Biodiversity research in relation to crop improvement and conservation genetics

J. Russell, A. Booth, M. Woodhead, K. Caldwell, H.V. Davies, D.F. Marshall, J.R. Hillman & W. Powell

Biological diversity – biodiversity – has many meanings, ranging from the number of species in an ecosystem, a country or the world, through to the levels of genetic diversity within a species, family, or Phylum. It tends to refer to the phenotype - the biological form we see, in essence, the manifestation of gene expression in the organism as influenced by various components of the environment.

The conservation of wildlife and habitats is integral to the public debate and government policies on biodiversity. In 1994, the UK ratified the international Convention on Biological Diversity (www.conbio.org.uk), building on the 1971 Ramsar Convention on Wetlands of International Importance especially as Waterfowl Habitat, which entered into force in the UK in 1976. A national Biodiversity Action Plan was published in 1994 (www.ukbap.org.uk) followed in 1995 by a report of the UK Biodiversity Steering Group proposing the monitoring of 1,252 species to gauge biodiversity trends. In March 2001, *Sustaining the Variety of Life: 5 years of the UK Biodiversity Action Plan* was published. At present, there are nearly 170 local biodiversity action plans in preparation or in progress in the UK. Other relevant international conventions include the 1973 Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) which came into force in the UK in 1975; the 1979 Convention on Conservation of Migratory Species of Wild Animals, which came into force in the UK in the same year; and the 1979 Bern Convention on the Conservation of European Wildlife and Natural Habitats which came into force in the UK in 1982. The application of CITES is reinforced by the European Council Regulation on the Protection of Species of Wild Fauna and Flora by Regulating Trade Therein, which came into force in the UK in 1997. In terms of UK legislation, the Wildlife and Countryside Act 1981 offers legal protection to a number of specified wild animals and plants; any variations to the listing requested by the Secretary of State for the Environment, Food and Rural Affairs are dependent on parliamentary approval.

Maintenance of biodiversity is generally regarded as an important aspect of the quality of life, fundamental to the natural heritage, and the basis of important economic and social activities. Biodiversity is therefore regarded as natural capital, but vulnerable to the impact of humans on the natural environment. Even so, it is difficult to measure biodiversity, or identify the optimum level of biodiversity in a habitat, let alone assess the optimum level of genetic diversity within a species. Microbial biodiversity, especially in soils, is poorly understood, and there has been a tendency to underestimate the role and relevance of biodiversity studies in economically important species and in agriculture, horticulture, and forestry, despite the current focus on the concept of sustainability, as well as initiatives assessing the potential impacts of climate change and the introduction of invasive non-native ('alien') species.

Concerns about biodiversity have become increasingly relevant to the way we manage biological systems. This is both because the diversity within species or ecosystems plays a major role in determining both their resilience to short-term environmental pressures and their survival over evolutionary timescales. In addition, biodiversity has increasing relevance as a resource that can provide valuable and often novel products at scales ranging from the pharmaceutical to the commodity. Biodiversity, in term of intra-specific variability, has a key role in underpinning the ability of global agricultural systems to keep pace not only with the biological arms race imposed by pest and pathogen burdens but also with the increasing demands both for volume and processing quality associated with the food industry as well as the increasingly diverse range of industries that utilise plant products. Plant biologists must therefore not only face up to the biodiversity challenge but also consider creative ways to exploit biodiversity using the contemporary tools of genome science and informatics.

Genetic diversity within species has long been the mainstay of population genetics and in the last century an extensive body of mathematical theory, built initially on the elegant work of Fisher, Wright and Haldane described the key factors that determine the patterns of genetic variation within and between populations. It is however only with the recent development of 'omics' technologies that the potential to

