From phenotype to genotype: whole tissue profiling for plant breeding

Royston Goodacre,[a,b] Luned Roberts,[d] David I. Ellis,[a,b,c] Danny Thorogood,[d] Stephen M. Reader,[e] Helen Ougham,[d,*] and Ian King[d]

[a]*Institute of Biological Sciences, University of Wales, Aberystwyth, Ceredigion, SY23 3DD, UK*
[b]*School of Chemistry and Manchester Interdisciplinary Biocentre, The University of Manchester, 131 Princess Street, Manchester, M1 7ND, UK*
[c]*School of Chemistry, The University of Manchester, PO Box 88, Sackville Street, Manchester, M60 1QD, UK*
[d]*Plant Genetics & Breeding Department, Institute of Grassland & Environmental Research, Plas Gogerddan, Aberystwyth, SY23 3EB, Wales, UK*
[e]*John Innes Centre, Norwich Research Park, Colney, Norwich, NR4 7UH, UK*

Received 3 January 2007; Accepted 23 April 2007

Fourier transform infrared spectroscopy (FT-IR) was used to obtain 'holistic' metabolic fingerprints from a wide range of plants to differentiate species, population, single plant genotype, and chromosomal constitution differences. Sample preparation simply entailed the maceration of fresh leaves with water, and these samples were then dried and analysed by reflectance FT-IR where spectral acquisition was typically 10 s. All samples gave reproducible, characteristic biological infrared absorption spectra and these were analysed by chemometric methods. FT-IR is not biased to any particular chemical species and thus the whole tissue profiles produced measure the total biochemical makeup of the test sample; that is to say it represents a plant phenotype. We show that by simple cluster analysis these phenotypic measurements can be related to the genotypes of the plants and can reliably differentiate closely related individuals. We believe that this approach provides a valuable new tool for the rapid metabolomic profiling of plants, with applications to plant breeding and the assessment of substantial equivalency for genetically-modified plants.

KEY WORDS: artificial neural network; hierarchical cluster analysis; discriminant function; Lolium; principal components; Triticum.

1. Introduction

Plant identification and differentiation at the species, population and individual genotype levels is of major importance for plant scientists and breeders. Until the advent of genetic markers the most commonly used approach for discriminating between plants had been the analysis of a range of both vegetative and reproductive morphological traits using gravimetric parameters and visual scores. The use of such traits has enabled plant scientists to differentiate plant types with some degree of accuracy. Furthermore, the science of quantitative genetics developed by pioneers such as Fisher (Fisher *et al.*, 1932) and Mather (Mather, 1949) has enabled an understanding of the genetical control of such morphological traits. Analysis is complicated by the fact that the observed phenotype is influenced by genotype and environment and interaction of these two factors. In order to separate the effects of the genotype and environment it is necessary to take measurements from complex pedigrees and to repeat experiments over a number of environmentally diverse sites or time periods.

Much interest has arisen in the last two decades in using genetic markers as a tool to enhance our understanding of the genetic control of morphological traits and our ability to differentiate plant types. Genetic markers have applications in determining genetic diversity and distinctness in natural populations (Schoen and Brown, 1991; Breyne *et al.*, 1999) and also in determining phylogeny of species and populations (Wang *et al.*, 1992). Similarly, they can be used to characterise and compare genetic stocks (Virk *et al.*, 1995) and cultivated plant varieties (Everaert *et al.*, 1993; Rafalski and Tingey, 1993; De Riek *et al.*, 2001; Roldan-Ruiz *et al.*, 2001). Furthermore, markers have a role to play in helping to develop superior cultivars of crop plants through (multiple) QTL selection within breeding populations (Dudley, 1993; Varshney *et al.*, 2006) or introgression of specific chromosome segments from related species (Islam and Shepherd, 1992; King *et al.*, 1993; Tanksley and Nelson, 1996; Korzun *et al.*, 1997; Hernandez, 2005).

Thus genetic markers provide a faster, more discriminating and more cost effective method of plant identification than the analysis of morphological characters. Storage proteins and isozymes were one of the first marker systems used. These biochemical markers are still used for monitoring genetic purity, testing of parentage and as additional aids in trials. However, the use of these markers is rather laborious and, in addition, does not provide sufficient discrimination. This has led to the use of molecular DNA markers which sample a far greater proportion of the genome and, in many

[*] To whom correspondence should be addressed.
E-mail: helen.ougham@bbsrc.ac.uk

cases, are far faster to use than protein-based markers. Molecular DNA markers include the relatively labour intensive but reliable restriction fragment length polymorphisms (RFLPs) and a whole battery of PCR based marker systems such as randomly amplified polymorphic DNA (RAPDs), DNA amplification fingerprinting (DAF), amplified fragment length polymorphisms (AFLPs), microsatellites (SSRs) and single nucleotide polymorphisms (SNPs).

The major advantages of molecular markers are that very large numbers are available and many are highly polymorphic. However, although molecular markers have provided a significant advance in plant identification, plants still need to be screened with a battery of markers which ideally need to be evenly distributed throughout the genome. Thus a rapid automatable method of discriminating between plants, which involves a single test that provides a metabolic snapshot representing the integrated effect of the whole expressed part of the genome, would have considerable advantages over both morphological and marker analysis.

In the post-genomic era, there is greater emphasis in evaluating the functional roles of genes and gene products for cellular characterisation, and this will enable us to understand and ultimately define the organism's phenotype. In such instances, analysis at the level of gene products, such as mRNAs, proteins and metabolites could be of greater relevance (Oliver, 2002). However, whilst each of these individually, and indeed combined, can be very valuable in defining an organism's phenotype we consider that a 'holistic' phenotypic approach would provide a very useful solution to fingerprinting plant tissues and that these phenotypic profiles can be correlated via mathematical transformations to the plant's genotype.

Fourier transform infrared (FT-IR) spectroscopy (Griffiths and de Haseth, 1986) was first proposed in 1991 as a 'whole-organism fingerprinting' (Magee, 1993) method for characterising microorganisms by Naumann and colleagues (Naumann *et al.*, 1991). This non-destructive rapid analytical technique has since been shown to be a valuable tool for the rapid and accurate characterisation of axenically cultured bacteria (Goodacre *et al.*, 1998; Maquelin *et al.*, 2002), including the detection of physiological changes in microorganisms (Goodacre *et al.*, 2000), and single gene knockout strains (Oliver *et al.*, 1998). With particular reference to the analysis of plant materials, FT-IR has been used to screen for cell-wall mutants (Chen *et al.*, 1998), for pollen identification (Pappas *et al.*, 2003) to study the effect of salinity on tomato fruit quality (Johnson *et al.*, 2000) and within *Arabidopsis thaliana* (Yang and Yen, 2002), whilst near IR has been used to investigate the gene regulator *lys3a* in barley (Munck *et al.*, 2001).

For FT-IR, a particular bond absorbs electromagnetic (EM) radiation at a specific wavelength. For example, the infrared spectra of proteins exhibit strong amide I and II absorption bands at 1640 and 1550 cm^{-1} associated with the characteristic stretching of C=O and C–N and the bending of the N–H bond (Stuart, 1997). Therefore by interrogating a plant tissue with EM radiation of many wavelengths in the mid-IR range (usually defined as 4000–600 cm^{-1}) one can construct an infrared absorbance spectrum which is a complex composite of many different vibrational modes of the plethora of diverse biomolecules of the cell wall, membrane, cytoplasm and extracellular polymeric substances (Udelhoven *et al.*, 2000). However these holistic spectral fingerprints are multivariate in nature and so generally uninterpretable to the naked eye and thus require simplification (Chatfield and Collins, 1980) or dimensionality reduction (Tukey, 1977) via some multivariate analysis computational-based approach [see (Martens and Næs, 1989; Manly, 1994; Ripley, 1996; Quackenbush, 2001)]. The goal of these methods is to summarise a large body of data by means of *relatively* few parameters, preferably the two or three which lend themselves to graphical display, with minimal loss of information. This may be achieved by unsupervised clustering methods or supervised pattern recognition techniques (Beavis *et al.*, 2000).

