Detection of large sequence insertions by a hybrid approach that combine *de novo* assembly and resequencing of medium-coverage genome sequences

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>gen-2018-0027.R1</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Article</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>18-May-2018</td>
</tr>
</tbody>
</table>
| Complete List of Authors: | Lee, Taeyoung; Seoul National University, Department of Plant Science and Research Institute of Agriculture and Life Sciences  
Kim, Moon Young; Seoul National University, Department of Plant Science and Research Institute of Agriculture and Life Sciences; Seoul National University, Plant Genomics and Breeding Institute  
Ha, Jungmin; Seoul National University, Department of Plant Science and Research Institute of Agriculture and Life Sciences; Seoul National University, Plant Genomics and Breeding Institute  
Lee, Suk-Ha; Seoul National University, Department of Plant Science and Research Institute of Agriculture and Life Sciences; Seoul National University, Plant Genomics and Breeding Institute |
| Keyword:         | large sequence insertion, orphan contig, *de novo* assembly, medium-coverage sequencing, soybean |
| Is the invited manuscript for consideration in a Special Issue?: | Not applicable (regular submission) |
Detection of large sequence insertions by a hybrid approach that combine de novo assembly and resequencing of medium-coverage genome sequences

Taeyoung Lee¹, Moon Young Kim¹, ², Jungmin Ha¹, ², Suk-Ha Lee¹, ²,*

¹ Department of Plant Science and Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul 08826, Republic of Korea
² Plant Genomics and Breeding Institute, Seoul National University, Seoul 08826, Republic of Korea

Address: Crop Genomics Lab. Rm. 4105 Bldg. 200, Department of Plant Science, College of Agriculture and Life Science, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

*Correspondence: Suk-Ha Lee
E-mail address: sukhalee@snu.ac.kr
Tel: +8228804545
Fax: +8228774550
Abstract

Large sequence insertion (LSI) is one of the structural variations (SVs) that may cause phenotypic differences in plants. To identify the LSIs using medium-coverage sequencing data of four wild soybean (*Glycine soja*) genotypes, we designed a hybrid approach combining *de novo* assembly and read mapping. Total reads and reads with both ends unmapped were independently assembled into "ordinary contigs" and "orphan contigs", respectively, and subjected to pairwise alignment and stringent filtering. This approach predicted 24 LSIs averaging 2,682 bp in size, with no overlap with SVs detected by Pindel, BreakDancer, or ScanIndel, and they were validated by PCR. Compared with the soybean (*Glycine max*) reference genome, 20 LSIs were located outside genic regions. One of the four LSIs within a genic region, LSI05, is located in the coding DNA sequence region of a protein kinase superfamily gene (Glyma.08G123500). It caused delayed translation initiation and loss of 24 amino acids in the wild soybean genotype CW12. LSI05 was more frequently observed in 29 *G. soja* accessions than in 34 *G. max* accessions. Identified LSIs would be genomic resources harboring novel gene contents for studying SVs and improving crops. Moreover, our cost-efficient approach may be applicable to other plant species.

Keywords: large sequence insertion, orphan contig, *de novo* assembly, medium-coverage sequencing, soybean, domestication bottleneck
1 **Introduction**

2 Structural variations (SVs) caused by genomic rearrangements (>50 bp) are considered to play important roles in evolution, adaptation, and phenotypic variation (Saxena et al. 2014). Types of SVs include deletions, novel sequence insertions, inversions, translocations, tandem repeats, mobile element transpositions, and copy number variations (Tattini et al. 2015). Among human genomes, single-nucleotide polymorphisms (SNPs) represent only 0.1% of total variation, whereas SVs account for 1.2% (Pang et al. 2010).

3 Historically, SV detection was primarily performed using array comparative genome hybridization and SNP microarray, but array platforms have a limited ability to detect novel sequence insertions in target samples. Consequently, next-generation sequencing (NGS) has largely replaced hybridization-based technologies (Henrichsen et al. 2009; Quinlan et al. 2010). In recent efforts to detect SVs using NGS, four strategies are commonly applied: read-pair, read-depth, split-read, and *de novo* assembly strategies (Alkan et al. 2011; Tattini et al. 2015). Read-pair, split-read, and *de novo* assembly can theoretically detect all types of SV, but each has a different bias depending on insert size and read length. On the other hand, the read-depth strategy can only detect deletion and duplication (Tattini et al. 2015). A dozen computational tools based on read-pair, read-depth, and split-read strategies have been reported, including BreakDancer (Fan et al. 2014) and Pindel (Ye et al. 2009). However, the *de novo* assembly approach has the highest resolution and therefore the greatest power to detect all SV types (Saxena et al. 2014).

4 The well-characterized soybean (*Glycine max* [L.]) reference genome has made it
possible to identify SVs among \textit{G. max} genotypes (Cook et al. 2012; McHale et al. 2012). Meanwhile, several efforts to resequence the whole genome of wild soybean (\textit{Glycine soja} Sieb. and Zucc.) have been conducted to identify SVs relative to \textit{G. max} (Kim et al. 2010; Lam et al. 2010). Among the types of SV, novel sequence insertions represent promising sources of novel genes involved in traits such as environmental tolerance and disease resistance. However, resequencing cannot detect novel sequence insertions because they are not mapped to the reference genome. Moreover, high-quality \textit{de novo} genome assembly for prediction of SVs, including novel sequence insertions, still requires high-depth sequencing coverage (about 100-fold) (Li et al. 2014; Qi et al. 2014). Third-generation sequencers such as the PacBio platform can produce read sequences long enough to detect SVs, but these methods remain expensive (Rhoads and Au 2015). Therefore, development of economic and efficient approaches is essential for identification of novel sequences among diverse soybean genotypes.

