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 определили 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 изучения 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, могут 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)*, аллооктоплоида (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 u Fragaria chiloensis)*, множественными последовательными интрогрессиями генетического разнообразия подвидов *F. virginiana* и *F. chiloensis* в последующих поколениях и искусственным отбором 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].

Многие подходы к фенотипированию плодов основаны на том, что видит человеческий глаз при сортировке плодов по дискретным описательным категориям для плоских (2D) форм (например, ромбических и почковидных) [14–19]. Категории являются либо номинальными [11, 20, 21], существующими только по названию, либо порядковыми, относящимися к позиции в упорядоченном ряду или по градиенту [15, 16, 21]. Классификация по категориям часто

является трудоемкой и склонна к предвзятости человека, которая может возрасти в зависимости от сложности задачи и временных затрат [22, 23]. Альтернативные подходы к оценке полагаются на морфометрию и машинное обучение для автоматизации классификации; например, сортировка фруктов по категориям формы как в томате [11], так и в клубнике [20]. Неконтролируемые методы машинного обучения (например, кластеризация k-средних), в отличие от контролируемых методов, полезны для обнаружения и кластеризации образов, в то время как контролируемые методы машинного обучения (например, машины опорных векторов) полезны для прогнозирования и классификации [24, 25]. Неконтролируемая кластеризация позволяет рассчитать несколько показателей производительности и переоснащения модели, чтобы сбалансировать сжатие и точность. Однако категории, полученные из этих методов, не имеют порядка, что приводит к необходимости соответствующего преобразования в порядковую шкалу, более подходящую для количественного генетического анализа [26–30]. В этом контексте порядковые категории дают интерпретацию взаимосвязи или расстояния от других категорий фигур в серии. Чтобы сделать возможной эту интерпретацию, мы разработали метод для подтверждения прогрессии по категориям форм фруктов, полученных из неконтролируемых методов машинного обучения. *Принципиальная прогрессия k кластеров* (PPKC) позволила нам произвольно определить подходящий градиент формы для статистического анализа с использованием эмпирических данных. Преимущества PPKC по сравнению с порядковой шкалой, определенной вручную, заключаются в том, что она не требует произвольных априорных решений и не контролируется, что устраняет дополнительные ошибки оператора. Здесь мы опишем подходы для перевода цифровых изображений клубники в вычисленные фенотипические переменные для идентификации и классификации форм фруктов.

> 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.

Форма и анатомия плода являются сложными, многомерными и, возможно, абстрактными фенотипами, которые часто не полностью или интуитивно не описываются плоскими дескрипторами и отдельными качественными или количественными переменными. Помимо качественных определений, используемых в систематике растений [18, 20], ссылки на форму плодов охватывают широкий спектр математических параметров и геометрических показателей, которые устанавливают количественные измерения органов растений [19, 31–33]. Так же, как человеческие лица или урожай зерна, форма и анатомия плода являются продуктами основных генетических и негенетических детерминант фенотипической изменчивости в популяции [34, 35]. Количественные фенотипические измерения позволили исследователям раскрыть некоторые генетические основы формы плодов у томатов [36, 37], перца [38, 39], груши [40], дыни [35], картофеля [41] и клубники [9 , 42]. Тем не менее, основные генетические детерминанты формы плодов остаются неясными или недостаточно изученными у клубники, отчасти потому, что исследователи еще не перевели атрибуты формы плодов в целостные количественные переменные, которые могут дать возможность идентификации основных генов или локусов количественных признаков через *genome-wide association studies (GWAS)* и другие количественные генетические подходы [43–46]. Количественные характеристики часто основаны на линейных метриках расстояния (например, высота, ширина и периметр) и обычно модифицируются в составные дескрипторы, которые устраняют влияние размера (например, соотношение сторон или округлость) [40, 42, 47]. Однако составные линейные дескрипторы часто имеют ограниченное разрешение по сравнению с более полными, многомерными дескрипторами [33]. *Эллиптический анализ Фурье (EFA)* количественно определяет форму плода по замкнутому контуру путем преобразования замкнутого контура в взвешенную сумму гармонических функций [12, 48–51]. *Generalized Procrustes analysis (GPA)* количественно определяет расстояние между наборами биологически гомологичных или математически

сходных ориентиров на поверхности объекта [48, 51–57]. Форма плода также может быть описана с использованием линейных комбинаций интенсивности пикселей на цифровых изображениях, экстраполируемых на основе анализа, обычно используемого для количественного определения цветовых моделей и распознавания лиц [13, 58–63]. Подобные основанные на пикселях дескрипторы недавно были названы *«скрытыми пространственными фенотипами»* и возникают в результате неконтролируемого анализа (т.е. *Анализа главных компонентов [PCA]* и *auto-encoding neural networks*), который позволяет компьютеру создавать новые, независимо распределенные функции непосредственно из изображения [64, 65]. Здесь мы создаем словарь из 68 количественных признаков, включая дескрипторы на основе линейных, контурных, ориентирных и пиксельных элементов, чтобы исследовать качество различных признаков при подготовке к количественному генетическому анализу.

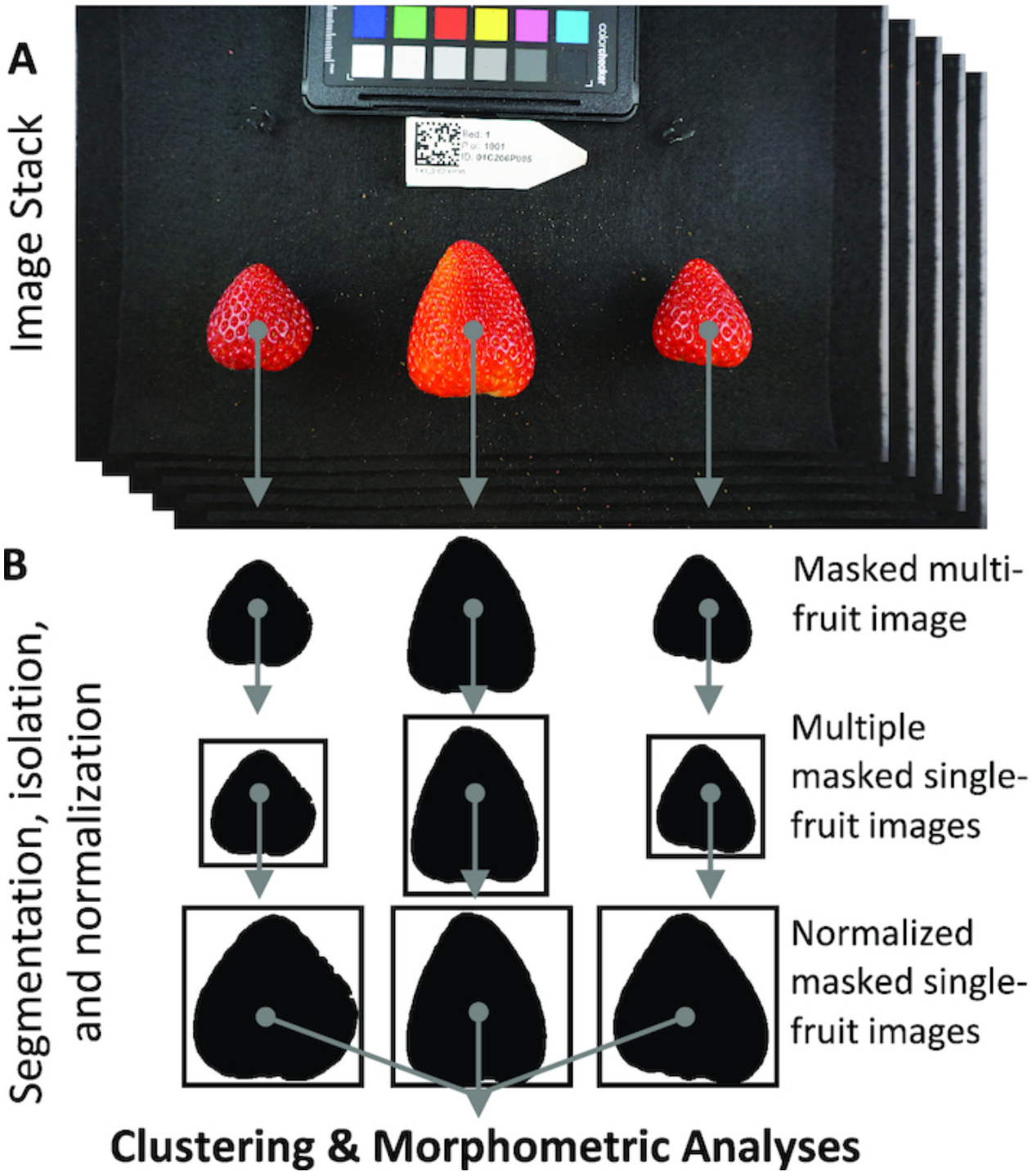
> Fruit shape and anatomy are complex, multi-dimensional, and, potentially, abstract phenotypes that are often not completely or intuitively described by planar descriptors and individual qualitative or quantitative variables. Beyond the qualitative definitions used in plant systematics [18, 20], references 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.

Конечная цель нашего исследования заключалась в разработке наследственных фенотипических переменных для описания формы плодов, которые затем можно было бы использовать для выявления генетических факторов, лежащих в основе фенотипических различий в форме плодов. Фенотипирование и аналитический рабочий процесс для этого исследования суммированы на рисунках 1 и 2. Сначала мы опишем и продемонстрируем применение PPKC, который преобразует категории, обнаруженные из неконтролируемых методов машинного обучения, в более удобный и аналитически управляемый порядковый масштаб [26, 28, 29]. Затем мы исследуем связь между категориями, полученными машиной, и 68 количественными характеристиками, извлеченными из цифровых изображений. Затем мы применяем *регрессию случайных лесов* для выбора критических наборов количественных признаков для классификации и используем контролируемые методы машинного обучения, включая_ регрессию опорных векторов (SVR)_ и *линейный дискриминантный анализ (LDA)*, для определения точности классификации форм. Мы обнаружили, что в сильно одомашненной размножающейся популяции есть только несколько категорий, представляющих интерес, и что для точной классификации формы по обнаруженным категориям необходимо небольшое количество признаков. Мы также нашли, что категории порядковых форм очень наследуемы, и что признаки, необходимые для точной классификации, также наследуемы.

> 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:

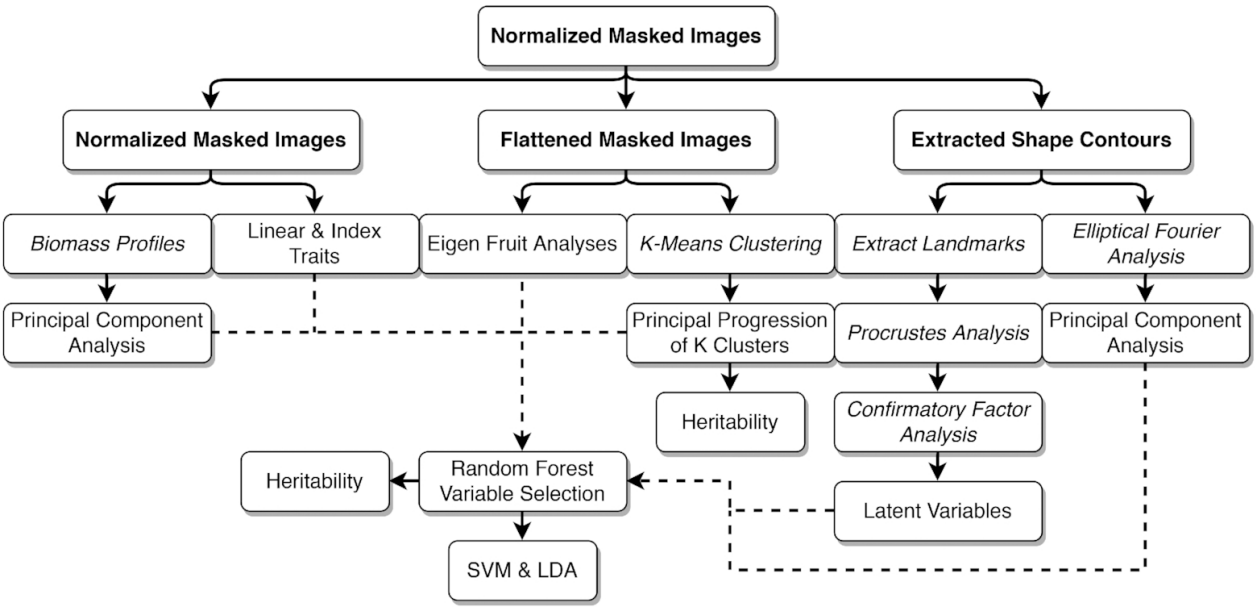


(1)

Пример обработки конвейера. **(A)** Пользователь собирает стопку изображений, содержащих несколько клубник и уникальный QR-код. **(B)** Все изображения затем сегментируются с использованием алгоритма SIOX, реализованного в ImageJ. Затем каждый объект вырезается из исходного изображения на основе координат его ограничительного прямоугольника в R 3.5.3. Белые пиксели затем добавляются к краям каждого кадра, пока все изображения не станут 1000 × 1000 пикселей. Интересующие области затем масштабируются таким образом, что главная ось каждого объекта становится 1000 пикселей в ImageJ. Выходные изображения не зависят от масштаба и сохраняют исходное соотношение сторон.

> 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:



(2)
Анализ конвейера для этого исследования. Все изображения начинаются как нормализованные, бинарные изображения с рисунка 1. Изображения затем следуют по каждому из путей через различные морфометрические извлечения признаков, включая линейные геометрические особенности, *анализ профиля биомассы (BPA)*, анализ *EigenFruit*, анализ *Procrustes* и *эллиптический анализ Фурье* как нормализованные или сплющенные изображения (например, линейный анализ, *анализ BPA* и *EigenFruit*) или в виде контуров формы (например, *GPA* и *EFA*). Сглаженные двоичные изображения используются для кластеризации k-средних, а затем для PPKC.

> 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.

