Многомерные подходы машинного обучения для фенотипирования формы плодов в клубнике

> # Multi-dimensional machine learning approaches for fruit shape phenotyping in strawberry

Mitchell J. Feldmann, Michael A. Hardigan, Randi A. Famula, Cindy M. Lopez, Amy Tabb, Glenn S. Cole and Steven J. Knapp

Аннотация

Предпосылки: Форма является критическим элементом визуального представления плодов клубники и зависит от генетических и негенетических факторов. Современные подходы к фенотипированию плодов для внешних характеристик клубники часто основаны на том, что видит человеческий глаз, для того, чтобы сделать категориальные оценки. Тем не менее, форма плода по своей сути - многомерная, постоянно переменная черта и адекватно не описывается одной категориальной или количественной характеристикой. Морфометрические подходы позволяют изучать сложные, многомерные формы, но часто абстрактны и их трудно интерпретировать. В этом исследовании мы разработали математический подход для преобразования классификации форм фруктов из цифровых изображений в порядковый масштаб, называемый *Принципиальная прогрессия k кластеров (PPKC)*. Мы используем эти распознаваемые человеком категории форм, чтобы выбрать количественные особенности, извлеченные из многочисленных морфометрических анализов, которые лучше всего подходят для генетического разбора и анализа.
Результаты: Мы преобразовали изображения клубники в узнаваемые человеком категории с помощью самообучающегося машинного обучения, обнаружили 4 основных категории формы и вывели прогрессию с использованием PPKC. Мы извлекли 68 количественных признаков из цифровых изображений клубники с использованием набора морфометрических анализов и многомерного статистического подхода. Эти анализы определи информативные наборы признаков, которые эффективно фиксируют количественные различия между классами фигур. Точность классификации варьировалась от 68% до 99% для вновь созданных фенотипических переменных для описания формы.
Выводы: Наши результаты показали, что формы плодов клубники можно надежно определить количественно, точно классифицировать и эмпирически упорядочить с использованием анализа изображений, машинного обучения и PPKC. Мы создали словарь количественных признаков для изучение и прогнозирования классов формы и выявления генетических факторов, лежащих в основе фенотипической изменчивости формы плода в клубника. Методы и подходы, которые мы применяли в клубнике, должны применяться к другим фруктам, овощам и специальным культурам.

> ## Abstract
>
> **Background:** Shape is a critical element of the visual appeal of strawberry fruit and is influenced by both genetic and non-genetic determinants. Current fruit phenotyping approaches for external characteristics in strawberry often rely on the human eye to make categorical assessments. However, fruit shape is an inherently multi-dimensional, continuously variable trait and not adequately described by a single categorical or quantitative feature. Morphometric approaches enable the study of complex, multi-dimensional forms but are often abstract and difficult to interpret. In this study, we developed a mathematical approach for transforming fruit shape classifications from digital images onto an ordinal scale called the *Principal Progression of k Clusters (PPKC)*. We use these human-recognizable shape categories to select quantitative features extracted from multiple morphometric analyses that are best fit for genetic dissection and analysis.
> **Results:** We transformed images of strawberry fruit into human-recognizable categories using unsupervised machine learning, discovered 4 principal shape categories, and inferred progression using PPKC. We extracted 68 quantitative features from digital images of strawberries using a suite of morphometric analyses and multivariate statistical approaches. These analyses defined informative

> feature sets that effectively captured quantitative differences between shape classes.
> Classification accuracy ranged from 68% to 99% for the newly created phenotypic variables for describing a shape.
> **Conclusions:** Our results demonstrated that strawberry fruit shapes could be robustly quantified, accurately classified, and empirically ordered using image analyses, machine learning, and PPKC.
> We generated a dictionary of quantitative traits for studying and predicting shape classes and identifying genetic factors underlying phenotypic variability for fruit shape in strawberry.
> The methods and approaches that we applied in strawberry should apply to other fruits, vegetables, and specialty crops.

Предпосылки

Во время одомашнивания садовой земляники (Fragaria × ananassa), аллооктоплоид (Having eight sets of chromosomes, four from each parent) (2n = 8x = 56) гибридного происхождения, селекционеры активно выбирали несколько морфологических и качественных фенотипов [1–3]().
F. × ananassa был создан в начале 1700-х годов путем межвидовой гибридизации между экотипами диких видов октоплоидных (Fragaria virginiana и Fragaria chiloensis), множественными последовательными интрогрессиями генетического разнообразия подвидов F. virginiana и F. chiloensis в последующих поколениях и искусственным отбором для важных для садоводства черт среди межвидовых гибридных потомков. Одомашнивание и размножение изменили морфологию, развитие и метаболизм плодов садовой земляники, отделяя современные сорта от их диких предшественников [4–9]().
Приблизительно 300 лет размножения в смешанной гибридной популяции привели к появлению высокоурожайных сортов с крупными, крепкими, визуально привлекательными, с длительным сроком хранения фруктами, которые могут противостоять трудностям сбора, обработки, хранения и доставки на большие расстояния. [10](). Форма плода является существенной чертой сельскохозяйственной продукции, особенно той, которая специализируется на сельскохозяйственных культурах, благодаря воспринимаемой и осознанной взаимосвязи с качеством и ценностью продукции. Фенотипирование плодов, основанное на изображениях, может увеличить объем, пропускную способность и точность количественных генетических исследований за счет снижения влияния предвзятости пользователей, анализа больших выборок и более точного разделения генетической дисперсии от среды, управления и других негенетических источников вариации [11–13]().

> ## Background
>
> Fruit breeders actively selected several morphological and quality phenotypes during the domestication of the garden strawberry (Fragaria × ananassa), an allo-octoploid (2n = 8x = 56) of hybrid origin [1–3](). F. × ananassa was created in the early 1700s by interspecific hybridization between ecotypes of wild octoploid species (Fragaria virginiana and Fragaria chiloensis), multiple subsequent introgressions of genetic diversity from F. virginiana and F. chiloensis subspecies in subsequent generations, and artificial selection for horticulturally important traits among interspecific hybrid descendants. Domestication and breeding have altered the fruit morphology, development, and metabolome of the garden strawberry, distancing modern cultivars from their wild progenitors [4–9](). Approximately 300 years of breeding in the admixed hybrid population has led to the emergence of high-yielding cultivars with large, firm, visually appealing, long shelf-life fruit that can withstand the rigors of harvest, handling, storage, and long-distance shipping [10](). Fruit shape is an essential trait of agricultural products, particularly those of specialty crops, owing to perceived and realized relationships with the quality and value of the products. Image-based fruit phenotyping has the potential to increase scope, throughput, and accuracy in quantitative genetic studies by reducing the effects of user bias, enabling the analysis of larger sample sizes, and more accurate partitioning of genetic variance from environments, management, and other non genetic sources of variation [11–13]().

