

# Metabolic analyses of interspecific tomato recombinant inbred lines for fruit quality improvement

Mariana G. López<sup>1,8</sup>, María I. Zanor<sup>2,8</sup>, Guillermo R. Pratta<sup>3,8</sup>, Georgina Stegmayer<sup>4,8</sup>, Silvana B. Boggio<sup>2,8</sup>, Mariana Conte<sup>1</sup>, Luisa Bermudez<sup>1,8</sup>, Carla Coluccio Leskow<sup>1,8</sup>, Gustavo R. Rodríguez<sup>3,8</sup>, Liliana A. Picardi<sup>3</sup>, Roxana Zorzoli<sup>3</sup>, Alisdair R. Fernie<sup>5</sup>, Diego Milone<sup>6,8</sup>, Ramón Asís<sup>7,8</sup>, Estela M. Valle<sup>2,8</sup>, Fernando Carrari<sup>1,8(\*)</sup>

<sup>(1)</sup> Instituto de Biotecnología, INTA, Argentina

<sup>(2)</sup> Instituto de Biología Molecular y Celular de Rosario, CONICET-UNR, Argentina

<sup>(3)</sup> Cátedra de Genética, FCA, UNR, Argentina

<sup>(4)</sup> Centro de Investigación en Ingeniería en Sistemas de Información, CONICET, Argentina

<sup>(5)</sup> Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany

<sup>(6)</sup> Research Center for Signals, Systems and Computational Intelligence, FICH-UNL, CONICET, Argentina

<sup>(7)</sup> Departamento de Bioquímica Clínica, FCQ, UNC, Argentina

<sup>(8)</sup> Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina

(\*) corresponding author:

Dr. Fernando Carrari

[carrari.fernando@inta.gob.ar](mailto:carrari.fernando@inta.gob.ar)

Instituto de Biotecnología - CICVyA INTA

Dr. Nicolás Repetto y De Los Reseros S/N°

(B1686IGC) Hurlingham. Buenos Aires. Argentina

Tel.: 005411 4621-1447/1676/1127/1278 int. 172 - Fax: 005411 4621-0199

Abbreviated title: RILs analysis for quality improvement

## Abstract

Elucidating the determinants of tomato nutritional value and fruit quality to introduce improved varieties on the international market represents a major challenge for crop biotechnology. Different strategies can be undertaken to exploit the natural variability of *Solanum* to re-incorporate lost allelic diversity into commercial varieties. One of them is the characterization of selected germplasm for breeding programs. To achieve this goal, 18 RILs (*S. lycopersicum* x *S. pimpinellifolium*) were comprehensively phenotyped for fruit polar metabolites and quality associated traits. Metabolites were quantified by GC-MS and <sup>1</sup>H-NMR. Integrative analyses by neuronal clustering and network construction revealed that fruit properties are strongly associated with the metabolites aspartate, serine, glutamate and 2-oxoglutarate. Shelf life and firmness appeared to be linked to malate content. By a comparative analysis of the whole data set, ten RILs presented higher number of traits with positive effect than the *S. lycopersicum* x *S. pimpinellifolium* hybrid. Thus, these lines can be proposed as promising candidates for breeding programs aimed to improve fruit quality.

**Keywords:** tomato, fruit quality, *Solanum lycopersicum*, metabolomics

# 1 Introduction

Tomato (*Solanum lycopersicum* L.) is an important horticultural crop world-wide and the second most consumed non-cereal vegetable after potato (<http://faostat3.fao.org/>). Its fruits are a source of essential minerals and nutrients for the human diet, such as vitamins A and C and other antioxidants (Willcox et al. 2003; Fitzpatrick et al. 2012).

Cherry tomato (*S. lycopersicum* var *cerasiforme* (Alef.) Voss.) was probably domesticated from *S. pimpinellifolium* L. and is likely the ancestor of cultivated big-fruited tomato. Domestication was initiated by indigenous people of the Andes who kept and propagated seeds from wild plants with bigger and tastier fruits. During this process 186 sweeps were selected representing 8.3% of the genome (Lin et al. 2014).

Tomato breeding, on the other hand, began in Europe when improved cultivars were generated to meet several needs including fresh market and processing industries (Foolad 2007). At the beginning of the 20th century, public institutions from USA and new private companies became interested in this practice. Since then until now, vast research and economic efforts have been invested in tomato improvement. It was recently reported that selection during improvement process affected 4,807 genes, representing 7 % of the genome, of which 1 % has also undergone double selection because of previous domestication process (Lin et al. 2014).

Over the history of tomato breeding four clear periods can be defined according to the major target of improvement (Bai and Lindhout 2007). Initially, efforts were focused on yield (*i*), which increased seven-fold in processing tomato between the 20s and 30s. This was due to incorporation of diseases resistances, a good performance of the selected F1 hybrids to fertilization, the use of pesticides (Warren 1998), enhanced tolerance to abiotic stresses and an increase in fruit sugar content (Foolad 2007). In fresh tomato, yield enhancement was achieved by crosses with wild relative species (Swamy and Sarla 2008). On this regard, many quantitative trait loci (QTL) for yield and related traits as fruit weight, total soluble solids or lycopene content were found non-randomly distributed in the genome (Fulton et al. 1997; 2000).

The second period began in the '80s, when shelf life (*ii*) became the most important issue for fresh market tomatoes. In that time, the effort was focused on elucidating the mechanism of ripening since this process directly affects shelf life. Many studies were concentrated in identifying the principal components involved in fruit ripening and softening including the role of ethylene (Alba et al. 2005) and the enzyme polygalacturonase (Giovannoni 2001). As a result, several ripening-related genes and QTLs were characterized and mapped. One of them, *rin* (ripening inhibitor) has been used in marker assisted selection (MAS) programs with promising results (Foolad 2007). In the '90s, flavor (*iii*) became the new target for breeding programs. This is a very complex trait mainly because it is determined by quite a lot of genetic and non-genetic factors, not all of them are identified or well characterized (Causse et al. 2002; Klee 2013). It has been postulated that the ratio between sugars and organic acids is the main determinant of tomato

flavor (Bucheli et al. 1999) but several aromatic volatile compounds also have major influences on this trait (Goff and Klee 2006; Tieman et al. 2012; Rambla et al. 2014). Even though the great effort to improve fruit flavor, only increments in soluble solid and decreases of acidity were obtained in few cases (Foolad 2007). Finally, in the last and current stage, the breeding goal has been re-orientated to the fruit nutritional quality (*iv*) (Bai and Lindhout 2007).

Over the history of tomato breeding diverse techniques have been applied. Initially they were based on phenotypic selection and progeny testing and later, with the advent of molecular markers and linkage maps, MAS improved the efficiency and reduced the time of the traditional programs (Yang et al. 2004; Foolad 2007) and QTL allow identifying complex traits in F2 and backcross populations (Paterson et al. 1988). However, these materials resulted disadvantageous for genetic mapping (Foolad 2007). This problem was overcome a few years ago with the development of recombinant inbred lines (RILs) and introgression lines (ILs) that harbor specific genomic regions of wild relative species. Nowadays, with the full tomato genome sequence available (Sato et al. 2012), candidate genes for important traits are being rapidly identified (Causse et al. 2004; Price 2006; Bermúdez et al. 2008; Kamenetzky et al. 2010; Goulet et al. 2012; Lee et al. 2012; Sauvage et al. 2014) and breeders may obtain novel and improved varieties by introgressing useful alleles from wild relative species to the cultivated varieties. Since these species represent a source of natural genetic variation, of which only 5% is harbored by cultivated tomato (Miller and Tanksley 1990; Bretó et al. 1993), introgression lines represent powerful tools for crop improvement maximizing the potential of wild germplasm. Moreover, RILs constitute an original germplasm source for exploiting favorable new genetic combinations involved in fruit quality through the generation of Second Cycle Hybrids (SCHs) (Liberatti et al. 2013). Because *S. pimpinellifolium* displays high sweetness and vitamin C contents, disease-resistance and improved stress tolerance (Foolad et al. 1998; Foolad 2007; Tigchelaar 1986), and also its genome has been recently sequenced (Sato et al. 2012); the comprehensive metabolic profiling of RILs derived from this species is of major interest to assess fruit nutritional quality.

Metabolite profiling alone or combined with other approaches has been used to identify key compounds involved in development, stress tolerance, and nutritional metabolites in many agricultural important plants (Hu et al. 2014). Those metabolites are useful and helpful for plants and either human health and diets (Hall et al. 2008; Gechev et al. 2014). This approach was likewise applied to explore natural variability in wild related species in order to find valuable source for the improvement of agriculturally important traits (Schauer et al. 2005; Rambla et al. 2014). It was also used to discover enzyme function, reconstruct important pathways and define its regulation (Bermúdez et al. 2014; Araújo et al. 2012). Additionally, metabolic profiling coupled with GWA studies allow the identification of 44 SNP loci associated to 19 metabolic traits and provide 5 candidate genes involved in the genetic architecture of fruit metabolic traits (Sauvage et al. 2014).

