Euphytica (2014) 200:429–439
DOI 10.1007/s10681-014-1179-0

Modeling of tomato fruits into nine shape categories using elliptic fourier shape modeling and Bayesian classification of contour morphometric data

Sofia Visa · Chunxue Cao · Brian McSpadden Gardener · Esther van der Knaap

Received: 24 March 2014 / Accepted: 3 June 2014 / Published online: 19 June 2014

Abstract Classification and characterization of the shape of plant organs are important tools for plant biologists, breeders and growers. Here we use boundary measurements, i.e. contour morphometric data, of scanned tomato fruits in conjunction with elliptic Fourier shape modeling and Bayesian classification techniques to find the optimum number of shape categories. Our findings show that there are nine computationally and visually distinct tomato shape categories: ellipsoid, flat, heart, long, long rectangular, rectangular, round, obovoid, and oxheart. Analyses of fruits from a diverse set of tomato accessions demonstrate that some varieties carry fruits that conform to predominantly one shape category while others carry fruits that conform to multiple shape categories. In particular the categories oxheart and long rectangular feature fruit that tend to equivalently fit several categories of shape, while the flat and obovoid categories contain fruit that consistently conform exclusively to a single category. The findings show that elliptic Fourier shape modeling and Bayesian classification provide an excellent tool for further in depth analyses of fruit shape variation that may occur across varieties and/or result from growth under different environmental conditions.

Keywords Classification · Contour morphometric analysis · Fruit shape · Modelling · Tomato · Uniformity

Electronic supplementary material The online version of this article (doi:10.1007/s10681-014-1179-0) contains supplementary material, which is available to authorized users.

S. Visa
Department of Mathematics and Computer Science, The College of Wooster, Wooster, OH, USA

C. Cao · B. M. Gardener
Department of Plant Pathology, The Ohio State University, Wooster, OH, USA

E. van der Knaap (✉)
Department of Horticulture and Crop Science, The Ohio State University, Wooster, OH, USA
e-mail: Vanderknaap.1@osu.edu

Introduction

Selections of fruit and vegetable crops resulted in numerous varieties that differ in shape and size of the produce (Paran and van der Knaap 2007; Pickersgill 2007). The dimensions of the produce are important selection criteria when developing new varieties that are targeted to fill specific market needs. For example, rectangular and blocky tomato fruit are typically used in the processing industry for the production of tomato paste and sauce, as well as canned and diced tomatoes. Those fruit are harvested mechanically and the shapes of the produce are critical to prevent the fruit from rolling from conveyer belts. Other fruit shapes, such as ellipsoid, round and heart, are primarily found in

varieties targeted to the fresh market industry to be eaten fresh as in salads. Flat and large tomatoes are also used in the fresh market industry and are typically used as slicing tomatoes for sandwiches and hamburgers.

Classification of the tomato plants based on fruit shape is critically important for the correct categorization of the different varieties (Union for the Protection of New Varieties of Plants (UPOV) and the International Plant Genetic Resources Institute (IPGRI) (IPGRI 1996; UPOV 2001). In recent years, several genes that control tomato fruit shape have been cloned (Cong et al. 2008; Liu et al. 2002; Munos et al. 2011; Xiao et al. 2008). The majority of fruit shape diversity found in the germplasm is controlled by these four known genes: SUN and OVATE regulating fruit elongation; and LOCULE NUMBER and FASCIATED regulation locule number and flat shape (Rodriguez et al. 2011). However, potentially many more genes with subtle effects on organ shape underlie the entire tomato morphological diversity (Rodriguez et al. 2013).

Fruit shape can be analyzed using morphometric studies, which are defined as the quantitative analysis of a biological form (Bookstein 1982; Rohlf 1990). Morphometric analysis uses the position of and distance between landmarks of the object as the source of morphological data. These modeling methods have been used to investigate phenotypic variations (Chandler and Crisp 1998; Henderson 2006; Klingenberg and Monteiro 2005; Lihova et al. 2004; Sonibare et al. 2004; Weight et al. 2008), as well as evolutionary analyses (Borba et al. 2007; Langlade et al. 2005). Morphometric analyses have also been applied in genetic studies of anatomy in animals (Cheverud 1996; Klingenberg et al. 2001; Weber et al. 1999) as well as plants (Dryden and Mardia 1998; Langlade et al. 2005; Perez–Perez et al. 2002). The advantage of morphometric analyses is that they require neither prior knowledge nor predetermined notions of the shape features that will be measured. Therefore, morphometric analyses offer less biased assessment compared to other evaluations of shape.

Previously, we proposed a revised tomato variety classification scheme found at UPOV and IPGRI. The revised classification was based on visual inspection and fruit shape analyses using the Tomato Analyzer software (Brewer et al. 2006; Gonzalo et al. 2009; Rodriguez et al. 2010). This resulted in the classification of tomato varieties into eight categories: flat, round, heart, oxheart, long, rectangular, obovoid and ellipsoid. This classification was supported by the analysis of 36 shape attributes implemented in the Tomato Analyzer software (Rodriguez et al. 2011). The most important attributes controlling the shape were: fruit shape index (ratio of fruit length over width), distal end protrusion, widest width position, proximal end blockiness at 20 %, distal angle at 20 %, rectangular and proximal eccentricity. However, the classification may have been biased since the categorization was determined manually and the Tomato Analyzer software captures only a limited number of shape attributes. Moreover, varieties were classified into one category of shape, even though a particular variety may have carried differently shaped fruits. It is also likely that certain tomato varieties may carry more uniformly shaped fruit, conforming to one shape category, whereas others carry variably shaped fruit, conforming to several shape categories. The uniformity of the produce is highly desirable, for example when the produce is harvested mechanically and should fit certain dimensions that are optimized for a particular production system.

