Theor Appl Genet (2012) 124:947–956
DOI 10.1007/s00122-011-1759-5

ORIGINAL PAPER

Bin mapping of tomato diversity array (DArT) markers to genomic regions of *Solanum lycopersicum* × *Solanum pennellii* introgression lines

Antoinette Van Schalkwyk · Peter Wenzl · Sandra Smit · Rosa Lopez-Cobollo · Andrzej Kilian · Gerard Bishop · Charles Hefer · Dave K. Berger

Received: 29 March 2011 / Accepted: 24 November 2011 / Published online: 13 December 2011
© The Author(s) 2011. This article is published with open access at Springerlink.com

Abstract Marker-trait association studies in tomato have progressed rapidly due to the availability of several populations developed between wild species and domesticated tomato. However, in the absence of whole genome sequences for each wild species, molecular marker methods for whole genome comparisons and fine mapping are required. We describe the development and validation of a diversity arrays technology (DArT) platform for tomato using an introgression line (IL) population consisting of wild *Solanum pennellii* introgressed into *Solanum lycopersicum* (cv. M82). A tomato diversity array consisting of 6,912 clones from domesticated tomato and twelve wild tomato/Solanaceous species was constructed. We successfully bin-mapped 990 polymorphic DArT markers together with 108 RFLP markers across the IL population, increasing the number of markers available for each *S. pennellii* introgression by tenfold on average. A subset of DArT markers from ILs previously associated with increased levels of lycopene and carotene were sequenced, and 44% matched protein coding genes. The bin-map position and order of sequenced DArT markers correlated well with their physical position on scaffolds of the draft tomato genome sequence (SL2.40). The utility of sequenced DArT markers was illustrated by converting several markers in both the *S. pennellii* and *S. lycopersicum* phases to cleaved amplified polymorphic sequence (CAPS) markers. Genotype scores from the CAPS markers confirmed the genotype scores from the DArT hybridizations used to construct the bin map. The tomato diversity array provides additional "sequence-characterized" markers for fine mapping of QTLs in *S. pennellii* ILs and wild tomato species.

Communicated by G. Bryan.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-011-1759-5) contains supplementary material, which is available to authorized users.

Present Address:
A. Van Schalkwyk
Inqaba Biotechnical Industries (Pty) Ltd, P.O. Box 14356, Hatfield 0028, South Africa

A. Van Schalkwyk · D. K. Berger (✉)
Department of Plant Science, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Private Bag X20, Hatfield 0028, South Africa
e-mail: Dave.berger@fabi.up.ac.za

P. Wenzl · A. Kilian
Diversity Arrays Technology P/L, GPO Box 7141, Yarralumla, ACT 2600, Australia

Present Address:
P. Wenzl
Centro Internacional de Mejoramiento de Maiz y Trigo (CIMMYT), Apdo. Postal 6-641, 06600 Mexico, DF, Mexico

S. Smit
Applied Bioinformatics, Plant Research International and Laboratory of Bioinformatics, Wageningen University, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands

R. Lopez-Cobollo · G. Bishop
Imperial College London, London, UK

C. Hefer
Bioinformatics and Computational Biology Unit, Department of Biochemistry, University of Pretoria, Private Bag X20, Hatfield 0028, South Africa

Introduction

Tomato is a commercially important food crop of global importance. Breeding of new varieties is consumer driven with the latest demand calling for fruit rich in nutritional qualities with properties beneficial to human health (Bai and Lindhout 2007). Tomatoes contain various carotenoids, which have been implicated in reducing the risk of certain cancers (Yeh et al. 2009). The protective properties of carotenoids are mainly due to them being potent antioxidants, but a non-oxidative mode of action has also been proposed for lycopene (Agarwal and Rao 2000). Lycopene is an acyclic isomer of β-carotene and the most abundant antioxidant present in red tomatoes. Unlike β-carotene, lycopene cannot be converted to vitamin A in the body, but it has twice the antioxidant potential of β-carotene (Agarwal and Rao 2000; Miller et al. 1996). Recently, tomatoes selected for their high lycopene levels have been marketed as specialty fruits (Bai and Lindhout 2007; Butelli et al. 2008). It is therefore a challenge for breeders to generate fruits containing high levels of antioxidants such as lycopene and carotene without compromising the taste or placing additional demands on the environment during cultivation (Zamir 2001). It has been shown that utilizing wild species during breeding contributes not only a source of genetic diversity but also can improve fruit quality traits, such as increased levels of lycopene (Liu et al. 2003).

An effective tool for investigating valuable agronomic traits offered by wild species in the genetic background of domesticated varieties is the construction of an introgression line (IL) population. An IL population is a set of individuals each containing a defined chromosomal segment from the donor wild species in the uniform genetic background of the domesticated recipient (Eshed and Zamir 1995; Zamir 2001). This study investigated the genetic diversity of an IL population comprising 75 *S. lycopersicum* (cv. M82) lines each harboring a single RFLP-defined chromosomal introgression originating from *S. pennellii* (Eshed and Zamir 1995). The ILs are nearly isogenic to the domesticated tomato except for each *S. pennellii* introgression. This population covers the whole genome with 107 marker-defined mapping bins (average size 12 cM) (Pan et al. 2000).

Molecular maker technologies contribute to the effective utilization of crop genetic diversity through marker assisted selective breeding (Moose and Mumm 2008). Several techniques have been used for marker-trait association studies, such as restriction fragment length polymorphism (RFLP), randomly amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), simple sequence repeats (SSR) and single nucleotide polymorphisms (SNP) (Jones et al. 2009). The efficiency of each method is based on throughput, reproducibility, time, cost, and dependency on sequence information. Diversity array technology (DArT) is a hybridization-based approach that utilizes the simultaneous analysis ability of DNA microarrays without the prerequisite of complete genome or marker sequence information. DArT is a low cost, high-throughput technology that has been applied to the analysis of at least a dozen crop species including rice (Jaccoud et al. 2001), eucalyptus (Lezar et al. 2004), barley (Wenzl et al. 2004), cassava (Xia et al. 2005), wheat (Akbari et al. 2006; White et al. 2008), pigeonpea (Yang et al. 2006), *Sorghum* (Mace et al. 2008), banana (Risterucci et al. 2009), sugarcane (Heller-Uszynska et al. 2011), as well as *Arabidopsis* (Wittenberg et al. 2005), fern and moss (James et al. 2008). However, to date there have been no published developments of DArT for members of the Solanaceae.

