Germplasm collections: Gaining new knowledge from old datasets

Robert Davey (1,2), George Savva (1,2), Runchun Jing (3), Martin Lott (4), TH Noel Ellis (1), Michael Ambrose (1), Vincent Moulton (4), Andrew Flavell (3), Ian Roberts (2) and Jo Dicks (1)

1) John Innes Centre, Norwich Research Park, Colney, Norwich, NR4 7UH, UK.
2) Institute of Food Research, Norwich Research Park, Colney, Norwich, NR4 7UA, UK.
3) University of Dundee at SCRI, Invergowrie, Dundee, DD2 5DA, UK.
4) University of East Anglia, Norwich, NR4 7TJ, UK.

Abstract

Over several decades, germplasm collections have been developed across the world to capture the genetic diversity of crop plants vital to food and agriculture. Recently, the genetic characterisation of many of these collections has begun, using a variety of genetic marker technologies. Here, we describe some first attempts at uncovering the genetic structure of a single collection characterised by high-throughput marker techniques. This research both hints at the knowledge that may be gained by analysing such datasets and identifies areas of research that should be targeted for the future.

Introduction

The production of crops, a large part of the worldwide food supply, relies on intensive agricultural practices that can lead to genetic uniformity. Such uniformity creates risks regarding maintaining protection against pests, disease and environmental change. Plant breeders wish to develop new crop varieties that can overcome old adversities and deal with new ones as they arise. In order to achieve this, the breeder must have access to a wealth of genetic diversity in their crop of interest. The development of germplasm collections, which capture this diversity, has been ongoing for decades. Collections have been developed though national projects and through international collaborations. For example, the work of the Consultative Group on International Agricultural Research (CGIAR; http://www.cgiar.org/) has led to the systematic collection of specimens of landraces, old cultivars, wild species, advanced cultivars and breeders lines, for cassava and sweet potato to rice and maize. The eleven CGIAR international genebanks currently maintain over 600,000 crop, forage and agroforestry samples in the public domain, providing massive datasets to be analysed in the coming years.

The historical driver for the development of germplasm collections was to preserve and document crop genetic diversity, thus ensuring future food security, whilst also evaluating and distributing the germplasm. Recent advances in genetic marker technology are leading to the growing genetic characterisation of germplasm collections, allowing for informed exploitation of the germplasm in future breeding studies. Germplasm collections may hold important *alleles* (or versions of a gene) for agronomic traits such as disease resistance, yield and tolerance to a broad range of environmental conditions. With the molecular characterisation of germplasm collections comes the ability to carry out detailed analyses on their genetic structure. For example, we may wish to search for associations between traits and alleles or perhaps traits and haplotypes (allelic combinations of adjacent genes). We may wish to understand the evolutionary history of the species, in particular the balance between the different processes of genetic marker evolution (vertical evolution) and introgression (introduction of a gene or haplotype from one variety to another via hybridisation - horizontal evolution). We may wish to know how closely related two members of a collection, or *accessions*, are to one another. We may wish to examine all relationships within the collection and use this to develop a *core collection* that maximises the diversity for a small, fixed number of accessions.

In this article, we will introduce ongoing research that attempts to answer some of these questions. We will begin by describing the molecular characterisation of a germplasm collection for pea by a recently developed high-throughput marker technique. Such techniques bring with them new challenges for determining marker scores from the resultant raw datasets. We will then discuss how we can use the marker scores to assess the genetic difference between accession and estimate the structure of an entire germplasm collection. Finally, we will touch upon ongoing research into the estimation of efficient core collections.

Marker prediction from high-throughput datasets

With the desire to analyse the genetic structure of a germplasm collection, there comes an interesting debate as to which type of molecular marker is most appropriate for the task. The last decade has seen the rapid development of marker technologies in the plant domain, from RFLPs, SSRs and AFLPs through to SNPs (single nucleotide polymorphisms), SSCPs (single stranded conformation polymorphisms) and RBIPs (retrotransposon-based insertion polymorphisms), with some marker types targeting genic regions of the genome and others deriving from alternative genomic features.