This study presents a new pipeline to identify large sequence insertions (LSIs) >500 bp with a relatively small amount of sequence data (~20x), compared with the \textit{G. max} reference (Williams 82). We used wild soybean (\textit{G. soja}) as a target sample for LSI identification due to its higher genetic distance from the reference genome than that of cultivated soybean (Kim et al. 2010; Li et al. 2014). To this end, we developed a hybrid approach combining \textit{de novo} assembly with resequencing. This hybrid approach consists of two independent assemblies, yielding “ordinary contigs” derived from total paired-end reads and “orphan contigs” assembled only from reads with both ends unmapped (orphan reads). In addition, we used three different SV prediction programs to predict insertion events and to evaluate our pipeline. We validated the novel LSI
candidates by PCR, and determined the frequency of one of the validated LSIs, LSI05, in other soybean accessions.

Materials and Methods

Plant materials and DNA isolation

Four wild soybean genotypes, CW12 (PI464934), CW14 (PI407307), KW4 (IT183024), and KW5 (IT183030), were subjected to genome sequencing followed by LSI identification. CW12 and CW14, which originate from China, were obtained from the USDA Germplasm Resources Information Network (https://www.ars-grin.gov/). KW4 and KW5, collected in Korea, were obtained from the National Agrodiversity Center of Rural Development Administration, Republic of Korea (http://genebank.rda.go.kr). Williams 82 (G. max, accession no. PI 518671) was used as a null control for PCR validation of the identified LSIs. To evaluate the frequency of LSIs identified in this study, soybean germplasms comprising of 29 G. soja and 34 G. max accessions were used (Table S1). Genomic DNA was isolated from healthy young leaves of each soybean genotype, as previously described (Shure et al. 1983). The concentration of each DNA sample was determined on an ND-3000 spectrophotometer (NanoDrop Technologies Inc., Wilmington, DE, USA) and diluted to the working concentration (50 ng/μl) with Tris-EDTA buffer (pH 8.0).

Whole-genome sequencing, de novo assembly, and read alignment

Genomic DNAs of four G. soja genotypes were prepared in paired-end libraries with an insert size of 350 bp, and subjected to whole-genome sequencing on an
Illumina HiSeq 2000 (Illumina, Co., San Diego, CA, USA). De novo assembly of the reads was performed to construct contigs (herein referred to as “ordinary contigs”) using AbySS assembler 1.5.1 (Simpson et al. 2009) with a q-value of 20 and a k-mer option of 51. Meanwhile, to extract reads with both ends unmapped (“orphan reads”) from G. soja genotypes, the paired-end reads from each genotype were aligned against the G. max reference genome (Wm82.a2) (Schmutz et al. 2010), downloaded from Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html), using Burrows-Wheeler Aligner (BWA) v0.7.12 (Li and Durbin 2009) with default options. After the orphan reads of each genotype were retrieved using Samtools v1.3 (Li et al. 2009), they were assembled into “orphan contigs” using the same program as for de novo assembly of ordinary contigs.

Prediction of insertions using Pindel, BreakDancer, and ScanIndel

To compare the LSIs identified by our approach, three SV prediction tools, Pindel (Ye et al. 2009), BreakDancer (Fan et al. 2014), and ScanIndel (Yang et al. 2015), were used with default settings to call insertion variants in the four G. soja genotypes.

Ab initio gene prediction of LSIs and comparisons with protein databases

We predicted genes on large sequences inserted in the G. soja genotypes with the AUGUSTUS 3.3 software, using ab initio methods with default options except for species (Stanke and Morgenstern 2005); the species option was set as Arabidopsis. The proteins predicted by AUGUSTUS were compared with G. max (Gm) reference v2.0 protein database (Wm82.a2) (Schmutz et al. 2010), the TAIR 10 Arabidopsis thaliana (At) protein database (https://www.arabidopsis.org), and the non-redundant

**PCR validation of LSIs**

To validate the LSIs identified in this study, sequence-specific dominant primer sets were designed using in-house Python scripts. Target amplicon size and annealing temperature were set at 300–600 bp and 56 °C, respectively. ePCR was performed using these primer sets to estimate product sizes and multiple primer binding against Wm82.a2 (Rotmistrovsky et al. 2004) (Table S2). For PCR, the targeted regions were amplified from 50 ng of genomic DNA using Taq Plus DNA polymerase (Vivagen, Daejeon, Korea) in a T100™ Thermal Cycler (BIO-RAD, Richmond, CA, USA) under the following conditions: 94 °C for 3 min, followed by 30 cycles of 94 °C for 30 s, 56 °C for 30 s, and 72 °C for 30 s, and a final cycle of 72 °C for 5 min. Co-dominant primer sets were designed for six randomly selected LSIs, and their product sizes were predicted by ePCR (Table S3). Optimal annealing temperatures were determined by gradient PCR ranging from 50 to 65 °C. For these reactions, PCR conditions were as described above, except with elongation times of 1 min per 1 kb of product size.

