

## ORIGINAL ARTICLE

# Stability across environments of the coffee variety near infrared spectral signature

H Posada<sup>1</sup>, M Ferrand<sup>2</sup>, F Davrieux<sup>3</sup>, P Lashermes<sup>4</sup> and B Bertrand<sup>2</sup>

<sup>1</sup>CENICAFE, Plant Breeding, Manizales, Caldas, Colombia; <sup>2</sup>CIRAD, Centre de Coopération Internationale en Recherche Agronomique pour le développement, UMR RPB, Montpellier, France; <sup>3</sup>CIRAD, UMR QUALISUD, Montpellier, France and <sup>4</sup>IRD, UMR RPB, Montpellier, France

Previous study on food plants has shown that near infrared (NIR) spectral methods seem effective for authenticating coffee varieties. We confirm that result, but also show that inter-variety differences are not stable from one harvest to the next. We put forward the hypothesis that the spectral signature is affected by environmental factors. The purpose of this study was to find a way of reducing this environmental variance to increase the method's reliability and to enable practical application in breeding. Spectral collections were obtained from ground green coffee samples from multi-location trials. Two harvests of bean samples from 11 homozygous introgressed lines, and the cv 'Caturra' as the control, supplied from three different sites, were compared. For each site, squared Euclidean distances among the 12 varieties were estimated from the NIR spectra. Matrix

correlation coefficients were assessed by the Mantel test. We obtained very good stability (high correlations) for inter-variety differences across the sites when using the two harvests data. If only the most heritable zones of the spectrum were used, there was a marked improvement in the efficiency of the method. This improvement was achieved by treating the spectrum as succession of phenotypic variables, each resulting from an environmental and genetic effect. Heritabilities were calculated with confidence intervals. A near infrared spectroscopy signature, acquired over a set of harvests, can therefore effectively characterize a coffee variety. We indicated how this typical signature can be used in breeding to assist in selection.

*Heredity* (2009) **102**, 113–119; doi:10.1038/hdy.2008.88; published online 29 October 2008

**Keywords:** NIRS; coffee; heritability; chemometric; selection

## Introduction

The Timor hybrid, a natural hybrid originating from an interspecific cross between *Coffea arabica* and *Coffea canephora* (Lashermes *et al.*, 2000) is being extensively used as the main source of resistance to pests and diseases for arabica breeding. Exploitation of Timor hybrid populations has thus far relied on conventional procedures, in which a hybrid is produced with an outstanding arabica variety, and the progeny is self and selected more than 4–8 generations (Castillo, 1990). Undesirable genes from the donor parent are expected to be gradually eliminated by selection. On the basis of this strategy, improved cultivars have already been released in several important coffee-producing countries, such as Brazil, Colombia, India and Kenya. In doing so, coffee-breeding programmes face considerable difficulties. In particular, there are strong limitations because of the length of time needed for coffee tree generation, the high cost of field trials, and the lack of accuracy in the current strategy to restore the genetic background of the recipient cultivar, and thereby ensure the good quality of the improved variety.

Molecular markers are widely used to study alien gene introgression in a large number of crop species (Fulton *et al.*, 1997; King *et al.*, 1998). In coffee, amplified fragment length polymorphism and simple sequence repeats have been successfully used to analyze the introgression of DNA fragments from *Coffea canephora* and *Coffea liberica* into *C. arabica* (Lashermes *et al.*, 2000; Prakash *et al.*, 2002; Herrera *et al.*, 2002a). Currently, it is not possible to rely those markers with undesirable genes. Furthermore, alien introgression in plants can induce extensive alterations in DNA methylation and transcription patterns (Liu *et al.*, 2004). Consequently, there is a strong need for inexpensive and high throughput evaluation methods to provide applications suitable for coffee breeding programmes. The purpose of this study was to assess the application of a near infrared spectroscopy (NIRS) methodology to assist in the selection of introgressed lines.

Near infrared spectroscopy is based on the absorption of electromagnetic radiation by matter (Osborne *et al.*, 1993). When applied to food products, this technique is of analytical use (Bertrand and Qannari, 2006), as it can extract a large amount of information concerning biochemical composition. Calibration equations are established by regression techniques taking into account a limited number of predictors, which are absorbances at certain wavelengths. Another possible use of NIRS is to authenticate varieties. For instance, this technique has been used to authenticate asparagus cultivars (Perez

Correspondence: Dr B Bertrand, UMR RPB, CIRAD-IRD TA80/IRD, 34398 Montpellier, Cedex 5, France.

E-mail: benoit.bertrand@cirad.fr

Received 3 April 2008; revised 20 June 2008; accepted 18 July 2008; published online 29 October 2008

*et al.*, 2001). The chemometric discrimination of the two coffee species *C. arabica* and *Coffea canephora* with NIRS approaches (Downey and Boussion, 1996; Esteban-Diez *et al.*, 2007) or Raman spectroscopy (Rubayiza and Meurens, 2005), have been tested with success. On the basis of NIR spectra treatment, Bertrand *et al.*, 2005 investigated ways of discriminating between modern introgressed varieties and traditional ones. They demonstrated approximately 90% of varieties could be successfully classified in these two categories through combined principal component analysis and discriminant analysis of spectral data.