Molecular breeding, on the other hand, has been adopted in most biotechnological strategies to develop new crops (Moose and Mumm 2008; Saito and Matsuda 2010). For functional understanding of phenotypes it is essential to integrate genomic information to characterize gene-to-metabolite associations (Carreno-Quintero et al. 2013). Metabolic analysis has been increasingly and successfully used to assist elite germplasm selection (Fernie and Schauer 2009; Rao et al. 2014). In tomato, the first study using metabolite profiling showed that metabolic traits correlated with phenotypic traits such as yield or harvest index, exposing the challenge to use metabolites as biomarkers (Hermann and Schauer 2013). In this regard, the major aim of this work was to characterize metabolic based (Wahyuni et al. 2012) and related agronomic traits of 18 select tomato RILs (*S. lycopersicum* x *S. pimpinellifolium*) to provide valuable information to improve fruit quality and metabolic-based traits. To achieve this goal we firstly applied a well-established gas chromatography coupled to mass spectrometry (GC-MS) platform, examining polar extracts of tomato fruit pericarp (Roessner-Tunali et al. 2003) complemented with proton nuclear magnetic resonance (<sup>1</sup>H-NMR) metabolite profiling method (Mattoo et al. 2006; Sorrequieta et al. 2013). Secondly, we evaluated trait relations using clustering methods and network analyses (Guimerà and Nunes Amaral 2005).

## 2 Materials and methods

### 2.1 Plant material

Plant material consisted of *S. lycopersicum* (cv Caimanta), the red-fruited wild species *S. pimpinellifolium* (LA722), F1 hybrids between both species and eighteen tomato RILs. RILs were obtained after seven generations of selfing and five cycles of antagonistic and divergent selection for fruit shelf life and weight (Zorzoli et al. 2000; Rodríguez et al. 2006); they represent a novel source of public germplasm.

Seeds of the 18 RILs, their parents and F1 were germinated in seedling trays at the end of June and transplanted to greenhouse after a month according to a completely randomized design. Plants were grown at the Experimental Station "José F. Villarino" (33° SL and 61° WL), Argentina. Plant density in the field was approximately 4 plants per square meter (70 cm between grooves, 40 cm between plants into each groove).

### 2.2 Fruit phenotype analyses

Twelve agronomic, both morphological and biochemical traits were measured according to Gallo et al. (2011). Six plants and ten fruits per plant were evaluated for morphological fruit traits including: diameter, weight, height, shape (height/diameter ratio), pericarp thickness, locule number, firmness, shelf life, color index (a/b) and reflectance percentage. Biochemical traits: pH, titratable acidity, soluble solids and soluble solids/acidity ratio, were measured in three independent fruit pools per line harvested from six plants. All characters were measured at ripe stage except

(when fruit has fully acquired its final color) except for shelf life that was evaluated at breaker stage (10 % of fruit surface has acquired red (or final) color). for shelf life that was evaluated at breaker stage.

## 2.3 Metabolite profile analyses

Metabolite profiles were obtained by both GC-time of flight-(tof)-MS and <sup>1</sup>H-NMR analyses (as described below) from pericarps of six fruits harvested from six independent plants at ripe stage. Fruits were collected from the 2<sup>nd</sup> and 3<sup>rd</sup> floors.

### 2.3.1 Sample preparation, extraction and GC-MS analyses

The relative levels of metabolites were determined from frozen pericarp samples following the protocol established by Roessner-Tunali et al. (2003) for tomato tissue. Fresh tomato tissue pericarps were harvested at ripe stage, rapidly frozen in liquid nitrogen and stored at -80 °C until analysis. For sample extraction ~250 mg of frozen pericarps were manually grounded in liquid nitrogen. All the powder obtained was extracted in 3000 µl of cold methanol and 120 µl of internal standard (0.2 mg/ml ribitol in water) was added for quantification. The mixture was incubated for 15 min at 70°C, mixed vigorously with 1500 µl of water and centrifuged at 2,200 g. The methanol/water supernatant was reduced to dryness under vacuum. Samples were stored at -80 °C until GC-MS analysis. The dried extract was re-dissolved and derivatised for 120 min at 37°C (in 60 µl of 30 mg/ml methoxyamine hydrochloride in pyridine) followed by a 30 min treatment at 37°C with a mixture of 100 µl of *N*-methyl-*N*-[trimethylsilyl] trifluoroacetamide and 20 µl of retention time standard mixture composed by 0.4 ml ml<sup>-1</sup> of the same 13 fatty acid methy esters (FAMES) used by (Lisec et al. 2006). Sample volumes of 1 µl were then injected in the GC-MS using a splitless mode and a hot needle technique.

The GC-tof-MS system was composed of an AS 2000 autosampler, a GC 6890N gas chromatographer (Agilent Technologies, Santa Clara, CA, USA), and a Pegasus III time-of-flight mass spectrometer (LECO Instruments, St. Joseph, MI, USA), provided with an Electron Impact ionization source. GC was performed on a MDN-35 capillary column, 30 m in length, and 0.32 mm in inner diameter, 0.25 mm in film thickness (Macherey-Nagel). The injection temperature was set at 230°C, the interface at 250°C, and the ion source adjusted to 200°C. Helium 5.0 was used as the carrier gas at a flow rate of 2 ml/min. The analysis was performed under the following temperature program: 2 min of isothermal heating at 80°C, followed by a 15°C per min ramp to 330°C, and holding at this temperature for 6 min. Mass spectra were recorded at 20 scans per sec with a scanning range of 70 to 600 m/z. The experimental set was composed by 126 samples (18 RILs, both parents and F1 hybrid x 6 replicates each), they were separated in three runs of 54 samples each one, including 6 RILs, parents and F1 hybrids for comparison in each run. To

homogenize variation through a single run, samples were injected in the following order: replicate #1, #2...#6 of all lines. A sample of *Arabidopsis* leaves were included in each run, randomly distributed to check machine sensitivity, one of the most important GC-MS quality control (see Lisec et al. 2006). Both chromatograms and mass spectra were evaluated using ChromaTOF software, version 3.00 (LECO Instruments, St. Joseph, MI, USA), .peg files were exported to .cdf using a baseline off set of 1('just above the noise'), an average of 5 points for smoothing, a peak width of 10 and a signal to noise ratio of 10. Identification and semi-quantitation of the compounds detected in the GC-TOF-MS metabolite profiling experiment were performed with TagFinder 4.0 software (Luedemann et al. 2008). Then .cdf files were converted to .txt with the Pick Apex finding tool of TagFinder, considering a smooth width apex finder of 10 and an intensity threshold of 50. Retention time index were calculated using data of the standards included. Data matrix was obtained using a time scan width of 400 s and a Max-Intensity aggregation of peaks. Mass pairs were automatically extracted with the pBuilder.MassPairFinder tool of TagFinder. Finally, metabolites were identified by comparison with spectral data from the public library GMD@CSB.DB (The Golm Metabolome Database; Kopka et al. 2005), this library includes RIs, molecular weight ( $m/z$ ), and the associated MS spectra (see online resources Table S1 for metabolites annotation data).

The standardization of the complete metabolite profiling experiment was made in accordance with Lisec et al. (2006). All samples were measured in three independent GC-MS runs, since datasets measured at different times are not directly comparable because of varying tuning parameters of the GC-MS machine over time; we therefore normalized the data by using *S. lycopersicum* parental species of each measured batch as a reference (Roessner et al. 2001).

### 2.3.2 Sample preparation and $^1H$ -NMR analyses

Absolute levels of metabolites were determined from frozen pericarp samples following the protocol published by Sorrequieta et al. (2013). One gram of fresh weight of pericarp frozen in liquid nitrogen without peel was extracted in 0.3 ml of 1 M phosphate buffer (pH 7.4) prepared in  $D_2O$ . The solution was centrifuged at 13,500 g for 15 min at 4°C and the supernatant filtered to remove any insoluble material. One mM of internal standard (TSP: 3-(trimethylsilyl) propionic-2,2,3,3- $d_4$  acid sodium salt) was added to the resulting transparent soluble fraction and the solution was subjected to spectral analysis at 600.13 MHz on a BrukerAvance II spectrometer. Proton spectra were acquired at 298 K by adding 512 transients of 32K data points with a relaxation delay of 5 s. A 1D-NOESY pulse sequence was utilized to remove the water signal. The 90° flip angle pulse was always ~10  $\mu$ s. TSP was used for both, chemical shift calibration and quantitation, that is, proton spectra were referenced to the TSP signal ( $\delta = 0$  ppm) and their intensities were scaled to that of TSP. Spectral assignment and identification of specific metabolites

was established by fitting the reference  $^1\text{H-NMR}$  spectra of several compounds using the software Mixtures, developed *ad hoc* as an alternative to commercial programs (Abriata 2012). Briefly, the programme allows easy visualization and basic editing of spectra. It also provides a wizard that aids in fitting spectra from a database of known compounds to the signals in the spectrum. Fitted signals are integrated and integrals are exported to a standard spreadsheet file for further analysis (see online resources Table S1 and Fig. S1 for metabolites annotation data). Further confirmation of the assignments for some metabolites was obtained by acquisition of new spectra after addition of authentic standards. Analysis of the  $^1\text{H-NMR}$  spectra of the pericarp of different mature fruits were performed as previously described (Sorrequieta et al. 2013).

