

University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

Agronomy & Horticulture -- Faculty Publications

Agronomy and Horticulture Department

2016

LETTER TO THE EDITOR *Ws-2* Introgression in a Proportion of *Arabidopsis thaliana Col-0* Stock Seed Produces Specific Phenotypes and Highlights the Importance of Routine Genetic Verification

Mon-Ray Shao

University of Nebraska-Lincoln, mshao2@unl.edu

Vikas Shedge

University of Nebraska-Lincoln, vshedge2@unl.edu

Hardik Kundariya

University of Nebraska-Lincoln, kundariyahardik@unl.edu

Fredric R. Lehle

LEHLE SEEDS, Round Rock TX

Sally Ann MacKenzie

University of Nebraska-Lincoln, sam795@psu.edu

Follow this and additional works at: <https://digitalcommons.unl.edu/agronomyfacpub>



Part of the [Agricultural Science Commons](#), [Agriculture Commons](#), [Agronomy and Crop Sciences Commons](#), [Botany Commons](#), [Horticulture Commons](#), [Other Plant Sciences Commons](#), and the [Plant Biology Commons](#)

Shao, Mon-Ray; Shedge, Vikas; Kundariya, Hardik; Lehle, Fredric R.; and MacKenzie, Sally Ann, "LETTER TO THE EDITOR *Ws-2* Introgression in a Proportion of *Arabidopsis thaliana Col-0* Stock Seed Produces Specific Phenotypes and Highlights the Importance of Routine Genetic Verification" (2016). *Agronomy & Horticulture -- Faculty Publications*. 868.

<https://digitalcommons.unl.edu/agronomyfacpub/868>

This Article is brought to you for free and open access by the Agronomy and Horticulture Department at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Agronomy & Horticulture -- Faculty Publications by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

LETTER TO THE EDITOR

***Ws-2* Introgression in a Proportion of *Arabidopsis thaliana Col-0* Stock Seed Produces Specific Phenotypes and Highlights the Importance of Routine Genetic Verification**

Arabidopsis thaliana is an important model organism with a robust network of resources that has been of enormous value to the plant science research community. The use of isogenic material as a reference point or control is critical for many types of experiments in plant molecular biology and genetics. Recently, we noticed that some seed from a common source of the widely-used *Columbia-0* (*Col-0*) strain gave rise to plants showing features atypical for this strain. Whole genome DNA-sequencing and allele-specific PCR assays confirmed that the abnormal individuals contain multiple introgressions from the ecotype *Wassilewskija-2* (*Ws-2*), as described below. This emphasizes the importance of practices necessary to maintain the integrity of seed stocks and other biological collections. We urge research groups to evaluate whether they may have been affected and to revisit their materials if needed.

PHENOTYPIC VARIANTS WITHIN A COMMON *Col-0* SEED STOCK CONTAIN CHROMOSOMAL SEGMENTS FROM *Ws-2*

Relative to other *Arabidopsis* ecotypes, *Col-0* is characterized by a medium rosette size, slightly serrated leaf margins, intermediate height, and an intermediate flowering time (The Arabidopsis Information Resource [TAIR], www.arabidopsis.org). However, in *Col-0* seed (lot #214-509) obtained from LEHLE SEEDS Company, we observed that some plants were larger than others grown in the same tray. The abnormal plants showed an increase in leaf area and displayed broader, flatter, and more serrated leaves, and tended to flower earlier than other individuals (particularly when grown under long-day cycles; Supplemental Figure 1). Based on phenotypic scoring, we observed that approximately 6-10% (Table 1) of plants from the indicated lot were phenotypically abnormal compared to the majority of the *Col-0* plants.