The goal of this research was to model the tomato fruit shapes by using contour morphometric data without taking into account the classification of the variety. The fruit boundary information was used to derive elliptic shape descriptors which model closed contours and, in addition, were invariant to scale and rotation. These descriptors were further classified to identify the shape categories, thereby facilitating the identification of distinct classes represented by tomato varieties from a phenotypically diverse germplasm collection. The contour morphometric analysis showed that the classes identified previously were largely upheld. Additionally, this new approach allowed for the identification of shape classes and tomato varieties that carried fruit that varied significantly in uniformity.

Materials and methods

Plant materials

Forty eight phenotypically distinct tomato varieties that represented the eight classifications (Rodriguez et al. 2011) (supplemental Table 1) were grown in three locations: Mountain Horticultural Crops Research and Extension Center in Mills River, North Carolina; Ohio

Table 1 Statistics of the final nine classes from Fig. 3b

Class	No. of examples in class	Error	Variance
1 Round	189	0.13	0.0036
2 Rectangular	99	0.06	0.0018
3 Ellipsoid	96	0.10	0.0051
4 Flat	94	0.15	0.0060
5 Obovoid	80	0.23	0.0059
6 Oxheart	76	0.06	0.0145
7 Long rectangular	60	0.12	0.0054
8 Heart	54	0.09	0.0046
9 Long	36	0.32	0.0110

The class error and variance are defined by Eqs. (2) and (5)

Agricultural Research and Development Center in Wooster, Ohio; and New York State Agricultural Experiment Station in Geneva, New York in summer of 2010. At each site, three seedlings per variety were transplanted in separate field plots. Two fruits were collected from each plant, cut longitudinally and scanned. Strangely shaped fruit and those that were immature were removed from the dataset resulting in a total of 784 longitudinal sections for the analysis.

Obtaining morphometric data from scanned fruit images

Fruit images were saved as jpeg and analysed using the latest version of Tomato Analyzer (Rodriguez et al. 2010). The morphometric data used were a sample of (x,y) coordinates from the contour of longitudinally cut fruit sections. We choose the setting of 200 2-dimensional boundary points. The (x,y) pairs were collected in a counter-clockwise direction starting from the proximal end, which was manually corrected in Tomato Analyzer if needed, with 100 points equally distributed on the left side of the fruit down to the distal end, and 100 points equally distributed from the distal end to the proximal end on the right side of the fruit.

Deriving shape descriptors from morphometric data

From each fruit contour, the elliptic Fourier-normalized coefficients were computed as described previously (Kuhl and Giardina 1982) and implemented in supplemental Algorithm 1. The normalization ensured that the model was invariant to spatial translation and rotation. The model depended on how many harmonics (H) were considered, i.e. the number of terms in the Fourier series. Since the optimal number of harmonics to be used for a particular problem is not known a priori, we investigated several including H = 5, 15, 20, 25, and 30. Further, from these coefficients, $2 \times H-1$ normalized shape descriptors were obtained by using the algorithm shown in the supplemental Algorithm 2. These shape descriptors were scale-invariant and used as input to the AutoClass Bayesian clustering system (Achcar et al. 2009).

Identifying clusters of similar shape descriptors through AutoClass

The AutoClass Bayesian clustering program is available at NASA (http://ti.arc.nasa.gov/tech/rse/synthesis-projects-applications/autoclass/autoclass-c/) and at the Institute Jacques Monod (http://ytat2.ijm.univ-paris-diderot.fr/AutoclassAtIJM.html). AutoClass has been used in various scientific applications (Goebel et al. 1989; Kanefsky et al. 1994) including in biology for the prediction of novel intron–exon boundaries in genome sequences (Cheeseman and Stutz 1996). The AutoClass Bayesian classification system uses a finite mixture model and an Expectation–Maximization algorithm to find the optimal classes within the data. In essence, AutoClass fits a mixture of Gaussians over the data. AutoClass started with random parameters for the Gaussians and, after many iterations, these parameters were updated to minimize the fit error. The algorithm also varied the number of Gaussians in search for the best model that fitted the data.

AutoClass is an unsupervised clustering technique, which means that it finds clusters within data according to certain attributes, in this case the shape descriptors. However, the validity of these clusters needed to be further investigated by subject matter experts in a metaclustering step, as illustrated below.

Evaluating clustering performance

The following notations were used: $C_k, k = \overline{1,K}$ was the class k;

$|C_k|$ denoted the number of examples in class k;

H denoted the number of harmonics;

$s_i^k, i = \overline{1,|C_k|}, k = \overline{1,K}$ was the shape descriptor vector of example i in class k;

$s_i^k(j), j = \overline{1, 2 \cdot H - 1}$ was the jth component of the shape descriptor vector s_i^k.

