

Chapter 10:

Diversity Arrays Technology: A Novel Tool for Harnessing the Genetic Potential of Orphan Crops

Eric Huttner, Peter Wenzl, Mona Akbari, Vanessa Caig, Jason Carling, Cyril Cayla, Margaret Evers, Damian Jaccoud, Kaiman Peng, Sujin Patarapuwadol, Grzegorz Uszynski, Ling Xia, Shiyong Yang, and Andrzej Kilian

Introduction

Genetic diversity is the raw material available to plant breeders. By productively recombining genetic diversity, plant breeders have been successfully producing, year after year, improved cultivars of the major domesticated species used in the world's diverse agricultural systems. Molecular genetic markers offer a powerful tool to accelerate and refine this process. Existing genetic marker (genotyping) technologies, mostly developed for applications in human health, have also been applied successfully to agricultural species, but their cost remains prohibitive for most agricultural applications. This is particularly true for species for which no molecular data and very limited resources are available.

Because of the limitations of existing marker technologies, we have developed Diversity Arrays Technology (DArT), a novel method to discover and score genetic polymorphic markers. DArT is a sequence-independent, high-throughput method, able to discover hundreds of markers in a single experiment. DArT markers are typed in parallel, using high-throughput platforms, with a low cost per data point. With DArT, plant breeders, plant ecologists, as well as the managers of the germplasm collections, will be able to perform genetic analysis in a cost-effective and high-throughput manner. DArT fingerprints will be useful for accelerating plant breeding, and for the characterization and management of genetic diversity in domesticated species as well as in their wild relatives.

We have developed DArT successfully for rice, barley, wheat, and cassava. We have also produced a dedicated data management and analysis package, a key part of the technology, entirely built from Open Source components. Work is in progress to establish DArT for other species of importance to tropical agriculture: pigeonpea, sorghum, and chickpea. We have a strong interest in developing partnerships to establish DArT for many species, and we are developing a network model for the delivery of technology to users.

Whole Genome Profiling

The genetic diversity present within species is one of the components of biological systems. In many cases, a high level of diversity provides robustness to natural ecosystems and maximizes their opportunity for further diversification. Natural ecosystems are increasingly managed by humans with the objective of maintaining the existing genetic diversity. This diversity is considered an insurance against catastrophic damage and a resource for future human use. In agricultural ecosystems, management by the farmer is the key determinant of genetic diversity. It is well established that biological diversity contributes to the robustness and sustainability of agricultural

production systems, particularly in developing countries where societal support to farmers in time of crisis is limited or nonexistent (Conway, 1999).

A well-known example of the danger of limited diversity within a cultivated species is the 1970 epidemic of southern leaf blight in the Corn Belt of the USA. The quasi-universal adoption in the 1960s of hybrid maize cultivars produced using T cytoplasmic male sterility led to the loss of about 15% of the US maize crop in the early 1970s, because the cultivars were susceptible to the new race T of *Helminthosporium maydis*. Although the economic loss was large for the commercial farmers involved, a similar event in a subsistence agricultural system would have had more devastating consequences.

Measuring genetic diversity for many species is not an easy task. Molecular genetic markers, based on DNA sequence polymorphism, are increasingly used to complement phenotypic and protein-based markers. Over the past 20 years, DNA-based markers have been established in many species, mostly agricultural crops. Molecular markers linked to desirable traits have been used to accelerate plant breeding (Ribaut and Hoisington, 1998), for example by replacing phenotypic assays with single-marker assays when possible and cost effective (Bonnett *et al.*, 2004). Many traits of interest to plant breeders, however, are complex and polygenic. Therefore the creation of an adapted elite variety will increasingly involve the deliberate combination of various genomic regions from many different individuals (Peleman and van der Voort, 2003). Comprehensive knowledge of genetic diversity in the cultivated and wild germplasm—the source of novel genomic regions, novel alleles, and novel traits (Xiao *et al.*, 1998; Li *et al.*, 2003)—is very important. Applying molecular markers in this context requires moving from single marker assays to genome-wide marker profiles: genomic fingerprints covering genetic diversity at hundreds of loci. For genetic diversity analysis also, a reliable measure of the differences and the relatedness between individuals will require whole-genome profiling. We briefly review the marker systems suitable for whole-genome profiling. We then present our current work on the development of DArT, a novel marker system invented by one of us (A. Kilian), and particularly suitable for the analysis of genetic diversity in orphan crops and wild species (Jaccoud *et al.*, 2001).

