TITLE

FTIR spectra from grass pollen: a quest for species-level resolution of Poaceae and Cerealia-type pollen grains

AUTHORS

Katsi, F.[1], Kent M.S.[1], Jones M.[2], Fraser W.T.[3], Jardine P.E.[4], Eastwood W.[5], Mariani M.[3], Osborne C.[6], Edwards S.[1], and Lomax B.H.[1]

ABSTRACT

Palynological analysis based on spore and pollen morphology is well established in the field of palaeo-environmental reconstruction but is currently not fully exploited for understanding the history and development of cereal cultivation due to difficulties in visually differentiating between grass species (Poaceae). Here we employ a chemotaxonomic approach, by examining the chemical differences amongst Poaceae taxa, based on Fourier-transform infrared (FTIR) microspectroscopy data to overcome problems associated with morphological similarities across the Poaceae family. FTIR spectra of untreated and acetolysed pollen from 19 Poaceae taxa were used in our study. We used both populations and individual pollen grains to explore how we can minimize the effect of Mie scattering (spectral distortions caused by scattering of the incident IR beam) on spectra from individuals. Random forest classification algorithms were applied to explore our ability to differentiate

taxa at the species level. We found that pollen grains treated with acetolysis yield better classification results (86% for individuals and 97% for populations) compared to untreated samples (65.7% for individuals and 83% for populations), since they are less affected by Mie scattering. The high classification success at species level on acetolysed individual pollen grains suggests that our chemotaxonomic method holds substantial promise in numerous areas of grass and in particular cereal pollen research, including elucidating the history of agriculture.

Key words: Chemotaxonomy, Poaceae pollen, individual grains, acetolysis, Random Forest

HIGHLIGHTS

- Chemotaxonomy applied on Poaceae populations and individual pollen grains.
- Chemical spectra of individual pollen grains are comparable to population spectra.
- Individual grains of acetolysed pollen were scanned without embedment matrix.
- Spectra from individual acetolysed pollen grains. showed minimal Mie scattering.
- Chemotaxonomy can be useful tool for fossil pollen classification.

1. INTRODUCTION

Agricultural tradition came along with the need to manage and adapt cultivation practices during periods of instability or environmental stress, which is still a major challenge for humanity (Altieri et al., 2015; Riehl et al., 2015, 2014). Therefore, understanding how past societies adapted their cultivation practices can help us develop more resilient agricultural systems in the future. In this work, we test a novel application based on pollen biochemistry to aid discrimination between wild grasses and Cerealia-type pollen which would allow us to use pollen records for a more holistic understanding of cereal cultivation history.

The study of ancient agriculture has developed alongside the analysis of archaeobotanical remains preserved in archaeological contexts, primarily charred plant microremains and phytoliths (Fuller, 2007; Fuller and Lucas, 2014; Piperno, 2011). Nevertheless, these attempts to reconstruct past agricultural systems are usually incomplete, since even the most informative archaeological contexts tend to represent a limited range of past floral diversity (Fuller & Lucas, 2014). Pollen data can however reveal numerous examples of past landscape management, including clearance for pastoral and agricultural activities (arboriculture, cultivation of cereal and legume crops) and the emergence of secondary forests following abandonment of farmlands (England et al., 2008, Li et al., 2008; Marquer et al., 2017; Morrison et al., 2018; Roberts, 2015, 2002; Trondman et al., 2015). Despite the presence of pollen from cereal crops in sediment cores (e.g., Williams et al., 2018), their use for uncovering past agriculture practises is not always straightforward (see Eastwood et al., 2018). One of the shortcomings of pollen analytical data are the morphological similarities among pollen grains of different species within the Poaceae family, which

includes not only domesticated cereal crops but also wild grasses, and causes them to be near-indistinguishable under the light microscope and/or scanning electron microscope (Fægri and Iversen, 1989; Mander et al., 2013; Schüler and Behling, 2011). However, accurate identification of Poaceae pollen and in particular discrimination of Cerealia-type pollen in the palaeoenvironmental record is important when trying, for example, to reconstruct changes in agricultural practices, the introduction of new crops into existing farming systems, the adaptation of local societies to climate change, understanding the initial exploitation of "proto-domesticated" cereals and tracing the beginning of cereal domestication (de Vareilles et al., 2021; Marston, 2021)

1.1 Poaceae palynological studies

Different analytical approaches have been applied to discriminate between grass pollen of native wild plants (not cultivated nor exploited by humans) and cereal crops (domesticated plants): i) analysis of the morphological characteristics and the size of the pollen grains under the light microscope (Andersen, 1979; Bottema, 1992; Dickson, 1988; Fægri and Iversen, 1989; Hapsari and Ballauff, 2022; Joly et al., 2007; Küster, 1988, Rowley, 1960; Schüler and Behling, 2011), ii) analysis of the surface patterns of the pollen grains using observations from scanning electron microscopy (SEM) (Andersen & Bertelsen, 1972; Grohne, 1951, Köhler & Lange, 1979; Mander et al., 2013), and iii) confocal microscopy which can be used to study the sculpture of the exine (Salih et al., 1997). Of those methods, the most commonly used considers the size difference between cereals and wild grasses (Eastwood et al., 2018, Bottema et al., 1992, Küster, 1988) and the eccentrical position of the pore in *Secale cereale* grains (Beug, 1961). The main disadvantage of this approach is

