

NMR profiling of transgenic peas

Adrian Charlton^{*1}, Theo Allnut², Stephen Holmes¹, James Chisholm¹, Samantha Bean², Noel Ellis², Phil Mullineaux² and Sarah Oehlschlager¹

¹Department for Environment, Food and Rural Affairs, Central Science Laboratory, Sand Hutton, York, YO41 1LZ, UK

²John Innes Centre, Colney, Norwich, NR4 7UH, UK

Received 31 July 2003;

revised 12 August 2003;

accepted 12 August 2003.

*Correspondence (fax: +44 1904 462133;

e-mail: adrian.charlton@csl.gov.uk)

Abbreviations: ¹H, proton; ²H, deuterium; FID, free induction decay; GM, genetically modified; LDA, linear discriminant analysis; PC, principal component; PCA, principal components analysis; PLS, partial least squares; TSP, sodium 3-(trimethylsilyl) propionate-d₄; WT, wild-type, AMOVA, analysis of molecular variance; ANOVA, analysis of variance.

Keywords: fingerprinting, GMO, metabolite profiling, multivariate statistics, NMR spectroscopy, pea, *Pisum sativum*, transgenic.

Summary

A high throughput proton nuclear magnetic resonance spectroscopy method for the metabolite fingerprinting of plants was applied to genetically modified peas (*Pisum sativum*) to determine whether biochemical changes, so called 'unintended effects', beyond those intended by incorporation of a transgene, were detectable. Multivariate analysis of ¹H NMR (nuclear magnetic resonance) spectra obtained from uniformly grown glasshouse plants revealed differences between the transgenic and control group that exceeded the natural variation of the plants. When a larger data set of six related transgenic lines was analysed, including a null segregant in addition to the wild-type control, multivariate analysis showed that the distribution of metabolites in the transgenics was different from that of the null segregant. However, the profile obtained from the wild-type material was diverse in comparison with both the transgenics and the null segregant, suggesting that the primary cause of the observed differences was that the transformation process selects for a subset of individuals able to undergo the transformation and selection procedures, and that their descendants have a restricted variation in metabolite profile, rather than that the presence of the transgene itself generates these differences.

Introduction

Metabolomics is concerned with the investigation of the chemical processes occurring in living organisms that result in the group of small molecules collectively known as metabolites. The metabolome is defined by the entire complement of low molecular weight, non-peptide metabolites within a cell, tissue or organism at a particular physiological state. Whilst there is much current interest in the genome-wide analysis of cells at the level of transcription (to define the transcriptome), and translation and protein modification (to define the proteome), the metabolome as a whole has been less extensively investigated.

The wide range of chemical processes that occur inside the cell suggest that there is a need for a profiling technique that can obtain large amounts of decipherable data for analysing biological matter (Hall *et al.*, 2002). Such a method would facilitate the investigation of highly correlated and interdependent biochemical pathways, providing an overview of metabolism. By establishing baseline data to describe the

'normal' physiological state for a given genotype, the effect of external factors, such as genetic manipulation, environmental stress or disease, could be detected as fluctuations from this baseline.

Proton nuclear magnetic resonance (¹H NMR) spectroscopy is often used to determine chemical structural information and make quantitative measurements of the concentration of proton containing molecules. As the ¹H NMR spectrum of a complex mixture contains quantitative information about all of the proton-containing molecules present in the mixture, provided that they are above the limit of detection and have a molecular weight of less than $\approx 30\,000$ Da, ¹H NMR is of great value for metabolic studies (Charlton, 2001; Fernandez and Clark, 1987; Fan, 1996; Ratcliffe and Shachar-Hill, 2001). The highly reproducible nature of the NMR spectrum permits rapid comparisons between sample groups to be made (Charlton *et al.*, 2002). The spectrum generated can be used as a specific (and often unique) profile of the matrix that is being studied, providing a fingerprint to identify a particular type of organism (Kelly-Borge *et al.*, 1994; Rycroft, 1996)

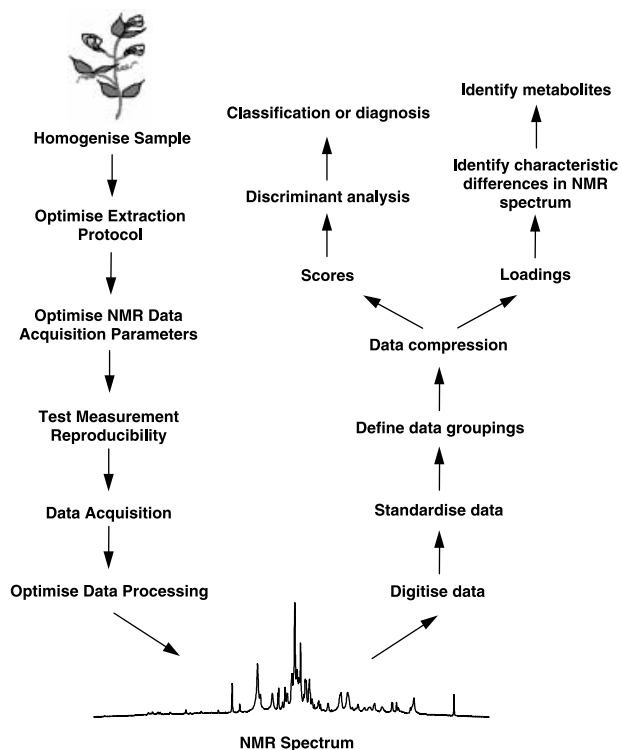


Figure 1 Methods for metabolite profiling by NMR.

or unexplained variation within the studied sample set (Noteborn *et al.*, 2000). The development of chemical profiling methods allows the detection of unexpected or unpredicted changes that could be missed if only a few, pre-selected target compounds are analysed (Fiehn, 2002).