	RFLPs	RAPDs	Microsatellites (SSRs)	AFLPs
Pros	Effectively infinite in number Single protocol for all markers Codominant Detects homeologous loci Robust Transferable	Quick Easy Primers available cheaply Many markers per run Simple Unlimited number of markers	Simple PCR based Codominant Highly polymorphic Quick Robust	Detects large numbers of loci PCR based Highly polymorphic Robust
Cons	Requires large quantities of DNA Slow-hybridisation based Variable level of polymorphism Difficult in large genomes Generally uses isotope	Dominant markers Experimental conditions are critical Poor transferability Poor reliability	Long development time Need specific primers Expensive to establish May use isotope	Dominant May use isotope

Table 1 Associated advantages and disadvantages with 4 popular DNA marker technologies

both measure and understand these processes at a genome-wide level has become possible. This together with the second generation 'functional genomics' approaches built on a range of technologies such as microarrays, knockout, proteomics, metabolomics etc., is providing us with unique resources with which to build our understanding of 'functional biodiversity'. The purpose of this article is to review some of the key developments of relevance to biodiversity research in relation to crop improvement and conservation genetics.

Intra-specific diversity The discovery of DNA-based genetic markers by Botstein *et al*[1] in 1980 fundamentally changed our ability to detect, analyse, describe and manipulate sources of genetic variation. Since that time, various molecular assays have been developed (Table 1) each possessing a specific set of attributes ranging from simplicity e.g. RAPDs, informativeness e.g. SSRs, through to high multiplex ratios e.g. AFLPs. All the methods outlined in Table 1 have one common feature in that they represent indirect approaches to detecting DNA sequence differences. Developments in genome science, particularly high-capacity DNA sequencing technology coupled with powerful informatics tools have allowed the genomes of many organisms to be either fully or partially sequenced. A re-focusing of effort in the intra-specific diversity of crop plants is now emerging in both private and public research organisations (Rafalski [2]). The extent of polymorphism differs substantially between species and sampled loci. For example, each copy of the human genome differs from any other copy in the population by roughly 1 in 1,250 nucleotides. Although far less advanced, data on the extent and distribution of sequence polymorphism in

Genome	Common Name	Haploid size (Mb)	Intra-specific sequence diversity
2ea Mays	Maize/Corn	2,292	1 SNP per 83bp
Glycine max	Soybean	1,115	1.64 SNPs in coding and 4.85 SNPs in non-coding regions per kb.
Arabidopsis thaliana	Thale Cress	125	1 SNP per 3.3 kb

Table 2 Estimates of sequence diversity in some crop plants

Haplotype	The sequence configuration of two or more alleles on a single chromosome of a given individual.
Linkage Disequilibrium	Non-random association of a particular haplotype for two or more loci.
Nucleotide Diversity	A measure of DNA sequence variation for a given region influenced by the number of variable sites and their population allele frequency.

Table 3 Definition of some terms used in this article.

Cultivated

```
CGGCGCTGCCTCTGCCGCTGCCGCTGCCTCAGCAGCTTGTCCTGTGCTTTTTCGTTCGTGAAAACTATGTTCGTTC   Haplotypes
CGGCGCTGCCTCTGCCGCTGCCGCTGCCTCAGCAGCTTGTCCTGTGCTTTTTCGTTCGTGAAAACTATGTTCGTTC
CGGCGCTGCCTCTGCCGCTGCCGCTGCCTCAGCAGCTTGTCCTGTGCTTTTTCGTTCGTGAAAACTATGTTCGTTC
CGGCGCTGCCTCTGCCGCTGCCGCTGCCTCAGCAGCTTGTCCTGTGCTTTTTCGTTCGTGAAAACTATGTTCGTTC
CGGCGCTGCCTCTGCCGCTGCCGCTGCCTCAGCAGCTTGTCCTGTGCTTTTTCGTTCGTGAAAACTATGTTCGTTC   1
CGGCGCTGCCTCTGCCGCTGCCGCTGCCTCAGCAGCTTGTCCTGTGCTTTTTCGTTCGTGAAAACTATGTTCGTTC
CGGCGCTGCCTCTGCCGCTGCCGCTGCCTCAGCAGCTTGTCCTGTGCTTTTTCGTTCGTGAAAACTATGTTCGTTC
CGGCGCTGCCTCTGCCGCTGCCGCTGCCTCAGCAGCTTGTCCTGTGCTTTTTCGTTCGTGAAAACTATGTTCGTTC
```