that the distributions of pollen sizes can overlap considerably between some cereals and wild grass species (Bottema et al. 1992, Faegri and Iversen, 1989; Joly et al., 2007) and this has resulted in the misclassification of “Cerealia”-type pollen as wild grass pollen (Hapsari and Ballauff, 2022). Joly et al. (2007) suggested that up to 41% of “Cerealia”-type species could be misclassified as wild species, while 30% of wild grasses could potentially be misclassified as “Cerealia”-type. Köhler & Lange (1979) introduced broad sub-categories for the cereal crops (e.g., Hordeum-type, Triticum-type, Avena-type, Setaria-type, etc.) (see also Wei et al. 2023), when SEM images of the pollen surface ornamentation patterns are used in combination with size/shape criteria. However, those categories include multiple genera and are therefore not appropriate when investigating the diversification of cereal cultivation, which requires species specific identifications. Additionally, exine ornamentation is not always visible under the light microscope (Mander et al. 2013), while preservation issues could be a complicating factor for robust identifications (Bottema et al. 1992, Eastwood et al., 2018). Computational image-based methods have also proven successful (Mander et al. 2013), but SEM microscopy is not only very expensive and time-consuming, but also requires extensive sample preparation and high level of expertise, so its use has been relatively limited in the fossil record (Julier et al. 2016).

1.2 Studies of Fourier-transform infrared spectra of pollen

Recent studies (Diehn et al., 2020; Jardine et al., 2021, 2019; Julier et al., 2016) have successfully classified modern Poaceae pollen grains using chemical spectra obtained by Fourier-transform infrared (FTIR) spectroscopy, yielding classification accuracies above 80% at the subfamily level. Jardine et al. (2019) and Julier et al.

(2016) used pollen spectra obtained by scanning populations (group of pollen grains of the same species) of 8 different extant Poaceae taxa. Their results showed that the region below 1800cm^{-1} in the FTIR spectra, known as the “fingerprint region”, represented the most information-dense region in terms of chemistry and contained a disproportionate amount of chemical variation amongst their pollen samples. Subfossil sediments, however, contain a mixture of different pollen, and therefore this necessitates the scanning of individual pollen grains rather than populations. One chemotaxonomic study of single pollen grains of modern wild grass species grown in greenhouses by Diehn et al. (2020), accomplished an 83% success rate at the species level. Diehn et al. (2020) reported species-specific classification successes between 63% and 94% despite complications arising from spectral scattering. Their findings showed that chemotaxonomy surpasses the taxonomical resolution of most optical techniques. However, to obtain “scatter-free” spectra from individual pollen grains, which usually exhibit Mie scattering (Bassan et al., 2009) due to the spherical shape and the small size of the grains that coincides with the size of IR beam, the authors embedded the pollen in paraffin. Since the produced spectra included peaks related to the paraffin the researchers tried to distinguish the pollen spectra from the paraffin related signal, which not only complicated the analysis but also meant that part of the fingerprint region between 1500cm^{-1} to 1300cm^{-1} was omitted. The omitted spectral region is part of the fingerprint which carries the most variation among Poaceae species and therefore diagnostic potential was reduced (Jardine et al. 2019). Additionally, this approach adds an extra time-consuming stage on the analysis undermining the potential of FTIR for high-throughput data generation.

These previous Poaceae-FTIR studies focused on either a limited number of species (Diehn et al. 2020) or untreated pollen (Diehn et al. 2020, Jardine et al. 2019), which contain organic compounds that do not survive in fossil or sub-fossil samples. Additionally, fossil and sub-fossil pollen is routinely treated with acetolysis (a 9:1 mixture of acetic anhydride and sulphuric acid, Erdtman, 1960) to remove any extraneous compounds derived from the fossil matrix, isolate the sporopollenin and stain the grains to facilitate identification. Acetolysis is also used to isolate the sporopollenin from fresh pollen. However, acetolysis not only isolates the sporopollenin by removing protein related peaks (at 1550 cm^{-1} and 1650 cm^{-1}), reducing the height of aliphatic peaks (at 2925 cm^{-1} and 2850 cm^{-1}) but also alters the pollen chemistry with respect to “pure” sporopollenin. Those alterations include the reduction of non-aliphatic peaks (the 1710 cm^{-1} carboxyl peak (in *Lycopodium*), the 1510 cm^{-1} aromatic peak, and the aliphatic C–O peaks at 1100 cm^{-1} and 980 cm^{-1}), the increase of others (eg. the 1710 cm^{-1} carboxyl peak (in Angiosperms), peaks at 1230 cm^{-1}, 1175 cm^{-1} and 1025 cm^{-1}), while it can add extra peaks in the spectra (eg. 1170 cm^{-1} and 1030 cm^{-1}) (Domínguez et al., 1998; Jardine et al., 2021, 2015). Yet the chemistry of acetolysed sporopollenin itself has been shown to be useful for UV-B reconstructions (Jardine et al., 2016) and ideally classifications (Jardine et al, 2021), and therefore the application of chemotaxonomic methods on sub-fossil pollen treated with acetolysis could be beneficial for taxonomic purposes. Here we compare and contrast the results of chemotaxonomic analyses of untreated and acetolysed pollen for the first time.

The FTIR spectra of pollen are inherently high dimensional data where each dimension represents transmission or absorbance of infrared at a particular wavenumber or group of wavenumbers (dependent on the resolution of the scan).

