the plant journal

SEB Society for Experimental Biology

*The Plant Journal* (2012) **69**, 1094–1101 doi: 10.1111/j.1365-313X.2011.04852.x

TECHNICAL ADVANCE

# DNA fingerprinting and new tools for fine-scale discrimination of *Arabidopsis thaliana* accessions

**Matthieu Simon[1,*], Adeline Simon[2], Fréderic Martins[3], Lucy Botran[1], Sébastien Tisné[1], Fabienne Granier[1], Olivier Loudet[1] and Christine Camilleri[1]**

[1]*Institut Jean-Pierre Bourgin, UMR1318 INRA-AgroParisTech, F-78000 Versailles, France,*

[2]*BIOGER, UR1290 INRA-AgroParisTech, 78850 Thiverval-Grignon, France, and*

[3]*Plateforme Génomique, UMR INRA/ENVT Laboratoire de Génétique Cellulaire, INRA, 31326 Castanet-Tolosan, France*

Received 27 September 2011; revised 8 November 2011; accepted 9 November 2011; published online 28 December 2011.
*For correspondence (fax +33(0)130833319; e-mail matthieu.simon@versailles.inra.fr).

## SUMMARY

**One of the main strengths of *Arabidopsis thaliana* as a model species is the impressive number of public resources available to the scientific community. Exploring species genetic diversity – and therefore adaptation – relies on collections of individuals from natural populations taken from diverse environments. Nevertheless, due to a few mislabeling events or genotype mixtures, some variants available in stock centers have been misidentified, causing inconsistencies and limiting the potential of genetic analyses. To improve the identification of natural accessions, we genotyped 1311 seed stocks from our Versailles Arabidopsis Stock Center and from other collections to determine their molecular profiles at 341 single nucleotide polymorphism markers. These profiles were used to compare genotypes at both the intra- and inter-accession levels. We confirmed previously described inconsistencies and revealed new ones, and suggest likely identities for accessions whose lineage had been lost. We also developed two new tools: a minimal fingerprint computation to quickly verify the identity of an accession, and an optimized marker set to assist in the identification of unknown or mixed accessions. These tools are available on a dedicated web interface called ANATool (https://www.versailles.inra.fr/ijpb/crb/anatool) that provides a simple and efficient means to verify or determine the identity of *A. thaliana* accessions in any laboratory, without the need for any specific or expensive technology.**

**Keywords: *Arabidopsis* thaliana, natural accessions, single nucleotide polymorphisms, fingerprinting, stock center.**

## INTRODUCTION

More than 10 years after its whole genome was sequenced, *Arabidopsis thaliana* is used more than ever as a model species in plant biology (Buell and Last, 2010). In particular, the study of the natural genetic variation in *A. thaliana* is important for identifying the genetic basis of complex traits, including those that contribute to adaptive variation (Bergelson and Roux, 2010). This kind of study can now be undertaken in part due to the availability of a multitude of biological resources throughout the world, and through new technical advances. Through the emergence of new technologies such as next-generation sequencing, genome-wide sequence variation of thousands of strains can now be analyzed in international collaborative projects, such as the 1001 Genomes Project (Cao *et al.*, 2011). These new genomic technologies, combined with strategies that include classical mapping populations (Simon *et al.*, 2008), genome-wide association mapping (Atwell *et al.*, 2010), or the development of promising approaches such as nested association mapping (Buckler *et al.*, 2009), will increase our understanding of the natural genetic variation present in different genetic backgrounds.

Due to the open and collaborative nature of the Arabidopsis community, publicly available resources abound, including stock centers in several countries for archiving and dispatching germplasms (Koornneef and Meinke, 2010). Although each of these stock centers offers specific resources, a large number of natural variants are found simultaneously in several collections under the same name, but which actually correspond to different batches of seeds.

This complex situation greatly increases the risk of genetic differentiation among accessions due to spontaneous mutation, segregation of residual heterozygosity or intra-population heterogeneity or genotype mixtures.

A large number of molecular variation data are already available for many accessions. After several pioneering studies (Nordborg *et al.*, 2005; Clark *et al.*, 2007; Atwell *et al.*, 2010), millions of single nucleotide polymorphisms (SNPs) are now available on a genome-wide scale. From all the variation revealed, 149 SNP markers have been selected to genotype thousands of plants (Platt *et al.*, 2010); these markers made it possible to demonstrate that, in its Eurasian range, *A. thaliana* shows a continuous isolation-by-distance pattern. This particular genetic structure has been used to identify accessions that deviate significantly from this profile and that may have been misidentified, due to incorrect location-of-origin data (Anastasio *et al.*, 2011). However, to date, none of the Arabidopsis stock centers have developed a minimum molecular pattern (fingerprint) to identify each accession as has been done for other species, such as the bacterium *Mycobacterium tuberculosis* (Filliol *et al.*, 2006) or Australian barleys (Hayden *et al.*, 2010).

With the goal of developing an identifying fingerprint, we genotyped the entire collection found at the Versailles Arabidopsis Stock Center (VASC) (http://www-ijpb.versailles.inra.fr/en/cra/cra_accueil.htm) as well as several other collections, using a set of 384 SNP markers and Illumina VeraCode technology (Lin *et al.*, 2009). We also created two new tools: (i) a fingerprint was defined for each accession and (ii) we suggest several optimized sets of markers to identify unknown variants. These small subsets of SNPs make it possible to quickly check the genotype of an accession from just a few DNA sequences. We built a dedicated web interface ANATool (https://www.versailles.inra.fr/ijpb/crb/anatool) where molecular profiles can be directly retrieved.

