Final Report: 18th January 2008

Inaugural Mathematics in the Plant Science Study Group
University of Nottingham, 17th – 20th December 2007

MEASURING PLANT GENETIC DIVERSITY USING INTER-SIMPLE SEQUENCE REPEATS (ISSRs)

Jane Wishart[1], Kim Evans[2], Theodore Kypraios[2], Simon White[2], Simon Preston[2], Graham Begg[1], Ian Dryden[2], Pietro Iannetta[1, *]

[1] Ecosystem Plant Interactions, SCRI, Invergowrie Dundee DD2 5DA, Scotland UK.

[2] School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD England UK.

* Principle author and to whom correspondence should be addressed: Pete.Iannetta@scri.ac.uk

Summary

Simple sequence repeats (SSRs) have great utility as they are conserved and present in all eukaryotic genomes. Here we report the use of a simple PCR with fluorescently-labelled primers to amplify inter-SSR markers (ISSRs) for diversity assessments. The use of ISSR markers does not rely upon specific genetic sequence information, or prolonged method development and may be measured rapidly using the automated equipment. The major restriction of the ISSR method is at the analysis stage, as the markers are dominant it is not possible to distinguish heterozygotes as loci. We obtained ISSR data from *ca.* 60 phenotypically characterised *Capsella bursa pastoris* L. Medic (shepherds purse) accessions that had been isolated from a diverse mix of arable field sites throughout the UK. We developed mathematical scripts for use with the free statistical software tool R (http://www.r-project.org/), that processed the molecular data in a binary format to estimate genetic diversity (using the Jaccard co-efficient), and that related genotype to the plant phenotypic and environmental (site specific) traits. The methodology established has the power to predict the relationship between environmental and plant morphological characteristics

Introduction

Effective measurement of ecologically meaningful plant-biological variation demands that the unit of diversity be defined at the level of within-species variation (Iannetta *et al.*, 2007). Biodiversity is usually defined taxonomically, at the level of species or genus, often only inferring function. However, species-level diversity descriptors do not accommodate functional differences within species, nor do they correlate with measures of ecosystem productivity; and therefore cannot provide a prediction of ecosystem functional-efficiency. Intra-specific functional diversity quantified from the relative abundance of ecologically significant plant life-history traits may be the best predictor of whole ecosystem productivity (Diaz & Cabido, 2001; Hawes *et al.*, 2005). Towards that goal, we have isolated *ca.* 150 individual accessions of the common annual weed *Capsella bursa pastoris* L. Medic (shepherds purse), from characterised environments. From among these accessions 56 maternal lines have provided 156 individual plants that have been phenotypically characterised *ex situ* (Iannetta *et al.*, 2007). Arable field management traits associated with the different environments from which the *Capsella* accessions were isolated include: longitude and latitude, current crop (when weed was isolated), previous crops and soil pH. Attributes of the weeds isolated from those environments include, leaf shape, time to flowering, reproductive duration and germination-character of the seed yielded. In particular, and from the covariates, it is clear that several site traits and only one plant trait emerged as significant. However, this only represents the data from one primer that must be combined with information gathered using the remaining 5 primers.

In addition, molecular genotyping using Inter-Simple Sequence Repeats (ISSRs), has also been carried out to provide information that complements the site and phenotypic characterisation data. This dataset is to be used to test for relationships between genetic markers and the functional attributes of wild plants at an intraspecific level and in the context of the (agricultural) environment, (including site-management and cultivation history), from which the individuals were isolated. Success in correlating the genetic marker data with functional plant and environmental traits could provide a model that predicts intraspecific-diversity from environmental variables.

Whilst a variety of genetic marker systems exist, most of these require long development times for effective application to plant diversity analysis. We have aimed to establish a robust method that uses genetic data acquired in a (potentially), high throughput approach, that does not rely upon the discovery of specific genetic information, or prolonged method development, ahead of the application. In this respect, inter-simple sequence repeats (ISSRs) emerged as a suitable tool. SSRs are regions of DNA that comprise repeated units of di, or tri, *etc* -nucleotide units; these repeat regions are also referred to as microsatellite sequences. A simple PCR is used to amplify the nucleotide sequence between two such SSR regions; hence the resultant PCR product is termed an 'inter-SSR' (ISSR) marker. No specific sequence information is required to apply this PCR based approach, as only a single PCR primer constructed to be universal (in that it will bind to specific SSRs), may be applied with relative ease to low amounts of even 'crude' genomic DNA. Other molecular marker systems, such as SSRs and AFLPs can demand long development times: test DNA may need screened for suitable polymorphic sites (SSRs). Also, AFLP requires relatively large quantities of high-quality test DNA to accommodate the necessary manipulations; for example, the addition of PCR primer-specific regions (linkers). The ISSR method applied here however, may be applied rapidly and in a high throughput fashion as no prior DNA manipulation, or 'discovery' approaches are necessary before the standard PCR amplification. Fluorescently-labelled primers are used to generate similarly-labelled PCR products whose size (number of nucleotide base pairs) is determined against internal size-standards for each test sample. PCT Product quantity is also measured rapidly and simultaneously quickly using the automated laser-based genotyping equipment. This method is universal in its applicability to all eukaryotic genomes as SSRs are conserved and abundant in a wide range of organisms. Furthermore, ISSRs are inherited in a Mendelian fashion, are hyper-variable (in length) and occur with sufficient frequency to allow a high mapping density that is representative of the whole genome. It is also highly significant that the ISSS method possessing a higher reproducibility across different species and laboratories than other dominant, and non-dominant, marker techniques.