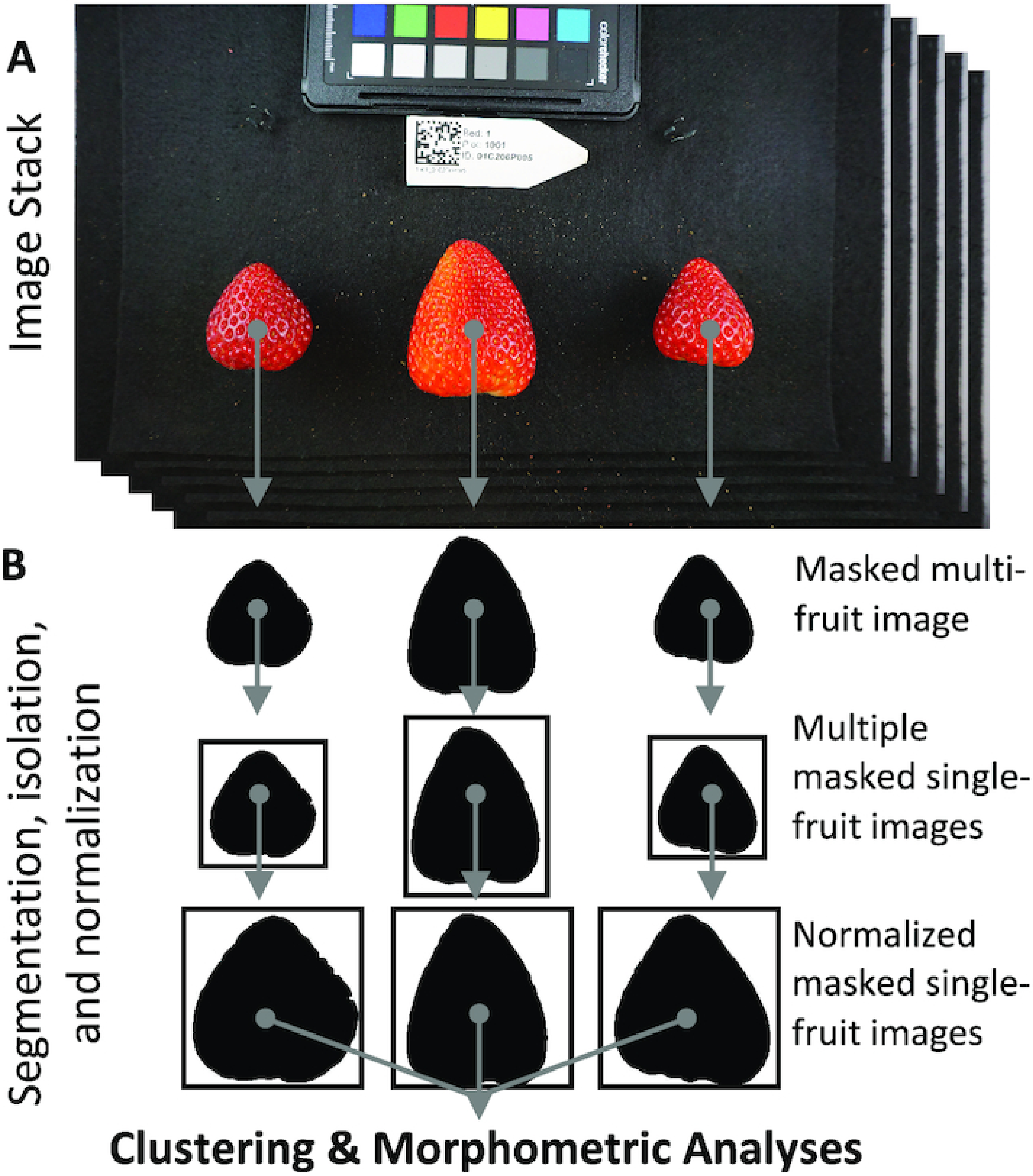
> Many fruit phenotyping approaches rely on the human eye to sort fruit into discrete, descriptive categories for planar (2D) shapes (e.g., rhombic and reniform) [14–19]. Categories are either nominal [11, 20, 21], existing in name only, or ordinal,

> referring to a position in an ordered series or on a gradient [15, 16, 21]. Classification into categories is often labor-intensive and prone to human bias, which can increase with task complexity and time requirements [22, 23]. Alternative scoring approaches rely on morphometrics and machine learning to automate classification; e.g., sorting fruit into shape categories in both tomato [11] and strawberry [20]. Unsupervised machine learning methods (e.g., k-means clustering), unlike supervised methods, are useful for pattern detection and clustering, while supervised machine learning methods (e.g., support vector machines) are useful for prediction and classification [24, 25]. Unsupervised clustering enables the calculation of several measures of model performance and overfitting to balance compression and accuracy. However, the categories derived from these techniques are without order, resulting in the need for a suitable transformation to an ordinal scale more appropriate for quantitative genetic analyses [26–30]. In this context, ordinal categories give the interpretation of relationship with, or distance from, other shape categories in a series. To enable this interpretation, we developed a method for asserting the progression through fruit shape categories derived from unsupervised machine learning methods. The Principal Progression of k Clusters (PPKC) allowed us to non-arbitrarily determine the appropriate shape gradient for statistical analyses using empirical data. The advantages of PPKC, relative to a manually determined ordinal scale, are that it does not require arbitrary, a priori decisions and is unsupervised, which obviates additional operator bias. Here, we describe approaches for translating digital images of strawberries into computationally defined phenotypic variables for identifying and classifying fruit shapes.

> Fruit shape and anatomy are complex, multi-dimensional, and, potentially, abstract phenotypes that are often not completely or intuitively described by planar descriptors and individual qualitative or quantitative variables. Beyond the qualitative definitions used in plant systematics [18, 20], references to fruit shape encompass a wide variety of mathematical parameters and geometric indices that establish quantitative measurements of plant organs [19, 31–33]. Much like human faces or grain yield, fruit shape and anatomy are products of the underlying genetic and non-genetic determinants of phenotypic variability in a population [34, 35]. Quantitative phenotypic measurements have allowed researchers to uncover some of the genetic basis of fruit shape in tomato [36, 37], pepper [38, 39], pear [40], melon [35], potato [41], and strawberry [9, 42]. However, the major genetic determinants of fruit shape remain unclear, or understudied, in octoploid strawberry, in part because researchers have not yet translated fruit shape attributes into holistic, quantitative variables, which may empower the identification of underlying genes or quantitative trait loci through genome-wide association studies (GWAS) and other quantitative genetic approaches [43–46]. Quantitative features often rely on linear metrics of distance (e.g., height, width, and perimeter) and are generally modified into compound descriptors that remove the effects of size (e.g., aspect ratio or roundness) [40, 42, 47]. However, compound linear descriptors often have limited resolution compared to more comprehensive, multivariate descriptors [33]. Elliptical Fourier analysis (EFA) quantifies fruit shape from a closed outline by converting a closed contour into a weighted sum of harmonic functions [12, 48–51]. Generalized Procrustes analysis (GPA) quantifies the distance between sets of biologically homologous, or mathematically similar, landmarks on the surface of an object [48, 51–57]. Fruit shape can also be described using linear combinations of pixel intensities from digital images extrapolating from analyses generally used to quantify color patterns and facial recognition [13, 58–63]. Similar pixel-based descriptors have recently been referred to as ”latent space phenotypes” and arise from unsupervised analyses (i.e., principal component analysis [PCA] and auto-encoding neural networks) that allow a computer to produce novel, independently distributed features directly from images [64, 65]. Here, we generate a dictionary of 68 quantitative features, including linear-, outline-, landmark-, and pixel-based descriptors to investigate the quality of different features in preparation for quantitative genetic analyses.