Limitations of Existing Technologies

Current molecular marker technologies include RFLP, AFLP®, SSR (microsatellites), and SNP. All have at least one of the following limitations:

- The discovery of a sufficiently large number of polymorphic markers to achieve genome coverage is slow because it is a sequential process.
- Some marker systems are based on sequence information. For many plant species it may not be practical or economical to determine a large amount of genomic sequence, particularly if different alleles have to be identified by sequencing, for example when sources of new alleles are discovered.
- Once markers are identified, the cost of scoring the markers (“genotyping”) is high and the throughput low. For SSR and SNP markers an assay has to be developed first. This often results in only a fraction of the candidates identified being retained for routine use. Markers are then typically scored one by one (or at a modest level of multiplexing), usually using gel-based systems.

AFLP® is a proprietary, cost-effective method to discover and type polymorphic DNA sequences. The main limitations to its use in its current format are:

- Scoring is done by electrophoresis on gels (limited throughput).
- An allele is represented by the size of a band on the gel, which can be difficult to assess objectively.
- Typing of new varieties of a species can be done under the hypothesis that bands of identical size represent the same allele of the same locus, which is not always true.
- Cloning the polymorphic bands is a labor-intensive and sequential procedure.

Simple Sequence Repeat markers (Microsatellites; SSR) also have limitations:

- The marker discovery phase is expensive and involves DNA sequencing. A standard team of two persons in a well-supported environment is able to develop approximately 100 reliable SSR markers per year. Although availability of genome or EST sequence data would accelerate the discovery process, the development of a reliable subset of polymorphic SSR markers remains a laboratory-based empirical task.
- A high-resolution gel equipment system is required for genotyping. This still is an expensive piece of equipment to purchase and service. The throughput is limited by the reliance on electrophoresis and gels.
- Marker scoring usually requires one amplification reaction per marker; therefore the analysis is sequential not parallel. After extensive development work, some markers can be multiplexed up to 10 markers per reaction in the best case, thereby reducing the number of reactions required. However, the cost of the assay development to achieve this level of multiplexing is such that it is unlikely to be performed for species other than humans and a few major crops.
- The cost of producing an SSR data point is about US\$1, once the polymorphic SSR is discovered and the scoring protocol is well established. A genome-wide genotype of 500 established SSR markers for 100 plants would therefore cost about US\$50,000.

Single Nucleotide Polymorphisms are the preferred markers in human genotyping. Their application to other species, however, is limited by two important factors:

- Although high-throughput methods are being developed for scoring of SNP markers, most methods still require a marker-specific amplification reaction, or marker-specific primers, oligonucleotides, or probes.
- Although the cost per data point for well-formatted SNP markers is expected to decrease to approximately US\$0.20, the initial investment required for marker discovery (sequencing of allelic variants) and assay development remains prohibitive for many agricultural species, at least for the foreseeable future.

Diversity Arrays Technology

DArT was developed to provide a practical and cost-effective whole-genome fingerprinting tool. DArT has three key attributes of interest to plant breeders and scientists studying and managing genetic diversity: (a) it is independent from DNA sequence, (b) the genetic scope of analysis is defined by the user and easily expandable,

and (c) the method provides for high throughput and low-cost data production. The principle of DArT is presented in Figure 10. 1.

Sequence independence

The discovery of polymorphic DArT markers and their scoring in subsequent analysis does not require any DNA sequence data. This makes the method applicable to all species, regardless of how much DNA sequence information is available for that species. However, DArT markers are sequence-ready clones of genomic DNA.