While ISSR markers have been used to detect hybridization in natural populations of plants and animals (Wolfe *et al.*, 1998, Fritz *et al.*, 2005); the major restriction of the ISSR method is that the generated markers are said to be 'dominant'. This is that across the two-chromosome unit that is the basis of eukaryotic genomes, an ISSR marker in not polymorphic in size. Consequently, ISSR markers are either present or absent, and differences between loci cannot be discerned. Consider that an ISSR detected as 'present', may exist on either one, or both, chromosomes. The absence of the marker from one chromosome in heterozygotes cannot be assessed, as the 'present' marker is "dominant" across this locus as it masks the absent band.

Questions and Results

Before any questions could be addressed it was necessary to establish whether we would use the raw analogue output (see example raw data output, Fig. 1., Appendix 1 (below) generated form GeneMapper® software.), where the fluorescent ISSR-PCR products at specific positions on the x-axis (denoting their size in nucleotide base pairs), and amplitude, visualised as peaks. Alternatively, the different PCR products could be converted to binary (presence/absence) information data. The potential utility of using the full analogue output is that the data could be used to give an estimate of the relative quantities of ISSR-products which could prove useful to distinguish the individuals. However, since the ISSR products were not amplified using a validated relative-quantification PCR approach, and issues such as size-homoplasy (peaks occurring to close together), provide exaggerated fluorescent values: the data was processed as a binary data-set.

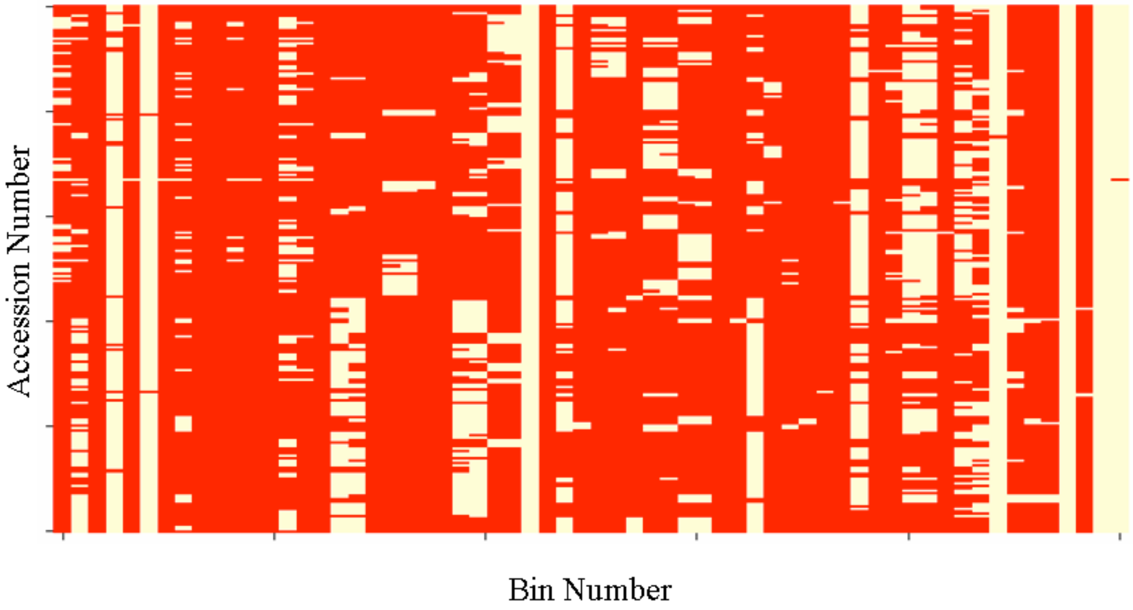


Figure 1. A visual display of genotypes by band presence (cream) and absence (red). The variation within and across all the 109 individuals analysed, demonstrate the relative abundance of particular PCR products (measured as size in nucleotide base pairs) and defined below by consecutive PCR product size (in nucleotide base-pairs, and defined as 'bin number' (1 to 62)). The data shown has been acquired from only one ISSR primer (primer 1417).