166 Due to the number of dimensions in the data, it quickly becomes inefficient for visual comparisons to be made between increasing numbers of samples. As such, 168 computational methods such as dimensionality reduction techniques and supervised machine learning algorithms are commonly used for classification purposes 170 Supervised machine learning (ML) algorithms, for example *k*-nearest neighbour (*k*-NN) (Dell'Anna et al., 2009; Jardine et al., 2019; Julier et al., 2016; Woutersen et al., 172 2018) and partial least squares regression (PLS) (Diehn et al., 2020; Zimmermann, 2018, 2010; Zimmermann et al., 2017, 2016, 2015) analyses are widely used for 174 chemotaxonomic classification in studies using FTIR spectra of pollen. These "supervised" algorithms generate predictive models trained on labelled classes of 176 observations in training datasets. These models can be used to predict the class of observations which might otherwise have been withheld or not known. ML-178 classification results are usually compared with unsupervised ML methods (e.g., principal components (PCA) and hierarchical clustering (HCA) analyses) to cross-180 reference results. Although both *k*-NN and PLS models have achieved high classification accuracies when using pollen spectra, *k*-NN performs more poorly as 182 the dimensionality of the data increase (Murphy, 2012, pp. 18-19), and the PLS algorithm has been criticized regarding its ability to process multi-class, imbalanced 184 data or datasets larger than 200 observations (Lee et al., 2018). Some studies have highlighted that random forests (RF) is a very robust ML algorithm that performs well 186 on multi-dimensional large datasets (Singh et al., 2016; Sobol and Finkelstein, 2018; Ziegler and König, 2014). Irrespective of which ML algorithm researchers have used, 188 the main drawback of these methods was the time spent training the models (Sobol and Finkelstein, 2018) and there is also the possibility of overfitting a model causing 190 an inflation of model accuracy (Murphy, 2012). Yet, the risk of overfitting can be

reduced by training the model with an appropriately large, well-balanced (in terms of
192 the number of observations per class) and representative dataset, the use of cross-validation (*k*-folds, leave-one-out) approaches during training, and by testing the
194 prediction accuracy of the algorithm with a separate dataset (Murphy, 2012). Cross-validation is a particularly effective method of assessing the potential performance of
196 a model on unseen data. It is a resampling approach that splits the training data into *k* folds of approximately the same size and uses all but one fold for training a model,
198 and testing the model which predicts values for the last fold. Train and test repeats are performed *k* times for each *k* fold to be tested. Performance measures (such as
200 accuracy, precision and recall) indicate how well the model might generalise on unseen data (Murphy, 2012), reducing the chance of a model overfitting its training
202 data and yielding overoptimistic estimates of a model's utility.

1.3 This study

204 Here we expand on previous work by using the largest grass dataset to date, comprising 19 taxa that grew in a variety of environmental conditions and regions,
206 analysing both populations of pollen and individual grains which were analysed as untreated and acetolysed samples. We test the following hypotheses: a) the
208 chemical spectra from sporopollenin (acetolysed samples) contain enough information for taxonomic classification, similar to chemical spectra from untreated
210 pollen grains and b) spectral classification of individual Poaceae pollen grains from untreated and acetolysed pollen is possible without emending the grains in any
212 medium.

214 **2. MATERIALS AND METHODS**

2.1 Pollen collection

Anthers of 19 Poaceae taxa were used in this study, comprising 7 domesticated cereal and 12 wild grass species (Figure S1). Pollen was harvested from plant populations in Greece (Larisa), Germany (University of Münster Botanical Garden) and UK (University of Nottingham Sutton Bonington glasshouses, Sheffield Botanical Garden, Wollaton Park in Nottingham and James Hutton Institute in Scotland), though not all species at each location. For each taxon, at least three plants were sampled during their flowering seasons of 2018 and 2019. Additionally, *Sorghum halepense* and *Secale cereale* pollen was purchased from Sigma-Aldrich and *Poa pratensis* pollen from Allergon (Table 1).

2.2 Chemical processing

For the acetolysis, the standard procedures described by Fægri and Iversen (1989), with some modifications, were followed: to remove labile compounds from the pollen grains, an acetolysis treatment (a solution of 90% of acetic anhydride ($C_4H_6O_3$) and 10% sulphuric acid (H_2SO_4)) was added to the pollen samples and heated for 3 minutes in hot water bath at 80-85°C temperature. Prior to and after acetolysis samples were treated with glacial acetic acid to remove water from the samples and avoid explosive reactions. The samples were then washed with deionised water and stored in Eppendorf tubes until being scanned.

2.3 FTIR spectra acquisition

Pollen samples were pipetted directly onto clean CaF_2 windows. The FTIR spectra were measured using an Agilent Cary 670 FTIR spectrometer fitted with a KBr beamsplitter coupled with a Cary 610 FTIR imaging microscope with a liquid

238 nitrogen-cooled focal plane array detector, at the School of Biosciences, University of Nottingham. Spectra were collected in transmission mode with a resolution of
240 4cm^{-1} at 128 scans per replicate. Background spectra were collected using 256 scans prior to the data generation for each taxon and every ten replicates thereafter.
242 The background scan was automatically subtracted from the sample scan by the Resolutions Pro software (Agilent Technologies). For the population scans 20
244 replicates were scanned per taxon using a 352 x 352 μm^2 aperture size, and for the analysis of individual pollen grains 30 grains per taxon were scanned using a 72 x 72
246 μm^2 aperture size. Each population scan consisted of a contiguous cluster of at least 20 pollen grains. The scan range was limited to 4000 to 950 cm^{-1}.

248 *2.4 Spectral analysis and classification*

Only the fingerprint region (wavenumbers below 1800 cm^{-1}) which carries most of
250 the chemical information useful in taxonomic classification was used in the analysis, as this decreases model training time approximately 3-fold without significantly
252 impacting the achievable taxonomic resolution. The spectra were corrected using extended multiplicative scatter correction (EMSC) to correct for baseline differences,
254 scaling effect, minimise the Mie scattering and aid the classification. EMSC is frequently used on spectral data to reduce absolute absorbance differences among
256 spectra and the variation between samples that could be due to FTIR beam scattering effects (Rinnan et al., 2009). To examine which species carried the most
258 intersample variation in their spectra we plotted the mean spectrum and corresponding standard deviation for each wavenumber. We also used the pooled
260 variance estimate to quantify the variability in the entire spectrum, because the rate of change of mean spectrum was higher than the rate of change of the standard

262 deviation (Dodge, 2008). We calculated the pooled variance, by multiplying the square of species' standard deviation per wavenumber by the number of samples,
264 and dividing them by the number of species multiplied by the number of wavenumbers with Bessel correction (Dodge, 2008; Radziwill, 2017). Additionally,
266 the first derivatives of the spectra were used in order to inspect spectral details from broad peaks and aid classification (Jardine et al. 2019, Zimmermann and Kohler,
268 2013).