The aim of this study was to use FT-IR spectroscopy to analyse homogenates of whole plant tissues and to elucidate whether these phenotypic fingerprints can be used to discriminate between different plant species, varieties and genotypes and hence provide plant breeders and biologists with a novel screening approach.

2. Materials and methods

2.1. Plants and sample preparation

Three sets of plant experiments were conducted, consisting of FT-IR differentiation of (i) species/genera, (ii) varieties and (iii) wheat nullisomic/tetrasomic lines. All the plants in this study were grown and maintained in humax compost under natural glasshouse conditions. Details of these and how the plants were cultivated are given below:

Set 1 contained 10 individual plants of each of *Lolium multiflorum*, *Festuca glaucescens* and *Festuca pratensis*. All three species are highly heterogeneous outbreeders with an effective self-incompatibility mechanism such that all 10 individuals of each species are unique yet related genotypes. The three species are phylogenetically related and can hybridise with each other. The plants were grown and leaf material harvested from mature 9-month-old plants.

Set 2 contained 10 individual plants of 15 different varieties of *Lolium multiflorum*, *Lolium perenne* and hybrids between these two species. Full details of the varieties, species (or hybrids) and whether the grass was bred for amenity or forage are given in table 1.

Varieties of these species and the hybrids are produced as synthetic populations derived by poly-crossing a small number of original mother plants each with a unique but related genotype. Subsequent generations are then produced by poly-crossing the plants of the previous generation to produce a stable heterogeneous population in Hardy–Weinberg equilibrium. The plants were obtained from seeds grown in 77 square multi-pot trays. Leaf material was harvested from these plants 2 months after germination.

Set 3 contained *Triticum aestivum* nullisomic/tetrasomic lines from chromosomes 1 and 3. For each chromosome six different wheat lines were analysed. For chromosome 1 these were (codes for plots in parenthesis where numbers 1–3 refer to chromosomes A (*Triticum urartu*), B (*Aegilops speltoides*) and D (*Aegilops squarrosa*), respectively): N1AT1B (042), N1AT1D (024), N1BT1A (402), N1BT1D (204), N1DT1A (420), and N1DT1B (240). Likewise, six wheat lines for chromosome 3 were investigated: N3AT3B (042), N3AT3D (024), N3BT3A (402), N3BT3D (204), N3DT3A (420), and N3DT3B (240). For each wheat line six individual plants were grown in a single 10 inch pot and leaf material harvested 6 weeks after germination.

The sample preparation was rapid and straightforward. Upon harvesting, fresh undamaged leaf material was immediately transferred to liquid nitrogen. Leaves were then macerated with a pestle and mortar in liquid N_2. A small aliquot of distilled H_2O was then added to the resultant slurry. The final approximate density of the samples was 800 μg ml^{-1}, and they were stored at −80°C prior to analysis.

2.2. FT-IR spectroscopy

Ten microlitre aliquots of the above plant materials were evenly applied onto an aluminium plate. Prior to analysis the samples were oven-dried at 50°C for 30 min. Samples were run in replicate (Sets 1 and 2 were collected in triplicate, and for Set 3 five replicates were collected). The FT-IR instrument used was the Bruker IFS28 FT-IR spectrometer (Bruker Spectrospin Ltd., Banner Lane, Coventry, UK) equipped with an MCT (mercury–cadmium–telluride) detector cooled with liquid N_2. The aluminium plate was then loaded onto the motorised stage of a reflectance TLC accessory (Timmins *et al.*, 1998). The IBM-compatible PC used to control the IFS28, was also programmed (using OPUS version 2.1 software running under IBM O/S2 Warp provided by the manufacturers) to collect spectra over the wavenumber range 4000–600 cm^{-1}. Spectra were acquired at a rate of 20 s^{-1}. The spectral resolution used was 4 cm^{-1}. To improve the signal-to-noise ratio, 256 spectra were co-added and averaged. Each sample was thus represented by a spectrum containing 882 points and spectra (see figure 1 for examples) were displayed in terms of absorbance as calculated from the reflectance–absorbance spectra using the Opus software [which is based on the Kubelka–Munk theory (Griffiths and de Haseth, 1986)]. These conditions were used for all experiments. To minimise problems arising from baseline shifts the smoothed first derivatives of these spectra were calculated using the Savitzky–Golay algorithm (Savitzky and Golay, 1964) with 5-point smoothing.

2.3. Cluster analysis

The initial stage involved the reduction of the multidimensional FT-IR data by principal components analysis (PCA) (Jolliffe, 1986). PCA is a well-known

Table 1
Details of the 15 grass varieties analysed

Variety	Code	Species	Amenity/forage
Dancer	a	*L. perenne*	A
AberSprite	b	*L. perenne*	A
AberElan	c	*L. perenne*	F
AberCraigs	d	*L. perenne* + some *L. multiflorum* hybrid	F
Barclay	e	*L. perenne*	A
Elka	f	*L. perenne*	A
AberSilo	g	*L. perenne*	F
AberDart	h	*L. perenne*	F
AberComo	i	*L. multiflorum*	F
AberMara	j	*L. perenne*	F
AberElf	k	*L. perenne*	A
AberExcel	l	*L. multiflorum*× *L. perenne* hybrid	F
AberGold	m	*L. perenne*	F
AberImp	n	*L. perenne*	A
Lex86	o	*L. perenne*	A

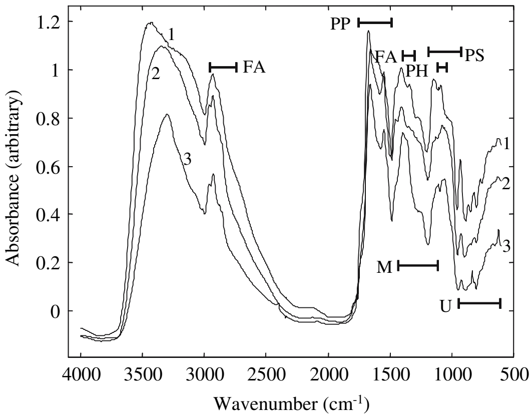

Figure 1. Typical raw FT-IR spectra of *L. multiflorum* (1), *F. pratensis* (2) and *F. glaucescens* (3). Some biochemical regions are highlighted and refer to: FA, fatty acids; PH, phosphates; PP, peptides; PS, polysaccharides; U, unassigned; M, mixed region of proteins, fatty acids, phosphate rich.

technique for reducing the dimensionality of multivariate data whilst preserving most of the variance, and PCA was programmed according to the NIPALS algorithm (Wold, 1966). PCA was also used to cluster ryegrass varieties on the basis of Distinctness, Uniformity and Stability (DUS) character scores. Discriminant function analysis (DFA) then discriminated between groups on the basis of retained principal components (PCs) and some *a priori* knowledge (details are given in the text) (Manly, 1994). Finally, hierarchical cluster analysis (HCA) was performed where the Euclidean distance between *a priori* group centres in DFA space was used to construct a similarity measure, with the Gower general similarity coefficient S_G (Gower, 1966), and these distance measures were then processed by an agglomerative clustering algorithm to construct a dendrogram (Manly, 1994).

