doi: 10.1111/j.1365-313X.2004.02034.x

TECHNICAL ADVANCE

Nested core collections maximizing genetic diversity in *Arabidopsis thaliana*

Heather I. McKhann[1,2,*], Christine Camilleri[1,2], Aurélie Bérard[1], Thomas Bataillon[3], Jacques L. David[3], Xavier Reboud[4], Valérie Le Corre[4], Christophe Caloustian[1], Ivo G. Gut[1] and Dominique Brunel[1,2]

[1]Centre National de Génotypage, 2 rue Gaston Crémieux, 91057 Evry Cedex, France,
[2]Station de Génétique et Amélioration des Plantes, Route de Saint-Cyr, Institut National de la Recherche Agronomique (INRA), 78026 Versailles Cedex, France,
[3]UMR 1097 Diversité et Génomes des Plantes Cultivées, INRA, Domaine de Melgueil, 34130 Mauguio, France, and
[4]UMR Biologie et Gestion des Adventices, INRA, B.P. 86510, 21065 Dijon Cedex, France

Received 10 October 2003; revised 18 December 2003; accepted 6 January 2004.
*For correspondence (fax +33 1 60 87 83 83; e-mail mckhann@cng.fr).

Summary

The successful exploitation of natural genetic diversity requires a basic knowledge of the extent of the variation present in a species. To study natural variation in *Arabidopsis thaliana*, we defined nested core collections maximizing the diversity present among a worldwide set of 265 accessions. The core collections were generated based on DNA sequence data from a limited number of fragments evenly distributed in the genome and were shown to successfully capture the molecular diversity in other loci as well as the morphological diversity. The core collections are available to the scientific community and thus provide an important resource for the study of genetic variation and its functional consequences in *Arabidopsis*. Moreover, this strategy can be used in other species to provide a rational framework for undertaking diversity surveys, including single nucleotide polymorphism (SNP) discovery and phenotyping, allowing the utilization of genetic variation for the study of complex traits.

Keywords: natural genetic diversity, SNPs, core collection, association studies, *Arabidopsis thaliana*, linkage disequilibrium.

Introduction

Naturally occurring variation within a species can be exploited to identify quantitative trait loci (QTLs), to perform genotype/phenotype association studies and to explore the ecological and evolutionary forces shaping currently observable genetic diversity. In plants, the wild crucifer *Arabidopsis thaliana* is an ideal candidate for such studies as it is well suited for molecular biological, and genetic analyses (Alonso-Blanco and Koornneef, 2000) and the whole genomic sequence of a reference accession is known in this species (*Arabidopsis* Genome Initiative, 2000). Until now, most natural variation studies in *Arabidopsis* have focused on about 20–40 accessions (ecotypes) for one or two genes (e.g. Aguade, 2001; Caicedo *et al*., 1999; Kawabe *et al*., 2000; Kuittinen and Aguade, 2000; Purugganan and Suddith, 1999; Savolainen *et al*., 2000; Schmid *et al*., 2003) or a particular morphological character (Li *et al*., 1998). Other works have used RFLP (Bergelson *et al*., 1998; Ullrich *et al*., 1997), microsatellite (Innan *et al*., 1997; Loridon *et al*., 1998; Todokoro *et al*., 1995), amplified fragment length polymorphism (AFLP) (Breyne *et al*., 1999; Erschadi *et al*., 2000; Miyashita *et al*., 1999) or recently, single nucleotide polymorphism (SNP) markers (Schmid *et al*., 2003; Törjék *et al*., 2003) on a comparable number of accessions, with the exception of one more exhaustive AFLP study of 142 accessions (Sharbel *et al*., 2000) and the development of genome-wide SNP markers on 96 accessions (http://walnut.usc.edu/2010.html). The accessions used for these studies were not chosen to cover the genetic diversity of the species as this information was not available. Further, each study surveyed different accessions, using different markers, making cross-study comparisons difficult. Clearly, a small reference group of accessions demonstrated to be highly polymorphic was needed to best exploit the *A. thaliana* natural variation, not only for molecular studies but also for the laborious task of phenotyping. This is particularly relevant as there are increasing numbers of studies aiming to characterize complex traits.

We therefore aimed to generate in *Arabidopsis* a core collection of accessions, i.e. a subset of a larger germplasm collection that contains the maximum possible genetic diversity of the species with a minimum of repetitiveness (Frankel, 1984). Such a sample of reasonable size could be intensely studied within the scientific community.

Results

Survey of genetic diversity

Single nucleotide polymorphisms offer a means of characterizing the range of DNA variation on a genomic scale. In order to build up the core collection, we performed a polymorphism survey by sequencing in a worldwide collection of accessions 10 fragments of genes of diverse functions including exons and introns, distributed throughout the genome. Two such fragments, hereafter markers, were chosen on each of the five *A. thaliana* chromosomes. Of the roughly 300 *A. thaliana* accessions available at the time from the stock centers, a collection of 265 was available in Versailles. From this collection, 95 accessions were chosen that covered the range of known ecological and geographical habitats and which eliminated apparent redundancies (e.g. one Columbia (Col-0) accession was chosen of the available four). The 10 markers were then sequenced on the set of 95. For 4 of these 10 markers, the entire collection of 265 accessions was sequenced to verify that significant variation had not been missed in the initial choice of 95 accessions. Indeed, we observed nearly no additional polymorphism as compared to the set of 95 accessions. No additional SNPs were found for one marker, and one, four, and four additional rare SNPs, mostly singletons (polymorphisms only present in a single accession), were found for the other three markers.

