsis* accession. LR-PCR was used to amplify overlapping genomic fragments

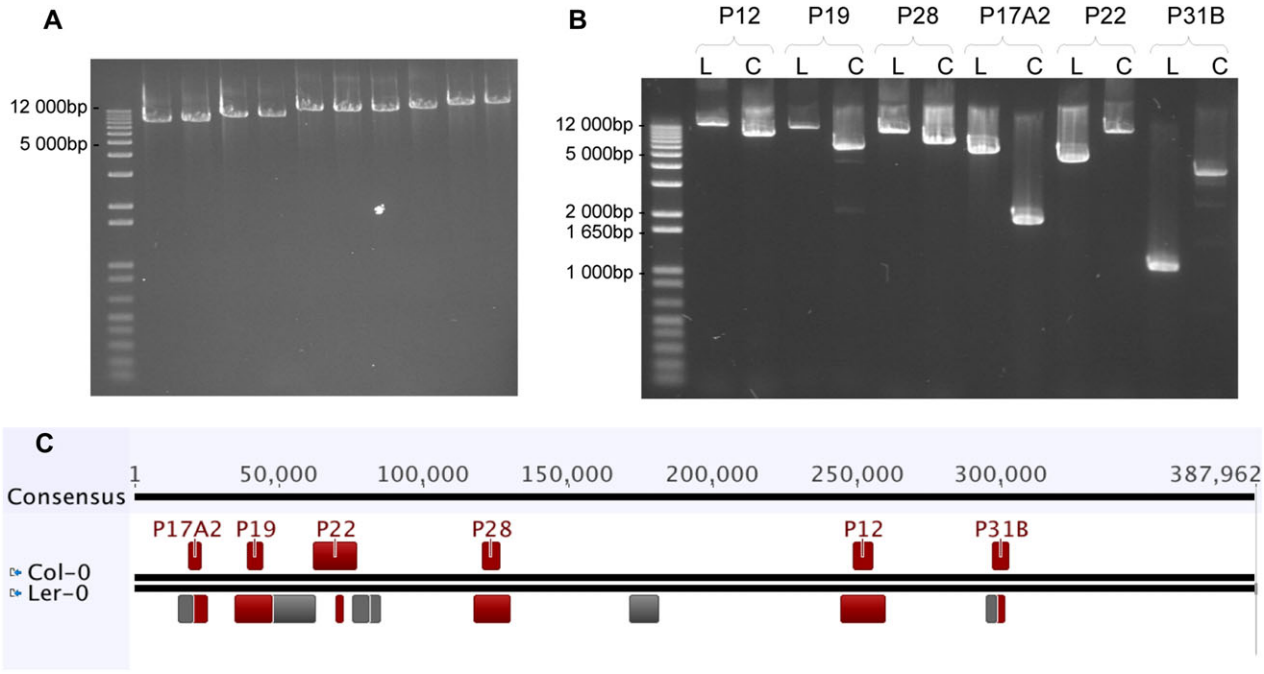

Fig. 1.—LR-PCR amplification and large indel polymorphisms between L*er*-0 and Col-0. L*er*-0 locus that corresponds to 16653794-17025087 positions on Col-0 chromosome 3 is amplified in 49 overlapping fragments using LR-PCR. (*A*) Several examples of LR-PCR amplicons are shown on the gel. (*B*) Gel image depicts large indel polymorphisms between L*er*-0 and Col-0. Amplicon identifier: L, L*er*-0 allele; C, Col-0 allele; P*X*, primer identifier. (*C*) Illustration of locus-specific genetic architecture between L*er*-0 and Col-0. Putative large indels are represented as red blocks and six unamplified gaps as gray blocks.

within the locus with amplicon sizes ranging from 647 to 13,702 bp (fig. 1*A*). LR-PCR was performed in two steps. In the first step, 40 primer pairs were used for amplification, and we were able to obtain 29 out of 40 amplicons. In the second step, an additional 26 primer pairs were designed to divide regions that were not obtained in the first round into 2 or 3 smaller fragments (supplementary table 2A, Supplementary Material online). The second round of amplification is crucial to rule out chances of obtaining no amplicons due to misannealing of the first primer pairs to polymorphic L*er*-0 sites because Col-0 is used as the reference for primer design. From the second round, 20 additional amplicons were obtained. The entire locus is spanned by 49 amplicons, including six gaps that were not covered by PCR (fig. 1*C*). In an attempt to bridge the gaps, additional primers that spanned those gaps were designed. However, we were still unable to obtain any amplicon for the six gaps, suggesting the presence of large insertions in these regions that are beyond amplifiable range, that is, larger than 25 kb (data not shown). The locus of study is partitioned into 49 amplicons that represent genomic fragments obtained from known locations and thus having positional information. Comparison between Col-0 and L*er*-0 amplicon lengths revealed that at least six amplicon pairs harbor large indel polymorphisms (fig. 1*B*). We hypothesized that in addition to these six large indels, a considerable number of indels of significant size remained undetected due to limited gel resolution. Overall, the PCR results suggest that large SVs exist within the selected genomic region between the two *Arabidopsis* accessions.

High-Coverage Sequencing of Amplicons to Detect Interaccession SVs

Illumina Sequencing and Read Mapping to the Col-0 Reference. To precisely capture the sequence context of these SVs, we proceeded to sequence the ~371-kb contiguous locus in L*er*-0 using the Illumina Genome Analyzer II platform. L*er*-0 amplicons were pooled in equal molarity (supplementary fig. 1*A*, Supplementary Material online) and sequenced to generate a 75-bp single-read data set (supplementary fig. 1*B*, *C*, and *D*, Supplementary Material online) with positional information. The filtered and quality-trimmed reads were assembled to the Col-0 reference locus by allowing up to four mismatches and gaps of up to 50 bp to permit small indel detection. Using the Geneious software (Drummond et al. 2010), misassembly signatures, indicative of SVs, were identified. Deletions in L*er*-0 were seen as gaps in the assembly and insertions as arrays of consecutive mismatches (supplementary fig. 2, Supplementary Material online). To locate the region of the six large indels as observed from differences in Col-0/L*er*-0 amplicon

lengths (fig. 1*B*), all primer sequences were aligned to the Col-0 reference. By tracking the flanking primers for the six large indel amplicons, misassembly signatures found at those regions confirmed the occurrences of indels in L*er*-0. In addition, 123 non-SNP misassembly signatures were found (excluding the six unamplified gaps) that corroborated our initial speculation on the presence of additional indels that fall below the range of gel-based detection.

Read Mapping to the Ler-0 Draft. The Wellcome Trust Centre for Human Genetics (WTCHG) has generated an L*er*-0 draft genome from 36- to 51-bp paired-end Illumina libraries of approximately 40-fold coverage. As part of our analysis, we subsequently used the L*er*-0 draft as the reference for read mapping based on the assumption that the draft sequence would be a better reference than Col-0. In parallel, our analysis will also serve as an indicator of L*er*-0 draft sequence quality. Reads were assembled to the L*er*-0 draft by allowing up to six mismatches and 5 bp gaps. The number of non-SNP misassembly signatures was reduced from 123 misassemblies down to 61 misassemblies when the draft sequence was used. However, the large SVs detected from PCR amplicon sizing were not represented in the L*er*-0 draft sequence (table 1). The L*er*-0 draft was generated by a combination of de novo assembly and reference-based mapping. Hence, a large pool of de novo contigs could not be incorporated in the draft due to lack of sequence context from the Col-0 reference and the lack of positional information for these contigs. Therefore, it is expected that SV sequence information remained in the pool of unmapped contigs.