We used the random forest (RF) algorithm in each of the four datasets
270 (untreated/acetolysed populations and individual pollen grains) for species classification, creating four sets of classification models; one set of classification
272 models for untreated populations, one for acetolysed populations, one for untreated individual pollen grains and one for acetolysed individual pollen grains. The RF
274 algorithm uses multiple decision trees on the training dataset and for each observation outputs the most popular prediction (Breiman, 2001). Additionally, we
276 implemented the varImp() function included in the caret R package that reports the importance of each wavenumber for data classification (variable importance). For
278 each set of models, we randomly split the data (untreated/acetolysed populations or individual pollen grains) in two groups: a training dataset (80% of the total spectra of
280 each species) and a test dataset (20% of the total spectra of each species). For each of the four sets of classification models we used the default parameters of 500 trees,
282 mtry equal to the square root of the number of wavenumbers in the data, a minimum node size of 1 and which tested both "gini" and "extratrees" split rules. Ten-fold
284 cross-validation was used on the training set to establish the model parameters with the best classification accuracy. This validation process randomly split the training
286 dataset into 10 parts. The model was then trained with observations from 9 of the 10

parts and the remaining observations were used as validation data to choose the best model parameters. Once the model is trained and validated it was then asked to predict the class (species) of the test dataset and the accuracy of the predictions was reported.

The training, including the ten-fold cross validation, and test procedure was repeated 100 times for each of the four datasets. Every time a model was repeated it was trained with a different training dataset randomly subset (80% of the total spectra of each species) and tested in a different withheld test subset. Therefore, we generated 4 sets of 100 models per dataset: a) the untreated populations, b) the untreated individual pollen grains c) the acetolysed populations, and d) the acetolysed individual pollen grains. Classification success rates were reported as a range of values for each of the four datasets: untreated/acetolysed populations/individual pollen grains. The median classification success predictions of the test subset for each dataset were presented in a confusion matrix.

Principal components analysis (PCA) was used for ordinating the data and evaluate differences among pollen spectra of different species (Diehn et al. 2019, Jardine et al. 2019). Initially we used the whole fingerprint region (1800 to 950 cm^{-1}) for the analysis. However, since there were 443 wavenumber divisions in the fingerprint region, many were autocorrelated, the first two principal components (PCs) typically explained approximately 50% of the variation in the four datasets and clustering by taxon was not particularly pronounced (Supplementary Fig. S4a-d). Instead, we used the wavenumbers variable importance(s) greater than 60% respectively for the 100 RF runs of each training dataset. Following this step, the clustering of like-taxa in the PCA was clearer for all datasets, and the amount of variation explained by the first

two PCs typically increased by ~30% with respect to the PCA of all fingerprint region
312 wavenumbers, so that most of data variation is now explained by the first two principal components (Anderson, 2003).

314 Data analysis was performed in R (R Core Team, 2021) via RStudio version 4.0.4/2023.06.0+421 (RStudio Team 2021), using the packages *EMSC* version 0.9.2
316 (Liland, 2017) for data processing, and the packages *gplots* version 3.3.4 (Warnes et al., 2020), *scico* version 1.2.0 (Pedersen and Crameri, 2020) and *corrplot* version
318 0.89 (Taiyun and Viliam, 2017) for data visualisation. The *caret* package version 6.0-88 (Kuhn, 2019) was used for training, validating and testing the classification
320 models.

322 **3. RESULTS**

3.1 FTIR spectra

324 Spectra from individual pollen grains (either untreated or acetolysed pollen) exhibit the same main chemical peaks as the respective spectra from the population scans.
326 However, spectra from individual pollen grains exhibit more variation than those from populations (Fig. 1-2 and S2). Acetolysis, through removal of labile compounds,
328 results in a change of the pollen chemistry, with several peaks in the FTIR spectra reducing in size (eg. 1230 cm^{-1}), others appearing much stronger (1705 cm^{-1}), and
330 the protein related peaks disappearing (1650cm^{-1}) completely. Spectra from untreated pollen show absorbance peaks at 1743 cm^{-1}, at 1705 cm^{-1} and 1460 cm^{-1}
332 representing lipids, at 1650 cm^{-1} identified as proteins and sporopollenin associated peaks at 1515cm^{-1}, 1230cm^{-1} and 1161cm^{-1}. While spectra from acetolysed pollen

334 exhibit peaks at 1705 cm^{-1} assigned to carboxylic acid *v*(C==O), at 1680 cm^{-1}, 1580 cm^{-1}, 1430 cm^{-1} and 1230 cm^{-1} corresponding to sporopollenin-related compounds
336 and a very pronounced peak at 1034 cm^{-1} and one at 1170 cm^{-1} which have previously interpreted as artificial peaks formed during acetolysis (Bağcıoğlu et al.,
338 2017, 2015; Domínguez et al., 1998; Jardine et al., 2021, 2019; Lutzke et al., 2020).