We suspected that the abnormal individuals could be the result of seed contamination, and chose six plants (four abnormal and two typical) for whole genome DNA-sequencing. Relative to the reference *Col-0* sequence, the two typical plants each had fewer than 700 homozygous single nucleotide polymorphisms (SNPs) detected by the software SHORE

(Schneeberger et al., 2009). The four abnormal individuals, however, each had over 131,000 homozygous SNPs (Supplemental Figure 2A), concentrated on Chromosomes 1, 3, and 5. Except for a comparatively small segment on Chromosome 5, the distributions of SNPs were essentially identical across all four abnormal individuals (Supplemental Figure 2B), with an introgression pattern and number consistent with expectations if outcrossing followed by recombination occurred (Giraut et al., 2011). The apparent lack of segregation among the four abnormal samples indicates that the genetic contamination in this seed lot likely arose from a single hybridization event followed by self-pollination for several additional generations, leading to genetic fixation in these individuals.

Table 1. Estimated percentage of affected plants in indicated *Col-0* seed lots, based on phenotypic scoring or PCR genotyping.

Lot	Year	Method	Plants Grown	Affected Plants	Frequency
#197-089	1998	PCR	23	0	0%
#203-280	2003	PCR	84	8	9.5%
#206-440	2007	PCR	84	5	6.0%
#210-485	2010	Phenotype	275	17	6.2%
#214-509	2014	Phenotype	213	22	10.3%
#215-511	2015	PCR	96	15	15.6%

A comparison of the SNPs in abnormal plants to variant files generated by the *Arabidopsis* 1001 Genomes Project revealed that approximately 84% (or 115,000) of SNPs called from each abnormal plant were also found in the *Ws-2* ecotype. The remaining 16% of abnormal sample SNPs unmatched to *Ws-2* follow a similar chromosomal distribution to the SNPs matching *Ws-2* (Supplemental Figure 2B), suggesting that they may be the result of technical differences in software and parameter settings, as opposed to an additional genetic source. The abnormal sample SNPs that matched *Ws-2* also had nearly an equally strong match to the ecotype *Rag1-1*. Indeed, the publicly available *Ws-2* and *Rag1-1* variant files share over 91% SNPs in common, an extremely high proportion for two *Arabidopsis* ecotypes purportedly from geographically distant regions (*Ws-2* from Russia, *Rag1-1* from UK). A comparison between 5965 accessions by Anastasio *et al.* (2011) placed *Rag1-1* within the same haplogroup as *Ws-2* and identified it as one of the hundreds of accessions having potentially mis-identified geographic origins. No other ecotype from the 838 accessions we tested accounted for more than 65% of the abnormal sample SNPs, and most accounted for only

18.78% to 50.81% of the abnormal sample SNPs, which is consistent with comparisons between different ecotypes (Salomé and Weigel, 2015).

SNPhylo (Lee et al., 2014) was used to perform SNP-based phylogenetic analysis with variant files of *Ws-2*, additional ecotypes (primarily from the same sequencing project as *Ws-2*), and our samples. For these samples and ecotypes, SNPs within a subset of coordinates contained in the large introgressions on Chromosomes 1, 3, and 5 were included. Corroborating the prior SNP matching approach, these results indicated that the genomic blocks within the abnormal samples are similar to *Ws-2* (Supplemental Figure 2C), while the typical-appearing *Col-0* samples are likely pure *Col-0*. Based on these results, we suggest that *Ws-2*, a relatively common laboratory ecotype provided by several seed distributors, is the donor of the observed genetic variation. We found no compelling evidence of additional genetic history involving artificial mutagenesis in the abnormal samples.

Recently, phenotypes thought to be caused by a mutant allele for the auxin-binding protein *ABP1* were called into doubt when it was discovered that the *abp1-5* mutant line harbored multiple second-site mutations, as well as a large *Ws-2* introgression (Enders et al., 2015) that may have resulted from backcrossing to a contaminated *Col-0* plant. Although both have introgressions on Chromosome 3, the overall distribution of *Ws-2* SNPs in our abnormal samples is not the same as the *abp1-5* line, suggesting that these events occurred independently.