## RESULTS

For clarity, we defined four levels of organization for the *A. thaliana* natural accessions that are managed at the VASC (Figure 1). The terms 'population', 'natural variant', 'line'

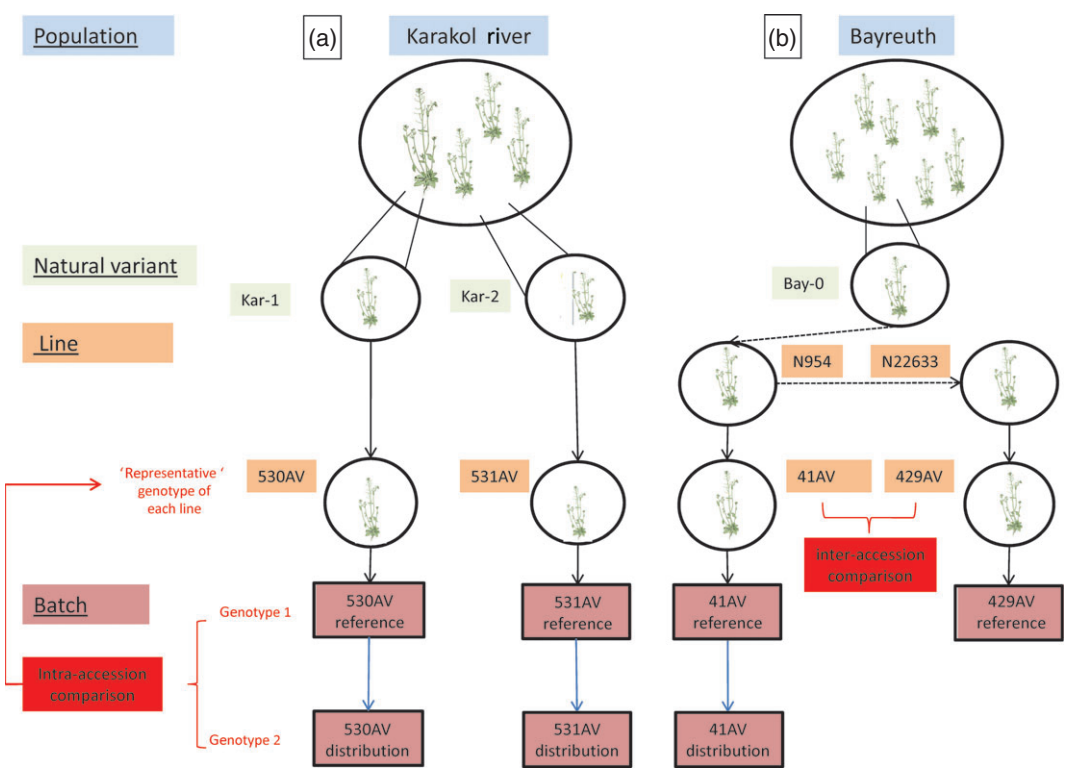


**Figure 1.** Acquisition process and management of natural accessions at the Versailles Arabidopsis Stock Center.
We distinguish between resources collected in the field by collaborators (a) and those originally obtained from a stock center (b). There are four levels of organization: the first level is the population level (in blue) and corresponds to the site where individuals were sampled in nature. The second level is the individual level; one or more individuals (in green) from the original population are assigned a name. At this level, 'natural variant' is preferred over 'ecotype' because the latter term implies detailed knowledge of the original habitat, which is rarely available for *Arabidopsis thaliana*. At the third level, variants are available from stock centers as selfing lines (in orange) identified by accession numbers. The fourth and last level is the batch level (in red) and corresponds to the progeny of one line. The 'reference' batch refers to the progeny of an original line after two consecutive single-seed descent procedures in a greenhouse and the 'distribution' batch is the bulk progeny of the reference stock. The term 'accession' is used generically and can refer to any of the four levels depending on context. The genotyping experiment was carried out at the batch level. Genotypes from all batches of each line were analyzed through an intra-accession comparison to check conformity. When a natural variant is represented by multiple lines (e.g. Bay-0), we performed an inter-accession comparison on their genotypes.

and 'batch' are used below in the meaning defined in Figure 1, otherwise, the generic term 'accession' is used.

**Genotyping results**

We obtained the genotypes for 1311 batches corresponding to 598 different lines (Table S1). From the original 384 SNP marker set (Table S2), we selected 341 markers that gave the best results in terms of scoring quality (see Experimental Procedures). These markers are evenly distributed along the chromosomes with one SNP every 350 kb on average (Figure S1).

To check the conformity of batches managed at the VASC, we compared genotypes of all reference and distribution batches for the same line (see Experimental Procedures). We found only three differences among our reference and distribution batches (Table S3-A). After removing these discrepancies from subsequent analyses, each line was represented by a single genotype: the genotype of its reference batch. Heterozygous results were then examined carefully. Two or more consecutive heterozygous markers were found for 38 lines. These results were checked by sequencing 10 loci for 11 batches and heterozygosity was confirmed in all cases.

Because 45 natural variants are found with two or more different lines (such as Bay-0, Figure 1), it was possible to investigate their conformity on an 'inter-accession' scale. We found 38 'true' duplicates that shared exactly the same genotype (Table S3-B). Nevertheless, we also revealed seven discrepancies where two lines from the same natural variant had two different genotypes (Table S3-C). For these variants, users should be particularly cautious of the line used.