The polymorphisms found, all referred to as SNPs (Table 1), are primarily single-nucleotide changes, but also include small insertions and deletions (7.5% of the polymorphic sites). The level of variation greatly depends on the locus considered. In most cases, two different nucleotides were present at one polymorphic site, but 2.5% of the SNPs were found to have three variants. Thus, the 197 polymorphisms identified gave rise to 399 variants. Twenty-one per cent of the SNPs were only found in a single accession. Of 78 polymorphisms present in exons, 35 (45%) lead to an amino acid change. Variants present at

Table 1 Markers sequenced and their nucleotide diversity

Marker name	Gene	Chromosome position (Mb)[b]	Sequence length (bp)	No. of polymorphic sites (SNPs)	No. of haplotypes	No. of associated SNPs	SNPs/kbp in exons	SNPs/kbp in introns	Watterson θ (per site)	Nucleotide diversity π	Tajima's D	P-value of $D = 0$
125C F	Acetyl CoA synthetase	1–7.1	650	9	8	–	12	34	0.00272	0.00082	−1.76	<0.1
133C R[a]	Geranylgeranyl reductase	1–27.6	660	25	10	17	21	77	0.00629	0.00870	1.04	>0.1
49A R	Cycloartenol synthase	2–2.9	630	49	16	14	61	95	0.01550	0.00662	−1.85	<0.05
67D R[a]	Auxin responsive protein	2–19.5	630	9	9	–	14	15	0.00240	0.00180	−0.58	>0.1
92B R[a]	Long-chain fatty acid CoA ligase	3–5.4	650	14	14	–	10	26	0.00457	0.00260	−1.11	>0.1
100D F	Chorismate mutase precursor	3–11.1	659	12	10	–	10	25	0.00378	0.00100	−1.99	<0.05
76D F[a]	Methionyl tRNA synthetase like	4–6.9	517	21	19	–	27	67	0.00799	0.00439	−1.23	>0.1
82A F	Enoyl CoA hydratase	4–14.3	592	7	9	–	12	12	0.00233	0.00086	−1.5	>0.1
115A R	Peptidase	5–9.4	610	20	10	11	16	59	0.00640	0.00861	1.01	>0.1
DFR	Dihydroflavonol reductase	5–16.8	650	31	21	7	27	105	0.00930	0.01196	0.87	>0.1
Total			6248	197								
Average					13		21	51				

The 10 markers were sequenced on a common worldwide set of 95 accessions. Four of them (indicated by [a]) were sequenced on the whole collection of 265 accessions with only few additional polymorphisms found. The results displayed are from the 95 accessions for every marker. For the four loci with two highly differentiated groups of haplotypes, the number of associated SNPs is indicated.

[b] Chromosome position determined with SEQVIEWER available at the TAIR home page: http://www.arabidopsis.org

polymorphic sites within one fragment are often correlated, defining a limited number of haplotypes (see the distribution of linkage disequilibrium (LD) within fragments in Figure S1). The haplotype structure and the patterns of LD are also greatly dependent on the fragment considered, with several fragments exhibiting clear footprints of recombination events. For four loci, two highly differentiated sets of haplotypes are found, but such dimorphism is not the rule, as already discussed by Aguade (2001). Among the 95 accessions studied, four pairs of accessions, two groups of three accessions, and one group of four accessions were found to be identical for the 197 observed polymorphisms. It is possible that, in some cases, there has been a duplication of the most commonly used accessions (Col-0, Wassilewskija (WS), Landsberg *erecta* (L*er*)).

Values of nucleotide diversity as estimated by Watterson's θ (Nei, 1987) are of comparable magnitude to estimates obtained in other surveys of sequence diversity in *A. thaliana*. The average pair-wise nucleotide diversity π varies from about 1 to 10 SNP per kilobase according to the locus considered. Tajima's *D* statistics are overall negative, which is consistent with recent studies (Innan and Stephan, 2000). This reflects the excess of rare polymorphisms in most loci sequenced here relative to what is expected in a stable non-subdivided population at mutation drift equilibrium (Tajima, 1989). We do not find widespread evidence for dimorphism, which has been sometimes interpreted as the footprint of balancing selection (Hanfstingl *et al*., 1994). A number of non-exclusive selective and demographic scenarios can explain these results. Overall, this suggests that recent demographic history (probably a rapid demographic expansion) is shaping much of the currently observable sequence variation. Future detailed studies examining the joint patterns of detected polymorphism will allow the identification of loci that may have been affected by events such as recent selective sweeps (Galtier *et al*., 2000) while accounting for the complex demographic history of *Arabidopsis*.

Generation of the core collections

The SNPs detected were used to generate core collections using the MSTRAT software (Gouesnard *et al*., 2001). It has been shown (Bataillon *et al*., 1996) that the maximization (M) strategy is expected to perform particularly well when accessions are from populations with restricted gene flow or when accessions are primarily selfing, as is the case of *Arabidopsis*. Figure 1 shows that, in our case, the sampling efficiency of the M strategy is always superior to a random strategy and that the relative efficiency is highest for small-size samples.

In addition, cross-validation using different subsets of the 10 markers (see Experimental procedures) demonstrated that despite the small number of predictor markers used, MSTRAT outperformed a random sample by 10% in the range of sample sizes of 15–25 (see Figure S2).

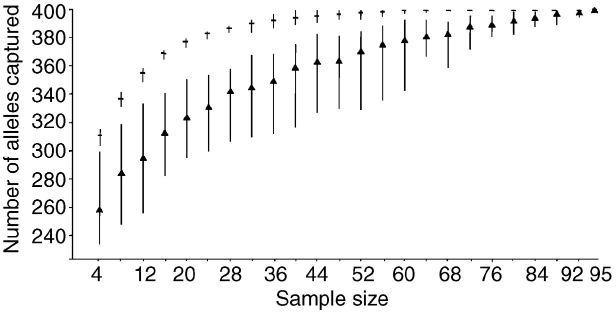

Figure 1. Sampling efficiency (the ability to capture genetic diversity) of the M strategy (upper curve) and a random strategy (lower curve) in increasingly large samples up to 95 accessions.
Vertical bars indicate the interval between the minimum and maximum values obtained over 50 independent replicates for each method. Dashes and triangles represent the mean values of the allelic richness present in a given size subset obtained by each method.