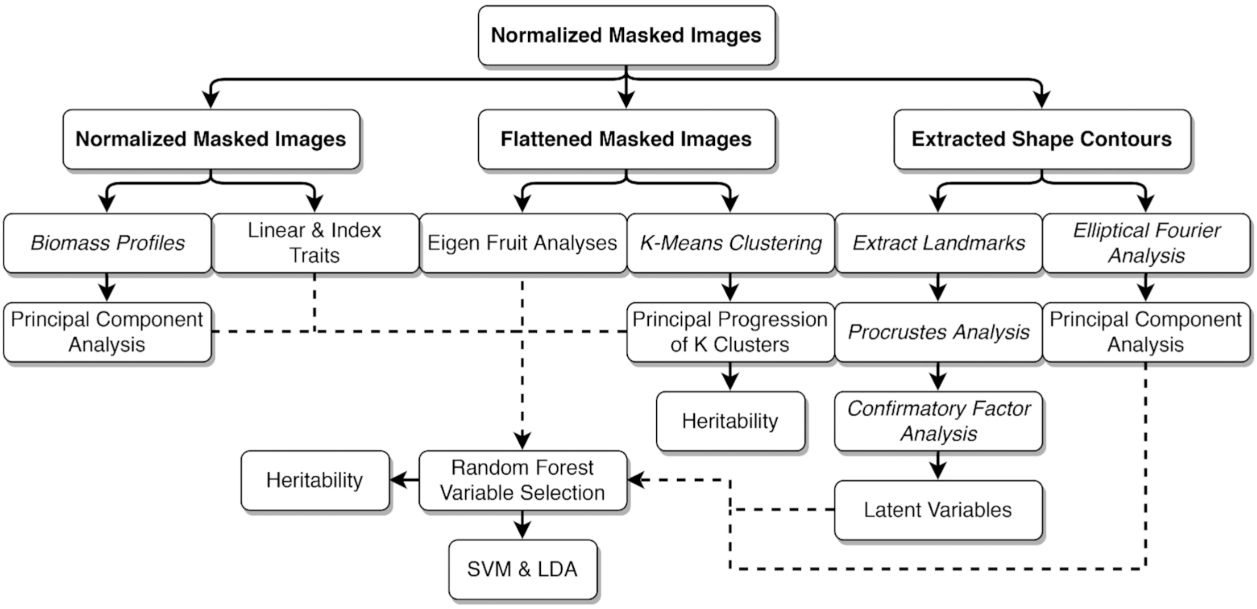
> The ultimate goal of our study was to develop heritable phenotypic variables for describing fruit shape, which could then be used to identify the genetic factors underlying phenotypic differences in fruit shape. The phenotyping and analytic workflow for this study are summarized in Figs 1 and 2. We first describe and demonstrate the application of PPKC, which transforms categories discovered from unsupervised machine learning methods to a more convenient and analytically tractable ordinal scale [26, 28, 29]. We then explore the relationship between machine-acquired categories and 68 quantitative features extracted from digital images. Next, we apply random forest regression to select critical sets of quantitative features for classification and use supervised machine learning methods, including support vector regression (SVR) and linear discriminant analysis (LDA), to determine the accuracy of shape classification. We discovered that there are only a few categories of interest in a highly domesticated breeding population and that a small number of features are needed to classify shape into the discovered categories accurately. We also find that ordinal shape categories are highly heritable and that the features needed for accurate classification are also heritable.

Figure 1:


A
Image Stack
B
Segmentation, isolation, and normalization
Masked multi-fruit image
Multiple masked single-fruit images
Normalized masked single-fruit images
Clustering & Morphometric Analyses

> An example of the processing pipeline. (A) A user collects a stack of images containing multiple strawberries and a unique QR code. (B) All images are then segmented using the SIOX algorithm implemented in ImageJ. Each object is then cut from its original image based on the coordinates of its bounding rectangle in R 3.5.3. White pixels are then added to the edges of each frame until all images are 1,000 × 1,000 pixels. Regions of interest are then scaled such that the major axis of each object becomes 1,000 pixels in ImageJ. Output images are scale invariant and maintain the original aspect ratio.

Figure 2:



Analysis pipeline for this study. All images start as normalized, binary images from Fig 1. Images then follow each of the paths through different morphometric feature extractions including linear geometric features, biomass profile analysis (BPA), EigenFruit analysis, Procrustes analysis, and elliptical Fourier analysis as either normalized or flattened images (e.g., linear, BPA, and EigenFruit analysis) or as shape contours (e.g., GPA and EFA). Flattened binary images are used to perform k-means clustering and subsequently PPKC.

Data Description

The data released with this article contain digital images of 6,874 strawberry fruit from 572 hybrids originating from the University of California, Davis, Strawberry Breeding Program. The data for this article, including pre-processed images (Fig. 1A), processed images (Fig. 1B), and extracted features (see Methods, Fig. 2), are available on Zenodo [66]. The pre-processed images typically contained multiple berries per image along with a data matrix bar code indicating the genotype ID and other elements of the experiment design. The processed images are 1,000 × 1,000 pixels-scaled binary images of individual fruit. The extracted features data set is provided as a CSV file. The code to replicate the analyses in this article is provided in a GitHub repository [67]. Additionally, snapshots of the code and data supporting this work are available in the GigaScience repository, GigaDB [68]. We hope that the release of these data assists others in developing novel morphometric approaches to better understand the genetic, developmental, and environmental control of fruit shape in strawberry, and more broadly in other fruits, vegetables, and specialty crops.