Landraces

```
CGGCGCTGCCTCTGCCGCTGCCGCTGCCTCAGCAGCTTGTCCTGTGCTTTTTCGTTCGTGAAAACTATGTTCGTTC  1
CGGCGCTGCCTCTGCCGCTGCCGCTGCCTCAGCAGCTTCTCCTGTGCTTTTTCGTTCGTGAAAATTATGTTCGTTC  2
CGGCGCTGCCTCTGCCGCTGCCGCTGCCTCAGCAGCTTGTCCTGTGCTTTTTCGTTCGTGAAAACTATGTTCGTTC  1
CGGCGCTGCCTCTGCCGCTGCCGCTGCCTCAGCAGCTTCTCCTGTGCTTTTTCGTTCGTGAAAATTATGTTCGTTC  2
CGGCGCTGCCTCTGCCGCTGCCGCTGCCTCAGCAGCTTCTCCTGTGCTTTTTCGTTCGTGAAAATTATGTTCGTTC  2
CGGCGCTGCCTCTGCCGCTGCCGCTGCCTCAGCAGCTTGTCCTGTGCTTTTTCGTTCGTGAAAACTATGTTCGTTC  1
CGGCGCTGCCTCTGCCGCTGCCGCTGCCTCAGCAGCTTCTCCTGTGCTTTTTCGTTCGTGAAAATTATGTTCGTTC  2
CGGCGCTGCCTCTGCCGCTGCCGCTGCCTCAGCAGCTTGTCCTGTGCTTTTTCGTTCGTGAAAACTATGTTCGTTC  1
```

H.spontaneum

```
CGGCGCTGCCTCTGCCCCTGCCGCTGCCTCAGCAGCTTCTCCTGTGCTTTTTCGTTCGTGAAAACTATGTTCGTTC  3
CGGCGCTGCCTCTGCCGCTGCCGCTGCCTCAGCAGCTTCTCTTGTGCTTTTTCGTTCGTGAAAACTATGTTCGTTC  4
CGGCGCTGCCTCTGCCGCTGCCGCTGCCTCAGCAGCTTGTCCTGTGCTTTTTCGTTCGTGAAAACTATGTTCGTTC  1
CGGCGCTGCCTCTGCCGCTGCGGCTGCCTGAGCAGCTTCTCCTGTGCTTTTTCGTTCGTGAAAATTATGTTCGTTC  5
CGGCGCTGGCGCTGCCGCTGCCGCTGCCTCAGCAGCTTCTCCTGTGCTTTTTCGTTTGTGAAAACTATGTTCGTTC  6
CGGCGCTGCCTCCGCCGCTGCCGCTGCCTCAGCAGCTTCTCCTGTGCTTTTTCGTTCGTGAAAATTATGTTCGTTC  7
CGGCGCTGCCTCTGCCGCTGCCGCTGCCTCAGCAGCTTCTCCTGTGCTTTTTCGTTCGTGAAAACTATGTTCGTTC  8
CGGTGCTGCCTCTGCCGCTGCCGCTGCCTCAGCAGCTTCTCCTGTGCTTTTTCGTTCGTGAAAACTATGTTCGTTC  9
```

Figure 1 Sequence polymorphism at Best608, with homology to a jasmonate-induced protein

crop plants is progressing and estimates of sequence diversity for maize, *Arabidopsis* and soybean are given in Table 2. Recent studies cited by Rafalski [2] have identified for elite US maize germplasm 1 SNP per 48 bp in non-coding regions of the maize genome.