Where distinct PCR products were generated, these appeared as clearly visible peaks of fluorescence at specific size locations defined as a 'bin number' (in the case from 1 to 62). Each bin number has associated with it a PCR product whose specific length is given in 'number of nucleotide base pairs' (nbps). The products ranged in length from *ca.* 50 to 1000 nucleotide base pairs. Linear regression of the specific ISSR-PCR products against a set of known length standards (in nbps), allowed the relative length of the products to be determined. Therefore, within an accession, each bin number had associated with it (in addition to size information), either 'no data', or an amount of product in 'fluorescence units' (FUs). The data set for each individual was then standardised by summing the fluorescence across all the positive bins, the FUs for each bin were then converted to a proportion of the total, and this data is expressed at relative fluorescence units (RFUs). Any bins containing 'no data', or less that 1% RFUs, were scored as '0'. All other bins containing positive RFU values were scored as '1'.

Also, before any key questions could be addressed it was necessary to describe the data in more detail so that our understanding of it, and its consequent utility could be better assessed. Towards that end, we quantified the variability of the number of positive PCR products (binary scores), and variation among the peak amplitudes (analogue data), from within each bin and across all the accessions.

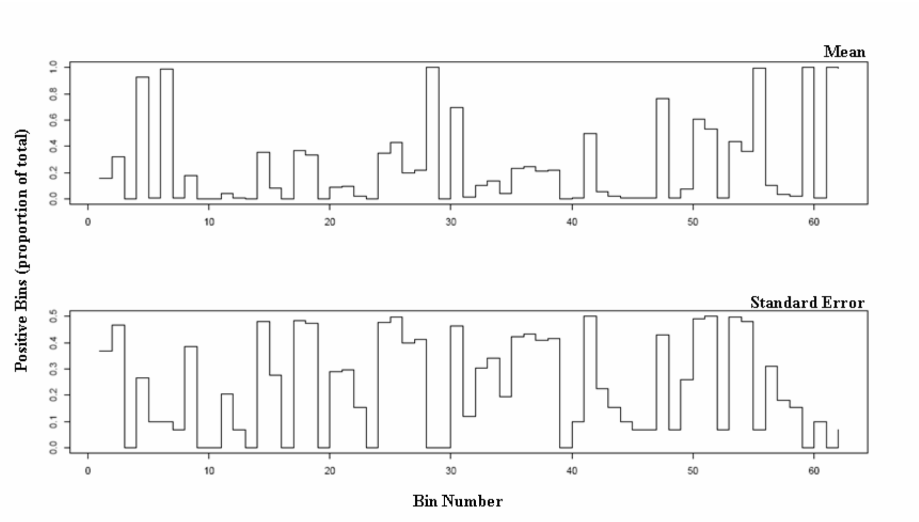


Figure 2. The upper 'mean' graph: the number of positive ('1'), PCR products scored as a proportion of the 109 accessions for number 'bin-number'. Bin number denotes specific PCR-product lengths (in nucleotide base pairs). The lower graph ('standard error'), encapsulates the variability in the raw analogue output: where relative fluorescence for each positive PCR product bin is also scored as a proportion of the total RF for that bin.

Note that low standard errors were confined largely to those bins where only a small proportion of accessions had scored as either positive, or negative. Most variability in PCR-product amplitude was therefore associated with bins that scored positive (or negative), with intermediate frequency.

Key Question 1: Can the data provide information on biodiversity?

Traditionally, population genetic analysis using dominant markers determines allele frequencies from the number of band absences and assumes the Hardy-Weinberg Equilibrium. In this way, the absence of a peak is interpreted as loss of a locus (Wolfe & Liston, 1998). However, the Hardy-Weinburg can be easily violated and band absences may not be due to loss of a locus *per se*, but loss of a primer annealing site perhaps due to nucleotide substitutions, or changes in the PCR product size due to DNA insertions or deletions in the inter-SSR region. Therefore, the absence of bands may overestimate relatedness and (without very cautious application), provides no evidence of common ancestry. The Jaccard Coefficient is a measure of the similarity between two binary strings where we are not interested in a zero occurring in each string,

$$J = a / (a+b+c), \qquad \ldots (1)$$

where

a = number of two individuals proving positive for a specific PCR product 'bin number': *1:1*
b = mismatch *1:0*,
c = mismatch *0:1*.

To create a distance measure between two strings we calculated: *1-J*.

For comparison we also used a Euclidean metric, with the binary strings thought of as a point in a Euclidean space of n dimensions, where n is the number of 'bins'. We compared the output of these two matching coefficients (Figure 3), and found that the Jaccard measure gave the most marked differences. The Jaccard was therefore selected for further analysis. However, it may be beneficial to compare the Jaccard output with other non-Euclidean similarity coefficients such as Nei and Li (also known as the Dice, or Sørensen's coefficient: $2a / (2a+b+c)$; again excluding the '*d*' (*0:0*) component), that places more weight on positive matches.