## 2.4 $^1\text{H-NMR}$ and GC-MS methods comparison

A comparative analysis between  $^1\text{H-NMR}$  and GC-MS data was performed in order to evaluate the confidence between both technologies. The normalized values for the 16 compounds quantified in common by both methods were correlated by applying Pearson's coefficient (Table 1). Although both methods use different extractions solvents the general correlation was a moderate and highly significant ( $r= 0.60$ ,  $p<0.0001$ ). Few metabolites, namely fructose, glucose, glutamine, sucrose and tryptophan, displayed non-significant correlation between both technologies. Particularly, tryptophan and sucrose show the low  $^1\text{H-NMR}$  signals (see online resources Fig. S1); this makes difficult the calculation of both metabolites contents by  $^1\text{H-NMR}$ . We then decided to keep data from  $^1\text{H-NMR}$  and GC-MS separately and evaluate the association between those metabolites with the \*omeSOM tool (see below).

## 2.5 Statistical analyses

In order to compare agronomic and metabolic trait differences between RILs and the parental line *S. lycopersicum* (cv. Caimanta), collected data were analyzed using the t-test algorithm embedded into Microsoft Excel software (Microsoft Corporation, Redmond, WA, USA) considering a  $p$  value  $<0.01$  as significant.

## 2.6 Data integration

Agronomic and metabolic data were firstly analyzed by HC (Hierarchical clustering) and visualized using MeV software (Saeed et al. 2006). Secondly, they were integrated into a self-organizing map for \*omic data (\*omeSOM) developed by Milone et al. (2010). Self-organizing maps (SOM) represent a special class of neural networks that use competitive learning, which is based on the idea of units (neurons) that compete to respond to a given subset of inputs. Each neuron corresponds to a cluster and is associated with a prototype or weight vector. Given an input

pattern, its distance to the weight vectors is computed and only the neuron closest to the input becomes activated. Results were visualized using the software's graphic interface (\*omeSOM, available on: <http://sourcesinc.sourceforge.net/omesom/>). This tool trains a two-dimensional SOM for clustering, allowing representing complex high-dimensional input patterns in the form of a simple low-dimensional discrete map. Therefore, SOMs can be appropriate for cluster analysis when looking for underlying or so-called hidden patterns in data. Once trained, the software allows the visualization of coordinated variations of all integrated elements in order to easily reveal relations among the different kind of included data. A visualization neighbourhood (Vn) can be set, which defines the radius of adjacent neurons that will be considered as a unique group. Data matrix was constructed with all agronomic and metabolic data. Each data point, the *i*-th agronomic or metabolic trait for the *j*-th RIL, was normalized by mean ( $\mu$ ) and standard deviation ( $\sigma$ ) as follows:

$$\bar{x}_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i}$$

Metabolite data obtained by both <sup>1</sup>H-NMR and GC-MS technology are expressed relative to one of those measured in the parental *S. lycopersicum* (cv Caimanta). In order to reveal all the possible associations, directed and inverted patterns were used to train the \*omeSOM model as described by Milone et al. (2010).

With the aim to define the Vn value to be considered as informative, obtain a better visualization of neurons that were related among them and also compare the SOM method (innovatively applied to breeding here) with the traditional used analysis (correlation), a network reconstruction was performed using NetDraw (Analytic Technologies, Lexington, KY). The same data matrix used for SOM analysis was employed to compute the correlations. Each nodes (agronomic –diamonds- or metabolic –circles- traits) were represented as edges and correlations (Pearson with  $p < 0.001$ ) between traits were represented by connector lines. Correlation coefficient and significances between traits were calculated with InfoStat software (Di Rienzo et al. 2011). Positive and negative correlations and coefficient values are indicated by different color and thickness of the lines, respectively.

## 2.7 Mode of inheritance assessments

For both metabolic and agronomic traits measured in the parents and in their offspring (F1 hybrid), differences in mean values were analyzed by ANOVA and Tuckey test. Those traits showing significant differences ( $p < 0.05$ ) were classified into the following mode-of-inheritance categories and classes defined by Lisec et al. (2011), in which the effect of the *S. pimpinellifolium* allele is compared with the *S. lycopersicum* allele: *recessive* (only *S. pimpinellifolium* is significantly different from *S. lycopersicum* whereas the offspring is similar to *S. lycopersicum*), *additive* (the F1 is between the parents, which are significantly different from each other), *dominant* (both the

homozygous *S. pimpinellifolium* and the hybrid showed similar values but differed significantly from *S. lycopersicum*), or *overdominant* (the F1 is significantly higher or lower than both parents).

### 3 Results and Discussion

#### 3.1 Variation in the metabolic fruit composition in parental species, F1 and in the RIL population

The RIL population used in this study was obtained by divergent-antagonistic selection for fruit weight and shelf life traits (Zorzoli et al. 2000; Rodríguez et al. 2006). It was previously characterized with molecular markers (Pratta et al. 2011b) and evaluated for different traits related to tomato crop productivity (Gallo et al. 2011; Pratta et al. 2011a; 2011b). Since many efforts have been invested in recording agronomical valuable information, it constitutes an excellent source of public germplasm for breeding programs. Here, we focus the analysis on variations of the metabolite content in mature fruits by applying two standardized methods (<sup>1</sup>H-NMR and GC-MS) for metabolite profiles. A total of 60 different metabolic traits (16 with <sup>1</sup>H-NMR and 59 with GC-MS) were quantified in ripe fruits harvested from the two parental lines, their interspecific F1 hybrid and 18 selected RILs (see online resource Table S1). These metabolites correspond to amino (23) and organic acids (9), TCA cycle intermediates (5), soluble sugars (6), sugar alcohols (4), phosphorylated intermediates (3), few fatty acids (4), alkaloids (2), nucleotides (1), amides (1) and others (2) (for details see online resource Table S2).

A high level of divergence in terms of primary metabolism was evident when comparing metabolite contents between the parental lines. Most of the measured amino acids (9), all tricarboxylic acid (TCA) cycle intermediates (with the exception of succinate), dehydroascorbate, glucarate, nicotinate, turanose, xylose, erythritol, galactinol, inositol, glycerol-3P and the two measured alkaloids (calystegine A3 and calystegine B2) were significantly more abundant in *S. pimpinellifolium* than in cultivated tomato (see online resource Table S2, p<0.01). By contrast, only the levels of the amino acids methionine, serine, threonine, tyrosine and galacturonate were significantly lower in the fruits of the *S. pimpinellifolium* than in cultivated species (see online resource Table S2, p<0.01). It has been reported that these two tomato species have similar fruit water content (Schauer et al. 2005). Thus, the variations observed in the metabolite levels would indeed reflect differences in fruit composition. Our results are in good agreement with a previous report on the free amino acid composition over the same parental accessions but under different environmental conditions (Pratta et al. 2011a). This observation suggests strong heritability of these traits as was reported for many of the metabolite measured here by Schauer et al. (2008). Since, the existence of variation between parental species of a RIL population is a key pre-requisite for any breeding program; our results suggest that the chosen material and the approach selected is valid for improving nutritional quality.

**Fig 1**

A subset of metabolites displayed highly variable levels in the RILs with respect to their contents in the parental *S. lycopersicum* fruits (Fig. 1a, online resource Table S2). Among them, the amino acids  $\beta$ -alanine, GABA, glutamate, proline and 5-oxoproline showed the highest increases in a considerable number of RILs (Fig. 1b, online resource Table S2). The organic acids glucarate and quinate together with the TCA cycle intermediates malate, pyruvate and 2-oxoglutarate, displayed variable levels in the RIL population. In contrast, only two soluble sugars showed significant variations namely sucrose and, to a lesser extent, xylose. Besides, the contents of glycerol-3P, galactinol and erythritol also varied significantly among the RILs. Remarkably, the levels of calystegine B2, an alkaloid present in a wide range of *Solanaceae* species (Asano et al. 1997; Bekkouche et al. 2001), displayed a variation between four and 13-fold. Evaluating the proportion of every metabolite category varying in the RIL population (and also in the F1 hybrid) with respect to the cultivated parent (Fig. 1a, online resource Table S2), we found that five RILs (6, 7, 8, 9 and 10) displayed significant differences in all metabolite categories. These genotypes, together with the RILs 4, 17 and 18, showed the highest number of significant changes. The remaining lines also revealed significant variations; however, the proportion of altered metabolites in each category was considerably lower. Zooming into each category, the amino acid proline and the alkaloid calystegine B2 were the metabolites displaying the highest values of relative changes in all analysed lines, followed by the organic acid quinate and the sugar sucrose. In contrast, all fatty acids showed very similar values to those measured in the cultivated parent (Fig. 1b).

