Journal of Experimental Botany, Vol. 59, No. 10, pp. 2875–2890, 2008
doi:10.1093/jxb/ern146 Advance Access publication 13 June, 2008
This paper is available online free of all access charges (see http://jxb.oxfordjournals.org/open_access.html for further details)

Journal of Experimental Botany
www.jxb.oxfordjournals.org

RESEARCH PAPER

A candidate gene survey of quantitative trait loci affecting chemical composition in tomato fruit

L. Bermúdez[1,*], U. Urias[1,2,*], D. Milstein[1], L. Kamenetzky[2], R. Asis[3], A. R. Fernie[4], M. A. Van Sluys[1], F. Carrari[2,†] and M. Rossi[1,†]

[1] *GaTE Lab, Departamento de Botânica-IB-USP, Brasil. Rua do Matão, 277, 05508-900, São Paulo, SP, Brazil*

[2] *Instituto de Biotecnología, Instituto Nacional de Tecnología Agrícola (IB-INTA), PO Box 25, B1712WAA Castelar, Argentina (partner group of the Max Planck Institute for Molecular Plant Physiology, Potsdam-Golm, Germany)*

[3] *Facultad de Ciencias Químicas Universidad Nacional de Córdoba, CC 5000, Haya de la Torre y Medina Allende, Córdoba, Argentina*

[4] *Max Planck Institute for Molecular Plant Physiology, Wissenschaftspark Golm, Am Mühlenberg 1, Potsdam-Golm, D-14 476, Germany*

Received 18 February 2008; Revised 3 April 2008; Accepted 29 April 2008

Abstract

In tomato, numerous wild-related species have been demonstrated to be untapped sources of valuable genetic variability, including pathogen-resistance genes, nutritional, and industrial quality traits. From a collection of *S. pennellii* introgressed lines, 889 fruit metabolic loci (QML) and 326 yield-associated loci (YAL), distributed across the tomato genome, had been identified previously. By using a combination of molecular marker sequence analysis, PCR amplification and sequencing, analysis of allelic variation, and evaluation of co-response between gene expression and metabolite composition traits, the present report, provides a comprehensive list of candidate genes co-localizing with a subset of 106 QML and 20 YAL associated either with important agronomic or nutritional characteristics. This combined strategy allowed the identification and analysis of 127 candidate genes located in 16 regions of the tomato genome. Eighty-five genes were cloned and partially sequenced, totalling 45 816 and 45 787 bases from *S. lycopersicum* and *S. pennellii*, respectively. Allelic variation at the amino acid level was confirmed for 37 of these candidates. Furthermore, out of the 127 gene-metabolite co-locations, some 56 were recovered following correlation of parallel transcript and metabolite profiling. Results obtained here represent the initial steps in the integration of genetic, genomic, and expressional patterns of genes co-localizing with chemical compositional traits of the tomato fruit.

Key words: Candidate genes, introgressed lines, metabolite content, quantitative trait loci, *Solanum lycopersicum*, *Solanum pennelli*, tomato.

Introduction

Tomato (*Solanum lycopersicum* = *Lycopersicum esculentum*) is a horticultural crop of major economic importance, displaying several characteristics which have established it as a model system for dissection of genetic determinants of quantitative trait loci. In tomato, numerous wild-related species have been demonstrated to be untapped sources of valuable genetic variability, including pathogen-resistance genes, and nutritional and industrial quality traits (Fernie *et al*., 2006). Despite the fact that the tomato genome sequence is not yet complete, there is an extensive amount of genetic data on this species comprising relatively comprehensive genetic maps, expressed sequence tag (EST) collections, as well as precious germoplasm collections and mapping populations (including recombinant inbred and introgression lines), from which many quantitative trait loci (QTL) have already been reported

* These authors contributed equally to this work.
† To whom correspondence should be addressed. E-mail: magda1708@yahoo.com. Correspondence may also be addressed to F. Carrari. Email: fcarrari@cnia.inta.gov.ar

(Van der Hoeven *et al*., 2002; Mueller *et al*., 2005*a*; Lippman *et al*., 2007; Paran and Van der Knaap, 2007).

Historically in plant genetics, traits of interest have been genetically dissected through physical mapping followed by positional cloning (Salvi and Tuberosa, 2005). The advent of genomics and the increase of gene expression and mapping information that became available on its application have, however, recently facilitated the candidate gene approach. Following this approach the co-location of course map positions of genes with genomic regions conferring a trait of interest are regarded as 'candidates' that contribute, if not determine, changes in the trait (Tabor *et al*., 2002). Given that relatively few tomato QTL have been cloned or accurately tagged (see, for example, Frary *et al*., 2000; Fridman *et al*., 2004; Galpaz *et al*., 2006; Chen *et al*., 2007), and this is currently a laborious and slow process, requiring many generations of crossings and the screening of thousands of segregants, the candidate gene approach represents an attractive alternative as a way to start QTL characterization (Causse *et al*., 2004; Price, 2006). When studying populations resulting from inter-specific crosses the first step of this process is to identify co-location of course map position with trait variation associated with genomic regions harbouring QTL of interest. However, several further steps can be taken to support the candidacy of the genes in question. It is important to determine whether the genes are expressed in a spatial–temporal pattern that is consistent to that under which the QTL is detected. In addition, it is now relatively easy to determine whether the parental alleles differ in sequence identity or their level of expression.

In a recent study, Schauer *et al*. (2006) identified 889 fruit metabolic loci (QML) and 326 yield-associated loci (YAL) distributed across the tomato genome. These QTL were identified using the *S. pennelli* introgression line (ILs) population (Eshed and Zamir, 1995), that had previously been utilized by several groups to identify a further 1000 QTL (Lippman *et al*., 2007). However, despite producing an enormous amount of QTL data, the level of genetic resolution of these traits is currently somewhat limited since each IL harbours hundreds to thousands of genes, and, despite the availability of dense genetic maps for tomato, the number of metabolism-associated genes currently mapped is relatively low (in the region of 200–300). In a previous study by Causse *et al*. (2004), some 100 genes associated with primary metabolism were mapped and associations with fruit weight, and sugar and organic acid contents in fruits were examined. More recently, a map-based approach revealed few co-locations between candidate genes and QTL involved in the metabolism of ascorbic acid. Remarkable are the cases of the monodehydroascorbate reductase and the GDP-mannose epimerase genes that co-locate with two distinct QTL for ascorbic acid on chromosome 9 (Stevens *et al*., 2007). However, these studies not withstanding and the analysis of all genes associated with metabolism currently mapped failed to yield candidate genes for the vast majority of QML identified by Schauer *et al*. (2006).

In the current study, the aim was to provide a more comprehensive list of candidate genes following a slightly different strategy. Rather than taking the top–down approach of pre-selecting genes of interest and mapping their positions by means of multi-parallel Southern hybridizations, it was decided to identify all candidates within specific genomic regions of interest. The focus was on a subset of 106 QML and 20 YAL reported by Schauer *et al*. (2006), specifically those associated either with important agronomic or nutritional characteristics. It was possible to identify a total of 88 metabolism-associated and 39 non-metabolism (transport, signalling, protein processing or degradation, and DNA/RNA–protein metabolism) -associated candidate genes for these QTL. To validate these further, two additional experiments were performed: (i) sequence analysis of allelic variation between *S. lycopersicum* and *S. pennellii*; and (ii) evaluation of the correlation between the expression of these genes and the trait of interest within a dataset obtained from the assessment of tomato fruit development (Carrari *et al*., 2006). The combined results are discussed with respect both to the use of multiple association approaches and select sequencing for the cross-validation of candidate genes, and the ultimate utility of IL breeding in crop compositional improvement.

Materials and methods

QML selection and identification of candidate genes

All the molecular markers mapped onto the selected genomic regions (BINs: 1J, 2F, 4E, 4I, 5D/E/F, 7B, 7F, 7H, 9B/D/E, 9J, 10B, 11C), selected on the basis of the data presented in Schauer *et al*. (2006), were obtained from the Solanaceae Genomic Network (http://www.sgn.cornell.edu/). Marker sequences were compared by WU-BLAST algorithm (http://blast.wustl.edu) to the NCBI protein database (http://www.ncbi.nlm.nih.gov/). The pipeline designed for selection and analysis of candidate genes is shown in Fig. 1. The functions of selected gene products within metabolic pathways were predicted by mapping them using the KEGG database (http://www.genome.jp/kegg/; Kanehisa *et al*., 2008).

Plant material and DNA extraction

Seeds from 75 independent ILs, were kindly provided by CM Rick, Tomato Genetics Resource Center (TGRC). This resource is composed of a tomato variety, *Solanum lycopersicum* (inbred variety M82, Acc LA3475), which includes single introgressed genomic regions from the wild green-fruited species *Solanum pennellii* (LA716). Amongst the ILs there is a complete coverage of the wild-species genome. The ILs have been produced through successive introgression backcrossing and marker-assisted selection to generate a set of recurrent parent lines with single introgressed segments (Eshed and Zamir, 1995). Plants were grown in a greenhouse and DNA extraction was performed from fresh leaf material following the method described by Hoisington *et al*. (1994).