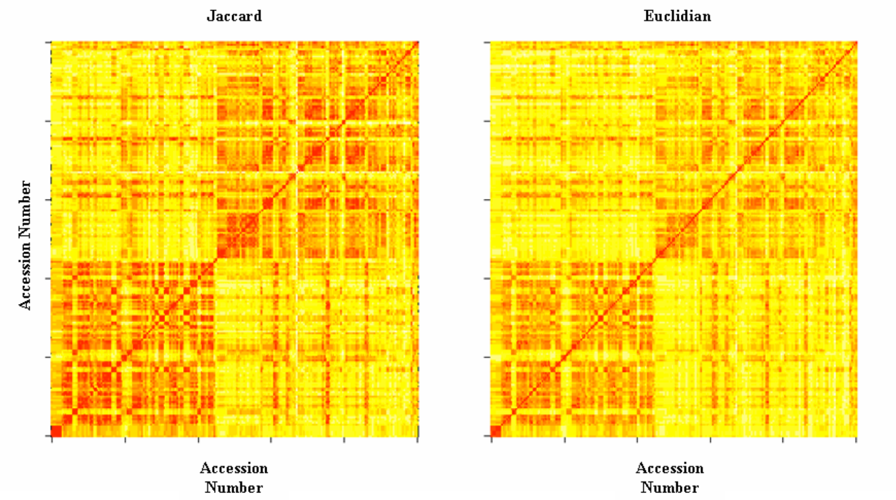


Figure 3. A 'heat-map' of distance matrices generated using Jaccard and Euclidean distances. Small distances are shown as red, and large distances as yellow. The darker values (high values and greater differences) produced by the coefficient of Jaccard, (relative to the Euclidean output) is apparent. The red diagonal centre-line consists of all 0's which corresponds to individuals compared with themselves, where obviously the distance is zero. Note that the heat-maps are symmetric about these midlines.

This approach may be selected as preferable to that suggested by other researchers who use ISSRs for a genetic survey of community composition. Whitlock *et al.,* (2007), used a binary DNA marker data (ISSR) to assess the probability of identity ($P_{(\mathrm{ID})}$): that individuals selected at random for a population are different. However, whilst the $P_{(\mathrm{ID})}$ assesses relative genotype abundance to assess community structure, it does not test the correlation between genetic diversity and functional attributes of the plant or environment, which is our aim here.

Key Question 2: Can the molecular data predict site or phenotypic traits?

Genetic-distance (biodiversity) estimates for pairs of observations may be calculated using *1-J*; where *J* is the Jaccard coefficient, and used to construct phylogenetic-trees for the 109 accessions that divided the observations into four distinct clusters.