Here, we assumed that each introgressed coffee variety could be characterized by a NIR spectral signature. As a phenotypic trait that signature is influenced by the environment and possibly by 'variety  $\times$  environment' interactions, which result from the differential performance of the variety across environments. For breeders, strong interactions are considered to be random nuisance factors. In this study, based on spectral collections acquired for ground green coffee samples from a mating design, (i) we confirmed the effectiveness of the NIRS signature in discriminating among varieties and (ii) we tried to demonstrate that the coffee variety NIR spectral signature was stable across environments.

We first used the total spectral information in a multivariable analysis, assuming that all the points of the whole spectrum were good predictors. We also tested a new method on the basis of the heritability concept, to choose the best predictors among the spectra. Heritability is known to be a function of genotypic interaction and error variances. This is the degree of confidence when evaluating genotypic value through the phenotypic value and gives a measurement of genetic stability for a given trait. We considered that a NIR spectrum was a discrete sequence, each point of which was a trait used as a phenotypic predictor of the genotypic spectral signature influenced by environmental effects. By keeping only those predictors for which heritability was significant, we sought whether it was possible to improve the stability of the coffee variety NIR spectral signature.

## Materials and methods

### Plant material

The experimental design employed in this study, included three Colombian locations in full combination with 12 *C. arabica* L. genotypes (11 introgressed lines and Caturra) and two field replications (total of 72 coffee bean samples). The variety Caturra was selected as representing high-quality traditional varieties. The 11 introgressed lines (A to K) were advanced lines (at least generation F5) derived from crosses between Caturra and the Timor hybrid accession CIFIC-1343. They were selected for their high yield, quality, and resistance to rust. The three locations studied represented the main coffee growing regions in Colombia, namely Naranjal at 1381 m a.s.l., Paraguaicito at 1250 m a.s.l and Rosario at 1635 m a.s.l. They exhibited contrasting agroclimatic characteristics. Samples were collected during the harvest peak, using healthy ripe cherries. For each sample, 1 kg of cherries was processed by the wet method

(pulping, fermentation and drying) to obtain approximately 250 g of green coffee beans. The samples of green coffee were screened through a size 17/64 inches sieve and the most defective beans were discarded.

### NIRS phenotyping

NIR reflectance spectra were collected using a scanning monochromator NIR systems spectrophotometer (model 6500, Perstrop Analytical Inc., 1201 Tech Road Silver Spring, MD, USA) driven by NIRS2 (4.0) software (Intrasoft Intl., LLC, RD109, Sellers Lane, Port Matilda, PA, USA). The analyses were performed on green coffee (50 g) after grinding (ground to  $<0.5$  mm). For each sample, a NIR spectrum was acquired in reflectance ( $R$ ) mode, where  $R$  represented reflectance energy in the 4000–9090  $\text{cm}^{-1}$  range in 7  $\text{cm}^{-1}$  step (Downey and Boussion, 1996). The log ( $1/R$ ) absorbance spectrum was obtained by the mean of those measurements and comparison with the reference. The mean quadratic error estimated from two subsamples (two distinct samplings of the same sample) on the basis of the raw spectrum (log  $1/R$ ) was under 300  $\mu\text{abs}$ ; that error was below the manufacturer's specifications and indicated satisfactory repeatability of the spectral measurement. Given these results, a single spectrum was acquired per sample.

The NIR spectra of the following main components of green coffee were also studied: caffeine, trigonelline, sucrose and chlorogenic acids (namely caffeoylquinic 5-CQA), which came from commercial standards. Total lipids were extracted from 2-g samples of dried ground green coffee using a modified Folch method (Folch *et al.*, 1957).

### Data analysis

As the spectrum consisted of 699 wavelengths, the same number of heritability estimations was performed. Heritability, the regression of unobservable genotypic values on observable phenotypic values, is a function of genotypic, environmental interaction and error variances. The predictor genotypic value  $G_i$  was the mean of the variety over all the test locations ( $P_i \dots P$ ). A multisite analysis was carried out over three locations, using the following statistical model:

$$Y_{ik} = \mu + F_i + s_k + (Fs)_{ik} + E_{ik}$$

Where  $Y_{ik}$  was the phenotypic value of the  $i$ th variety based on plot means at the  $k$ th site,  $\mu$  was the overall mean,  $F_i$  was the random effect because of the  $i$ th variety,  $s_k$  was the fixed effect because of the  $k$ th station,  $(Fs)_{ik}$  was the interaction between the  $i$ th variety and site  $k$ , and  $E_{ik}$  was the random error of plot means. The corresponding variances of random effects were:  $\sigma_F^2$ ,  $\sigma_{Fs}^2$ , and  $2\sigma_e^2$ , respectively. Estimated broad sense heritability ( $H_b^2$ ) for inbred lines (two replications for each) evaluated at three locations was given by the formula:

$$H_b^2 = \sigma_F^2 / (\sigma_F^2 + \frac{1}{3} \sigma_{(Fs)}^2 + \frac{1}{(2 \times 3)} \sigma_e^2)$$

This estimation was performed on harvests 1 and 2. The confidence intervals for heritabilities were assessed by using an 'all but one' jack-knife, by removing one individual at a time from each subsample with a circular permutation pattern. The sampling variance of the estimated parameters was given by Tukey's formula. Trait heritability and confidence intervals were calculated

using DIOGENE software, an extension of OPEP software (Baradat *et al.*, 1995). The heritability was compared with the spectra of the major coffee bean compounds.