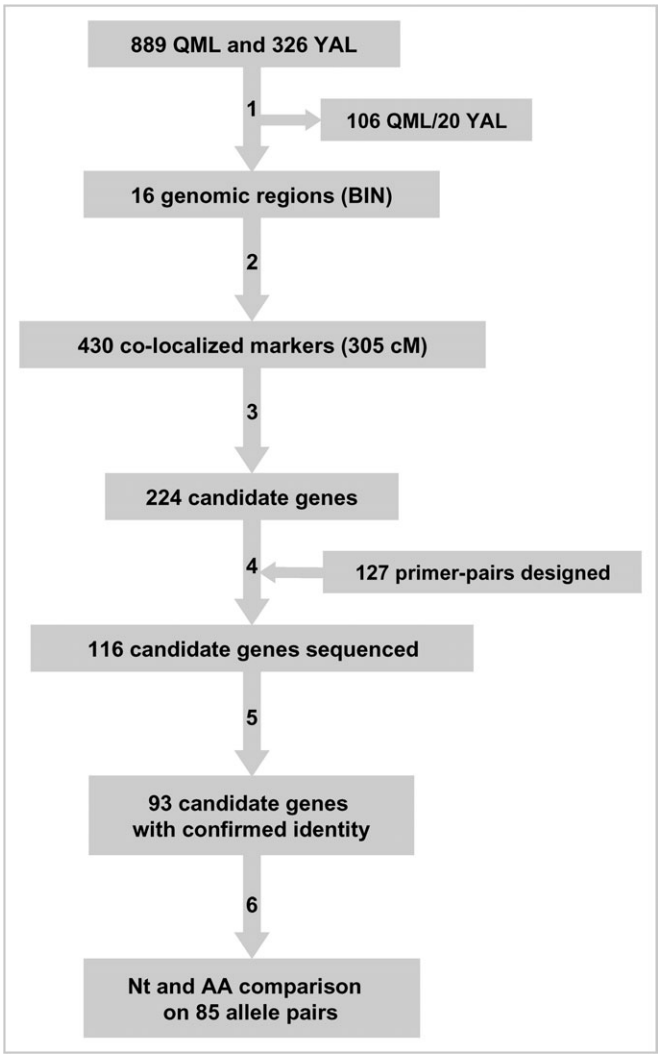

Fig. 1. QML selection and candidate genes identification pipeline. Schematic representation of the process designed to identify candidate genes co-localizing with previously detected QML onto tomato genomic regions. (1) At least 2-fold variation in metabolite content relative to *S. lycopersicum* and precise genome localization by at least two overlapped introgressed regions. (2) Retrieval of all mapped markers onto the selected genomic regions from the comparison between the Tomato-EXPEN2000, the Tomato-EXPEN1992, and the Tomato IL map by using the comparative map web interface from SGN (Mueller *et al*., 2008). (3) Sequence analysis by comparison with NCBI protein data base by using the Blastx algorithm. (4) Selection of complete *Solanum* cDNA sequences deposited onto SGN data repository or NCBI for primer design. PCR amplification and cloning from *S. lycopersicum* (M82 cultivar) and from the corresponding IL. End-sequencing of three independent clones from each genotype. (5) Sequence quality trimming and identity evaluation against the sequence used for primer design. (6) Identification of exons and introns by alignment with the corresponding sequence used for primer design. Allele comparison by identification of nucleotide and amino-acid polymorphisms. Output results from these analyses can be downloaded from URL: http://gracilaria.ib.usp.br/services/tomato/index.html.

Candidate gene amplification and cloning

Primers were designed with the Vector NTI 10.0 software package (Invitrogen) based on the unigene sequences available at the SGN (www.sgn.cornell.edu) or NCBI cDNA accessions (http://www.ncbi.nlm.nih.gov/) (Table S2 in Supplementary data available at *JXB* online). Candidate genes were amplified by PCR using Elongase® DNA polymerase (Invitrogen). The PCR reactions were performed using 0.2 mM of each dNTPs, 0.2 mM of each primer, 1.5 mM of $MgSO_4$, 100 ng of genomic DNA, and 2 units of enzyme. The PCR programme was 94 °C for 3 min; 35 cycles of 94 °C for 30 s, primer-specific annealing temperature for 30 s, 68 °C for 4 min; and a final period of 68 °C for 10 min. Amplification products were purified with GFX purification Kit (Amersham Biosciences) and cloned using the pMOSBlue blunt-ended cloning kit (Amersham Biosciences), following the manufacturer's instructions. Clones were end sequenced using vector universal primers, and reactions were read either with an ABI3700 or ABI3100 (Applied Biosystems).

Sequence and co-expression analyses

Vector sequences were trimmed using the VecScreen (www.ncbi.nlm.nih.gov/VecScreen/VecScreen.htm) software at the NCBI (www.ncbi.nlm.nih.gov). After quality trimming, all accepted sequences reached a Phred value ⩾20 (Gordon *et al*., 1998). Intron/exon prediction was performed by comparing the *S. pennelli* or *S. lycopersicum* sequences obtained with the corresponding unigene or marker sequence from SGN (www.sgn.cornell.edu), or NCBI cDNA accessions (http://www.ncbi.nlm.nih.gov/) using the Blast2 Sequences algorithm (Tatusova and Madden, 1999). Polymorphisms were detected at nucleotide and amino acid levels aligning *S. pennellii* and *S. lycopersicum* sequenced alleles (excluding primer regions) using the MULTALIN program (http://www-archbac.u-psud.fr/genomics/multalin.html; Corpet, 1988). The nucleotide diversity, which estimates the average number of substitutions between any two sequences, was determined using the software DNAsp version 4.10.9 (Rozas *et al*., 2003). The rate of synonymous and non-synonymous substitutions was determined using Nei and Gojobori's method (Nei and Gojobori, 1986) with the Jukes–Cantor correction, calculated using the MEGA 2.1 software (Kumar *et al*., 2001). Codon-based tests of selection (Fisher's exact test) were performed using the same software.

Developmental microarray expression data and metabolite data had been previously described in Carrari *et al*. (2006). In that study a combined analysis of metabolite and gene expression profiles from tomato fruits harvested through development and ripening stages (10, 15, 20, 21, 35, 49, 56, and 70 d after anthesis) was carried out. Although, the previous study reported extensive correlation analysis, this was performed in a targeted manner and did not include the candidate genes identified in the current study. For this reason, the expression data from 56 candidate genes, out of the 127 selected, which were spotted on the TOM1 microarray were correlated against the metabolite data of 66 metabolites determined in the ILs, using the Spearman algorithm (Urbanczyk-Wockniak *et al*., 2003).

Results and discussion

QML selection and identification of candidate genes

As a starting point for this study, we relied on the recent identification of 889 fruit metabolic loci (QML) and 326 yield-associated loci (YAL) in the *S. pennellii* IL population (Schauer *et al*., 2006). In order to select QML and identify candidate genes putatively responsible for those metabolite variations, a pipeline was established (Fig. 1). Out of those QML, 106 were selected based on the following criteria: they exhibited (i) at least 2-fold

variation in metabolite content relative to M82 variety of *S. lycopersicum* and (ii) a clear chromosomal position using the BIN mapping method. The selected QML were localized on 16 BINs (1J, 2F, 4E, 4I, 5D, 5E, 5F, 7B, 7F, 7H, 9B, 9D, 9E, 9J, 10B, 11C) across 8 of the 12 tomato chromosomes and comprised 52 different metabolites and nine different yield-associated traits. In addition, some QML for a range of traits were selected despite the fact that they did not fulfil the second criterion. Specifically, citrate, palmitate, stearate, fructose, GABA (γ-aminobutyric acid), glycine, tyrosine, and threonate QML (mapped onto chromosome 5), and phosphate and dehydroascorbate QML (mapped onto chromosome 9) could not be unambiguously defined to any of the BINs of these chromosomes. In these instances, candidate genes were grouped within BINs 5D/E/F and 9B/D/E for chromosomes 5 and 9, respectively (see Figs 3 and 4 and Table S1 in Supplementary data available at *JXB* online).

The selected regions carry a total of 430 mapped molecular markers present on the Tomato-EXPEN 2000 and Tomato-EXPEN 1992 maps (*S. lycopersicum* LA925×*S. pennellii* LA716) (http://www.sgn.cornell.edu/) spanning 305 cM. Sequences of the 430 available molecular markers, as well as previously described genes and cDNAs (Ganal *et al*., 1998; Causse *et al*., 2004; Zou *et al*., 2006), mapping onto the 16 selected genome regions were compared with the NCBI protein database. This survey resulted in a catalogue of 224 candidate genes (not shown) that presented sequence homology to previously characterized expressed sequences (reference proteins), whose functions have been experimentally demonstrated and could be involved in the observed metabolic changes. Out of these 224 putative genes, for 127 genes, it was possible to identify complete *Solanum* cDNA sequences (unigenes or markers from the Solanaceae Genome Network, or NCBI accessions) and to design primers that facilitated genomic-based PCR of a significant portion of the coding regions. Detailed information of these 127 candidates as well as the entire dataset of all 16 genomic regions studied is provided in Table S1 in Supplementary data available at *JXB* online. Identity between the *Solanum* cDNA sequences and the reference proteins varied between 32% and 100%.

The 127 candidate genes were positioned with respect to metabolic pathways (using the KEGG database) where their products are predicted to be involved, to visualize better their putative contributions to the described QML. Figures 2–5 provide an overview of the central metabolic pathways where each colour represents a selected genomic region, or BIN, with its corresponding QML and the candidate genes. For each gene, results of the amplification, cloning, and allele mining are also indicated. These genes were grouped, according to putative function, into six categories: carbon and nitrogen metabolism, transport, photosynthesis and oxidative phosphorylation, protein processing and degradation, DNA/RNA–protein metabolism, and signalling and regulation. The most abundant gene category was carbon and nitrogen metabolism (59%). This observation is

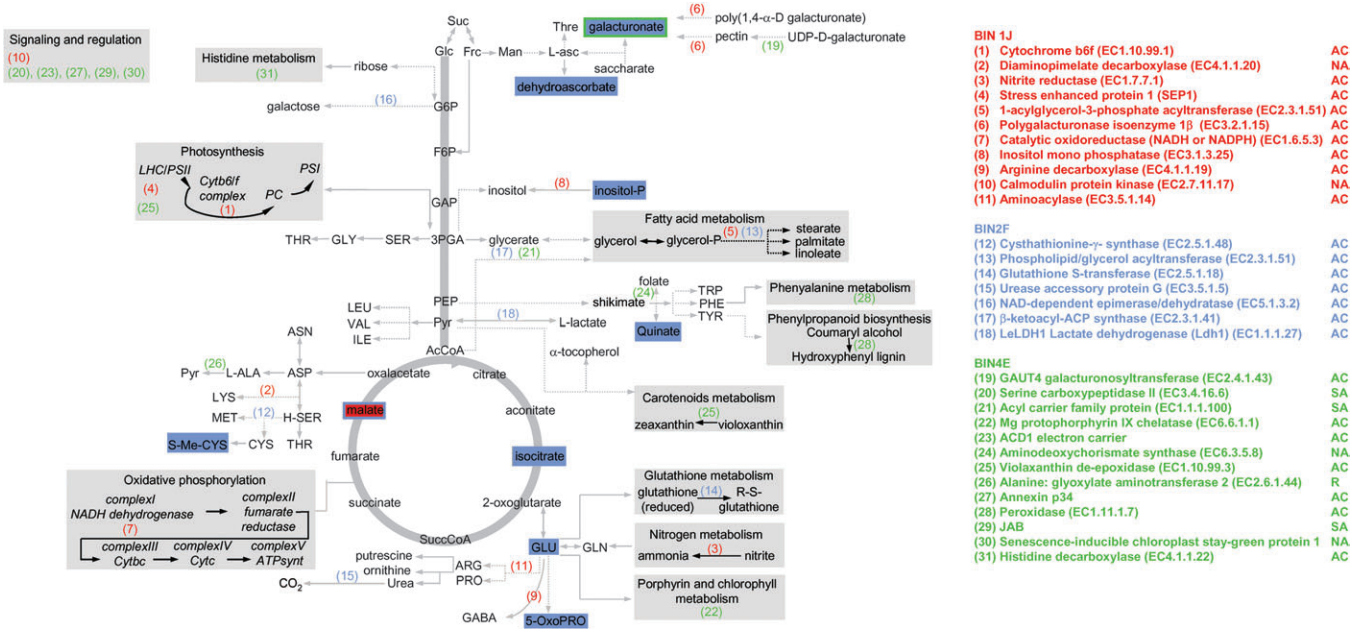

Fig. 2. Metabolic role of candidate genes in BINs 1J, 2F, and 4E. BINs are identified by colours. Candidate genes are identified by numbers and both metabolites and genes are highlighted in the corresponding BIN colour. The KEGG Accession Map Code and the results of the amplification, cloning, and allele mining are also indicated. NA, No amplification product; SA, spurious amplification product; R, sequence rearrangements; AC, alleles comparison (Table 1).

