New developments in genotyping and phenotyping for breeding improved sustainability in temperate forage grasses

MO Humphreys, I Armstead, I Donnison, MW Humphreys, I King, S Thain, J Taylor, H Roderick, L Turner, L. Skøt, P Wilkins and R Yadav
Institute of Grassland and Environmental Research (IGER), Plas Gogerddan, Aberystwyth, Ceredigion, SY23 3EB UK.

Abstract. New genes for ‘sustainability’ traits are being incorporated into ryegrass (Lolium sp.) breeding populations using marker-assisted intra- and inter-specific introgression from a wide range of genetic resources. Mapping families, substitution and introgression lines have been produced to facilitate gene transfer between Festuca and Lolium species. Advanced metabolomic phenotyping techniques (e.g. NIRS and FT-IR) can help close genotype-phenotype gaps to increase the environmental sustainability of grassland and improve the nutritional value of genotypes for forage or their chemical suitability for biofermentation/biofuel.

Use of genetic markers
Marker Assisted Selection (MAS) improves the efficiency of breeding particularly where traits are difficult or expensive to evaluate; are expressed late in crop development; or where several components need to be combined. QTL mapping which identifies trait-associated genetic markers has been carried out for a wide range of agronomic ‘sustainability’ traits including seasonal patterns of production, regrowth, root characteristics and forage quality including components of water-soluble carbohydrate (WSC) content (Turner et al. 2005).

Linkage disequilibrium-based association mapping offers potential to identify functional SNP markers within candidate gene sequences or extremely closely linked markers. Candidate genes have been identified for flowering time, WSC and tolerance to abiotic stresses and SNPs are being detected. For example a 68 base-pair indel has been identified in the 5’ upstream region of an alkaline invertase gene, which is located on LG 6 in L. perenne, underlying a QTL for glucose and fructose content (Skøt, unpublished).

Gene introgression in Lolium/Festuca
The Lolium/Festuca species complex is unique in that high levels of recombination are combined with the ability to distinguish homoeologous chromosomes through genomic in situ hybridisation (GISH) (King et al. 2002). Thus Lolium-Festuca hybrids are highly amenable to introgression-mapping to identify gene complexes involved in trait expression (Humphreys et al 2003) and major advances in breeding for stress resistance have been achieved. QTL for survival and recovery following severe drought were found along the entire length of Festuca chromosome 3 (Alm et al. 2005) which provides a genetic resource for improving drought tolerance in Lolium (Humphreys et al. 2005). F. arundinacea and F. glaucescens genes for drought resistance, introgressed onto different sites on Lolium chromosome 3 were also found to confer improved salinity tolerance (Lattore, unpublished). Chromosome 3 in Lolium and Festuca shares considerable synteny with rice chromosome 1 (Jones et al. 2002) which has QTL for traits known to contribute to drought resistance such as root development and osmotic adjustment. F. glaucescens genes for drought resistance have also been introgressed into
*L. multiflorum* close to the *F. pratensis* stay-green locus on the long arm of chromosome 5 (Ghesquière, pers. comm. Using MAS, different combinations of the *Festuca* genes for drought resistance on chromosomes 3 and 5 have now been combined in advanced breeding lines and are being assessed.

Inefficient use of feed nitrogen by grazing ruminants is a major problem in livestock agriculture. Of the nitrogen ingested in forage as little as 35% maybe retained and used in the production of milk and meat due to the rapid degradation of plant protein in the rumen. The remaining nitrogen is excreted and contributes to agricultural pollution. Plant proteases have a significant role in the initial phase of proteolysis in the rumen (Kingston-Smith *et al.* 2005). Chromosomal segments from *F. glaucescens* (protein half life 19.2 h) have been transferred to *L. multiflorum* (protein half life 2.3h) resulting in a 3x increase in protein half life (Gardner 2003). A combination of GISH, AFLP and biochemical analysis suggests that the trait is associated with at least three genome locations.

**Gene isolation**

Introgression between *Lolium* and *Festuca* species is also a powerful tool to assist gene isolation and plants containing small *F. pratensis* substitutions provide useful markers for screening a *F. pratensis* BAC library. Markers or candidate genes derived from rice can also be used to screen *L. perenne* and *F. pratensis* BAC libraries to identify corresponding genes in forage grasses. We have screened our *F. pratensis* BAC library using both genomic-DNA derived AFLP markers and cDNA sequences (Donnison *et al.* 2005, Moore *et al.* 2005). For about 50% of the AFLP markers it was possible to identify BAC clones, but others were too repetitive to enable reliable identification of locus-specific BACs. However none of the *F. pratensis* BACS we have identified using genomic DNA derived AFLPs contained obvious coding regions following partial shot-gun sequencing. Given the size of the *F. pratensis* genome at approximately 4.5x that of rice or 12x that of *Arabidopsis thaliana* this is not too surprising as a comparatively small proportion of the genome is expected to be taken up by gene sequences. In contrast, BACs identified using primers for cDNA sequences contained sequences of other genes on the same BAC (e.g. Armstead *et al.* 2004). This indicates that our library contains a mixture of BACs which are either gene rich, or contain no genes at all.