Improving Local Assembly to Reveal SVs Using Targeted Iterative Read Mapping. The PCR-based approach provides us with the information that reads obtained originate from the target locus and not from other genomic regions. We assumed that the pool of unmapped reads (~7%) accounted for SVs. Geneious assembler was used to perform a targeted iterative read-mapping step to map these reads to their designated regions. Each misassembled region was flagged, and their left and right flanking sequences were extracted for iterative mapping. Iterative read mapping consists of the following five steps (fig. 2): 1) Extract 200-bp sequences that flank the misassembled region (these flanks serve as reference sequences for subsequent iterative mapping). 2) Map all reads to both flanks independently. 3) After each round of iteration, reads that assembled to the border of the flank will have sequences extended beyond this flank. The extended sequence is then incorporated to the border of the initial flank to produce a longer flank that is the combination of the initial flank and the assembled read sequence (approximately 45–50 bp for each iteration). Reads are then remapped to the new reference flank. 4) Repeat steps 2 and 3 until the left and right iteratively "extended" flanks overlap and can be aligned. 5) Incorporate the new local consensus sequence into the reference sequence followed by realigning all original reads to the modified reference.

Table 1
L*er*-0 Amplicon Size Estimates Correlate with the Actual Lengths in the L*er*-0 Revised Sequence

Primer ID	Gel-Estimated L*er*-0 Amplicon Length (bp)	Col-0 Length (bp)	L*er*-0 Draft Length (bp)	L*er*-0 Revised Length (bp)
P12	13,500	9,816	9,816	14,526
P19	12,000	6,811	6,989	13,389
P28	11,000	8,683	8,755	12,285
P17A2	7,000	2,084	2,084	6,782
P22	5,000	13,702	13,741	5,691
P31B	1,300	4,282	4,340	1,248

NOTE.—L*er*-0 amplicon lengths were estimated on an agarose gel, and Col-0 lengths were obtained from TAIR. The corresponding lengths of these amplicons were determined by mapping flanking primer sequences to the L*er*-0 draft and L*er*-0 revised sequence. P*X*, primer identifier.

Manual iterative steps allowed us to pinpoint problematic regions that could not be resolved by automated assembly programs. In a particular iterative step when there was more than one possible read option for subsequent contig extension (fig. 3*A*), we could not proceed onto the next iteration. Instead, iterative read mapping was performed from the opposite flank until it could be aligned to the previous flank. Regions were flagged as unresolved when more than one read option was obtained from both left and right flank extensions, as selecting any one of these possible read options would ultimately result in an incorrect final consensus sequence. This step is crucial to prevent the generation of incorrect chimeric contigs that occur when attempting to assemble duplicated or conserved regions. By referring to each amplicon size, the newly assembled sequence can be cross-checked with the estimated PCR product length.

Out of the 61 misassembled regions, 57 were resolved (supplementary table 1, Supplementary Material online) by local iterative mapping, whereas the remaining four regions could not be confidently determined. For the first three regions, more than one option in iterative extensions from both flanks was present (Fig. 3*A*). Nevertheless, initial iterative results suggested the presence of duplications in these regions. We subsequently attempted to resolve these regions by making use of de novo contigs generated by SSAKE (Warren et al. 2007) using only unmapped reads. In the first two regions, a single de novo contig mapped to each of the corresponding iterative flanks. The contigs were incorporated into the flanks, and iterative mapping was performed to validate the contig sequence. In the third region, more than one contig mapped to the flanks, and the correct one could therefore not be confidently identified without further analysis. Thus, 2 out of the 3 regions were resolved by de novo contig mapping combined with an iterative validation step. In the fourth region, we encountered stretches of long CT-AG inverted repeat sequences from

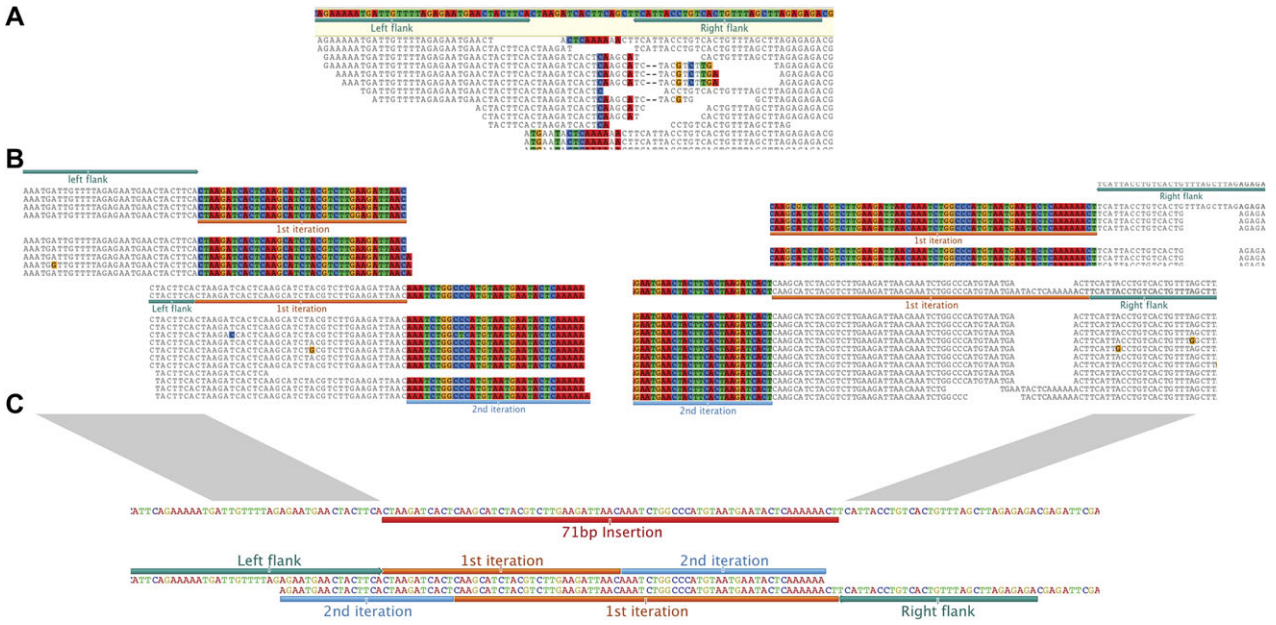

Fig. 2.—Draft sequence correction by iterative read mapping. (*A*) An insertion site is identified by detecting misassembly patterns as described in supplementary figure 2 (Supplementary Material online). Left and right sequences that flank the incorrect region on the draft are used as references for local iterative read mapping. (*B*) In this particular case, two rounds of iterative mapping from both flanks are sufficient to span the insertion. (*C*) Alignment between iteratively extended left and right flanks.

both left and right flank directions (fig. 3*B*). This region is estimated to be 2 kb in length by cross-checking to its corresponding amplicon size (P19 in table 1). An ~1-kb de novo contig flanked by CT and AG sequences was identified and was confirmed to be present within the region by restriction digestion on the PCR amplicon from this region. Because the CT-GA repeats extended beyond the read length (reads that consist entirely of these dinucleotide repeat sequences were identified), the actual length of the repeats could not be deduced. Nevertheless, the results suggest that the total length of the combined CT and AG repeats is close to 1 kb. Repeat expansion has been found