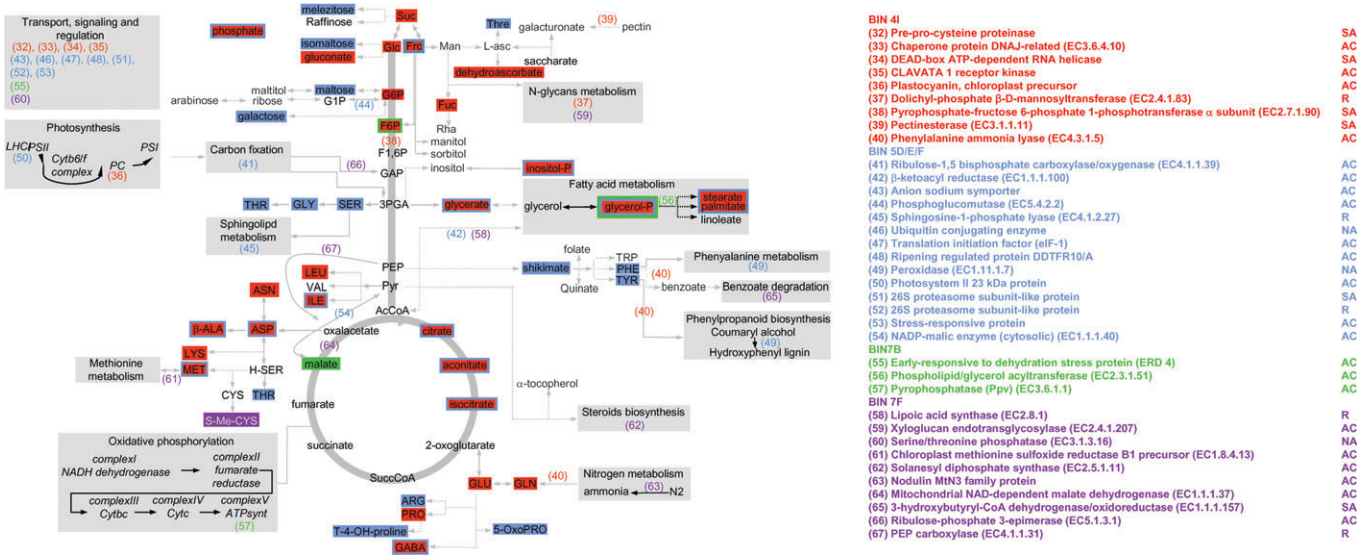

Fig. 3. Metabolic role of candidate genes in BINs 4I, 5D/E/F, 7B, and 7F. BINs are identified by colours. Candidate genes are identified by numbers and both metabolites and genes are highlighted in the corresponding BIN colour. The KEGG Accession Map Code and the results of the amplification, cloning, and allele mining are also indicated. NA, No amplification product; SA, spurious amplification product; R, sequence rearrangements; AC, alleles comparison (Table 1).

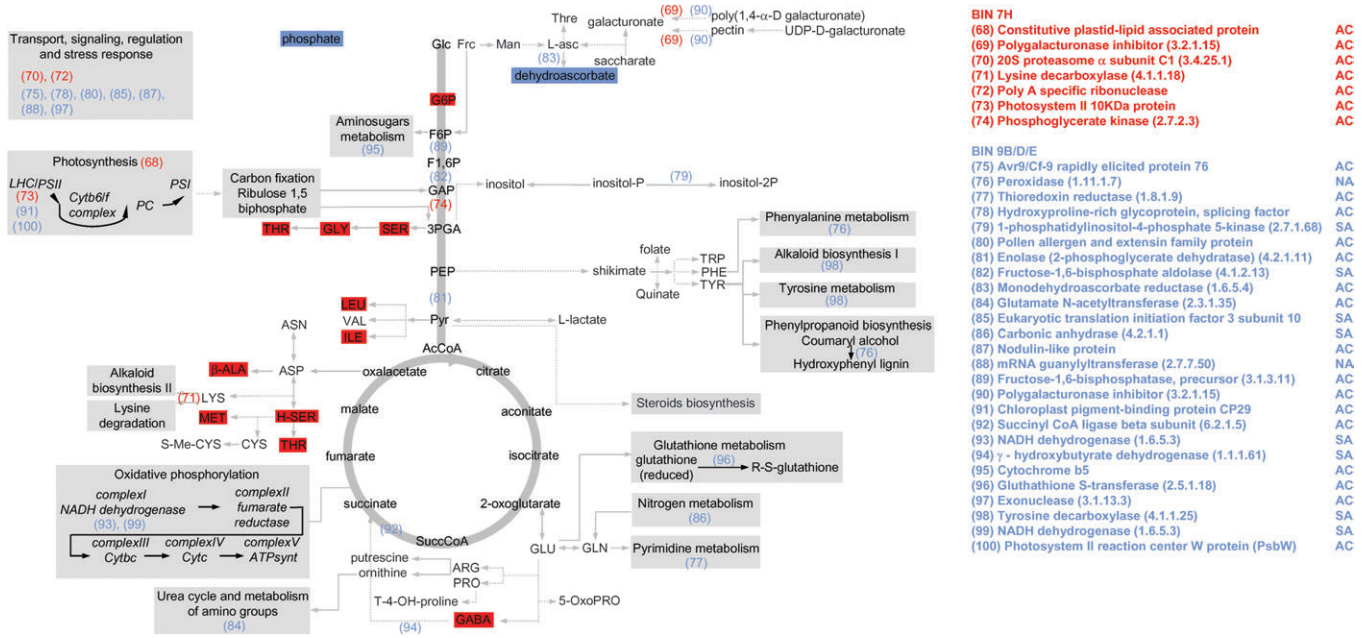

Fig. 4. Metabolic role of candidate genes in BINs 7H and 9B/D/E. BINs are identified by colours. Candidate genes are identified by numbers and both metabolites and genes are highlighted in the corresponding BIN colour. The KEGG Accession Map Code and the results of the amplification, cloning, and allele mining are also indicated. NA, No amplification product; SA, spurious amplification product; AC, alleles comparison (Table 1).

somehow predictable since regulatory factors control entire carbon and nitrogen metabolic networks. In the same way, transport effectors re-distribute products of those metabolic pathways. Within carbon and nitrogen metabolism, 23% corresponds to genes involved in amino acid metabolism and 24% to those implicated on central carbon metabolism. Only three candidates are genes related to nitrogen metabolism and the rest, 49%, distributed along different secondary pathways.

Candidate gene cloning and allele mining

Even though the candidature of some of the 127 identified genes is questionable in terms of the control they exert on the selected QML, given that metabolic variation within

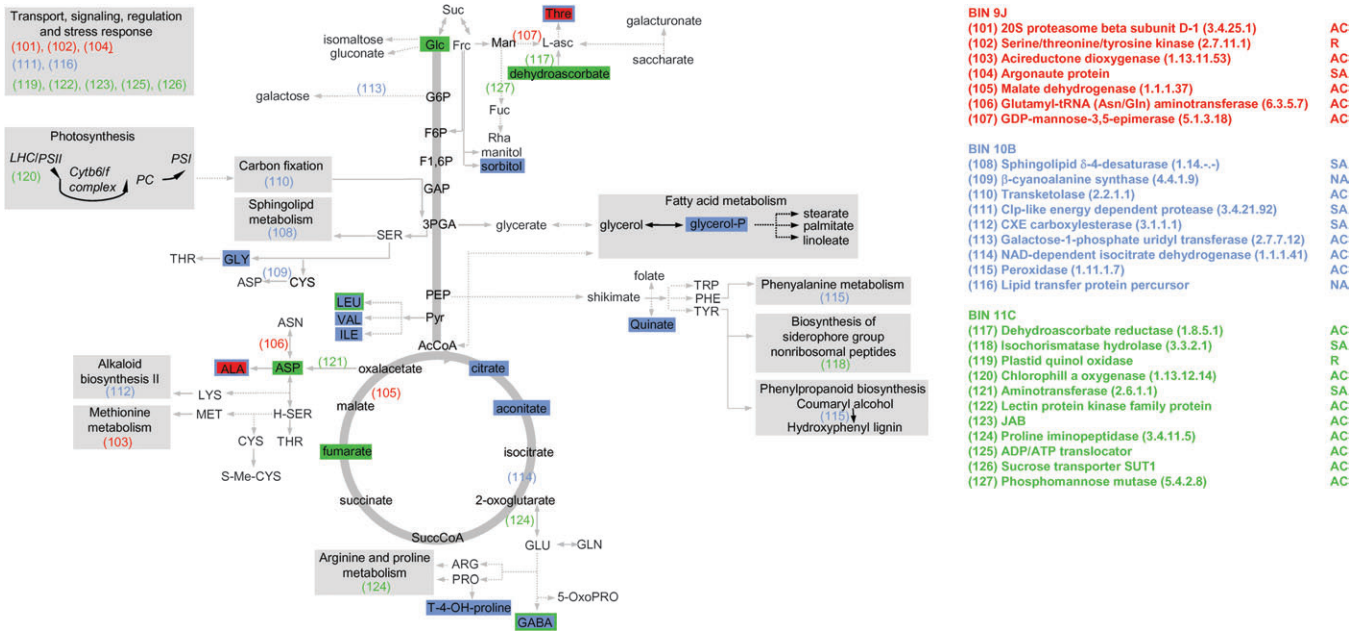

Fig. 5. Metabolic role of candidate genes in BINs 9J, 10B, and 11C. BINs are identified by colours. Candidate genes are identified by numbers and both metabolites and genes are highlighted in the corresponding BIN colour. The KEGG Accession Map Code and the results of the amplification, cloning and allele mining are also indicated. NA, No amplification product; SA, spurious amplification product; R, sequence rearrangements; AC, alleles comparison (Table 1).

the ILs is likely to arise from the *S. pennellii* introgressed genomic fragments, the comparative analysis of both alleles adds valuable information about polymorphisms between *S. lycopersicum* and *S. pennellii*. For this reason, a pair of primers was designed for each of the 127 candidate genes to amplify the alleles from the M82 variety and the corresponding IL (Table S2 in Supplementary data available at *JXB* online). PCR products were obtained for 116 pairs of alleles, with the remaining 11 genes being recalcitrant for amplification (Fig. 1). It is conceivable that the absence of amplification products of these alleles might be indicative of allele polymorphism, resulting in dominant molecular markers; however, this conclusion cannot be drawn from the present study alone. Larger genomic rearrangements also need to be considered in the chromosomal region encompassing the allele position.

