

1 **Persistent homology demarcates a leaf morphospace**

2
3 Mao Li¹, Hong An², Ruthie Angelovici², Clement Bagaza², Albert Batushansky², Lynn Clark³,
4 Viktoriya Coneva¹, Michael Donoghue⁴, Erika Edwards⁵, Diego Fajardo⁶, Hui Fang⁷, Margaret
5 Frank¹, Timothy Gallaher³, Sarah Gebken², Theresa Hill⁸, Shelley Jansky^{9,10}, Baljinder Kaur⁷,
6 Philip Klahs³, Laura Klein¹¹, Vasu Kuraparthi⁷, Jason Londo¹², Zoë Migicovsky¹³, Allison Miller¹¹,
7 Rebekah Mohn¹⁴, Sean Myles¹³, Wagner Otoni¹⁵, J. Chris Pires², Edmond Riffer², Sam
8 Schmerler^{5,16}, Elizabeth Spriggs⁴, Christopher Topp¹, Allen Van Deynze⁸, Kuang Zhang⁷, Linglong
9 Zhu⁷, Braden M. Zink², Daniel H. Chitwood^{17,*}

10

11 ¹Donald Danforth Plant Science Center, St. Louis MO USA

12 ²Division of Biological Sciences, University of Missouri-Columbia, Columbia, MO USA

13 ³Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA
14 USA

15 ⁴Department of Ecology and Evolutionary Biology, Yale University, New Haven CT USA

16 ⁵Department of Ecology and Evolutionary Biology, Brown University, Providence, RI USA

17 ⁶National Center for Genome Resources (NCGR), Santa Fe NM USA

18 ⁷Department of Crop and Soil Sciences, North Carolina State University, Raleigh, NC USA

19 ⁸Department of Plant Sciences, University of California-Davis, Davis CA USA

20 ⁹Vegetable Crops Research Unit, USDA-Agricultural Research Service, Madison WI USA

21 ¹⁰Department of Horticulture, University of Wisconsin-Madison, Madison WI USA

22 ¹¹Department of Biology, Saint Louis University, St. Louis MO USA

23 ¹²Grape Genetics Unit, USDA-Agricultural Research Service, Geneva NY USA

24 ¹³Department of Plant, Food, and Environmental Sciences, Dalhousie University, Truro, Nova
25 Scotia, Canada

26 ¹⁴Miami University, Oxford OH USA

27 ¹⁵Departamento de Biologia Vegetal, Universidade Federal de Viçosa, Viçosa, Minas Gerais,
28 Brasil

29 ¹⁶American Museum of Natural History, New York, NY USA

30 ¹⁷Independent Researcher, Santa Rosa CA USA

31

32

33 Short title: Persistent homology and leaf shape

34

35 *To whom correspondence should be addressed:

36

37 Daniel H. Chitwood

38 Independent Researcher

39 Santa Rosa, CA 95409

40 dhchitwood@gmail.com

41

42

43

44 **Abstract**

45

46 Current morphometric methods that comprehensively measure shape cannot compare the
47 disparate leaf shapes found in seed plants and are sensitive to processing artifacts. We explore
48 the use of persistent homology, a topological method applied across the scales of a function, to
49 overcome these limitations. The described method isolates subsets of shape features and
50 measures the spatial relationship of neighboring pixel densities in a shape. We apply the
51 method to the analysis of 182,707 leaves, both published and unpublished, representing 141
52 plant families collected from 75 sites throughout the world. By measuring leaves from
53 throughout the seed plants using persistent homology, a defined morphospace comparing all
54 leaves is demarcated. Clear differences in shape between major phylogenetic groups are
55 detected and estimates of leaf shape diversity within plant families are made. This approach
56 does not only predict plant family, but also the collection site, confirming phylogenetically
57 invariant morphological features that characterize leaves from specific locations. The
58 application of a persistent homology method to measure leaf shape allows for a unified
59 morphometric framework to measure plant form, including shape and branching architectures.

60

61 **Introduction**

62

63 As generally flattened structures, leaves provide a unique opportunity to quantify morphology
64 as a two-dimensional shape. Local features (such as serrations and lobes) and general shape
65 attributes (like length-to-width ratio) can be measured, but numerous methods also exist to
66 measure leaf shape more globally and comprehensively. A popular method to quantify leaf
67 shape is to place (x, y) coordinates, known as landmarks, on homologous features that are
68 related by descent from a common ancestor on every sample (Bookstein, 1997). The set of
69 landmarks from each leaf can be superimposed by translation, rotation, and scaling using a
70 Generalized Procrustes Analysis (Gower, 1975). Once superimposed, the Procrustes-adjusted
71 coordinates of each shape can be used directly for statistical analyses. Landmark analysis excels

72 in its interpretability, because each landmark is an identifiable feature with biological meaning
73 imparted by the shared homology between samples. Because landmarks are homologous
74 features, their use often reveals genetic and developmental patterns in shape variation
75 (Chitwood et al., 2016a).

76
77 Not all leaves have obvious homologous features that can be used as landmarks. Further, when
78 comparing leaves with disparate morphologies (e.g., simple vs. compound leaves), there may
79 not be identifiable homologous points. Nearly all leaves have homologous landmarks at the tip
80 and base, but if there are no other identifiable landmarks, an equal number of equidistant
81 points on each sample between the landmarks can be placed (Langlade et al., 2005). The denser
82 and more numerous such pseudo-landmarks are, the closer they come to approximating the
83 contour itself.

84
85 Another method, the use of Elliptical Fourier Descriptors (EFDs), measures shape as a
86 continuous closed contour, and can also be used when homologous features are absent. EFD
87 analysis begins with a lossless data compression method called chain-code, in which the
88 direction to move from one pixel to the next is recorded as a chain of numbers (where each link
89 in the chain a is an integer between 0 and 7 specifying the pixel direction $\left(\frac{\pi}{4}\right) a$) so that from
90 this chain of numbers the closed contour can be faithfully reproduced (Freeman, 1974). The
91 chain code is decomposed by a Fourier analysis into a harmonic series that is used to quantify
92 an approximate reconstruction of the shape (Kuhl and Giardina, 1982).

93
94 Both pseudo-landmarks and EFDs measure leaf shapes for which homologous features that can
95 be used as landmarks are lacking (Bensmihen et al., 2008; Chitwood and Otoni, 2017). Still,
96 when comparing disparate leaf shapes, unless major sources of shape variance in the data (such
97 as the number of lobes or leaflets) are present in every sample, individual pseudo-landmarks or
98 harmonic coefficients will not correspond between samples in a comparable way useful for
99 analysis. Recently, a computer vision method coupled with machine learning was used to

100 classify leaves, with diverse vascular patterns and leaf shapes, into plant families and orders
101 (Wilf et al., 2016). This method uses a visual descriptor to train a classifier. Since cleared leaves
102 are used, this method relies on both internal features like branch points in the vasculature as
103 well as features on the leaf margin, instead of just leaf shape alone as in traditional
104 morphometric approaches. Nonetheless, the method overcomes a central problem in the
105 morphometric analysis of leaves: comparing leaves with very different morphologies.

106

107 To develop a morphometric method that 1) comprehensively measures shape features in
108 leaves, both locally and globally, 2) can compare disparate leaves shapes, 3) is robust against
109 noise commonly found in leaf shape data (e.g., internal holes because of overlapping leaflets or
110 small defects introduced during imaging and thresholding), and 4) is potentially compatible
111 with other plant phenotyping needs (e.g., measuring the branching architectures of roots and
112 trees, the spatial distributions of plants in ecosystems, or the texture of different pollen types;
113 Mander et al., 2013; 2017; Li et al., 2017b) we used a persistent homology approach. Persistent
114 homology is a topological data analysis method. Topology is the field of mathematics concerned
115 with properties of space preserved under deformations (e.g., bending) but not tearing or re-
116 attaching. Persistent homology measures topological features across the scales of a function
117 (Edelsbrunner and Harer, 2008; Weinberger, 2011; Li et al., 2017b). The compatibility of
118 persistent homology with numerous functions makes it a versatile method that can be tailored
119 for diverse uses (Li et al., 2017a).

120

121 Here, we present a morphometric technique based on topology, using a persistent homology
122 framework, to measure the outlines of leaves and classify them by plant family and region in
123 which they were collected. We analyze 182,707 leaves (freely available to download; Chitwood,
124 2017a), from both published studies and shapes analyzed for the first time, from 141 plant
125 families and 75 sites throughout the world. We first compare the diverse shapes represented in
126 a common morphospace using persistent homology, which captures traditional shape
127 descriptors in a non-linear fashion. Major phylogenetic groups of plants occupy distinct regions

128 of the morphospace and we estimate plant families that have the most and least diverse leaf
129 shapes. Using persistent homology, we then use a linear discriminant analysis to classify leaves
130 by plant family and collection site. Persistent homology predicts both family and collection site
131 at a rate above chance, and predicts leaf family at 2.7 times and collection site at 1.5 times the
132 rate of traditional shape descriptors. Persistent homology is a topological method that can
133 measure and compare diverse leaf shapes from across seed plants and outperforms traditional
134 shape descriptors in classifying plant families and geographic locations.

135

136 **Results**

137

138 *Dataset and a morphospace defined using traditional shape descriptors*

139

140 To broadly analyze seed plant leaf shape diversity collected from sites throughout the world,
141 we used both published and unpublished data. In total, 182,707 leaves were analyzed (**Table 1**).
142 Many of these datasets address specific genetic and developmental questions, focusing on
143 genetic variability within a group or closely related species. Leaves were analyzed from the
144 following publications, pre-prints, and authors focusing on specific groups of plants:
145 *Alstroemeria* (2,392 leaves; Chitwood et al., 2012a), apple (9,619 leaves; Migicovsky et al.,
146 2017), *Arabidopsis* (5,101 leaves; AB, RA, CB, ER, BZ), *Brassica* (1,832 leaves; HA, SG, JCP),
147 *Capsicum* (3,277 leaves; TH, AVD), *Coleus* (34,607 leaves; VC, MF, ML), cotton (2,885 leaves;
148 Andres et al., 2017), grapevine and wild relatives (20,121 leaves; Chitwood et al., 2014; 2016a;
149 2016b; VC, MF, LK, JL, AM), *Hedera* (common ivy, 865 leaves; Martinez et al., 2016), *Passiflora*
150 (3,301 leaves; Chitwood and Otoni, 2017), Poaceae (866 leaves; LC, TG, PK), wild and cultivated
151 potato (1,840 leaves; DF, SJ), tomato and wild relatives (82,034 leaves; Chitwood et al., 2012b;
152 2012c; 2013), and *Viburnum* (2,422 leaves; Schmerler et al., 2012; MD, EE, SS, ES). We also
153 analyzed two datasets that sample broadly across seed plants and from sites throughout the
154 world. The Leafsnap dataset, with 5,733 leaves, represents mostly tree species of the
155 Northeastern United States, but other groups of plants as well (Kumar et al., 2012). The Climate

156 dataset, with 5,812 leaves total, analyzes the relationship between leaf shape and present
157 climates as indicators of paleoclimate (Huff et al., 2003; Royer et al., 2005; Peppe et al., 2011).