Similar studies are being conducted for barley and an example of SNPs identified in a 246 bp amplified region of the barley genome is shown in Figure 1 for a sample of 24 accessions. There are a number of key features of this figure, foremost of which is the organisation of SNPs into haplotypes (Table 3). Overall there is a limited number of haplotypes (Figure 2) in the cultivated gene-pool and it would appear that barley breeders are shuffling a restricted range of haplotypes in current barley germplasm. The structure of haplotype blocks in crop germplasm is important since it represents a fossil record of the history and structure of ancestral populations together with a

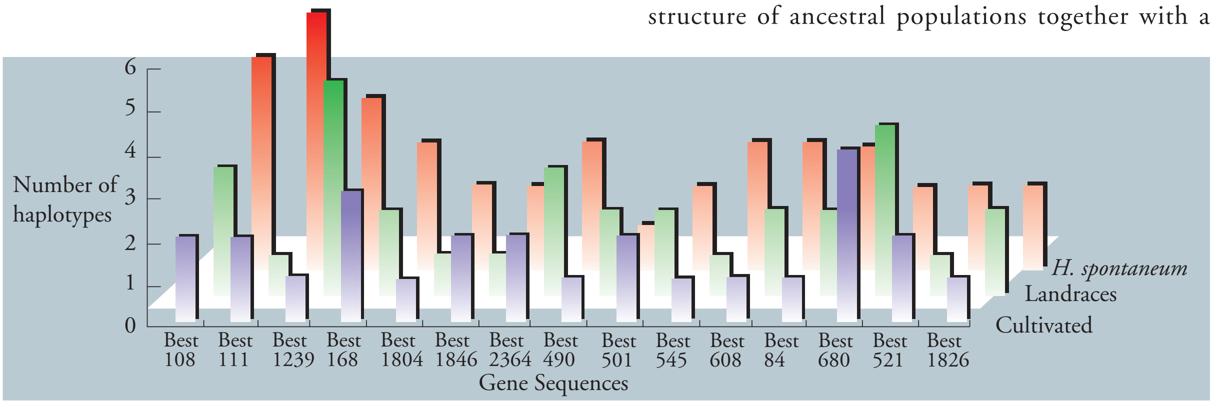

Figure 2 Comparison of haplotypes among the three gene pools of barley.

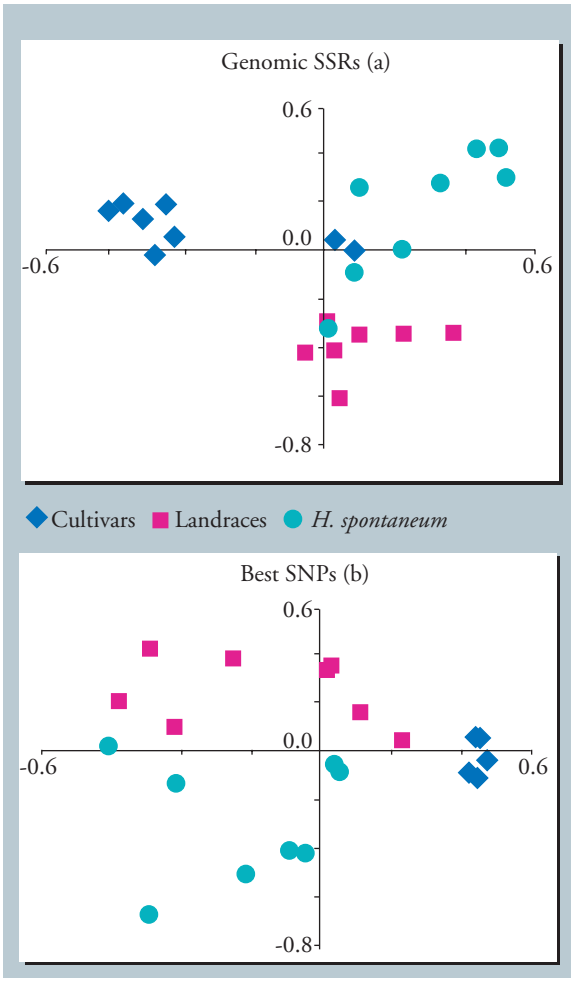

Figure 3. Principal Coordinate Analysis (PCO) analysis using distance matrices generated by genomic SSRs (a) and SNP haplotypes (b).