### 3.2 Data integration and clustering using the \*omeSOM model (SOM) to expose associations between traits

Having established the metabolic and agronomic variation into the studied material (online resource Table S2) we next focused our attention on integrating all data aiming to detect connections between yield-associated and metabolic traits. For this propose the \*omeSOM model was used (Stegmayer et al. 2009; Milone et al. 2010). The first step for clustering with \*omeSOM is the definition of the map size. In order to obtain the best value, we assayed a range of map sizes and evaluated the relative distance between the same metabolites measured by GC-MS and <sup>1</sup>H-NMR methods (Table 2). The relative distance is defined as the number of neurons between the position of a metabolite in the map measured with GC-MS and the same compound measured with <sup>1</sup>H-NMR, divided by the total number of neurons in the map. All map sizes evaluated (from 7x7 to 11x11) displayed similar relative distance values (Table 2, online resource Fig. S2). On the other hand, we explored the neuron components looking for well-established associations to find a biologically meaningful map size. The 9x9 map displayed neurons grouping all characters associated to fruit shape. This result is in agreement with previously reported analyses performed with the same RIL population (Pratta et al. 2011b). Moreover, malate content, fruit firmness and shelf life traits fell into the

**Table 2**

same neuron showing a relationship that is in accordance with the findings reported by Centeno et al. (2011). In addition, the 9x9 map also performed a high cohesion of the elements, meaning that the pattern of variation of the components into the same neuron was very similar. Consequently, the 9x9 map was chosen for further analyses.

The obtained map (Fig. 2, online resource Table S3) contained a total of 81 neurons grouping the 14 agronomic and 60 metabolic traits. We firstly evaluated if metabolites measured by both methods (<sup>1</sup>H-NMR and GC-MS) clustered together or in close proximity. For those located in the same neuron (or in close proximity  $V_n \leq 2$ ; see online resource Table S3) and showing significant Pearson correlation coefficients ( $\geq 0.65$ , Table 1) both measurements were considered equally confident and thus, kept separately for the rest of the study. For the other compounds we choose the methodology showing the lowest standard deviation values among replicates of the same lines (see online resource Table S2).

Analyzing the upper-left region of the map, 32 neurons integrated all measured characters. Within these, six were integrative neurons grouping metabolic with agronomic traits (Fig. 2, black squares); 25 neurons grouped only metabolic characters (Fig. 2, blue squares) and one neuron contained only a single agronomic trait (shape index) (Fig. 2, red squares). In order to identify metabolite regulators of the complex agronomical traits evaluated, we next concentrated our further analyses on integrative neurons and also on groups of neighboring neurons ( $V_n = 1$  and  $2$ ), which also harbour informative associations between characters (Fig. 2; neurons 3, 4 and 12; neurons 7 and 8).

**Fig. 2**

**3.3 Metabolic and agronomic traits relations**

From the evaluation of integrative neurons (Fig. 2) it could be proposed the existence of different associations between metabolic with agronomic traits. Although these links must be further investigated and experimentally validated, they represent the first step toward new biomarkers as selection tools in breeding programs. Inverse associations between fruit morphology traits (diameter, height, weight, pericarp thickness and locule number) and the primary and secondary metabolites erythritol, glutamate (<sup>1</sup>H-NMR), dodecanoate, calystegine A3 and calystegine B2 are exposed in neuron 1. The group including neighbor neurons 3, 4 and 12 associated acidity, arginine and ethanol contents (neuron 3) with phenylalanine, 5-oxoproline, succinate, juice pH and soluble solids/acidity ratio, with the last three characteristics being inversely associated (neuron 4). Last neuron of this group (neuron 12) included soluble solids and proline contents. Another group, composed by neurons 7 and 8 showed strong association between fruit firmness, shelf-life, glycerate and malate contents (measured both by GC-MS and <sup>1</sup>H-NMR), and inversely with fructose contents (neuron 7). Finally, color index was inversely associated with reflectance, asparagine and threonine contents (neuron 8) (Fig. 2). This integrative analysis of the data suggests novel relations between the analyzed components, exposing those traits which can be considered to design breeding programs. The SOM model applied appears as a valuable tool representing complex high-dimensional input patterns

into a simpler low-dimensional discrete map, easing the results interpretation. Since, it was the first time that this method is used, instead of the correlation analysis commonly performed to evaluate trait relations, we decided to compare both results –SOM and a network correlation analysis- in order to evaluate the consistency between them. Additionally, this comparison gives us a statistical frame to establish an informative neighborhood value (i.e. how many neighbor neurons that contain significant associations must be considered) and also simplify the visualization of the clusters. Correspondence between Pearson and Spearman correlation matrices was assayed with a Mantel test (999 permutations,  $R_{xy}=0.91$ ,  $p<0.001$ ). Since both matrices were highly correlated, we selected the most widely used Pearson correlation coefficient for the network reconstruction. Figure 3 shows the resulting unrooted network where nodes indicate metabolic (circles) and agronomic (diamonds) traits; the number inside the node designates the corresponding  $\omega$ SOM neuron number where this trait was located; and connector lines depict significant correlations between characters ( $p<0.001$ ). Different nodes colors denote metabolic pathways according to KEGG categories (<http://www.genome.jp/kegg/>) and classes of agronomic traits according to Guimera` and Nunes Amaral (2005).

Both SOM and network analysis (Fig. 3) display highly similar results, since correlation among elements of the same and close neighbor neurons were significant. This indicates that  $\omega$ SOM is an accurate tool to reveal trait relations and also that in cases where Pearson correlation significances are not high, the SOM method can expose novel links between variables which are easy to visualize.

Interestingly, the topology of the constructed network is in agreement with the SOM map analyzed. But, it reveals that integrative neurons are not the only informative ones but they must be considered as clusters with their neighbors. Results of network reveals that neuron 1, instead considered as separated neuron, is highly correlated with the elements of neighbor neurons 2 and 10, indicating that  $V_n=1$  is a very informative neighbor value. They are included in one of the major clusters named C1-10-2. On the other hand, neurons 7, 8 and 9 should be integrated with their neighbors 18 and 27 ( $V_n=2$  and 3 respectively), since the elements of all of them are highly correlated. They compose the other major cluster named C7-8-9-18-27. On the contrary, elements in neurons 3, 4 and 12 displayed dissimilar results in SOM and correlation analysis. Even when they are in close proximity in the SOM map they do not form a cluster in the network, indicating that this group is not supported by both analytical methods. This result indicates that informative  $V_n$  value is variable along the map, but the value threshold of 3 is the higher to be considered.

The most cohesive cluster, C1-10-2, represents a core-group with the strongest and most numerous relations between its components. It comprises those traits defining the morphology of the fruits (diameter, weight, height, locule number and pericarp thickness) strongly correlated (negatively) with the amino acids glutamate and aspartate, with the TCA cycle intermediate 2-oxoglutarate, with the fatty acid dodecanoate and also with the contents of the

two measured alkaloids; calystegine A3 and B2; they are all putative regulators of morphology traits. Evidence about the influence of glutamate and 2-oxoglutarate in fruit development via GABA shunt have been proposed by Kisaka et al. (2006). Additionally, downregulation of the 2-oxoglutarate dehydrogenase has major consequences on this process which extend to the end of ripening (Araújo et al. 2012). On the other hand, calystegines exhibit selective inhibition of glycosidases which is universally required for normal cell function (Kvasnicka et al. 2008). Based on their structural similarities to sugars, it has been suggested that in mammal cells calystegines may interact with enzymes of carbohydrate metabolism and could be functional in diets preventing a steep increase in blood glucose after a carbohydrate-rich meal (Jocković et al. 2013). These hints toward mechanistic links, therefore suggest that the relation between these traits found in the current study, will likely provide a valuable tool for tomato breeding programs.

The other central cluster, C7-8-9-18-27, includes agronomic characters related to fruit attributes such as color, firmness and shelf life, traits for which this RIL population was initially selected. They showed strong connections between each other (Fig. 3) and with the vast majority of amino acids and notably with malate and glycerate contents. In this regard, Centeno et al. (2011) previously reported that changes in malate metabolism result in an early water-loss phenotype of the tomato fruits with a consequent effect on post-harvest shelf life. Transgenic lines displaying reduced levels of malate dehydrogenase and fumarase activities presented elevated levels of soluble sugars, suggesting that osmotic potential may be a contributing factor to the water-loss phenotype. Our finding although likely operating by a different mechanism adds to the link between malate metabolism and shelf life of tomato fruits. Additionally, association between asparagine and threonine (neuron 18) was also observed by Sauvage et al. (2014), who found that variations in the contents of these two amino acids were linked to the same SNP locus, annotated as a Copine-like protein. They proposed that this locus is close to one or several pleiotropic effect gene(s) directly involved in the biosynthetic pathway of these amino acids.

Somehow unexpected, fruit biochemical characters (juice pH, acidity and soluble solids) showed few connections with the metabolic complement of the fruits. However, the association of proline with soluble solids resulted particularly interesting because of nitrogenous compounds -in the form of amino acids, peptides and proteins, minerals and pectic substances- are also considered part of the soluble solids. Additionally, in grape, this amino acid is proposed as an indicator of the berry ripeness (Carnevillier et al.1999; Mulas et al.2011), therefore it can be also considered the use of proline as an indicator of fruit ripeness in tomato.