3.2 Classification

340 The predictive models trained on population scans performed better than those trained on scans of individual pollen grains (Fig. 3). Additionally, the models trained
342 on acetolysed pollen, in general, performed better than those trained on untreated pollen spectra. In detail, the classification success values range from 67% to 93% for
344 the untreated populations, while the median classification success was 83%. The untreated individual pollen grains presented a classification range of 48% to 74.5%,
346 with 65.7% being its median value. For the acetolysed populations the classification success range was 91% to 100%, while the median success value was 97%. The
348 acetolysed individual pollen grains models yielded classification success values between 79% to 92%, with median value of 86%.

350 In the next section the confusion matrices of the prediction results on the withheld test subset of each dataset are presented. Most misclassifications in all datasets
352 occur within Triticum cereal species and their wild relatives (Figs. 4a-d). Additionally, on models trained on spectra of individual pollen grains, both untreated and
354 acetolysed, *Hordeum vulgare* was sometimes misclassified as the wild *Hordeum spontaneum* (Figs. 4b). In the models that used individual, untreated pollen grains,
356 three *Avena sativa* samples were wrongly predicted as Triticum. crops and *Secale cereale* (Fig. 4b). Moreover, in the same model a single *S. cereale* grain was

358 confused as *Aegilops caudata*. The most misclassifications pertaining to the untreated, individual pollen grain data were among the wild relatives of Triticum
360 cereals, with *Thinopyrum elongatum* having the lowest classification success of 17%. The confusion matrix for the acetolysed individual pollen data presents
362 misclassifications mainly between Triticum cereals and their wild relatives, among the wild relatives of Triticum crops, and limited misclassifications of *S. cereale*
364 (towards *Triticum timopheevii* and *Festuca drymeja*), *Avena sativa* (towards *Z. mays*) and *Z. mays* (towards *T. durum* and *H. spontaneum*) (Fig. 4d).

366 To further visualise and understand how taxa are related chemotaxonomically (Fig. 5) we used the wavenumbers with variable importance(s) greater than 60% on the
368 PCA (Fig. S3 a-d). The PCA plots show tighter taxonomic clustering for populations spectra (both untreated and acetolysed pollen) compared to the spectra of individual
370 pollen grains. PCs 1 and 2 of the untreated populations spectra together explain 83.6% of the variation in the dataset. The plot shows broad groups in subfamily level
372 (e.g., Triticum cereal species and their wild relatives); Fig. 5a). There is no clear separation between wild and domesticated species, instead clustering is more
374 dictated by species (or genus in the case of most Triticum species). The compactness of the clusters is variable among taxa: some species like *Poa*
376 *pratensis, Avena fatua, Sorghum halepense, T. timopheevii* and *T. aestivum* form tight clusters, whilst others are more diffuse (e.g., *H. spontaneum*, *A. sativa* and *Z.*
378 *mays*) or split into several sub-clusters (e.g., *Ph. pratense* and *Th. elongatum*). The first two PCs of the untreated, individual pollen grain spectra account for 80.7% of
380 the total variation in that dataset (Fig. 5b). In this plot, most taxa overlap and a few wild grass species spread on the left side of the plot forming diffused taxa clusters
382 (*H. lanatus*, *Ph. pratense and P. pratensis*).

PCs 1 and 2 account for 76.8% of the total variation in the acetolysed populations data (Fig. 5c). In this plot there is a clear separation on species level, while the Triticum cereals and their wild relatives cluster is also apparent (as it is in the PCA of untreated populations; Fig. 5a). The 77.6% of the total variation of the acetolysed, individual pollen grain dataset is explained by the first two PCs (Fig. 5d). The majority of species overlap in ordination space in the centre of the plot making it difficult to discern any trend in taxonomic clustering at species or subfamily level. Exceptions to this undefined clustering were the spectra of *Ph. pratense*, *A. fatua, S. cereale, S. halepense, P. pratensis, T. dicoccoides* and *H. vulgare*. The "Triticum cereal crops and wild relatives" cluster is once again present, however in this instance, *H. lanatus, F, drymeja, H spontaneum, A. sativa* and *Z. mays* are also clustered with the aforementioned Triticum group.

4. DISCUSSION

4.1 Implications of sample preparation

To examine the Mie scattering effect on individual pollen grains of untreated and acetolysed samples we compared those spectra against spectra from population samples, which are not susceptible to scattering (Jardine et al., 2019; Muthreich et al., 2020; Pappas et al., 2003; Zimmermann, 2010). Our study also showed that the averaged EMSC-corrected spectra of individual pollen grains exhibited peaks at the same locations as the spectra obtained from populations of pollen, respectively in both untreated and acetolysed samples. However, spectra from individual pollen grains presented more variation compared to population spectra especially towards the lower wavenumbers, between 1750 cm^{-1} and 1700 cm^{-1}, around 1400 cm^{-1} and

1200 cm^{-1} (Figs 1 and 2). Spectral variability on acetolysed individual pollen grains compared to population samples was more pronounced on a few species (*Z. mays*, *T. timopheevii*, *T. urartu*), contrary to untreated samples where variability was present in most species. We expect that the Mie scattering likely introduced unpredictable distortions to varying regions of the spectra of individual pollen grains rendering them too randomly variable. Despite this spectral variability, the broadly consistent taxon-specific chemistries between spectra of individual pollen grains and populations indicate that classifying spectra from individual pollen grains can yield taxonomically meaningful results.

In our study, we avoided the time-intensive procedures of sample preparation and the complicated pre-processing spectral analysis or elimination of wavenumbers within the targeted fingerprint region that other studies have employed (Diehn et al., 2020; Zimmermann et al., 2015). The spectra from individual pollen grains were generated simply by pipetting the pollen directly onto FTIR microscope (in this case, CaF_2) slides without the use of paraffin or any other mounting medium to reduce scattering (Diehn et al. 2019, Zimmerman et al. 2016). We only corrected the spectra with EMSC and took the 1^{st} derivatives to eliminate spectral inconsistences and aid classification. We found that Mie scattering on individual pollen grains was limited and more pronounced on the untreated individual pollen grains, hence they exhibited the lowest classification success of all datasets. As such, minimal sample preparation or simpler, more commonplace pre-processing techniques that capitalise on the high-throughput potential of FTIR do not always appear practicable in terms of classifying spectra with evidence of some scatter distortion obtained from individual (grass) pollen grains.