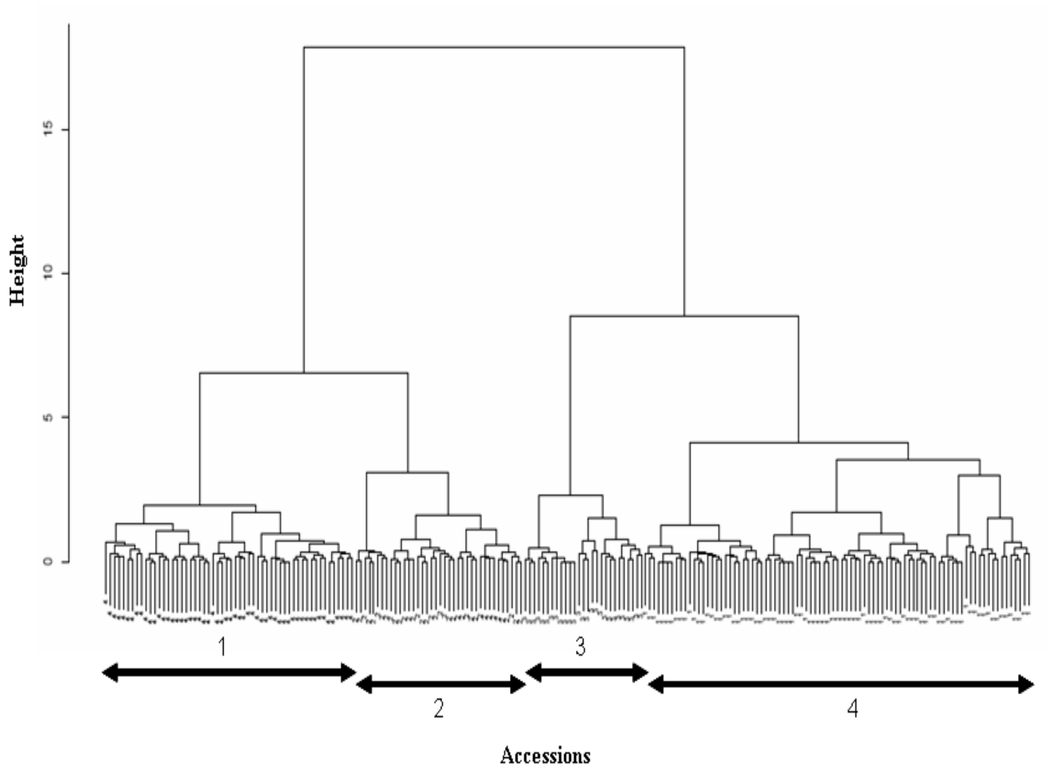


Figure 4: A hierarchical cluster analysis of the binary data from 109 accessions using the Jaccard statistic was used to generate the hierarchical tree below. Individuals belonging to each of the four clusters that were distinguished are indicated using the arrowed-lines numbered 1-4.

Principle coordinate analysis was also used to assess the degree to which individuals belonging to the four clusters could be separated. This was done by finding a set of mutually perpendicular axes such that the first axis (or component), described the greatest possible proportion of the variability among the Jaccard estimated from specific PCR products (‘bins’), of the binary molecular data set, the second axis (or component), described the greatest possible proportion of the variability among other remaining bins, and so on through the remaining axes. The ability of the first three components to isolate each of the four distinct genotypic groups is shown in Fig. 5A. Also, the proportion of the variability explained by each component is also shown (Fig. 5B).

.

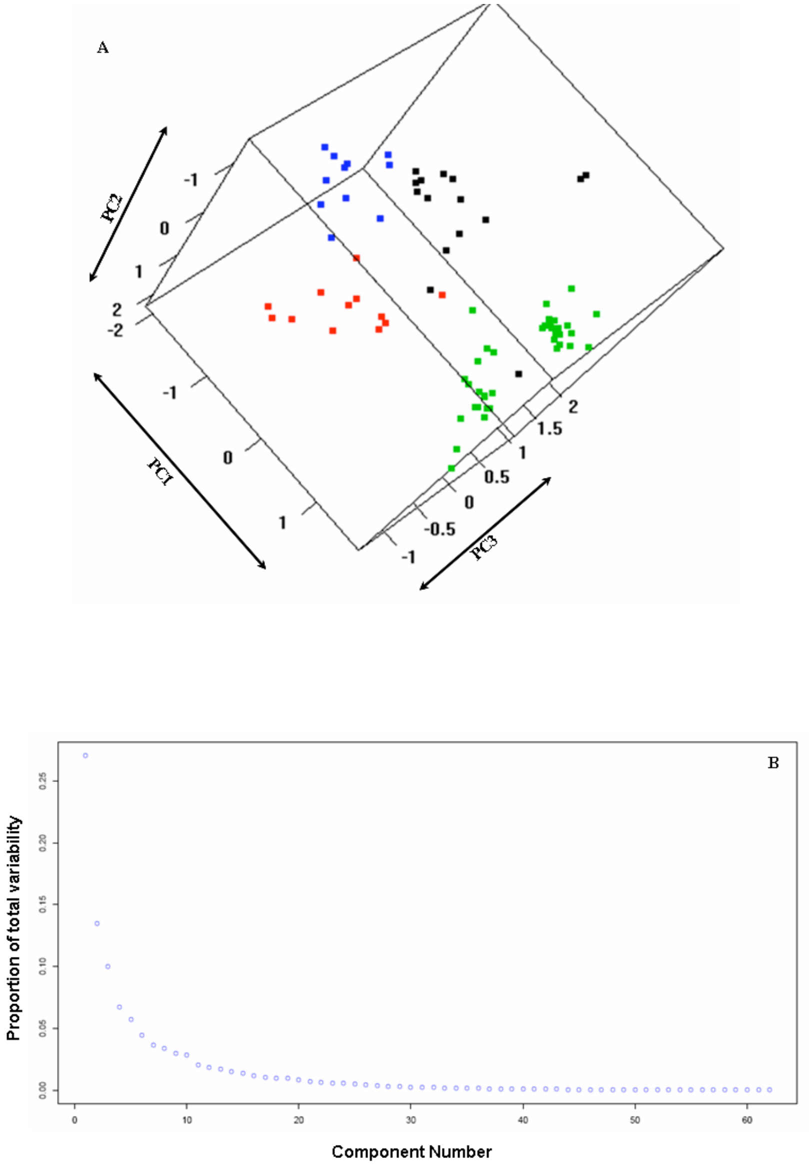


Figure 5. PCA of individual accessions identified as belonging to one of four clusters from genetic-distance (Jaccard) estimates. **A**, three-dimensional PCA plot separates the four clusters over three dimensions. **B**, a ranking of the percentage variability accounted for by each dimension of the PCA analysis

Regression analysis was carried out with the first, second and third dimension PCA scores on functional site and trait covariates. The first PC distinguishes clusters 1 and 3 from clusters 2 and 4 (the two super-clusters), and the significant covariates are **current crop** (with GM beet having a significantly greater affect than GM rape or GM maize), **latitude** and **previous crop two years ago**. The second PC distinguishes cluster 3 from the other clusters and the significant covariates are **soil pH**, **latitude**, **previous crop two years ago** and (‘mid’) **leaf-shape**. In particular, and from the covariates, it is clear that several site traits and only one plant trait emerged as significant. However, this only represents the data from one primer that must be combined with information gathered using the remaining five primers.