means of describing and quantifying functional polymorphism. Haplotype analysis may therefore be used to analyse the genetic structure and relationships of various gene pools. In Figure 3 we present data comparing the genetic relationship between *H.vulgare* cultivars, landraces and *H.spontaneum* accessions based on SNiP-derived haplotype data and genomic SSRs. Although there are similarities between both plots, the hapolotype-based analysis clearly discriminates between the three gene pools with the *H.vulare* cultivars being represented by a very similar and distinctive cluster. In particular, Figure 3a graphically illustrates the narrowing of the genetic base or "bottle neck" experienced during the domestication of cultivated barley. Well-characterised and sampled products of the domestication process may provide an experimental framework to identify and eventually validate the association of genes with phenotypes that have been exposed to selection during crop domestication. The products of the domestication process together with their ancestors therefore provide a rich resource to isolate genes involved in domestication and adaptation. Exploiting the products of domestication in conjunction with population-based genetic analysis of relevant genes can give rise to new insights into the evolutionary processes that sculptured the formation of crop (and animal) species.

The frequency of SNPs in some plant genomes together with their occurrence in the vicinity of plant genes make this class of polymorphic marker an ideal choice for meiotic map construction. We have initiated a programme to map both SNP and other functional classes of polymorphism in double-haploid (DH) populations of barley. The highly successful BBSRC/SEERAD-funded cereal transcriptomics programme has generated a wealth of EST sequences providing a substrate for SNP discovery. Amplification by PCR followed by direct sequencing of DNA regions from the parents of the mapping population has been used to identify SNPs and insertion/deletion events. Preliminary data on the map location of ESTs are shown in Figure 4 for the Steptoe x Morex mapping populations. Microsatellites also occur in reasonable high frequency in ESTs (Cardle *et al* 2000 [3], Morgante *et al* 2002 [4]) providing an opportunity to map ESTs based on length polymorphism. The advantages of this marker system are that the polymorphism is physically associated with a coding region and the amplification products are robust with a high degree of reproducibility.

Conservation Genetics The tools of genetic science can also be applied to species of high conservation priority, by constructing transcriptome libraries from rare and scarce species, to offer novel insights into biodiversity, including sequence diversity and its relationship to biological function. As part of a collaborative project with the Royal Botanic Gardens, Edinburgh, we have generated transcriptome libraries from three rare species, *Anastrophyllum joergensenii*, a dioecious leafy liverwort restricted to cool montane high-rainfall areas; *Athyrium distentifolium*, a diploid out-crossing fern of montane areas and *Koenigia islandica*, a diminutive annual, which in the UK is restricted to Skye and Mull. Over 1000 ESTs have been sequenced from two of the three species and functional SSR markers have been developed (Table 4), to allow a detailed genetic assessment of populations from different environments and across a major continental disjunct (Figure 5). The patterns of variability detected with this type of functional gene analysis will provide the baseline scien-

	A. jorgensenii	*A. distentifolium*	*K. islandica*
No. of Clones Sequenced	1152	1152	384
No. of Quality Sequences (>100bp)	1050	1065	133
Total No. of SSRs	41	165	24
Percentage of SSRs	3.8%	15.5%	18%

Table 4 Details of the three cDNA libraries and the overall percentage of quality sequences containing SSRs ≥ 11 base pairs.

tific background to develop informed conservation management and species recovery programmes.

This approach is also being applied to important genetic resource collections held at SCRI e.g. Commonwealth Potato Collection (CPC) and soft fruit (*Rubus* and *Ribes*) collections.

Metabolomics. This relatively new field of scientific discovery deals with chemical processes in living organisms which result in a metabolite production. The term "metabolome" usually refers to the entire complement of all low molecular mass, non-peptide metabolite molecules in cells and tissues at a particular physiological and developmental state. To place metabolomics in context, four types of metabolic investigation have been defined (Fiehn, 2001 [5]):

Target compound analysis (analysis of specific compounds most directly affected by a modification or experiment).

Metabolic profiling (analysis of selected compounds from the same chemical group or compounds linked by known metabolic relationships).