### Analysis of Euclidean distances between homozygous lines

In an initial analysis, all the spectral data were used. Chemometric processing consisted initially of a principal component analysis on the basis of the spectra on 4000–9090  $\text{cm}^{-1}$  segment. In an initial approach, the efficiency of NIRS applied to green coffees was assessed as a way of discriminating between varieties for harvests 1 and 2. Factorial scores of PCs showing an eigenvalue higher than one were used to calculate the discriminant function models using varieties as criteria. In a second approach, we tried to test the stability of the NIR signature across harvests and environments. On the basis of the same factorial scores of PCs as previously, chemometric processing then consisted in calculating the matrices of the square Euclidean distances between the 12 inbred lines for each site and each harvest, then for each site taking both harvests into account. The matrices were compared with a Mantel test (Mantel, 1967) using GENETIX software version 4.01 (Laboratoire Génome et Populations, CNRS UPR 9060, Université de Montpellier II, Montpellier, France). That nonparametric test was used to compare two distance matrices by determining whether or not there was any correlation between the two matrices (Piepho, 2005).

In a second analysis, only those wave numbers for which absorbance heritability ( $\log 1/R$ ) was significant were kept. These were heritabilities whose confidence interval had a lower limit greater than zero. The subset of wave numbers was therefore defined by the heritability value. The same procedure as earlier was applied to that subset. To compare the two approaches (that is, with the whole spectrum or with the most heritable predictors), we compared the pseudo  $r$  of correlation obtained by comparing the two distance matrices by a Mantel test, along with a comparison of the associated probabilities. The procedure that gave the highest pseudo  $r$  associated with the greatest probability was considered to be the most efficient. Hierarchical cluster classifications were then performed using the square Euclidean distance to illustrate the stability of the relations among varieties from one site to the next.

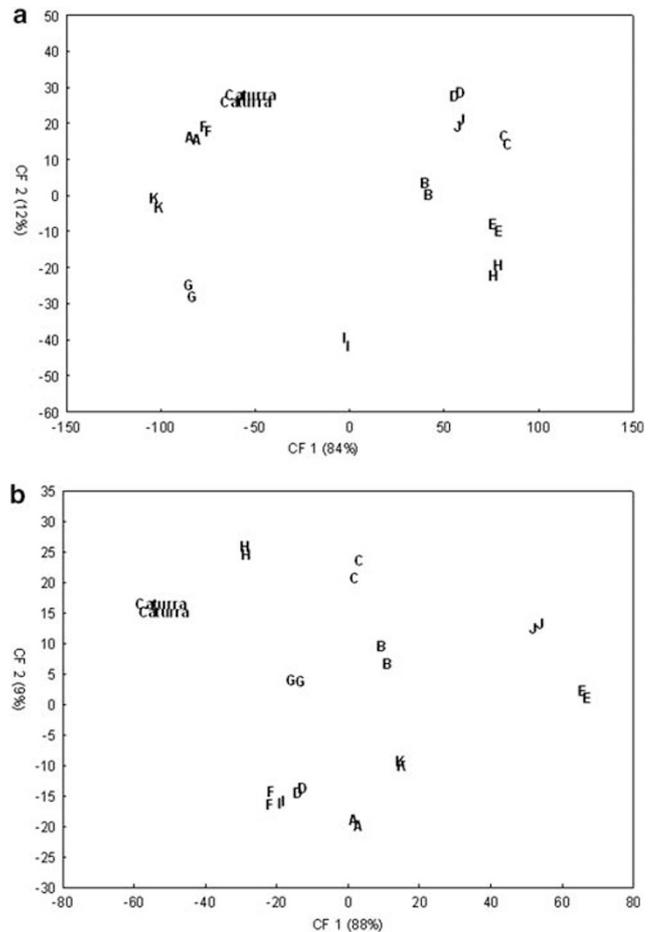
## Results

### Principal component analysis for harvests 1 and 2

Principal component analysis was used to set up noncorrelated variables that contained the maximum of the initial variance. For the two harvests, principal component analysis provided a similar pattern for the cumulative percentage of variance explained by the first principal components. The first 12 factors that explained more than 99% of the total variance and had eigenvalues more than one were used to calculate the discriminant function models.

### Discriminant analysis for varieties

Significant classifications were obtained for both harvests 1 and 2, as estimated by the  $P$ -value associated with the

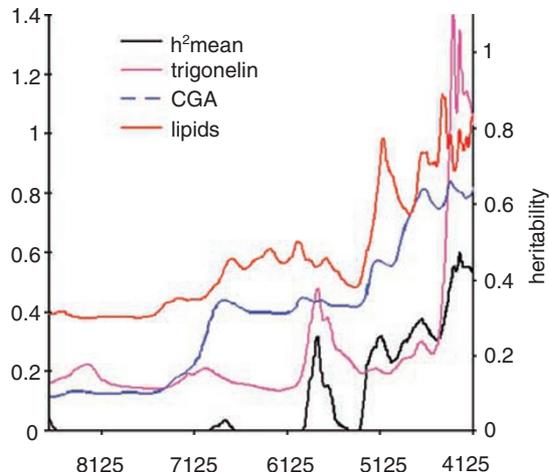


**Figure 1** Scatterplot of canonical scores for the first two canonical functions resulting from the discriminant analysis of the 12 varieties studied (Caturra, A, B, C, D, E, F, G, H, I, J, K), based on NIRS acquired on ground green coffee. Figure 1a corresponds to harvest 1 and Figure 1b to harvest 2.