In the present study we demonstrate the use of NMR and multivariate statistics to investigate the effects of transgene insertion in pea (*Pisum sativum*). Figure 1 shows a schematic representation of the stages required when using NMR spectroscopy to produce metabolic fingerprints.

Results and discussion

In this study a total of six independent lines of transgenic pea were used. These plants were transformed using an *Agrobacterium tumefaciens*-mediated procedure with a plasmid called pSLJ1561. The method and plasmid used and the specific preliminary description of these plants has been reported previously (Bean *et al.*, 1997; Bishop *et al.*, 1992). A map of the T-DNA of pSLJ1561 is shown in Figure 2A.

Plants harbouring this T-DNA were chosen because the T-DNA is more complex, having five transgenes and a *Ds* transposable element, than that is normally used to generate most transgenic plants (see Figure 2A). We reasoned that this was a rigorous test to challenge a null hypothesis that no differences would be detected between transgenic and wild-type lines. Peas were chosen because this species lends itself well to the development of methodology in plant biochemistry (Casey and Davies, 1994), which was the primary aim of this study.

The six lines were chosen out of a larger group of 26 as being representative of the types of T-DNA insertion events observed in this larger group. A summary of these insertion events is depicted in Figure 2B. In five out of six cases, the positions of the T-DNA inserts in these lines were mapped on

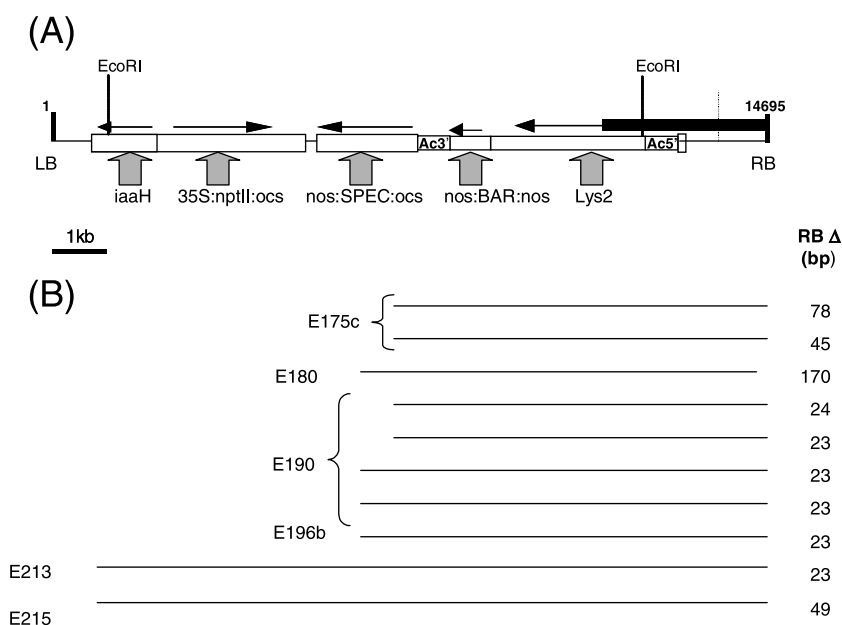


Figure 2 (A) A map of the T-DNA construct of pSLJ1561, containing five transgenes and a *Ds* transposable element. (B) A summary of the six insertion events selected as being representative of the types of T-DNA insertion events observed.

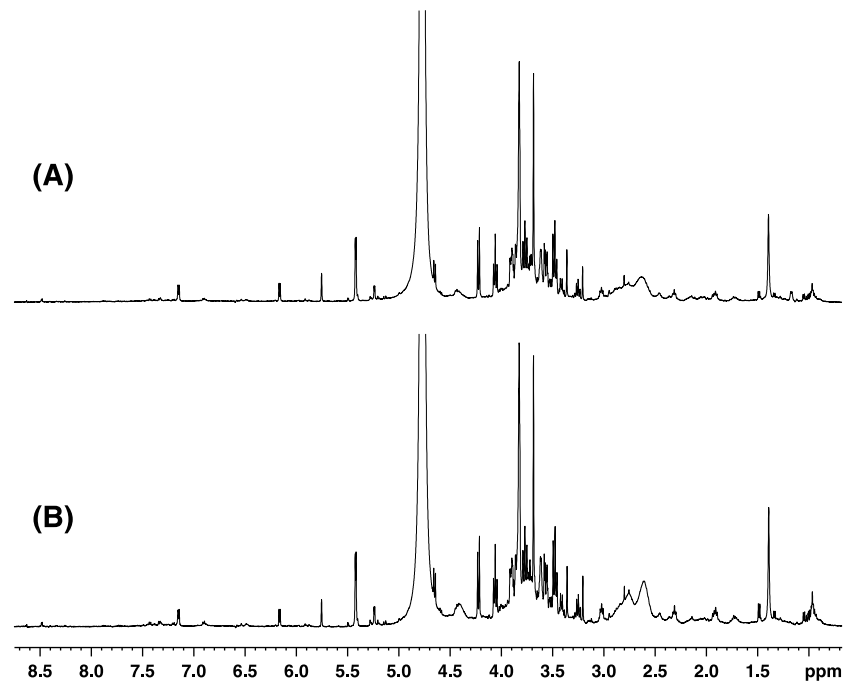


Figure 3 A typical ^1H NMR spectrum obtained from aqueous extracts of (A) wild-type and (B) transgenic pea leaves.

to the pea genome and confirmed that each line contained independent insertion events. The sixth line (E213; Figure 2) harboured its unmapped T-DNA locus in highly repetitive DNA (data not shown). A detailed description of how these data were derived will appear in a separate publication.