Fingerprinting (rapid screening for sample classification, e.g. by global analysis of spectroscopic data, not identification and quantification of individual compounds).

Metabolomics (identification and quantification of as many individual compounds as possible across all compound classes).

Several hundred metabolites can now be screened using appropriate analytical tools such as GC-MS, LC-MS, NMR and FTIR. The metabolites cover many metabolic pathways, increasing the chances of

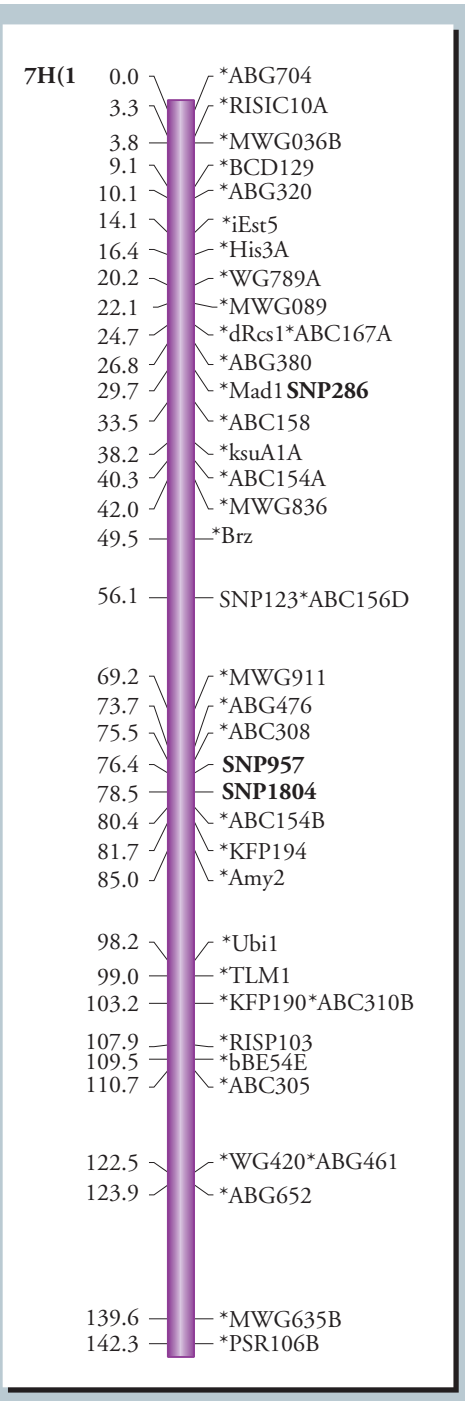

Figure 4 Map location of EST-SNPs on chromosome 7H (Steptoe x Morex)

detecting significant perturbations in metabolic networks that give rise to phytochemical and functional diversity. Because the approach is broad-based, additional functionality properties may be detected even in cultivated plant species which have undergone intensive selection for specific traits of commercial value.

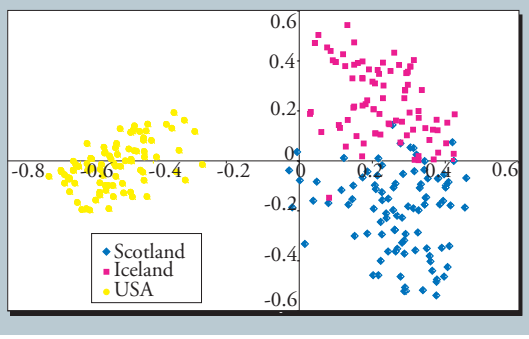

Figure 5 Principal Coordinate Analysis (PCO) of Diversity detected with EST-SSR in *Athyrium distentifolium* populations from Scotland, Iceland and North America.

Work to date has demonstrated that the application of data-mining tools to metabolic profiling analysis allows insights into the relatedness of certain genetic situations. Correlation analysis allows confirmation of established hypotheses concerning metabolic interactions within these systems. As an example, Principal Component Analysis (PCA) of metabolomic data derived from GC-MS has been used to differentiate, through hierarchical data clustering approaches, *Arabidopsis* ecotypes and mutants (Fiehn *et al.*, 2000[6]). However, complete elimination of environmental effects on metabolite profiles is never possible, thus data clusters for a genotype will have a spread of points determined by the 'environmental' influence.