For the discrimination of *Lolium multiflorum*, *Lolium perenne* and the two hybrids PC-DFA was performed on training sets (containing the first nine plants analysed) with the *a priori* knowledge of the species and hybrids, and the 'unknown' test data (the 10th plant) were projected into this PC-DFA space as described elsewhere (Radovic *et al*., 2001). Briefly, PCA followed by DFA were carried out on only the training set, the test set spectra were first projected into the PCA space and then the resultant PCs projected into the DFA space. Finally, the resultant training set DFs and the projected DFs from the test set were used to construct a dendrogram by HCA.

All cluster analyses were performed using Matlab version 5.0.0.4069 (The Math Works, Inc., 24 Prime Par Way, Natick, MA, USA), which runs under Microsoft Windows NT on an IBM-compatible PC.

2.4. *Artificial neural network (ANN) analysis of the Lex86 and Dancer grass varieties*

The Lex86 and Dancer grass varieties were chosen because DUS trials fail to separate these two varieties convincingly (Michael S. Camlin, personal communication). The input data for the ANN comprised the first 10 principal component (PC) scores from the FT-IR data and these were partitioned into training and test sets randomly. The use of PCs as inputs to ANNs rather than the full high dimensional spectra has been shown to produce more robust models (Blanco *et al*., 1995; Goodacre *et al*., 2002) as the fewer inputs used lead to the production of a more parsimonious model (Seasholtz and Kowalski, 1993; Kell and Sonnleitner, 1995; Bo and Jonassen, 2002). The training set contained the replicate PC scores from seven plants from each variety and the test set comprised the PC scores from the remaining three plants (details of training and test sets are given in table 2). The output data were binary encoded such that Dancer variety coded as 0 and Lex86 variety as 1. Correct identity for Dancer was taken as output < 0.2, whilst for Lex86 the output must be > 0.8.

Table 2
Identity of the Lex86 and Dancer *L. perenne* grass varieties used in the training and test sets as judged by MLP analysis of their FT-IR data

Training/test set	Variety	Plant	Output
Training	Dancer	a3	0.0
Training	Dancer	a4	−0.1
Training	Dancer	a5	0.0
Training	Dancer	a6	0.1
Training	Dancer	a7	0.0
Training	Dancer	a8	−0.1
Training	Dancer	a9	0.0
Test	Dancer	a0	0.0
Test	Dancer	a1	−0.1
Test	Dancer	a2	−0.1
Training	Lex86	o3	1.1
Training	Lex86	o4	1.0
Training	Lex86	o5	0.9
Training	Lex86	o6	1.0
Training	Lex86	o7	1.1
Training	Lex86	o8	1.1
Training	Lex86	o9	1.0
Test	Lex86	o0	0.9
Test	Lex86	o1	1.1
Test	Lex86	o2	0.8

The ANN used was a multilayer perceptron (MLP) employing log sigmoidals as the transfer functions and standard back-propagation (Rumelhart *et al*., 1986; Bishop, 1995) and was trained using a user-friendly, neural network simulation program, NeuFrame version 3,0,0,0 (Neural Computer Sciences, Lulworth Business Centre, Nutwood Way, Totton, Southampton, Hants), which runs under Microsoft Windows NT on an IBM-compatible personal computer as detailed by us elsewhere (Goodacre *et al*., 1998, 2000).

3. Results

All plant samples gave reproducible characteristic biological infrared absorption spectra with recognisable amide I and II protein vibration bands, acyl vibration bands from fatty acids and polysaccharide and nucleic acid vibration bands (see figure 1 for examples). However, these FT-IR spectra (and indeed all the others collected) show broad and complex contours highlighting there was very little *qualitative* difference between the spectra, although at least some complex *quantitative* differences between them were observed. Such spectra, essentially uninterpretable by the naked eye, readily illustrate the need to employ multivariate statistical techniques for their analyses.

3.1. *FT-IR analysis of L. multiflorum, F. glaucescens and F. pratensis*

Discriminant function analyses (DFA) was used to observe the relationships between the three different species of plants in the first set as judged from their derivatised FT-IR spectra, and DFA was performed as detailed above. The *a priori* class structure that was used in DFA was each individual plant (that is to say 30 groups) and this allowed any natural relationships between the 30 plants to be elucidated. The resulting DFA plot is shown in figure 2 which clearly shows that the *L. multiflorum*, *F. glaucescens* and *F. pratensis* replicated genotypes fall into three clusters with each cluster being comprised of genotypes from a single species. In addition it shows that *F. pratensis* and *F. glaucescens* genotypes are more closely related to each other (being discriminated in the second DF rather than the first) than to *L. multiflorum*. However, this is not unexpected as *F. pratensis* and *F. glaucescens* both belong to the same genus *Festuca* L. (the *Festuceae*) whilst *L. multiflorum* belongs to *Lolium* L.

In a blind experiment FT-IR spectra from a single genotype were selected at random from each of the three species and PC-DF-HCA projection analysis was performed. It was found that it was not only possible to determine which species the sample derived from but also the genotype of the plant selected (data not shown).

3.2. FT-IR analysis of the 15 Lolium varieties

Initial DFA cluster analysis used *a priori* coding according to the 150 plants analysed and although some separation was seen according to species, the results were not clear cut because of the plant-to-plant variability. Whilst there is an inherent background biological variation in plants grown in controlled environments which has been observed by Fiehn and colleagues using GC-MS on isogenic plants (Fiehn *et al.*, 2000), this variation was exacerbated by the fact that these plants are outbreeders and hence show large amounts of natural heterogeneity. There is therefore a need to compensate for this variability in the cluster analyses in order to observe the true underlying phenotypic, and hopefully genotypic, structure.

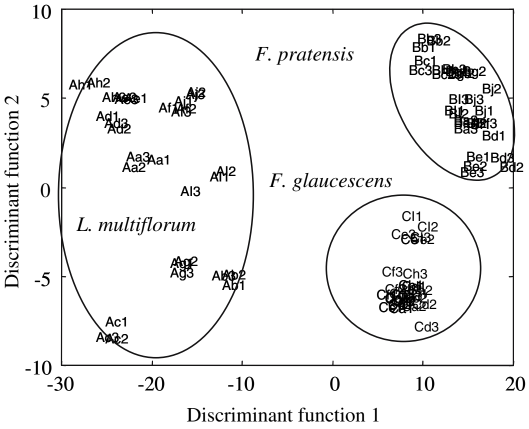

Figure 2. Discriminant function analysis on FT-IR data from 10 *L. multiflorum*, 10 *F. glaucescens* and 10 *F. pratensis* plants. The *a priori* knowledge used in the construction of PC-DFA was the 30 plants and not whether they were *L. multiflorum*, *F. glaucescens* or *F. pratensis*.

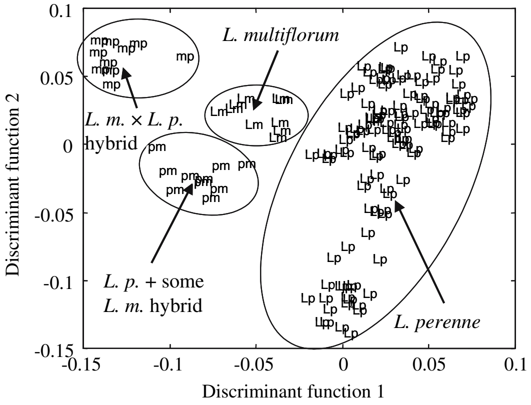

Figure 3. Discriminant function analysis on FT-IR data from 15 grass species showing the speciation into *L. multiflorum*, *L. perenne* and two different hybrids. The *a priori* knowledge used in the construction of PC-DFA was the 15 varieties and not the species or hybrid status of those varieties.

