somatic diploids. Fruiting bodies can form in somatic diploid strains of *A. nidulans*. However, most of the ascospores they contain are sterile, and some asci contain sixteen rather than the standard eight ascospores. This has been interpreted as a tetraploid meiosis in which two diploid nuclei had fused and then undergone meiosis [1], but cytological observations failed to detect nuclear fusions and are incompatible with this neat model [18]. Thus, somatic diploids can 'switch' in the right developmental environment to germinal diploids, which will engage instantly in meiosis. Why this meiosis should be aberrant at all and sometimes lead to 16 ascospores asci remains a mystery.

These old problems can now be examined through new eyes. I propose that *A. nidulans* is epigenetically heterothallic. In this model, within specialized structures and preceding karyogamy, alternative mating type loci would be activated in the pairs of nuclei destined to fuse, leading to diploids that are epigenetically heterozygous at each of the mating loci. The proportion of selfed and crossed cleistothecia would depend on the ability of the different pairs of nuclei to switch to opposite mating types, which in turn might depend on the genetic markers segregating and/or subtle and uncontrolled environmental cues. In somatic diploids both mating types would be in the same state (possibly off). When somatic diploid nuclei find themselves in fruiting bodies, different combinations of mating gene switching would be possible. Only the switching patterns that mimic exactly that of germinal diploids would allow the completion of meiosis (Figure 2b), the others leading to aberrant meioses. The availability of the genome sequences, rapid gene replacement techniques and the fluorescent tagging of proteins involved in karyogamy and meiosis should enable the verification or falsification of this hypothesis.

References

1 Pontecorvo, G. *et al.* (1953) *The Genetics of* Aspergillus nidulans. *Adv. Genet.* 5, 141–238
2 Arst, H.N., Jr and Cove, D.J. (1973) Nitrogen Metbolite Repression in *Aspergillus nidulans*. *Mol. Gen. Genet.* 126, 111–141
3 Morris, N.R. and Enos, A.P. (1992) Mitotic gold in a mold. *Trends Genet.* 8, 32–37
4 Timberlake, W.E. (1990) Molecular genetics of *Aspergillus* development. *Annu. Rev. Genet.* 24, 5–36
5 Galagan, J.E. *et al.* (2005) Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature* 438, 1105–1115
6 Machida, M. *et al.* (2005) Genome sequencing and analysis of *Aspergillus oryzae*. *Nature* 438, 1157–1161
7 Nierman, W.C. *et al.* (2005) Genomic sequence of the pathogenic and allergenic fungus *Aspergillus fumigatus*. *Nature* 438, 1151–1156
8 Cai, J.J. *et al.* (2006) Accelerated evolutionary rate may be responsible of lineage-specific genes in Ascomycota. *J. Mol. Evol.* 63, 1–11
9 Geiser, D.M. *et al.* (1996) Loss of meiosis in *Aspergillus*. *Mol. Biol. Evol.* 13, 809–817
10 Pöggeler, S. (2002) Genomic evidence for mating abilities in the asexual pathogen *Aspergillus fumigatus*. *Curr. Genet.* 42, 153–160
11 Varga, J. (2003) Mating type gene homologues in *Aspergillus fumigatus*. *Microbiology* 149, 816–819
12 Paoletti, M. *et al.* (2005) Evidence for sexuality in the opportunistic fungal pathogen *Aspergillus fumigatus*. *Curr. Biol.* 15, 1242–1248
13 Bruggeman, J. *et al.* (2003) Male and female roles in crosses of *Aspergillus nidulans* as revealed by vegetatively incompatible parents. *Fungal Genet. Biol.* 39, 136–141
14 Hoffman, B. *et al.* (2001) Sexual diploids of *Aspergillus nidulans* do not form by random fusion of nuclei in heterokaryons. *Genetics* 157, 141–157
15 Pontecorvo, G. and Kafer, E. (1958) Genetic analysis base on mitotic recombination. *Adv. Genet.* 9, 71–104
16 Kafer, E. (1958) An 8-chromosome map of *Aspergillus nidulans*. *Adv. Genet.* 9, 105–145
17 Kafer, E. (1977) Meiotic and mitotic recombination in *Aspergillus* and its chromosomal aberrations. *Adv. Genet.* 19, 33–131
18 Elliott, C.G. (1960) The cytology of *Aspergillus nidulans*. *Gen. Res. Cam.* 1, 462–476

0168-9525/$ – see front matter
doi:10.1016/j.tig.2006.08.004

Shall I compare thee to a GM potato?