This study reports on the successful construction of a diversity array from domesticated and wild tomato species, validation of the DArT markers by bin mapping in a *S. pennellii* × *S. lycopersicon* IL population, and verification of DArT marker sequences by physical mapping to the tomato genome sequence. Our tomato diversity array platform can serve as a useful resource for genetic mapping and genotyping wild tomato species on a whole genome level, or fine mapping of QTLs by conversion of DArT markers to CAPS markers.

Materials and methods

Plant material

The tomato accessions used to prepare DArT libraries represented cultivated species *Solanum lycopersicum* (cv. M82) (LA3475), *S. lycopersicum*-E-6203 (LA4024), *S. lycopersicum*-Florida 7547 (LA4025), *S. lycopersicum*-Florida 7481 (LA4026), *S. lycopersicum* (LA1491), ecotypes *S. lycopersicum*-Santorimis, *S. lycopersicum*-Chiou and *S. lycopersicum*-Limnou from Greece, wild tomato species *Solanum pennellii* (LA0716), *S. galapagense* (LA1410), *S. pimpinellifolium* (LA1586), *S. habrochaites* (LA1775), *S. arcanum* (LA2153), *S. neorickii* (LA2615) and *S. chmielewskii* (LA2695). Solanum species that produce berries *S. africanum, S. pseudocapsicum, S. catabolense, S. chenopodioides, S. retroflexum* and *S. nigrum* were collected in South Africa, and were added to enhance the genetic diversity of the DArT libraries. Equal quantities of DNA of each genotype were combined for the construction of each DArT library. Germplasm for all the tomato accessions except the three Greek ecotypes were obtained from the C.M. Rick Tomato Genetics Resource Center (TGRC), California, USA. The three Greek ecotypes were kindly donated by A. Kanellis, Aristotle University of Thessaloniki, Greece. Three 786-clone DArT libraries previously

constructed at Diversity Array P/L, Canberra, Australia were also included in the final tomato diversity array (*S. lycopersicum* near isogenic line (NIL) library from a proprietary source, *S. lycopersicum cv.* West Virginia 700 library, Wild tomato species (additional) library from a proprietary source; A. Kilian and P.Wenzl, unpublished).

DNA extractions

Seeds were germinated on MS-media and transplanted to potting soil 10 days post germination. One or two of the youngest leaves were crushed and DNA was extracted using a modified cetyltrimethylammonium bromide (CTAB) method (Doyle and Doyle 1987). In order to reduce high phenol contamination, the samples were treated with 5% Polyvinylpolypyrrolidone (PVPP) during the initial cell lysis. DNA was dissolved in ultra high quality water and the concentration adjusted to 100 ng μl^{-1}.

Diversity array construction

Genomic representations were produced by digesting 100 ng of tomato DNA with 2U *Pst*I and 2U *Taq*I (NEB, Beverly, MA). A *Pst*I adapter (5′-CAC GAT GGA TCC AGT GCA-3′ annealed to 5′-CTG GAT CCA TCG TGC CA-3′) was simultaneously ligated to the genomic fragments using 4U T4 ligase (NEB, Beverly, MA). A 1 µl aliquot of the ligation mixture was the template in a subsequent 50 µl amplification reaction using primer DArT-*Pst*I (5′-GAT GGA TCC AGT GCA G-3′) according to the following specifications: 94°C for 1 min, followed by 30 cycles of 94°C for 20 s, 58°C for 40 s, 72°C for 1 min and finally 72°C for 7 min. Libraries of genomic representations were constructed from amplified fragments as described by (Jaccoud et al. 2001). Individual colonies were grown in 384-well plates containing Luria Broth supplemented with 4.4% [v/v] glycerol and 100 µg ml^{-1} ampicillin. Aliquots of the cultures were used as templates to amplify inserts according to (Jaccoud et al. 2001). Amplicons were dried, resuspended in DArT spot buffer and spotted in duplicate randomly on polylysine-coated slides (Erie Scientific) using a MicroGrid II arrayer (Genomics Solution; Lincoln). These tomato diversity array slides were dried at room temperature for 2 days followed by incubation in hot water (95°C) for 2 min and drying by centrifugation.

Hybridization of DNA samples

Genomic representations were generated from *S. lycopersicum, S. pennellii* and 75 ILs using the same complexity reduction method described for library construction (*Pst*I/*Taq*I). The DArT-*Pst*I amplified fragments were concentrated tenfold by precipitation using one volume isopropanol. The fragments were subsequently heat denatured and labeled with 0.1 µl 25 nmoles Cy3- or Cy5-labeled dUTP using random decamers and exo-Klenow DNA polymerase I (NEB, USA). The labeled representations were mixed with 6-carboxy fluorescein (FAM) labeled polylinker fragment of the pCR2.1-TOPO vector and 50 µl ExpressHyb buffer (Clonetech, Palo Alto) containing 10 µg ml^{-1} herring sperm DNA (Jaccoud et al. 2001; Wenzl et al. 2006). Following heat denaturation, target DNA was hybridized to the microarray slides in a humidified hybridization chamber at 65°C overnight. The slides were washed according to (Jaccoud et al. 2001) and scanned using a Tecan LS300 confocal laser scanner (Grödig, Austria).