Cluster analysis of the 15 *Lolium* varieties (table 1) using one group per variety resulted in the production of four clusters (figure 3). The largest comprised all the straight *L. perenne* genotypes, including both amenity and forage varieties. The three remaining clusters were composed exclusively of either the AberCraig genotypes (*L. perenne* plus some *L. multiflorum* hybrid), AberExcel (the *L. perenne*×*L. multiflorum* hybrid) or AberComo (the pure *L. multiflorum* variety). Thus the analysis was able to discriminate between *L. perenne*, *L. multiflorum* and the partial and complete *L. perenne*×*L. multiflorum* hybrids. Although the recovery of *L. perenne* into one cluster was encouraging (since this was encoded as 12 separate groups in DFA), the other three clusters might be construed in a rather subjective way. Therefore, in order to test the ability of the FT-IR approach to identify an unknown plant to the species level in an objective fashion, PC-DFA was performed on plants 1–9 for each of the 15 *Lolium* varieties and the *a priori* knowledge used was four classes (one for each of *L. perenne*, *L. multiflorum* and the partial and complete hybrids). The 10th plant was used as an 'unknown' test set then projected into this PC-DFA space and the resultant DFs used to construct a dendrogram. The dendrogram (figure 4) shows that four groups are recovered; one containing all the *L. multiflorum* spectra, including the projected plant spectrum marked with a <; another two small clusters were recovered which contained the two hybrids and the projected plant spectra were recovered in their prospective clusters. Finally, all 12 *L. perenne* clustered in one single large group along with the 12 test spectra that were projected into this dendrogram. This confirmed that the cluster analysis was reproducible as each of the 15 genotypes were placed into the correct species or hybrid clusters. Note that we chose to randomly select one-tenth of the data as the test set because preliminary analysis showed

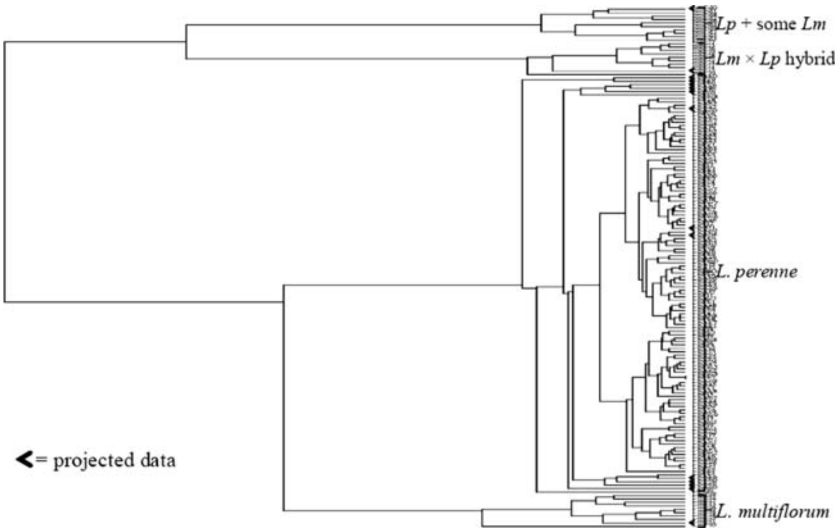

Figure 4. Projection of the 10th plant from each of the 15 grass species into PC-DF-HCA space. The *a priori* knowledge in the training data was four groups relating to *L. multiflorum*, *L. perenne* and the two different hybrids.

that constructing the cluster analysis on fewer biological replicates led to less robust predictions for the test data. We believe this is due to the inherent plant biological variability because these grasses were grown in non-standardised environments (i.e., natural greenhouse, where light and water were not rigorously controlled). One would expect the clustering to improve if highly controlled growth conditions were used, but we believe the environments we used are closer to normal field conditions.

The next stage was to analyse the 12 *L. perenne* genotypes alone using DFA with the *a priori* knowledge being 12 classes for each different variety. Figure 5 shows that the separation of the *L. perenne* genotypes into forage and amenity varieties was not possible using cluster analysis. Two of the forage varieties, AberElan (coded as 'c' in figure 5) and AberDart (h), were separated from the remaining forage and amenity varieties (Cluster II in figure 5b). The remaining three forage varieties, AberGold (m), AberMara (j) and AberSilo (g), were clustered together (Cluster I in figure 5b) and surrounded by clusters of amenity varieties. The probable reason for this grouping is discussed later. Projection analysis of the 10th plant from each of the varieties into PC-DFA calibrated with the *a priori* knowledge of whether the plant was amenity (class 1) or forage (class 2) successfully identified the amenity or forage status of the *L. perenne* (data not shown).

In contrast to the FT-IR data, morphological and other plant phenotype characters scored for Distinctiveness, Uniformity and Stability (DUS) assessment (table 3) did not provide effective resolution between any of the forage perennial ryegrass varieties, though they separated them well from the amenity grasses, forage Italian ryegrass and hybrid ryegrass varieties

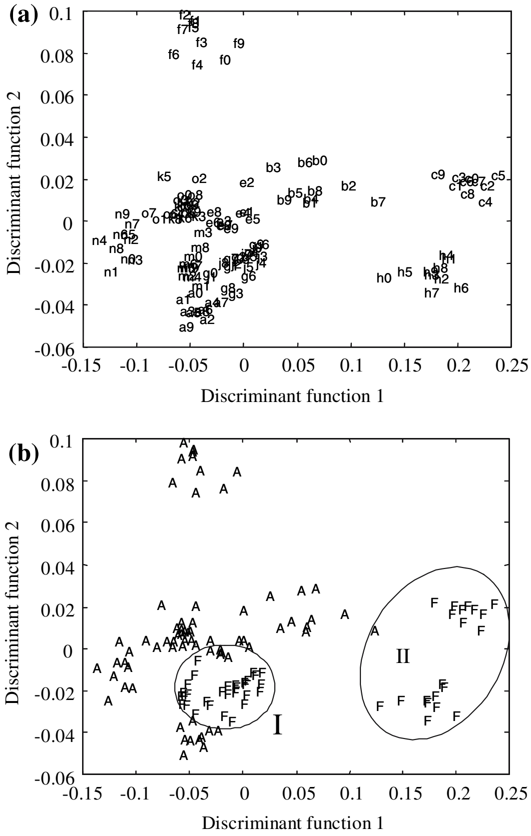

Figure 5. Discriminant function analysis on FT-IR data from the 12 *L. perenne* varieties (a) shows the codes as detailed in table 1, whilst (b) shows whether the plants were grown for amenity (A) or forage (F). The *a priori* knowledge used in the construction of PC-DFA was the 12 varieties. It is notable in (b) that the higher yielding lower ground cover forage varieties (cluster I) cluster separately from those forage varieties breed for ground cover (cluster II). These ground cover forage grasses cluster with the amenity grasses which are also bred for ground cover.

