EXPLORING THE MATHEMATICAL SHAPE OF PLANTS

By

Erik J. Amézquita

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computational Mathematics, Science, and Engineering—Doctor of Philosophy

2023

ABSTRACT

Shape plays a fundamental role across all organisms at all observable levels. Molecules and proteins constantly fold and wrap into intricate designs inside cells. Cells arrange into elaborate motifs to form sophisticated tissues. Layers of different tissues come together to form delicate vascular systems that sustain leaves. Each of these tissues evolved as part of a distinct branch of the ever-growing tree of life. From micro-biology to macro-evolutionary scales, shape and its patterns are foundational to biology. Measuring and understanding the shape is key to extracting valuable information from data, and push further our insights.

Shape is too complex to be comprehensively tackled with traditional methods. Landmark-based morphometrics fail if there are not enough homologous points shared across all sampled individuals. Elliptical Fourier Descriptors are not suitable for 3D data. These limitations are especially pressing when we combine our plant visualizations with X-ray Computed Tomography (CT) technology to also record the sophisticated internal structure of stems, seeds, and fruits.

Here, I study the potential of Topological Data Analysis (TDA) for plant shape quantification. TDA is a combination of different mathematical and computational disciplines that seeks to describe concisely and comprehensively the shape of data in a general setting. In very succinct terms, TDA consists of two basic ingredients. First we think the data as a collection of points. Second, we define a notion of distance between every pair of points. The points could be atoms, biomolecultes, cells' nuclei, image pixels, or an organism itself. Distances between points could be the Euclidean, geodesic, genetic, or correlation-based. Once we have data points and distances, we merge systematically the points, starting with those that are closer to each other. The key idea is to keep track of distinct blobs, loops, and voids that form and disappear as we merge several points. This versatile idea is not constrained to a particular dimension or set of landmarks, which makes it ideal to compare a vast array of possible different shapes.

In this work, we will explore new techniques for mathematical plant phenotyping by studying three concrete cases. For the first case, we digitally extract the totality of shape information from X-ray CT scans of tens of thousands of barley seeds. With the Euler Characteristic Transform,

topological and traditional morphological descriptors of the seeds are then used to successfully characterize different barley varieties based solely on the grain shape. This result later enables us to deduce potential genes that contribute to distinct morphology, bridging the phenotype with its genotype. A future goal is to link these genes to climate adaptability to breed better crops for an ever changing weather. For the second case, we use directional statistics and persistence homology to model the shape and distribution of citrus and their oil glands. This leads us to a novel path to explore developmental constraints that govern novel relationships between fruit dimensions from both evolutionary and breeding perspectives. For the third case, we comprehensively measure the shape of walnut shells and kernel. Combining novel size- and shape-specific descriptors, we explore the relationship between shell morphology and traits of commercial interest such as the easiness to remove the kernel intact or the integrity of the shell after being cracked. From the perspective that all data, whether phenotypic or genotypic, has shape, TDA can extract the totality of morphological information. We have interest applying this approach to more crops, to more plant biology inspired datasets, and to large-scale gene expression and population genetic data.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Un viento que no deja crecer ni a las dulcamaras: esas plantitas tristes que apenas si pueden vivir un poco untadas en la tierra, agarradas con todas sus manos al despeñadero de los montes.

—from *El Llano en Llamas*

JUAN RULFO

Demeter was the Greek goddess of the harvest, agriculture, fertility and religious law. She, as the rest of the Greek deities, was an embodiment of multitasking and multidisciplinarity. In fact, most of the Olympus had to supervise multiple tasks each, as it seems that economic trouble and budget constraints have plagued Greece since mythological times. Demeter must have had a precise knowledge of yields and cycles of all the different crops, along with fine-grained details on how each variety of cereal reacts to different soils, climates and farming practices. On top of that, she had to supervise and assist the proper following of sacred rules. With such a packed schedule, she needed to identify quickly the important traits of numerous cultivated plants. Perhaps a thorough, informed glance at the shape of each grain and spike revealed to her all the information required to understand the yield and resistance properties of different plant species. Maybe Demeter used mathematics, especially algebraic topology, and some directional statistics to spice things up, to comprehensively quantify and compare shape.

Shape plays a fundamental role across all organisms at all observable levels. Molecules and proteins constantly fold and wrap into intricate designs inside our cells. Cells arrange into elaborate motifs to form sophisticated tissues. Layers of different tissues come together to form delicate vascular and nervous systems that sustain hands, wings, or fins. Each of these limbs evolved as part of a distinct branch of the ever-growing tree of life. From micro-biology to macro-evolutionary scales, shape and its patterns are foundational to biology. Measuring and understanding the shape is key to extracting valuable information from data, and push further our insights.

Even if we limit our scope to plant biology, a simple glance outdoors reveals a large diversity

of shape among flowers, leaves, fruits, and branches. A first attempt to characterize the shape of a plant could use traditional morphometrics, describing the shape of the plant in terms of height, stem thickness, or number of branches. With these measures at hand, we could look for allometry —the relative growth of parts of an organism to the whole— and thus linearly transform biological shapes between each other. However, any given plant shape is too complex to measure it simply in terms of length, width, and branching angles. A more careful attempt, as suggested by Bookstein (1997), could use geometric modern morphometrics (GMM) instead, where we first define homologous landmark points on every sample and then measure shape similarity by overlapping all these landmarks and computing their Euclidean differences. The computation of differences can be refined by rotating, translating, and scaling appropriately the landmark coordinates prior to the general overlap. This procedure, known as generalized Procrustes analysis, defines a morpho-space, or a space of all possible shapes based on all the possible landmark configurations, which allows us to define overall shape distance (Gower, 1975). The GMM approach can produce distorted results if there are not enough landmarks shared across all sampled individuals, which could occur if we attempt to compare tissues from different families. In the absence of corresponding sets of coordinates, we may attempt to describe the outline of the shape using Fourier analysis (Lestrel, 1997), by considering the outline as a harmonic series, or the sum of wave-like curves (Kuhl and Giardina, 1982). Both morphometrics or Fourier analysis have proven to be extremely insightful to uncover hidden patterns that mold diverse organism shapes at genetic, developmental, evolutionary, and environmental levels (Chitwood and Sinha, 2016).

The methods described above, face enormous challenges when we combine our plant visualizations with X-ray CT (computed tomography) technology to also record the sophisticated internal structure of stems, seeds, and roots. Even with high-resolution images, it is hard to comprehensively and simultaneously measure the vast array of external patterns and internal structures across all possible tissues. Complex branching architectures multiply tenfold in front of our eyes with CT technology, as we now have access to vascular and nervous systems, hyphae, and a better insight of the overall evolutionary tree. X-ray CT imaging has provided new insights into root architecture

(Atkinson et al., 2019; Booth et al., 2020; Griffiths et al., 2022; Zhou et al., 2020), xylem vessel mechanics (Choat et al., 2018; Gauthey et al., 2020), and internal voids in flowers and berries (Xiao et al., 2021). But at the same time in this CT setting, shared coordinates and common outlines are simply not enough to capture the rich morphology we observe both with our eyes and with X-rays (Li et al., 2020). A new strategy is needed. We focus thus on topology and as famously stated by David Kendall (1984),

> *As topologists already have a theory of* shape*, I must apologize for using the word again with an entirely different meaning. In this paper 'shape' is used in the vulgar sense, and means what one would normally expect it to mean.*

Topological Data Analysis (TDA) is a combination of different mathematical and computational disciplines that seeks to describe concisely and comprehensively the shape of data in a general setting (Amézquita et al., 2020; Lum et al., 2013; Munch, 2017). In extremely succinct terms, TDA consists of two basic ingredients and a key idea. The first ingredient is to think of the data as a collection of points, and the second is to define a notion of distance between every pair of points. The points could be atoms, biomolecultes, cells' nuclei, image pixels, or an organism itself. Distances between points could be the Euclidean, geodesic, genetic, or correlation-based. Once we have data points and distances, known formally as a metric space, we can connect these points starting with those that are closer to each other first. The key idea is to keep track of distinct shape features that form and disappear as we connect and merge several points. These ingredients and idea, albeit simple, are extremely versatile and can be adapted to a myriad of contexts and data collections. Moreover, the notion of shape presented by TDA is limited solely by the data itself, unleashing it from possible selection biases. This very adaptability and impartiality makes TDA a powerful data analysis tool that can further our insights in a variety of plant biology scenarios.

Plant biology faces mountains of genetic and phenotypic data that must be efficiently analyzed and summarized. The solutions to grand challenges we face in the plant sciences, including predicting phenotype from molecular profiles, lie in mathematics (Autran et al., 2021; Bucksch et al., 2017). Despite emerging interdisciplinary research networks, the active collaboration between

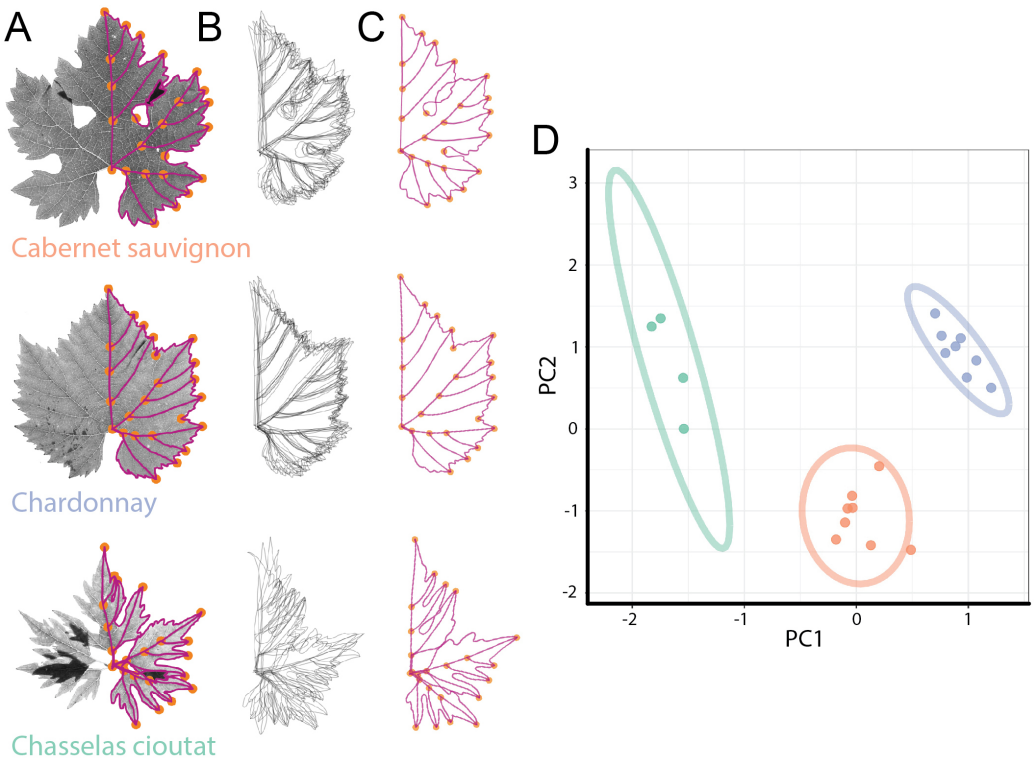

Figure 1.1: **An example of geometric morphometrics. A)** 24 landmarks (orange dots) and pseudo-landmarks (6,000 evenly spaced vertices between landmarks, magenta dots) on grapevine leaves of Cabernet sauvignon (orange), Chardonnay (blue), and Chasselas cioutat (green) varieties. Every grapevine leaf has five major veins, allowing corresponding landmarks to be placed throughout every leaf. **B)** Corresponding vertices allows replicates to be superimposed on each other and **C)** mean leaves calculated using Procrustean methods that translate, rotate, reflect, and scale. **D)** A Principal Component Analysis (PCA) and other statistics can be performed on the Procrustes-adjusted vertices (95% confidence ellipses for each variety are shown).

these two domains remains limited. We propose the mathematical study of shape as one of the many potential bridges between mathematics and plant biology.

In recent years, TDA has produced promising results in diverse biological problems, like histological image analysis (Kovacev-Nikolic et al., 2016), viral phylogenetic trees description (Chan et al., 2013), and active-binding sites identification in proteins (Qaiser et al., 2019). In plant biology specifically, it has been used to characterize the morphospace of all possible leafs as in Figure 1.1 (Li et al., 2018), the 3D structure of grapevine inflorescence (Li et al., 2019), the shape of different apple accessions (Migicovsky et al., 2019). The Euler characteristic has also been used to study the genetic basis of cranberry shape (Diaz-Gárcia et al., 2018), the hairiness and shape

of spikelets —arrangements of grass flowers— (McAllister et al., 2019) and patterns of vegetation from satellite imagery (Mander et al., 2017). The power of TDA lies on its versatility to produce different "topological signatures" for different shapes.

In Chapter 2 we present the main mathematical concepts behind TDA at an intuitive level. We provide a middle ground between mathematics and biology: for mathematicians, a review of ways TDA has been successfully used to study biology and for biologists, an accessible introduction to topological thinking. We begin with examples from structural biology, evolution, and cellular architecture that lend themselves to simple but powerful TDA representations. We then examine shapes, focusing on the outlines of leaves, and the use of Euler characteristic curves as convenient topological signatures that enable statistical analyses. Next, we highlight ways that TDA can measure branching architecture and the use of bottleneck distance to calculate the overall topological similarity between objects. We end with a discussion about future trends in TDA: measuring dynamic shapes and time series as well as using topology to convert data to graphs representing its structure.

In Chapter 3, as a proof of concept, we quantify the morphology of X-ray CT scans of barley spikes and seeds using both traditional and topological shape descriptors. By combining both sets of descriptors, we can successfully train a support vector machine to distinguish and classify 28 different varieties of barley based solely on the 3D shape of their grains. Rather than being an alternative, we propose TDA as a powerful complement to traditional shape analysis, where the topological descriptors pick up morphological information that is usually missed. We propose then that this shape characterization will allow us later to link genotype with phenotype, furthering our understanding on how the physical shape is genetically specified in DNA.

In Chapter 4, we investigate the shape of citrus fruit. We analyze 3D X-ray CT scan reconstructions corresponding to 51 citrus accessions. Based on the centers of the oil glands, overall fruit shape is approximated with an ellipsoid. Possible oil gland distributions on this ellipsoid surface are explored using directional statistics. Our observations point to the existence of biophysical developmental constraints that govern novel relationships between fruit dimensions from both

evolutionary and breeding perspectives. Understanding these biophysical interactions prompt an exciting research path on fruit development and breeding.

In Chapter 5, we explore the shape of walnut shells and kernels. We analyze almost 1300 individual 3D X-ray CT scan reconstructions of walnuts, corresponding to 171 walnut accessions. Like in the previous chapter, we exploit the nondestructiveness of X-rays to digitally segment and measure the 4 main tissues of interest for each walnut, namely shell, kernel, packing tissue, and sealed air. From these we extract a total of 38 size- and shape-specific descriptors, many of them unexplored in the current literature. We focus on several allometric relationships of interest, from which we draw theoretical upper and lower bounds of possible walnut and kernel sizes. We then study correlations and variations of these morphological descriptors with qualitative data corresponding to traits of commercial interest like ease of kernel removal and shell strength.

Finally, Chapter 6 poses several exciting potential future directions we could take for each of the previous three chapters.

CHAPTER 2

BACKGROUND

En qué momento habían resucitado, cómo había sido la sensación de
pasar del polvo a la forma y de la forma a la vida y se pellizcaban
para ver si les salía sangre
—from *El Tiempo Principia en Xibalbá*
Luis de Lión

Biologists are accustomed to thinking about how the shape of biomolecules, cells, tissues, and organisms arise from the effects of genetics, development, and the environment. Less often do we consider that data itself has shape and structure, or that it is possible to measure the shape of data and analyze it. Here, we review applications of Topological Data Analysis (TDA) to biology in a way accessible to biologists and applied mathematicians alike. TDA uses principles from algebraic topology to comprehensively measure shape in datasets. Using a function that relates the similarity of data points to each other, we can monitor the evolution of topological features ——connected components, loops, and voids. This evolution, a topological signature, concisely summarizes large, complex datasets.

This chapter is derived from the review paper

- **E.J. Amézquita, M.Y. Quigley, T. Ophelders, E. Munch, D.H. Chitwood** (2020). The shape of things to come: Topological data analysis and biology, from molecules to organisms. *Developmental Dynamics* 249(7): 816–833.

2.1 Topological Data Analysis (TDA): a primer

2.1.1 The Vietoris-Rips Complex

Topology is the branch of mathematics concerned with mathematical properties that are preserved under continuous transformations. With some mathematical framework, described below, topology offers powerful tools that can precisely describe the overall shape and structure of the data encoded by a given network. Informally, we can think the topology of a network as the collection

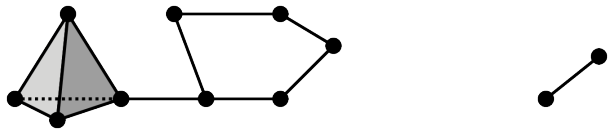

Figure 2.1: An example of a simplicial complex. It has two connected components, one loop, and one void.

of its features that remain unchanged whenever the data "varies smoothly". For example, scaling, centering, translation, and rotation are all smooth operations that do not alter the topology (i.e., the core shape) of our data. However, partitioning, merging, and attaching are not smooth operations and may significantly alter topology.

In a mathematical context, networks are referred to as graphs. Nodes or points are referred to as vertices, while links between nodes as edges. We can generalize the idea of graphs by adding triangles that link edges or even tetrahedrons that link triangles. More formally, we can think of our data as composed of different building blocks, called simplices. Vertices, edges and triangles are 0-, 1- and 2-dimensional simplices, respectively. A collection of multiple simplices makes a simplicial complex, or complex, for short. For example, in Figure 2.1 we have a complex made of vertices, edges, and triangles.

We can describe the topology of a complex based on the number of its connected components, loops, and voids. For example, in Figure 2.1 we can see two distinct, separate pieces, each of them being a connected component. We see that five edges in the left component form the frame of a pentagon. We say then that these five edges form a loop. Also on the left, we see a collection of four triangular faces that form a tetrahedron. We can assume that this tetrahedron is hollow, so that the complex contains a void.

Many times, our data or network cannot be thought immediately as a complex. However, we can generate a complex based on a collection of data points and a notion of similarity or distance between these points. Formally, a collection of individual points and positive distances between every pair of points is referred to as a metric space. The Vietoris-Rips (VR) complex is a versatile method to define a complex from a network. The VR complex starts with data in a metric space and a fixed nonnegative parameter r, often referred to as the radius. If two vertices are close enough,

that is, the distance between them is less than r, then the VR complex will have an edge between those two vertices. Similarly, if there are three vertices close enough, that is, the distance between every pair of them is less than r, then the VR complex will have a triangle between those three vertices. Following these two rules, every time we have a triangle, we also have the three edges that make the frame or border of such triangle. Conversely, every time a trio of vertices form a triangular frame, the VR complex will also contain the corresponding triangle.

2.1.1.1 Walking through an example

Notice that the same data and metric can produce different VR complexes by using a different parameter r each time. We can consider a sequence of increasing radii and its corresponding sequence of VR complexes. First observe that the distance between any two different data points is always a positive quantity. If we start with $r = 0$, then the corresponding VR complex will consist solely of separate vertices, one for each data point. As the radius r increases, the corresponding VR complex will now have edges that link the pairs of vertices that are close to each other. If r keeps increasing, we may then have triangles that link trios of close vertices.

For example, consider the five data points in Figure 2.2.A, which we can take as vertices of a complex. The distance between the points will be simply the Euclidean distance. Consider seven different positive radii. For the first three radii, the shape remains the same: just five separate components. Suddenly, as soon as we increase the radius a fourth time, four pairs of vertices are finally close to each other so we draw edges between them to form a square. There is also a fifth vertex that remains isolated, as it is still distant from the rest. When the radius increases a fifth time, the isolated vertex is finally close enough to one of the square vertices. We draw one more edge at this point. The radius increases a sixth time, so that the pair of diagonal vertices in the square are close enough. We then draw the diagonals of the square, which also draws the four possible triangles in the square. The radius finally increases a seventh time, so that the fifth vertex is closer to another vertex in the square. We then add an edge and a triangle including this fifth vertex. As the radius keeps increasing beyond this, the overall shape of the VR complex will not have any significant changes: it will always remain a single component with no holes.

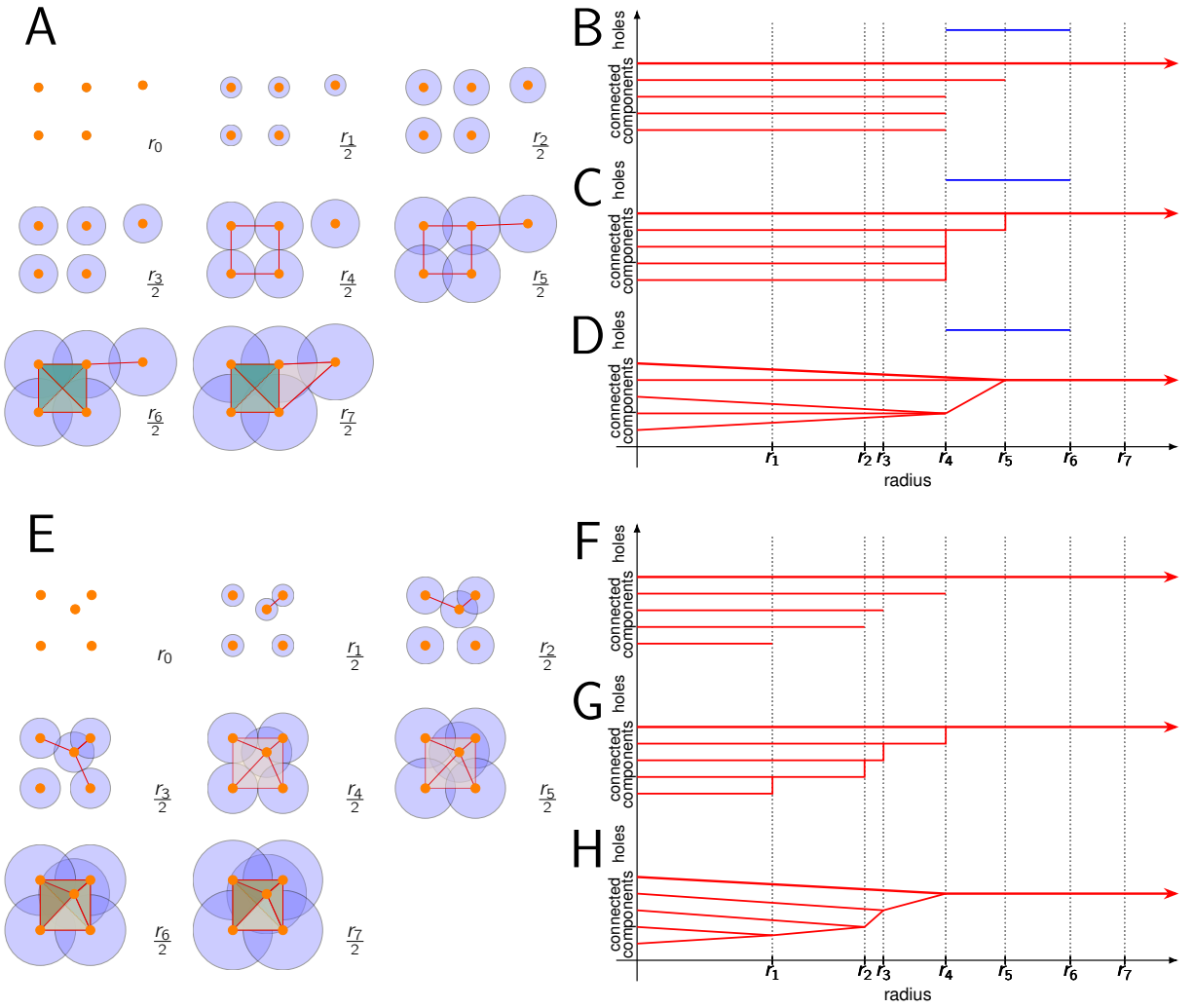

Figure 2.2: An example of two different Vietoris-Rips complexes with resulting persistence barcodes. (A) Evolution of a VR complex with 7 vertices as Euclidean distance increases. (B) Persistence barcode corresponding to topological changes in the previous VR complex. (C) Alternative visualization of the persistence barcode B as a dendrogram. (D) Alternative visualization of the persistence barcode B as a tree. (E) Moving one vertex in A yields a different VR complex as Euclidean distance increases. (F) Persistence barcode corresponding to topological changes in the previous complex E. (G) Alternative visualization of the persistence barcode F as a dendrogram. (H) Alternative visualization of the persistence barcode F as a tree.

2.1.2 Representing persistent features

All the observations described above can be summarized using two topological features: connected components and holes. For connected components we need to keep track of which snapshot each connected component appeared (was born) and in which snapshot two separate components merged (died). Similarly, we can keep track of when each hole is formed (born), and when it is filled (dies). These life spans of topological birth and death can be drawn as life bars, the length of which indicates for how long a component persisted before it merged or how long a hole persisted before being filled. Putting all the bars together, we obtain a *persistence barcode*, in which each bar corresponds to a topological feature and the horizontal axis indicates at which radius value these features are born and die. Note that the vertical order of these life bars is irrelevant.

For the persistence barcode in Figure 2.2.B, we observe that we start with five different vertices, all of which remain separate (the components persist) until the fourth radius. By the fourth radius, we only have two connected components: one square and one distant vertex. We also observe the birth of the hole in the square (indicated in blue). By the fifth radius, the distant vertex has merged with the square so we have only one connected component. By the sixth radius, we observe that the square hole has been filled with triangles. From this point onwards, as radius keeps increasing, our VR complex will be essentially a single connected component with no holes. We say then this components dies at infinity and it is represented by the continuing red arrow. We can alternatively display the persistence of components and holes as a dendrogram (Figure 2.2.C) or a tree (Figure 2.2.D), keeping track of which components merge. A particularly useful display of persistence barcodes are persistence diagrams, as illustrated in Figure 2.3. Simply, the birth start point and death end point of a bar in a persistence barcode are transformed as (x, y) coordinates in a death-vs-birth plane in a persistence diagram. Persistence diagrams have a convenient visual and mathematical representation which has allowed further theoretical developments in TDA.

Barcodes are a useful way to illustrate and summarize prominent topological features, such as the distant fifth vertex or the hole enclosed by a square. Consider now "obstructing" this hole by moving the distant fifth vertex inside the square, as in Figure 2.2.E. We observe in Figures 2.2.F–H

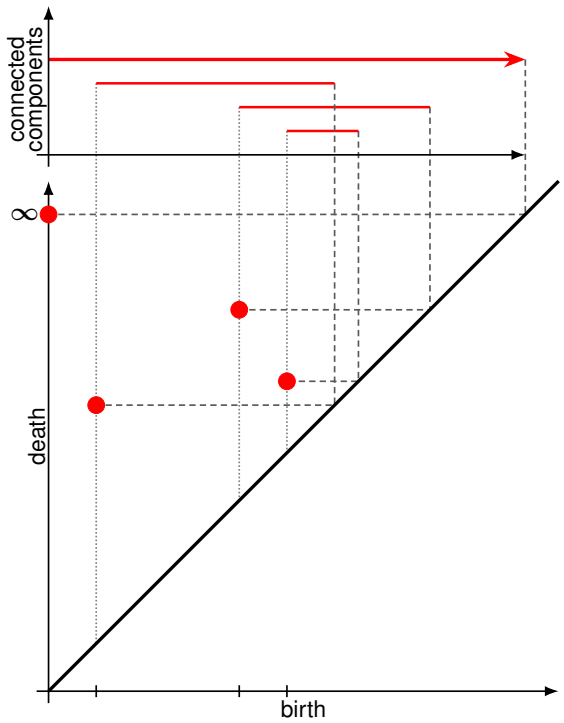

Figure 2.3: Translating a persistence barcode into a persistence diagram. Birth and death times in the persistence barcode are interpreted as (x, y) coordinates on a death-vs-birth plane. This planar display is referred to as a persistence diagram.

a different persistence barcode, dendrogram, and tree, respectively. The barcode now shows that all five vertices merge into a single connected component at earlier stages compared to Figure 2.2.A. Also notice that we have now filled the square's hole, so that the barcode in Figure 2.2.F registers no holes, unlike Figure 2.2.B.

2.1.3 Filters: Beyond spatial distances

As mentioned before, the Vietoris-Rips complex is constructed from a set of vertices and a sense of distance or similarity between these. Usually we refer to such a measure of similarity as a filter function. This function can determine when the vertices and their edges in between are observed. For example, in 2.2.A and 2.2.E, our filter function was the Euclidean distance between vertices. Given a filter function, we can consider a series of snapshots, where in each snapshot, we consider larger and larger filter values, called thresholds. Going back to 2.2.A and 2.2.E, each snapshot considers increasing radius lengths around each vertex. In this case, we say that our collection of data points have been filtered by Euclidean distance with six thresholds. Notice that if we increase

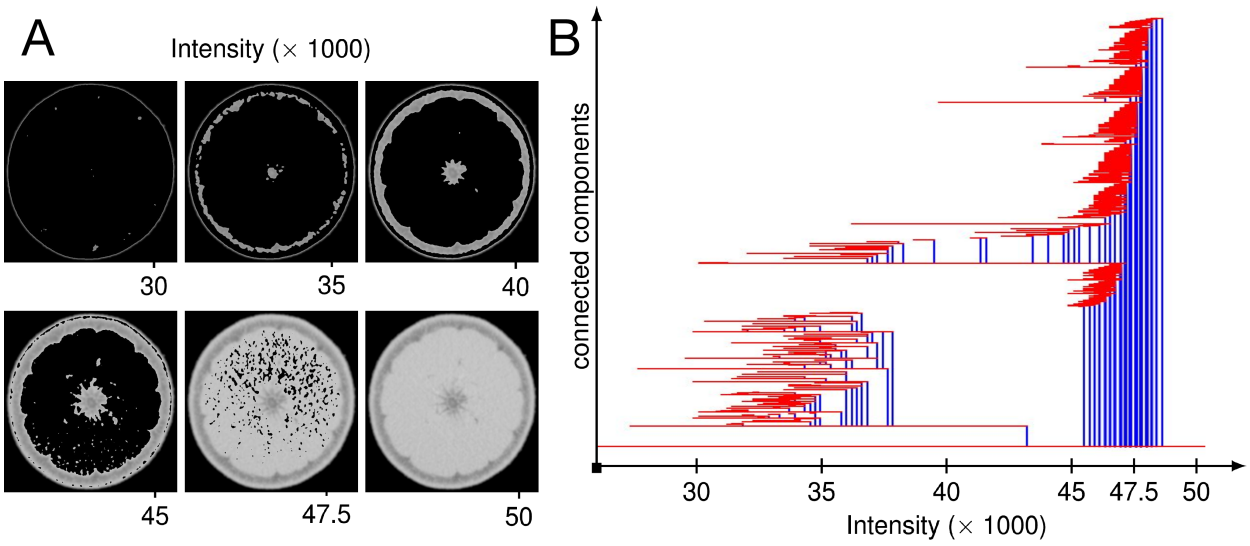

Figure 2.4: An example of a persistence barcode. (A) Snapshots of an X-ray CT image of an orange. Only the pixels with intensity lower than indicated are displayed. (B) Persistence barcode of connected components of such image. Observe that the barcode distinguishes the existence of exocarp, rind, and pith as separate components at lower intensities.

the number of thresholds, we may be able to capture finer topological changes which may in turn produce richer persistence barcodes.

Filter functions are extremely flexible and we can use more than a spatial distance. For example, consider a grayscale image. We will consider each pixel a separate vertex and use an intensity filter, resulting in the distance between two pixels simply being the difference of their intensities. We then consider each possible intensity value as a threshold. Figure 2.4 shows the persistence barcode of connected components from an X-ray Computed Tomography (CT) scan of an orange where we consider more than 50,000 threshold values (Figure 2.4.A). In each snapshot, we only display the voxels (vertices) whose intensity value is less than the value of the threshold. At 30,000, we only observe the contour of the exocarp with some separate bits of rind. At 35,000, more bits (connected components) of rind appear, and some of these rind bits merge into each other. Additionally, we observe the appearance of the pith. By 40,000, we have 3 clear separate connected components, namely exocarp, rind and pith. By 45,000, the rind and the exocarp have merged while numerous bits of endocarp have appeared. By 50,000, the appearance of the endocarp has merged the pith to the exocarp, yielding a single connected orange.

2.2 Applied topology: examples from structural biology, evolution, cellular architecture, and brain networks

The Vietoris-Rips complex framework introduced above, filtering on the Euclidean distance between data points, can be used to study a wide range of complex phenomena in biology. A metric space might be the 3D coordinates of amino acids in a protein, could represent species or virus variants separated from each other by genetic distance, or might be defined by the nuclei of cells in a cross-section of tissue. Below, we provide examples where the Vietoris-Rips complex has successfully been applied to structural biology, evolution, and cellular architecture (Figure 2.5).

2.2.1 Structural biology

This prototypical example of TDA —a metric space consisting of points where the filter is Euclidean distance—can be extended to biomolecules, where the points are atoms or residues (Figure 2.5.A). Proteins are comprised of a linear polymer of amino acids. The primary structure of a protein is the sequence of amino acids. The polypeptide chain of a protein folds upon itself, stabilized by interactions between amino acids, first forming a secondary structure (local structures such as alpha helices and beta sheets formed by hydrogen bonds in the peptide backbone) followed by the tertiary structure. The overall 3D structure of a protein is determined by the interactions of amino acid side chains within the protein. This overall structure, or conformation, of a protein is the basis of protein function: metabolism, transport, signaling, structure, and movement, among many others. The conformation of a protein can change depending upon binding ligands, signaling, or the chemical environment.

Kovacev-Nikolic et al. (2016) use TDA to distinguish the open and closed conformations of maltose-binding protein (MBP). Each of the 370 amino acids of MBP is treated as a vertex, its 3D coordinates reflecting the spatial location of the residue. Euclidean distance is used to create a filtered Vietoris-Rips complex for each protein studied. As the Vietoris-Rips complex evolves, persistence barcodes record the birth and death of connected components, loops, and voids, which within the context of the tortuous folding of a protein backbone, yield complex topological signatures unique to distinct conformations. These persistence barcodes are transformed

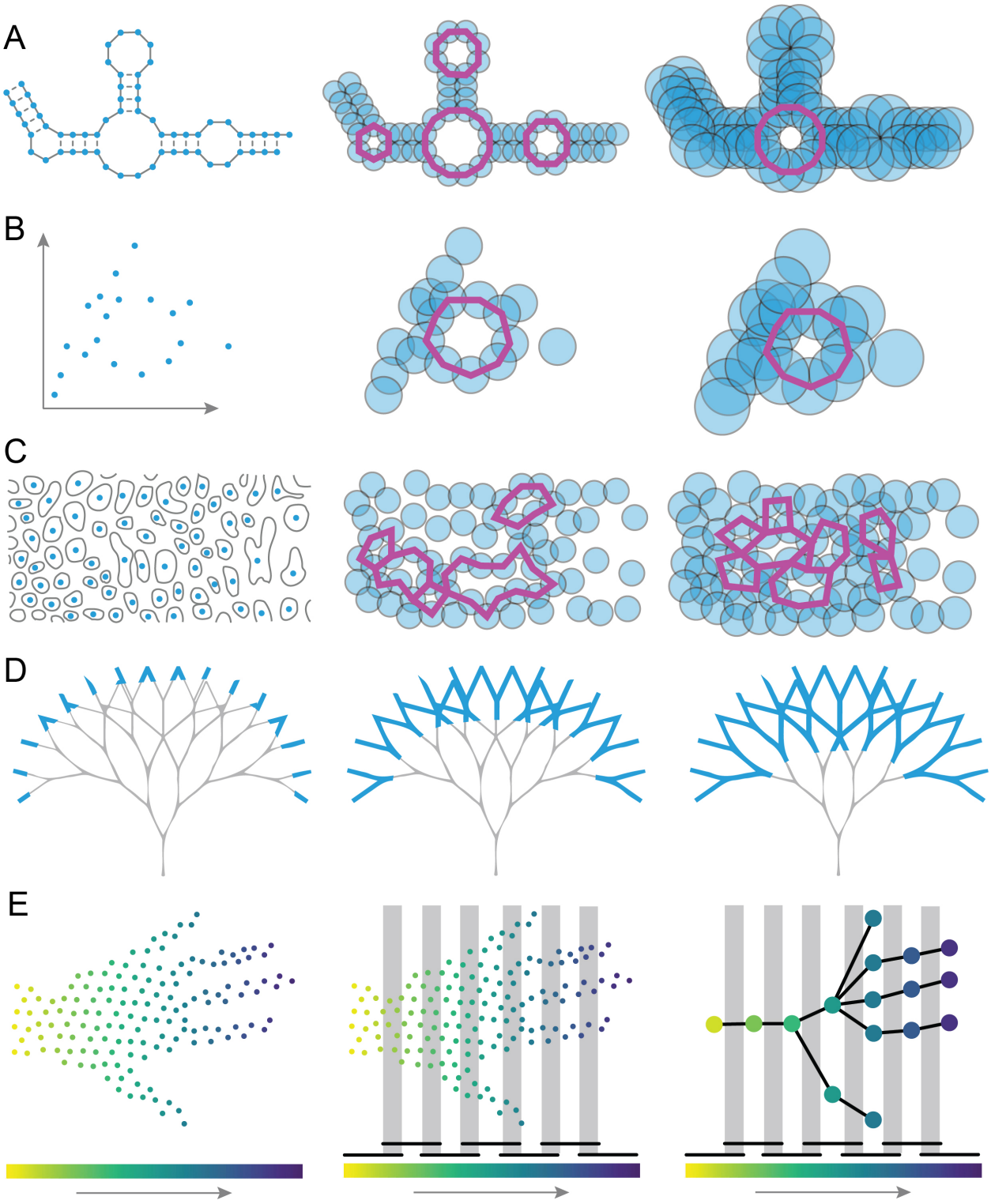

Figure 2.5: Applications of Topological Data Analysis (TDA) to biology, from protein structure, to individual cell patterns, and phylogenetic tree analyses.

Figure 2.5 (cont'd):
Applications of Topological Data Analysis (TDA) to biology. (A) Structural biology. A diagram of RNA secondary structure (left; solid lines covalent bonds, dashed lines hydrogen bonds). Increasing radii of vertices (middle, right; blue points) is used to visualize filtration on Euclidean distance. As radii merge, connected components die. Purple lines indicate the formation of loops that eventually fill in as the radius threshold increases. (B) Evolution. A plot showing the genetic distance of samples (left). As radius threshold value increases (middle, right) the birth and death of connected components (blue) represents vertical evolution (a tree) while that of loops (purple) horizontal evolution events (such as hybridization, gene transfer, or recombination; modified from Chan et al. (2013)). (C) Cellular architecture. Modification of a part of the original Gleason guide to prostate cancer changes in cellular architecture (left). Nuclei (blue) increase in radius (middle, right) and connected components (blue) and loops (purple) are born and die. (D) Branching architecture. A theoretical tree where the filter is geodesic distance to the base (blue). Branching tips are separate connected components that merge as the filter progresses to the base of the tree (left to right). (E) Mapper. Point cloud of a hand where the filter is the axes from the wrist to fingertips (left). Cover intervals (bars on top of color scale) and their overlap (gray bars) divide points into bins (middle). Points that cluster together over each cluster are assigned to a vertex, and if the points are shared between clusters in an overlap, then they are assigned to an edge connecting the corresponding vertices (modified Lum et al. (2013)).

into *persistence landscapes* (Bubenik, 2015) that allow statistics, hypothesis testing, and machine learning to be applied to differentiate the shapes captured by topological signatures. The authors successfully differentiate open- and closed-conformation states of MBP. They also note that the active site residues (the amino acids responsible for ligand binding) lie at the edge of the most persistent loop of the Vietoris-Rips complex, indicating that TDA is sensitive to the relationship between structure and function. Beyond structure, electrostatic and other chemical properties of atoms can be incorporated into topological signatures that, when analyzed using machine learning methods, can predict protein-ligand binding affinities (Cang et al., 2018; Cang and Wei, 2018).

2.2.2 Evolution

Evolution is typically depicted as a tree, which in mathematical terms is an acyclic graph (a graph with no loops). Each node and its descendant branches represent a common ancestor of a taxonomic group and its members as a hierarchy of similarity or relatedness. Evolutionary trees depict *vertical evolution*, random mutations that accumulate within a specific lineage that lead to phenotypic changes. But genetic material can be exchanged between lineages as well. This

process, known as *horizontal evolution*, is depicted as a reticulate graph (with loops), in which genetic information is exchanged by recombination, hybridization, horizontal gene transfer, or viral reassortment. Extensive phylogenetic theory models vertical evolutionary processes using trees, but the study of horizontal evolution is often limited to detecting reticulate phylogenetic events, and a theory unifying vertical and horizontal evolution had remained elusive.

Chan et al. (2013) reconsider evolution from the perspective of topology. Using influenza as an example, they begin by considering that every sample has a genetic distance to every other, a metric space. From this genetic space, they construct a Vietoris-Rips complex, just as in the previous examples (Figure 2.5.B). The resulting persistence barcode for connected components can be converted into a dendrogram, which is the phylogenetic tree that biologists are accustomed to. Influenza viruses extensively exchange genetic material in a process known as reassortment. If the persistence barcode highlights loops, then this represents a topological signature of horizontal evolution. For example, a lower bound of recombination rate can be calculated from the number of loops (recombination or reassortment events) for a given time frame (in this example, the filter of genetic distance which can be calibrated to time). In higher dimensional spaces, voids can be detected, and the authors show that persistence barcodes for voids detect more complicated reassortment events, such as the triple reassortment that gave rise to the 2013 avian influenza outbreak and complex reassortment events within HIV. Using loops as an estimator of recombination rate has been extended to large-scale genomic analysis of human biology (Cámara, 2017; Cámara et al., 2016), expanded upon by evaluating different topological features (Humphreys et al., 2019), applied to coalescent theory to estimate ancestral recombination events (Emmett et al., 2014), and used to study lateral gene transfer of protein families and its implications for the evolution of antibiotic resistance (Emmett and Rabadán, 2014).

2.2.3 Cellular architecture

Tissues are comprised of cells, the organization of which is determined by cell division, differentiation, growth, movement, migration, and death. Within a tissue each cell takes up a finite volume, often in close contact with neighbors. When a tissue is finely cross-sectioned and micro-

scopically examined, a tessellated array of cells emerges: an aggregated mixture of parenchyma, stroma, and glands (Figure 2.5.C). Staining can differentiate nuclei, cytoplasm, and extracellular matrix. To a trained eye, these complex patterns can indicate disease or abnormalities, but the process takes time and is subjective. The emergent organization of cells reflects developmental processes as well. TDA provides an objective way to classify these patterns, potentially removing the subjectivity of histopathological diagnosis enabling a rigorous way to define cellular anatomy.

Lawson et al. (2019) explore the cellular architecture of prostate cancer. The Gleason grading system is a one to five scale that is a powerful prognostic indicator based on increasingly neoplastic tissue organization of the prostate: a uniform cellular architecture becomes disrupted forming glands that eventually form solid cell types. Tissue sections are stained with blue-purple hematoxylin and pink eosin which indicate nuclei and cytoplasm/extracellular matrix, respectively. The authors use these stains to isolate cell nuclei from surrounding structures (Figure 2.5.C). They then use thresholding as a filter on histological images of prostate cancer to create binary images, where connected components and loops are recorded as persistence barcodes. Creating vectors of the most persistent features, they use a variety of statistical techniques including Principal Component Analysis (PCA), hierarchical clustering, and t-distributed Stochastic Neighbor Embedding (t-SNE) to successfully classify images according to the Gleason grading system (Lawson et al., 2019). The strategy of reducing cells to data points of a Vietoris-Rip complex to classify cellular architecture works for predicting epithelial organization from cell centroids (Atienza et al., 2019) and in other cancers as well (Chung et al., 2018; Singh et al., 2014).

2.2.4 Brain networks

The complex architecture of neurons and their numerous connections in the brain, formally referred to as the connectome, is of particular interest. The brain architecture, its activity and connectivity, is usually presented as square pair-wise correlation matrix each row (and column) represents different neurons or encoding brain regions when the subject is performing a fixed task. Usually negative correlations are treated as zero, and the rest of matrix entries are thresholded so that only neurons or regions with strong connectivity are considered. We can then consider a metric

space where the points are different neurons, anatomical regions of interest, or imaging voxels. The distance between these points is given by the correlation between them (or 1 minus correlation to be mathematically consistent). With this setup, it is possible to produce Vietoris-Rips complexes and persistence diagrams that summarize the brain network model.

