Molecular population genetics and agronomic alleles in seed banks: searching for a needle in a haystack?

Dario Prada*

Seed Conservation Department, Royal Botanic Gardens Kew, Wakehurst Place, West Sussex RH17 6TN, UK

Received 14 January 2009; Revised 31 March 2009; Accepted 31 March 2009

Abstract

Seed banking has been the single most significant reaction of the research community to the alarming rates of plant genetic erosion occurring in the wild. One enduring challenge for a wiser utilization of the resources enclosed in seed banks, however, has been the estimation of their genetic potentials for agriculture’s benefit. Key to detecting in landraces and/or wild relatives of modern crops any allelic variant lost during domestication and crop improvement is the use of molecular information to determine structure, evolution, and function of the genes harbouring these alleles. This paper reviews some of the theoretical and statistical issues surrounding the use of molecular population genetics tools for the detection of agronomical valuable alleles in seed banks. Emphasis is made on the technical limitations imposed by seed banking that may lessen the success of integrated and multi-disciplinary molecular approaches. The influence that population stratification and linkage disequilibrium exert on specific experimental designs for a better understanding of the evolutionary history of potential agronomic-related genes is also examined.

Key words: Association mapping, linkage disequilibrium, plant genetic resources, plant molecular genetics, population stratification, seed banking, selective genes.

Seed banking and its impact on modern agriculture

Plant improvement relies on selective propagation of genotypes containing favourable combinations of alleles of genes controlling desirable agronomic traits. Over the last 10,000 years plant domestication produced numerous landrace populations that served as the founder material for further genetic improvement through more recent selective breeding in the last 150 years. Both early domestication and later crop improvement have caused several genetic bottlenecks presumably reducing the levels of genetic diversity in modern crops. In fact, most of the contemporary crop varieties descend from a relatively small number of founder landraces. It follows that those genes underlying agronomic traits in modern cultivated genotypes retain decreased levels of diversity compared with the entire gene pool of landrace populations and the closest wild relatives from which they derive ( Tanksley and McCouch, 1997). For example, in the case of grasses, the world’s most important crop family, cultivated varieties maintain, on average, 70% of the diversity found in their closest wild progenitors ( Buckler et al., 2001).

Nowadays there is a long-standing concern in the international community about the disappearance of genetic variants ( Esquinas-Alcazar, 2005). The historic narrowing of the genetic basis for enhanced agronomic performance might make, for instance, modern varieties more susceptible to newly emerging diseases ( Harlan, 1975). Global climate change over the last ~30 years has produced directional shifts in the distribution and abundance of wild plant communities, representing a major cause of widespread reduction of the biological diversity ( Parmesan and Yohe, 2003). In reaction to these threats of genetic erosion, the global strategy adopted by the research community has been to increase efforts to warehouse wild plant species in seed banks ( Schoen and Brown, 2001). Seed banking prolongs seed viability ex situ under cold and dry conditions and thus safeguards plants for future use. The collections...
held in seed banks are usually formed by gene pools of individual crops at the interspecies level, that is, cultivated crops, their founder ancestors as well as the closest-relative species that still survive in the wild. In 1997, for instance, FAO reported that the seed bank collections maintained by the Consultative Group on International Agricultural Research (CGIAR) consisted of 59% landraces and old cultivars, 27% modern and historical breeders’ cultivars, and 14% wild relatives (FAO, 1997). In the case of major crops held in seed banks, there are many examples of wild relatives organized in the form of regional accessions, such as wild barley (*Hordeum spontaneum*), wild maize (five recognized species of teosinte for the genus *Zea*), wild wheat (*Triticum urartu*, *T. boeoticum*, and *T. dicoccoides*), wild rice (*Oryza barthii*, *O. glumaepatula*, *O. meridionalis*, *O. nivara*, and *O. rufipogon*), or wild rye (*Secale vavilioid*). There can also be found in seed banks some collections derived from crosses between taxonomically-related crops at the interspecies level; one of the most significant examples of this type of germplasm is the ‘Veery wheat’ lines which were obtained after the introgression of an entire chromosomal segment of rye into wheat (Rajaram et al., 1983; Hoisington et al., 1999). As a whole, the seed bank collections may quantitatively represent only a small fraction of the global biodiversity, but they are undoubtedly one of the world’s richest stocks of plant genetic diversity and offer a source of alleles for future genetic improvement of crops.