### \*Fig. 3

## 3.4 Assessment of the mode of inheritance of metabolic and agronomic traits

Given that a number of metabolic and agronomic traits differed significantly either between the parents or with respect to the F1 offspring used in this study, we then analyzed these differences aiming to assess the mode of

inheritance of the measured traits. Based on these comparisons we classified each trait into the following categories: recessive, additive, dominant or overdominant (see Material and Methods section). Out of all 12 possible classes proposed by Lisec et al. (2011) nine were identified in our data set (see online resource Fig. S3). Sixty traits could be classified within the mentioned classes: 47 and 13 of the metabolic and agronomic traits measured, respectively (online resource Fig. S3). However, a minor proportion of the traits could not be classified because they showed no significant differences between both parents and the F1 hybrid. The mode of inheritance of all classified traits can be seen in online resource Table S2. A high number of metabolic traits appeared to be recessive (29), where the F1 mean value is equal to the mean value of the *S. lycopersicum* parent. This class includes a large number of amino acids and TCA cycle intermediates. By contrast, the vast majority of the agronomic traits exhibited either dominant or additive mode of inheritance (see online resource Table S2). No agronomic traits were detected to exhibit overdominance and only five metabolic traits showed this mode of inheritance: the amino acids alanine, aspartate, GABA (<sup>1</sup>H-NMR) and glycine and the monosaccharide xylose (online resource Table S2).

Mode of inheritance of yield associated traits in tomato has been analyzed in detail by Semel et al. (2006). By using *S. pennellii* near isogenic lines (NIL) in heterozygosity these authors have shown that four traits exhibited heterosis (overdominance): seed number per plant, fruit number, total yield, and biomass. Other phenotypes, however, such as fruit weight, plant weight, soluble solids content and seed morphology, showed no heterotic effects. Our results are coincident with these findings, even when we use an interspecific F1 hybrid to evaluate the inheritance mode, which can expose more epistatic effects than those detected in NILs. Regarding the mode of inheritance of the metabolic traits, evidences for changes in metabolites in hybrids have been documented in Arabidopsis (Gärtner et al. 2009; Lisec et al. 2009), maize (Lisec et al. 2009; Riedelsheimer et al. 2012) and also in tomato fruits from the above mentioned NIL heterozygotes (Schauer et al. 2008). More than 60% of the metabolic traits analyzed here shared the same mode of inheritance reported by Schauer et al. (2008), but notably none of them presented overdominance. Since our analysis involves a two-parental-hybrid system, the identification of overdominant traits could result from epistatic allelic interactions in the *S. lycopersicum* x *S. pimpinellifolium* hybrid, even considering that these two genomes are evolutionary closer than the donor parent (*S. pennellii*) of the NIL system used by Schauer et al. (2008) (Kamenetzky et al. 2010; Sato et al. 2012). On the other hand, the potential of tomato MAGIC populations as source of genetic diversity was recently reported (Pascual et al. 2014). Another study, using a genome-wide association approach, identifies putative candidate genes involved in the genetic architecture of fruit metabolic traits (Sauvage et al. 2014).

Attractively, from the extensive metabolic characterization conducted here, the link between different traits and the heritability evidenced in this paper, it can be proposed several candidate RILs for the generation of SCH. In this regard, the heat map presented in Fig. 4 displays a summary of the changes for each RIL relative to the cultivated

tomato. When taking the analyzed traits displaying positive effects, RILs 1, 2, 3, 4, 7, 9, 10, 16, 17 and 18 can be considered as candidate lines. All of them displayed higher number of traits with positive effect than the *S. lycopersicum* x *S. pimpinellifolium* hybrid. In agreement with this selection, a recent work (Liberatti et al. 2013) showed that the SCH produced with three of these lines (RILs 1, 9 and 18) display longest shelf life and higher soluble solid values; among the few traits in common between both reports.

On another hand, RIL 4 displayed increased levels of chlorogenate. This antioxidant is considered a very important compound for human health since it can limit low-density lipid (LDL) oxidation protecting against degenerative age-related diseases, and it also removes particularly toxic reactive species by scavenging alkylperoxyl radicals and may prevent carcinogenesis by reducing the DNA damage they cause (Niggeweg et al. 2004).

pH and titratable acidity contribute to tomato flavor intensity, being particularly responsible for the sourness of the fruit. Both measures are, however, very complex and uncertain since the profile and concentration of compounds defining them vary greatly among accessions and varieties (Fernandez-Ruiz et al. 2004). Even when the best pH value for fresh tomato consumers was not determined, commercial lines have a pH range from 4.0 to 4.6. Conversely, pH values below 4.5 are required for processing tomatoes, since they reduce pathogens growth preserving product storability (Bucheli et al. 1999; Foolad 2007). The majority of lines showed pH values lower than 4.5 (RILs 1, 4, 5, 7, 9, 10, 11, 13, 16, 17). Thus, any of them would be appropriate for processing industry.

Overall, assessments presented in this paper demonstrate that wild tomato species offer breeders a great potential for exploiting genetic variability and enhancing fruit weight, shelf life and chemical composition of the fruits. Moreover, <sup>1</sup>H NMR SOM analyses and network construction illustrate the power of these tools when applied in combination into tomato breeding programs. The approach used here opens new routes to the understanding of the nutritional traits which have been undervalued in classical breeding.

## 4 Concluding remarks

In this paper we provide metabolic information from a group of selected RILs, evidenced links among metabolites and agronomic traits and established the inheritance of the traits evaluated. We have highlighted the potential of the metabolic analyses to assess agronomical important traits in tomato. Additionally, this study revealed that *S. pimpinellifolium*, a wild tomato species, introduced a broad metabolic variability which can be of importance for tomato breeding programs. All these information assessed with modern analytical tools, facilitated the identification of candidate lines for breeding programs towards the tomato nutritional quality improvement.

## Acknowledgments

M.G.L was recipient of a fellowship of Consejo Nacional de Investigaciones Científicas y Técnicas (Argentina). This work was partially supported with grants from Instituto Nacional de Tecnología Agropecuaria, Consejo Nacional de Investigaciones Científicas y Técnicas, Agencia Nacional de Promoción Científica y Tecnológica (Argentina) and from the Max Planck Society (Germany).

**Disclosures** This work was carried out in compliance with current laws governing genetic experimentation in Argentina. The authors declared no conflict of interest.

## References

- Abriata, L. A. (2012). Utilization of NMR spectroscopy to study biological fluids and metabolic processes: Two introductory activities. *Concepts in Magnetic Resonance Part A*, 40A(4), 171–178. doi:10.1002/cmr.a.21235
- Alba, R., Payton, P., Fei, Z., McQuinn, R., Debbie, P., Martin, G. B., et al. (2005). Transcriptome and selected metabolite analyses reveal multiple points of ethylene control during tomato fruit development. *Plant Cell*, 17(11), 2954–65. doi:10.1105/tpc.105.036053
- Araújo, W. L., Tohge, T., Osorio, S., Lohse, M., Balbo, I., Krahnert, I., et al. (2012). Antisense inhibition of the 2-oxoglutarate dehydrogenase complex in tomato demonstrates its importance for plant respiration and during leaf senescence and fruit maturation. *Plant Cell*, 24(6), 2328–51. doi:10.1105/tpc.112.099002
- Asano, N., Kato, A., Matsui, K., Watson, A. A., Nash, R. J., Molyneux, R. J., et al. (1997). The effects of calystegines isolated from edible fruits and vegetables on mammalian liver glycosidases. *Glycobiology*, 7(8), 1085–8. <http://www.ncbi.nlm.nih.gov/pubmed/9455909>
- Bai, Y., & Lindhout, P. (2007). Domestication and breeding of tomatoes: what have we gained and what can we gain in the future? *Annals of Botany*, 100(5), 1085–94. doi:10.1093/aob/mcm150
- Bekkouche, K., Daali, Y., Cherkaoui, S., Veuthey, J. L., & Christen, P. (2001). Calystegine distribution in some solanaceous species. *Phytochemistry*, 58(3), 455–62. <http://www.ncbi.nlm.nih.gov/pubmed/11557078>
- Bermúdez, L., de Godoy, F., Baldet, P., Demarco, D., Osorio, S., Quadrana, L., et al. (2014). Silencing of the tomato sugar partitioning affecting protein (SPA) modifies sink strength through a shift in leaf sugar metabolism. *The Plant Journal*, 77(5), 676–87. doi:10.1111/tbj.12418
- Bermúdez, L., Urias, U., Milstein, D., Kamenetzky, L., Asis, R., Fernie, A. R., et al. (2008). A candidate gene survey of quantitative trait loci affecting chemical composition in tomato fruit. *Journal of Experimental Botany*, 59(10), 2875–90. doi:10.1093/jxb/ern146
- Bretó, M. P., Asins, M. J., & Carbonell, E. A. (1993). Genetic variability in *Lycopersicon* species and their genetic relationships. *Theoretical and Applied Genetics*, 86(1). doi:10.1007/BF00223815
- Bucheli, P., Voirol, E., de la Torre, R., López, J., Rytz, a, Tanksley, S. D., & Pétiard, V. (1999). Definition of nonvolatile markers for flavor of tomato (*Lycopersicon esculentum* Mill.) as tools in

selection and breeding. *Journal of Agricultural and Food Chemistry*, 47(2), 659–64.  
<http://www.ncbi.nlm.nih.gov/pubmed/10563949>