Overall, the '598 data set' (598 lines genotyped with 341 markers, therefore excluding multiple batches for a given line) contained 502 different haplotypes, of which 442 are composed of single line (unique haplotypes), and 60 are found in several lines (shared haplotypes). Due to the strong population structure previously described in *A. thaliana*, most lines sharing the same genotype were expected to have very close geographical origins (Platt *et al.*, 2010). We confirmed this hypothesis in several cases, such as Dja-1 and Dja-5 (Kyrgyzstan; Table S4). Nevertheless, we revealed some inconsistencies where lines sharing the same genotype had distant geographical origins. Apart from long-distance transport of individuals in nature, this discrepancy may reflect seed-stock contaminations or mislabeling in stock centers. Many discrepancies involved standard laboratory strains (Col, L*er* and Ws) or corresponded to cases already described, such as for Tsu-0/Tu-0 or Ct-1/En-1 (Anastasio *et al.*, 2011). These well-known cases aside, we nonetheless revealed 39 new discrepancies. We also solved the question of the nature of two lines of unknown or lost origin: XXX-0 is actually a Col-0 lineage, and, interestingly, we identified accession JIC240 (from the John Innes Centre), used as the male parent in the Ts-5 × JIC240 recombinant inbred line (RIL) population (O'Neill *et al.*, 2008), as a line derived from the Mz-0 variant.

**Clustering analysis**

A clustering analysis was performed on the '598 dataset' using AWClust implemented in R (Figure 2). We think AWclust is quite suitable for this because it is based on a non-parametric analysis which does not violate common assumptions of population genetics models and linkage disequilibrium among loci for a predominantly selfing species. The gap statistic indicated an optimal cluster number of $K = 4$. Nevertheless, we also used $K = 12$ to describe these four groups in more detail. As expected, except for some outliers, accessions were well separated (at both $K$ levels) according to their geographical locations. We also examined clade membership in the core collection of 48 accessions selected to maximize its representation of the genetic diversity in *A. thaliana* (McKhann *et al.*, 2004). These core collection accessions fell into the same four clusters at $K = 4$ and in 10 of the 12 clades at $K = 12$ (i.e. all clades except the two 'artificial' Col and Ws clades). This core collection thus does indeed span the genetic structure of available *A. thaliana* resources.

**A fingerprint for each accession**

We searched for the smallest number of markers that uniquely defined each haplotype. This pattern was called a fingerprint and is composed of a combination of different SNPs for each haplotype. Using these molecular barcodes, researchers will be able to quickly verify the identity of a particular accession by sequencing just a few loci. Interestingly, all haplotypes could be fingerprinted with a combination of only two to five SNPs (Figure 3).

**Optimized SNP sets for distinguishing accessions**

We also searched for a minimal set of common markers that would allow the discrimination of as many haplotypes as possible (Figure 4). We identified sets of 4–12 markers using the single nucleotide polymorphism typing data analysis tool (SNPT; see Experimental Procedures). These sets can be used as a modular tool to identify an unknown or ambiguous batch of seeds, according to the needs of the researcher (Table S5). With four markers, we separated 598 accessions into 16 haplotypes, each composed of 23–44 accessions. With two more markers, groups included 2–19 accessions. With eight markers, 73 haplotypes could be unambiguously identified, and the remaining accessions were grouped in 142 different eight-marker haplotypes, each composed of two to nine accessions. Finally, we were able to identify 363 (nearly 70%) and 396 (nearly 80%) haplotypes with 10 and 12 markers, respectively. The remaining accessions often had very similar genotypes, but can be separated on a case-by-case basis.