RBIPs [1] are based on a genomic element known as a *retrotransposon*. Such an element can be thought of as a mobile piece of DNA that inserts itself within a genome and subsequently jumps to a new genomic location, whilst leaving a copy of itself behind. Thus retrotransposons accumulate within the genome, leading to an observed growth in plant genome size. As these elements can only be gained, and not lost, through their normal mode of evolution, they can help us to understand the direction of evolution by the order of their accumulation. However, introgression can lead to the appearance of an element being lost or gained without a deletion or a jump taking place. Each retrotransposon type possesses a number of locations within a genome at which it can be inserted. Therefore each plant accession can be characterised by a particular pattern of presence and absence of the retrotransposon at each of these locations. Formally, for a particular retrotransposon, there is a fixed order i (along the genome, if this information is known, otherwise a conceptual order for clarity only) and number N (i.e. $i = 1,\ldots,N$) of locations at which the retrotransposon may be present or absent. Thus, for each plant j each value $m_{\mathrm{i,j}}$ denotes the presence or absence of a copy of the retrotransposon at position i in plant j. We say that $m_{\mathrm{i,j}} = 0$ when the retrotransposon is absent and $m_{\mathrm{i,j}} = 1$ when the retrotransposon is present.

A high-throughput experimental technique for the assaying of a single RBIP marker in a large number of plant accessions (e.g. several thousand) has recently been developed [2]. This technique is known as the tagged microarray marker (TAM) approach. TAM microarrays have recently been used to characterise the John Innes *Pisum* Collection (http://www.jic.bbsrc.ac.uk/germplas/pisum/) using the PDR1 retrotransposon. This characterisation has taken the form of 76 experiments (one for each insertion site) over 3,029 *Pisum* accessions together with 171 positive and negative controls. The experiment measures the relative levels of red and green fluorescently labelled probes for each plant. The probes are specific to each genomic location such that, at a particular insertion site, the red probe is designed to be indicative of the absence of the retrotransposon at that particular location and the green probe is designed to be indicative of its presence. Thus for each experiment i, we are given a measure of intensities of green and red for each plant j, $g_{\mathrm{i,j}}$ and $r_{\mathrm{i,j}}$ respectively. The first problem presented to us is to use these values of $g_{\mathrm{i,j}}$ and $r_{\mathrm{i,j}}$ to predict the corresponding values of $m_{\mathrm{i,j}}$.

The recent widespread use of gene expression microarrays in biological research has taught us many lessons about analysing such datasets. We know that, prior to comparison of our red and green intensities, we need to *normalise* the raw data. In particular, we have chosen to use the *vsn* routine [3] within the BioConductor suite [4] of the R statistical package (http://www.r-project.org/). This algorithm both calibrates the red and green values (i.e. brings them onto a common scale so that the intensities can be directly compared) and stabilises their variance (i.e. transforms the values so that their variance is no longer a function of intensity but is more or less constant across the intensity scale). Figure 1 shows two distributions of the ratios of red and green intensity levels after they have been analysed with *vsn*, one for each of two markers. The left distribution is bimodal, with the left peak representing accessions where the retrotransposon is present (i.e. the green intensity level is significantly higher than the red intensity level) and the right peak representing accessions where the retrotransposon is absent (i.e. the red intensity level is significantly higher than the green intensity level). The right distribution is more difficult to analyse, with four significant peaks. In this example, it is most likely (comparing the distribution to that of other markers) that the leftmost peak represents low intensity values that cannot be analysed with any certainty. The second left peak represents the accessions with the retrotransposon present and the rightmost peak accessions with the retrotransposon absent. The remaining peak represents "yellow" spots where the red and green intensity levels are comparable. At first sight, one would presume that these were plants that were *heterozygous* for the retrotransposon insertion (i.e. on one copy of the relevant chromosome the retrotransposon was present and on the other it was absent). However, it is known that the plants within the *Pisum* collection are *homozygous* for these retrotransposons (i.e. both copies are present or both are absent). What appears to be happening is that some insertions reside in repeated sequences, so a plant containing an 'occupied' signal from a locus might nevertheless produce an 'unoccupied' signal from another copy of the repeat elsewhere in the genome. In some cases such problems can be solved but further research must be done to resolve this issue.