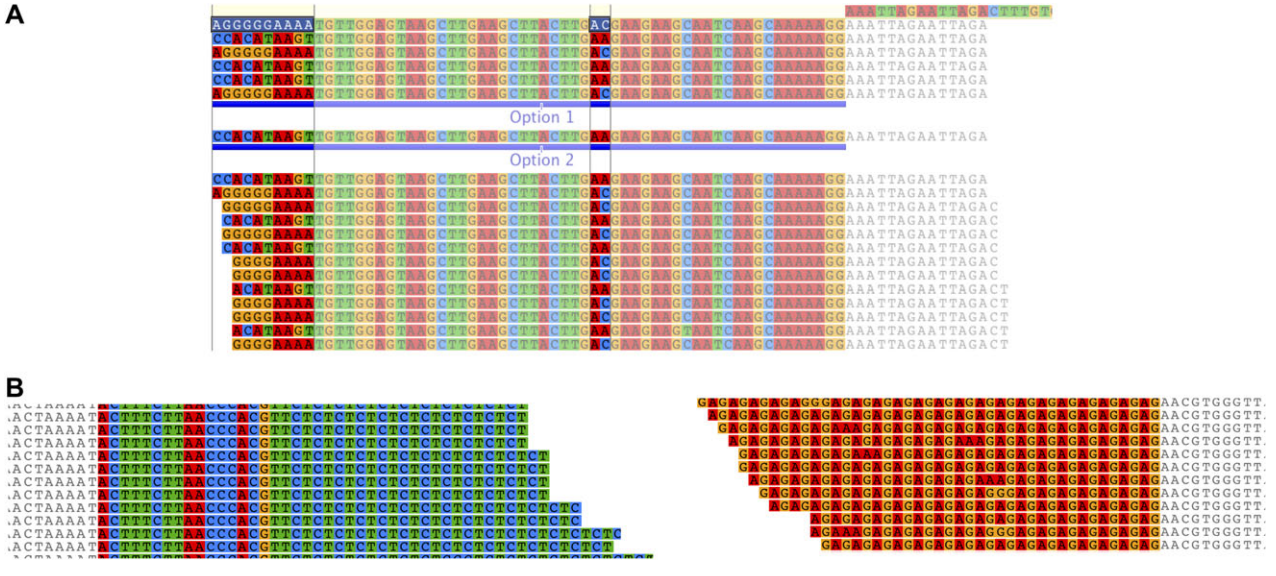

Fig. 3.—Limitations of iterative read mapping. (*A*) Figure illustrates more than one possible read option obtained during iterative mapping. Iterative extension is then performed from the opposite flank to prevent the generation of chimeric contigs. (*B*) Figure depicts a stretch of long inverted dinucleotide repeat in L*er*-0 that is absent from the Col-0 genome. Further iterative steps are not possible in this region as repeat length is longer than the read length. This region is estimated to be 2 kb in length based on PCR amplicon size.

Table 2
Comparative Analysis of Col-0, Ler-0 Draft (WTCHG), and Ler-0 Revised Sequence Using Locus-Specific and Whole-Genome Ler-0 Reads

	No. of Aligned Locus-Specific Ler-0 Reads (Mean Coverage)	No. of Aligned WTCHG Whole-Genome Ler-0 Reads (Mean Coverage)
Col-0	2,002,286 (375.4)	161,312 (16.1)
Ler-0 draft (WTCHG)	3,096,868 (595.5)	210,349 (21.9)
Ler-0 revised	3,432,240 (643.6)	220,178 (22.5)

NOTE.—Locus-specific reads and whole-genome reads are aligned to the Ler-0 draft and revised sequences.

to cause environment-dependent genetic defects in *Arabidopsis* (Sureshkumar et al. 2009). Interestingly, TAIR Blast (http://arabidopsis.org/Blast/index.jsp) revealed that this stretch of long inverted repeats (CT and GA) is not found anywhere in the Col-0 genome, hence not represented in the Ler-0 draft sequence either. In total, 60 (98.4%) out of 61 misassembled sites were resolved to generate a revised Ler-0 sequence of 375,893 bp in length. The largest insertion and deletion resolved by iterative read mapping were 4,819 and 5,139 bp, respectively. By accounting for the size of the six unamplified gaps, the Ler-0 locus was estimated to be considerably larger than its Col-0 counterpart.

To evaluate the accuracy of the Ler-0 revised sequence, locus-specific reads were mapped to all three sequences (Col-0, Ler-0 draft, and Ler-0 revised) using the most stringent parameters (no gaps and no ambiguities were allowed). Because only reads that have no errors were included, the mean coverage decreased from ~930-fold (when one error is allowed) to ~643-fold (only perfect reads allowed). The same process was repeated using Ler-0 whole-genome reads from WTCHG. In comparison to the Ler-0 draft, the Ler-0 revised sequence is a better reference (table 2). From the stringent alignment of locus-specific reads to the Ler-0 revised sequence, seven gaps were identified (six gaps corresponding to unamplified regions and one gap to the aforementioned unresolved region). Similarly, seven gaps were present when Ler-0 whole-genome reads were aligned to the revised sequence. However, size differences were observed in the gaps when either locus-specific reads or whole-genome reads were used. This is due to the fact that PCR primers were designed to amplify regions that are spanned by the forward and reverse primers. However, variations in Ler-0 do not always start and end at the primer-binding sites. The absence of misassembly signatures overall demonstrates that the Ler-0 revised sequence is superior to the draft sequence. Moreover, Sanger sequencing on 16 random corrections subsequently confirmed that all were

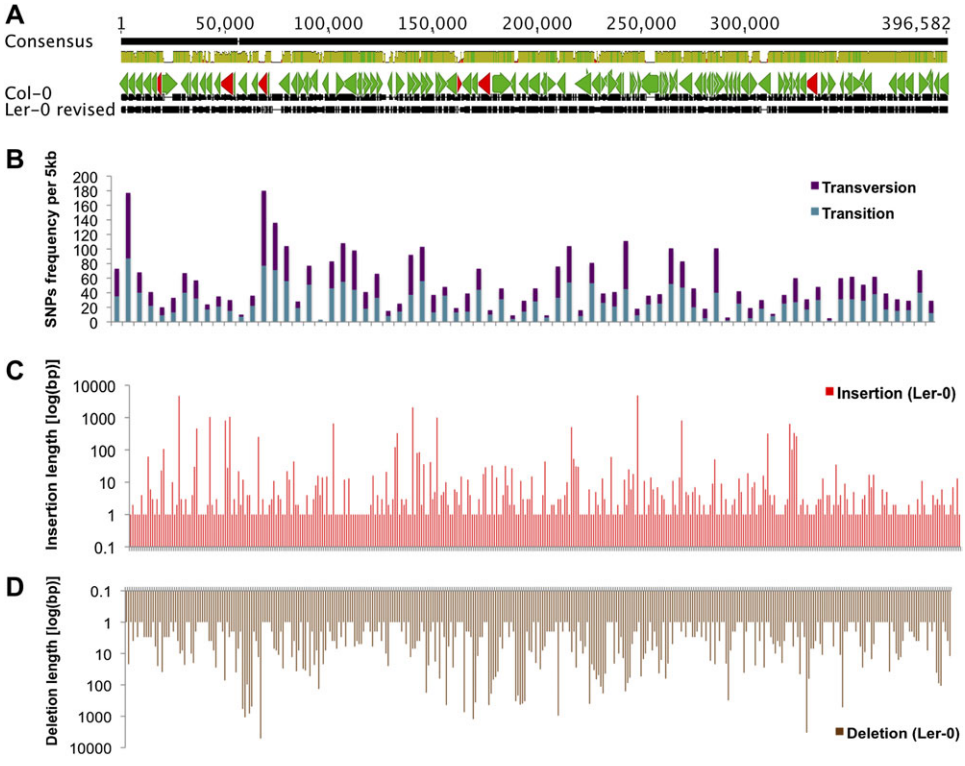

FIG. 4.—Schematic diagram of polymorphisms on the selected Ler-0 locus. Variations between Ler-0 and Col-0 are indicated on the diagram. (*A*) Figure illustrates the pairwise alignment between Ler-0 and Col-0. TAIR10 annotated genes (green arrows) and transposable element genes (red arrows) are indicated, respectively. Detailed representations of the variations between Ler-0 and Col-0, (*B*) SNPs, (*C*) insertions, and (*D*) deletions, are indicated in the diagram.