**Results**

**LSI prediction approach**

Our approach for LSI prediction in soybean uses two independent *de novo* assembled contigs with medium-coverage paired-end read sequence data from four *G. soja* genotypes. The *de novo* assembly using total reads produced general ordinary
contigs (Fig. 1). The other assembly, using reads with both ends unmapped, resulted in orphan contigs, indicating the presence of LSIs in the *G. soja* genotypes (Fig. 1). To improve the reliability of LSI prediction, we applied three stringent filtering steps (a to c in Fig. 1). The first condition is that the full length of the orphan contig should map within an ordinary contig with 100% identity when only orphan contigs of ≥2 kb are BLASTed against ordinary contig data (Fig. 1a). The second condition is that both tails of the ordinary contigs flanking the orphan contig should be aligned on the same chromosome of Wm82.a2 in the same direction (Fig. 1b). For the first and second steps, we used BLASTN with an e-value of 1e⁻¹⁰⁰. The third filtering step was performed to predict the breakpoints of LSI on Wm82.a2 and to remove false positives (Fig. 1c). The breakpoints of the LSI candidates were determined by BLAST of the ordinary contigs against Wm82.a2. By manual inspection of sequence alignments in BAM files, we removed LSI false positives in the unsequenced regions of Wm82.a2 because unsequenced gaps (with multiple ambiguous nucleotides, Ns) on Wm82.a2 could generate LSI artifacts. Additionally, we selected only LSI candidates supported by multiple short reads with low mapping quality on the breakpoints, possibly caused by mismatches with inserted sequence. These three stringent filtering steps enabled prediction of reliable candidate LSIs in the wild soybeans genotypes. The candidates were validated by PCR using dominant (insertion-specific) and co-dominant (common to wild soybean and reference, Williams 82) primer sets (Fig. 1).

**De novo assembly of ordinary and orphan contigs**

To predict LSIs longer than 500 bp using medium-coverage sequencing data, we
generated paired-end read sequences of 18.4–21.4 Gb in four G. soja genotypes, CW12, CW14, KW4, and KW5 (Table S4). Using these total reads, the de novo assembly yielded ordinary contigs covering 460–498 Mbp in CW12, CW14, KW4, and KW5, with an average N50 length of 2,124 bp (Table 1).

For the orphan assembly, an average of 88.7% of G. soja short reads were properly mapped to Wm82.a2, spanning 94.7% of the G. max genome at 19.6-fold mapping depth (Table S4). The number of orphan reads of the four wild soybean genotypes ranged from 3.6 (KW5) to 5.1 (CW14) million with 2.2% of total reads on average; these reads were assembled into 49,345 (KW5) to 63,147 (CW12) orphan contigs (Table 1). The numbers of orphan contigs longer than 2 kb were 781, 992, 507, and 569 in CW12, CW14, KW4, and KW5, respectively (Table 1).

Identification of LSI through three filtering steps

Three hundred and eleven ordinary contigs that contained the full sequences of the orphan contigs passed the first filtering step: 100, 135, 34, and 42 ordinary contigs in CW12, CW14, KW4, and KW5, respectively (Fig. 1a). The second step retrieved 49 candidate insertion regions where both tails of the ordinary contigs were aligned in the same direction to the same chromosome of Wm82.a2 (Fig. 1b). That is to say, if the left end of an ordinary contig was aligned in the forward direction, the right end should be aligned in the forward direction when ordinary contigs were blasted against Wm82.a2. Using our pipeline, we identified 10 LSI candidates from CW12, 16 from CW14, 4 from KW4, and 0 from KW5 (Fig. 1c, Table 2). Four LSIs (04, 06, 07, and 22) were predicted in more than one G. soja genotype (Table 2). Thus, in total we predicted 24 unique
insertion candidates.

The positions of LSIs that could affect gene function or expression level are shown schematically in Fig. 2. Ten candidates were located in 1 kb downstream regions, and nine were in intergenic regions (Table 2). Only one LSI (LSI01) was predicted within 1 kb upstream of a transcription start site (Glyma.18G280400). LSI11 and LSI23 are located in the 3’ UTR of Glyma.16G022500 and 5’ UTR of Glyma.12G23900, respectively, and LSI20 is located in an intron of Glyma.15G246800. Most notably, LSI05, with a length of 1,389 bp, replaces parts of the 5’ UTR, intron, and the first CDS of Glyma.08G123500 which encodes a protein kinase superfamily protein (Table 2, Fig. 2).

**Gene contents affected by LSIs**

We next investigated whether the LSI candidates harbored meaningful gene content. AUGUSTUS predicted five genes on five LSIs (LSI04, LSI06, LSI07, LSI13, and LSI21). Among them, three (LSI04, LSI07, and LSI21) yielded hits in the At, Gm, and NR protein databases (Table 2). A gene predicted on LSI04 matched with a leucine-rich repeat (LRR) family protein (Glyma.18G231200) based on the Gm protein models. A gene predicted on LSI07 was mapped to a pseudouridine synthase family protein (AT3G06950 and Glyma.13G119200). A gene predicted in LSI21 was annotated as a putative ribonuclease H protein (KHN07237.1) based on the NR database. Meanwhile, LSI05 from CW12 was predicted to break the first exon of Glyma.08G123500 (Fig. 2), resulting in amino acid changes. Sequence analysis of the ordinary contig containing LSI05 from CW12 revealed deletions of both 14 amino acids.
from 1st to 14th and 10 amino acids from 29th to 38th of Glyma.08G123500 in CW12. It indicated that the CW12 variant produces a shorter polypeptide than that of Williams 82 (Fig. S1).