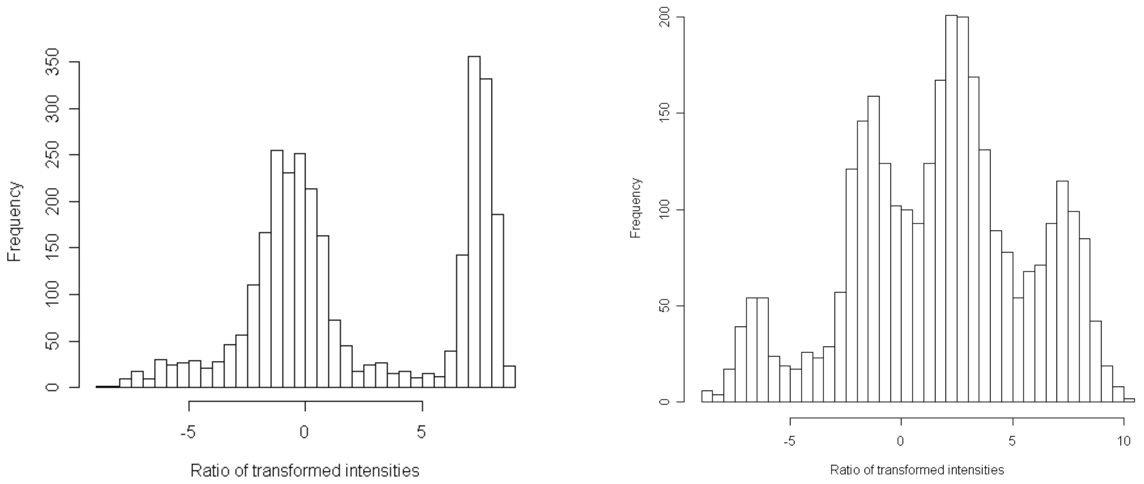

Figure 1: Two distributions of transformed intensity ratios for a single RBIP marker, each assayed by a TAM microarray in 3,029 varieties of *Pisum*

Once we have determined the meaning of these peaks, we can use mathematical techniques to predict the status of each marker within each plant accession. We have recently fitted Gaussian (normal) mixtures to the distribution of transformed intensity ratios to predict marker presence or absence. Furthermore, we are currently automating the analysis of TAM microarrays within our MPP (microarray-to-phylogeny pipeline) software (http://cbr.jic.ac.uk/dicks/software), which was originally developed for the analysis of Comparative Genomic Hybridisation (CGH) microarrays. By combining the results of each array analysis, we produce a table of 76 x 3,029 elements, with each element being a 1 or a 0. We can now use this table to find out more about the relatedness of our accessions and the overall structure of the germplasm collection.

Measures of distance

When comparing the marker scores of two or more accessions, we need to have some measure of comparison. Usually, we will use a measure of the distance between two sets of marker scores, where we would like this distance to be strongly correlated to the real evolutionary time separating them. There are many distance measures in the biological domain that are used to compare sets of binary characters, such as our RBIP marker scores. We will now discuss briefly two distance measures: the Jaccard distance and the retrotransposon distance.

