LETTER TO THE EDITOR

A Proposal Regarding Best Practices for Validating the Identity of Genetic Stocks and the Effects of Genetic Variants^{OPEN}

Colleagues from the medical field have estimated that up to one-third of cell lines are contaminated with other cell lines or are misidentified; in addition, repeated passaging substantially changes cell line properties (reviewed in Hughes et al., 2007). The medical community has therefore begun to establish standards for verification of cell lines and genetic stocks, and NIH has announced efforts to require validation and to aid researchers in validating their biological material (Lorsch et al., 2014). Plant biologists should do the same. Even though the propagation of seed stocks cannot be directly compared with animal cell culture, contamination is a real possibility, and it is not uncommon that the same genetic stock produces different phenotypes in different laboratories. Confirming the genetic identity of research material is necessary to know whether such phenotypic differences reflect gene-by-environment (GxE) interactions or whether they are simply due to apples being compared to oranges.

There are two principal areas of concern in this regard. The first one is that transgenic lines, mutants, or accessions are not what they are supposed to be; that is, they carry a different transgene, allele, or mutation, or they are a different accession than assumed. Likely sources are inadvertent seeding of soil, mislabeling of plants, and mix-ups during seed collection. Those who work with self-fertilizing species such as *Arabidopsis thaliana* must be aware of the risk of outcrossing. In the field, the rate of outcrossing is often a few percent (Bergelson et al., 1998; Bomblies et al., 2010; Platt et al., 2010). In addition, specific genetic backgrounds with altered floral morphologies or reduced male fertility can greatly increase outcrossing rates (Peng et al., 2006; Luo and Widmer, 2013).

^{OPEN}Articles can be viewed online without a subscription.
www.plantcell.org/cgi/doi/10.1105/tpc.15.00502

Systematic analyses have indicated the scope of these issues. For example, genotyping with single nucleotide polymorphism markers revealed that up to 5% of Arabidopsis accessions in the stock centers were mislabeled (Anastasio et al., 2011; Simon et al., 2012). Another example comes from grapevine (*Vitis vinifera*), where an inbred Pinot Noir derivative was targeted for genome sequencing, but it was not noticed until later that there was either an uncontrolled outcrossing event, or a complete mix-up, so that in the end a Pinot Noir × Helfensteiner hybrid, or possibly a selfed Helfensteiner derivative, was sequenced (Jaillon et al., 2007). Similarly, wild-type strains of *Chlamydomonas reinhardtii* have complex histories, with spontaneous mutations and several examples of misidentification (Gallaher et al., 2015). In at least one case, a mutant line was more closely related to other wild-type lines than its supposed isogenic parent (Blaby et al., 2013). There are also plenty of anecdotes of mix-ups of T-DNA insertion lines in Arabidopsis; one published example is a line for which a specific, published insertion subsequently could not be detected again (Richter et al., 2010).

A range of issues across the entire research and breeding community affects the consistent preservation of germplasm identity. The visually obvious hybrid vigor of outcrossed progeny in maize (*Zea mays*) or distinct morphological defects in many mutants help reduce the problem, but genetic contamination or mix-ups may occur easily, especially if one is not familiar with subtle phenotypic variation. The standard approaches for maintaining germplasm identity in most model species are good, but even if errors are reduced to 1% per generation, roughly 10% of stocks will be mislabeled after 10 generations without genetic validation, which should be common practice.

A second area of concern is that phenotypic effects can be due to additional, unrecognized genetic variants segregating in

a stock. Initially pure genetic lines can diverge through secondary, spontaneous mutations, while progeny of lines that were initially not pure can diverge through fixation of segregating variation. One such example is the Arabidopsis accession Landsberg *erecta*, where one set of stocks being used in the community has a *hua2* mutation, while another set of lines does not (Doyle et al., 2005). The origin of this mutation is unclear. There are also several examples of spontaneous mutations in Arabidopsis accessions identified by their phenotypic effects (Loudet et al., 2008; Laitinen et al., 2010), including one that arose in the Col-0 reference strain (Coustham et al., 2014). Hundreds of mutations have been documented in another Col-0 lab stock by whole-genome sequencing (Cubillos et al., 2014), and thousands of single base pair mutations along with several new transposon insertions have apparently occurred since a standard Chlamydomonas line was established in the laboratory (Gallaher et al., 2015). In Arabidopsis, genome-wide mutation rates are ~1 per generation for single nucleotide polymorphisms and 0.5 per generation for indels (Ossowski et al., 2010; Jiang et al., 2014). Guidelines for how many differences are deemed acceptable before a stock is given a distinct name or identifier are needed and should be developed by each community.