The cloning and end-sequencing of three independent clones of each allele, enabled the confirmation of the identities of 93 pairs, while for 23 either one or both alleles did not present detectable homology to the sequence used for primer design and were considered as spurious amplification. Out of these 93 pairs, eight pairs were considered as possible rearrangements because, even when both alleles presented homology to the corresponding reference sequence, they did not overlap each other. After quality trimming, 85 pairs of genes were *in silico* spliced and translated, and the nucleotide and amino acid sequences of both alleles were compared (Table 1).

In order to provide the full information of these sequences together with the derived analyses, a database was created that can be accessed via a web interface (http://gracilaria.ib.usp.br/services/tomato/ILs.html). This resource allows both sequences and raw chromatograms, as well as the analyses of the results discussed in this paper, to be downloaded.

A total of 17 857 intron and 27 959 exon bases from *S. lycopersicum* and 17 974 intron and 27 813 exon bases from *S. pennellii* were sequenced. *In silico* translation of these sequences resulted in 9229 and 8994 protein amino acids from *S. lycopersicum* and *S. pennellii*, respectively. Out of those numbers, 15 261 intron bases, 23 716 exon bases, and 8007 protein amino acids overlap between both genotypes. The comparison between these overlapping regions revealed some interesting observations. The overall nucleotide polymorphism frequency was 4%, with an expected statistically significant greater variation in introns (7%) than in exons (1%) (Fisher's exact test $P < 0.05$). Most of the detected modifications corresponded to single nucleotide polymorphisms. INDELs (insertion/deletion) were found in 27 of the genes fragments analysed and almost all were located within intron regions (25 out of 27). Exon fragments were obtained and analysed for 81 out of the 85 genes amplified. From those, 56 contained nucleotide polymorphisms and 37 of these resulted in an amino acid change. Within each pair of alleles, a comparison between the ratio of non-synonymous (K_a) and synonymous (K_s) substitutions showed

Table 1. *Allele analysis of candidate genes from* S. lycopersicum *(Lyc) and* S. pennellii *(Pen)*

Marker (unigene)[a]	Size (b) exon/intron[b]	Nucleotide polymorphism (exon)[c]	Nucleotide polymorphism (intron)[d]	Amino acid coverage[e]	Analysed fragment[f]	Amino acid polymorphism[g]
(1) T0646 (U316058)	Lyc: 356/– Pen: 356/–	4/313	–	118/123	5–122	$T_{17}\rightarrow I$ $T_{29}\rightarrow P$ $K_{70}\rightarrow R$
(3) T1006 (U317524)	Lyc: 626/– Pen: 734/–	5/605	–	208/584	376–583	$E_{531}\rightarrow G$ $V_{569}\rightarrow I$
(4) C2_At4g34190 (U216629)	Lyc: 136/172 Pen: 93/467	1/93	6/128	31/141	16–46	$T_{32}\rightarrow P$
(5) CLET-1-A11 (U324336)	Lyc: 512/– Pen: 512/–	1/469	–	170/186	12–181	$I_{24}\rightarrow M$
(6) T1782 (U319301)	Lyc: 714/47 Pen: 585/68	9/585	2/47	194/405	16–209	$L_{92}\rightarrow H$ $S_{133}\rightarrow N$ $E_{151}\rightarrow G$ $Q_{157}\rightarrow R$ $A_{161}\rightarrow E$ $N_{185}\rightarrow D$
(7) C2_At4g34700 (U216646)	Lyc: 196/355 Pen: 264/315	1/196	20/320	58/119	1–58	–
(8) T1749 (U326864)	Lyc: 72/448 Pen:72/278	0/49	0/278	24/180	3–26	
(9) T1368 (U312881)	Lyc:459/– Pen:742/–	5/437	–	153/707	1–153	–
(11) T1306 (U319133)	Lyc: 749/– Pen: 614/–	3/609	–	202/448	36–237	$F_{186}\rightarrow L$ $D_{204}\rightarrow H$
(12) T0869 (AY508112[h])	Lyc: 335/300 Pen: 335/293	8/311	27/294	111/540	429–539	$P_{483}\rightarrow S$
(13) T1768 (U321585)	Lyc: 291/234 Pen: 291/352	0/267	7/234	96/189	93–188	–
(14) T1698 (U315881)	Lyc: 560/173 Pen: 523/173	4/504	7/173	174/367	32–217	$A_{50}\rightarrow S$ $V_{107}\rightarrow I$ $T_{118}\rightarrow M$
(15) C2_At2g34470 (U219076)	Lyc: 190/– Pen: 179/–	1/179	–	59/277	25–83	–
(16) T1516 (U317147)	Lyc:149/581 Pen:150/549	0/149	19/550	49/252	20–68	–
(17) cTOB-9-H18U315474	Lyc:349/276 Pen:441/137	0/349	0/137	116/469	36–151	–
(18) TC128325U326680	Lyc:459/– Pen: 534/39	0/459	–	153/350	35–187	–
(19) T0891 (U320717)	Lyc:290/35 Pen:303/267	0/278	1/35	95/679	585–679	–
(22) T0635 (U313864)	Lyc: 618/– Pen: 746/–	4/532	–	177/722	31–207	–
(23) T1054 (U319327)	Lyc: 512/75 Pen: 409/109	1/409	0/75	135/222	41–175	$H_{85}\rightarrow Y$
(25) T1317 (AK247081[h])	Lyc:465/131 Pen: 465/131	5/445	2/131	149/478	1–149	$H_{27}\rightarrow Q$ $F_{30}\rightarrow Y$
(27) C2_At1g35720 (U314161)	Lyc:465/142 Pen: 505/248	5/465	143/248	154/316	148–301	$K_{256}\rightarrow$ [h] $N_{288}\rightarrow S$
(28) T1719A (L1365[h])	Lyc: 567/152 Pen: 561/158	24/537	47/169	187/329	5–191	$C_{14}\rightarrow Y$ $V_{17}\rightarrow L$ $A_{20}\rightarrow V$ $I_{51}\rightarrow V$ $N_{53}\rightarrow K$ $A_{59}\rightarrow P$ $S_{85}\rightarrow R$ $V_{113}\rightarrow L$
(31) T0883 (U313818)	Lyc: 540/70 Pen: 556/27	31/540	0/27	179/413	228–406	$S_{249}\rightarrow P$ $G_{251}\rightarrow D$ $V_{252}\rightarrow I$ $R_{257}\rightarrow K$ $S_{258}\rightarrow T$ $L_{263}\rightarrow H$ $A_{272}\rightarrow T$ $L_{292}\rightarrow I$ $T_{333}\rightarrow S$

Continued

Table 1. *Continued*

Marker (unigene)[a]	Size (b) exon/intron[b]	Nucleotide polymorphism (exon)[c]	Nucleotide polymorphism (intron)[d]	Amino acid coverage[e]	Analysed fragment[f]	Amino acid polymorphism[g]
						$R_{346}\rightarrow H$ $I_{361}\rightarrow V$ $N_{382}\rightarrow K$ $Y_{383}\rightarrow –$ $K_{384}\rightarrow R$ $Y_{386}\rightarrow F$ $D_{388}\rightarrow Y$ $V_{389}\rightarrow G$ $A_{391}\rightarrow T$ $L_{392}\rightarrow Q$
(33) T0739 (U321142)	Lyc: 140/393 Pen:140/424	1/118	27/403	44/146	4–47	$K_{11}\rightarrow R$
(35) cLEW-8-J19 (U324703)	Lyc: 431/170 Pen: 431/140	4/412	13/151	121/285	165–285	$V_{241}\rightarrow I$
(36) cLET-5-D13 (U312690)	Lyc:427/– Pen: 379/–	3/379	–	124/170	35–158	–
(40) LED50 (LED50[h])	Lyc: 728/– Pen: 632/–	0/611	–	210/704	485–694	–
(41) T0778 (U317221)	Lyc: 411/209 Pen: 467/54	0/383	0/54	127/488	33–159	–
(42) T1174 (U321882)	Lyc: 536/18 Pen: 208/–	0/208	–	69/234	12–80	–
(43) T0328 (U315874)	Lyc: 118/6 Pen: 241/157	0/93	0/6	39/407	2–40	–
(44) T1601 (U333333)	Lyc: 473/25 Pen: 473/41	4/451	0/25	157/191	17–173	$S_{50}\rightarrow G$ $T_{96}\rightarrow A$ $R_{108}\rightarrow G$
(47) cTOS-7-03 (U314198)	Lyc: 175/446 Pen: 175/300	3/148	90/354	58/145	85–142	$V_{124}\rightarrow D$
(48) cLEX-13-G5 (U315595)	Lyc: 588/– Pen: 710/–	1/316	–	105/314	104–208	$M_{194}\rightarrow V$
(50) T0837 (U312572)	Lyc: 404/134 Pen: 124/39	0/124	0/39	41/258	37–77	–
(53) C2_At3g17210 (U214933)	Lyc:142/294 Pen:142/410	6/122	18/302	47/106	2–48	$E_{18}\rightarrow K$
(54) cLES-1-A11 (U312789)	Lyc: 459/350 Pen: 432/352	4/432	13/352	141/579	438–578	$V_{503}\rightarrow M$
(55) T1355 (U323609)	Lyc: 300/239 Pen: 272/131	0/272	0/131	73/312	28–100	–
(56) C2_At4g30580 (U229764)	Lyc: 25/514 Pen: –/557	–	68/514	–/284	–	–
(57) cLER–17P11 (U313426)	Lyc: 390/383 Pen: 467/239	5/390	10/239	129/765	83–211	–
(59) C2_At4g03210 (DQ098654[h])	Lyc: 169/228 Pen: 102/316	1/102	4/162	34/266	24–57	–
(61) C2_At1g53670 (U216219)	Lyc: 169/231 Pen: 75/317	1/75	6/231	24/189	33–56	$S_{34}\rightarrow R$
(62) T1624 (T1624[h])	Lyc: 285/136 Pen: 285/274	2/262	4/138	94/398	3–96	–
(63) C2_At3g14770 (U231080)	Lyc: 363/222 Pen:240/209	9/217	5/209	37/235	199–235	–
(64) T1171 (U313128)	Lyc: 247/338 Pen:247/345	1/226	13/338	82/345	5–86	–
(66) cLET-14-A10 (U313308)	Lyc: 148/419 Pen: 148/306	0/127	0/306	39/282	244–282	–
(68) T0966 (U313029)	Lyc: 249/437 Pen: 191/411	0/191	1/411	63/192	25–87	–
(69) T1255 (U315727)	Lyc: 427/– Pen: 726/–	1/415	–	138/327	60–201	–
(70) cLEX-13-I15 (U316193)	Lyc:597/– Pen:543/–	0/528	–	175/224	41–215	–
(71) C2_At1g50575 (U222777)	Lyc:218/220 Pen:241/473	1/218	8/220	62/202	115–176	–