Jaccard distance

A distance measure widely used with genetic markers is the Jaccard distance which, for our data, can be calculated as follows between two accessions *x* and *y*:

$$d_J(x, y) = a / (a + b + c)$$

where
a = no. of markers i where $m_{i,x} = 1$ and $m_{i,y} = 1$
b = no. of markers i where $m_{i,x} = 0$ and $m_{i,y} = 1$
c = no. of markers i where $m_{i,x} = 1$ and $m_{i,y} = 0$

The Jaccard distance is very easy to calculate, even for large datasets. Furthermore it is widely used and understood and may be used flexibly for many types of marker. However, because it is widely applicable it may not maximise the information contained within a particular type of dataset. For this reason, we are looking to develop a custom distance measure for RBIP datasets.

Retrotransposon distance
We have recently begun looking at ways of modelling the retrotransposon insertion process. If we suppose that retrotransposons arise according to a simple birth process and, furthermore, that a proportion of insertion sites ρ are invariant (i.e. always empty) and that rates of insertion per site vary across the genome according to a Gamma distribution with shape parameter α then the maximum likelihood estimates (Savva, manuscript in preparation) of the retrotransposon distance between an accession x and the *reference* accession *0* (where all sites are empty) and between two accessions x and y are as follows:

$$d_R(x,0) = \left(\frac{1-\rho}{\frac{N_x}{N} - \rho}\right)^{\frac{1}{\alpha}} - 1$$

$$d_R(x,y) = \begin{cases} 2\left(\frac{1-\rho}{\frac{N_{xy}}{N}-\rho}\right)^{\frac{1}{\alpha}} - \left(\frac{1-\rho}{\frac{N_x}{N}-\rho}\right)^{\frac{1}{\alpha}} - \left(\frac{1-\rho}{\frac{N_y}{N}-\rho}\right)^{\frac{1}{\alpha}} & \text{if } \frac{N_x N_y}{N^2} < \frac{N_{xy}}{N} \\ \left(\frac{1-\rho}{\frac{N_x}{N}-\rho}\right)^{\frac{1}{\alpha}} + \left(\frac{1-\rho}{\frac{N_y}{N}-\rho}\right)^{\frac{1}{\alpha}} - 2 & \text{otherwise} \end{cases}$$

where N is the number of insertion sites, N_x is the number of empty sites in x, N_y is the number of empty sites in y and N_{xy} is the number of sites empty in both x and y.

At present, this distance is a very simple model of the retrotransposon insertion process and we have yet to compare its performance to that of the Jaccard distance. In the future, we intend to validate and extend the model. For example, we would like to be able to analyse more than one retrotransposon type simultaneously. Furthermore, we need to take into account that most crop plant germplasm collections contain strongly conserved, fragmented haplotypes, which have been distributed across the species by introgression (i.e. horizontal evolution) Thus, two apparently highly diverged plants might be almost identical for a large fraction of a particular chromosome(s). The new model should take into account the introgression process such that we will be able to formally estimate the relative contributions of insertion and introgression to the evolution of a group of accessions, while analysing more than one retrotransposon type simultaneously.

Deducing network-like structures
Once we have established methods of calculating distances between pairs of accessions, we can use these values to analyse the pattern of genetic diversity within the collection as a whole. Traditionally, many types of biological dataset have been viewed as tree structures, after the “tree of life” thought to connect all living organisms. However, it has become apparent in recent years that trees will not always describe adequately a biological dataset and that network-like evolutionary events may play an

important role in shaping such datasets. Several algorithmic methods have been developed to estimate some type of network from a matrix of distances. One such method is the NeighborNet [5], which is implemented in the SplitsTree4 software [6]. NeighborNet essentially extends the widely used neighbor-joining algorithm [7], one of the most popular methods of tree construction in the biological domain, to one capable of deducing a planar phylogenetic network. NeighborNet allows the researcher to visualise areas of the graph that are inconsistent with a tree-like structure, via “box-like” features. For a germplasm collection, such features may represent introgression events, which do not follow a treelike evolutionary mode.