Analyses

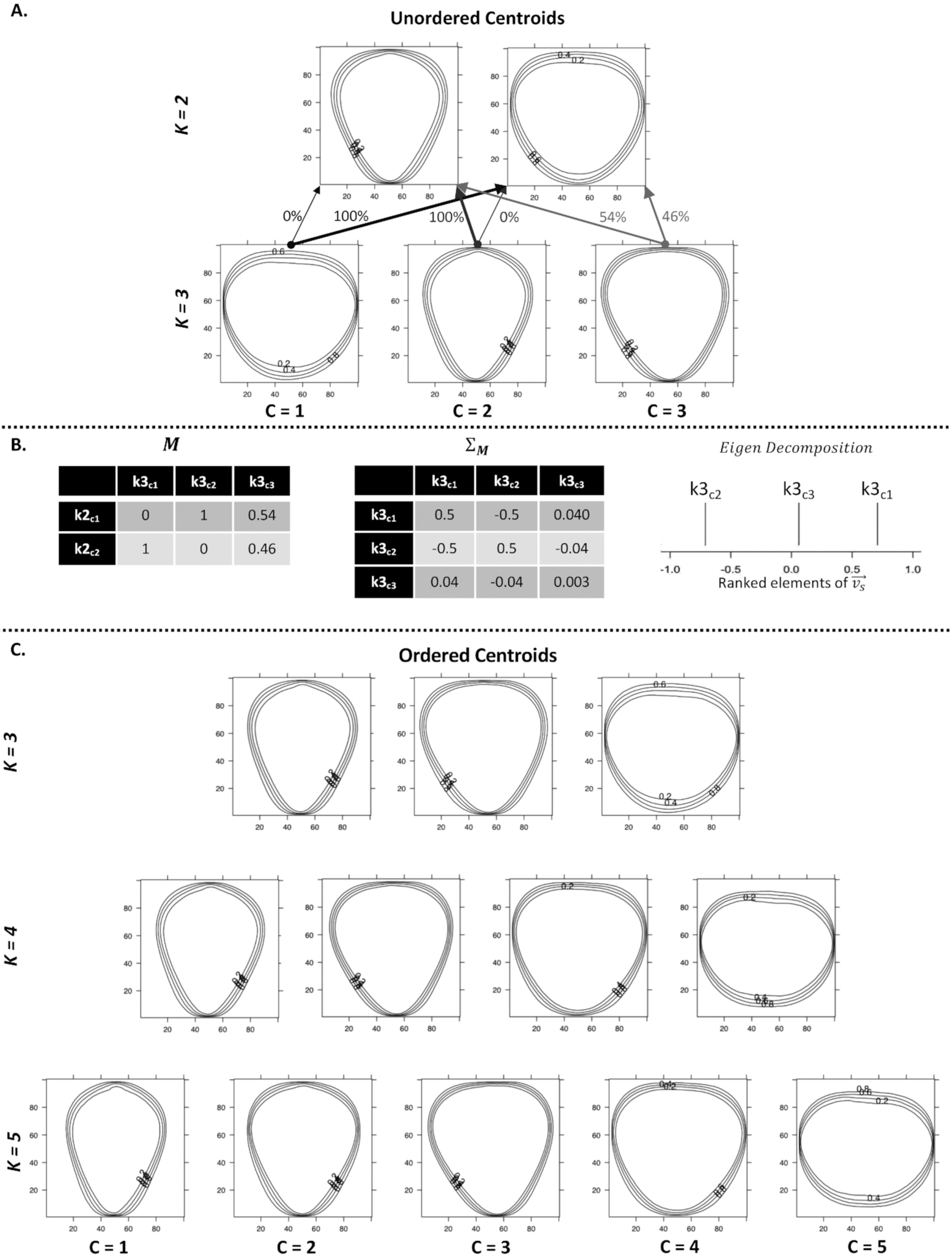
k-Means clustering

k-Means clustering rapidly detects patterns in large, multi-dimensional data sets used for clustering, decision making, and dimension reduction [24, 69, 70]. It is an iterative algorithm that partitions a data set into a pre-defined number of non-overlapping clusters, k, by minimizing the sum of squared distances from each data point to the cluster centroid. A centroid corresponds to the mean of all points assigned to the cluster. Here, we used k-means to cluster flattened binary images (Fig. 1; see Methods). Individual fruits were segmented from the image background as a binary mask, normalized by the major axis, resized to 100 × 100 pixels, and flattened into a vector (Figs 1 and 2; see Methods). We represented each image as a 10,000-element vector containing binary pixel values. We were able to rapidly and reliably assign images to classes using k-means clustering. In this experiment, we allowed k, the number of permitted categories, to range from 2 to 10. This range was chosen because we anticipate that a human-based classification system would not have the speed or reliability needed for this task, particularly for larger values of k.

Principal progression of k clusters

> k-Means clustering does not assign a progression or gradient to discovered classes. However, score and ordinal traits are typically more useful and are more common in quantitative genetic studies than nominal scales [26, 28, 29, 71]. We developed a new method to transform the categories derived from k-means onto an ordinal scale, which we call PPKC (Fig. 3; Algorithm 1). This method relies on k-means clustering to categorize images and can be used to discover an appropriate ordinal scale in nominal data empirically. k-Means supports several metrics for evaluating model performance and overfitting, including adjusted R2, Akaike information criterion (AIC), and Bayesian information criterion (BIC), which allows users to determine the most appropriate value of k given the observed data. The gradient between clusters was estimated by performing PCA on a covariance matrix reflecting the structured relationship between a focal cluster and all previously discovered clusters.

Figure 3:



	$k3_{c1}$	$k3_{c2}$	$k3_{c3}$
$k2_{c1}$	0	1	0.54
$k2_{c2}$	1	0	0.46

	$k3_{c1}$	$k3_{c2}$	$k3_{c3}$
$k3_{c1}$	0.5	-0.5	0.040
$k3_{c2}$	-0.5	0.5	-0.04
$k3_{c3}$	0.04	-0.04	0.003

> An example use of PPKC. (A) After k-means clustering is performed clusters are randomly assigned a numeric value (1,2,…,k). When k > 2, this value becomes nominal. PPKC relies on the fact that the order through clusters when k = 2 has identical interpretations in either direction. The lines representing each clusters centroid reflect the 20th, 40th, 60th, and 80th quantiles, moving out from the center of each image. (B)Left, A table representation of the resultant matrix from Equation (1). Each cell represents the proportion of images in the column class and in the row class, normalized by the number of images in the column class. (B)Middle, A table representation of ΣM. (B)Right, The ranked elements of v⃗s shown on a number line. (C) After using PPKC, the order of groups is explicitly identified. In this example, showing k = [3,

> 5], the order discovered seems to trend from tall and thin berries, through more triangular shapes, ending with berries that are short and wide.

> We first assign each flattened binary image (Fig. 1) to a category using a k-means approach. We assign a cluster to each image and allow the number of clusters, k, to range from 2 through 10. The order is subsequently inferred using PPKC (Fig. 3, Algorithm 1). When k = 2, the order of relatedness is considered arbitrary, and both k2c1 → k2c2 and k2c2 → k2c1 have the same meaning, where ”→” indicates the progression of discovered categories. Any given order and its reverse are considered equivalent, and this applies to higher levels of k as well; e.g., the hypothetical ranking of clusters 1, 4, 2, 3 is considered equivalent to 3, 2, 4, 1 because the relative relationship between the k clusters is identical in both (e.g., c3 is more related to c2 than either c1 or c4). For each cluster of interest (e.g., k4c1, k4c2, k4c3, and k4c4), we calculate the proportion of each cluster that came from k3c1, k3c2, or k3c3 and k2c1 or k2c2 (i.e., all former classifications). These proportions enable the estimation of similarity between a focal cluster (e.g., k4c1) and the clusters of all prior values of k. We then normalize the proportions by the total number of images in the focal cluster (e.g., k4c1, k4c2, k4c3, and k4c4) (Equation 1).

> For every level of k > 2, we construct M, a rectangular matrix of size (k2−k)/2−1 × k(Algorithm 1 line 13). The sum of each column should equal k − 2. The proportions are continuous values in the range [0, 1] that described the origin of a particular focal cluster (e.g., k4c1) as it relates to the clusters of k = 3 and k = 2 or all clusters [2, k − 1]. In the following example, k = 4:

$$\mathbf{M} = \begin{bmatrix} \frac{|k4_{c1} \wedge k3_{c1}|}{|k4_{c1}|} & \frac{|k4_{c2} \wedge k3_{c1}|}{|k4_{c2}|} & \frac{|k4_{c3} \wedge k3_{c1}|}{|k4_{c3}|} & \frac{|k4_{c4} \wedge k3_{c1}|}{|k4_{c4}|} \\ \frac{|k4_{c1} \wedge k3_{c2}|}{|k4_{c1}|} & \frac{|k4_{c2} \wedge k3_{c2}|}{|k4_{c2}|} & \frac{|k4_{c3} \wedge k3_{c2}|}{|k4_{c3}|} & \frac{|k4_{c4} \wedge k3_{c2}|}{|k4_{c4}|} \\ \frac{|k4_{c1} \wedge k3_{c3}|}{|k4_{c1}|} & \frac{|k4_{c2} \wedge k3_{c3}|}{|k4_{c2}|} & \frac{|k4_{c3} \wedge k3_{c3}|}{|k4_{c3}|} & \frac{|k4_{c4} \wedge k3_{c3}|}{|k4_{c4}|} \\ \frac{|k4_{c1} \wedge k2_{c1}|}{|k4_{c1}|} & \frac{|k4_{c2} \wedge k2_{c1}|}{|k4_{c2}|} & \frac{|k4_{c3} \wedge k2_{c1}|}{|k4_{c3}|} & \frac{|k4_{c4} \wedge k2_{c1}|}{|k4_{c4}|} \\ \frac{|k4_{c1} \wedge k2_{c2}|}{|k4_{c1}|} & \frac{|k4_{c2} \wedge k2_{c2}|}{|k4_{c2}|} & \frac{|k4_{c3} \wedge k2_{c2}|}{|k4_{c3}|} & \frac{|k4_{c4} \wedge k2_{c2}|}{|k4_{c4}|} \end{bmatrix}$$

> We then calculate the variance-covariance matrix of Equation (1) (Algorithm 1 line 18). The variance-covariance matrix, ΣM, represents the relationship between each focal cluster (e.g., k4c1, k4c2, k4c3, or k4c4).