Ian J. Colquhoun, Gwénaëlle Le Gall, Katherine A. Elliott, Fred A. Mellon and Anthony J. Michael

Institute of Food Research, Norwich Research Park, Colney, Norwich, NR4 7UA, UK

A fundamental issue in the safety assessment of genetically modified crops is the question of whether unintentional changes have occurred in the crop plant as a consequence of the genetic modification. This question was addressed recently by using a powerful metabolite fingerprinting and metabolite profiling method to assess whether genetically modified potatoes are substantially similar to their corresponding conventional cultivars.

Corresponding author: Michael, A.J.. (tony.michael@bbsrc.ac.uk)
Available online 9 August 2006.

Introduction

One of the stages in the safety assessment of genetically modified (GM) crop plants is the compositional comparison of the GM line with its corresponding traditionally bred cultivar. This is to identify any unintended changes resulting from the genetic modification (such as insertion of the transgene into another gene, or the production of new metabolites), a process formalized as the Principle of Substantial Equivalence [1,2]. Any changes detected in the GM line are assessed in the context of the range of values for a given variable found within different conventionally bred cultivars. In the case of metabolites, if a GM line has an

Box 1. Metabolomics data analysis methods

Univariate and multivariate statistics

Every sample analyzed by metabolomics is characterized by the same variables. The variables are data points taken directly from the instrumental output (e.g. FIE-MS) or intensities related to concentrations of individual compounds (e.g. GC/MS). Univariate statistical tests are carried out on one variable at a time. There are advantages to using multivariate methods in which all variables are considered simultaneously.

Principal component analysis

PCA generates a rotated set of axes using linear combinations of the original axes (variables), which reduces the number of variables needed to describe the variance in the dataset. The **scores** are the co-ordinates of the samples in the new axis system. When groups of samples have systematic differences, the **scores plots** on one or more principal components show spatial clustering. The PC **loadings** show the contribution of the original variables to each PC. **Loadings plots** identify the data points or compounds that are responsible for the differences between groups.

Supervised and unsupervised classification methods

In a supervised method [15] the experimental data and sample group are input together. The group information affects the outcome. In an unsupervised method (e.g. PCA) the experimental data alone is analyzed [15]. The sample group is not provided. **Discriminant factor analysis, linear discriminant analysis** and **decision tree analysis** are supervised 'machine learning' methods. They take the sample vectors from known classes to build a model or system of rules that provides optimal separation of classes. Interpretable models identify variables (compounds) responsible for the discrimination and unknown samples can be submitted to the model for classification.

The model must be validated by classifying a test set (samples not used to build the model). The results of the test set predictions are summarized in a **confusion matrix** (Table I). In this example a model is tested with 30 samples, 10 in each class. All class A samples are correctly predicted; two class B samples are incorrectly assigned to A; two class C samples are incorrectly assigned, one to each of A and B.

Table I. Example of a confusion matrix

	Predicted class		
True class	**A**	**B**	**C**
A	10	0	0
B	2	8	0
C	1	1	8

unexpected change in the concentration of a particular metabolite but that change is within the natural range of concentrations found amongst natural cultivars, then the change would be regarded as 'safe'.

Transcriptomic, proteomic and metabolomic analyses have begun to be used recently to assess unintended effects in GM crops [3] and, of these approaches, metabolomics is currently the most useful owing to its independence from the requirement for pre-existing genome or expressed sequence tag data. The metabolome is defined as the quantitative content of all low molecular weight metabolites in a cell in a specific physiological state [4]. Although there is some semantic uncertainty about the term metabolomics as an analytical approach or as an area of integrative biology concerning metabolites [5], it is generally taken that metabolomics means measuring as many detectable metabolites as possible. Metabolomics is beginning to have a role in functional genomics [6] and in optimising metabolic engineering in plants [7]. The potential of metabolomics for comparing GM plants with conventional cultivars and detecting qualitative and quantitative changes was shown recently in an important paper by Catchpole and colleagues [8]. This is the first report of a large scale metabolomic analysis of field grown GM potato plants, and it highlights the importance of analytical technology and appropriate data analysis for the safety assessment of GM crops.