Based on the sequence data from the 10 markers, nested core collections were generated in which eight accessions were added incrementally up to a total of 48 (Table 2). It was found that 24–32 accessions are sufficient to capture almost all of the DNA sequence diversity of the markers present in the initial worldwide collection.

Capture of unknown diversity

The ability of these core collections to capture unknown sequence diversity was then assessed in a set of four additional loci different from the markers used to build the cores and compared to random samples of the same size (Figure 2). For CRT-binding factor (CBF)3 and LUMINIDEPENDENS, the core collections perform better than most of 1000 random choices of the same number of accessions. For DRE-binding protein (DREB)2A, even if the capture is a little less efficient, the core collections also do better than most of the random samples in every case. For CBF2, for which the efficiency is the lowest, the capture of diversity by the core collections of 24–40 accessions is nonetheless superior to the capture in the majority of the 1000 random same-size samples. Moreover, the average frequency of capture of singletons for the four genes is higher than predicted by chance alone in every size core collection, with the highest relative efficiency for smaller size cores (data not shown). The absolute number of variants captured at these four loci in the nested core collections is shown in Table S4. Given the nature of core collections, it is impossible to guarantee the complete capture of all alleles for each gene. However, on average, the core collections eliminate the redundancy in the worldwide collection and yet succeed in capturing the majority of diversity in a set of genes of interest.

Table 2 Nested core collections of 8 to 48 accessions

					Alleles	
Core size	No. of stock or reference	Versailles number	Name	Country	nb	%
8	N1094	162AV	Ct-1	ITA		
8	N1436	224AV	Oy-0	NOR		
8	N929	236AV	Shakdara	TJK		
8	N1030	180AV	Blh-1	CSK		
8	N1028	172AV	Bur-0	IRL		
8	1, 2	25AV	JEA	FRA		
8	N902	166AV	Cvi-0	CPV		
8	N1244	157AV	Ita-0	MAR	337	84
16	N1186	101AV	Ge-0	CHE		
16	N1656	178AV	Alc-0	ESP		
16	N1534	62AV	St-0	SWE		
16	N1064	163AV	Can-0	ESP		
16	1, 2	8AV	PYL-1	FRA		
16	N22491	266AV	N13 Konchezero	RUS		
16	N1380	94AV	Mt-0	LBY		
16	N1368	215AV	Mh-1	POL	367	92
24	3	257AV	Sakata	JPN		
24	N1210	200AV	Gre-0	USA		
24	N1122	83AV	Edi-0	UK[a]		
24	N1286	70AV	Kn-0	LIT		
24	N1564	91AV	Tsu-0	JPN		
24	3	252AV	Akita	JPN		
24	N968	42AV	Bl-1	ITA[a]		
24	N1538	92AV	Stw-0	RUS	384	96
32	N905	93AV	Ms-0	RUS		
32	N970	76AV	Bla-1	ESP		
32	N1641	229AV	Rld-2	RUS		
32	N1500	233AV	Sah-0	ESP		
32	N1548	56AV	Ta-0	CSK		
32	N927	231AV	Rubezhnoe-1	UKR		
32	N22492	267AV	Sampo Mountain	RUS		
32	3	253AV	Ishikawa	JPN	392	98
40	N1550	68AV	Te-0	FIN		
40	N1492	160AV	Ri-0	CAN		
40	N916	190AV	Kondara	TJK		
40	N1258	206AV	Jm-0	CSK		
40	N1622	250AV	Yo-0	USA		
40	N1506	234AV	Sap-0	CSK[b]		
40	N1514	235AV	Sav-0	CSK[b]		
40	N1530	53AV	Sp-0	DEU	398	100
48	N1438	50AV	Pa-1	ITA		
48	N1454	40AV	Pi-0	AUT		
48	N1400	95AV	Nok-1	NLD		
48	N22484	262AV	N6 Karelian	RUS[a]		
48	N921	197AV	En-T	DEU		
48	N22485	263AV	N7 Pinguba	RUS		
48	1	21AV	RAN	FRA		
48	N1336	63AV	Lip-0	POL	399	100

Alleles captured: number of alleles present in the initial worldwide collection at the ten sequenced loci (399 in total) that were captured in the different size nested core collections.

[a] Botanic garden.

[b] Plant breeding station.

All accessions are available from the NASC (http://nasc.nott.ac.uk; indicated by N in stock center number). The accessions are also available from the ABRC. 'Versailles number' refers to accessions that have undergone a single seed descent step at the INRA of Versailles. Such seeds are available upon request (Matthieu Simon: msimon@versailles.inra.fr) and will be deposited as a set to the stock centers.

1: Le Corre *et al.* (2002).

2: Lavigne *et al.* (2001).

3: Todokoro *et al.* (1995).

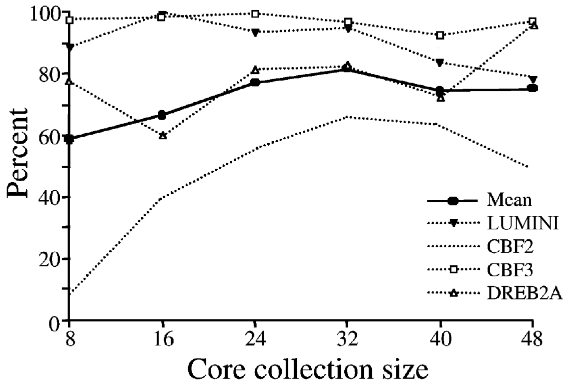

Figure 2. Comparison of the M strategy core collections to random samples of equal size for the capture of SNP variants present in 95 accessions at four loci (CBF2, CBF3, DREB2A, and LUMINIDEPENDENS (LUMINI)).
The lines represent the percentage of 1000 random samples that have a score below or equal to that of the same size core collection.