To validate the DNA sequencing results, single-nucleotide amplified polymorphism (SNAP) primers were designed based on known *Ws-2* SNPs that were also identified in our abnormal samples by SHORE. All the plants that were visually scored as abnormal harbored *Ws-2*-specific SNPs in the introgressed regions and *Col-0*-specific SNPs in the introgression-free region, while all the plants that were visually scored as normal harbored *Col-0* specific SNPs in all the regions tested (Supplemental Figure 3). Using two PCR markers, random sampling of 96 seedlings from lot #215-511 found that 15 seedlings (15.6% of total) were positive for *Ws-2* SNPs (Table 1), a proportion generally consistent with phenotypic scoring. PCR assays of additional earlier lots indicate that seed lots going back to at least 2003 are affected.

Thus, whole genome DNA-sequencing and allele-specific PCR assays confirm that a genetic mixture between *Col-0* and *Ws-2* is present in a proportion of this *Col-0* seed stock. Groszmann et al. (2014) found that *Col-0* x *Ws* hybrids show modest (~15%) heterosis in

rosette diameter up to the first 28 days of growth. Thus the phenotypic changes in our abnormal samples may reflect *Col-0/Ws-2* heterosis.

LEHLE *Col-0* PEDIGREE AND PROTOCOLS

LEHLE SEEDS Company propagates seed for 21 *Arabidopsis* ecotypes. Historically, different ecotypes have been planted at different times at four locations in Tucson, AZ and four locations in Round Rock, TX. *Col-0* bulk seed has been grown continuously since 1985, with 48 generations as of the most recent 2015 bulk. *Ws-2* bulk seed has been grown every 1-2 years since 1989, with the last bulk in 2003, for a total of 13 generations. In 1995-1998, 2000, and 2003, *Col-0* and *Ws-2* bulks were grown in the same location, providing the opportunity for possible cross-pollination.

The standard practice by LEHLE SEEDS is to perform simple sequencing length polymorphism (SSLP) analysis using five separate markers on DNA preparations from several pooled plants and from 20 individual plants. Two of these markers appear to fall within the *Ws-2* introgressions, suggesting that the contamination was at sufficiently low frequency (ca. 6%) when last tested.

From now on, LEHLE SEEDS will abandon the practice of using raw seed from a previous propagation for a high density planting of the next propagation. Rather, ecotypes will be bulk propagated only from selfed progeny of individual plants, all of which will have passed a growth-stage phenotypic analysis for conformity to ecotype under low density planting and easily reproducible conditions. A recent hydroponic platform for *Arabidopsis* looks promising for this purpose (Conn et al. 2013). As there seems to be no solution for eliminating out-crossing completely in confined spaces, LEHLE SEEDS will conduct future bulk propagations of *Col-0* in complete physical isolation from other ecotypes.

RECOMMENDATIONS

Given typical high-density growth conditions of *Arabidopsis*, along with its branching habit, numerous flowers, fecundity and small seed size, accidental outcrossing or seed mixtures can sometimes occur despite good practices. Thus, careful observation and molecular characterization are recommended, particularly when a mutation of interest could conceal other phenotypes epistatically, or when conducting experiments such as genetic screens where the large number of plants used increases the chances of inadvertently including a contaminant (Greene et al., 2003). If multiple *Arabidopsis* ecotypes are grown at a single facility, steps taken

to ensure genetic purity can include staggered planting, physical separation, staking, floral sleeves, careful seed collection habits, secure seed drying, organized storage and periodic genotyping.

Recent advancements in bioinformatics tools enable SNP calling from a variety of sequencing applications (Ossowski et al., 2008; Van der Auwera et al., 2013) as a confirmation of genetic background and purity. Since an increasing number of research projects leverage some form of sequencing data, one recommendation is that, whenever appropriate, SNP analysis be routinely included in the bioinformatics pipeline of high-throughput sequencing experiments.

Mon-Ray Shao
Department of Agronomy and Horticulture
University of Nebraska

Vikas Shedge
Department of Agronomy and Horticulture
University of Nebraska, Lincoln, NE 68588

Hardik Kundariya
Department of Agronomy and Horticulture
University of Nebraska, Lincoln, NE 68588

Fredric R. Lehle
LEHLE SEEDS
PO Box 2366, Round Rock TX 78681-2366
ORCID ID: [0000-0002-9019-6099](https://orcid.org/0000-0002-9019-6099)

Sally A. Mackenzie
Department of Agronomy and Horticulture
University of Nebraska, Lincoln, NE 68588
ORCID ID: [0000-0003-2077-5607](https://orcid.org/0000-0003-2077-5607)
smackenzie2@unl.edu

SUPPLEMENTAL DATA

Supplemental Figure 1. Phenotype of abnormal plants in LEHLE SEEDS Company *Col-0* lot #214-509.