Despite the preponderant role of seed banks in sheltering novel alleles with potential use for crop enhancement, the reality is that seed bank accessions still have a very limited impact on current crops (FAO, 1997). Some technical restrictions associated with seed banking may have lessened the use of wild plant genetic resources into agriculture. First, since seed bank collections are stored outside the natural habitat, typically in such a way as to minimize genetic change, they show a lack of evolutionary response to changing environmental conditions with its resulting low adaptation rate. Second, because of practical (small) collection sizes, founder effects may possibly induce changes in frequency distribution of the alleles present in seed bank accessions—samples are usually obtained from fewer than 100 individuals in the wild. Third, seed viability declines during *ex situ* storage, hence the collections require regeneration in order to replenish stocks; however, regeneration strategies too often contribute to a shift in the genetic composition (genetic drift) of the accessions. Fourth, the ability to implement and design strategies for the identification and isolation of useful novel genes in wild donors has been proven unsuccessful in many ways. In this context, the development of efficient strategies that can facilitate the active incorporation of wild genetic resources into agricultural systems still remains an active area of research.

**Strategies for the isolation of agronomical desired genes in seed banks**

Traditionally, the approach for the utilization of the genetic material maintained in seed banks has been to screen accessions for a phenotypic appearance. Individual wild progenitors or landraces having the desired phenotype are repeatedly backcrossed with an elite genotype eventually to introduce the beneficial wild allele into the cultivated genetic background. Remarkable examples of recurrent backcross have been primarily produced for *Lycopersicon*. Within this genus, some wild species can be effectively crossed with the cultivated tomato *L. esculentum* and have been successfully used as donors of fungus- and insect-resistant genes (*L. hirsutum* and *L. peruvianum*), genes for fruit quality improvement (*L. chmielewskii*), and genes for adaptation to adverse environments (*L. cheesmaniae*) (Esquinas-Alcazar, 2005). Crosses between cultivated forms of sugar beet and its wild relative species have also been made to enhance disease-resistance and broaden the genetic base of the crop in general (Fig. 1; sugar beet). The repeated backcross strategy looks first at phenotypic observations to characterize the underlying genetic architecture: it starts the analysis at high levels of biological organization (phenotype) and descends to lower levels in order to gain knowledge on how genetic variation is arrayed within crops and relatives (the so-called ‘top-down’ approach). However, this approach has serious limitations because it is only applicable to genes easily observable in the phenotype which are mostly controlled by one, or a few, set(s) of genes.

In contrast to ‘top-down’ genetics, alternative approaches that begin the analyses at the genomic sequence level and then work back up to the phenotype (so-called ‘bottom-up’ strategies) have emerged recently with real promise for a more comprehensive understanding of plant genetic variation. Tanksley and McCouch (1997), in a seminal paper, stressed the poor prediction capacity of the phenotype and pointed out the genetic composition at molecular level as the best indicator of genetic potentials in seed bank accessions. The approach they suggested centred on linkage mapping analysis over populations derived from crosses between a crop variety and one of its wild relatives: once a target quantitative trait locus (QTL) is identified in the mapping population, advanced molecular marker-assisted backcrossing facilitates the isolation of the causal wild beneficial allele within a homogeneous elite genetic background (Fig. 1; tomato). This method may theoretically allow more accurate resolution of any specific QTL, in principle, from typical 10 cM QTL intervals into 1 cM intervals, identifying targets for further positional cloning. Nevertheless, the progress in isolating, cloning, and characterizing novel genes following this strategy has been restricted to a small number of noteworthy cases, such as *th1* in maize (Doebley et al., 1995), *fw2.2* in tomato (Frary et al., 2000), and *Hdl* in rice (Yano et al., 2000). What has prevented this pioneering approach from fulfilling its initial expectations? The lack of more positive results could be attributed to different facts. On the one hand, the modest degree of recombination in practical population sizes may have limited the power of the statistical tests for QTL detection and, on the other hand, since only two alleles are sampled in a biparental population, there is a clear underrepresentation of the putative pool of allelic variants at
a locus. Moreover, even if statistically significant evidence of linkage is obtained, extensive positional cloning should still be required to progress from a broad linkage region containing millions of bases to the causal gene(s) within the genomic region.