Image analysis and polymorphism scoring

Each microarray image was analyzed with DArTsoft (DArT P/L, Australia http://www.diversityarrays.com/software.html#dartsoft). The program localizes features, rejects those with a weak reference signal, calculates and normalizes relative hybridization intensities (Cy3/FAM or Cy5/FAM) of each DArT marker across all slides. It uses a C mean fuzzy-clustering algorithm to score the hybridizing DNA fragment as present or absent (1 or 0, respectively) when compared across all slides (Wenzl et al. 2004). Polymorphic DArT markers were evaluated based on additional quality control (QC) parameters (P, call rate, PIC value and discordance). Parameter P measures the fraction of the total variation in relative signal intensity across all individuals due to bimodality. The call rate indicates the frequency at which a marker is scored as clearly falling into a present or absent group across replicates. The polymorphism information content (PIC) value is a measurement of how informative a marker is. The discordance indicates the lack of reproducibility of an allelic state score between replications of the same genotype (Wenzl et al. 2004, 2006). The thresholds for each of the QC parameters were determined empirically by regression analysis. The slope of the regression line was used to determine the cutoff value for each QC parameter. The threshold values for *P* value, call rate, PIC and discordance ranged from 78 to 100, 94 to 100%, 0.04 to 0.5, and <0.01, respectively.

Genetic bin mapping

Genomic representations from the 75 ILs were labeled and hybridized to the tomato diversity arrays in dye swap pairs using the methods described above. Markers were selected if they were polymorphic between *S. pennellii* and *S. lycopersicum*, and passed the quality control parameters for *P* value, call rate, PIC and discordance. Polymorphic mark-

ers were those that had the same absence or presence score as *S. pennellii* in only one or more overlapping ILs, and had the same score as *S. lycopersicum* in all the other ILs. The presence/absence scores for these markers in the ILs were converted into genotype codes (A for *S. pennellii*/B for *S. lycopersicum*), depending which parents score they matched in each IL. For example, if a DArT marker was scored "present" in *S. pennellii* and scored "present" in an IL, then the genotype code in the IL would be "A". Following the same convention, if a DArT marker was scored "absent" in *S. pennellii* and scored "absent" in another IL, then the genotype code in the other IL would also be "A". A table of the genotype codes was prepared in Microsoft Excel in which the ILs were arranged in chromosome order as column headings, and the DArT markers were sorted together with the RFLP markers from Eshed et al. (1992) in rows so that the *S. pennellii* genotype belonging to one or overlapping ILs were clustered together, therefore bin mapping the markers according to the chromosomal localization of the individual *S. pennellii* introgression fragments (Online resource 1).

Sequence analysis of polymorphic clones

Diversity arrays technology clones located on ILs associated with increased levels of carotene and lycopene (Liu et al. 2003) were selected for sequencing (Macrogen, USA). These IL's were IL1-2, IL2-3, IL2-6, IL3-2, IL3-4, IL4-1, IL4-1-1, IL4-3, IL4-4, IL5-1, IL7-3, IL9-1, IL9-2-5, IL11-1 and IL12-1. Sequences of hybridization probes used for RFLP analysis (Eshed et al. 1992) were obtained from the SGN website (http://solgenomics.net/tomato). The sequence dataset was uploaded to the custom built database SSHdb which automatically performs vector sequence clipping, clustering of redundant clone sequences, selection of a representative clone from each cluster, and annotation using BLASTN, BLASTX, and Blast2GO against GenBank (Coetzer et al. 2010). The data can be visualized and exported in a series of tab-delimited tables, and are publicly accessible (http://sshdb.bi.up.ac.za, login: tomatoguest; password dart_tomato). DArT clone and RFLP sequences were aligned with the *S. lycopersicum* cv. Heinz tomato genome sequence SL2.40 pseudomolecules (http://solgenomics.net/genomes/Solanum_lycopersicum/index.pl) using BLAST+ (v2.2.24+, applications MEGABLAST and BLASTN; identity >80%; *E* value $<1E^{-10}$) (Camacho et al. 2009). BLASTN results were used when there were no MEGABLAST results that passed the filter. BLASTX of the DArT clone sequences was carried out against the gene models predicted from the *S. lycopersicum* cv. Heinz SL2.40 genome sequence (ITAG version 2.3; 34,727 gene models; released on April 29, 2011; downloaded from http://solgenomics.net/itag/release/2.3/list_files).

Cleaved amplified polymorphic sequence (CAPS) analysis

Solanum lycopersicum cv. Heinz and *S. pennellii* DNA were used as template DNA in PCRs to amplify DArT marker sequences using primers as shown in Online resource 5. PCR products were sequenced and compared to identify polymorphisms that could be used in CAPS analysis. Primers were used to amplify products from parental and the most informative IL lines. PCR products were then digested with the appropriate restriction enzyme according to manufacturer's conditions, namely: DArT438438, *Mnl*I; DArT440552, *Bpm*I; DArT441173, *Taq*I; DArT436990, *Mnl*I; DArT437279, *Mbo*II; DArT437970, *Sau*3AI. Digestion products were separated on 2% (w/v) TAE agarose gels.

Results

Construction of tomato diversity array

The *Pst*I/*Taq*I complexity reduction method has been compared with other restriction enzyme combinations in several plant genomes where it resulted in the highest level of polymorphisms for DArT analysis (Akbari et al. 2006; Wenzl et al. 2004). In addition, previous studies indicated that *Pst*I/*Taq*I was a suitable complexity reduction technique for the *Solanaceae* (A.Kilian and P.Wenzl, unpublished), and thus this method was chosen for this study. Nine DArT libraries were constructed using the *Pst*I/*Taq*I complexity reduction method from (i) several *S. lycopersicum* accessions; (ii) wild tomato species *S. pennellii*, *S. habrochaites*, *S. galapagense*, *S. pimpinellifolium*, *S. arcanum*, *S. neorickii* and *S. chmielewskii*; and (iii) related *Solanum* species that produce berries *S. africanum, S. pseudocapsicum, S. catabolense, S. chenopodioides, S. retroflexum* and *S. nigrum* (Table 1). A total of 6,912 clones from the nine libraries were combined and PCR products of each insert were spotted in random order in duplicate on each polylysine slide to produce the tomato diversity array.