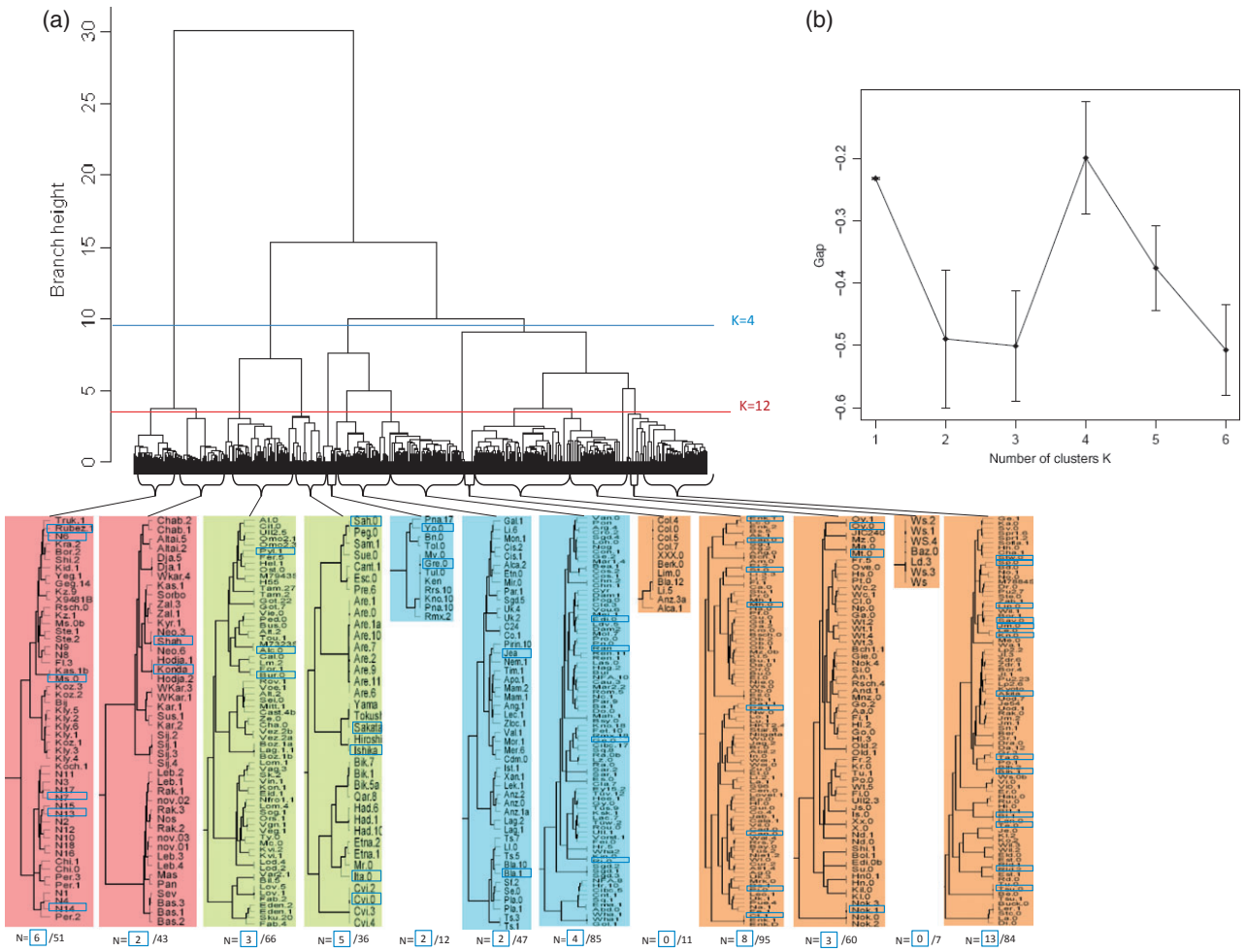


**Figure 2.** Clustering results for the 598 lines analyzed using 341 single nucleotide polymorphisms.
(a) Dendrogram obtained from the matrix of 598 lines and 341 markers. The blue line cuts across the clades for the optimal cluster number $K = 4$. The arbitrary red line cuts across the $K = 12$ clades and corresponds to a more detailed analysis of the clustering. Each of the 12 groups is shown in the insets color-coded according to the $K = 4$ clustering. Accessions from the core collection are framed in blue, and for each cluster, the total number of accessions and those belonging to the core collection are indicated.
(b) Gap statistics calculated from the complete matrix indicating an optimal cluster number of $K = 4$. This figure is also available through ANATool (https://www.versailles.inra.fr/ijpb/crb/anatool).

### The ANATool interface

The ANATool (Arabidopsis Natural Accession Tool) web interface (https://www.versailles.inra.fr/ijpb/crb/anatool) was developed to provide dynamic access to the fingerprints and genotypes of the accessions. From ANATool, four kinds of queries are available: (i) 'Accession Fingerprints' displays the smallest SNP set that will confirm a given accession; (ii) 'Find Genotypes' displays genotypes from a selection of accessions and markers; (iii) 'Find Accessions' identifies accessions corresponding to genotypes chosen from a selection of markers; (iv) 'Distinguish Between Accessions' displays the minimal marker set that will discriminate between selected accessions (maximum 50). Moreover, the set of 12 optimized SNPs with primer sequences (Table S6) is also provided in ANATool. Three typical applications of ANATool are described in Figure 5.

## DISCUSSION

We used Illumina VeraCode technology to genotype the 1311 batches available at the VASC. Our 341 SNP set included 54 of the 149 markers developed in a previous genetic structure analysis (Platt *et al.*, 2010) and used in a recent verification of the sources of misidentified accessions (Anastasio *et al.*, 2011). Therefore, it is possible to determine the links between genotypes conserved in different stock centers and detect lines that are not identical to those that have the same name in other stock centers. Although 341 SNPs suffice largely for fingerprinting purposes, they are probably dense enough to identify close relatives (i.e. two

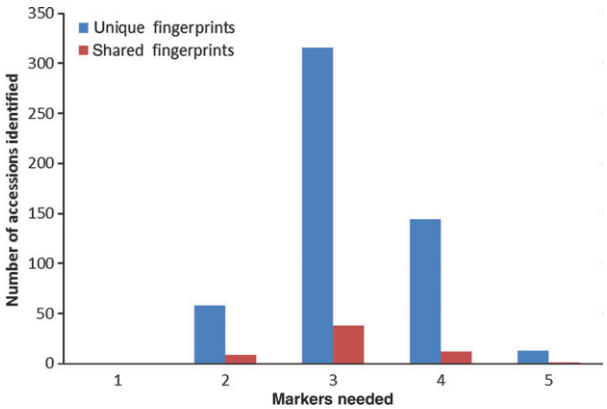


**Figure 3.** Number of markers used in fingerprints.
Determination of the smallest number of markers that uniquely define each haplotype (fingerprint). We distinguish fingerprints that identify only one line (unique fingerprint) from those that identify several lines (shared fingerprints). All lines are identified using only two to five single nucleotide polymorphisms and 90% of fingerprints identify unique lines. All minimal fingerprints are available at https://www.versailles.inra.fr/ijpb/crb/anatool.