For each of the k classes outputted by AutoClass, the class prototype was defined as

$$S^k = \frac{\sum_{i=1}^{|C_k|} s_i^k}{|C_k|} \quad k = \overline{1, K} \tag{1}$$

The individual class error was defined in Eq. (2) as the mean of the Euclidean distance between the class prototype S^k and the class shape descriptors s_i^k.

$$\mathrm{err}(C_k) = \frac{\sum_{i=1}^{|C_k|} \sqrt{\sum_{j=1}^{2H-1} (s_i^k(j) - S^k(j))}}{|C_k|} \tag{2}$$

with $k = \overline{1, K}$, and $j = \overline{1, 2 \cdot H - 1}$.

The overall clustering error was obtained by

$$\mathrm{err} = \sum_{k=1}^{K} \mathrm{err}(C_k) \tag{3}$$

Shape visualization, meta-classification, and class-variance evaluation

To visualize the shape corresponding to a particular cluster, the class average of the non-normalized Fourier coefficients were computed. The non-normalized coefficients were used to obtain a shape model in the original space, i.e. the raw contour space. Then 200 points were reconstructed from the averaged elliptic Fourier vectors as shown in the supplemental Algorithm 3 and were plotted in the original space.

In the applications described above, a metaclustering step was used after AutoClass to further refine the unsupervised findings. This is explained as follows: the unsupervised learning is meant to find natural patterns in data, which human experts could not observe otherwise, mainly due to size and high dimensionality of the data sets. An unsupervised learning model (here a mixture of Gaussians through the AutoClass system) will find the hidden similarities in the data. However, not all resulting classes were meaningful to the researchers. In the unsupervised learning and data mining areas of research, it is well known that new knowledge is generated only from the interaction between computational models and the knowledge of the experts in the corresponding field. Therefore in the metaclassification step, we selected the Fourier model having a minimum number of

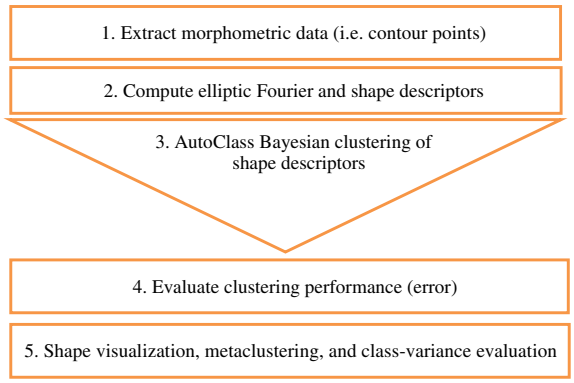

Fig. 1 Computational steps used to derive shape categories from the morphometric data of the fruit boundary. For step 1, Tomato Analyzer was used (Rodriguez et al. 2010). MATLAB 7.9 was used to implement the algorithms from steps 2, 4, and 5. For step 3 we used the AutoClass Bayesian clustering system (Achcar et al. 2009)

harmonics that fit our data subjective to the following two constraints:

(I) The shape categories must be distinguishable by human eye because plant biologists and agronomists must be able to distinguish fruit shapes without relying on scanned and computer-analyzed images. Therefore, too subtle shapes which were distinguishable only through a computer algorithm were merged together.

(II) The final set of shape categories must contain long, oxheart, flat, obovoid, round, and ellipsoid categories.

These basic shapes are obvious to the tomato experts and have been documented previously (IPGRI 1996; Rodriguez et al. 2011; UPOV 2001).

Once the optimal number of harmonics was determined, the AutoClass classification with the lowest error as defined in Eq. (3) was selected. Further, manual evaluation of the resulting shape categories (visualized as explained above) and merged those that were indistinguishable by the human eye. This last step is necessary because the final number of shape categories will be used by plant biologists and agronomists, and a distinguishable set of shapes is desired for this purpose.

Finally, the variance within each final shape class was evaluated. We denote by $\mathrm{var}^k(j)$ the variance per column j within the class k computed as

$$\mathrm{var}^k(j) = \frac{1}{|C_k| - 1} \sum_{i=1}^{|C_k|} \left(s_i^j(j) - S^k(j)\right)^2 \tag{4}$$

with $k = \overline{1, K}$, and $j = \overline{1, 2 \cdot H - 1}$.

Then the overall class variance was obtained from Eq. (5). In other words, the variance within each class k is computed as the sum of the variances per each dimension.

$$\mathrm{var}^k = \sum_{j=1}^{2H-1} \mathrm{var}^k(j) \quad \text{with} \quad k = \overline{1, K} \tag{5}$$

Figure 1 shows the five steps taken to derive the final classification.

Analysis of environmental influence on number of fruits per shape class

To evaluate the influence of the environment on fruit development in the different shape classes, we conducted a balanced one-way analysis of variance (ANOVA) to compare the means of independent samples containing independent observations. This analysis returned the p value under the null hypothesis that all samples are drawn from populations with the same mean.