- Carnevillier, V., Schlich, P., Guerreau, J., Charpentier, C., Feuillat, M. (1999). Characterization of the production regions of Chardonnay wines by analysis of free amino acids. *Vitis* 38(1), 37–42.
- Carreno-Quintero, N., Bouwmeester, H. J., & Keurentjes, J. J. B. (2013). Genetic analysis of metabolome-phenotype interactions: from model to crop species. *Trends in Genetics* 29(1), 41–50. doi:10.1016/j.tig.2012.09.006
- Causse, M., Duffé, P., Gomez, M. C., Buret, M., Damidaux, R., Zamir, D., et al. (2004). A genetic map of candidate genes and QTLs involved in tomato fruit size and composition. *Journal of Experimental Botany*, 55(403), 1671–85. doi:10.1093/jxb/erh207
- Causse, M., Saliba-Colombani, V., Lecomte, L., Duffé, P., Rousselle, P., & Buret, M. (2002). QTL analysis of fruit quality in fresh market tomato: a few chromosome regions control the variation of sensory and instrumental traits. *Journal of Experimental Botany*, 53(377), 2089–2098. doi:10.1093/jxb/erf058
- Centeno, D. C., Osorio, S., Nunes-Nesi, A., Bertolo, A. L. F., Carneiro, R. T., Araújo, W. L., et al. (2011). Malate plays a crucial role in starch metabolism, ripening, and soluble solid content of tomato fruit and affects postharvest softening. *Plant Cell*, 23(1), 162–84. doi:10.1105/tpc.109.072231
- Di Rienzo, J. A., Casanoves, F., Balzarini, M. G., Gonzalez, L., Tablada, M., & Robledo, C. W. (2011). InfoStat version 2011. <http://www.infostat.com.ar>
- Fernandez-Ruiz, V., Sanchez-Mata, M. C., Camara, M., Torija, M. E., Chaya, C., Galiana-Balaguer, L., et al. (2004). Internal quality characterization of fresh tomato fruits. *HortScience*, 39(2), 339–345. <http://hortsci.ashspublications.org/content/39/2/339.abstract>. Accessed 1 July 2013
- Fernie, A. R., & Schauer, N. (2009). Metabolomics-assisted breeding: a viable option for crop improvement? *Trends in Genetics*, 25(1), 39–48. doi:10.1016/j.tig.2008.10.010
- Fitzpatrick, T. B., Basset, G. J. C., Borel, P., Carrari, F., DellaPenna, D., Fraser, P. D., et al. (2012). Vitamin deficiencies in humans: can plant science help? *Plant Cell*, 24(2), 395–414. doi:10.1105/tpc.111.093120
- Foolad, M. R. (2007). Genome mapping and molecular breeding of tomato. *International Journal of Plant Genomics*, 2007, 64358. doi:10.1155/2007/64358
- Foolad, M. R., Chen, F. Q., & Lin, G. Y. (1998). RFLP mapping of QTLs conferring salt tolerance during germination in an interspecific cross of tomato. *Theoretical and Applied Genetics*, 97(7), 1133–1144. doi:10.1007/s001220051002
- Foolad, M. R., & Lin, G. Y. (1997). Absence of a genetic relationship between salt tolerance during seed germination and vegetative growth in tomato. *Plant Breeding*, 116(4), 363–367. doi:10.1111/j.1439-0523.1997.tb01013.x
- Fulton, T. M., Beck-Bunn, T., Emmatty, D., Eshed, Y., Lopez, J., Petiard, V., et al. (1997). QTL analysis of an advanced backcross of *Lycopersicon peruvianum* to the cultivated tomato and comparisons with QTLs found in other wild species. *Theoretical and Applied Genetics*, 95(5-6), 881–894. doi:10.1007/s001220050639
- Fulton, T. M., Grandillo, S., Beck-Bunn, T., Fridman, E., Frampton, A., Lopez, J., et al. (2000). Advanced backcross QTL analysis of a *Lycopersicon esculentum* X *Lycopersicon parviflorum* cross. *Theoretical and Applied Genetics*, 100(7), 1025–1042. doi:10.1007/s001220051384

- Gallo, M., Rodríguez, G. R., Zorzoli, R., & Pratta, G. R. (2011). Ligamiento entre caracteres cuantitativos de calidad de fruto y perfiles polipeptídicos del pericarpio en dos estados de madurez en tomate. *Revista de la Facultad de Ciencias Agrarias Universidad Nacional de Cuyo*, 43(2), 145–156.
- Gärtner, T., Steinfath, M., Andorf, S., Liseč, J., Meyer, R. C., Altmann, T., et al. (2009). Improved heterosis prediction by combining information on DNA- and metabolic markers. *PLoS One*, 4(4), e5220. doi:10.1371/journal.pone.0005220
- Gechev, T. S., Hille, J., Woerdenbag, H. J., Benina, M., Mehterov, N., Toneva, V., et al. (2014). Natural products from resurrection plants: potential for medical applications. *Biotechnology Advances*, 32(6), 1091–101. doi:10.1016/j.biotechadv.2014.03.005
- Giovannoni, J. (2001). Molecular biology of fruit maturation and ripening. *Annual Review of Plant Physiology and Plant Molecular Biology*, 52, 725–749. doi:10.1146/annurev.arplant.52.1.725
- Giovannoni, J. (2004). Genetic regulation of fruit development and ripening. *Plant Cell*, 16, S170–S180
- Goff, S. a, & Klee, H. J. (2006). Plant volatile compounds: sensory cues for health and nutritional value? *Science*, 311(5762), 815–9. doi:10.1126/science.1112614
- Goulet, C., Mageroy, M. H., Lam, N. B., Floystad, A., Tieman, D. M., & Klee, H. J. (2012). Role of an esterase in flavor volatile variation within the tomato clade. *Proceedings of the National Academy of Sciences of the United States of America*, 109(46), 19009–14. doi:10.1073/pnas.1216515109
- Guimerà, R., & Nunes Amaral, L. A. (2005). Functional cartography of complex metabolic networks. *Nature*, 433(7028), 895–900. doi:10.1038/nature03288
- Hermann, A. & Schauer, N. (2013) Metabolomics-assisted plant breeding. In Handbook of Plant Metabolomics (ed. Wolfram Weckwerth, Guenter Kahl) John Wiley & Sons, pp 247-254
- Hall, R. D., Brouwer, I. D., & Fitzgerald, M. A. (2008). Plant metabolomics and its potential application for human nutrition. *Physiologia Plantarum*, 132(2), 162–75. doi:10.1111/j.1399-3054.2007.00989.x
- Hu, C., Shi, J., Quan, S., Cui, B., Kleessen, S., Nikoloski, Z., et al. (2014). Metabolic variation between japonica and indica rice cultivars as revealed by non-targeted metabolomics. *Scientific Reports*, 4, 5067. doi:10.1038/srep05067
- Jocković, N., Fischer, W., Brandsch, M., Brandt, W., & Dräger, B. (2013). Inhibition of human intestinal  $\alpha$ -glucosidases by calystegines. *Journal of Agricultural and Food Chemistry*, 61(23), 5550–7. doi:10.1021/jf4010737
- Kamenetzky, L., Asís, R., Bassi, S., de Godoy, F., Bermúdez, L., Fernie, A. R., et al. (2010). Genomic analysis of wild tomato introgressions determining metabolism- and yield-associated traits. *Plant Physiology*, 152(4), 1772–86. doi:10.1104/pp.109.150532
- Kisaka, H., Kida, T., & Miwa, T. (2006). Antisense suppression of glutamate decarboxylase in tomato (*Lycopersicon esculentum* L.) results in accumulation of glutamate in transgenic tomato fruits, *Plant Biotechnology*, 274, 267–274.
- Klee, H. J. (2013). Purple tomatoes: longer lasting, less disease, and better for you. *Current Biology*, 23(12), R520–R521. doi:10.1016/j.cub.2013.05.010
- Kvasnicka, F., Jockovic, N., Dräger, B., Sevcík, R., Cepl, J., & Voldrich, M. (2008). Electrophoretic determination of calystegines A3 and B2 in potato. *Journal of Chromatography. A*, 1181(1-2), 137–44. doi:10.1016/j.chroma.2007.12.037
- Lee, J. M., Joung, J.-G., McQuinn, R., Chung, M.-Y., Fei, Z., Tieman, D., et al. (2012). Combined transcriptome, genetic diversity and metabolite profiling in tomato fruit reveals that the ethylene

response factor SIERF6 plays an important role in ripening and carotenoid accumulation. *The Plant Journal*, 70(2), 191–204. doi:10.1111/j.1365-313X.2011.04863.x

Liberatti, D., Rodriguez, G., Zorzoli, R., & Pratta, G. R. (2013). Tomato second cycle hybrids differ from parents at three levels of genetic variation. *International Journal of Plant Breeding*, 7(1), 1–7. [http://www.globalsciencebooks.info/JournalsSup/images/Sample/IJPB\\_7\(1\)1-6o.pdf](http://www.globalsciencebooks.info/JournalsSup/images/Sample/IJPB_7(1)1-6o.pdf).