Forty-four T_3 pea plants (cv. Puget) were grown to first pod set. Eleven were wild-type (WT) plants grown from the stock of seed that was originally used to generate the transformants (Bean *et al.*, 1997). Thirty-three plants were from line E213 (Figure 2B) and were the T_3 progeny of selfed T_2 parents derived from an original T_1 primary transformant (see Experimental procedures). All T_3 individuals of line E213 were confirmed as containing the T-DNA insertion using a Southern blotting procedure. Aqueous extracts of the T_3 plants were prepared and their ^1H NMR spectra were recorded. Figure 3 shows a typical NMR profile obtained from both the T_3 transgenic and the wild-type pea samples. The 33 T_3 transgenic pea plants were allocated to a single group and compared to the 11 wild-type controls grown under the same conditions. Principal component analysis (PCA) was used to compress the NMR data and the principal component (PC) scores were used to classify all of the extracts into one of the two groups during model building and validation. Linear discriminant analysis (LDA) using the squared Mahalanobis distance metric was used to classify the compressed NMR data.

Figure 4 (inset) shows a three-dimensional principal component scores plot, calculated from the ^1H NMR spectrum of aqueous extracts obtained from the leaves of the T_3 plants.

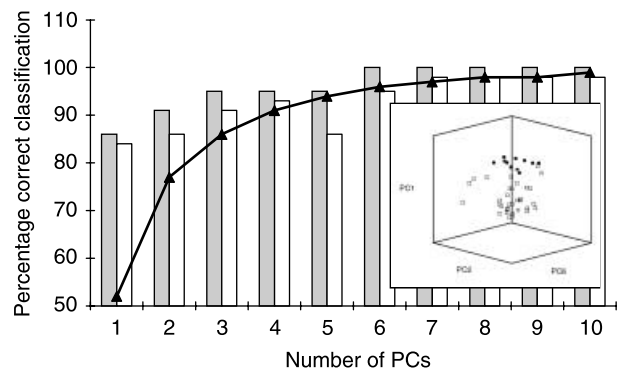


Figure 4 The classification success rate at both the model construction (shaded plot) and model validation (\square) stage of an LDA calculation performed on the PCA compressed NMR data derived from the T_3 pea extracts. The cumulative variance is also shown (\blacktriangle). Inset: Principal components analysis scores 1, 2 and 6 plotted to illustrate the clustering of the NMR data derived from the transgenic (\circ) and wild-type (\bullet) T_3 plants.

Good separation between the wild-type and the transgenic pea extracts was achieved using PC scores 1, 2 and 6, although a two-dimensional plot using only the first two principal components also demonstrated the presence of clustered data. When LDA was applied to these data, all of the samples were correctly classified at the model building stage using six principal components. During model validation using the blind 'leave one out' method, 95% of the samples (42/44) were correctly classified. The maximum number of correct classifications was achieved by using seven or more principal components, 98% (43/44 correct). By this process

it was determined that 99% of the variation in the NMR spectrum was represented within the first 10 principal components, with 96% represented in the first six (Figure 4).

The data presented in Figure 4 clearly demonstrated that the NMR profiles of the T_3 transgenic and the wild-type control plants were consistently different. It was not clear whether the differences observed in this principal components analysis resulted from the presence of the transgene insertion, from the transformation process, or from an undefined property. To investigate this further, a larger comparison between all six lines (Figure 2B) was carried out. However, by using a null segregant control from which the transgene has been lost, the effect of the presence of the insertion can be separated from effects resulting from the transformation process. Comparison between the null segregant and the genetically modified lines was carried out using T_4 plants, i.e. the third selfed generation from the primary transformant (T_1). As before, all individuals were confirmed as harbouring the expected T-DNA insertion event for that line, and null segregants were confirmed as having no insertion. Two hundred and twenty-two pea plants, comprising eight groups; six transgenic lines obtained from the separate transformation events listed in Figure 2B, one wild-type and one null segregant group were grown under greenhouse conditions in a randomised block design.

Principal components analysis was used to construct a model containing up to 12 principal components using all of the NMR data from the T_4 plant extracts. Linear discriminant analysis was used to simultaneously classify each of the samples into one of eight groups (Figure 2B). In contrast to the earlier initial experiment using T_3 generation plants, it was not possible to achieve an acceptable classification rate at the model building stage, and only 50% of the samples were correctly classified using the principal component scores. This was attributed to the similarity of the NMR profiles of the extracts derived from the different transgenic lines and the apparent wider diversity of the NMR profiles of the wild-type samples.