Capture of morphological diversity

An independent verification of the core collection sampling efficiency was performed using morphological data, by comparing the phenotypic variability found in each core collection with the variability of same-size subsets taken at random among the worldwide collection. Similar to the approach of Ungerer *et al.* (2002), morphological characters with high heritability were examined. Of 23 characters, 17 were retained, which mainly describe plant size and architecture as well as reproductive capacity.

The distribution of morphological characters encompassed by the core collections was initially assessed using a principal component analysis. Figure 3 shows that not only does the core collection of 24 accessions span the range of all the accessions but it also shows an even distribution along each of the two main axes. The ability of the M strategy to comprehensively sample the available morphological variability was then assessed using first, the number of Mahalanobis classes represented in each core collection, and second, the evenness of the distribution. Figure 4a shows that for core sizes up to 24, the M method based on molecular data is more efficient than random sampling in capturing the phenotypic variation observed. Furthermore, as shown in Figure 4b, the morphological variability is more evenly distributed among accessions from the core collections than among accessions chosen randomly. The reduced size of the core collections thus allows the capture of most of the variation while eliminating a number of similar phenotypes from the most represented classes. The highest morphological diversity is found in the core collection of size 24, additional accessions within cores of higher sizes adding, on average, more redundancy than novelty.

Similar analyses were also conducted trait by trait, and revealed that the most efficiently captured variability was that of traits associated with flowering time (e.g. number of rosette leaves at flowering) and of other traits also found to have a high genetic heritability (number of bracts at flowering, main flowering stem diameter). Finally, the conservation of phenotypic associations between traits, which may be under genetic control, was investigated by computing the correlation between the phenotypic variance–covariance matrices in the worldwide collection and in each core collection. Correlation values were not significantly different between core collections and random samples, indicating that no bias was associated with any of the core collections, which overall retained the pattern of co-variation among traits.

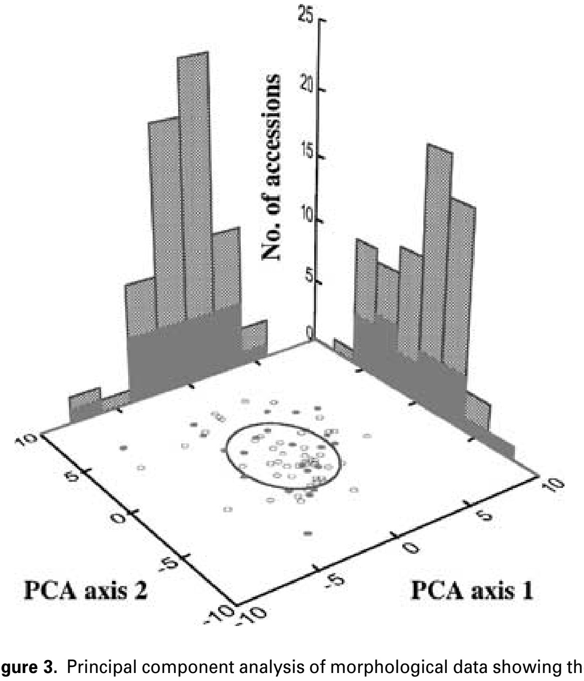

Figure 3. Principal component analysis of morphological data showing the core collection of 24 (black circles) within the distribution of the accessions (open circles).
Projection along the primary and secondary axes shows the evenness of the distribution of the 24 accessions (in black). Only considering the component loadings for variables higher than 0.5, axis 1 is positively correlated with the flowering time, main stem diameter, number of leaves on the rosette at flowering, and number of cauline leaves, and axis 2 is negatively correlated with time to mature siliques; axis 1 and axis 2 tend to have opposite contributions according to treatment (axis 1 for the vernalized and axis 2 for the non-vernalized treatment, respectively) on the following variables: maximum plant height, height from soil to the first silique at maturity, number of primary and secondary branches, number of flowering heads, and total number of siliques produced.

Discussion

Our approach is successful in capturing most of the DNA diversity at sequenced loci in a small set of individuals and permits an enhanced capture of diversity throughout the genome, allowing in turn the capture of the phenotypic variation of traits that are not controlled *a priori* by the genes used to sample the core collection. The efficiency of

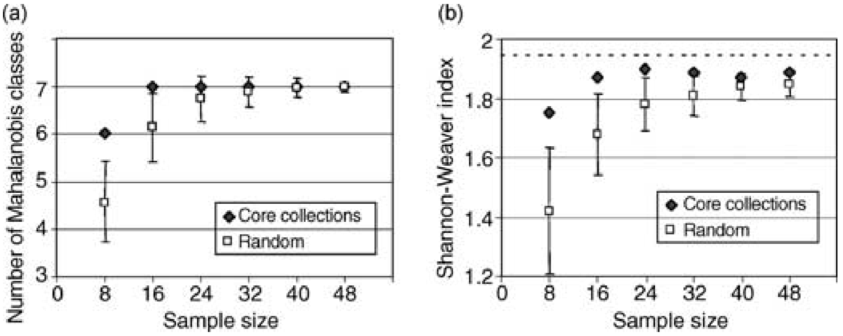

Figure 4. Morphological diversity captured by the nested core collections of 8–48 accessions as compared to random samples with the same sizes.
(a) Range of morphological variation captured, as indicated by the number of Mahalanobis distance classes represented within samples, where distances are calculated between each accession and the mean for a worldwide collection. A score of 7 indicates that the whole range of available variation is covered.
(b) Morphological diversity captured within samples, as measured by the Shannon–Weaver index. The maximum theoretical value of the Shannon–Weaver index, corresponding to all phenotypic classes being equally represented, is 1.946 (horizontal dashed line).
Diamonds are observed values for the nested core collections. Squares are mean values over 2000 random samples. Vertical bars indicate SD of the 2000 random values.