Supplemental Figure 2. DNA sequencing of abnormal plants reveals large *Ws-2* introgressions.

Supplemental Figure 3. Allele-specific PCR assays confirm presence of *Ws-2* introgression in abnormal plants.

Supplemental Table 1. SNAP PCR primer sequences used for genotyping shown in Supplemental Figure 3.

ACKNOWLEDGEMENTS

We thank the University of Nebraska Beadle Center Greenhouse Facility for assistance in tracing the seed contamination and identifying seed lots. Funding for this analysis was provided by a grant from the Bill and Melinda Gates Foundation to S.M.

AUTHOR CONTRIBUTIONS

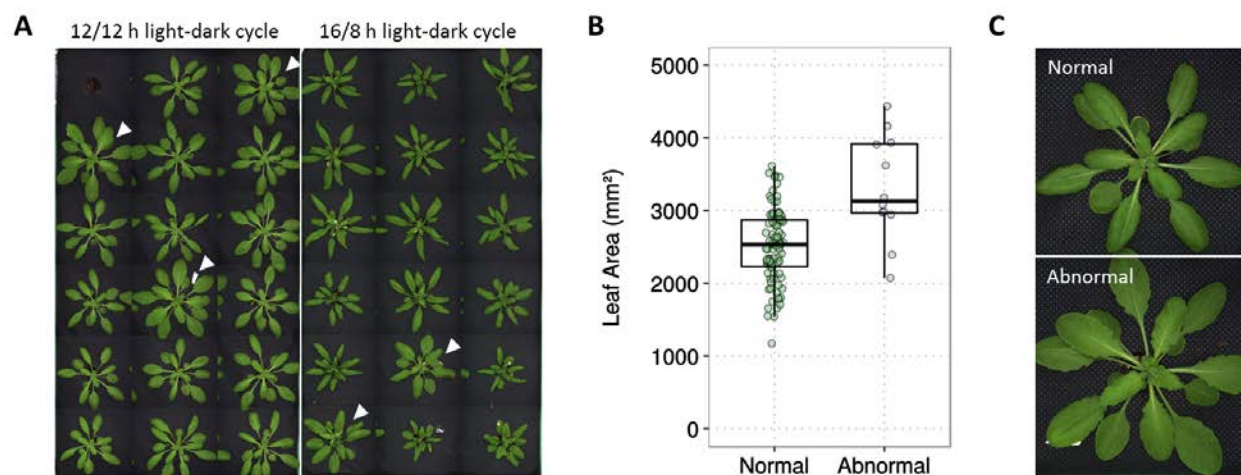
M.R.S., V.S., and S.A.M. designed the experiments. M.R.S. performed DNA-sequencing analysis. V.S. performed SNAP PCR assays. H.K. provided phenotypic measurements. M.R.S., F.L., and S.A.M. wrote the article. First authors M.R.S and V.S. contributed equally.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Sally A. Mackenzie (smackenzie2@unl.edu).