Results

Classification of tomato fruit based on contour morphometric data

To initiate the identification of the shape classes, we first sought to identify the optimum number of harmonics for the clustering of the morphometric contour data. In general, the first harmonic terms in the Fourier series describe global shape, and the subsequent terms in the series model the finer curvatures of a contour. When using H = 5, 20 classes were distinguished of which many were variations of the round and ellipsoid shape classes (Fig. 2; supplemental Fig. 1). We attributed this redundancy to the manner in which AutoClass operates in a 9-dimensional space: the use of 5 harmonics leads to 9 (2 × 5−1) shape descriptors which were generated for each of the 784 tomato fruits. The 784 9-dimensional data points were more uniformly distributed than in a larger space. For H = 30, AutoClass used 59 (2 × 30−1) shape descriptors for each of the 784 examples resulting in only eight shape classes because fewer Gaussians were needed to cover the data

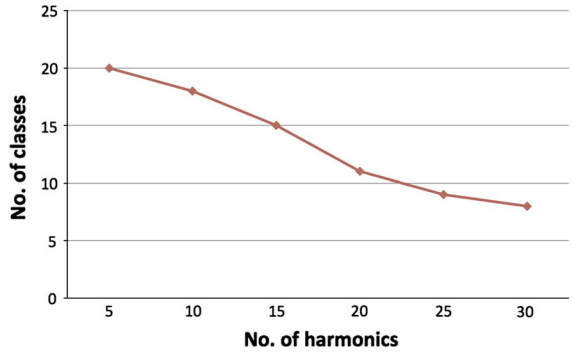

Fig. 2 Establishing the optimum number of harmonics. The number of harmonics is plotted as a function of the number of clusters obtained. Using the morphometric data of the tomato fruit, H = 5 resulted in 20 clusters whereas H = 30 resulted in eight clusters

(supplemental Fig. 2). However, for H = 30, the oxheart and long classes were merged (see shape class 8 in supplemental Fig. 2) which violated our constraint (II) since both are valid shape classes that should remain distinct. Thus by using too many harmonics, distinct shapes that were detected by the human eye were not discerned anymore.

The minimum number of harmonics that best modeled the contour morphometric data while still subjected to constraints (I) and (II) was found to be H = 20, resulting in ten shape classes (Fig. 3a). For all other cases (H = 5, 10, 15, 25, 30) one or both of the two constraints were violated (supplemental Figs. 1, 2). AutoClass is a stochastic unsupervised learning method, and therefore, different runs with the same data may result in slightly different classifications each time. For instance, for different runs of the H = 20 model three different sets of results consisting of 10, 11, and 12 shape classes are produced. Their respective corresponding errors obtained using Eq. (3) are 1.31, 1.61, and 1.54. Since the lowest error was achieved for 10 shape classes, this model was selected for the metaclustering step.

At the metaclustering step (a necessary step after unsupervised clustering), the human expert evaluates the 10 shape classes and incorporates additional knowledge. As briefly discussed in the Materials and methods section, this step is necessary because the unsupervised learning techniques cluster data according to mathematical models. However, it might be that not all the resulting clusters are useful and practical. By inspecting the shape classes in Fig. 3a the human