the marker-assisted sampling strategy used relies on the existence of statistical correlations between alleles segregating at different loci in the genome, i.e. on LD. These correlations can affect physically linked as well as unlinked loci, as genetic differentiation between populations comprising our collection creates some LD at the level of the whole genome irrespective of the physical distance between loci. Locally (within population), LD is generated by selection, drift, and founder effects and is eroded by recombination, and mostly concerns linked loci. Although the efficiency of recombination is greatly reduced in highly selfing species such as *A. thaliana*, suggesting that LD between linked loci could be maintained over greater distances than in outcrossing species (Nordborg, 2000), recent surveys of LD within small genomic regions suggest a decay of LD over 250 kbp depending on local rates of recombination and gene conversion (Haubold *et al.*, 2002; Nordborg *et al.*, 2002). If this is the case, the M strategy should only have been efficient within a window of about 250 kbp around the sequenced loci to capture diversity. Here, we found some linkage disequilibrium between SNPs located on independent fragments (see Figure S1). Indeed, although we surveyed a minute portion of the genome to guide the sampling of the core collection, the global level of LD in the collection could explain the gain in diversity captured at the four loci sequenced for validation and at unknown loci governing the morphological variation. In that respect, even if LD components are difficult to disentangle in samples coming from worldwide collections, LD depends here on differences in the past evolutionary history of migration and isolation of the populations when accessions from the global geographical range of the species are taken together as a whole. In *Arabidopsis*, as well as in other strong selfing species, identifying population genetic boundaries is challenging as neighboring locations can be as much differentiated as much more distant locations. Here, maximizing allelic variation at a set of unlinked markers ensures that most divergent ‘populations/units’ (whose boundaries are generally not known) represented in the initial collection are included in the core collection. Thus, this strategy can be efficient even using a reduced number of markers scattered through the genome.

The use of these nested core collections allows flexibility in experimentation, making it possible to rationally choose both the accessions to analyze and the size of the working collection. Although it may be possible to sequence large numbers of accessions, the ability to determine phenotypes is often severely limiting and the effort and cost of large-scale experiments are often prohibitive. In this case, a small-size core collection is preferable. The core collection of 24 accessions captured the majority of all the SNPs in the genes examined as well as maximized the morphological diversity, and thus represents the optimal size core collection for most applications. On the other hand, if one wishes to find the maximum of rare alleles, the core collection of 48 is considered the best choice for SNP detection. It should be noted that, when using such core collections, caution must be exercised in making inferences about the evolutionary history of *A. thaliana* from contemporary levels of naturally occurring variation. In particular, statistics related to the spectrum of nucleotide frequencies (such as Tajima’s *D*) are likely to be skewed because of the highly non-random sampling inherent in building such collections and standard tests for the neutrality of polymorphism will likely not apply.

No reference accession was included in the core; rather anyone can add their own reference(s), for example L*er* or Col-0 for which the genomic sequence is available, or accessions of particular interest in each experiment. Until now, the only study of *A. thaliana* natural genetic variation in which a large number of accessions was examined (Sharbel *et al.*, 2000) provided a geographical context for choosing genetically distant accessions. Here we show that, although the diversity of the original worldwide collection is significant, nearly all this variation can be represented in a small subset of accessions of the initial collection, pointing out its high redundancy. Not surprisingly, the core collection of eight includes accessions collected at the edge of the natural distribution of the species, such as Cvi-0, Ita-0, or Shakdara. Some of these accessions have been shown to have highly diverged alleles in some genes, and were suggested to be distantly related to other accessions (Schmid *et al.*, 2003; Sharbel *et al.*, 2000).

To date, large-scale efforts for SNP discovery have been carried out mostly in humans (Sachidanandam *et al.*, 2001) and the mouse. The sampling of human diversity until now has mostly relied on the knowledge of the history of ethnic groups by including individuals representing each 'area' of the human gene pool and not on molecular data. The method we have employed could be applied fruitfully to humans and other species. Specifically, this approach consists of: (i) surveying a worldwide collection of individuals for a limited number of genomic fragments (loci) evenly distributed throughout the genome; and (ii) using the polymorphisms detected to guide the sampling of a core set of individuals capturing the bulk of the variation present in the initial large set.

The core collection strategy provides a rational framework for undertaking studies on genetic diversity, SNP discovery, and the phenotyping of traits in the absence of prior knowledge on the distribution of a character or the polymorphism in a target gene. This is particularly pertinent given the increasing interest in natural genetic variation and its exploitation in studying complex traits, from human diseases to quality in crop plants. This approach opens the possibility of performing large-scale association studies to study such complex traits. Once SNP detection has been performed on the accessions of the core collection, the entire collection can be genotyped for association studies. The utility of genotype/phenotype association studies has been documented in humans (Hugot *et al.*, 2001; Ogura *et al.*, 2001; Roses, 1997), and is emerging in plants (Maloof *et al.*, 2001; Thornsberry *et al.*, 2001). By maximizing the diversity studied in a reduced number of individuals through the use of core collections, the probability of identifying variants of interest for association studies involving complex traits is increased. Furthermore, the knowledge gained through the core collections allows the choice of optimal crosses for generating QTL mapping populations. To date, 17 of the 48 core accessions have been used as parents with Col-0 as male in a set of recombinant inbred lines, generated at the INRA of Versailles. These collections thus constitute a chance to foster an international coordination on *Arabidopsis* diversity studies. Currently, projects involving a number of European countries are underway that utilize the core collections and the derived recombinant inbred lines to study numerous diverse aspects of *Arabidopsis* development and architecture, metabolism, and abiotic and biotic interactions. Such characters, being polygenic, represent ideal subjects for the utilization of natural diversity, either by means of QTL identification or association studies. To facilitate additional research, seeds of the core collection accessions that have been subjected to single-seed descent are available upon request and will be deposited to the Nottingham Arabidopsis Stock Center (NASC) and Arabidopsis Biological Resource Center (ABRC).