An example of fixation of segregating variants is the soybean reference cultivar Williams 82. Several different lines are in use, apparently because Williams 82 was distributed before it was fully inbred and thus completely homozygous throughout the genome (Haun et al., 2011). Residual heterozygosity giving rise to different sublines among inbreds with the same name has also been described in maize (Romay et al., 2013). This even extends to the maize reference genome B73, where 1.5 Mb (0.07%) of the genome differs between sublines that are nearly 50 years old now (Gore et al., 2009). Similarly, it has only recently been recognized,

based on whole-genome resequencing, that many common Chlamydomonas laboratory strains are segregants apparently derived from a single cross (Gallaher et al., 2015).

A related problem is that phenotypes are attributed to a known mutation in a particular stock, but these are instead caused by second-site mutations; several such examples from Arabidopsis are listed in Table 1. Similarly, insufficient characterization of the exact molecular defect at a mutant locus can lead to confusion, for example, when a mutation is more complex than initially realized and affects more than one gene. This was the case for tomato (*Solanum lycopersicum*) *fas*, Arabidopsis *abp1-1*, and Chlamydomonas *sta6* mutants, where at least some aspects of the reported phenotypes were caused by genes adjacent to the ones initially thought to be responsible (Blaby et al., 2013; Dai et al., 2015; Xu et al., 2015). These problems are not restricted to mutant lines; a trisomic Arabidopsis stock that was otherwise considered wild type turned out to have a complex history involving introgression of a mutation into a different background, with that mutation explaining an important aspect of the phenotype (Salomé and Weigel, 2015).

Because of these issues, we suggest a series of best practices for verifying both the identity of genetic stocks and the causal relationship between a genetic variant and a phenotype.

1. State in an article as clearly and exactly as possible the origin of mutants, transgenics, and accessions, including genetic stocks that were used as background for random mutagenesis, genome editing, or generation of transgene insertions. This should minimally include the stock center or the scientist and laboratory who donated the stock and preferably also the approximate date when the stock was acquired and, if available, seed lot information.

2. State if and how the identity of genetic stocks and the effects of specific mutations were verified. For example, if no further molecular or genetic validation was undertaken, please state this, potentially with an appropriate reason (example: "root phenotype is diagnostic for this mutant"). Examples of genetic validation that can link a phenotype to a specific genetic variant include backcrossing and cosegregation analysis, transgenic complementation, examination of a second allele, and recreation of a mutant using gene editing. In the case of mutants, best practice is often to phenotype a segregating population, using homozygous wild-type siblings as control, to account for genetic background differences that may not be represented in the original parent line. Examples of molecular validation include PCR amplification of transgene sequences (O'Malley et al., 2015), targeted analysis of selected polymorphic markers, array-based genotyping or genotyping-by-sequencing (Salathia et al., 2007; Anastasio et al., 2011; Elshire et al., 2011; Simon et al., 2012), shallow resequencing, or even complete resequencing. Ensure that seeds of key stocks for a published article are saved, and consider resubmitting such stocks to stock centers.

While large-scale genotyping or resequencing is not yet practical as a standard, such efforts are already going on for the accessions from the Arabidopsis 1001 Genomes project, in conjunction with curation of the 1001 Genomes collection of accessions by the ABRC, similar to large-scale genotyping efforts at the U.S. national maize inbred seed bank (Romay et al., 2013). Each community focused on a specific organism should also consider how to empower all laboratories, regardless of their size, resources, and bioinformatics sophistication, to harness the power of ever cheaper genome sequencing and genotyping for stock validation, along the lines suggested by NIH for vertebrate cell lines (Lorsch et al., 2014). A key role should be played by stock centers, which should implement routines for validating newly submitted as well as propagated stocks and which may consider offering services that go beyond their core species and provide access to colleagues working with less popular, non-model organisms. Useful in this regard are Web resources that allow facile matching of patterns of genome-wide polymorphisms found in a specific stock with a database of sequenced genomes, as they have been implemented for Chlamydomonas (https://bitbucket.org/gallaher/custom-chlamy-generator) and Arabidopsis (http://tools.1001genomes.org/strain_id/). Other communities are encouraged to develop similar resources.

Of course, efforts must be commensurate with the number of lines analyzed or the goal of a specific experiment. That is, greater care needs to be taken in a study that primarily focuses on a single genetic stock than, say, a GWAS study with hundreds or thousands of lines. In the latter case, an estimate of the number of misidentified lines should suffice, e.g., by genotyping a subsample. As another example, a mutant that has an easily recognized, distinctive phenotype and that is being used merely for a complementation cross needs to be less rigorously validated than a mutant that is used for detailed phenotypic comparison with an isogenic wild-type control.

We would like to emphasize that spending a few hundred dollars and a week or so on validating genetic stocks is appropriate, since experimental articles typically report on research that has been performed over years and has cost tens to hundreds of thousands of dollars (including salaries). Of course, all of us should expend more effort to avoid mix-ups in the first place, for example, by using better practices, such as strict separation of areas for seed production and harvest from areas for sowing seeds.