REFERENCES

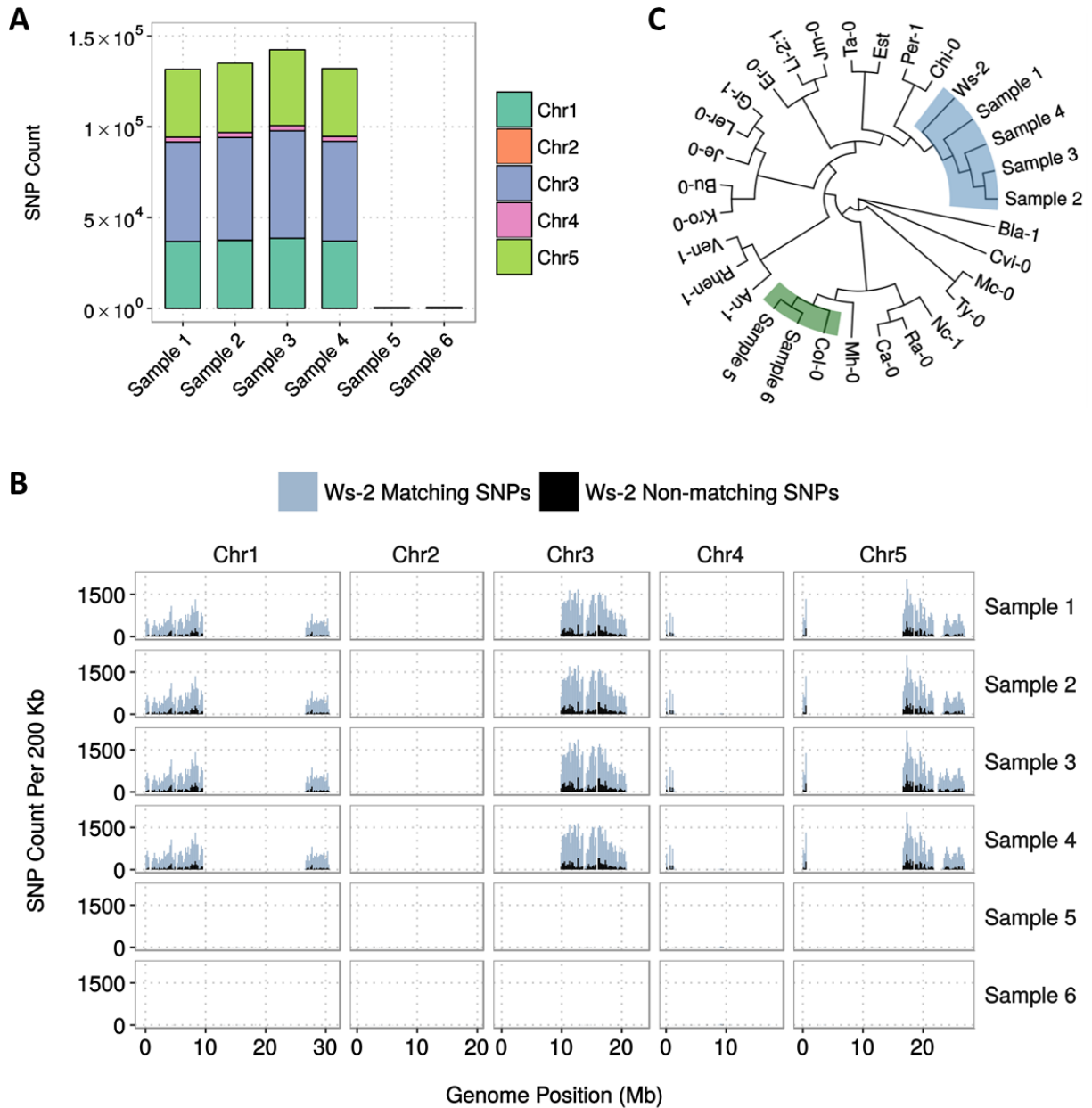
- Anastasio, A.E., Platt, A., Horton, M., Grotewold, E., Scholl, R., Borevitz, J.O., Nordborg, M., Bergelson, J.** (2011). Source verification of mis-identified *Arabidopsis thaliana* accessions. *Plant J.* **67**: 554-66.
- Conn, S.J., Hocking, B., Dayod, M., Xu, B., Athman, A., Hederson, S., Auket, L., Conn, V., Shearer, M.K., Fuentes, S., Tyerman, S.D., and Gillham, M.** (2013) Protocol: optimizing hydroponic growth systems for nutritional and physiological analysis of *Arabidopsis thaliana* and other plants. *Plant Methods* **9**:4.
- Enders, T.A., Oh, S., Yang, Z., Montgomery, B.L., Strader, L.C.** (2015). Genome Sequencing of *Arabidopsis* *abp1-5* Reveals Second-Site Mutations That May Affect Phenotypes. *Plant Cell* **27**: 1820-6.
- Greene, E.A., Codomo, C.A., Taylor, N.E., Henikoff, J.G., Till, B.J., Reynolds, S.H., Enns, L.C., Burtner, C., Johnson, J.E., Odden, A.R., Comai, L., Henikoff, S.** (2003). Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in *Arabidopsis*. *Genetics* **164**: 731-40.
- Giraut, L., Falque, M., Drouaud, J., Pereira, L., Martin, O.C., Mézard, C.** (2011). Genome-wide crossover distribution in *Arabidopsis thaliana* meiosis reveals sex-specific patterns along chromosomes. *PLoS Genet.* **7**: e1002354.
- Groszmann, M., Gonzalez-Bayon, R., Greaves, I.K., Wang, L., Huen, A.K., Peacock, W.J., Dennis, E.S.** (2014). Intraspecific *Arabidopsis* hybrids show different patterns of heterosis despite the close relatedness of the parental genomes. *Plant Physiol.* **166**: 265-80.
- Lee, T.H., Guo, H., Wang, X., Kim, C., Paterson, A.H.** (2014). SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics* **15**: 162.
- Salomé, P.A., and Weigel, D.** (2014). Plant genetic archaeology: whole-genome sequencing reveals the pedigree of a classical trisomic line. *G3* **5**: 253-9.
- Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A.H., Nielsen, K.L., Jørgensen, J.E., Weigel, D., and Andersen, S.U.** (2009). SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat. Methods* **6**: 550-551.
- Ossowski, S., Schneeberger, K., Clark, R.M., Lanz, C., Warthmann, N., Weigel, D.** (2008). Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* **18**: 2024-33.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K.V., Altshuler, D., Gabriel, S., DePristo, M.A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **11**: 11.10.1-11.10.33.