In the hope of overcoming some of the limitations of linkage analysis, the plant research community has only recently started to exploit the natural genetic diversity of germplasm collections as an additional means to identify marker-trait associations. This type of (association) analysis is performed across highly diverse sets of genotypes which contain many more historical recombinational events than the biparental populations used in linkage analysis, allowing much higher mapping resolution. The target germplasm collections for association tests often include elite and historic commercial cultivars, as well as landraces and the wild relatives of crops (Zhu et al., 2008). Different accessions of teosinte (Zea mays ssp. parviglumis), the wild progenitor of maize, have been used, for instance, in association mapping panels for the study of major regulatory genes controlling plant growth and development (Fig. 1; maize). One reason for optimism regarding ‘bottom-up’ approaches used in agricultural research (also called ‘reverse’ genetics). In the sections that follow, some of the theoretical, statistical, and practical issues surrounding the efficient deployment of molecular population genetics tools for the detection of agronomically valuable alleles in seed banks are presented. The stepwise integration of these techniques may signal the advent of a new and promising era for a better understanding and use of the genetic variation enclosed in seed banks. Figure 2 summarizes possible interconnections among different molecular genetics applications in order to gain better genetic knowledge of the novel and (valuable) functional variation present in seed banks. This general methodology, based on combined tests of selection and association mapping, is being used in humans for the isolation and characterization of specific genomic sequences; in plant research it has also been applied but with various levels of integration and resolution.

Cultivated genotypes are the direct result of the accumulation of beneficial alleles at key genes controlling traits of agronomic interest, but considering the high number of loci expected to participate in the phenotypic expression of such traits, it is very unlikely that modern genotypes retain beneficial allelic variants at all of the agronomically related sites ( Tanksley and McCouch, 1997). It is thought that sets of interesting classes of alleles, in many cases with minor effects, could have been missed during domestication and/or crop improvement. The evolutionary history of agronomically related genes has been shaped primarily by human-mediated selection events, either unintentionally through domestication or intentionally through more recent crop improvement activities. The resulting genetic bottlenecks

Molecular population genetics and plant genetic potentials in seed banks

Plant researchers are beginning to benefit from some of the specific areas of research falling under the molecular population genetics framework, with evolutionary biology, association mapping, and comparative genomics being three of its most representative examples. These methods embody ‘bottom-up’ approaches used in agricultural research (also called ‘reverse’ genetics). In the sections that follow, some of the theoretical, statistical, and practical issues surrounding the efficient deployment of molecular population genetics tools for the detection of agronomically valuable alleles in seed banks are presented. The stepwise integration of these techniques may signal the advent of a new and promising era for a better understanding and use of the genetic variation enclosed in seed banks. Figure 2 summarizes possible interconnections among different molecular genetics applications in order to gain better genetic knowledge of the novel and (valuable) functional variation present in seed banks. This general methodology, based on combined tests of selection and association mapping, is being used in humans for the isolation and characterization of specific genomic sequences; in plant research it has also been applied but with various levels of integration and resolution.

Cultivated genotypes are the direct result of the accumulation of beneficial alleles at key genes controlling traits of agronomic interest, but considering the high number of loci expected to participate in the phenotypic expression of such traits, it is very unlikely that modern genotypes retain beneficial allelic variants at all of the agronomically related sites (Tanksley and McCouch, 1997). It is thought that sets of interesting classes of alleles, in many cases with minor effects, could have been missed during domestication and/or crop improvement. The evolutionary history of agronomically related genes has been shaped primarily by human-mediated selection events, either unintentionally through domestication or intentionally through more recent crop improvement activities. The resulting genetic bottlenecks
created in this course have very likely altered the allelic frequencies at loci with agronomical significance, as well as inducing overall genetic variation to decrease and linkage disequilibrium (LD) to increase (Fig. 2A). LD refers to the non-zero correlation between alleles at different loci, even unlinked, and it is inversely proportional to levels of allelic recombination (Flint-Garcia et al., 2003). The extent of LD is expected to vary within individual genetic pools forming the seed bank accessions; LD seems to decrease gradually for modern genotypes, landraces, and wild relatives. Such differences are mainly explained by the diverse mating histories of each genetic pool. The decay of LD across different germplasm pools has been extensively characterized for major annual crops (Gupta et al., 2005). For example, for modern cultivated varieties of barley, complete LD has been observed across contiguous genomic sequences of up to 212 kb in length (Piffanelli et al., 2004) while in landraces LD decays over 90 kb and in wild barley LD does not extend beyond a single genic region (Caldwell et al., 2006). Analogous patterns of LD decay have also been observed in maize: it extends 100 kb for commercial elite inbred lines but declines to levels of 1.5 kb for maize landraces (Yu and Buckler, 2006). There are only a few studies that have estimated the extent of LD in perennials, in part, because these species have more complex life histories and mating systems than annual crops. Examples can be found for woody plants—European aspen (Populus tremula) (Ingvarsson, 2005), Scots pine (Pinus sylvestris) (Dvornyk et al., 2002), Douglas fir (Pseudostuga menziensii) (Krutovsky and Neale, 2005), and loblolly pine (Pinus taeda) (Brown et al., 2004)—as well as for grapevine (Vitis vinifera) (Barraud et al., 2006) and perennial ryegrass (Lolium perenne) (Ponting et al., 2007).