As an alternative approach, 14 partial least squares (PLS) models were constructed using 10 PLS factors calculated from each of the T_4 transgenic lines in turn and either the null segregant or the wild-type group. The first seven of these PLS models were constructed using the wild-type as the control group and the remaining seven models used the null segregant samples as controls. Each of the T_4 transgenic samples were placed in a group with others derived from the same T_1 transformant. Hence, eight groups were defined, six transgenic lines (Figure 2B), null segregant group and a wild-type group. The transgenic lines were individually compared to

Table 1 The number of PLS factors required to correctly classify all of the samples during model construction using the ^1H NMR spectrum derived from aqueous extracts of the T_3 pea population

Control group ^a	Genetic line	Number of PLS scores used	Number of samples
WT	1	8	49
WT	2	8	47
WT	3	10	51
WT	4	7	47
WT	5	7	50
WT	6	7	48
WT	NS	6	32
NS	1	7	49
NS	2	8	47
NS	3	7	51
NS	4	6	47
NS	5	5	50
NS	6	7	48
NS	WT	6	32

^aWT = Wild-type, NS = Null segregant.

both the null segregant and the wild-type plants. Linear discriminant analysis, using the squared Mahalanobis distance for classification, was performed on each pair-wise comparison of the compressed NMR data. The LDA was applied to sufficient PLS factors to achieve 100% correct classification at the model building stage. The number of PLS scores used in each model is summarized in Table 1.

The PLS models that were constructed to compare the null-segregant and the wild-type group to each of the transgenic lines were then subjected to blind 'leave one out' validation. The validation reclassification results using nine PLS factors are shown in Figure 5 for each line using the null segregant as the control group. Each of the transgenic lines could be distinguished from the null segregant with a success rate of between 71 and 84%.

There was no general correlation between the number of T-DNA insertions and the classification rate for each of the transgenic lines. This suggests that T-DNA insertion is sufficient to allow significant discrimination of a single transgenic line and the null segregant when considering the T_4 plants.

To assess the significance of the classification rates achieved during the validation stage of the PLS models, the NMR data were placed at random into eight groups and seven new PLS models were constructed using up to 10 PLS factors. One of the groups was arbitrarily chosen as a control group, against which each of the remaining seven groups were compared. During model building, classification rates of between 97 and 100% were achieved. However, following validation, the classification rate fell to between 44 and 54%,

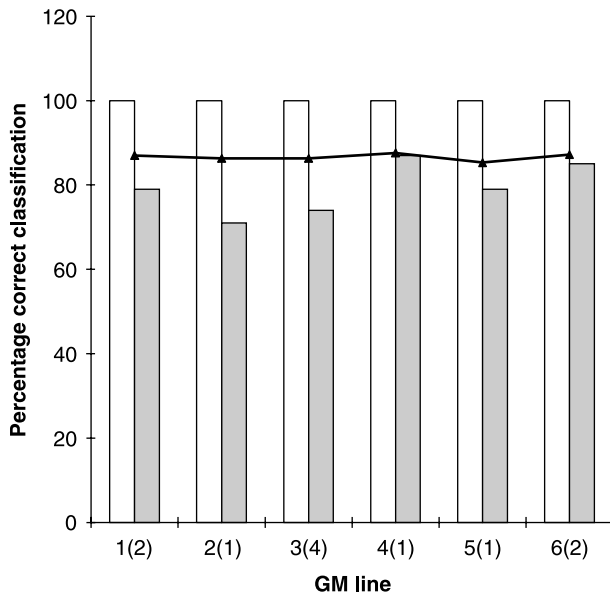


Figure 5 The classification success rate at both the model construction (□) and validation (shaded plot) stage when LDA was applied to a seven PLS factor model of the T_4 genetically modified and null segregant lines. The variances for the PLS calculations are also shown (▲) along with the number of T-DNA insertions for each genetically modified line (in parentheses).

with an overall correct classification rate of 50%. Theory suggests that the probability of classifying each sample into the correct group in a two-group model is 50%. The range of 71–84% correct classifications of the transgenic lines and null segregant, suggests that significant characteristics were present in the NMR spectrum of the transgenic lines that could be used to distinguish them from the control population. However, further analysis of the significance of these groups in the data was performed using AMOVA (analysis of molecular variance).

Nested AMOVA showed that there was no significant ($P = 0.5165$) variance between non-transgenic (wild-type and null segregants considered as a single data set) and transgenic plants (all T_4 lines). Therefore, AMOVA showed that the effect of the presence or absence of the transgene on the ^1H NMR profiles could not distinguish the data sets. These data suggest that the presence of the transgene does not impart a consistent difference upon the composition of the pea leaf extracts. When all pair-wise comparisons between each transgenic line and the null segregant group were considered, the NMR profiles were found to be different with confidence limits ranging from 79.02% to greater than 99.99%. Four of the six genetically modified lines were found to have NMR profiles that were different to those of the null segregant group with a confidence limit of 98.3% or above. Significant variance was also often present between different transgenic lines.