**Prediction of insertion events by Pindel, BreakDancer, and ScanIndel**

Three SV prediction tools, Pindel, BreakDancer, and ScanIndel were used to predict insertion events in the four *G. soja* genotypes. The 24 LSIs we identified ranged in length from 851 (CW14) to 6,211 bp (CW12), with an average length of 2682 bp (Table S6, Fig. 3). The three SV prediction tools detected many more insertions than our approach. Totally, Pindel, BreakDancer and ScanIndel predicted 1,364,680, 13,153 and 886,025 insertions, respectively, not including repeated ones among the *G. soja* genotypes (Table S5). The maximum insertion sizes detected by Pindel, BreakDancer, and ScanIndel were 82, 254, and 2,212 bp, respectively (Table S5). Pindel called the largest number of insertions in all genotypes, but predicted only small insertions of <100 bp (Table S5, Fig. 3). BreakDancer called a much smaller number of insertions of around 200 bp in two genotypes, KW4 and KW5, than Pindel and ScanIndel (Table S5). ScanIndel predicted the second largest number of insertions, ranging in length from 1 to 2,212 bp. ScanIndel likely detected insertions across a wider size spectrum than the other two tools, but maximum sizes and average sizes of insertions predicted by ScanIndel are shorter than our LSIs (Fig. 3). Within the sizes of 851 bp (the shortest insertion predicted by our pipeline) to 2,212 bp (the longest insertion predicted by ScanIndel), the number of insertions detected by ScanIndel was 34. Among them, three insertions overlapped with LSI01, LSI04 and LSI12 predicted by our pipeline,
respectively. LSI01 corresponded to the largest insertion of 2,212 bp even if their breakpoints showed the difference of 1 bp (Table S6).

4 **PCR validation of the identified LSIs**

We performed PCR to validate the 24 LSI candidates. To this end, we used dominant primer sets specific for each inserted sequence to amplify genomic DNA from the *G. soja* genotypes and the reference soybean genotype, Williams 82 (Table S2). For all LSIs, specific PCR products were amplified from the corresponding *G. soja* genotypes, but not from Williams 82 (Fig. S2). We then designed co-dominant primer sets targeting the six LSIs (05, 10, 12, 16, 17, and 21) to confirm the presence of the insertions based on differences in PCR product sizes between the *G. soja* genotypes and Williams 82 (Table S3, Fig. 4a, b, and Fig. S3). The LSI05c primers amplified the LSI with the expected size (1,876 bp) from CW12, but it produced a shorter amplicon (496 bp) from Williams 82 (Fig. 4b). Longer DNA fragments corresponding to the other five LSI sizes, 3.0–4.8 kb, were amplified from the *G. soja* genotypes, whereas shorter DNA bands ranging from 0.8 to 2.7 kb were amplified from Williams 82, as expected (Fig. S3).

19 **Distribution of LSI05 in other *G. soja* and *G. max* accessions**

Next, we investigated the distribution of LSI05, present in CW12 and located in a CDS region, in a group of other soybean genotypes comprising of 29 *G. soja* and 34 *G. max* accessions (Fig. 4c). As expected, a higher frequency of LSI05 was observed in *G. soja* than *G. max* accessions. Five *G. soja* accessions (CW6, CW8, CW15, CW16, and
KW11) contained LSI05. Of these five wild soybeans, CW15 and CW16 were homozygous for LSI05, yielding a single band. CW6, CW8, and KW11 produced double DNA bands, including the shorter band produced by Williams 82, indicating that they were heterozygous. In 33 G. max accessions, only one genotype, CC16, had the LSI05 amplicon.

**Discussion**

To identify LSIs using significantly less sequence data, we developed a hybrid approach that combines de novo assembly and resequencing (Fig. 1). In three G. soja genotypes, CW12, CW14, and KW4, 24 reliable LSIs were detected through further stringent three-step filtering to remove false positives (Fig. 1, Table 2). In KW5, no LSI was detected using our pipeline (Fig 1, Table 2). The numbers of orphan reads and orphan contigs from KW5 was fewer than those of CW12, CW14 and KW4 (Table S4, Table 1), and most of KW5 orphan contigs were filtered out at the second filtering step (Fig. 1b), which removed ordinary contigs harboring orphan contigs aligned to the Wm82.a2 in different directions such as inversion. The polymorphisms of LSIs, both presence-absence variations (PAVs) and size differences between the G. soja genotypes and the G. max reference (Williams 82), were validated by PCR (Fig. S2 Fig. S3, Fig. 4).

The identified LSIs were compared with the SVs predicted by three computational tools, Pindel, BreakDancer, and ScanIndel. The insertions shorter than 200 bp called by these programs were highly predominant over larger insertions (Fig. 3), although ScanIndel detected 34 insertions larger than 851 bp (the shortest LSI detected by our
pipeline). In addition, our three LSIs, LSI01, LSI04 and LSI12 only overlapped with three of 34 insertions larger than 851 bp predicted by ScanIndel (Table S6). It indicates higher specificity of our pipeline to the prediction of large insertions over 500 bp than three other SV prediction tools. In Arabidopsis and foxtail millet, large number of SVs were predicted using these resequencing-based computational tools, but most of them were indels shorter than 100 bp (Bai et al. 2013; Santuari et al. 2010). In soybean, de novo assembly strategy has been applied to discover PAVs over 500 bp between G. soja genotypes and Wm82.a2; in this approach, sequencing data with 80–100-fold coverage were analyzed to identify PAVs (Lam et al. 2010; Li et al. 2014). Despite a reduction in the expense of sequencing due to technological advances, high-coverage genome sequencing (>100x) still remains expensive. Furthermore, additional stringent filtering steps are required to remove false positives caused by unsequenced gaps in a reference genome (Lam et al. 2010; Li et al. 2014). Our hybrid approach using medium-coverage genome sequencing successfully discovered LSIs present in the G. soja genotypes, except KW5, ranging from 851 to 6,211 bp (Fig. 3, Table S5), further novel gene contents were identified within the LSIs (Table 2). Even though we filtered out orphan contigs shorter than 2 kb in the pipeline, because the orphan contigs contained not only inserted sequences but also sequences from Wm82.a2 on both their ends, the LSIs shorter than 2kb could be predicted in our pipeline.