Wilk's lambda coefficient ( $\lambda_{\text{Wilk}} = 0.001$ ,  $P = 0.0001$  and  $\lambda_{\text{Wilk}} = 0.0001$ ,  $P = 0.000$  for harvests 1 and 2, respectively). The percentage of correct variety classifications was perfect, as it reached 100% for both harvests 1 and 2. For each of the two harvests, two discriminant functions accounting for more than 90% of the variance were obtained using this discriminant model. Scatterplots presented in Figure 1, on the basis of canonical scores of the 72 samples analyzed, show that the spectral signature for harvest 1 (Figure 1a) and harvest 2 (Figure 1b) enabled very good variety discrimination. However, the inter-variety relations are not the same in the two figures (Figures 1a and b). For example, if we consider the variety Caturra, which was the control variety as it is the most widely planted variety in the world, it can be seen that the distances from the other varieties were not the same between the two harvests. Consequently, it is evident that it is not possible to propose selection according to relations of proximity between the variety Caturra and introgressed lines on the basis of a single harvest.

### Heritability as a function of wave number

The NIR diffuse reflection spectra ( $\log 1/R$ ) for the ground green coffee obtained for harvest 1 were very

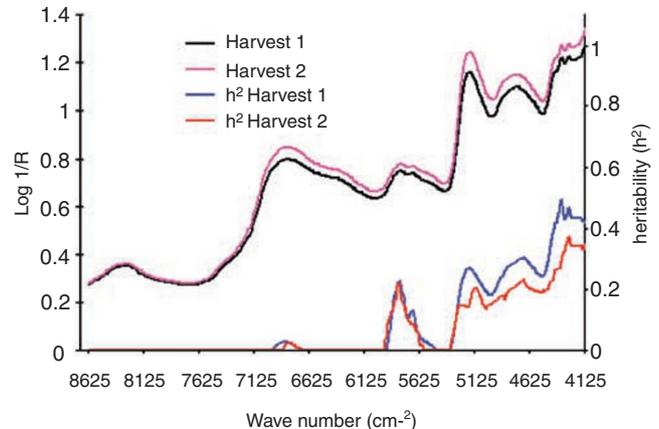


**Figure 2** Near infrared diffuse reflection mean spectra ( $\log 1/R$ ) for ground green coffees for harvests 1 and 2 and corresponding heritabilities ( $h^2$ ). The heritability of  $\log(1/R)$  was calculated for the three sites and 12 varieties for harvests 1 and 2 ( $h^2$  harvest 1 and  $h^2$  harvest 2, respectively). The heritability is represented as a function of wave number. The values of  $h^2$  above 0.2 are significantly different from zero (confidence intervals not passing zero).

similar to those obtained for harvest 2 (Figure 2). The heritabilities calculated for harvests 1 and 2 were also very similar, making it possible to calculate a mean heritability ( $h^2$  mean). When  $h^2 < 0.2$ , it was found that the lower limit of the confidence interval was always less than zero, which would seem to be nonsignificant heritability under the experimental conditions. Heritabilities were null or nonsignificant for the spectrum interval ranging from 8700 to 5925  $\text{cm}^{-1}$ . The heritabilities increased and became significant (confidence intervals not passing zero) between 5800 and 5767  $\text{cm}^{-1}$ . From 5700 to 5235  $\text{cm}^{-1}$ , the heritabilities were null or nonsignificant. Lastly, in the 5230–4000  $\text{cm}^{-1}$  interval, the heritabilities were significant and reached high values ( $h^2 = 0.49$ ). It appeared that the heritabilities at adjacent wave numbers were correlated. Spectral regions where absorbance was more or less heritable thus appeared.

#### Relation between the spectra of the major bean coffee compounds and the heritability

Ground green coffee is a complex solid containing a mixture of numerous molecules that absorb light simultaneously and in different ways. For this type of complex mixture, reflections at a wave number cumulate. The absorption bands of organic products observed in NIR are harmonics and combination bands, whose fundamentals are located in the mid infrared. It is therefore illusory to link the diffuse reflexion at a wave number with a particular compound. However, in infrared, spectra display few broad reflection peaks (namely 'series of peaks'). We tried empirically to compare the spectra for the major coffee bean compounds with the heritability values obtained depending on the wave number (Figure 3). The series of peaks occurring on the mean heritability curve ( $h^2$  mean) seemed to be linked to the trigonelline series of peaks around 5790  $\text{cm}^{-1}$ . On the interval (5270–4125  $\text{cm}^{-1}$ ), the