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., et al. (2014). Genomic analyses provide insights into the history of tomato breeding. *Nature Genetics*, 46(11), 1220–1226. doi:10.1038/ng.3117

Lisec, J., Römisch-Margl, L., Nikoloski, Z., Piepho, H.-P., Giavalisco, P., Selbig, J., et al. (2011). Corn hybrids display lower metabolite variability and complex metabolite inheritance patterns. *The Plant Journal*, 68(2), 326–36. doi:10.1111/j.1365-313X.2011.04689.x

Lisec, J., Schauer, N., Kopka, J., Willmitzer, L., & Fernie, A. R. (2006). Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nature Protocols*, 1(1), 387–96. doi:10.1038/nprot.2006.59

Lisec, J., Steinfath, M., Meyer, R. C., Selbig, J., Melchinger, A. E., Willmitzer, L., & Altmann, T. (2009). Identification of heterotic metabolite QTL in *Arabidopsis thaliana* RIL and IL populations. *The Plant Journal*, 59(5), 777–88. doi:10.1111/j.1365-313X.2009.03910.x

Luedemann, A., Strassburg, K., Erban, A., & Kopka, J. (2008). TagFinder for the quantitative analysis of gas chromatography--mass spectrometry (GC-MS)-based metabolite profiling experiments. *Bioinformatics*, 24(5), 732–7. doi:10.1093/bioinformatics/btn023

Mattoo, A. K., Sobolev, A. P., Neelam, A., Goyal, R. K., Handa, A. K., & Segre, A. L. (2006). Nuclear magnetic resonance spectroscopy-based metabolite profiling of transgenic tomato fruit engineered to accumulate spermidine and spermine reveals enhanced anabolic and nitrogen-carbon interactions. *Plant Physiology*, 142(4), 1759–70. doi:10.1104/pp.106.084400

Miller, J. C., & Tanksley, S. D. (1990). RFLP analysis of phylogenetic relationships and genetic variation in the genus *Lycopersicon*. *Theoretical and Applied Genetics*, 80(4). doi:10.1007/BF00226743

Milone, D., Stegmayer, G., Gerard, M., Kamenetzky, L., López, M., & Carrari, F. (2010). Métodos de agrupamiento no supervisado para la integración de datos genómicos y metabólicos de múltiples líneas de introgresión. *Inteligencia Artificial*, 13(44), 56–66. doi:10.4114/ia.v13i44.1046

Moose, S. P., & Mumm, R. H. (2008). Molecular plant breeding as the foundation for 21st century crop improvement. *Plant Physiology*, 147(3), 969–77. doi:10.1104/pp.108.118232

Mulas, G., Galaffu, M. G., Pretti, L., Nieddu, G., Mercenaro, L., Tonelli, R., & Anedda, R. (2011). NMR analysis of seven selections of vermentino grape berry: metabolites composition and development. *Journal of Agricultural and Food Chemistry*, 59(3), 793–802. doi:10.1021/jf103285f

Niggeweg, R., Michael, A. J., & Martin, C. (2004). Engineering plants with increased levels of the antioxidant chlorogenic acid. *Nature Biotechnology*, 22(6), 746–54. doi:10.1038/nbt966

Pascual, L., Desplat, N., Huang, B. E., Desgroux, A., Bruguier, L., Bouchet, J.-P., et al. (2014). Potential of a tomato MAGIC population to decipher the genetic control of quantitative traits and detect causal variants in the resequencing era. *Plant Biotechnology Journal*. doi:10.1111/pbi.12282

Paterson, A. H., Lander, E. S., Hewitt, J. D., Peterson, S., Lincoln, S. E., & Tanksley, S. D. (1988). Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature*, 335(6192), 721–6. doi:10.1038/335721a0

Pratta, G. R., Rodríguez, G. R., Zorzoli, R., Picardi, L. A., & Valle, E. M. (2011a). Biodiversity in a tomato germplasm for free amino acid and pigment content of ripening fruits. *American Journal of Plant Sciences*, 02(02), 255–261. doi:10.4236/ajps.2011.22027

- Pratta, G. R., Rodriguez, G. R., Zorzoli, R., Valle, E. M., & Picardi, L. A. (2011b). Phenotypic and molecular characterization of selected tomato recombinant inbred lines derived from the cross *Solanum lycopersicum* x *S. pimpinellifolium*. *Journal of Genetics*, *90*(2), 229–37. <http://www.ncbi.nlm.nih.gov/pubmed/21869471>
- Price, A. H. (2006). Believe it or not, QTLs are accurate! *Trends in Plant Science*, *11*(5), 213–6. doi:10.1016/j.tplants.2006.03.006
- Rambla, J. L., Tikunov, Y. M., Monforte, A. J., Bovy, A. G., & Granell, A. (2014). The expanded tomato fruit volatile landscape. *Journal of Experimental Botany*, *65*(16), 4613–23. doi:10.1093/jxb/eru128
- Rao, J., Cheng, F., Hu, C., Quan, S., & Lin, H. (2014). Metabolic map of mature maize kernels. *Metabolomics*, doi:10.1007/s11306-014-0624-3
- Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., et al. (2012). Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nature Genetics*, *44*(2), 217–20. doi:10.1038/ng.1033
- Rodríguez, G. R., Pratta, G. R., Zorzoli, R., Picardi, L. A., & Divergent-antagonistic, L. A. (2006). Recombinant lines obtained from an interspecific cross between *Lycopersicon* species selected by fruit weight and fruit shelf life. *Journal of the American Society for Horticultural Science*, *131*(5), 651–656.
- Roessner, U., Luedemann, a, Brust, D., Fiehn, O., Linke, T., Willmitzer, L., & Fernie, a. (2001). Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell*, *13*(1), 11–29.
- Roessner-Tunali, U., Lytovchenko, A., Carrari, F., Bruedigam, C., Granot, D., & Fernie, A. R. (2003). Metabolic profiling of transgenic tomato plants overexpressing hexokinase reveals that the influence of hexose phosphorylation diminishes during fruit development. *Plant Physiology*, *133*(2), 84–99. doi:10.1104/pp.103.023572.84
- Saeed, A. I., Bhagabati, N. K., Braisted, J. C., Liang, W., Sharov, V., Howe, E. A., et al. (2006). TM4 microarray software suite. *Methods in Enzymology*, *411*(null), 134–93. doi:10.1016/S0076-6879(06)11009-5
- Saito, K., & Matsuda, F. (2010). Metabolomics for functional genomics, systems biology, and biotechnology. *Annual Review of Plant Biology*, *61*, 463–89. doi:10.1146/annurev.arplant.043008.092035
- Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., et al. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, *485*(7400), 635–41. doi:10.1038/nature11119
- Sauvage, C., Segura, V., Bauchet, G., Stevens, R., Do, P. T., Nikoloski, Z., et al. (2014). Genome-Wide association in tomato reveals 44 candidate loci for fruit metabolic traits. *Plant Physiology*, *165*(3), 1120–1132. doi:10.1104/pp.114.241521
- Schauer, N., Semel, Y., Balbo, I., Steinfath, M., Reptsilber, D., Selbig, J., et al. (2008). Mode of inheritance of primary metabolic traits in tomato. *Plant Cell*, *20*(3), 509–23. doi:10.1105/tpc.107.056523
- Schauer, N., Zamir, D., & Fernie, A. R. (2005). Metabolic profiling of leaves and fruit of wild species tomato: a survey of the *Solanum lycopersicum* complex. *Journal of Experimental Botany*, *56*(410), 297–307. doi:10.1093/jxb/eri057
- Semel, Y., Nissenbaum, J., Menda, N., Zinder, M., Krieger, U., Issman, N., et al. (2006). Overdominant quantitative trait loci for yield and fitness in tomato. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(35), 12981–6. doi:10.1073/pnas.0604635103

- Sorrequieta, A., Abriata, L., Boggio, S., & Valle, E. (2013). Off-the-vine ripening of tomato fruit causes alteration in the primary metabolite composition. *Metabolites*, 3(4), 967–978. doi:10.3390/metabo3040967
- Stegmayer, G., Milone, D., Kamenetzky, L., Lopez, M., & Carrari, F. (2009). Neural network model for integration and visualization of introgressed genome and metabolite data. *2009 International Joint Conference on Neural Networks*, 2983–2989. doi:10.1109/IJCNN.2009.5179039
- Swamy, B. P. M., & Sarla, N. (2008). Yield-enhancing quantitative trait loci (QTLs) from wild species. *Biotechnology Advances*, 26(1), 106–20. doi:10.1016/j.biotechadv.2007.09.005
- Tieman, D., Bliss, P., McIntyre, L. M., Blandon-Ubeda, A., Bies, D., Odabasi, A. Z., et al. (2012). The chemical interactions underlying tomato flavor preferences. *Current Biology*, 22(11), 1035–1039. doi:10.1016/j.cub.2012.04.016
- Tigchelaar, E. C. (1986). *Breeding vegetable crops*. (M. J. Basset, Ed.) (pp. 135–170). Westport, CT, USA: AVI Publishing Company.
- Wahyuni, Y., Ballester, A.-R., Tikunov, Y., Vos, R. C. H., Pelgrom, K. T. B., Maharijaya, A., et al. (2012). Metabolomics and molecular marker analysis to explore pepper (*Capsicum* sp.) biodiversity. *Metabolomics*, 9(1), 130–144. doi:10.1007/s11306-012-0432-6
- Warren, G. F. (1998). Spectacular increases in crop yields in the United States in the twentieth century. *Weed Technology*, 12(4), 752–760.
- Willcox, J. K., Catignani, G. L., & Lazarus, S. (2003). Tomatoes and cardiovascular health. *Critical Reviews in Food Science and Nutrition*, 43(1), 1–18. doi:10.1080/10408690390826437
- Yang, W., Bai, X., Kabelka, E., Eaton, C., Kamoun, S., Knaap, E. Van Der, & Francis, D. (2004). Discovery of single nucleotide polymorphisms in *Lycopersicon esculentum* by computer aided analysis of expressed sequence tags. *Molecular Breeding*, 14(1), 21–34.
- Zorzoli, R., Pratta, G. R., & Picardi, L. A. (2000). Variabilidad genética para la vida postcosecha y el peso de los frutos en tomate para familias F3 de un híbrido interespecífico. *Pesquisa Agropecuária Brasileira*, 35(12), 2423–2427. doi:10.1590/S0100-204X2000001200013

### Electronic supplementary material

**Supplementary Tables** List of properties for metabolite (60) determination (Table S1). Relative values of mature fruit metabolic contents (60) and agronomic (14) trait from all the material analyzed are provided (Table S2). Details of the components of each neuron in the \*omeSOM 9x9 map are also presented (Table S3).