The onset of ‘agriculture’ occurred in the wake of unconscious selection made by humans, leading to the fixation of relevant alleles in a reduced set of key traits with drastic impact on the phenotype, such as seed size, ear rachis stiffness, and the ease with which the seed was released from its enclosing leaf-like structures. Afterwards, there was a long phase of human-mediated selection over primitive wild plant forms that allowed the expansion of agriculture into new environments, in this case mainly by modifying different, but again relatively small, sets of traits with polygenic inheritance (i.e. seed weight, seed dormancy) which resulted in landrace populations adapted to local conditions (Salamini et al., 2005). Common patterns of selection seem to have been applied to a limited group of traits/genes throughout more recent breeding activities. Interestingly, for the most important crop families, target loci for plant improvement show high levels of homology across species. For example, in the case of cereals, the homologue sequences of the genes RhtD1a of wheat and
Gai of rice, both affecting plant height and flowering time, had impressive effects on yield during the ‘Green Revolution’ in the 1960s and 1970s (Peng et al., 1999). Flowering time has had a central role in selection for local adaptation across several major crops; landrace populations of cereals that flowered in short days have been transformed during selection into crops in which flowering time is unaffected by day length (Putterill et al., 2004). This is the case for barley, in which the gene Ppd-H1 that controls flowering time had a strong influence in the domestication and further Neolithic spread of cultivation due to human-mediated selection of non-responsive ppd-H1 phenotypes (Jones et al., 2008). This convergent selection in cereals may have been centred on a set of genes homologous to the PHY family of Arabidopsis thaliana (Sawers et al., 2005).

Modern crops constitute good models to evaluate imprints of human-mediated selection at specific genomic sites, frequently sharing homology across species, especially for the traits where domestication and/or breeding have acted more strongly (Wright and Gaut, 2005). The reduction in allelic diversity and enhancement of LD have been extensively used in plant molecular population genetics as primary signals of selection on random sequence data (Fig. 2B). Wright et al. (2005) suggest that if genomic sites showing clear evidence of selection are identified in modern crops they may provide a substrate for the amplification of homologous alleles in sets of landraces and wild relatives. These sequences ultimately represent allelic forms missed in modern crops because of selection at different time scales, but that still exist in the wild as candidates for novel variation in agronomically related genes. Primer design, however, can be very difficult for the very heterogeneous germplasm kept in seed banks because priming in more conserved regions may amplify paralogous regions while priming in less conserved regions may fail if there is extensive polymorphism in the primer region across the germplasm (Flint-Garcia et al., 2005). In spite of these well-known technical drawbacks, amplification of homologous sequences across different genetic pools at interspecies level has proved to be successful on the maize genome. Common primer sequences that are maintained for modern corn varieties with respect to teosinte, its wild common ancestor, have been amplified in the wild ancestor background and have been subsequently flagged for introgression into the elite germplasm pool of maize (Wright et al., 2005).

The influence of selective genes on phenotype varies tremendously. As a result, not all selective alleles should necessarily have a real impact over agronomical traits. In order to avoid novel but non-functional (non-valuable) genetic variation being reintroduced into modern crops, the effect of candidate loci on the phenotype needs to be investigated. At present, association analysis is the most powerful method in molecular population genetics to establish the link between genotype and phenotype in highly diverse panels (Fig. 2C). If candidate sequences for introgression are eventually identified and validated, any molecular marker-assisted backcross strategy can then facilitate the reintroduction process by monitoring the presence of target primer sequences as backcross generations advance (Fig. 2D). Crosses between cultivated genotypes and their wild relatives within the same genus, thus overcoming the sexual barriers for inter-specific crossing, have been consistently produced for several major crops like tomato, sugar beet, maize, and barley. Molecular marker-assisted backcross schemes are routinely used in modern agricultural research; a classical example is Marker-Assisted Recurrent Selection (MARS) which refers to the improvement of an F2 population by combining several cycles of phenotypic and marker selection (Johnson, 2004).