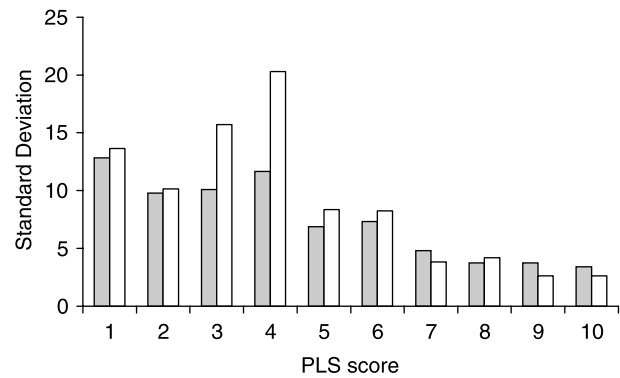


Figure 6 The standard deviation of the PLS scores for both the T_4 genetically modified (shaded plot) and wild-type plants (□).

The AMOVA results agree with the multivariate analyses in suggesting that the major factor contributing to the difference in metabolite profile is the selection and reduction in variation of the transformation and culture processes. Furthermore, the null segregant group was significantly different from the wild-type, indicating that differences due to factors other than the presence of the transgene have a larger effect on metabolite profile.

Further statistical analysis using PLS to compress the NMR data permitted the variation of the PLS factors for the wild-type samples to be used as a measure of the variation in the NMR spectrum as a whole. Figure 6 shows the standard deviation from the mean value for PLS factors 1–10 when the transgenic samples were considered as a single group. These were plotted with the standard deviation of the PLS scores calculated for the wild-type samples. It was clear that the spread of the PLS scores for the wild-type samples was considerably greater than that recorded for the genetically modified (GM) groups when considering PLS factors 1 to 6. This variation could be rationalized by considering the phenotypic diversity present in each of the two groups. The wild-type group was grown from Puget, a standard cultivar and derived from more than one plant. Therefore, this group was assumed to show the maximum range of variation in metabolites for this genotype grown under these environmental conditions. In contrast, the descendants of each of the transgenic lines were derived from a single individual, the primary T_1 transformant. Therefore, it was concluded that the T_4 plants still appeared to be exhibiting restricted profiles that could be associated with a history of single seed descent from a single individual. However, this did not explain why the profiles obtained from plants produced by different transformation events did not overall have a similar variation to the wild-type group. One possible explanation is that the transformation process itself selected for a subpopulation of cells that were

susceptible to *Agrobacterium* transformation and were subsequently able to give rise to regenerated plants.

Inspection and further statistical evaluation (Student *t*-tests and ANOVA [data not shown]) of the NMR spectra of the transgenic and null segregant groups, showed a number of differences between transgenic and control groups. However, due to the natural variation in signal intensity within a single transgenic line and within the null segregant and wild-type groups, single consensus differences between two groups could not be identified in all samples. Pair-wise Student *t*-tests (null segregant vs. each transgenic line) showed single point differences between the mean NMR intensity of the two data groups ($P < 0.01$), but considerable overlap was present when individual measurements were considered. This suggests that the statistical separation of the group members using LDA was achieved by considering the discrete contributions of many compounds to the overall metabolite profile, leading to combinations of metabolites that were indicative of the transgenic or null segregant group. Further work would be required to deconvolute these spectra.

We have shown that the NMR profiles of transgenic pea leaf extracts differ consistently from their non-transgenic counterparts for both the second and third generation after the original transformation. These differences were more pronounced for the T_3 plants than for the T_4 generation illustrated by their NMR profiles and represented by decreasing LDA classification rates with increasing generation.

We have also shown that the wild-type group, when compared to the T_4 plants, had much wider variation in metabolite profile. The T_1 transgenic plants were derived from a single transformed cell that was then regenerated under selective conditions. The T_1 lines were then selfed to produce the subsequent generations used in this study (Bean *et al.*, 1997). In all cases in subsequent generations, seed lots from individual plants were kept separate to create lineages. We propose that the step-wise bulking up from a single original transformant may have been the biggest contribution to the changing effects on NMR profiles observed in the T_3 vs. the T_4 generations. In addition, the similarity of the profiles obtained from all the transgenic lines derived from different transformation events suggests that the diversity between the transgenic plants within each line is small. One theory to explain this may be that each of the plants that successfully underwent the transformation procedure shared some commonality. Such commonality could take two, not mutually exclusive forms: Firstly, somaclonal variation, or secondly, selection of cells most able to undergo the transformation, tissue culture and regeneration process.

Phenotypic and genotypic change has been documented many times in plants regenerated from tissue culture and

is often termed somaclonal variation. (Bregitzer *et al.*, 2002; Hossain *et al.*, 2003; Jalignot *et al.*, 2002; Joyce and Cassells, 2002; Kubis *et al.*, 2003). The most likely cause is the many stresses such cells undergo during tissue culture, combined with the plasticity of their genomes in response to such conditions (Cassells and Curry, 2001; Joyce *et al.*, 2003). The observation that the period in culture is linked to the frequency of appearance of somoclonal abnormalities in regenerants is consistent with this view (Cote *et al.*, 2001; Pluhar *et al.*, 2001). Such plasticity is brought about by epigenetic changes such as histone acetylation, methylation of DNA and chromatin remodelling (Cassells and Curry, 2001; Joyce *et al.*, 2003; Kaeppler *et al.*, 2000; Peraza-Echeverria *et al.*, 2001). Such changes are usually most manifested in primary transformants and in any vegetative propagants, and are often, but not always, lost in subsequent generations (Bregitzer *et al.*, 2002; Hossain *et al.*, 2003). In these transgenic peas, searches for alterations in methylation patterns and new polymorphisms in retro-element flanking sequences failed to reveal any differences from plants of the same genotype (cv. Puget) that had never been through the transformation procedures (T. Allnutt; unpublished data). Therefore, while changes in NMR profiles may have been sensitive enough to pick up the consequences of any epigenetic variation, such metabolic variation was not caused by massive changes in the methylation of gene sequences or movements of retroelements. This epigenetic behaviour reflects variation at the cellular level in the T_0 ; the association of this variation with totipotency and a propensity to interact with *Agrobacterium* is reminiscent of the distinct regulatory states that genetically identical yeast cells exhibit in expressing or repressing *HML α* (Pillus and Rine, 1989).