Table 3
Ten DUS characters measured for National Listing of Cultivars. Variety codes as in table 1 (Plant Testing Station, 1995)

Variety	Code	Growth habit in year of sowing	Height of plant in Spring	Mean date of ear emergence	Height of plant at ear emergence	Width of plant at ear emergence	Length of flag leaf at ear emergence	Width of flag leaf at ear emergence	Length of longest stem 30 days after ear emergence	Length of ear	Number of spikelets per ear
AberSprite (Ba11972)	b	9	1	9	5	1	1	1	1	1	1
AberImp (Ba11986)	n	7	1	7	5	1	1	1	1	1	1
AberExcel (BaB455)	l	5	5	5	1	7	5	1	3	7	1
AberGold (Ba11368)	m	9	7	5	7	5	5	3	7	7	7
AberElf (Ba11773)	k	5	1	9	3	1	1	1	3	3	1
AberCraigs (Ba11927)	d	5	7	7	9	7	7	7	9	9	5
AberMara (Ba10879)	j	9	7	5	7	5	7	5	7	7	7
AberSilo (Ba11317)	g	9	9	5	7	7	5	5	7	7	7
AberElan (Ba10761)	c	9	7	5	7	5	7	5	7	7	7
Dancer	a	5	3	5	3	1	1	1	3	3	1
Lex86	o	5	3	5	3	1	1	1	3	3	1
AberDart (Ba11778)	h	9	7	5	7	5	5	5	7	7	7
AberComo (Bb2042)	i	5	5	5	5	3	1	1	5	3	5
AberCraigs (Ba11927)	d	5	7	7	9	7	7	7	9	9	5
AberMara (Ba10879)	j	9	7	5	7	5	7	5	7	7	7
AberSilo (Ba11317)	g	9	9	5	7	7	5	5	7	7	7
AberElan (Ba10761)	c	9	7	5	7	5	7	5	7	7	7
AberGold (Ba11368)	m	9	7	5	7	5	5	3	7	7	7
AberDart (Ba11778)	h	9	7	5	7	5	5	5	7	7	7

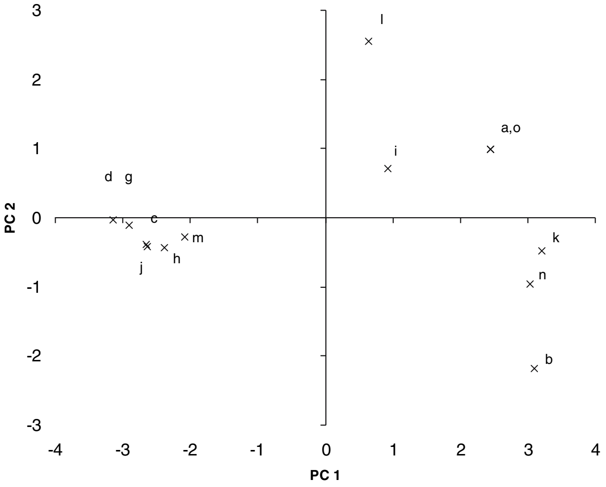

Figure 6. First two principal components for 10 DUS characters measured for National Listing of Cultivars. Variety codes as in table 1. First component accounts for 70.7% and second component, 12.9% of total variation. Characters measured were: Growth habit in year of sowing, Height of plant in Spring, Mean date of ear emergence, Height of plant at ear emergence, Width of plant at ear emergence, Length of flag leaf at ear emergence, Width of flag leaf at ear emergence, Length of longest stem 30 days after ear emergence, Length of ear, Number of spikelets per ear (Plant Testing Station, 1995).

(figure 6). A plot of the first two principal components, which together accounted for over 93% of the total variation, derived from ten DUS characters clustered all forage perennial ryegrasses (including Abercraigs (d), which contains some *L. multiflorum* genetic material) into a single group.

Since it was now possible to discriminate between the amenity and forage varieties a further analysis was performed on the *L. perenne* amenity varieties and this revealed essentially three clusters (figure 7): One comprised Lex86 and Dancer (two North American varieties), one was composed of only Elka (a European variety), and the remaining cluster contained Barclay (another European variety) and three closely related varieties, AberImp, AberElf and AberSprite, all of which contained germplasm derived from Barclay.

In order to test the resolution of the system a further attempt was made to determine if FT-IR could be used to distinguish between the two closely related North American amenity varieties, Lex86 and Dancer. The initial PC-DFA performed, encoding as 10 groups per variety, provided no real discrimination between the two varieties (figure 8a). However, when two groups for each variety were used there was clear differentiation between Lex86 and Dancer in the first DF (figure 8b), suggesting that FT-IR could be used to discriminate between these two varieties but only when some of the background (due to outbreeding) biological plant-to-plant variation was disregarded in the calculation of the discrimination function. Finally, ANNs trained with seven of the plants (details are given in table 2, and protocol in the Materials and Methods) successfully predicted whether the three remaining plants in the test set were Lex86 or Dancer (table 2).

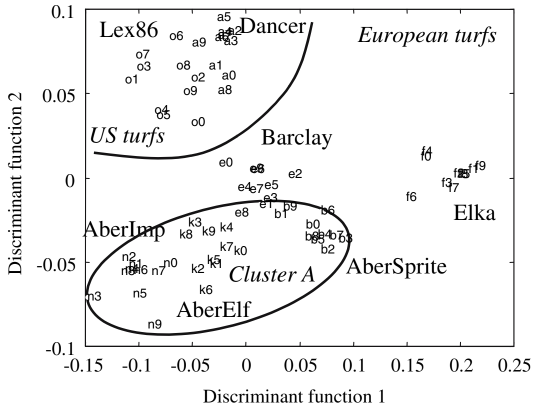

Figure 7. Discriminant function analysis on FT-IR data from the 7 *L. perenne* amenity varieties. The codes used are those as detailed in table 1. The *a priori* knowledge used in the construction of PC-DFA was the seven varieties.

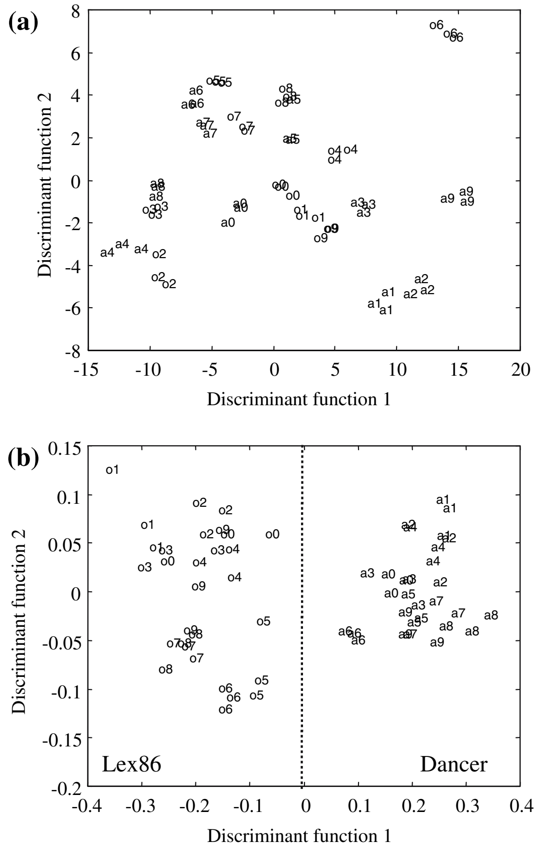

Figure 8. Discriminant function analysis on FT-IR data from *L. perenne* amenity varieties Lex86 and Dancer. The *a priori* knowledge used in the construction of PC-DFA was the 10 individual plants (a) or two groups (b) for each variety. See also table 1 for ANN classification of these plants.

3.3. FT-IR analysis of the Triticum aestivum nullisomic/tetrasomic lines

In a third experiment wheat groups 1 and 3 nullisomic/tetrasomic lines were analysed by FT-IR. Wheat is an allohexaploid ($2n = 6\times = 42$; $1\times = 7$), i.e., it is composed of three genomes from three different species. The three genomes are given the symbols A (*Triticum urartu*), B (*Aegilops speltoides*) and D (*Aegilops tauschii*) with each genome contributing 14 chromosomes (diploid chromosome number). The chromosomes of the three genomes can be divided up into seven homologous groups of six chromosomes. Thus homologous group 1 contains two 1A, two 1B and two 1D chromosomes. All the group 1 chromosomes are syntenic, i.e., they carry the same genes (or alleles of the same genes) in the same order along the chromosome. The loss of a pair of group 1 chromosomes such as the 1A chromosomes would be detrimental to plant vigour and possibly lead to death. However, as a result of the syntenic relationship of the chromosomes belonging to the same homologous group, the addition of an extra pair of 1B or 1D chromosomes will compensate for the loss of the pair of 1A chromosomes. Thus a plant missing a pair of 1A chromosomes, for example, but carrying an extra pair of compensating 1B chromosomes (i.e., four 1B chromosomes) is known as nullisomic 1A tetrasomic 1B or N1AT1B.