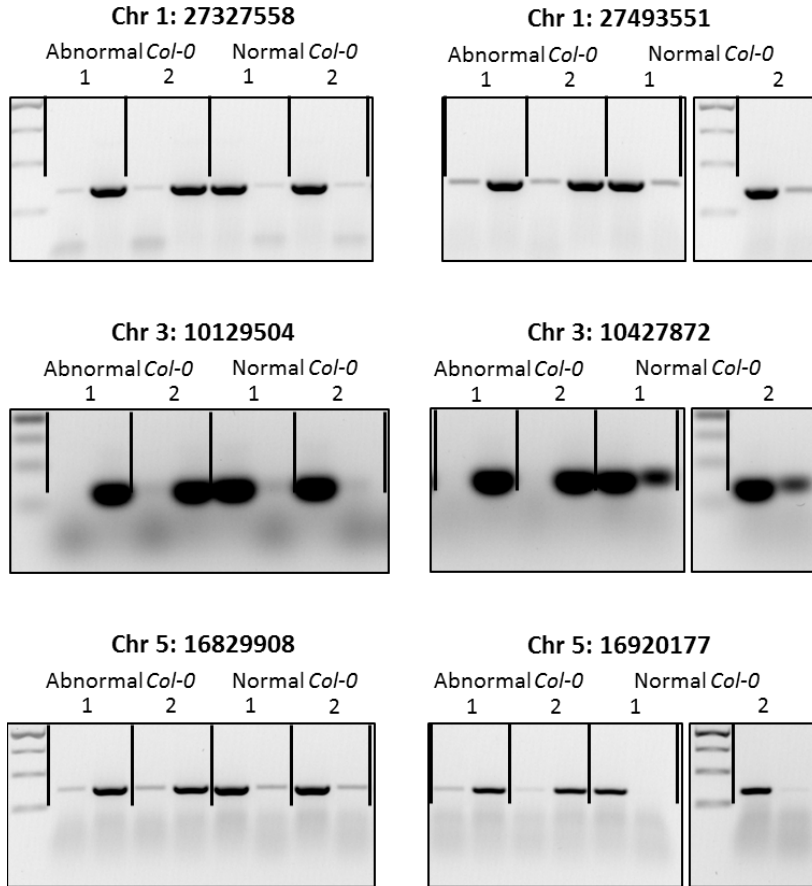
Supplemental Figure 1. Phenotype of abnormal plants in LEHLE SEEDS Company *Col-0* lot #214-509.