Описание данных

Данные, опубликованные в этой статье, содержат цифровые изображения 6874 плодов клубники от 572 гибридов, полученных из Калифорнийского университета в Дэвисе, программа по выращиванию клубники. Данные для этой статьи, включая предварительно обработанные изображения (рис. 1A), обработанные изображения (рис. 1B) и извлеченные элементы (см. Методы, рис. 2), доступны на Zenodo [66]. Предварительно обработанные изображения обычно содержали несколько ягод на изображение вместе со штрих-кодом матрицы данных, указывающим идентификатор генотипа, и другими элементами дизайна эксперимента. Обработанные изображения представляют собой двоичные изображения в масштабе 1000 × 1000 пикселей отдельных фруктов. Извлеченный набор данных объектов предоставляется в виде файла CSV. Код для репликации анализов в этой статье предоставлен в репозитории GitHub [67]. Кроме того, снимки кода и данных, поддерживающих эту работу, доступны в репозитории GigaScience, GigaDB [68]. Мы надеемся, что публикация этих данных поможет другим в разработке новых морфометрических подходов для лучшего понимания генетического, фруктового и экологического контроля формы фруктов в клубнике и, в более широком смысле, в других фруктах, овощах и специальных культурах.

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.

Анализ

Кластеризация k-средних

Кластеризация k средних позволяет быстро обнаруживать закономерности в больших многомерных наборах данных, используемых для кластеризации, принятия решений и уменьшения размеров [24, 69, 70]. Это итеративный алгоритм, который разбивает набор данных на заранее определенное количество непересекающихся кластеров k, минимизируя сумму квадратов расстояний от каждой точки данных до центроида кластера. Центроид соответствует среднему значению всех точек, назначенных кластеру. Здесь мы использовали k-means для кластеризации сглаженных бинарных изображений (рис. 1; см. Методы). Отдельные плоды были отделены от фона изображения в виде бинарной маски, нормализованы по главной оси, изменены до 100 × 100 пикселей и сведены в вектор (Рис. 1 и 2; см. Методы). Мы представили каждое изображение как вектор из 10000 элементов, содержащий двоичные значения пикселей. Мы смогли быстро и надежно назначить изображения классам, используя кластеризацию k-средних. В этом эксперименте мы допустили, чтоб k, количество разрешенных категорий, варьировалось от 2 до 10. Этот диапазон был выбран, потому что мы ожидаем, что система классификации на основе человека не будет иметь скорости или надежности, необходимых для этой задачи, особенно для больших значения k.

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.

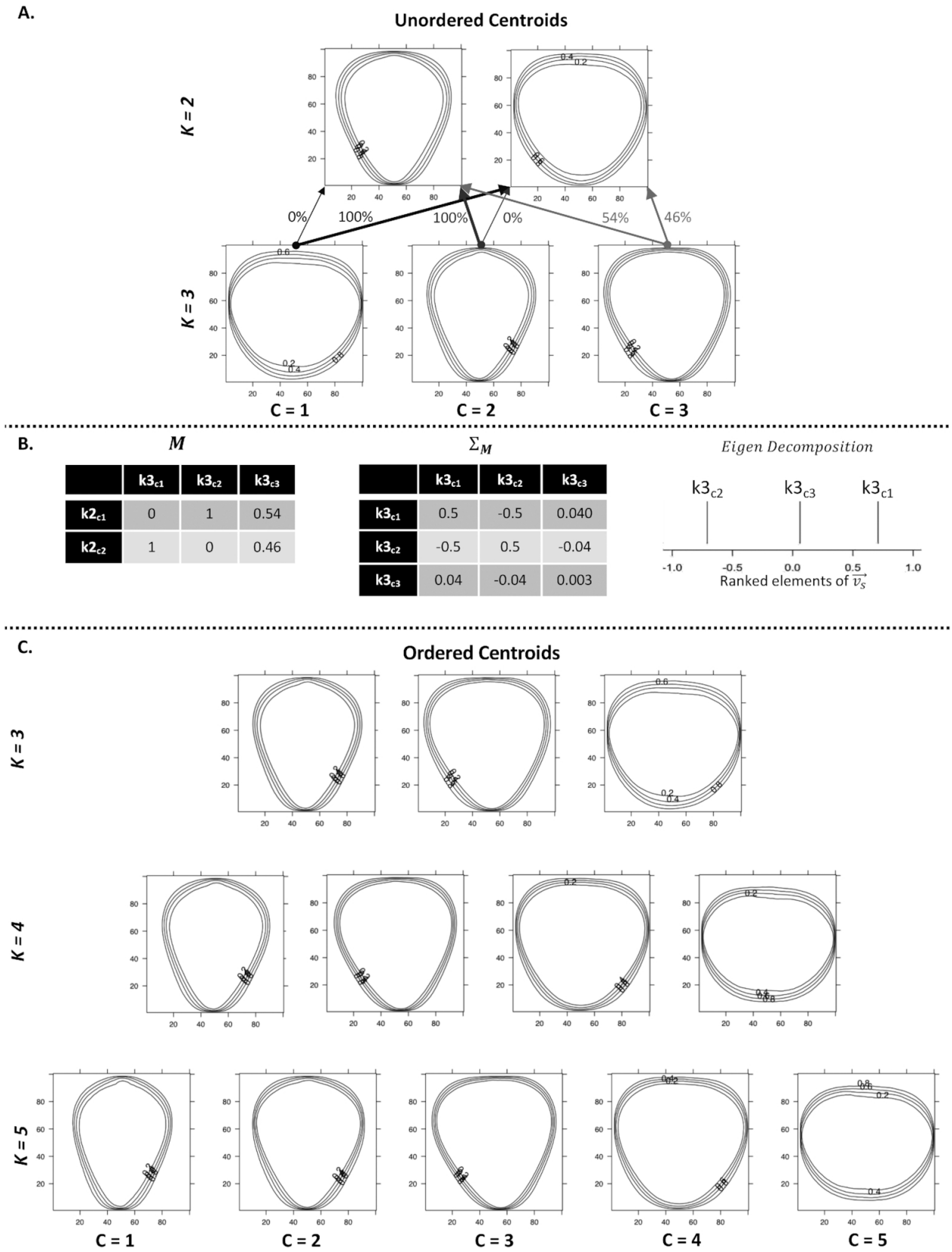
Главная прогрессия k кластеров

Кластеризация k-Means не назначает прогрессию или градиент обнаруженным классам. Тем не менее, оценка и порядковые черты, как правило, более полезны и чаще встречаются в количественных генетических исследованиях, чем номинальные шкалы [26, 28, 29, 71]. Мы разработали новый метод для преобразования категорий, полученных из k-средних, в порядковый масштаб, который мы называем PPKC (Рис. 3; Алгоритм 1). Этот метод основан на кластеризации k-средних для категоризации изображений и может быть использован для эмпирического определения соответствующего порядкового масштаба в номинальных данных. k-Means поддерживает несколько метрик для оценки производительности и переоснащения модели, включая *скорректированный* R^2, *информационный критерий Акаике (AIC)* и *байесовский*

информационный критерий (BIC), который позволяет пользователям определять наиболее подходящее значение k с учетом наблюдаемых данных. Градиент между кластерами оценивали путем выполнения PCA на ковариационной матрице, отражающей структурированные отношения между фокусным кластером и всеми ранее обнаруженными кластерами.

> ## 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 R^2, 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:



(3)

Пример использования ППКК. **(A)** После выполнения кластеризации k-средних кластерам случайным образом присваивается числовое значение (1,2,…,k). Когда k > 2, это значение становится номинальным. PPKC опирается на тот факт, что порядок через кластеры при k = 2 имеет идентичные интерпретации в любом направлении. Линии, представляющие центроид каждой группы, отражают 20-й, 40-й, 60-й и 80-й квантили, выходя из центра каждого изображения. **(B)** Слева, табличное представление результирующей матрицы из уравнения (1). Каждая ячейка представляет долю изображений в классе столбцов и в классе строк, нормализованную по количеству изображений в классе столбцов. **(B)** Середина, табличное представление $\mathbf{\Sigma_M}$. **(B)** Справа: ранжированные элементы $\vec{v}_s$ показаны в

числовой строке. **(C)** После использования PPKC порядок групп четко идентифицирован. В этом примере, показывающем k = [3, 5], обнаруженный порядок, кажется, имеет тенденцию от высоких и тонких ягод к более треугольным формам, заканчивающимся ягодами, которые являются короткими и широкими.

> 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 $\mathbf{\Sigma_M}$. **(B)**Right, The ranked elements of $\bar{v}_s$ s 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.

Сначала мы назначаем каждое сплющенное двоичное изображение (рис. 1) категории, используя метод k-средних. Мы назначаем кластер каждому изображению и позволяем количеству кластеров, k, колебаться от 2 до 10. Порядок впоследствии определяется с помощью PPKC (Рис. 3, Алгоритм 1). Когда k = 2, порядок родства считается произвольным, и оба $k2_{c1} \rightarrow k2_{c2}$ и $k2_{c2} \rightarrow k2_{c1}$ имеют одинаковое значение, где «→» указывает прогрессию обнаруженных категорий. Любой данный порядок и его обратное считаются эквивалентными, и это также относится к более высоким уровням k; например, гипотетическое ранжирование кластеров 1, 4, 2, 3 считается эквивалентным 3, 2, 4, 1, потому что относительные отношения между k кластерами идентичны в обоих (например, c3 больше относится к c2, чем либо c1, либо c4). Для каждого интересующего кластера (например, $k4_{c1}$, $k4_{c2}$, $k4_{c3}$ и $k4_{c4}$) мы рассчитываем долю каждого кластера, который произошел от $k3_{c1}$, $k3_{c2}$ или $k3_{c3}$ и $k2_{c1}$ или $k2_{c2}$ (то есть, все прежние классификации). Эти пропорции позволяют оценить сходство между фокусным кластером (например, $k4_{c1}$) и кластерами всех предыдущих значений k. Затем мы нормализуем пропорции по общему количеству изображений в фокусном кластере (например, $k4_{c1}$, $k4_{c2}$, $k4_{c3}$ и $k4_{c4}$) (уравнение 1).

> 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 $k2_{c1} \rightarrow k2_{c2}$ and $k2_{c2} \rightarrow k2_{c1}$ 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., $k4_{c1}$, $k4_{c2}$, $k4_{c3}$, and $k4_{c4}$), we calculate the proportion of each cluster that came from $k3_{c1}$, $k3_{c2}$, or $k3_{c3}$ and $k2_{c1}$ or $k2_{c2}$ (i.e., all former classifications). These proportions enable the estimation of similarity between a focal cluster (e.g., $k4_{c1}$) 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., $k4_{c1}$, $k4_{c2}$, $k4_{c3}$, and $k4_{c4}$) (Equation 1).

Для каждого уровня k > 2 построим M - прямоугольную матрицу размера $(k^2 - k)/2–1 \times k$ (алгоритм 1, строка 13). Сумма каждого столбца должна равняться k - 2. Пропорции - это непрерывные значения в диапазоне [0, 1], которые описывают происхождение конкретного фокусного кластера (например, $k4_{c1}$), так как он относится к кластерам k = 3 и k = 2 или все кластеры
[2, k - 1]. В следующем примере k = 4:

> For every level of k > 2, we construct M, a rectangular matrix of size $(k^2–k)/2–1 \times k$(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., $k4_{c1}$) as it relates to the clusters of k = 3 and k = 2 or all clusters [2, k − 1]. In the following example, k = 4:

$$\mathbf{M} = \begin{bmatrix} \frac{|k4_{c1} \wedge k3_{c1}|}{|k4_{c1}|} & \frac{|k4_{c2} \wedge k3_{c1}|}{|k4_{c2}|} & \frac{|k4_{c3} \wedge k3_{c1}|}{|k4_{c3}|} & \frac{|k4_{c4} \wedge k3_{c1}|}{|k4_{c4}|} \\ \frac{|k4_{c1} \wedge k3_{c2}|}{|k4_{c1}|} & \frac{|k4_{c2} \wedge k3_{c2}|}{|k4_{c2}|} & \frac{|k4_{c3} \wedge k3_{c2}|}{|k4_{c3}|} & \frac{|k4_{c4} \wedge k3_{c2}|}{|k4_{c4}|} \\ \frac{|k4_{c1} \wedge k3_{c3}|}{|k4_{c1}|} & \frac{|k4_{c2} \wedge k3_{c3}|}{|k4_{c2}|} & \frac{|k4_{c3} \wedge k3_{c3}|}{|k4_{c3}|} & \frac{|k4_{c4} \wedge k3_{c3}|}{|k4_{c4}|} \\ \frac{|k4_{c1} \wedge k2_{c1}|}{|k4_{c1}|} & \frac{|k4_{c2} \wedge k2_{c1}|}{|k4_{c2}|} & \frac{|k4_{c3} \wedge k2_{c1}|}{|k4_{c3}|} & \frac{|k4_{c4} \wedge k2_{c1}|}{|k4_{c4}|} \\ \frac{|k4_{c1} \wedge k2_{c2}|}{|k4_{c1}|} & \frac{|k4_{c2} \wedge k2_{c2}|}{|k4_{c2}|} & \frac{|k4_{c3} \wedge k2_{c2}|}{|k4_{c3}|} & \frac{|k4_{c4} \wedge k2_{c2}|}{|k4_{c4}|} \end{bmatrix} \quad (1)$$

Затем мы рассчитываем дисперсионно-ковариационную матрицу уравнения (1) (алгоритм 1, строка 18). Матрица дисперсии-ковариации $\mathbf{\Sigma_M}$ представляет отношение между каждым фокусным кластером (например, $k4_{c1}$, $k4_{c2}$, $k4_{c3}$ или $k4_{c4}$).

> We then calculate the variance-covariance matrix of Equation (1) (Algorithm 1 line 18). The variance-covariance matrix, $\mathbf{\Sigma_M}$, represents the relationship between each focal cluster (e.g., $k4_{c1}$, $k4_{c2}$, $k4_{c3}$, or $k4_{c4}$).

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

Затем мы выполняем собственное разложение по уравнению (2), используя следующее уравнение (алгоритм 1, строка 19).