The SCRI is using similar approaches to assess the capacity of metabolic profiling to differentiate between potato accessions maintained in the Commonwealth Potato Collection. Preliminary work using LC-MS of leaf extracts coupled with PCA analysis is already indicating the utility of this approach (Figure 6); In this particular example (an analysis) of 90 accessions of the CPC, two accessions, TBR MPG 501 and CPH 5844, cluster away from the main body of germplasm analysed. This differentiation can be achieved with a dataset based on as few as ten variables (metabolites). When analysis is repeated with all 160 visible metabolites from LC-MS, near total differentiation is achieved for all 90 accessions. The ability to capture such information by observing the broadest possible class of metabolites and relating phytochemical diversity to DNA sequence and transcriptional /translational profiles remains a key challenge.

Connecting sequence Diversity to heritable phenotypic differences An avalanche of DNA information derived from both model and crop plant genomes is

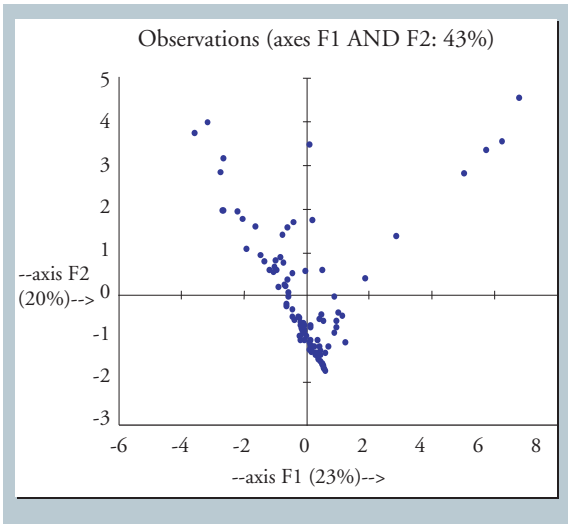

Figure 6 Principal Component Analysis (PCA) of data acquired (LC-MS) from potato leaves sampled from the Commonwealth Potato Collection (CPC)

currently being released into the public domain (http://www.ncbi.nlm.nih.gov/dbEST_summary.htm) As more plant and crop genomes are either being fully or partially sequenced more emphasis is being placed on the discovery and exploitation of intra-specific sequence diversity. Associating and validating the occurrence of sequence polymorphism with phenotypes of strategic relevance to UK agriculture therefore represents a major scientific opportunity of immense relevance to our future.

Access to more molecular markers or ideally fully or partially sequenced genomes provides an opportunity to examine the global patterns of polymorphism. This allows the correlation of allele frequency or non-random distribution of alleles at adjacent loci to be determined (linkage disequilibrium (LD) see Table 3). The extent and magnitude of LD is a prime determinant of the feasibility of association mapping. Association studies based on LD can provide a novel route to the identification and isolation of genes controlling quantitative sources of phenotypic variation. Candidate gene and whole genome scans have both been advocated. For the former approach, a limited number of candidate genes whose biochemical functions hint at a role in controlling a given phenotype is used to identify and test gene-phenotype association. In contrast, genome scans attempt to survey the entire genome for regions implicated in the control of a phenotypic trait. The success of either approach will depend on the extent and distribution of LD together with haplotype structure. Thus, regions with high lev-