Genetic scope

For each species, the method is developed on the “metagenome,” the pooled genomes from the germplasm of interest to the user. For example, the metagenome may include DNA from the cultivated varieties of a particular region or the lines used in a breeding program. Alternatively, the metagenome may cover the genetic diversity within the entire species and even extend to its wild relatives. Importantly, the diversity surveyed by DArT can be expanded if new individuals with marked genetic differences are incorporated into the analysis at a later stage.

High-throughput, low-cost data production

In DArT, several hundred polymorphic markers are identified in parallel. The efficiency of this marker discovery effort is only dependent on the level of genetic diversity within the species. For example, 5-10% of wheat and barley DArT clones and 25-30% of cassava DArT clones were polymorphic. The same platform is used for both discovery and scoring of markers, therefore no assay development, apart from consolidating all polymorphic markers into a single genotyping array, is required after the marker discovery. The microarray platform we currently use enables a high level of multiplexing: approximately 5000-8000 genomic loci are typically surveyed in parallel in single-reaction assays to discover polymorphic markers. For routine genotyping, several hundred markers are typed in parallel using only 50-100 ng of genomic DNA. We project that our data production service will soon deliver data for less than US\$0.10 per data point.

DArT markers

A DArT marker is a segment of genomic DNA, *the presence of which* is polymorphic in a defined *genomic representation* (see Figure 10.1). DArT markers are biallelic and behave in a dominant (present vs absent) or co-dominant (2 doses vs 1 dose vs absent) manner.

Identification of polymorphic DArT markers

To identify the polymorphic markers, a *complexity reduction method* is applied on the metagenome, a pool of genomes representing the germplasm of interest (Figure 10.1). The genomic representation obtained from this pool is then cloned and individual inserts are arrayed on a microarray resulting in a “discovery array.” Labeled genomic representations prepared from the individual genomes included in the pool are hybridized to the discovery array. Polymorphic clones (DArT markers) show variable hybridization signal intensities for different individuals. These clones are subsequently assembled into a “genotyping array” for routine genotyping.

Complexity reduction methods

Many complexity reduction methods can be used (Jaccoud *et al.*, 2001; Peng *et al.*, 2002). A suitable complexity reduction method produces genomic representations that are sufficiently large and contain a sufficient fraction of polymorphic clones to enable the production of a genotyping array containing several hundred markers. Our currently preferred method is based on digestion of genomic DNA with *Pst*I and a frequent cutter, followed by ligation of an adapter to the *Pst*I ends and amplification of *Pst*I fragments using a primer complementary to the adapter. This method was shown to work well in barley, a species with a 5000-Mbp genome (Wenzl *et al.*, 2004).

Genotyping by hybridization

For each individual DNA sample being typed, a genomic representation is prepared using a defined complexity reduction method. The representation is labeled and hybridized to a genotyping array, and a microarray printed with copies of the DArT markers. The hybridization signal for each marker is measured and converted into a score.

Technical platform

The platform we are currently using to discover and score polymorphic markers comprises a standard molecular biology laboratory, a microarray printer and scanner, and computer infrastructure to analyze, store, and manage the data produced. Platforms other than printed microarrays – for example color-encoded beads or self-assembling arrays - could be used for the routine typing of samples. These platforms offer good opportunities to reduce further the cost of routine genome profiling.

Software

We have written DArTsoft, a dedicated software for automatic data extraction, which is capable of producing up to 200,000 scores from discovery arrays in less than two hours (Cayla *et al.*, in preparation). With this sort of throughput, sample tracking and data management become essential. We are building DArTdb, a laboratory information management system for barcode-facilitated sample tracking, data storage, and data management (Uszynski *et al.*, in preparation). As a matter of principle we only use Open Source components for all our DArT-related software products.

Current Status of DArT Development

In the last 4 years, we have established the proof of concept of DArT (Jaccoud *et al.*, 2001), and we have developed the technology for a range of species. A list of species we are currently working on is given in Table 10.1, with the number of clones we have assayed from each species, to identify the best complexity reduction method for each species.