Continued

Table 1. *Continued*

Marker (unigene)[a]	Size (b) exon/intron[b]	Nucleotide polymorphism (exon)[c]	Nucleotide polymorphism (intron)[d]	Amino acid coverage[e]	Analysed fragment[f]	Amino acid polymorphism[g]
(72) C2_At1g55870 (U228097)	Lyc: 481/– Pen: 312/–	23/291	–	104/355	255–354	$H_{267}\rightarrow Y$ $R_{309}\rightarrow G$ $-_{315}\rightarrow V$ $-_{315}\rightarrow C$ $-_{315}\rightarrow V$ $-_{315}\rightarrow E$ $R_{320}\rightarrow S$ $N_{323}\rightarrow D$ $I_{330}\rightarrow M$
(73) CT223 (U143214)	Lyc:100/326 Pen:153/340	1/100	44/311	32/138	20–51	–
(74) cLEB-3-N22 (U313176)	Lyc:415/45 Pen:415/160	3/394	0/45	138/482	2–140	$T_{47}\rightarrow A$ $V_{64}\rightarrow L$
(75) cLEX-3-N24 (U3208109)	Lyc: 660/– Pen: 415/–	11/415	–	138/251	11–148	$K_{20}\rightarrow N$ $C_{74}\rightarrow F$ $L_{83}\rightarrow F$ $V_{100}\rightarrow L$ $D_{115}\rightarrow E$ $N_{120}\rightarrow Y$
(77) C2_At2g41680 (U221908)	Lyc: 248/362 Pen: 248/362	0/248	0/362	82/256	12–93	–
(78) C2_At2g32600 (U218453)	Lyc: 266/207 Pen: 332/371	3/245	15/214	87/252	155–241	$T_{217}\rightarrow I$
(80) T1673 (U327399)	Lyc: 109/319 Pen: 82/60	0/82	25/84	27/173	27–53	–
(81) T0532 (U312379)	Lyc: 255/289 Pen: 254/287	1/232	14/290	82/444	353–434	–
(83) cLET-3-C15 (U315877)	Lyc: 299/182 Pen: 299/80	1/299	2/81	99/433	328–426	$P_{416}\rightarrow A$
(84) C2_At2g37500 (U231168)	Lyc: 134/363 Pen: 134/454	0/112	1/361	44/234	217–233	–
(87) T1617 (U321884)	Lyc: 334/358 Pen: 348/340	6/328	14/345	110/388	273–382	$V_{309}\rightarrow I$ $P_{366}\rightarrow L$ $S_{377}\rightarrow L$
(89) T1212 (U316424)	Lyc: 282/295 Pen: 380/232	0/282	0/231	93/403	45–137	–
(90) cLET-2-D4 (U315727)	Lyc: 556/– Pen: 442/–	2/322	–	106/327	96–201	$A_{101}\rightarrow T$
(91) cLET-7-N21 (U312661)	Lyc: 241/– Pen: 384/144	2/241	–	80/285	38–117	–
(92) T0443 (U315467)	Lyc:105/9 Pen: 229/339	1/105	0/9	34/421	76–109	–
(95) T1785 (U318473)	Lyc: 199/328 Pen: 180/303	29/179	186/328	59/137	49–107	$D_{76}\rightarrow E$ $A_{80}\rightarrow S$ $K_{85}\rightarrow S$ $T_{86}\rightarrow V$ $Q_{95}\rightarrow H$ $S_{102}\rightarrow T$ $V_{105}\rightarrow I$ $V_{106}\rightarrow I$
(96) cLEX-13-I3 (U324385)	Lyc: 318/246 Pen: 322/243	0/236	0/243	65/229	42–106	–
(97) cTOA-30-C21 (U327971)	Lyc: 22/425 Pen: 22/374	–	109/374	–	–	–
(100) T0556 (U314531)	Lyc: 269/496 Pen: 269/381	1/246	1/381	89/132	32–120	$R_{51}\rightarrow K$
(101) cLET-7-D17 (U316001)	Lyc: 312/284 Pen: 312/351	0/291	1/284	102/198	89–191	–
(103) cLET-42–02 (U313367)	Lyc: 263/240 Pen: 182/239	1/160	17/240	59/200	142–200	–
(105) T1190 (U312385)	Lyc: 192/602 Pen: 190/463	0/97	21/448	32/583	271–302	–
(106) T1519 (U332457)	Lyc: 455/131 Pen: 505/–	5/230	–	76/219	50–125	$G_{79}\rightarrow V$

Continued

Table 1. *Continued*

Marker (unigene)[a]	Size (b) exon/intron[b]	Nucleotide polymorphism (exon)[c]	Nucleotide polymorphism (intron)[d]	Amino acid coverage[e]	Analysed fragment[f]	Amino acid polymorphism[g]
(107) cTOF-18-B12 (BG128005[h])	Lyc: 262/439 Pen: 254/315	1/254	9/316	84/219	54–137	$V_{77} \rightarrow A$
(110) cLES-2-K4 (U312319)	Lyc: 312/16 Pen: 258/–	0/258	–	85/760	77–161	–
(113) T1164 (U320574)	Lyc: 397/344 Pen: 223/344	1/222	13/344	73/340	237–309	$Y_{284} \rightarrow F$
(114) T0308 (U316154)	Lyc: 230/138 Pen: 350/138	1/218	0/138	76/373	257–332	–
(115) cLEY-13-H6 (U315415)	Lyc: 585/150 Pen: 603/150	4/565	4/150	200/300	21–220	$N_{164} \rightarrow D$
(117) C2_At5g16710 (U214041)	Lyc: 89/452 Pen: 89/263	1/68	20/267	28/268	241–268	$E_{246} \rightarrow D$
(120) C2_At1g44446 (U220686)	Lyc: 29/560 Pen: 29/562	–	32/562	9/461	8–16	–
(122) cLEX-4-G10 (U346954)	Lyc: 681/– Pen: 658/–	11/634	–	219/233	14–233	$A_{75} \rightarrow V$ $N_{82} \rightarrow D$ $P_{87} \rightarrow Q$ $Y_{119} \rightarrow C$
(123) cTOE-7-B4 (U315480)	Lyc: 171/488 Pen: 171/354	0/151	13/354	54/367	313–366	–
(124) C2_At2g14260 (U220663)	Lyc: 24/613 Pen: 24/634	–	0/613	7/380	1–7	–
(125) CT55 (U143394)	Lyc: 561/110 Pen: 303/–	1/303	–	101/386	36–136	$H_{55} \rightarrow Q$
(126) cLED-7-H11 (U315661)	Lyc: 147/252 Pen: 147/381	1/126	42/269	48/511	455–502	–
(127) cLEC-68-J21 (BI421979[h])	Lyc: 182/185 Pen: 209/204	0/182	0/185	60/241	171–230	–

[a] Marker and unigene according to the Sol Genomics Network (www.sgn.cornell.edu). Genes are numbered according to Figs 2–5 and Table S1 (in Supplementary data available at *JXB* online).
[b] Total number of trimmed bases for each genotype, exon/intron.
[c] Number of nucleotides along the exon showing polymorphisms between genotypes/total of exon bases compared (primer sequences were not considered, a dash means no exon fragment sequenced).
[d] Number of nucleotides along the intron showing polymorphisms between genotypes/total of intron bases compared (a dash means no intron fragment compared).
[e] Number of compared amino acids between alleles/total number of amino acids of the corresponding unigene translated protein.
[f] Analysed amino acid interval of the corresponding translated unigene.
[g] Polymorphic amino acids between amplified alleles. The numbers indicate the position of changes corresponding to the translated unigene. When there is no number it means that there is a frame shift between the predicted proteins for Lyc and Pen and the unigene protein. A dash means insertion or deletion.
[h] When there was no unigene, or the unigene was uncompleted, the sequence used for the analysis was taken from the GenBank (NCBI accession number) or the marker sequence according to Sol Genomic Network (www.sgn.cornell.edu).

values lower than 1 for 51 out of the 56 polymorphic genes. For only the eight following genes, out of the 51, the ratio was statistically significant ($P < 0.05$): arginine decarboxylase (gene 9) on BIN 1J; cystathionine-γ-synthase (gene 12) on BIN 2F; Mg-protophorphyrin IX chelatase (gene 22) and peroxidase (gene 28) both located on BIN 4E; pyrophosphatase (Ppv) (gene 57) on BIN 7B; poly(A)-specific ribonuclease (gene 72) on BIN 7H; cytochrome b_5 (gene 95) on BINs 9B/D/E; and lectin protein kinase family protein (gene 122) on BIN 11C. Although caution should be taken in order not to over-interpret these results, it is tempting to speculate the occurrence of purifying selection against non-synonymous substitutions in these genes indicative of a functional requirement for their products.

The analysis of the sequence divergence between *S. pennellii* and *S. lycopersicum* alleles across different candidate categories (Table 2) showed that the largest number of genes with polymorphisms resulting in changes at amino acid level were those belonging to signalling and regulation (seven out of nine), DNA/RNA–protein metabolism (three out of three), and transport (three out of five) categories. By contrast, those related to central carbon metabolism (3 out of 14), protein processing and degradation (one out of four), and photosynthesis and oxidative phosphorylation (3 out of 10) displayed only a few genes with amino acid changes. The rest of the categories presented intermediate numbers of polymorphism at the level of a protein amino acid sequence. Whilst it is important to point out that amino acid position, which is an important component, was not considered here. The observed trends are largely in accordance with results reported by Schauer *et al*. (2006). In this study, it had been noted that a large proportion

Table 2. *Distribution of candidate genes between metabolic categories*

n, Total number of genes in each category according to the 127 candidates identified.
p/np, Number of genes that presented amino acid polymorphisms on the analysed fragment sequence/number of genes that did not present amino acid polymorphisms on the fragment sequence analysed. In this case, the total is the 81 genes for which amino acid sequences were analysed.