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

$$\mathbf{\Sigma_M} = \mathbf{V\Lambda V}^{-1}. \quad (3)$$

В уравнении (3) $\mathbf{\Lambda}$ является диагональной матрицей со значениями, соответствующими k собственным значениям $\mathbf{\Sigma_M}$, а $\mathbf{V}$ является квадратной матрицей, содержащей собственные векторы, связанные с собственными значениями в $\mathbf{\Lambda}$. Затем мы извлекаем собственный вектор, связанный с наибольшим собственным значением, $\vec{\mathbf{v}}_{\boldsymbol{\lambda max}}$. Мы упорядочим элементы $\vec{\mathbf{v}}_{\boldsymbol{\lambda max}}$ так, чтобы результирующий вектор $\vec{\mathbf{v}}_{\mathbf{s}}$ обладал свойством $v_{s_1} \leq \ldots \leq v_{s_k}$. Мы не рассматриваем расстояние между элементами в $\vec{\mathbf{v}}_{\mathbf{s}}$, только их ранг. Затем кластеры индексируются для соответствия рангу связанных элементов в $\vec{\mathbf{v}}_{\mathbf{s}}$. Существует не более k собственных значений, связанных с собственными векторами длины k, поскольку $\mathbf{\Sigma_M}$ равно k × k. Собственное разложение используется для описания главной оси дисперсии в $\mathbf{\Sigma_M}$. Теоретически, эта перспектива ковариации должна быть способна эффективно разделять классы, потому что она описывает линейную ось, содержащую наибольшее количество независимых вариаций, а решения не являются произвольными. Поэтому значение, которое категория принимает на этой составной оси, наводит на мысль о ее линейном отношении к другим рассматриваемым k категориям. Тем не менее, мы отмечаем, что отношения, содержащие ветви, пузырьки и другие топологические особенности, не будут получены точно. В этом исследовании мы не можем сообщить о визуально

значимом порядке, когда k ≥ 9 (рис. S1) [66]. Изменение в прогрессии может быть отражением перенастройки числа групп в кластеризации k-средних. Значительное изменение наклона при k = 4 в общих внутригрупповых суммах квадратов, AIC и скорректированного R^2 свидетельствует о переоснащении (рис. S2) [66]. Наиболее убедительным доказательством для 4 кластеров является BIC, который минимизируется при k = 4 (рис. S2D) [66]. Элементы $\vec{\mathbf{v}}_{\mathbf{S}}$ имеют тенденцию сходиться друг с другом при увеличении k, что может указывать на небольшую биологическую информацию в новых кластерах и переоснащение (рис. S3) [66]. Учитывая, что в этом алгоритме рассматриваются только относительно небольшие ковариационные матрицы, время вычисления порядка k = [3, …, 10] в MacBook Pro 2,9 ГГц Core i5 в начале 2015 года с 8 ГБ памяти составляет <0,2 секунды.

> In Equation (3), $\mathbf{\Lambda}$ is a diagonal matrix with values corresponding to the k eigenvalues of $\mathbf{\Sigma_M}$ and $\mathbf{V}$ is a square matrix containing eigenvectors associated with the eigenvalues in $\mathbf{\Lambda}$. We then extract the eigenvector associated with the largest eigenvalue, $\vec{\mathbf{v}}_{\boldsymbol{\lambda max}}$. We order the elements of $\vec{\mathbf{v}}_{\boldsymbol{\lambda max}}$ such that the resultant vector, $\vec{\mathbf{v}}_{\mathbf{S}}$, has the property $v_{s_1} \leq \ldots \leq v_{s_k}$. We do not consider the distance between elements in $\vec{\mathbf{v}}_{\mathbf{S}}$, only their rank. The clusters are then indexed to match the rank of the associated elements in $\vec{\mathbf{v}}_{\mathbf{S}}$. There are at most k eigenvalues associated with eigenvectors of length k due to $\mathbf{\Sigma_M}$ being k × k. Eigen decomposition is used to describe the major axis of variance in $\mathbf{\Sigma_M}$. 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 ≥ 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 R^2 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 $\vec{\mathbf{v}}_{\mathbf{S}}$ 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, …, 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 $\Sigma_{\mathbf{M}}$
    $\Lambda = \lambda_{max}, ..., \lambda_k\mathbf{I}$ (▷λ_{max} is the largest eigenvalue of $\Sigma_{\mathbf{M}}$.
    $\vec{v}_{\lambda_{max}} = \mathbf{V}_{.,1}$ (▷$\vec{v}_{\lambda_{max}}$ is the eigenvector of λ_{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

Алгоритм главной прогрессии кластеров K (PPKC)

> Principal Progression of K Clusters (PPKC) Algorithm

Наследование в широком смысле упорядоченных категорий

Для каждого значения k наследственность в широком смысле (H^2) на основе среднего входа оценивалась с использованием общей линейной смешанной модели с кумулятивной функцией логит-линка (уравнения 4 и 5) [72]. Для

этого набора данных H^2 обычно был высоким, в диапазоне от H^2 = 0,80 до 0,98, даже при k → 10 (таблица 2). Эти оценки H^2 очень похожи на оценки, представленные Антанавичюте [16] (т.е. H2 = 0,84). Когда H^2 признака находится в этом диапазоне, это указывает на то, что независимые репликации одних и тех же индивидов имеют высокую степень сходства и что большая часть вариаций среди индивидов возникла из генетических вариаций между индивидами. Поскольку растительный материал, использованный в этом исследовании, происходил из генетических клонов, любое изменение формы плодов среди повторностей происходило от случайных, ненаблюдаемых эффектов. Ожидается, что при k ≥ 9 точность оценок H^2 будет ниже, чем при k ≤ 8, поскольку градиент фенотипа, по-видимому, задан неправильно. В этом наборе гермоплазмы мы предлагаем набор из 4 основных классов для классификации формы плода (рис. 3 и S2) [66]. При увеличении k от 5 до 10 визуальное сходство некоторых кластеров становится высоким (рис. S1) [66], что указывает на меньшее количество соответствующих границ (рис. S3) [66]. Как указано, в этих данных имеются убедительные доказательства того, что в этих данных имеется 4 различных кластера (рис. S2) [66].

Broad-sense heritability of ordered categories

For each value of k, broad-sense heritability (H^2) 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, H^2 was generally high, ranging from H2 = 0.80 to 0.98, even as k → 10 (Table 2). These estimates of H^2 are very similar to those reported by Antanaviciute [16] (i.e., H^2 = 0.84). When the H^2 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 H^2 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].

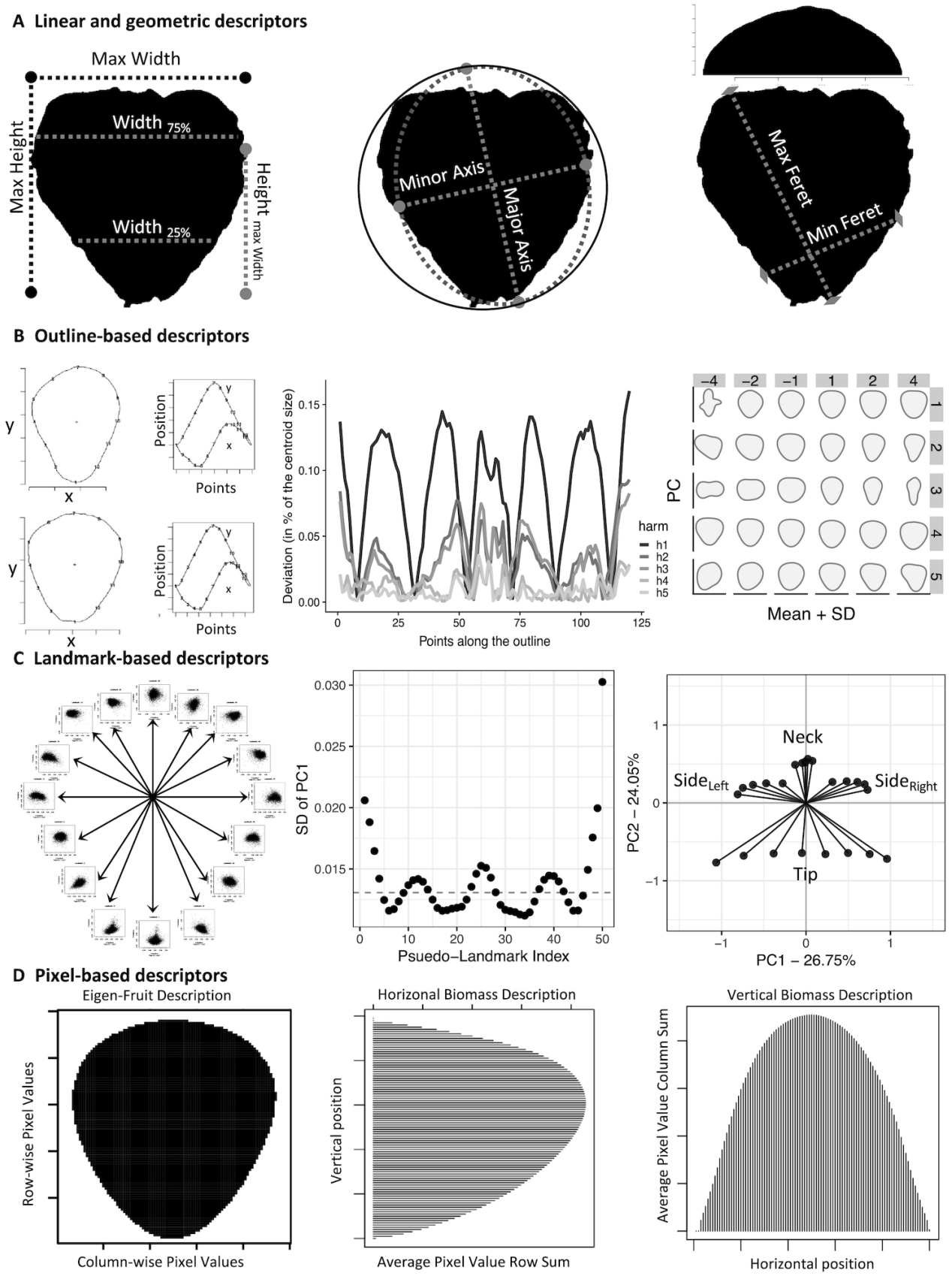
Выбор объектов с использованием случайных лесов

Чтобы выяснить, какие из 68 количественных признаков (суммированных на рисунках 4 и 5) охватывают и отражают различия в категориях форм, для оценки важности признаков использовалось контролируемое машинное обучение (см. Методы) [73]. Из 68 функций, используемых в качестве предикторов в регрессии случайных лесов (см. Методы), мы выбрали только 13. Ошибка «вне пакета» (OOB) - это оценка того, насколько плохо работают модели, когда конкретный объект исключен и похож на ошибку, оценивается по повторной выборке из jackknife resampling (рис. 6). Таким образом, характеристики с более высокими оценками имеют тенденцию быть более релевантными для классификации и прогнозирования. В этом эксперименте функции могут быть выбраны только 9 раз, один раз на значение k. Мы сохранили особенности, которые были выбраны на уровне ≥3 k для использования в качестве независимых переменных в классификации (Таблица 1). 13 выбранных объектов составляли > 80% важности, назначенной 68 признакам для всех значений k (рис. 6B). Здесь, использование EigenFaces, анализа 1980-х годов, предназначенного для классификации человеческих лиц, было переосмыслено для количественного определения и классификации формы плодов в клубнике [58–61]. Пиксельные элементы доминировали над выбранными функциями и включают в себя основные компоненты (ПК) 1-7 анализа EigenFruit ($EigenFruitPC_{[1, 6]}$), ПК 1 и 2 вертикального профиля биомассы ($BioVPC_{[1, 2]}$) и ПК 1 и 2 горизонтального профиля биомассы ($BioHPC_{[1, 3]}$) (таблица 1 и рисунки 6 и 7). Эти признаки возникли из того же типа данных, который использовался в кластеризации k-средних (то есть интенсивности пикселей), что, вероятно, является причиной того, что они составляют большинство выбранных признаков (таблица 1 и рисунки 6 и 7). Также было выбрано несколько геометрических дескрипторов, включая ограничивающее соотношение сторон (BAR), индекс формы (SI) и соотношение сторон эллипса (AR) (таблица 1 и рисунки 6 и 7). Мы создали подмножество из 5 признаков со средним OOB ≥ 0,047 (рис. 6A). OOB = 0,047 было медианной ошибкой OOB для всех объектов всех классов. Этот набор функций включает $EigenFruitPC_{[1, 2]}$, BioVPC1 и $BioHPC_{[1]}$ (Таблица 1). Мы также создали третий меньший набор, который включал только $EigenFruitPC_1$, $BioVPC_1$ и $BioHPC_1$ со средним OOB ≥ 0,12 (рис. 6A). OOB = 0,12 было средней ошибкой OOB для всех функций всех классов. Преобладание дескрипторов на основе пикселей в этих выбранных подмножествах указывало на величину соответствующей информации о форме, которую они описали.