158

159 We analyzed all leaves using the traditional shape descriptors circularity, aspect ratio, and
160 solidity (**Figure 1**). These shape descriptors are simple in the sense that they each measure a
161 very specific aspect of shape, but they are powerful in that they can be applied to any shape,
162 which is not necessarily true of other methods that measure shape more comprehensively
163 (such as landmarks, pseudo-landmarks, and Elliptical Fourier Descriptors). Circularity is a ratio
164 of area to perimeter (true perimeter, excluding holes in the middle of the object) measured as
165 $4\pi * (\frac{area}{perimeter^2})$ and is sensitive to undulations (like serrations, lobes, and leaflets) along the
166 leaf perimeter, but is also influenced by elongated shapes (like grass leaves) when comparing
167 leaves with such different shapes, as in this analysis. Aspect ratio is measured as $(major\ axis)/$
168 $(minor\ axis)$ of a fitted ellipse, and it is a robust metric of overall length-to-width ratio of a
169 leaf. Solidity is measured as $\frac{area}{convex\ hull}$ where the convex hull bounds the leaf shape as a
170 polygon. Leaves with a large discrepancy between area and convex hull (such as compound
171 leaves with leaflets, leaves with deep lobes, or leaves with a distinct petiole) can be
172 distinguished from leaves lacking such features using solidity.

173

174 Differences between groups were visualized as scatterplots and density diagrams (**Figure 1**),
175 using transformed values of aspect ratio ($1/(aspect\ ratio)$) and solidity ($solidity^8$) to create
176 more even distributions that allow the separation between groups to be better visualized. The
177 long leaves of grasses (Poaceae, lavender) are perhaps the most distinct group of leaf shapes.
178 The Brassicaceae (light green) are bimodal in their distribution, reflecting entire vs. highly lobed
179 and compound leaves, as well as differences in petiole length. *Passiflora* (dark orange),
180 Solanaceae (purple), and *Viburnum* (brown) exhibit broad, continuous distributions, which like
181 the Brassicaceae reflect the diversity of leaf shapes in these groups. *Alstroemeria* (light blue),
182 apple (light orange), *Coleus* (pink), cotton (dark green), grapevine (red), and common ivy (dark
183 blue) all have more localized distributions in the morphospace, indicating that shape variation is

184 expressed within a smaller range, relative to other groups, as measured using traditional shape
185 descriptors.

186

187 *Persistent homology and non-linear relationships with traditional shape descriptors*

188

189 Although traditional shape descriptors can describe important shape features across diverse
190 leaves, they do not measure shape comprehensively like landmarks, pseudo-landmarks, and
191 Elliptical Fourier Descriptors. Comprehensive morphometric methods, however, cannot be
192 applied across diverse shapes, only between leaves with similar shapes, as in natural variation
193 studies. We crafted a persistent homology method to quantify the features of leaves,
194 conceptualizing shape as a two-dimensional point cloud of an outline defined by pixels (Li et al.,
195 2017a; Migicovsky et al., 2017). The method begins by calculating a Gaussian density estimator,
196 assigning each pixel a value that indicates the density of neighboring pixels (**Figure 2**). In leaves,
197 high density pixels with lots of neighbors tend to reside in the sinuses of serrations or lobes or
198 at points of intersection, such as the attachment points of leaflets to the rachis of a compound
199 leaf. Using a Gaussian density estimator, rather than focusing on continuity of a closed contour
200 (as in pseudo-landmarks and Elliptical Fourier Descriptors), minimizes the impact of internal or
201 non-continuous features, such as holes or occlusions made by the overlap of leaflets and lobes
202 (see the bottom palmately-shaped leaf in **Figure 2**). Sixteen annuli emanating from the centroid
203 of the shape (**Figure 2A**) serve to partition the leaf into subsets of features, increasing the
204 ability to distinguish between shapes. An annulus kernel for each ring (**Figure 2C**) is multiplied
205 by the density estimator (**Figure 2B**) to isolate density features that intersect with the annulus
206 (**Figure 2D-E**). The resulting density function from each annulus is the function across which
207 topological space is measured. As shown in **Figure 2F**, beginning with the highest density level,
208 the number of connected features with densities above that level is recorded. Counting the
209 number of connected components minus the number of holes (which is a topological feature,
210 known as the Euler characteristic) continues across the function, proceeding to lower density
211 levels. The value of the curve (y axis in **Figure 2F**) at each density level (x axis in **Figure 2F**)

212 records the topological structure across the values of the function, the crux of persistent
213 homology. A curve is recorded for each annulus, so that using our method, the shape of a single
214 leaf is represented by 16 curves.

215

216 To analyze the persistent homology output, we discretize each Euler characteristic curve into
217 500 values (**Figure 2F**) and then concatenate these values over the 16 annuli, representing each
218 leaf shape as 8,000 values. A Principal Component Analysis (PCA) performed using the 8,000
219 values creates a leaf morphospace defined by persistent homology (**Figure 3**). To interpret this
220 morphospace, we colored data using traditional shape descriptor values. Although clear
221 patterns among aspect ratio (**Figure 3A**), circularity (**Figure 3B**), and solidity (**Figure 3C**) with
222 persistent homology data are evident, the relationships are non-linear compared to the
223 orthogonal PC axes. Aspect ratio, circularity, and solidity are similarly correlated with PC1 (rho
224 values of -0.72, 0.70, and 0.61, respectively) demonstrating that persistent homology PCs can
225 capture distinct attributes of shape simultaneously (**Figure 3D**). The correlations between
226 traditional shape descriptors and persistent homology PCs rapidly diminish among high order
227 PCs (**Figure 3D**). The non-linear relationship between traditional shape descriptors and
228 persistent homology PCs indicates that persistent homology captures differing combinations of
229 traditional shape descriptors in different ways among the represented leaf shapes. Such non-
230 linear relationships are influenced by the different groups represented in the dataset (**Figure**
231 **3E**). If the Leafsnap and Climate datasets are superimposed as black points on top of a density
232 diagram representing different groups (**Figure 3F**), then the overall shape of the persistent
233 homology space defined by specific groups is recapitulated. As the Leafsnap and Climate
234 datasets together represent 141 plant families and 75 sites throughout the world, the data
235 suggest that the overall shape and density of the persistent homology morphospace is partially
236 saturated. This does not mean that there is no other significant leaf shape variation to be
237 explored, only that some archetypal leaf shapes are well represented in our dataset. The
238 boundaries of the persistent homology morphospace allow for speculation. Likely the
239 morphospace is 1) bimodal, defined by elongated leaf shapes found in some Poaceae and

240 Gymnosperms (specifically Pinophyta in the Leafsnap and Climate datasets) compared to other
241 leaf shapes and 2) is defined by variation spanning entire to deeply lobed (or even compound)
242 leaf shapes, as represented by *Passiflora*, Solanaceae, and *Vibrunum* across PC1. Of course,
243 other leaf shape variation exists (and is even visually apparent in the plots of PC2 vs. PC1) and
244 other PCs in this dataset remain to be explored. The dataset does not come near to sampling
245 all existing leaf shapes.

246

247 *Differences in leaf shape between phylogenetic groups and the most diverse plant families*

248

249 We were interested in detecting difference in leaf shape between phylogenetic groups and
250 performed a Principal Component Analysis (PCA) for just the Leafsnap and Climate datasets
251 (**Table 1**), which together represent 141 plant families, but without the over-representation
252 from specific taxonomic groups presented earlier. Visualizing gymnosperms, magnoliids, rosids
253 I, rosids II, asterids I, and asterids II across PCs 1-10 (representing 73% of shape variance) clear
254 differences in persistent homology shape space can be detected (**Figure 4**). Differences in shape
255 are most easily detected for the earliest diverging lineages. For example, gymnosperms occupy
256 a distinct region of morphospace defined by PCs 1-6 (**Figure 4A-C**) compared to angiosperms.
257 Subtler differences between recently diverging groups can also be seen. Asterids II, for
258 example, are excluded from some regions of morphospace occupied by rosids I/II and asterids I
259 for PCs 1-4 (**Figure 4A-B**).

260

261 Differences in occupied morphospace between phylogenetic groups prompted us to ask: are
262 plant families diverse across all PCs or just some, and what are the most and least
263 morphologically diverse families? To answer the first question, we calculated variance across
264 PCs 1-179 (representing >95% of all shape variance) for each plant family and then ranked
265 families from most to least variable for each PC (**Figure 5A**). Visualizing the ranked variability of
266 families across PCs (the most variable ranked families for a PC depicted as yellow, the least
267 variable black, **Figure 5A**), it is apparent that the most diverse tend be the most diverse across

268 PCs. Increased variability in persistent homology PCs, though, might simply be due to more
269 leaves in some families compared to others. Indeed, the most diverse plant families are also the
270 most represented in our dataset, as seen when families are arranged by abundance (**Figures 5A**,
271 see bar graph of counts on the right side). Because highly variable families tend to be variable
272 across PCs, we took the median rank of variance across PCs as a measure of overall family leaf
273 shape diversity. The relationship between –median rank variance and $\log_{10}(\text{count})$ is linear
274 (**Figure S1**). Using linear regression, we took the residuals from the model as an estimate of
275 plant family leaf shape diversity, corrected for differences in sample size (**Figure 5B**). A wilcoxon
276 signed rank test on residuals indicates that asterids I are marginally significant ($p = 0.08$) for
277 lacking diversity (two sided, $\mu = 0$) but other groups (gymnosperms, $p = 0.25$; magnoliids, $p =$
278 0.20 ; rosids I, $p = 0.97$; rosids II, $p = 0.63$; asterids II, $p = 0.63$) show no detectable biases in
279 diversity. The overall results indicate that, for the current dataset, leaf shape diversity within
280 major phylogenetic plant groups is equivalent, but specific families have higher estimated leaf
281 shape diversity than others.

282

283 *Persistent homology predicts plant family and region and outperforms traditional shape*
284 *descriptors*

285

286 The separation of different groups in the traditional shape descriptor (**Figure 1**) and persistent
287 homology (**Figures 3-4**) morphospaces suggests the ability to predict the phylogenetic identity
288 of a leaf based on its shape. Previous computer vision approaches coupled with machine
289 learning have successfully predicted plant family and order using vein patterning and margin
290 features (Wilf et al., 2016). Can the same be done using a persistent homology analysis of the
291 outline alone? Using the Leafsnap and Climate datasets (**Table 1**) that together represent 141
292 plant families, we used a Linear Discriminant Analysis (LDA) on PCs 1-179, representing >95% of
293 the persistent homology morphospace variation, to create a classifier scheme. Leaves were
294 then reassigned to the linear discriminant space using a cross-validated “leave one out”
295 approach (Venables and Ripley, 2002) and the results visualized as a confusion matrix (**Figure**

296 **6**), plotting the actual family identity of leaves as a function of the proportion of their predicted
297 family identity. Using persistent homology, there was a 27.3% correct plant family assignment
298 rate of leaves. Using a bootstrapping approach permuting plant family identity against leaf
299 shape information, a 27.3% correct reclassification rate or higher was never achieved in 1,000
300 bootstrapped simulations, indicating that assignment is above chance. This outperforms
301 traditional shape descriptor prediction (at a rate of 10.2%) by 2.7 times (**Table 2**), and including
302 both persistent homology and traditional shape descriptor data only marginally increases the
303 prediction rate (to 29.1%) over that of persistent homology alone (27.3%), indicating that
304 persistent homology largely captures the same shape features as traditional descriptors, but
305 provides additional information as well.