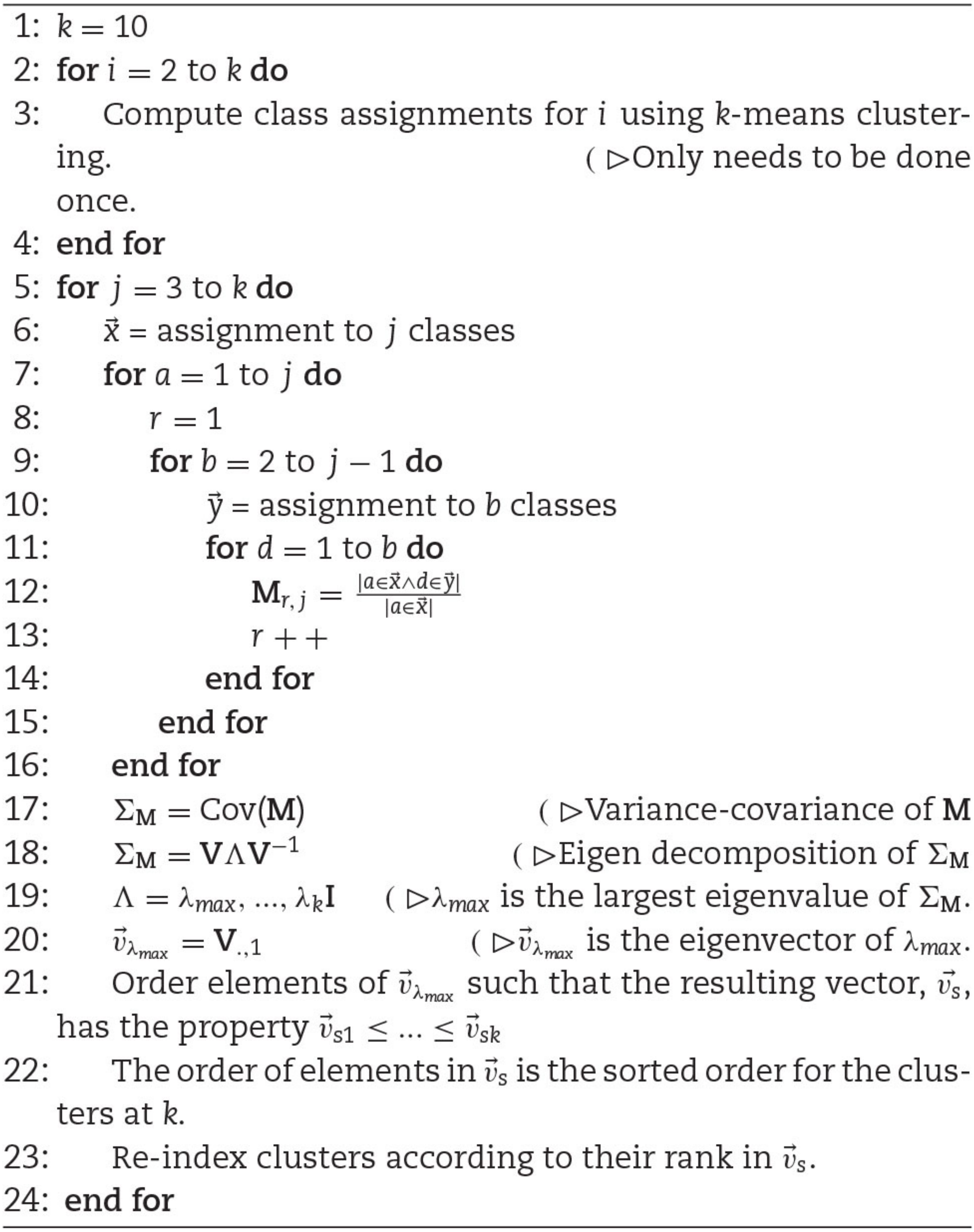
$$\mathbf{\Sigma_M} = \begin{bmatrix} \sigma^2_{k4_{c1}} & \sigma_{k4_{c2},k4_{c1}} & \sigma_{k4_{c3},k4_{c1}} & \sigma_{k4_{c4},k4_{c1}} \\ \sigma_{k4_{c1},k4_{c2}} & \sigma^2_{k4_{c2}} & \sigma_{k4_{c3},k4_{c2}} & \sigma_{k4_{c4},k4_{c2}} \\ \sigma_{k4_{c1},k4_{c3}} & \sigma_{k4_{c2},k4_{c3}} & \sigma^2_{k4_{c3}} & \sigma_{k4_{c4},k4_{c3}} \\ \sigma_{k4_{c1},k4_{c4}} & \sigma_{k4_{c2},k4_{c4}} & \sigma_{k4_{c3},k4_{c4}} & \sigma^2_{k4_{c4}} \end{bmatrix}$$

> We then perform eigen decomposition on Equation (2) using the following equation (Algorithm 1 line 19).

$$\mathbf{\Sigma_M} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}.$$

> In Equation (3), Λ is a diagonal matrix with values corresponding to the k eigenvalues of ΣM and V is a square matrix containing eigenvectors associated with the eigenvalues in Λ. We then extract the eigenvector associated with the largest eigenvalue, $\vec{v}\lambda max$. We order the elements of $\vec{v}\lambda max$ such that the resultant vector, $vs\rightarrow$, has the property $vs1 \leq \ldots \leq vsk$. We do not consider the distance between elements in $vs\rightarrow$, only their rank. The clusters are then indexed to match the rank of the associated elements in $vs\rightarrow$. There are at most k eigenvalues associated with eigenvectors of length k due to ΣM being $k \times k$. Eigen decomposition is used to describe the major axis of variance in ΣM. In theory, this perspective of covariance should be able to separate the classes effectively because it describes a linear axis containing the greatest amount of independent variation and solutions are non-arbitrary. The value a category takes on this composite axis is therefore suggestive of its linear relationship to other the k categories being considered. However, we note that relationships containing branches, bubbles, and other topological features will not be captured accurately. In this study, we are unable to report a visually meaningful order when $k \geq 9$ (Fig. S1) [66]. The change in progression could be reflective of overfitting the number of groups in k-means clustering. The large change of slope at $k = 4$ in the total within-group sums of squares, AIC, and adjusted R2 evidenced overfitting (Fig. S2) [66]. The strongest evidence for 4 clusters is in the BIC, which is minimized when $k = 4$ (Fig. S2D) [66]. The elements of $vs\rightarrow$ tend to converge on one another as k increases, which may be indicative of little biological information in the new clusters and overfitting (Fig. S3) [66]. Given that only relatively small covariance matrices are considered in this algorithm, the computational time to order $k = [3, \ldots, 10]$ on an early 2015 MacBook Pro 2.9 GHz Core i5 with 8GB memory is <0.2 seconds.