**Figure 3** Near infrared diffuse reflection spectra ( $\log 1/R$ ) for three major components of the coffee bean. Total lipids (lipids), chlorogenic acids (CGA), and trigonelline, accounting for 10–17, 6–8 and 0.9–1% of dry matter respectively are represented. Additionally, the mean heritability ( $h^2$  mean) that is the mean of  $h^2$  harvest 1 and  $h^2$  harvest 2 (see legend Figure 2) is also represented.

trigonelline curve was very significantly correlated to those of the heritabilities ( $r = 0.96$ ,  $P < 0.00001$ ). A strong correlation was found between the chlorogenic acids spectrum and that of the heritabilities on the interval (5270–4125  $\text{cm}^{-1}$ ) ( $r = 0.80$ ,  $P < 0.0001$ ). For sucrose and caffeine (data not shown) there were no series of peaks linked to those of the heritabilities. The coefficients of correlation between the sucrose and caffeine spectra on the one hand, and that of heritability on the other hand, were highly significant and amounted to  $r = 0.75$  and  $r = 0.71$  respectively on the interval (5270–4125  $\text{cm}^{-1}$ ). Total lipids displayed a major series of peaks approximately 5080  $\text{cm}^{-1}$  corresponding to high heritability ( $h^2 > 0.25$ ). The second highest series of peaks for lipids approximately 4625  $\text{cm}^{-1}$  corresponded to a series of peak that was also substantial for heritability ( $h^2 > 0.29$ ). Lastly, the series of peaks approximately 4255  $\text{cm}^{-1}$  corresponded to very high heritabilities ( $h^2 > 0.46$ ). On the interval (5270–4125  $\text{cm}^{-1}$ ), the total lipids spectrum was significantly correlated to heritability ( $r = 0.62$ ).

#### NIR spectral signature across environments

Bean samples of 11 introgressed lines and the variety Caturra supplied from three different sites were analyzed to assess how the NIRS-based distances were affected by the environment. We compared distance matrices between accessions for harvest 1, 2 and the mean of harvest 1 and 2 (harvests 1–2), determined at three different sites. For each site and each harvest, square Euclidean distances among the 12 accessions were estimated on the basis of NIR spectra. The distance matrices among varieties obtained in that way were compared with a Mantel test to determine whether there existed any correlation between two matrices (Table 1). A comparison of the three sites for harvest 1 (Table 1a) showed that the matrices were identical between site 1 and site 2 ( $r = 0.72$  and  $P < 0.001$ ). In contrast, site 3 was not identical to sites 1 and 2.

Keeping only the heritable predictors, corresponding to the 5800–5767  $\text{cm}^{-1}$  and the 5700–5235  $\text{cm}^{-1}$  intervals,

**Table 1** Pairwise comparisons of distance matrices between varieties for harvest 1, 2 and the mean of harvests 1 and 2 (harvests 1–2), determined at three different sites (numbered 1, 2 and 3)

Site	1	2
<i>(a) Harvest 1 and all wave numbers</i>		
3	−0.07 (0.55)	0.21 (0.11)
2	0.72 (0.00)	
<i>(b) Harvest 1, (5800–5767 cm<sup>−1</sup>) and (5235–4000 cm<sup>−1</sup>) intervals</i>		
3	0.63 (0.00)	0.56 (0.00)
2	0.55 (0.00)	
<i>(c) Harvest 2 and all wave numbers</i>		
3	0.05 (0.31)	0.03 (0.56)
2	0.04 (0.42)	
<i>(d) Harvest 2, (5800–5767 cm<sup>−1</sup>) and (5235–4000 cm<sup>−1</sup>) intervals</i>		
3	0.02 (0.58)	0.03 (0.53)
2	0.18 (0.16)	
<i>(e) Mean of harvest 1–2 and all wave numbers</i>		
3	0.49 (0.00)	0.21 (0.06)
2	0.42 (0.01)	
<i>(f) Mean of harvest 1–2, (5800–5767 cm<sup>−1</sup>) and (5235–4000 cm<sup>−1</sup>) intervals</i>		
3	0.81 (0.00)	0.56 (0.00)
2	0.55 (0.00)	

The tables indicate correlations between sites for harvest 1 (Tables 1a and b), for harvest 2 (Tables 1c and d) and for the mean of the two harvests (Tables 1e and f). In Tables 1a, c and e, the correlations were calculated from the distance matrices obtained from the whole spectrum; in Tables 1b, d and f, the correlations were calculated from the distance matrices obtained from the heritable values, that is, the 5800–5767 cm<sup>−1</sup> and the 5235–4000 cm<sup>−1</sup> intervals. For each site, square Euclidean distances among the 12 varieties were estimated on the basis of NIR spectra. Matrix correlation coefficients among matrices are indicated and assessed according to the Mantel test. The probability of error when rejecting the null hypothesis (that is, absence of correlation among matrices) is indicated in brackets.

very good identity was found for the matrices among the three sites (Table 1b). The distances among varieties calculated according to those intervals were significantly identical from one site to the next.

For harvest 2, the relations among the distances of the varieties were different from one site to the next when all the wave numbers were considered or when only the heritable absorbances were kept. However, when considering both harvests, the analysis performed for each site showed that the distance relations were highly significantly stable among varieties (Table 1e). That stability was even greater when only considering a wave number for which absorbance was significantly heritable (Table 1f).