Observing the change in number of connected components, we can differentiate the abnormal glucose metabolism associated to neuronal activity between attention-deficit hyperactivity disorder (ADHD) children, autism spectrum disorder (ASD) children, and pediatric control subjects (Lee et al., 2012). By keeping track of persistent loops, we can distinguish the effects of psilocybin in the human brain functional patterns (Petri et al., 2014). Persistent loops also highlight that mental imagery shares the same neurophysiological bases with perceptual and motor experience (Ibáñez-Marcelo et al., 2019). TDA has also revealed previously ignored anatomical loops and voids in the connectome, which might explain both spatial and nonspatial behaviors both in mice (Giusti et al., 2015) and humans (Sizemore et al., 2018).

2.3 Shape, texture, and the Euler characteristic curve

The examples above rely on point-based representations of biological data to which a filtered Vietoris-Rips complex on Euclidean (or genetic) distance produces zero-dimensional (connected components) and one-dimensional (loops) persistence barcodes. The persistence of the barcodes, representing prominent topological features, is the focus of analysis. However, the concept of filter can be extended to any real values that can be associated with structures: for instance, different choices of metric space between data points, or even using filter functions based on the data points instead of the edges between the data points. For the same data, a number of filters might be applied, yielding a new lens to reveal different facets of shape. Persistence barcodes can always be calculated, but there are other ways to record topological signatures as well.

Below, we first describe the Euler characteristic curve (ECC) as a convenient complement to persistence barcodes to capture topological signatures that can be used with traditional statistical methods. We then describe TDA frameworks to measure shape (in the traditional sense of a closed contour) focusing on leaf outlines and the usefulness of ECCs to measure genetic and environmental

effects that determine phenotype.

2.3.1 Euler characteristic curve (ECC)

The Euler characteristic, often denoted by the Greek letter χ, was originally defined by the equation:

$$\chi = \#(\text{Vertices}) - \#(\text{Edges}) + \#(\text{Faces}).$$

The Euler characteristic is the first example of a topological invariant; that is, a quantity that can be calculated and returns the same value on many different representations of the same topological shape. For convex polyhedra (e.g., the Platonic solids), the Euler characteristic always equals two since all platonic solids are topologically spheres. For example, a tetrahedron has four vertices, six edges, and four faces ($4 - 6 + 4 = 2$); a cube has eight vertices, 12 edges, and six faces ($8 - 12 + 6 = 2$).

What is even more surprising is that this quantity can also be obtained by counting some intrinsic properties of a given shape. The Euler-Poincaré formula establishes that the formula above is the same as:

$$\chi = \#(\text{Connected Components}) - \#(\text{Loops}) + \#(\text{Voids}).$$

Since all convex polyhedra have one connected component and one void, the Euler characteristic is indeed $1 - 0 + 1 = 2$. On the other hand, a doughnut, mathematically known as a solid torus, has one connected component, one loop, and no voids so its Euler characteristic is $1 - 1 + 0 = 0$. By keeping track of the number of building blocks of our simplicial complex, we can indirectly summarize its topological features.

Similar to persistence barcodes, given a data set, for each sample we define vertices, a filter function, and a number of thresholds. For example, consider a 3D (voxel-based) image of a barley seed (Figure 2.6.A). Each voxel is a vertex in our simplicial complex. One type of filter we can apply is the 3D axes of the coordinate system, which is oriented with respect to the depth, width, and height of the seed. For each of these filters, the voxels take the real number value of their coordinate for the particular axis. We then choose a number of thresholds, or equivalently, we choose how many times to "slice" through the seed along the given axis. Each time we take a slice, we compute

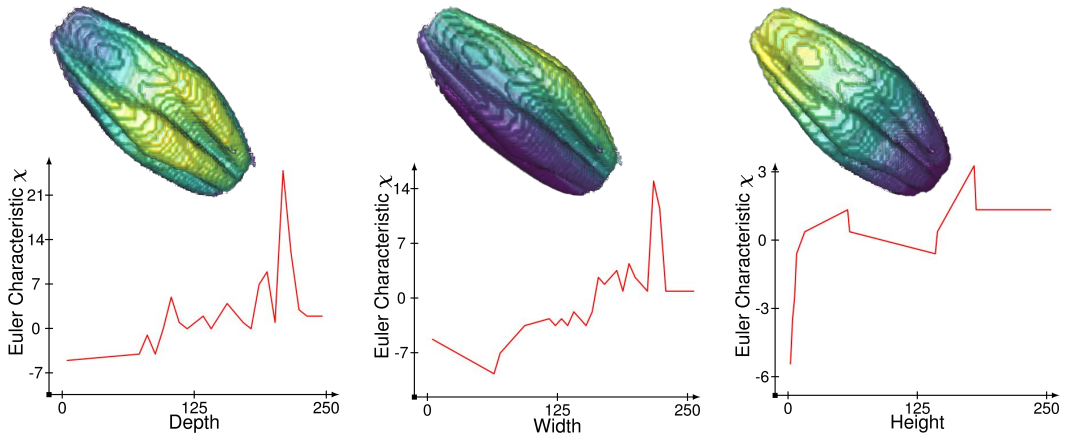

Figure 2.6: Three different Euler Characteristic Curves (ECCs) from three different filters. (A) X-ray CT scan of a barley seed. The symmetry of the seed encourages a filter by depth, width and height values, i.e., the three main axis directions with respect to the seed scan. Slicing the barley seed in different directions produce (B) different corresponding ECCs. Notice that the three curves end with Euler characteristic equal to 1, which corresponds to the Euler characteristic of a solid sphere.

the Euler characteristic of the seed. We continue to add slices one-by-one and recalculate the Euler characteristic each time as we continue through the axis, which is the filter function. Adding all the slices together yields the original seed. Finally, we summarize our computation as an ECC (Figure 2.6.B), where the x-axis is the threshold while the y-axis is the Euler characteristic of the complex at that particular threshold value.

Persistence barcodes tend to be notoriously expensive and difficult to compute since they must keep track of all the possible component merges and hole fillings for every threshold value (Otter et al., 2017). Most of the available software to compute persistence barcodes is incapable of handling truly large data sets effectively, especially when each sample consists of millions of vertices. (That being said, there have been recent breakthroughs to efficiently compute persistent diagrams. Ripser is very a promising software capable of handle more and more data with each new release (Bauer, 2021)). Euler characteristic curves are a convenient way to summarize a topological signature of an object as a sequence of numbers, a curve, or numerical vector. Computing and storing these vectors is quite efficient, and it is especially convenient since it allows us to perform standard statistical analysis techniques and test hypotheses about the shape of our data. This simplicity makes it ideal to process objects that may have more than a billion vertices (Heiss and Wagner, 2017; Wang et al.,

2022).

2.3.2 Shapes and textures

Sometimes leaves have corresponding coordinates, as in the case of grapevine where every leaf has five major veins and numerous landmark features (Chitwood and Sinha, 2016). In these instances, geometric morphometrics is a powerful tool. Besides the base of the petiole and tip, though, most leaves do not have coordinates that correspond in a way that analysis by geometric morphometrics is possible. In order to compare the outlines of 182,707 leaves from 141 plant families and 75 sites throughout the world, Li et al. (2018) used TDA. The pixel outline of each leaf is treated as a point cloud. The filter applied to each pixel is a Gaussian density estimator, sensitive to the number of neighboring pixels around each pixel. Straighter edges of the leaf blade will have low density values while pixels in serrations, lobes, or other undulations will have higher values. The number of connected components is monitored and the respective ECCs are computed.

For so many leaves, an ECC curve serves as a succinct, computationally feasible topological signature that allows downstream statistical analyses. These signatures can define the morphospace for all leaves, which reveals not only the leaf shapes that exist, but those that do not, either because of developmental constraint or negative selection. This morphospace can be used then to predict plant family and location (Li et al., 2018). Others have used the same filter and ECCs to determine the genetic basis of leaf shape in apple (Migicovsky et al., 2018) and tomato (Li et al., 2018) as well as the genetic basis of cranberry shape (Diaz-Gárcia et al., 2018). ECCs are sensitive enough to complex and subtle changes in shape to measure the effects of rootstock and climate on grapevine leaf shape (Migicovsky et al., 2019). ECCs have also been used to measure the hairiness and shape of spikelets (arrangements of grass flowers) (McAllister et al., 2019), patterns of vegetation from satellite imagery (Mander et al., 2017), and flow cytometry features (Smith and Zavala, 2021). Moreover, by considering all the ECCs corresponding to all directional filters, we can successfully encode important morphological information of barley seeds (Amézquita et al., 2021) and protein structure (Tang et al., 2022).

2.4 Branching architectures and bottleneck distances

2.4.1 Persistence diagrams

The Euler characteristic allows us to monitor a topological summary as a function of the filter we choose. The resulting curve enables statistical analyses. In some cases, we might not want a summary, though; we may want to keep track of each topological feature separately, as we do in a barcode. The bottleneck distance is a convenient way to determine the overall topological similarity of two barcodes with each other. If we compute the bottleneck distance of all barcodes to all other barcodes, we can determine the overall topological similarity of samples to each other, in which case statistical analyses can be performed. To understand the meaning of bottleneck distance, we need a better display of topological information than persistence barcodes. We thus turn to persistence diagrams.

As mentioned previously, each topological feature in the barcode has a birth and death time. Instead of representing a topological feature with a life bar as in persistence barcodes, we can simply represent it with a point in a plane; the x-coordinate of this point is the birth time of the topological feature, while the y-coordinate is its death time. All of our topological information is then displayed in a death-vs-birth plane, referred to as a persistence diagram. Certainly a topological feature cannot die before it is born, so all our points will lie above the diagonal line. We also agree that the top of the plane will represent infinite time, for those features that persist until infinity.

Consider a very simple persistence barcode as shown in Figure 2.3. The birth and death times of each life bar are read as (x, y) coordinates on the plane below. Observe that the barcode presents a component that persists until infinity. Thus, we define an "infinite death time" at the top of our diagram below

2.4.2 Bottleneck distance

For ease of exposition, we will describe the bottleneck distance in terms of the persistence diagrams rather than the persistence barcodes as they are equivalent. Intuitively, the bottleneck distance between two diagrams measures how much change the first sample must undergo so that its resulting persistence diagram matches the diagram of the second sample. More formally, think

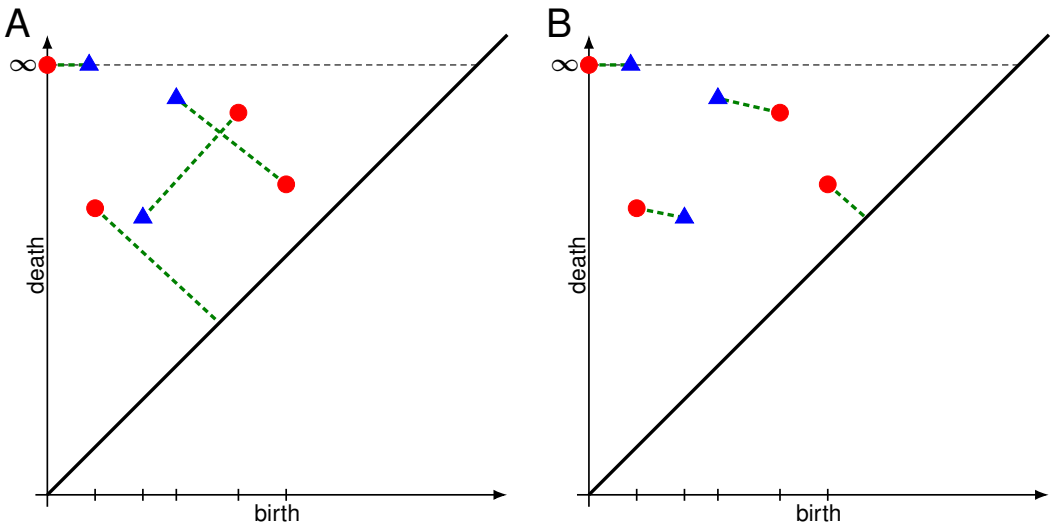

Figure 2.7: Computing the bottleneck distance between two persistence diagrams. (A) A possible pairing of points is suggested. Observe that it produces a large maximum distance between pairs. (B) An alternate pairing that yields a considerably smaller maximum distance between pairs.

of bottleneck distance as follows: we overlap the persistence diagrams of two samples, so both diagrams are actually on the same plane. Next, we are tasked to pair topological features between the diagrams. Every point from the first diagram must be either paired to an unmatched point from the second diagram, or matched with the diagonal. Given a pairing, we define its score as the maximum distance between pairs. The bottleneck distance is then defined as the minimum score when considering all possible pairings.

For example, consider the two different persistence diagrams drawn on top of each other in Figure 2.7, the first one being represented with red circles while the second with blue triangles. In Figure 2.7.A we pair each triangle with another circle, taking care to match the infinite triangle with the infinite circle. Observe that one circle is matched to the diagonal. The score of this pairing is the length of the longest green, dashed line. A different pairing is considered in Figure 2.7.B, which in turn produces a considerably smaller score as the green lines are all considerably shorter. After considering all possible pairings between diagrams, we realize that Figure 2.7.B is optimal in the sense that it produces the smallest score. The bottleneck distance between these barcodes is then this minimum score.

2.4.3 Branching architectures

Branching architecture is one example where bottleneck distance is useful. Traditional morphometric approaches fail to measure branching, despite it being a common architectural motif throughout life (Li et al., 2017). Li et al. (2019) measure the branching architecture of X-ray Computed Tomography (CT) scans of grapevine rachises, the branching stem structure that remains after removing the berries from the cluster. The filter they choose is geodesic distance of each voxel to the rachis base (Figure 2.5.D). The geodesic distance is the shortest distance between two vertices on a surface, in this case, the grapevine rachis itself. Starting with those voxels with the furthest geodesic distance from the base and filtering towards those closest, if zero-dimensional features are monitored, connected components at the tip of the branching structure are first born and then die as they merge at their parent node. Connected components continue to arise at branch tips and die at parent nodes in a hierarchical fashion. Each topological feature corresponds to a bar, and the record of merging can create a dendrogram that recapitulates the branching. There are a number of filters sensitive to branching that have been used in both plants and other organisms (Belchi et al., 2018; Bendich et al., 2010; Bendich et al., 2016; Kanari et al., 2018; Stolz et al., 2022).

Branching is an instance where calculating bottleneck distance might be preferred to Euler characteristic curves, because the topological features more directly correspond to the feature of interest (branches). The bottleneck distances of each grapevine rachis to the other create a metric space from which the rachises hierarchically cluster based on morphology (Li et al., 2019). Comparing morphological similarity to evolutionary history, rates of evolution along branches of the phylogenetic tree can be modeled. The morphological similarity matrix calculated from bottleneck distances can also be compared to traditional measurements (such as number of branches, median branch length and width, convex hull) and the ability to classify rachises from different species.

2.5 The structure of data: Mapper and biological networks

2.5.1 Mapper

Topological signatures and TDA outputs —barcodes, Euler characteristic curves, and bottleneck distances— measure the shape of data comprehensively but lack a correspondence to the original

data. This is known as the inverse problem: from data we can calculate a topological signature, but from a topological signature we cannot reproduce the original data. Biological data is noisy, and if the shape of the underlying structure in data could be visualized, individual data points that contribute to the overall shape of data could be isolated and studied in detail. For this reason, we now turn our attention to the mapper graph, which does provide some information in the reverse direction. By delimiting an underlying structure to our data and assigning correspondence of data points to this structure, complex and noisy datasets are simplified in a way similar to data reduction techniques.

Mapper is a tool from TDA that skeletonizes and summarizes the shape and structure of data as a graph Singh et al. (2007). The mapper algorithm is comprised of three main steps. 1) We choose a filter function on the data (Figure 2.8.A), this time associated to vertices rather than edges, and project all data points onto a line according to their filter function values (Figure 2.8.B). 2) Next, we split the real line into a fixed number of bins called covers. Each cover is an interval over the filter and, additionally, there is overlap between the covers. 3) Finally, we cluster the original data points in each of these bins to form graph vertices. Edges are drawn between nodes in the mapper graph if two clusters share some data points (Figure 2.8.C).

Imagine a 3D point cloud of data shaped like your hand, where the filter function is the distance of each data point to your wrist Lum et al. (2013) (Figure 2.5). If we created overlapping intervals, or covers, along this axis, then points at the fingertips would each form a vertex, and points towards the base of the fingers would form their own vertices as well. Because there is overlap between the covers, then vertices along each finger, but not between fingers, would share points, and we would draw in edges between these groups of points that would recapitulate the structure of fingers. The most proximal finger vertices would converge with vertices representing the palm, as well as vertices of the thumb. In the case of a hand, it is easy to see how a mapper graph summarizes and recapitulates the structure of the actual data. When applied to real world data, such as volumetric images like an X-ray CT scan, mapper graphs recapitulate shape in intuitive ways Chitwood et al. (2019).

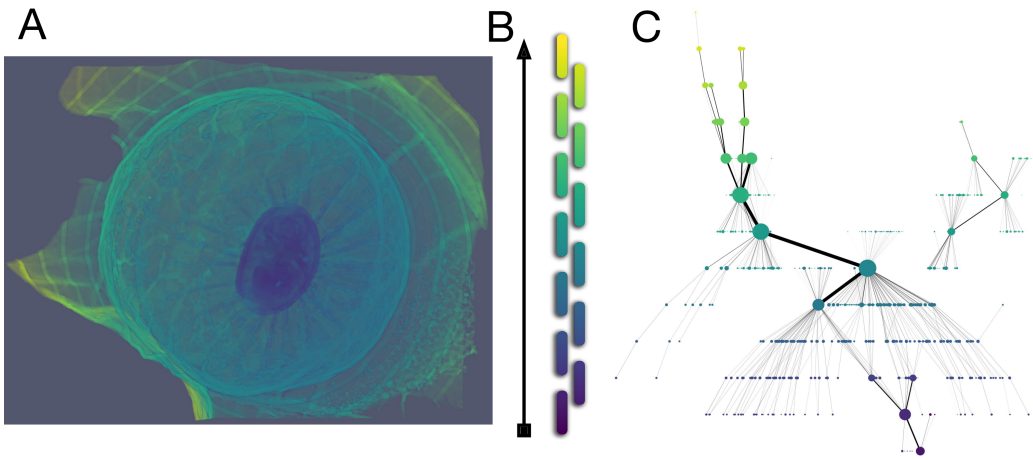

Figure 2.8: An example of mapper graphs. (A) X-Ray CT scan of a gall filtered by distance from the center. (B) These filter values are projected to a real line. The real line is then covered by a collection of overlapping intervals. For each interval, we then form different clusters of voxels whose filter value is in such interval. These clusters then yield the vertices and edges of (C) a mapper graph. Formally, the vertices are connected components within a certain range of radius from the center and edges correspond to overlap. Size of vertices and edges corresponds to the size of the component or overlap.

Let us consider a voxel-based X-ray CT scan of a gall, a swollen plant growth induced by an insect for its own benefit (Figure 2.8.A). Each voxel is a data point that takes on the value of the filter function, which in this case is its distance from the center of the gall. The Mapper algorithm clusters the data into vertices based on their filter function value (Figure 2.8.B), and if two vertices share some voxels between them (based on the cover intervals assigned and the physical location of the voxels), then they are connected by an edge. Bigger vertices in the mapper graph (Figure 2.8.C) correspond to a larger number of clustered voxels. Thicker edges correspond to a larger number of voxels in the bin overlap. The color of the vertices corresponds to the average filter function values of its voxel members. At the bottom of the graph, we can see a purple cluster corresponding to the core of the gall. As we move up, we eventually find larger turquoise vertices corresponding to the outer layers of the gall. Notice the small vertices that stem from these large turquoise vertices which represent the vasculature of the gall. Finally, as we reach the top of the mapper graph, we find green and yellow vertices that represent the leaf. From this example, two important features of mapper can be seen: its ability to serve as a data reduction technique that summarizes structure and the correspondence of the graph to the original data.

2.5.2 Biological networks

We have focused on Euclidean distances up until this point. But just like genetic distances can be used to create metric spaces to study evolution, other distance metrics can be used to create graphs that can be studied with TDA as well. Nicolau et al. (2011) used mapper to identify breast cancer subtypes using gene expression microarray data. The filter they use decomposes their data into separate normal and disease components. The resulting mapper graph reveals three distinct arms that, upon subsequent analysis, reveal a distinct genetic subtype of tumors. The architecture of the mapper graph corresponds to disease progression and its vertices to the expression of genes linked to breast cancer subtypes. By choosing an appropriate filter, the mapper graph reveals a structure of the data that might have been missed otherwise and is linked to prognosis. A recent study by Jeitziner et al. (2019) presented a new two-tiered version of the Mapper algorithm, which is particularly useful for small genomic sample sizes. Topological approaches to RNA-sequencing (RNA-seq) data have also been applied to study the in vitro differentiation of murine embryonic stem cells into neurons (Rizvi et al., 2017), specific genomic differences across various lung cancer patients (Amézquita et al., 2022), and to link gene expression to morphological outcomes across all flowering plants (Palande et al., 2022). All this suggests the translatability of Mapper-based topological analysis to many biological contexts.

2.6 A word on statistical caution

Most of the time, our data is subject to different kind of errors and we must address the statistical robustness of our topological signals. One foundational result by Cohen-Steiner et al. (2007) is the stability of persistence diagrams with respect to the bottleneck distance. Intuitively, this stability result implies that if all our data points wiggle only a little bit (possibly due to noise), then the resulting points in the persistence diagram will only wiggle a little bit as well. We must be careful with outliers though, since a single outlier can significantly alter our persistence diagram (Figure 2.2). Nonetheless, there has been a number of ideas to address this lack of robustness with respect to outliers, such as using a distance to measure (Chazal et al., 2017) or multi-parameter persistence (Lesnick and Wright, 2015, 2022). Intuitively, since an outlier is distant from every other point, it

will lie in a low density region, so we then proceed to discard such regions. Alternatives exploit bootstrap-like ideas, where the original point cloud is resampled and its topological signature computed numerous times (Reani and Bobrowski, 2022). Intuively, a single, isolated, outlying point will very rarely appear in resamples, so its distorting effect will be a rare event in the long run.

It is worth to warn that the space of all possible persistence diagrams is a mathematically complicated space to work with. For instance, given a collection of persistence diagrams, there might not be a unique "mean diagram" (Mileyko et al., 2011). The space of persistence diagrams presents a number of difficulties to define p-values, or confidence intervals, which are crucial in any statistical analysis. However, there has been a growing number of ideas and research to address such pitfalls, such as modifying the bottleneck distance to explicitly construct "mean diagrams" (Munch et al., 2015; Turner et al., 2014), adapting randomized null hypothesis tests (Robinson and Turner, 2017), or defining a confidence interval line along the diagonal of the diagrams (Fasy et al., 2014). Other alternatives include transforming diagrams to a simpler and more sound space, where the usual statistics, parameter estimation and hypothesis testing can be carried out as usual. This can be done with persistence landscapes (Bubenik, 2015), persistence images (Adams et al., 2017), or tent functions (Perea et al., 2022; Tymochko et al., 2019) to name a few examples.

Another caution to make is the interpretability of topological signatures. While summaries as persistence landscapes and ECCs are powerful when combined with machine learning techniques, it is hard to directly identify phenotypes from them. For instance, it is difficult to deduce seed's length, height and width based solely on the ECCs from Figure 2.6. Turner et al. (2014) mathematically prove that the collection of all ECCs corresponding to all possible directions effectively summarizes all the morphological information for 3D and 2D shapes. Moreover, with such collection we would be able to reconstruct the original object. Nonetheless, in practice we cannot consider an infinite number of directions. A finite bound on the number of necessary directions for general 3D shapes has been proven (Curry et al., 2022), although the idea of efficiently reconstructing large objects solely from ECCs remains elusive (Betthauser, 2018; Fasy et al., 2019; Micka, 2020).

2.7 Conclusion

We have seen how given data, a summary of the topological shape and structure of the space can be computed. For instance, data could come as a metric space of any distance—whether Euclidean, geodesic, genetic, functional, or correlative—and we can return a Vietoris-Rips complex and corresponding persistence barcode, which measure the shape of our data. By monitoring connected components, loops, or higher dimensional features, the barcode captures shape comprehensively, by monitoring the evolution of these features as a function of the filter. Such a framework has been used to measure the shapes of proteins, model evolution, and classify tissue architecture. The filter that we choose is arbitrary: it is merely a lens through which we can view relationships between our data points. The ability to choose a filter tailored to the hypothesis at hand is what confers the versatility of TDA to measure the shape of nearly any dataset, often in multiple ways. Gaussian density estimators applied to the pixels defining leaf outlines measures shape, allowing the genetic basis of the plant form to be studied. Geodesic distance captures the branching patterns of grapevine clusters, permitting the analysis of their evolution and modeling of berry development. We can analyze and compare the most persistent features in our barcodes, summarize them using the Euler characteristic, or truly calculate the overall topological similarity between barcodes using bottleneck distance. Using mapper, we can summarize the structure of data as a graph, and upon visualizing nodes of interest, identify the data points —whether voxels of an X-ray CT scan or nodes corresponding to gene expression— for further study and interpretation.

The promise of the application of TDA to biology is still in its infancy. Unlike any other method in biology, TDA provides a way to measure topological features and shape in a comprehensive way. The versatility of filter function selection allows TDA to be applied to any number of datasets across sub-disciplines: structural biology, evolution, molecular biology, medicine, neuroscience, and developmental biology. The methods described here can be applied to higher dimensional datasets that are dynamic or evolve over time (Myers et al., 2019; Perea, 2019; Topaz et al., 2015; Tymochko et al., 2020), easily accommodating biological complexity. Regardless of data size, complexity, or dimensionality, TDA provides concise summaries of the information content of any

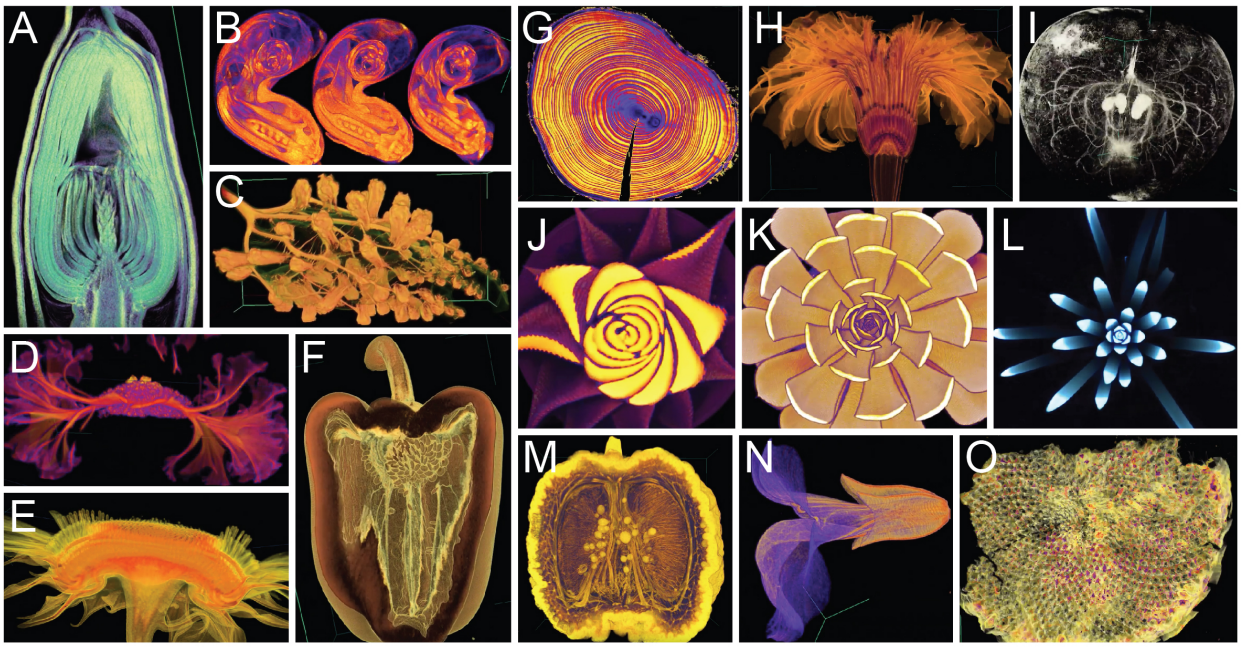

Figure 2.9: Endless forms most beautiful. X-ray Computed Tomography (CT) scans of biological specimens showing the diversity of morphology in the natural world. (A) Magnolia bud, (B) bean flowers, (C) grapevine leaf with phylloxera galls, (D) the fasciated meristem of a velvet flower, (E) side view of a sunflower disc, (F) bell pepper, (G) treerings, (H) marigold flower, (I) vasculature within an apple, (J) Haworthia, (K) Echeveria, (L) Agave hybrid, (M) citrus fruit, (N) monkeyflower, (O) archaeological sunflower disc specimen.

dataset from the perspective of shape and structure. Given the spectacular diversity of form across biology (Figure 2.9), a method like TDA, that can be customized to measure shape using a tailored filter function, will allow previously unstudied phenomena to be analyzed from the perspective of shape. The vision of TDA, that data is shape and shape is data, will be relevant as biology transitions into a data-driven era where meaningful interpretation of large datasets is a limiting factor.

CHAPTER 3

THE SHAPE OF BARLEY

> Go and measure for me the barley which is in the storehouse, that which remains from last year's barley." Then he set out for her six measures of barley. Then the peasant said to his wife, "Behold, there are twenty measures of barley as food for you and your children. Now make these six measures of barley into bread and beer for me as daily rations, that I may live on them.
>
> —from *The Tale of the Eloquent Peasant*
>
> Anonymous tale from the Egyptian Middle Kingdom
>
> (1991–1786 BC)

There is a discrepancy between the information embedded in biological forms that we can discern with our senses versus that which we can quantify. Methods to comprehensively quantify phenotype are not commensurate with the thoroughness and speed with which genomes can be sequenced. High-throughput phenotyping has enabled us to collect large amounts of phenotyping data (Andrade-Sanchez et al., 2013; Araus and Cairns, 2014; Tanabata et al., 2012); nonetheless, we are not maximizing the information extracted from the data we collect.

To extract, compare, and analyze this information embedded in a robust and concise way, we turn to Topological Data Analysis (TDA), specifically the Euler Characteristic Transform (ECT). Here we show the use of ECTs to correctly summarize the shape of barley seeds as a proof of concept. We scanned a collection of barley panicles comprising 28 different accessions with X-ray CT technology at 127 micron resolution. These scans were later digitally processed to isolate 3121 individual grains, and their morphology was quantified using both traditional and topological shape descriptors. We then explored both qualitatively and quantitatively the descriptiveness of these measurements. To aid both assessments, we used KPCA and UMAP separately to aggressively reduce the dimension of the traditional and ECT vectors. We observe that traditional shape descriptors tend to cluster seeds based on their accession, while KPCA-reduced topological shape

descriptors tend to cluster them based on panicles. UMAP-reduced topological descriptors balance both approaches and draw shape distinctions at both accession and spike level. This in turn shows that KPCA and UMAP draw from different pieces of ECT information. This observation suggests that the ECT effectively summarizes both spike-specific and accession-specific morphological information which can be then highlighted with an appropriate dimension reduction technique. To quantify the descriptor correctness, we trained a support vector machine (SVM) to determine the accession of individual grains based on their shape alone. Our experiments show that SVMs perform better whenever topological information is taken into account, which suggests that the ECT measures shape that is "hidden" from traditional shape descriptors.

This chapter is derived from the original research paper

- **E.J. Amézquita, M.Y. Quigley, T. Ophelders, J.B. Landis, D. Koenig, E. Munch, D.H. Chitwood** (2021). Measuring hidden phenotype: Quantifying the shape of barley seeds using the Euler Characteristic Transform. *in Silico Plants* 4(1): diab033.

3.1 Introduction

Topological Data Analysis (TDA) is a set of tools that arise from the perspective that all data has shape and that shape is data (Amézquita et al., 2020; Lum et al., 2013; Munch, 2017). TDA treats the data as if made of elementary building blocks: points, edges, squares, and cubes, referred to as 0-, 1-, 2-, and 3-dimensional *cells* respectively (Figure 3.1.A). A collection of cells is referred to as a *cubical complex*, or complex, for short.

Cubical complexes are both a natural and consistent way to represent image data (Kovalevsky, 1989). Given a grayscale image, we follow a strategy similar to Wagner et al. (2012) to construct a cubical complex: a nonzero pixel will correspond to a vertex in our complex. If two pixels are adjacent —in the 4-neighborhood sense— we say that there is an edge between the corresponding vertices in the complex. If 4 pixels in the image form a 2×2 square, we will consider a square in our complex between the corresponding 4 vertices (Figure3.1.A). Additionally, for the 3D image case, if 8 voxels —the 3D equivalent of pixels— make a $2 \times 2 \times 2$ cube, we will draw a cube in our

complex between the corresponding 8 vertices.

TDA seeks to describe the shape of our data based on the number of relevant topological features found in the corresponding complex. For instance, the complex in Figure 3.1.A has two distinct, separate pieces colored in blue and red respectively, formally referred to as *connected components*. This complex also has 8 edges forming the outline of a square without an actual red block filling it —edges thickened for emphasis— this is referred to as a *loop*. In higher dimensions, we could also consider hollow blocks containing *voids*. We can even go a step further and summarize these topological features with a single value known as the *Euler characteristic*, represented by the Greek letter χ, defined for voxel-based images as

$$\chi = \#(\text{connected components}) - \#(\text{loops}) + \#(\text{voids}).$$

The Euler characteristic is a topological invariant; that is, it will remain unchanged under any smooth transformation applied to our shape. The well-known but surprising Euler-Poincaré formula states that χ can be computed easily as

$$\chi = \#(\text{Vertices}) - \#(\text{Edges}) + \#(\text{Faces}) - \#(\text{Cubes}).$$

This equivalence can be seen in the cubical complex in Figure 3.1.A, where

$$\begin{aligned}\chi &= 20 \text{ vertices} - 22 \text{ edges} + 3 \text{ faces} \\ &= 2 \text{ connected components} - 1 \text{ loop} + 0 \text{ voids} = 1.\end{aligned}$$

The Euler characteristic by itself might be too simple. Nonetheless, we can extract more information out of our data-based complex if we think of it as a dynamic object that grows in number of vertices, edges, and faces across time. As our complex grows, we may observe significant changes in χ. The changes in χ can be thought as a topological signature of the shape, referred to as an *Euler characteristic curve (ECC)*. The growth of the complex is defined by a *filter function* which assigns a real number value to each voxel. For reasons discussed later, we will focus on directional filters which assign to each voxel its height as if measured from a fixed direction.

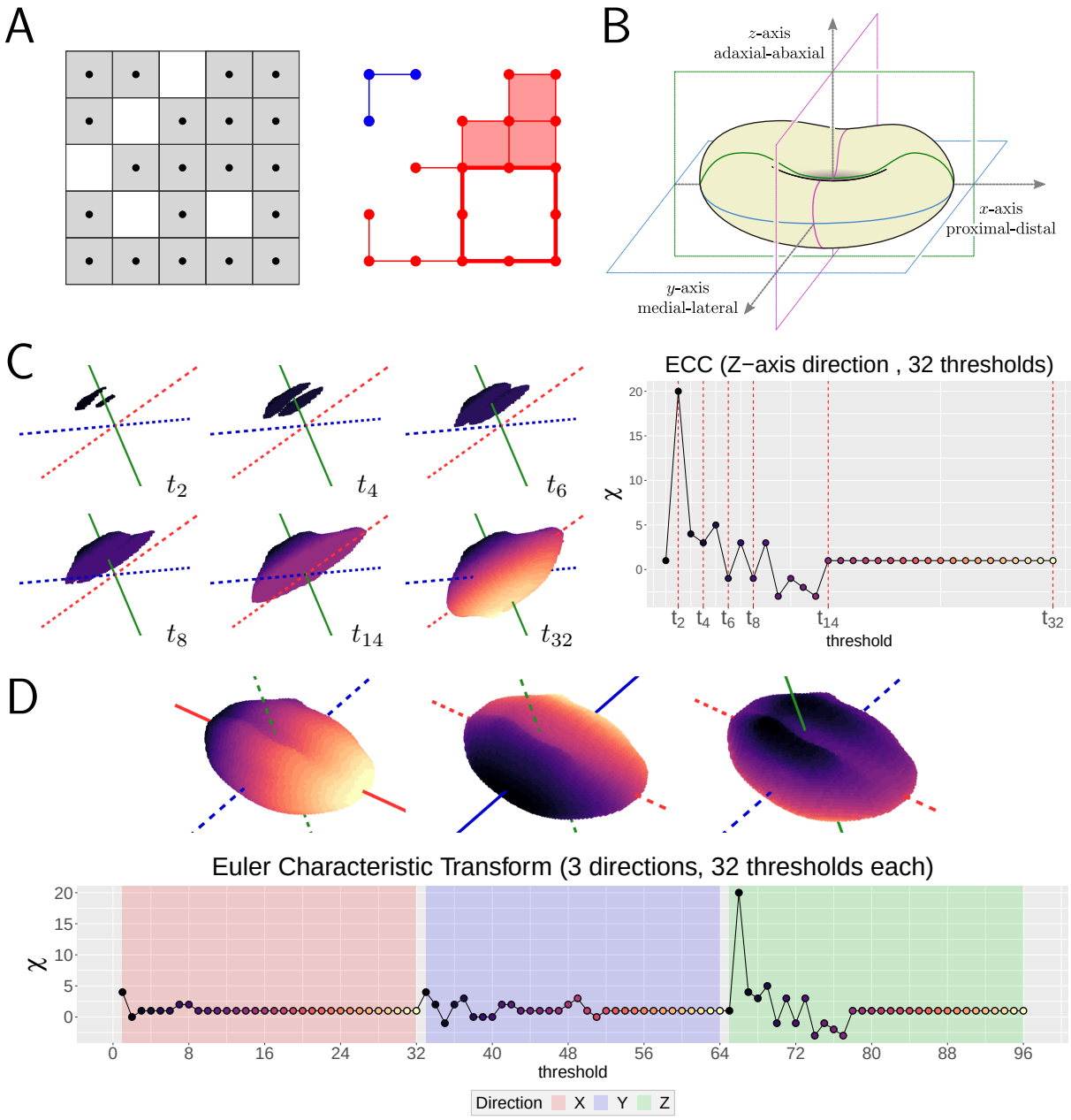

Figure 3.1: Extracting topological shape signatures from barley seeds. (A) A binary image (left) is treated as a cubical complex (right). This cubical complex has 2 connected components, 1 loop, 0 voids. The distinct connected components are colored in blue and red respectively. The loop is emphasized with thicker edges. (B) The barley seeds were aligned so that their proximal-distal, medial-lateral, and adaxial-abaxial axes corresponds to the X, Y, Z-axes in space. (C) Example of an Euler Characteristic Curve (ECC) as we filter the barley seed across the adaxial-abaxial axis (depicted as a solid, green line) through 32 equispaced thresholds. (D) The Euler Characteristic Transform (ECT) consists of concatenating all the ECCs corresponding to all possible directions. In this example, we concatenate 3 ECCs corresponding to the X, Y, Z directions, represented by the solid lines respectively.

As an example, consider the cubical complex of a barley seed and the direction corresponding to the adaxial-abaxial axis (Figure 3.1.B). Voxels at the top of the seed will be assigned the lowest values, while voxels at the bottom will obtain the highest values. We then consider 32 equispaced, increasing thresholds $t_1 < t_2 < \ldots < t_{32}$ which define 32 different slices of equal thickness along the adaxial-abaxial axis. We start by computing the Euler characteristic of the first slice, that is, all the voxels with filter value less than t_1. Next we aggregate the second slice, which are all the voxels with filter value less than t_2, and recompute the Euler characteristic. We repeat the procedure for the 32 slices. For instance in Figure 3.1.C, we observe that we started with scattered voxels which are thought of as many connected components which may explain the high Euler characteristic values. As we keep adding slices, we connect most of the stray voxels into fewer but larger connected components, and simultaneously, we might have created loops as seen in t_4 and t_6. This merging of connected components, and formation and closing of loops might explain the fluctuation of the Euler characteristic between positive and negative values. Finally, after more than half of the slices have been considered, at t_{14}, we observe that no new loops are formed, and every new voxel will simply be part of the single connected component. Thus, the Euler characteristic remains constant at 1. The ECC is precisely the sequence of different Euler characteristic values as we add systematically individual slices along the chosen direction.

To get a better sense of how the Euler characteristic changes overall, we can compute several ECCs corresponding to different directional filters. For example, in Figure 3.1.D we choose three directions in total corresponding to the proximal-distal, medial-lateral, and adaxial-abaxial axes respectively. Each filter produces an individual ECC, which we later concatenate into a unique large signal known as the *Euler Characteristic Transform (ECT).*

There are two important reasons to use ECT over other TDA techniques. First, the ECT is computationally inexpensive, since it is based on successive computations of the Euler characteristic, which is simply an alternating sum of counts of cells. This inexpensiveness is especially relevant as we are dealing with thousands of extremely high-resolution 3D images. Assuming that we have already treated the image as a cubical complex, we can compute a single ECC in linear

time with respect to the number of voxels in the image (Richardson and Werman, 2014). We can thus compute the ECT of a 50,000-voxel seed scan with 150 directions in less than two seconds on a traditional PC. The second reason to use the ECT is its provable invertibility and statistical sufficiency. As first proved by Turner et al. (2014), and later extended separately by Curry et al. (2022) and Ghrist et al. (2018), if we compute all possible directional filters we would have sufficient information to reconstruct the original shape. Moreover, this ECT is a sufficient statistic that effectively summarizes all information regarding shape. Although there are infinite possible directional filters, there is ongoing research into defining a sufficient finite number of directions such that we can effectively reconstruct shapes based solely on their finite ECT (Belton et al., 2020; Betthauser, 2018; Curry et al., 2022; Fasy et al., 2019). Nonetheless, a computationally efficient reconstruction procedure for large 3D images remains elusive.

Another computational consideration is the fact that the ECT produces a vector of topological information of #(directions) $\times$ #(thresholds) dimensions, which is usually above 2000 dimensions. In general, high-dimensional vectors tend to produce distorted prediction and regression results (Köppen, 2000), and it is advised to denoise and summarize these vectors by using different dimension reduction techniques. One such standard technique is principal component analysis (PCA), which seeks to project the high-dimensional vectors unto the orthogonal directions that capture the greatest variability of the data. These linear directions are referred to as the *principal components* of the data. Sometimes, the data cannot be properly summarized as a collection of lines. A more flexible approach is to consider *kernel PCA (KPCA)* (Schölkopf et al., 1998), a *nonlinear* alternative. By specifying a *kernel function*, we can instead project the high-dimensional samples unto the polynomial, trigonometric, or circular curves that capture the most variance of the data. A completely different dimension reduction strategy is the *uniform manifold approximation and projection (UMAP)* (McInnes et al., 2020), which also draws several ideas from TDA. Intuitively, UMAP seeks to project the high-dimensional data unto a low-dimensional space while preserving the most prominent topological local features. That is, if the original data contains large connected components, wide loops, and ample voids, its low-dimensional UMAP projection should also

exhibit several connected components, loops, and voids. If two sample points are in the same connected component in the high-dimensional space, these two should remain in the same cluster when projected to the low-dimensional space.

3.2 Materials and Methods

We selected 28 barley accessions with diverse spike morphologies and geographical origins for our analysis (Harlan and Martini, 1929, 1936, 1940). In November of 2016, seeds from each accession were stratified at 4C on wet paper towels for a week, and germinated on the bench at room temperature. Four day old seedlings were transferred into pots in triplicate and arranged in a completely randomized design in a greenhouse. Day length was extended throughout the experiment using artificial lighting —minimum 16h light / 8h dark. After the plants reached maturity and dried, a single spike was collected from each replicate for scanning at Michigan State University. The scans were produced using the North Star Imaging X3000 system and the included efX software, with 720 projections per scan, with 3 frames averaged per projection. The data was obtained in continuous mode. The X-ray source was set to a voltage of 75 kV, current of 100 µA, and focal spot size of 7.5 µm. The 3D reconstruction of the spikes was computed with the efX-CT software, obtaining a final voxel size of 127 µm. The intensity values for all raw reconstructions was standardized as a first step to guarantee that the air and the barley material had the same density values across all scans. Next, the air and debris were thresholded out, and awns digitally pruned (Figures 3.2.A-D). Finally, the seed coat of the caryopses was digitally removed, leaving only the embryo and endosperm due to their high water content (Figure 3.2.E). We did not have enough resolution in the raw scans to distinguish clearly the endosperm from the embryo. Hereafter, we will refer to these embryo-endosperm unions simply as seeds. Thus, we digitally isolated all the seeds and obtained a collection of 3438 seeds in total. Due to the large volume of data, we used an in-house scipy-based python script to automate the image processing pipeline for all panicles and grains.