All six nullisomic/tetrasomic group 1 lines, N1AT1B, N1AT1D, etc., were analysed by FT-IR and the plants were separated into three clusters (figure 9a). The three clusters were each composed of gentoypes from two nullisomic/tetrasomic lines, i.e., N1AT1B (coded on figure 8a as '042') and N1AT1D (024); N1BT1A (402) and N1BT1D (204); N1DT1A (420) and N1DT1B (240). Thus the analysis performed was able to discriminate between plants which had either lost chromosomes 1A, 1B or 1D. Further work using ANNs revealed that 50% of the time the analysis could distinguish which compensating group 1 chromosome was tetrasomic (data not shown).

An additional experiment to analyse the wheat group 3 nullisomic/tetrasomic lines was also performed. However, in contrast to the three clusters observed in the group 1 nullisomic/tetrasomic experiment, five clusters were seen (figure 9b). The genotypes lacking chromosomes 3A, i.e., N3AT3B and N3AT3D, clustered together as expected. However, the N3BT3A, N3BT3D, N3DT3A and N3DT3B genotypes formed four separate clusters. This result was initially unexpected as the previous work would have predicted that these four genotypes would have formed two separate clusters, one composed of the plants lacking chromosomes 3B, i.e., N3BT3A and N3DT3B and the other lacking chromosome 3D, i.e., N3DT3A and N3DT3B. The fact that these four genotypes formed four distinct clusters indicated that these plants had been discriminated by some other factor.

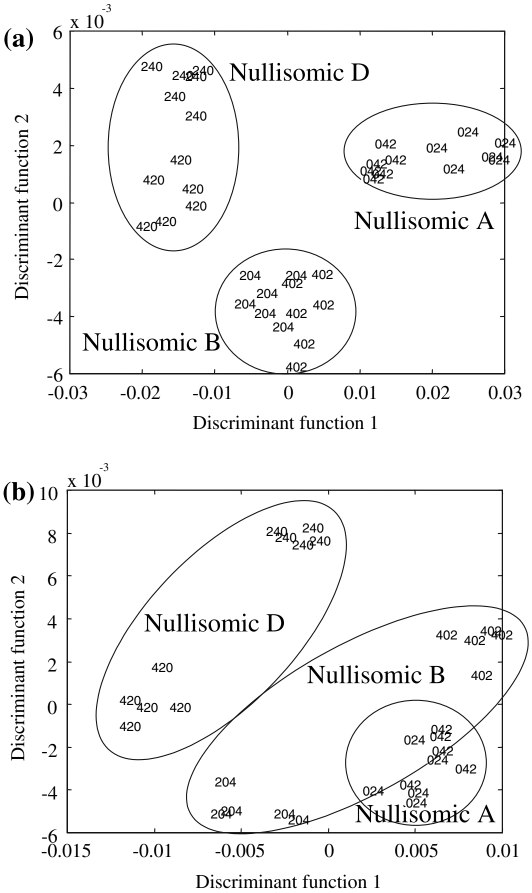

Figure 9. Discriminant function analysis on FT-IR data from wheat nullisomic/tetrasomic lines from chromosomes 1 (a) and 3 (b). The *a priori* knowledge used in the construction of PC-DFA was 2 groups for each line.

4. Discussion

The work described clearly showed that it was possible to distinguish between the three species *L. multiflorum*, *F. glaucescens* and *F. pratensis*. The fact that FT-IR was able to differentiate between the three species is not unexpected considering the level of discrimination possible as described in the other experiments. However, FT-IR did place the two most clearly related species belonging to the genus *Festuca* L closer together than to *Lolium* belonging to the genus *Lolium* L. FT-IR analysis is dependent on phenotypic markers (proteins, nucleic acids, lipids, polysaccharides and other primary and secondary metabolites), which depend on the genes that code for them. Furthermore, gene expression can be of course altered in various ways, e.g., plant age, environment, etc., unlike the gene sequences which remain fixed until a mutation occurs. Thus because the FT-IR fingerprints could potentially be very variable because they are phenotypic (i.e., genotypic + environment) the relationship identified by FT-IR of *F. glaucescens* and *F. pratensis* was encouraging. However, it remains to be seen if FT-IR could be a complementary strategy for determining species relationships and evolution, although within microbiology dendrograms constructed from FT-IR and 16S rDNA sequences are very similar (Tintelnot *et al*., 2000; Kirschner *et al*., 2001).

FT-IR clearly discriminated between *L. perenne*, *L. multiflorum*, and the two *L. perenne*/*L. multiflorum* hybrids. The fact that *L. perenne* and *L. multiflorum* clustered closely together is not unexpected as the species are closely related. In contrast, it is difficult to draw conclusions on the clustering of the partial and complete *L. perenne*×*L. multiflorum* hybrids as it is not known what effect *L. perenne* and *L. multiflorum* genes together in the same plant would have on gene expression and hence the phenotype. That FT-IR was able to place additional samples of the 15 varieties used in this work by projection into PC-DF-HCA into the correct clusters additionally demonstrates the system's ability to discriminate plants at the species level, i.e., between *L. perenne*, *L. multiflorum*, and hybrids thereof.

The primary goal of plant breeders (of all crop species) is to develop new varieties, which in some way out perform existing ones. New varieties of many crops have to go through tests for distinctiveness, uniformity and stability (DUS). Distinctiveness involves the demonstration that a new variety is genetically different from a pre-existing variety. Thus, this test provides a safeguard which prevents anyone from taking a pre-existing variety and marketing it as their own under another name. Whilst genetic markers are one way by which the time for testing distinctiveness can be reduced, they can be a time consuming process.

In the work here, we have used perennial ryegrass (*L. perenne*) to demonstrate how FT-IR can greatly reduce the time required to test for distinctiveness. *L. perenne* is a temperate grass used as both a forage (grazed by ruminants in the field, and used in silage production) and an amenity grass. The latter category encompasses lawn grass, sports turf for golf courses and stadia, and other non-feed uses. Desirable traits for a variety will vary depending on its intended use. For example, good ground cover may be a trait of interest in both a forage and an amenity variety, which may be superficially similar in phenotype though underlying traits such as cell wall digestibility may differ greatly. The development and release of new *Lolium* varieties is particularly difficult with regard to determining the distinctiveness and hence obtaining breeders rights. The problem of variety identification in *L. perenne* is, first, that it is an outbreeder and hence each variety is heterogeneous. Second, variety improvement, particularly with the amenity varieties, has been based on a limited gene pool. These two factors have made it difficult to distinguish different varieties by conventional morphological and field analysis. For example, the amenity grasses Lex86 and Dancer required 3 years of field trials before they could be identified as distinct varieties, and cannot be discriminated on the basis of morphological and yield characters measured for DUS assessment (table 3, figure 6). By contrast, FT-IR was able to distinguish between these two varieties in a single experiment in less than a day, post-growth. The fact that FT-IR is far quicker at determining distinctiveness means that its application could complement DUS procedures and ultimately reduce the need for extensive field assessment of every candidate variety submitted to the European Union Common Catalogue of Varieties of Agricultural Species.