Table 3
Large Indels that Overlap Genes and Regulatory Regions

Figure ID	Col-0 Gene	Gene Description[a]
Figure 5*A*	(TAIR:At3G45500)	RING/U-box protein with C6HC-type zinc finger
Figure 5*B*	(TAIR:At3G45490)	RING/U-box superfamily protein
Figure 5*C*	(TAIR:At3G45840)	Protein binding/zinc ion binding
Figure 5*D*	(TAIR:At3G45955)	tRNA-Val
Figure 5*E*	(TAIR:At3G46110)	Unknown protein
Figure 6*A*	(TAIR:At3G45990)	Cofilin/tropomyosin-type actin-binding protein
Supplementary figure 4*A* (Supplementary Material online)	(TAIR:At3G46060)	Small GTP-binding protein
Supplementary figure 4*B* (Supplementary Material online)	(TAIR:At3G45910)	Unknown protein
Supplementary figure 4*C* (Supplementary Material online)	(TAIR:At3G45540)	RING/U-box protein with C6HC-type zinc finger
	(TAIR:At3G45550)	Non-LTR retrotransposon family (LINE)
	(TAIR:At3G45555)	Zinc finger (C3HC4-type RING finger) family protein
Supplementary figure 4*D* (Supplementary Material online)	(TAIR:At3G45750)	Nucleotidyltransferase family protein
	(TAIR:At3G45755)	Transposable element gene
	(TAIR:At3G45760)	Nucleotidyltransferase family protein
Supplementary figure 4*E* (Supplementary Material online)	(TAIR:At3G45673)	Unknown protein

NOTE.—LTR, long terminal repeat.
[a]Information obtained from the TAIR10 genome annotation

accurate (supplementary table 2B, Supplementary Material online).

To investigate the feasibility of our method for low-coverage whole-genome data sets, targeted iterative read assembly was performed on a random unmapped contig obtained from WTCHG's L*er*-0 N50 de novo contigs. Using the L*er*-0 whole-genome reads from WTCHG that has a modest coverage of 40-fold, a selected 478-bp contig was iteratively extended to a 2,091-bp sequence. This sequence was validated by Sanger sequencing (Lai AG, Dijkwel PP, unpublished data) and does not align to any region of the L*er*-0 draft, suggesting that it is present within a haplotype-specific insertion in L*er*-0. In an attempt to fill in the six unamplified gaps in the locus of interest, iterative read mapping was done using WTCHG L*er*-0 whole-genome reads as well as the de novo contigs. However, we were mostly unsuccessful for several reasons. The relatively low-coverage data set along with the lack of read positional information did not allow accurate iterative mapping particularly when the region is duplicated or is highly repetitive. Furthermore, because of the lack of positional information, the correct de novo contig that aligns to the border of the gap could not be selected when there is more than one possible match.

The previously predicted SVs were resolved by iterative read mapping using a high-coverage data set aided by PCR-based positional information. In total, 31 large indels (>100 bp), 52 smaller (<100 bp) indel-like misassemblies, and 722 novel SNPs that were not present in the L*er*-0 draft sequence were identified. On average, one SNP per 97 bp (10 SNPs/kb) and one indel per 507 bp (2 indels/kb) were detected between L*er*-0 and Col-0. Novel variations identified from this study were not represented in the L*er*-0 draft presumably because they occurred in duplicated or highly conserved regions where these regions can hamper accurate variant calling. Alignment between the L*er*-0 and Col-0 loci yielded a pairwise identity of 84.2%. In addition, we provide a snapshot of variations between L*er*-0 and Col-0 at this small genomic interval (fig. 4, supplementary fig. 3 and table 1, Supplementary Material online).

Biological and Evolutionary Significance of SVs

We next determined whether the SVs could have effects on genes. According to TAIR10 annotation, the 371-kb locus on chromosome 3 comprises 102 genes, 3 transfer RNA genes, and 6 transposable element genes (fig. 4*A*). Fifteen novel large indels were found to be present within genes and regulatory regions (table 3). These large indels are grouped into three categories: 1) SVs that alter predicted open reading frames, 2) SVs located in regulatory regions, and 3) SVs affecting clusters of genes with similar functions (fig. 5; supplementary fig. 4, Supplementary Material online).

In the first category, SVs were found to either disrupt genes or, as observed in several cases, predicted to produce new transcripts. Figure 5*A* depicts a copia-like retrotransposon insertion within the second intron of (TAIR:At3G45500), whereas in figure 5*B*, a transposon insertion before the first exon of (TAIR:At3G45490) is illustrated. In another example, an 812-bp deletion was identified in a 3.4-kb Col-0 cofilin/tropomyosin-type actin-binding gene (TAIR:At3G45990) (fig. 6*A*). Interestingly, TAIR10 Gbrowse (http://gbrowse.arabidopsis.org/cgi-bin/gbrowse/arabidopsis/) revealed that no expressed sequence tag was found for this gene. A gene prediction program (Stanke and Morgenstern 2005) predicted a 1.1-kb gene from the revised L*er*-0 sequence that was subsequently validated by PCR amplification and Sanger sequencing. TAIR BLASTP results of the putative L*er*-0 allele suggest that it is an *ACTIN-DEPOLYMERIZING FACTOR 4*–like gene (fig. 6*B*). In addition, insertions within