A *L. perenne* BAC library produced for the EU GRASP project ([http://www.grasp-euv.dk](http://www.grasp-euv.dk)) consists of approximately 95,000 clones with an average insert size of 112 kb and provides five genome equivalents coverage. DNA pools from this library have been screened with PCR primers for candidate genes for traits of interest including cell wall composition, protein stability, control of flowering time, and carbohydrate metabolism. Upstream and other non-coding regions of these genes are being sequenced and PCR primers designed to amplify alleles from 20 *L. perenne* genotypes which are then sequenced and aligned to derive allele-specific single nucleotide polymorphism (SNP) markers.

Sequences from every 5th mapped rice BAC are being used to develop *Lolium* SNP markers which are placed in the *Lolium/Festuca* introgression bins. The bin mapping of rice BACs requires that a coding region is identified on each BAC which can be used to screen other monocot databases. Primers are designed from complementary regions and used to amplify equivalent sequences in *Lolium* and *Festuca* via PCR. The products are sequenced and SNP markers that discriminate between *Lolium* and *Festuca* are designed. Once identified, SNPs can be mapped to a
specific bin on a specific chromosome. This strategy is proving successful for *Festuca*
chromosome 3 where 51 BACs from rice chromosome 1 have been bin mapped to the
*Festuca* physical map.

**Developments in Phenotyping**

To match progress in the development of molecular genetic techniques, advanced
phenotyping techniques can be exploited to close the genotype-to-phenotype gap
(Jenkins et al. 2004) for sustainability traits such as nutritional value and chemical
suitability for biofermentation/biofuel. Infrared spectroscopy, near-infrared (“NIR”) and
a mid-infrared method termed Fourier-transform Infrared (“FT-IR”) are applicable for
solids or liquids. NIR light carries more energy per quanta and can thus be more robust
than FT-IR for solids or course-ground material. Already NIRS is used routinely to
determine digestibility, protein and carbohydrate contents in forage breeding and to
assess the clover content of mixed swards. However, in NIR spectral peaks are much
less defined reducing the qualitative information compared to FT-IR data. IR and
particularly FT-IR spectroscopy is highly versatile using a wide variety of transmission
and reflective methods (including microscopy) appropriate for different sample types.
IR light excites particular molecular bond types between atoms at specific “fingerprint”
wavelengths. These are recorded as peaks in the spectrum trace scan. In biological
samples the mix of molecules produce quantitative but highly complex information rich
spectra. Such data require extensive processing and statistical analysis in order to bring
out the underlying chemical information.

Metabolomics is developing as an important functional genomic tool to assist in
understanding the complex molecular interactions in biological systems (Bino et al.
2004). The metabolome represents a collection of all metabolites arising from gene
expression and metabolic profiling provides an instantaneous 'snapshot' of an
organism’s physiology. “Metabolomic research” has two characteristics that set it apart
from “analytical chemistry”:- (1) metabolites are profiled across a wide range of
chemical classes and (2) data is summarized and analysed by multivariate statistics.
Estimates suggest that 200,000 different metabolites exist within the plant kingdom and
a “typical” plant leaf may have 5000 primary and secondary metabolites. A variety of
approaches are required to identify all these metabolites simultaneously including Mass
Spectroscopy (MS) coupled to either Gas Chromatography (GC-MS) or Liquid
Chromatography (LC-MS). Nuclear Magnetic Resonance (NMR) is increasingly used
particularly when coupled to liquid chromatography (LC-NMR) (Bino et al. 2004).
However it is difficult to assign specific chemical identities to the majority of
metabolites measured in a typical profiling experiment. Also non-random data “noise”
introduced by growth environments, sample processing or analytical machine effects is
unavoidable. Multivariate statistical approaches may interpret this as “real” biological
differences so carefully designed experimental protocols; adequate sample numbers;
controls, replication and randomisations are essential at every stage to allow for this.
The correct use of statistical methods including independent “training” and “validation”
data sets helps to prevent “over fitting” models and promote robust predictions.

All spectroscopy data require some level of pre-processing prior to statistical
analysis but there are no standard agreed protocols. For example, Figure 1a indicates
that there are base line shifts between consecutive sample measurements that require
mathematical normalisation. Multiple overlapping peaks between ~1800-700cm⁻¹ may
require separation by various deconvolution functions before further quantification.
Even after these are performed the data remain complex and difficult to interpret by eye. Simple data reduction methods such as Principle Component Analysis (PCA) help to visualise simple structure in the data. Figure 1b shows a PCA of the 18 samples after normalisation of the data from the signal rich region (1800-700cm\(^{-1}\)). Despite one outlier the 2 grass varieties can be separated clearly indicating significant cell wall biochemical differences.

One way to visualise very large data sets is in an “Array” format. Figure 1c shows the main carbohydrate spectral region of the traces from Fig 1a. Grass varieties \( A \) and \( B \) have visually differing absorbance patterning. The “outlier” (black arrow) from the \( B \) variety is immediately obvious. These statistical methods are based on linear modelling. However “machine learning” approaches such as neural networks or genetic algorithms may offer new insights as metabolomic data sets grow in size and complexity.

Figure 1 FT-IR data of cell wall residues water/methanol extracted for nine samples from each of 2 grass varieties. (a) Reflectance FT-IR Spectra. (b) PCA analysis, plotting the first 3 principle components in x,y,z after vector normalisation on 1800-700cm\(^{-1}\) data. (c) “Metabolite Array” plot of the carbohydrate region after 0-1 scaling.
References