Hierarchical cluster analyses were performed for each of the sites, using the spectral data from the two harvests, on intervals where heritability was significant. For site 1, three major clusters, called C1–C3, were obtained as shown on the dendrogram (Figure 4, graph 1). For site 2, three major clusters, called C1'–C3' were also obtained (Figure 4, graph 2). The variety composition of each cluster of site 1 was strictly conserved in the clusters of site 2. For site 3, three clusters (called C1''–C3'') were obtained (Figure 4, graph 3). Compared with sites 1 and 2, a small discrepancy was observed. Variety B was classed in cluster C3 when it was initially classed in C2

for site 1 and 2. When considering the spectrum intervals where absorbance was heritable, the same hierarchical clusters were obtained without any discrepancy among sites (Figure 4, graphs 4, 5 and 6).

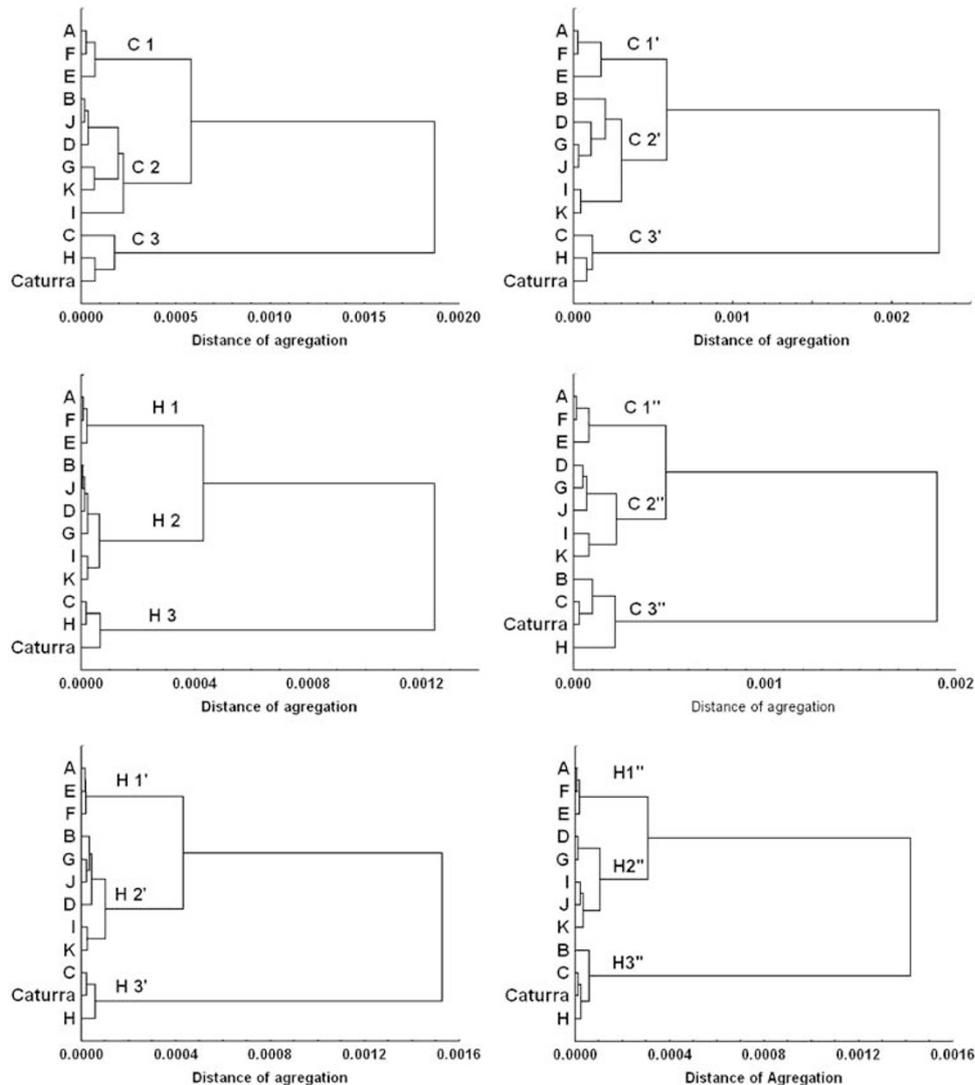
The proximity of varieties C and H with cv 'Caturra' suggested great proximity of those two varieties to the control. The distance of cluster C3 from cluster C1 (lines A, F and E) was always large. That large distance might have meant that the biochemical composition of those three lines was substantially different from that of the lines in cluster C1. In all, the results suggested stability across environments for inter-variety relationships determined using NIRS-based square Euclidean distances.

## Conclusion

The successful transfer of agriculturally valuable traits from wild species into elite breeding lines is highly dependent on the ability to manage the undesirable effects potentially associated with alien DNA introgression. Exotic germplasm resources often carry many undesirable genes in addition to valuable traits (Lee, 1998). Furthermore, alien introgression in plants can induce extensive alterations in DNA methylation and transcription patterns (Liu *et al.*, 2004). Recurrent backcrossing with selection for the presence of a target trait is a well-established breeding method for introgressing desirable genes from a donor to the genetic background of a recipient variety used as a recurrent parent. DNA markers could be used efficiently to reduce the number of backcross generations (Melchinger, 1990). However, marker-assisted introgression is not a panacea, and similar to any breeding method, its utility is affected by many factors (Lee, 1998). There is a strong need to add inexpensive and high throughput evaluation methods to established-breeding programmes. In view of the low genetic diversity existing in *C. arabica*, coffee improvement programmes are highly dependent on the ability to make efficient use of the diploid gene pool. Fortunately, there is great potential in exploiting genetic variation from diploid species, as most coffee species can hybridize readily with *C. arabica*, produce hybrids and recombine because of poor differentiation among homologous chromosomes (Herrera *et al.*, 2002b, 2004).