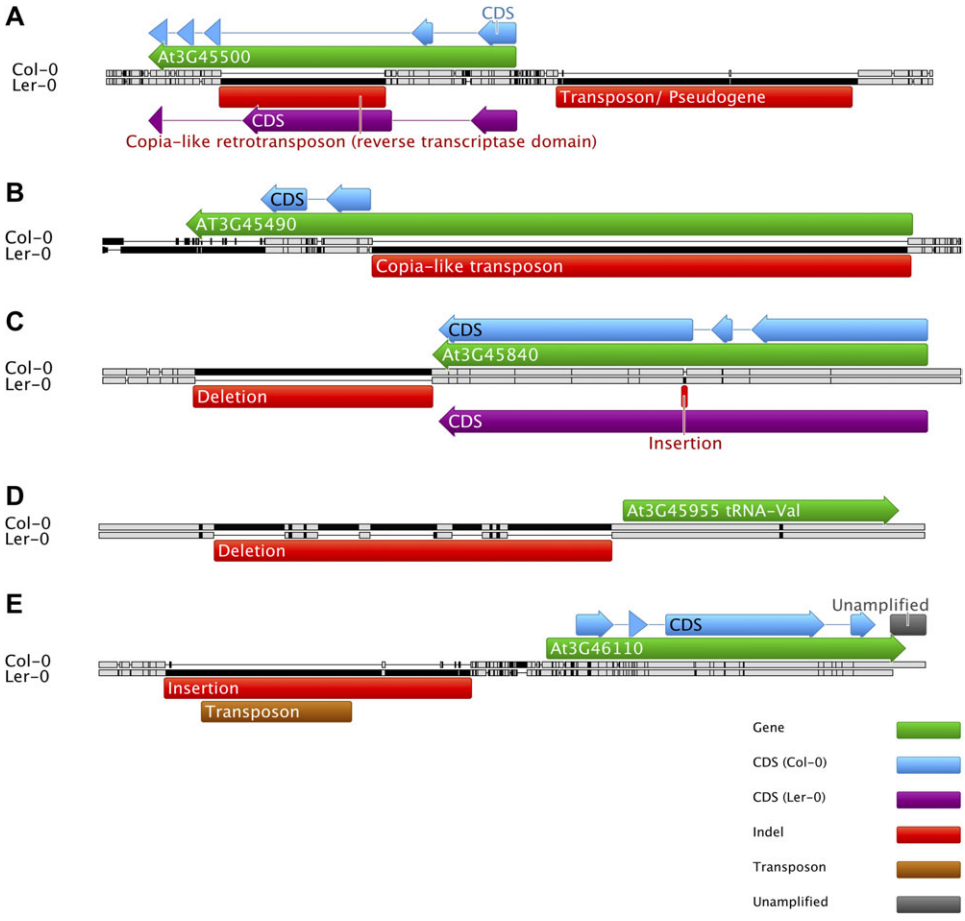

Fig. 5.—SVs that overlap genes. (*A* and *B*) depict copia-like retrotransposon sequences inserted in the corresponding Ler-0 allele. (*C*) A 10-bp insertion within a gene resulted in an inferred intronless transcript variant. (*D*) illustrates a deletion and (*E*) a transposon-like insertion in regulatory regions. Augustus program (Stanke and Morgenstern 2005) is used to predict coding sequences (CDS) of the Ler-0 alleles.

genes can also lead to the formation of inferred new transcripts, for example, an intronless variant (fig. 5*C*). A further noteworthy observation is the insertion in the third intron of At3G46060 (supplementary fig. 4*A*, Supplementary Material online) encoding a GTP-binding protein involved in ethylene signaling (Zimmerli et al. 2008).

SVs occurring in regulatory regions can influence gene expression through numerous positional effects (Feuk et al. 2006). Deletion of regulatory elements (fig. 5*D*) or insertion within such elements (fig. 5*E*) might affect expression of the immediate downstream gene and also the successive gene if both genes share the same *cis*-regulatory

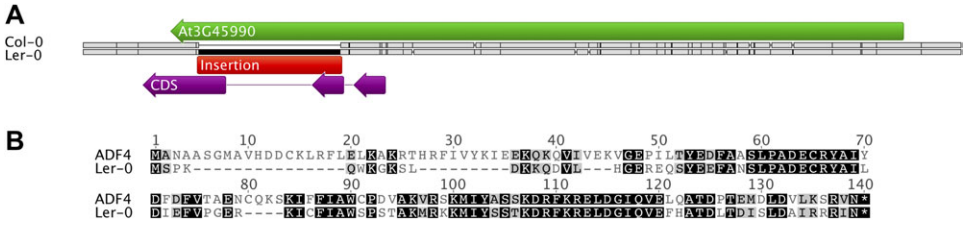

Fig. 6.—Large deletion within a putative Col-0 gene suggests the formation of a novel allelic variant in Ler-0. (*A*) An 812-bp Col-0 deletion is found to be located within a cofilin/tropomyosin-type actin-binding gene (TAIR:At3G45990). Augustus program predicted a 1.1-kb gene model from the Ler-0 allele, which differs from the 3.4-kb Col-0 gene. (*B*) BLASTP revealed that the protein encoded by the Ler-0 allele has 45% pairwise amino acid identity to known ACTIN-DEPOLYMERIZING FACTOR 4 (ADF4) protein encoded by (TAIR:At5G59890). Identical amino acid motifs are highlighted in black and similar motifs in gray.

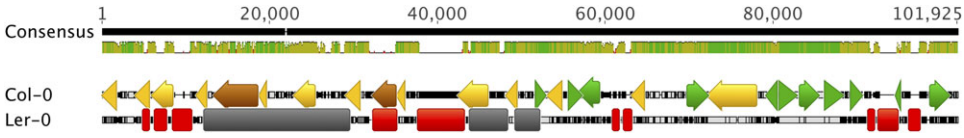

Fig. 7.—Zinc-binding protein gene cluster is enriched in SVs. Diagram illustrates the presence of large indels (red blocks) in a region where a cluster of zinc-binding protein genes (yellow arrows) is located. Gray blocks, brown arrows, and green arrows indicate unamplified regions, transposable element genes, and other protein-coding genes, respectively.

elements (Cordaux and Batzer 2009). In the third category, a high number of SVs were found in a region enriched with genes that encode zinc-binding proteins (fig. 7). Col-0 has two transposable element genes within this region, and we hypothesize that additional transposons are present in the unamplified gaps (data not shown).

Our findings confirm that transposable elements do not merely cause genetic perturbations; they participate in gene regulatory networks in ways that SNPs could not achieve (Heard et al. 2010). In this work, fundamental challenges in SV detection were tackled using an LR-PCR–based resequencing approach that yielded valuable read positional information. Genome assemblies could be improved to show SVs if the experiment is planned in a way that incorporates positional information to the reads. This work emphasizes the importance of detecting SVs as they can have significant implications on downstream biological inferences, particularly on the identification and the study of evolutionarily shared allelic variants.

Discussion

Many agree that the real excitement in whole-genome–sequencing experiments only starts when another genome of a closely related individual is sequenced (Ossowski et al. 2008; Hoberman et al. 2009; McKernan et al. 2009). The human 1000 genomes project (Collins et al. 2003) and the *Arabidopsis* 1001 genomes project (Weigel and Mott 2009) are two examples of joint international collaborations to create a catalogue of intraspecific genetic variations. Together with the rapid advancement in NGS technology and the reduction in sequencing costs, there has been a massive proliferation in the number of drafted genomes produced. However, the inability of current assembly programs to address problematic areas has resulted in the generation of many poor quality drafts (Chain et al. 2009). Capturing large genomic SVs has been particularly challenging (Chen et al. 2009; Kidd et al. 2010).

In an attempt to identify problematic regions and find methods for improving draft genomes, we performed a resequencing experiment at a selected genomic interval of the *Arabidopsis* L*er*-0 accession. Fine-scale sequence analysis at this target locus suggests that *A. thaliana* L*er*-0 and Col-0 genomes are highly variable. From our analysis, it was observed that the L*er*-0 draft sequence accurately incorporates Col-0/L*er*-0 polymorphisms if they are short in length and/or located in regions that are not conserved, duplicated, or repetitive. On the contrary, large SVs that lie in conserved, duplicated, or repetitive regions such as variations in gene families and transposon-like indels were not incorporated in the L*er*-0 draft. Nevertheless, those SVs may affect gene integrity and expression. Over 700 indels (supplementary table 1, Supplementary Material online) between L*er*-0 and Col-0 and 15 large indels (figs. 5, 6, and 7; supplementary fig. 4, Supplementary Material online) present in genes and regulatory regions were identified. Seven of these indels involve transposon-like sequences. Although once thought to be "junk" DNA, an increasing number of studies have shed new light on the functional role of these jumping genes (Lippman et al. 2004; Wheelan et al. 2005). Transposons represent a dynamic portion of genomes, where some can mediate rearrangements of adjacent DNA (Bennetzen 2005), present new regulatory effects on nearby genes (Michaels et al. 2003; Blewitt et al. 2005; Weil and Martienssen 2008; Lisch 2009), and contribute to gene expression divergence between closely related species (Hollister et al. 2011). The presence of transposons could also affect recombination in adjacent genes by heterochromatic effects (He and Dooner 2009).