The identified LSIs were widely distributed throughout the soybean genome, and were mainly (20 out of 24) located in non-genic regions, e.g., in the 1 kb up- or downstream and intergenic regions (Table 2, Fig. 2). Although LSIs outside genic regions may have weaker effects on gene functions, and thus evade selection pressure,
changes in up- or downstream sequences containing regulatory elements can alter transcription levels (Krimpenfort et al. 1988; Morton et al. 2014). In soybean, an insertion of about 1500 bp in promoter region of GmNFYA (Glyma02g47380) encoding the nuclear transcription factor Y subunit A was reported to influence seed oil content as well as expression level (Lu et al. 2016). Four LSIs positioned in genic regions, including UTRs, introns, or CDSs, were identified in multi-gene families, including the protein kinase and NB-ARC (Nucleotide-Binding adaptor shared by Apaf-1, Resistance proteins, and CED-4) families (Table 2). Insertions may be frequent in multi-gene families because other copies of the gene can mask phenotypic effects. For instance, Multiple SVs, identified in the plant genomes including soybean, localize in gene families involved in responses to biotic or abiotic stresses (Cao et al. 2011; Gordon et al. 2014; McHale et al. 2012).

*In silico* gene prediction revealed that three LSIs, LSI04, 07, and 21, contained the genes encoding a member of the LRR family, a pseudouridine synthase, and a ribonuclease H protein, respectively (Table 2). The LRR domain is prevalent in plant resistance proteins (Dangl and Jones 2001). Pseudouridine synthases are the enzymes responsible for the most abundant post-transcriptional modification of cellular RNAs (Hamma and Ferre-D'Amare 2006), and they are also involved in mRNA splicing, and can induce novel splicing events under stress conditions (Adachi and Yu 2014). Ribonuclease H has cleavage activity during DNA replication and repair, which is necessary to stabilize DNA (Cerritelli and Crouch 2009) (Table 2).

LSI05, which is located within a coding region, had a large effect on a gene (Glyma.08G123500) encoding a member of the protein kinase superfamily (Table 2).
Even though *G. soja* genotype CW12 has 1,380 additional sequences, comparison of amino acid sequences revealed delayed translation initiation and a loss of 24 amino acids in CW12, resulting in a shorter polypeptide (Fig. S1). No significant functional domain or signal peptide was detected by *in silico* analysis of the missing amino acids. Meanwhile, we found that 6 of the 29 *G. soja* accessions and 1 of the 34 *G. max* accessions contained the LSI05 fragment (Fig. 4), indicating that frequency of LSI05 in *G. max* might be diminished, perhaps owing to the domestication bottleneck, as suggested by previous studies (Lam et al. 2010; Li et al. 2014). Further studies of such large variations within genic regions between *G. soja* and *G. max* may provide insight into whether and how they have yielded deleterious or selective outcomes during the evolution of two species and soybean domestication.

To investigate variations between samples of a target and a reference, mapping short resequencing reads to a reference genome requires high levels of sequence similarity. In such an approach, however, valuable information inferred from more divergent genomic regions is lost due to removal of unmapped reads. In this study, using a hybrid approach that exploits medium-coverage sequencing data, we demonstrated that orphan reads also are important resources for the identification of large sequences harboring novel gene content. The identified LSIs would be useful resources for improving crops, as well as for studies of evolution and domestication. In addition, our approach may provide new guidelines for detection of SVs using cost-efficient NGS data in other species.
Data accessibility

The DNA raw sequencing reads of CW12, CW14, KW4 and KW5 were deposited in NCBI SRA (https://www.ncbi.nlm.nih.gov/sra/) under the SRA accession code SRP142397. The pipeline scripts used in this research were deposited in GitHub (https://github.com/alima90/LSI).

Acknowledgements

This work was supported by a grant from the Next Generation BioGreen 21 Program (Code No. PJ01322401), Rural Development Administration, Republic of Korea.

Conflict of interest statement

The authors declare that there is no conflict of interest associated with this publication.
References


Park, D., Ma, J., Kim, W.Y., Kim, B.C., Park, S., Lee, K.A., Kim, D.H., Kim, K.H.,
Shin, J.H., Jang, Y.E., Kim, K.D., Liu, W.X., Chaisan, T., Kang, Y.J., Lee, Y.H.,
2010. Whole-genome sequencing and intensive analysis of the undomesticated

Krimpenfort, P., de Jong, R., Uematsu, Y., Dembic, Z., Ryser, S., von Boehmer, H.,
Steinmetz, M., and Berns, A. 1988. Transcription of T cell receptor beta-chain
genes is controlled by a downstream regulatory element. EMBO J. 7(3): 745-750.

Lam, H.M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F.L., Li, M.W., He, W., Qin, N.,
doi:10.1038/ng.715.