To make the collection of different directional filters comparable across seeds, all the seeds were aligned with respect to their first three principal components. Since all the seeds are oblong in

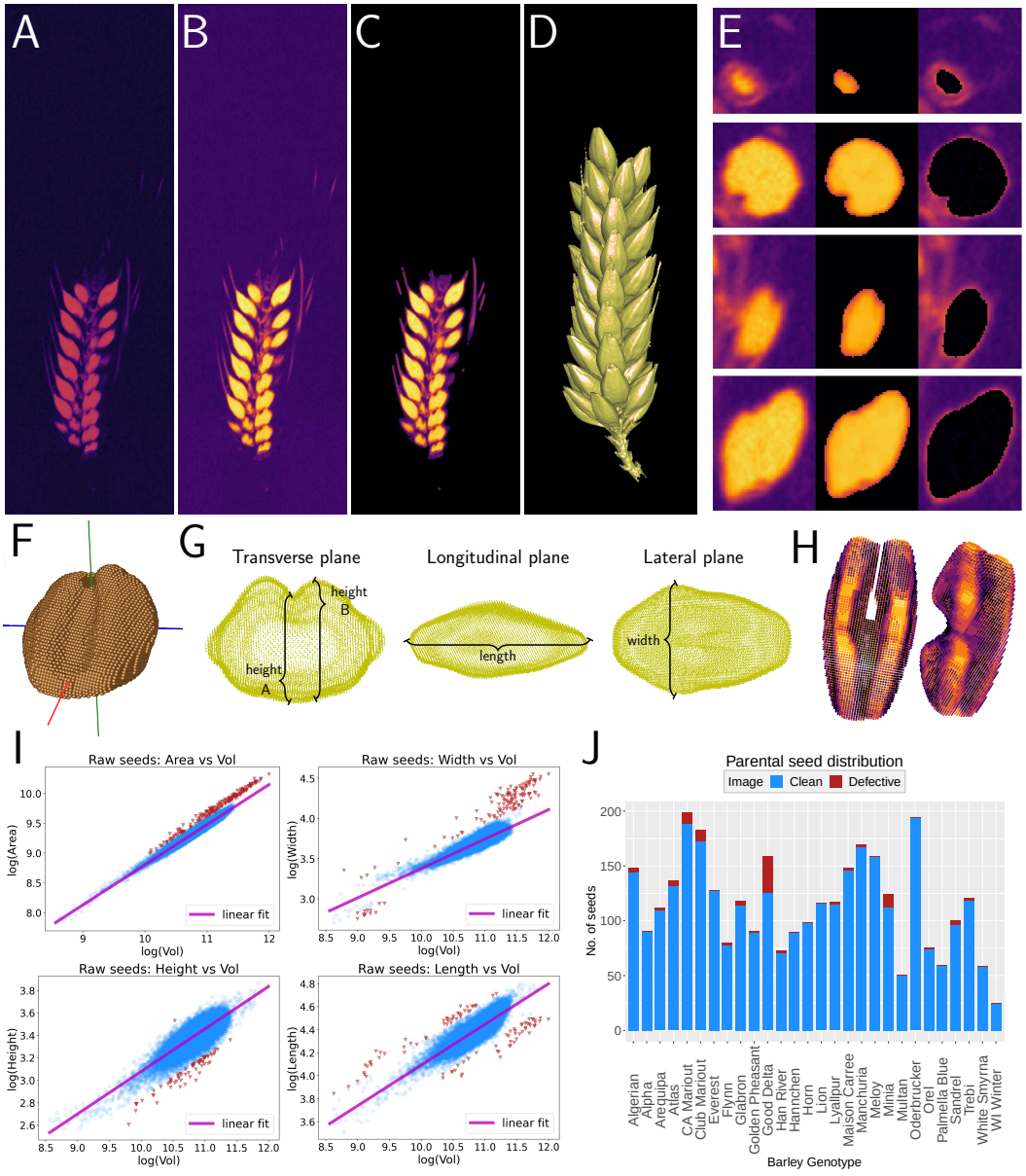

Figure 3.2: Barley image processing. The morphology measurements were extracted from 3D voxel-based images of the barley panicles. Before any analysis was done, the (A) raw X-ray CT scans of the panicles had their (B) densities normalized, (C) air and other debris removed, and awns pruned. (D) After automating these image processing steps, we could finally work with a large collection of clean, 3D panicles. (E) An extra digital step segmented the individual seeds —embryo and endosperm— for each barley spike. The left shows the original raw scan, the center shows the isolated seed, while the right side shows part of the coat that was removed while segmenting. (F) The seeds were aligned according to their principal components, which allowed us to (G) measure a number of traditional shape descriptors. (H) Incomplete or broken seeds were later removed from the data set. (I) These defective seeds were identified by manually examining the outliers of different allometry plots. Outliers depicted as red triangles. (J) The total number of clean and defective seeds measured from each accession. Defective seeds were not concentrated in a particular accession.

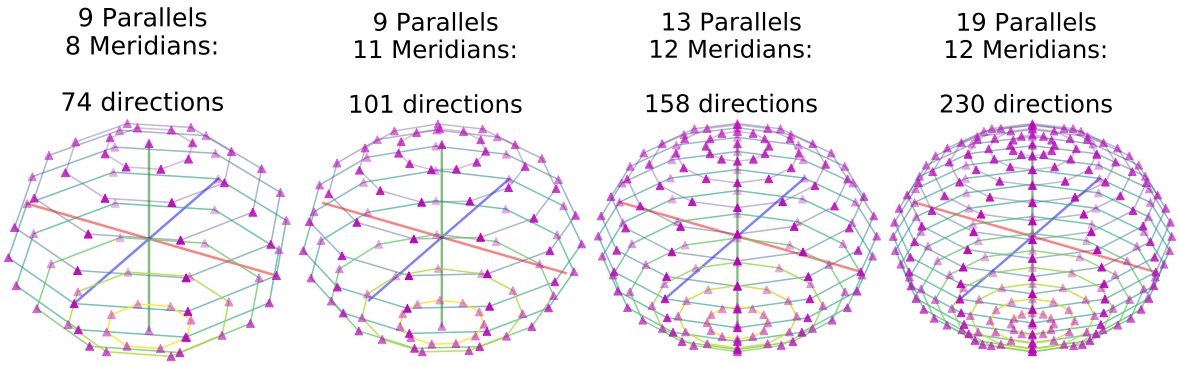

Figure 3.3: Directions chosen to compute the ECT. The sphere was split into a equispaced fixed number of parallels and meridians in each case. The directions were the taken from the intersections.

Table 3.1: Sample size of seed scans used for each individual accession. **N** equals the number of panicles from which seeds are derived.

Accession	**N**	**seeds**	**Accession**	**N**	**seeds**	**Accession**	**N**	**seeds**
Algerian	3	144	Golden Pheasant	3	89	Minia	3	112
Alpha	3	90	Good Delta	3	126	Multan	1	50
Arequipa	3	110	Han River	2	71	Oderbrucker	3	194
Atlas	3	132	Hannchen	3	89	Orel	3	74
California Mariout	3	189	Horn	3	98	Palmella Blue	3	59
Club Mariout	3	173	Lion	3	116	Sandrel	2	96
Everest	3	128	Lyallpur	3	115	Trebi	2	119
Flynn	3	78	Maison Carree	3	146	White Smyrna	3	58
Glabron	3	114	Manchuria	3	167	Wisconsin Winter	1	25
			Meloy	3	159			
TOTAL	83	3121	mean		111.5	standard dev.		42.2

shape, this PCA-based alignment corresponds to the proximal-distal, medial-lateral, and adaxial-abaxial axes respectively (Figures 3.1.B, 3.2.F). The orientation of the principal components is arbitrary with every run, so we did keep track of the crease and the tip of seed and flipped the axes accordingly so that the tip would always be located as the rightmost point of the image and the crease would always point north. With this uniform alignment we were able to measure the length, width, heights, surface area and volume of each seed (Figure 3.2.G). We also computed the convex hull for each seed and measured its surface area and volume, as well as the ratios with respect to seed surface area and volume. In total, 11 different traditional shape descriptors were measured. Damaged and incomplete seeds (Figure 3.2.H) were removed by evaluating allometry

plots along their best linear fits and residuals (Figure 3.2.I). Points with residuals 4 times larger than the standard deviation were deemed as outliers and the associated seed was manually examined further. Outliers usually corresponded to either defective seeds —which were discarded— or to a cluster of seeds that failed to be individually segmented. In the latter case, we repeated our image processing scripts with more aggressive parameters to segment the seeds and re-examined the result. A final visual assessment of the remaining images was conducted to ensure the removal of all damaged seeds. These outliers did not represent a significant portion of the seeds of any accession (Figure 3.2.J). In total we obtained 3121 cleanly segmented seeds. Every accession is represented on average by 111 seeds, with ±42 seeds as standard deviation. All the accession numbers are within 2 standard deviations from this empirical mean (Figure A.1; Table 3.1).

As a proof of concept, we explored how topological descriptors varied as we varied both the number of different directions and the number of uniformly spaced thresholds. In total, for every seed we computed the ECT considering 74, 101, 158, and 230 different directions. We emphasized directions toward the seed's crease, which correspond to directions close to both north and south poles (Figure 3.1.B; Figure 3.3). For each direction, we produced ECCs with 4, 8, 16, 32, and 64 thresholds.

Recall that the ECT is a record of how topology changes at every single slice taken at every direction (Figure 3.1.C). We performed Kruskal-Wallis one-way analyses (Kruskal and Wallis, 1952) to determine if the Euler characteristic inter-accession variance was significantly different from the intra-accession variance at a particular slice and direction. This way, we observed which parts of the seed anatomy were of particular relevance to the ECT. Accessions and individual spikes were both considered as possible classes when performing the Kruskal-Wallis tests. These results follow a conservative 10^{-10} false discovery rate after considering a multiple test Benjamini-Hochberg correction (Benjamini and Hochberg, 1995).

For every seed we computed a very high-dimensional vector of topological information, usually above 2000 dimensions, which were later reduced in dimension independently with KPCA and UMAP to prevent high-dimensionality distortions. A non-linear KPCA with a $\sigma = 1$ Laplacian

kernel reduced the ECT dimension based on its largest source of variance. UMAP on the other hand was used to preserve the prominent, high-dimensional topological features of the ECT in an unsupervised fashion. We fixed the use of 50 nearest neighbors, 0.1 minimum distance, and Manhattan distance as the rest of key UMAP hyperparameters. For all dimension reduction techniques, the ECT dimension was reduced to just 2, 3, 6, 12, and 24 dimensions. We focused on an aggressive 2-dimensional reduction for visualization purposes both with KPCA and UMAP.

To evaluate the descriptiveness, we trained three non-linear support vector machines (SVM) with radial kernel $\sigma = 0.1$ (Burges, 1998) to characterize and predict the seeds from 28 different accessions based on three different collections of descriptors: traditional, topological, and combining both traditional and topological descriptors. In every case, the descriptors were centered and scaled to variance 1 prior to classification. Given that SVM is a supervised learning method, we partitioned our data into training and testing sets. In our case, we randomly sampled 75% of the seeds from every accession as our training data set, labeled according to their accession. The remaining 25% was used to test the accuracy of our prediction model. We repeated this SVM setup 100 times and considered the average accuracy and confusion matrices as final results. This was done for all possible combinations of directions, thresholds, and dimensionality reductions mentioned above. The SVM was our classifier of choice since it is quick to train and it does not require vast amounts of training data to produce reasonable results.

3.3 Results

Topological and combined shape descriptors tend to produce more accurate shape-based classification results, provided that the ECT is computed with sensible parameters and an adequate dimension reduction technique. The best SVM classification results were yielded by topological and combined shape descriptors based on a 2568-dimensional ECT —158 directions and 16 thresholds (Figure A.3). Based on the highest F_1 classification scores, these high-dimensional vectors were best parsed after being reduced to just 2 dimensions with KPCA, or to 12 dimensions with UMAP. Hereafter, the rest of topology-related results are based on these specific choice of directions, thresholds, and dimensionality reduction.

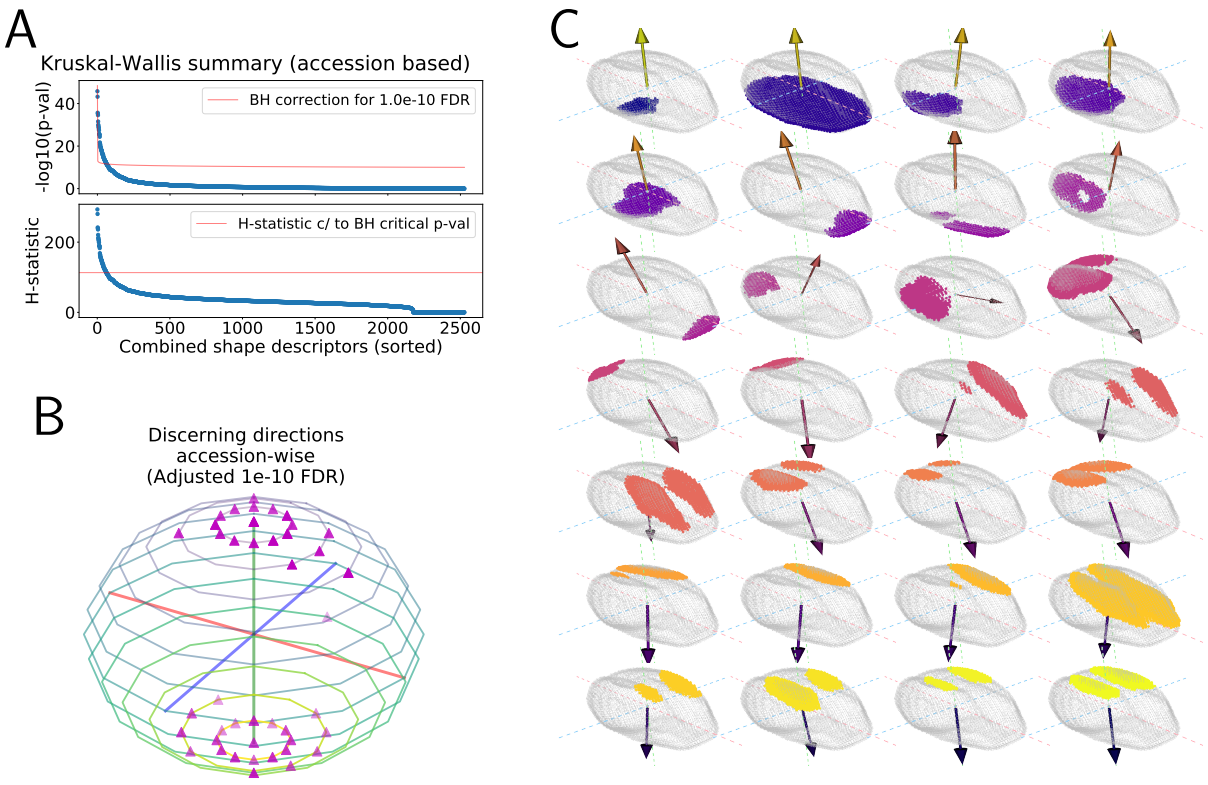

Figure 3.4: Relevant ECT directions and slices. (A) We examine the inter-accession and intra-accession variance differences of the Euler characteristic for each direction and threshold. A Kruskal-Wallis analysis combined with a Benjamini-Hochberg multiple test correction suggests a handful of particularly discerning slices across accessions. (B) These directions and thresholds are mostly concentrated around the poles, and (C) correspond to the seed's crease and bottom morphology. Colors bear no special meaning.

A Kruskal-Wallis one-way analysis of the ECT vectors, combined with a Benjamini-Hochberg correction admitting a 10^{-10} FDR, reveals 55 features that explain the most of inter-accession variance (Figure 3.4.A). The most accession-discerning slices and directions correspond to the north and south poles (Figure 3.4.B). As discussed in the seed alignment heuristics in the Methods section, these pole directions in turn correspond to the morphology of the crease and the bottom of the seed (Figure 3.4.C). Similar results were observed when analyzing for the most spike-discerning directions (Figure A.4). In other words, the topological shape descriptors do measure the crease and bottom shape of the seed, a morphological feature not explicitly measured by our traditional setting.

Turning back to the traditional shape descriptors, these share similar distributions across the 28

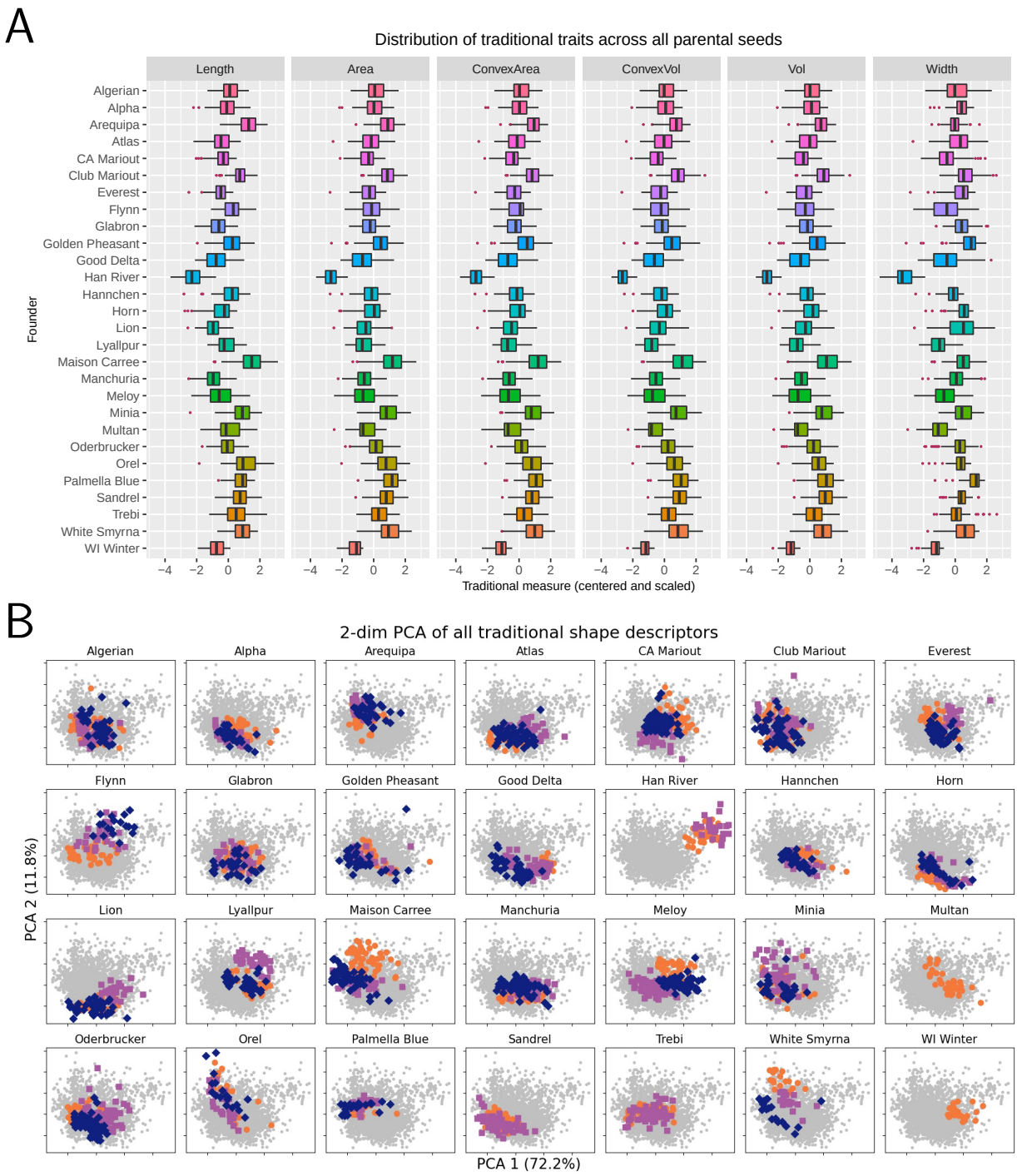

Figure 3.5: Distribution of traditional shape descriptors. (A) Distribution of six of the 11 traditional seed shape descriptors across the 3121 seeds. These measurements were first centered at 0 and scaled to have variance 1. (B) Plot of the first 2 principal components of the 11 shape descriptors. The first PC describes more than 70% of the total variance. Different marker and color indicate seeds from different spikes.

accessions, provided they are all centered and scaled to variance 1 (Figure 3.5.A). Kruskal-Wallis analyses suggest that the seed length, surface area, and volume related measures explain the most inter-accession variance (Figure A.5.A–B). Reducing the descriptors to a 2D representation with PCA suggests that these traditional descriptors tend to group the seeds based on their accession (Figure 3.5.B). These two components explain 84.0% of the total variance, with the first principal component explaining a considerable 72.2% alone. A similar grouping-by-accession behavior was observed whenever we reduced the traditional shape descriptors to 2 dimensions with UMAP instead. KPCA dimension reduction did not yield insightful results.

Topological shape descriptors on the other hand can provide a more spike-specific morphology encoding, depending on the dimension reduction technique used to parse the ECT. KPCA summarizes the topological information as a loop, with sharply defined clusters corresponding to seeds from individual spikes (Figure 3.6.A). On the other hand, the UMAP projection produces a large, round cluster. Notice that seeds of different spikes tend to lie on different locations, while these locations overlap partially for spikes of the same accession (Figure 3.6.B). This behavior suggests that UMAP dimension reduction tries to balance both spike-specific and accession-specific shape features.

Another round of Kruskal-Wallis analyses on the combined shape descriptors reinforce the idea that traditional descriptors cluster based on accession, KPCA-reduced topological descriptors do so based on spike, while UMAP-reduced ones provide a balanced clustering. The most inter-accession variance is explained predominantly by the traditional shape descriptors, with just a few topological features as complement (Figures A.5.A–B). However, most of the inter-spike variance is predominantly captured by the dimension-reduced topological descriptors. The first two KPCA components do explain most of this inter-spike variance, which agrees with the tight panicle clusters seen before (Figure 3.6; Figure A.5.C). On the other hand, UMAP distributes regularly the spike variance across most of its components, complemented by a few traditional shape descriptors (Figure A.5.D). In other words, traditional shape descriptors capture accession-specific shape features, KPCA highlights spike-specific features, and UMAP provides a balance between both of

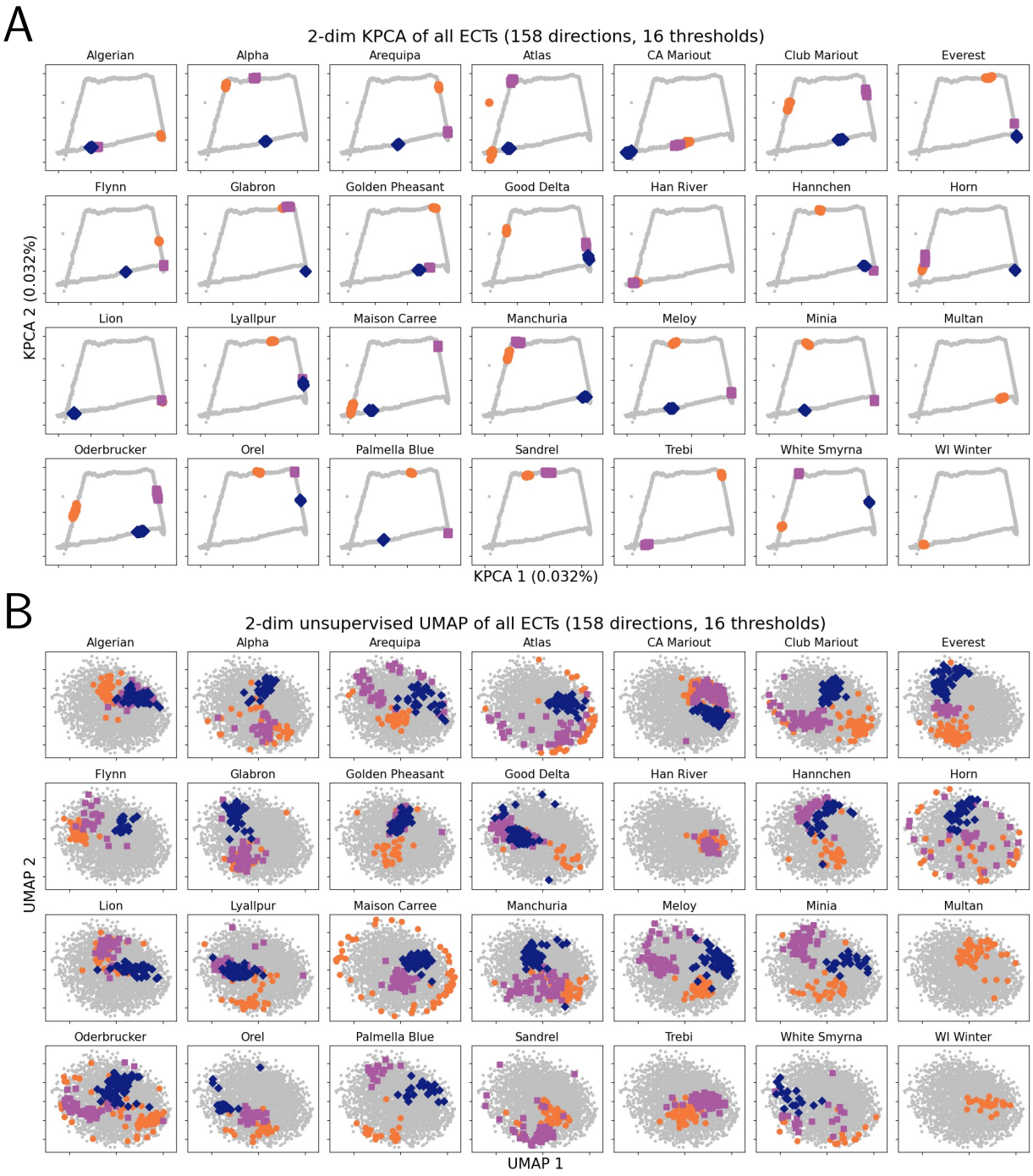

Figure 3.6: Dimension reduction of the ECT vectors. The ECT can produce a high-dimensional topological signature for each seed. To better visualize this topological information, we can reduce it to just two dimensions with (A) kernel PCA or (B) unsupervised UMAP. The seeds of individual accessions are highlighted in every frame. Different marker and color indicate seeds from different spikes.

them.

When evaluating quantitatively the descriptiveness of these cluster differences, we observed that topological shape descriptors are able to produce much better SVM classification results than traditional shape descriptors (Table 3.2). Using exclusively traditional descriptors, the machine is able to correctly determine the grain variety roughly 57% of the time. For comparison, by simply randomly guessing the variety, we would expect to be correct just $1/28 \times 100 \approx 4\%$ of the time. The classification could not be improved by reducing the dimension of the traditional vector (Figure A.2). If we use exclusively topological shape descriptors instead, the machine can classify different accessions with more than 75% accuracy. These results depend on the dimension reduction technique of choice (Figure A.3.A). We observe that KPCA provides a powerful 2-dimensional summary of the ECT, which later can be used to predict grain accession with 85% classification accuracy. This accuracy diminishes considerably as more nonlinear principal components are considered. This drop of classification performance can be offset by combining the KPCA summary with traditional shape descriptors, which keep the classification accuracy above 70% (Figure A.3.B).

The 2-dimensional UMAP summary (Figure 3.6.B) exhibits difficulties and discerning accessions, where classification accession does not go above 25%. Nonetheless, if a 12-dimension UMAP summary is considered, it is possible to classify accessions with 75% accuracy using exclusively topological information. Moreover, these UMAP-summary classification results can be further improved by combining them with traditional shape descriptors, where classification accuracy goes beyond 88%. The ECT thus captures important morphological patterns that can be complemented by size features which are provided by the traditional shape descriptors.

Additionally, for both KPCA and UMAP cases, a small p-value produced by Friedman tests (Friedman, 1937) suggests that the three SVM classifiers, corresponding to the three sets of shape descriptors, are statistically different. Since we are comparing only three classifiers at a time, we can rely better on a Quade post-hoc pairwise test (Quade, 1979) as suggested in (Conover, 1998) (Table 3.3).

Table 3.2: SVM classification accuracy of barley seeds from 28 different founding lines after 100 randomized training and testing sets. Since we are in a multi-class classification setting we first computed the precision, recall, and F_1 scores for each founding line. Later, we computed the weighted average for each score, where the weight depended on the number of test seeds for each of the barley lines. Observe that the use of either topological or combined descriptors outperforms the use of exclusively traditional descriptors.

Shape descriptors	Dimension reduction	No. of dims	Scores (weighted average ± standard deviation) Precision	Recall	F_1
Traditional	*	11	0.58 ± 0.050	0.58 ± 0.016	0.57 ± 0.016
Topological	KPCA	2	0.88 ± 0.031	0.87 ± 0.010	0.87 ± 0.011
Topological	UMAP	12	0.75 ± 0.047	0.75 ± 0.016	0.74 ± 0.016
Combined	KPCA	13	0.73 ± 0.052	0.72 ± 0.017	0.71 ± 0.017
Combined	UMAP	23	0.89 ± 0.028	0.89 ± 0.010	0.89 ± 0.010

Table 3.3: Small Friedman and Quade post-hoc p-values (using t-distribution approximation with Bonferroni correction) suggest that different descriptors produce statistically different SVM results.

ECT + KPCA			ECT + UMAP		
Friedman p-val		1.4×10^{-5}	Friedman p-val		4.4×10^{-10}
	Traditional	Topological		Traditional	Topological
Topological	1.8×10^{-11}	*	Topological	8.0×10^{-4}	*
Combined	5.9×10^{-5}	4.4×10^{-4}	Combined	4.4×10^{-13}	7.8×10^{-7}

3.4 Discussion

Traditional morphometrics has been used to reveal fundamental trends in morphological changes across space and time in ancient cereal grains (Bouby, 2001; Tanno and Willcox, 2012). Historical evidence shows that barley seeds became smaller as the crop moved from Mediterranean climates to Northwest Europe due to colder temperatures and higher sunlight variance, shedding insight on the timeline of barley domestication in Central Asia (Motuzaite Matuzeviciute et al., 2018). Similarly, grains became rounder and the spikes more compact as they moved to higher altitude sites in Nepal (Fuller and Weisskopf, 2014). Differences are more subtle if we compare accessions originating from similar regions and time periods. Geometric Morphometrics (GMM) has provided a more quantitative characterization of the grains. For example, GMM can successfully tell apart barley grains from einkorn (*Triticum monococcum*) and emmer (*T. dicoccum*) accessions (Bonhomme et al., 2017); it can be used to distinguish two-row vs six-row barley seeds (Ros et al., 2014); and it

can establish unique morphological characteristics of land races to deduce their possible historical origins (Wallace et al., 2019).

The PCA of the traditional shape descriptors tends to group seeds based on accession as the largest source of variance. This observation is further supported by the Kruskal-Wallis analyses of variance (Figure A.5). The Euler characteristic however encodes additional important shape information missed by traditional descriptors. We observe that the topological shape descriptors provide better classification than the traditional shape descriptors (Table 3.2). Recall that we can mathematically prove that the ECT captures all the shape information, to the point that a finite topological signature can be used to reconstruct the original object (Curry et al., 2022; Fasy et al., 2019). This vast amount of information is best parsed with dimension reduction techniques, which highlight different morphology features encoded by the ECT. The biggest source of variation encoded by the ECT, rendered through KPCA, are individual panicles. This high degree of spike distinction may ignore underlying shape similarities between panicles of the same accession. In contrast, with UMAP we reduce the ECT's dimensionality taking into account overall topology and geometry, and produce a clustering that balances both panicle-specific nuances with more general accession-based traits. This accession-vs-panicle balance is further aided by combining traditional and UMAP-reduced descriptors. In other words, the ECT is capable of capturing both panicle- and accession-specific morphological descriptors, but different dimension reduction techniques emphasize some nuances over others. The addition of traditional shape descriptors aids accession-based clustering, by supplying size-related measurements.

The majority of the accessions studied are more easily distinguished with the topological lens but not with traditional measures, with few exceptions (Figure 3.7). Exceptions like Hannchen, Han River and Palmella Blue have slightly distinctive traditional trait distributions, so seed size does matter and it is important to take it into account (Figure 3.5.A). At the same time, we observe accessions like Alpha, Glabron, Minia, and Wisconsin Winter, that are poorly differentiated with traditional information but report considerably higher classification accuracies whenever using topological information. When looking at a more robust dimension reduction technique like

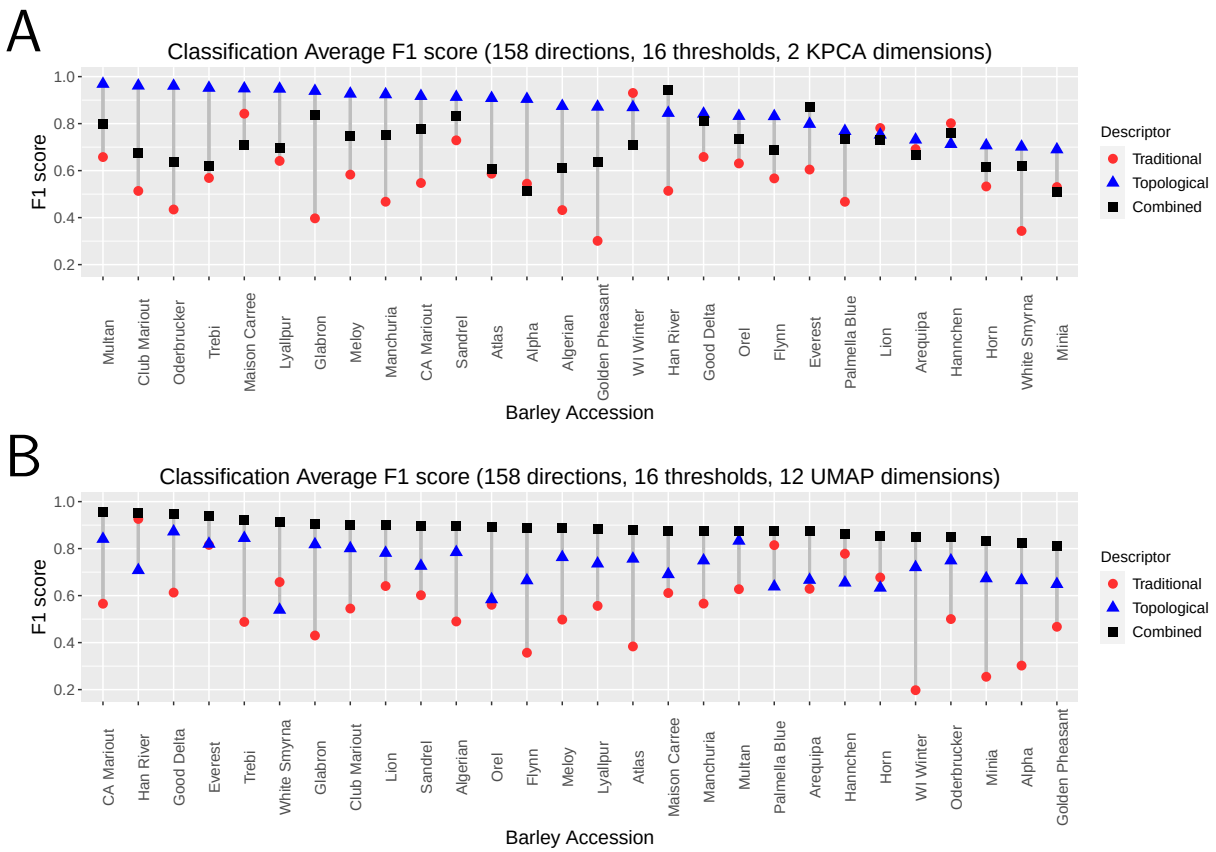

Figure 3.7: SVM classification results for individual accessions. (A) Results when using a KPCA 2-dimension reduced topological vector. Accessions ordered according to their classification accuracy determined by the topological shape descriptors. (B) Results when using a UMAP 12-dimension reduced topological vector. Accessions ordered according to their classification accuracy determined by the combined shape descriptors.

UMAP, classification accuracy is increased when combined with size-related information.

An exploration on the directions used to compute the ECT reveals that the shape of the crease and bottom discriminate accessions the most (Figure 3.4). These features are not directly measured with our traditional setting. By analyzing inter- vs. intra-accession variance of a large number of ECT axes and thresholds, we effectively isolate complex morphological features responsible for distinguishing selected groups. Although the ECT comprehensively measures the information content of an object, different dimension reduction techniques highlight different aspects of that shape information (Figure 3.6). A more systematic exploration of other dimension reduction algorithms, and classification techniques afterward is warranted moving forward.

3.5 Software and data availability

The processed and cleaned barley panicles and barley seeds X-ray CT 3D reconstructions can be found in the Dryad repository https://doi.org/10.5061/dryad.rxwdbrv93.

All of our code is available at the https://github.com/amezqui3/demeter/ repository. This includes the image processing pipeline to clean the raw scans and segment the seeds (python), the computation of the ECTs (python), and the SVM classification and analysis (R). A collection of Jupyter notebook tutorials is also provided in order to ease the usage and understanding of the different components of the data processing and data analyzing pipelines.

CHAPTER 4

EXPLORING THE SHAPE OF AROMA AND CITRUS OIL GLANDS

Fairest of all God's trees, the orange came and settled here,
Its leaves of green and pure white blossoms delight the eye of the beholder,
And the thick branches and spines so sharp, and the fine round fruits,
Green ones with yellow intermingling to make a pattern of gleaming brightness,
—from *In Praise of the Orange-Tree (Ju song)*
Qu Yuan (340–278 BC?)

Citrus come in diverse sizes and shapes, and play a key role in world culture and economy. Citrus oil glands in particular contain essential oils which include plant secondary metabolites associated with flavor and aroma. Capturing and analyzing nuanced information behind the citrus fruit shape and its oil gland distribution provides a morphology-driven path to further our insight into phenotype-genotype interactions.

We study the shape of citrus fruits and fruits from close citrus relatives based on 3D X-ray CT (computed tomography) scan reconstruction of 166 different samples comprising 51 different accessions, including samples of the three fundamental citrus species (*C. medica*, *C. reticulata*, and *C. maxima*), accessions from related genera (*P. trifoliata* and *F. margarita*), and several interspecific hybrids. First, using the power of X-rays and image processing, we compared volume ratios between different tissues, including exocarp, endocarp, and oil gland tissue. Second, since citrus oil glands contain essential oils which include plant secondary metabolites associated with flavor and aroma, we examined the number of individual oil glands, their density, and their overall distribution across all fruits. We determine that the average distance between neighboring oil glands follows a square root model, which indicates that gland distribution might follow normal diffusion dynamics (Vlahos et al., 2008). Finally, based off a point cloud defined by the center of all individual oil glands, we model the fruit shape as an ellipsoidal surface, a sphere with its three main axes shrunk or

stretched accordingly. Once the glands are considered points on an ellipsoid, we are able to apply multiple tools from directional statistics (Ley and Verdebout, 2017; Mardia and Jupp, 1999; Pewsey and García-Portugués, 2021), which allows us to study and infer possible statistical distributions on spherical surfaces. As an example of this mathematical machinery, we test whether the oil glands either follow a uniform or symmetric distribution across the fruit surface. To the best of our knowledge, the shape of citrus has not been explored with similar scanning technologies, nor has it been analyzed with ellipsoidal and directional approximations. This morphological modeling will allow us to set a new exciting path to explore further the phenotype-genotype relationship in citrus.

This chapter is derived from the original research paper

- **E.J. Amézquita, M.Y. Quigley, T. Ophelders, D. Seymour, E. Munch, D.H. Chitwood** (2022). The shape of aroma: Measuring and modeling citrus oil gland distribution. *Plants, People, Planet*.

4.1 Introduction

Citrus fruits and leaves have played a fundamental role across multiple aspects of human history including the development of modern nutrition and medical sciences. The aromatic and medicinal properties of mandarins and oranges have inspired delicate poetry since ancient times (Tseng, 1999; Vovin, 2016). Etrog citrons represent "the fruit of a goodly tree" during the Sukkot celebrations in the Jewish community (Isaac, 1959). The bael tree is considered sacred and it is generally grown near Hindu temples (Sharma et al., 2007). The fruits, peels, and leaves of diverse citrus have been used as traditional medicine for millennia for a diverse array of maladies (Mahomoodally and Mooroteea, 2021; Shrestha and Dangol, 2019). Sour oranges and lemons inspired the first modern clinical trials in the 18th and 19th centuries to determine the causes and cure of scurvy —"the plague of the sea, and the spoyle of mariners" (Hawkins, 1986)— thus paving the way to the eventual isolation and synthesis of the first vitamin, vitamin C. (Baron, 2009; Magiorkinis et al., 2011).

Currently there is a rising trend in global citrus production, with more than 143 million tonnes

produced in 2019 alone (FAO, 2021). Citrus production is valued for more than 3.3 billion US dollars in the US alone. (NASS, 2021). Citrus derived products are vital for other multi-billion dollar industries as well, from orange juice in the food industry, to essential oils in the perfume and cosmetics industry (Spreen et al., 2020). Essential oils in particular are extracted for their aromatic, flavoring, medicinal, and preservation properties useful in a variety of contexts (Mahato et al., 2019).

Before any human intervention, current paleobotanical evidence suggests that the common ancestor of citrus species originated more than 8 million years ago in the triangle defined by modern day northeastern India, northern Myanmar, and northwestern Yunnan (Talon et al., 2020). As monsoons weakened in southeastern Asia and climate transitioned to drier conditions, citrus radiated and diversified over the next 5 million years across the southeast Asian peninsula, Australia, New Caledonia, the western Indian coast, and even Japan (Wu et al., 2018). Early civilizations in India and China domesticated some of these species and their interspecific hybrids, even as early as during the Xia Dynasty (2100-1600 BC) in Southern China (Deng et al., 2020). Through tribute, trade, and invasion, different cultures contributed to spread many of these citrus across the rest of the world over the next 3000 years (Langgut, 2017).

Citrus species are sexually compatible and their ability to hybridize, combined with constant displacement and cultivation in multiple environments, produced a diversity of admixed accessions with a vast array of phenotypic traits (Gmitter et al., 2020; Luro et al., 2017; Wu et al., 2021). Asexual propagation is common in citrus and the interaction between grafted individuals has led to novel phenotypes, including through the formation of graft chimeras, conglomerations of cells that originated from separate zygotes (Caruso et al., 2020). The first reported plant chimera, known as *Bizarria*, arose from a fortuitous graft junction of a Florentine citron and a sour orange in 1674 (Nati, 1674). Since then, chimeras have proved to be more common than originally thought, transforming our perception of the genetic heterogeneity of individuals and its impact on plant development and phenotype (Frank and Chitwood, 2016).

A phenotype of particular interest is shape. Specific combinations of shape features are used to

distinguish diverse citrus varieties, and have motivated various citrus taxonomic systems (Ollitrault et al., 2020). Leaf shape has been used to distinguish pummelos from sweet oranges among other different citrus genotypes and their respective environment interactions (Iwata et al., 2002). Root architecture is indicative of soil deficiencies for sour orange rootstocks (Mei et al., 2011). Morphological traits, such as fruit size and oil gland density, are used to infer genetic similarities between various mandarin cultivars (Pal et al., 2013). Oil gland size, structure and distribution are associated with the fruit development of navel oranges (Knight et al., 2001) and grapefruits (Voo et al., 2012).

4.2 Materials and Methods

4.2.1 Plant material and scanning

We selected 51 different accessions of citrus and citrus relatives with diverse fruit morphologies and geographical origins for our analysis. Fruits were sampled from a single tree for each selected accession maintained in the University of California Riverside Givaudan Citrus Variety Collection. 166 different individuals in total were sent for scanning at Michigan State University in December 2018 (Figure 4.1.A; Table B.1.) These 166 samples were arranged into 63 raw scans, one scan per citrus variety containing all the replicates. Pummelos and citrons samples were scanned individually due to the fruit size. The scans were produced using the North Star Imaging X3000 system and the included efX software, with 720 projections per scan, at 3 frames per second and with 3 frames averaged per projection. The data was obtained in continuous mode. The X-ray source was set to a current of 70 µA, voltage ranging from 70 to 90 kV, and focal spot sizes ranging from 4.9 to 6.3 µm. The 3D reconstruction of the fruit was computed with the efX-CT software, obtaining final voxel sizes ranging from 18.6 to 110.1 µm for different scans (Figure 4.1.B; Table B.2.)