In addition, from the PC eigenvectors (Fig. 6), the bins that contribute most to the PC scores are those whose weightings deviate most from zero. For example, for the first PC score (Fig, 6A) bins 17, 18, 42 all show large negative weightings, and 30, 50 and 51 show large positive weightings.

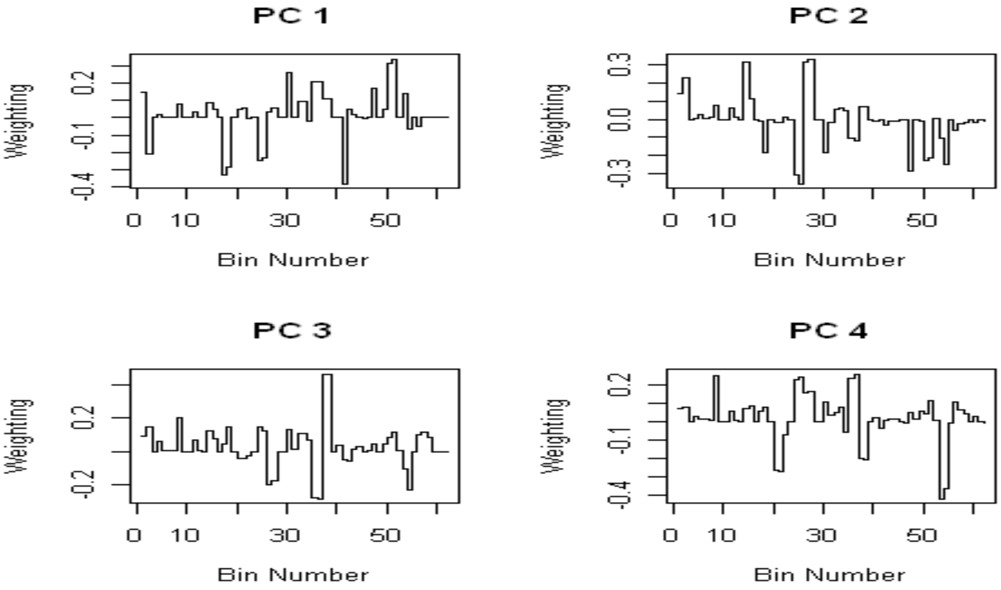


Figure 6. PC vectors showing the weightings associated with each bin for the first four principal components. Large positive and negative bins indicate peaks that may have predictive value.

We may be able to predict diversity from the genetic data. We could use the large positive, or negative, weightings to identify key bins, or key-bin positive/negative patterns that are indicative of particular functional attributes. Additional data would also improve the accuracy of this exercise and is in hand: individual plant samples have been gathered from a variety of different field sites that encompass different farming-intensity regimes from across Scotland. In the interim the threshold at which bins or bin-patterns are to be either included or excluded with maximum utility still can be assessed by iterative testing with a modelling approach on the current (full) data set.

With this information in mind we re-examined the genetic diversity within each of the clusters using the Jaccard Coefficient (*J*) with *D(i ; j)=1-J* being the Jaccard distance between individual *i* and individual *j*. From a matrix of inter-point distances for the individuals of each cluster, the average value of *D(i,j)* was calculated. Population-average Jaccard distances were 0.360, 0.323, 0.339, 0.304 for clusters 1 to 4 respectively. The average Jaccard distance between the whole population was 0.493. It would appear that diversity is not equal between the four clusters: with cluster 1 > 3 > 2 > 4. Whether these differences are statistically significant requires more work using ANOVA of the Jaccard statistics of each group, the traits that relate to each collection of accessions and their respective source field-sites. This shall be an objective of future analysis.

Future Work

The analysis of the molecular data reported here and in the figure above relates only to the products of a single ISSR primer of six. There are therefore a further five genetic data-sets to be analysed. The traits that correlate with each primer-output need compared between primers. It may be that the variability associated with each primer will describe the same suite of significantly correlated plant and site attributes. Otherwise, there will be either no correlation, or that different primers may identify different attributes. It is worth noting at that stage, that the correlations distinguished and reported above used data filtered to remove those PCR products that contributed less that 1% of the total RFU. The findings using more stringent threshold could also be assessed (*e.g.* using only those products that contribute over 5% of the RFU cut-off), which may also give the same predictive information of greater robustness as the products are more-obviously contributory.