Experimental procedures

Plant material

The *A. thaliana* collection of Versailles, France, was used. At the time this study began, it consisted of 265 accessions, obtained from the ABRC and Nottingham stock centers or derived from collections carried out by French (Lavigne *et al.*, 2001; Le Corre *et al.*, 2002) and Japanese (Todokoro *et al.*, 1995) groups. From this collection, 95 accessions were chosen to span the range of eco-geographical distribution. The lists of these 265 and 95 accessions can be found in Table S1. Passport (eco-geographical) data are accessible at http://dbsgap.versailles.inra.fr/vnat/.

Genomic DNA was extracted from leaves or from seedlings using a cetyltrimethylammonium bromide (CTAB) protocol in microtiter plates (Loudet *et al.*, 2002).

DNA sequencing

Sequences of the primers used for PCR and sequencing are available in Table S2, except for dihydroflavonol reductase (DFR) (Konieczny and Ausubel, 1993). PCR reactions were performed according to Fourmann *et al.* (2002) or Konieczny and Ausubel (1993) (DFR), and then the products were purified and sequenced using the Big Dye Sequencing kit according to the manufacturer's specifications (ABI, Courtaboeuf, France). Sequence products were purified and loaded onto ABI3700 96 capillary sequencers.

Polymorphisms at 85 loci were confirmed by independent duplicate PCR amplifications and sequencing reactions. No error was detected. The probability of finding a false SNP was then estimated to be less than one in 1.8×10^4 bp surveyed.

Sequence alignment and SNP detection were performed using the software GENALYS, available at http://software.cng.fr. The genomic sequence of the ecotype Col-0 (Arabidopsis Genome Initiative, 2000) was used as a reference. Polymorphism data is available at http://dbsgap.versailles.inra.fr/cngdata/.

Nucleotide diversity statistics (π, θ, and Tajima's *D*) were computed using DNASP Version 3.51 (Rozas and Rozas, 1999).

Core collection sampling

The M strategy (Schoen and Brown, 1993) was used for generating core collections that maximize the number of observed alleles at the marker loci. As the actual exploration of all possible core collections is not feasible (the number of collections to examine grows factorially with the size of the collection), we used a heuristic algorithm implemented in the MSTRAT software (Gouesnard *et al.*, 2001). The efficiency of this algorithm was previously checked on different sets of data where an exhaustive search was possible (Bataillon *et al.*, 1996); web site for MSTRAT: http://www.ensam.inra.fr/gap/MSTRAT/mstratno.htm).

The efficiency of the sampling strategy was assessed by comparing the total number of alleles captured using MSTRAT in samples of increasing sizes to the number of alleles captured in randomly chosen collections of the same sizes (50 independent samplings of core collections were made in each case).

All the SNPs detected were used in MSTRAT to generate nested core collections (from 8 to 48 with a step of eight accessions). Putative core collections exhibiting the same allelic richness were ranked using Nei's diversity index (Nei, 1987) as a second criterion of M. Three hundred core collections were generated independently for each sampling size, and the accessions that were most often present in these 300 replicates were retained as the final core collection.

A cross-validation was performed in which all combinations of 5 out of the 10 original 600-bp fragments were used to capture the allelic richness in the remaining five fragments in an exhaustive manner (252 combinations). For every combination, three independent samplings of core collections (both random and using the marker information) were performed. The relative efficiency of the M strategy was then assessed (see Figure S2).

Validation of the core collections based on gene data

Four loci, CBF2, CBF3, DREB2A (Shinozaki and Yamaguchi-Shinozaki, 2000), and LUMINIDEPENDENS (Lee *et al.*, 1994; between 1 and 1.5 kbp each) were chosen for validation. For CBF2, CBF3, and DREB2A, which contain no introns, the entire coding region was sequenced as well as part of the promoter and 3′ non-coding region. For LUMINIDEPENDENS, 1440 bp encompassing exon 7 to exon 11 were sequenced. The DNA polymorphism present at these four loci (4913 bp in total) in the 95 accessions was evaluated. The proportion of the diversity captured in the core collections was then compared to that captured in 1000 samples of the same size chosen randomly.

Validation of the core collections based on morphological data

Two hundred and forty accessions were morphologically characterized. The accessions were assayed in two controlled short-day environments (after 3 weeks, 4°C vernalization versus non-vernalization) using a randomized block design with three plants per accession in each environment. The VARCOMP procedure (SAS) was used to estimate variance components considering the block effect as fixed and the accessions as a random factor. Broad-sense heritabilities were computed independently for each treatment, as V(G)/(V(G) + V(E)), where V(G) is the between-accession variance component and V(E) is the residual (error) variance. Among 23 morphological traits of plant size architecture, 17 presenting high heritability values in both environments (see Table S3) were used to characterize the phenotypic diversity present in the collection, for example flowering time and maximum height. A set of 70 accessions analyzed had the molecular SNP information and no missing morphological data. This set was used for comparison and random re-sampling.

A cluster tree of variables was constructed using the Pearson correlation coefficient as a measure of pair-wise distance between variables. A single variable having a high heritability was then selected within each branch of the tree, leading to a set of 17 variables with little redundancy. This set of variables was used to perform a principal component analysis (PCA). First and second axes in the PCA accounted for, respectively, 28.2 and 17.3% of the total variance.