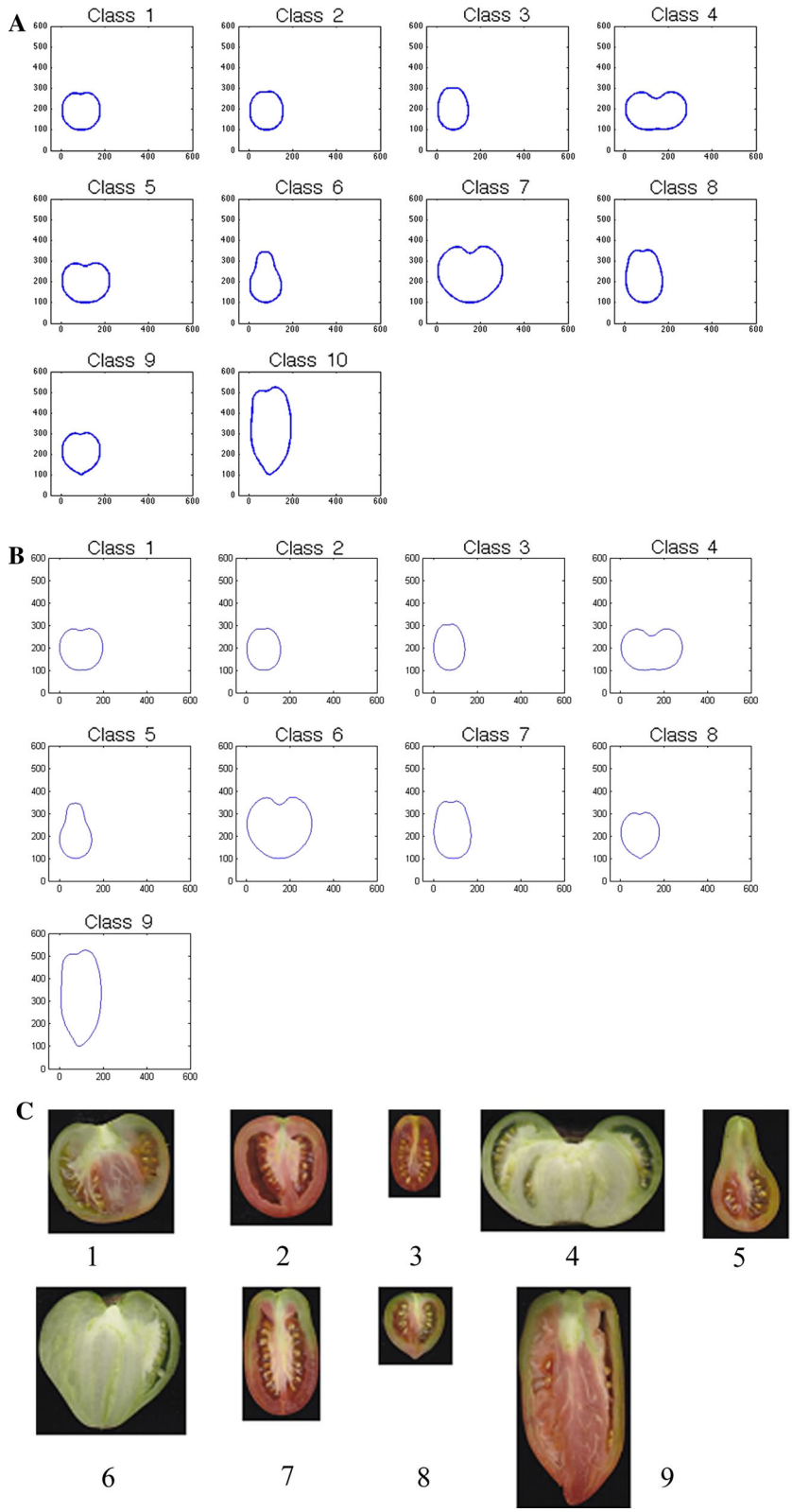

Fig. 3 **a** Ten classes obtained for H = 20. Since class 1 and 5 look similar to the human eye (they are both round shaped) they were merged giving the final nine classes. **b** The final nine classes obtained for H = 20 after the metaclustering step. These nine classes are distinguishable by the human eye and correspond to 1: round; 2: rectangular; 3: ellipsoid; 4: flat; 5: obovoid; 6: oxheart; 7: long rectangular; 8: heart; 9: long. **c** Representative fruit images of the H = 20 and 9 resulting classes. Labeled as 1–9: 1 = round; 2 = rectangular; 3 = ellipsoid; 4 = flat; 5 = obovoid; 6 = oxheart; 7 = long rectangular; 8 = heart; 9 = long

expert easily identified class 4 as flat, class 6 as obovoid, class 7 as oxheart, class 9 as heart, class 10 as long, class 8 as long rectangular, class 3 as ellipsoid, and class 2 as rectangular. However, it was difficult to visually discern class 1 from class 5 (Fig. 3a). This implied that these two shape classes could be merged, as all other classes were visually distinct. This metaclustering step was also consistent with constraint I which emphasizes that the final shapes must be distinguishable by human eye. More importantly, the cumulated error for the resulting nine shape categories decreased further from 1.31 to 1.27. These final 9 categories largely corresponded to those developed previously (Rodriguez et al. 2011), except that the contour morphometric analysis identified a new category, namely "long rectangular" which was derived from the original rectangular class (Fig. 3b; Table 1). This new category was also proposed recently using attribute analysis (Cao 2012). The "long rectangular" category represented 7.5 % of the analysed fruit (60 of 784), suggesting that this shape occurs quite frequently in the tomato germplasm. The tomato fruits in Fig. 3b are assigned by our method to the following categories: class 1 as round; class 2 as rectangular; class 3 as ellipsoid; class 4 as flat; class 5 as obovoid; class 6 as oxheart; class 7 as long rectangular; class 8 as heart; class 9 as long. The uneven distribution of fruit numbers in each shape class and the high number of fruit in the "round" class (Table 1) was due in part to the fact that several tomato varieties were initially classified as rectangular while they carried fruit of a round shape (supplemental Table 1).

Fruit shape and number variability in the contour morphometrics-defined classes

We determined whether certain shape categories were more variable than others by using Eq. (5) (Table 1). The categories with the largest variation were the oxheart and long, and to a lesser extent the flat and obovoid classes. The number of fruit in the long category was relatively low, due to the fact that only three cultivars from this class were grown which may explain the high variation observed (supplemental Table 1). Varying degrees of obovoid fruit were also noted such that some fruits were very strongly pear shaped and others were weakly pear shaped which would explain the large error associated with this category. The rectangular category exhibited the lowest amount of variability, further implying that it was critical to separate the rectangular from the long rectangular class.

To evaluate whether certain shape classes produced similar numbers of fruit under different environmental conditions, we ran a one-way ANOVA for the nine shape classes relative to the three growing locations in Ohio, North Carolina, and New York. Because some classes were much larger than others, we normalized the data relative to each class prior to the ANOVA. The analysis showed a significant difference ($p = 0.0205$) in fruit number per location indicating that more Ohio-grown fruit were analysed (supplemental Fig. 3). Further insights into whether the difference in fruit number was associated with certain shape categories indicated that the number of fruit in the flat and oxheart category was lower in North Carolina whereas the number of fruit in the long, long rectangular and obovoid category was lower in New York compared to the other locations. This could be due to environmental factors such that not enough fruits were ripe (cool summer), were malformed (those wedged between two stems or laying on the ground), or were damaged (animal or bird damage) to be included in this study. This notion was consistent with the shape variability analysis that showed higher error and variance in the flat, long, obovoid and oxheart categories (Table 1). The most robust shape classes that featured similar numbers of fruit across the three locations were ellipsoid, heart, and round categories (supplemental Fig. 3).