BIN (total candidates)	Carbon and nitrogen metabolism					Transport	Photosynthesis and oxidative phosphorylation	Protein processing and degradation	DNA/RNA/ protein metabolism	Signalling and regulation	Total
	n *p/np*					*n* (%) *p/np*	*n* (%) *p/np*	*n* (%) *p/np*	*n* (%) *p/np*	*n* (%) *p/np*	*n* *p/np*
	Amino acids	Central carbon	Nitrogen	Others (secondary metabolism)	Total (%)						
1J (11)	3 1/1	–	1 1/–	3 2/1	7 (64) 4/2	–	2 (18) 1/1	–	–	2 (18) 1/–	11 6/3
2F (7)	2 1/1	1 –/1	–	4 1/3	7 (100) 2/5	–	–	–	–	–	7 2/5
4E(13)	2 1/–	–	–	5 ½	7 (54) 2/2	1 (8) 1/–	1 (8) 1/–	2 (15)	–	2 (15) 1/–	13 5/2
4I (9)	1 –/1	1	–	2	4 (44) –/1	–	1 (11) –/1	2 (22) 1/–	1 (11)	1 (11) 1/–	9 2/2
5D/5E/5F (14)	–	3 2/1	–	3 –/1	6 (43) 2/2	1 (7) –/1	1 (7) –/1	3 (21)	1 (7) 1/–	2 (14) 2/–	14 5/4
7B (3)	–	–	–	1	1 (33)	–	1 (33) –/1	–	–	1 (33) –/1	3 –/2
7F (10)	1 1/–	3 –/2	1 –/1	4 –/2	9 (90) 1/5	–	–	–	–	1 (10)	10 1/5
7H (7)	1 –/1	1 1/–	–	1 –/1	3 (43) 1/2	–	2 (29) –/2	1 (14) –/1	1 (14) 1/–	–	7 2/5
9B/9D/9E (26)	2 –/1	5 –/3	1	7 3/2	15 (58) 3/6	1 (4) 1/–	4 (15) 1/1	–	4 (15) 1/–	2 (8) 1/1	26 7/8
9J (7)	2 1/1	1 –/1	–	1 1/–	4 (57) 2/2	–	–	1 (14) –/1	–	2 (29)	7 2/3
10B (9)	1	2 –/2	–	4 1/1	7 (78) 1/3	1 (11)	–	1 (11)	–	–	9 1/3
11C (11)	2	1 –/1	–	2 1/–	5 (45) 1/1	2 (18) 1/1	1 (9)	1 (9) –/1	–	2 (18) 1/–	11 3/3
Total n p/np	17 5/6	18 3/11	3 1/1	37 10/13	75 19/31	6 3/2	12 3/7	12 1/3	7 3/–	15 7/2	127 81

of the fruit QML were strongly associated with variation in yield-associated traits (Table S1in Supplementary data available at *JXB* online), in particular with the harvest index which is obviously closely related to assimilate partitioning. Thus, one could rationalize that allelic variations on genes of the first groups (signalling and regulation, DNA/RNA–protein metabolism, and transport) may well play a more major role affecting the final fruit metabolite content than those of the second group (central carbon metabolism, protein processing and degradation, photosynthesis, and oxidative phosphorylation). It should be borne in mind, however, that the failure in the present study to detect polymorphism between *S. pennellii* and *S. lycopersicum* alleles does not preclude the candidacy of the genes for two reasons: (i) since only partial sequences were analysed it cannot be excluded that the alleles were polymorphic in the non-sequenced regions of their reading frames; and (ii) because regulatory sequences, upstream of the amplified coding region, could be responsible for differential expression levels or pattern of the alleles.

Co-response and integrative analyses

The evaluation of the co-response pattern of transcription in relation to the variations in metabolite contents of interest supports the candidacy of the selected genes and may provide hints about epistatic interactions of the candidates identified with QML localized in other BINs. Then, a correlation analysis was performed between the expression profile of the candidates and the metabolite variations along fruit development and ripening in *S. lycopersicum*. Expression data of 56 of the selected candidate genes that were present on the TOM1 micro-array were correlated against the content variation of 66 metabolites quantified across a fruit development and

ripening time course (Carrari *et al*., 2006). Out of the 3696 pairs analysed, 724 positive (blue) and 307 negative (red) significant correlations were observed (Fig. 6). This number of correlations is well above of that expected merely by chance (185 at $P < 0.05$).

In the following section, only those candidate genes, which it was possible to provide supporting evidence from metabolic mapping, sequence analysis, and their correlative behaviour from the developmental time series experiment, will be dealt with in detail.

The gene encoding cystathionine-γ-synthase (gene 12), localized on BIN 2F, correlated with the contents of many metabolites including several for which QML mapped to this BIN (correlating positively with malate, quinate, and inositol-P; and negatively with galacturonate and dehydroascobate). Allelic variation, at amino acid level, between *S. lycopersicum* and the corresponding *S. pennellii* introgressed line was also found (Table 1). This enzyme participates in the conversion of homo-Ser to Met (Fig. 2) and could be involved in the variation of S-Me-Cys found in this BIN. A role for this pathway during tomato fruit ripening has been assessed recently by Katz *et al.* (2006). The massive production of ethylene during ripening requires an increase in the *de novo* Met synthesis through up-regulation of this enzyme. In the present correlation analysis, mRNA levels of this gene correlate negatively with Cys, a precursor in the biosynthesis of S-Me-Cys (Fig. 6). In addition, a glutathione *S*-transferase encoding gene (gene 14) was found co-locating with variations in Glu and 5-OxoPro QML onto this BIN (Fig. 2). Together with the amino acid

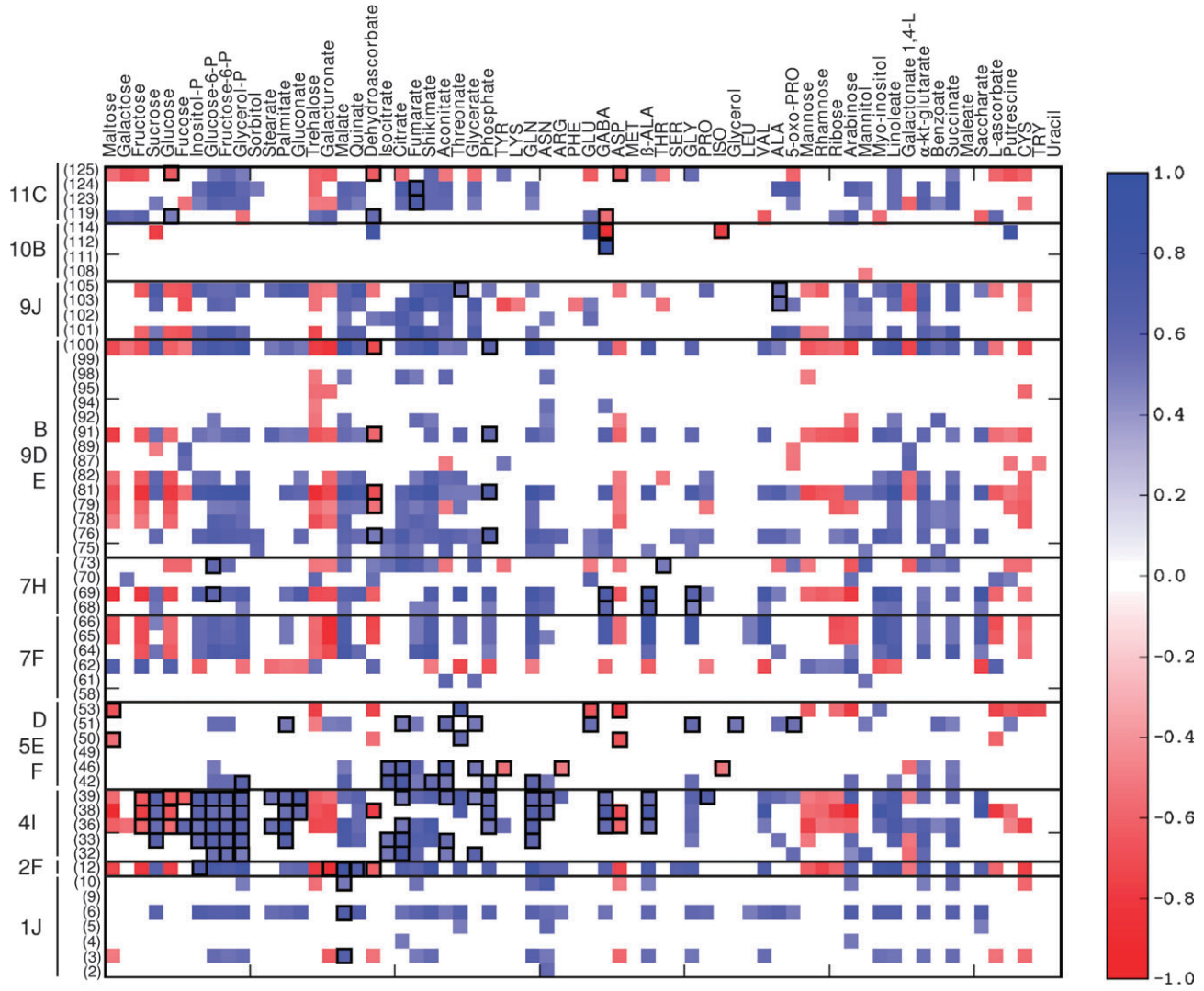

Fig. 6. Correlations between candidate gene transcription profile and metabolite contents through *S. lycopersicum* fruit development. Correlation coefficients (two tailed) and significances ($P < 0.05$) were calculated by applying Spearman algorithm using SSPS software. Each dot indicates a given *r*-value, resulting from a Spearman correlation analysis, in a false colour scale. Blue and red represent significant positive and negative correlations, respectively; white indicates a lack of significant correlation. The genes are indicated with the same number order as in Figs 2–5. Dots demarcated by a bold border indicate those that exhibit significant correlations between a given gene and the metabolite corresponding to the QML to which it co-localizes.

changes found in the coding region of this gene (Table 2), results mentioned above make it a good candidate to test using functional approaches.

The most supported candidate gene mapped on BIN 4E is GAUT4 galacturonosyltransferase (gene 19, Fig. 2), which co-localized with galacturonate QML, since this enzyme participates in pectin biosynthesis from galacturonate (Sterling *et al*., 2006). However, no polymorphisms at the amino acid level were detected on the protein fragment analysed.