Table 1. Second-Site Mutations Responsible for Major Phenotypes in Arabidopsis

Mutant Stock	Major Phenotype	Candidate for Causal Second-Site Mutation	Reference
abp1-5	Long hypocotyl	*phyB*	Enders et al. (2015)
avp1-1	Auxin transport defect	*gnom*[a]	Kriegel et al. (2015)
coi-16	Compromised non-host resistance	*pen2*[a]	Westphal et al. (2008)
mca2-1	Abnormal touch-induced root responses	*axr4*	Frontiers in Plant Science Editorial Office (2015)
abi1-3	Increased ABA sensitivity	*mkk1*[a]	Wu et al. (2015)
tt4(2YY6)	Bushy	*max4*[a]	Bennett et al. (2006)

ABA, abscisic acid.
[a]Confirmed by genetics.

Joy Bergelson
University of Chicago
Chicago, Illinois 60637

Edward S. Buckler
USDA-ARS/Cornell University
Ithaca, New York 14850
ORCID ID: 0000-0002-3100-371X

Joseph R. Ecker
HHMI-The Salk Institute for
Biological Studies
La Jolla, California 92037
ORCID ID: 0000-0001-5799-5895

Magnus Nordborg
Gregor Mendel Institute
1030 Vienna, Austria
ORCID ID: 0000-0001-7178-9748

Detlef Weigel
Max Planck Institute for
Developmental Biology
72076 Tübingen, Germany
weigel@weigelworld.org
ORCID ID: 0000-0002-2114-7963

REFERENCES

Anastasio, A.E., Platt, A., Horton, M., Grotewold, E., Scholl, R., Borevitz, J.O., Nordborg, M., and Bergelson, J. (2011). Source verification of mis-identified *Arabidopsis thaliana* accessions. Plant J. **67:** 554–566.

Bennett, T., Sieberer, T., Willett, B., Booker, J., Luschnig, C., and Leyser, O. (2006). The *Arabidopsis MAX* pathway controls shoot branching by regulating auxin transport. Curr. Biol. **16:** 553–563.

Bergelson, J., Stahl, E., Dudek, S., and Kreitman, M. (1998). Genetic variation within and among populations of *Arabidopsis thaliana*. Genetics **148:** 1311–1323.

Blaby, I.K., et al. (2013). Systems-level analysis of nitrogen starvation-induced modifications of carbon metabolism in a *Chlamydomonas reinhardtii* starchless mutant. Plant Cell **25:** 4305–4323.

Bomblies, K., Yant, L., Laitinen, R.A., Kim, S.T., Hollister, J.D., Warthmann, N., Fitz, J., and Weigel, D. (2010). Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of *Arabidopsis thaliana*. PLoS Genet. **6:** e1000890.

Coustham, V., Vlad, D., Deremetz, A., Gy, I., Cubillos, F.A., Kerdaffrec, E., Loudet, O., and Bouché, N. (2014). SHOOT GROWTH1 maintains Arabidopsis epigenomes by regulating *IBM1*. PLoS One **9:** e84687.

Cubillos, F.A., Stegle, O., Grondin, C., Canut, M., Tisné, S., Gy, I., and Loudet, O. (2014). Extensive cis-regulatory variation robust to environmental perturbation in *Arabidopsis*. Plant Cell **26:** 4298–4310.

Dai, X., Zhang, Y., Zhang, D., Chen, J., Gao, X., Estelle, M., and Zhao, Y. (2015). Embryonic lethality of *Arabidopsis abp1-1* is caused by deletion of the adjacent *BSM* gene. Nat. Plants **1:** 15183.

Doyle, M.R., Bizzell, C.M., Keller, M.R., Michaels, S.D., Song, J., Noh, Y.S., and Amasino, R.M. (2005). *HUA2* is required for the expression of floral repressors in *Arabidopsis thaliana*. Plant J. **41:** 376–385.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One **6:** e19379.

Enders, T.A., Oh, S., Yang, Z., Montgomery, B.L., and Strader, L.C. (2015). Genome sequencing of *Arabidopsis abp1-5* reveals second-site mutations that may affect phenotypes. Plant Cell **27:** 1820–1826.

Frontiers in Plant Science Editorial Office (2015). Retraction: Mechanosensitive channel candidate MCA2 is involved in touch-induced root responses in *Arabidopsis*. Front. Plant Sci. **6:** 153.

Gallaher, S.D., Fitz-Gibbon, S.T., Glaesener, A.G., Pellegrini, M., and Merchant, S.S. (2015). *Chlamydomonas* genome resource for laboratory strains reveals a mosaic of sequence variation, identifies true strain histories, and enables strain-specific studies. Plant Cell **27:** 2335–2352.