We believe that the Mie scattering effect was more distinct on untreated grains compared to acetolysed ones, because acetolysis removes the labile intercellular pollen components making the grains slightly deflated and less spherical (and frequently collapsed) allowing the IR beam to penetrate them more easily increasing their spectral resolution. Untreated pollen grains consist of the intercellular material and an outer pollen wall that is divided into the inner intine (consisting of cellulose and pectin) and the outer exine (sporopollenin). The structure of the Poaceae pollen wall includes, also, the Zwischenkörper, a thin gel foaming pectin layer around and below the aperture, between the intine and the exine of the pollen wall as described in detail by Heslop-Harrison (1979). The Zwischenkörper along with the rest labile compounds of the pollen are eliminated after the use of acetolysis (Domínguez 1998, Jardine et al. 2021, Heslop-Harrison 1979, Li et al. 2019, Lutzke et al. 2020). These pollen materials provide structural support to the pollen grain and therefore after their removal with acetolysis treatment the sporopollenin could explain the slightly deflated appearance of the grains (see S1). Objects with less spherical shape result in less scatter or scatter free spectra (Bassan et al., 2009), similar to the spectra from individual pollen grains of acetolysed pollen grains in this study. For this reason, the most noticeable change on the classification accuracy between untreated and acetolysed pollen appears on the individual pollen grains (Fig. 3 and 4), since the quality of the spectra from populations- acetolysed or untreated- is generally a lot better, as they do not suffer from the scattering affect. Therefore, we suggest that simplifying the preparation of the samples will not affect our ability to classify acetolysed pollen, but there are still some challenges when untreated pollen individual pollen grains are concerned.

4.2 Chemotaxonomic classification on Poaceae pollen

We generated 100 RF models per pollen dataset (untreated/acetolysed, populations/individual grains) which were respectively trained and tested on randomly selected subsets. Classification accuracy of the untreated populations ranged from 67% to 93% (median = 83%) and the individual pollen grains ranged from 48% to 74.5% (median = 65.7%). The range of classification accuracies from the models trained and tested with untreated populations is comparable with the classification accuracies of other published studies on the chemotaxonomy of grass pollen and their spectra (Jardine et al 2019, Julier et al. 2016). However, we increased the number of taxa to be classified- from 8 to 19 species- in our study whereby a correct classification by chance would have been significantly lower, yet the methods employed in this study have shown that our approach yields adequate, comparable results. The classification accuracy range and median using the spectra of untreated, individual pollen grains was lower than those of Diehn et al. (2020) in their study of individual grass pollen grains. In this study, we included 19 species, nearly 4-times as many as the 5 species used by Diehn et al. (2020), and this added dataset complexity and increased difficulty for the RF models might explain our lower successful classification rate. In addition to more species in our study, we also used more closely-related taxa and hence we would expect their spectra to be more difficult to distinguish. It is, however, more likely that the higher classification accuracy in the Diehn et al. (2020) study was because spectral distortions that result from IR beam scattering were reduced by using paraffin mounting and corrected using sophisticated machine learning techniques. As explained above, we did not attempt to reduce or control IR beam scatter or its effects beyond using commonly employed spectral pre-processing techniques and, as a result, limited scatter likely affected their correct classification.

To date, there is only one study (Jardine et al., 2021) that has investigated whether Poaceae acetolysed pollen grains can be classified below subfamily level, considering also the possibility that even finer taxonomic levels could be achieved. Our results pertaining to individual pollen grains of Poaceae agree with those of Jardine et al. (2021). With regards to acetolysed pollen, the clustering in the PCA plots of populations shows clear taxonomical signal, with the plot of individual pollen grains presenting a more complicated story. However, the classification success rates of both acetolysed individual pollen grains (median = 86%) and populations (median = 97%) indicates that a strong chemotaxonomic signal is recoverable from acetolysed pollen, even for species-level classification. Models trained on spectra from acetolysed pollen performed considerably better than untreated pollen samples, especially on individual pollen grains (Fig. 4c-d). Those results suggest that the spectra of acetolysed pollen can be reliably classified in a chemotaxonomic perspective, despite the chemical alterations to the chemistry involved with acetolysis which results to removal of peaks related with labile compounds and addition of peaks (1170 cm^{-1} and 1034 cm^{-1}).

Frequent misclassifications among Triticum cereal species and their wild relatives in all pollen datasets indicate that more work or sophisticated techniques are needed to discriminate such closely related taxa. In all the PCA plots (Fig. 5), irrespective of using acetolysed or untreated samples, there was a large cluster of Triticum cereal crops and their wild relatives, whilst other wild grasses usually plotted in the periphery of this cluster (Fig. 5b and d). The clustering is clearer on the populations, where taxon clusters are also more pronounced, while on individual pollen grains a more diffused cluster with overlap of various Triticum species is present. The RF showed very limited misclassifications of wild grasses as cereal crops, although

506 there was a lot of confusion among the wild grasses, among cereal crops and between wild Triticum relatives and domesticated crops. There may be more
508 misclassifications among Triticum cereals or between cereals and their wild relatives because their chemistries are very similar since they belong to the same genus.
510 Common misclassifications between *H. vulgare* and *H. spondaneum* could reinforce the argument that our data show a phylogenetic signal strong enough to reliably
512 distinguish between genera, but only in some cases at the species level. However, even if we cannot achieve species specific classifications for closely related taxa, our
514 results show that Triticum cereals can be distinguished from common wild grasses and vice versa. Another point highlighted from this study is that misclassifications
516 (e.g., of *T. aestivum, T. durum and T. dicoccum*) did not appear on more chemically variable taxa *(T. urartu*, *F. drymeja*, *T. timopheevii* and *Th. elongatum*) (Figs. 1-2 and
518 S2), as has been suggested in other studies (Jardine et al. 2019).