306
307 Previous studies analyzing correlations between leaf shape with present and ancient climates
308 debated the presence of “phylogenetic invariant” features that vary by climate, not
309 phylogenetic context. The Climate dataset includes leaves from 75 sites throughout the world
310 (**Table 1**). Like the phylogenetic prediction above, we sought to determine the degree that
311 geographic location (regardless of plant family) can be predicted from shape alone. An LDA
312 performed on PCs 1-191, representing >95% of the persistent homology morphospace variation
313 for the Climate dataset, can predict the site where a leaf was collected (**Figure 7**) at a rate of
314 14.5% (**Table 2**). Although much lower than the overall prediction rate by plant family (27.3%),
315 a rate of 14.5% or higher was never achieved in 1,000 bootstrapped simulations, indicating that
316 assignment is above chance. Persistent homology outperforms traditional shape descriptors (at
317 a rate of 9.5%) by 1.5 times (**Table 2**), and including both persistent homology and traditional
318 shape descriptor data only marginally increases the prediction rate (to 16.2%) over that of
319 persistent homology alone (14.5%).

320
321 Although the overall prediction rates of 27.3% for plant family and 14.5% for site collected are
322 relatively low (**Table 2**), it is important to remember that they are above the level of chance
323 (determined by bootstrapping, 1,000 simulations) and that the rates are not evenly distributed

324 across factor levels. Plant family prediction rates vary from 0-100%, and site collected
325 prediction rates vary from 0-40% (**Figure 8**). The variability in rates is not overly influenced by
326 sampling depth or variation within a group. For example, prediction rate of plant family and
327 abundance are correlated at $\rho = 0.37$, and the correlation with median rank PC variance is ρ
328 $= -0.24$. Although comprehensive, our dataset does not begin to encompass the total shape
329 variation present in a plant family or region and there are undoubtedly collection biases in the
330 data influencing prediction. Other factors than diversity within a group or the degree to which it
331 is sampled, though, likely influence prediction rate too.

332

333 **Discussion**

334

335 We have presented a new morphometric method using persistent homology, a topological
336 approach, that can comprehensively measure leaf shape. Other methods measure leaf shape
337 comprehensively, including traditional landmarks, pseudo-landmarks, and Elliptical Fourier
338 Descriptors (EFDs). However, no method comparatively analyzes the diverse *shapes* of leaves in
339 seed plants (simple leaves, deeply lobed leaves, compound leaves of different shapes, leaves
340 with differing numbers of leaflets or lobes, or large variation in petiole length and shape), only
341 naturally varying leaves among related plant species. Other morphometric methods that only
342 analyze the external contour of shapes are sensitive to artifacts, such as internal holes made by
343 the overlap of leaflets or lobes, or small errors during thresholding and isolation. Finally,
344 although appropriate for plant organs that can be represented by discrete shapes—like leaves,
345 petals, seeds, or other lateral organs—current morphometric techniques fail to capture other
346 attributes of plant architecture, like the branching patterns of roots, shoots, and inflorescences.
347 A framework that can not only measure shape, but other features that are important to the
348 plant form, is currently lacking.

349

350 By converting shapes into a topological space, as defined by a function that isolates subsets of
351 the shape and describes it in terms of neighboring pixel density (**Figure 2**), the described

352 persistent homology approach can compare disparate leaf shapes across seed plants, allowing
353 for the approximation of the overall leaf morphospace (**Figure 3**). By estimating pixel density,
354 the method accommodates internal features (such as holes caused by leaflet overlap) or small
355 processing artifacts, that do not unduly influence the output compared to the absence of such
356 imperfections. The ability to compare shapes broadly and be robust against processing artifacts
357 will enable large scale data analyses in the future, such as the analysis of digitized herbarium
358 vouchers, ecological studies, or genetic and developmental insights into complex morphologies,
359 for which current morphometric approaches are not designed. We detected clear differences in
360 leaf shape between major phylogenetic groups (**Figure 4**) and estimated leaf shape diversity
361 across plant families (**Figure 5**), demonstrating that a persistent homology approach is relevant
362 for large-scale morphometric studies across plant evolution. The ability to comprehensively
363 measure shapes permits alternative statistical approaches, moving beyond descriptive statistics
364 used with traditional shape descriptors (**Figure 1**) and allowing for classifier and prediction
365 approaches (**Figures 6-8; Table 2**). Theoretically, a unifying morphometric framework that can
366 accommodate not only shapes but the branching architectures of plants, is lacking. As we have
367 previously described, persistent homology functions are ideal to apply to branching plant
368 structures as topological spaces (Li et al., 2017b). The morphometric approach described here
369 applied to leaf shapes is compatible with similar persistent homology methods, creating a
370 shared framework in which the plant form can be measured (Li et al., 2017a).

371

372 **Materials and Methods**

373

374 *Leaf shapes*

375

376 The 182,707 leaf outlines from 141 plant families from 75 sites throughout the world used in
377 this manuscript are available to download (Chitwood, 2017a). This file directory includes x,y
378 coordinates that form the outlines of the leaves. Separate folders contain text files with x,y
379 coordinates for the leaves from each of the indicated groups in **Table 1**. Within each folder,

380 original x,y coordinates and scaled coordinates are provided. This dataset contains leaves from
381 both published and unpublished sources (see text for details; Andres et al., 2017; Chitwood et
382 al., 2012a; 2012b; 2012c; 2013; 2014; 2016a; 2016b; Chitwood and Otoni, 2017; Huff et al.,
383 2003; Kumar et al., 2012; Li et al., 2017a; Martinez et al., 2016; Migicovsky et al., 2017; Peppe
384 et al., 2011; Royer et al., 2005; Schmerler et al., 2012; *Arabidopsis* BA, RA, CB, ER, BZ; *Brassica*
385 HA, SG, JCP; *Capsicum* TH, AVD; *Coleus* VC, MF, ML; grapevine and wild relatives VC, MF, LK, JL,
386 AM; Poaceae LC, TG, PK; wild and cultivated potato DF, SJ; *Viburnum* MD, EE, SS, ES).

387

388 *Persistent homology*

389

390 The MATLAB code necessary to recapitulate the persistent homology analysis in this manuscript
391 can be found in the following GitHub repository (Li, 2017):

392 <https://github.com/maoli0923/Persistent-Homology-All-Leaf>

393

394 Persistent homology is a flexible method to quantify branching structures (Edelsbrunner and
395 Harer, 2008; Weinberger, 2011; Li et al., 2017b), point clouds (Ghrist, 2008), two-dimensional
396 and three-dimensional shapes (Gamble and Heo, 2010), and textures (Mander et al., 2013;
397 2017). Each of these different phenotypes can be described by a multidimensional vector (e.g.
398 Euler characteristic curve), integrating how homology (e.g. path-connected components)
399 persists across the scales of a tailored mathematical function.

400

401 Leaf contours are represented as two-dimensional point clouds extracted from binary images
402 (**Figure 2A**). We use a Gaussian density estimator, which can be directly derived from the point
403 cloud and is also robust to noise, to estimate the neighborhood density of each pixel. Denser
404 point regions, such as serrations, lobes, or the attachment points of leaflets, have higher
405 function values (**Figure 2B**). Formally, the Gaussian density estimator is defined as

406
$$\varphi(x) := \frac{1}{n} \sum_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-y_i}{h}\right)^2},$$
 where y_1, \dots, y_n are the data points and h is a bandwidth

407 parameter. Because a set of local and regional topologies may often be more effective to
408 represent shapes, we use a local persistent homology technique to subset the density estimator
409 into 16 concentric annuli centered around the centroid of the leaf (**Figures 2A, D**). To achieve
410 this, we multiply this function by a “bump” function K which highlights an annulus, defined as
411 $K_{\sigma,t,y}(x) := e^{-\frac{(d(x,y)-t\sigma)^2}{2\sigma^2}}$, where y is the center of the annulus, $t\sigma$ determines its radius, and the
412 parameter σ is its width (**Figure 2C**). Each local function emphasizes the density function falling
413 in the annulus. Given a threshold and a local function, the points whose function values are
414 greater than this threshold form a subset (superlevel set). Changing this threshold value from
415 the maximum function value to its minimum value, we can get an expanding sequence of
416 subsets, or a superlevel set filtration. **Figure 2E** shows the shapes above a plane, an example of
417 a superlevel set filtration. For each subset, we calculate the Euler characteristic, which equals
418 the number of connected components minus the number of holes. Thus, for a sequence of
419 subsets, we get a sequence of numbers (a multidimensional vector). All 16 annuli derive 16
420 multidimensional vectors which are concatenated into an overall vector used for analysis.
421 Principal Component Analysis (PCA) was performed in MATLAB on the vectors and PC scores
422 and percent variance explained by each PC used in subsequent analyses.

423

424 *Statistical analysis and visualization*

425

426 The R code (R Core Team, 2017) and data necessary to recapitulate the statistical analyses and
427 figures in this manuscript can be found as a zipped folder directory on figshare (Chitwood,
428 2017b): <https://figshare.com/articles/LeafMorphospace/4985561/1>

429

430 Unless otherwise specified, all graphs were visualized using ggplot2 (Wickham, 2016).

431 Scatterplots were visualized using the `geom_point()` function, density plots were visualized with
432 the `geom_density2d()` function, heatmaps were visualized using the `geom_tile()` function, and
433 colors were selected from ColorBrewer (Harrower and Brewer, 2003) and viridis (Garnier,

434 2017). Other visualization functions used and specific parameters that can be found in the code
435 used to generate the figures (Chitwood, 2017b).

436

437 Variance was calculated for each plant family for each principal component using `var()` and
438 families ranked for each principal component using `rank()` (**Figure 5**). Linear regression was
439 performed using `lm()` and residuals retrieved to estimate leaf shape diversity for each plant
440 family (**Figure S1**). The Wilcoxon signed rank test was performed using `wilcox.test()` to test for
441 higher or lower than expected phylogenetic diversity using a two-sided test with $\mu = 0$. Linear
442 Discriminant Analysis (LDA) was performed using the `lda()` function in the package MASS
443 (Venables and Ripley, 2002). LDA was performed using the number of principal components
444 that contributed at least 95% of all variance in each analysis (PCs 1-179 for phylogenetic
445 prediction and PCs 1-191 for site prediction). The Leafsnap and Climate datasets were used for
446 phylogenetic prediction (**Figure 6**) whereas just the Climate dataset was used for site prediction
447 (**Figure 7**). Prediction using the discriminant space was performed using `CV = TRUE` for a “leave
448 one out” cross-validated jack-knifed approach and the priors set equal across factor levels. Both
449 the phylogenetic and site LDA prediction rates were bootstrapped over 1,000 simulations. A for
450 loop was used, permuting family or site identity against leaf identity, performing an LDA on the
451 permuted data, and recording the correct prediction rate for each permuted simulation. For
452 both the phylogenetic and site predictions, a permuted correct prediction rate (out of 1,000
453 simulations) higher than the actual correct prediction rate was never detected.