The tomato diversity array has a high polymorphic frequency between domesticated tomato and the wild species *S. pennellii*

The frequency of polymorphisms in the 6,912 clone tomato diversity arrays was evaluated by hybridizing the genomic DNA of *S. lycopersicum* (cv. M82) and *S. pennellii* on eight replicate arrays, half of which were dye swaps. A total of 1,645 clones were classified as polymorphic markers since they showed different hybridization scores between the two tomato species and met the prerequisite criteria for the four quality control parameters: *P* value, discordance, PIC and

Table 1 DArT libraries used to construct the tomato diversity array, and number of clones scored as polymorphic when the array was hybridized with the two parents *S. pennellii* and *S. lycopersicum* (first experiment), and the introgression lines (second experiment)

DArT library	Clones	Hybridization of *S. pennellii* and *S. lycopersicum* to diversity array		Hybridization of *S. pennellii* × *S. lycopersicum* IL lines to diversity array	
		Polymorphic[d]	Polymorphic (%)	Polymorphic[e]	Polymorphic (%)
Solanum pennellii (LA0716)	1,152	407	35	192	17
Solanum lycopersicum cv. M82	1,152	239	21	154	11
S. lycopersicum NIL	768	235	31	157	18
*S. lycopersicum cv.*W.Virginia 700	768	186	24	134	15
S. lycopersicum accessions[a]	384	87	23	61	13
Solanum habrochaites	384	79	21	46	12
Wild tomato species[b]	768	140	18	83	9
Wild tomato species (additional)	768	208	27	130	16
Related *Solanum* species[c]	768	64	8	33	4
Total	6,912	1,645	24	990	14

[a] *S. lycopersicum* E-6203 (LA4042), *S. lycopersicum*-Florida 7547 (LA4025), *S. lycopersicum*-Florida 7481 (LA4026), *S. lycopersicum* (LA1419), *S. lycopersicum* (cv. Santorimis), *S. lycopersicum* (cv. Chiou), and *S. lycopersicum* (cv. Limnou)

[b] *S. galapagense* (LA1410), *S. pimpinellifolium* (LA1586), *S. habrochaites* (LA1775), *S. arcanum* (LA2153), *S. neorickii* (LA2615) and *S. chmielewskii* (LA2695)

[c] *S. africanum, S. pseudocapsicum, S. catabolense, S. chenopodioides, S. retroflexum* and *S. nigrum*

[d] Number of clones polymorphic between *S. pennellii* and *S. lycopersicum*

[e] Number of clones that had the same presence/absence score as the donor parent *S. pennellii* in only one or overlapping ILs, and had the opposite presence/absence score in the remaining ILs (i.e. the same as the recurrent parent *S. lycopersicum*)

call rate (see "Materials and methods"). The selected polymorphic markers had a *P* value greater than 86 (average *P* value, 94). The majority of the markers (1,160) had a call rate of 100%, while the call rate of another 359 markers was 94.4% (average call rate: 97.7%). A total of 1,143 markers had a PIC value of 0.5 (average PIC value, 0.49) and the average discordance of all the polymorphic markers was 0.008. The *S. pennellii* library produced the largest number of polymorphic markers, whereas the library constructed from related *Solanum* species contributed the least number of polymorphic markers (Table 1). The tomato diversity array had a total frequency of polymorphisms between the two tomato species of 24%, which was deemed sufficient for subsequent analysis of the *S. lycopersicum* × *S. pennellii* IL population.

Tomato diversity array analysis of *S. lycopersicum* × *S. pennellii* IL population

Genomic representations of the *S. lycopersicum* × *S. pennellii* ILs were hybridized to the diversity array in duplicate using a dye swap strategy. Four replicate dye swaps of the parents, *S. lycopersicum* and *S. pennellii,* were also hybridized as controls. Only markers that were scored polymorphic between the parents in this particular experiment were used for subsequent analysis. Within this group of polymorphic markers, a marker was selected if it had the same presence/absence score as the donor parent *S. pennellii* in only one or overlapping ILs, and had the opposite presence/absence score in the remaining ILs (i.e. the same as the recurrent parent *S. lycopersicum*). Nine of the ILs did not contain any polymorphic markers unique to them or overlapping IL. A total of 990 polymorphic markers were identified that adhered to the above mentioned criteria (Table 1). This was a subset of the 1,645 markers identified to be polymorphic between *S. lycopersicum* and *S. pennellii* in the first experiment (Table 1). There was a good representation of markers that gave "present" scores from each parent library (11% from *S. lycopersicum* and 17% from *S. pennellii*). The average call rate was 96% and the average *P* value was 78. PIC values ranged from 0.04 to 0.35 (average PIC value of 0.11) and the average discordance was 0.008. The polymorphic frequencies generated by each of the nine libraries are indicated in Table 1. The highest numbers of polymorphic markers were derived from the *S. lycopersicum* NIL library and the *S. pennellii* library, while the related *Solanum* species contributed the least number of polymorphic markers (Table 1).

Genetic bin mapping of DArT markers in IL population

The presence and absence scores for the 990 polymorphic DArT markers in 66 *S. lycopersicum* × *S. pennellii* IL lines were translated to the corresponding allelic phases of the

two parents (i.e. "A" if the score in the IL matched the score in *S. pennellii*; and "B" if the score in the IL matched the score in *S. lycopersicum*). The DArT marker genotypes (rows) for the ILs were rearranged in a Table in which the ILs were sorted in chromosome order (columns). This clusters those DArT markers that were in the same phase as *S. pennellii* for each individual IL (Online resource 1). The DArT markers were therefore binned into groups belonging either to one or more overlapping IL's. The genotype scores of the known RFLP markers from Eshed and Zamir (1995) used to delimit the ILs were also sorted with the DArT marker scores. The RFLP markers, which have previously been mapped, therefore delimit the bins (Eshed and Zamir 1995). A total of 101 bins were assigned, with 42 belonging to single introgression lines, 43 overlapping between 2 adjacent ILs, 14 shared between 3 overlapping ILs and 3 bins had positive hits between 4 ILs. The result was visualized as a "graphical genotype" in which the markers in *S. pennellii* phase are color coded orange and markers in *S. lycopersicum* phase are green (Fig. 1; Online resource 1). Figure 1 shows that the DArT markers can be clearly associated with the *S. pennellii* introgressions in specific introgression lines.