The analysis of the five forage grasses is of particular interest. All the five forage varieties were derived from similar germplasm. However, AberElan and AberDart were selected for high yield. By contrast, AberGold, AberMara and AberSilo were selected for higher ground cover. It is interesting to note that these three varieties clustered with the amenity varieties which are also selected for ground cover rather than yield (% dry matter). This result indicates that it may be possible to use FT-IR to distinguish plants that have been bred for different characters. This is particularly significant since the DUS characters traditionally measured for confirmation of variety uniqueness—as legally required, for example, under the UK's National Listing procedures—do not provide good discrimination between cultivars. As figure 6 showed, DUS characters could successfully discriminate between forage perennial ryegrasses on the one hand, and amenity varieties, forage Italian ryegrasses and forage hybrid ryegrasses on the other, with the latter two categories each also well separated. However, all six forage perennial ryegrass varieties clustered together in a PCA analysis based on 10 DUS characters (figure 6), consistent with the similarity in scores between these varieties; all had scored highly for height in spring, height and width at leaf emergence, length of ear and longest stem, and number of spikelets. While the DUS characters did not provide a sufficient basis for separating the high-yielding AberElan and AberDart from the high ground cover group, FT-IR analysis was very effective in this regard (figure 5). This result suggests that the biochemical composition of the plant as analysed by FT-IR is more useful than DUS scores in discriminating high-yielding from high ground cover varieties within a species. Further work is now required to determine if FT-IR can be used as a selection tool. Of particular interest will be the possibility of selecting parental plants based on FT-IR to develop new higher yielding or higher ground cover varieties.

In addition to being distinct new varieties, new plants also have to be shown to be uniform before they are accepted onto recommended lists. A fast and accurate method of assessing the uniformity of a variety is therefore of interest to both breeders and national organisations that test for uniformity. The FT-IR analysis revealed in cluster analysis that the distribution of AberSprite genotypes was very heterogeneous as compared to AberElf and AberImp. It is therefore important to note that AberSprite, although found to be satisfactory during uniformity testing showed a relatively wide distribution of heading date times compared to AberElf and AberImp. This raises the possibility that FT-IR analysis can be used as a test for uniformity by both breeders wishing to assess the status of their varieties prior to submitting them for testing and also for national testing organisations.

The analysis of the group 1 nullisomic/tetrasomic wheat lines revealed that FT-IR can not only distinguish between genotypes from different species and varieties but also discriminate plants lacking a specific chromosome. In the case of the chromosome group 1 nulli/tetrasomic genotypes, each nullisomic type (lacking chromosome 1A, 1B or 1D) could be clearly distinguished from the other two types. The analysis of the group 3 nullisomic/tetrasomic genotypes revealed that FT-IR could in some cases also discriminate between plants carrying an extra copy of a specific chromosome. While all plants lacking chromosome 3A were clustered, those lacking either 3B or 3D were further sub-classified according to the chromosome for which they were tetrasomic, so that the six nulli/tetrasomic combinations were grouped into five distinct clusters. The fact that the N3BT3A, N3BT3D, N3DT3A and N3DT3B genotypes were placed into four discrete groups indicates that these plants are being discriminated by some factor in addition to the chromosome which they lack. PC-DFA projection analysis of the sixth plant in each category into PC-DFA space calibrated on the first five plants confirmed this (data not shown). The basis for this additional discrimination is not known, but the results suggest that additional copies of each of the other group 3 chromosomes in backgrounds nullisomic for 3B or 3D have unique effects on gene expression as manifested in the FT-IR data. Further work will be required to test this hypothesis.

Wheat nullisomic/tetrasomic, ditelocentric, double ditelocentric, isochromosome, monosomic substitution and alien introgression lines are important tools in wheat genetics. However, they require cytological maintenance which is a skilled and very time consuming process. In addition, cytological analysis frequently needs verification using genetic markers. Thus the analysis described of nullisomic/tetrasomic lines demonstrates that FT-IR has the potential to prove an important tool for the identification of cytological stocks and alien introgression lines in wheat and, by extension, in other plant species.

5. Concluding remarks

We have shown that FT-IR analysis of plant tissues presents a novel rapid technological approach for use by plant biologists and breeders for differentiating between plants at the species, variety, genotype and, in some cases, chromosomal levels by using their metabolic fingerprints. The system described has two major advantages over 'traditional' molecular marker techniques. First the sample preparation is simple and does not require DNA extraction, which can be laborious in labs lacking high-throughput facilities. Second, a single run results in the generation of an information rich spectrum (containing ~900 data points) which is a metabolic fingerprint of the plant material analysed representing a wide range of compounds. By contrast, genetic markers are usually not multiplexable and following DNA extraction the lengthy laborious screening of *many* single markers is undertaken.

Acknowledgements

RG thanks the UK BBSRC Engineering and Biological Systems Committee for financial support. IGER is a BBSRC-sponsored institute.

References

Beavis, R.C., Colby, S.M., Goodacre, R., Harrington, P.B., Reilly, J.P., Sokolow, S. and Wilkerson, C.W. (2000). Artificial intelligence and expert systems in mass spectrometry in Meyers, R.A. (Ed), *Encyclopedia of Analytical Chemistry*. John Wiley, Chichester, pp. 11558–11597.

Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

Blanco, M., Coello, J., Iturriaga, H., Maspoch, S. and Redon, M. (1995). Artificial neural networks for multicomponent kinetic determinations. *Anal. Chem.* **67**, 4477–4483.

Bo, T.H. and Jonassen, I. (2002). New feature subset selection procedures for classification of expression profiles. *Genome Biol.* **3**, research0017.0011-0017.0011.

Breyne, P., Rombaut, D., Van Gysel, 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.

Chatfield, C. and Collins, A.J. (1980). *Introduction to Multivariate Analysis*. Chapman & Hall, London.

Chen, L.M., Carpita, N.C., Reiter, W.D., Wilson, R.H., Jeffries, C. and McCann, M.C. (1998). A rapid method to screen for cell-wall mutants using discriminant analysis of Fourier transform infrared spectra. *Plant J.* **16**, 385–392.

De Riek, J., Calsyn, E., Everaert, I., Van Bockstaele, E. and De Loose, M. (2001). AFLP based alternatives for the assessment of Distinctness, Uniformity and Stability of sugar beet varieties. *Theor. Appl. Genet.* **103**, 1254–1265.

Dudley, J.W. (1993). Molecular markers in plant improvement – manipulation of genes affecting quantitative traits. *Crop Sci.* **33**, 660–668.

Everaert, I., De Riek, J., De Loose, M., Van Waes, J. and Van Bockstaele, E. (1993). Most similar variety grouping for distinctness evaluation of flax and linseed (*Linum usitatissimum* L.) varieties by means of AFLP and morphological data. *Plant Var. Seeds* **14**, 69–87.

Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R.N. and Willmitzer, L. (2000). Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* **18**, 1157–1161.

Fisher, R.A., Immer, F.R. and Tedin, O. (1932). The genetical interpretation of statistics of the third degree in the study of quantitative inheritance. *Genetics* **17**, 107–124.

Goodacre, R., Timmins, E.M., Burton, R., Kaderbhai, N., Woodward, A.M., Kell, D.B. and Rooney, P.J. (1998). Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks. *Microbiol. UK* **144**, 1157–1170.

Goodacre, R., Shann, B., Gilbert, R.J., Timmins, É.M., McGovern, A.C., Alsberg, B.K., Kell, D.B. and Logan, N.A. (2000). The detection of the dipicolinic acid biomarker in *Bacillus* spores using Curie-point pyrolysis mass spectrometry and Fourier transform infrared spectroscopy. *Anal. Chem.* **72**, 119–127.

Goodacre, R., Radovic, B.S. and Anklam, A. (2002). Progress toward the rapid non-destructive assessment of the floral origin of European honey using dispersive Raman spectroscopy. *Appl. Spectrosc.* **56**, 521–527.

Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**, 325–338.