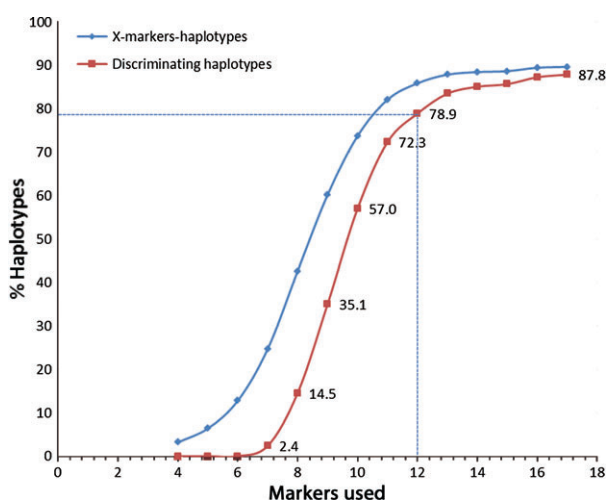


**Figure 4.** Determination of the optimized single nucleotide polymorphism (SNP) sets.
We selected 17 optimized markers using the single nucleotide polymorphism typing data analysis tool (SNPT; see Experimental Procedures) and we constructed genotypes for 14 optimized *X*-marker sets where *X* ranges from 4 to 17 SNPs. For each set we calculated (i) the number of *X*-marker haplotypes; and (ii) the number of discriminating haplotypes (*X*-marker haplotypes identifying a full-length-haplotype). We identified nearly 80% of haplotypes with the 12-marker set. To identify an unknown accession, these sets should be used as a modular tool with 4–12 markers that can progressively discriminate between up to nearly 90% of haplotypes (Table S5). The remaining haplotypes can be separated on a case-by-case basis with the help of ANATool.

individuals from the same population) as for Neo-3 and Neo-6 (Tajikistan) or Zal-1 and Zal-3 (Kyrgyzstan) that differ by only 15 and 21 SNPs, respectively. Overall, this SNP set provides very few unusable markers (see Experimental Procedures) and should prove to be useful for other genotyping applications.

The homozygous status of the genotyped accessions is one of the questions that our study could easily address. Only 38 accessions out of 598 showed from 1 to 36 heterozygous regions (up to nine consecutive markers). Among them, half are newly collected accessions and half are from stock center collections. Due to the outcrossing rates in the wild, heterozygous individuals are often found in natural populations (Le Corre, 2005; Nordborg *et al.*, 2005; Bomblies *et al.*, 2010). Although the two single-seed descent (SSD) steps performed on each newly collected accession before its integration into the VASC collection decrease the number of heterozygous loci, some residual heterozygosity found in the original plant can remain. On the other hand, accessions originating from old collections such as the European Arabidopsis Stock Centre in Nottingham, UK are expected to be highly inbred. Interestingly, all of the remaining heterozygous accessions derive from the Arabidopsis Information Service (AIS) bulk collection: they were not generated by SSD steps but on the contrary by sowing a large number of the original seeds in order to maintain variation. This could explain the residual heterozygosity that was found.

At the intra-accession level, we carried out 455 comparisons between our reference and distribution batches (see Experimental Procedures) and we revealed only three discrepancies that were removed from the VASC collection. In particular, we confirmed the identity of batches used as parents in our RIL populations relative to other batches of the same line (Table S3-D). Moreover, when these parents are found with two or more lines, they always had the same haplotype. Therefore, these resources are reliable, although individual stocks can always differ by punctual mutations (Ossowski *et al.*, 2010), some of which may possibly be functional (Loudet *et al.*, 2008).

Overall, the inter-accession analysis identified seven potential discrepancies (Table S3-C). Most of them were close relatives, as shown by clustering, e.g. lines of Ms-0, No-0, Ra-0 or Wei-0. These differences can be attributed to either differential fixation of segregating polymorphisms during a recent laboratory SSD procedure and/or recent isolation followed by genetic drift in natural populations. On the other hand, Edi-0, Mz-0 and Ws-0 lines were found to be distant from each other after clustering. Looking for genotype homologies, we revealed that the two Edi-0 lines (83AV and 453AV) belonged to different haplotypes and Edi-0 (453AV) was identical to Su-0, another UK accession. Likewise, both Mz-0 lines showed a different haplotype and Mz-0 (138AV) was identical to JIC240. Similarly, one Ws-0 line (419AV) was very similar to the German line Vi-0 (only three contrasted SNP genotypes) and we suggest that this is a putative mislabeling event. However, except for these three variants that should be used with caution, all accession duplicates were appropriate controls and confirmed the overall quality of the VASC collection.