Gore, M.A., Chia, J.M., Elshire, R.J., Sun, Q., Ersoz, E.S., Hurwitz, B.L., Peiffer, J.A., McMullen, M.D., Grills, G.S., Ross-Ibarra, J., Ware, D.H., and Buckler, E.S. (2009). A first-generation haplotype map of maize. Science **326:** 1115–1117.

Haun, W.J., Hyten, D.L., Xu, W.W., Gerhardt, D.J., Albert, T.J., Richmond, T., Jeddeloh, J.A., Jia, G., Springer, N.M., Vance, C.P., and Stupar, R.M. (2011). The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. Plant Physiol. **155:** 645–655.

Hughes, P., Marshall, D., Reid, Y., Parkes, H., and Gelber, C. (2007). The costs of using unauthenticated, over-passaged cell lines: how much more data do we need? Biotechniques **43:** 575–586, 577–578, 581–582 passim.

Jaillon, O., et al.; French-Italian Public Consortium for Grapevine Genome Characterization (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449:** 463–467.

Jiang, C., Mithani, A., Belfield, E.J., Mott, R., Hurst, L.D., and Harberd, N.P. (2014). Environmentally responsive genome-wide accumulation of de novo *Arabidopsis thaliana* mutations and epimutations. Genome Res. **24:** 1821–1829.

Kriegel, A., et al. (2015). Job sharing in the endomembrane system: Vacuolar acidification requires the combined activity of V-ATPase and V-PPase. Plant Cell **27:** 3383–3396.

Laitinen, R.A., Schneeberger, K., Jelly, N.S., Ossowski, S., and Weigel, D. (2010). Identification of a spontaneous frame shift mutation in a nonreference *Arabidopsis* accession using whole genome sequencing. Plant Physiol. **153:** 652–654.

Lorsch, J.R., Collins, F.S., and Lippincott-Schwartz, J. (2014). Cell Biology. Fixing problems with cell lines. Science **346:** 1452–1453.

Loudet, O., Michael, T.P., Burger, B.T., Le Metté, C., Mockler, T.C., Weigel, D., and Chory, J. (2008). A zinc knuckle protein that negatively controls morning-specific growth in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **105:** 17193–17198.

Luo, Y., and Widmer, A. (2013). Herkogamy and its effects on mating patterns in *Arabidopsis thaliana*. PLoS One **8:** e57902.

O'Malley, R.C., Barragan, C.C., and Ecker, J.R. (2015). A user's guide to the Arabidopsis T-DNA insertion mutant collections. Methods Mol. Biol. **1284:** 323–342.

Ossowski, S., Schneeberger, K., Lucas-Lledó, J.I., Warthmann, N., Clark, R.M., Shaw, R.G., Weigel, D., and Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science **327:** 92–94.

Peng, P., Chan, S.W., Shah, G.A., and Jacobsen, S.E. (2006). Plant genetics: increased outcrossing in hothead mutants. Nature **443:** E8.

Platt, A., et al. (2010). The scale of population structure in *Arabidopsis thaliana*. PLoS Genet. **6:** e1000843.

Richter, R., Behringer, C., Müller, I.K., and Schwechheimer, C. (2010). The GATA-type transcription factors GNC and GNL/CGA1 repress gibberellin signaling downstream from DELLA proteins and PHYTOCHROME-INTERACTING FACTORS. Genes Dev. **24:** 2093–2104.

Romay, M.C., et al. (2013). Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. **14:** R55.

Salathia, N., Lee, H.N., Sangster, T.A., Morneau, K., Landry, C.R., Schellenberg, K., Behere, A.S., Gunderson, K.L., Cavalieri, D., Jander, G., and Queitsch, C. (2007). Indel arrays: an affordable alternative for genotyping. Plant J. **51:** 727–737.

Salomé, P.A., and Weigel, D. (2015). Plant genetic archaeology: whole-genome sequencing reveals the pedigree of a classical trisomic line. G3 (Bethesda) **5:** 253–259.

Simon, M., Simon, A., Martins, F., Botran, L., Tisné, S., Granier, F., Loudet, O., and Camilleri, C. (2012). DNA fingerprinting and new tools for fine-scale discrimination of *Arabidopsis thaliana* accessions. Plant J. **69:** 1094–1101.

Westphal, L., Scheel, D., and Rosahl, S. (2008). The *coi1-16* mutant harbors a second site mutation rendering PEN2 nonfunctional. Plant Cell **20:** 824–826.

Wu, Y., Li, Y., Liu, Y., and Xie, Q. (2015). Cautionary notes on the usage of *abi1-2* and *abi1-3* mutants of *Arabidopsis ABI1* for functional studies. Mol. Plant **8:** 335–338.

Xu, C., et al. (2015). A cascade of arabinosyltransferases controls shoot meristem size in tomato. Nat. Genet. **47:** 784–792.