The introgressed region delineated as BIN 4I harbours QML for elevated sugars, as well as elevations in the metabolites belonging to the pathway linking citrate to glutamate, and was, therefore, defined as a pathway QTL (Schauer and Fernie, 2006). Moreover, about 60% of all QML mapped onto this BIN have been defined as morphology-dependent QML as they could be associated with phenotypic traits by correlation analyses (Table S1 in Supplementary data available at *JXB* online). In fact, this pathway QTL showed significant association with plant weight, *Brix* levels, fruit width, and harvest index.

Genes mapped onto this BIN for which it would be possible to evaluate transcriptional behaviour (gene 32, pre-pro-cysteine proteinase; gene 33, chaperone protein DnaJ-related; gene 36, plastocyanin chloroplast precursor; gene 38, pyrophosphate-fructose 6-phosphate 1-phosphotransferase α-subunit; and gene 39, pectinesterase) displayed a wide range of significant correlations with most of the co-localizing QML including sugars, phosphates, and amino acids. Their co-localization, together with the similar patterns of correlation they showed, might be indicative of a coordinated mechanism of regulation operating at this same position on the genome. Phenylalanine ammonia lyase (gene 40) is involved in a wide range of metabolic pathways (Fig. 3) including nitrogen releases of NH_4^+ for Gln and Glu biosynthesis. Thus, this enzyme, previously mapped by Causse *et al*. (2004) onto the same BIN, may be involved in the variations of these two amino acids observed in *S. pennellii* IL. However, no amino acid polymorphisms were detected in the gene fragment analysed here (Table 1) and further investigations are needed to evaluate the candidature of this gene. When looking for the genetic determinants for these pathway QTL there are two alternatives: either (i) is not controlled by variation at a single genetic locus or (ii), more likely, the gene responsible for the entire pathway variation encodes not an enzyme activity protein but a regulatory one. In this direction, a chaperone protein DnaJ-related (gene 33) mapped onto this BIN and classified within regulatory categories emerged as an interesting candidate. DnaJ-chaperone proteins constitute a wide family both in prokaryote and eukaryote organisms and participate in protein folding, assembly, disassembly, and translocation into organelles through a mechanism involving interaction with Hsp70 chaperones (Qiu *et al*., 2006). It has been demonstrated recently that a mutation on a cauliflower locus encoding a member of this family (*Or* locus) leads to chromoplast differentiation and consequently a deposit of β-carotene in the affected tissues (Lu *et al*., 2006). Moreover, transgenic potato plants over-expressing this allele in a tuber-specific manner result in the production of orange-yellow tubers associated with high levels of carotenoids (Lu *et al*., 2006). The chaperone protein DnaJ-related described here also possesses the Cys-rich zinc finger domains characteristics of DnaJ chaperones (not shown) and an amino acid polymorphism in the coding region ($K_{11} \rightarrow R$). Additionally, the levels of β-carotene in the fruit correlated positively (r=0.76; P <0.0001) with the expression of this gene during the developmental time-series experiment. These results, when taken together, make this gene a good candidate to be tested functionally for its putative role in the control of fruit metabolism.

The other BINs exhibiting patterns of metabolite and whole-plant phenotypic variations similar to those described for 4I are 5D, E, and F (Fig. 3). For reasons explained above, these last three BINs are here considered as a single entity. Except for the case of the peroxidase (gene 49), the other five candidates mapped on BIN 5D/E/F for which a transcriptional pattern was analysed, correlate with three to seven QML localized onto these BINs. β-Ketoacyl reductase (gene 42) plays a key role in fatty acid biosynthesis and positively correlates with glycerol-P, rendering it indirectly linked to stearate and palmitate QML. Phosphoglucomutase (gene 44) has long been considered a key enzyme in starch biosynthesis of potato tubers. Although the role of this enzyme has not been directly assessed in tomato fruits, it has been demonstrated that its activity declines during early developmental stages in accordance with the expression level of its cytosolic isoform (Kortstee *et al*., 2007). The co-localization of this gene with a YAL for *Brix* variation, and QML for maltose and galactose, renders it a good candidate to follow-up. Whilst, starch in its own right is not a highly important quality trait in tomato, its accumulation is only transient and there is increasing evidence that the biosynthesis and degradation of starch plays an important role in determining *Brix* at harvest time (Baxter *et al*., 2005).

Sphingosine-1-phosphate lyase (gene 45) catalyses one of the first steps of sphingolipid biosynthesis. Interestingly, this enzyme co-maps with one of the precursors, Ser, and the metabolically related Gly and Thr QML (Fig. 3).

On BIN 7B, the phospholipid/glycerol acyltransferase gene (gene 56, Fig. 3) could be related to the decrease in glycerol-P levels, a QML localized into this BIN. Glycerol-P is an important intermediate metabolite in the fatty acids biosynthesis pathway in which this enzyme is involved.

The QML mapped onto BIN 7F was a variation in *S*-methyl-Cys content; since this metabolite was not measured through a developmental time-series experiment, it was not possible to analyse any correlation

between candidate gene expression and the QML (Fig. 6). However, three candidates deserve to be highlighted in view of their positions within the metabolic pathways (Fig. 3). First, a methionine sulphoxide reductase (gene 61), thought to participate in the protection of chloroplasts against oxidative damage (Vieira Dos Santos *et al*., 2005), may be involved in the alterations found in the levels of *S*-methyl-Cys increasing the free Met pool by reducing Met-S-oxide in the reverse reaction. The amino acid variation found at residue 34 ($S_{lyc} \rightarrow R_{pen}$) of the sequence analysed lies within the signal peptide that directs this protein into the chloroplast (Vieira Dos Santos *et al*., 2005). Secondly, a mitochondrial malate dehydrogenase (mMDH) also mapped onto this BIN (gene 64). This protein has been implicated in modifying photosynthetic activity and aerial growth in tomato under ambient growth conditions (Nunes-Nesi *et al*., 2005). mMDH-silenced tomato plants were characterized by a decreased partitioning into organic and amino acids, an altered redox state and dramatic alterations in foliar ascorbic acid levels (Nunes-Nesi *et al*., 2005). No allelic variations were observed between *S. lycopersicum* and *S. pennellii*. Thirdly, phosphoenolpyruvate carboxylase (gene 67), which also mapped to this BIN, is involved in malate assimilation and at post-transcriptional level it is regulated by this compound which could eventually lead to the modification of the fluxes from the TCA cycle (through oxalacetate) to amino acid biosynthesis.

The expression of a constitutive plastid lipid-associated protein (gene 68), a polygalacturonase inhibitor gene (gene 69), and a photosystem II 10 kDa protein (gene 73), all mapped on BIN 7H presented a co-response with 3, 4, and 2 of the co-located QML observed in *S. pennellii* ILs, respectively (Fig. 6). Interestingly, the polygalacturonase inhibitor expression displays a positive correlation with the QML for glucose-6-P, mapped onto this region, as well as with the sucrose content; and a negative correlation with fructose and glucose contents. A gene encoding a lysine decarboxylase protein (gene 71) could possibly be involved in the variation in β-Ala, Met, H-Ser, and Thr co-localizing to this BIN. Another gene with the potential to be directly involved with variations of Thr, Gly, Ser, and glucose-6-P is the phosphoglycerate kinase (gene 74), linked to these QML, which showed two amino acid polymorphisms in the fragment analysed (Table 1). This observation is in line with the finding of two other linked genes related with photosynthesis: a chloroplast-associated (gene 68) and the photosystem II 10 kDa proteins (gene 73) that could play in the mentioned variations. A poly(A)-specific ribonuclease (gene 72) also mapped onto this region showed a high level of amino acid polymorphisms (Table 1). As an alternative to the involvement of the other candidates mentioned above, it is conceivable that this gene has a regulatory role that contributes to, or indeed even causes, the observed metabolic variations.

Out of the 15 genes profiled from BIN 9B/D/E only five (gene 76, peroxidase; gene 79, 1-phosphatidylinositol-4-phosphate 5-kinase; gene 81, enolase; gene 91, chloroplast pigment-binding protein; and gene 100, photosystem II reaction centreW protein) displayed a co-response with dehydroascorbate and phosphate, both co-located QML (Fig. 6). An obvious candidate associated with the increment observed in the levels of dehydroascorbate was the gene encoding monodehydroascorbate reductase (gene 83; Fig. 4), wherein three single nucleotide polymorphisms were found; two in an intron and one that resulted in an amino acid change in the coding region analysed.

On BIN 9J, an acireductone dioxygenase (gene 103), involved in Met metabolism, and a malate dehydrogenase (gene 105; Fig. 5), positively correlated with a co-located QML observed for Ala (Fig. 6). Another gene that could putatively be involved in the variation of this amino acid was a glutamyl-tRNA aminotransferase (gene 106). Despite the fact that the correlative behaviour of this gene could not be assessed, an amino acid polymorphism was found in its coding region (Table 1), so its candidature cannot be discarded. Similarly, variation in threonate levels mapped on this BIN could be linked to the presence of a GDP-mannose-3,5-epimerase gene (107), where polymorphisms between the two alleles were observed.

In BIN 10B, a CXE carboxylesterase (gene 112) presents a positive correlation with a GABA-co-localized QML, while an NAD-dependent isocitrate dehydrogenase (gene 114) negatively correlates with both GABA and Ile QML. It is conceivable that an increment in NAD-dependent isocitrate dehydrogenase mRNA levels negatively affects the GABA and T-4-OH-Pro contents by diverting the flux of 2-oxoglutarate towards Glu metabolism. In addition, β-cyanoalanine synthase (gene 109), a key enzyme involved in the detoxification of HCN, co-localizes with QML for Ala and Gly. This enzyme has previously been characterized as playing an important role in the detoxification of HCN, a side product of ethylene biosynthesis during climateric fruit ripening (Han *et al*., 2007).

Four genes which mapped to BIN 11C (gene 119, plastid quinol oxidase; gene 123, JAB; gene 124, proline iminopeptidase; and gene 125, ADP/ATP translocator), displayed correlation with several of the metabolites whose QTL co-localized to glucose, dehydroascorbate, fumarate, GABA, and Asp. Intriguingly, the dehydroascorbate QML co-localizes to a dehydroascorbate reductase (gene 117). Whilst no expression data are available for this gene from the previous developmental series experiment, allelic variation was found at the amino acid level, highlighting this as an interesting candidate for further study. Since ascorbic acid-associated genes have been deeply surveyed in tomato, it is unlikely that the gene identified in this work localized into BIN 11C is different from that previously mapped by Zou *et al*. (2006) onto BIN 11D, being an inaccurate localization. Another

obvious candidate for the dehydroascorbate QML mapped into this BIN is the phosphomannose mutase (gene 127) that was also mapped by Zou *et al*. (2006). Finally, the co-localization of the sucrose transporter *SUT1* gene with glucose QML is highly interesting, particularly in light of the fact that antisense inhibition of this gene resulted in modification of this metabolite content, as well as dramatic morphological changes (Hackel *et al*., 2006).