All molecular genetics techniques described in Fig. 2 have been effectively applied in plant research but with different levels of integration. Maize and tomato are the crops on which most research has been devoted in recent years. There are two central aspects yet to consider when linking up any molecular genetics technique through multi-disciplinary approaches, the existence of stratification due to population admixture of the sample and the degree of LD (population-specific) between alleles. In the next sections, the implications that these two phenomena may have in the detection of signals of selection, and in functional validation, are examined.

**Identification of alleles with imprints of selection in seed banks**

Two types of selection have been targeted in crop evolution, i.e. either directional or balancing selection. Directional selected genes are normally genes with high influence on the phenotype and, in many cases, have been easily selected by obvious morphological and developmental phenotypic observations. This is the case for the tb1 gene in maize that governs lateral branching (Doebly et al., 1997) and the RhtD1a gene in wheat that controls plant height and flowering time (Peng et al., 1999). For direct deployment in breeding programmes, directional-selected genes that have undergone the most stringent selection (and thereby the greatest reduction in diversity) have little remaining genetic variation and cannot easily be further improved by breeding. In contrast, genes bearing signatures of balancing selection may be of greater importance because valuable allelic variants having small impact on the phenotype could still reside in the wild. Directional selection, also referred to as selective sweep, favours one allele over others and can lead to the fixation of the favoured variant in the entire population. Balancing selection implies the long-term selective maintenance of multiple allelic variants at intermediate frequencies not resulting in allele fixation. The simplest model to explain most evolutionary change in the wild through selection was proposed by Kimura (1968). His neutral equilibrium theory (NE) suggests that plant adaptation is the result of genetic drift acting on selectively-equivalent (neutral) mutant alleles. NE serves then as the null hypothesis to evaluate imprints of selection through modification of allelic frequencies from equilibrium expectations. Following Kimura’s work, many statistical tests have been developed to screen sequence data for differences
in allelic distributions relative to NE expectations. The most popular statistical indicator is the mean value of Tajima’s D statistic (Tajima, 1989) which compares the number of nucleotide polymorphisms with the mean pairwise difference between sequences. Other alternative statistics relying upon the same principle are Fay and Wu’s H statistic (Fay and Wu, 2000) that evaluates the number of derived nucleotide variants at low and high frequencies with the number of variants at intermediate frequencies; and Fu and Li’s D, D*, F, and F* statistics (Fu and Li, 1993) that compare the number of derived nucleotide variants observed only once in a sample with either the total number of derived nucleotide variants (D and D*) or the mean pairwise difference between sequences (F and F*). In addition to allele distribution, signatures of selection can also be evaluated by comparisons of the rates of divergence between different classes of mutations because selection causes a reduction in levels of nucleotide diversity. The Hudson–Kreitman–Aguade’s HKA statistic (Hudson et al., 1987) is the most popular example of this methodology. HKA compares the degree of polymorphism within and between species at two or more loci.

One of the difficulties in applying studies for the detection of selection at specific genomic sites is the estimation of how demographic history (population structure) affects genetic variation in the entire genome. The inference of population demographic history (population structure) affects genetic of selection at specific genomic sites is the estimation of how rapidly. HKA compares the degree of polymorphism within and between species at two or more loci. Selective genes for major crops have been found in maize (Yamasaki et al., 2005), Arabidopsis (Nordborg et al., 2005), barley (Morrell et al., 2003), soybean (Hyten et al., 2006), and sorghum (Hamblin et al., 2006). Notwithstanding these remarkable examples, much still remains to be learned about selection. One of the most important aspects to be considered is the sampling strategy of the genes for scrutiny. In fact, many genes are studied because they are hypothesized a priori to be under selection. To avoid overestimation of discovery rates due to sampling biases, random genome-wide surveys become critical. This type of analysis computes statistical tests for sets of genes evenly distributed across the genome thereby lessening false discovery rates. These statistical methods leave unresolved the question of multiple testing so they are still in development for humans (McVean and Spencer, 2006) and effective progress in plants has been restricted to maize (Wright et al., 2005).