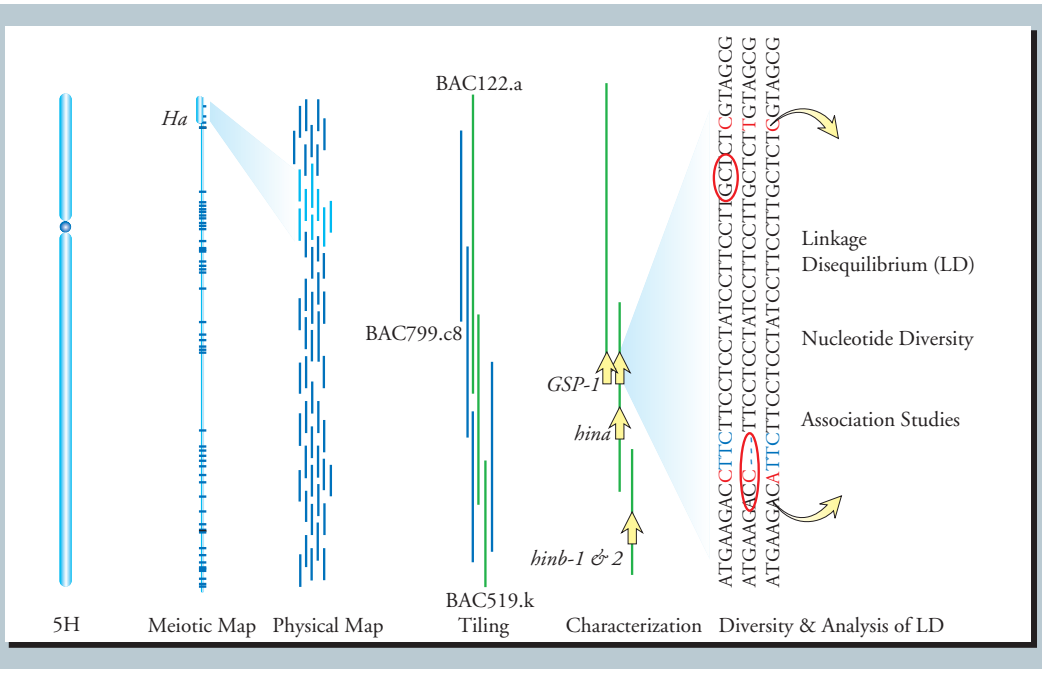

Figure 7 Schematic representation of candidate gene approach for grain texture in barley

els of LD will require few markers but the resulting resolution is likely to be low favouring a whole genome scan to association mapping (Rafalski [2]). In contrast regions, with low LD are likely to require large numbers of markers with the potential for high resolution. A major goal for our research must therefore be to better understand LD distribution and recombination in crop plants.

An example of a candidate gene approach for grain texture in barley is shown in Figure 7. In this case candidate genes hordoindoline-a (hin-a), hordoindoline-b (hin-b) and grain softness protein (GSP-1) have been proposed based on homology to puroindolines genes identified in wheat. A comparative genomic approach has therefore been taken to identify barley BACs clones that harbour the three candidate genes and these have been mapped meiotically to the distal end of chromosome 5H. In the first instance, this approach allows us to anchor BACs to meiotic maps and initiate a forward genetics approach. Surveys of nucleotide diversity across the three genes have also been undertaken to determine haplotype content as a prelude to determining the feasibility of association mapping for quality traits in barley.

In conclusion, we have outlined examples of where biodiversity research is benefiting from opportunities arising from new areas of genome science. However, to fully capture these opportunities the fundamental role of genetics as a tool to analyse biological processes must be reinforced. Genetics together with expertise in whole-organism biology, analytical chemistry, genetic resources and plant breeding are key competencies for the future.

References

[1] Botstein, D., White, R.L., Skolnick, M. & Davis, R.W. (1980). *American Journal of Human Genetics* 32, 314-331.

[2] Rafalski, J.A. (2002). *Current Opinions in Plant Biology* 5, 94-100.

[3] Cardle, L., Ramsay, L., Milbourne, D., Macaulay, M., Marshall, D. & Waugh, R. (2000). *Genetics* **156**, 847-854.

[4]Morgante, M., Hanafey, M. & Powell, W. (2002). *Nature Genetics* **30**, 194-200

[5] Fiehn, O. (2001). *Comparative and Functional Genomics.*, **2,** 155-168.

[6] Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R.N. & Willmitzer, L. (2000). *Nature Biotechnology*, **18,** 1157-1161.