The air and debris were thresholded out of each raw scan, and individual replicates segmented into separate images. Based on density and location, for each fruit we further segmented 3D voxel-based reconstructions of its central column, endocarp, mesocarp, exocarp, and oil glands (Figure 4.1(c)–(g)). The center of each oil gland was calculated as the center of mass of the voxels

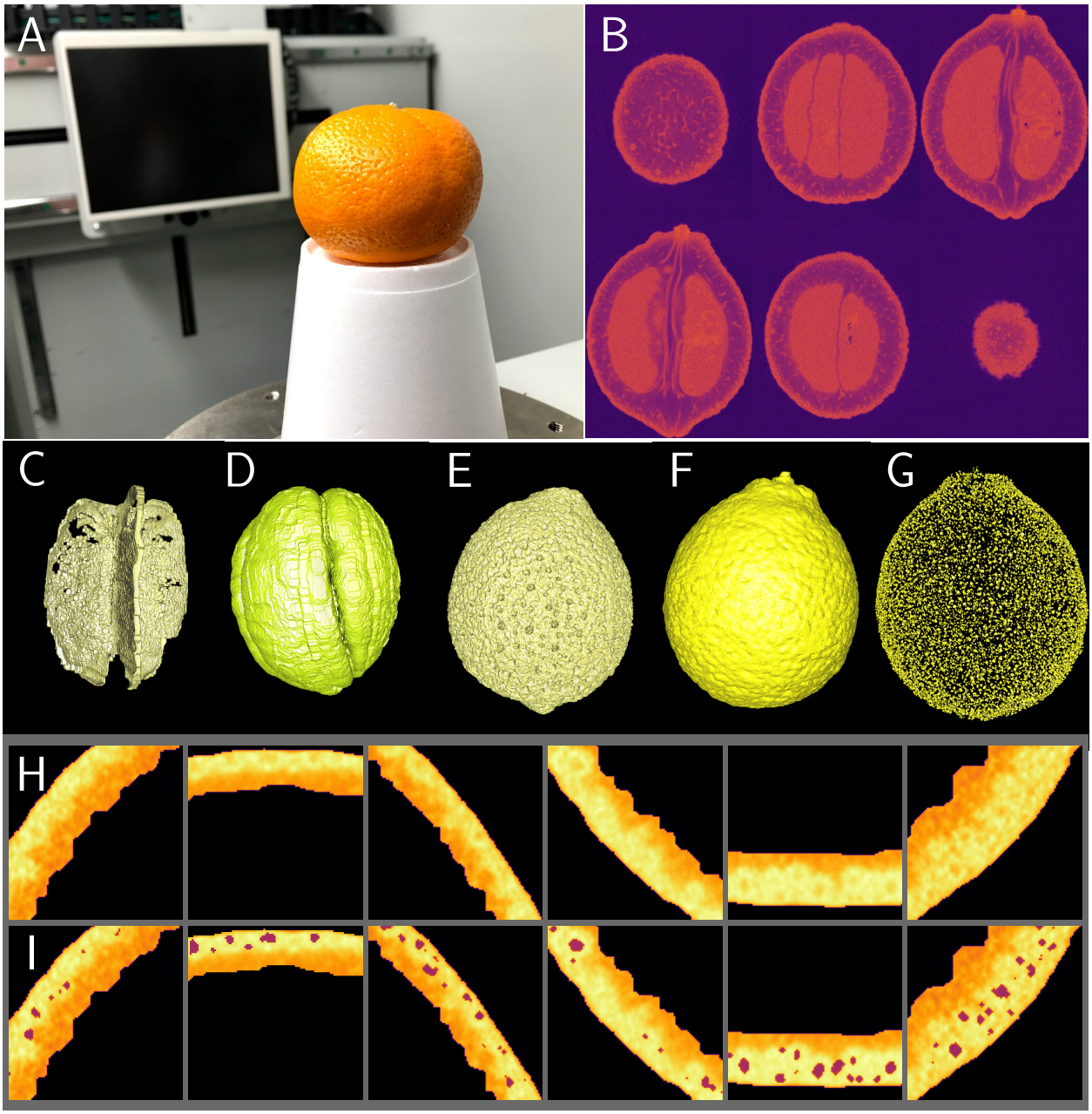

Figure 4.1: Citrus scanning and image processing. (A) A diverse collection was scanned using X-ray CT technology. (B) Slices of a raw scan. The image processing steps involved segmenting individual fruits and removing air and other debris. Then, individual tissues for each fruit were segmented such as the (C) central spine, (D) endocarp, (E) mesocarp, (F) exocarp, and (G) oil glands. (H) Close-up of some X-ray slices of the exocarp. (I) Same figure as above, with the segmented oil gland tissue darkened for emphasis. A Willowleaf sour orange is used as an example for Figures (B)–(I). All the figures are for illustrative purposes only.

composing such gland. An in-house *scipy.ndimage*-based python script was used to process the images for all fruits and their tissues. These were later visually inspected to verify their correctness. All the Chinese box oranges (*Severinia buxifolia*) scans were discarded due to their poor quality. To highlight nuanced differences among certain citrus groups, scans were partly split into sensible clusters of morphological interest (Table 4.1).

4.2.2 Allometric relationships

The total volume of fruits and their separate tissues was measured from the scans, as well as the number of individual oil glands. We studied the allometric relationships between these measurements; that is, the relative size of different tissues with respect to each other. These relationships in plants often follow a power law, so all the measurements were first log-transformed (Niklas, 2004) and a reduced major axis linear regression (Smith, 2009) was fitted considering all fruits. The slope, intercept, and R^2 correlation coefficient were recorded (Figures 4.2, B.3). The distribution of the residuals was compared against a normal distribution to determine the adequacy of the linear fit (Figures B.2, B.4).

4.2.3 Oil gland distribution

For each fruit, a point cloud, a collection of (x, y, z) coordinates in the space, was defined by the centers of all its individual oil glands. The 25 nearest neighbors, based on Euclidean distance, were computed for each point, so that distances are not affected by the fruit skin curvature. The average distance between each gland and its nearest neighbor, its second nearest neighbor, and so on were computed. The oil gland density was determined both in terms of volume and surface area, by dividing the number of glands by the volume of the whole fruit, and by the surface area of the best-fit ellipsoid (discussed later in Section 4.2.4) respectively. As in the previous section, all these measurements were log-transformed, linear regressions fitted, parameters recorded, and residuals compared to a normal distribution (Figures 4.3.A, B.5). A root square model was fitted between the average nearest neighbor distance and the nearest neighbor index to describe how far apart glands spread from each other (Figure 4.3.B).

Table 4.1: Selected citrus groups and varieties. **N** equals the number of pseudo-replicates. The full list of citrus fruits and relatives scanned is found in Table B.1. Names according to the University of California Givaudan Citrus Variety Collection (CVC). Asterisk denotes not available.

CVC Name	Scientific Name	N	-	CVC Name	Scientific Name	N
Kumquats						
Nagami	*F. margarita*	3				
Lemons and lemon hybrids						
Limon Real	*C. excelsa*	4		Lamas	*C. limon*	3
Interdonato	*C. limon*	3		Volckamer	*C. volkameriana*	3
Eureka	*C. limon*	3				
Mandarins and mandarin hybrids						
Emperor	*C. reticulata*	3		*	*C. reticulata*	3
Lee	*C. reticulata*	3		Som Keowan	*C. reticulata*	3
Koster	*C. reticulata*	3		Cleopatra	*C. reshni*	3
Beledy	*C. reticulata*	4		Fremont	*C. reticulata*	3
USDA 88	*C. reticulata*	3		Kinkoji	*C. neo-aurantium*	3
Microcitrus						
Finger lime	*M. australasica*	4				
Papedas						
*	*C. hanayu*	3		Kalpi	*C. webberii*	2
Makrut	*C. hystrix*	4				
Pummelos and pummelo hybrids						
Star Ruby	*C. paradisi*	3		Pomelit	*C. maxima*	3
Kao Pan	*C. maxima*	3		Hassaku	*C. hassaku*	3
Egami Buntan	*C. maxima*	3				
Sour oranges and sour orange hybrids						
Willowleaf	*C. aurantium*	3		Standard	*C. aurantium*	3
Konejime	*C. neoaurantium*	4		Olivelands	*C. aurantium*	3
Sweet oranges						
Valencia	*C. sinensis*	3		Shamouti	*C. sinensis*	3
Navel	*C. sinensis*	3		Argentina	*C. sinensis*	3
Cara Cara	*C. sinensis*	3				
Trifoliates and trifoliate hybrids						
Little-leaf	*P. trifoliata*	3		Rubidoux	*P. trifoliata*	4
C-35	*X Citroncirus*	3		Carrizo	*X Citroncirus*	5
Swingle	*X Citroncirus*	5				

4.2.4 Modeling the whole fruit as an ellipsoid and computing its sphericity

The surface of most of citrus fruits and their relatives can be approximated by an ellipsoid, a sphere with its three main axes possibly shrunk or stretched.The three axes of symmetry of an ellipsoid delimit three line segments from the center of the ellipsoid to its surface. These are referred to as the *ellipsoid semi-axes*. Notice that a sphere is an ellipsoid with its three semi-axes of the same length. We will consider triaxial ellipsoids, where the length of each semi-axes can be different. An ellipsoid can also be represented as a quadratic equation surface which is both mathematically simple to manipulate (Harris and Stöcker, 1998, Ch. 8.12), and versatile enough to represent both the shapes of nearly-spherical Valencia oranges and elongated finger limes given the right semi-axes lengths.

Each fruit is defined by a point cloud made by the centers of all its individual oil glands. The parameters of the best-fit ellipsoid for this point cloud are computed following the algorithm by Li and Griffiths (2004), from which the semi-axes lengths, rotations, and center are determined (Panou et al., 2020). The fruit point cloud is then rotated and translated such that the best-fit ellipsoid is centered at the origin and its semi-major axes coincide with the proximal-distal axis of the fruit. Finally, the centers of the oil glands are projected to this ellipsoid via geocentric projection, where a ray from the center of the ellipsoid to the gland is drawn and its intersection with the ellipsoid is considered (Figure 4.4.A–F).

This ellipsoid model summarizes important information of the overall shape of the fruit. As an example, we measure how sphere-like different citrus are. There is no unique way to measure sphericity, however, most of the commonly used formulas are based on the semi-axes lengths of the object (Blott and Pye, 2008; Clayton et al., 2009). We measured the sphericity of the resulting fruit-based ellipsoids using 6 different sphericity indices, all of them taking values between 0 (planes and lines) and 1 (perfect spheres) (Figure 4.4.G; Table 4.2).

4.2.5 Revisiting the distribution of the oil glands

The projected gland center locations on the ellipsoid were described in terms of longitude and latitude coordinates with respect to the ellipsoid (Diaz-Toca et al., 2020). We tested whether the

Table 4.2: Common sphericity indices based off best-fit ellipsoid. The indices values are bounded between 0 (line or plane) and 1 (perfect sphere). The surface area, volume, and the largest, intermediate, and smallest semi-axes lengths of the ellipsoid are denoted by A_e, V_e, a, b, c respectively. Also, A_s denotes the surface area of a sphere of volume V_e.

Name	**Formula**	**Reference**
True sphericity	$A_s/A_e = \sqrt[3]{36\pi V_e^2}/A_e$	(Wadell, 1932)
Intercept sphericity	$\sqrt[3]{bc/a^2}$	(Krumbein, 1941)
Corey shape factor	$c/\sqrt{ab}$	(Corey, 1949)
Maximum projection sphericity	$\sqrt[3]{c^2/ab}$	(Sneed and Folk, 1958)
Janke form factor	$c/\sqrt{\frac{1}{3}(a^2+b^2+c^2)}$	(Janke, 1966)
Degree of equancy	c/a	(Blott and Pye, 2008)

gland point cloud follows a uniform distribution, where every unit area of the skin has the same probability of containing oil glands; or if the underlying distribution is rotationally symmetric, where the oil glands pattern is symmetrical around a fixed direction. Uniformity was tested with Projected Anderson-Darling (PAD) test (García-Portugués et al., 2023) with the R package *sphunif* (García-Portugués and Verdebout, 2021). The rotational symmetry was tested with a scatter-location hybrid test with an unspecified direction of symmetry (García-Portugués et al., 2020) with the R package *rotasym* (García-Portugués et al., 2021). Additionally, we visually examined the distribution of oil glands for most fruits and compared them to simulated uniform distributions by projecting them to 2D via Lambert azimuthal equal-area projections (Mardia and Jupp, 1999, Ch. 9.1) from the North and South poles. Intuitively, these two projections flatten the sphere on a plane by pushing it from the North pole and South pole while minimizing the distortion seen on the north and south hemisphere respectively (Figure 4.5).

4.3 Results

4.3.1 Allometric relationships

The estimated volume of each tissue type and fruit follows the expected average fruit size of each genetic group, with the smallest fruit in the bottom left corners (microcitrus, kumquats) and large fruit in the top right corners (pummelos). Strong linear trends are observed when comparing most of the volume-related features of all the fruits, indicated by high R^2 correlation coefficient

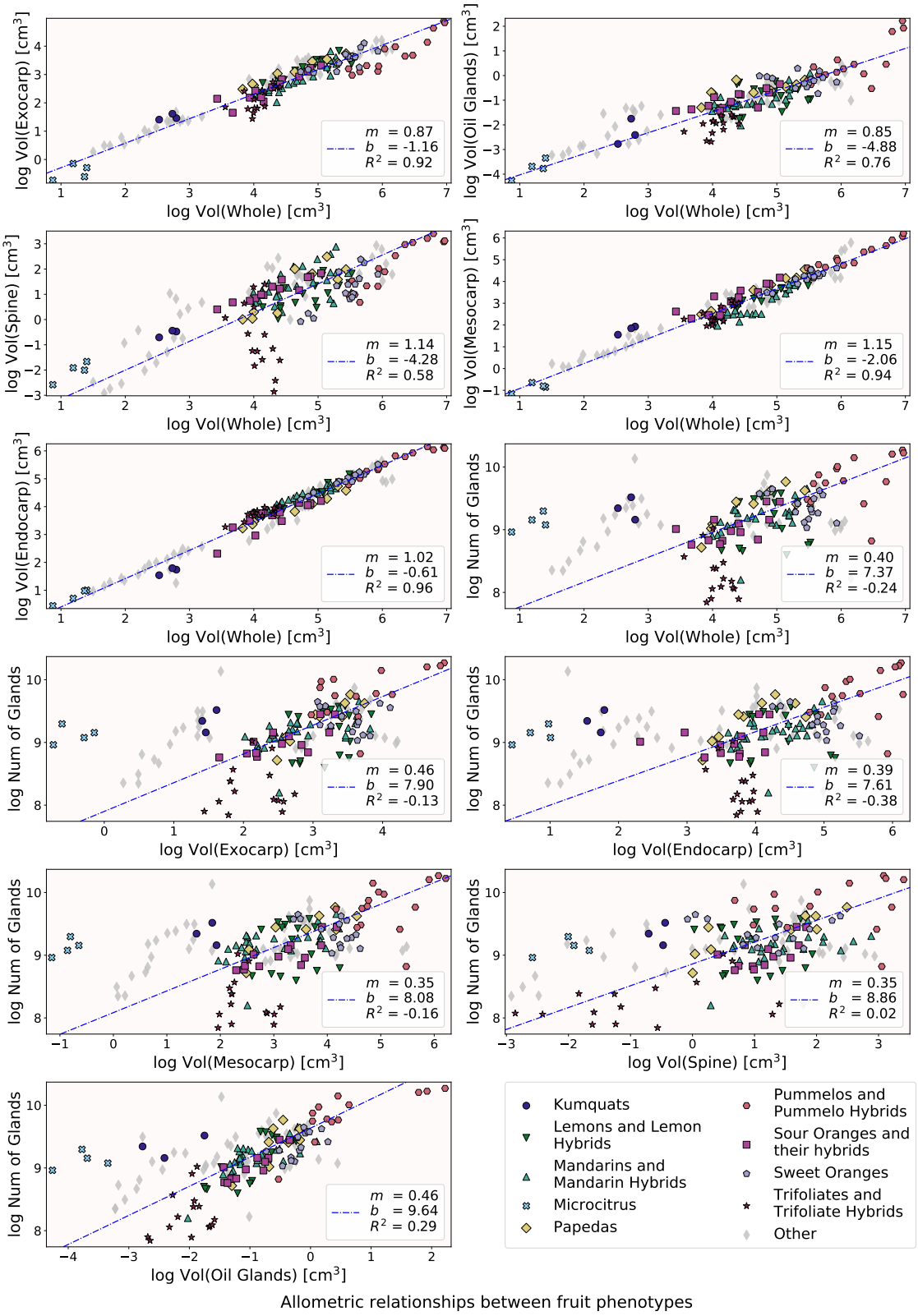

Figure 4.2: Various allometry plots between different tissue volumes compared to the total fruit volume.

Figure 4.2 (cont'd):
Various allometry plots between different tissue volumes compared to the total fruit volume across all fruits. The best fit line is depicted by a dashed line in blue. For each plot, the slope, intercept, and correlation coefficient are recorded as m, b, and R^2 respectively. The linear relationship in the log-log plots suggests that fruit tissues may grow following a power law.

values, usually above 0.75 except for the central column tissue (Figures 4.2, B.3). Due to their thin size and scanning quality, these columns were difficult to identify and isolate, especially in some trifoliates, which might explain lower R^2 values. The residuals of the fitted linear regression tend to follow a normal distribution for the majority of measurement pairs, suggesting that the linear fit is adequate (Figures B.2, B.4). This linearity indicates that the tissues across all citrus fruits grow relative to each other following a power rule. For example, looking at the slope m values, both the exocarp and the oil glands grow in volume at the same relative rate with respect to the volume of the whole fruit ($m = 0.85$). On the other hand, the total number of oil glands appears to be decoupled from all the measured size-related traits, as shown by much lower R^2 values (Figure 4.2). In this case, a power law may not be an adequate model to describe the oil gland number with respect to tissue volume.

4.3.2 Oil gland distribution

There is a strong positive linear relationship between the volume of the fruit, and the average distance between an oil gland and its nearest neighbor, with R^2 correlation coefficients above 0.65. There is a stronger negative linear relationship when considering the overall oil gland density, reflected by R^2 coefficients above 0.85 (Figure 4.3.A). The residuals follow normal distributions, indicating that the linear model is adequate (Figure B.5). These allometric relationships suggest that for all citrus and relative fruits, the average distance between nearest oil glands follows a power law with respect to fruit volume and gland density. When considering fruit size, as expected, the samples distribute in a similar pattern as with most of the previous allometry plots. However, an inverse pattern is observed when considering oil gland density. In this case, the smallest fruits tend to report the highest number of oil glands per unit volume or unit area. Other than microcitrus and kumquats, the rest of highlighted citrus groups form a tighter cluster.

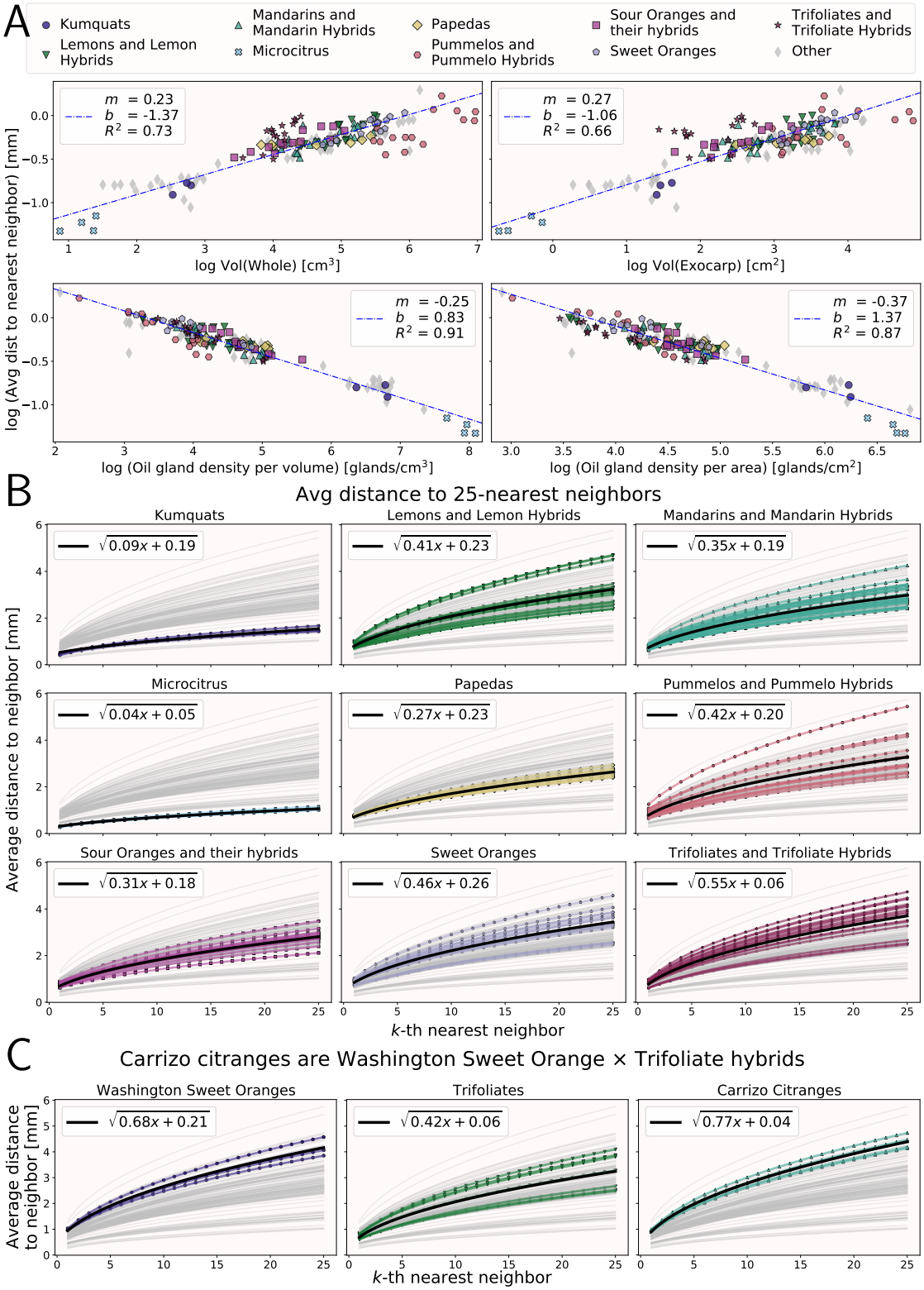

Figure 4.3: Studying the average distance from each gland to its first 25 nearest neighbors, as measured by the Euclidean distance.

Figure 4.3 (cont'd):
Studying the average distance from each gland to its nearest neighbors. (A) Allometric relationships are observed across all fruits when comparing the average distance between each gland to its closest neighbor with the overall size of the fruit. The overall linear trend is depicted by a dashed blue line. The slope, intercept, and correlation coefficient are denoted by m, b, and R^2 respectively. (B) For each oil gland, the average distance to its nearest neighbors follows a square root relationship. The average for each group is plotted as black, thick line. These square root models follow different parameters depending on the citrus group. The fruits that deviate the most from the average usually correspond to hybrids. (C) Carrizo citranges are Washington sweet orange × Trifoliate hybrids. The average distance between oil glands increases at a faster rate for citranges than for their parents, which suggests hybrid vigor.

The average distance between an oil gland and its k-th nearest neighbor is modeled as

$$\text{Average distance}(k) = \sqrt{Mk + B},$$

where M is the rate of distance growth and B the line intercept (Figure 4.3.B). As expected from their higher oil gland density, the average distance to the gland's nearest neighbors increases the slowest for microcitrus, followed distinctly by kumquats. On the other hand, sweet oranges and trifoliates report the largest average distances between neighboring oil glands. This higher gland density could be partly affected by differences in scanning resolutions (Figure B.1).

In general, all the samples of every accession follow the same growth model. Outliers in growth models are typically associated with hybrid accessions. Increased growth rates are found when the second parent is a large fruited accession. For example, consider the Carrizo citrange, a trifoliate x Washington sweet orange hybrid where sweet orange fruits are much larger than trifoliate fruits. We observed that the oil glands in the citrange grow on average farther apart from each other than in any of the parents (Figure 4.3.C), which suggests that hybrid vigor might be at play. Similarly, hybrids derived from crosses with small-fruited accessions have reduced growth rates.

4.3.3 Ellipsoid modeling and sphericity of fruits

The best-fit ellipsoid successfully captures the overall shape of the citrus and relatives, with a negligible portion of gland centers differing by more than 0.2cm from their ellipsoidal approximation (Figure 4.4.G). This ellipsoidal model is flexible enough to capture both spherical and elongated fruit shapes, from sweet oranges to Australian finger limes (Figures 4.4.A–F).

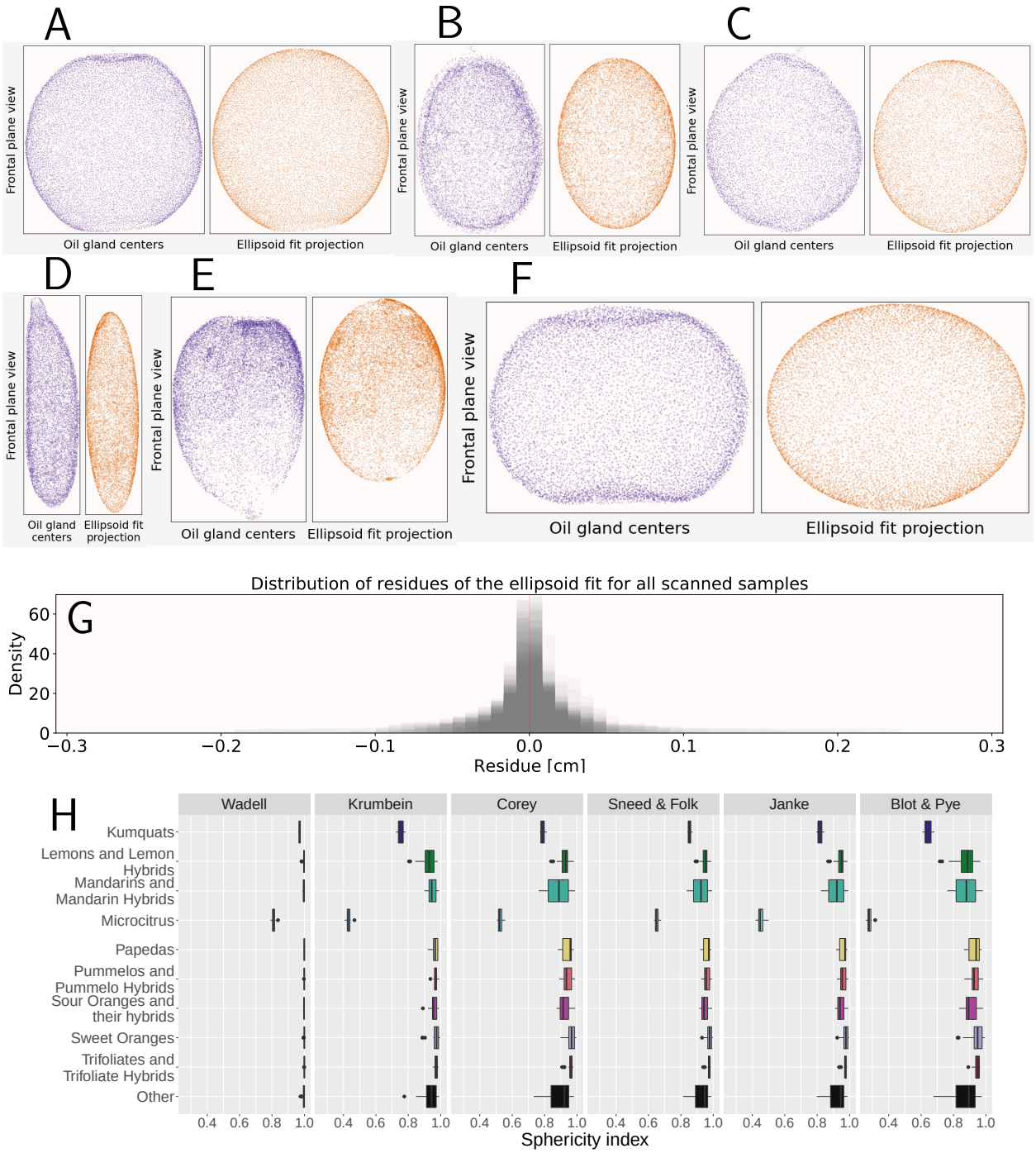

Figure 4.4: Modeling citrus fruit surface as tri-axial ellipsoids. The glands were centered at the origin, and the ellipsoid aligned with the proximal-distal, medial-lateral, and adaxial-abaxial axes. Then the oil glands were projected to this best-fit ellipsoid. Examples of a (A) Valencia orange, (B) Nagami kumquat, (C) Willowleaf sour orange, (D) Australian finger lime, (E) South Coast Field Station citron, and a (F) Cleopatra mandarin. (G) Distribution of the residuals of the centers of the oil glands to the best-fit ellipsoid. The distributions for all the fruit scans are overlaid. (H) Various sphericity indices are computed and compared across different citrus groups. The indices are named according to their original reference in Table 4.2. Figures (A)–(F) are not scaled.

Most of the fruits report highly spherical indices, more than 0.9 for every sphericity index. Unsurprisingly, the elongated Australian finger limes are the least spherical and their shape is very distinct from the rest of the samples. The kumquats are less elongated than the finger limes, but they also remain highly distinguishable for most of the sphericity indices. Mandarins and their hybrids also tend to be slightly less spherical than the remaining groups of interest (Figure 4.4.H).

4.3.4 Oil glands revisited

Although the tools from directional statistics assume that the data points lie on a sphere rather than an ellipsoid, most of scanned fruits were very sphere-shaped according to a variety of sphericity indices (Figure 4.4.H). Thus shape information is not significantly altered when translating longitude and latitude coordinates from the best-fit ellipsoid to a sphere.

The uniform oil gland distribution hypothesis was strongly rejected for all scans, with all p-values below 0.015 for the PAD test, and below 2.5×10^{-7} for 95% of all the point clouds (Figure B.6.A). The scatter-location hybrid test strongly rejected the rotationally symmetric oil gland distribution hypothesis for most of the point clouds as well. More than 90% of all the scans reported p-values smaller 0.02. The 10 samples for which the rotationally symmetric hypothesis was not rejected were not concentrated in any citrus groups (Table B.3; Figure B.6.B). Upon closer visual examination, differences arise between the uniform distribution on a sphere and oil gland distributions. The oil gland distributions tend to have defined clusters and empty spots, which are not seen in typical uniform distributions. The northern and southern hemispheres are noticeably different from each other for the oil gland distributions, while these look roughly the same in uniform distributions (Figure 4.5).

4.4 Discussion

Measuring and understanding the shape is fundamental to extracting valuable information from data, and push further our insights. A vast number of biological-inspired shapes are intrinsically 3 dimensional, like citrus fruit, and capturing their shape as 3D voxel-based provides a faithful shape representation that allows accurate measurement of tissue volumes and modeling of gland distributions in space. Better fruit modeling is key to provide more accurate descriptions of fruit

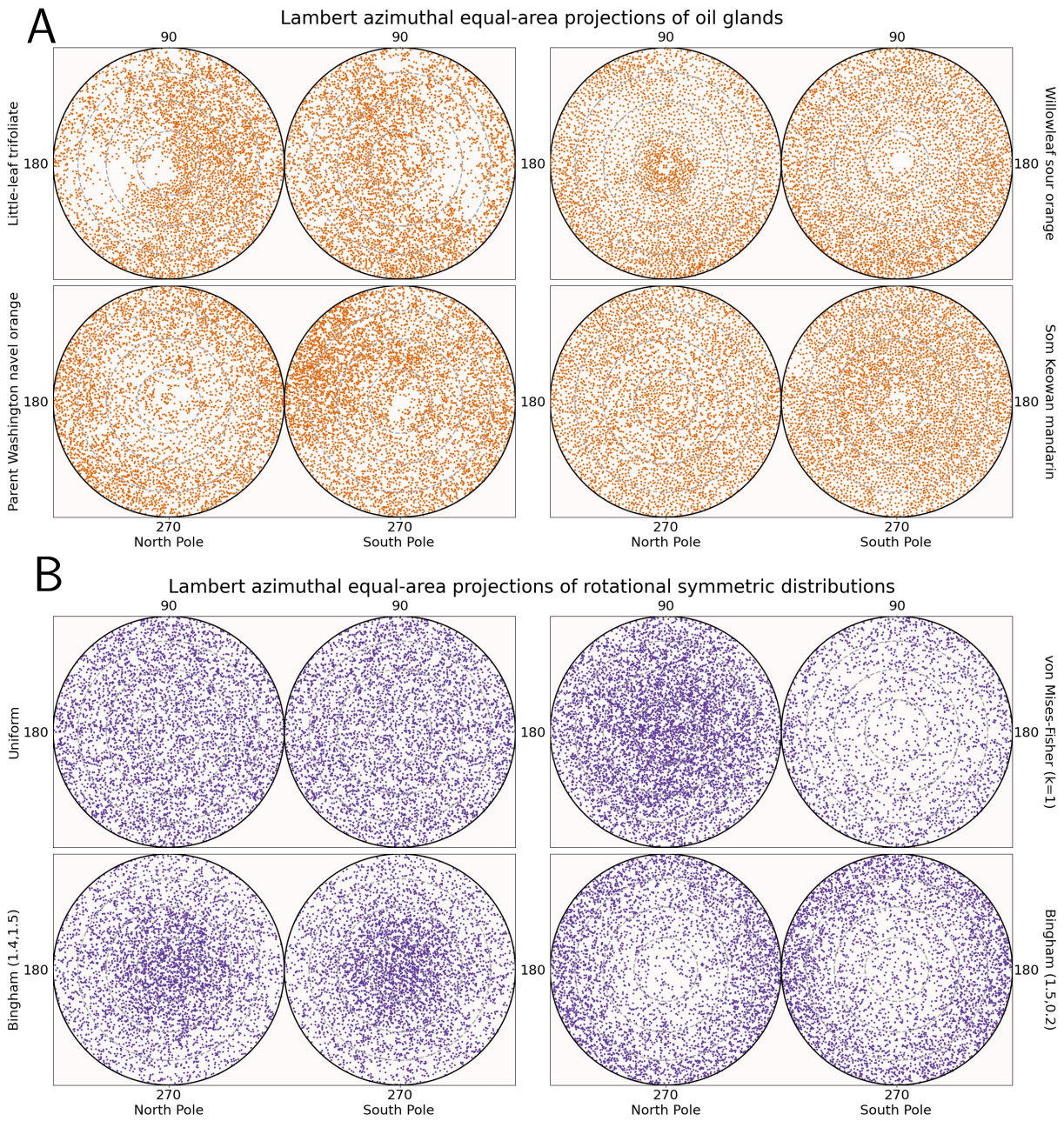

Figure 4.5: Distribution of oil glands is not uniform nor rotationally symmetric across the citrus exocarp. After modeling the fruits as ellipsoids and projecting their oil glands onto the ellipsoidal surface, longitude and latitude coordinates are computed as in Figure 4.4. These coordinates can be better visualized using two Lambert azimuthal equal-area projections, from the north and south poles which represent the northern and southern hemispheres respectively with minimal distortion. A battery of statistical tests strongly rejects the hypothesis of these glands being uniformly or symmetrically distributed over the ellipsoid surface. (A) Examples of oil gland distribution of a little leaf trifoliate, a parent Washington navel orange, a Willowleaf sour orange, and a Som Keowan mandarin. (B) For comparison, a similar number of points are simulated following uniform, low-concentration von Mises-Fisher, and low-concentration Bingham distributions. These three distributions are rotationally symmetric.

shape and oil gland content, as both are important traits for citrus scion improvement (Barry et al., 2020). Citrus shape impacts oil gland abundance and distribution, as the shape of the rind, the exocarp, and other tissues, along the distribution of the oil glands affects the physics of citrus essential oil extraction and aroma dispersion (Smith et al., 2018).

When observing overall fruit tissue size trends, these correspond to known citrus genealogy. For example, when comparing the size of exocarp against size of endocarp, most of the sour and sweet oranges tend to lie between mandarins and pummelos, with sour oranges lying closer to mandarins, while the sweet oranges are closer to pummelos (Figure 4.2). A similar arrangement of citrus groups is observed when comparing the average distance between neighboring oil glands to either fruit volume or gland density. Moreover, it is observed that oil glands distance themselves from each other following a square root rate in general. The exact degree to which they push each other apart depends on the oil gland density, which in turn is partly affected by the citrus genealogy. For example, the average distances between oil glands in Carrizo citranges increase at higher rates than in either the Washington sweet oranges or the trifoliates, the citrange parents (Figure 4.3). The square root suggests that the mechanics of oil gland displacement across the fruit could be partly governed by based on Brownian motion and normal diffusion interactions (Vlahos et al., 2008). However, the hypothesis of oil gland locations following either a uniform or symmetric distribution on the fruit surface is strongly rejected for all scans by the PAD and location-scatter hybrid tests respectively. This discrepancy between our gland distribution observations could be explained by the fact that uniform distributions and diffusion processes assume that the data consists of point particles with no volume that can stand arbitrarily close to each other. For oil glands this is obviously not the case, as they have volume and there are physical limitations on the proximity between glands, which requires a more complex diffusion modeling. Higher resolution scans might be able to capture better individual oil gland shape, rather than just its center. Individual oil glands then could be approximated by individual minimum volume enclosing ellipsoids (Todd and Yıldırım, 2007), which could then pose more advanced distribution and diffusion models.

All the studied citrus and related accessions exhibit allometric behavior in general across

both their tissue volumes, and average distances between neighboring glands. This relative growth relationships suggest that tissue sizes are deeply linked, as the size of oil gland tissue in general may not be able to change without changing volume of both the endocarp and mesocarp. Moreover, there might be biophysical principles at play that govern different tissue development across all citrus fruits in general, just like normal diffusion might govern oil gland distribution. The determination of such biophysical constraints prompt future lines of exciting research.

The quality of input data for our models is equally important. Through X-ray CT scanning technology we have a novel way to observe, quantify, and analyze all the shape of citrus and their tissues in a comprehensive, automated, non-invasive, and non-destructive manner. With the right voltage and current, the 3D X-ray CT reconstructions can discern small, individual tissues, like oil glands, which enables us to analyze tissue shape and distribution at very granular levels.

4.5 Software and data availability

The processed and cleaned citrus X-ray CT 3D reconstructions can be found in the Dryad repository https://doi.org/10.5061/dryad.34tmpg4n6, along with their separated tissues and associated point clouds and ellipsoidal approximations.

All our code is available at the https://github.com/amezqui3/vitaminC_morphology repository. This includes the image processing pipeline to clean the raw scans and segment the fruit tissues, the computation of tissue volume, the best-fit ellipsoid, and the hypothesis testing of uniform and symmetric distributions on a unit sphere. A collection of Jupyter notebook tutorials is also provided to ease the usage and understanding of the different components of the data processing and data analyzing pipelines. All the image-related scripts are available in python, while the statistical analyses are in R.

CHAPTER 5

THE SHAPE OF KERNELS AND CRACKS, IN A NUTSHELL

Don't depend on the world's friendship,
For friends can turn into foes.
Although walnut has a round shape,
Not every round object is a walnut

Flower and Soil

Parvin E'tesami

There is more than meets the eye whenever we look at walnuts (*Juglans regia*). Civilizations originary from modern day Iran have used and traded walnut tree products since the 7000 BC (Vahdati, 2014). The walnut fruit offers plenty of nutrients and health benefits (Chudhary et al., 2020), the wood is strong and lustrous (Voulgaridis and Vassiliou, 2005), the essential oils of the leaves are moisturizing (Verma et al., 2013). Walnuts traveled far and wide as they were actively traded through the Silk Road, reconquering the Eurasian continent (Pollegioni et al., 2014). Moreover, spatial genetic partitions among walnut populations coincide with large differences in human language; similarly, areas with similar human languages coincide with areas were walnut populations have been homogenized (Pollegioni et al., 2015).

The trade of walnuts remains an important part of the global economy. In 2021, the California and the US produced more than 725,000 tons of walnuts valued in more than $1.0B, following a historically increasing trend of both bearing acreage and bearing trees per acre (NASS, 2022). World demand for walnut keeps increasing and it is estimated that the world will consume a record 2.5M tons of walnuts for 2023, and the US is forecasted to satisfy 25% of the global demand (FAS, 2022). The trade is not limited to the food industry, as there is also growing research on additional uses for walnut shell material for more durable batteries (Wahid et al., 2017), lower-cost concrete (Hilal et al., 2020), and stronger epoxy composites (Lala et al., 2018). As climate change alters weather patterns, and the demand for walnut and its byproducts increases, we must breed walnuts with more suitable traits (Bernard et al., 2017). Quantitative analyses and comprehensive pheontyping can

accelerate current breeding programs by quickly identifying varieties and individuals with desirable characteristics (Fiorani and Schurr, 2013; Rahaman et al., 2015). The rapid selection of potentially desirable progenitors for breeding programs is especially crucial for walnuts, as seedlings are hard to propagate, it takes at least 2 years for trees to bear fruit for the first time, and at least 5 more to yield fruit at a commercial scale (Lopez, 2004; Popa et al., 2023; Verma, 2014).

Most of the current walnut phenotyping follows the measuring guidelines set by the International Plant Genetic Resources Institute (IPGRI, 1994). The morphological phenotyping of the fruit is mainly done using calipers to measure length, width, and height, combined with visual assessments to describe more complicated traits such as texture and curvature. These simple measurements have proved to be insightful to evaluate and identify promising genotypes. Moderate correlations have been reported between these traditional morphological traits of the walnut tree and fruit with commercial and horticultural traits of interest such as pollen release strategy, yield, shell thickness, kernel weight, and pathogen resistance (Akca and Şen, 1995; Kelc et al., 2007; Khadivi-Khub et al., 2015; Rezaei et al., 2018; Shah et al., 2021; Solar et al., 2003).

However, this caliper- and eye-based approach is time consuming, prone to human error and subjectivity, and fails to capture richer shape nuance observed in the shells and kernels. As next generation sequencing (NCS) technology advances, we observe an explosion in genomics data collection that must be matched by equally powerful and encompassing phenomics (Andrade-Sanchez et al., 2013; Araus and Cairns, 2014; Tanabata et al., 2012). We have to look deeper than just nut lengths and widths. To that end, X-ray computed tomography (CT) scanning has proved to be a powerful tool to accurately capture intricate, internal features of a vast array of plant data in a nondestrutive manner. High-resolution, X-ray CT 3D reconstruction have been successfully used to capture and quantify the complex branching architecture of inflorescence in grapevines (Li et al., 2019) and sorghum panicles (Li et al., 2020), digitally segment and phenotype all the individual seeds in a barley panicle (Amézquita et al., 2021), determine nuances in soil porosity for diverse wheat root-soil interactions (Zhou et al., 2020), and measure exact volumes and distribution of oil glands across multiple citrus exocarps (Amézquita et al., 2022). To the best of our knowledge,

Bernard et al. (2020) is the first study that exploits X-ray CT 3D reconstructions to automatically, accurately, and systematically quantify multiple nut shape phenotypes from a germplasm diversity panel maintained by INRIA. Given the nature of X-rays, Bernard et al. are able to fully measure the volume and percentage of different nut tissues, namely shell, kernel, and air contained, as well as shape descriptors related to the whole nut, like total volume and surface area. Based on these results, they are able to propose the selection of genotypes with higher kernel filling ratio and thinner shells. In particular, they observe that larger fruits are correlated with rougher shell shape and smaller kernel filling ratio. We also highlight that micro-CT imaging has been recently used to document morphological changes of flower bud development (Gao, 2022), and to explore the puzzling diversity and structure of the cell tesselations that conform the hard shell tissue for multiple nuts (Huss et al., 2020). In both of these cases, the micro-CT imaging plays a more exploratory role rather than a quantifying one.

Here, we study the shape of walnut fruits based on the X-ray CT 3D reconstruction of 1256 different samples comprising 173 accessions maintained by the Walnut Improvement Program at the University of California Davis. We exploit the nondestructiveness of X-rays to isolate individual walnuts and segment out shell, kernel, and packing tissues, as well as the air contained inside every walnut. We then compute 38 different shape- and size-related traits for each walnut. This includes side lengths, surface areas, and volumes of the whole nut and individual tissues, filling ratios, and sphericity and convexity indices. This image processing task was done with an in-house, python-based, open-source script inspired by the procedure described by Bernard et al. (2020). We include the computation of the 14 traits used by them. Second, we look for allometric relationships of interest across the whole population —the growth rate of a tissue relative to another. Third, we perform Kruskal-Wallis analyses of variance (Kruskal and Wallis, 1952) to determine which morphological traits contribute the most to qualitative traits of interest, such as kernel ease of removal, kernel plumpness, and shell strength and integrity. Finally, we singled out the only Himalayan, heterozygous accession and studied more carefully its morphology, as it is reportedly extremely hard to crack open and extract its kernel. We noticed that this particular accession is

no different in size or tissue distribution compared to other easier to crack accessions, except for subtle differences in the kernel's main cavity at the proximal end. This morphological modeling will allow us to set a new exciting path to explore further the phenotype-genotype relationship in walnuts.

5.1 Materials and methods

5.1.1 Plant material and scanning

All plant materials represent walnut breeding lines, germplasm, and cultivars maintained by the Walnut Improvement Program at the University of California, Davis. A total of 149 walnuts accessions were harvested into mesh bags at hull split, oven-dried overnight at 95F, and then air-dried for several weeks before moving into cold storage at 35F. 5 to 16 individuals were selected for each accession, for a total of 1301 individual walnuts to be scanned at Michigan State University (Table C.1). The walnuts were scanned in 173 batches. The scans were produced using the the North Star X3000 system and the included efX-DR software, with 720 projections per scan, at 3 frames per second and with 3 frames averaged per projection. The data was obtained in continuous mode. The 3D X-ray CT reconstruction was computed with the efX-CT software, obtaining voxel-based images with voxel size of 75.9 μm.