Comparisons between manual cultivar and single fruit morphometric classification

The classification of tomato fruit shapes was similar whether conducted manually or computationally (Fig. 3b) (Rodriguez et al. 2011). Based on our data and our constraints the types of shapes formed by tomato fruit are largely delimited to the 9 defined categories. We next sought to determine how well the varieties used in this study conform to one or another shape class. Does the original variety classification fit with the classification of the fruit based on contour morphometric analysis? Do all fruit of a given variety classify in the same shape category? The varieties that conformed the least to their assigned category were those that were initially classified as rectangular and ellipsoid (Table 2, column 1). Fruit from only 21 and

Table 2 Comparisons between visual cultivar classification (first column) and morphometric classification of the fruit

Original variety classification[a]	Overall fit of the original variety classification with morphometric modeling classification	Variety Classification based on morphometric modeling[a]	Overall fit of the reclassified varieties	Best fitting varieties of the given shape class	Varieties carrying the most variable shaped fruit in a given shape class[b]
Ellipsoid	34 % Ellipsoid	Ellipsoid	59 % Ellipsoid	C4, C8	C3, C7, C31
8 varieties (C1-8)	46 % Rectangular	5 varieties (C3,4,7,8, 31)	20 % Rectangular		
137 fruits	12 % Round	87 fruits	15 % Long rectangular		
	8 % Long rectangular		6 % Obovoid		
Flat	88 % Flat	Flat	Unchanged	C9, C10, C11, C13	
6 varieties (C9-14)	7 % Round	6 varieties (C9-14)			
91 fruits		91 fruits			
Heart	61 % Heart	Heart	78 % Heart	C17, C18	C15
4 varieties (C15-19)	15 % Round	3 varieties (C15-18)	6 % Round		
66 fruits	12 % Ellipsoid	49 fruits	12 % Ellipsoid		
Long	79 % Long	Long	Unchanged	C22, C23	C21
3 varieties (C21-23)	12 % Ellipsoid	3 varieties (C21-23)			
43 fruits		43 fruits			
Long Rectangular	57 % Long rectangular	Long Rectangular	57 % Long rectangular	C24 and C25	C24, C25, C26, C41
3 varieties (C24-26)	31 % Ellipsoid	4 varieties (C24-26, 41)	30 % Ellipsoid		
51 fruits		67 fruits	6 % Oxheart		
Obovoid	78 % Obovoid	Obovoid	89 % Obovoid	C27, C29, C30	C32
6 varieties (C27-32)	7 % Ellipsoid	5 varieties (C27-30, 32)	7 % Oxheart		
99 fruits	7 % Oxheart	81 fruits			
Oxheart	64 % Oxheart	Oxheart	Unchanged	C33, C34, C36	C33, C34, C36, C37
5 varieties (C33-37)	13 % Flat	5 varieties (C33-37)			
76 fruits	11 % Heart	76 fruits			
Rectangular	21 % Rectangular	Rectangular	64 % Rectangular	C1, C5	C1, C2, C40
5 varieties (C38-42	45 % Round	5 varieties (C1,2,5,6,40)	21 % Round		
82 fruits	16 % Long Rectangular	85 fruits	6 % Long Rectangular		
			8 % Ellipsoid		
Round	80 % Round	Round	75 % Round	C38, C43, C45, C46, C47	C19, C44, C49, C51
8 varieties (C43-51)	6 % Oxheart	12 varieties (C19, 38,39,42-51)	5 % Oxheart		
139 fruits	6 % Rectangular	205 fruits	10 % Rectangular		

Only values above 5 % are reported

[a] Cluster name, number of varieties in the cluster, total number of fruit

[b] Fruit from one plant that fit in three or more shape categories of which two categories are represented by two or more fruit

34 % of the rectangular and ellipsoid varieties respectively, were categorized in the corresponding classes based on contour morphometric data analysis. The most uniform varieties were those that were initially classified as flat, long, obovoid and round (with 78–88 % fit). This may be due to the fact that the latter classes are easily discerned by eye and by certain Tomato Analyzer attributes: fruit shape index, proximal eccentricity and widest width position which were used to facilitate the original variety classification done manually by experts (Rodriguez et al. 2011).

To determine whether certain tomato varieties should be reclassified or whether they carry fruit that is highly variable in shape, we evaluated those that conformed poorly with respect to their initial classification. Based on the morphometric computations performed in this study, four of the eight varieties originally categorized as ellipsoid should be reclassified as rectangular because the largest group of the fruit (in this case 50 % or more) from those varieties conform to the rectangular class (supplemental Table 1). Conversely, four of the five varieties initially classified as rectangular should be reclassified: three to the round category and one to the long rectangular category. For other shape categories from the original study, no reclassification is required except for the heart variety LYC2406 to be reclassified as round, and the obovoid variety LYC1918 to be reclassified as ellipsoid (supplemental Table 1).

After reclassification of the varieties whose fruit did not conform to the original classification, the fit increased clearly except for the round class (Table 2 columns 2 and 4). This may be due to the inclusion, after reclassification, of the M82 variety that carries more variably shaped fruit than most other varieties that are classified as round. The largest improvement in uniformity of classification was found for the reclassified rectangular and ellipsoid classes, which improved from 21 to 64 % and from 34 to 59 % respectively (Table 2).

For each computationally defined shape category, certain varieties carry similarly shaped fruit that conform well to a single shape category. The best performing category in this regard is the obovoid, where LA0330 (C27), LYC449 (C29), and Yellow Pear (C30) produced 100 % obovoid fruits (Table 2, column 5). Except for LYC438 (C9), none of the other varieties produced fruit that were 100 % in the same shape category. However, many varieties performed quite well and produced fruit that were largely classified in one shape or another category. For example, in the flat category, five out of six varieties produced only flat fruit with the exception of one (T764, T864, Goliath) or two (T954) fruit from that variety (supplemental Table 1).