Box 2. The transgene enzymes and their metabolic products

The sucrose:sucrose 1-fructosyltransferase (*1-SST*) gene encodes an enzyme that transfers a fructosyl residue from one sucrose molecule to another to form the trisaccharide 1-kestose (S series). A second transgene encodes 1-fructan:fructan 1-fructosyltransferase (*1-FFT*), an enzyme that forms inulin polymers from 1-kestose and other oligofructans. Transgenic potato plants with both transgenes form the **SF series**. Fructans are oligomers and polymers of fructose and are soluble in water: inulins are linear fructans. In the transgenic potato lines expression of 1-SST leads to production of oligofructans (DP3 and DP4 resulting from addition of fructose to sucrose, where DP is the degree of polymerization) in potato tubers and expression of both genes produces oligofructans and inulin polymers of high DP.

Hierarchical metabolomic analysis of GM potato lines

To assess metabolite compositional changes in GM potato lines, Catchpole *et al.* [8] used a two stage non-targeted analytical approach that they termed 'hierarchical metabolomics'. The first stage was a rapid metabolome fingerprinting using flow injection electrospray ionization mass spectrometry (FIE-MS), in which samples were analyzed directly by MS without prior chromatographic separation of metabolites. This fingerprinting method was used to guide a more detailed second-stage profiling by gas chromatography time-of-flight MS (GC-ToF-MS). Data were analyzed by multivariate statistics: the unsupervised method of principal components analysis (PCA) followed by the supervised classification methods of either linear discriminant analysis or decision tree analysis (Box 1).

Catchpole *et al.* [8] chose to analyze two series of transgenic potato lines [9] expressing genes from globe artichoke (Box 2), either the sucrose:sucrose 1-fructosyltransferase (*1-SST*) gene alone (S series), or the *1-SST* and 1-fructan:fructan 1-fructosyltransferase (*1-FFT*) genes together (SF series). Catchpole *et al.* [8] compared the tuber composition across 12 genotypes: three independent transgenic lines from each of the S and SF series, two lines from the parent cultivar, Désirée, plus four other conventional cultivars, included to provide a wider background of natural variation than given by Désirée alone. The Désirée controls were a wild-type line produced through tuber propagation (De1) and another non-transgenic line produced by tissue culture (De2). Plants were field grown over two years.

First-stage data analysis: FIE-MS

FIE-MS metabolite fingerprinting data were analyzed by PCA (Box 1). The first principal component (PC) axis of the scores plot showed a separation into three main groups corresponding to the two series of transgenic lines and the conventional cultivars including Désirée. The first PC loading indicated that fructans (DP3–DP7; see Box 2) were responsible for the separation. As the substantial

equivalence test is meant to look at compounds other than the primary products of the modification, the columns corresponding to these compounds were removed from the data table and the PCA was repeated. It still showed a separation between transgenic and non-transgenic lines (including the Désirée controls), although it was less pronounced and only appeared on the second PC axis. Does this mean that the Désirée-based transgenic and control lines are not substantially equivalent? The loadings for the second PC were not shown [8], so we can only speculate on the reason for the separation. One possibility is that not all the 'primary products' have been removed from the data table.

Confusion matrices (Box 1) were used to ask whether different groups of samples were substantially equivalent or not. When fructans were omitted from the decision tree analysis (see Ref. [8]), there was confusion between the S series transgenic lines and the Désirée controls, suggesting equivalence, whereas the conventional cultivars were clearly separated from each other and the transgenic lines. However, the SF series transgenic lines were fully distinguished from the Désirée background controls. This suggests that the SF series transgenic lines are not equivalent to the Désirée control lines, De1 and De2. Also, based on the full FIE-MS dataset, linear discriminant analysis correctly separated De2 from De1 in 16 out of 16 cases, suggesting that the tuber-propagated Désirée line is not equivalent to the tissue culture generated Désirée line (see Ref. [8]). This has ramifications for the equivalence of the transgenic lines, as they were produced by tissue culture procedures [9].