The results of the LDA comparing the wild-type plants to the T_4 transgenic plants clearly showed a reduction in the number of correct classifications when compared to the same comparison using the T_3 lines. This suggests that generations of the transgenic plants that were further from the original transformation event appeared to be (at least in part) exhibiting metabolite profiles that were more similar to those of the non-transformed samples than earlier generations. Therefore, these results suggest that the metabolite profiles of the transgenic plants were beginning to return back to the degree of variation normally found in this genotype and that, if responsible, any subtle epigenetic changes were gradually being lost through successive generations. Such variation in metabolite profiles detected in extracts prepared from all the leaves of a single wild-type plant may reflect similar, if not greater variation at the cellular level. Such metabolic variation may be the basis of totipotency (i.e. able to regenerate) and

competence for transformation (Gaspar *et al.*, 2002; Ribnicky *et al.*, 2002). Furthermore, the necessary insertion of a selectable marker gene in a region of the genome capable of expressing under such conditions could be a further level of discrimination (Koncz *et al.*, 1989; Scott *et al.*, 1998). The net effect of such hurdles for a cell to overcome in order to give rise to a transformed regenerated plant may be the requirement to be in a particular metabolic state most appropriate to each stage of the procedure. The observed compression in variance and its subsequent restoration to wild-type levels in such lines is consistent with this interpretation.

In conclusion, variation brought about by epigenetic processes most likely lie at the heart of our observations on changes in the metabolome of these transgenic peas. However, it was not possible to determine if such epigenetic changes were brought about by the transformation procedure *per se* or were caused by intrinsic epigenetic variation in explant material that selected for totipotent, transformation-competent cell types; but given that variance expands in progressive generations of the transgenic lines back towards the maximum variance in wild-type plants, we tend to favour the latter explanation.

Finally, it should be stated that the plants used in the study were developed for experimental purposes only, as part of method development. The materials used in this work were grown under uniform conditions in a controlled glasshouse with the express purpose of minimizing the environmental variation, which would be expected to confound the separation of transgenic effects from, for example, the effects of herbivory, disease and abiotic stress that such plants would encounter in the field. It should be noted that these data cannot, at this time, be set into the context of normal variation encountered in the genus *Pisum* grown under various environmental conditions. However, the wider variation encountered even within the wild-type group compared with those descended from individuals that arose out of the transformation process suggest that the differences between transgenic and null-segregant individuals is relatively minor. Provided a database of spectra that describe the variation within a species can be assembled, examination of the natural variation of metabolite pools, should provide new methods for rapidly demonstrating that a GM plant destined for human or animal consumption lies within the spectrum of normal variation.

Experimental procedures

Plant material

The procedures used to generate these transformed plants and the lines used in this project have been described

previously (Bean *et al.*, 1997; Bishop *et al.*, 1992). A detailed analysis of the six lines described and the development of methodology will be published elsewhere. Plants were pot-grown in a controlled environment glasshouse at 15 °C (± 2 °C) with a maximum 16 h photoperiod, 70% relative humidity, twice daily watering and diluted feed once per week in John Innes no. 1 compost with 30% extra grit. Both generations were grown in October–December and supplemental lighting (Na lamps; 250 $\mu\text{mol}/\text{m}^2/\text{s}$ at 1 m from the lamps) was used as necessary. All plants, control and the transgenic lines, were grown together in a randomised block design surrounded by three lines of wild-type plants to eliminate edge effects within the group. Total leaf material was harvested from each plant to be analysed on the same day when the first flowers were observed on several individuals. The leaf material was flash frozen in liquid nitrogen and stored at -70 °C prior to freeze-drying. Since large numbers of plants were harvested simultaneously, we had previously determined that materials could be stored for periods of up to 10 weeks prior to freeze-drying without significantly influencing the metabolite profile (data not shown). Lyophilized leaf material was stored in darkness at ambient temperature.

Samples of the frozen leaf material from each plant were taken and their DNA extracted using Plant DNAeasy columns (Qiagen), according to the manufacturer's instructions. Each sample was analysed by Southern blotting after digestion with Hpa1 and probed with a Lys2-BAR fragment (see Figure 2A) to confirm the presence of the correct T-DNA insertion event for that individual.

In the nomenclature we have used here, the T_0 plants are the individuals that were infected by the *Agrobacterium* strain, the T_1 plants are the regenerated primary transgenics; these are genetically different from the T_0 in that they carry a T-DNA, but have not gone through a sexual generation. The T_2 plants are the first selfed generation from the T_1 , and so on.