The binary output, [1-Jaccard Coefficient (J)], was identified as a suitable diversity statistic and was used to produce genetic-dissimilarity trees, identifying genotypically distinct individuals associated within clusters. Variation among the diversity estimates correlated with site specific and accession-specific functional attributes. We can now assess the genetic diversity (J) of particular sub-populations, or clusters. Multivariate analysis suggested genetic peaks that may predict functional attributes. However, the utility of this approach has still to be modelled. For example, remaining to be assessed is the genetic data from the five remaining primers and the combined output modelled to address other questions that relate the choice of ISSR-primer to the number of loci, or individual accessions, that should considered when assessing diversity in this way.

Successful application of the analysis described demands familiarity with the mathematical and statistically software package ‘R’. The may be downloaded for free from a variety of different internet sources. The specific scripts for the ISSR data analysis described above are not presented here but are available through contact with the corresponding author.

References

Diaz & Cabido (2001) *Trends in Ecology Evolution* **16**, 646-655.
Fritz *et al.* (2005) *Molecular Phylogeny Evolution* April.
Hawes *et al.* (2005) *Oikos* **109**, 521-534.
Iannetta *et al.* (2007) *Physiologia Plantarum* **129**, 542-554.
Waits *et al.* (2001) *Molecular Ecology* **10**, 249-256.
Whitlock *et al.* (2007) *Journal of Ecology* **95**, 895-907.
Wolfe & Liston (1998) *Plant Molecular Systematics II.* Chapman Hall, New York, pp43-86.
Wolfe *et al.* (1998) *Molecular Ecology* **7**, 1107-1125.

Email Contact Details

Ian Dryden:	pmzild@exmail.nottingham.ac.uk
Kim Evans:	Kim.Evans@nottingham.ac.uk
Pietro Iannetta:	Pete.Iannetta@scri.ac.uk
Theodore Kypraios:	theodore.kypraios@nottingham.ac.uk
Simon Preston:	pmzspp@exmail.nottingham.ac.uk
Simon White:	pmxsw@exmail.nottingham.ac.uk

APPENDIX I

THE MATHS WORKSHOP PROPOSAL AS ORIGINALLY SUBMIT IN NOVEMBER 2007

Measuring genetic diversity? Wishart J. & Iannetta P.P.M (Nov. 2007)

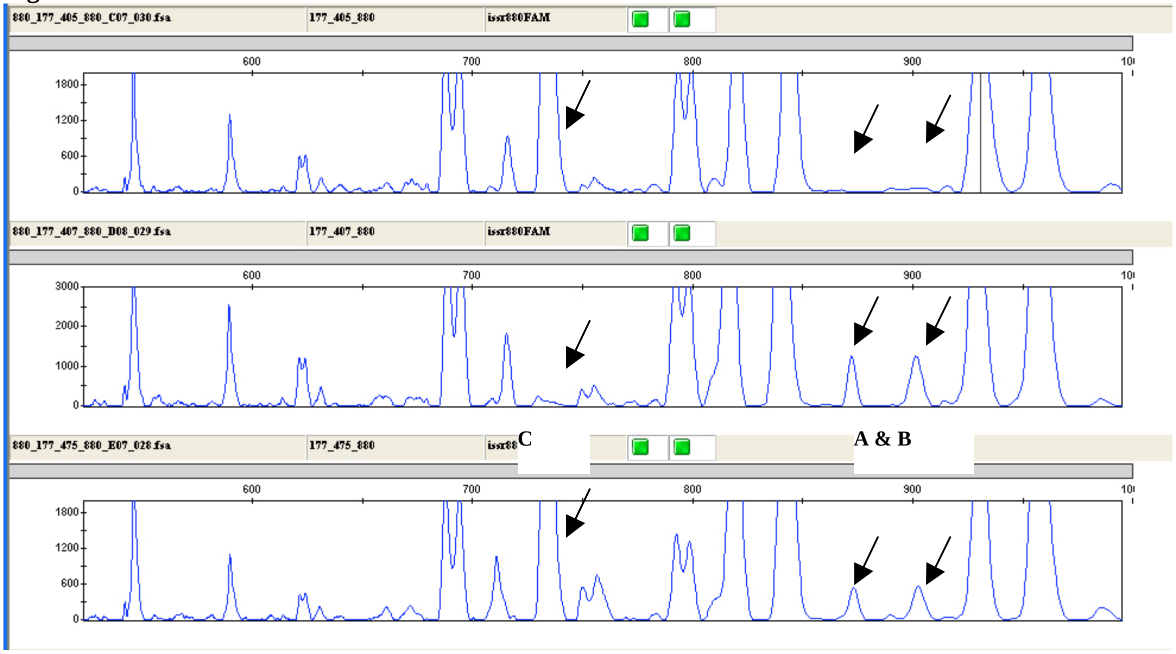
Effective measurement of ecologically meaningful plant-biological variation demands that the unit of diversity be defined (Iannetta *et al* 2007). Diversity in the arable flora has been defined taxonomically by species or genus, often inferring function. Levels of intra-specific variability are largely unknown and mainly ignored (Hawes *et al* 2005). To understand ecosystem function it is necessary not only to identify species diversity but also to monitor individuals within populations. Loss of diversity can occur at different scales and it could be that simply observing species diversity we miss large scale loss of diversity occurring at the population level. The effect of biodiversity upon system function could be better considered by describing the relative abundances of ecologically significant plant life-history traits (Diaz & Cabido, 2001) and there is a need to quantify the extent of within and between species variation in traits linking plants with environment (Iannetta *et al* 2007).