All the individual walnuts were manually separated with the efX-CT software (Figure 5.1.A). Densities were rescaled so that all scans share similar air, kernel, and shell densities. Once densities were comparable across samples, the external air and other debris was removed through thresholding and mathematical morphology operations (Figure 5.1.B). Rough estimates for the location of shell, air, kernel, and packing tissues were obtained based on density and object thickness information. These tissues were fully segmented using the watershed segmentation algorithm (Falcao et al., 2004) (Figure 5.1.C, E-H). We took particular care of tissue labeled as shell, where we distinguished voxels on or close to the walnut surface, to voxels protruding into the internal cavity. We also labeled apart voxels that were of similar exterior shell density and that comprised extraneous bulges (Figure 5.1.D). Some of the scanned walnuts contained incomplete or no kernel at all. These were discarded from further morphological analysis, leaving us with a total of 1264 individual walnuts. All the

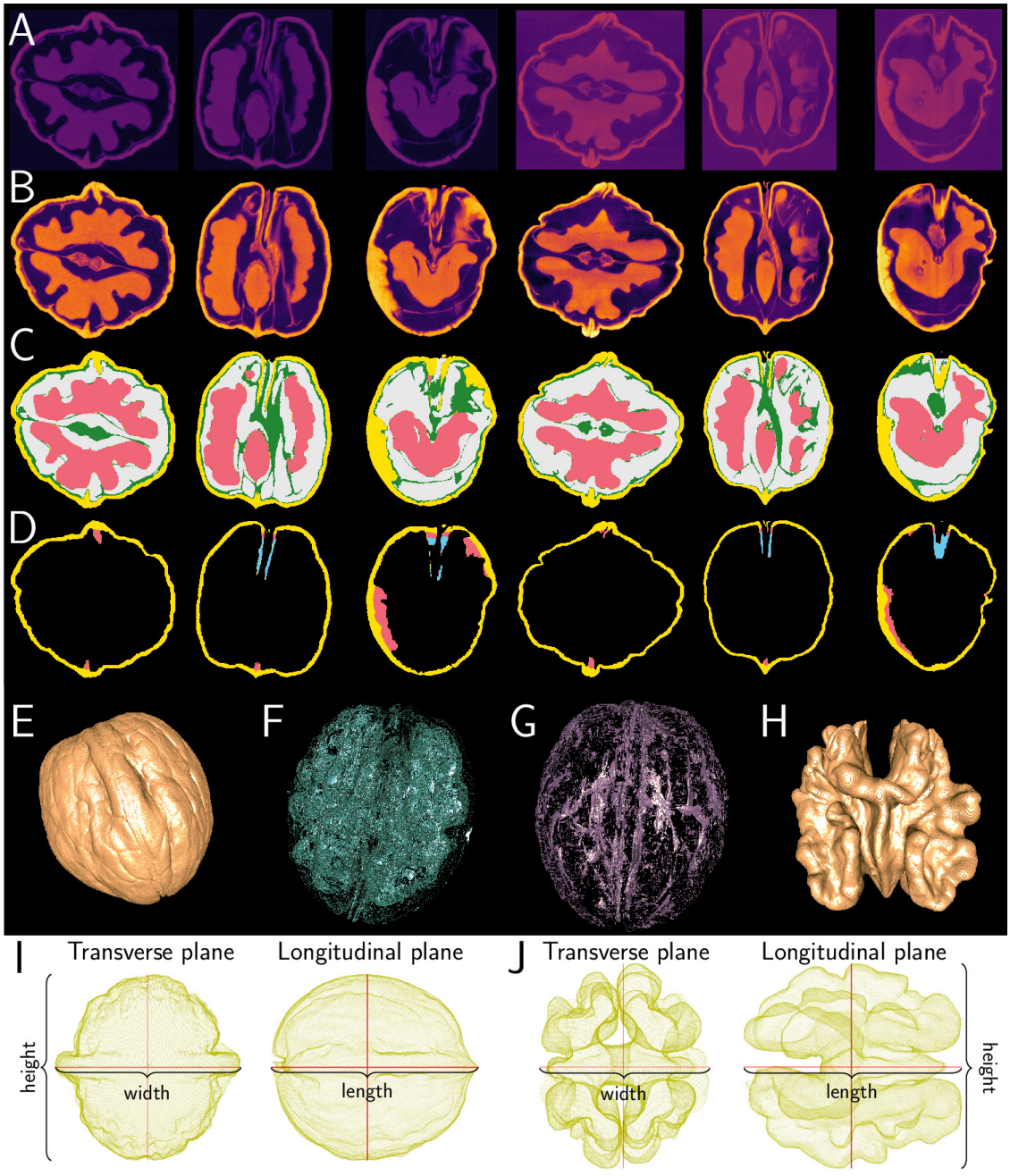

Figure 5.1: Walnut scanning, image processing, and phenotyping. (A) Raw scans of individual walnuts. (B) Densities were standardized across all samples and the external air removed. (C) Shell, air, kernel, and packing tissue were automatically labeled with a combination of basic image morphology operations and watershed segmentation. (D) The tissue labeled as shell was further broken down into external shell, bulging, and protruding tissue. (E) 3D renders of shell, (F) air, (G) packing tissue, and (H) kernel. (I) All the walnuts were centered on their center of mass and aligned. (J) The same centering and alignment was applied to the kernels. All the figures above are for illustration purposes only and are not scaled.

image processing above was done automatically with in-house, scipy-based, python scripts.

To make some measurements comparable, all the walnuts were centered on their centers of mass and rotated such that the lateral plane goes through the walnut seal, and the shell tip is the rightmost point of the longitudinal plane (Figure 5.1.I.) The same center and rotation was immediately applied to the kernel (Figure 5.1.J).

5.1.2 Walnut morphological traits and evaluation

For each individual we computed the same 14 morphological traits as in Bernard et al. (2020): nut length, height, width, total surface area, total volume, rugosity, sphericity, shape VA3D, equancy, shell volume, shell thickness, kernel volume, kernel filling ratio, and the empty space volume. We computed an additional collection of 24 morphological traits for a total of 38 measurements per sample. We computed the length, width, and height of the kernel, the volume of protruding shell, inner bulging shell, packing tissue, the percentage of each tissue volume, and Krumbein and Sneed sphericity indices (Blott and Pye, 2008). We also computed surface area and volume of the nut's convex hull, and their ratio between actual nut surface area and volume respectively as a proxy for lobeyness. This computation was repeated for the kernel. (Table 5.1)

Since all the 38 morphological traits are nonnegative, we used the quartile coefficient of variation (QCD) to measure the numerical variability of each of them across the 1301 scans. We preferred the QCD as it only depends on the 25th and 75th quartiles, making it robust against outliers compared to the coefficient of variation (CV) (Bonett, 2006). We studied allometric relationships between all size-specific morphological traits, that is, the relative growth of one feature with respect to another one. It is a well-documented phenomenon that different tissues grow relative to each other following a power law rather than a simple linear relationship (Niklas, 2004; West et al., 1999), so we plotted our data in log-log plots (Figure 5.2). Due to this nonlinear relationships between traits, we favored the computation of Spearman rather Pearson correlation coefficients between different morphological phenotypes (Figure 5.3).

Ten walnuts were cracked open from each sample using a hammer. Ease of removal was scored onto that each sample on an ordinal scale (1-9) as the ease with which intact kernel halves could be

Table 5.1: Morphological traits measured. Traits organized by unit and walnut tissue involved. Exp indicates formula or expression. Asterisk denotes not applicable.

Trait	Exp	Description	Unit
Whole walnut			
Length	L_w	Distance from base to tip	mm
Width	W_w	Longest distance across the seal	mm
Height	H_w	Longest distance perpendicular to the seal	mm
Surface	A_w	Surface area of the actual nut	mm^2
Convex surface	A_{cw}	Surface area of the convex hull of the nut	mm^2
Volume	V_w	Total volume of the actual nut, including air	mm^3
Convex volume	V_{cw}	Volume of the convex hull of the nut	mm^3
VA3D	$A_w^3/(36\pi V_w^2)$	Shape factor	*
Feret Ratio	H_w/L_w	Inverse index of roundness	*
Krumbein	$\sqrt[3]{W_w H_w/L_w^2}$	Index of roundness	*
Sneed	$\sqrt[3]{H_w^2/W_w L_w}$	Index of roundness	*
Sphericity	$\sqrt[3]{36\pi V_w^2}/A_w$	Wadell's index of roundness	*
Rugosity	$1/\text{sphericity}$	Index of surface roughness	*
Convex area ratio	A_{cw}/A_w	Index of nonconvexity	*
Convex vol ratio	V_w/V_{cw}	Index of nonconvexity	*
Shell			
Total Volume	V_s	Total volume of the shell	mm^3
External Volume	V_e	Volume of shell without protrusions or bulges	mm^3
Bulging Volume	V_b	Volume of shell-like tissue bulging into the walnut	mm^3
Protruding Vol	V_p	Volume of shell-like bits that protrude into cavity	mm^3
Thickness	T_s	Average thickness of the external section of the shell	mm
Volume ratio	V_s/V_w	Percentage of shell with respect to the whole walnut	‰
External ratio	V_e/V_s	Percentage of the external section of the shell	‰
Bulging ratio	V_b/V_s	Percentage of shell-like tissue bulging into the walnut	‰
Protruding ratio	V_p/V_s	Percentage of tissue protruding into the cavity	‰
Kernel			
Length	L_k	Distance perpendicular to the transverse plane	mm
Width	W_k	Distance perpendicular to the longitudinal plane	mm
Height	H_k	Distance perpendicular to the lateral plane	mm
Surface	A_k	Surface area of the actual kernel	mm^2
Convex surface	A_{ck}	Surface area of the convex hull of the kernel	mm^2
Volume	V_k	Total volume of the actual kernel	mm^3
Convex volume	V_{ck}	Volume of the convex hull of the kernel	mm^3
Volume ratio	V_k/V_w	Percentage of kernel with respect to the whole walnut	‰
Convex area ratio	A_{ck}/A_k	Index of nonconvexity	*
Convex vol ratio	V_k/V_{ck}	Index of nonconvexity	*

Table 5.1 (cont'd)

Trait	Exp	Description	Unit
Packing tissue			
Volume	V_t	Total volume of the packing tissue	mm^3
Volume ratio	V_t/V_w	Percentage of packing tissue	‰
Air			
Volume	V_a	Total volume of the air contained by the nut	mm^3
Volume ratio	V_a/V_w	Percentage of air with respect to the whole walnut	‰

extracted, with lower numbers representing easier removal. At the same time, the physical walnuts were cracked open and qualitative score was assigned to the ease of removal, shell integrity, shell strength, and shell texture. We computed Kruskal-Wallis one-way analyses (Kruskal and Wallis, 1952) to determine which morphological phenotypes contribute the most to differentiate ease of removal and shell strength scores. These results follow a conservative 10^{-10} false discovery rate after performing a multiple test Benjamini-Hochberg correction (Benjamini and Hochberg, 1995). A similar analysis of variation was performed to determine the morphological traits that are the most different between one walnut accession and the rest of the scanned collection.

5.2 Results

Our scanned walnut collection reported overall stable values of sphericity indices, and surface area and volume ratios with respect to their convex hull. This suggests that walnuts by and large have similar overall shell shape, rugosity, and lobeyness. More specifically, we observe a very high ratio of total nut volume to convex hull volume (0.95±0.01), so overall nut is very close to being convex. However, nuts also exhibit a much lower ratio of total surface area to convex hull area (0.63±0.01), which indicates that their shell surface is covered by numerous, thin grooves. At the same time, walnuts reveal a especially large variability of shell tissue. In particular, the amount of shell tissue that protrudes into the walnut inner cavity ranges from 0 to 206 mm^3. This corresponds to a QCD of almost 0.5, where the 75th quartile (41.5 mm^3) is almost 3 times as large as the 25th one (15.1 mm^3). There is a similar large variation of reported values of shell tissue that bulges inwards (Table 5.2).

Table 5.2: Morphological trait values. Standard deviation, 25th quartile, 75th quartile, quartile coefficient of dispersion, and coefficient of variance are indicated by SD, $\mathbf{Q_{25}}$, $\mathbf{Q_{75}}$, QCD, and CV respectively. Traits sorted by QCD. Asterisk denotes not applicable.

Trait	**Units**	**Mean + SD**	**Range**	$\mathbf{Q_{25}}$	$\mathbf{Q_{75}}$	**QCD**	**CV**
Nut Cvex Vol Ratio	*	0.95 ± 0.01	0.86 — 0.98	0.95	0.97	0.01	0.02
Nut Cvex Area Ratio	*	0.63 ± 0.01	0.49 — 0.66	0.63	0.64	0.01	0.02
External Shell Ratio	%	0.94 ± 0.03	0.81 — 0.99	0.93	0.96	0.02	0.03
Shell Rugosity	*	1.66 ± 0.06	1.56 — 2.30	1.62	1.69	0.02	0.03
Nut Sphericity	*	0.60 ± 0.02	0.44 — 0.64	0.59	0.62	0.02	0.03
Sneed Index	*	0.92 ± 0.03	0.83 — 1.00	0.90	0.94	0.02	0.03
Krumbein Index	*	0.90 ± 0.05	0.75 — 1.00	0.87	0.93	0.04	0.05
Kernel Cvex A Ratio	*	0.38 ± 0.02	0.30 — 0.53	0.36	0.40	0.04	0.06
Nut Feret Ratio	*	1.21 ± 0.10	1.00 — 1.56	1.14	1.27	0.05	0.08
Kernel Cvex V Ratio	*	0.56 ± 0.05	0.38 — 0.69	0.54	0.60	0.05	0.09
Nut Width	mm	32.1 ± 2.87	23.1 — 45.4	30.2	33.7	0.06	0.09
Nut Height	mm	33.4 ± 2.94	25.6 — 44.6	31.4	35.2	0.06	0.09
Nut VA3D	*	4.59 ± 0.50	3.78 — 12.2	4.27	4.83	0.06	0.11
Kernel Height	mm	28.2 ± 2.72	20.8 — 39.9	26.4	29.9	0.06	0.10
Kernel Length	mm	30.4 ± 2.97	18.2 — 40.4	28.4	32.4	0.06	0.10
Kernel Width	mm	24.8 ± 2.8	15.7 — 39.8	23.0	26.4	0.07	0.11
Nut Length	mm	38.5 ± 4.10	26.3 — 53.2	35.7	41.2	0.07	0.11
Packing Vol Ratio	%	0.13 ± 0.02	0.07 — 0.23	0.11	0.14	0.08	0.15
Kernel Vol Ratio	%	0.34 ± 0.05	0.18 — 0.48	0.31	0.37	0.09	0.15
Nut Cvex Area	mm^2	3675 ± 609	2057 — 6067	3257	4044	0.11	0.17
Kernel Cvex Area	mm^2	2572 ± 427	1277 — 3934	2281	2847	0.11	0.17
Nut Area	mm^2	5799 ± 1011	3316 — 9978	5116	6401	0.11	0.17
Air Vol Ratio	%	0.37 ± 0.07	0.14 — 0.61	0.32	0.41	0.12	0.18
Shell Thickness	mm	0.88 ± 0.16	0.51 — 1.57	0.77	0.99	0.12	0.18
Kernel Area	mm^2	6773 ± 1276	2827 — 10991	5905	7595	0.13	0.19
Shell Vol Ratio	%	0.16 ± 0.03	0.09 — 0.33	0.14	0.18	0.13	0.20
Nut Volume	mm^3	19560 ± 4757	7905 — 41132	16300	22235	0.15	0.24
Kernel Vol	mm^3	6565 ± 1589	2087 — 12232	5542	7650	0.16	0.24
Nut Cvex Vol	mm^3	20511 ± 5102	8470 — 43346	16932	23441	0.16	0.25
Packing Vol	mm^3	2476 ± 676	954 — 6321	2032	2815	0.16	0.27
Kernel Cvex Vol	mm^3	11680 ± 2900	4007 — 21761	9710	13501	0.16	0.25
External Shell Vol	mm^3	3001 ± 811	1020 — 6369	2418	3501	0.18	0.27
Shell Volume	mm^3	3188 ± 886	1139 — 7446	2568	3724	0.18	0.28
Air Volume	mm^3	7332 ± 2628	2126 — 21812	5520	8681	0.22	0.36
Bulge Shell Ratio	%	0.05 ± 0.02	0.00 — 0.18	0.03	0.06	0.32	0.52
Prot Shell Ratio	%	0.01 ± 0.01	0.00 — 0.07	0.01	0.01	0.41	0.81
Bulge Shell Vol	mm^3	155 ± 108	9.94 — 1073	81.2	199	0.42	0.70
Prot Shell Vol	mm^3	32.5 ± 26.8	0.00 — 206	15.1	41.5	0.47	0.82

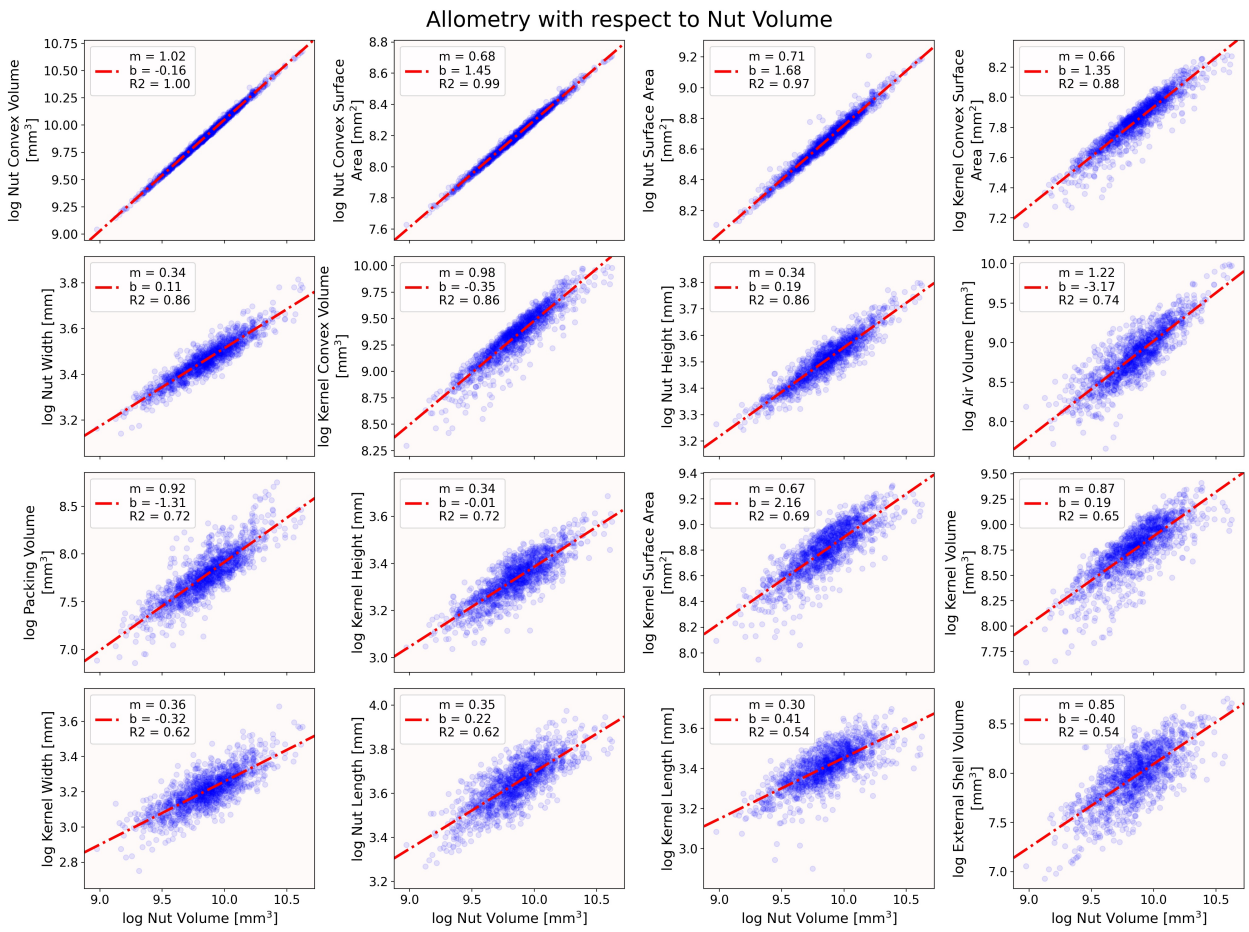

Figure 5.2: Various allometry plots between different logarithmic values of tissue volumes, areas, and lengths compared to the total walnut volume. An ordinary least squares linear model was computed for each case. The slope, intercept, and coefficient of determination for each linear model is indicated by m, b, and R^2 respectively.

We observe that most of the size-specific traits follow power laws with respect to the total nut volume V_w, as our allometric log-log plots exhibit large R^2 coefficients of determination (Figure 5.2). The size-related measurement that exhibits the most superlinear growth rate is the total air volume contained inside the nut V_a. Our data suggests that these two volumes follow the power law $V_a \approx \exp(-3.17)V_w^{1.22}$. That is, as the nut volume increases, biophysical constraints require that the air volume increases by a larger factor. However, the air volume must always be lower than the total nut volume. Evaluating the extreme case of a hypothetical walnut consisting entirely of air, we find that $V_w = \exp(-3.17)V_w^{1.22}$ when $V_w \approx 2.3 \times 10^6$ mm^3. This is the same volume of a 16cm diameter sphere. We also highlight that with a R^2 coefficient of determination very close to 1, the volume

of the convex hull of the nut V_{cw} follows a superlinear growth rate with respect to V_w, the power law $V_{cw} = \exp(-0.16)V_w^{1.02}$. Do notice that the constant factor, $\exp(-0.16) = 0.85$ is less than 1, so that if $V_w < 2063\text{mm}^3$, then $V_{cw} < V_w$, which is impossible. In other words, our allometric power law only holds for walnuts whose total volume is comparable to that of a 1.6cm diameter sphere. This in turn suggests interesting biophysical growth patterns at the early developmental stages of the walnut (Pinney and Polito, 1983; Zhao et al., 2016). Of important note is the fact that kernel volume grows at a slightly sublinear rate with respect to total nut volume, with a power law $V_k = \exp(0.19)V_w^{0.87}$. For instance, if the total nut volume is duplicated, then the kernel volume will only increase by a factor of $2^{0.87} \approx 1.8$, which already indicates that larger walnuts tend to have smaller kernel percentages, while they also tend to contain higher air percentages.

The last observation is also supported by a high, negative Spearman correlation index (-0.78) between the kernel and air volume ratios , and a smaller, negative index (-0.25) with nut volume. We also observe that kernel ratio is positively correlated with its convex volume inverse ratio (0.84) while negatively correlated with the convex area ratio (-0.65). There is also a small positive correlation with walnut sphericity (0.25) and walnut convex volume inverse ratio (0.23) (Figure 5.4). This implies that walnuts that are smaller in volume, with smoother shells, and less grooves on their surface tend to have a larger percentage of kernel with respect to its total size. This observation agrees with Bernard et al. (2020). At the same time, kernel percentage is higher when its overall shape is more convex and it present numerous deep but thin grooves. (Figure 5.4). For shell thickness, we observe unsurprising high correlations with shell volume percentage (0.91), shell total volume (0.78), and other shell-related measurements. There are also negative correlations with air percentage (0.47) and nut sphericity (-0.20), which implies that nuts with thinner shells tend to have smoother shells but a higher content of air. (Figure 5.4).

The morphological traits that explained the most variance for ease of removal were shell thickness and shell, air, and packing tissue volume percentage. The amount of shell-like tissue that bulges into the walnut cavity also appears to influence the ease of removal, as well as the volume and surface area ratio of the kernel with respect to its convex hull. In general, unbroken kernel

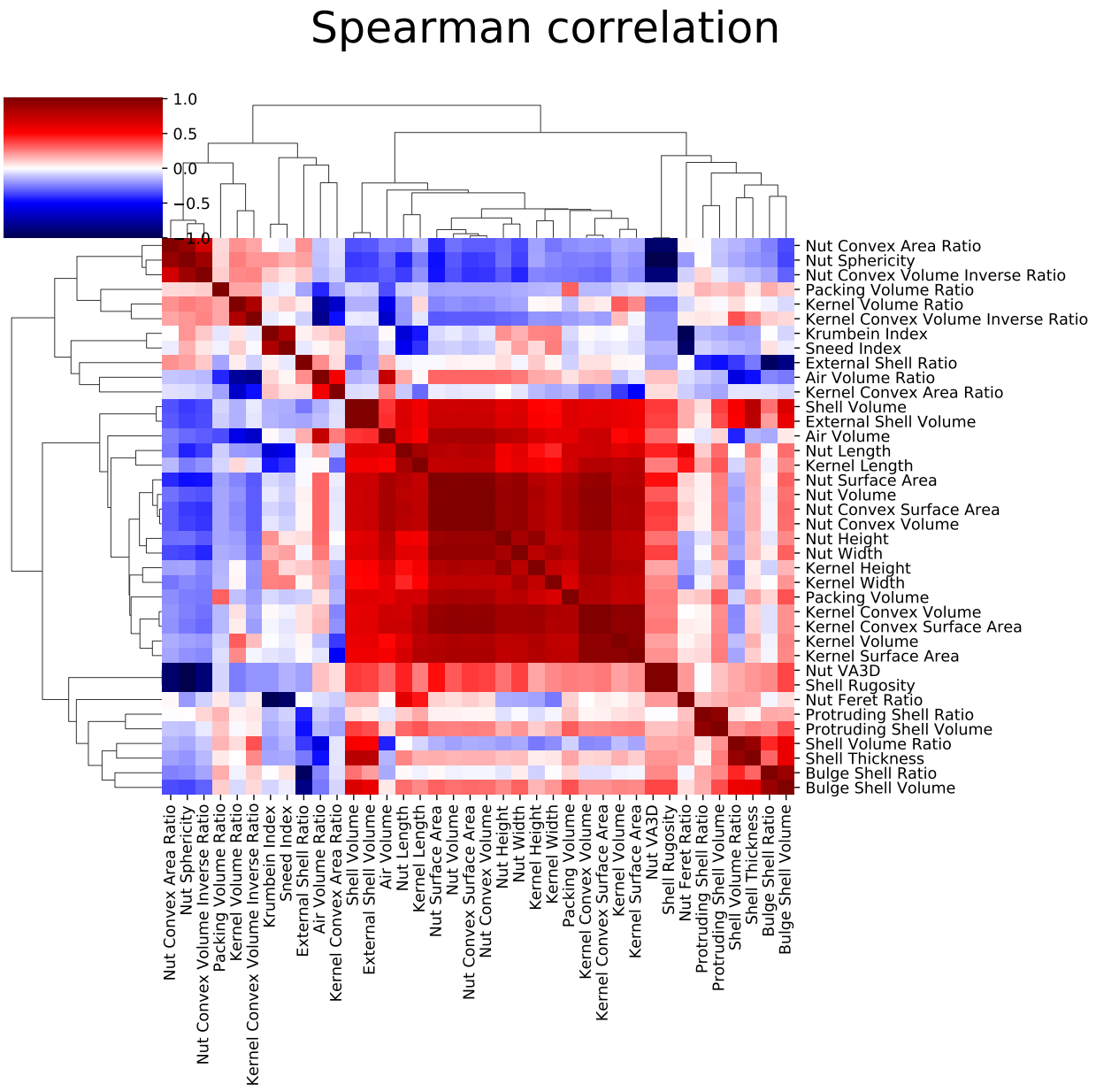

Figure 5.3: Spearman correlation for all phenotypes. Most of the overall nut size-related traits are only positively correlated with kernel size-related traits as expected. All the shell-specific traits are only positively correlated between themselves. All the sphericity, aspect ratio, and rugosity are highly correlated only among themselves.

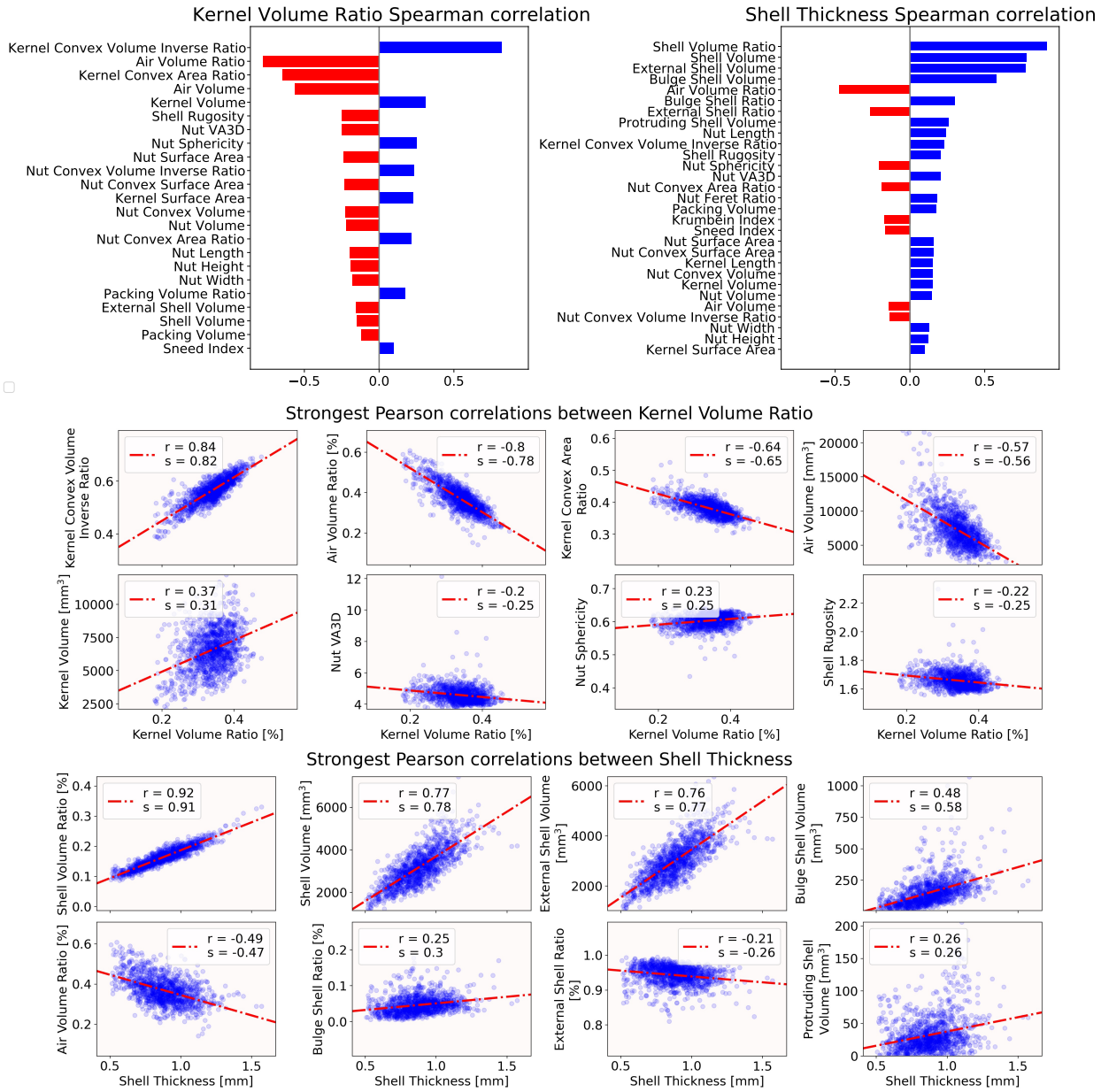

Figure 5.4: Highest Spearman correlation traits for kernel volume ratio and shell thickness. (A) All the correlation coefficients highlighted in the barplots are statistically significant (p-value $> 10^{-3}$). (B) Different traits versus kernel filling ratio and (C) shell thickness. The Pearson and Spearman correlation coefficients are denoted by r and s respectively.

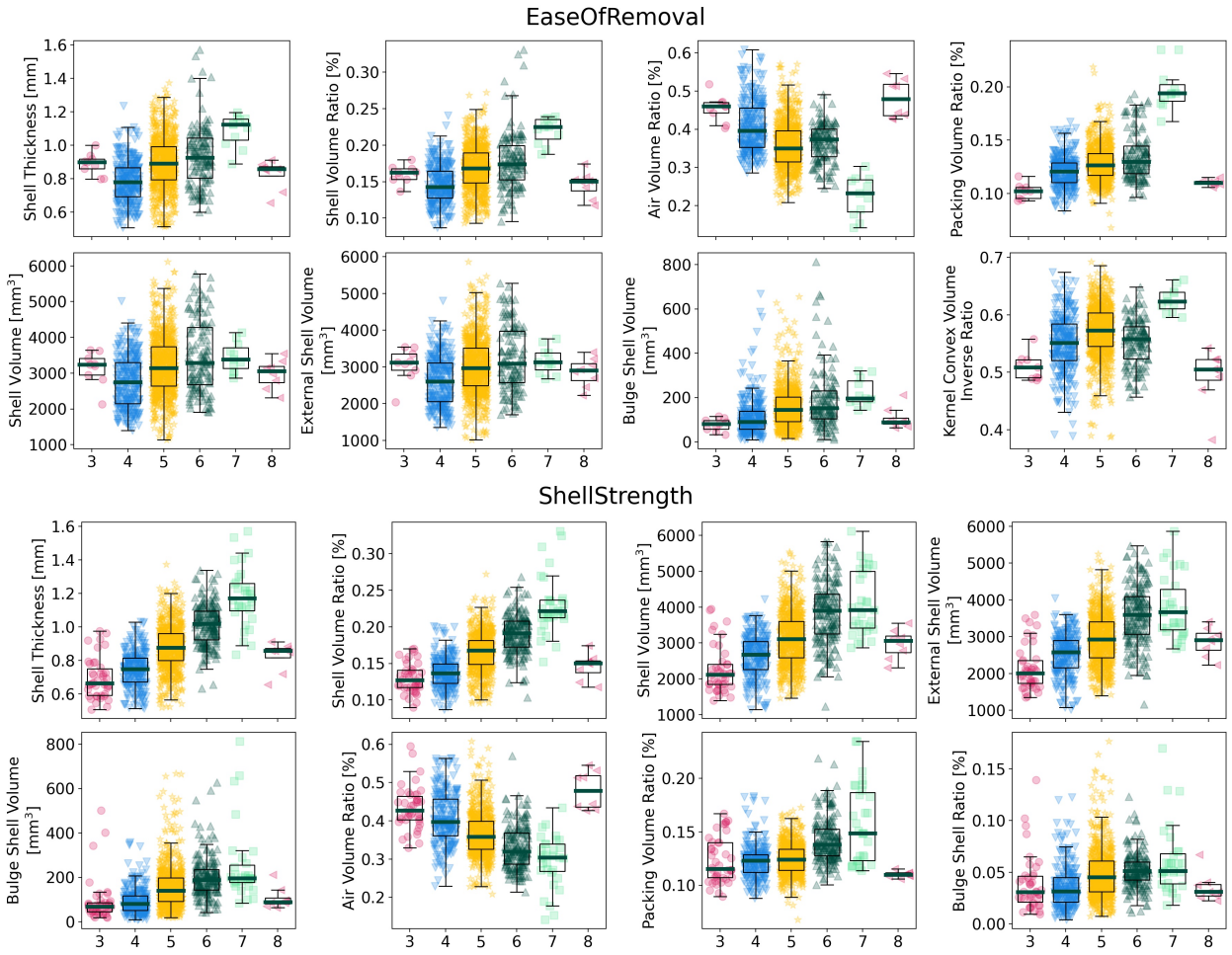

Figure 5.5: Morphological traits that explain the most variance across different qualitative groups according to Kruskal-Wallis analyses. (A) Boxplots for different trait values according to the ease of kernel removal and (B) shell strength of the walnut. Both qualitative traits are scored in an increasing scale 3–8, where 3 indicates the easiest walnut to remove both kernel halves intact and the least strong shell. All the highlighted traits are statistically significant after a 10^{-10} false discovery rate correction.

halves are easier to remove for walnuts with thinner shells, with relatively little external and internal shell and packing tissue content. The kernels are easier to extract when the nut contains a higher percentage of air, and when they have deeper and wider grooves (Figure 5.5). For shell strength, the morphological traits that explained it the most were unsurprisingly related to shell thickness, volume, or relative percentage. We also noticed that shell is stronger for walnuts with lower relative content of air and higher content of packing tissue. Up to a point, shell strength also seems to be affected by kernel volume and walnut length (Figure 5.5).

For both ease of removal and shell strength, we notice that in general the Himalayan Earliest accession (the only one scoring **8** for both evaluations) consistently breaks a visual trend followed by the rest of accessions and scores. Comparing the morphological features distributions of this Himalayan accession with the rest of the collection reveals that this Himalayan accession has on average a larger percentage of air content and a lower percentage of kernel and packing tissue volume. This accession on average also has a higher kernel area ratio and lower kernel volume inverse ratio with respect to its convex hull. This indicates that kernels of Himalayan accession tend to have wider, deeper cavities, which most likely are filled with air. We did not find significant distribution differences for the rest of morphological traits between the Himalayan accession and the rest of the collection.

5.3 Discussion

The diversity, propagation, and diffusion of walnut populations offer an important window to past climates and civilizations. Current genome sequencing data suggests that the common walnut originated as an ancient hybrid in the late Pliocene, 3.45 million years ago, as a cross between American and Asian *Juglans* lineages (Zhang et al., 2019). However, the Last Glacial Maximum extinguished most of the walnut populations except for some located in select glacial refugia 18 thousand years ago (Aradhya et al., 2017). This cataclismic event imposed a severe bottleneck effect on the walnut germplasm, reducing dramatically its effective population size (Ding et al., 2022). These refugia were mostly separate pockets surrounded by mountainous terrain between Southwestern China, the Qinghia-Tibet Plateu, and the Himalayas, regions that represent the core of walnut genetic diversity (Luo et al., 2022). From there, humans propagated walnut populations through trade following the Silk Road, dispersing the walnut all the way from the Iberian peninsula to Southeastern China (Beer et al., 2008). There is plenty of observed phenotypic diversity within a fixed walnut population, but surprisingly these variations are usually not sufficient to distinguish geographically distant populations (Roor et al., 2017). It could be that morphological differences across regions are too subtle to be captured by simple caliper measurements.

This difficulty to comprehensively measure walnut morphology is more pressing when trying

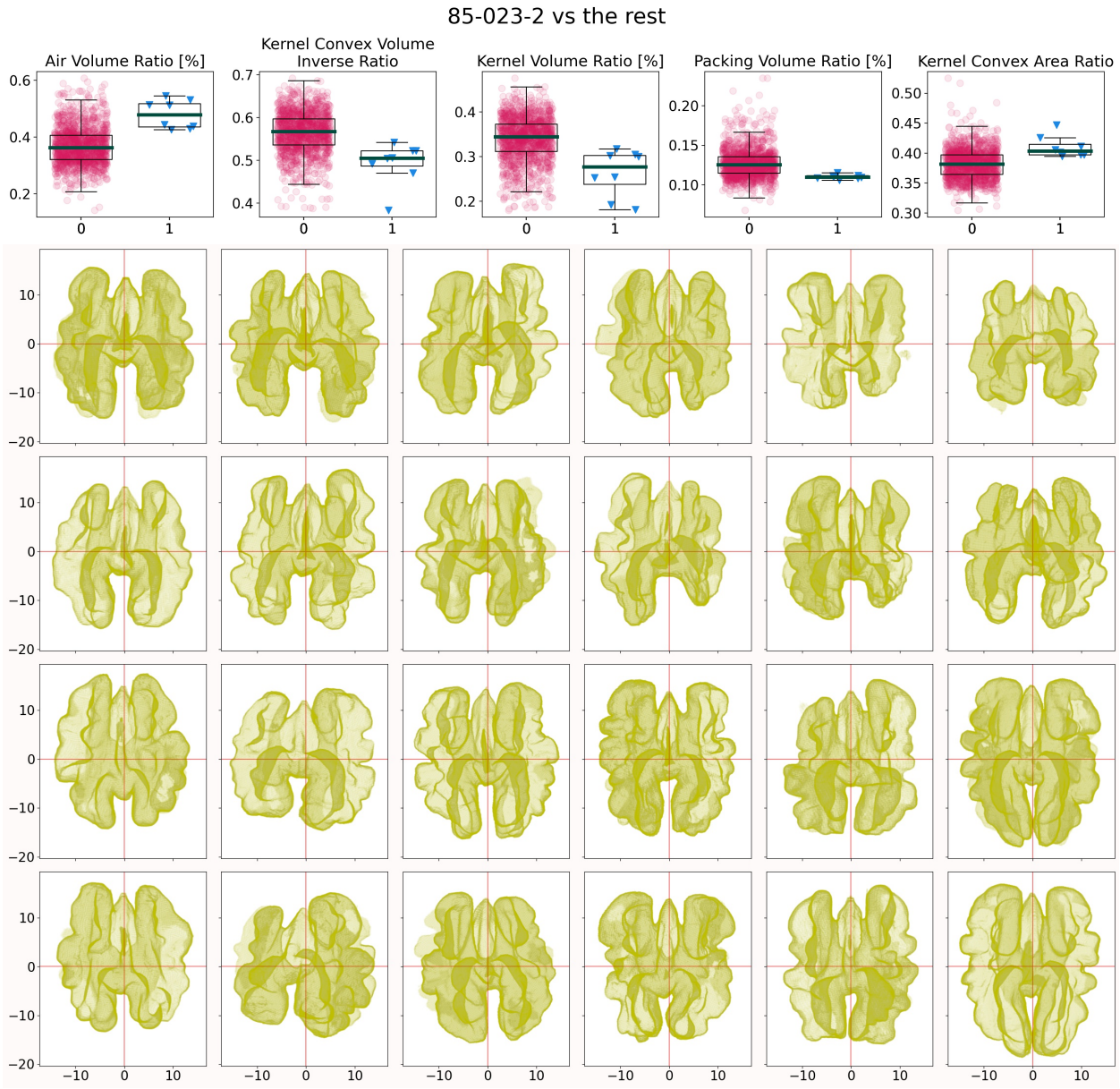

Figure 5.6: Main morphological differences between the Earliest Himalayan accession (**1**) and the rest of the collection (**0**). (A) Boxplots of morphologhical traits that are the most different for the Earliest accession. (B) Longitudinal plane view of kernel halves. The top two rows depict Earliest kernels, while the bottom two rows depict other accessions. Notice that the bottom cavity tends to be deeper and wider for the Earliest accession. All the highlighted traits are statistically significant according to Kruskal-Wallis analyses of variance after a 10^{-5} false discovery rate correction.

to understand the fine-grained details that determine important traits of commercial concern (Du and Tan, 2021). One such trait is the ease of removal of the kernel, how easy is to remove the main two halves of walnut intact. A related trait is the shell strength and the fracture patterns suffered by shells under pressure (Gülsoy et al., 2019). Multiple-pronged strategies have been proposed and developed to unravel the underlying mechanisms that regulate shell thickness, shape, and strength. Research on these traits has been approached from a genetics standpoint, such as quantitative trait loci (QTL) and genome-wide association studies (GWAS) to identify transcription factors and pathways that affect the shell and seal formation (Sideli et al., 2020; Wang et al., 2022). Walnut shell behavior has been explored with numerical simulations, where walnuts are modeled as thin spheres and biophysical mechanical properties are tested under unidirectional loads based on finite-element analyses (Bao et al., 2022; Koyuncu et al., 2004). Recent exciting work has focused on the shape of the polylobate sclerid individual cells that tesselate and conform the walnut shell while forming intricate puzzles that confer remarkable toughness and strength (Antreich et al., 2019; Zhang et al., 2014). This approach is then complemented with biochemical techniques that seek to unravel the structural and compositional changes during walnut shell development, as individual cells go from a soft to hard state (Antreich et al., 2021; Xiao et al., 2020).

X-ray CT scans allows us to accurately extract more nuanced shape and size features from our sampled walnuts, providing new avenues to explore subtle morphological changes and implications. For example, with more size-related features, we can compute better allometric relationships that point to biophysical constraints in walnut growth development. In particular, we are able to draw theoretical upper and lower bounds on walnut size. The growth of empty space within a walnut outpaces the overall nut growth rate, which indicates that larger walnuts tend to contain a higher proportion of air. At the same time, we observed that for small nuts, their total volume was almost identical to the volume of their convex hulls. This implies that smooth, groove-free nuts must be small. Moreover, this allometric relationship only holds for nut that are larger than a certain size, which indicates that the growth dynamics of the nut undergo a regime change as the nut develops. We also observe that the kernel volume grows at a slower rate than the total nut volume, which

suggests that larger walnuts tend to contain less kernel tissue (Figure 5.2). Walnuts with more relative kernel content tend to be smoother but the kernels themselves present more grooves that are deeper and narrower. Walnuts with thinner shells tend to have smoother shells but a higher content of air (Figure 5.4). All of the allometric observations and correlations above suggest that walnut and kernel sizes and smoothness are not just dependent on genes and environment, but there are also unavoidable biophysical constraints at play that should be explored further and considered by breeding programs (Niklas and Hammond, 2019).