Griffiths, P.R. and de Haseth, J.A. (1986). *Fourier Transform Infrared Spectrometry*. John Wiley, New York.

Hernandez, P. (2005). Comparison among available marker systems for cereal introgression breeding: a practical perspective. *Euphytica* **146**, 95–100.

Islam, A.K.M.R. and Shepherd, K.W. (1992). Production of wheat-barley recombinant chromosomes through induced homoeologous pairing. 1. Isolation of recombinants involving barley arm-3HL and arm-6HL. *Theor. Appl. Genet.* **83**, 489–494.

Johnson, H.E., Gilbert, R.J., Winson, M.K., Goodacre, R., Smith, A.R., Rowland, J.J., Hall, M.A. and Kell, D.B. (2000). Explanatory analysis of the metabolome using genetic programming of simple, interpretable rules. *Genet. Program. Evolvable Mach.* **1**, 243–258.

Jolliffe, I.T. (1986). *Principal Component Analysis*. Springer-Verlag, New York.

Kell, D.B. and Sonnleitner, B (1995). GMP – Good Modelling Practice: an essential component of Good Manufacturing Practice. *Trends Biotechnol.* **13**, 481–492.

King, I.P., Purdie, K.A., Rezanoor, H.N., Koebner, R.M.D., Miller, T.E., Reader, S.M. and Nicholson, P. (1993). Characterization of *Thinopyrum bessarabicum* chromosome segments in wheat using random amplified polymorphic DNAs (RAPDs) and genomic *in situ* hybridization. *Theor. Appl. Genet.* **86**, 895–900.

Kirschner, C., Maquelin, K., Pina, P., Ngo Thi, N.A., Choo-Smith, L.-P., Sockalingum, G.D., Sandt, C., Ami, D., Orsini, F., Doglia, S.M., Allouch, P., Mainfait, M., Puppels, G.J. and Naumann, D. (2001). Classification and identification of enterococci: a comparative phenotypic, genotypic, and vibrational spectroscopic study. *J. Clin. Microbiol.* **39**, 1763–1770.

Korzun, V., Borner, A., Worland, A.J., Law, C.N. and Roder, M.S. (1997). Application of microsatellite markers to distinguish inter-varietal chromosome substitution lines of wheat (*Triticum aestivum* L.). *Euphytica* **95**, 149–155.

Magee, J.T. (1993). Whole-organism fingerprinting in Goodfellow, M. and O'Donnell, A.G. (Eds), *Handbook of New Bacterial Systematics*. Academic Press, London, pp. 383–427.

Manly, B.F.J. (1994). *Multivariate Statistical Methods: A Primer*. Chapman & Hall, London.

Maquelin, K., Kirschner, C., Choo-Smith, L.-P., van den Braak, N., Endtz, H.P., Naumann, D. and Puppels, G.J. (2002). Identification of medically relevant microorganisms by vibrational spectroscopy. *J. Microbiol. Meth.* **51**, 255–271.

Martens, H. and Næs, T. (1989). *Multivariate Calibration*. John Wiley, Chichester.

Mather, K. (1949). *Biometrical Genetics*. Dover Publications, New York.

Munck, L., Nielsen, J.P., Moller, B., Jacobsen, S., Sondergaard, I., Engelsen, S.B., Norgaard, L. and Bro, R. (2001). Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics. *Anal. Chim. Acta* **446**, 171–186.

Naumann, D., Helm, D. and Labischinski, H. (1991). Microbiological characterizations by FT-IR spectroscopy. *Nature* **351**, 81–82.

Oliver, S.G. (2002). Functional genomics: lessons from yeast. *Philos. T. Roy. Soc. B* **357**, 17–23.

Oliver, S.G., Winson, M.K., Kell, D.B. and Baganz, F. (1998). Systematic functional analysis of the yeast genome. *Trends Biotechnol.* **16**, 373–378.

Pappas, C.S., Tarantilis, P.A., Harizanis, P.C. and Polissiou, M.G. (2003). New method for pollen identification by FT-IR spectroscopy. *Appl. Spectrosc.* **57**, 23–27.

Quackenbush, J. (2001). Computational analysis of microarray data. *Nat. Rev. Genet.* **2**, 418–427.

Radovic, B.S., Goodacre, R. and Anklam, E. (2001). Contribution of pyrolysis mass spectrometry (Py-MS) to authenticity testing of honey. *J. Anal. Appl. Pyrol.* **60**, 79–87.

Rafalski, J.A. and Tingey, S.V. (1993). Genetic diagnostics in plant-breeding – RAPDs, microsatellites and machines. *Trends Genet.* **9**, 275–280.

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Roldan-Ruiz, I., van Eeuwijk, F.A., Gilliland, T.J., Dubreuil, P., Dillmann, C., Lallemand, J., De Loose, M. and Baril, C.P. (2001). A comparative study of molecular and morphological methods of describing relationships between perennial ryegrass (*Lolium perenne* L.) varieties. *Theor. Appl. Genet.* **103**, 1138–1150.

Rumelhart, D.E., McClelland, J.L. and The PDP Research Group (1986). *Parallel Distributed Processing, Experiments in the Microstructure of Cognition, Vol I and II*. MIT Press, Cambridge, MA.

Savitzky, A. and Golay, M.J.E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **36**, 1627–1633.

Schoen, D.J. and Brown, A.H.D. (1991). Intraspecific variation in population gene diversity and effective population-size correlates with the mating system in plants. *Proc. Natl. Acad. Sci. USA* **88**, 4494–4497.

Seasholtz, M.B. and Kowalski, B. (1993). The parsimony principle applied to multivariate calibration. *Anal. Chim. Acta* **277**, 165–177.

Stuart, B. (1997). *Biological Applications of Infrared Spectrocopy*. John Wiley & Sons, Chichester.

Tanksley, S.D. and Nelson, J.C. (1996). Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. *Theor. Appl. Genet.* **92**, 191–203.

Timmins, É.M., Howell, S.A., Alsberg, B.K., Noble, W.C. and Goodacre, R. (1998). Rapid differentiation of closely related *Candida* species and strains by pyrolysis mass spectrometry and Fourier transform infrared spectroscopy. *J. Clin. Microbiol.* **36**, 367–374.

Tintelnot, K., Haase, G., Seibold, M., Bergmann, F., Staemmler, M., Franz, T. and Naumann, D. (2000). Evaluation of phenotypic markers for selection and identification of *Candida dubliniensis*. *J. Clin. Microbiol.* **38**, 1599–1608.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

Udelhoven, T., Naumann, D. and Schmitt, J. (2000). Development of a hierarchical classification system with artificial neural networks and FT-IR spectra for the identification of bacteria. *Appl. Spectrosc.* **54**, 1471–1479.

Varshney, R.K., Hoisington, D.A. and Tyagi, A.K. (2006). Advances in cereal genomics and applications in crop breeding. *Trends Biotechnol.* **24**, 490–499.

Virk, P.S., Ford-Lloyd, B.V., Jackson, M.T. and Newbury, H.J. (1995). Use of RAPD for the study of diversity within plant germplasm collections. *Heredity* **74**, 170–179.

Wang, Z.Y., Second, G. and Tanksley, S.D. (1992). Polymorphism and phylogenetic-relationships among species in the genus *Oryza* as determined by analysis of nuclear RFLPs. *Theor. Appl. Genet.* **83**, 565–581.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares in Krishnaiah, K.R. (Eds), *Multivariate Analysis*. Academic Press, New York, pp. 391–420.

Yang, J. and Yen, H.E. (2002). Early salt stress effects on the changes in chemical composition in leaves of ice plant and Arabidopsis. A Fourier transform infrared spectroscopy study. *Plant Physiol.* **130**, 1032–1042.