We recently showed (Bertrand *et al.*, 2008) that determining the fatty acid composition of the coffee bean is an effective tool for distinguishing among varieties. Here, we show on samples from different sites that the spectral signature is much more efficient in discriminating among varieties, as 100% correct classification is achieved as opposed to 79% with the fatty acid composition. Consequently, following many other authors, we show the power of this tool, which proves to be much less expensive than traditional chemical analyses for discriminating among varieties. The introgressed varieties were easily distinguished from the standard arabica variety (that is, Caturra), used as the reference variety in the trials and the recurrent parent in the backcrosses. Therefore, NIRS analyses proved to be extremely valuable for detecting in coffee beans, the biochemical modifications associated with *Coffea canephora* DNA introgressions, as shown previously (Bertrand *et al.*, 2005).

Unfortunately, we also show that the differences among varieties prove not to be stable from one harvest



**Figure 4** Hierarchical clustering analysis (square Euclidean distance, Ward grouping method) performed on NIR spectra acquired on ground green coffee. Graphs 1, 2, 3 representing locations 1, 2 and 3 respectively were obtained from the whole spectra of the 12 varieties. Graphs 4, 5 and 6 representing locations 1, 2 and 3 respectively were obtained from the heritable zones of the whole spectra of the 12 varieties. For each graph, distance-based clustering enabled the definition of three main groups (that is, C1, C2, and C3 for graph 1) of varieties sharing a similar NIR signature, and (that is, H1, H2, and H3 for graph 4) of varieties sharing similar NIR heritable zones in their signature.

to another. Obviously, NIRs profiles are strongly affected by environmental factors. Therefore, we sought to minimize environmental effects. We first compared matrices obtained for a single harvest at a single site. In that case, the distances between lines were not significantly comparable. However, when calculating the distances over two harvests, the relationships among lines were very stable from one site to the next. Although dynamic (that is, subject to environmental effects), the established relationships among lines seemed to be well conserved across locations. The NIRs-based inter-variety relationships determined at one site seemed to be stable across environments. Therefore, we concluded that there was stability across environments for inter-variety relationships.

The efficiency of the method was greatly increased if only certain zones of the spectrum were kept. To select those zones of the spectrum, we treated infrared spectra as a sequence of discrete variables that displayed a

genetic variance and an environmental variance. The ratio of genetic variance to the sum of variances amounted to heritability. Spectral regions where absorbance displayed high heritability, and spectral regions where heritabilities were not significant or null were revealed. Selecting from the spectrum, only those wave numbers for which absorbance was significantly heritable, amounted to choosing the best predictors, which is a conventional approach in breeding. That approach proved to be effective, as we found a significant increase in stability across environments. The study of heritable predictor distribution within the spectrum did not seem to be random. We attempted to show that some of the statistically significant heritabilities might correspond to certain major compounds of the coffee bean. It would be interesting to go into those studies in greater detail.

This stability of the NIR spectral signature is very interesting for breeders. The signature becomes a trait on which selection can be based through comparison with

ideotypes. For example, in the dataset, we analyzed one of the possible applications was to select lines closest to cv 'Caturra', which is an acknowledged standard in terms of cup quality. Indeed, it can be assumed that the closer a line is to cv 'Caturra', the closer its biochemical composition will be and the less it will display sensory differences from the reference variety. This type of approach can prove to be particularly fruitful, where introgressed lines are involved as in our study. Indeed, it has been shown that poorly mastered introgression of major chromosome fragments of this species, with a view to introducing resistance genes, could have a substantial negative impact on beverage quality (Bertrand *et al.*, 2003). We concluded that the NIRS methodology will be an efficient tool to assist in the selection of introgressed lines.

NIRS applications seem promising in several fields linked to breeding study, notably for genetically improving quality. In fact, in addition to cultivar authentication and traceability, NIRS can be used to assist in selection. Our results confirm that infrared spectroscopy is an inexpensive high through-put phenotyping tool that can be used by breeders for indirect selection on the basis of biochemical composition, and consequently on cup quality, in the coffee tree, but probably also in other species.

## Acknowledgements

The authors thank Philippe Baradat (Geneticist, INRA) for assistance with DIOGENE software. This study was supported by the National Federation of Coffee Growers of Colombia and the Colombia Ministry of Agriculture.