Validation of functional polymorphisms in heterogeneous genetic backgrounds

Allelic association refers to the relationship between a phenotypic trait and the genotype at a locus. There is a variety of statistical methods for association mapping routinely exploited to map genes of complex diseases in humans (Risch, 2000), which are now largely applied in plants (Mackay and Powell, 2007). In association mapping, LD can be the result not only of (physical) linkage but also of population admixture, genetic drift, and selection. The resolution power of association mapping ultimately depends on the structure of LD across the genome as well as how rapidly LD decays with physical distance (Hirschhorn and Daly, 2005). Association mapping is difficult in structured populations, leading to spurious results if this feature is not taken into consideration in the statistical tests. When the functional variants are unequally distributed among different subgroups for the trait under study the association analysis leads to false evidence for allelic association (Knowler et al., 1988). Several facts are likely to create high levels of population structure in very diverse panels of individuals maintained in seed banks. First, seed bank collections are more often than not organized in the form of regional accessions, in some cases the accessions are sampled in a single field-collecting trip thus accentuating even more stratification effects. Second, plant populations still inhabiting the wild have often had a limited gene flow, which makes them more susceptible to population differentiation (Sharbel et al., 2000). Third, population structure can become highly trait-dependent for wild forms and landraces, especially for traits playing a pivotal role in local plant adaptation, such as seed dormancy.

The first methodologies implemented for marker-trait associations relied upon comparisons of trait mean shifts for the different allelic states at a single locus by using classical forms of t tests (parametric or non-parametric), Pearson tests or Fisher exact tests (Balding, 2006). Alternative methods to control for population admixture search for
evidences of background structure and account for it directly into the association statistic test. Genomic Control and Structured Association, both extensively used in animal and plant systems, are examples of these methods (Mackay and Powell, 2007). For Genomic Control analysis a set of random markers is used to assess the bias of the statistical tests explained by population structure (Devlin and Roeder, 1999). The general strategy of Structured Association is first to classify individuals into subpopulations according to the evolutionary history of a large number of independent genetic markers across the genome, and later to perform marker-trait association tests within the established subgroups. Subpopulation membership has been largely explored by a popular Bayesian-based model developed by Pritchard et al. (2000a, b) although less computationally demanding models, i.e. genetic distance-based methods like principal component or cluster analysis, can also be considered for this type of analysis (Zhao et al., 2007). Advanced statistical methods, that incorporate pedigree relationships and population structure at the same time in the models, have recently emerged in plant research. The mixed-model framework offers a high degree of flexibility for this purpose because it can account for multiple levels of relatedness by using a genotypic relationship matrix to structure the variance–covariance matrix between individuals. Studies in maize and potato have confirmed the value of this approach, showing improved control of false positives compared to classical forms of t tests performed within prior identified subgroups (Yu et al., 2006; Malosetti et al., 2007). Achievements in LD mapping for plants, including specific examples of both candidate-gene testing and genome-wide surveys, can be found in the comprehensive reviews of Flint-Garcia et al. (2003), Gupta et al. (2005), and Zhu et al. (2008). Most of these studies have been performed with highly diverse collections of annual crops, but, recently, several cases with positive marker-trait associations for perennial species have also been published, such as loblolly pine (Pinus taeda) (Gonzalez-Martinez et al., 2007), grapevine (Vitis vinifera) (This et al., 2007), eucalyptus (E. nitens) (Thumma et al., 2005), and perennial ryegrass (Skøt et al., 2007).

**Evaluation of phenotypes across different germplam pools**

Seed banks normally possess passport data that include taxonomy, life history, ethnobotanical knowledge or eco-geographic patterns of the collecting sites for the seed accessions that they maintain. This basic information serves for primary characterization and classification of the collections. Nevertheless, the phenotypic evaluation of the seed bank entries for potentially valuable agronomic traits results very daunting due to the actual sizes of the whole collections, more often than not reaching tens of thousands of entries. In the interest of cost-effective characterization of the plant genetic resources held in seed banks, Frankel (1984) proposed the development of core collections; these are subsets of the whole collection chosen as representing most of the genetic diversity found in the collection sample. The phenotypic screening is initially restricted to the core collection, and if desirable phenotypes are found, then only those accessions of the whole collection sharing similar characteristics to the flagged individuals of the core subset (i.e. common ecogeographic origin, genetic resemblance) are evaluated.

Phenotypic testing of core collections largely responds, apart from the economical obstacles or space limitations (FAO, 1997), to the level of genetic complexity of the trait of interest. Simple phenotypes can be directly scored over seed lots, with no need to grow plants in the field (Fig. 3). Direct measurements in kernels generally concern strongly
heritable characters which are largely independent of the environment. These phenotypes usually result from the accumulation of specific metabolites controlled by single inherited genes, like those responsible for biosynthetic enzymes (Doebley et al., 2006). An example of seed attributes publicly available for applied research is the seed information database of the Royal Botanic Gardens Kew (http://www.kew.org/data/sid). The impact that this class of phenotypes has on modern agriculture is, however, very limited, and has been restricted to several traits with explicit use to humans, such as new sources of medicines or nutrition.