Li, H., and Durbin, R. 2009. Fast and accurate short read alignment with Burrows-

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G.,
Abecasis, G., Durbin, R., and Genome Project Data Processing Subgroup.
2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics,

Li, Y.H., Zhou, G., Ma, J., Jiang, W., Jin, L.G., Zhang, Z., Guo, Y., Zhang, J., Sui, Y.,


Qi, X., Li, M.W., Xie, M., Liu, X., Ni, M., Shao, G., Song, C., Kay-Yuen Yim, A., Tao, Y.,
Wong, F.L., Isobe, S., Wong, C.F., Wong, K.S., Xu, C., Li, C., Wang, Y., Guan,
R., Sun, F., Fan, G., Xiao, Z., Zhou, F., Phang, T.H., Liu, X., Tong, S.W., Chan,
of a novel salt tolerance gene in wild soybean by whole-genome sequencing.
Quinlan, A.R., Clark, R.A., Sokolova, S., Leibowitz, M.L., Zhang, Y., Hurles, M.E., Mell,
J.C., and Hall, I.M. 2010. Genome-wide mapping and assembly of structural
Santuari, L., Pradervand, S., Amiguet-Vecher, A.M., Thomas, J., Dorcey, E.,
deletion overlap among divergent Arabidopsis genomes revealed by
Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L.,
Sakurai, T., Umezawa, T., Bhattacharyya, M.K., Sandhu, D., Valliyodan, B.,
Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell-
Sethuraman, A., Zhang, X.C., Shinozaki, K., Nguyen, H.T., Wing, R.A., Cregan,
P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R.C., and
Shure, M., Wessler, S., and Fedoroff, N. 1983. Molecular identification and isolation of
ABySS: a parallel assembler for short read sequence data. Genome Res. 19(6):
Stanke, M., and Morgenstern, B. 2005. AUGUSTUS: a web server for gene prediction in
eukaryotes that allows user-defined constraints. Nucleic Acids Res. 33(Web
ScanIndel: a hybrid framework for indel detection via gapped alignment, split
015-0251-2.
## Tables

**Table 1.** Summary of statistics of ordinary and orphan contigs generated using total reads and reads with both ends unmapped, respectively, in four *G. soja* genotypes.

| G. soja genotype | Ordinary contigs | | | | | | Orphan contigs | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Total | N50 (bp) | >N50 | Max (bp) | Sum (Mb) | Total | N50 (bp) | >2 kb | Max (bp) | Sum (Mb) |
| CW12 | 6,883,181 | 2,105 | 67,892 | 24,419 | 480.0 | 63,147 | 1,184 | 781 | 14,703 | 5.1 |
| CW14 | 9,218,369 | 2,534 | 58,431 | 31,989 | 497.6 | 61,727 | 1,286 | 992 | 19,172 | 6.2 |
| KW4 | 6,480,316 | 1,902 | 72,280 | 28,528 | 459.8 | 52,614 | 1,075 | 507 | 19,167 | 3.8 |
| KW5 | 8,579,424 | 1,956 | 69,364 | 24,312 | 459.9 | 49,345 | 1,084 | 569 | 14,701 | 4.4 |
**Table 2.** Sizes, locations, and annotations of LSI candidates detected in four *G. soja* genotypes, CW12, CW14, KW4 and KW5, compared with Wm82.a2.