454

455 **References**

456

457 Andres RJ, Coneva V, Frank MH, Tuttle JR, Samayoa LF, Han SW, Kaur B, Zhu LL, Fang H,
458 Bowman DT, Rojas-Pierce M. Modifications to a *LATE MERISTEM IDENTITY1* gene are
459 responsible for the major leaf shapes of Upland cotton (*Gossypium hirsutum* L.). Proceedings of
460 the National Academy of Sciences of the United States of America. 2017 Jan 3;114(1):E57-66.

461

462 Bensmihen S, Hanna AI, Langlade NB, Micol JL, Bangham A, Coen ES. Mutational spaces for leaf
463 shape and size. HFSP Journal. 2008 Apr 1;2(2):110-20.

464

465 Bookstein FL. Morphometric tools for landmark data: geometry and biology. Cambridge

466 University Press; 1997 Jun 28.
467
468 Chitwood DH, Naylor DT, Thammapichai P, Weeger AC, Headland LR, Sinha NR. Conflict
469 between intrinsic leaf asymmetry and phyllotaxis in the resupinate leaves of *Alstroemeria*
470 *psittacina*. *Frontiers in Plant Science*. 2012a Aug 10;3:182.
471
472 Chitwood DH, Headland LR, Filiault DL, Kumar R, Jiménez-Gómez JM, Schragar AV, Park DS,
473 Peng J, Sinha NR, Maloof JN. Native environment modulates leaf size and response to simulated
474 foliar shade across wild tomato species. *PLoS One*. 2012b Jan 12;7(1):e29570.
475
476 Chitwood DH, Headland LR, Kumar R, Peng J, Maloof JN, Sinha NR. The developmental
477 trajectory of leaflet morphology in wild tomato species. *Plant Physiology*. 2012c Mar
478 1;158(3):1230-40.
479
480 Chitwood DH, Kumar R, Headland LR, Ranjan A, Covington MF, Ichihashi Y, Fulop D, Jiménez-
481 Gómez JM, Peng J, Maloof JN, Sinha NR. A quantitative genetic basis for leaf morphology in a
482 set of precisely defined tomato introgression lines. *The Plant Cell*. 2013 Jul 1;25(7):2465-81.
483
484 Chitwood DH, Ranjan A, Martinez CC, Headland LR, Thiem T, Kumar R, Covington MF, Hatcher T,
485 Naylor DT, Zimmerman S, Downs N. A modern ampelography: a genetic basis for leaf shape and
486 venation patterning in grape. *Plant Physiology*. 2014;164(1):259-72.
487
488 Chitwood DH, Klein LL, O'Hanlon R, Chacko S, Greg M, Kitchen C, Miller AJ, Londo JP. Latent
489 developmental and evolutionary shapes embedded within the grapevine leaf. *New Phytologist*.
490 2016a Apr 1;210(1):343-55.
491
492 Chitwood DH, Rundell SM, Li DY, Woodford QL, Tommy TY, Lopez JR, Greenblatt D, Kang J,
493 Londo JP. Climate and developmental plasticity: interannual variability in grapevine leaf
494 morphology. *Plant Physiology*. 2016b Mar 1;170(3):1480-91.
495
496 Chitwood DH, Otoni WC. Morphometric analysis of *Passiflora* leaves: the relationship between
497 landmarks of the vasculature and elliptical Fourier descriptors of the blade. *Gigascience*. 2017
498 Jan 1;6(1):1-13.
499
500 Chitwood DH. Leaf_coordinates. Figshare. 2017a. Accessed June 17, 2017.
501 <https://doi.org/10.6084/m9.figshare.5056441.v1>
502
503 Chitwood DH. LeafMorphospace. Figshare. 2017b. Accessed May 29, 2017.
504 <https://doi.org/10.6084/m9.figshare.4985561.v1>
505
506 Edelsbrunner H, Harer J. Persistent homology-a survey. *Contemporary mathematics*. 2008 Feb
507 29;453:257-82.

508
509 Freeman H. Computer processing of line-drawing images. *ACM Computing Surveys (CSUR)*.
510 1974 Mar 1;6(1):57-97.
511
512 Gamble J, Heo G. Exploring uses of persistent homology for statistical analysis of landmark-
513 based shape data. *Journal of Multivariate Analysis*. 2010 Oct 31;101(9):2184-99.
514
515 Garnier S. viridis: Default Color Maps from 'matplotlib'. R package version 0.4.0. 2017
516 <https://CRAN.R-project.org/package=viridis>
517
518 Ghrist R. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical*
519 *Society*. 2008;45(1):61-75.
520
521 Gower JC. Generalized procrustes analysis. *Psychometrika*. 1975 Mar 27;40(1):33-51.
522
523 Harrower M, Brewer CA. ColorBrewer. org: an online tool for selecting colour schemes for
524 maps. *The Cartographic Journal*. 2003 Jun 1;40(1):27-37.
525
526 Huff PM, Wilf P, Azumah EJ. Digital future for paleoclimate estimation from fossil leaves?
527 Preliminary results. *Palaios*. 2003 Jun;18(3):266-74.
528
529 Kuhl FP, Giardina CR. Elliptic Fourier features of a closed contour. *Computer Graphics and*
530 *Image Processing*. 1982 Mar 1;18(3):236-58.
531
532 Kumar N, Belhumeur P, Biswas A, Jacobs D, Kress WJ, Lopez I, Soares J. Leafsnap: A computer
533 vision system for automatic plant species identification. *Computer Vision–ECCV 2012*.
534 2012:502-16.
535
536 Langlade NB, Feng X, Dransfield T, Copsey L, Hanna AI, Thébaud C, Bangham A, Hudson A, Coen
537 E. Evolution through genetically controlled allometry space. *Proceedings of the National*
538 *Academy of Sciences of the United States of America*. 2005 Jul 19;102(29):10221-6.
539
540 Li M, Frank MH, Coneva V, Mio W, Topp CN, Chitwood DH. Persistent homology: a tool to
541 universally measure plant morphologies across organs and scales. *bioRxiv*. 2017a
542 <https://doi.org/10.1101/104141>
543
544 Li M, Duncan K, Topp CN, Chitwood DH. Persistent homology and the branching topologies of
545 plants. *American Journal of Botany*. 2017b 104(3):349-353.
546
547 Li M. Persistent-Homology-All-Leaf. GitHub. 2017. Accessed May 29, 2017.
548 <https://github.com/maoli0923/Persistent-Homology-All-Leaf>
549

550 Mander L, Li M, Mio W, Fowlkes CC, Punyasena SW. Classification of grass pollen through the
551 quantitative analysis of surface ornamentation and texture. *Proceedings of the Royal Society of*
552 *London B: Biological Sciences*. 2013 Nov 7;280(1770):20131905.
553
554 Mander L, Dekker SC, Li M, Mio W, Punyasena SW, Lenton TM. A morphometric analysis of
555 vegetation patterns in dryland ecosystems. *Royal Society Open Science*. 2017 Feb
556 1;4(2):160443.
557
558 Martinez CC, Chitwood DH, Smith RS, Sinha NR. Left–right leaf asymmetry in decussate and
559 distichous phyllotactic systems. *Philosophical Transactions of the Royal Society* 2016 Dec
560 19;371(1710):20150412.
561
562 Migicovsky Z, Li M, Chitwood DH, Myles S. Morphometrics reveals complex and heritable apple
563 leaf shapes. *bioRxiv*. 2017 <https://doi.org/10.1101/139303>
564
565 Peppe DJ, Royer DL, Cariglino B, Oliver SY, Newman S, Leight E, Enikolopov G, Fernandez-Burgos
566 M, Herrera F, Adams JM, Correa E. Sensitivity of leaf size and shape to climate: global patterns
567 and paleoclimatic applications. *New Phytologist*. 2011 May 1;190(3):724-39.
568
569 R Core Team. R: A language and environment for statistical computing. R Foundation for
570 Statistical Computing. 2017. Vienna, Austria. Accessed May 29, 2017. [https://www.R-](https://www.R-project.org/)
571 [project.org/](https://www.R-project.org/)
572
573 Royer DL, Wilf P, Janesko DA, Kowalski EA, Dilcher DL. Correlations of climate and plant ecology
574 to leaf size and shape: potential proxies for the fossil record. *American Journal of Botany*. 2005
575 Jul 1;92(7):1141-51.
576
577 Schmerler SB, Clement WL, Beaulieu JM, Chatelet DS, Sack L, Donoghue MJ, Edwards EJ.
578 Evolution of leaf form correlates with tropical–temperate transitions in *Viburnum* (Adoxaceae).
579 *Proceedings of the Royal Society of London B: Biological Sciences*. 2012 Oct 7;279(1744):3905-
580 13.
581
582 Venables WN, Ripley BD. *Modern Applied Statistics with S*. Springer; 2017 New York
583
584 Weinberger S. What is... persistent homology? *Notices of the AMS*. 2011 Jan;58(1):36-9.
585
586 Wickham H. *ggplot2: elegant graphics for data analysis*. Springer; 2016 Jun 8 New York
587
588 Wilf P, Zhang S, Chikkerur S, Little SA, Wing SL, Serre T. Computer vision cracks the leaf code.
589 *Proceedings of the National Academy of Sciences of the United States of America*. 2016 Mar
590 7:201524473.
591

592

593 **Figure Legends:**

594

595 **Figure 1: Traditional shape descriptors delimit leaves from different taxonomic groups. A)**
596 **Circularity vs. 1/Aspect Ratio, B) Solidity⁸ vs. 1/Aspect Ratio, and C) Circularity vs. Solidity⁸.** Left:
597 Scatter plots of 182,707 leaves analyzed, from 141 plant families from 75 sites throughout the
598 world. Right: For select taxonomic groups, density plots showing ability of traditional shape
599 descriptors to delimit different leaf shapes and distributions of different groups. Solidity and
600 Aspect Ratio values have been transformed to yield more even distributions. Taxonomic groups
601 are indicated by color and silhouettes of representative leaves close to the overall mean of
602 descriptor values provided.