Mature stage

Our work on barley has resulted in the identification of approximately 1000 polymorphic markers from two different genomic representations. A DArT genetic map has been built for a population derived from a cross between cultivars Steptoe and Morex (Wenzl *et al.*, 2004). We are now in a position to deliver whole-genome profiles

of barley. Similarly, we have identified several hundred DArT markers in rice, and we are developing a genotyping tool in collaboration with the Australian rice industry.

Establishment stage

We are currently establishing the technology on wheat, cassava, apple, pigeonpea, and sorghum, in all cases in partnership with interested users. Together with colleagues from Plant Research International we also established DArT for the model species *Arabidopsis thaliana* (Wittenberg *et al.*, 2004).

Applications of DArT Markers

DArT markers can be used as any other genetic marker. With DArT, comprehensive genome profiles are becoming affordable for virtually any crop, regardless of the level of molecular information available for the crop. We anticipate that DArT genome profiles will be used for the recognition and management of biodiversity, for example in germplasm collections. Identification of duplicate accessions and a better understanding of the genetic relationships between the accessions could help to control the costs of maintaining these collections.

In plant breeding, DArT genome profiles will enable breeders to map QTL in one week, thereby allowing them to focus on the most crucial factor in plant breeding: reliable and precise phenotyping. Once many genomic regions of interest are identified in many different lines, DArT profiles accelerate the introgression of a selected genomic region into an elite genetic background (for example by Marker Assisted Back Crossing). Furthermore DArT profiles can be used to guide the assembly of many different regions into improved varieties. For that purpose, dense genome cover is essential in order to follow many regions simultaneously. Because of the large number of lines to be typed, high throughput and affordability are critical factors in this context.

A New Model for Technology Delivery

We believe that the social and environmental benefits of applying DArT could be quite substantial in developing countries, both as a result of accelerated plant breeding and better management of biodiversity. We are keen to see these benefits bring profit to users, but we also realize that developing DArT for the hundreds of species will require substantial resources. We will first explain why we think the “classical” path for the delivery of biotechnological inventions appears to be unable to deal with the peculiarities of a diverse and decentralized agricultural sector. We will then present our vision for the continuing improvement and the delivery of DArT.

In developing countries, agriculture encompasses subsistence farming that is not integrated in the global economy. In many industrial countries, agriculture is often a subsidized activity, with low economic margins. Although agricultural research results in large socioeconomic benefits (Alston *et al.*, 2000), the benefits are not easily captured by the private sector. For this reason it is becoming more difficult to attract venture capital in agricultural R&D. Early investors in agricultural biotechnology (“ag-biotech”), mostly agrochemical companies, tried to ensure a return on investment by appropriating the tools of innovation as well as its products. They used the same intellectual property (IP) protection mechanisms as in biomedical biotechnology. This left only a handful of multinational companies able to operate in this field, and certainly contributed to the negative public perception of agbiotech in Europe and elsewhere.

We believe that other models are more appropriate for agriculture. Rather than denying access to technologies as a way to realize their value, it may be possible to develop a framework of open access to the tools of innovation (O’Neill, 2004). Janet

Hope's web site at the Research School of Social Sciences of the Australian National University presents this open access concept for biotechnology in great detail: rsss.anu.edu.au/~janeth/home.html.

We have initiated the establishment of a network of DArT users, who will contribute their scientific expertise and resources to develop and improve the technology further. Key requirements to join the network are:

- Willingness to share improvements.
- Acceptance that financial rewards from the delivery of DArT services will not be derived from access to protected IP but from providing efficient “value for money” services to customers.

The growing success of the Open Source model in the software industry may provide some guidance toward establishing a sustainable system for open access biotechnology, where competition would take place, and profits would be made at the level of products and services. In this context, we hope that providing DArT services will be a sustainable activity for our organization and its partners, allowing us to develop and deliver improved genome profiling methods and to apply them to biological research and crop improvement.