The Mahalanobis distance (Mahalanobis, 1936) was used as a measure of multivariate distance between each accession and the multitrait mean value. The observed range of Mahalanobis distance values was divided into several discrete classes of equal length. The ability of the M strategy to comprehensively sample the available morphological variability was then assessed using two statistics: first, the number of classes represented in each core collection; second, the evenness of the distribution of the accessions comprising the core among classes using the Shannon–Weaver (Shannon and Weaver, 1949) diversity index ($H = -\sum_i p_i \ln(p_i)$, where p_i is the proportion of accessions found in the i^{th} class). For each core size, observed values of the two statistics were compared with those obtained from 2000 random samples. As morphological data were available for only seven accessions from the core collection of size 8, seven Mahalanobis distance classes were used in all computations so that even in the smallest core, the number of classes and the Shannon–Weaver index could (potentially) reach the maximum value.

Acknowledgements

Génoplante (http://www.genoplante.org/, rapeseed program GOPT4) provided PCR primers. Financial support was provided by the French Minister of Research (INRA programs AIP 258, 272 and 1618). We are grateful to F. Roux and N. Scarcelli for their help in collecting morphological data and to F. Matsuda and M. Takahashi for their support in sequencing and analysis. We thank M. Lathrop and D. Schoen for critical reading of the manuscript.

Supplementary Material

The following material is available from http://www.blackwellpublishing.com/products/journals/suppmat/TPJ/TPJ2034/TPJ2034sm.htm

Table S1 List of the 265 accessions used for this study with their origins

Table S2 Primers used to amplify marker genes

Table S3 Morphological characters and their heritabilities in two controlled environments (3 weeks, 4°C vernalization versus non-vernalization)

Table S4 Capture by the core collections of SNP alleles in four independent loci

Figure S1. Patterns of pair-wise LD in the initial collection.

Figure S2. Relative efficiency of M strategy compared to random sampling in a cross-validation.

References

Aguade, M. (2001) Nucleotide sequence variation at two genes of the phenylpropanoid pathway, the FAH1 and F3H genes, in *Arabidopsis thaliana*. *Mol. Biol. Evol.* **18**, 1–9.

Alonso-Blanco, C. and Koornneef, M. (2000) Naturally occurring variation in *Arabidopsis*: an underexploited resource for plant genetics. *Trends Plant Sci.* **5**, 22–29.

***Arabidopsis* Genome Initiative** (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Bataillon, T.M., David, J.L. and Schoen, D.J. (1996) Neutral genetic markers and conservation genetics: simulated germplasm collections. *Genetics*, **144**, 409–417.

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

Breyne, P., Rombaut, D., Van Geysel, A., Van Montagu, M. and Gerats, T. (1999) AFLP analysis of genetic diversity within and between *Arabidopsis thaliana* ecotypes. *Mol. Gen. Genet.* **261**, 627–634.

Caicedo, A.L., Schaal, B.A. and Kunkel, B.N. (1999) Diversity and molecular evolution of the RPS2 resistance gene in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA*, **96**, 302–306.

Erschadi, S., Haberer,G., Schoniger, M. and Torres-Ruiz, R.A. (2000) Estimating genetic diversity of *Arabidopsis thaliana* ecotypes with amplified fragment length polymorphism (AFLP). *Theor. Appl. Genet.* **100**, 633–640.

Fourmann, M., Barret, P., Froger, N., Baron, C., Charlot, F., Delourme, R. and Brunel, D. (2002) From *Arabidopsis thaliana* to *Brassica napus*: development of amplified consensus genetic markers (ACGM) for construction of a gene map. *Theor. Appl. Genet.* **105**, 1196–1206.

Frankel, O.H. (1984) Genetic perspectives of germplasm conservation. In *Genetic Manipulation: Impact on Man and Society* (Arber, W., Llimensee, K., Peacock, W.J. and Starlinger, P., eds), pp. 161–170. Cambridge: Cambridge University Press.

Galtier, N., Depaulis, F. and Barton, N.H. (2000) Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics*, **155**, 981–987.

Gouesnard, B., Bataillon, T.M., Decoux, G., Rozale, C., Schoen D.J. and David, J.L. (2001) MSTRAT: An algorithm for building germ plasm core collections by maximizing allelic or phenotypic richness. *J. Hered.* **92**, 93–94.

Hanfstingl, U., Berry, A., Kellogg, E.A., Costa, J.T., III, Ruediger, W. and Ausubel, F.M. (1994) Haplotypic divergence coupled with lack of diversity at the *Arabidopsis thaliana* alcohol dehydrogenase locus: roles for both balancing and directional selection? *Genetics*, **138**, 811–828.

Haubold, B., Kroymann, J., Ratzka, A., Mitchell-Olds, T. and Wiehe, T. (2002) Recombination and gene conversion in a 170-kb genomic region of *Arabidopsis thaliana*. *Genetics*, **161**, 1269–1278.

Hugot, J.P., Chamaillard, M., Zouali, H. *et al.* (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, **411**, 599–603.

Innan, H. and Stephan, W. (2000) The coalescent in an exponentially growing metapopulation and its application to *Arabidopsis thaliana*. *Genetics*, **155**, 2015–2019.

Innan, H., Terauchi, R. and Miyashita, N.T. (1997) Microsatellite polymorphism in natural populations of the wild plant *Arabidopsis thaliana*. *Genetics*, **146**, 1441–1452.

Kawabe, A., Yamane, K. and Miyashita, N.T. (2000) DNA polymorphism of the cytosolic phosphoglucose isomerase (PgiC) locus of the wild plant *Arabidopsis thaliana*. *Genetics*, **156**, 1339–1347.

Konieczny, A. and Ausubel, F.M. (1993) A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers. *Plant J.* **4**, 403–410.