Our results also suggest the occurrence of a potential synteny break (Al-Shahrour et al. 2010) between L*er*-0 and Col-0 within a zinc-binding protein gene cluster (fig. 7). Ten large indels that include three transposon-like insertions, one transposon deletion in L*er*-0, and three unamplified gaps further imply that this neighborhood has been dynamically reorganized in L*er*-0. Indeed, functional clusters in mammals are significantly enriched by SINE elements as they contribute to the rearrangement process (Zhao et al. 2004). The prevalence of SVs in genic regions can potentially lead to the formation of natural allelic variants or alter gene expression and function altogether. Thus, it is imperative for drafted genomes to incorporate SVs in both coding and noncoding regions so that accurate biological and evolutionary inferences can be drawn from comparative genomics studies on closely related individuals.

Positional Information Allows Correct Assembly of SVs

Whole-genome sequencing is now a routine practice, thanks to the advancements in sequencing technology.

Assembling large and complex genomes is unfortunately a less straightforward task. For example, it is particularly challenging to deduce large insertions in nonreference accessions, variations within conserved or duplicated regions, and variations in microsatellite repeat lengths. Moreover, if the reference accession has a reduced genome (Schmuths et al. 2004), it can significantly impair insertion-based SV detection in nonreference accessions (supplementary fig. 5, Supplementary Material online). Using a combination of wet lab and dry lab approaches, we demonstrated the feasibility in resolving regions that have marked deviations from the reference genome. Amplicon size information was employed to identify the location of large SVs. Once the approximate location was identified, it can be narrowed down to the point where the variation starts by looking for misassembly signatures. Local iterative read mapping was then performed to resolve the variation in question, and the length of the newly deduced sequence was then compared with its respective amplicon size. Algorithms for iterative gap closure have been described elsewhere (Tsai et al. 2010). However, these algorithms detect gaps in assemblies and are not suitable for insertions that do not manifest as assembly gaps (supplementary fig. 2*B*, Supplementary Material online).

Conserved or duplicated regions can affect variant detection, for example, large deletions in conserved regions, transposon-like indels, and polymorphisms within gene families. Santuari and colleagues have recently demonstrated the combined use of tiling array hybridizations with NGS to detect large deletions by identifying regions that have weak hybridization signals along with the absence of short reads (Santuari and Hardtke 2010; Santuari et al. 2010). Here we show that fine-scale manual inspection can resolve regions that are conserved, duplicated, or repetitive. Information contained in a single read is significantly limited by its length and can result in ambiguous placement of reads to homologous regions (Young et al. 2010). Aberrant alignments of homologous reads may inflate the number of false-positive detections (Pool et al. 2010). In particular, we observed ambiguous placements of transposons in the L*er*-0 draft. Although deletions are easier to detect, we have nevertheless identified large deletions absent from the L*er*-0 draft. Deletions that lie in conserved regions will be missed (false negatives) as reads from homologous regions can map to the reference sequence although it is not present in the study accession. Because our work was targeted to a specific locus, regions that are duplicated elsewhere in the genome will not interfere with the iterative mapping step, unless a particular region is duplicated within the locus itself. Therefore, we emphasize the importance of having positional information that assists sorting of reads to their respective locations and allows the resolution of duplications independently without interference from other homologous sequence reads.

Previously, an indel prediction has been performed using the 2-fold redundant L*er*-0 shotgun contigs generated by Cereon Genomics (Ziolkowski et al. 2009). Thirteen out of the 19 predicted indels that fall within the locus of interest were found to be false positives, the largest being a 7.8-kb insertion. The high rate of false-positive predictions can be attributed to the assignment of incorrect chimeric Cereon contigs (Lai AG, Dijkwel PP, unpublished data) that have partial sequence homology to a particular region. The incorrect placement of contigs is therefore exacerbated by the absence of positional information.

Another challenge in whole-genome assembly is the accurate deductions of microsatellite repeat lengths from short-read data sets (supplementary table 1E, Supplementary Material online). In theory, paired-end mapping should mitigate this problem if the gap spanned by the paired reads is larger than the repeat itself. Most paired-end libraries, however, lack sufficient coverage to enable reliable sequence predictions (Schatz et al. 2010). Therefore, positional information is useful for the sorting of repetitive sequences to their respective genomic locations. Furthermore, accurate deduction of repeat length is crucial in order to reveal rare allelic variants (Sureshkumar et al. 2009). In short, significant progress can be made on genome assembly if the experimental design prior to sequencing is modified such that positional information is incorporated into data sets.

Genes Associated with SVs May Evolve New Functions

The organization of SVs has two implications on genome evolution. First, structural changes can be observed in regions that have high rates of evolutionary turnover and second it allows genes that are duplicated or transposed to new chromosomal regions to be free from selective constraints and evolve independently, giving rise to genes with altered functions or altered regulation (Samonte and Eichler 2002). The most common type of SVs that affect genes are segmental duplications where a likely outcome would be the accumulation of partial gene structures or pseudogenes (Lynch and Conery 2000; Zhang 2003). These paralogous genomic copies have been often treated as "dead on arrival." Recent studies on whole-genome tiling analysis, however, revealed that pseudogenes can be expressed (Akama et al. 2009). Expressed pseudogenes also play a role in the regulation of the messenger RNA stability of its homologous coding gene (Hirotsune et al. 2003). Gene density greatly correlates with segmental duplication density and in comparison to unique genes; genes in segmental duplicated regions are more likely to display inter- and intraspecific CNV (Tuzun et al. 2005) along with signatures of positive selection (Johnson et al. 2001; Birtle et al. 2005). Although genes affected by SVs are most likely associated with subtle

phenotypic alterations due to selective constraints, they can nevertheless have an influence on the phenotype by altering gene dosage (Sharp et al. 2006). Genes involved in environmental interaction and host defense have been found to be enriched with SVs (Emes et al. 2003; Tuzun et al. 2005). Examining structurally dynamic regions of the genome may provide clues on lineage-specific adaptation patterns (Emes et al. 2003; Sharp et al. 2006) that are under diversifying positive selection pressure.

Harnessing Positional Information to Boost Comparative Genomics

Our work suggests that positional information is important for obtaining reliable ordering of scaffolds on chromosomes and improving genome assembly to unveil dynamic genome architectures. Likewise, the development of high-resolution physical maps (Lewin et al. 2009) are indispensible to the ordering of contigs in whole-genome alignments and also for the discovery of evolutionary break point regions based on comparative physical maps (Larkin et al. 2009). A comparison between two forms of genome assembly, that is, hierarchical sequencing of large insert clones and whole-genome shotgun sequence assembly (WGSA) of reads, revealed that the WGSA method yields a 20-Mb shorter sequence than the clone-based assembly (Marques-Bonet et al. 2009). Length discrepancy is caused by the failure of many whole-genome shotgun reads to map to a locus containing a highly duplicated and rapidly evolving gene family (Johnson et al. 2006). This problem will be further aggravated when significantly shorter NGS reads are used (Marques-Bonet et al. 2009).

A fail proof method that accurately detects SVs is still potentially missing. We envisage genome-sequencing experiments to proceed in a clone-based manner that allows the incorporation of positional information to the generated reads. This technique is comparable to the "first-map, then sequence" strategy that uses a BAC-based scaffolding method (Kuhl et al. 2010), which has been successfully implemented in various sequencing projects (Fujiyama et al. 2002; Larkin et al. 2009; Lewin et al. 2009). Construction of large DNA insert libraries will be useful for genome-sequencing projects. This form of genome partitioning will undoubtedly require more work than generating reduced representation libraries from restriction digestions (Young et al. 2010). Although reduced representation libraries can simplify assembly and potentially yield larger contigs, it lacks the positional information required to tease out duplicated regions.