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 $EigenFruitPC_1$, $BioVPC_1$, and $BioHPC_1$ 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:



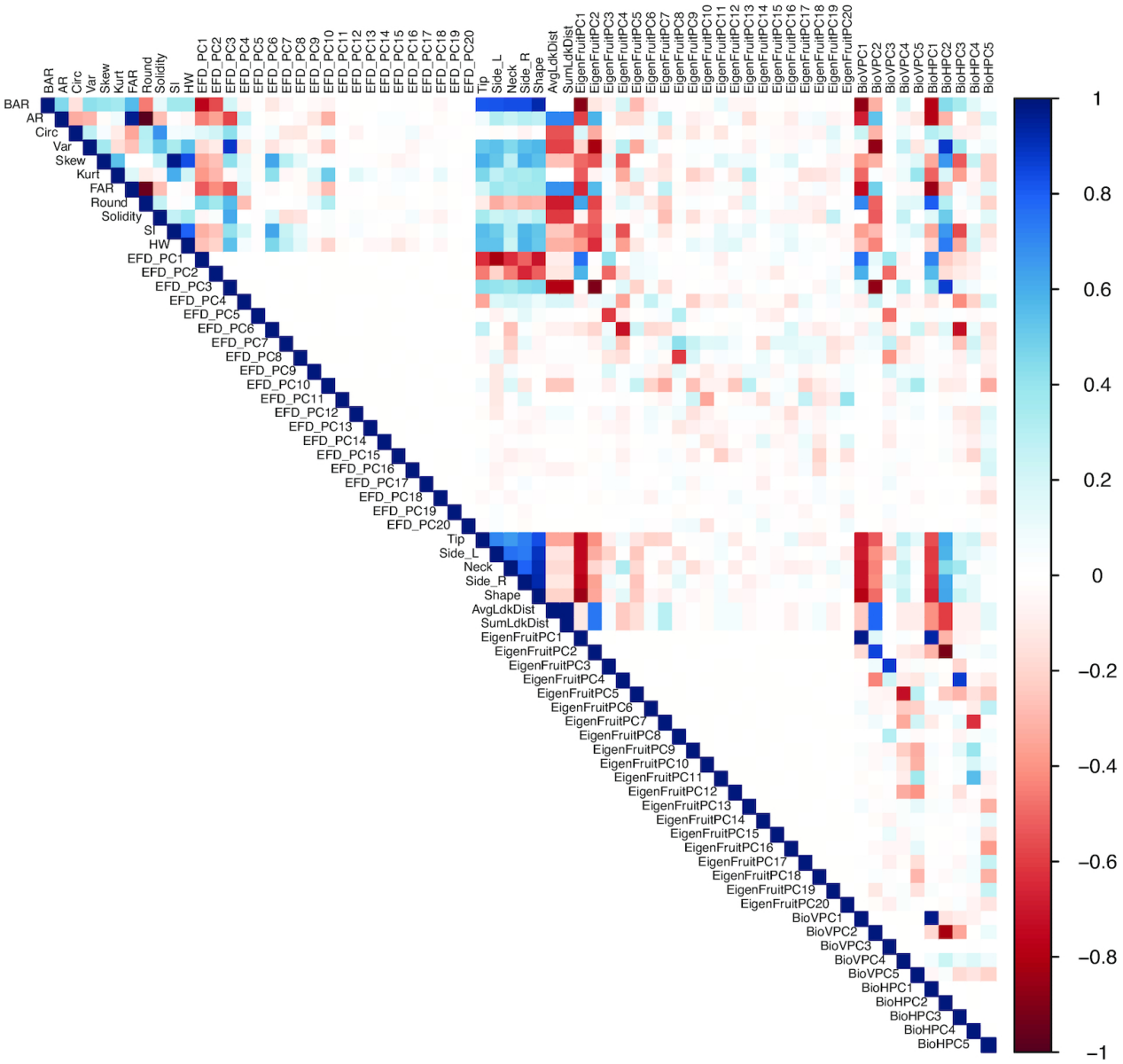
(4)

Словарь черт для этого исследования. **(A)** Линейные дескрипторы. Слева, Простые линейные измерения. Центр, наиболее подходящие оси эллипса. Для круга Round и Circ = 1. Правый, максимальный и минимальный коэффициент. Гистограмма представляет предельное распределение по горизонтальной оси, используемое для вычисления Var, Skew и Kurt. **(B)** Контур дескрипторов. Слева, 2 крайних левых изображения - это контуры 2 клубник с 12 равномерно расположенными точками. Графики справа показывают исходный замкнутый контур как две осциллирующие функции.

Центр, отклонения от замкнутого контура с увеличением гармоник (вред = [h1, h5]). Справа: график показывает эффекты ПК [1,5] (по вертикали) с размерами эффектов, [−4, 4] (по горизонтали) на средней фигуре. **(C)** дескрипторы ориентиров. Слева, 50 равномерно распределенных ориентиров извлекаются и обрабатываются как переменные объекты. Центр, Стандартное отклонение PC1 для каждого ориентира строится в последовательности. Пунктирная горизонтальная линия - среднее стандартное отклонение (SD). Справа, псевдо-ориентиры были выбраны для представления каждого региона с высокой дисперсией. Используя значения на первой главной оси в качестве наблюдаемых переменных, был проведен подтверждающий факторный анализ, чтобы вывести скрытые отношения к кончику, левой и правой стороне и форме шеи. **(D)** Пиксельные дескрипторы. Слева, Mean EigenFruit с использованием сплющенных двоичных изображений. Центр, Средняя горизонтальная биомасса, используя суммы строк изображения. Справа - средняя вертикальная биомасса с использованием сумм столбцов изображений.

> 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:

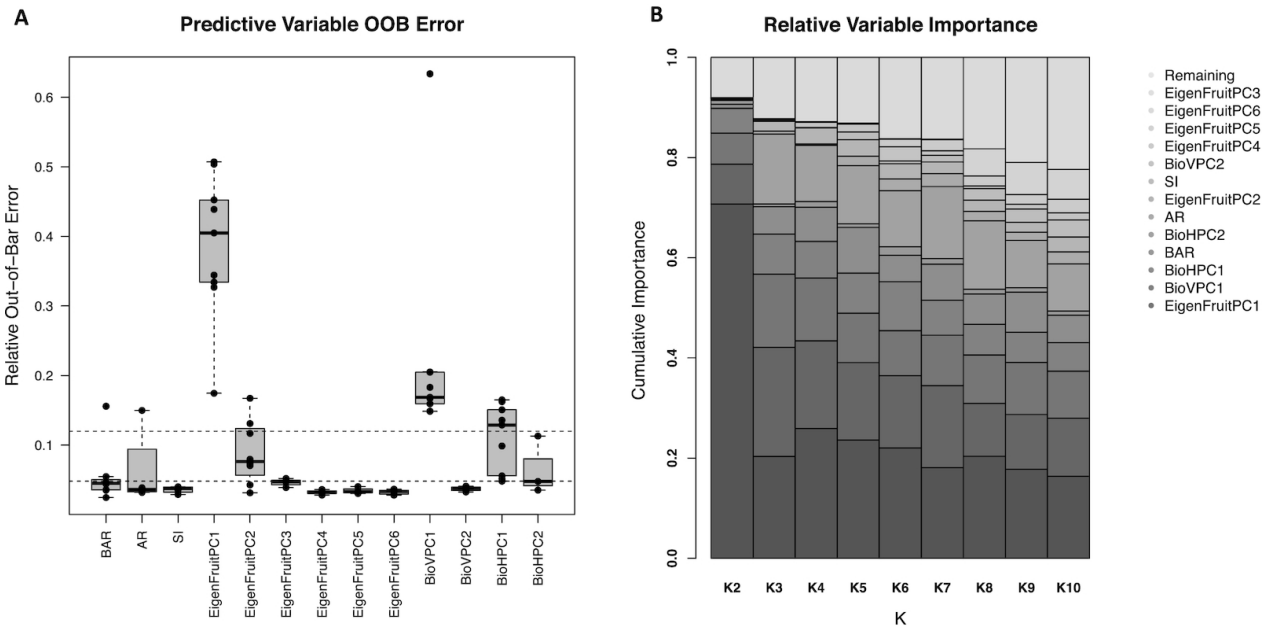


(5)

Корреляции между всеми 68 функциями, используемыми в этом исследовании. Синий цвет указывает на положительную корреляцию, а красный - на отрицательную.

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

Figure 6:

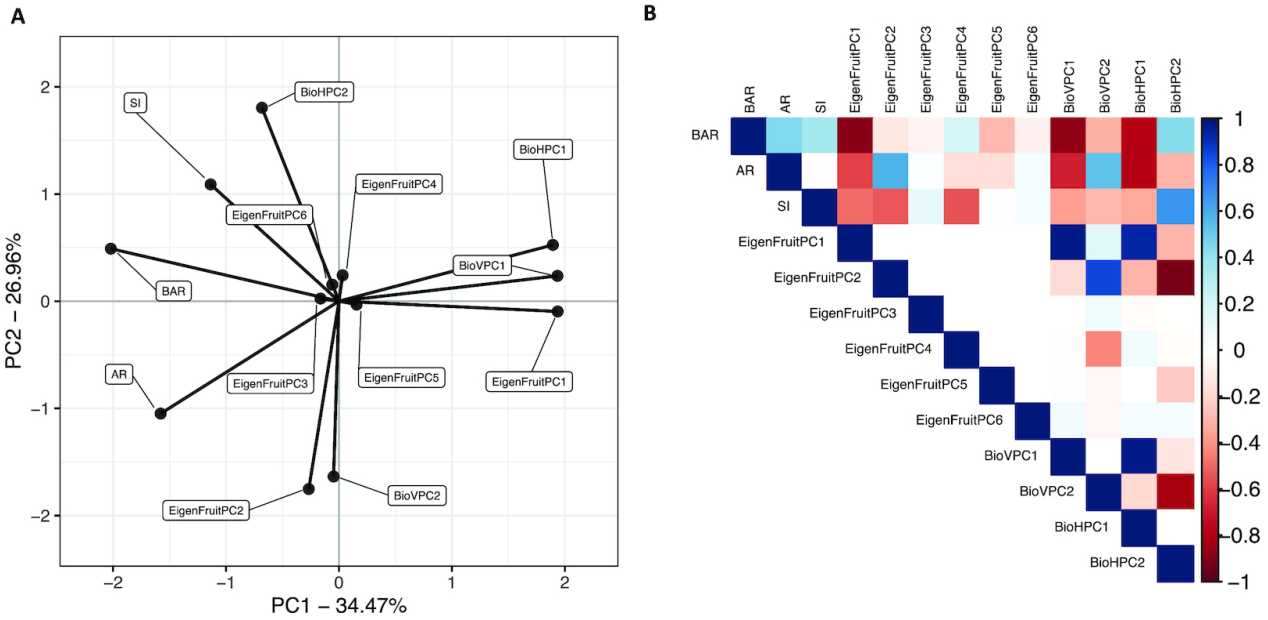


(6)

Результаты выбора функций. (A) Ошибка «вне пакета» для каждой из 13 выбранных функций. Горизонтальные пунктирные линии - медиана (0,047) и средняя (0,12) OOB. Для каждой показанной черты нижняя вертикальная пунктирная линия является первым квартилем, нижняя граница серого прямоугольника с горизонтальной черной линией - вторым квартилем, горизонтальная черная линия до верхней границы серого прямоугольника - третьим квартилем, и верхняя пунктирная линия - четвертый квартиль. Точки, не входящие в квартильный диапазон, считаются выбросами. **(B)** Относительная важность каждого признака на каждом уровне k. 13 выбранных признаков объясняют> 80% веса, приписанного всем признакам, за исключением K = 9 и 10.

> **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:



(7)

Связь между выбранными функциями. **(A)** Основные направления матрицы дисперсии-ковариации признаков среди 13

признаков, выбранных для классификации. **(B)** корреляционная матрица Пирсона из 13 выбранных признаков. Синий цвет указывает на положительную корреляцию, а красный - на отрицательную.

> 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:

Наследуемость отдельных функций в широком смысле

> Broad-sense heritability of selected features

Feature	H^2	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

Наследственность в широком смысле (H^2) оценивается на основе каждой строки.
k selected - это число моделей классификации, в которых был выбран элемент, из 9 (то есть k = [2, 10]).
Нормализованные собственные значения - это собственные значения, связанные с конкретным PC, деленные на сумму всех собственных значений.
Большое значение является нормализованным собственным значением из полного набора данных. Значения в скобках содержат нормализованные собственные значения для тренировочных наборов 80%, 50% и 20% соответственно.
Набор функций указывает, в какой из 3 наборов была включена данная функция.

> Broad-sense heritability (H^2) 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.

Наследственность в широком смысле и взаимосвязь выбранных признаков

Хотя ожидается, что непрерывный характер морфометрических признаков будет более благоприятным и обеспечит более высокое разрешение для количественного генетического анализа по сравнению с их категориальными аналогами, также важно, чтобы эти признаки были наследуемыми. H^2 для каждой характеристики был оценен на основе среднего клона с использованием линейной модели смешанных эффектов (см. Уравнения 5 и 6) [74]. H^2 для каждого признака представлен в таблице 1. Оценки H^2 для количественных признаков варьировались от низкого (> 0,3) до высокого (> 0,7). Оценки наследуемости согласуются с ранее сообщенными для фенотипов формы клубники и других видов растений [12, 42, 75].

Broad-sense heritability and relationship of selected features

While the continuous nature of the morphometric features is expected to be more conducive and provide higher resolution to quantitative genetic analyses compared to their categorical counterparts, it is also vital that these features be heritable. The H^2 for each feature was estimated on a clone-mean basis using a linear mixed-effects model (see Equations 5 and 6) [74]. The H^2 for each feature is reported in Table 1. Estimates of H^2 for the quantitative features ranged from low (>0.3) to high (>0.7). Heritability estimates were consistent with those previously reported for shape phenotypes in strawberry and other plant species [12, 42, 75].

На фиг.7A показаны направления матрицы дисперсии-ковариации признаков с признаками, помеченными, как на фиг.6. На фиг.7B показана матрица корреляции между 13 выбранными признаками. Для 5 признаков, выбранных по ошибке OOB (рис. 6), обозначенных «5» в Таблице 1, оценка H^2 составила ≥0,58. Поскольку большинство выбранных функций представляют собой ПК с различным анализом пикселей (рис. S5) [66], было много слабых корреляций (рис. 7B). Мы предполагаем, что важность этих функций отчасти обусловлена сходством необработанных данных (т. Е. Интенсивности двоичных пикселей), используемых в кластеризации k-средних для получения категорий форм и для анализа форм EigenFruit. Хотя ПК не коррелированы, мы наблюдали сильную корреляцию между ПК из различных анализов (рис. 7). $EigenFruitPC_1$ имеет сильную положительную корреляцию как с $BioVPC_1$, так и с $BioHPC_1$ (ρ = 0,98; P <2E-16 и ρ = 0,93; P <2E-16 соответственно), как и $EigenFruitPC_2$ с $BioVPC_2$ (ρ = 0,86; P <2E-16). $BioHPC_2$ отрицательно коррелировал как с $EigenFruitPC_2$, так и с $BioVPC_2$ (ρ = -0,92; P <2E-16 и ρ = -0,81; P <2E-16 соответственно). $BioHPC_3$ отрицательно коррелировал с $EigenFruitPC_4$ (ρ = -0,87; P <2E-16). BAR отрицательно коррелировал с $EigenFruitPC_1$, $BioVPC_1$ и $BioHPC_1$ (ρ = -0,89; P <2E-16, ρ = -0,87; P <2E-16 и ρ = -0,78; P <2E-16 соответственно). Указанные значения P были скорректированы по Бонферрони для всех 78 парных сравнений между 13 выбранными функциями). Корреляция между этими признаками показала, что основанные на пикселях дескрипторы описывают сопоставимые образцы фенотипического изменения.