603

604 **Figure 2: Persistent homology and leaf shape. A)** Contours of a simple leaf (top), compound
605 pinnate leaf (middle), and compound palmate leaf with a hole and overlap in leaflets (bottom).
606 16 annuli used to isolate pixel density are shown, with annulus 10 used in subsequent panels
607 indicated in green. **B)** Colormap of a Gaussian density estimator that is robust to noise. Red
608 indicates a larger density of neighboring pixels and blue less density. **C)** An annulus kernel is
609 used to localize and smoothen data. **D)** Multiplication of the annulus kernel with the density
610 estimator isolates density features of the leaf contour. **E)** Side view of the annulus kernel-
611 isolated density features of the leaf. The high peaks in red indicate higher pixel density. **F)** A
612 plane traverses the density function from the highest to lowest densities (x axis). As the plane
613 traverses the function, the topological space is recorded as the number of connected
614 components above the plane at any given point, the Euler characteristic (y axis). Three pink
615 dotted lines correspond to the plane at three points along the density function, which are
616 visualized below the graphs. Together, similar curves from the 16 annuli comprise the
617 persistent homology description of leaf shape.

618

619 **Figure 3: Principal Component Analysis (PCA) of persistent homology results.** Principal

620 Component 2 (PC2) vs. PC1 based on persistent homology results for 182,707 leaves colored by
621 **A) 1/Aspect Ratio, B) Circularity, and C) Solidity**⁸. Aspect Ratio and Solidity values have been
622 transformed to yield more even distributions. Note non-linear relationships between traditional
623 shape descriptors and persistent homology PCs. **D) Correlations between aspect ratio,**
624 **circularity, and solidity and PCs 1-69 (representing 90% of variation).** Positive and negative
625 Spearman's rho values are indicated as blue and yellow, respectively. **E) Density plots show**
626 **distributions of selected taxonomic groups in persistent homology PCA and F) Climate and**
627 **Leafsnap datasets, representing 141 plant families from 75 sites throughout the world, are**
628 **superimposed as black dots. Taxonomic groups are indicated by color and silhouettes of**
629 **representative leaves close to the overall mean of descriptor values provided.**

630

631 **Figure 4: Differences in leaf shape between phylogenetic groups.** Gymnosperm, magnoliid,
632 rosid I, rosid II, asterid I, and asterid II leaves (left to right) are each plotted in blue against all
633 samples (gray) for **A) PC2 vs. PC1, B) PC4 vs. PC3, C) PC6 vs. PC5, D) PC8 vs. PC9, and E) PC10 vs.**
634 **PC9.** Percent variance explained by each PC is indicated.

635

636 **Figure 5: Highly variable plant families are variable across Principal Components (PCs) and**
637 **estimates of leaf shape diversity by family. A) Variance was measured for each plant family**
638 **and then ranked from most variable (yellow) to least variable (black) for each PC. Plant families**
639 **are ordered by abundance, as seen in the bar graph (right) indicating count number in the**
640 **dataset. The most abundant plant families in the dataset tend to be the most variable. B) Linear**
641 **regression was used to model the -median variance ranking for each plant family as a function**
642 **of $\log_{10}(\text{count})$. The residuals are estimates of plant family leaf shape diversity, as corrected for**
643 **representation in the dataset. Higher residual values indicate higher estimated leaf shape**
644 **diversity. Gymnosperms, orange; magnoliids, yellow; rosids I, light blue; rosids II, dark blue;**
645 **asterids I, light green; asterids II, dark green; other groups, gray.**

646

647 **Figure 6: Predicting plant family using persistent homology.** Using persistent homology data

648 from the Climate and Leafsnap datasets, a Linear Discriminant Analysis (LDA) was used as a
649 classifier to predict plant family, cross-validated using a jackknifed “leave one out” approach.
650 The vertical axis indicates actual plant family and the horizontal axis predicted plant family.
651 Color indicates proportion of leaves from each actual plant family assigned to each predicted
652 family, such that proportions across the horizontal axis sum to 1. Black indicates no assignment.
653 A phylogeny indicating key taxonomic groups is provided.

654

655 **Figure 7: Predicting collection site using persistent homology.** Using persistent homology data
656 from the Climate dataset, a Linear Discriminant Analysis (LDA) was used as a classifier to predict
657 collection site, cross-validated using a jackknifed “leave one out” approach. The vertical axis
658 indicates actual collection site and the horizontal axis predicted collection site. Color indicates
659 proportion of leaves from each actual collection site assigned to each predicted collection site,
660 such that proportions across the horizontal axis sum to 1. Black indicates no assignment. Sites
661 are grouped into nine different regions that are indicated by color on a map.

662

663 **Figure 8: Prediction rates using persistent homology data across plant families and collection**
664 **sites. A)** Proportion of leaves from each family correctly assigned. Red line indicates overall
665 correct prediction rate of plant family of 27.3%. Phylogeny and major taxonomic groups are
666 indicated. **B)** Proportion of leaves from each collection site correctly assigned. Red line indicates
667 overall correct prediction rate of collection site of 14.5%. Collection sites are grouped by region,
668 indicated by color.

669

670 **Supplemental figure legend:**

671

672 **Supplemental Figure 1: Linear relationship between median ranked variability and count.**

673 Linear regression was used to model -median rank variability (higher values indicated more
674 variability within a plant family) as of function of the abundance of each plant family in the
675 dataset, as measured by \log_{10} (leaf count). The model (shown in blue) was used to estimate

676 overall leaf shape variance in plant family, as corrected for sampling depth, by using the
 677 residuals from the model as an indication of diversity.

678

679 **Table 1:** Leaf counts of datasets.

Leaf type	Count
<i>Alstroemeria</i>	2,392
Apple	9,619
<i>Arabidopsis</i>	5,101
<i>Brassica</i>	1,832
<i>Capsicum</i>	3,277
Climate	5,812
<i>Coleus</i>	34,607
Cotton	2,885
Grapevine	20,121
<i>Hedera</i>	865
Leafsnap	5,733
<i>Passiflora</i>	3,301
Poaceae	866
Potato	1,840
Tomato	82,034
<i>Viburnum</i>	2,422
Total	182,707

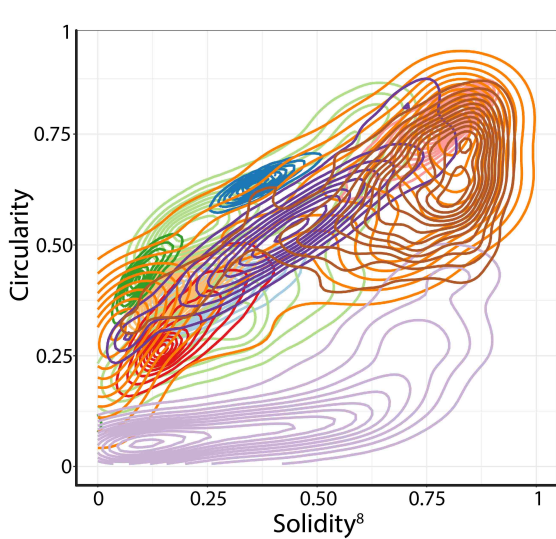
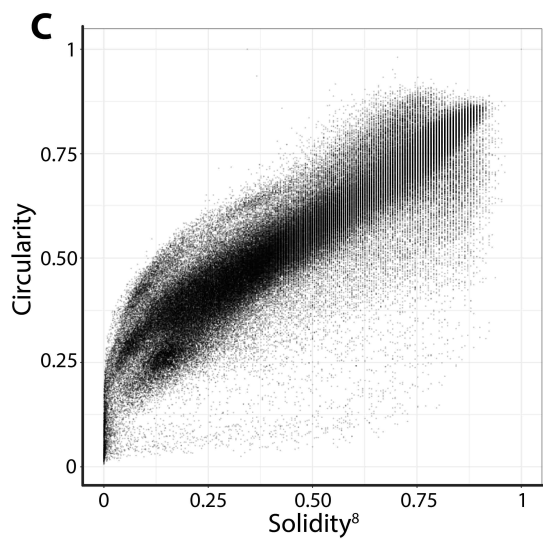
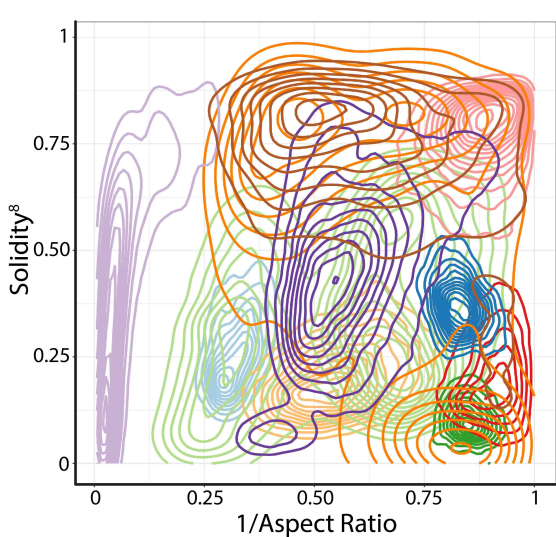
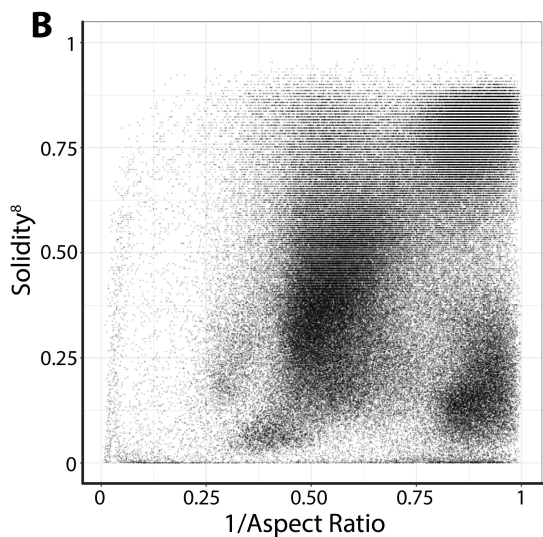
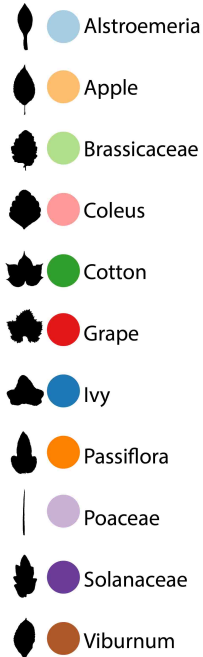
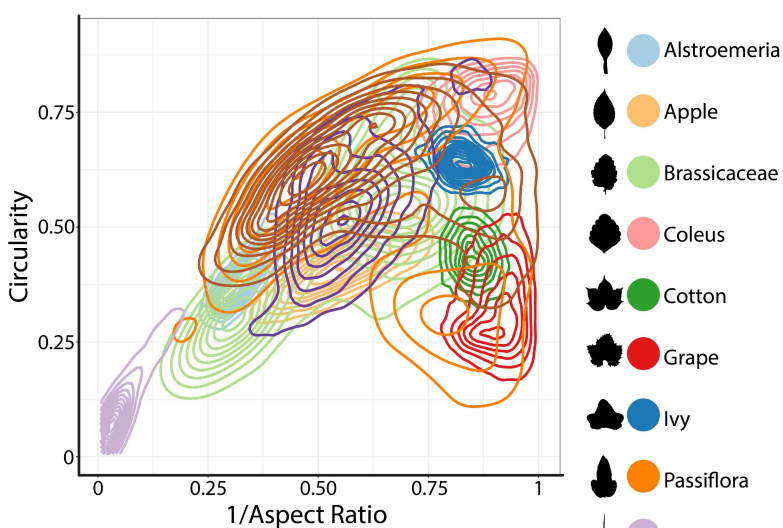
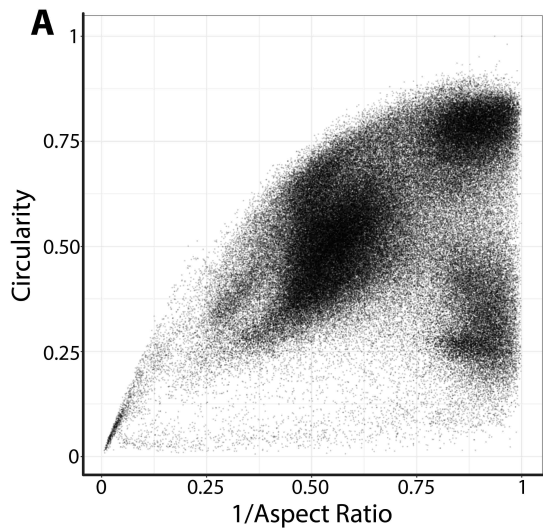
701