Conclusions

In this article, a combination of molecular marker sequence analysis, PCR amplification and sequencing, analysis of allelic variation, and evaluation of co-responses between gene expression and metabolite composition traits was used in order to identify candidate genes responsible for a sub-set of the previously reported metabolic QTL (Schauer *et al*., 2006). Using this combined strategy, 127 candidate genes located in 16 regions of the tomato genome were identified, 85 genes were cloned and partially sequenced from both *S. lycopersicum* and *S. pennellii*, and allelic variation at the amino acid level was confirmed in 37 of these candidates. Furthermore, of the 127 gene-metabolite co-locations, some 56 were recovered following correlation of parallel transcript and metabolite profiling. It is likely that the combined approaches taken here would allow the detection of both expression QTL (wherein the mechanism underlying the metabolic change is an alteration in transcript and by implication in protein amount), as well as change in function mutations in which the level of expression is unaltered (for example, the modified enzymatic activity of the *S. pennellii* LIN5 isoform invertase; Fridman *et al*., 2004). The candidate genes discussed here fit into both categories.

The work presented here represents the initial steps in the integration of genetic, genomic, and expressional patterns of genes co-localizing with chemical compositional traits of the fruit. Whilst, in the present study were mapped a similar number of genes as by Causse *et al*. (2004), due to the nature of the present approach it was possible to map a higher density of candidate genes. Depending on the gene nature, different strategies are being used for functional analyses in order to gather information about the role of these candidates. Moreover, a physical map of some of the genomic regions studied is under construction using *S. pennellii* BAC and COS libraries in order to facilitate future sequencing initiatives. Once complete it is likely that this work will allow the identification of novel candidate genes but will also be useful for BAC sorting and sequence assembly in the nascent tomato genome sequencing programme (Mueller *et al*., 2005*b*).

Supplementary data

The complete candidate genes information is detailed in Table S1. The primer sequences used to amplify all selected candidate genes are provided in Table S2. Supplementary data may be found at *JXB* online.

Acknowledgements

This work was partially supported with grants from FAPESP (Brazil), CNPq (Brazil), Max Planck Society (Germany), INTA (Argentina), CONICET (Argentina), and under the auspices of the EU SOL Integrated Project FOOD-CT-2006-016214. UU was the recipient of PIBIC (Brazil) and CONICET (Argentina) fellowships. LB was the recipient of a FAPESP (Brazil) fellowship. RA, FC, and LK are members CONICET. This work was carried out in compliance with current laws governing genetic experimentation in Brazil and in Argentina.

References

Baxter CJ, Carrari F, Bauke A, Overy S, Hill SA, Quick PW, Fernie AR, Sweetlove LJ. 2005. Fruit carbohydrate metabolism in an introgression line of tomato with increased fruit soluble solids. *Plant Cell Physiology* **46,** 425–437.

Carrari F, Baxter C, Usadel B, *et al*. 2006. Integrated analysis of metabolite and transcript levels reveals the metabolic shifts that underlie tomato fruit development and highlight regulatory aspects of metabolic network behaviour. *Plant Physiology* **142,** 1380–1396.

Causse M, Duffe P, Gomez MC, Buret M, Damidaux R, Zamir D, Gur A, Chevalier C, Lemaire-Chamley M, Rothan C. 2004. A genetic map of candidate genes and QTLs involved in tomato fruit size and composition. *Journal of Experimental Botany* **55,** 1671–1685.

Chen KY, Cong B, Wing R, Vrebalov J, Tanksley SD. 2007. Changes in regulation of a transcription factor lead to autogamy in cultivated tomatoes. *Science* **318,** 643–645.

Corpet T. 1988. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Research* **16,** 10881–10890.

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.

Fernie AR, Tadmor Y, Zamir D. 2006. Natural genetic variation for improving crop quality. *Current Opinion in Plant Biology* **9,** 196–202.

Frary A, Nesbitt TC, Grandillo S, Knaap E, Cong B, Liu J, Meller J, Elber R, Alpert KB, Tanksley SD. 2000. fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. *Science* **289,** 85–88.

Fridman E, Carrari F, Liu YS, Fernie A, Zamir D. 2004. Zooming in on a quantitative trait for tomato yield using interspecific introgressions. *Science* **305,** 1786–1789.

Galpaz N, Ronen G, Khalfa Z, Zamir D, Hirschberg J. 2006. A chromoplast-specific carotenoid biosynthesis pathway is revealed by cloning of the tomato white-flower locus. *The Plant Cell* **18,** 1947–1960.

Ganal MW, Czihal R, Hannappel U, Kloos DU, Polley A, Ling HQ. 1998. Sequencing of cDNA clones from the genetic map of tomato (*Lycopersicon esculentum*). *Genome Research* **8,** 842–847.

Gordon D, Abajian C, Green P. 1998. Consed: a graphical tool for sequence finishing. *Genome Research* **8,** 195–202.

Hackel A, Schauer N, Carrari F, Fernie AR, Grimm B, Kühn C. 2006. Sucrose transporter LeSUT1 and LeSUT2

inhibition affects tomato fruit development in different ways. *The Plant Journal* **45,** 180–192.

Han S, Seo YS, Kim D, Sung S-K, Kim WT. 2007. Expression of MdCAS1 and MdCAS2, encoding apple β-cyanoalanine synthase homologs, is concomitantly induced during ripening and implicates MdCASs in the possible role of the cyanide detoxification in Fuji apple (*Malus domestica* Borkh.) fruits. *Plant Cell Reports* **26,** 1321–1331.

Hoisington D, Khairallah M, Gonzalez de Leon D. 1994. *Laboratory protocols*. El Baton, Mexico: CIMMYT Applied Molecular Genetics Laboratory.

Kanehisa M, Araki M, Goto S, *et al.* 2008. KEGG for linking genomes to life and the environment. *Nucleic Acids Research* **36,** 480–484.

Katz YS, Galili G, Amir R. 2006. Regulatory role of cystathionine-γ-synthase and de novo synthesis of methionine in the ethylene production during tomato fruit ripening. *Plant Molecular Biology* **61,** 255–268.

Kortstee AJ, Appeldoorn NJ, Oortwijn ME, Visser RG. 2007. Differences in regulation of carbohydrate metabolism during early fruit development between domesticated tomato and two wild relatives. *Planta* **226,** 929–939.

Kumar S, Tamura K, Jakobsen IB, Nei M. 2001. MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17,** 1244–1245.

Lippman ZB, Semel Y, Zamir D. 2007. An integrated view of quantitative trait variation using tomato interspecific introgression lines. *Current Opinion in Genetics & Development* **17,** 1–8.

Lu S, Van Eck J, Zhou X, *et al.* 2006. The cauliflower *Or* gene encodes a Dnaj cysteine-rich domain-containing protein that mediates high levels of β-carotene accumulation. *The Plant Cell* **18,** 3594–3605.

Mueller L, Mills A, Skwarecki B, Buels R, Menda N, Tanksley S. 2008. The SGN comparative map viewer. *Bioinformatics* **24,** 422–423.

Mueller LA, Solow TH, Taylor N, *et al.* 2005*a*. The SOL genomics network: a comparative resource for Solanaceae biology and beyond. *Plant Physiology* **138,** 1310–1317.

Mueller LA, Tanksley SD, Giovannoni JJ, *et al.* 2005*b*. The Tomato Sequencing Project, the first cornerstone of the International Solanaceae Project (SOL). *Comparative and Functional Genomics* **6,** 153–158.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* **5,** 418–426.

Nunes-Nesi A, Carrari F, Lytovchenko A, Smith AMO, Loureiro ME, Ratcliffe RG, Sweetlove LJ, Fernie AR. 2005. Enhanced photosynthetic performance and growth as a consequence of decreasing mitochondrial malate dehydrogenase activity in transgenic tomato plants. *Plant Physiology* **137,** 611–622.

Paran I, Van der Knaap E. 2007. Genetic and molecular regulation of fruit and plant domestication traits in tomato and pepper. *Journal of Experimental Botany* **58,** 3841–3852.

Price A. 2006. Believe it or not, QTLs are accurate! *Trends in Plant Science* **11,** 213–216.

Qiu XB, Shao YM, Miao S, Wang L. 2006. The diversity of the DnaJ/Hsp40 family, the crucial partners for Hsp70 chaperones. *Cell and Molecular Life Science* **63,** 2560–2570.

Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19,** 2496–2497.

Salvi S, Tuberosa R. 2005. To clone or not to clone plant QTLs: present and future challenges. *Trends in Plant Science* **10,** 297–304.

Schauer N, Fernie A. 2006. Plant metabolomics: towards biological functions and mechanism. *Trends in Plant Science* **11,** 508–516.

Schauer N, Semel Y, Roessner U, *et al.* 2006. Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nature Biotechnology* **24,** 447–454.

Sterling JD, Atmodjo MA, Inwood SE, Kumar Kolli VS, Quigley HF, Hahn MG, Mohnen D. 2006. Functional identification of an *Arabidopsis* pectin biosynthetic homogalacturonan galacturonosyltransferase. *Proceedings of National Academy of Science, USA* **103,** 5639–5640.

Stevens R, Buret M, Duffé P, Garchery C, Baldet P, Rothan C, Causse M. 2007. Candidate genes and QTLs affecting fruit ascorbic acid content in three tomato populations. *Plant Physiology* **143,** 1943–1953.

Tabor HK, Risch NJ, Myers RM. 2002. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Reviews Genetics* **3,** 1–7.

Tatusova TA, Madden TL. 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiology Letters* **174,** 247–250.

Urbanczyk-Wochniak E, Luedemann A, Kopka J, Selbig J, Roessner-Tunali U, Willmitzer L, Fernie AR. 2003. Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Reports* **4,** 989–93.

Van der Hoeven R, Ronning C, Giovannoni J, Martin G, Tanksley S. 2002. Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. *The Plant Cell* **14,** 1441–1456.

Vieira Dos Santos C, Cuiné S, Rouhier N, Rey P. 2005. The Arabidopsis plastidic methionine sulfoxide reductase B proteins: sequence and activity characteristics, comparison of the expression with plastidic methionine sulfoxide reductase A, and induction by photooxidative stress. *Plant Physiology* **138,** 909–922.

Zou L, Li H, Ouyang B, Zhang J, Ye Z. 2006. Cloning and mapping of genes involved in tomato ascorbic acid biosynthesis and metabolism. *Plant Science* **170,** 120–127.