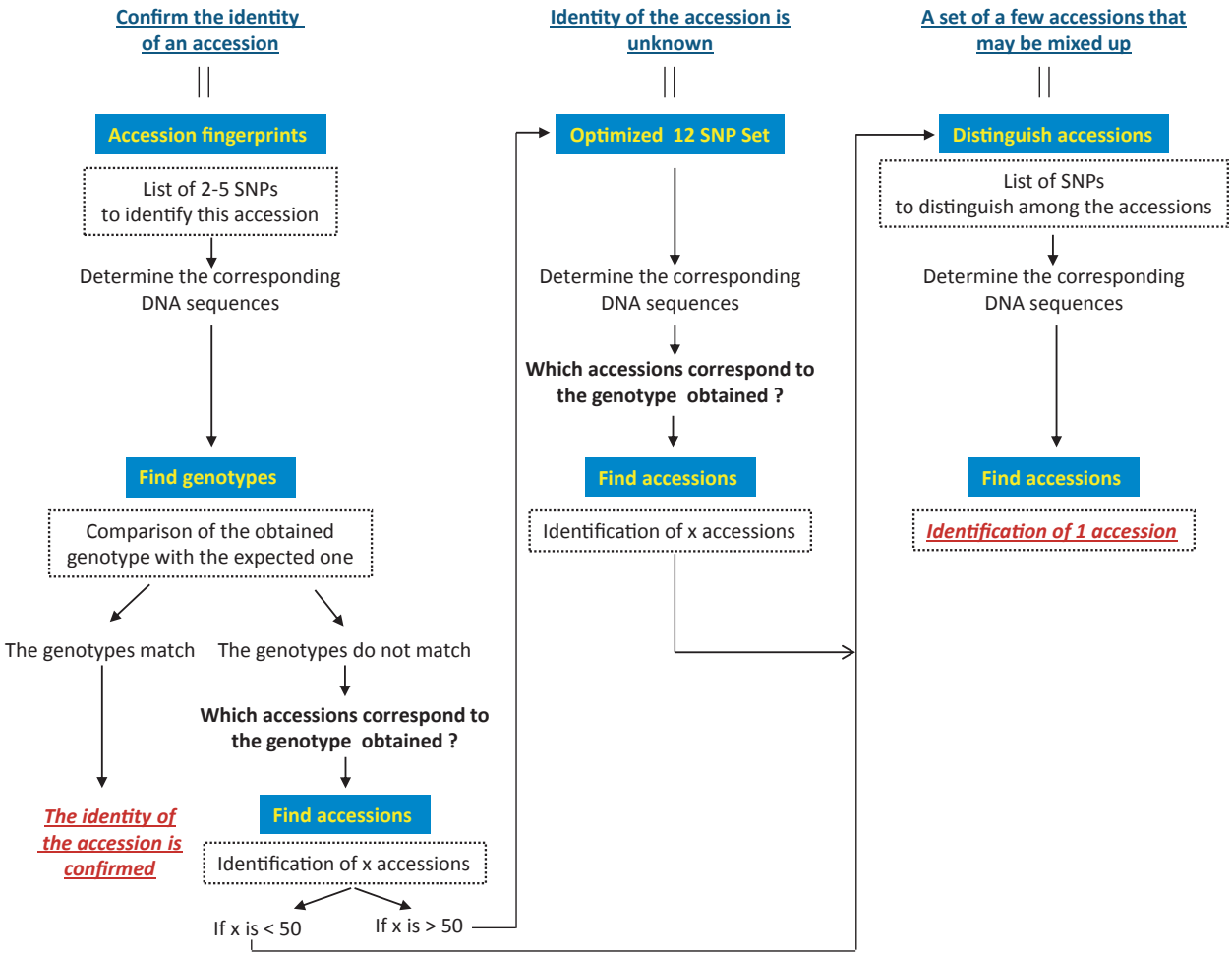


**Figure 5.** Examples of use of the ANATooL web interface for treating three common queries.
Blue boxes correspond to ANATool modules.

On investigating shared haplotypes in detail, three types of case arose (Table S4). First, we found haplotypes composed of different individuals from the same population, as for Dja-1 and Dja-5 (Kyrgyzstan). Secondly, we found haplotypes composed of at least one already red-listed accession (Anastasio *et al.*, 2011), thus confirming these discrepancies. Third, we discovered haplotypes that probably represent new misidentified accessions. These may be either new mislabeling events, as for Kl-2/Ko-2 and Is-0/Js-0, contamination from widely used standard strains (Col-0, L*er* or Ws), or other discrepancies, such as for Est-1/Rd-0 and Ca-0/Fl-1. For the 39 new discrepancies revealed, we recommend using geographical information with caution and we suggest that they be added to the red list. The emergence of misidentified accessions demonstrates the frequency with which errors or uncertainty can occur, making the availability of fast and cheap fingerprints of accessions even more important.

It would be very useful to quickly confirm the identity of accessions, not only before performing long and expensive experiments, but also for routine procedures where genotype confirmation is crucial (e.g. before performing an important cross or transgenesis, or bulking to generate a lab seed stock, or using a resource generated from one stock center's batch following information obtained from another source's batch of – theoretically – the same line). The four modules developed in ANATool are complementary and should result in a fine-scale identification of accessions (Figure 5). To confirm the identity of one accession, the appropriate markers can be determined from the 'Accession Fingerprints' module. Genomic positions of SNPs are directly available, thus facilitating primer design. After amplification and simple Sanger sequencing, the obtained genotypes can be easily compared with the expected genotypes in the 'Find Genotypes' module. Either the two genotypes are identical and the identification of the accession is confirmed, or genotypes are different and it is necessary to continue the identification process. In the latter case, the accession(s) corresponding to the genotype sequenced can be easily

determined using the 'Find Accessions' query. Depending on the number of accessions returned, the accessions can be determined in one of two ways. With fewer than 50 accessions (probably the most frequent case), the 'Distinguish Accessions' module can be used to detect discriminatory SNPs. With more than 50 accessions, the 'Optimized 12 SNP set' should be used to continue genotyping using a few DNA sequences until identification is complete.

Finally, fingerprinting will improve the quality of accession stocks by quickly revealing inconsistencies and thereby preventing their spread. However, to be truly useful, fingerprints must be readily available. This is now possible thanks to the ANATool portal, which offers the Arabidopsis community a new resource that will help ensure the integrity of *A. thaliana* accessions.