Even though all varieties were classified into a single shape category, the fruits collected from some varieties fell into three or more shape categories. Collectively, the worst performing classes were the long rectangular and the oxheart categories where nearly all varieties produced fruits that could be classified in three or more categories (Table 2, column 4; supplemental Table 1). The best performing classes were the flat and obovoid categories where only one or no variety produced fruits in more than two shape categories.

Discussion

In this study, we used contour morphometric data obtained from scanned longitudinally cut fruit. The analysis showed that the variation is largely confined to nine shape categories which are near-identical to the previously defined categories (IPGRI 1996; Rodriguez et al. 2011; UPOV 2001). The morphometric data obtained from the Tomato Analyzer application combined with the Bayesian analysis using AutoClass provided a classification scheme that allowed for more uniform separation of fruit and varieties into defined shape categories than visual inspection by experts. Therefore, this approach provides a more useful tool to evaluate variety uniformity and classification for tomato and other fruits and vegetables.

Classification of shape and variability within a shape category were not always correlated. For example, variability was high in flat and obovoid, yet varieties that feature fruit with those shapes often conformed very highly to the class (Tables 1, 2). This suggests that those fruit are easily classified into those categories but that shape is not entirely uniform. On the other hand, oxheart shapes showed high variability and also most varieties classified as oxheart featured more than two fruit shapes. In this case, the oxheart fruit is variable and the corresponding varieties are difficult to classify.

Certain varieties carried fruit that fit in only two shape classes and for which the lowest class contained four or more fruit. The rectangular variety T1355

carried fruit that was rectangular or round; the round variety UPV24629 carried fruit that was round or rectangular; the ellipsoid variety UPV24514 carried fruit that was ellipsoid or rectangular; the flat variety T1121 carried fruit that was flat or round; the oxheart variety Orange Strawberry carried fruit that was oxheart or flat. Thus some shapes like round and rectangular appear to some extent to be interchangeable. This finding suggests that certain accessions produce a shape that is intermediate of the nine defined shape classes perhaps due to the effect of modifier genes. Although most varieties produce fruit that fall in one or a few categories, LYC2406 is most unusual since it produced fruit that conforms to six out of nine categories (supplemental Table 1).

The data support the previous conclusion that *OVATE* controls obovoid, long rectangular and ellipsoid shapes (supplemental Table 1) (Rodriguez et al. 2011). Also, the effect of *LC* and *FAS* on flat and oxheart shape is clearly demonstrated as is the effect of *SUN* on long and oxheart. The *lc* mutation is found in round as well as flat, long and oxheart tomatoes, suggesting that the increase in locule number does not always lead to a change in overall shape. We did not perform statistical analysis regarding the effects of the four fruit shape genes on uniformity because of relatively low numbers. However, the data suggest that neither fruit shape gene leads to increased variability per cultivar or better uniformity (supplemental Table 1). Instead the data suggest there are a number of loci that modify the effect of the fruit shape genes. For example, Spitz produced ellipsoid, obovoid and oxheart fruit in addition to long. Howard German, carrying the same fruit shape alleles as Spitz, produced 17 long and only one obovoid fruit. In addition to genetic modifiers, this suggests that Howard German may carry a "uniformity" gene that is not present in Spitz. In all, our research demonstrates the usefulness of contour morphometric analysis and modelling to classify tomato and other fruits, and to identify varieties with high and low variability within and among plants of the same cultivar.

Acknowledgments This work is supported by the National Science Foundation grant IOS 0922661. We thank Dr. Gustavo Rodriguez, Spencer Debenport and Jenny Moyseenko for help with dispensing the seeds, and collecting and scanning of the fruit. We also thank Dr. Dilip Panthee at Mills River, NC and Drs. Joanne Labate and Larry Robertson in Geneva NY; and John Elliot in Wooster OH for field preparation and plant care.

References

Achcar F, Camadro J-M, Mestivier D (2009) AutoClass@IJM: a powerful tool for Bayesian classification of heterogeneous data in biology. Nucl Acid Res 37:W63–W67. doi:10.1093/nar/gkp430

Bookstein FL (1982) Foundations of morphometrics. Annu Rev Ecol Syst 13:451–470. doi:10.1146/annurev.es.13.110182.002315

Borba EL, Funch RR, Ribeiro PL, Smidt EC, Silva-Pereira V (2007) Demography, and genetic and morphological variability of the endangered *Sophronitis sincorana* (*Orchidaceae*) in the Chapada Diamantina, Brazil. Plant Syst Evol 267:129–146

Brewer MT, Lang L, Fujimura K, Dujmovic N, Gray S, van der Knaap E (2006) Development of a controlled vocabulary and software application to analyze fruit shape variation in tomato and other plant species. Plant Physiol 141:15–25

Cao C (2012) Characterization of management and environment effects on cultivated tomatoes. Masters thesis. The Ohio State University. Retrieved from. http://rave.ohiolink.edu/etdc/view?acc_num=osu1352998717

Chandler GT, Crisp MD (1998) Morphometric and phylogenetic analysis of the *Daviesia ulicifolia* complex (*Fabaceae*, Mirbeliae). Plant Syst Evol 209:93–112