Our extended list of measured phenotypes also offers new insight to qualitatively assessed traits. We highlight traits related to packing tissue, contained air, or convexity ratios, as they are difficult or impossible to measure with traditional tools. To the best of our knowledge, this is the first time that such traits are completely quantified. Packing tissue is commercially relevant, as its filling ratio seems to affect the ease of kernel removal and shell strength. Nuts with high relative content of packing tissue tend to have stronger shells and present more difficulties to remove kernels (Figure 5.5). This confirms previous reports on moderate correlations between packing tissue thickness and ease of kernel removal (Fallah et al., 2022; Kouhi et al., 2020). The convexity of kernel also appears to play an important role in the cracking mechanics of the nut. The hardest walnut to crack in the collection reported very average size and shape measurements except for subtle differences in the kernel's main cavity at the proximal end. This cavity might act as a clamp under pressure, which could explain the harness to crack this particular accession open. This might suggest that the easiness of kernel removal and shell integrity might not be solely dependent on shell thickness and volume, but also on shape characteristics of the rest of nut tissues. Moreover, this uniquely clamp-shaped kernels are closely related to wild type accessions from the Himalayas, one of the areas with the highest diversity of walnut germplasm (Shah et al., 2021).

Walnuts offer a especially unique opportunity to analyze domestication in perennial crops. Despite their long history with humans, current research suggests that walnut domestication happened less than 100 years ago (Mapelli et al., 2018). Even today, due to economical and horticultural reasons, walnut is propagated through via seeds and not grafting throughout most of Southwest

Asia (Rezaee et al., 2008; Thapa et al., 2021). This makes walnut an exciting organism to study the immediate effects of domestication and breeding in real time across multiple populations. A careful, nuanced study of kernel morphology might provide us key insights into domestication-induced morphological changes, and accelerate the selection of progenitors in breeding programs.

CHAPTER 6

CONCLUDING REMARKS

All the plants out here are malevolent, heavy and sharp. The parts of the palms above the fronds are tufted in sick stuff like coconut-hair. Roaches and other things live in the trees. Rats, maybe. Loathsome high-altitude critters of all kinds. All the plants either spiny or meaty. Cacti in queer tortured shapes. The tops of the palms like Rod Stewart's hair, from days gone by.

—from *Infinite Jest*

David Foster Wallace

6.1 Conclusion

In this dissertation we presented three different project, each with a novel way to comprehensively encode and compare the diverse morphology found in the plant biology domain. The three of our applications focus on the quantification of shape based from high-resolution 3D X-ray CT scan reconstructions. The non-destructive and thorough nature of X-rays allowed us to fully extract barley seeds from their panicles without worrying about occlusion; it allowed us to examine different fruit tissues from citrus; and it was key to measure the walnut kernels before cracking the shell open. Traditional and modern morphometrics has a number of drawbacks with respect to X-ray CT images. Landmark-based morphometrics requires homologous points and, although 3D and higher dimensional analysis is possible, it is usually applied to 2D images (Dryden and Mardia, 2016). Further, a geometric framework is limited to the relationship of data points to each other. Landmark-based approaches reduce the shape information to a relatively small and possibly subjective collection of points, which can be further restricted if there are no obvious homologous landmarks across all samples. Fourier-based outlines are limited to the analysis of 2D images and are not suitable for inputs in higher dimensions. We thus turned to algebraic topology and directional statistics.

Our results on the quantification of barley seed morphology shows that the Euler characteristic is a simple yet powerful way to reveal features not readily visible to the naked eye. There is "hidden" morphological information that traditional and geometric morphometric methods are missing. The Euler characteristic, and Topological Data Analysis (TDA) in general, can be readily computed from any given image data, which makes it a versatile tool to use in a vast number of biology-related applications. TDA provides a comprehensive framework to detect and compare morphological nuances, nuances that traditional measures fail to capture and that remain unexplored using simple geometric methods. In the specific case of barley seeds presented here, these "hidden" shape nuances provide enough information to not only characterize specific accessions, but the individual spikes from which seeds are derived. Our results suggest a new exciting path, driven by morphological information alone, to explore further the phenotype-genotype relationship.

TDA is just one of the novel mathematical domains that can be used to further our biological insight. There is rich shape information in the natural world to be captured, analyzed, and linked to biophysical developmental and evolutionary principles. Sound mathematical models are key to uncover these biophysical interactions at work. For example, even with a limited number of points, overall fruit shapes can be approximated with various quadratic surfaces like ellipsoids. Given the appropriate parameters, an ellipsoid can represent both nearly spherical navel oranges, and elongated finger limes. This quadratic surface approximation is mathematically versatile and computationally simple, and can be applied to other round-shaped biologically-motivated data. Moreover, ellipsoid coordinates can be translated naturally to longitudes and latitudes on a sphere, which opens the door to a wide array of mathematical tools from directional statistics. Some of those tools, like density estimations, hypothesis testing, and distribution fitting, allow us to quantify shape in a mathematically rigorous and comprehensive way.

Our results for each of the three projects presented suggest extremely exciting future research, in both mathematical and biological lines.

6.1.1 Further exploration with ECT

The ECT has proved to be an extremely powerful morphological descriptor of grain shape. There are a number of topics to explore with respect to the implementation of this novel approach, both theoretically and empirically. For instance, exploring how our predictions might change if we pick uniformly randomly distributed directions —or according to any other probability distribution— instead of polar-biased ones. There is also more research to be done into alternatives to determine which 3D shape features are the most relevant to distinguish inter-accession characteristics. One possibility is running our results through established statisitical pipelines like SINATRA to uncover such features (Wang et al., 2021). We can also explore how our barley classification results might change with variants of ECT, such as the Smooth ECT (Crawford et al., 2020), Weighted ECT (Jiang et al., 2020), or Euler Characteristic Surfaces (Beltramo et al., 2022).

6.1.2 Brewing barley genomics into the topological party

There is extensive literature to understand the underlying genetic mechanisms that allow barley's tremendous versatility (Hockett and Nilan, 1985; Mascher et al., 2017; Sato, 2020). The combination of genomics with archaeology has revealed important patterns of its domestication and its intimate relationship with ancient civilizations (Mascher et al., 2016; Russell et al., 2016). The historical and geographical diversity makes barley an ideal organism to understand the genetic adaptations to tillering (Komatsuda et al., 2007), UV intensity, changes in sunlight availability, and flowering time (Dawson et al., 2015), and more importantly, to understand how these adaptations relate to total grain weight, which is a crucial trait to develop better cereal crops (Liller et al., 2015).

For our historical composite cross barley population, we have access to genome-wide genotypic data for 850 progeny using a RADseq reduced representation approach (Baird et al., 2008), allowing us to score over 200,000 polymorphic sites in each individual. RADseq has proved to be a powerful approach to sequence organisms with large genomes with limited sequence data. Combined with high-density linkage maps, RADseq has been able to map QTLs responsible for chlorophyl-related traits in soybean (Wang et al., 2020), fiber quality in hemp (Petit et al., 2020), and agronomic traits (such as panicle length and average grain weight) in grasses such as foxtail millet (Wang et al.,

2017). RADseq has been especially powerful when it comes to study how a particular trait has evolved among large and diverse populations (Davey and Blaxter, 2011). RADseq combined with genome-wide association studies (GWAS) has unveiled important polymorphic regions responsible for caffeine content during the domestication of multiple tea plants (Yamashita et al., 2020), seed size during domestication of soybean (Zhou et al., 2015), and has further our knowledge of the population structure of sesame accessions across four continents (Basak et al., 2019).

There is a fine relationship between barley genome and barley morphology, especially about the specifics on the domestication of barley shape. Important ongoing work explores patterns and relationships between our computed topological signatures and different genetic loci across multiple filial generations using RADseq reads, heritability analyses, and GWAS. Current preliminary results have already yielded candidate loci that can be further explored for synteny among other barley accessions and even among other related grasses like wheat or rye.

6.1.3 If life gives you lemons, determine distances between their oil gland distributions

As discussed at the end of Chapter 4, directional statistical tests strongly reject the hypothesis that citrus oil glands follow a uniform, rotationally symmetric, or any other well-studied distribution. An alternative to mathematically characterize oil gland distributions is to turn to non-parametric approaches based on spherical kernel density estimators (Di Marzio et al., 2019; Vuollo and Holmström, 2018). Then we can measure numerically the distance between two nonparametric spherical density functions (Boente et al., 2014) and compute a pairwise distance matrix. Another alternative to comprehensively describe oil gland distributions is to compute persistent homology directly on the point cloud defined by the oil gland centers (Figure 4.4). This way we do not have to resort to deformations induced by ellipsoidal approximations. Once every fruit oil gland distribution is summarized as a persistence diagram, there are many well established pipelines to compute a matrix of pairwise distances. Specifically, we can compute bottleneck distances between diagrams, or transform such diagrams into more mathematically amenable objects like persistence landscapes (Bubenik, 2015), persistence images (Adams et al., 2017), or tent functions (Perea et al., 2022; Tymochko et al., 2019) to name a few examples. In either case, we can then compare

citrus phylogenetic distances with oil gland distribution distances. This would ultimately provide us with novel insight into fruit development.

There is also future work to be done in terms of citrus breeding. We can link our morphological characterization of citrus oil gland distribution with commercially important traits, such as acidity, sweetness, skin response to mechanical injuries, or amount of oil extracted to name a few examples.

6.1.4 The morphology of domestication: a hard nut to crack

As discussed in Chapter 5, we were initially puzzled by the fact that Earliest Himalayan accession initial results. It was the hardest accession to crack open and it is essentially impossible to extract their kernel intact. However, their morphological trait values are close to the population average, which clearly breaks visual trends when relating shape phenotypes with qualitative data (Figure 5.5). Moreover, current literature highlights that individual phenotype variation of walnuts is high within a given population, but not so much between physically distant populations (Mapelli et al., 2018; Roor et al., 2017). At the same time, the Himalayas are reported to be one of the main hotspots for walnut germplasm diversity (Luo et al., 2022; Shah et al., 2021), while the rest of the collection is mainly homozygous (Aradhya et al., 2010). A closer look at trait distributions reveals that this Earliest Himalaya accession can be distinguished from the rest of the collection by relatively lower content of packing tissue, a relatively higher content of air, and distinct convexity indices (Figure 5.6). Preliminary results suggest that the Earliest accession kernel has a distinctive wide, deep cavity at its proximal end. To the best of our understanding, this specific traits, packing tissue volume and convexity indices, have never been measured. The cavity at the proximal end offers new research directions aimed at unraveling its specific development mechanics from both genetical and biophysical points of view.

This new cavity also suggests that it is possible to characterize the morphology of physically separate populations. We just need more comprehensive and fine-tuned morphological descriptors. Walnut kernels have tortuous, polylobed morphologies that no set of traditional shape descriptors will every fully capture. An exciting future direction is to analyze these intricate shapes with TDA, specifically with the ECT like in Chapter 3 or a similar technique. This in turn might shed new light

into broader morphological changes for perennial crops in general when they are domesticated.

In general, capturing and analyzing this nuanced shape information for a wide array of data sets provides a morphology-driven path to further our insight into phenotype-genotype relationships. As stated by D'Arcy Thompson in his seminal biomathematical treatise *On Growth and Form* (1942),

> *An organism is so complex a thing, and growth so complex a phenomenon, that for growth to be so uniform and constant in all the parts as to keep the whole shape unchanged would indeed be an unlikely and an unusual circumstance. Rates vary, proportions change, and the whole configuration alters accordingly.*

BIBLIOGRAPHY

Adams H, Emerson T, Kirby M, Neville R, Peterson C, Shipman P, Chepushtanova S, Hanson E, Motta F, Ziegelmeier L (2017). Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research* **18**(8): 1–35.

Akca Y, Şen SM (1995). The relationship between dichogamy and yield–nut characteristics in *Juglans regia* L. *In Acta Horticulturae*, Volume 442, pp. 215–216. Leuven, Belgium: International Society for Horticultural Science (ISHS).

Amézquita EJ, Nasrin F, Storey KM, Yoshizawa M (2022). Genomics data analysis via spectral shape and topology. Preprint.

Amézquita EJ, Quigley MY, Ophelders T, Landis JB, Koenig D, Munch E, Chitwood DH (2021). Measuring hidden phenotype: quantifying the shape of barley seeds using the Euler characteristic transform. *in silico Plants* **4**(1): diab033.

Amézquita EJ, Quigley MY, Ophelders T, Munch E, Chitwood DH (2020). The shape of things to come: Topological data analysis and biology, from molecules to organisms. *Developmental Dynamics* **249**(7): 816–833.

Amézquita EJ, Quigley MY, Ophelders T, Seymour D, Munch E, Chitwood DH (2022). The shape of aroma: Measuring and modeling citrus oil gland distribution. *Plants, People, Planet* **0**: 1–14.

Andrade-Sanchez P, Gore MA, Heun JT, Thorp KR, Carmo-Silva AE, French AN, Salvucci ME, White JW (2013). Development and evaluation of a field-based high-throughput phenotyping platform. *Functional Plant Biology* **41**(1): 68–79.

Antreich SJ, Xiao N, Huss JC, Gierlinger N (2021). A belt for the cell: cellulosic wall thickenings and their role in morphogenesis of the 3D puzzle cells in walnut shells. *Journal of Experimental Botany* **72**(13): 4744–4756.

Antreich SJ, Xiao N, Huss JC, Horbelt N, Eder M, Weinkamer R, Gierlinger N (2019). The puzzle of the walnut shell: A novel cell type with interlocked packing. *Advanced Science* **6**(16): 1900644.

Aradhya M, Velasco D, Ibrahimov Z, Toktoraliev B, Maghradze D, Musayev M, Bobokashvili Z, Preece JE (2017). Genetic and ecological insights into glacial refugia of walnut (*Juglans regia* L.). *PLOS ONE* **12**(10): 1–27.

Aradhya M, Woeste K, Velasco D (2010). Genetic diversity structure and differentiation in cultivated walnut (*Juglans Regia* L.). *In Acta Horticulturae*, Number 861, pp. 127–132. International Society for Horticultural Science (ISHS), Leuven, Belgium.

Araus JL, Cairns JE (2014). Field high-throughput phenotyping: the new crop breeding frontier. *Trends in Plant Science* **19**(1): 52–61.

Atienza N, Escudero LM, Jimenez MJ, Soriano-Trigueros M (2019). Characterising epithelial tissues using persistent entropy. *In* Marfil R, Calderón M, Díaz del Río F, Real P, Bandera A (Eds.), *Computational Topology in Image Context*, pp. 179–190. Cham: Springer International Publishing.

Atkinson JA, Pound MP, Bennett MJ, Wells DM (2019). Uncovering the hidden half of plants using new advances in root phenotyping. *Current Opinion in Biotechnology* **55**: 1–8. Analytical Biotechnology.

Autran D, Bassel GW, Chae E, Ezer D, Ferjani A, Fleck C, Hamant O, Hartmann FP, Jiao Y, Johnston IG et al. (2021). What is quantitative plant biology? *Quantitative Plant Biology* **2**: e10.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLOS ONE* **3**(10): 1–7.

Bao X, Chen B, Dai P, Li Y, Mao J (2022). Construction and verification of spherical thin shell model for revealing walnut shell crack initiation and expansion mechanism. *Agriculture* **12**(9).

Baron JH (2009). Sailors' scurvy before and after James Lind — a reassessment. *Nutrition Reviews* **67**(6): 315–332.

Barry GH, Caruso M, Gmitter FG (2020). Chapter 5 - Commercial scion varieties. *In* Talon M, Caruso M, Gmitter FG (Eds.), *The Genus Citrus*, pp. 83–104. Woodhead Publishing.

Basak M, Uzun B, Yol E (2019). Genetic diversity and population structure of the Mediterranean sesame core collection with use of genome-wide SNPs developed by double digest RAD-Seq. *PLOS ONE* **14**(10): e0223757.

Bauer U (2021). Ripser: efficient computation of Vietoris–Rips persistence barcodes. *Journal of Applied and Computational Topology* **5**(3): 391–423.

Beer R, Kaiser F, Schmidt K, Ammann B, Carraro G, Grisa E, Tinner W (2008). Vegetation history of the walnut forests in Kyrgyzstan (Central Asia): natural or anthropogenic origin? *Quaternary Science Reviews* **27**(5): 621–632.

Belchi F, Pirashvili M, Conway J, Bennett M, Djukanovic R, Brodzki J (2018). Lung topology characteristics in patients with chronic obstructive pulmonary disease. *Scientific Reports* **8**(5341).

Belton RL, Fasy BT, Mertz R, Micka S, Millman DL, Salinas D, Schenfisch A, Schupbach J, Williams L (2020). Reconstructing embedded graphs from persistence diagrams. *Computational*

Geometry **90**: 101658.

Beltramo G, Skraba P, Andreeva R, Sarkar R, Giarratano Y, Bernabeu MO (2022). Euler characteristic surfaces. *Foundations of Data Science* **4**(4): 505–536.

Bendich P, Edelsbrunner H, Kerber M (2010). Computing robustness and persistence for images. *IEEE Transactions on Visualization and Computer Graphics* **16**(6): 1251–1260.

Bendich P, Marron JS, Miller E, Pieloch A, Skwerer S (2016). Persistent homology analysis of brain artery trees. *Ann. Appl. Stat.* **10**(1): 198–218.

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**(1): 289–300.

Bernard A, Hamdy S, Le Corre L, Dirlewanger E, Lheureux F (2020). 3D characterization of walnut morphological traits using X-ray computed tomography. *Plant Methods* **16**(1): 115.

Bernard A, Lheureux F, Dirlewanger E (2017). Walnut: past and future of genetic improvement. *Tree Genetics & Genomes* **14**(1): 1.

Betthauser LM (2018). *Topological reconstruction of grayscale images*. Ph. D. thesis, University of Florida, Gainesville, Florida.

Blott SJ, Pye K (2008). Particle shape: a review and new methods of characterization and classification. *Sedimentology* **55**(1): 31–63.

Boente G, Rodriguez D, Manteiga WG (2014). Goodness-of-fit test for directional data. *Scandinavian Journal of Statistics* **41**(1): 259–275.

Bonett DG (2006). Confidence interval for a coefficient of quartile variation. *Computational Statistics & Data Analysis* **50**(11): 2953–2957.

Bonhomme V, Forster E, Wallace M, Stillman E, Charles M, Jones G (2017). Identification of inter- and intra-species variation in cereal grains through geometric morphometric analysis, and its resilience under experimental charring. *Journal of Archaeological Science* **86**: 60–67.

Bookstein FL (1997). *Morphometric Tools for Landmark Data: Geometry and Biology*. Geometry and Biology. Cambridge: Cambridge University Press.

Booth S, Kurtz B, de Heer MI, Mooney SJ, Sturrock CJ (2020). Tracking wireworm burrowing behaviour in soil over time using 3D X-ray computed tomography. *Pest Management Science* **76**(8): 2653–2662.

Bouby L (2001). L'orge à deux rangs (*Hordeum distichum*) dans l'agriculture gallo-romaine :

données archéobotaniques. *ArchéoSciences, revue d'Archéométrie* **25**: 35–44.

Bubenik P (2015). Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research* **16**(3): 77–102.

Bucksch A, Atta-Boateng A, Azihou AF, Battogtokh D, Baumgartner A, Binder BM, Braybrook SA, Chang C, Coneva V, DeWitt TJ et al. (2017). Morphological plant modeling: Unleashing geometric and topological potential within the plant sciences. *Frontiers in Plant Science* **8**.

Burges CJ (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* **2**(2): 121–167.

Cámara PG (2017). Topological methods for genomics: Present and future directions. *Current Opinion in Systems Biology* **1**: 95–101. Future of Systems Biology • Genomics and epigenomics.

Cámara PG, Rosenbloom DI, Emmett KJ, Levine AJ, Rabadán R (2016). Topological data analysis generates high-resolution, genome-wide maps of human recombination. *Cell Systems* **3**(1): 83–94.

Cang Z, Mu L, Wei GW (2018). Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLOS Computational Biology* **14**(1): 1–44.

Cang Z, Wei GW (2018). Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *International Journal for Numerical Methods in Biomedical Engineering* **34**(2): e2914. e2914 cnm.2914.

Caruso M, Smith MW, Froelicher Y, Russo G, Gmitter FG (2020). Chapter 7 - Traditional breeding. *In* Talon M, Caruso M, Gmitter FG (Eds.), *The Genus Citrus*, pp. 129–148. Woodhead Publishing.

Chan JM, Carlsson G, Rabadán R (2013). Topology of viral evolution. *Proceedings of the National Academy of Sciences* **110**(46): 18566–18571.

Chazal F, Fasy B, Lecci F, Michel B, Rinaldo A, Rinaldo A, Wasserman L (2017). Robust topological inference: Distance to a measure and kernel distance. *J. Mach. Learn. Res.* **18**(1): 5845–5884.

Chitwood D, Sinha N (2016). Evolutionary and environmental forces sculpting leaf development. *Current Biology* **26**(7): R297–R306.

Chitwood DH, Eithun M, Munch E, Ophelders T (2019). Topological mapper for 3D volumetric images. *In* Burgeth B, Kleefeld A, Naegel B, Passat N, Perret B (Eds.), *Mathematical Morphology and Its Applications to Signal and Image Processing*, pp. 84–95. Cham: Springer International

Publishing.

Choat B, Nolf M, Lopez R, Peters JMR, Carins-Murphy MR, Creek D, Brodribb TJ (2018). Non-invasive imaging shows no evidence of embolism repair after drought in tree species of two genera. *Tree Physiology* **39**(1): 113–121.

Chudhary Z, Khera RA, Hanif MA, Ayub MA, Hamrouni L (2020). Chapter 49 – Walnut. *In* Hanif MA, Nawaz H, Khan MM, Byrne HJ (Eds.), *Medicinal Plants of South Asia*, pp. 671–684. Elsevier.

Chung Y, Hu C, Lawson A, Smyth C (2018). Topological approaches to skin disease image analysis. *In 2018 IEEE International Conference on Big Data (Big Data)*, pp. 100–105. Seattle, WA: IEEE.

Clayton CRI, Abbireddy COR, Schiebel R (2009). A method of estimating the form of coarse particulates. *Géotechnique* **59**(6): 493–501.

Cohen-Steiner D, Edelsbrunner H, Harer J (2007). Stability of persistence diagrams. *Discrete & Computational Geometry* **37**(1): 103–120.

Conover WJ (1998). *Practical Nonparametric Statistics* (3rd ed.). Wiley Series in Probability and Statistics. New York: Wiley.

Corey AT (1949). *Influence of shape on fall velocity of sandgrains*. Master's thesis, Colorado Agricultural and Mechanical College, Fort Collins, CO.

Crawford L, Monod A, Chen AX, Mukherjee S, Rabadán R (2020). Predicting clinical outcomes in glioblastoma: An application of topological and functional data analysis. *Journal of the American Statistical Association* **115**(531): 1139–1150.

Curry J, Mukherjee S, Turner K (2022). How many directions determine a shape and other sufficiency results for two topological transforms. *Transactions of the American Mathematical Society Series B* **9**: 1006–1043.

Davey JW, Blaxter ML (2011). RADSeq: next-generation population genetics. *Briefings in Functional Genomics* **9**(5-6): 416–423.

Dawson IK, Russell J, Powell W, Steffenson B, Thomas WTB, Waugh R (2015). Barley: a translational model for adaptation to climate change. *New Phytologist* **206**(3): 913–931.

Deng X, Yang X, Yamamoto M, Biswas MK (2020). Chapter 3 - Domestication and history. *In* Talon M, Caruso M, Gmitter FG (Eds.), *The Genus Citrus*, pp. 33–55. Woodhead Publishing.

Di Marzio M, Fensore S, Panzera A, Taylor CC (2019). Kernel density classification for spherical data. *Statistics & Probability Letters* **144**: 23–29.

Diaz-Gárcia L, Covarrubias-Pazaran G, Schlautman B, Grygleski E, Zalapa J (2018). Image-based phenotyping for identification of QTL determining fruit shape and size in american cranberry (*Vaccinium macrocarpon L.*). *PeerJ* **6**(e5461): e5461.

Diaz-Toca GM, Marin L, Necula I (2020). Direct transformation from Cartesian into geodetic coordinates on a triaxial ellipsoid. *Computers & Geosciences* **142**: 104551.

Ding YM, Cao Y, Zhang WP, Chen J, Liu J, Li P, Renner SS, Zhang DY, Bai WN (2022). Population-genomic analyses reveal bottlenecks and asymmetric introgression from Persian into iron walnut during domestication. *Genome Biology* **23**(1): 145.

Dryden IL, Mardia KV (2016). *Statistical Shape Analysis with Applications in R* (2 ed.). Chichester, West Sussex, England: John Wiley & Sons Ltd.

Du F, Tan T (2021). Recent studies in mechanical properties of selected hard-shelled seeds: A review. *JOM* **73**(6): 1723–1735.

Emmett K, Rosenbloom D, Cámara P, Rabadán R (2014). Parametric inference using persistence diagrams: A case study in population genetics. *In Proceedings of the 31st International Conference on Machine Learning*, Volume 32. Beijing, China: W&CP.

Emmett KJ, Rabadán R (2014). Characterizing scales of genetic recombination and antibiotic resistance in pathogenic bacteria using topological data analysis. *In* Ślęzak D, Tan AH, Peters JF, Schwabe L (Eds.), *Brain Informatics and Health*, pp. 540–551. Cham: Springer International Publishing.

Falcao AX, Stolfi J, de Alencar Lotufo R (2004). The image foresting transform: theory, algorithms, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(1): 19–29.

Fallah M, Vahdati K, Hasani D, Rasouli M, Sarikhani S (2022). Breeding of Persian walnut: Aiming to introduce late-leafing and early-harvesting varieties by targeted hybridization. *Scientia Horticulturae* **295**: 110885.

FAO (2021). Citrus fruit. Fresh and processed. Statistical bulletin 2020. Technical report, Food and Agriculture Organization of the United Nations, Rome, Italy.

FAS (2022). Tree nuts: World markets and trade. Technical report, Foreign Agricultural Service. US Department of Agriculture, Washington, DC.

Fasy BT, Lecci F, Rinaldo A, Wasserman L, Balakrishnan S, Singh A (2014). Confidence sets for persistence diagrams. *Ann. Statist.* **42**(6): 2301–2339.

Fasy BT, Micka S, Millman DL, Schenfisch A, Williams L (2019). The first algorithm for reconstructing simplicial complexes of arbitrary dimension from persistence diagrams. Preprint.

Fiorani F, Schurr U (2013). Future scenarios for plant phenotyping. *Annual Review of Plant Biology* **64**(1): 267–291.

Frank MH, Chitwood DH (2016). Plant chimeras: The good, the bad, and the 'Bizzaria'. *Developmental Biology* **419**(1): 41–53. Plant Development.

Friedman M (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* **32**(200): 675–701.

Fuller DQ, Weisskopf A (2014). Barley: Origins and development. *In* Smith C, Smith C (Eds.), *Encyclopedia of Global Archaeology*, pp. 763–766. New York, NY: Springer New York.

Gao Z (2022). Three-dimensional virtual flower bud of walnut tree based on micro-computed tomography. *Agronomy Journal* **114**(4): 1935–1943.

García-Portugués E, Navarro-Esteban P, Cuesta-Albertos JA (2023). On a projection-based class of uniformity tests on the hypersphere. *Bernoulli* **29**(1): 181–204.

García-Portugués E, Paindaveine D, Verdebout T (2020). On optimal tests for rotational symmetry against new classes of hyperspherical distributions. *Journal of the American Statistical Association* **115**(532): 1873–1887.

García-Portugués E, Paindaveine D, Verdebout T (2021). *rotasym*: Tests for rotational symmetry on the hypersphere. R package version 1.1.3.

García-Portugués E, Verdebout T (2021). *sphunif*: Uniformity tests on the circle, sphere, and hypersphere. R package version 1.0.1.

Gauthey A, Peters JMR, Carins-Murphy MR, Rodriguez-Dominguez CM, Li X, Delzon S, King A, López R, Medlyn BE, Tissue DT et al. (2020). Visual and hydraulic techniques produce similar estimates of cavitation resistance in woody species. *New Phytologist* **228**(3): 884–897.

Ghrist R, Levanger R, Mai H (2018). Persistent homology and Euler integral transforms. *Journal of Applied and Computational Topology* **2**(1): 55–60.

Giusti C, Pastalkova E, Curto C, Itskov V (2015). Clique topology reveals intrinsic geometric structure in neural correlations. *Proceedings of the National Academy of Sciences* **112**(44): 13455–13460.

Gmitter FG, Wu GA, Rokhsar DS, Talon M (2020). Chapter 1 - The citrus genome. *In* Talon M, Caruso M, Gmitter FG (Eds.), *The Genus Citrus*, pp. 1–8. Woodhead Publishing.

Gower JC (1975). Generalized procrustes analysis. *Psychometrika* **40**: 33–51.

Griffiths M, Mellor N, Sturrock CJ, Atkinson BS, Johnson J, Mairhofer S, York LM, Atkinson JA, Soltaninejad M, Foulkes JF et al. (2022). X-ray CT reveals 4D root system development and lateral root responses to nitrate in soil. *The Plant Phenome Journal* **5**(1): e20036.

Gülsoy E, Kuş E, Altıkat S (2019). Determination of physico-mechanical properties of some domestic and foreign walnut (*Juglans regia* L.) varieties. *Acta Scientiarum Polonorum Hortorum Cultus* **18**(6): 67–74.

Harlan HV, Martini ML (1929). A composite hybrid mixture. *Agronomy Journal* **21**(4): 487–490.

Harlan HV, Martini ML (1936). *Problems and results in barley breeding*. Washington, DC: US Department of Agriculture.

Harlan HV, Martini ML (1940). A study of methods in barley breeding. Technical Report 720, US Department of Agriculture, Washington, DC.

Harris JW, Stöcker H (1998). *The Handbook of Mathematics and Computational Science* (1st ed.). Berlin, Heidelberg: Springer-Verlag.

Hawkins R (1986). The Observations of Sir Richard Hawkins, Knight, in his Voyage into the South Sea, Annodomini, 1593. *Nutrition Reviews* **44**(11): 370–371. Facsimile reproduction of excerpts.

Heiss T, Wagner H (2017). Streaming Algorithm for Euler Characteristic Curves of Multidimensional Images. *In* Felsberg M, Heyden A, Krüger N (Eds.), *Computer Analysis of Images and Patterns*, pp. 397–409. Cham: Springer International Publishing.

Hilal N, Mohammed Ali TK, Tayeh BA (2020). Properties of environmental concrete that contains crushed walnut shell as partial replacement for aggregates. *Arabian Journal of Geosciences* **13**(16): 812.

Hockett E, Nilan R (1985). Genetics. *In Barley*, Chapter 8: pp. 187–230. John Wiley & Sons, Ltd.

Humphreys DP, McGuirl MR, Miyagi M, Blumberg AJ (2019). Fast estimation of recombination rates using topological data analysis. *Genetics* **211**(4): 1191–1204.

Huss JC, Antreich SJ, Bachmayr J, Xiao N, Eder M, Konnerth J, Gierlinger N (2020). Topological interlocking and geometric stiffening as complementary strategies for strong plant shells. *Advanced Materials* **32**(48): 2004519.

Ibáñez-Marcelo E, Campioni L, Phinyomark A, Petri G, Santarcangelo EL (2019). Topology highlights mesoscopic functional equivalence between imagery and perception: The case of hypnotizability. *NeuroImage* **200**: 437–449.

IPGRI (1994). *Descriptors for Walnut (*Juglans *Spp.)*. Rome, Italy: International Plant Genetic Resources Institute.

Isaac E (1959). Influence of religion on the spread of citrus. *Science* **129**(3343): 179–186.

Iwata H, Nesumi H, Ninomiya S, Takano Y, Ukai Y (2002). The Evaluation of Genotype × Environment Interactions of Citrus Leaf Morphology Using Image Analysis and Elliptic Fourier Descriptors. *Breeding Science* **52**(4): 243–251.

Janke NC (1966). Effect of shape upon the settling vellocity of regular convex geometric particles. *Journal of Sedimentary Research* **36**(2): 370–376.

Jeitziner R, Carrière M, Rougemont J, Oudot S, Hess K, Brisken C (2019). Two-Tier Mapper, an unbiased topology-based clustering method for enhanced global gene expression analysis. *Bioinformatics* **35**(18): 3339–3347.

Jiang Q, Kurtek S, Needham T (2020). The weighted Euler curve transform for shape and image analysis. *In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3685–3694.

Kanari L, Dłotko P, Scolamiero M, Levi R, Shillcock J, Hess K, Markram H (2018). A topological representation of branching neuronal morphologies. *Neuroinformatics* **16**(1): 3–13.

Kelc D, Štampar F, Solar A (2007). Fruiting behaviour of walnut trees influences the relationship between the morphometric traits of the parent wood and nut weight. *The Journal of Horticultural Science and Biotechnology* **82**(3): 439–445.

Kendall DG (1984). Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces. *Bulletin of the London Mathematical Society* **16**(2): 81–121.

Khadivi-Khub A, Ebrahimi A, Sheibani F, Esmaeili A (2015). Phenological and pomological characterization of Persian walnut to select promising trees. *Euphytica* **205**(2): 557–567.

Knight TG, Klieber A, Sedgley M (2001). The Relationship Between Oil Gland and Fruit Development in Washington Navel Orange (*Citrus sinensis* L. Osbeck). *Annals of Botany* **88**(6): 1039–1047.

Komatsuda T, Pourkheirandish M, He C, Azhaguvel P, Kanamori H, Perovic D, Stein N, Graner A, Wicker T, Tagiri A et al. (2007). Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene. *Proceedings of the National Academy of Sciences* **104**(4): 1424–1429.

Köppen M (2000). The curse of dimensionality. *In 5th Online World Conference on Soft Computing in Industrial Applications*, Volume 1, pp. 4–8. WSC5.

Kouhi M, Rezaei A, Hassani D, Sarikhani S, Vahdati K (2020). Phenotypic Evaluation and Identification of Superior Persian Walnut (*Juglans regia* L.) Genotypes in Mazandaran Province, Iran. *Journal of Nuts* **11**(4): 315–326.

Kovacev-Nikolic V, Bubenik P, Nikolić D, Heo G (2016). Using persistent homology and dynamical distances to analyze protein binding. *Statistical Applications in Genetics and Molecular Biology* **15**(1): 19–38.

Kovalevsky V (1989). Finite topology as applied to image analysis. *Computer Vision, Graphics, and Image Processing* **46**(2): 141–161.

Koyuncu MA, Ekinci K, Savran E (2004). Cracking characteristics of walnut. *Biosystems Engineering* **87**(3): 305–311.

Krumbein WC (1941). Measurement and geological significance of shape and roundness of sedimentary particles. *Journal of Sedimentary Research* **11**(2): 64–72.

Kruskal WH, Wallis WA (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* **47**(260): 583–621.

Kuhl FP, Giardina CR (1982). Elliptic Fourier features of a closed contour. *Computer Graphics and Image Processing* **18**(3): 236 – 258.

Lala SD, Deoghare AB, Chatterjee S (2018). Mechanical and morphological characterization of walnut shell reinforced epoxy composite. *IOP Conference Series: Materials Science and Engineering* **377**(1): 012011.

Langgut D (2017). The citrus route revealed: From Southeast Asia into the Mediterranean. *HortScience* **52**(6): 814–822.

Lawson P, Schupbach J, Fasy BT, Sheppard JW (2019). Persistent homology for the automatic classification of prostate cancer aggressiveness in histopathology images. *In* Tomaszewski JE, Ward AD (Eds.), *Medical Imaging 2019: Digital Pathology*, Volume 10956, pp. 109560G. SPIE.

Lawson P, Sholl AB, Brown JQ, Fasy BT, Wenk C (2019). Persistent homology for the quantitative evaluation of architectural features in prostate cancer histology. *Scientific Reports* **9**(1139).

Lee H, Kang H, Chung MK, Kim B, Lee DS (2012). Persistent brain network homology from the perspective of dendrogram. *IEEE Transactions on Medical Imaging* **31**(12): 2267–2277.

Lesnick M, Wright M (2015). Interactive visualization of 2-D persistence modules. Preprint.

Lesnick M, Wright M (2022). Computing minimal presentations and bigraded Betti numbers of 2-parameter persistent homology. *SIAM Journal on Applied Algebra and Geometry* **6**(2): 267–298.

Lestrel PE (Ed.) (1997). *Fourier Descriptors and their Applications in Biology*. Cambridge: Cambridge University Press.

Ley C, Verdebout T (2017). *Modern Directional Statistics* (1st ed.). Interdisciplinary Statistics. Boca Raton: Chapman and Hall/CRC.

Li M, An H, Angelovici R, Bagaza C, Batushansky A, Clark L, Coneva V, Donoghue MJ, Edwards E, Fajardo D et al. (2018). Topological data analysis as a morphometric method: Using persistent homology to demarcate a leaf morphospace. *Frontiers in Plant Science* **9**: 553.

Li M, Duncan K, Topp CN, Chitwood DH (2017). Persistent homology and the branching topologies of plants. *American Journal of Botany* **104**(3): 349–353.

Li M, Frank MH, Coneva V, Mio W, Chitwood DH, Topp CN (2018). The persistent homology mathematical framework provides enhanced genotype-to-phenotype associations for plant morphology. *Plant Physiology* **177**(4): 1382–1395.

Li M, Klein LL, Duncan KE, Jiang N, Chitwood DH, Londo JP, Miller AJ, Topp CN (2019). Characterizing 3D inflorescence architecture in grapevine using X-ray imaging and advanced morphometrics: implications for understanding cluster density. *Journal of Experimental Botany* **70**(21): 6261–6276.

Li M, Shao MR, Zeng D, Ju T, Kellogg EA, Topp CN (2020). Comprehensive 3D phenotyping reveals continuous morphological variation across genetically diverse sorghum inflorescences. *New Phytologist* **226**(6): 1873–1885.

Li Q, Griffiths J (2004). Least squares ellipsoid specific fitting. *In Geometric Modeling and Processing, 2004. Proceedings*, pp. 335–340. IEEE.

Liller CB, Neuhaus R, von Korff M, Koornneef M, van Esse W (2015). Mutations in barley row type genes have pleiotropic effects on shoot branching. *PLOS ONE* **10**(10): 1–20.

Lopez JM (2004). Walnut tissue culture: research and field applications. *In* Michler CH, Pijut PM, Van Sambeek JW, Coggeshall MV, Seifert J, Woeste K, Overton R, Ponder FJ (Eds.), *Black walnut in a new century*, 6th Walnut Council Research Symposium, pp. 146–152. St. Paul, MN: USDA, North Central Research Station.

Lum PY, Singh G, Lehman A, Ishkanov T, Vejdemo-Johansson M, Alagappan M, Carlsson J, Carlsson G (2013). Extracting insights from the shape of complex data using topology. *Scientific Reports* **3**(1236).

Luo X, Zhou H, Cao D, Yan F, Chen P, Wang J, Woeste K, Chen X, Fei Z, An H et al. (2022). Domestication and selection footprints in Persian walnuts (*Juglans regia*). *PLOS Genetics* **18**(12): 1–19.

Luro F, Curk F, Froelicher Y, Ollitrault P (2017). Recent insights on Citrus diversity and phylogeny. *In* Zech-Matterne V, Fiorentino G (Eds.), *AGRUMED: Archaeology and history of citrus fruit in the Mediterranean: Acclimatization, diversification, uses*. Naples: Publications du Centre Jean Bérard.

Magiorkinis E, Beloukas A, Diamantis A (2011). Scurvy: Past, present and future. *European Journal of Internal Medicine* **22**(2): 147–152.

Mahato N, Sharma K, Koteswararao R, Sinha M, Baral E, Cho MH (2019). Citrus essential oils: Extraction, authentication and application in food preservation. *Critical Reviews in Food Science and Nutrition* **59**(4): 611–625.

Mahomoodally MF, Mooroteea K (2021). A comparative ethno-religious study of traditionally used medicinal plants employed in the management of cardiovascular diseases. *Journal of Herbal Medicine* **25**: 100417.

Mander L, Dekker SC, Li M, Mio W, Punyasena SW, Lenton TM (2017). A morphometric analysis of vegetation patterns in dryland ecosystems. *Royal Society Open Science* **4**(2): 160443.

Mapelli S, Pollegioni P, Woeste KE, Chiocchini F, Lungo SD, Olimpieri I, Tortolano V, Clark J, Hemery GE, Malvolti ME (2018). Spatial genetic structure of common walnut (*Juglans regia* L.) in central Asia. *In Acta Horticulturae*, Number 1190, pp. 27–34. International Society for Horticultural Science (ISHS), Leuven, Belgium.

Mardia KV, Jupp PE (1999). *Directional Statistics* (1st ed.). Probability and Statistics. Chichester, West Sussex: John Wiley & Sons, Ltd.

Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J et al. (2017). A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**(7651): 427–433.

Mascher M, Schuenemann VJ, Davidovich U, Marom N, Himmelbach A, Hübner S, Korol A, David M, Reiter E, Riehl S et al. (2016). Genomic analysis of 6,000-year-old cultivated grain illuminates the domestication history of barley. *Nature Genetics* **48**(9): 1089–1093.

McAllister CA, McKain MR, Li M, Bookout B, Kellogg EA (2019). Specimen-based analysis of morphology and the environment in ecologically dominant grasses: the power of the herbarium. *Philosophical Transactions of the Royal Society B: Biological Sciences* **374**(1763): 20170403.

McInnes L, Healy J, Melville J (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint.

Mei L, Sheng O, Peng S, Zhou G, Wei Q, Li Q (2011). Growth, root morphology and boron uptake by citrus rootstock seedlings differing in boron-deficiency responses. *Scientia Horticulturae* **129**(3): 426–432.

Micka SA (2020). *Searching and Reconstruction: Algorithms with Topological Descriptors*. Ph. D. thesis, Montana State University, Bozeman, Montana.

Migicovsky Z, Harris ZN, Klein LL, Li M, McDermaid A, Chitwood DH, Fennell A, Kovacs LG, Kwasniewski M, Londo JP et al. (2019). Rootstock effects on scion phenotypes in a '*Chambourcin*' experimental vineyard. *Horticulture Research* **6**(64).

Migicovsky Z, Li M, Chitwood DH, Myles S (2018). Morphometrics reveals complex and heritable apple leaf shapes. *Frontiers in Plant Science* **8**: 2185.

Mileyko Y, Mukherjee S, Harer J (2011). Probability measures on the space of persistence diagrams. *Inverse Problems* **27**(12): 124007.

Motuzaite Matuzeviciute G, Abdykanova A, Kume S, Nishiaki Y, Tabaldiev K (2018). The effect of geographical margins on cereal grain size variation: Case study for highlands of Kyrgyzstan. *Journal of Archaeological Science: Reports* **20**: 400–410.

Munch E (2017). A User's Guide to Topological Data Analysis. *Journal of Learning Analytics* **4**: 47–61.

Munch E, Turner K, Bendich P, Mukherjee S, Mattingly J, Harer J (2015). Probabilistic Fréchet means for time varying persistence diagrams. *Electron. J. Statist.* **9**(1): 1173–1204.

Myers A, Munch E, Khasawneh FA (2019). Persistent homology of complex networks for dynamic state detection. *Phys. Rev. E* **100**: 022314.

NASS (2021). Citrus fruits. 2021 Summary. Technical report, National Agricultural Statistics Service. United States Department of Agriculture., Washington, DC.

NASS (2022). 2022 California Walnut. Objective Measurement Report. Technical report, National Agricultural Statistics Service. US Department of Agriculture, Sacramento, CA.

Nati P (1674). *Florentina phytologica observatio de malo Limonia citrata-aurantia, Florentiae vulgo la bizzaria*. Florentia, Italy: Hippolyti de Naue.

Nicolau M, Levine AJ, Carlsson G (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences* **108**(17): 7265–7270.

Niklas KJ (2004). Plant allometry: is there a grand unifying theory? *Biological Reviews* **79**(4): 871–889.

Niklas KJ, Hammond ST (2019). Biophysical effects on the scaling of plant growth, form, and ecology. *Integrative and Comparative Biology* **59**(5): 1312–1323.

Ollitrault P, Curk F, Krueger R (2020). Chapter 4 - Citrus taxonomy. *In* Talon M, Caruso M, Gmitter FG (Eds.), *The Genus Citrus*, pp. 57–81. Woodhead Publishing.

Otter N, Porter MA, Tillmann U, Grindrod P, Harrington HA (2017). A roadmap for the computation of persistent homology. *EPJ Data Science* **6**(17).