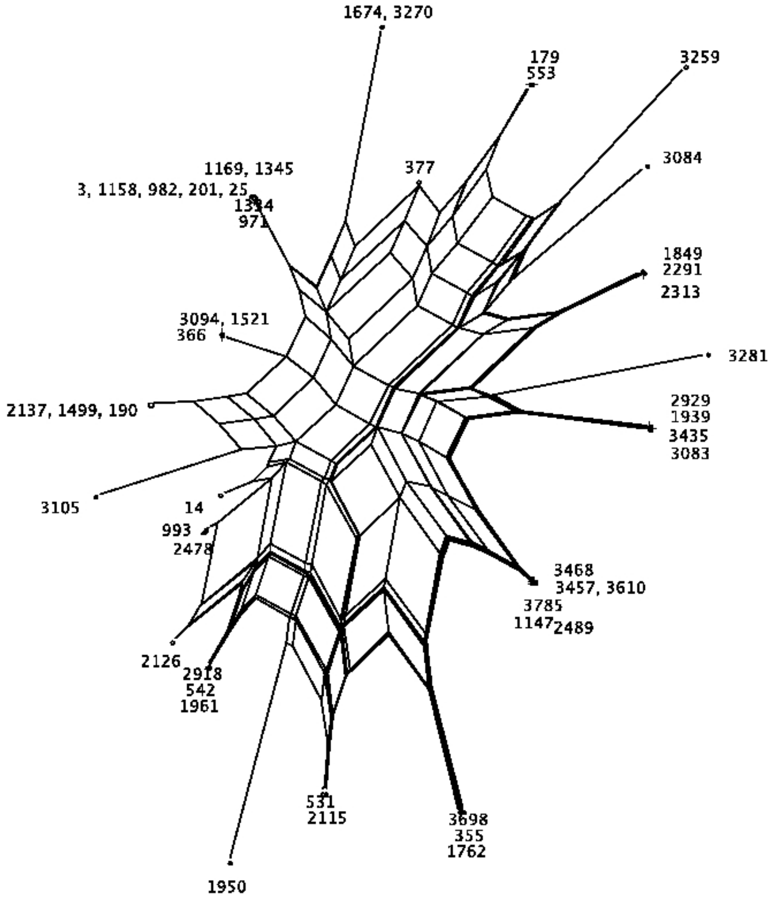

Figure 2: A NeighborNet of 50 *Pisum* accessions scored over a subset of 7 RBIP markers

Figure 2 above shows a NeighborNet of 50 *Pisum* accessions, assayed over 7 RBIP markers and with evolutionary distances estimated using the Jaccard distance. It will be interesting to see whether or not the apparently significant network-like structure seen within this figure remains when all 76 markers have been analysed.

Estimating a core collection

Having calculated both a distance matrix and a neighbor-joining tree or NeighborNet (whichever is most appropriate) for a set of accessions within a germplasm collection, we would then like to use this information further to find an efficient *core collection*. A core collection may be thought of as a subset of the overall collection that describes most of its diversity (whether genetic, geographical or phenotypic) for a fraction of its size. Typically, a core collection comprises 10% of the number of accessions seen in the whole collection. Therefore, for a fixed number *c*, the required size of our core collection, we need to find the network that displays the maximum amount of diversity over *c* accessions.

We have recently begun to develop new approaches for the estimation of core collections based on genetic diversity. The starting point of this research is the greedy algorithm proposed by Steel [8]. This algorithm proceeds as follows, either from a distance matrix or a neighbor-joining tree, where *c* is the number of accessions within the core collection and G is the set of all accessions:

```
Choose the pair of accessions most diverged from one another within set G
Add both accessions to the current accession set S
While ($|S| < C$) do
    (Choose the accession from G that is most diverged from S
     Add this accession to S)
```