Algorithm 1



```
$k = 10$
for $i = 2$ to $k$ do
    Compute class assignments for $i$ using $k$-means clustering.    ( ▷Only needs to be done once.
end for
for $j = 3$ to $k$ do
    $\vec{x}$ = assignment to $j$ classes
    for $a = 1$ to $j$ do
        $r = 1$
        for $b = 2$ to $j - 1$ do
            $\vec{y}$ = assignment to $b$ classes
            for $d = 1$ to $b$ do
                $\mathbf{M}_{r,j} = \frac{|a \in \vec{x} \wedge d \in \vec{y}|}{|a \in \vec{x}|}$
                $r++$
            end for
        end for
    end for
    $\Sigma_{\mathbf{M}} = \mathrm{Cov}(\mathbf{M})$    ( ▷Variance-covariance of $\mathbf{M}$
    $\Sigma_{\mathbf{M}} = \mathbf{V}\Lambda\mathbf{V}^{-1}$    ( ▷Eigen decomposition of $\Sigma_{\mathbf{M}}$
    $\Lambda = \lambda_{max}, ..., \lambda_k \mathbf{I}$    ( ▷$\lambda_{max}$ is the largest eigenvalue of $\Sigma_{\mathbf{M}}$.
    $\vec{v}_{\lambda_{max}} = \mathbf{V}_{.,1}$    ( ▷$\vec{v}_{\lambda_{max}}$ is the eigenvector of $\lambda_{max}$.
    Order elements of $\vec{v}_{\lambda_{max}}$ such that the resulting vector, $\vec{v}_s$, has the property $\vec{v}_{s1} \leq ... \leq \vec{v}_{sk}$
    The order of elements in $\vec{v}_s$ is the sorted order for the clusters at $k$.
    Re-index clusters according to their rank in $\vec{v}_s$.
end for
```

Principal Progression of K Clusters (PPKC) Algorithm

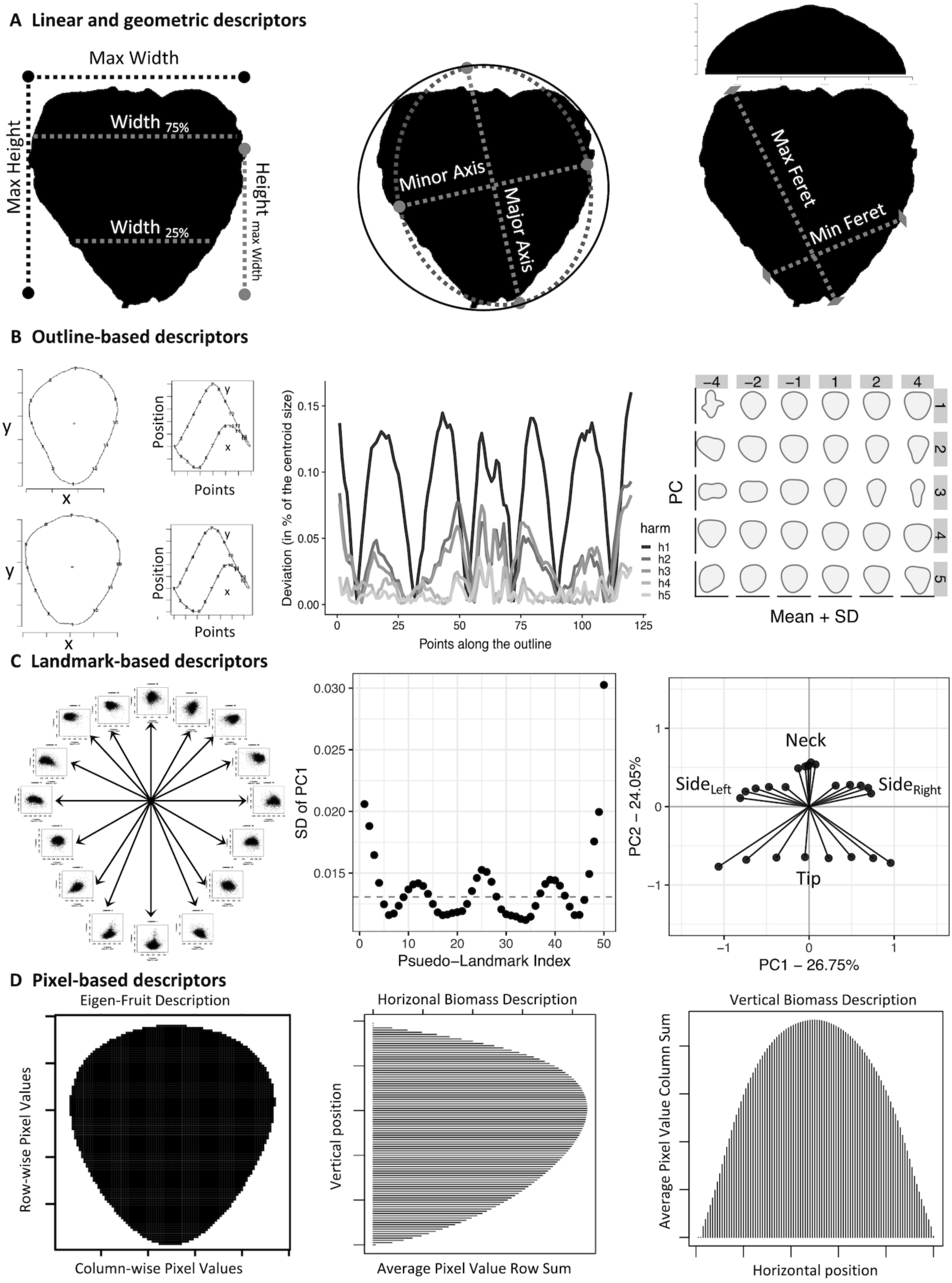
Broad-sense heritability of ordered categories

For each value of k, broad-sense heritability (H2) on an entry-mean basis was assessed using a general linear mixed model with a cumulative logit link function (Equations 4 and 5) [72]. For this data set, H2 was generally high, ranging from H2 = 0.80 to 0.98, even as k → 10 (Table 2). These estimates of H2 are very similar to those reported by Antanaviciute [16] (i.e., H2 = 0.84). When the H2 of a trait is in this range, it indicates that independent replications of the same individuals share a high degree of similarity and that most of the variation among individuals originated from genetic variation among individuals. Because the plant material used in this study came from genetic clones, any variation in fruit

> shape among replicates originated from random, unobserved effects. For k ≥ 9, the accuracy of H2 estimates is expected to be lower than for k ≤ 8 because the gradient of the phenotype seems to be improperly specified. In this set of germplasm, we propose a set of 4 primary classes for categorizing fruit shape (Fig. 3 and S2) [66]. As k increases from 5 to 10, the visual similarity of some clusters is high (Fig. S1) [66], thus indicating fewer relevant delineations (Fig. S3) [66]. As indicated, there is strong evidence in these data that there are 4 distinct clusters in these data (Fig. S2) [66].