Acknowledgments

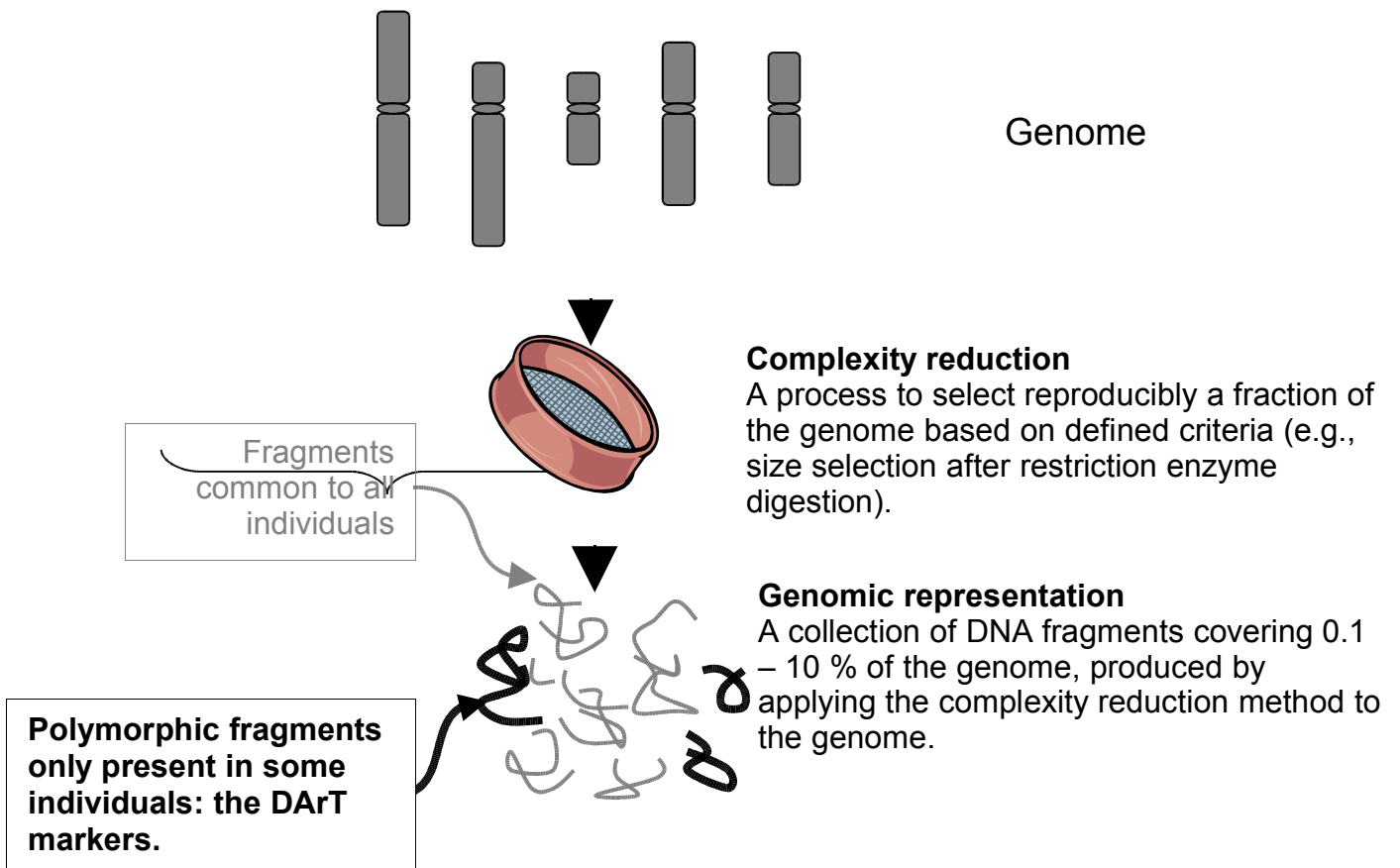
The development of DArT has been supported by CAMBIA, Rockefeller Foundation, Grains Research and Development Corporation (Australia), Cooperative Research Centre for Value Added Wheat, Rural Industries Research and Development Corporation (Australia), Horticulture Australia, Monticello Research Australia, Australian Wool Innovation, Gardiner Foundation, the Australian Federal Government Biotechnology Innovation Fund, and the Government of the Australian Capital Territory. Their support is gratefully acknowledged.