This algorithm is simple to implement and has been shown to give a guaranteed solution to the problem of finding an optimal genetically-based core collection from a distance matrix or a neighbor-joining tree, with no constraints on collection members or their properties. However, for a computational solution to be of real practical benefit to germplasm collection managers, other factors need to be considered. For example, it would be useful to be able to place constraints on datasets, as different managers will have different priorities for selecting core collections such as requiring allelic variation at a particular site or only including accessions with particular characteristics. Furthermore, many traditionally created core collections attempt to maximise variation not only genetically but also geographically and phenotypically and this needs to be taken into account in algorithmic approaches. We also need to develop techniques to account for missing data. For example, in our current marker analysis of the JIC *Pisum* dataset, roughly 10% of the dataset is uninterpretable, an unfortunate but common downside to high-throughput techniques. If we were to use basic techniques for dealing with missing data, we would ignore any marker where could not determine a score for one or more accessions. In some datasets this could mean ignoring a large proportion of the dataset. Clearly, more research in this area is required to maximise the information gained from germplasm collections with missing marker scores (or indeed any other type of missing data).

Discussion

Germplasm collections are essential resources for maintaining and documenting crop diversity and for developing efficient and targeted plant breeding studies. They contain useful alleles and allelic combinations that may help to combat crop disease and to overcome environmental pressures. High-throughput marker technologies present us with large, complex datasets that describe these collections in much greater genetic detail than has been available before now. Such information will enable us to understand the structure of crop plant species and therefore help us to develop strategies for the development of new varieties. Here, we have presented recent research in the analysis of such datasets, describing how marker scores may be predicted, evolutionary distances be estimated, and collection structures and core collections be determined. These approaches are essentially first efforts at understanding these datasets. In addition to the approaches touched upon here, other techniques may also be of considerable value. For example, a pilot study on the use of data mining algorithms (C4.5 and simulated annealing) for rule-based classification of trait-allele associations, in particular the association of marker scores with disease status, has been promising [9]. For all our methods, we need to evaluate formally their efficiency and utility, possibly through simulation. Ultimately, we aim to develop more sophisticated methodologies that will allow us and others to exploit these datasets to their full potential.

Acknowledgements

RD was funded by a BBSRC Committee Studentship. GS was funded by a John Innes Foundation PhD Studentship. RJ was funded by a grant from the European Union. THNE, MA, AF, IN and JD acknowledge support from the Biotechnology and Biological Sciences Research Council and the European Union Framework 6 project Grain Legumes (FOOD-CT-2004-506223).

References

1. Flavell AJ, Knox MR, Pearce SR and Ellis THN (1998) Retrotransposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. **Plant J.** 16: 643-650.
2. Flavell AJ, Bolshakov VN, Booth A, Jing R, Russel J, Ellis THN and Isaac P (2003) A microarray-based high throughput molecular marker genotyping method – The Tagged Microarray marker (TAM) approach. **Nucleic Acids Res**. 31: e115.
3. Huber W, von Heydebreck A, Sultmann H, Poustka A and Vingron M (2002) Variance stabilisation applied to microarray data calibration and to the quantification of differential expression. **Bioinformatics** 18(1): S96-S104.
4. Gentleman RC, Carey VJ and Bates DM (2004) BioConductor: Open software development for computational biology and bioinformatics. **Genome Biology** 5: R80
5. Bryant D and Moulton V (2002) NeighborNet: an agglomerative method for the construction of planar phylogenetic networks. **Molecular Biology and Evolution** 21:255-265.
6. Huson DH and D Bryant D (2006) Application of Phylogenetic Networks in Evolutionary Studies. **Molecular Biology and Evolution** 23(2): 254-267.
7. Saitou N and Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees **Molecular Biology and Evolution** 4: 406-425.
8. Steel M (2005) Phylogenetic diversity and the greedy algorithm. **Systematic Biology** 54(4): 527-529.
9. Davenport G, Ellis THN, Ambrose M and Dicks J. (2004) Using bioinformatics to analyse germplasm collections. **Euphytica** 137(1): 39-54.