Sequencing and physical mapping of selected DArT markers

Introgression line populations provide an opportunity to investigate the function of a single fragment of donor parent DNA (e.g. *S. pennellii)* in the genetic background that is similar to the recipient parent (e.g. *S. lycopersicum*) (Zamir 2001). Various phenotypes (traits) have been quantified in the 75 ILs derived from the *S. pennellii* × *S. lycopersicum* cross. Among these traits is the ability of each IL to produce and accumulate antioxidants, such as carotenoids (Liu et al. 2003). The RFLP delimited bins associated with ILs previously shown to have altered levels of the carotenoids lycopene and carotene were chosen, and the DArT markers in these bins were sequenced (ILs are listed in "Materials and methods"). A total of 431 DArT clones were sequenced of which 99 were identified as redundant clones. A total of 332 non-redundant clone sequences were searched against Genbank (8 September 2010), and 44% matched predicted protein coding sequences either in GenBank or gene models predicted from the *S. lycopersicum* cv. Heinz SL2.40 genome sequence (ITAG version 2.3) (BLASTX *E* value < $1E^{-10}$) and an additional 26% had significant hits to nucleotide sequences in Genbank (BLASTN *E* value < $1E^{-10}$) (Online resource 2). The sequence dataset, including Blast2GO annotations, can be visualized and exported from the custom built database SSHdb (http://sshdb.bi.up.ac.za; login: tomatoguest; password dart_tomato) (Coetzer et al. 2010).

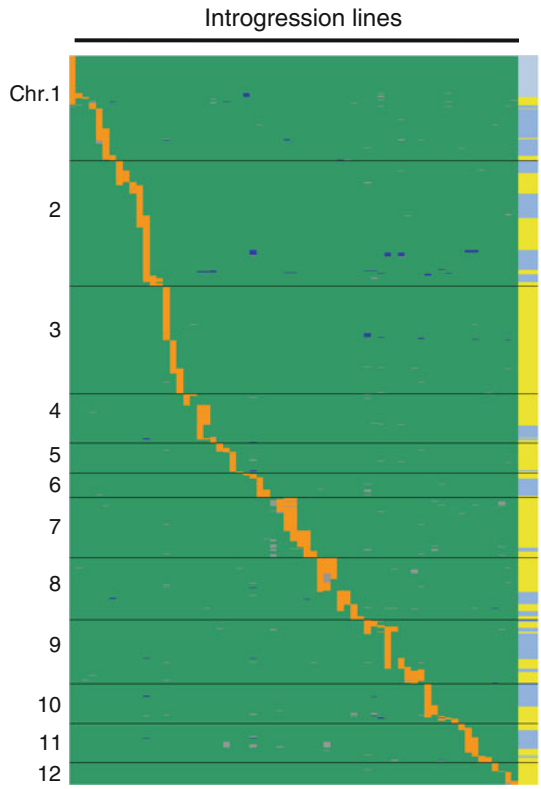

Fig. 1 Bin map of DArT and RFLP markers in the *S. lycopersicum* × *S. pennellii* introgression line population. *Each column* represents one of the 66 introgression lines, whereas *each row* represents the genotype scores for either DArT or RFLP markers. The data have been sorted to illustrate bins which contain markers that exhibit the same genotype, and thus represent the fragments introgressed from the wild parent (*orange*) into the domesticated parent (*green*) for all 12 chromosomes. Missing data or unknown genotype scores are indicated by *gray* or *dark blue*, respectively. Bins represented by DArT or RFLP markers are indicated in *light blue* or *yellow*, respectively (*right hand column*). In the print version of Fig. 1, the colors are represented as follows: *orange* = *white*, *green* = *gray*, *gray* = *light gray*, *dark blue* = *black*; and in the *right hand column*: *light blue* = *gray*, *yellow* = *white*

Several DArT clones mapping to ILs were predicted to code for proteins with putative roles in fruit quality traits, for example DArT_195128 (neutral invertase; *E* value = $1.3E^{-80}$; mapping to IL3-2), DArT_439268 (pectinesterase; $2E^{-31}$; IL3-2), DArT_90792 (glucosylceramidase; $9E^{-14}$; IL3-2), DArT_440443 (lectin/glucanase; $9E^{-50}$; IL4-3), DArT_437740 (glucan synthase; $9E^{-20}$; IL4-1-1), DArT_438110 (glucosyltransferase; $2E^{-19}$; IL4-4), DArT_440552 (phosphatidylinositol *N*-acetylglucosaminyltransferase; $7E^{-24}$; IL2-3/2-4), and DArT_439313 (sucrose phosphatase; $2E^{-46}$; IL10-2-2) (Online resource 2).

Physical mapping of DArT markers to the tomato genome sequence

Draft un-annotated releases of the genome sequence of *S. lycopersicum* cv. Heinz are currently being made available

through the SGN website (http://solgenomics.net/tomato). Sequenced DArT markers, as well as RFLP markers, were aligned to the latest genome build (release SL2.40), which is made up of 91 scaffolds on pseudomolecules 1 to 12, plus an additional 3,132 scaffolds on pseudomolecule 0. A total of 76 RFLP and 114 DArT markers showed significant top MEGABLAST/BLASTN hits (sequence identity >80%; $E < 1E^{-10}$) to 38 sequence scaffolds on the same chromosomes to which the markers had been bin mapped (Online resource 3). Furthermore, the order of the scaffolds based on sequence matches of the bin-mapped markers was the same as the order of scaffolds in release SL2.40 (Online resource 3). Importantly, physical mapping of the marker sequences to the scaffold sequences indicated that the order of the marker sequences on each scaffold correlated well with the order of the markers based on the bin map (Online resource 3).

It was expected that not all DArT markers would show top BLASTN hits to the same chromosome as predicted by bin mapping, since a DArT marker will only be visualized if it is flanked by *Pst*I sites, which are separated by not more than ~2.0 kb, which allows it to be amplified in the genomic representation of at least one parental genotype. The largest insert size in the sequenced clones was 1.6 kb (DArT_89524). If PstI fragments also contain a *Taq*I site, they will also be removed from the genomic representation. This means that a DArT clone may have multiple BLAST hits in the SL2.40 genome sequence, but only one of them corresponds to a DArT marker. This might also be the case for some RFLP markers in tomato species that are not highly heterozygous—the probe may have sequence identity to multiple sites in the genome, but only one position provides the restriction site polymorphism between the genotypes under study.