**Supplementary Fig. S1** Representative <sup>1</sup>H-NMR spectrum of tomato pericarp extract in buffered D<sub>2</sub>O. Known compounds are annotated according to Table S1. The <sup>1</sup>H chemical shifts used for identification/quantification of metabolites were determined at pH 7.4 and expressed as relative values to that of TSP at 0 ppm. Ala: alanine (doublet at 1.45 ppm), Asn: asparagine (multiplet at 2.82), Asp: aspartate (doublet of doublet at 2.76), Citrate (doublet of doublets at 2.51), Ethanol (triplet at 1.15), Frc: fructose (multiplet at 4.08), GABA (multiplet at 1.84), Glc: glucose (doublet at 4.62), Glu: glutamate (multiplet at 2.05), Gln: glutamine (multiplet at 2.44), Ile: Isoleucine (doublet at 0.98), Malate (doublet of

doublets at 4.27), Methanol (singlet at 3.31), Phe:phenylalanine (multiplet at 7.38), Pyr: Pyruvate (singlet at 2.35), Suc: sucrose (doublet at 5.38), Thr: Threonine (doublet at 1.3), Trp: Tryptophan (doublet at 7.51), Val: valine (doublet at 1.01).

**Supplementary Fig. S2** Values of relative distance between the same metabolite (16) measured by GC-MS and <sup>1</sup>H-NMR across different map sizes evaluated are plotted.

**Supplementary Fig. S3** Different mode of inheritance of metabolic (60) and agronomic (14) traits. Traits (metabolic –blue bars- and agronomic –red bars-) were classified according their mode of inheritance following the analysis proposed by Liseć et al. (2011) (see Material and Method section).

### Figure captions

**Fig. 1** Metabolic variations in polar extracts of ripe tomato pericarps from a RIL population derived from a cross of *S. lycopersicum* (cv. Caimanta) x *S. pimpinellifolium* (LA722). **a.** Proportion of significantly altered metabolites (increase or decrease relative to the cultivated species) for each line (and the wild parent and F1 progeny) for each chemical category. **b.** Hierarchical cluster analysis (represented by a heat map) of the quantitative variation detected in all metabolic and agronomic traits measured in the RIL population. Amino acids are depicted in the three-letter code. DHA: dehydroascorbate. PT: pericarp thickness. SSC: soluble solid contents. TA: titratable acidity

**Fig. 2** \*omesom model of 81 neurons grouping 60 different metabolic and 14 agronomic traits from the 18 analyzed RIL, the parent *S. pimpinellifolium* (LA722) and the F1 hybrid. Metabolites were measured in polar extracts from tomato pericarps. Directed and inverted relations are shown in the left and right quadrants respectively. Black neurons group metabolic and agronomic traits, blue and red neurons group only metabolic and only agronomic traits, respectively. Histograms showing components variation along the lines analyzed are presented for those neurons grouping at least one agronomic trait. References: Neuron 1: fruit diameter (inv, blue line), height (inv, green line), weight (inv, red line), pericarp thickness (inv, light blue line), locule number (inv, violet line), calystegine A3 (blue dotted line), calystegine B2 (green dotted line), dodecanoate (red dotted line), erythritol (light blue dotted line) and glutamate (<sup>1</sup>H-NMR) (violet dotted line). Neuron 3: acidity (blue line), arginine (blue dotted line) and ethanol (green dotted line) (<sup>1</sup>H-NMR). Neuron 4: pH (inv, blue line), soluble solids (SSC)/acidity (inv, green line), phenylalanine (blue dotted line), 5-oxoproline (green dotted line) and succinate (inv, red dotted line). Neuron 7: Fruit shelf life (blue line), Fruit firmness (green line), fructose (inv, blue dotted line), glycerate (green dotted line), malate (red dotted line) and malate (<sup>1</sup>H-NMR) (light blue dotted line). Neuron 8: reflectance (blue line),

color index (inv, green line), asparagine (blue dotted line) and threonine (green dotted line). Neuron 12: Soluble solids (blue line) and proline (blue dotted line). Neuron 31: Fruit shape (blue line)

**Fig. 3** Network analysis of metabolic and agronomic traits from the 18 analyzed RIL, the parent *S. pimpinellifolium* (LA722) and the F1 hybrid. Metabolites were measured in polar extracts from tomato pericarps. Circles and diamonds indicate metabolic and agronomic traits, respectively. Connections represent significant Pearson correlation ( $p < 0.001$ ) between edges, blue and black lines depict negative and positive correlations, respectively. Line thickness is proportional to the correlation coefficient values (ranging between -0.94 to 0.97). Numbers inside the nodes indicate \*omesom neuron number and Vn value indicate neighborhood between neurons calculated as described in Milone et al, (2010). Metabolic nodes are coloured according to KEGG pathways categories ([www.genome.jp/kegg/](http://www.genome.jp/kegg/))

**Fig. 4** Differences in metabolic or agronomic traits relative to the domesticated parental line (*S. lycopersicum* cv. Caimanta). Metabolites were measured in polar extracts from tomato pericarps. Red and blue squares represent significant ( $p < 0.01$  by t-test) increase or decrease of each trait relative to *S. lycopersicum* values. Black squares indicate missing data. Values below the graph indicate the number of total traits presenting changes and those with positive effects (increments)

1

**Table 1** Comparative analysis of metabolite profiles from tomato pericarps obtained by GC-MS -polar extracts- and <sup>1</sup>H-NMR methods. Comparison between GC-MS and <sup>1</sup>H-NMR methods was performed using Pearson correlation analysis. Values represent the correlation coefficient between both methods considering all data and partitioned by metabolites. p values indicate significant correlation (p<0.01)

	Pearson coefficient	P value
<sup>1</sup> H-NMRGC-MS	0.60	<0.0001
Alanine	0.75	0.0002
Asparagine	0.82	<0.0001
Aspartate	0.90	<0.0001
GABA	0.61	0.0040
Citrate	0.62	0.0035
Fructose	0.39	0.0857
Glucose	0.49	0.0267
Glutamate	0.84	<0.0001
Glutamine	0.53	0.0165
Isoleucine	0.67	0.0012
Malate	0.83	<0.0001
Phenylalanine	0.60	0.0051
Sucrose	-0.03	0.9167
Threonine	0.79	<0.0001
Tryptophan	-0.17	0.5609
Valine	0.75	0.0001

**Table 2** Relative distance ( $Vn_r$ ) between the same metabolite measured by GC-MS –polar extracts- and  $^1H$ -NMR and analyzed by \*omeSOM with different map sizes. Relative distance was calculated as  $Vn / \text{total neurons in the map}$ . Metabolites were obtained from tomato pericarps.

Metabolite	Map size				
	7x7	8x8	<b>9x9</b>	10x10	11x11
Alanine	0.08	0.08	<b>0</b>	0.07	0.00826
Asparagine	0.02	0.02	<b>0.025</b>	0.01	0.00826
Aspartate	0	0	<b>0.012</b>	0.01	0.01653
GABA	0	0.02	<b>0.037</b>	0.02	0.00826
Citrate	0.04	0.09	<b>0.086</b>	0.07	0.06612
Fructose	0.16	0.14	<b>0.136</b>	0.07	0.04959
Glucose	0.04	0.02	<b>0.012</b>	0.05	0.04959
Glutamate	0.02	0.02	<b>0.012</b>	0.01	0.00826
Glutamine	0.02	0.02	<b>0.025</b>	0.02	0.02479
Isoleucine	0.02	0.02	<b>0.025</b>	0.01	0.00826
Malate	0	0	<b>0</b>	0	0
Phenylalanine	0.08	0.08	<b>0.086</b>	0.06	0.05785
Sucrose	0.08	0.09	<b>0.086</b>	0.08	0.05785
Threonine	0.04	0.03	<b>0.025</b>	0.02	0.01653
Tryptophan	0.06	0.08	<b>0.074</b>	0.04	0.03306
Valine	0.02	0.02	<b>0.025</b>	0.01	0.02479