<table>
<thead>
<tr>
<th>Insertion name</th>
<th>Location on Wm82.a2</th>
<th>Expected insertion size (bp)</th>
<th>Type</th>
<th>Nearest gene model and its annotation</th>
<th>Annotation of inserted sequences using At, Gm, and NR protein databases</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSI01</td>
<td>Chr18:56,122,528..</td>
<td>2,212</td>
<td>1 kb upstream of Glyma.18G280400</td>
<td>Glyma.18G280400 (NB-ARC domain-containing disease resistance protein)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>56,122,529</td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>LSI02</td>
<td>Chr19:42,431,909..</td>
<td>2,311</td>
<td>1 kb downstream of Glyma.19G163400</td>
<td>Glyma.19G163400 (unknown)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>42,431,910</td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>LSI03</td>
<td>Chr20:33,621,276..</td>
<td>2,548</td>
<td>intergenic</td>
<td>Glyma.20G092800 (tetraticopeptide repeat [TPR]-like superfamily protein)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>33,621,279</td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>LSI04</td>
<td>Chr09:48,088,244..</td>
<td>1,987 1,867 1,867</td>
<td>intergenic</td>
<td>Glyma.09G263400 (major facilitator superfamily with SPX [SYG1/Pho81/XPR1] domain-containing protein)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>48,088,245</td>
<td></td>
<td></td>
<td></td>
<td>Leucine-rich repeat (LRR) family protein (Glyma.18G231200); hypothetical protein (PHAVU_008G066600g, XP_007139884.1)</td>
</tr>
<tr>
<td>LSI05</td>
<td>Chr08:9,502,835..</td>
<td>1,380</td>
<td>Intron, UTR, CDS of Glyma.08G123500</td>
<td>Glyma.08G123500 (protein kinase superfamily protein)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9,503,279</td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>LSI06</td>
<td>Chr05:8,655,766..</td>
<td>6,211 6,209</td>
<td>intergenic</td>
<td>Glyma.05G074100 (unknown)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8,655,771</td>
<td></td>
<td></td>
<td></td>
<td>No hit</td>
</tr>
<tr>
<td>LSI07</td>
<td>Chr12:6,656,077..</td>
<td>3,267 3,269 3,225</td>
<td>1 kb downstream of Glyma.12G083500</td>
<td>Glyma.12G083500 (Argonaute family protein)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6,656,079</td>
<td></td>
<td></td>
<td></td>
<td>tRNA pseudouridine synthase family protein, putative, expressed (AT3G06950); Pseudouridine synthase family protein (Glyma.13G119200); hypothetical protein (glysoja_006575; KHN21096.1)</td>
</tr>
<tr>
<td>LSI08</td>
<td>Chr06:20,274,770..</td>
<td>2,040</td>
<td>intergenic</td>
<td>Glyma.06G207900 (glycine-rich protein)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20,274,782</td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>LSI09</td>
<td>Chr15:13,282,320..</td>
<td>1,982</td>
<td>1 kb downstream of Glyma.15G158400</td>
<td>Glyma.15G158400 (ribosomal L27e protein family)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>13,282,325</td>
<td></td>
<td></td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>LSI</td>
<td>Chromosome</td>
<td>Start</td>
<td>End</td>
<td>Length</td>
<td>Type</td>
</tr>
<tr>
<td>------</td>
<td>-------------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------</td>
<td>------------------------</td>
</tr>
<tr>
<td>LSI10</td>
<td>Chr18:49,428,493-49,428,511</td>
<td>2,228</td>
<td>-</td>
<td>-</td>
<td>1 kb downstream of Glyma.18G209400</td>
</tr>
<tr>
<td>LSI11</td>
<td>Chr16:2,117,015-2,117,023</td>
<td>-</td>
<td>2,869</td>
<td>-</td>
<td>3' UTR of Glyma.16G022500</td>
</tr>
<tr>
<td>LSI12</td>
<td>Chr11:2,464,611-2,464,630</td>
<td>-</td>
<td>2,196</td>
<td>-</td>
<td>intergenic</td>
</tr>
<tr>
<td>LSI13</td>
<td>Chr20:45,545,707-45,545,714</td>
<td>-</td>
<td>3,406</td>
<td>-</td>
<td>1 kb downstream of Glyma.20G220100</td>
</tr>
<tr>
<td>LSI14</td>
<td>Chr16:3,078,472-3,078,475</td>
<td>-</td>
<td>2,392</td>
<td>-</td>
<td>1 kb downstream of Glyma.16G032500</td>
</tr>
<tr>
<td>LSI15</td>
<td>Chr08:13,242,396-13,242,396</td>
<td>-</td>
<td>2,334</td>
<td>-</td>
<td>1 kb downstream of Glyma.08G168000</td>
</tr>
<tr>
<td>LSI16</td>
<td>Chr15:48,833,791-48,833,859</td>
<td>-</td>
<td>1,897</td>
<td>-</td>
<td>intergenic</td>
</tr>
<tr>
<td>LSI17</td>
<td>Chr15:44,701,940-44,703,102</td>
<td>-</td>
<td>1,669</td>
<td>-</td>
<td>1 kb downstream of Glyma.15G237300</td>
</tr>
<tr>
<td>LSI18</td>
<td>Chr03:45,381,895-45,381,899</td>
<td>-</td>
<td>1,973</td>
<td>-</td>
<td>1 kb downstream of Glyma.03G259700</td>
</tr>
<tr>
<td>LSI19</td>
<td>Chr14:44,559,332-44,559,824</td>
<td>-</td>
<td>2,732</td>
<td>-</td>
<td>1 kb downstream of Glyma.14G182600</td>
</tr>
<tr>
<td>LSI20</td>
<td>Chr15:46,980,382-46,981,481</td>
<td>-</td>
<td>851</td>
<td>-</td>
<td>Intron of Glyma.15G246800</td>
</tr>
<tr>
<td>LSI21</td>
<td>Chr18:274,330-275,368</td>
<td>-</td>
<td>2,192</td>
<td>-</td>
<td>intergenic</td>
</tr>
<tr>
<td>LSI22</td>
<td>Chr04:34,418,473-34,418,477</td>
<td>-</td>
<td>4,431</td>
<td>4,433</td>
<td>intergenic</td>
</tr>
<tr>
<td>LSI23</td>
<td>Chr12:39,854,106-39,854,164</td>
<td>-</td>
<td>2,478</td>
<td>-</td>
<td>5' UTR of Glyma.12G239700</td>
</tr>
<tr>
<td>LSI24</td>
<td>Chr13:37,907,681-37,907,781</td>
<td>-</td>
<td>2,003</td>
<td>-</td>
<td>intergenic</td>
</tr>
</tbody>
</table>

Total: 24, 10, 16, 4, 0
Figure captions

Fig. 1. Schematic workflow of the large sequence insertion (LSI) prediction approach used in this study. Two independent de novo assemblies were conducted using total read set and reads with both ends unmapped, yielding “ordinary contigs” and “orphan contigs”, respectively. Three filtering steps (a to c) were used to identify LSI candidates. The numbers of contigs satisfying the conditions for each step are shown in three tables. Step (a) retrieved ordinary contigs containing a full-length orphan contig with 100% identity in the mid-region. Step (b) ensured that both tails of an ordinary contig harboring an orphan contig are anchored to the G. max reference genome. Step (c) eliminated LSI artifacts located in unsequenced regions of the reference genome and selected only LSIs candidates supported by multiple reads with low mapping quality at the breakpoints due to mismatches.

Fig. 2. Schematic diagram presenting the predicted positions of LSIs based on the G. max reference genome. (a) genic region spanning the 3’ UTR, an intron, and the first CDS of Glyma.08G123500, (b) intron of Glyma.15G246800, (c) 5’ UTR of Glyma.12G239700, (d) 3’ UTR of Glyma.16G022500, (e) 1 kb upstream of Glyma.18G280400, and (f) 1 kb downstream of Glyma.16G032500.