Pal D, Malik SK, Kumar S, Choudhary R, Sharma KC, Chaudhury R (2013). Genetic variability and relationship studies of mandarin (*Citrus reticulata* Blanco) using morphological and molecular markers. *Agricultural Research* **2**(3): 236–245.

Palande S, Kaste JAM, Roberts MD, Abá KS, Claucherty C, Dacon J, Doko R, Jayakody TB, Jeffery HR, Kelly N et al. (2022). The topological shape of gene expression across the evolution of flowering plants. *bioRxiv* **0**.

Panou G, Korakitis R, Pantazis G (2020). Fitting a triaxial ellipsoid to a geoid model. *Journal of Geodetic Science* **10**(1): 69–82.

Perea JA (2019). Topological time series analysis. *Notices of the American Mathematical Society* **66**(5): 686–694.

Perea JA, Munch E, Khasawneh FA (2022). Approximating continuous functions on persistence diagrams using template functions. *Foundations of Computational Mathematics*.

Petit J, Salentijn EMJ, Paulo MJ, Denneboom C, van Loo EN, Trindade LM (2020). Elucidating the genetic architecture of fiber quality in hemp (*Cannabis sativa L.*) using a genome-wide association study. *Frontiers in Genetics* **11**: 1101.

Petri G, Expert P, Turkheimer F, Carhart-Harris R, Nutt D, Hellyer PJ, Vaccarino F (2014). Homological scaffolds of brain functional networks. *Journal of The Royal Society Interface* **11**(101): 20140873.

Pewsey A, García-Portugués E (2021). Recent advances in directional statistics. *TEST* **30**(1): 1–58.

Pinney K, Polito VS (1983). English walnut fruit growth and development. *Scientia Horticulturae* **21**(1): 19–28.

Pollegioni P, Woeste KE, Chiocchini F, Del Lungo S, Olimpieri I, Tortolano V, Clark J, Hemery GE, Mapelli S, Malvolti ME (2015). Ancient humans influenced the current spatial genetic structure of common walnut populations in Asia. *PLOS ONE* **10**(9): 1–16.

Pollegioni P, Woeste KE, Chiocchini F, Olimpieri I, Tortolano V, Clark J, Hemery GE, Mapelli S, Malvolti ME (2014). Landscape genetics of Persian walnut (*Juglans regia* L.) across its Asian range. *Tree Genetics & Genomes* **10**(4): 1027–1043.

Popa RG, Bălăcescu A, Popescu LG (2023). Organic Walnut Cultivation in Intensive and Super-Intensive System—Sustainable Investment. Case Study: Gorj County, Romania. *Sustainability* **15**(2).

Qaiser T, Tsang YW, Taniyama D, Sakamoto N, Nakane K, Epstein D, Rajpoot N (2019). Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features. *Medical Image Analysis* **55**: 1 – 14.

Quade D (1979). Using weighted rankings in the analysis of complete blocks with additive block effects. *Journal of the American Statistical Association* **74**(367): 680–683.

Rahaman MM, Chen D, Gillani Z, Klukas C, Chen M (2015). Advanced phenotyping and phenotype data analysis for the study of plant growth and development. *Frontiers in Plant Science* **6**.

Reani Y, Bobrowski O (2022). Cycle registration in persistent homology with applications in topological bootstrap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*: 1–15.

Rezaee R, Vahdati K, Grigoorian V, Valizadeh M (2008). Walnut grafting success and bleeding rate as affected by different grafting methods and seedling vigour. *The Journal of Horticultural Science and Biotechnology* **83**(1): 94–99.

Rezaei Z, Khadivi A, ValizadehKaji B, Abbasifar A (2018). The selection of superior walnut (*Juglans regia* L.) genotypes as revealed by morphological characterization. *Euphytica* **214**(4): 69.

Richardson E, Werman M (2014). Efficient classification using the Euler characteristic. *Pattern Recognition Letters* **49**: 99–106.

Rizvi AH, Cámara PG, Kandror EK, Roberts TJ, Schieren I, Maniatis T, Rabadán R (2017). Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nature Biotechnology* **35**(551).

Robinson A, Turner K (2017). Hypothesis testing for topological data analysis. *J Appl. and Comput. Topology* **1**: 241–261.

Roor W, Konrad H, Mamadjanov D, Geburek T (2017). Population Differentiation in Common Walnut (*Juglans regia* L.) across Major Parts of Its Native Range—Insights from Molecular and Morphometric Data. *Journal of Heredity* **108**(4): 391–404.

Ros J, Evin A, Bouby L, Ruas MP (2014). Geometric morphometric analysis of grain shape and the identification of two-rowed barley (*Hordeum vulgare subsp. distichum L.*) in southern France. *Journal of Archaeological Science* **41**: 568–575.

Russell J, Mascher M, Dawson IK, Kyriakidis S, Calixto C, Freund F, Bayer M, Milne I,

Marshall-Griffiths T, Heinen S et al. (2016). Exome sequencing of geographically diverse barley landraces and wild relatives gives insights into environmental adaptation. *Nature Genetics* **48**(9): 1024–1030.

Sato K (2020). History and future perspectives of barley genomics. *DNA Research* **27**(4).

Schölkopf B, Smola A, Müller KR (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation* **10**(5): 1299–1319.

Shah RA, Bakshi P, Sharma N, Jasrotia A, Itoo H, Gupta R, Singh A (2021). Diversity assessment and selection of superior Persian walnut (*Juglans regia* L.) trees of seedling origin from North-Western Himalayan region. *Resources, Environment and Sustainability* **3**: 100015.

Sharma PC, Bhatia V, Bansal N, Sharma A (2007). A review on Bael tree. *Natural Product Radiance* **6**(2): 171–178.

Shrestha KB, Dangol DR (2019). Crops with medicinal, religious and cultural values. *In* Joshi BK, Shrestha R (Eds.), *Working Groups of Agricultural Plant Genetic Resources (APGRs) in Nepal*, pp. 198–204. Kathmandu, Nepal: NAGRC, Khumaltar. Proceedings of National Workshop, 21-22 June 2018.

Sideli GM, Marrano A, Montanari S, Leslie CA, Allen BJ, Cheng H, Brown PJ, Neale DB (2020). Quantitative phenotyping of shell suture strength in walnut (*Juglans regia* L.) enhances precision for detection of QTL and genome-wide association mapping. *PLOS ONE* **15**(4): 1–21.

Singh G, Mémoli F, Carlsson G (2007). Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. *In* Botsch M, Pajarola R, Chen B, Zwicker M (Eds.), *Eurographics Symposium on Point-Based Graphics*. The Eurographics Association.

Singh N, Couture HD, Marron JS, Perou C, Niethammer M (2014). Topological descriptors of histology images. *In* Wu G, Zhang D, Zhou L (Eds.), *Machine Learning in Medical Imaging*, pp. 231–239. Cham: Springer International Publishing.

Sizemore AE, Giusti C, Kahn A, Vettel JM, Betzel RF, Bassett DS (2018). Cliques and cavities in the human connectome. *Comput Neurosci* **44**: 115–145.

Smith A, Zavala VM (2021). The Euler characteristic: A general topological descriptor for complex data. *Computers & Chemical Engineering* **154**: 107463.

Smith NM, Ebrahimi H, Ghosh R, Dickerson AK (2018). High-speed microjets issue from bursting oil gland reservoirs of citrus fruit. *Proceedings of the National Academy of Sciences* **115**(26): E5887–E5895.

Smith RJ (2009). Use and misuse of the reduced major axis for line-fitting. *American Journal of Physical Anthropology* **140**(3): 476–486.

Sneed ED, Folk RL (1958). Pebbles in the Lower Colorado River, Texas a Study in Particle Morphogenesis. *The Journal of Geology* **66**(2): 114–150.

Solar A, Ivančič A, Štampar F (2003). Morphometric characteristics of fruit bearing shoots in persian walnut (*Juglans regia* l.) — potential selection criteria for breeding. *European Journal of Horticultural Science* **2**.

Spreen TH, Gao Z, Fernandes W, Zansler ML (2020). Chapter 23 - Global economics and marketing of citrus products. *In* Talon M, Caruso M, Gmitter FG (Eds.), *The Genus Citrus*, pp. 471–493. Woodhead Publishing.

Stolz BJ, Kaeppler J, Markelc B, Braun F, Lipsmeier F, Muschel RJ, Byrne HM, Harrington HA (2022). Multiscale topology characterizes dynamic tumor vascular networks. *Science Advances* **8**(23): eabm2456.

Talon M, Wu GA, Gmitter FG, Rokhsar DS (2020). Chapter 2 - The origin of citrus. *In* Talon M, Caruso M, Gmitter FG (Eds.), *The Genus Citrus*, pp. 9–31. Woodhead Publishing.

Tanabata T, Shibaya T, Hori K, Ebana K, Yano M (2012). SmartGrain: High-throughput phenotyping software for measuring seed shape through image analysis. *Plant Physiology* **160**(4): 1871–1880.

Tang WS, da Silva GM, Kirveslahti H, Skeens E, Feng B, Sudijono T, Yang KK, Mukherjee S, Rubenstein B, Crawford L (2022). A topological data analytic approach for discovering biophysical signatures in protein dynamics. *PLOS Computational Biology* **18**(5): 1–42.

Tanno Ki, Willcox G (2012). Distinguishing wild and domestic wheat and barley spikelets from early Holocene sites in the Near East. *Vegetation History and Archaeobotany* **21**(2): 107–115.

Thapa R, Thapa P, Ahamad K, Vahdati K (2021). Effect of grafting methods and dates on the graft take rate of Persian walnut in open field condition. *International Journal of Horticultural Science and Technology* **8**(2): 133–147.

Thompson DW (1942). *On Growth and Form* (2nd ed.). Cambridge, UK: Cambridge University Press.

Todd MJ, Yıldırım EA (2007). On Khachiyan's algorithm for the computation of minimum-volume enclosing ellipsoids. *Discrete Applied Mathematics* **155**(13): 1731–1744.

Topaz CM, Ziegelmeier L, Halverson T (2015). Topological data analysis of biological aggregation models. *PLOS ONE* **10**(5): e0126383.

Tseng CC (1999). An Allegory on Allegory: Reading "Ju song" as Qu Yuan's *Ars Poetica*. *Dong Hwa Journal of Humanistic Studies* **1**: 69–101.

Turner K, Mileyko Y, Mukherjee S, Harer J (2014). Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry* **52**: 44–70.

Turner K, Mukherjee S, Boyer DM (2014). Persistent homology transform for modeling shapes and surfaces. *Information and Inference* **3**(4): 310–344.

Tymochko S, Munch E, Dunion J, Corbosiero K, Torn R (2020). Using persistent homology to quantify a diurnal cycle in hurricanes. *Pattern Recognition Letters* **133**: 137–143.

Tymochko S, Munch E, Khasawneh FA (2019). Adaptive partitioning for template functions on persistence diagrams. *In 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 1227–1234.

Vahdati K (2014). Traditions and folks for walnut growing around the Silk Road. *In Acta Horticulturae*, Volume 1032, pp. 19–24. Leuven, Belgium: International Society for Horticultural Science (ISHS).

Verma MK (2014). Walnut production technology. *In Training manual on teaching of post-graduate courses in horticulture (Fruit Science)*, pp. 281–287. New Delhi, India: Indian Agricultural Research Institute.

Verma RS, Padalia RC, Chauhan A, Thul ST (2013). Phytochemical analysis of the leaf volatile oil of walnut tree (*Juglans regia* L.) from western Himalaya. *Industrial Crops and Products* **42**: 195–201.

Vlahos L, Isliker H, Kominis Y, Hizanidis K (2008). Normal and anomalous diffusion: A tutorial. Preprint.

Voo SS, Grimes HD, Lange BM (2012). Assessing the biosynthetic capabilities of secretory glands in *Citrus* peel. *Plant physiology* **159**(1): 81–94.

Voulgaridis V, Vassiliou VG (2005). The walnut wood and its utilisation to high value products. *In Acta Horticulturae*, Volume 705, pp. 69–81. Leuven, Belgium: International Society for Horticultural Science (ISHS).

Vovin A (Ed.) (2016). *Man'yōshū (Book 18): A New English Translation Containing the Original Text, Kana Transliteration, Romanization, Glossing and Commentary*. Leiden, The Netherlands: Brill.

Vuollo V, Holmström L (2018). A scale space approach for exploring structure in spherical data. *Computational Statistics & Data Analysis* **125**: 57–69.

Wadell H (1932). Volume, shape, and roundness of rock particles. *The Journal of Geology* **40**(5): 443–451.

Wagner H, Chen C, Vuçini E (2012). Efficient computation of persistent homology for cubical data. *In* Peikert R, Hauser H, Carr H, Fuchs R (Eds.), *Topological Methods in Data Analysis and Visualization II: Theory, Algorithms, and Applications*, pp. 91–106. Berlin, Heidelberg: Springer Berlin Heidelberg.

Wahid M, Gawli Y, Puthusseri D, Kumar A, Shelke MV, Ogale S (2017). Nutty Carbon: Morphology Replicating Hard Carbon from Walnut Shell for Na Ion Battery Anode. *ACS Omega* **2**(7): 3601–3609.

Wallace M, Bonhomme V, Russell J, Stillman E, George TS, Ramsay L, Wishart J, Timpany S, Bull H, Booth A et al. (2019). Searching for the origins of *Bere* barley: a geometric morphometric approach to cereal landrace recognition in archaeology. *Journal of Archaeological Method and Theory* **26**(3): 1125–1142.

Wang B, Sudijono T, Kirveslahti H, Gao T, Boyer DM, Mukherjee S, Crawford L (2021). A statistical pipeline for identifying physical features that differentiate classes of 3D shapes. *The Annals of Applied Statistics* **15**(2): 638–661.

Wang F, Wagner H, Chen C (2022). GPU Computation of the Euler Characteristic Curve for Imaging Data. *In* Goaoc X, Kerber M (Eds.), *38th International Symposium on Computational Geometry (SoCG 2022)*, Volume 224, pp. 64:1–64:16. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Wang J, Wang Z, Du X, Yang H, Han F, Han Y, Yuan F, Zhang L, Peng S, Guo E (2017). A high-density genetic map and QTL analysis of agronomic traits in foxtail millet [*Setaria italica (L.) P. Beauv.*] using RAD-seq. *PLOS ONE* **12**(6): 1–15.

Wang J, Ye H, Zhou H, Chen P, Liu H, Xi R, Wang G, Hou N, Zhao P (2022). Genome-wide association analysis of 101 accessions dissects the genetic basis of shell thickness for genetic improvement in Persian walnut (*Juglans regia* L.). *BMC Plant Biology* **22**(1): 436.

Wang L, Conteh B, Fang L, Xia Q, Nian H (2020). QTL mapping for soybean (*Glycine max L.*) leaf chlorophyll-content traits in a genotyped RIL population by using RAD-seq based high-density linkage map. *BMC Genomics* **21**(1): 739.

West GB, Brown JH, Enquist BJ (1999). A general model for the structure and allometry of plant vascular systems. *Nature* **400**(6745): 664–667.

Wu GA, Sugimoto C, Kinjo H, Azama C, Mitsube F, Talon M, Gmitter FG, Rokhsar DS (2021). Diversification of mandarin citrus by hybrid speciation and apomixis. *Nature Communications* **12**(1): 4377.

Wu GA, Terol J, Ibanez V, López-García A, Pérez-Román E, Borredá C, Domingo C, Tadeo FR, Carbonell-Caballero J, Alonso R et al. (2018). Genomics of the origin and evolution of citrus. *Nature* **554**(7692): 311–316.

Xiao N, Bock P, Antreich SJ, Staedler YM, Schönenberger J, Gierlinger N (2020). From the soft to the hard: Changes in microchemistry during cell wall maturation of walnut shells. *Frontiers in Plant Science* **11**.

Xiao Z, Stait-Gardner T, Willis SA, Price WS, Moroni FJ, Pagay V, Tyerman SD, Schmidtke LM, Rogiers SY (2021). 3D visualisation of voids in grapevine flowers and berries using X-ray micro computed tomography. *Australian Journal of Grape and Wine Research* **27**(2): 141–148.

Yamashita H, Uchida T, Tanaka Y, Katai H, Nagano AJ, Morita A, Ikka T (2020). Genomic predictions and genome-wide association studies based on RAD-seq of quality-related metabolites for the genomics-assisted breeding of tea plants. *Scientific Reports* **10**(1): 17480.

Zhang BW, Xu LL, Li N, Yan PC, Jiang XH, Woeste KE, Lin K, Renner SS, Zhang DY, Bai WN (2019). Phylogenomics Reveals an Ancient Hybrid Origin of the Persian Walnut. *Molecular Biology and Evolution* **36**(11): 2451–2461.

Zhang H, Lan H, Tang Y, Li Y (2014). Study on fracture mechanism of walnut shell according to brittle fracture area. *In 2014 Fifth International Conference on Intelligent Systems Design and Engineering Applications*, pp. 954–957.

Zhao S, Wen J, Wang H, Zhang Z, Li X (2016). Changes in lignin content and activity of related enzymes in the endocarp during the walnut shell development period. *Horticultural Plant Journal* **2**(3): 141–146.

Zhou H, Whalley WR, Hawkesford MJ, Ashton RW, Atkinson B, Atkinson JA, Sturrock CJ, Bennett MJ, Mooney SJ (2020). The interaction between wheat roots and soil pores in structured field soil. *Journal of Experimental Botany* **72**(2): 747–756.

Zhou L, Wang SB, Jian J, Geng QC, Wen J, Song Q, Wu Z, Li GJ, Liu YQ, Dunwell JM et al. (2015). Identification of domestication-related loci associated with flowering time and seed size in soybean with the RAD-seq genotyping method. *Scientific Reports* **5**(1): 9350.

APPENDIX A

BARLEY APPENDIX

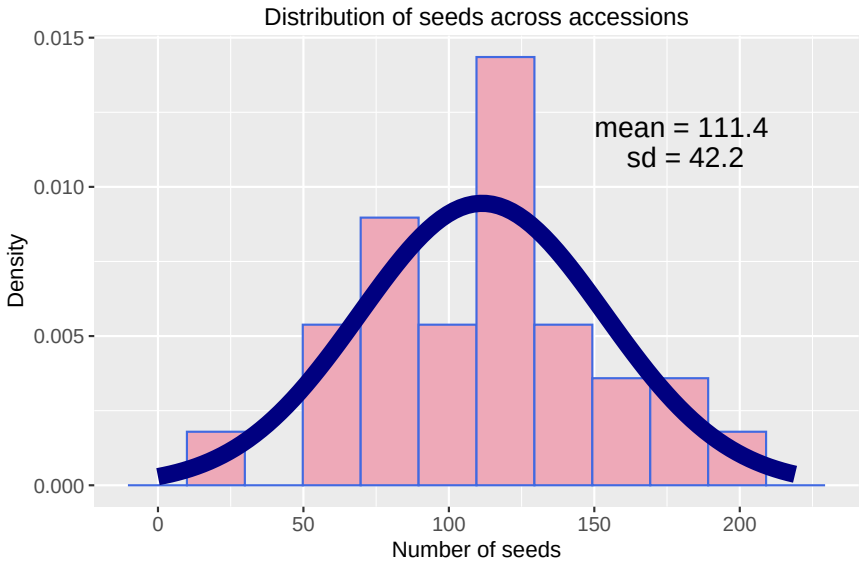

Figure A.1: Distribution of the 3121 seeds according to their accession. The seed number values as in Table 3.1 have empirical mean $\bar{\mu} = 111.46$ and empirical standard deviation $\bar{\sigma} = 42.21$. A normal distribution with these parameters is drawn on top of the histogram. Observe that all the accession seed numbers are within two standard deviations.

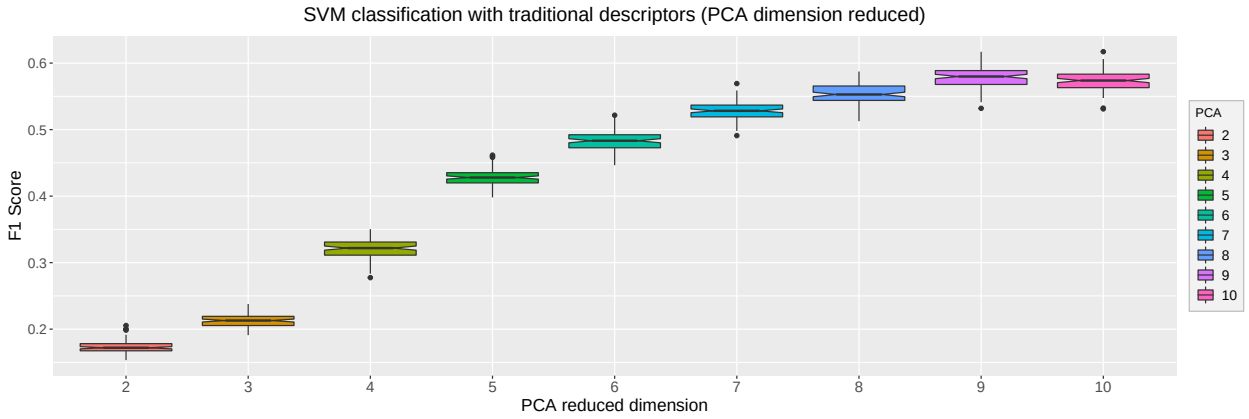

Figure A.2: Classification results for traditional shape descriptors. After centering and scaling the traditional shape descriptors, we used PCA to reduce their dimension and then performed an SVM classification with these dimension-reduced vectors. We observe that the highest classification F_1 scores correspond to the use of almost all the traditional dimensions.

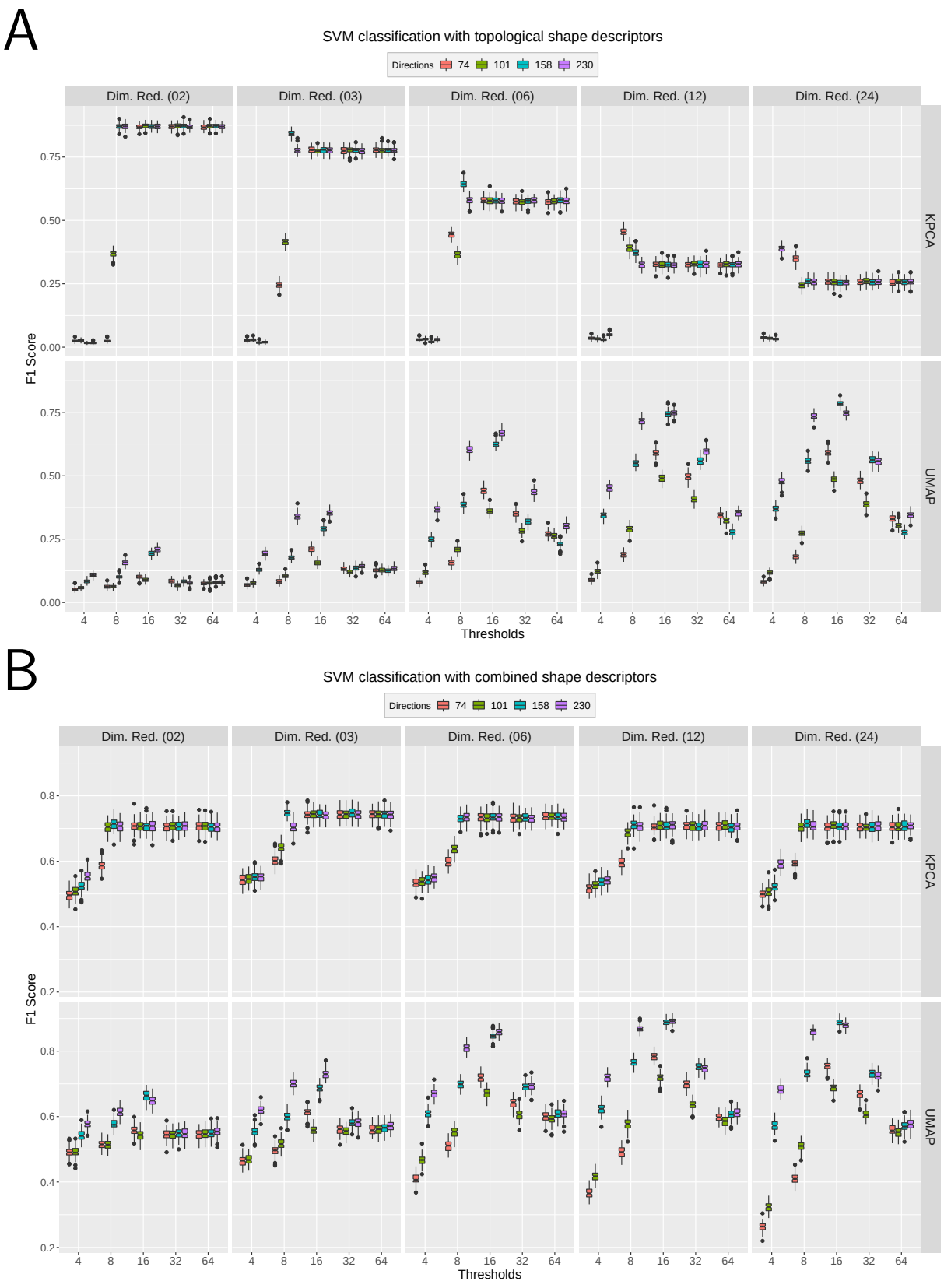

Figure A.3: Classification results for combined and topological shape descriptors computed for different choices of parameters.

Figure A.3 (cont'd):
Classification results for combined and topological shape descriptors computed for different choices of parameters. To evaluate the ECT descriptiveness, we sought to use these ECT vectors to classify 28 different barley accessions based solely on seed morphology. The ECT was computed for different number of directions and thresholds. These high-dimensional vectors were later reduced to different number of dimensions using both KPCA and UMAP. Observe that both dimension reduction techniques summarize the ECT information in very different ways, as evidenced by the different SVM classification F1 scores when using (A) exclusively topological information or (B) combining both topological and traditional seed shape descriptors.

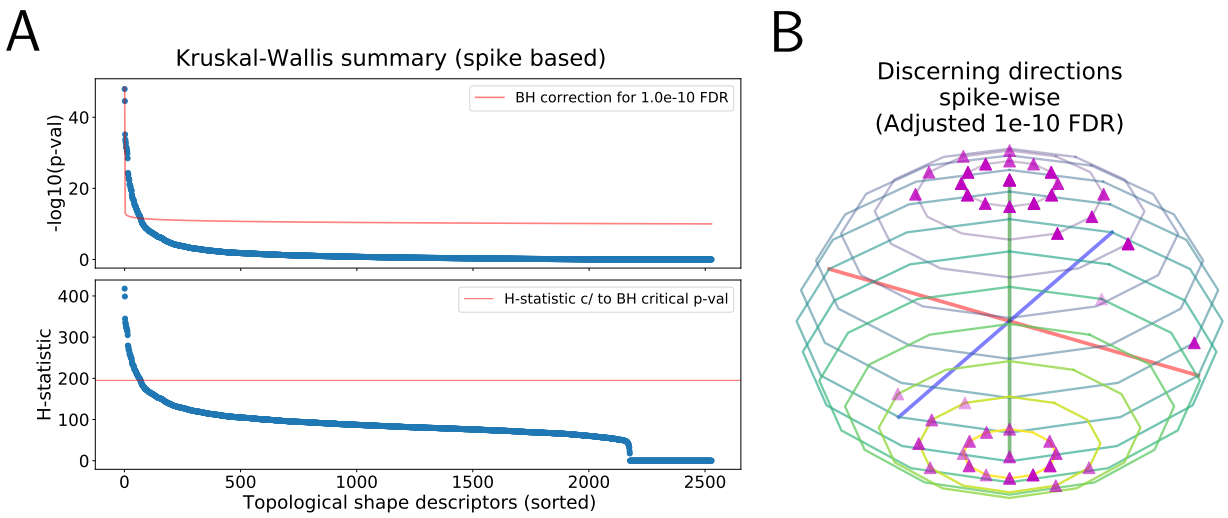

Figure A.4: Relevant ECT directions and slices. (A) We examine the inter-spike and intra-spike variance differences of the Euler characteristic for each direction and threshold. A Kruskal-Wallis analysis combined with a Benjamini-Hochberg multiple test correction suggests a number of discerning slices across accessions. (B) These directions and thresholds are mostly concentrated around the poles, similar to the case of inter- and intra-accession variance case (Figure 3.4).

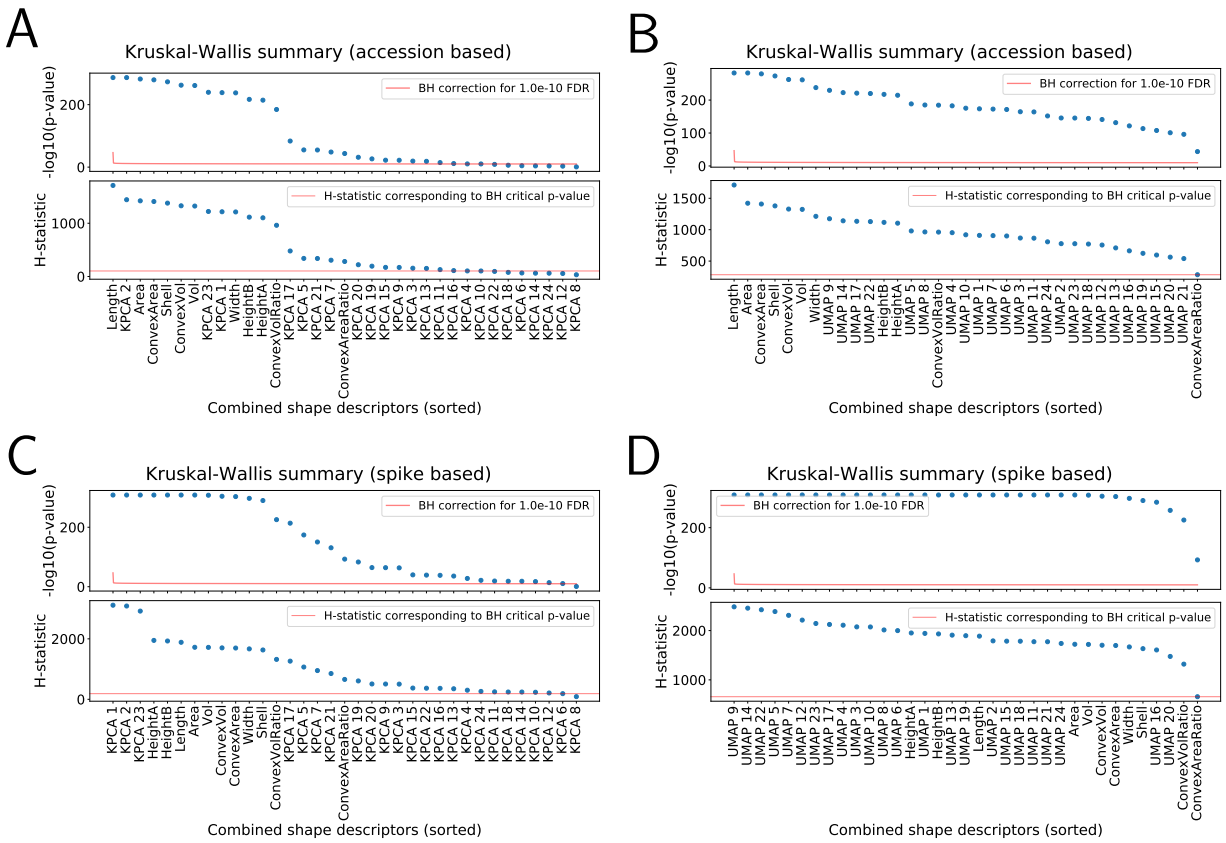

Figure A.5: Relevant combined descriptors. Dimension-reduced topological vectors were concatenated with tradtional shape descriptors to produce combined descriptors. Kruskal-Wallis analyses reveal which descriptors explain the most inter-accession variance when the ECT was reduced in dimension with (A) KPCA, and (B) UMAP. Similar analyses also reveal which features contribute the most to inter-spike variance when the ECT vector was reduced with (C) KPCA, and (D) UMAP.

APPENDIX B

CITRUS APPENDIX

B.1 Supporting figures

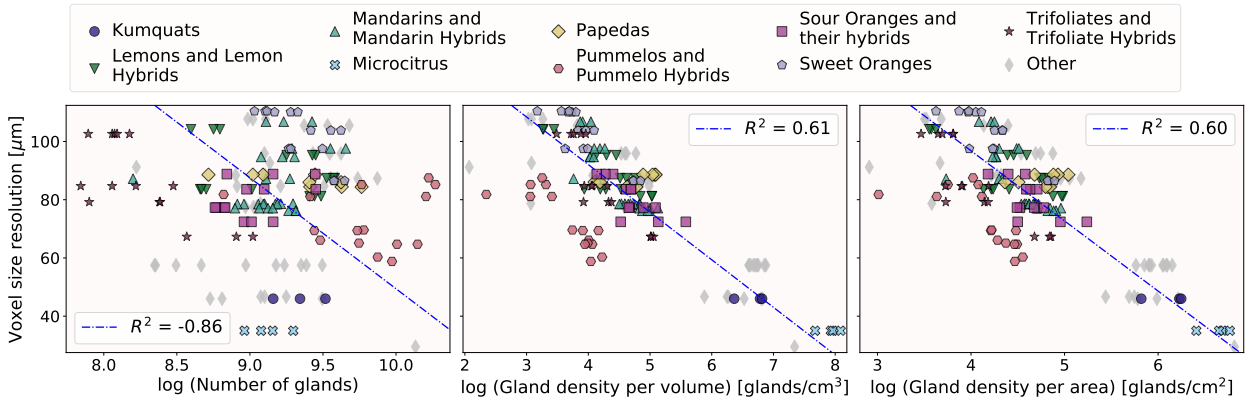

Figure B.1: Comparing the effect of different scan resolutions. A negative R^2 value suggests no correlation between the scan resolution and the number of oil glands extracted. Nonetheless, there is a possible correlation between scan resolution and oil gland density.

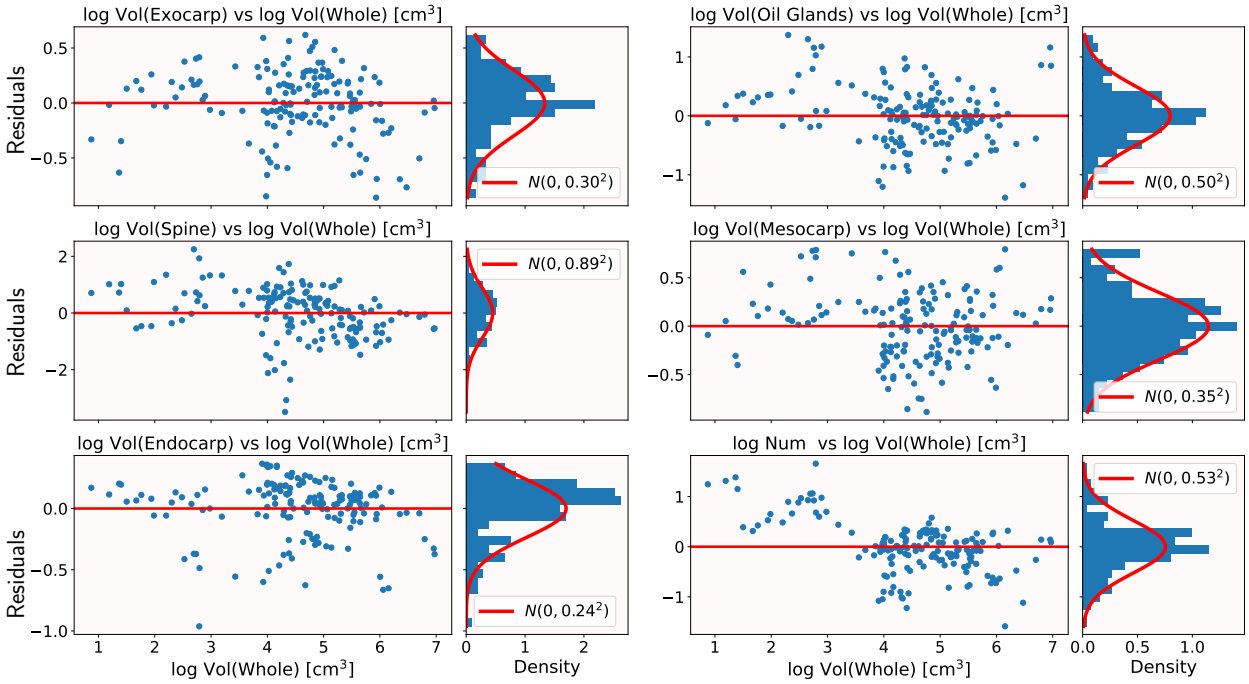

Figure B.2: Analysis of the distribution of residuals. The left side of each column are the residuals of the fitted linear regression from Figure 4.2. The right side of each column shows the distribution of these residuals. For some of these measurement pairs of traits, the residuals follow a normal distribution, suggesting that the linear fit in the log-log plots is adequate.

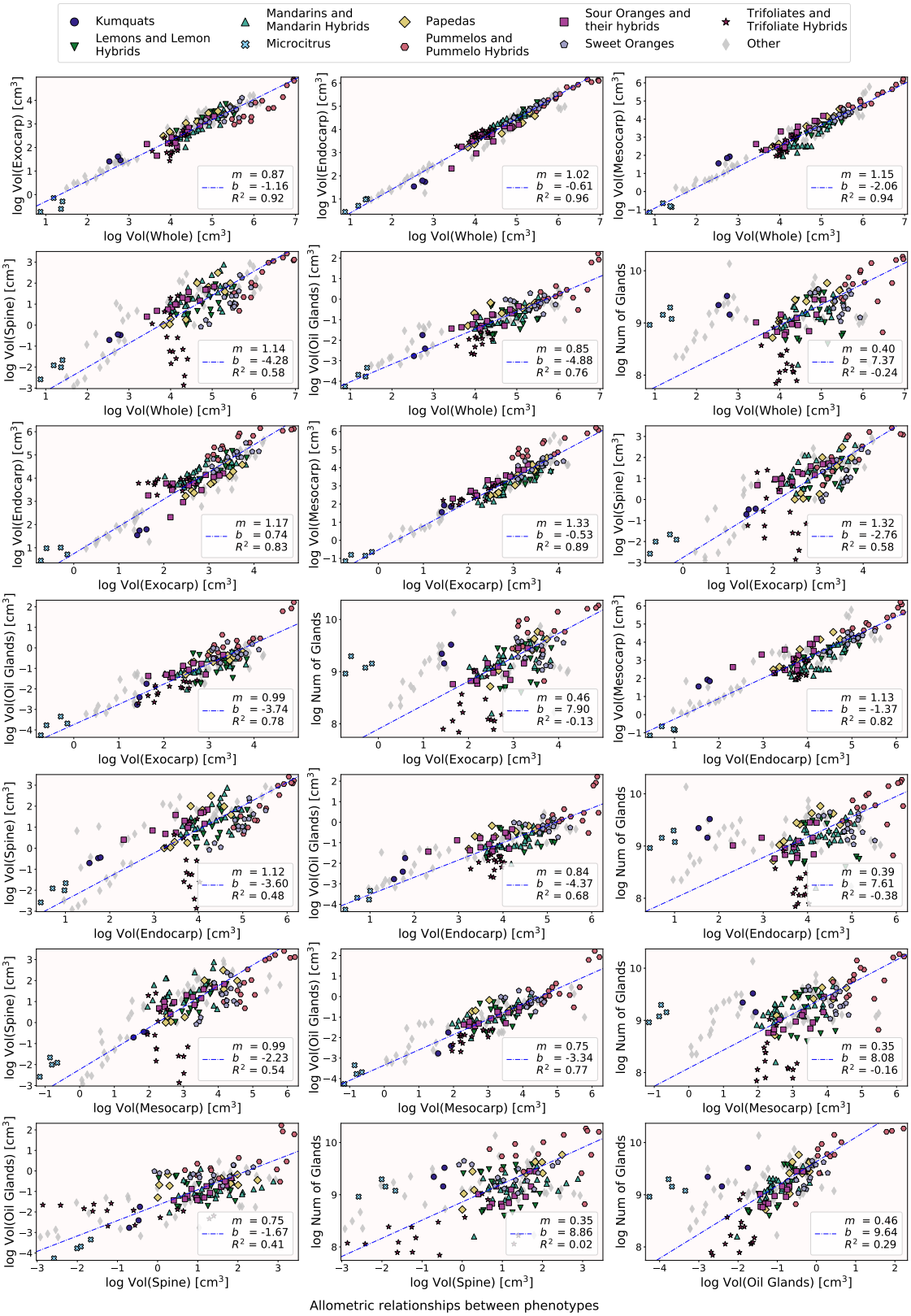

Figure B.3: Allometry plots for all possible pairs of measured phenotypes. Different citrus species are denoted with different markers.

Figure B.3 (cont'd):
Allometry plots for all possible pairs of measured phenotypes. Different citrus species are denoted with different markers.We observe a strong allometric relationship between different tissue volumes. However, this relationship is missing when comparing the total number of oil glands to all tissue volumes, suggesting that the number of glands is decoupled from these volume traits. The best fit line is depicted by a dashed line in blue. For each plot, the slope, intercept, and correlation coefficient are recorded as m, b, and R^2 respectively. The linear relationship in the log-log plots suggests that fruit tissue volumes may grow following a power law.

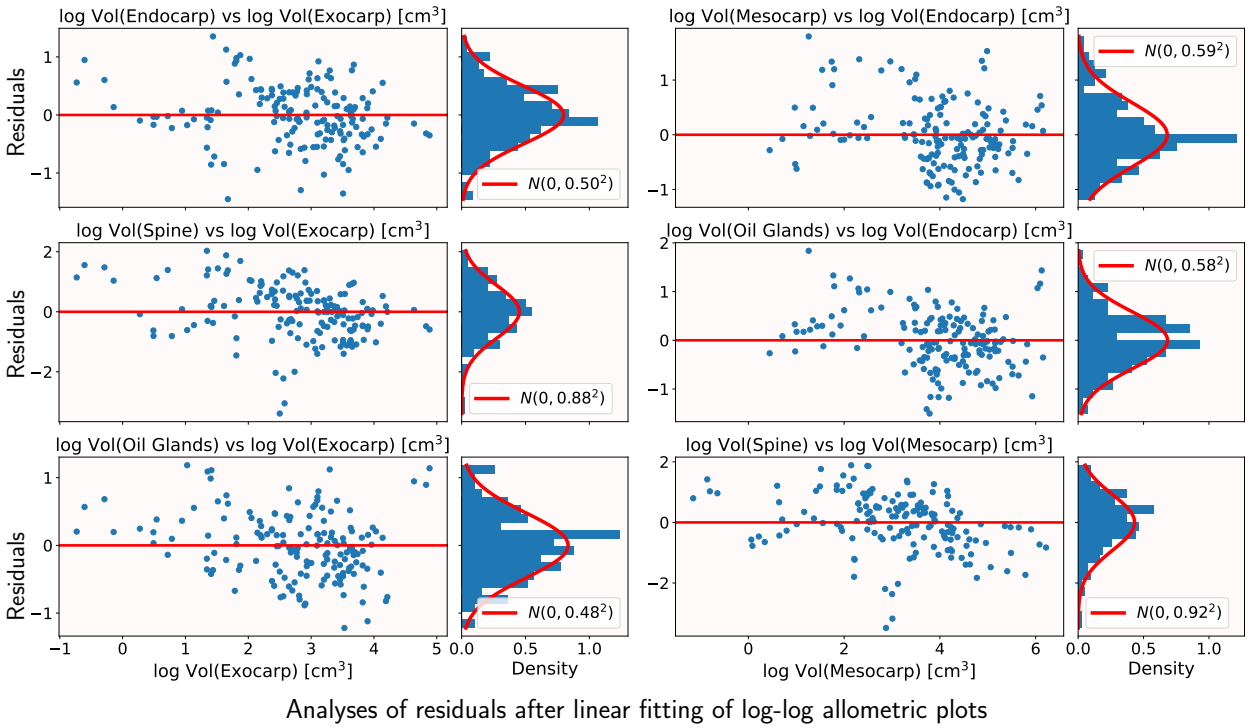

Figure B.4: Analysis of the distribution of residuals as in Figure B.2. The left side of each column depicts residuals of the linear fit as in Figure B.3. The right side of each column shows the distribution of these residuals. For some of these measurement pairs of traits, the residuals follow a normal distribution, suggesting that the linear fit in the log-log plots is adequate.