Preparation of NMR samples

Lyophilized pea leaves were ground into a fine powder using a coffee grinder and sieved to remove large residual particles. 150 ± 1 mg of sample was placed in a crimp top vial and 3 mL of $^2\text{H}_2\text{O}$ (99.9% ^2H ; Goss Scientific) containing 1 mM sodium 3-(trimethylsilyl) propionate- d_4 (TSP; Goss Scientific) added. The samples were mixed on a vibrating platform for 90 min at room temperature. The vials were centrifuged at 2328 **g** for 15 min and the supernatant removed. The supernatant was filtered using a 0.45 μm and a 0.2 μm syringe filter. The pH of the filtered extract was measured to

be 5.75 ± 0.05 . 540 μL of the extract was placed in a 5 mm NMR tube and 60 μL of 10 mM NaN_3 added to inhibit microbial growth.

NMR data acquisition

NMR data were collected on a Bruker ARX500 NMR spectrometer using a 5 mm broad band probe tuned to detect ^1H signals at 500.13 MHz. NMR parameters and the magnetic field homogeneity were optimized using an arbitrary pea leaf extract. The magnetic field was locked on the deuterium signal of the $^2\text{H}_2\text{O}$ and adjusted to homogeneity. The free induction decay (FID) was recorded using a 30° ^1H flip angle determined from a 90° pulse length of 12.1 μs . A relaxation delay of 3.5 s was inserted into the pulse sequence to ensure that quantitative data were acquired. 1024 repetitions of 32 768 complex points were collected over a spectral width of 5154.6 Hz, with the centre of the spectrum at 500.1388035 MHz. The NMR probehead was maintained at a temperature of 300 K and the sample remained static during data collection.

NMR data processing

The FID was reduced to 1024 real points using a sine-bell shaped window function, phase shifted by 90° (cosine). The data were Fourier transformed and an interactive phase correction applied to the spectrum. A baseline correction was applied and the spectrum referenced to the TSP peak at 0 p.p.m., the area of which was set to unity for all processed spectra. Spectral data were saved as ASCII formatted text.

Multivariate statistics

Using the statistical package WINDAS (Kemsley, 1998), covariance PCA and PLS were used to calculate the first 10–15 PC scores and PLS factors for the NMR data. LDA, using the squared Mahalanobis distance between each sample and the group centres, was used to classify each sample using the compressed ^1H NMR data. All samples were included in the PCA/PLS calculation, except one. This was used to validate the model and a classification was made without prior indication of sample identity. This 'leave one out' method was repeated, omitting each of the samples in turn, using the blind samples to determine a success rate for classification during the validation phase. Where PC scores are presented they have been calculated using all of the available data.

A nested analysis of molecular variance (AMOVA) (Excoffier et al., 1992) was performed to investigate the significance of

groups within the T_4 data. A pair-wise sum of squared differences matrix was calculated from the 86 highest variance standardized NMR data points. The nested AMOVA consisted of the following groups. (1) Wild-type (i.e. non-transformed Puget plants) with null-segregants (i.e. all transformed Puget individuals that did not contain the transgenic construct). (2) All transgenic individuals. All eight populations (six transgenic, one null segregant and one wild-type) were also analysed in all pair-wise combinations. The significance of all variance components was tested by a 1000-iterate bootstrap.

Reproducibility

The reproducibility of the extraction protocol and the NMR measurement were assessed. Six extracts of the same pea sample were prepared and the ^1H NMR spectrum recorded. The NMR spectrum of one of the extracts was recorded six times. This was repeated using extracts from two plants of the same genotype. The data was compressed using PCA and the PC scores plot of the first two PCs showed discrete clusters of data for each of the two plants, illustrating that the interplant variation is more significant than that due to analytical error. The variation introduced due to the extraction protocol is greater than that for repeat measurements of the same sample, but both were negligible compared to the variation between the profiles generated from different plants.

Acknowledgements

We wish to thank the UK Food Standards Agency for funding this work (grant number G01017) and Dr J. Godward for help in the preparation of this manuscript. PMM and NE are supported by Biotechnology and Biological Sciences Research Council (BBSRC) Core Strategic Grant to the John Innes Centre.

References

- Bean, S.J., Gooding, P.G., Mullineaux, P.M. and Davies, D.R. (1997) A simple system for pea transformation. *Plant Cell Rep.* **16**, 513–519.
- Bishop, G.J., Carland, F., English, J., Harrison, K., Jones, J.D.G., Scofield, S.R. and Shlumukov, L. (1992) Effective vectors for transformation, expression of heterologous genes, and assaying transposon excision in transgenic plants. *Transgenic Res.* **1**, 285–297.
- Bregitzer, P., Zhang, S.B., Cho, M.J. and Lemaux, P.G. (2002) Reduced somaclonal variation in barley is associated with culturing highly differentiated, meristematic tissues. *Crop Sci.* **42**, 1303–1308.
- Casey, R. and Davies, D.R. (1994) *Peas: Genetics Molecular Biology and Biotechnology*. Wallingford, UK: CAB International.