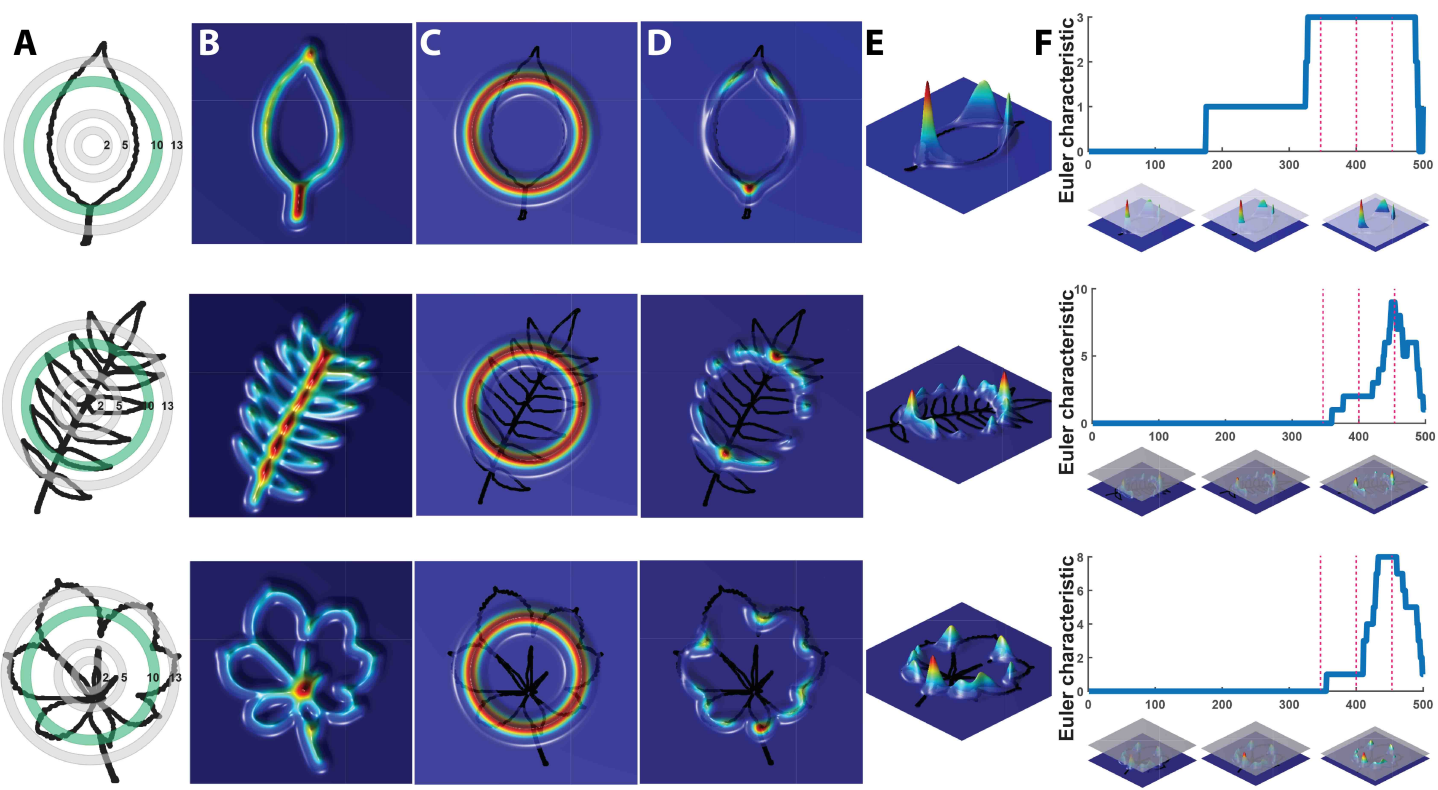
702 **Table 2:** Overall prediction rates of plant family and collection site using different
 703 morphometric methods

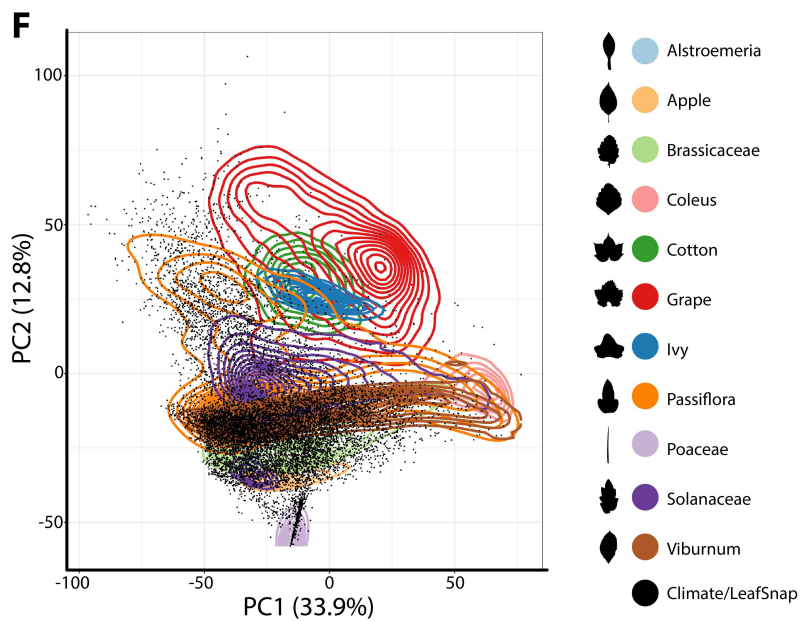
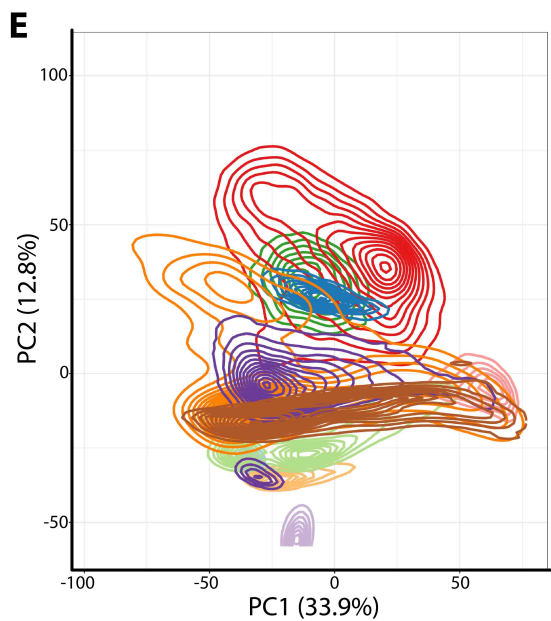
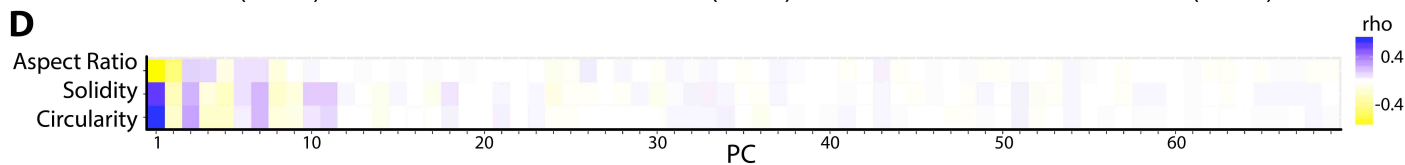
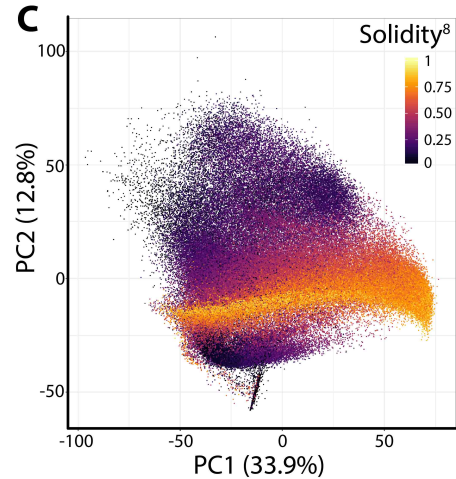
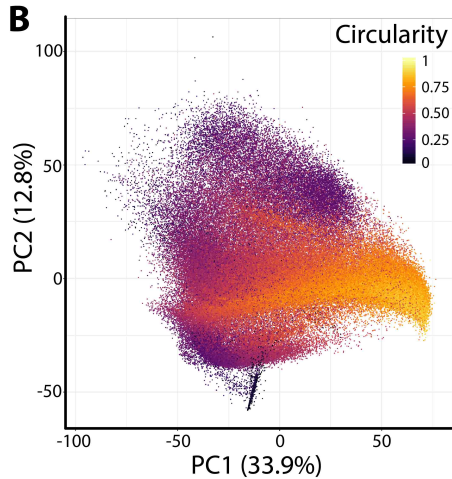
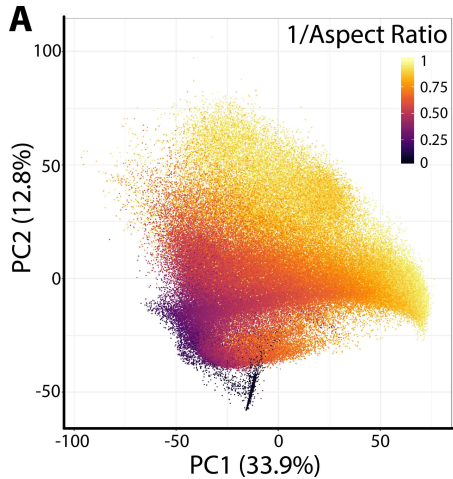
704