We have just completed a study, using ISSR combined with high-throughput, fluorescent technology, to measure the genetic diversity found in an arable weed species, *Capsella bursa pastoris*. By using a system such as ISSR markers we have been able to identify individual variation within our collected ecotypes of *Capsella* and relate these broadly to function. We have been able to distinguish genetic identities among our collected ecotypes of *Capsella* and can correlate these to four distinct plant functional groups. We have found statistical evidence that certain loci (peaks) are associated with certain phenotypic characters and that these markers are heritable.

ISSR uses a single PCR primer binding to di or tri-nucleotide repeats (Microsatellite sequences) which are abundant in eukaryotic genomes. Microsatellite sequences are conserved over a wide range of organisms and therefore the PCR can use universal primers to amplify the nucleotide sequence between two simple sequence repeats (SSR) with the priming sites oriented on opposite DNA strands. SSR regions are scattered throughout the genome and produce highly polymorphic bands which vary within and between species giving a measure of intraspecific variation as well as the possibility of determining species specific patterns and measuring genetic differences dispersed across the entire nuclear genome. Absence of a peak is interpreted as primer site divergence or loss of a locus (Wolfe & Liston, 1998). ISSR markers have been used to detect hybridization in natural populations of plants and animals (Wolfe *et al* 1998, Fritz *et al* 2005).

Fig 1, see below, shows traces from three individuals with the same parent. Peaks A & B & C show presence in 2 of the 3 offspring. Differences between siblings are fewer than between individuals from different parents. Peaks correspond to PCR products of different sizes (top axis is base pair length) and therefore to amplified nuclear sequences. Each peak area is calculated as a relative proportion of the total fluorescence for that individual sample. There may be 80 “bins” (peak positions) scored for one primer with each individual having around 20 peaks.

Fig 1.



Most interesting to us, however has been the use of ISSR for a genetic survey of community composition (Whitlock *et al* 2007); “the DNA marker data (ISSR) were used to create , for the first time, a genotype abundance hierarchy describing the structure of a community at the level of genotypes”. We have extended and adapted this approach by relating the genotype to the phenotype thus confirming a functional link. While our data has provided clear genotypic identities for the individuals, the information has not yet been extended to provide a general measure of molecular diversity.

Objective:
Could ISSR data be used to define a useful measure for biodiversity which would enable ecosystem diversity to be modelled and predictions reliably made about future viability of populations and species within changing environments based on the link between genotypic and phenotypic variability.

The questions:

1. *What statistic based on the ISSR data would provide a suitable measure of genetic diversity? [Is the method (developed from Waits et al, 2001) used by Whitlock et al (2007) of calculating Pid (probability of genetic identity) the best method for defining overall genetic diversity?]*
2. *How would the statistic be estimated? [Could a model be defined which would enable testing of fewer individuals to give an estimate of genetic variation?]*
3. *How would the statistic be compared between populations?*
4. *How would the statistic be related to functional diversity within and between populations?*
5. *could this be modelled in a way which includes predictions based on knowledge of the selective effects of environmental changes.*

To achieve these

6. *How many loci would have to be considered?*
7. *How many individuals from each population would have to be tested?*

References

Diaz, S., & Cabido, M. 2001. Trends in Ecology and Evolution 16, 646-655.
Fritz, U., Siroky, P., Kami, H., Wink, M. 2005. Molecular Phylogeny and Evolution, April 2005.
Hawes, C., .Begg, GS., Squire, GR., Iannetta, PPM. 2005. Oikos 109, 521-534.
Iannetta, PPM., Begg, G., Hawes, C. Young, M., Russell, J., Squire, GR. 2007.
Waits, LP., Luikart, G., Taberlet, P. 2001. Molecular Ecology, 10, 249-256.
Whitlock, RAJ., Grime, JP., Booth, R., Burke, T. 2007. Journal of Ecology 95, 895-907.
Wolfe, AD., & Liston, A. 1998. Plant Molecular Systematics II. Chapman Hall, New York, pp43-86.
Wolfe, AD., Xiang, Y., Kephart, SR. 1998. Molecular Ecology, 7, 1107-1125.