More complex phenotypes in core collections must be measured in plant populations grown under specific field experiments, or even indoor pots (Fig. 3). The extent to which core collections are planted and characterized for traits with agronomic importance is widely variable, and mainly relates to the specific focus of each seed bank (FAO, 1997). Van Hintum et al. (2000) provide a comprehensive list of traits for which some core collections have been screened in the field, including, among others, various disease-resistance traits and abiotic stress tolerance. But when core collections are tested in the field it is important to consider that, for a target species, different populations sampled in different habitats often exhibit local (ecotypic) adaptation to site conditions (Schoen and Brown, 2001). Common garden experiments are the classical designs for the analysis of ecotypic adaptation. In this type of experiment seed lots of two (or more) populations of the same species, but having different geographical origins, are planted in a common environment to allow the distinction of heredity from local adaptation. Traditionally, these designs have been deployed for the evaluation of plant growth and plant architecture in populations resulting from the natural hybridization of crops and their wild relatives (Jarvis and Hodgkin, 1999).

A series of trials containing sets of genotypes tested across different years and locations is the method of choice for the evaluation of those traits with high degrees of genetic complexity, normally subjected to strong genotype-by-environment interactions (Fig. 3). These networks of experiments have been historically managed by public or private breeding programmes, requiring advanced field designs and statistical methods to gain a deeper insight into the genetic bases of the traits under study. The assessment of phenotypic adaptation in multi-environment traits has often relied upon empirical methods (e.g. yield per se), based on the differential genotypic responses to environmental changes. Nevertheless, as new and more refined methods for linkage and association mapping are emerging, the analysis of genotype-by-environment interactions is being moved towards the dissection of QTL-by-environment interactions. Furthermore, advanced statistical models have also been used as an aid to the introduction of relevant environmental (climatic/edaphic) factors into statistical linkage mapping models for in-depth analysis of genomic regions that show an environmental-dependent contribution to the phenotypes (Yin et al., 2004).

**Rare alleles in seed banks: can they sensibly impact on agrononomical traits?**

Alleles at low frequencies in seed bank accessions may represent, if identified in the original pool, interesting variants conferring local or wide adaptation for crop improvement. Two questions must be properly addressed before considering the contribution that rare alleles can have for crop improvement. First, are rare alleles really represented in the germplam collection? To respond to this question many studies have focused on practical considerations of seed collection strategies (Way, 2003). In fact, the sampling strategies seek to balance the risk of failing to collect rare alleles against the daunting challenge of collecting very large sample sizes. Second, can the functional variation of rare alleles be efficiently identified with the resolution exhibited by the association mapping approaches? This feature must be seen not only in terms of allele frequencies but also in the proportion of individuals that, despite having the allele in question, do not express its phenotype.

The scope of the seed collecting strategies is to maximize the genetic diversity sampled in the wild. Since the genetic variation present among and within target populations is largely unknown in advance of sampling, the general guidelines for collecting usually rely on the analysis of theoretical models for genetic variation in the population sample. From theoretical breakthroughs, several population genetic models based on molecular marker information have been developed to assist in determining minimal sample requirements (Crosso, 1989; Schoen and Brown, 2001). Most of these models are built around the infinite, selectively neutral allele model of Kimura, assuming Hardy–Weinberg equilibrium and, when possible, incorporating breeding system and population distribution (Brown and Briggs, 1991). Brown (1989), for instance, using the sampling theory for selectively neutral alleles showed that the number of alleles captured in a sample was approximately proportional to the natural logarithm of its size. To estimate the cost-effectiveness of sampling rare alleles, Marshall and Brown (1975) classified the allelic variants present in wild populations into four classes on the basis of their frequencies and geographic distributions. (i) Common-widespread alleles: they are almost certainly included even in small samples collected from only a few populations. (ii) Rare-widespread alleles: the target populations containing this type of alleles behave as a single, large and unstructured population. (iii) Common-localized alleles: they occur in only one or a few habitats reaching a high frequency in each of them. (iv) Rare-localized alleles: the inclusion of an allele of this class will be unusual and serendipitous, even in very large samples taken from a large number of populations.