> Feature selection using random forests
> To discover which of 68 quantitative features (summarized in Figs 4 and 5) capture and reflect differences in shape categories, supervised machine learning was used to estimate feature importance (see Methods) [73]. Of the 68 features used as predictors in a random forest regression (see Methods), we selected only 13. Out-of-bag (OOB) error is an estimate of how poorly models perform when a specific feature is excluded and is akin to error estimated from jackknife resampling (Fig 6). In this way, features with higher estimates tend to be more relevant for classification and prediction. In this experiment, features could only be selected up 9 times, once per value of k. We maintained features that were selected in ≥3 levels of k to use as independent variables in classification (Table 1). The 13 selected features accounted for >80% of importance assigned to the 68 features across all values of k (Fig 6B). Here, the use of ”EigenFaces,” an analysis from the 1980s, designed to classify human faces, was re-purposed for the quantification and classification of fruit shape in strawberry [58–61]. Pixel-based features dominated the selected features and include principal components (PCs) 1−7 of the EigenFruit analysis (EigenFruitPC[1, 6]), PCs 1 and 2 of the vertical biomass profile (BioVPC[1, 2]), and PCs 1 and 2 of the horizontal biomass profile (BioHPC[1, 3]) (Table 1 and Figs 6 and 7). These features originated from the same data type as used in k-means clustering (i.e., pixel intensities), which is likely the reason they make up the majority of the selected features (Table 1 and Figs 6 and 7). Several geometric descriptors were also selected, including the bounding aspect ratio (BAR), shape index (SI), and ellipse aspect ratio (AR) (Table 1 and Figs 6 and 7). We generated a subset of 5 features with mean OOB ≥ 0.047 (Fig. 6A). OOB = 0.047 was the median OOB error for all features across all classes. This subset of features included EigenFruitPC[1, 2], BioVPC1, and BioHPC[1] (Table 1). We also generated a third smaller set that included only EigenFruitPC1, BioVPC1, and BioHPC1 with mean OOB ≥ 0.12 (Fig. 6A). OOB = 0.12 was the mean OOB error for all features across all classes. The prevalence of pixel-based descriptors in these selected subsets indicated the magnitude of relevant shape information that they described.

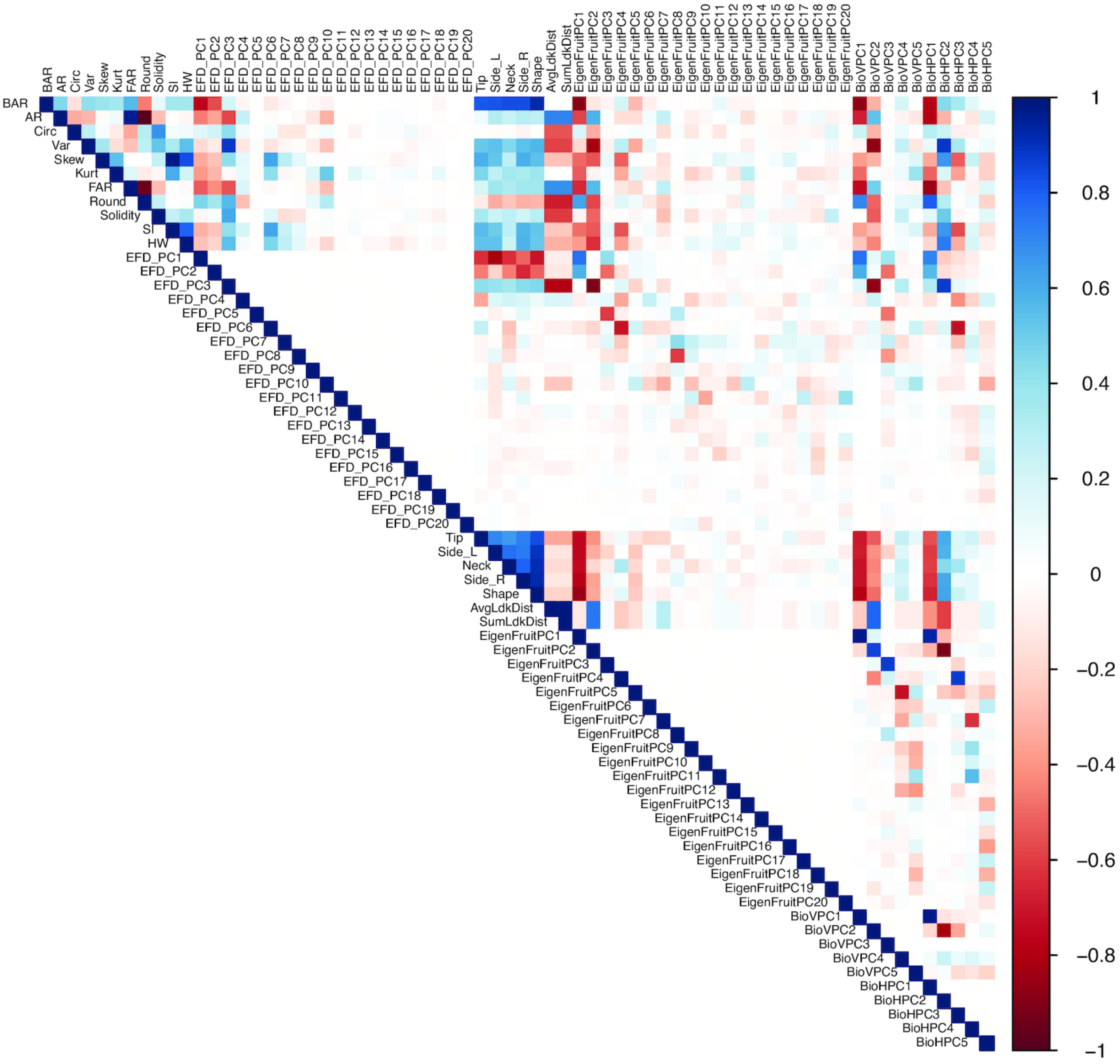
Figure 4:



Trait dictionary for this study. (A) Linear descriptors. Left, Simple linear measurements. Center, Best-fit ellipse axes. For the circle, Round and Circ = 1. Right, Maximum and minimum Feret. Histogram represents the marginal distribution on the horizontal axis used to calculate Var, Skew, and Kurt. (B) Outline descriptors. Left, The 2 leftmost images are the outlines of 2 strawberries with 12 evenly spaced points. The graphs on the right show the original closed outline as 2 oscillating functions. Center, Deviations from the closed outline with increasing harmonics (harm = [h1, h5]).Right, The plot shows the effects of PC 1,5 with effect sizes, −4, 4 on the mean shape. (C) Landmark descriptors. Left, 50 evenly spaced