Fig. 7A shows the directions of the feature variance-covariance matrix with the traits labeled as in Fig. 6. Fig. 7B shows the correlation matrix between the 13 selected features. For the 5 features selected by OOB error (Fig. 6), indicated with a "5" in Table 1, the estimated H^2 was ≥0.58. Because the majority of selected features are PCs of different pixel-based analyses (Fig. S5) [66], there were many weak correlations (Fig. 7B). We hypothesize that the importance of these features is partly driven by the similarity of the raw data (i.e., binary pixel intensities) used in k-means clustering to acquire shape categories and for EigenFruit shape analysis. Although PCs are uncorrelated, we observed strong correlations between PCs from different analyses (Fig. 7). $EigenFruitPC_1$ shared a strong positive correlation with both $BioVPC_1$ and $BioHPC_1$ (ρ = 0.98; P < 2E−16 and ρ = 0.93; P < 2E−16, respectively), as did $EigenFruitPC_2$ with $BioVPC_2$ (ρ = 0.86; P < 2E−16). $BioHPC_2$ was negatively correlated with both $EigenFruitPC_2$ and $BioVPC_3$ (ρ = −0.92;P < 2E−16 and ρ = −0.81;P < 2E−16, respectively). $BioHPC_3$ was negatively correlated with $EigenFruitPC_4$ (ρ = −0.87; P < 2E−16). BAR was negatively correlated with $EigenFruitPC_1$, $BioVPC_1$, and $BioHPC_1$ (ρ = −0.89;P < 2E−16, ρ = −0.87; P< 2E−16, and ρ = −0.78; P < 2E−16, respectively). Reported P-values were Bonferroni adjusted for all 78 pairwise comparisons between the 13 selected features). The correlations between these features indicated that the pixel-based descriptors describe comparable patterns of phenotypic variation.

Классификация изображений с использованием выбранных функций

Точность классификации или прогнозирования, как правило, оценивается перекрестной проверкой [24, 76]. Мы создали обучающие наборы, которые состояли из 80% (5500), 50% (3437) или 20% (1374) изображений. Назначение на тренировочный или тестовый набор было случайным и без стратификации. Возможно, что стратификация будет необходима для большего количества итераций, > 10, меньших размеров выборки или очень неравных изображений для каждой категории. Кластеризация k-средних была выполнена с использованием обучающих наборов, и k было разрешено варьировать от 2 до 10. Мы присвоили изображения тестового набора ближайшему соседнему кластеру для каждого уровня k. Мы выполнили PPKC для кластеров, полученных из обучающего набора, и визуально оценили сходство между полным набором и обучающими наборами. Кластеры, полученные из разных наборов, оказались практически идентичными (рис. S6) [66]. Порядок кластеров, полученных из сокращенного набора данных, также выглядит идентичным описанным в полном наборе (рис. S6) [66]. Функции на базе ПК были пересчитаны с использованием наборов обучающих данных и соответствующих изображений тестовых наборов, спроецированных в новое пространство. Мы только извлекли 13 выбранных функций. К ним относятся $EigenFruitPC_{[1, 6]}$, $BioVPC_{[1, 2]}$ и $BioHPC_{[1, 2]}$ (Таблица 1). Выбранные геометрические особенности, включая BAR, SI и AR, не были пересчитаны, поскольку они не изменяются в отношении других образцов, в отличие от k-средних и PCA, которые основаны и изменяются на основе данных наблюдений. Для $EigenFruitPC_{[1, 6]}$, $BioVPC_{[1, 2]}$ и $BioHPC_{[1, 2]}$ процентная дисперсия, объясняемая каждой функцией, была аналогична дисперсии в полном наборе данных (Таблица 1), что указывает на то, что PC получены из сокращенного набора, описывает сходные черты формы как те, которые получены из полного набора.

> ## Image classification using selected features
>
> The accuracy of classification, or prediction, is typically assessed by cross-validation [24, 76]. We generated training sets that consisted of 80% (5,500), 50% (3,437), or 20% (1,374) of the images. Assignment to either training or test set was random and without stratification. It is possible that stratification would be needed for more iterations, >10, smaller sample sizes, or very unequal images per k category. k-means clustering was performed using the training sets, and k was allowed to range from 2 to 10. We assigned the test set images to the nearest neighboring cluster for each level of k. We performed PPKC on the clusters derived from the training set, and the similarity between the full set and training sets was visually assessed. The clusters derived from the different sets appeared to be nearly identical (Fig. S6) [66]. The order of clusters derived from the reduced data set also appears identical to those described in the full set (Fig. S6) [66]. The PC-based features were recalculated using the training data sets and the corresponding test set images projected into the new space. We only extracted the 13 selected features. These included $EigenFruitPC_{[1, 6]}$, $BioVPC_{[1, 2]}$, and $BioHPC_{[1, 2]}$ (Table 1). The selected geometric features, including BAR, SI, and AR, were not recalculated because they do not change concerning the other samples, unlike k-means and PCA, which both rely on and change on the basis of observed data. For $EigenFruitPC_{[1, 6]}$, $BioVPC_{[1, 2]}$, and $BioHPC_{[1, 2]}$, the percent variance explained by each feature was similar to that in the full data set (Table 1), indicating that the PCs derived from the reduced set describe similar features of shape as those derived from the full set.

SVR и LDA были использованы для классификации (см. Методы). Мы выполнили 10 итераций каждого размера набора и набора функций на всех уровнях k. Результаты этого эксперимента приведены в Таблице 2. В целом, модели выполнены с высокой точностью классификации. Как правило, поскольку мы использовали меньшее количество функций для классификации, производительность модели была снижена, особенно при больших значениях k. Действительно, при k = 2 точность немного улучшилась с меньшим количеством функций в разных моделях. В целом было установлено, что SVR постоянно превосходят LDA. LDA только превосходил SVR с очень маленькими тренировочными наборами относительно тестового набора (Таблица 2). Используя 5 признаков для классификации, мы достигли максимальной точности (99,5%) для k = 2. В интересующем диапазоне k = [2, 4] модели не упали ниже точности 90,0% для любого размера тренировочного набора.

> SVR and LDA were both used for classification (see Methods). We performed 10 iterations of each set size and feature set across all levels of k. The results of this experiment are reported in Table 2. Overall, the models performed with high accuracy of classification. Generally, as we used fewer features for classification, model performance was reduced, most notably for larger values of k. Indeed, when k = 2 accuracy improved slightly with fewer features in the different models. In general, SVR was found to outperform LDA consistently. LDA only outperformed SVR with very small training sets relative

> to the test set (Table 2). Using 5 features for classification, we achieved the highest accuracy (99.5%) for k = 2. In the range of interest, k = [2, 4], the models did not fall below 90.0% accuracy for any training set size.

Table 2:

Эксперимент по валидации оценок классификационной модели

> Classification model evaluations validation experiment

Set (Train/Test)	*k*	H^2	$Accuracy_{13}$	$Precision_{13}$	$Recall_{13}$	FPR_{13}	$Accuracy_5$	$Precision_5$	$Recall_5$	FPR_5	$Accuracy_3$	$Precision_3$	$Recall_3$	FPR_3
80/20	2	0.98	0.990/0.978	0.990/0.978	0.990/0.978	0.010/0.022	0.995/0.982	0.995/0.983	0.995/0.981	0.005/0.019	0.990/0.983	0.990/0.983	0.990/0.985	0.010/0.015
	3	0.87	0.985/0.963	0.985/0.962	0.982/0.957	0.009/0.019	0.990/0.971	0.990/0.973	0.989/0.969	0.003/0.014	0.941/0.910	0.938/0.906	0.926/0.904	0.030/0.046
	4	0.85	0.982/0.949	0.982/0.953	0.981/0.943	0.008/0.020	0.982/0.950	0.983/0.952	0.981/0.951	0.008/0.018	0.946/0.921	0.950/0.935	0.934/0.896	0.019/0.029
	5	0.81	0.973/0.942	0.979/0.949	0.975/0.941	0.009/0.013	0.976/0.955	0.977/0.962	0.980/0.954	0.008/0.010	0.932/0.893	0.939/0.917	0.928/0.879	0.020/0.030
	6	0.83	0.973/0.943	0.976/0.947	0.973/0.940	0.008/0.010	0.965/0.926	0.966/0.934	0.965/0.919	0.010/0.015	0.898/0.852	0.903/0.876	0.889/0.835	0.020/0.031
	7	0.83	0.966/0.941	0.968/0.947	0.966/0.940	0.006/0.010	0.951/0.910	0.952/0.922	0.950/0.904	0.010/0.016	0.870/0.824	0.880/0.857	0.866/0.815	0.021/0.030
	8	0.82	0.963/0.928	0.964/0.934	0.962/0.926	0.009/0.010	0.866/0.825	0.856/0.828	0.858/0.812	0.018/0.027	0.790/0.748	0.790/0.765	0.778/0.731	0.028/0.038
	9	0.80	0.954/0.920	0.956/0.926	0.954/0.917	0.009/0.010	0.828/0.789	0.825/0.801	0.827/0.781	0.021/0.030	0.745/0.715	0.751/0.736	0.741/0.707	0.030/0.038
	10	0.81	0.951/0.909	0.952/0.915	0.951/0.906	0.008/0.010	0.798/0.752	0.798/0.770	0.802/0.752	0.024/0.026	0.708/0.679	0.718/0.704	0.706/0.676	0.034/0.036
50/50	2		0.990/0.978	0.990/0.978	0.990/0.979	0.010/0.021	0.993/0.983	0.992/0.983	0.993/0.983	0.007/0.017	0.990/0.990	0.990/0.990	0.990/0.990	0.010/0.010
	3		0.981/0.961	0.981/0.963	0.980/0.958	0.010/0.022	0.988/0.972	0.989/0.974	0.987/0.971	0.006/0.016	0.943/0.907	0.940/0.902	0.934/0.909	0.030/0.047
	4		0.979/0.951	0.980/0.953	0.979/0.944	0.010/0.019	0.981/0.952	0.981/0.955	0.980/0.954	0.010/0.018	0.943/0.920	0.947/0.933	0.927/0.896	0.020/0.030
	5		0.969/0.941	0.972/0.945	0.969/0.938	0.010/0.014	0.969/0.948	0.972/0.955	0.969/0.945	0.010/0.012	0.922/0.885	0.931/0.912	0.916/0.869	0.020/0.032
	6		0.966/0.941	0.967/0.945	0.966/0.939	0.010/0.010	0.961/0.928	0.961/0.935	0.960/0.917	0.010/0.014	0.887/0.856	0.896/0.882	0.879/0.835	0.022/0.030
	7		0.961/0.934	0.961/0.939	0.960/0.931	0.010/0.010	0.933/0.897	0.932/0.906	0.931/0.887	0.011/0.017	0.851/0.818	0.861/0.848	0.845/0.805	0.025/0.031
	8		0.955/0.928	0.957/0.931	0.955/0.923	0.010/0.010	0.872/0.831	0.861/0.832	0.861/0.808	0.017/0.027	0.794/0.759	0.790/0.772	0.776/0.738	0.028/0.037
	9		0.950/0.918	0.950/0.923	0.949/0.910	0.010/0.010	0.836/0.793	0.830/0.799	0.829/0.779	0.021/0.029	0.746/0.718	0.747/0.731	0.731/0.704	0.030/0.038
	10		0.947/0.909	0.949/0.915	0.947/0.904	0.010/0.010	0.802/0.762	0.798/0.774	0.804/0.755	0.022/0.027	0.707/0.693	0.716/0.713	0.705/0.687	0.031/0.034
20/80	2		0.987/0.977	0.987/0.977	0.986/0.977	0.014/0.023	0.990/0.983	0.990/0.983	0.990/0.983	0.010/0.017	0.990/0.986	0.990/0.986	0.990/0.986	0.010/0.014
	3		0.973/0.955	0.975/0.958	0.973/0.950	0.013/0.024	0.982/0.967	0.982/0.966	0.981/0.967	0.010/0.018	0.942/0.906	0.943/0.907	0.939/0.911	0.028/0.047
	4		0.967/0.944	0.971/0.951	0.964/0.938	0.010/0.020	0.973/0.953	0.977/0.955	0.971/0.953	0.010/0.017	0.940/0.921	0.949/0.939	0.921/0.892	0.020/0.030
	5		0.959/0.941	0.963/0.946	0.954/0.936	0.010/0.017	0.953/0.931	0.954/0.939	0.948/0.923	0.012/0.016	0.899/0.875	0.912/0.899	0.883/0.849	0.026/0.033
	6		0.953/0.935	0.958/0.940	0.951/0.935	0.010/0.012	0.937/0.909	0.938/0.917	0.935/0.899	0.012/0.020	0.851/0.835	0.864/0.861	0.837/0.819	0.030/0.034
	7		0.945/0.928	0.950/0.933	0.944/0.925	0.010/0.010	0.902/0.876	0.901/0.885	0.901/0.866	0.016/0.021	0.812/0.804	0.827/0.829	0.799/0.789	0.032/0.034
	8		0.937/0.913	0.938/0.920	0.933/0.914	0.010/0.011	0.829/0.804	0.825/0.810	0.822/0.792	0.023/0.030	0.736/0.733	0.755/0.753	0.722/0.720	0.039/0.040
	9		0.930/0.908	0.933/0.915	0.927/0.903	0.010/0.010	0.808/0.780	0.802/0.788	0.798/0.767	0.024/0.028	0.706/0.707	0.724/0.727	0.688/0.692	0.038/0.038
	10		0.927/0.901	0.930/0.905	0.926/0.896	0.010/0.010	0.794/0.758	0.796/0.781	0.796/0.760	0.023/0.028	0.677/0.681	0.701/0.706	0.670/0.676	0.038/0.036

(2)

Набор относится к 80/20, 50/50 или 20/80 обучающему набору / разделению тестового набора. k - номер категории. H^2 - наследственность в широком смысле и была оценена с использованием полного набора данных. Метрика SVR / LDA представлена для точности, точности, отзыва и ложноположительного показателя (FPR). Подписи 13, 5 и 3 указывают количество выбранных признаков в соответствии с моделью классификации.

> Set refers to the 80/20, 50/50, or 20/80 training set/test set split. k is the number of categorie. H^2 is the broad-sense heritability and was estimated using the full data set. SVR metric/LDA metric is presented for accuracy, precision, recall, and false-positive rate (FPR). Subscripts 13, 5, and 3 indicate the number of selected features in the classification model fit.