## EXPERIMENTAL PROCEDURES

### Plant material

This project was carried out on the 598 lines referenced at the VASC (Table S1). Initially, our stocks mainly originated from the European Arabidopsis Stock Centre in Nottingham, UK (http://arabidopsis.info/) or from individual collections from under-represented geographical areas. The term 'reference stock' refers to the progeny of an original line after two consecutive SSD procedures in a greenhouse, and 'distribution stock' is the bulk progeny of the reference stock (Figure 1). Each line was analyzed from all the batches available at the beginning of the study, including original batches used as male or female parents in our mapping populations (Simon *et al.*, 2008). For each stock, about 50 seeds were sterilized and sown on standard Arabidopsis MS media and grown for 15 days under a long-day photoperiod (16-h light/8-h dark) at 20°C and 65% relative humidity. Approximately 200 mg of pooled plantlets for each stock were arranged on a 96-well plate, and one metal bead was added to each well. Several controls were included in each plate to help in the assignment of genotypes: three known genotypes (natural variants Col-0, Cvi-0 and Shahdara, corresponding to lines 186AV, 166AV and 236AV, respectively), and two synthetic heterozygotes made from a 1:1 mix of two accessions (Col-0:Cvi-0 and Col-0:Shahdara). All plates were stored at −80°C before freeze-drying them. Seedlings were then ground and genomic DNA was recovered using a protocol adapted for the Qiagen DNeasy 96® Kit (http://www.qiagen.com/). DNA was diluted in 100 μl $H_2O$ and stored at −20°C until it was sent to the genotyping facility.

### Single nucleotide polymorphism set

Our 384 SNP set (Table S2) was based on the 149 SNP set from Platt *et al.* (2010) and on other publicly available SNP data (http://walnut.usc.edu/2010/2010-project) (Nordborg *et al.*, 2005; Clark *et al.*, 2007; Kim *et al.*, 2007; Warthmann *et al.*, 2007). From these sets, we preferentially selected SNPs to ensure a homogeneous distribution across the five chromosomes of *A. thaliana* and with intermediate allele frequencies. The SNP-harboring sequences were then submitted for processing by the Illumina® assay design tool (ADT). The ADT generates scores for each SNP that vary from 0 to 1. We selected only SNPs with scores above 0.6 and a high probability of being converted into a successful genotyping assay. Within this final list, 62 SNPs called 'Iplex' are common to those developed previously (Platt *et al.*, 2010) with the SEQUENOM company, USA (http://www.sequenom.com/).

### Genotyping and data cleaning

The genotyping assay was performed at the 'Genomics' facility at INRA in Toulouse, France (http://genomique.genotoul.fr/). The DNA concentrations were first normalized to 50 ng $\mu l^{-1}$. Genotyping was then carried out using the GoldenGate Genotyping Assay which interrogated the 384 SNPs simultaneously and VeraCode technology was used for the reading (Lin *et al.*, 2009). Data analysis was performed using GenomeStudio software v. 2010.2 (http://www.illumina.com/software/genomestudio_software.ilmn) following the recommendations of the INRA genomics facility. Those SNPs with a GenTrain score (quality score of clustering for one SNP) ≤0.8 were manually checked and curated. We culled 19 markers with a GenTrain score <0.5 (based on Illumina® recommendations) and 23 markers with a high failure rate (undetermined genotype for more than 100 lines), leaving a total of 341 informative markers. Each batch was then defined by the genotype defined by the alleles at 341 loci. For each line with at least two seed stocks, we checked whether all batches shared the same genotype. If 'A' is the reference allele, 'B' the alternative allele, 'U' an unknown allele and 'H' stands for heterozygous, identical genotypes are characterized for each SNP that has no A/B differences, but possible A/U, B/U, A/H, B/H and U/H differences. From nine discrepancies originally found, we extracted DNA from the same batches and sequenced them at two loci that are polymorphic in these batches. Six inconsistencies were not confirmed by sequencing, indicating a probable mistake/contamination during plate handling and/or DNA extraction steps. All confirmed suspicious batches were definitively removed from the collection and from subsequent analyses. Each line was then defined by a unique genotype, that of its reference batch. Heterozygous loci were identified as a minimum of two consecutive heterozygous SNP markers.

### Clustering analysis

We performed a non-parametric population structure analysis, since *A. thaliana* violates common assumptions of population genetics models and linkage disequilibrium among loci. Clustering of 598 lines and 341 SNPs was performed using AWClust implemented in R (Gao and Starmer, 2008). AWClust was also used to calculate gap statistics to estimate optimal cluster number (Tibshirani *et al.*, 2001; Gao and Starmer, 2008).

### Fingerprints and optimized marker sets

All scripts developed on ANATool were written in Python. The 'Accession Fingerprints' script sorts markers according to their discrimination rate relative to the target accession. Markers are added in a forward stepwise procedure in four rounds with thresholds of 60% to 10% until the target accession is fully discriminated. Similarly, the 'Distinguish Between Accessions' script sorts markers according to their allele frequencies (the more intermediate the frequency is, the more discriminating the marker is) and keeps the best one. Then, each round incorporates a new marker that better discriminates the remaining list until all accessions are fully discriminated.

Determination of the optimized SNP sets able to distinguish all accessions was performed using the SNPT software developed at Michigan State University (http://www.shigatox.net/stec/cgi-bin/snpt) (Filliol *et al.*, 2006). Because SNPT is based on haploid genomes and because ambiguities (U and H) are treated as an additional base in the original Perl program, we ran the program on a modified matrix: we culled genotypes and loci with heterozygous markers and low failure rates, resulting in a final matrix of 226 markers and 484 lines. As no marker set allowed complete

discrimination, we choose the first 17 markers according to their SNPT output rank and we constructed genotypes with *X* markers (where *X* ranges from 4 to 17 markers). For each marker set, we calculated the total number of *X*-marker haplotypes and the number of *X*-marker haplotypes identifying a full length haplotype (i.e. discriminating haplotypes).

## ACKNOWLEDGEMENTS

We thank Jacqueline Babilliot, Béatrice Bouhedi, Liliane Laroche, Bernadette Trouvé and Christine Sallé for their excellent technical assistance. This work was supported by French IBiSA grants to Christine Camilleri and Matthieu Simon.

## SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

**Figure S1**. Physical map showing the genomic positions of the 341 validated single nucleotide polymorphisms (SNPs). The SNPs from the optimized 12 SNP set are in red. This figure is also available through ANATool (https://www.versailles.inra.fr/ijpb/crb/anatool).

**Table S1**. List of the 598 lines genotyped including passport information and membership in other resources (core collection, Versailles recombinant inbred line populations).

**Table S2**. List of the 384 single nucleotide polymorphisms (SNPs) used. This table is also available through ANATool (https://www.versailles.inra.fr/ijpb/crb/anatool).

**Table S3**. Genotype confirmation and discrepancies revealed after analysis of 341 single nucleotide polymorphisms.

**Table S4**. List of lines with the same haplotype at 341 single nucleotide polymorphism markers (shared haplotypes).

**Table S5**. Identification of haplotypes with the optimized marker sets designed using the single nucleotide polymorphism typing data analysis tool SNPT. Each set is composed of *X* markers (where *X* ranges from 4 to 12 markers). Accession(s) with the same genotype are separated by a black horizontal line. Accessions identified by a discriminating *X*-marker haplotype are in red.

**Table S6**. Primer sequences for the optimized set of 12 single nucleotide polymorphisms.

Please note: As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missingfiles) should be addressed to the authors.

## REFERENCES

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

**Atwell, S., Huang, Y.S., Vilhjalmsson, B.J. *et al.*** (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, **465**, 627–631.

**Bergelson, J. and Roux, F.** (2010) Towards identifying genes underlying ecologically relevant traits in *Arabidopsis thaliana*. *Nat Rev Genet*, **11**, 867–879.

**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 Genetics*, **6**, e1000890.

**Buckler, E.S., Holland, J.B., Bradbury, P.J. *et al.*** (2009) The genetic architecture of maize flowering time. *Science*, **325**, 714–718.

**Buell, C.R. and Last, R.L.** (2010) Twenty-first century plant biology: impacts of the Arabidopsis genome on plant biology and agriculture. *Plant Physiol*, **154**, 497–500.

**Cao, J., Schneeberger, K., Ossowski, S. *et al.*** (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genet*, **43**, 956–963.

**Clark, R.M., Schweikert, G., Toomajian, C. *et al.*** (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science*, **317**, 338–342.

**Filliol, I., Motiwala, A.S., Cavatore, M. *et al.*** (2006) Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: Insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. *J. Bacteriol*, **188**, 759–772.

**Gao, X. and Starmer, J.D.** (2008) AWclust: point-and-click software for non-parametric population structure analysis. *BMC Bioinformatics*, **9**, 77.

**Hayden, M.J., Tabone, T.L., Nguyen, T.M., Coventry, S., Keiper, F.J., Fox, R.L., Chalmers, K.J., Mather, D.E. and Eglinton, J.K.** (2010) An informative set of SNP markers for molecular characterisation of Australian barley germplasm. *Crop & Pasture Science*, **61**, 70–83.

**Kim, S., Plagnol, V., Hu, T.T., Toomajian, C., Clark, R.M., Ossowski, S., Ecker, J.R., Weigel, D. and Nordborg, M.** (2007) Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.*, **39**, 1151–1155.

**Koornneef, M. and Meinke, D.** (2010) The development of Arabidopsis as a model plant. *Plant J.* **61**, 909–921.

**Le Corre, V.** (2005) Variation at two flowering time genes within and among populations of *Arabidopsis thaliana*: comparison with markers and traits. *Mol. Ecol.* **14**, 4181–4192.

**Lin, C.H., Yeakley, J.M., McDaniel, T.K. and Shen, R.** (2009) Medium- to high-throughput SNP genotyping using VeraCode microbeads. *Methods Mol. Biol.* **496**, 129–142.

**Loudet, O., Michael, T.P., Burger, B.T., Le Mette, 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.

**McKhann, H.I., Camilleri, C., Berard, A., Bataillon, T., David, J.L., Reboud, X., Le Corre, V., Caloustian, C., Gut, I.G. and Brunel, D.** (2004) Nested core collections maximizing genetic diversity in *Arabidopsis thaliana*. *Plant J.* **38**, 193–202.

**Nordborg, M., Hu, T.T., Ishino, Y. *et al.*** (2005) The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol*, **3**, 1289–1299.

**O'Neill, C.M., Morgan, C., Kirby, J. *et al.*** (2008) Six new recombinant inbred populations for the study of quantitative traits in *Arabidopsis thaliana*. *Theor. Appl. Genet.*, **116**, 623–634.

**Ossowski, S., Schneeberger, K., Lucas-Lledo, 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.

**Platt, A., Horton, M., Huang, Y.S. *et al.*** (2010) The scale of population structure in *Arabidopsis thaliana*. *PLoS Genetics*, **6**, e1000843.

**Simon, M., Loudet, O., Durand, S., Berard, A., Brunel, D., Sennesal, F.X., Durand-Tardif, M., Pelletier, G. and Camilleri, C.** (2008) Quantitative trait loci mapping in five new large recombinant inbred line populations of *Arabidopsis thaliana* genotyped with consensus single-nucleotide polymorphism markers. *Genetics*, **178**, 2253–2264.

**Tibshirani, R., Walther, G. and Hastie, T.** (2001) Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. series B (Stat Methodol)*, **63**, 411–423.

**Warthmann, N., Fitz, J. and Weigel, D.** (2007) MSQT for choosing SNP assays from multiple DNA alignments. *Bioinformatics*, **23**, 2784–2787.

 