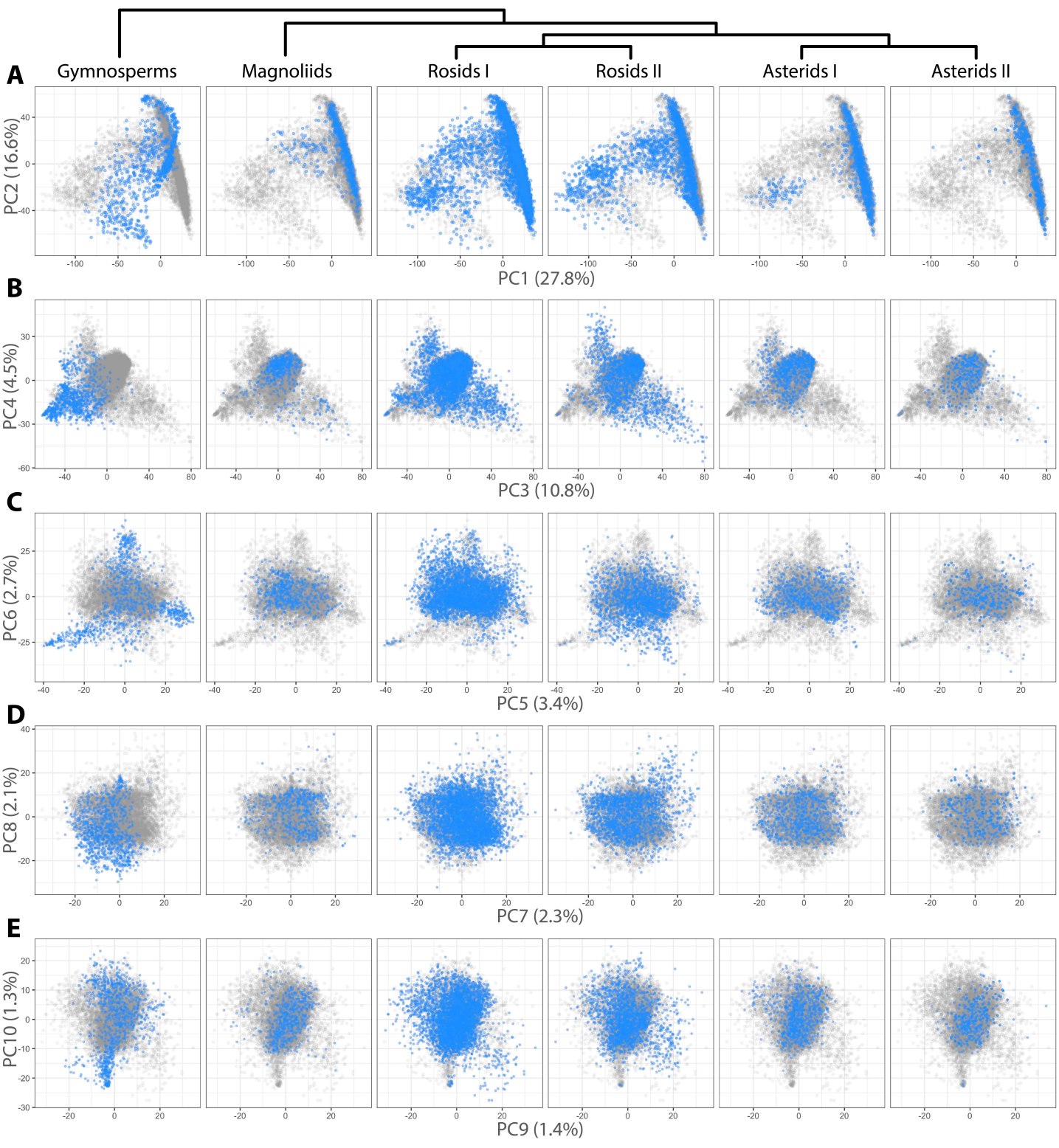
Prediction	Datasets	Method	Correct
Plant family	Climate, Leafsnap	Persistent homology	27.3%
Plant family	Climate, Leafsnap	Traditional descriptors	10.2%
Plant family	Climate, Leafsnap	Both methods	29.1%
Site	Climate	Persistent homology	14.5%
Site	Climate	Traditional descriptors	9.5%
Site	Climate	Both methods	16.2%

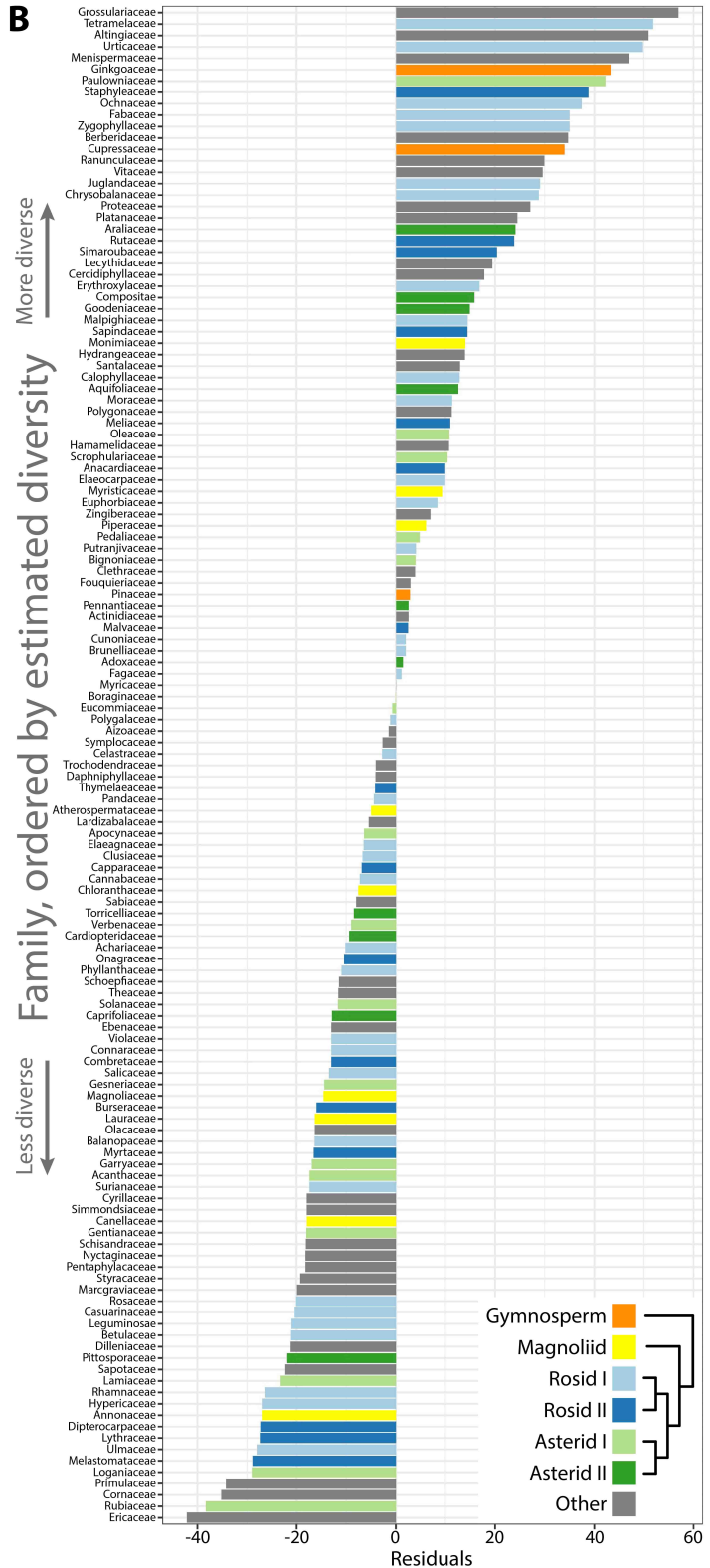
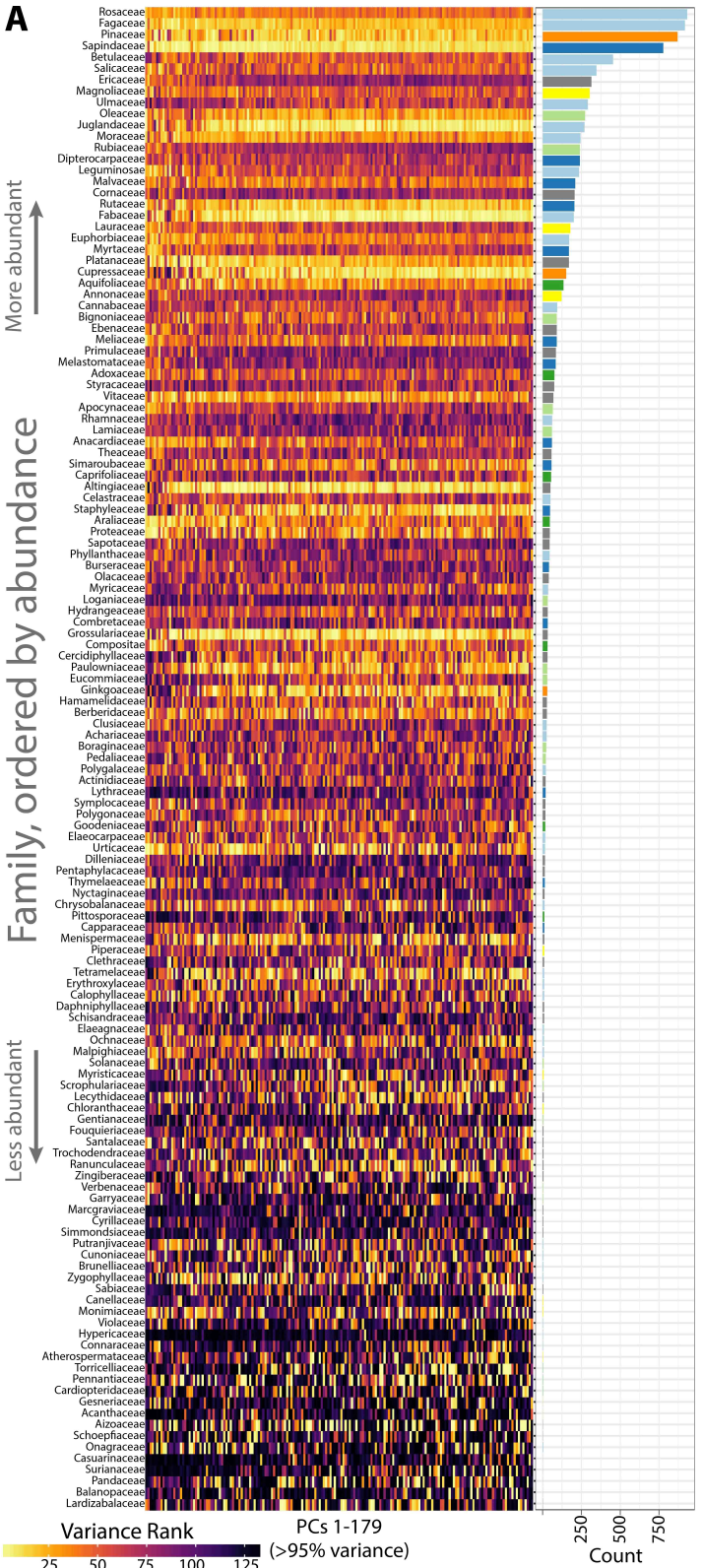
705