To evaluate the sensitivity of the association mapping methods to detect the effects of rare variants, both allele class frequency and allelic effect size (allelic penetrance), should be considered in a joint analysis. The two phenomena affect the statistical power of any association mapping approach although they are inversely related (Morton,
The identification of potentially valuable agronomic alleles in seed banks in studies with wild forms and landraces maintained in seed collections that have to be made for good experimental designs many small, but crucially important, practical and statistical aspects. There are consequently age disequilibrium, sample size, allelic penetrance, and allele frequency distribution (Table 1). There are consequently age disequilibrium, sample size, allelic penetrance, and allele frequency distribution (Table 1). There are consequently age disequilibrium, sample size, allelic penetrance, and allele frequency distribution (Table 1). There are consequently age disequilibrium, sample size, allelic penetrance, and allele frequency distribution (Table 1).

Conclusions and future prospects

Major genetic divergences between ‘wild’ and ‘cultivated’ genetic pools are largely explained by human-mediated selection through domestication, founding events, and breeding practices. Theses processes have presumably created different genetic bottlenecks which have resulted in decreasing rates of genetic diversity, changes in allele frequencies, increases in LD, and reduction of rare alleles in modern crops (Halliburton, 2004). Therefore any discovery initiative for the characterization of agronomically relevant alleles in seed banks should be much affected by analogous features, such as population stratification, linkage disequilibrium, sample size, allelic penetrance, and allele frequency distribution (Table 1). There are consequently many small, but crucially important, practical and statistical choices that have to be made for good experimental designs in studies with wild forms and landraces maintained in seed banks. Plant molecular population genetics seems to possess many of the scientific and technological potentials required for such experimental designs. This area of research has undergone drastic innovations in the last 10 years due to the steady deposition of informative single-nucleotide polymorphisms (SNP) into large panels, partly because of the rapid decrease of genotyping costs, and the ongoing improvements in algorithms for sequence data analyses (i.e. more refined methods to deal with population structure and better characterization of background LD patterns). The deployment of these techniques with increasing rates of interconnection holds real promise for a better understanding on how genetic variation is arrayed in modern crops and wild forms. Far too often genetic research for these two plant genetic backgrounds has been undertaken separately by the scientific community.

Acknowledgements

The author thanks Dr JPA Heuts (Eindhoven University of Technology) and two anonymous reviewers for very valuable comments.

References


---

**Table 1.** Effects that linkage disequilibrium and population structure may have on activities associated with seed banking and identification of potentially valuable agronomic alleles in seed banks

<table>
<thead>
<tr>
<th>Molecular marker-based activity</th>
<th>Linkage disequilibrium (LD): correlation among alleles</th>
<th>Population stratification (PS): over/underestimation of allele frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collecting strategies maximizing molecular marker diversity</td>
<td>High LD between target markers and causative (adaptive) genes may help to capture the phenotypic variation present in the wild</td>
<td>High PS determined in the population sample through marker-based estimations (i.e. $F_{st}$) must be contrasted with phenotypic differentiation due to geographic subpopulations</td>
</tr>
<tr>
<td>Identification of selective genes in seed bank collections</td>
<td>High LD between target markers and selective genes may create hitchhiking effects in tests of selection</td>
<td>High PS in the germplasm collection may confound low allele frequencies with signals of positive selection, thus enhancing false positives in tests of selection</td>
</tr>
<tr>
<td>Association mapping across highly diverse collections of germplasm</td>
<td>Low LD in genic regions facilitates candidate-gene testing</td>
<td>High PS in the germplasm collection requires association models that correct for it, showing then a better control of false positive discovery rates</td>
</tr>
</tbody>
</table>

1998). Major (Mendelian) genes having a large impact on the phenotype are mostly present at very low frequencies in wild populations, while polygenes with small effects on the phenotype account mostly for all alleles underlying the expression of a complex trait. In humans, several models have been implemented to parameterize the combined effect of allelic frequency and size at a single locus, either by correcting these coupled effects in a joint pooled statistic (Zondervan and Cardon, 2004) or by designing enriched mating schemes to increase population size and relative allelic contribution at the same time (Antoniou and Easton, 2003). In plant research, however, these strategies have deserved little attention, in part because most agronomic traits show no clear segregation patterns comparable to those found for Mendelian disorders in humans (Hirschhorn and Daly, 2005). Risch (2000) indicated that highly penetrant alleles with intermediate frequencies are realistically expected to be detected in association mapping, whereas alleles present at the same frequencies but with a more modest contribution to phenotype are practically impossible to detect.


Savitsky H. 1960. Meiosis in an F1 hybrid between a Turkish wild beet (Beta vulgaris ssp. maritima) and Beta procumbens. Journal of the American Society for Sugar Beet Technology 11, 49–67.