Обсуждение

Поскольку высокопроизводительное фенотипирование для внешних характеристик плодов становится интересным для исследователей специальных культур, мы ожидаем, что эта работа будет иметь различные применения как в прикладных, так и в фундаментальных исследованиях растений [12, 13, 51, 64, 65], защите интеллектуальной собственности и документации [77, 78], и сокращение отходов [20, 79]. Наше исследование показало, что формы плодов

клубники можно надежно определить количественно и точно классифицировать по цифровым изображениям. Самое главное, что в результате нашего анализа были получены количественные фенотипические переменные, которые описывают форму плода (рис. 4), являются результатом непрерывного распределения и являются наследственными от умеренного до очень высокого (таблица 1). Мы достигли этого путем перевода двумерных цифровых изображений фруктов в категориальные и непрерывные фенотипические переменные с использованием неконтролируемого машинного обучения и морфометрии. Мы обнаружили, что математические подходы, разработанные для распознавания человеческого лица [58, 59], были эффективными для фенотипирования формы плодов клубники (Таблица 1), что кластеризация форм без присмотра была устойчивой к отклонениям размера образца (Рис. S6) [66], и что только Для точной классификации форм по изображениям требуется несколько количественных характеристик (таблица 2), что указывает на подходящую для генетического анализа парадигму.

Discussion

As high-throughput phenotyping for external fruit characteristics becomes of interest to specialty crop researchers, we expect that this work will have various applications in both applied and basic plant research [12, 13, 51, 64, 65], intellectual property protection and documentation [77, 78], and waste reduction [20, 79]. Our study showed that strawberry fruit shapes could be robustly quantified and accurately classified from digital images. Most importantly, our analyses yielded quantitative phenotypic variables that describe fruit shape (Fig. 4), arise from continuous distributions, and are moderately to highly heritable (Table 1). We accomplished this by translating 2D, digital images of fruit into categorical and continuous phenotypic variables using unsupervised machine learning and morphometrics. We found that mathematical approaches developed for human face recognition [58, 59] were powerful for strawberry fruit shape phenotyping (Table 1), that unsupervised shape clustering was robust to sample size deviations (Fig. S6) [66], and that only a few quantitative features are needed to accurately classify shapes from images (Table 2), indicating a paradigm appropriate for genetic dissection.

Цифровое фенотипирование растений способно расширить возможности количественного генетического анализа, предоставляя наследственные и биологически релевантные скрытые фенотипы экономически эффективным образом [13, 64, 65, 80, 81]. Во многих случаях эти скрытые признаки получены из PCA, многомерного масштабирования (MDS), моделирования структурированных уравнений (SEM), устойчивой гомологии (PH) или автоматического кодирования сверточных нейронных сетей, которые могут быть чрезвычайно абстрактными и трудными для интерпретации. биологически, но может также выявить неожиданные паттерны фенотипической и генетической изменчивости [12, 13, 19, 24, 51, 59, 61, 75, 82–84]. Многие из особенностей, описанных в этом исследовании, наряду с описанными Turner et al. [13] (т.е. профиль биомассы) [12] (т.е. эллиптические Фурье-PC и PC с постоянной гомологией) и Gage et al. [65] (т.е. PC с изображениями и сверточные кодировки), обладали высокой наследуемостью (таблица 1) и являются интересными целями для будущих количественных генетических анализов, включая GWAS и геномное прогнозирование, которые, как было показано, оказались успешными для особенностей формы в недавней работе рис (Oryza sativa L.) [85], яблоко (Malus domestica) [86] и груша (Pyrus spp.) [87]. Однако выбранная особенность H^2 из 1 в этом исследовании, $EigenFruitPC_3$, была оценена как 0,00 (таблица 1 и рис. S4) [66]. Аналогичные результаты были получены у моркови (Daucus carota L.) для основанных на пикселях корней и побегов [13], яблока (Malus domestica) для эллиптических элементов формы листьев Фурье [12] и кукурузы (Zea mays) для побегов на основе пикселей особенности [65]. Turner et al. [13] приписали нулевой H^2 характеристик формы корня низкому фенотипическому изменению между инбредными родителями и взаимодействиями генотип-среда. Эта картина, хотя и присутствует, по-видимому, подробно не обсуждалась ни Migicovsy et al. [12] или Gage et al. [65]. Хотя для этого шаблона может быть много факторов, мы предполагаем, что нулевая оценка может возникать на основе пиксельных дескрипторов, описывающих более сложные аспекты формы плода или корня. Если негенетический компонент многомерного фенотипа велик, то выполнение PCA по этому многомерному признаку может привести к появлению ведущих PC, которые описывают в основном негенетическую дисперсию (например, окружающая среда, управление и остаточный). Тем не менее, имеется слишком мало сообщений для адекватного определения вероятности и причинного источника этого явления.

Digital plant phenotyping is able to empower quantitative genetic analyses by providing heritable and biologically relevant, latent phenotypes in a cost-effective manner [13, 64, 65, 80, 81]. In many cases, these latent traits are derived from PCA, multi-dimensional scaling (MDS), structured equation modeling (SEM), persistent homology (PH), or auto-encoding convolutional neural networks, which can be exceedingly abstract and difficult to interpret biologically but may also reveal unexpected patterns of phenotypic and genetic variation [12, 13, 19, 24, 51, 59, 61, 75, 82–84]. Many of the features described in this study, along with those reported by Turner et al. [13] (i.e., biomass profile) [12] (i.e., elliptical Fourier PCs and persistent homology PCs) and Gage et al. [65] (i.e., image PCs and convolutional encodings), had high heritability

> (Table 1) and are exciting targets for future quantitative genetic analyses, including GWAS and genomic prediction, which have been shown to be successful for shape features in recent work in rice (Oryza sativa L.) [85], apple (Malus domestica) [86], and pear (Pyrus spp.) [87]. However, the H^2 of 1 selected feature in this study, $EigenFruitPC_3$, was estimated to be 0.00 (Table 1 and Fig. S4) [66]. Similar results were reported in carrot (Daucus carota L.) for pixel-based root and shoot features [13], apple (Malus domestica) for elliptical Fourier leaf shape features [12], and corn (Zea mays) for pixel-based shoot features [65]. Turner et al. [13] attributed the null H^2 of root shape characteristics to low phenotypic variation between the inbred parents and genotype × environment interactions. This pattern, while seemingly present, was not discussed in detail by either Migicovsy et al. [12] or Gage et al. [65]. While there may be many drivers for this pattern, we hypothesize that the null estimate may arise from the pixel-based descriptors describing more complex aspects of fruit or root shape. If the non-genetic component of a multivariate phenotype is large, then performing PCA on that multivariate trait could produce leading PCs that describe mostly non-genetic variance (e.g., environment, management, and residual). However, there are too few reports to adequately determine the likelihood and causal source of this phenomenon.

Мы эмпирически получили прогрессию формы, полученную в настоящем исследовании, с помощью применения нового метода PPKC и использовали эти математические категории для интерпретации извлеченных элементов формы (алгоритм 1 и рис. 3). Порядковые категориальные черты являются обычным явлением в количественных генетических исследованиях [29, 71], нынешнем стандарте фенотипирования внешних характеристик плодов [14, 15, 42], и позволяют понять и объяснить сложные скрытые фенотипы формы растений (Рис. 6 и 7). PPKC, в частности, рассматривает взаимосвязь между кластером в k и всеми кластерами для значений 2 (таблица 2) и информационных критериев, рассчитанных для моделей k-средних (рис. . S2) [66]. Интересно, что PPKC может определять визуально приемлемый фенотипический градиент вплоть до k = 8 (рис. S3) [66], несмотря на убедительные доказательства переоснащения для k> 4 (рис. S2) [66]. Мы экстраполируем, что PPKC должен продолжать работать после k = 9, пока новые кластеры различны и не возникают как артефакт переоснащения k.

> We empirically derived the shape progression produced in the present study through the application of a new method, PPKC, and used these mathematical categories to interpret the extracted shape features (Algorithm 1 and Fig. 3). Ordinal categorical traits are commonplace in quantitative genetic studies [29, 71], a current standard for phenotyping external fruit characteristics [14, 15, 42], and enable understanding and explanation of complex, latent space plant phenotypes (Figs 6 and 7). PPKC specifically considers the relationship between a cluster at k and all clusters for values 2 estimates (Table 2), and the information criteria calculated for the k-means models (Fig. S2) [66]. Interestingly, PPKC can determine a visually, reasonable phenotypic gradient up to k = 8 (Fig. S3) [66] despite strong evidence of overfitting for k > 4 (Fig. S2) [66]. We extrapolate that PPKC should continue to work beyond k = 9 so long as new clusters are distinct and do not arise as an artifact of overfitting k.

Конкретные генетические факторы, которые приводят к изменению формы плодов у октоплоидов, земляники садовой, в настоящее время неясны или недостаточно изучены. Избирательное давление, оказываемое на форму плодов в клубнике, могло повлиять на локусы с большим эффектом, и в этом случае порядковые фенотипические оценки, вероятно, будут достаточными для выявления генетических факторов, влияющих на форму плодов. Мутации потери и усиления функции сыграли существенную роль в определении генов, влияющих на форму плодов у томатов, модель, которая была очень поучительной и важной для понимания генетики формы и увеличения плодов у растений [34–36, 89, 90]. Существуют яркие примеры у томатов и других растений, где идентифицированные гены регулируют развитие формы плода. Например, ген OVATE у томатов регулирует фенотипический переход от круглых к грушевидным плодам [91, 92]. Если мутации с большим эффектом лежат в основе различий в форме плодов клубники, предложенная здесь система порядковой классификации должна позволять обнаруживать такие эффекты. Кроме того, количественные фенотипы были связаны с генетическими особенностями, которые взаимодействуют с генами с большим эффектом, то есть с супрессорами OVATE (sov), посредством анализа объемных сегрегаций и картирования локусов количественных признаков [93]. В лесной землянике (Fragaria vesca) размер и форма плодов связаны с накоплением и сложным взаимодействием ауксина, гиббереллловой кислоты и абсцизовой кислоты, опосредованных экспрессией и активностью FveCYP707 и FveNCED, а также других генов [9]. Из-за высоких оценок H^2 для нескольких вновь созданных фенотипических переменных (Таблица 1), мы выдвигаем гипотезу, что количественные скрытые фенотипы пространства могут дать более полное понимание основных генетических механизмов формы плодов у земляники садовой через GWAS и другие количественные генетические анализы [44, 45, 94]. Мы ожидаем, что анализ этого исследования позволит нам обнаружить и изучить генетические детерминанты формы плодов клубники и других специальных культур.

> The specific genetic factors that give rise to variation in fruit shape in octoploid, garden strawberry are currently unclear or understudied. The selective pressure exerted on fruit shape in strawberry could have affected large-effect loci, in which case ordinal phenotypic scores are likely to be sufficient for identifying genetic factors affecting fruit shape. Loss- and

gain-of-function mutations have played an essential role in identifying genes affecting fruit shape in tomato, a model that has been highly instructive and important for understanding the genetics of fruit shape and enlargement in plants [34–36, 89, 90]. There are striking examples in tomato and other plants where identified genes regulate the development of fruit shape. For example, the OVATE gene in tomato regulates the phenotypic transition from round to pear-shaped fruit [91, 92]. If large-effect mutations underlie differences in strawberry fruit shape, the ordinal classification system proposed here should enable the discovery of such effects. Furthermore, quantitative phenotypes were linked to genetic features that interact with large-effect genes, i.e., suppressors of OVATE (sov), through bulk segregant analysis and quantitative trait locus mapping [93]. In woodland strawberry (Fragaria vesca), fruit size and shape are linked to the accumulation and complex interaction of auxin, gibberellic acid, and abscisic acid, mediated by the expression and activity of FveCYP707 and FveNCED, as well as other genes [9]. Because of the high H^2 estimates for several of the newly created phenotypic variables (Table 1), we hypothesize that quantitative, latent space phenotypes can yield a more comprehensive understanding of the underlying genetic mechanisms of fruit shape in garden strawberry through GWAS and other quantitative genetic analyses [44, 45, 94]. We anticipate that the analyses in this study will enable us to discover and study the genetic determinants of fruit shape in strawberry and other specialty crops.

Методы

Спаривание и дизайн поля

Семьдесят пять двойных родительских скрещиваний были получены путем контролируемого опыления 30 родителей в неполной (14 × 16) схеме факторного спаривания. Эти родители были выбраны, чтобы представлять широкий спектр фенотипического разнообразия в Университете Калифорнии, Дэвис, гермоплазмы клубники. В общей сложности 2800 гибридных потомков были посажены в экспериментальном саду Wolfskill в Уинтерсе, штат Калифорния, в наборах по 20 или 40 на семью, в зависимости от выживаемости рассады. Двадцать процентов посаженных материалов из каждого семейства были случайным образом отобраны для дальнейшего тестирования. Клоны 545 отобранного 560 потомства были успешно размножены. В ноябре 2017 года в Салинасе, штат Калифорния, было собрано и высажено 12 растений с голыми корнями для каждого из 545 потомства и 30 родителей на четырех участках растений в виде рандомизированной полной блочной конструкции с 3 повторностями каждого генотипа.

Methods

Mating and field design

Seventy-five bi-parental crosses were generated by controlled pollination of 30 parents in an incomplete (14 × 16) factorial mating design. These parents were chosen to represent a broad range of phenotypic diversity in the University of California, Davis, strawberry germplasm. A total of 2,800 hybrid progeny were planted at the Wolfskill Experimental Orchard in Winters, CA, in sets of 20 or 40 per family, depending on seedling survival. Twenty percent of the planted materials from each family were randomly selected for further testing. Clones of 545 of the selected 560 progeny were successfully propagated. Twelve bare-root runner plants of each of the 545 progeny and the 30 parents were collected and planted in November 2017 in Salinas, CA, in 4 plant plots as a randomized complete block design with 3 replicates of each genotype.

Получение изображения

Клубнику собирали со участков в Салинасе, штат Калифорния, один раз в апреле 2018 года и снова в мае 2018 года. Цифровые изображения до 3 плодов на участок получали с помощью цифровой зеркальной камеры Sony α-6000, установленной на переносной копировальной стойке с приоритетом диафрагмы, с диафрагмой, установленной на f / 8. Клубника со снятой чашечкой была помещена в раму на черном войлочном фоне вместе с QR-кодом, определяющим график, так что самое обширное лицо было перпендикулярно датчику. Ягоды были прикреплены к ряду скрепок, чтобы исключить скручивание или смолу ягод. Время для постановки данного набора фруктов и получения изображения варьировалось от 1 до 2 мин. Все изображения были получены с объективом 16-50 мм, установленным на 16 мм и расположенным на расстоянии ~16 см над основанием подставки для копирования, в результате чего были получены изображения с разрешением 97,4 пикселя на см. Всего за 2 даты сбора урожая было получено 2924 участка.