Fig. 3. Size distribution of the insertions predicted by our LSI approach and three other SV prediction tools, BreakDancer, Pindel, and ScanIndel. Left and right Y-axes represent the number of insertions predicted by three SV prediction tools and our pipeline according to sizes, respectively.

Fig. 4. PCR validation of LSI05 in CW12 (G. soja) and Williams 82 (G. max) and its distribution in 29 G. soja and 34 G. max accessions. The LSI05c primer set was
designed against the conserved region flanking LSI05 (a), and amplifies DNA fragments of 1876 bp and 496 bp in CW12 and Williams 82, respectively (b). (c) PCR products of LSI05c in 29 G. soja and 34 G. max accessions. Superscript L and H on the accession names indicate a homozygous type with a large single amplicon (1876 bp) and a heterozygous type with double bands, respectively. CW, Chinese wild soybean; KW, Korean wild soybean; CC, Chinese cultivated soybean; KC, Korean cultivated soybean; M, Size marker
List of captions for Supplementary Files

**Table S1** The 30 *G. soja* and 35 *G. max* genotypes used for LSI identification or PCR validation of LSI05

**Table S2** Dominant primer sets and expected amplicon sizes for PCR validation of the LSIs present in the *G. soja* genotypes and absent in Williams 82

**Table S3** Co-dominant primer sequences and expected amplicon sizes for PCR validation of six LSIs in *G. soja* genotypes and Williams 82

**Table S4** Sequencing statistics of four *G. soja* genotypes and results of mapping to the *G. max* reference genome sequence (Wm82.a2)

**Table S5** Comparison of identified LSIs with SVs predicted by three other tools: Pindel, BreakDancer, and ScanIndel.

**Table S6** Stat and end positions and sizes of three insertions predicted by both our LSI pipeline and ScanIndel

**Fig. S1.** Comparison of amino acid sequences of Glyma.08G123500 on the soybean reference genome and its homolog predicted from the ordinary contigs carrying LSI05 (1,380 bp) in CW12. Both 14 amino acids from 1st and 14th and 10 amino acids from 29th to 38th were deleted by the LSI05 insertion.

**Fig. S2.** Gel electrophoresis of PCR amplification to validate the presence and absence of 24 LSIs (LSI01 to 24) in *G. soja* genotypes and the *G. max* reference Williams 82. For each LSI, the DNA band on the left column was amplified from the wild soybean genotype in which the LSI was predicted, and the right column shows the negative amplification (no band) from Williams 82. For the LSIs predicted in multiple accessions (LSI04, LSI06, LSI07 and LSI22), CW14 was used as a representative accession for
DNA amplification. Primers amplifying the Tubulin gene were used as a positive control. M, size marker.

Fig. S3. PCR validation to ascertain size differences in PCR products caused by large insertions (LSI10, 12, 16, 17, and 21) between G. soja genotypes and the G. max genotype Williams 82. For each primer set, G. soja genotypes had longer PCR fragments (middle column) and Williams 82 had shorter DNA bands (right column). M, size marker; W, Wild soybean (G. soja) genotype; C, Cultivated soybean (G. max) Williams 82.
Fig. 1. Schematic workflow of the large sequence insertion (LSI) prediction approach used in this study. Two independent de novo assemblies were conducted using total read set and reads with both ends unmapped, yielding "ordinary contigs" and "orphan contigs", respectively. Three filtering steps (a to c) were used to identify LSI candidates. The numbers of contigs satisfying the conditions for each step are shown in three tables. Step (a) retrieved ordinary contigs containing a full-length orphan contig with 100% identity in the mid-region. Step (b) ensured that both tails of an ordinary contig harboring an orphan contig are anchored to the G. max reference genome. Step (c) eliminated LSI artifacts located in unsequenced regions of the reference genome and selected only LSIs candidates supported by multiple reads with low mapping quality at the breakpoints due to mismatches.
Fig. 2. Schematic diagram presenting the predicted positions of LSIs based on the G. max reference genome. (a) genic region spanning the 3’ UTR, an intron, and the first CDS of Glyma.08G123500, (b) intron of Glyma.15G246800, (c) 5’ UTR of Glyma.12G239700, (d) 3’ UTR of Glyma.16G022500, (e) 1 kb upstream of Glyma.18G280400, and (f) 1 kb downstream of Glyma.16G032500.
Fig. 3. Size distribution of the insertions predicted by our LSI approach and three other SV prediction tools, BreakDancer, Pindel, and ScanIndel. Left and right Y-axes represent the number of insertions predicted by three SV prediction tools and our pipeline according to sizes, respectively.
Fig. 4. PCR validation of LSI05 in CW12 (G. soja) and Williams 82 (G. max) and its distribution in 29 G. soja and 34 G. max accessions. The LSI05c primer set was designed against the conserved region flanking LSI05 (a), and amplifies DNA fragments of 1876 bp and 496 bp in CW12 and Williams 82, respectively (b). (c) PCR products of LSI05c in 29 G. soja and 34 G. max accessions. Superscript L and H on the accession names indicate a homozygous type with a large single amplicon (1876 bp) and a heterozygous type with double bands, respectively. CW, Chinese wild soybean; KW, Korean wild soybean; CC, Chinese cultivated soybean; KC, Korean cultivated soybean; M, Size marker

181x149mm (300 x 300 DPI)