Cheeseman P, Stutz J. (1996). Bayesian classification (AutoClass): theory and results, from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.44.7048

Cheverud JM (1996) Developmental integration and the evolution of pleiotropy. Am Zool 36:44–50

Cong B, Barrero LS, Tanksley SD (2008) Regulatory change in YABBY-like transcription factor led to evolution of extreme fruit size during tomato domestication. Nat Genet 40:800–804

Dryden IL, Mardia KV (1998) Statistical analysis of shape. Wiley, Chichester

Goebel J, Stutz J, Volk K, Walker H, Gerbault F, Self M, Taylor W, Cheeseman P (1989) A Bayesian classification of the IRAS LRS Atlas. Astron Astrophys 222:5–8

Gonzalo MJ, Brewer MT, Anderson C, Sullivan D, Gray S, Van der Knaap E (2009) Tomato fruit shape analysis using morphometric and morphology attributes implemented in Tomato Analyzer software program. J Am Soc Hortic Sci 134:77–87

Henderson A (2006) Traditional morphometrics in plant systematics and its role in palm systematics. Bot J Linnean Soc 151:103–111

IPGRI (1996) Descriptors for Tomato (Lycopersicon spp.). International Plant Genetic Resources Institute, Rome

Kanefsky B, Stutz J, Cheeseman P, Taylor W (1994) An improved automatic classification of a Landsat/TM Image from Kansas (FIFE). Technical report FIA-94-01, NASA Ames Research Center, Artificial Intelligence Branch

Klingenberg C, Monteiro L (2005) Distances and directions in multidimensional shape spaces: implications for morphometric applications. Syst Biol 54:678–688

Klingenberg CP, Leamy LJ, Routman EJ, Cheverud JM (2001) Genetic architecture of mandible shape in mice: effects of quantitative trait loci analyzed by geometric morphometrics. Genetics 157:785–802

Kuhl FP, Giardina CR (1982) Elliptic Fourier features of a closed contour. Computer Graphics and Image Processing 18:236–258. doi:http://dx.doi.org/10.1016/0146-664X(82)90034-X

Langlade NB, Feng X, Dransfield T, Copsey L, Hanna AI, Thebaud C, Bangham A, Hudson A, Coen E (2005) Evolution through genetically controlled allometry space. Proc Natl Acad Sci USA 102:10221–10226

Lihova J, Marhold K, Tribsch A, Stuessy TF (2004) Morphometric and AFLP re-evaluation of tetraploid *Cardamine amara* (*Brassicaceae*) in the Mediterranean. Syst Bot 29:134–146

Liu J, Van Eck J, Cong B, Tanksley SD (2002) A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. Proc Natl Acad Sci USA 99:13302–13306

Munos S, Ranc N, Botton E, Berard A, Rolland S, Duffe P, Carretero Y, Le Paslier MC, Delalande C, Bouzayen M, Brunel D, Causse M (2011) Increase in tomato locule number is controlled by two single-nucleotide polymorphisms located near WUSCHEL. Plant Physiol 156:2244–2254. doi:10.1104/pp.111.173997

Paran I, van der Knaap E (2007) Genetic and molecular regulation of fruit and plant domestication traits in tomato and pepper. J Exp Bot 58:3841–3852

Perez-Perez JM, Serrano-Cartagena J, Micol JL (2002) Genetic analysis of natural variations in the architecture of *Arabidopsis thaliana* vegetative leaves. Genetics 162:893–915

Pickersgill B (2007) Domestication of plants in the Americas: insights from Mendelian and molecular genetics. Ann Bot 100:925–940

Rodriguez GR, Moyseenko JB, Robbins MD, Morejon NH, Francis DM, van der Knaap E (2010) Tomato Analyzer: a useful software application to collect accurate and detailed morphological and colorimetric data from two-dimensional objects. J Vis Exp 16:37

Rodriguez GR, Munos S, Anderson C, Sim SC, Michel A, Causse M, McSpadden Gardener BB, Francis D, van der Knaap E (2011) Distribution of SUN, OVATE, LC, and FAS in the tomato germplasm and the relationship to fruit shape diversity. Plant Physiol 156:275–285. doi:10.1104/pp.110.167577

Rodriguez GR, Kim HJ, van der Knaap E (2013) Mapping of two suppressors of OVATE (sov) loci in tomato. Heredity 111:256–264. doi:10.1038/hdy.2013.45

Rohlf FJ (1990) Morphometrics. Annu Rev Ecol Syst 21:299–316. doi:10.1146/annurev.es.21.110190.001503

Sonibare MA, Jayeola AA, Egunyomi A (2004) A morphometric analysis of the genus *Ficus* Linn. (moraceae). Afr J Biotech 3:229–235

UPOV (2001) Guidelines for the conduct of tests for distinctness, uniformity and stability (Tomato).Geneva

Weber K, Eisman R, Morey L, Patty A, Sparks J, Tausek M, Zeng ZB (1999) An analysis of polygenes affecting wing shape on chromosome 3 in *Drosophila melanogaster*. Genetics 153:773–786

Weight C, Parnham D, Waites R (2008) LeafAnalyser: a computational method for rapid and large-scale analyses of leaf shape variation. Plant J 53:578–586

Xiao H, Jiang N, Schaffner EK, Stockinger EJ, Van der Knaap E (2008) A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. Sci 319:1527–1530