- (A) Randomly planted seeds grown under short day and long day conditions. White arrows indicate phenotypically abnormal individuals.
- (B) Leaf area of normal versus abnormal plants at 28 days old (n = 106).
- (C) Representative image under short day conditions showing the flatter, more serrated leaves in the abnormal individual when compared to a normal *Col-0* plant.



Supplemental Figure 2. DNA sequencing of abnormal plants reveals large *Ws-2* introgressions.

- (A) SNP counts in abnormal (Samples 1-4) and normal (Samples 5-6) *Col-0* plants.
 (B) Chromosomal distribution of SNPs within each sample.
 (C) Cladogram of sequenced samples plus 25 ecotypes, using SNP data from genomic regions showing high SNP density in the abnormal samples.



Supplemental Figure 3. Allele-specific PCR assays confirm presence of *Ws-2* introgression in abnormal plants.

Two abnormal and two normal sample results are shown for the indicated SNP genomic positions on Chromosomes 1, 3, and 5. For each PCR assay, the *Ws-2* specific primer result is in the right lane and the *Col-0* specific primer result is in the left lane. Primer sequences are listed in Table 1.

Supplemental Table 1. SNAP PCR primer sequences used for genotyping shown in Supplemental Figure 3.

Position	Primer name	Primer sequence
Chr1: 27327558	AT1G72560-Col0-F	TGTGACTGCCATCTTCAGTTCCCGT
	AT1G72560-Ws2-F	AAAATGTGACTGCCATCTTCAGTTCCACA
	AT1G72560-Rev	CTTTCCTTGATGATAGGGGTTTGCATCAC
Chr1: 27493551	AT1G73100-Col0-F	TGGTTCTTTCATATGTGAATATGCTGGTGAAGTTA
	AT1G73100-Ws2-F	TGGTTCTTTCATATGTGAATATGCTGGTGAAGTAG
	AT1G73100-Rev	GTCAATTCTGCCATAGGAGGGATATGACG
Chr3: 10129504	AT3G27360-Col0-F	CCTCTAATCCTCCTCGCAAGTTGAATATCCATA
	AT3G27360-Ws2-F	CTAATCCTCCTCGCAAGTTGAATATCCCTC
	AT3G27360-Rev	TGGCGACGAAAGCAGCAAGGA
Chr3: 10427872	AT3G28030-Col0-F	TCAGAAAGGGATAACCAAACAATTTGGTACAGA
	AT3G28030-Ws2-F	GTCAGAAAGGGATAACCAAACAATTTGGTACATT
	AT3G28030-Rev	CCCTTGATTTCAGTGACAAGACTCTGGATTTT
Chr5: 16829908	AT5G42100-Col0-F	CTTAACAGTAACGACCTTCTAGTCCTAAATGTCAAC
	AT5G42100-Ws2-F	CTTAACAGTAACGACCTTCTAGTCCTAAATGTGTAA
	AT5G42100-Rev	TCACCATCTTCGTCTTTGCTCTCTTTAACG
Chr5: 16920177	AT5G42320-Col0-F	CAAATATGGTTACCATTCTTCGTTCCACAAAGAACT
	AT5G42320-Ws2-F	AAATATGGTTACCATTCTTCGTTCCACAAAGATGA
	AT5G42320-Rev	TTTTGACAGATGATTCTTTTCTTCCAGACTTTT

SUPPLEMENTAL METHODS

Plant Growth Conditions and Phenotyping

Arabidopsis thaliana seeds were sown directly onto soil and stratified at 4°C for 2 days, then grown at 22°C under 12/12 hour light-dark cycles (short day) or 16/8 hour light-dark cycles (long day) as indicated. For leaf area, short day -grown plants were imaged and measured using a LemnaTec Scanalyzer HTS system.