Kuittinen, H. and Aguade, M. (2000) Nucleotide variation at the CHALCONE ISOMERASE locus in *Arabidopsis thaliana*. *Genetics*, **155**, 863–872.

Lavigne, C., Reboud, X., Lefranc, M., Porchet, E., Roux, F., Olivieri, I. and Godelle, B. (2001) Evolution of genetic diversity in metapopulations: *Arabidopsis thaliana* as an experimental model. *Genet. Sel. Evol.* **33**, S399–S423.

Le Corre, V., Roux, F. and Reboud, X. (2002) DNA polymorphism at the FRIGIDA gene in *Arabidopsis thaliana*: extensive nonsynonymous variation is consistent with local selection for flowering time. *Mol. Biol. Evol.* **19**, 1261–1271.

Lee, I., Aukerman, M.J., Gore, S.L., Lohman, K.N., Michaels, S.D., Weaver, L.M., John, M.C., Feldmann, K.A. and Amasino, R.M. (1994) Isolation of LUMINIDEPENDENS: a gene involved in the control of flowering time in *Arabidopsis*. *Plant Cell*, **6**, 75–83.

Li, B., Suzuki, J.I. and Hara, T. (1998) Latitudinal variation in plant size and relative growth rate in *Arabidopsis thaliana*. *Oecologia*, **115**, 293–301.

Loridon, K., Cournoyer, B., Goubely, C., Depeiges, A. and Picard, G. (1998) Length polymorphism and allele structure of trinucleotide microsatellites in natural accessions of *Arabidopsis thaliana*. *Theor. Appl. Genet.* **97**, 591–604.

Loudet, O., Chaillou, S., Camilleri, C., Bouchez, D. and Daniel-Vedele, F. (2002) Bay-0 x Shahdara recombinant inbred line population; a powerful tool for the genetic dissection of complex traits in *Arabidopsis*. *Theor. Appl. Genet.* **104**, 1173–1184.

Mahalanobis, P.C. (1936) On the generalized distance in statistics. *Proc. Natl. Acad. Sci. India*, **2**, 49–55.

Maloof, J.N., Borevitz, J.O., Dabi, T. *et al.* (2001) Natural variation in light sensitivity of *Arabidopsis*. *Nat. Genet.* **29**, 357–358.

Miyashita, N.T., Kawabe, A. and Innan, H. (1999) DNA variation in the wild plant *Arabidopsis thaliana* revealed by amplified fragment polymorphism analysis. *Genetics*, **152**, 1723–1731.

Nei, M. (1987) *Molecular Evolutionary Genetics*. New York: Columbia University Press.

Nordborg, M. (2000) Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics*, **154**, 923–929.

Nordborg, M., Borevitz, J.O., Bergelson, J. *et al.* (2002) The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* **30**, 190–193.

Ogura, Y., Bonen, D.K., Inohara, N. *et al.* (2001) A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature*, **411**, 603–606.

Purugganan, M.D. and Suddith, J.I. (1999) Molecular population genetics of floral homeotic loci: departures from the equilibrium-neutral model at the APETALA3 and PISTILLATA genes of *Arabidopsis thaliana*. *Genetics*, **151**, 839–848.

Roses, A.D. (1997) A model for susceptibility polymorphisms for complex diseases: apolipoprotein E and Alzheimer disease. *Neurogenetics*, **1**, 3–11.

Rozas, J. and Rozas, R. (1999) DnaSP version3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics*, **15**, 174–175.

Sachidanandam, R., Weissman, D., Schmidt, S.C. *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.

Savolainen, O., Langley, C.H., Lazzaro, B.P. and Freville, H. (2000) Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing *Arabidopsis lyrata* and the selfing *Arabidopsis thaliana*. *Mol. Biol. Evol.* **17**, 645–655.

Schmid, K.J., Sorensen, T.R., Stracke, R., Törjék, O., Altmann, T., Mitchell-Olds, T. and Weisshaar, B. (2003) Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. *Genome Res.* **13**, 1250–1257.

Schoen, D.J. and and. Brown, A.H.D. (1993) Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. *Proc. Natl. Acad. Sci. USA*, **90**, 10623–10627.

Shannon, C.E. and Weaver, W. (1949) *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press.

Sharbel, T.F., Haubold, B. and Mitchell-Olds, T. (2000) Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. *Mol. Ecol.* **9**, 2109–2118.

Shinozaki, K. and Yamaguchi-Shinozaki, K. (2000) Molecular responses to dehydration and low temperature: differences and cross-talk between two stress signaling pathways. *Curr. Opin. Plant Biol.* **3**, 217–223.

Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D. and Buckler, E.S., **IV** (2001) Dwarf8 polymorphisms associate with variation in flowering time. *Nat. Genet.* **28**, 286–289.

Todokoro, S., Terauchi, R. and Kawano, S. (1995) Microsatellite polymorphisms in natural populations of *Arabidopsis thaliana* in Japan. *Jpn. J. Genet.* **70**, 543–554.

Törjék, O., Berger, D., Meyer, R.C., Mussig, C., Schmid, K.J., Rosleff Sörensen, T., Weisshaar, B., Mitchell-Olds, T. and Altmann, T. (2003) Establishment of a high-efficiency SNP-based framework marker set for *Arabidopsis*. *Plant J.* **36**, 122–140.

Ullrich, H., Lattig, K., Brennicke, A. and Knoop, V. (1997) Mitochondrial DNA variations and nuclear RFLPs reflect different genetic similarities among 23 *Arabidopsis thaliana* ecotypes. *Plant Mol. Biol.* **33**, 37–45.

Ungerer, M.C., Halldorsdottir, S.S., Modliszewski, J.L., Mackay, T.F. and Purugganan, M.D. (2002) Quantitative trait loci for inflorescence development in *Arabidopsis thaliana*. *Genetics*, **160**, 1133–1151.