With the current capacity, an entire genome can be sequenced on a single flow cell by making pools of large insert clones and subsequently multiplexing these pools. These reads will have positional information and can then be assigned to their corresponding genomic intervals where de novo contigs can subsequently be generated from these region-specific reads. Using the combinatorial pooling and multiplexing strategy, tens of thousands of different samples can be analyzed with only several hundred appended barcodes (Erlich et al. 2009). Different levels of multiplexing can also be performed to achieve the desired resolution based on resource availability (Wood et al. 2010). With such positional information available, it is possible to elucidate more complex forms of polymorphisms that include segmental duplications, transversions, and transposition events. The size of each clone can be used to validate the accuracy of the assembled contigs. Indeed, several groups have started to follow the clone-based sequencing approach at a low-resolution scale in order to capture a more representative depiction of large intraspecific variations (Kidd et al. 2008; Hurwitz et al. 2010).

NGS platforms have been widely used in targeted resequencing experiments on selected genomic intervals (Martinez Barrio et al. 2009; Turner et al. 2010). Resequencing experiments are often required for the study of intraspecific polymorphisms in regions suspected to host vast amounts of variations. Discovering beneficial or heterotic genetic traits in crop species is primarily performed using a quantitative trait loci (QTL) mapping strategy. Because reference genomes per se may not contain the locus of interest, our approach can successfully identify such SV-related QTL. A majority of drafted genomes fail to provide sufficient granularity for comparative genomics in this sense. Hence, most studies on haplotypic variants still rely on clone-based Sanger shotgun sequencing (Alcázar et al. 2009; Heuer et al. 2009). An undistorted view of data quality is important for end users; hence, each genome should be independently assembled to reveal haplotypic differences. Furthermore, the accuracy of downstream gene annotations relies on the fidelity of the initial assembly. The resulting annotations will not only mislead end users but also defy the initial justification of comparative genomics. Elucidation of complex and dynamic regions of the genome should be the end goal of NGS projects apart from cataloguing small variations such as SNPs. The full benefit of comparative genomics can only be realized when high-quality genome sequences are available.

Supplementary Material

Supplementary figures 1–5 and tables 1 and 2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We are grateful to Professor Richard Mott of the WTCHG for granting access to the prereleased L*er*-0 draft and sequence, whose work was supported by the Biotechnology and Biological Sciences Research Council (BB/F022697/1). We thank Massey Genome Service staff, Lorraine Berry for performing the Illumina sample preparation and sequencing run, Maurice

Collins for running the Illumina pipeline analysis, and Dr Murray Cox for critically reviewing the manuscript. This work was supported by the Massey University Research Fund to P.P.D., Institute of Molecular Biosciences PhD studentship to A.G.L., and the Federal Ministry of Education and Research for Research Units for Systems Biology (BMBF FORSYS) Systems Biology Research Initiative Funding to B.M.-R. (FKZ 0313924).

Literature Cited

Akama T, et al. 2009. Whole-genome tiling array analysis of *Mycobacterium leprae* RNA reveals high expression of pseudogenes and noncoding regions. J Bacteriol. 191:3321–3327.

Al-Shahrour F, et al. 2010. Selection upon genome architecture: conservation of functional neighborhoods with changing genes. PLoS Comput Biol. 6:e1000953. doi:1000910.1001371/journal.pcbi.1000953

Alcázar R, Garcia AV, Parker JE, Reymond M. 2009. Incremental steps toward incompatibility revealed by *Arabidopsis* epistatic interactions modulating salicylic acid pathway activation. Proc Natl Acad Sci U S A. 106:334–339.

Alcázar R, et al. 2010. Natural variation at Strubbelig Receptor Kinase 3 drives immune-triggered incompatibilities between *Arabidopsis thaliana* accessions. Nat Genet. 42:1135–1139.

Alonso-Blanco C, et al. 2009. What has natural variation taught us about plant development, physiology, and adaptation? Plant Cell. 21:1877–1896.

Altshuler DL, et al. 2010. A map of human genome variation from population-scale sequencing. Nature. 467:1061–1073.

Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. 2001. Segmental duplications: organization and impact within the current human genome project assembly. Genome Res. 11:1005–1017.

Bennetzen JL. 2005. Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev. 15:621–627.

Bentley DR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 456:53–59.

Bikard D, et al. 2009. Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. Science. 323:623–626.

Birtle Z, Goodstadt L, Ponting C. 2005. Duplication and positive selection among hominin-specific *PRAME* genes. BMC Genomics. 6:120.

Blewitt ME, et al. 2005. An *N*-ethyl-*N*-nitrosourea screen for genes involved in variegation in the mouse. Proc Natl Acad Sci U S A. 102:7629–7634.

Bowman JL, Floyd SK, Sakakibara K. 2007. Green genes—comparative genomics of the green branch of life. Cell. 129:229–234.

Butler J, et al. 2008. ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res. 18:810–820.

Cai JJ, Macpherson JM, Sella G, Petrov DA. 2009. Pervasive hitchhiking at coding and regulatory sites in humans. PLoS Genet. 5:e1000336. doi:1000310.1001371/journal.pgen.1000336

Chain PSG, et al. 2009. Genome project standards in a new era of sequencing. Science. 326:236–237.

Chen K, et al. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 6:677–681.

Choi JH, et al. 2008. A machine-learning approach to combined evidence validation of genome assemblies. Bioinformatics. 24:744–750.

Collins FS, Morgan M, Patrinos A. 2003. The human genome project: lessons from large-scale biology. Science. 300:286–290.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. Nat Rev Genet. 10:691–703.

Darling A, Mau B, Perna N. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 5:e11147. doi:11110.11371/journal.pone.0011147

Drummond A, et al. 2010. Geneious v5.1. [cited 2011 Feb]. Available from: http://www.geneious.com/.

Ellegren H. 2008. Comparative genomics and the study of evolution by natural selection. Mol Ecol. 17:4586–4596.

Emes RD, Goodstadt L, Winter EE, Ponting CP. 2003. Comparison of the genomes of human and mouse lays the foundation of genome zoology. Hum Mol Genet. 12:701–709.

Erlich Y, et al. 2009. DNA Sudoku-harnessing high-throughput sequencing for multiplexed specimen analysis. Genome Res. 19:1243–1253.

Feuillet C, Keller B. 2002. Comparative genomics in the grass family: molecular characterization of grass genome structure and evolution. Ann Bot. 89:3–10.

Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. Nat Rev Genet. 7:85–97.

Fransz PF, et al. 2000. Integrated cytogenetic map of chromosome arm 4S of *A. thaliana*: structural organization of heterochromatic knob and centromere region. Cell. 100:367–376.

Frazer KA, Murray SS, Schork NJ, Topol EJ. 2009. Human genetic variation and its contribution to complex traits. Nat Rev Genet. 10:241–251.

Fujiyama A, et al. 2002. Construction and analysis of a human-chimpanzee comparative clone map. Science. 295:131–134.

Gresham D, Dunham MJ, Botstein D. 2008. Comparing whole genomes using DNA microarrays. Nat Rev Genet. 9:291–302.

He L, Dooner HK. 2009. Haplotype structure strongly affects recombination in a maize genetic interval polymorphic for Helitron and retrotransposon insertions. Proc Natl Acad Sci U S A. 106:8410–8416.

Heard E, et al. 2010. Ten years of genetics and genomics: what have we achieved and where are we heading? Nat Rev Genet. 11:723–733.

Heuer S, et al. 2009. Comparative sequence analyses of the major quantitative trait locus *phosphorus uptake 1 (Pup1)* reveal a complex genetic structure. Plant Biotechnol J. 7:456–471.

Hirotsune S, et al. 2003. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. Nature. 423:91–96.

Hoberman R, et al. 2009. A probabilistic approach for SNP discovery in high-throughput human resequencing data. Genome Res. 19:1542–1552.

Hollister JD, et al. 2011. Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. Proc Natl Acad Sci U S A. 108:2322–2327.