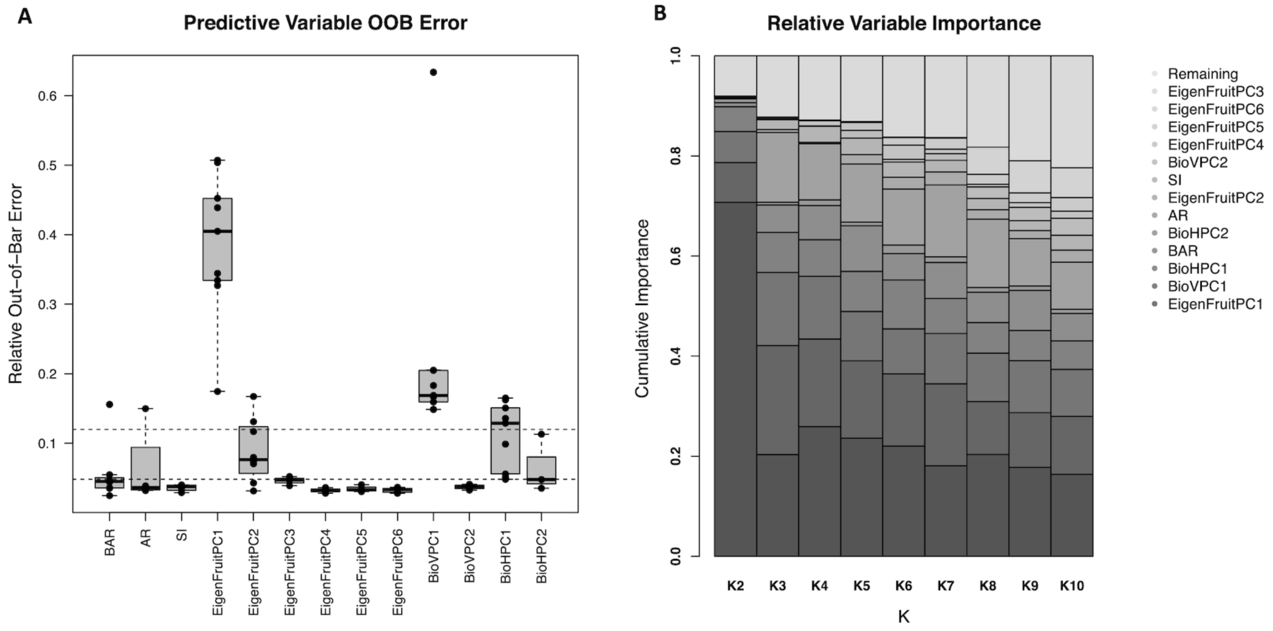
> landmarks are extracted and treated as bi-variate features.Center, Standard deviation of PC1 for each landmark is plotted in sequence. Dashed horizontal line is the median standard deviation (SD). Right, Pseudo-landmarks were selected to represent each region of high variance. Using the values on the first principal axis as observed variables, confirmatory factor analysis was performed to infer latent relationships to tip, left and right side, and neck shape. (D) Pixel descriptors. Left, Mean EigenFruit using flattened binary images. Center, Mean horizontal biomass using image row sums. Right, Mean vertical biomass using image column sums.

Figure 5:



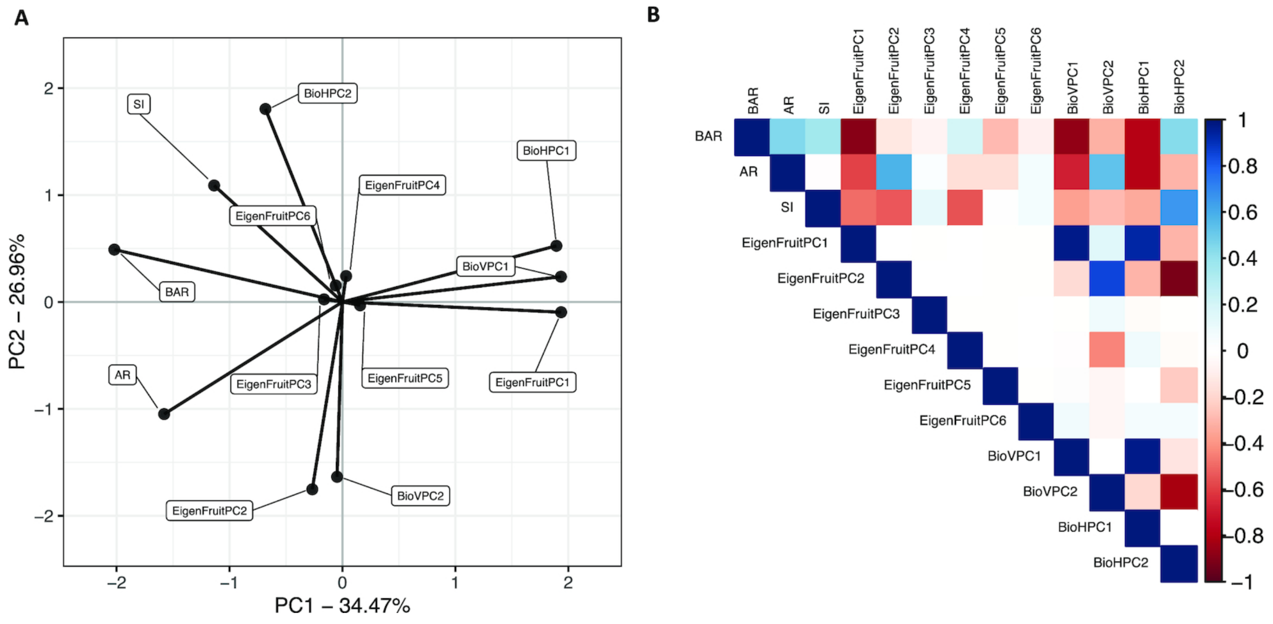
> Correlations between all 68 features used in this study. Blue indicates positive correlations, and red, negative correlations.

Figure 6:



> **Results from feature selection.** (A) Out-of-bag error for each of the 13 selected features. Horizontal dashed lines are the median (0.047) and mean (0.12) OOB. For each trait shown, the lower vertical dashed line is the first quartile, the lower boundary of the gray box to the horizontal black line is the second quartile, the horizontal black line to the upper boundary of the gray box is the third quartile, and the upper dashed line is the fourth quartile. Points not in the quartile range are considered outliers. (B) The relative importance of each feature within each level of k. The 13 selected features explain >80% of the weight attributed to all of the features, excluding K = 9 and 10.

Figure 7:



> Relationship between selected features. (A) Principal directions of the feature variance-covariance matrix among the 13 features selected for classification. (B) Pearson correlation matrix of the 13 selected features. Blue indicates positive correlations, and red, negative correlations.

Table 1:

> Broad-sense heritability of selected features

Feature	H2	k Selected	Normalized eigenvalue (80%,50%,20%)	Feature set
EigenFruit PC1	0.68	9	0.26 (0.27, 0.27, 0.26)	13, 5, 3
EigenFruit PC2	0.58	8	0.14 (0.14, 0.14, 0.14)	13, 5
EigenFruit PC3	0.00	3	0.05 (0.06, 0.05, 0.06)	13
EigenFruit PC4	0.69	5	0.04 (0.04, 0.05, 0.04)	13
EigenFruit PC5	0.43	4	0.03 (0.03, 0.04, 0.03)	13
EigenFruit PC6	0.47	5	0.03 (0.03, 0.03, 0.03)	13
Vertical biomass profile PC1	0.67	9	0.65 (0.66, 0.66, 0.66)	13, 5, 3
Vertical biomass profile PC2	0.49	4	0.17 (0.17, 0.16, 0.17)	13
Horizontal biomass profile PC1	0.65	9	0.44 (0.44, 0.46, 0.44)	13, 5, 3
Horizontal biomass profile PC2	0.62	3	0.36 (0.36, 0.35, 0.37)	13, 5
Bounding aspect ratio	0.71	8	NA	13
Shape index	0.72	4	NA	13
Ellipse aspect ratio	0.58	4	NA	13

> Broad-sense heritability (H2) estimated on a per-line basis.
> k selected is the number of classification models that a feature was selected in, out of 9 (i.e., k = [2, 10]).
> Normalized eigenvalues is the eigenvalue associated with a specific PC divided by the sum of all eigenvalues.
> The large value is the normalized eigenvalue from the full data set. Values in parentheses contain the normalized eigenvalues for the 80%, 50%, and the 20% training sets, respectively.
> Feature set indicates in which of the 3 sets a given feature was included.