Second-stage data analysis: GC-ToF-MS

GC-ToF-MS data were acquired for >2000 tubers (~180 per genotype). Altogether 252 metabolites were measured, including known and unknown compounds. Data analysis followed the same procedure as above but, in addition to multivariate analyses, a systematic univariate statistical analysis was carried out for every compound that could be measured. The large number of tubers analyzed per genotype enabled a population distribution to be plotted for metabolite levels. A separate distribution was plotted, showing the mean and spread in metabolite levels for each genotype for all 252 metabolites measured. From the combined data for the control genotypes (all regarded as 'safe'), a lower and upper limit can be set for each metabolite. The distributions calculated for the GM lines were then examined for breaches of these limits. In Catchpole *et al.* [8] the DP2 and DP3 fructans were the only compounds found to exceed the limits (mean values exceeded the upper limit for six peaks in several of the GM lines). Apart from these intentionally introduced differences, the GM lines could be judged substantially equivalent to conventional cultivars on the basis of almost 250 metabolites analyzed.

The results of the multivariate analysis on the GC data confirmed that the separation between transgenics and Désirée controls was much diminished with removal of the six fructan peaks, whereas the other cultivars remained distinct. The confusion matrices bear this out, although from the pattern of the misclassifications the two series of transgenics seem closer to each other than they are to the Désirée control lines (see Ref. [8]). This perceptible lack of equivalence could be due to tissue culture effects or higher glucose levels in some of the transgenic lines, as reported previously [9], and glucose is known to regulate a broad range of genes [10].

From science to policy

What are the practical limitations on wider scale adoption of this method of substantial equivalence testing? First, it should be recognized that a huge number of chromatograms had to be run. The whole procedure, from sample preparation through measurement and data evaluation, is a complex one calling for standard operating procedures and quality control checks at each step. It has to be demonstrated that different laboratories can produce reproducible results. The 'metabolite levels' used here to generate the population distributions were relative levels, not absolute concentrations, so inter-laboratory comparisons would be difficult. There is no doubt, however, that further development and validation of the technology, already being seen in FIE-MS [11,12] and LC/MS profiling [13], will continue apace. More subtle metabolome changes can be addressed by GC-ToF-MS to detect metabolic network connectivity [14], which could help explain the physiological basis for lack of substantial equivalence where it exists. It remains to be defined clearly how substantial 'substantial equivalence' should be, but the work of Catchpole and colleagues [8] demonstrates that even subtle changes between plants can be revealed by metabolomics.

References

1 OECD (1993) *Safety Evaluation of Foods Derived by Modern Biotechnology: Concepts and Principles.* Organisation for Economic Co-operation and Development, Paris (http://www.oecd.org/dataoecd/57/3/1946129.pdf)

2 Kok, E.J. and Kuiper, H.A. (2003) Comparative safety assessment for biotech crops. *Trends Biotechnol.* 21, 439–444

3 Cellini, F. *et al.* (2004) Unintended effects and their detection in genetically modified crops. *Food Chem. Toxicol.* 42, 1089–1125

4 Oliver, S.G. *et al.* (1998) Systematic functional analysis of the yeast genome. *Trends Biotechnol.* 16, 373–378

5 Villas-Boâs, S.G. *et al.* (2005) Metabolomics or metabolite profiles? *Trends Biotech.* 23, 385–386

6 Bino, R.J. *et al.* (2004) Potential of metabolomics as a functional tool. *Trends Plant Sci.* 9, 418–425

7 Trethewey, R.N. (2004) Metabolite profiling as an aid to metabolic engineering in plants. *Curr. Opin. Plant Biol.* 7, 196–201

8 Catchpole, G.S. *et al.* (2005) Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops. *Proc. Natl. Acad. Sci. U. S. A.* 102, 14458–14462

9 Hellwege, E.M. *et al.* (2000) Transgenic potato (*Solanum tuberosum*) tubers synthesize the full spectrum of inulin molecules naturally occurring in globe artichoke (*Cynara scolymus*) roots. *Proc. Natl. Acad. Sci. U. S. A.* 97, 8699–8704

10 Price, J. *et al.* (2004) Global transcription profiling reveals multiple sugar signal transduction mechanisms in *Arabidopsis*. *Plant Cell* 16, 2128–2150

11 Dunn, W.B. *et al.* (2005) Evaluation of automated electrospray-TOF mass spectrometry for metabolic fingerprinting of the plant metabolome. *Metabolomics* 1, 137–148

12 Wilson, I.D. *et al.* (2005) High resolution "Ultra Performance" Liquid Chromatography coupled to oa-TOF mass spectrometry as a tool for differential metabolic pathway profiling in functional genomics studies. *J. Proteome Res.* 4, 591–598