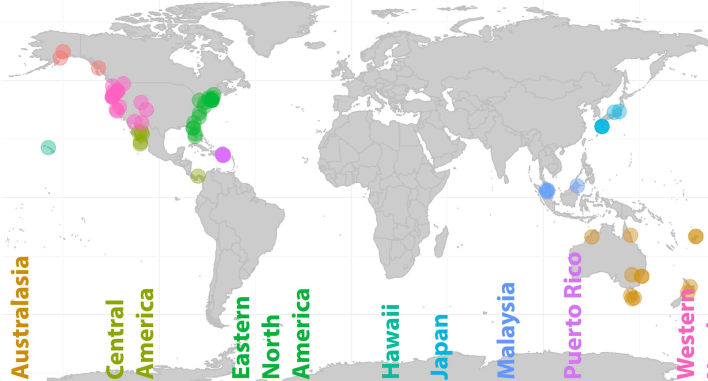
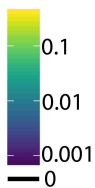








Proportion assigned



Alaska
Australasia
Central America
Eastern North America
Hawaii
Japan
Malaysia
Puerto Rico
Western North America

Actual site

Western North America

Puerto Rico

Malaysia

Japan

Hawaii

Eastern North America

Central America

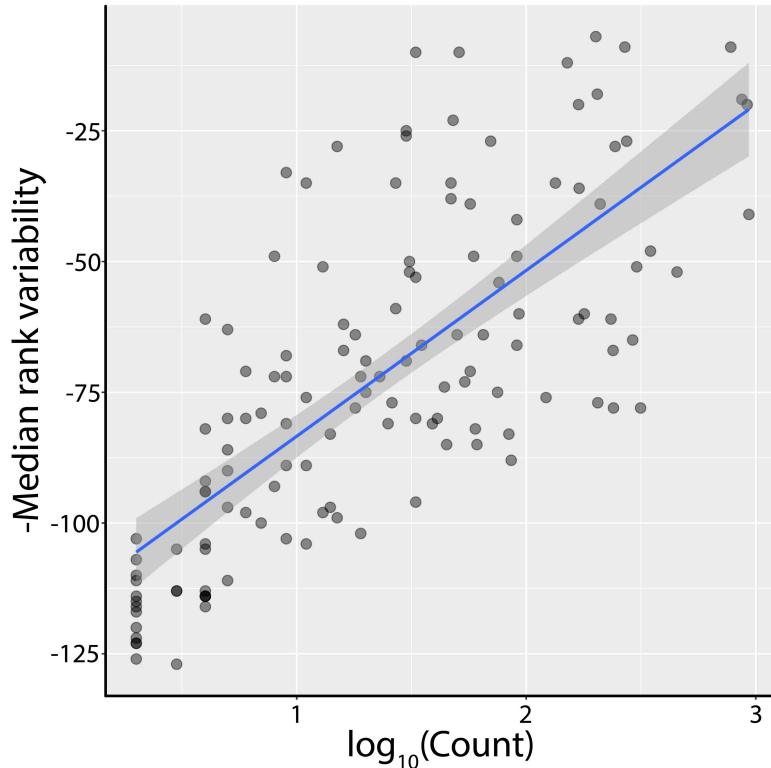
Australasia

Alaska

wolf_creek_colorado
three_lmn_clickamas_river_oregon
santa_cruz_new_brighton_state_beach_california
republic_north_fork_of_sanpoil_river_washington
red_feet_reservoir_utah
powers_coquille_river_oregon
portal_arizona
placerville_california
north_bend_oregon
nestucca_river_oregon
hood_river_oregon
half_moon_bay_california
government_camp_multitorpor_mountain_oregon
clearwater_washington
childs_arizona
cape_blanco_oregon
bandon_oregon
toro_negro_state_forest
monte_guilarte_in_guilarte_commonwealth_forest
guanica_state_forest
guajataca
cabo_rojo_in_borinquen_state_forest
buena_vista_in_maricao_state_forest
borinquen_in_guanica_state_forest
pasoh_forest_reserve
frim_kepong
danum_valley_field_center_borneo
yakusugi_800m_yakushima
yakusugi_260m_yakushima
yakusugi_1350m_yakushima
yakusugi_1080m_yakushima
suganuma_honshu
nagakubo_yakushima
chuzenji_ko_honshu
williwil_nui_ridge_trail_ohau
smithsonian_environmental_research_center_maryland
little_pee_dee_state_park_south_carolina
institute_for_ecosystem_studies_new_york
hubbard_brook_experimental_forest_new_hampshire
hawk_mountain_sanctuary_pennsylvania
harvard_forest_machusetts
florida_panther_national_wildlife_refuge_florida
e_n_huyck_preserve_and_biological_research_station_new_york
duke_forest_north_carolina
dilchers_woods_upland_gainesville_florida
dilchers_woods_lowland_gainesville_florida
connecticut_river_upland_near_middletown_connecticut
connecticut_river_swamp_near_middletown_connecticut
connecticut_river_near_middletown_connecticut
cockaponset_state_forest_connecticut
big_hammock_natural_and_wildlife_management_area_georgia
archbold_biological_station_florida
allegheeny_national_forest_pennsylvania
san_bartolo_baja_california_sur_mexico
nari_sonora_mexico
empalme_sonora_mexico
cabo_san_lucas_baja_california_sur_mexico
barro_colorado_island_panama
siamos_sonora_mexico
seqapa_cakaudrove_province_fiji
round_hill_nature_reserve_mallee_south_wales_australia
okataina_new_zealand
noah_creek_queensland_australia
mt_read_tasmania_australia
margaret_creek_western_australia
mt_read_tasmania_australia
noah_creek_queensland_australia
seqapa_cakaudrove_province_fiji
round_hill_nature_reserve_mallee_south_wales_australia
barro_colorado_island_panama
cabo_san_lucas_baja_california_sur_mexico
nari_sonora_mexico
empalme_sonora_mexico
san_bartolo_baja_california_sur_mexico
allegheeny_national_forest_pennsylvania
archbold_biological_station_florida
big_hammock_natural_and_wildlife_management_area_georgia
connecticut_river_near_middletown_connecticut
connecticut_river_swamp_near_middletown_connecticut
connecticut_river_upland_near_middletown_connecticut
dilchers_woods_upland_gainesville_florida
dilchers_woods_lowland_gainesville_florida
duke_forest_north_carolina
florida_panther_national_wildlife_refuge_florida
e_n_huyck_preserve_and_biological_research_station_new_york
harvard_forest_machusetts
hawk_mountain_sanctuary_pennsylvania
hubbard_brook_experimental_forest_new_hampshire
institute_for_ecosystem_studies_new_york
little_pee_dee_state_park_south_carolina
smithsonian_environmental_research_center_maryland
williwil_nui_ridge_trail_ohau
nagakubo_yakushima
nagakubo_yakushima
suganuma_honshu
yakusugi_1080m_yakushima
yakusugi_1350m_yakushima
yakusugi_260m_yakushima
yakusugi_800m_yakushima
danum_valley_field_center_borneo
pasoh_forest_reserve
frim_kepong
borinquen_in_guanica_state_forest
buena_vista_in_maricao_state_forest
cabo_rojo_in_borinquen_state_forest
guajataca
guanica_state_forest
monte_guilarte_in_guilarte_commonwealth_forest
toro_negro_state_forest
bandon_oregon
cape_blanco_oregon
clearwater_washington
government_camp_multitorpor_mountain_oregon
half_moon_bay_california
hood_river_oregon
nestucca_river_oregon
north_bend_oregon
placerville_california
powers_coquille_river_oregon
red_feet_reservoir_utah
republic_north_fork_of_sanpoil_river_washington
santa_cruz_new_brighton_state_beach_california
three_lmn_clickamas_river_oregon
wolf_creek_colorado

homer_kachemak_bay
ketchikan
talkeetna_australia_river
frederick_bonanza_australia
hobart_tasmania_australia
leka_cakaudrove_province_fiji
hobart_tasmania_australia
kuningal_chase_national_park_west_new_south_wales_australia
margaret_creek_western_australia
mt_read_tasmania_australia
noah_creek_queensland_australia
seqapa_cakaudrove_province_fiji
round_hill_nature_reserve_mallee_south_wales_australia
barro_colorado_island_panama
cabo_san_lucas_baja_california_sur_mexico
nari_sonora_mexico
empalme_sonora_mexico
san_bartolo_baja_california_sur_mexico
allegheeny_national_forest_pennsylvania
archbold_biological_station_florida
big_hammock_natural_and_wildlife_management_area_georgia
connecticut_river_near_middletown_connecticut
connecticut_river_swamp_near_middletown_connecticut
connecticut_river_upland_near_middletown_connecticut
dilchers_woods_upland_gainesville_florida
dilchers_woods_lowland_gainesville_florida
duke_forest_north_carolina
florida_panther_national_wildlife_refuge_florida
e_n_huyck_preserve_and_biological_research_station_new_york
harvard_forest_machusetts
hawk_mountain_sanctuary_pennsylvania
hubbard_brook_experimental_forest_new_hampshire
institute_for_ecosystem_studies_new_york
little_pee_dee_state_park_south_carolina
smithsonian_environmental_research_center_maryland
williwil_nui_ridge_trail_ohau
nagakubo_yakushima
nagakubo_yakushima
suganuma_honshu
yakusugi_1080m_yakushima
yakusugi_1350m_yakushima
yakusugi_260m_yakushima
yakusugi_800m_yakushima
danum_valley_field_center_borneo
pasoh_forest_reserve
frim_kepong
borinquen_in_guanica_state_forest
buena_vista_in_maricao_state_forest
cabo_rojo_in_borinquen_state_forest
guajataca
guanica_state_forest
monte_guilarte_in_guilarte_commonwealth_forest
toro_negro_state_forest
bandon_oregon
cape_blanco_oregon
clearwater_washington
government_camp_multitorpor_mountain_oregon
half_moon_bay_california
hood_river_oregon
nestucca_river_oregon
north_bend_oregon
placerville_california
powers_coquille_river_oregon
red_feet_reservoir_utah
republic_north_fork_of_sanpoil_river_washington
santa_cruz_new_brighton_state_beach_california
three_lmn_clickamas_river_oregon
wolf_creek_colorado

Predicted site



Supplemental Figure 1: Linear relationship between median ranked variability and count. Linear regression was used to model -median rank variability (higher values indicated more variability within a plant family) as a function of the abundance of each plant family in the dataset, as measured by $\log_{10}(\text{leaf count})$. The model (shown in blue) was used to estimate overall leaf shape variance in each plant family, as corrected for sampling depth, by using the residuals from the model as an indication of diversity.