4.3 Palaeoecological implications

520 Our results have shown that spectra from acetolysed pollen carry a distinct taxonomic signal and therefore this method could be used on sub-fossil samples that
522 are routinely treated with acetolysis. We also showed that it is not necessary to embed pollen grains in any medium to accomplish meaningful classification results,
524 so scanning sub-fossil pollen directly from the CaF_2 slides can readily provide spectra of sufficient quality for chemotaxonomic analysis. Avoiding the use of any
526 embedding medium for sub-fossil pollen scanning, will not only simplify the procedure but also speed the lab work and spectral analysis.

528 However, it should be noted that sub-fossil pollen grains are affected by sedimentation processes that may alter the sporopollenin chemical spectra. A few

studies suggested that sub-fossil late Quaternary sporopollenin (Jardine et al., 2021) and even well preserved Pennsylvanian fossil sporopollenin (Fraser et al., 2012) presented chemical similarities with spectra from extant plants. Therefore, chemotaxonomy could benefit at least well preserved fossil sporopollenin or relatively recent sub-fossil samples. Additionally, sub-fossil pollen, is treated with other chemicals (KOH, HCl) apart from acetolysis, which may leave residues on the chemical signature of the samples, although a lot less detectable (Wang et al., 2023). It is therefore important to include those chemicals in the treatment of extant pollen if we want to create a reference dataset comparable to the sub-fossil chemical spectra.

5. CONCLUSION

Here we have demonstrated that FTIR spectra from untreated and acetolysed pollen grains can be used for classification purposes to species level (or genus level for some taxa). We showed that acetolysis improves the classification accuracy especially on individual pollen grains (86% median classification accuracy), without embedding the grains in scatter-reducing media. As sub-fossils and fossils are frequently treated with acetolysis, we suggest our method is particularly suited to addressing palynological research questions related to the history of cereal cultivation.

6. ACKNOWLEDGEMENTS

552 This research was supported by the Palaeobenchmarking Resilient Agriculture Systems (PalaeoRAS) project funded by the Future Food Beacon of the University of
554 Nottingham. PEJ acknowledges funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number: 443701866. We thank Mr.
556 Antonopoulos A., Astrand J., Waugh R. and Hamilton R. for their assistance during pollen samples collection. We also thank Dr. Kadochnikova A. for her help and
558 suggestions on the statistical analysis of our data.

560

7. AUTHOR CONTRIBUTION

562 BHL, MJ and FK designed the project; FK scanned the pollen samples; FK and MSK analysed the data; FK wrote the manuscript with substantial help from MSK, PEJ,
564 WTF, MM, BHL, MJ and WE, CO and SE provided the pollen samples. All authors reviewed the manuscript.