References

- Alston, J.M., C. Chan-Kang, M.C. Marra, P.G. Pardey, and T.J. Wyatt. 2000. A Meta-Analysis of Rates of Return to Agricultural R&D: Ex Pede Herculem? *Research Report 113*, International Food Policy Research Institute, Washington, D.C.
- Bonnett, D.G., G.J. Rebetzke, and W. Spielmeyer. 2004. Strategies for efficient implementation of molecular markers in wheat breeding. *Molecular Breeding*, in press.
- Conway, G. 1999. *The Doubly Green Revolution: Food for All in the Twenty-First Century*. Cornell University Press, Ithaca, N.Y.
- Jaccoud, D., K. Peng, D. Feinstein, and A. Kilian. 2001. Diversity arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Research*, 29(4), e25.
- Li, Z.K., B.Y. Fu, Y.M. Gao, J.L. Xu, C.H.M. Vijayakumar, J. Ali, R. Lafitte, A. Ismail, S. Yanagihara, M.F. Zhao, J. Domingo, R. Maghirang, F.Y. Hu, and X. Q. Zhao. 2003. Discovery and exploitation of “hidden” genetic diversity in germplasm collections for genetic improvement of abiotic stress tolerances in rice. *XIX International Congress of Genetics*, Melbourne 6-11 July 2003, www.genonet1.org/IGC2003/abstracts/30MinSpeakers-HTML/Li,%20Zhi-Khang.htm

- O'Neill, G. 2004. Open-source biology (interview of Richard Jefferson), *Australian Life Scientist*, February 2004, 14-15.
www.cambia.org.au/downloads/Biotechnology_News_Dec_03.pdf
- Peleman, J.D., and J.R. van der Voort. 2003. Breeding by design. *Trends in Plant Science*, 8(7), 330-334.
- Peng, K.M., D. Jaccoud, D. Kudrna, Y.G. Cho, and A. Kilian. 2002. Diversity Array Technology (DArT) applications to plant and animal genomics. *Plant and Animal Genomes X Conference*, San Diego, 12-16 January 2002, www.intl-pag.org/pag/10/abstracts/PAGX_W180.html
- Ribaut, J.M., and D.A. Hoisington. 1998. Marker-assisted selection: new tools and strategies. *Trends in Plant Science* 3, 236-239.
- Wenzl, P., J. Carling, D. Kudrna, D. Jaccoud, E. Huttner, A. Kleinhofs, and A. Kilian. 2004. Diversity Arrays Technology (DArT) for whole genome-profiling of barley. *Proc. Natl. Acad. Sci. U S A* 101(26), 9915-9920.
- Wittenberg, A.H.J., A. Kilian, and H.J. Schouten. 2004. DArT™, a promising genetic fingerprinting technology. *Plant and Animal Genomes XII Conference*, San Diego, 10-14 January 2004. www.intl-pag.org/12/abstracts/W36_PAG12_168.html
- Xiao J., J. Li, S. Grandillo, S.N. Ahn, L. Yuan, S.D. Tanksley, and S.R. McCouch. 1998. Identification of trait-improving quantitative trait loci alleles from a wild rice relative, *Oryza rufipogon*. *Genetics* 150(2), 899-909.

Figure 10.1. Principle of DArT.



The genotype of an individual is determined by **detecting the presence or absence of the polymorphic fragments** in a genomic representation from that individual. This is achieved by hybridizing the genomic representation to a microarray containing copies of the polymorphic sequences.

Table 10.1. Ongoing DArT projects in different species.

Species	No. representations tested	No. clones assayed
Rice	14	26,112
Barley	10	21,504
Wheat	5	14,592
Apple	3	1,920
Cassava	4	9,216
Perennial rye grass	5	5,376
Pigeon pea	4	5,376
Sorghum	2	1,536
Fungal pathogens of barley	4	5,376
Arabidopsis	1	1,536
Mouse	2	1,536
Bovine	2	1,536
Sheep	5	3,840