> ## Image acquisition
>
> Strawberries were harvested from plots in Salinas, CA, once in April 2018 and again in May 2018. Digital images of up to 3 fruit per plot were imaged using a Sony α-6000 Mirrorless digital camera mounted on a portable copy stand in aperture priority, with the aperture set to f/8. Strawberries with the calyx removed were placed in the frame against a black felt backdrop, along with a QR code identifying the plot, such that the most extensive face was perpendicular to the sensor. Berries were mounted to a set of staples to eliminate any rolling or pitch of the berries. The time to stage a given set of fruit and acquire an image ranged from 1 to 2 min. All images were acquired with a 16−50 mm lens set to 16 mm and positioned ~16 cm above the base of the copy stand, resulting in images with 97.4 pixels per cm. In total, 2,924 plots were imaged over the 2 harvest dates.

Обработка изображений

Входные файлы представляли собой изображения в формате JPEG (3 008 × 1688 пикселей) с клубникой, расположенной в обычном положении в пределах сцены. Все изображения были сначала сегментированы и преобразованы в двоичный файл с использованием инструмента Simple Interactive Object Extraction (SIOX) в ImageJ 2.0.0 [95–97] с помощью пользовательских пакетных сценариев. Изображения, которые были неудачно сегментированы, помечались и обрабатывались индивидуально для обеспечения полноты. ImageJ был использован для получения ограничивающего прямоугольника каждого интересующего объекта. Каждый объект был извлечен на основе размеров его ограничительного прямоугольника с использованием R 3.5.3 [98] и пакета jpeg [99]. Белые пиксели были добавлены к краям каждого изображения таким образом, чтобы полученное изображение имело квадрат размера max (H, W) × max (H, W) с помощью пакета «magick :: image_border ()» [100]. «Magick :: image_resize ()» использовался для масштабирования изображений от макс. (H, W) × макс. (H, W) пикселей до 1000 × 1000 пикселей. Этот метод приводит к получению двоичных изображений, которые поддерживают исходное соотношение сторон с максимальным размером, равным 1000 пикселей, а затем изменяют размер до 100 × 100 (рис. 1). В целом, последующий анализ включал 6874 изображения отдельных ягод.

> ## Image processing
>
> Input files were JPEG images (3,008 × 1,688 pixels) with the strawberries placed in regular positions within a scene. All images were first segmented and converted to binary using the Simple Interactive Object Extraction (SIOX) tool in ImageJ 2.0.0 [95–97] through custom batch scripts. Images that were unsuccessfully segmented were flagged and handled individually to ensure completeness. ImageJ was used to acquire the bounding rectangle of each object of interest. Each object was extracted on the basis of the dimensions of its bounding rectangle using R 3.5.3 [98] and the jpeg package [99]. White pixels were added to the edges of each image such that the resulting image was a square of size max (H, W) × max (H, W) using the ”magick::image_border()” package [100]. ”magick::image_resize()” was used to scale the images from max (H, W) × max (H, W) pixels to 1,000 × 1,000 pixels. This method results in binary images that maintain the original aspect ratio with a maximum dimension equal to 1,000 pixels and then resized to 100 × 100 (Fig. 1). In total, the downstream analyses included 6,874 images of individual berries.

Извлечение функций

Категориальные особенности

Этот метод позволял кластеризовать решения на основе необработанных данных изображений вместо извлеченных количественных характеристик. Каждая матрица изображения была сведена в один вектор строки из 10 000 элементов; Затем все образцы были связаны столбцами. Результирующая матрица для всех образцов составляла 6874 × 10000. Функция ”stats :: kmeans ()” в R использовалась для выполнения кластеризации k-средних. Значения k (то есть количество кластеров) варьировались от 2 до 10. Назначенные кластеры были записаны для всех значений k. Обнаруженные кластеры были затем упорядочены с использованием PPKC (рис. 3). Упорядоченные категории на разных уровнях k стали ответом на классификационные эксперименты. Правильный выбор k часто неоднозначен, с интерпретациями, зависящими от формы и масштаба распределения точек в наборе данных и желаемого разрешения кластеризации пользователя. Кроме того, увеличение k без штрафа всегда будет уменьшать количество ошибок в результирующей кластеризации до крайнего случая нулевой ошибки, если каждая точка данных рассматривается как свой собственный кластер (то есть, когда k равно количеству точек данных, n). Интуитивно понятно, что оптимальный выбор k будет обеспечивать баланс между максимальным сжатием данных с использованием одного кластера и

максимальной точностью, назначая каждую точку данных своему кластеру. Оптимальное значение k было определено на основе 4 различных критериев оценки: общая сумма квадратов внутри кластера, скорректированная R^2, AIC и BIC.

Feature extraction

Categorical features

This method afforded clustering decisions based on raw image data instead of the extracted quantitative features. Each image matrix was flattened into a single 10,000 element row vector; all of the samples were then bound together by columns. The resulting matrix for all samples was 6,874 × 10,000. The ”stats::kmeans()” function in R was used to perform k-means clustering. Values of k (i.e., the number of clusters) ranged from 2 to 10. Assigned clusters were recorded for all values of k. Discovered clusters were then ordered using PPKC (Fig. 3). The ordered categories, across the various levels of k, became the response for classification experiments. The correct choice of k is often ambiguous, with interpretations depending on the shape and scale of the distribution of points in a data set and the desired clustering resolution of the user. In addition, increasing k without penalty will always reduce the amount of error in the resulting clustering, to the extreme case of zero error if each data point is considered its own cluster (i.e., when k equals the number of data points, n). Intuitively then, the optimal choice of k will strike a balance between maximum compression of the data using a single cluster, and maximum accuracy by assigning each data point to its own cluster. The optimal value of k was determined on the basis of 4 different evaluation criteria: total within-cluster sum of squares, adjusted R^2, AIC, and BIC.

Линейные и геометрические особенности

Линейные и геометрические особенности измеряют аспекты фруктов непосредственно из изображений и были обработаны с использованием ImageJ 2.0.0 [96, 97] и R 3.5.3 [98]. Извлеченные измерения включали индекс формы (SI) [40], округлость (Circ) [97], ограничивающее соотношение сторон (BAR) [97], соотношение сторон эллипса (AR) [97], округлость (Round) [97], твердость (Сплошной) [97], соотношение сторон по Фере (FAR) [97], соотношение высоты максимальной ширины и максимальной высоты (HW) [40], дисперсии (Var), асимметрии (Skew) [97] и эксцесса (Курт) [97] (рис. 4A). Для Var, Skew и Kurt анализ фокусируется на горизонтальной оси (Рис. 4A).

Linear and geometric features

Linear and geometric features measure aspects of the fruit directly from images and were processed using ImageJ 2.0.0 [96, 97] and R 3.5.3 [98]. Extracted measurements included shape index (SI) [40], circularity (Circ) [97], bounding aspect ratio (BAR) [97], ellipse aspect ratio (AR) [97], roundness (Round) [97], solidity (Solid) [97], Feret aspect ratio (FAR) [97], the ratio of the height of maximum width and maximum height (HW) [40], variance (Var), skewness (Skew) [97], and kurtosis (Kurt) [97] (Fig. 4A). For Var, Skew, and Kurt, the analyses focus on the horizontal axis (Fig. 4A).

Эллиптический анализ Фурье

EFA всесторонне описал замкнутые контуры как ряд осциллирующих гармонических функций и был рассчитан с использованием Momocs v1.2.9 [101] в R 3.5.3. Мы извлекли эллиптические особенности Фурье для первых 5 гармоник, в результате чего получили 20 коэффициентов, используя функцию «Momocs :: efourier ()». Каждый уровень гармоник состоит из 4 коэффициентов, которые соответствуют эффектам косинуса и синуса по оси x (коэффициенты A и B) и по оси y (коэффициенты C и D). Чтобы учесть различия между образцами, основанными на форме плода, анализ главных компонентов (PCA) был выполнен с использованием «Momocs :: PCA» от Momocs для EFF. Мы записали собственные векторы каждого изображения на 20 результирующих главных осях (рис. 4B).

Elliptical Fourier analysis

EFA comprehensively described closed outlines as a series of oscillating, harmonic functions and were calculated using Momocs v1.2.9 [101] in R 3.5.3. We extracted elliptical Fourier features for the first 5 harmonics, resulting in 20 coefficients using ”Momocs::efourier()” function. Each harmonic level is made up of 4 coefficients that correspond to the effects of the cosine and sine in the x-axis (coefficients A and B) and the y-axis (coefficients C and D). To allow for discrimination between accessions based on fruit shape, principal component analysis (PCA) was performed using the

> ”Momocs::PCA” from Momocs for EFFs. We recorded the eigenvectors of each image on the 20 resulting principal axes (Fig. 4B).

Обобщенный анализ прокрустов и выявленные скрытые признаки

GPA описывает форму как среднее расстояние между всеми измеренными ориентирами на целевом объекте и соответствующими ориентирами на эталонном объекте или центроиде. Контур каждого объекта был разложен на 50 равномерно расположенных псевдо-ориентиров, движущихся по часовой стрелке вокруг объекта. Функция «Momocs :: fgProcrustes ()» из Momocs v1.2.9 [101] использовалась для выравнивания фигур (рис. 4C, слева). Каждый из 50 выровненных псевдо-ориентиров рассматривался как отдельная многомерная особенность. Каждый из 50 объектов был отцентрирован таким образом, что среднее значение по обеим осям равно 0. Функция «stats :: prcomp ()» в R использовалась для выполнения PCA для каждого из 50 центрированных псевдо-ориентиров (рис. 4C, левое пятно). центр).

> ## Generalized Procrustes analysis and revealed latent features
>
> GPA describes shape as the average distance between all measured landmarks on a target object and the corresponding landmarks on a reference object or centroid. The outline of each object was decomposed into 50 evenly spaced pseudo-landmarks moving clockwise around the object. The ”Momocs::fgProcrustes()” function from Momocs v1.2.9 [101] was used to perform the alignment between shapes (Fig. 4C,left). Each of the 50 aligned pseudo-landmarks was considered as an individual multivariate feature. Each of the 50 features was centered such that the marginal mean of both axes is 0. The ”stats::prcomp()” function in R was used to perform PCA on each of the 50 centered pseudo-landmarks (Fig. 4C, leftand center).

Скрытые элементы из вычисленных PC ориентиров были сконструированы для описания 4 наиболее вариабельных областей контура клубники (т.е. Кончика, левой стороны, шеи и правой стороны) (Рис. 4C; в центре) с помощью «lavaan :: sem ()» используя пакет lavaan v0.6-5 [102]. Использование SEM обычно оправдано в социальных науках из-за его способности вменять отношения (то есть, ковариацию) между ненаблюдаемыми конструкциями (скрытыми переменными) из наблюдаемых переменных. Здесь мы рассмотрели различные псевдо-ориентиры как наблюдаемые переменные для изучения взаимосвязи между скрытыми компонентами формы. Только те псевдо-ориентиры с дисперсией на PC_1 больше, чем медиана были использованы для проявления 4 скрытых признаков (рис. 4C, в центре и справа). Затем для извлечения 5 скрытых переменных использовалась функция lavaan ::realt (). Tip, $Side_{Left}$, $Side_{Right}$, Neck и, наконец, Shape. Tip был проявлен комбинацией PC_1 псевдо-ориентиров 1, 2, 3, 4, 5, 48, 49 и 50; шея от ПК1 из ориентиров 24, 25, 26, 27, 28, a nd 29; $Side_{Left}$ от PC_1 из ориентиров 11, 12, 13, 14 и 15; и $Side_{Right}$ от PC_1 ориентиров 38, 39, 40, 41, 42 и 43. Форма тогда проявляется комбинацией Tip, Neck, $Side_{Left}$ и $Side_{Right}$. Дисперсии 5 скрытых переменных были установлены на 1 для идентификации модели. Подход модели был адекватным стандартизированным среднеквадратичным остатком = 0,095, среднеквадратичная ошибка аппроксимации = 0,071 ± 0,002, сравнительный индекс соответствия = 0,979) и индекс Такера-Льюиса = 0,977 [103]. Однако статистика теста χ2 была большой ($\mathbf{\chi^2_{df=271}}$ = **9,724.76.7**; P <2E-16), что, вероятно, связано с большим размером выборки. Мы не проводили сравнение моделей, потому что наша цель состояла в том, чтобы количественно оценить уменьшенное скрытое представление наблюдаемых псевдо-ориентиров, которое минимизирует разницу между подразумеваемой моделью и выборочной ковариационной матрицей. Каждая из 4 скрытых функций была рассчитана для всех изображений.

> Latent features from the calculated landmark PCs were constructed to describe the 4 most variable regions of the strawberry outline (i.e., tip, left side, neck, and right side) (Fig. 4C; center) with "lavaan::sem()” using the lavaan package v0.6−5 [102]. Use of SEM is commonly justified in the social sciences because of its ability to impute relationships (i.e., covariance) between unobserved constructs (latent variables) from observable variables. Here, we treated different pseudo-landmarks as observable variables to study the relationship between latent components of shape. Only those pseudo-landmarks with variance on PC_1 greater than the median were used to manifest the 4 latent features (Fig. 4C, center and right). The "lavaan::predict()" function was then used to extract 5 latent variables: Tip, $Side_{Left}$, $Side_{Right}$, Neck, and finally Shape. Tip was manifest by a combination of PC_1 of the pseudo-landmarks 1, 2, 3, 4, 5, 48, 49, and 50; Neck by PC_1 of landmarks 24, 25, 26, 27, 28, and 29; $Side_{Left}$ by PC_1 of landmarks 11, 12, 13, 14, and 15; and $Side_{Right}$ by PC_1 of landmarks 38, 39, 40, 41, 42, and 43. Shape is then manifest by a combination of Tip, Neck, $Side_{Left}$, and $Side_{Right}$. The variances of the 5 latent variables were set to 1 for model identification. The model fit was adequate standardized root mean squared residual = 0.095, root mean square error of approximation = 0.071 ± 0.002, comparative

> fit index = 0.979), and Tucker-Lewis index = 0.977 [103]. However, the χ2 test statistic was large ($\chi^2_{df=271}$=**9,724.76**;P < 2E−16), which likely resulted from the large sample size. We did not perform model comparisons because our goal was to quantify a reduced, latent-space representation of observed pseudo-landmarks that minimizes the difference between the model-implied and sample covariance matrices. Each of the 4 latent features was calculated for all images.