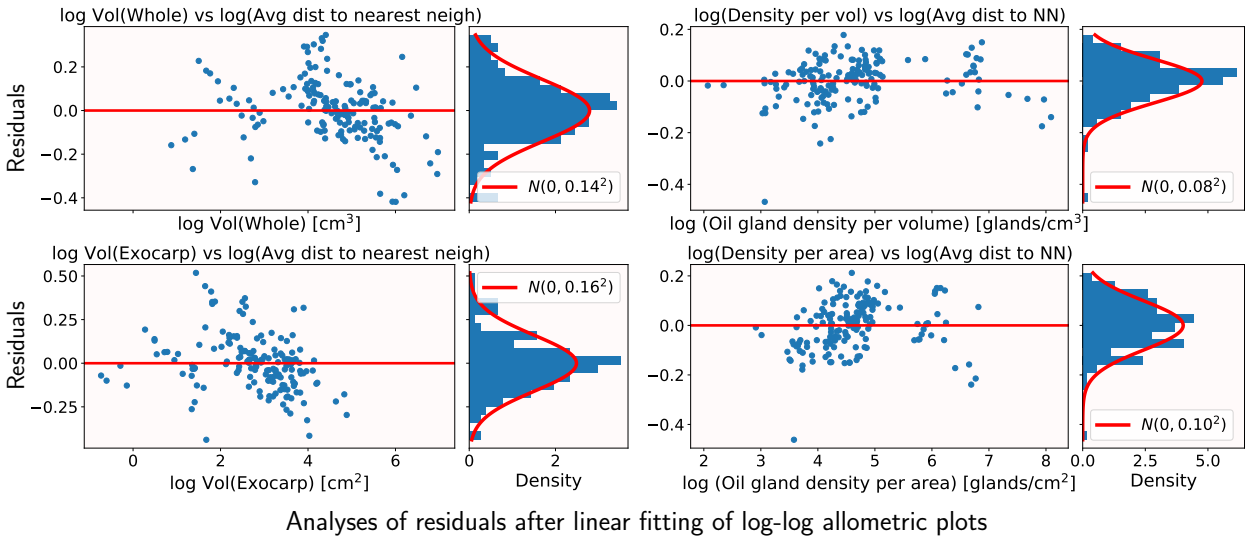

Figure B.5: Analysis of the distribution of residuals. The left side of each column depicts residuals of the linear fit as in Figure 4.3.A. The right side of each column shows the distribution of these residuals. For some of these measurement pairs of traits, the residuals follow a normal distribution, suggesting that the linear fit in the log-log plots is adequate.

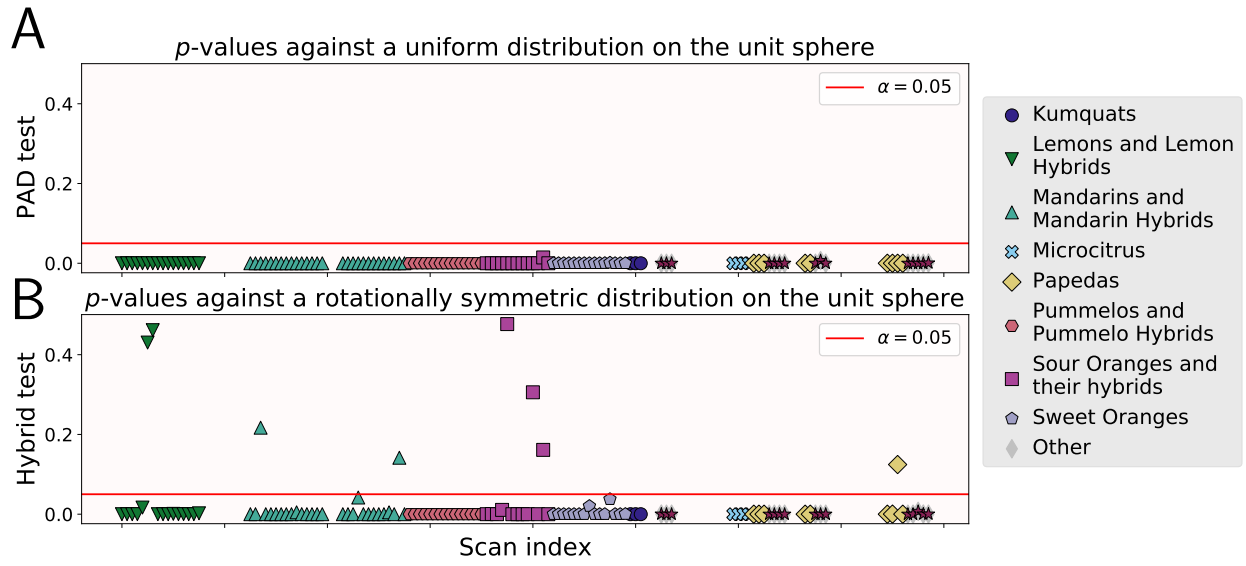

Figure B.6: Testing whether the oil glands are distributed uniformly or rotationally symmetric on the surface of the fruits. (A) p-values after testing if the underlying oil gland distribution is not uniform according to projected Anderson-Darling (PAD) test. (B) p-values after testing if the underlying oil gland distribution is not rotationally symmetric according to the scatter-location hybrid test. Red line at $\alpha = 0.05$ in all plots.

B.2 Supporting tables

Table B.1: A total of 166 different individual citrus fruits were scanned comprising 51 citrus accessions to represent modern cultivated citrus types as well as the species contributing ancestry to each group. All the shape and size analyses focused on a subset of these accessions, as indicated in Table 4.1. Names, identifiers, locations, and notes according to the University of California Givaudan Citrus Variety Collection (CVC). The **N** column indicates the number of pseudoreplicate fruits scanned per accession. Asterisk denotes not available.

ID	CVC_Name	Scientific Name	Type	Location	N
2317	Limon Real	*Citrus excelsa Wester*	Lemon	18B-18-9	4
3919	Lamas lemon	*Citrus limon L. Burm.f.*	Lemon	18A-24-1	3
3593	Interdonato lemon o.p. seedling	*Citrus limon L. Burm.f.*	Lemon-Hybrid	18A-8-3	3
3005	Frost nucellar Eureka lemon	*Citrus limon L. Burm.f.*	Eureka-type Lemon	18B-37-1	3
3050	Volckamer lemon o.p. seedling	*Citrus volkameriana*	Rough Lemon	18B-29-1	3
1482	Palestine sweet lime (Indian sweet lime)	*Citrus limettiodes Tan.*	Lime-Sweet	18B-27-7	3
661	Indian citron o.p. seedling (Zamburi)	*Citrus medica L.*	Citron-Hybrid	18B-13-7	3
3546	South Coast Field Station citron	*Citrus medica L.*	Citron	18A-5-3	3
3226	Scarlet Emperor mandarin o.p. seedling Pankan	*Citrus reticulata Blanco*	Mandarin	12B-27-1	3
3812	*	Citrus reticulata o.p. seedling	Mandarin	12B-29-1	3
3851	Lee mandarin (Clementine X Orlando)	*Citrus reticulata Blanco RUTACEAE*	Mandarin	12B-25-9	3
3752	Som Keowan o.p. seedling	*Citrus reticulata Blanco*	Mandarin	12B-28-7	3
3958	Koster tangor	*Citrus reticulata Blanco*	Tangor Mandarines	12A-23-13	3
3844	Cleopatra mandarin o.p. seedling	*Citrus reshni hort. ex Tanaka RUTACEAE*	Mandarin	12B-31-1	3
3363	Beledy mandarin o.p. seedling	*Citrus reticulata Blanco*	Mandarin	12B-23-1	4
3991	USDA 88-2 (Lee X Nova)	*Citrus reticulata Blanco*	Mandarin-Hybrid	12A-22-13	3
3816	Kinkoji Unshiu (graft hybrid of C. obovoidea + Satsuma)	*Citrus neo-aurantium*	Mandarin-Hybrid	12B-24-11	3
3558	Fremont mandarin	*Citrus reticulata Blanco RUTACEAE*	Mandarin	12B-23-15	3
3781	Tahitian pummelo X Star Ruby grapefruit	*Citrus paradisi Macfadyen*	Pummelo-Hybrid	12A-31-9	3

Table B.1 (cont'd)

ID	CVC_Name	Scientific Name	Type	Location	N
4026	Pomelit pummelo hybrid (Djeroek Deleema Kopjor)	*Citrus maxima (Burm.) Merr*	Pummelo-Hybrid	12A-19-7	3
3907	Hassaku (Citrus hassaku, Beni Hassaku)	*Citrus hassaku hort. ex Tanaka RUTACEAE*	Pummelo-Hybrid	12B-41-3	3
3959	Egami Buntan (Egami, Ogami) pummelo	*Citrus maxima (Burm.) Merr. RUTACEAE*	Pummelo	12B-48-13	3
2242	Kao Panne pummelo (Kao Pan)	*Citrus maxima (Burm.) Merr. RUTACEAE*	Pummelo	12A-35-3	3
3289	Willowleaf sour orange	*Citrus aurantium var. salicifolia*	Sour Orange	12B-19-9	3
628	Standard sour orange	*Citrus aurantium L.*	Sour Orange	12B-18-5	3
3611	Konejime sour orange o.p. seedling	*Citrus neo-aurantium*	Sour Orange-Hybrid	12B-20-11	4
2717	Olivelands Sour orange	*Citrus aurantium L.*	Sour Orange	12B-40-1	3
3030	Cutter Valencia nucellar seedling	*Citrus sinensis L. Osbeck*	Valencia Sweet Orange	12B-8-5	3
3746	Shamouti orange, Israeli seedling #1	*Citrus sinensis L. Osbeck*	Sweet Orange	12B-17-7	3
1241-B	Parent Washington navel	*Citrus sinensis (L.) Osbeck*	Early/mid-season Navel Orange	12B-4-3	3
2802	Argentina sweet orange o.p. seedling	*Citrus sinensis (L.) Osbeck RUTACEAE*	Sweet Orange	12B-11-5	3
3994	Cara Cara pink fleshed navel	*Citrus sinensis (L.) Osbeck*	Early/mid-season Navel Orange	12A-25-9	3
3163	Indian wild orange	*Citrus indica Tanaka*	Citrus species	12B-30-1	3
2485	Nasnaran	*Citrus amblycarpa (Hassk.) Ochse*	Citrus species	12B-29-9	4
3228	Korai tachibana	*Citrus nippokoreana hort ex. Tanaka*	Citrus species	12B-30-9	3
2320	Winged lime (Talimasan)	*Citrus longispina Wester*	Citrus species	18B-18-3	3
3877	Nagami kumquat	*Fortunella margarita (Lour.) Swingle*	Kumquat	12B-44-13	3
3661	Australian finger lime	*Microcitrus australasica*	Microcitrus	18B-16-5	4
3605	Samuyao	*Citrus micrantha var. microcarpa*	Papeda	18B-19-5	6

Table B.1 (cont'd)

ID	CVC_Name	Scientific Name	Type	Location	N
2327	Ichang papeda	*Citrus cavaleriei H. Lév. ex Cavalerie*	Papeda	18B-9-5	1
3469	*	*Citrus hanayu Siebold ex Shirai*	Papeda	12B-36-9	3
1455	Kalpi, Nogapog, Malayan lemmon	*Citrus webberii*	Papeda	18B-17-6	2
2454	Makrut lime	*Citrus hystrix DC.*	Papeda	18B-20-5	4
3842	Alemow (Colo)	*Citrus macrophylla Wester*	Papeda	18B-18-5	3
1491	Chinese box orange	*Severinia buxifolia*	Relative	12D-26-15	3
3140	Indian bael fruit	*Aegle marmelos (L.) Corrêa*	Relative	18B-9-1	3
4008	Little-leaf trifoliate	*Poncirus trifoliata (L.) Raf. RUTACEAE*	Trifoliate	12A-24-5	3
838	Rubidoux trifoliate	*Poncirus trifoliata (L.) Raf. RUTACEAE*	Trifoliate	12A-5-5	4
3912	C-35 citrange (Ruby orange x Webber-Fawcett trifoliate)	*X Citroncirus spp*	Trifoliate-Hybrid	12B-41-11	3
2863	Carrizo citrange o.p. seedling ('Washington' sweet orange X Poncirus trifoliata)	*X Citroncirus sp. RUTACEAE*	Trifoliate-Hybrid	12A-39-9	5
3771	Swingle citrumelo ('Duncan' grapefruit x Poncirus trifoliata)	*X Citroncirus spp. RUTACEAE*	Trifoliate-Hybrid	12A-42-5	4

Table B.2: Technical details for all the 63 citrus fruit scans produced comprising 166 individual fruits representing 51 citrus accessions (see Table B.1 for details). The scans were produced using the North Star Imaging X3000 system and the included efX software, with 720 projections per scan, at 3 frames per second and with 3 frames averaged per projection. The data was obtained in continuous mode. The fruits were placed as close as possible to the X-ray detector, provided all fruits of the same accession could be scanned completely at once, which resulted in varying voxel size resolutions after reconstruction. Pummelo and citron fruits were scanned individually due to their large size. The X-ray source was set to different current and voltages for different fruits, resulting in varying focal spot sizes. Voxel size, voltage, current, and focal size denoted by **vs**, **V**, **C**, and **fs** respectively.

scan_ID	**CVC_name**	**vs [µm]**	**V [kV]**	**C [µÅ]**	**fs [µm]**
C01_CRC2317_18B-18-9	Limon Real	83.6	75	70	5.25
C02_CRC3919_18A-24-1	Lamas lemon	81.2	75	70	5.25
C03_CRC3593_18A-8-3	Interdonato lemon o.p. seedling	104.2	75	70	5.25
C04_CRC3005_18B_37_1	Frost nucellar Eureka lemon	95.3	75	70	5.25
C05_CRC3050_18B-29-1	Volckamer lemon o.p. seedling	87.4	75	70	5.25
C06_CRC1482_18B-27-7	Palestine sweet lime (Indian sweet lime)	88.4	75	70	5.25
C07_CRC0661_18B-13-7	Indian citron o.p. seedling (Zamburi)	105.5	90	70	6.3
C08A_CRC3546_18A-5-3	South Coast Field Station citron	78.7	90	70	6.3
C08B_CRC3546_18A-5-3	South Coast Field Station citron	91.2	90	70	6.3
C08C_CRC3546_18A-5-3	South Coast Field Station citron	92.6	90	70	6.3
M01_CRC3226_12B-27-1	Scarlet Emperor mandarin o.p. seedling Pankan	78.6	70	70	4.9
M02_CRC3812_12B-29-1	*Citrus reticulata Blanco*	94.7	70	70	4.9
M03_CRC3851_12B-25-9	Lee mandarin (Clementine X Orlando)	87.2	70	70	4.9
M04_CRC3752_12B-28-7	Som Keowan o.p. seedling	78.5	70	70	4.9
M05_CRC3958_12A-23-13	Koster tangor	78.5	70	70	4.9
M06_CRC3844_12B-31-1	Cleopatra mandarin o.p. seedling	46.7	70	70	4.9
M07_CRC3363_12B-23-1	Beledy mandarin o.p. seedling	77.1	70	70	4.9
M08_CRC3991_12A-22-13	USDA 88-2 (Lee X Nova)	97.6	70	70	4.9
M09_CRC3816_12B-24-11	Kinkoji Unshiu (graft hybrid of C. obovoidea + Satsuma)	106.8	70	70	4.9
M10_CRC3558_12B-23-15	Fremont mandarin	76.2	70	70	4.9
P01A_CRC3781_12A-31-9	Tahitian pummelo X Star Ruby grapefruit	60.3	90	70	6.3
P01B_CRC3781_12A-31-9	Tahitian pummelo X Star Ruby grapefruit	58.8	90	70	6.3

Table B.2 (cont'd)

scan_ID	CVC_name	vs [µm]	V [kV]	C [µÅ]	fs [µm]
P01C_CRC3781_12A-31-9	Tahitian pummelo X Star Ruby grapefruit	66.1	90	70	6.3
P02A_CRC4026_12A-19-7	Pomelit pummelo hybrid (Djeroek Deleema Kopjor)	64.7	90	70	6.3
P02B_CRC4026_12A-19-7	Pomelit pummelo hybrid (Djeroek Deleema Kopjor)	65.1	90	70	6.3
P02C_CRC4026_12A-19-7	Pomelit pummelo hybrid (Djeroek Deleema Kopjor)	69.4	90	70	6.3
P03A_CRC3907_12B-41-3	Hassaku (Citrus hassaku, Beni Hassaku)	64.7	90	70	6.3
P03B_CRC3907_12B-41-3	Hassaku (Citrus hassaku, Beni Hassaku)	69.6	90	70	6.3
P03C_CRC3907_12B-41-3	Hassaku (Citrus hassaku, Beni Hassaku)	69.4	90	70	6.3
P04A_CRC3959_12B-48-13	Egami Buntan (Egami, Ogami) pummelo	85.2	90	70	6.3
P04B_CRC3959_12B-48-13	Egami Buntan (Egami, Ogami) pummelo	81.8	90	70	6.3
P04C_CRC3959_12B-48-13	Egami Buntan (Egami, Ogami) pummelo	81.1	90	70	6.3
P05A_CRC2242_12A-35-3	Kao Panne pummelo (Kao Pan)	85.2	90	70	6.3
P05B_CRC2242_12A-35-3	Kao Panne pummelo (Kao Pan)	87.5	90	70	6.3
P05C_CRC2242_12A-35-3	Kao Panne pummelo (Kao Pan)	81.1	90	70	6.3
SR01_CRC3289_12B-19-9	Willowleaf sour orange	72.4	75	70	5.25
SR02_CRC0628_12B-18-5	Standard sour orange	88.8	75	70	5.25
SR03_CRC3611_12B-20-11	Konejime sour orange o.p. seedling	77.3	75	70	5.25
SR04_CRC2717_12B-40-1	Olivelands Sour orange	83.6	75	70	5.25
SW01_CRC3030_12B-8-5	Cutter Valencia nucellar seedling	103.8	75	70	5.25
SW02_CRC3746_12B-17-7	Shamouti orange, Israeli seedling #1	110.1	75	70	5.25
SW03_CRC1241-B_12B-4-3	Parent Washington navel	110.4	75	70	5.25
SW04_CRC2802_12B-11-5	Argentina sweet orange o.p. seedling	86.5	75	70	5.25
SW05_CRC3994_12A-25-9	Cara Cara pink fleshed navel	97.5	75	70	5.25
WR01_CRC1491_12D-26-15	Chinese box orange	18.6	70	70	4.9
WR02_CRC3877_12B-44-13	Nagami kumquat	46	70	70	4.9
WR03_CRC3163_12B-30-1	Indian wild orange	46	70	70	4.9
WR04_CRC4008_12A-24-5	Little-leaf trifoliate	67.3	70	70	4.9
WR05_CRC3605_18B-19-5	Samuyao	57.5	70	70	4.9
WR06_CRC2485_12B-29-9	Nasnaran	57.5	70	70	4.9
WR07_CRC2327_18B-9-5	Ichang papeda	29.5	70	70	4.9
WR08_CRC3661_18B-16-5	Australian finger lime	35	70	70	4.9

Table B.2 (cont'd)

scan_ID	CVC_name	vs [µm]	V [kV]	C [µÅ]	fs [µm]
WR09_CRC3469_12B-36-9	*Citrus hanayu Siebold ex Shirai*	84.4	70	70	4.9
WR10_CRC0838_12A-5-5	Rubidoux trifoliate	84.7	70	70	4.9
WR11_CR3228_12B-30-9	Korai tachibana	84.7	70	70	4.9
WR12_CRC1455_18B-17-6	Kalpi, Nogapog, Malayan lemmon	86	70	70	4.9
WR13_CRC3912_12B-41-11	C-35 citrange (Ruby orange x Webber-Fawcett trifoliate)	79.2	70	70	4.9
WR14_CRC2320_18B-18-3	Winged lime (Talimasan)	95.9	70	70	4.9
WR15_CRC3140_18B-9-1	Indian bael fruit	77.5	70	70	4.9
WR16_CRC3771_12A-42-5	Swingle citrumelo ('Duncan' grapefruit x Poncirus trifoliata)	87.3	70	70	4.9
WR17_CRC2454_18B-20-5	Makrut lime	88.6	70	70	4.9
WR18_CRC2863_12A-39-9	Carrizo citrange o.p. seedling ('Washington' sweet orange X Poncirus trifoliata)	102.6	70	70	4.9
WR19_CRC3842_18B-18-5	Alemow (Colo)	107.7	70	70	4.9

Table B.3: We tested whether the oil glands on the citrus fruit skins are either uniformly or rotationally symmetrically distributed (see main text for details). Uniformity was tested with Projected Anderson-Darling (PAD) test. Rotational symmetry was tested with a scatter-location hybrid von Mises-Fisher (vMF) test with an unspecified direction of symmetry. The resulting statistics and p-values are reported below. Individual fruits are identified by their original scan_ID (as in Table B.2). Individual pseudoreplicates are identified with different labels.

scan_ID	label	kind	s_PAD	p_PAD	s_vMF	p_vMF
C01_CRC2317_18B-18-9	L00	Lemon	9.80E+01	0.00E+00	3.41E+02	1.70E-72
C01_CRC2317_18B-18-9	L01	Lemon	8.73E+01	0.00E+00	1.47E+02	8.00E-31
C01_CRC2317_18B-18-9	L02	Lemon	4.62E+01	1.20E-09	2.09E+02	5.12E-44
C01_CRC2317_18B-18-9	L03	Lemon	2.96E+01	4.12E-09	1.20E+02	6.29E-25
C02_CRC3919_18A-24-1	L00	Lemon	1.12E+01	1.28E-08	1.21E+01	1.66E-02
C02_CRC3919_18A-24-1	L01	Lemon	1.14E+01	0.00E+00	3.82E+00	4.30E-01
C02_CRC3919_18A-24-1	L02	Lemon	1.24E+01	2.93E-08	3.61E+00	4.62E-01
C03_CRC3593_18A-8-3	L00	Lemon	6.68E+01	2.73E-09	7.93E+02	2.47E-170
C03_CRC3593_18A-8-3	L01	Lemons	6.49E+01	1.69E-08	8.66E+02	3.06E-186
C03_CRC3593_18A-8-3	L02	Lemons	4.07E+01	0.00E+00	4.80E+02	1.48E-102
C04_CRC3005_18B_37_1	L00	Lemons	4.07E+01	1.01E-08	1.88E+01	8.52E-04
C04_CRC3005_18B_37_1	L01	Lemons	1.85E+01	1.94E-09	9.35E+01	2.41E-19
C04_CRC3005_18B_37_1	L02	Lemons	3.86E+01	0.00E+00	2.26E+01	1.55E-04
C05_CRC3050_18B-29-1	L00	Lemons	1.06E+02	0.00E+00	1.60E+02	1.41E-33
C05_CRC3050_18B-29-1	L01	Lemons	1.63E+01	4.43E-08	9.65E+01	5.38E-20
C05_CRC3050_18B-29-1	L02	Lemons	1.63E+01	3.47E-08	1.67E+01	2.22E-03
C06_CRC1482_18B-27-7	L00	Other	4.91E+01	0.00E+00	2.99E+02	1.54E-63
C06_CRC1482_18B-27-7	L01	Other	2.19E+01	0.00E+00	3.38E+01	8.33E-07
C06_CRC1482_18B-27-7	L02	Other	1.29E+01	2.29E-08	8.51E+01	1.47E-17
C07_CRC0661_18B-13-7	L00	Other	2.09E+01	0.00E+00	5.41E+01	4.92E-11
C07_CRC0661_18B-13-7	L01	Other	3.73E+01	0.00E+00	3.99E+01	4.60E-08
C07_CRC0661_18B-13-7	L02	Other	1.63E+01	1.37E-08	6.06E+01	2.14E-12
C08A_CRC3546_18A-5-3	L00	Other	3.98E+02	2.52E-07	9.76E+02	7.25E-210
C08B_CRC3546_18A-5-3	L00	Other	2.12E+02	9.15E-09	5.92E+01	4.37E-12
C08C_CRC3546_18A-5-3	L00	Other	5.22E+02	7.24E-09	1.38E+03	9.37E-298

Table B.3 (cont'd)

scan_ID	label	kind	s_PAD	p_PAD	s_vMF	p_vMF
M01_CRC3226_12B-27-1	L00	Mandarins	2.44E+01	0.00E+00	1.12E+02	3.40E-23
M01_CRC3226_12B-27-1	L01	Mandarins	2.48E+01	0.00E+00	4.75E+01	1.18E-09
M01_CRC3226_12B-27-1	L02	Mandarins	2.19E+01	0.00E+00	5.77E+00	2.17E-01
M02_CRC3812_12B-29-1	L00	Mandarins	5.45E+01	1.97E-09	6.60E+02	1.30E-141
M02_CRC3812_12B-29-1	L01	Mandarins	6.41E+01	1.22E-08	3.01E+02	6.71E-64
M02_CRC3812_12B-29-1	L02	Mandarins	2.58E+01	0.00E+00	1.54E+02	2.52E-32
M03_CRC3851_12B-25-9	L00	Mandarins	3.18E+01	0.00E+00	9.94E+01	1.29E-20
M03_CRC3851_12B-25-9	L01	Mandarins	1.72E+01	0.00E+00	4.52E+01	3.66E-09
M03_CRC3851_12B-25-9	L02	Mandarins	1.55E+01	4.30E-08	3.12E+01	2.74E-06
M04_CRC3752_12B-28-7	L00	Mandarins	8.15E+00	0.00E+00	1.50E+01	4.72E-03
M04_CRC3752_12B-28-7	L01	Mandarins	8.76E+00	1.50E-08	2.34E+01	1.04E-04
M04_CRC3752_12B-28-7	L02	Mandarins	7.81E+00	0.00E+00	2.45E+01	6.30E-05
M05_CRC3958_12A-23-13	L00	Mandarins	4.18E+01	1.38E-08	1.82E+02	2.56E-38
M05_CRC3958_12A-23-13	L01	Mandarins	1.68E+01	0.00E+00	6.30E+01	6.73E-13
M05_CRC3958_12A-23-13	L02	Mandarins	5.42E+01	1.40E-08	1.33E+02	8.60E-28
M06_CRC3844_12B-31-1	L00	Other	4.26E+01	1.32E-08	6.50E+02	2.60E-139
M06_CRC3844_12B-31-1	L01	Other	3.56E+01	3.15E-08	5.69E+02	8.75E-122
M06_CRC3844_12B-31-1	L02	Other	2.01E+01	0.00E+00	9.50E+00	4.98E-02
M07_CRC3363_12B-23-1	L00	Mandarins	2.19E+01	3.22E-10	2.67E+02	1.74E-56
M07_CRC3363_12B-23-1	L01	Mandarins	2.58E+01	0.00E+00	1.50E+02	2.27E-31
M07_CRC3363_12B-23-1	L02	Mandarins	2.01E+01	0.00E+00	1.98E+02	8.31E-42
M07_CRC3363_12B-23-1	L03	Mandarins	1.41E+01	8.69E-09	9.92E+00	4.18E-02
M08_CRC3991_12A-22-13	L00	Mandarins	3.53E+01	0.00E+00	1.33E+02	9.16E-28
M08_CRC3991_12A-22-13	L01	Mandarins	2.02E+01	0.00E+00	1.16E+02	3.65E-24
M08_CRC3991_12A-22-13	L02	Mandarins	3.28E+01	0.00E+00	4.83E+01	8.17E-10
M09_CRC3816_12B-24-11	L00	Mandarins	6.19E+00	2.19E-07	3.09E+01	3.19E-06
M09_CRC3816_12B-24-11	L01	Mandarins	6.58E+00	1.04E-07	3.30E+01	1.18E-06
M09_CRC3816_12B-24-11	L02	Mandarins	8.70E+00	2.54E-08	1.50E+01	4.70E-03
M10_CRC3558_12B-23-15	L00	Mandarins	4.92E+01	1.70E-09	3.14E+02	9.37E-67
M10_CRC3558_12B-23-15	L01	Mandarins	8.32E+00	1.70E-08	6.89E+00	1.42E-01

Table B.3 (cont'd)

scan_ID	label	kind	s_PAD	p_PAD	s_vMF	p_vMF
M10_CRC3558_12B-23-15	L02	Mandarins	1.44E+01	0.00E+00	2.75E+01	1.54E-05
P01A_CRC3781_12A-31-9	L00	Pummelos	3.32E+01	3.12E-08	1.01E+02	5.17E-21
P01B_CRC3781_12A-31-9	L00	Pummelos	1.64E+02	0.00E+00	3.68E+02	1.94E-78
P01C_CRC3781_12A-31-9	L00	Pummelos	2.67E+01	0.00E+00	8.44E+01	2.04E-17
P02A_CRC4026_12A-19-7	L00	Pummelos	5.88E+01	1.68E-08	4.76E+01	1.15E-09
P02B_CRC4026_12A-19-7	L00	Pummelos	1.50E+02	5.20E-06	1.54E+02	2.78E-32
P02C_CRC4026_12A-19-7	L00	Pummelos	7.30E+01	3.60E-09	3.16E+01	2.28E-06
P03A_CRC3907_12B-41-3	L00	Pummelos	6.88E+01	3.47E-08	3.42E+02	9.83E-73
P03B_CRC3907_12B-41-3	L00	Pummelos	7.32E+01	1.33E-08	7.47E+01	2.29E-15
P03C_CRC3907_12B-41-3	L00	Pummelos	2.78E+01	0.00E+00	1.46E+02	1.26E-30
P04A_CRC3959_12B-48-13	L00	Pummelos	1.14E+02	0.00E+00	5.19E+01	1.46E-10
P04B_CRC3959_12B-48-13	L00	Pummelos	2.27E+02	2.37E-09	7.96E+02	4.59E-171
P04C_CRC3959_12B-48-13	L00	Pummelos	1.37E+02	0.00E+00	2.68E+02	9.57E-57
P05A_CRC2242_12A-35-3	L00	Pummelos	1.16E+02	0.00E+00	5.92E+01	4.26E-12
P05B_CRC2242_12A-35-3	L00	Pummelos	3.83E+01	0.00E+00	6.26E+01	8.41E-13
P05C_CRC2242_12A-35-3	L00	Pummelos	8.52E+01	5.46E-09	8.52E+01	1.36E-17
SR01_CRC3289_12B-19-9	L00	Sour Oranges	2.65E+01	0.00E+00	2.72E+01	1.78E-05
SR01_CRC3289_12B-19-9	L01	Sour Oranges	1.85E+01	3.78E-08	2.00E+01	5.00E-04
SR01_CRC3289_12B-19-9	L02	Sour Oranges	1.55E+01	4.12E-08	3.58E+01	3.18E-07
SR02_CRC0628_12B-18-5	L00	Sour Oranges	3.72E+01	0.00E+00	1.31E+01	1.08E-02
SR02_CRC0628_12B-18-5	L01	Sour Oranges	2.31E+01	8.41E-10	3.51E+00	4.77E-01
SR02_CRC0628_12B-18-5	L02	Sour Oranges	1.59E+01	4.80E-08	9.22E+01	4.42E-19
SR03_CRC3611_12B-20-11	L00	Sour Oranges	1.67E+01	2.91E-08	5.73E+01	1.09E-11
SR03_CRC3611_12B-20-11	L01	Sour Oranges	1.21E+01	1.45E-08	4.14E+01	2.20E-08
SR03_CRC3611_12B-20-11	L02	Sour Oranges	6.64E+00	8.48E-08	2.51E+01	4.70E-05
SR03_CRC3611_12B-20-11	L03	Sour Oranges	9.85E+00	2.44E-08	4.83E+00	3.06E-01
SR04_CRC2717_12B-40-1	L00	Sour Oranges	2.36E+01	0.00E+00	3.54E+01	3.86E-07
SR04_CRC2717_12B-40-1	L01	Sour Oranges	2.28E+00	1.46E-02	6.56E+00	1.61E-01
SR04_CRC2717_12B-40-1	L02	Sour Oranges	9.70E+00	0.00E+00	1.07E+02	3.28E-22
SW01_CRC3030_12B-8-5	L00	Sweet Oranges	9.69E+01	0.00E+00	2.19E+02	3.07E-46

Table B.3 (cont'd)

scan_ID	label	kind	s_PAD	p_PAD	s_vMF	p_vMF
SW01_CRC3030_12B-8-5	L01	Sweet Oranges	1.38E+01	4.17E-08	1.76E+02	6.51E-37
SW01_CRC3030_12B-8-5	L02	Sweet Oranges	2.34E+01	1.47E-08	5.11E+01	2.12E-10
SW02_CRC3746_12B-17-7	L00	Sweet Oranges	2.19E+01	0.00E+00	1.65E+02	1.48E-34
SW02_CRC3746_12B-17-7	L01	Sweet Oranges	3.02E+01	5.90E-08	1.43E+02	5.54E-30
SW02_CRC3746_12B-17-7	L02	Sweet Oranges	2.10E+01	1.21E-08	1.42E+02	9.91E-30
SW03_CRC1241-B_12B-4-3	L00	Sweet Oranges	4.90E+01	1.46E-07	3.91E+01	6.80E-08
SW03_CRC1241-B_12B-4-3	L01	Sweet Oranges	2.45E+01	1.23E-08	1.16E+01	2.03E-02
SW03_CRC1241-B_12B-4-3	L02	Sweet Oranges	1.35E+01	0.00E+00	1.25E+02	4.22E-26
SW04_CRC2802_12B-11-5	L00	Sweet Oranges	9.34E+00	1.08E-08	3.02E+01	4.46E-06
SW04_CRC2802_12B-11-5	L01	Sweet Oranges	3.80E+00	1.99E-04	5.62E+01	1.86E-11
SW04_CRC2802_12B-11-5	L02	Sweet Oranges	3.17E+00	1.20E-03	1.02E+01	3.75E-02
SW05_CRC3994_12A-25-9	L00	Sweet Oranges	8.29E+01	0.00E+00	2.48E+02	1.95E-52
SW05_CRC3994_12A-25-9	L01	Sweet Oranges	3.29E+01	6.03E-09	1.01E+02	4.83E-21
SW05_CRC3994_12A-25-9	L02	Sweet Oranges	6.17E+01	6.62E-08	2.12E+01	2.83E-04
WR02_CRC3877_12B-44-13	L00	Kumquats	9.03E+01	0.00E+00	1.35E+02	2.99E-28
WR02_CRC3877_12B-44-13	L01	Kumquats	1.22E+02	1.62E-08	2.16E+02	1.59E-45
WR02_CRC3877_12B-44-13	L02	Kumquats	8.95E+01	0.00E+00	6.04E+02	1.67E-129
WR03_CRC3163_12B-30-1	L00	Other	1.14E+02	0.00E+00	5.09E+02	7.08E-109
WR03_CRC3163_12B-30-1	L01	Other	2.96E+01	9.10E-09	3.62E+01	2.65E-07
WR03_CRC3163_12B-30-1	L02	Other	1.82E+01	0.00E+00	2.03E+01	4.39E-04
WR04_CRC4008_12A-24-5	L00	Trifoliates	3.84E+01	5.98E-09	2.15E+01	2.54E-04
WR04_CRC4008_12A-24-5	L01	Trifoliates	2.06E+01	5.37E-09	8.45E+01	1.98E-17
WR04_CRC4008_12A-24-5	L02	Trifoliates	1.43E+01	0.00E+00	5.48E+01	3.59E-11
WR05_CRC3605_18B-19-5	L00	Other	1.40E+01	1.26E-08	2.54E+01	4.19E-05
WR05_CRC3605_18B-19-5	L01	Other	1.66E+01	0.00E+00	1.22E+01	1.62E-02
WR05_CRC3605_18B-19-5	L02	Other	7.77E+00	9.73E-09	3.67E+01	2.06E-07
WR05_CRC3605_18B-19-5	L03	Other	9.30E+00	1.48E-08	3.58E+01	3.17E-07
WR05_CRC3605_18B-19-5	L04	Other	7.41E+00	1.36E-08	9.91E+00	4.19E-02
WR05_CRC3605_18B-19-5	L05	Other	6.95E+00	3.44E-08	8.51E-01	9.32E-01
WR06_CRC2485_12B-29-9	L00	Other	2.26E+01	5.32E-09	1.90E+01	7.90E-04

Table B.3 (cont'd)

scan_ID	label	kind	s_PAD	p_PAD	s_vMF	p_vMF
WR06_CRC2485_12B-29-9	L01	Other	3.20E+01	0.00E+00	8.80E-01	9.27E-01
WR06_CRC2485_12B-29-9	L02	Other	2.21E+01	0.00E+00	2.28E+01	1.39E-04
WR06_CRC2485_12B-29-9	L03	Other	1.52E+01	3.36E-08	2.01E+01	4.83E-04
WR07_CRC2327_18B-9-5	L00	Other	9.28E+01	1.04E-08	1.31E+03	2.39E-281
WR08_CRC3661_18B-16-5	L00	Microcitrus	2.92E+02	0.00E+00	5.38E+03	0.00E+00
WR08_CRC3661_18B-16-5	L01	Microcitrus	3.66E+02	0.00E+00	6.30E+03	0.00E+00
WR08_CRC3661_18B-16-5	L02	Microcitrus	3.32E+02	9.20E-09	4.99E+03	0.00E+00
WR08_CRC3661_18B-16-5	L03	Microcitrus	2.60E+02	5.81E-10	6.40E+03	0.00E+00
WR09_CRC3469_12B-36-9	L00	Papedas	5.61E+01	0.00E+00	1.70E+02	9.47E-36
WR09_CRC3469_12B-36-9	L01	Papedas	2.34E+01	0.00E+00	2.26E+02	1.16E-47
WR09_CRC3469_12B-36-9	L02	Papedas	6.12E+00	2.67E-07	2.84E+01	1.02E-05
WR10_CRC0838_12A-5-5	L00	Trifoliates	2.28E+01	0.00E+00	1.56E+02	1.17E-32
WR10_CRC0838_12A-5-5	L01	Trifoliates	1.22E+01	1.58E-08	1.05E+02	7.15E-22
WR10_CRC0838_12A-5-5	L02	Trifoliates	2.76E+01	2.87E-08	2.27E+01	1.45E-04
WR10_CRC0838_12A-5-5	L03	Trifoliates	1.70E+01	0.00E+00	4.86E+01	6.94E-10
WR11_CR3228_12B-30-9	L00	Other	3.19E+01	6.61E-08	3.13E+01	2.69E-06
WR11_CR3228_12B-30-9	L01	Other	1.15E+01	8.61E-09	2.15E+01	2.54E-04
WR11_CR3228_12B-30-9	L02	Other	2.06E+01	2.24E-08	1.70E+02	1.25E-35
WR12_CRC1455_18B-17-6	L00	Papedas	5.62E+01	4.94E-09	4.69E+01	1.57E-09
WR12_CRC1455_18B-17-6	L01	Papedas	9.54E+01	0.00E+00	1.46E+02	1.72E-30
WR13_CRC3912_12B-41-11	L00	Trifoliates	6.16E+00	1.92E-07	6.81E+01	5.71E-14
WR13_CRC3912_12B-41-11	L01	Trifoliates	2.55E+00	6.76E-03	3.27E+01	1.35E-06
WR13_CRC3912_12B-41-11	L02	Trifoliates	2.57E+01	4.25E-07	2.58E+02	1.49E-54
WR14_CRC2320_18B-18-3	L00	Other	2.07E+01	0.00E+00	2.38E+01	8.62E-05
WR14_CRC2320_18B-18-3	L01	Other	1.28E+01	0.00E+00	2.99E+01	5.02E-06
WR14_CRC2320_18B-18-3	L02	Other	1.92E+01	1.60E-08	1.24E+01	1.47E-02
WR14_CRC2320_18B-18-3	L03	Other	1.00E+01	0.00E+00	3.38E+01	8.02E-07
WR15_CRC3140_18B-9-1	L00	Other	2.87E+01	3.52E-08	2.26E+01	1.54E-04
WR15_CRC3140_18B-9-1	L01	Other	2.23E+01	5.74E-09	1.43E+02	6.77E-30
WR15_CRC3140_18B-9-1	L02	Other	2.51E+01	1.90E-08	4.54E+01	3.21E-09

Table B.3 (cont'd)

scan_ID	label	kind	s_PAD	p_PAD	s_vMF	p_vMF
WR16_CRC3771_12A-42-5	L00	Other	1.80E+02	0.00E+00	1.09E+03	3.41E-235
WR16_CRC3771_12A-42-5	L01	Other	8.21E+01	3.55E-09	1.08E+02	2.04E-22
WR16_CRC3771_12A-42-5	L02	Other	1.16E+02	2.14E-08	9.25E+01	3.78E-19
WR16_CRC3771_12A-42-5	L03	Other	4.09E+01	0.00E+00	2.69E+02	6.05E-57
WR17_CRC2454_18B-20-5	L00	Papedas	2.22E+01	1.27E-09	8.35E+01	3.16E-17
WR17_CRC2454_18B-20-5	L01	Papedas	1.05E+01	2.96E-08	2.13E+01	2.76E-04
WR17_CRC2454_18B-20-5	L02	Papedas	7.45E+00	0.00E+00	7.22E+00	1.24E-01
WR17_CRC2454_18B-20-5	L03	Papedas	7.81E+00	5.49E-09	9.49E+01	1.19E-19
WR18_CRC2863_12A-39-9	L00	Trifoliates	3.07E+01	1.12E-09	3.97E+01	5.10E-08
WR18_CRC2863_12A-39-9	L01	Trifoliates	1.00E+01	5.03E-09	3.77E+01	1.32E-07
WR18_CRC2863_12A-39-9	L02	Trifoliates	2.00E+01	6.44E-09	1.40E+01	7.16E-03
WR18_CRC2863_12A-39-9	L03	Trifoliates	1.02E+01	3.62E-09	2.08E+01	3.50E-04
WR18_CRC2863_12A-39-9	L04	Trifoliates	3.38E+00	6.62E-04	2.17E+01	2.28E-04
WR19_CRC3842_18B-18-5	L00	Other	6.27E+01	0.00E+00	5.98E+02	4.79E-128
WR19_CRC3842_18B-18-5	L01	Other	3.48E+01	6.07E-09	1.05E+02	7.07E-22
WR19_CRC3842_18B-18-5	L02	Other	3.34E+01	1.33E-08	2.96E+02	7.33E-63

APPENDIX C

WALNUT APPENDIX

Table C.1: A total of 1301 individual walnuts were scanned, comprising 147 different accessions. The accessions are identified by their UCACCSD code, which is the identifying system used by the Walnut Improvement Program at the University of California, Davis. The number of pseudoreplicates scanned per accession are denoted by **N**. Asterisk denotes not available.

UCACCSD	**N**	**UCACCSD**	**N**
03-001-3395	7	12-042-3	9
06-004-4	7	12-042-4	7
06-005-27	9	12-042-5	7
06-030-18	5	12-042-7	8
08-001-28	6	12-042-8	9
08-002-4	7	12-042-9	8
08-006-11	8	12-045-11	9
08-019-11	9	12-045-12	4
09-003-20	7	12-045-13	7
09-025-107	9	12-045-14	7
09-025-117	9	12-045-15	5
09-025-123	8	12-045-3	7
09-025-13	9	12-045-7	7
09-025-24	8	12-045-8	9
09-025-60	8	12-045-9	9
09-025-62	8	12-048-11	6
09-025-64	8	12-048-13	7
09-025-69	7	12-048-14	7
09-025-72	9	12-048-3	8
09-025-78	9	12-048-4	7
09-025-99	9	12-048-5	4
1	15	12-048-6	8
10-001-6	9	12-048-7	8
10-005-3	9	12-048-8	8
10-008-16	9	12-048-9	5
10-008-23	9	12-053-19	9
10-008-37	7	12-053-20	9
10-008-63	9	12-053-22	8
10-008-71	5	12-053-23	8
10-008-73	6	12-053-24	9
10-008-76	6	12-053-3	9
10-016-34	9	12-053-4	9
10-018-1	8	12-053-5	7
10-019-84	7	12-054-10	9
10-020-10	9	12-054-11	8

Table C.1 (cont'd)

UCACCSD	N	UCACCSD	N
10-020-12	9	12-054-12	5
10-020-17	4	12-054-14	9
10-020-3	9	12-054-17	8
10-020-59	9	12-054-2	8
10-020-9	7	12-054-21	9
10-024-18	6	12-054-22	6
11-003-4	6	12-054-23	9
11-020-2	9	12-054-3	8
11-029-9	9	12-054-4	5
11-030-3	9	12-054-8	7
12-002-10	8	12-054-9	5
12-002-11	7	12-059-21	7
12-002-15	9	12-059-22	7
12-002-17	6	12-059-23	5
12-002-18	8	12-059-24	6
12-002-20	7	12-059-25	7
12-002-21	6	12-059-27	7
12-002-23	6	12-059-28	5
12-002-25	8	12-059-29	7
12-002-26	9	12-059-30	9
12-002-27	8	12-059-32	8
12-002-28	8	2	15
12-002-3	8	3	13
12-002-5	9	48	16
12-002-6	7	49-049	15
12-002-7	9	53-113	8
12-004-1	5	54	15
12-005-1	6	59-129	17
12-005-2	9	6	18
12-005-4	8	64-172	26
12-005-5	4	64-182	9
12-005-7	5	67-011	16
12-005-8	8	85-023-2	8
12-037-27	7	85-043-1	14
12-037-29	8	91-136	7
12-037-30	9	95-011-16	7
12-040-8	8	95-022-26	7
12-042-1	8	95-026-37	6
12-042-2	8	*	96