Hurwitz BL, et al. 2010. Rice structural variation: a comparative analysis of structural variation between rice and three of its closest relatives in the genus Oryza. Plant J. 63:990–1003.

Johnson ME, et al. 2001. Positive selection of a gene family during the emergence of humans and African apes. Nature. 413:514–519.

Johnson ME, et al. 2006. Recurrent duplication-driven transposition of DNA during hominoid evolution. Proc Natl Acad Sci U S A. 103:17626–17631.

Kidd JM, et al. 2008. Mapping and sequencing of structural variation from eight human genomes. Nature. 453:56–64.

Kidd JM, et al. 2010. Characterization of missing human genome sequences and copy-number polymorphic insertions. Nat Methods. 7:365–371.

Koonin EV. 2009. Darwinian evolution in the light of genomics. Nucleic Acids Res. 37:1011–1034.

Koornneef M, Alonso-Blanco C, Vreugdenhil D. 2004. Naturally occurring genetic variation in *Arabidopsis thaliana*. Annu Rev Plant Biol. 55:141–172.

Korbel JO, et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. Science. 318:420–426.

Krawitz P, et al. 2010. Microindel detection in short-read sequence data. Bioinformatics 26:722–729.

Kuhl H, et al. 2010. The European sea bass *Dicentrarchus labrax* genome puzzle: comparative BAC-mapping and low coverage shotgun sequencing. BMC Genomics. 11:68.
Larkin DM, et al. 2009. Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. Genome Res. 19:770–777.
Lewin HA, Larkin DM, Pontius J, O'Brien SJ. 2009. Every genome sequence needs a good map. Genome Res. 19:1925–1928.
Lippman Z, et al. 2004. Role of transposable elements in heterochromatin and epigenetic control. Nature. 430:471–476.
Lisch D. 2009. Epigenetic regulation of transposable elements in plants. Annu Rev Plant Biol. 60:43–66.
Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science. 290:1151.
Marques-Bonet T, Ryder OA, Eichler EE. 2009. Sequencing primate genomes: what have we learned? Annu Rev Genomics Hum Genet. 10:355–386.
Martinez Barrio A, et al. 2009. Targeted resequencing and analysis of the diamond-blackfan anemia disease locus RPS19. PLoS One. 4:e6172. doi:6110.1371/journal.pone.0006172
McKernan KJ, et al. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res. 19:1527–1541.
Meader S, Hillier LW, Locke D, Ponting CP, Lunter G. 2010. Genome assembly quality: assessment and improvement using the neutral indel model. Genome Res. 20:675–684.
Meyers BC, Kaushik S, Nandety RS. 2005. Evolving disease resistance genes. Curr Opin Plant Biol. 8:129–134.
Michaels SD, He Y, Scortecci KC, Amasino RM. 2003. Attenuation of FLOWERING LOCUS C activity as a mechanism for the evolution of summer-annual flowering behavior in *Arabidopsis*. Proc Natl Acad Sci U S A. 100:10102–10107.
Nishiyama T, et al. 2003. Comparative genomics of *Physcomitrella patens* gametophytic transcriptome and *Arabidopsis thaliana*: implication for land plant evolution. Proc Natl Acad Sci U S A. 100:8007–8012.
Ossowski S, et al. 2008. Sequencing of natural strains of *Arabidopsis thaliana* with short reads. Genome Res. 18:2024–2033.
Perry GH, et al. 2008. Copy number variation and evolution in humans and chimpanzees. Genome Res. 18:1698–1710.
Pool JE, Hellmann I, Jensen JD, Nielsen R. 2010. Population genetic inference from genomic sequence variation. Genome Res. 20:291–300.
Rieseberg LH, Whitton J, Gardner K. 1999. Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. Genetics. 152:713–727.
Rozen S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol. 132:365–386.
Samonte RV, Eichler EE. 2002. Segmental duplications and the evolution of the primate genome. Nat Rev Genet. 3:65–72.
Santuari L, Hardtke CS. 2010. The case for resequencing studies of *Arabidopsis thaliana* accessions: mining the dark matter of natural genetic variation. F1000 Biol Rep. 2:85. doi: 10.3410/B2-85.
Santuari L, et al. 2010. Substantial deletion overlap among divergent *Arabidopsis* genomes revealed by intersection of short reads and tiling arrays. Genome Biol. 11:R4.
Schatz MC, Delcher AL, Salzberg SL. 2010. Assembly of large genomes using second-generation sequencing. Genome Res. 20:1165–1173.
Schmuths H, Meister A, Horres R, Bachmann K. 2004. Genome size variation among accessions of *Arabidopsis thaliana*. Ann Bot. 93:317–321.
Sebat J, et al. 2004. Large-scale copy number polymorphism in the human genome. Science. 305:525–528.
Sharp AJ, Cheng Z, Eichler EE. 2006. Structural variation of the human genome. Annu Rev Genomics Hum Genet. 7:407–442.
Shen YF, et al. 2010. A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Res. 20:273–280.
Simpson JT, et al. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19:1117–1123.
Snyder M, Du J, Gerstein M. 2010. Personal genome sequencing: current approaches and challenges. Genes Dev. 24:423–431.
Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 33:W465–W467.
Sureshkumar S, et al. 2009. A genetic defect caused by a triplet repeat expansion in *Arabidopsis thaliana*. Science. 323:1060–1063.
Tsai IJ, Otto TD, Berriman M. 2010. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biol. 11:R41. doi:10.1186/gb-2010-1111-1184-r1141
Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV. 2010. Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. Nat Genet. 42:260–263.
Tuzun E, et al. 2005. Fine-scale structural variation of the human genome. Nat Genet. 37:727–732.
van der Biezen EA, Brandwagt BF, van Leeuwen W, Nijkamp HJJ, Hille J. 1996. Identification and isolation of the *FEEBLY* gene from tomato by transposon tagging. Mol Gen Genet. 251:267–280.
Warren RL, Sutton GG, Jones SJM, Holt RA. 2007. Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23:500–501.
Weigel D, Mott R. 2009. The 1001 Genomes Project for *Arabidopsis thaliana*. Genome Biol. 10:107. doi:110.1186/gb-2009-1110-1185-1107
Weil C, Martienssen R. 2008. Epigenetic interactions between transposons and genes: lessons from plants. Curr Opin Genet Dev. 18:188–192.
Wheelan SJ, Aizawa Y, Han JS, Boeke JD. 2005. Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. Genome Res. 15:1073–1078.
White MJD. 1978. Chain processes in chromosomal speciation. Syst Biol. 27:285–298.
Wood HM, et al. 2010. Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. Nucleic Acids Res. 38:e151. doi:110.1093/nar/gkq1510.
Young AL, et al. 2010. A new strategy for genome assembly using short sequence reads and reduced representation libraries. Genome Res. 20:249–256.
Zeller G, et al. 2008. Detecting polymorphic regions in *Arabidopsis thaliana* with resequencing microarrays. Genome Res. 18:918–929.
Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821–829.
Zhang F, Gu W, Hurles ME, Lupski JR. 2009. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet. 10:451–481.
Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol Evol. 18:292–298.
Zhao S, et al. 2004. Human, mouse, and rat genome large-scale rearrangements: stability versus speciation. Genome Res. 14:1851–1860.
Zimmerli L, et al. 2008. The xenobiotic B-aminobutyric acid enhances *Arabidopsis* thermotolerance. Plant J. 53:144–156.
Ziolkowski PA, Koczyk G, Galganski L, Sadowski J. 2009. Genome sequence comparison of Col and L*er* lines reveals the dynamic nature of *Arabidopsis* chromosomes. Nucleic Acids Res. 37:3189–3201.

Associate editor: Michael Purugganan