## References

- Baradat P, Labbé T, Bouvet JM (1995). Conception d'index pour la sélection réciproque récurrente. In: CIRAD (ed). *Traitements statistiques des essais de sélection. Stratégies d'amélioration des plantes pérennes*. CIRAD Publisher: Montpellier, France. pp 101–150.
- Bertrand B, Etienne H, Lashermes P, Guyot B, Davrieux F (2005). Can near infrared reflectance of green coffee be used to detect introgression in *Coffea arabica* cultivars? *J Sci Food Agric* **85**: 955–962.
- Bertrand B, Guyot B, Anthony F, Lashermes P (2003). Impact of introgression of *Coffea canephora* genes on Arabica Coffee Cup Quality. *Theor Appl Genet* **107**: 387–394.
- Bertrand B, Villarreal D, Laffargue A, Posada H, Lashermes P, Dussert S (2008). Comparison of the effectiveness of fatty acids, chlorogenic acids and elements for the chemometric discrimination of coffee (*Coffea arabica* L.) varieties and growing origins. *Journal of Agricultural and Food Chemistry* **56**: 2273–2280.
- Bertrand D, Qannari M (2006). Chimio-métrie appliquée à la spectroscopie proche infrarouge. In: Bertrand D, Dufour E (eds). *La spectroscopie infrarouge*, 2nd edn. Coll. Sciences et Techniques alimentaires, Lavoisier Presse: Paris, France. pp 309–343.
- Castillo Z (1990). Mejoramiento genético del café en Colombia. In: Centro Nacional de Investigaciones de Café CENICAFE (eds). *50 años de Cenicafe, 1938–1988*. Conferencias conmemorativas: Chinchina Colombia. pp 46–53.
- Downey G, Boussion D (1996). Authentication of coffee bean variety by near-infrared reflectance spectroscopy of dried extract. *J Sci Food Agric* **71**: 41–49.
- Esteban-Diez I, Gonzalez-Saiz JM, Saenz-Gonzales C, Pizzaro C (2007). Coffee varietal differentiation based on near infrared spectroscopy. *Talanta* **71**: 221–229.
- Folch J, Lees M, Sloane Stanley GH (1957). A simple method for the isolation and purification of total lipids from animal tissues. *J Biol Chem* **226**: 497–509.
- Fulton TM, Nelson JC, Tanksley SD (1997). Introgression and DNA marker analysis of *Lycopersicon peruvianum*, a wild relative of the cultivated tomato, into *Lycopersicon esculentum*, followed through three successive backcross generations. *Theor Appl Genet* **95**: 895–902.
- Herrera JC, Combes MC, Anthony F, Charrier A, Lashermes P (2002b). Introgression into the allotetraploid coffee (*Coffea arabica* L.): Segregation and recombination of the *C. canephora* genome in the tetraploid interspecific hybrid (*C. arabica* x *C. canephora*). *Theor Appl Genet* **104**: 661–668.
- Herrera JC, Combes MC, Cortina H, Alvarado G, Lashermes P (2002a). Gene introgression into *Coffea arabica* by way of triploid hybrids. *Heredity* **89**: 488–494.
- Herrera JC, Combes MC, Cortina H, Lashermes P (2004). Factors regulating gene introgression into the allotetraploid *Coffea arabica* L. from its diploid relatives. *Genome* **47**: 1053–1060.
- King IP, Morgan WG, Armstead IP, Harper JA, Hayward MD, Boyard A *et al.* (1998). Introgression mapping in the grasses. I. Introgression of *Festuca pratensis* chromosomes and chromosome segments into *Lolium perenne*. *Heredity* **81**: 462–467.
- Lashermes P, Andrzejewski S, Bertrand B, Combes MC, Dussert S, Graziosi G *et al.* (2000). Molecular analysis of introgressive breeding in coffee (*Coffea arabica* L.). *Theor Appl Genet* **100**: 139–146.
- Lee M (1998). Genome projects and gene pools: new germplasm for plant breeding? *Proc Natl Acad Sci USA* **95**: 2001–2004.
- Liu Z, Wang Y, Shen Y, Guo W, Hao S, Liu B (2004). Extensive alterations in DNA methylation and transcription in rice caused by introgression from *Zizania latifolia*. *Plant Mol Biol* **54**: 571–582.
- Mantel N (1967). The detection of disease clustering and generalized regression approach. *Cancer Res* **27**: 209–220.
- Melchinger AE (1990). Use of molecular markers in breeding for oligogenic disease resistance. *Plant Breed* **104**: 1–19.
- Osborne BG, Fearn T, Hindle PH (1993). *Practical NIR Spectroscopy with Applications in Food and Beverage Analysis*. Longman Scientific & Technical, Longman House, Burnt Hill, Harlow, Essex, UK. pp 31–37.
- Perez DP, Sanchez MT, Cano G, Garrido A (2001). Authentication of green asparagus varieties by near infrared reflectance spectroscopy. *J Food Sci* **66**: 323–327.
- Piepho HP (2005). Permutation tests for the correlation among genetic distances and measures of heterosis. *Theor Appl Genet* **111**: 95–99.
- Prakash NS, Combes MC, Naveen KS, Lashermes P (2002). AFLP analysis of introgression in coffee cultivars (*Coffea arabica* L.) derived from a natural interspecific hybrid. *Euphytica* **124**: 265–271.
- Rubayiza AB, Meurens M (2005). Chemical discrimination of Arabica and Robusta coffees by Fourier transform raman spectrometry. *J Agric Food Chem* **53**: 4654–4659.