EigenFruit анализ

Свойства EigenFruit были рассчитаны по EigenFaces и другим связанным с PCA методам [58–61, 65] и включали информацию о каждом пикселе в данном наборе изображений. Результирующая матрица векторов двоичного изображения составляла 6874 × 10000. Может быть только столько ненулевых PC, сколько было наблюдений (то есть, 6874). Функция «stats :: prcomp ()» использовалась для выполнения PCA. Мы записали собственные значения первых 20 ПК. Вместе эти 20 ПК объяснили 71,7% дисперсии. ПК1, ПК2 и ПК3 объяснили 26,8%, 12,6% и 5,24% соответственно (рис. 4D, слева).

> ### EigenFruit analysis
>
> EigenFruit features were calculated from the EigenFaces and other related PCA-based methods of [58–61, 65] and incorporated information about every pixel in a given set of images. The resulting matrix of binary image vectors was 6,874 × 10,000. There could only be as many non-zero PCs as there were observations (i.e., 6,874). The ”stats::prcomp()” function was used to perform PCA. We recorded the eigenvalues of the first 20 PCs. Together these 20 PCs explained 71.7% of the variance. PC1, PC2, and PC3 explained 26.8%, 12.6%, and 5.24%, respectively (Fig. 4D, left).

Особенности профиля биомассы

Характеристики профиля биомассы описывают форму как сумму пикселей в каждой строке или столбце данного изображения. Мы приняли этот метод от Turner et al. [13]. Мы создали горизонтальный профиль биомассы, записав количество черных пикселей в каждой из 100 строк. Вертикальный профиль биомассы был получен путем записи количества черных пикселей в каждом из 100 столбцов. Функция «stats :: prcomp ()» в R использовалась для выполнения PCA для каждого профиля (т.е. вертикального и горизонтального). Собственные векторы первых 5 ПК из каждого были сохранены. Вместе эти 5 ПК объяснили 95,9% и 95,4% общей симметричной дисперсии формы для горизонтального и вертикального профилей соответственно (рис. 4D, в центре и справа).

> ### Biomass profile features
>
> Biomass profile features described the shape as the sum of pixels in each row, or column, of a given image. We adopted this method from Turner et al. [13]. We generated the horizontal biomass profile by recording the number of black pixels in each of 100 rows. The vertical biomass profile was generated by recording the number of black pixels in each of the 100 columns. The function "stats::prcomp()" in R was used to perform PCA for each profile (i.e., vertical and horizontal). The eigenvectors of the first 5 PCs from each were retained. Together these 5 PCs explained 95.9% and 95.4% of the total symmetric shape variance for the horizontal and vertical profiles, respectively (Fig. 4D, center and right).

Оценка наследственности в широком смысле

Качественные особенности

Наследуемость в широком смысле по среднему клону (H^2) для каждого упорядоченного уровня k была оценена с использованием порядкового пакета v2019.3−9 [72] в R 3.5.3. Компоненты дисперсии оценивались с использованием кумулятивных смешанных моделей связей с кумулятивной функцией логит-линка и полиномиальной погрешностью,

$$y_{ijk_l} = \mu + G_i + H_j + B_k + E_{ijk} + F_{ijk_l}$$

(4)

где $\mathbf{y_{ij}k_l}$ - категориальная особенность, μ - великое среднее, G_i - случайный эффект i-го генотипа $\mathbf{[G_i \sim N(0, \sigma^2_G)]}$, H_j - фиксированный эффект j-го урожая, B_k - фиксированный эффект k-й блок, E_{ijk} - это остаточная ошибка графика ijkth $\mathbf{[E_{ijk} \sim N(0, \sigma^2_E)]}$, а F_{ijkl} - ошибка фрукта ijklth (подвыборка) $\mathbf{(F_{ijkl} \sim logit[P(Y \leq j)]\ _0^{k-1}}$, где k - количество кластеров. Функция

«clmm ()» реализует кумулятивные смешанные модели связей для порядковых данных. Порядковые GLMM считались наиболее подходящим и консервативным подходом, потому что мы не могли предположить, что категории формы были бы линейными. Оценка компонента дисперсии выполняется по максимальной вероятности и позволяет для множественных случайных эффектов со скрещенными и вложенными структурами [72]. H^2 для каждого признака рассчитывали как

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2/hr},$$

(5)

где σ^2_G - генетическая дисперсия, σ^2_E - остаточная дисперсия, h - среднее гармоническое из наблюдаемых дат сбора урожая по генотипу (1.66), а r - среднее по гармонике повторностей на урожай (2.50).

Количественные характеристики

Наследуемость в широком смысле на основе среднего клона (H^2) была оценена для признаков с пакетом lme4 v1.1−19 [74] в R 3.5.3. Компоненты максимальной дисперсии максимального правдоподобия были оценены с использованием линейной модели смешанных эффектов,

$$y_{ijk} = \mu + G_i + H_j + B_k + E_{ijk},$$

(6)

где y_{ijk} - количественная характеристика, μ - великое среднее, G_i - случайный эффект i-го генотипа **$[G_i \sim N(0, \sigma^2_G)]$**, H_j - фиксированный эффект j-го урожая, B_k - фиксированный эффект k-й блок, а E_{ijk} - это остаточная ошибка графика ijkth **$[E_{ijk} \sim N(0, \sigma^2_E)]$**. Наблюдались только 2 даты сбора урожая и 3 блока, и поэтому они рассматривались как фиксированные эффекты. H^2 для каждого признака рассчитывали, как в уравнении (5).

Broad-sense heritability estimation

Qualitative features

Broad-sense heritability on a clone-mean basis (H^2) for each ordered level of k was estimated using the ordinal package v2019.3−9 [72] in R 3.5.3. Variance components were estimated using cumulative link mixed models with a cumulative logit link function and a multinomial error,

$$y_{ijk_l} = \mu + G_i + H_j + B_k + E_{ijk} + F_{ijk_l}$$

(4)

where $\mathbf{y_{ij}k_l}$ is the categorical feature, μ is the grand mean, G_i is the random effect of the ith genotype **$[G_i \sim N(0, \sigma^2_G)]$**, H_j is the fixed effect of the jth harvest, B_k is the fixed effect of the kth block, E_{ijk} is the residual error of the ijkth plot **$[E_{ijk} \sim N(0, \sigma^2_E)]$**, and F_{ijkl} is the error of ijklth fruit (subsample) **$(F_{ijkl} \sim logit[P(Y \leq j)]\ _0{}^{k-1}$**, where k is the number of clusters. The "clmm()" function implements cumulative link mixed models for ordinal data. Ordinal GLMMs were considered the most appropriate, and conservative, approach because we could not assume that shape categories would be linear. Variance component estimation is performed via maximum likelihood and allows for multiple random effects with crossed and nested structures [72]. H^2 for each feature was calculated as

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2/hr},$$

(5)

where σ^2_G is the genetic variance, σ^2_E is the residual variance, h is the harmonic mean of observed harvest dates per genotype (1.66), and r is the harmonic mean of replicates per harvest (2.50).

Quantitative Features

Broad-sense heritability on a clone-mean basis (H^2) was estimated for features with the lme4 package v1.1−19 [74] in R 3.5.3. Restricted maximum likelihood variance components were estimated using the linear mixed effects model,

$$y_{ijk} = \mu + G_i + H_j + B_k + E_{ijk},$$

(6)

where y_{ijk} is the quantitative feature, μ is the grand mean, G_i is the random effect of the ith genotype **[$G_i \sim N(0, \sigma^2_G)$]**, H_j is the fixed effect of the jth harvest, B_k is the fixed effect of the kth block, and E_{ijk} is the residual error of the ijkth plot **[$E_{ijk} \sim N(0, \sigma^2_E)$]**. Only 2 harvest dates and 3 blocks were observed, and, because of this, they were treated as fixed effects. H^2 for each feature was calculated as in Equation (5).

Выбор функции

Модели случайной лесной регрессии были включены в R 3.5.3 с использованием пакета VSURF v1.0.4 [73]. Сто лесов, каждый из которых состоял из 2000 случайных деревьев, были подобраны с использованием 68 функций для прогнозирования кластерных назначений. Функция «VSURF :: VSURF ()» возвращает 2 набора функций. Первый включает важные функции с некоторой избыточностью, а второй, меньший набор соответствует модели, более сфокусированной на классификации и сокращающей избыточность [73]. появившиеся во втором наборе для> 3 уровней k были записаны и использованы для классификации для всех кластеров (набор функций 13) .В качестве набора функций 5 были использованы пять объектов, которые имели средние оценки OOB, превышающие медиану (OOB = 0,047). объекты, которые имели средние оценки OOB, превышающие среднюю оценку (OOB = 0,12), были зарегистрированы как набор функций 3.

Feature selection

Random forest regression models were fit in R 3.5.3 using the VSURF package v1.0.4 [73]. One hundred forests, each consisting of 2,000 random trees, were fit using 68 features to predict cluster assignments. The "VSURF::VSURF()" function returns 2 sets of features. The first includes important features with some redundancy, and the second, smaller set corresponds to a model focusing more closely on the classification and reducing redundancy [73]. Features that appeared in the second set for >3 levels of k were recorded and used for classification for all clusters (Feature Set 13). Five features that had mean OOB estimates greater than the median (OOB = 0.047) were used as Feature Set 5. Three features that had mean OOB estimates greater than the mean estimate (OOB = 0.12) were recorded as Feature Set 3.

Классификация производительности

Точность классификации была затем оценена с помощью функции ”MASS :: lda ()” из MASS v7.3−51.1 [104], а также функции “e1071 :: svm ()” из e1071 v1.7−0 [105]. Классификационные модели были обучены определению кластерных назначений из k-средних с использованием 3 различных наборов признаков в качестве переменных-предикторов.Все изображения были случайным образом отсортированы в обучающие и тестовые наборы без стратификации размера 80/20%, 50/50% и 20 /. 80% для изучения взаимосвязи между размером выборки и производительностью модели. Изображения обучающих наборов были сгруппированы с использованием функции «stats :: kmeans ()» в R. Как и раньше, для k в этом эксперименте было разрешено варьировать от 2 до 10. Изображения в тестовом наборе были назначены ближайшему кластеру для каждого значения k. Характеристики ПК (то есть EigenFruitPC [1,7], BioVPC [1,2] и BioHPC [1,3]) были рассчитаны с использованием только изображений обучающего набора, и тестовые изображения были спроецированы в это новое пространство. Максимальное количество ненулевых ПК в этом эксперименте для анализа EigenFruit составляло 5500, 3437 или 1374, в зависимости от размера набора обучающих данных. Процентная разница, объясненная для каждого ведущего ПК, была пересчитана. Геометрические дескрипторы (то есть, BAR, SI и Kurt) не были пересчитаны, потому что они получены из отдельной выборки, а не выборочной совокупности. Наконец, обе модели LDA и SVR были обучены с использованием всех 3 наборов функций для всех значений k с использованием функций «MASS :: lda ()» и «e1071 :: svm ()» в R. Обученные модели использовались для классификации изображения в соответствующем тестовом наборе. Эффективность модели оценивалась с использованием средней точности классификации, точности, отзыва и ложноположительного показателя (FPR) из 10 итераций перекрестной проверки.

Classification performance

The classification accuracy was then estimated using the ”MASS::lda()” function from MASS v7.3−51.1 [104] as well the "e1071::svm()” function from e1071 v1.7−0 [105]. Classification models were trained to delineate the cluster assignments from k-means using the 3 different feature sets as predictor variables. All images were randomly sorted into training and test sets without stratification of size 80/20%, 50/50%, and 20/80% to explore the relationship between sample size and model performance. The training set images were clustered using the "stats::kmeans()” function in R. As before, k was allowed to range from 2 to 10 for this experiment. The images in the test set were assigned to the nearest cluster for each value of k. The PC features (i.e., EigenFruitPC[1,7], BioVPC[1,2], and BioHPC[1,3]) were calculated using only the training set images, and the test images were projected into this new space. The maximum number of non-zero PCs in this experiment for the EigenFruit analysis was either 5,500, 3,437, or 1,374, depending on the size of the training data set. The percent variance explained of each leading PC was recalculated. Geometric descriptors (i.e., BAR, SI, and Kurt) were not recalculated because they are derived from an individual sample and not a sample population. Finally, both LDA and SVR models were trained using all 3 feature sets for all values of k using the "MASS::lda()” and "e1071::svm()” functions in R. The trained models were used to classify the images in the respective test set. The model performance was evaluated using the average classification accuracy, precision, recall, and false-positive rate (FPR) of 10 iterations of cross-validation.

Наличие подтверждающих данных и материалов

Данные, подтверждающие результаты этой статьи и дополнительные данные, доступны в репозитории Zenodo [66]. Код для воспроизведения этих анализов документирован и доступен на GitHub [67]. Данные, подтверждающие эту работу, доступны в репозитории GigaScience, GigaDB [68].

Availability of Supporting Data and Materials

The data supporting the results of this article and supplementary figures are available in the Zenodo repository [66]. The code to reproduce these analyses is documented and available on GitHub [67]. Data further supporting this work are available in the GigaScience repository, GigaDB [68].

Сокращения

AIC: информационный критерий Акайке; БИК: байесовский информационный критерий; BLUP: лучший линейный непредвзятый прогноз; CSV: значения через запятую; EFA: эллиптический анализ Фурье; FPR: ложноположительный показатель; GM: геометрическая морфометрия; ГПД: обобщенный анализ прокрустов; GWAS: исследования геномных ассоциаций; LDA: линейный дискриминантный анализ; МДС: многомерное масштабирование; OOB: ошибка вне пакета; ПК: основной компонент; PCA: анализ основных компонентов; PH: постоянная гомология; PPKC: основная прогрессия k кластеров; SEM: модель структурного уравнения; SIOX: простое извлечение интерактивных объектов; SVR: регрессия опорных векторов; VSURF: выбор переменной с использованием случайных лесов.

Abbreviations

AIC: Akaike information criterion; BIC: Bayesian information criterion; BLUP: best linear unbiased prediction; CSV: comma-separated values; EFA: elliptical Fourier analysis; FPR: false-positive rate; GM: Geometric Morphometrics; GPA: generalized Procrustes analysis; GWAS: genome-wide association studies; LDA: linear discriminant analysis; MDS: multi-dimensional scaling; OOB: out-of-bag error; PC: principal component; PCA: principal component analysis; PH: persistent homology; PPKC: Principal Progression of k Clusters; SEM: structure equation model; SIOX: Simple Interactive Object Extraction; SVR: support vector regression; VSURF: Variable Selection Using Random Forests.