Our analysis supported this expectation that not all RFLP or DArT markers would show top BLASTN hits to the same chromosome as predicted by mapping. A total of 86% of the RFLP markers had top BLAST hits to the same chromosome in SL2.40 as mapped in Eshed et al. (1992) (Online resource 3). The sequenced DArT markers that had top BLAST hits to the same chromosome in SL2.40 as the bin map were 40% of the total (Online resource 3). In this category, there were almost equal numbers of DArT markers in the *S. lycopersicum* and *S. pennellii* phases. Approximately half of the remaining DArT markers had significant BLASTN hits to several chromosome positions, one of which represents the DArT marker (Online resource 1). The remaining markers were derived from *S. pennellii*, wild tomato species, or other accessions of domesticated tomato, and thus the BLAST hit on SL2.40 did not represent the DArT marker (Online resource 1). Only 32 DArT markers had no significant hits to SL2.40, and these were either derived from *S. pennellii*, wild tomato species, or were cv. M82 sequences not present in *S. lycopersicum* cv. Heinz (Online resource 1). Alternatively, they may correspond to gapped regions in and between the scaffolds of SL2.40.

To confirm the chromosomal location of the DArT marker sequences several were developed further so that they could be used in cleaved amplified polymorphic sequence (CAPS) analysis. Figure 2 shows the results from CAPS analysis of two markers (DArT 441173 and DArT 436990) that map to chromosome 4. These markers are representative of the two phases of markers i.e. DArT 441173

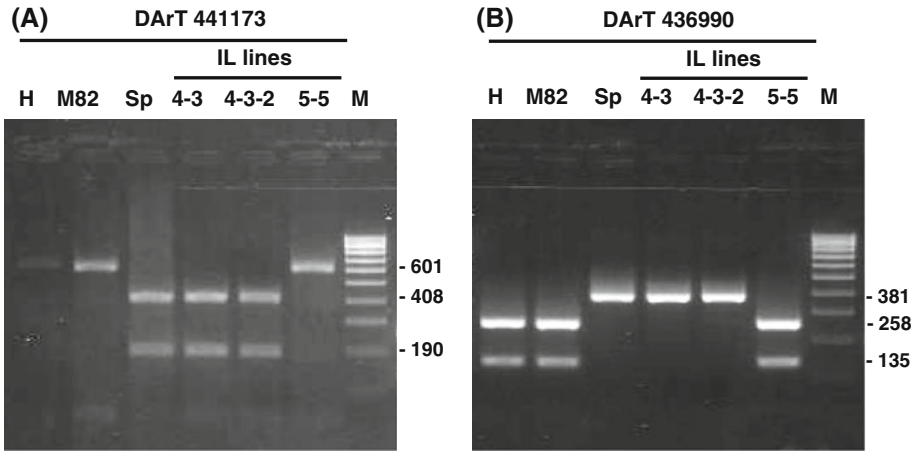

Fig. 2 Digested-PCR products of DArT marker sequences reveal predicted polymorphisms and chromosomal locations. CAPs analysis of DArT marker sequences from parental and introgression lines. SYBR Safe-stained 2% TAE agarose gels highlighting the polymorphisms observed between Heinz, M82, *S. pennellii* and different IL lines DNAs. **a** *Taq*I digestion of PCR-amplified DNA corresponding to S. *lycopersicum* DArT marker 441173 using primers DArT39 and DArT40. **b** *Mnl*I digestion of PCR-amplified DNA corresponding to *S. pennellii* DArT marker 436990 using primers DArT41 and DArT42. Template DNA; S. *lycopersicum* cv. Heinz (H), S. *lycopersicum* cv. M82 (M82), *S. pennellii* (*Sp*), Introgression lines (IL), molecular weight Hyperladder IV (BIOLINE) (M)

from *S. lycopersicum* and DArT 436990 from *S. pennellii*. Further examples of CAPS analysis that confirm the chromosomal location of DArT markers 438438, 440552, 437279 and 437970 can be found in Online resource 4.

Discussion

This study reports the development of a diversity array platform for wild and domesticated tomato species. The diversity array was constructed with *Pst*I/*Taq*I-derived genomic representation of not only domesticated and wild tomato species but also berry-producing related *Solanum* species. The latter contributed the least number of polymorphic fragments, probably due to their phylogenetic distance compared to wild and domesticated tomato species (Marshall et al. 2001). The tomato diversity array had 24% clones that were polymorphic between *S. lycopersicum* (cv. M82) and *S. pennellii* (Table 1). There was a good balance of positive hybridizations to these polymorphic clones from each parent (*S. pennellii*: 10.6%; *S. lycopersicum*: 13.2%) (Table 1). The 1,645 polymorphic markers were assessed based on their call rate, PIC value and scoring reproducibility and their quality was comparable with DArT markers previously identified for barley (Wenzl et al. 2004), cassava (Xia et al. 2005), pigonpea (Yang et al. 2006) and wheat (Akbari et al. 2006). Sequencing of a subset of the 1,645 DArT markers indicated that ~18% are likely to be redundant.

As proof of concept of the tomato diversity array, an IL population of *S. pennellii* in the recurrent background *S. lycopersicum* (cv. M82) was screened. This study identified 990 DArT markers that could be bin mapped in 66 of the ILs together with 108 RFLP markers that had previously been used to demarcate the *S. pennellii* introgressions in the recurrent *S. lycopersicum* (cv. M82) background (Eshed and Zamir 1995) (Fig. 1). The marker order was established by keeping the order of the RFLP markers fixed, and sorting the DArT markers so as to minimize the number of recombination events (Fig. 1). This resulted in a bin map of 1,098 markers across all 12 tomato chromosomes arranged in 102 bins, which represents a tenfold increase in the number of markers available for this population (Fig. 1, Online resource 1).