13 Overy, S.A. *et al.* (2005) Application of metabolite profiling to the identification of traits in a population of tomato introgression lines. *J. Exp. Bot.* 56, 287–296
14 Weckwerth, W. *et al.* (2004) Differential metabolic networks unravel the effects of silent plant phenotypes. *Proc. Natl. Acad. Sci. U. S. A.* 101, 7809–7814
15 Kemsley, E.K. (1998) *Discriminant Analysis and Class Modelling of Spectroscopic Data.* J. Wiley

0168-9525/$ – see front matter
doi:10.1016/j.tig.2006.08.002

Genome Analysis

In plants, highly expressed genes are the least compact

Xin-Ying Ren[1], Oscar Vorst[1], Mark W.E.J. Fiers[1], Willem J. Stiekema[2] and Jan-Peter Nap[1,2]

[1] Applied Bioinformatics, Plant Research International, Wageningen University and Research Centre, 6708 PB Wageningen, The Netherlands
[2] Centre for BioSystems Genomics, 6700 AA Wageningen, The Netherlands

In both the monocot rice and the dicot *Arabidopsis*, highly expressed genes have more and longer introns and a larger primary transcript than genes expressed at a low level: higher expressed genes tend to be less compact than lower expressed genes. In animal genomes, it is the other way round. Although the length differences in plant genes are much smaller than in animals, these findings indicate that plant genes are in this respect different from animal genes. Explanations for the relationship between gene configuration and gene expression in animals might be (or might have been) less important in plants. We speculate that selection, if any, on genome onfiguration has taken a different turn after the divergence of plants and animals.

Introduction

A major issue in relating genome structure to gene expression is the relationship between the relative activity of genes and their position and/or structure. In organisms as diverse as human [1–4] and *Caenorhabditis elegans* [1], highly expressed genes have fewer and shorter introns, shorter coding sequences and shorter intergenic regions [1–5]. This compact nature of highly expressed genes is explained by a selection for either transcriptional efficiency to reduce time and energy [1], a regional mutation bias that positions highly expressed genes in domains more prone to deletions [3] or by a genomic design into open chromatin [4]. We here present a whole genome analysis of the relationship between gene structure and gene expression for two widely diverged plant species, the monocotyledonous plant rice (*Oryza sativa*) and the dicotyledonous plant *Arabidopsis thaliana,* with data from two different expression platforms, massively parallel sequencing signature (MPSS) and microarrays. In both plant genomes, highly expressed genes have more and longer introns and a longer primary transcript. In short, they are less compact than the genes expressed at a low level. This contrasts with the relationship between gene expression and gene structure in human and *C. elegans*, although the absolute differences between plant genes are considerably smaller than for human genes. These findings could suggest that the outcome of selection has been different between animals and plants.

Analysis of plant gene expression in relationship to gene structure

The public domain MPSS expression data for *Arabidopsis* [6] (http://mpss.udel.edu/at/) and rice [7] (http://mpss.udel.edu/rice/) offer good genome-wide expression coverage in a range of different expression libraries and allow easy quantification. To correlate expression data with gene structure, we obtained *Arabidopsis* and rice genome sequences and annotations from The Institute of Genomic Research (TIGR). All genes annotated as either (retro)-transposons or pseudogenes were excluded from the analysis and, in cases of alternative splicing, the longest variant was used in the analyses. We mapped the MPSS expression data to their position in the *Arabidopsis* (TIGR5) and rice (TIGR version 3) genome and all 17 base MPSS tags with a unique position were taken into account. Genes without expression data were not included in the analysis.

To compare the levels of expression of genes in different expression libraries, we sorted the expression values in each library in an ascending order, then divided them into five groups, each containing 20% of the population, and assigned an expression rank from 1 (low expression) to 5 (high expression). Where the cutoff caused equal expression values to be in different rank groups (happening notably with zero expression), the expression values were placed in the lower rank group. For each gene, we averaged the expression ranks over all libraries. This averaged expression rank (rE) indicates the relative expression level of each gene under all conditions analyzed. Alternative methods of expression analysis (see the supplementary material online) give similar results as found for rE. As the rE can be influenced in part by the number of libraries

Corresponding author: Nap, J.-P. (janpeter.nap@wur.nl)
Available online 24 August 2006.