- Cassells, A.C. and Curry, R.F. (2001) Oxidative stress and physiological, epigenetic and genetic variability in plant tissue culture: implications for micropropagators and genetic engineers. *Plant Cell Tissue Organ Cult.* **64**, 145–157.
- Charlton, A.J. (2001) NMR – novel metabolite research? *LabPlus Int.* **15**, 10–12.
- Charlton, A.J., Farrington, W.H.H. and Brereton, P. (2002) Application of ¹H NMR and multivariate statistics for screening complex mixtures: Quality control and authenticity of instant coffee. *J. Agric. Food Chem.* **50**, 3098–3103.
- Cote, F.X., Teisson, C. and Perrier, X. (2001) Somaclonal variation rate evolution in plant tissue culture: Contribution to understanding through a statistical approach. *In Vitro Cell. Dev. Biol. – Plant*, **37**, 539–542.
- Excoffier, L., Smouse, P.E. and Quattro, J.M. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.
- Fan, T.W.M. (1996) Metabolite profiling by one- and two dimensional NMR analysis of complex mixtures. *Prog. Nucl. Mag. Res. Sp.* **28**, 161–219.
- Fernandez, E.J. and Clark, D.S. (1987) NMR spectroscopy: a non-invasive tool for studying intracellular processes. *Enzyme Microb. Technol.* **9**, 259–271.
- Fiehn, O. (2002) Metabolomics – the link between genotypes and phenotypes. *Plant Mol. Biol.* **48**, 155–171.
- Gaspar, T., Franck, T., Bisbis, B., Kevers, C., Jouve, L., Hausman, J.F. and Dommes, J. (2002) Concepts in plant stress physiology. Application to plant tissue cultures. *Plant Growth Regul.* **37**, 263–285.
- Hall, R., Beale, M., Fiehn, O., Hardy, N., Sumner, L. and Bino, R. (2002) Meeting report. Plant metabolomics: the missing link in functional genomics strategies. *Plant Cell*, **14**, 1437–1440.
- Hossain, M.A., Konisho, K., Minami, M. and Nemoto, K. (2003) Somaclonal variation of regenerated plants in chilli pepper (*Capiscum annuum* L.). *Euphytica*, **130**, 233–239.
- Jaligot, E., Beule, T. and Rival, A. (2002) Methylation-sensitive RFLPs: characterisation of two oil palm markers showing somaclonal variation-associated polymorphism. *Theor. Appl. Genet.* **104**, 1263–1269.
- Joyce, S.M. and Cassells, A.C. (2002) Variation in potato microplant morphology *in vitro* and DNA methylation. *Plant Cell Tissue Organ Cult.* **70**, 125–137.
- Joyce, S.M., Cassells, A.C. and Jain, S.M. (2003) Stress and aberrant phenotypes in *in vitro* culture. *Plant Cell Tissue Organ Cult.* **74**, 103–121.
- Kaepler, S.M., Kaepler, H.F. and Rhee, Y. (2000) Epigenetic aspects of somaclonal variation in plants. *Plant Mol. Biol.* **43**, 179–188.
- Kelly-Borge, M., Robinson, E.V., Gunasekara, S.P., Gunasekara, M., Gulavita, N.K. and Pomponi, S.A. (1994) Species differentiation in the marine sponge genus *Discodermia* (Demospongiae: Lithistida): the utility of ethanol extract profiles as species-specific chemotaxonomic markers. *Biochem. Syst. Ecol.* **22**, 353–365.
- Kemsley, E.K. (1998) *Discriminant Analysis and Class Modelling of Spectroscopic Data*. Chichester, UK: John Wiley & Sons.
- Koncz, C., Martini, N., Mayerhofer, R., Koncz-Kalman, Z., Korber, H., Redei, G.P. and Schell, J. (1989) High frequency T-DNA-mediated gene tagging in plants. *Proc. Natl. Acad. Sci. USA*, **86**, 8467–8471.
- Kubis, S.E., Castilho, A.A.M.F., Vershinin, A.V. and Heslop-Harrison, J.S. (2003) Retroelements, transposons and methylation status in the genome of oil palm (*Elaeis guineensis*) and the relationship to somaclonal variation. *Plant Mol. Biol.* **52**, 69–79.
- Noteborn, H.P.J.M., Lommen, A., van der Jagt, R.C. and Weseman, J.M. (2000) Chemical fingerprinting for the evaluation of unintended secondary metabolic changes in transgenic food crops. *J. Biotechnol.* **77**, 103–114.
- Peraza-Echeverria, S., Herrera-Valencia, V.A. and James-Kay, A. (2001) Detection of DNA methylation changes in micropropagated banana plants using methylation-sensitive amplification polymorphism (MSAP). *Plant Sci.* **161**, 359–367.
- Pillus, L. and Rine, J. (1989) Epigenetic inheritance of transcription states in *S. cerevisiae*. *Cell*, **59**, 637–647.
- Pluhar, S.A., Erickson, L. and Pauls, K.P. (2001) Effects of tissue culture on a highly repetitive DNA sequence (E180 satellite) in *Medicago sativa*. *Plant Cell Tissue Organ Cult.* **67**, 195–199.
- Ratcliffe, R.G. and Shachar-Hill, Y. (2001) Probing plant metabolism with NMR. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **52**, 499–526.
- Ribnicky, D.M., Cohen, J.D., Hu, W.S. and Cooke, T.J. (2002) An auxin surge following fertilization in carrots: a mechanism for regulating plant totipotency. *Planta*, **214**, 505–509.
- Rycroft, D.S. (1996) Fingerprinting of plant extracts using NMR spectroscopy: application to small samples of liverworts. *Chem. Commun.* 2187–2188.
- Scott, A., Woodfield, D. and White, D.W.R. (1998) Allelic composition and genetic background effects on transgene expression and inheritance in white clover. *Mol. Breed.* **4**, 479–490.