The reason that the number of DArT markers in the final bin map (990) was less than the number of markers that were scored as polymorphic between the parents (1,645) is that we applied stringent selection criteria for the bin map (Table 1; Online resource 1). Firstly, only markers that passed the DArTsoft quality control criteria of *P* value >70 were chosen. The *P* value was a measure of whether the hybridization signals across the ILs could be sorted by the clustering algorithm into the "absence" and "presence" scores (Wenzl et al. 2006). Clear examples of multi-locus markers with two alleles derived from the recurrent parent were removed, since, although these would have been scored as polymorphic between the parents, they would have scored the "present" across all the ILs. All wild phase markers with wild alleles in unexpected ILs were also removed, unless their segregation signature was confirmed by other co-segregating DArT markers in the opposite phase. This resulted in 990 robust DArT markers, made up of 323 having a *S. pennellii* allelic phase ("A") and 667 with a *S. lycopersicum* allelic phase ("B") in the final bin map (Online resource 1).

An advantage of DArT markers are that they are cloned and therefore can be readily sequenced. Furthermore, the *Pst*I/*Taq*I complexity reduction step enriches for non-methylated DNA, since *Pst*I sites are subject to methylation. This increases the probability of DArT markers in gene-rich regions (Peleg et al. 2008). For example, 30% of a subsample of DArT markers from sugarcane were from the actively transcribed regions of the genome (Heller-Uszynska et al. 2011). A sample of ~400 DArT markers were selected for sequencing, based on DArTs bin mapped to ILs associated with substantial increases in the production and accumulation of health quality metabolites lycopene and carotene (Liu et al. 2003). Only 18% of the sequenced clones were redundant, and 44% of the 322 non-redundant clones had BLASTX matches to protein coding genes (BLASTX *E* value $<1E^{-10}$), confirming that DArT markers are enriched in genes. Even though a large number of the DArT markers had BLASTX hits to sequences annotated as "hypothetical proteins", various metabolic enzymes were identified.

Comparison of DArT and RFLP marker sequences to the latest release of the *S. lycopersicum* cv. Heinz genome sequence (SL2.40) enabled a comparison between the bin map order of markers and the order of the marker sequences on the scaffolds (Online resource 3). This provided further verification of the DArT marker bin map since the order of 38 scaffolds across all 12 chromosomes correlated well, as well as the order of RFLP and DArT markers within scaffolds. Several DArT marker sequences were used to generate CAPS markers and confirmation of their location on the IL lines using CAPS analysis provided further molecular proof of chromosomal location (Fig. 2, Online resource 4). This proof included markers derived from both *S. lycopersicum* and *S. pennellii* phases and thus highlights the utility of the DArT marker system.

This work has made available sequenced DArT markers that are bin mapped in the *S. lycopersicum* × *S. pennellii* introgression line population, which provides a useful resource for researchers carrying out marker-trait association studies in domesticated and wild tomato species. For example, it can facilitate fine mapping of QTLs once a

researcher has delimited a QTL in a particular introgression line, but has used all available RFLP and other markers. The availability of several-fold more DArT markers for any particular introgression has facilitated fine mapping of QTLs within the EU SOL project (G. Seymour, J. Hirschberg, personal communications). Furthermore, the ease with which DArT markers can be converted to CAPS markers indicates how these markers can become a key resource in future mapping experiments.

In this study, we sequenced a selection of DArT markers that were bin mapped to introgressions with altered lycopene and carotene levels (Liu et al. 2003) with the aim of identifying markers linked to alleles that confer the altered traits. Future work, outside the scope of this study, would be to phenotype sub-ILs and genotype them using CAPS markers derived from the DArT markers, as well as RFLP markers. We also examined the BLASTX hits of the DArT markers to determine if they represented enzymes that may be involved in metabolism of lycopene or carotene (Online resource 2). DArT markers corresponding to enzymes with putative roles in fruit quality traits were identified (Online resource 2), however, since each introgression represents a large fragment of *S. pennelli* genomic DNA, it is unlikely that randomly cloned DArT markers will represent alleles responsible for the differences in lycopene or carotene levels.

The tomato diversity array was constructed from a range of Solanaceous species and shows sufficient polymorphisms between domesticated tomato and several wild tomato species (Table 1). Therefore it also provides a platform for construction of genetic maps and QTL identification in wild tomato species. These genetic maps can be made with DArT markers alone or in combination with other markers, as shown in barley, for example (Wenzl et al. 2006).

The tomato diversity array should be seen in context of future developments in marker technology for tomato, such as the recently announced SolCAP tomato SNP platform which has a similar number of markers (7720; http://solcap.msu.edu/). A limitation of the current SolCAP SNP platform is that it is heavily biased towards inbred tomato lines, whereas the tomato diversity array has a broader range of polymorphisms in wild tomato species (Table 1). In future, sequencing of bin-mapped DArT markers from the wild species can add significant value to the SNP array by increasing the number of SNPs between domesticated and wild tomato species.

Whole genome sequencing is predicted to play an increasing role in genetic mapping and genotyping in tomato, although costs currently preclude its implementation by tomato molecular breeders. The availability of the genome sequence for domesticated tomato was important for validation of the DArT markers in our study. The next development is likely to be release of the *S. pennellii* genome sequence, which together with the current RFLP markers and bin-mapped DArT markers will assist in defining the introgressions in the *S. pennelllii* IL lines. As described above, conversion of the DArT markers to CAPS or SNP markers will facilitate mapping of traits on a particular IL, especially in sub-IL populations.

The DArT platform is a robust, cost-effective and mature technology with standardized processing and analysis protocols, and thus can be implemented immediately by tomato breeders, much like its widespread adoption by wheat breeders (Akbari et al. 2006). Future work will include incorporating the DArT marker and sequence data into tomato genetic databases such as BreeDB (http://www.eu-sol.wur.nl/) and SGN (http://solgenomics.net).

Acknowledgments The research was supported by the European Commission (EU-SOL Project PL 016214) and the South African Department of Science and Technology (DST).