Whole Genome DNA Sequencing and Analysis

DNA for abnormal and typical samples was extracted using the DNeasy Plant Mini Kit (Qiagen) according to manufacturer protocol. Sequencing was performed by the University of Minnesota Genomics Center using a HiSeq 2500 to generate 100 bp paired-end reads. Samples were prepared using TruSeq Nano DNA libraries with an average insert size of approximately 350 bp. All six libraries were pooled and sequenced in a single lane. Each library had an average Q-score above 35, and yielded over 34 million reads per sample. De-multiplexed sequencing data is deposited at SRA accession SRX1630132.

Trim Galore (Babraham Bioinformatics) was used to remove remaining Illumina adapters and bases from the 5' and 3' ends with a PHRED score below 20. Reads were aligned against the *Col-0* reference sequence version TAIR10 using GenomeMapper, with a maximum edit

distance of 10% and a maximum gap of 7%. SNP calling was then performed using SHORE with the developer-included homozygous scoring matrix. Sample SNPs were further filtered to those within the nuclear genome with a quality score ≥ 35 , depth ≥ 10 , mapping-quality-zero $\leq 0.05 \times \text{depth}$, alternative allele frequency ≥ 0.95 , and unambiguous reference and alternative allele identities (A/T/G/C only). Filtering and calculation of SNP frequencies and genomic distributions were performed in R.

Variant files for *Arabidopsis* accessions sequenced by the Salk Institute, the Max Planck Institute for Developmental Biology (MPI), the George Mendel Institute of Molecular Plant Biology (GMI), and the Wellcome Trust Centre for Human Genetics (WTCHG) were obtained from the 1001 Genomes Project (<http://1001genomes.org>). Sample SNPs were considered to be matching those of another accession's if they shared the same position and alternative allele identity. To include the relatively common *C24* ecotype as a potential genetic donor, a variant file updated to reflect the TAIR10 reference was generated from publicly available sequencing data for *C24* (Schneeberger et al., 2011), using the same GenomeMapper and SHORE parameters as described above.

Phylogenetic analysis was based on introgression coordinates in the abnormal samples, and included regions Chr1: 2-8 Mb, Chr3: 10-20 Mb, and Chr5: 18-21 Mb. Twenty-five ecotypes, including *Col-0* and *Ws-2*, along with the abnormal samples were analyzed with SNPPhylo, using a maximum-likelihood based method. For SNP pruning, parameters were set at minor allele frequency ≥ 0.05 and a linkage disequilibrium threshold of 0.7, resulting in 30101 SNP sites kept for analysis. The cladogram was then visualized using the R/Bioconductor package *ggtree* (<http://bioconductor.org/packages/ggtree>).

Allele-Specific PCR Assay

SNAP PCR primers were designed for the introgressed regions on Chromosomes 1, 3 and 5, and introgression-free region on Chromosome 2 as recommended by the SNAP protocol/SNAP program (Drenkard et al., 2000) for *Col-0* and *Ws-2* specific SNPs. SNP sites within genes were preferentially chosen (indicated within primer name), and the genomic positions given are based on the *Col-0* reference sequence version TAIR10. PCR cycling was done as 5 min at 94°C, (30 sec at 94°C, 1 min at 62°C) 35 times, and finally 10 min at 72°C.

SUPPLEMENTAL REFERENCES

Drenkard, E., Richter, B.G., Rozen, S., Stutius, L.M., Angell, N.A., Mindrinos, M., Cho, R.J., Oefner, P.J., Davis, R.W., Ausubel, F.M. (2000). A simple procedure for the analysis of single nucleotide polymorphisms facilitates map-based cloning in *Arabidopsis*. *Plant Physiol.* **124**: 1483-92.

Schneeberger, K., Ossowski, S., Ott, F., Klein, J.D., Wang, X., Lanz, C., Smith, L.M., Cao, J., Fitz, J., Warthmann, N., Henz, S.R., Huson, D.H., Weigel, D. (2011). Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. *Proc. Natl. Acad. Sci. USA* **108**: 10249-54.