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

Agarwal S, Rao AV (2000) Tomato lycopene and its role in human health and chronic diseases. CMAJ 163:739–744

Akbari M, Wenzl P, Caig V, Carling J, Xia L, Yang S, Uszynski G, Mohler V, Lehmensiek A, Kuchel H, Hayden M, Howes N, Sharp P, Vaughan P, Rathmell B, Huttner E, Kilian A (2006) Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. Theor Appl Genet 113:1409–1420

Bai Y, Lindhout P (2007) Domestication and breeding of tomatoes: what have we gained and what can we gain in the future? Ann Bot 100:1085–1094

Butelli E, Titta L, Giorgio M, Mock HP, Matros A, Peterek S, Schijlen EGWM, Hall RD, Bovy AG, Luo J, Martin C (2008) Enrichment of tomato fruit with health-promoting anthocyanins by expression of select transcription factors. Nat Biotechnol 26:1301–1308

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden T (2009) BLAST+: architecture and applications. BMC Bioinform 10:421

Coetzer N, Gazendam I, Oelofse D, Berger DK (2010) SSHscreen and SSHdb, generic software for microarray based gene discovery: application to the stress response in cowpea. Plant Methods 6:10

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 19:11–15

Eshed Y, Zamir D (1995) An introgression line population of *Lycopersicon pennellii* in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. Genetics 141:1147–1162

Eshed Y, Abu-Abied M, Saranga Y, Zamir D (1992) *Lycopersicon esculentum* lines containing small overlapping introgressions from *L. pennellii*. Theor Appl Genet 83:1027–1034

Heller-Uszynska K, Uszynski G, Huttner E, Evers M, Carling J, Caig V, Aitken K, Jackson P, Piperidis G, Cox M, Gilmour R, D'Hont A, Butterfield M, Glaszmann JC, Kilian A (2011) Diversity arrays technology effectively reveals DNA polymorphism in a large and complex genome of sugarcane. Mol Breed 28:37–55

Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Res 29(4):e25 1–7

James KE, Schneider H, Ansell SW, Evers M, Robba L, Uszynski G, Pedersen N, Newton AE, Russell SJ, Vogel JC, Kilian A (2008) Diversity arrays technology (DArT) for pan-genomic evolutionary studies of non-model organisms. PLoS ONE 3:e1682

Jones N, Ougham H, Thomas H, Pasakinskiene I (2009) Markers and mapping revisited: finding your gene. New Phytol 183:935–966

Lezar S, Myburg AA, Berger DK, Wingfield MJ, Wingfield BD (2004) Development and assessment of microarray-based DNA fingerprinting in *Eucalyptus grandis*. Theor Appl Genet 109:1329–1336

Liu YS, Gur A, Ronen G, Causse M, Damidaux R, Buret M, Hirschberg J, Zamir D (2003) There is more to tomato fruit colour than candidate carotenoid genes. Plant Biotechnol J 1:195–207

Mace E, Xia L, Jordan D, Halloran K, Parh D, Huttner E, Wenzl P, Kilian A (2008) DArT markers: diversity analyses and mapping in Sorghum bicolor. BMC Genomics 9:26

Marshall JA, Knapp S, Davey MR, Power JB, Cocking EC, Bennett MD, Cox AV (2001) Molecular systematics of Solanum section Lycopersicum (Lycopersicon) using the nuclear ITS rDNA region. Theor Appl Genet 103:1216–1222

Miller NJ, Sampson J, Candeias LP, Bramley PM, Rice-Evans CA (1996) Antioxidant activities of carotenes and xanthophylls. FEBS Lett 384:240–242

Moose SP, Mumm RH (2008) Molecular plant breeding as the foundation for 21st Century Crop Improvement. Plant Physiol 147:969–977

Pan Q, Liu YS, Budai-Hadrian O, Sela M, Carmel-Goren L, Zamir D, Fluhr R (2000) Comparative genetics of nucleotide binding site-leucine rich repeat resistance gene homologues in the genomes of two dicotyledons: tomato and arabidopsis. Genetics 155:309–322

Peleg Z, Saranga Y, Suprunova T, Ronin Y, Röder MS, Kilian A, Korol AB, Fahima T (2008) High-density genetic map of durum wheat × wild emmer wheat based on SSR and DArT markers. Theor Appl Genet 117:103–115

Risterucci AM, Hippolyte I, Perrier X, Xia L, Caig V, Evers M, Huttner E, Kilian A, Glaszmann JC (2009) Development and assessment of diversity arrays technology for high-throughput DNA analyses in Musa. Theor Appl Genet 119:1093–1103

Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, Kilian A (2004) Diversity arrays technology (DArT) for whole-genome profiling of barley. Proc Natl Acad Sci USA 101:9915–9920

Wenzl P, Li H, Carling J, Zhou M, Raman H, Paul E, Hearnden P, Maier C, Xia L, Caig V, Ovesna J, Cakir M, Poulsen D, Wang J, Raman R, Smith K, Muehlbauer G, Chalmers K, Kleinhofs A, Huttner E, Kilian A (2006) A high-density consensus map of barley linking DArT markers to SSR, RFLP and STS loci and agricultural traits. BMC Genomics 7:206

White J, Law J, MacKay I, Chalmers K, Smith J, Kilian A, Powell W (2008) The genetic diversity of UK, US and Australian cultivars of *Triticum aestivum* measured by DArT markers and considered by genome. Theor Appl Genet 116:439–453

Wittenberg A, Lee T, Cayla C, Kilian A, Visser R, Schouten H (2005) Validation of the high-throughput marker technology DArT using the model plant *Arabidopsis thaliana*. Mol Gen Genomics 274:30–39

Xia L, Peng K, Yang S, Wenzl P, Carmen de Vicente M, Fregene M, Kilian A (2005) DArT for high-throughput genotyping of Cassava (*Manihot esculenta*) and its wild relatives. Theor Appl Genet 110:1092–1098

Yang S, Pang W, Ash G, Harper J, Carling J, Wenzl P, Huttner E, Zong X, Kilian A (2006) Low level of genetic diversity in cultivated Pigeonpea compared to its wild relatives is revealed by diversity arrays technology. Theor Appl Genet 113:585–595

Yeh M, Moysich KB, Jayaprakash V, Rodabaugh KJ, Graham S, Brasure JR, McCann SE (2009) Higher intakes of vegetables and vegetable-related nutrients are associated with lower endometrial cancer risks. J Nutr 139:317–322

Zamir D (2001) Improving plant breeding with exotic genetic libraries. Nat Rev Genet 2:983–989