566

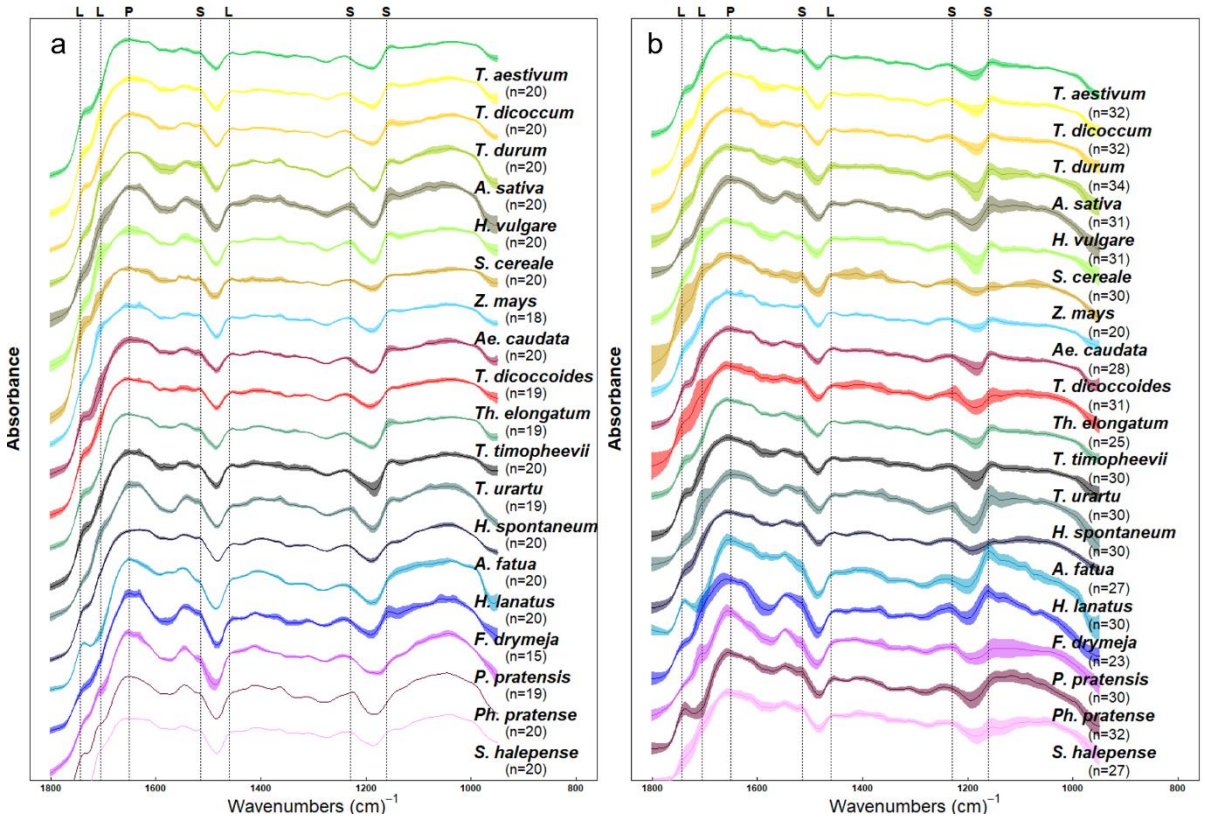
8. DATA AVAILABILITY STATEMENT

568 Data that supports this report and the R code for the analysis can be accessed at: 10.6084/m9.figshare.24190596 [NB For review please use this private link to access
570 the data: https://figshare.com/s/d1b0306f315468837088]

572 **9. STATEMENTS AND DECLARATIONS**

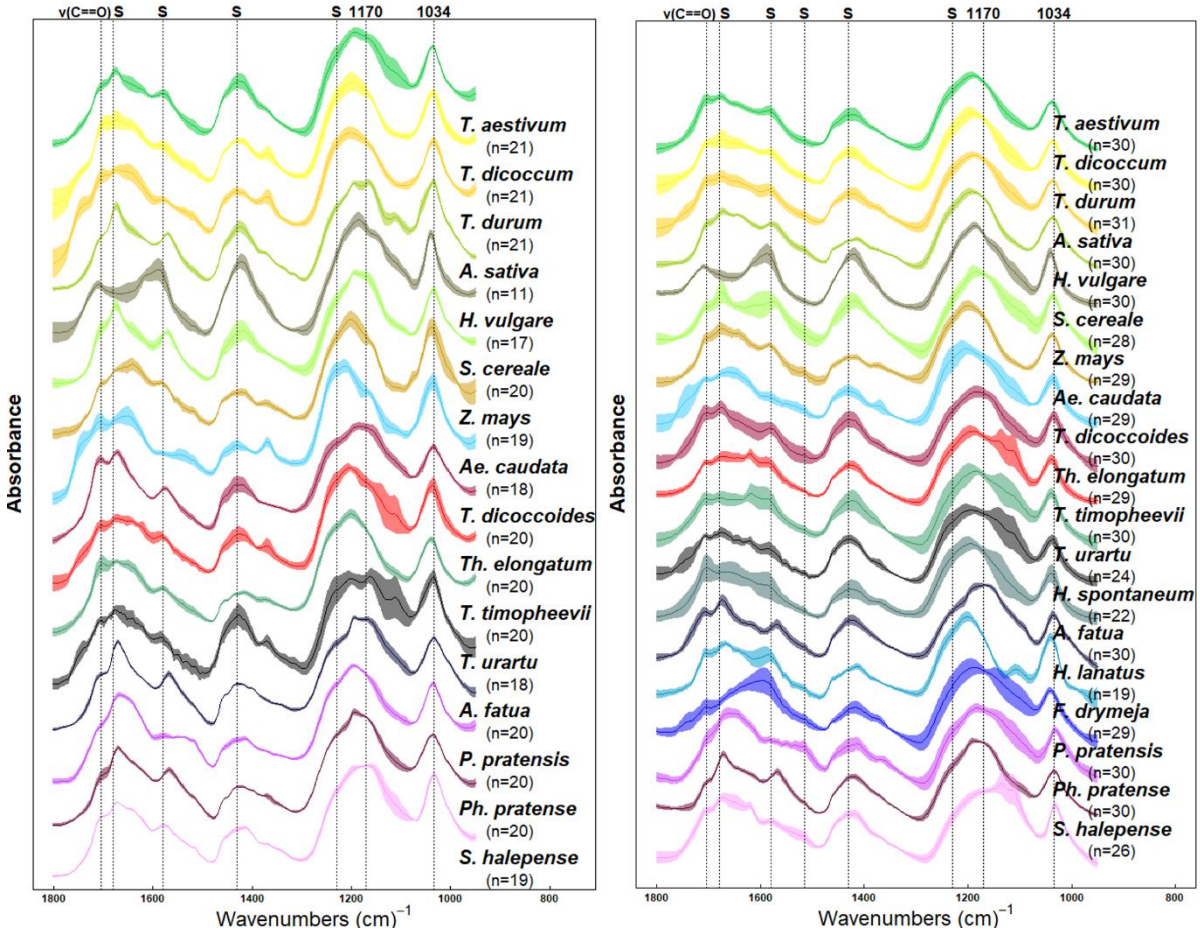
The authors have no relevant financial or non-financial interests to disclose. The authors have no competing interests to declare that are relevant to the content of this article.

576 **Figures:**



578 **Fig. 1** Mean FTIR spectra of pollen of untreated modern species of a) population scans and b) scans of individual pollen grains. The EMSC-corrected fingerprint 580 region of the spectra are plotted. The number of replicate scans used in the analysis are given as *n*. The shaded regions represent the mean ± 1 standard deviation. The 582 vertical dashed lines show the main peaks and their interpretation (L = lipids, P = protein and S = sporopollenin)

584



586 **Fig. 2** Mean FTIR spectra from acetolysed modern species of a) population scans and b) scans of individual pollen grains. The EMSC-corrected fingerprint region of
588 the spectra are plotted. The number of replicate scans used in the analysis are given as *n*. The shaded regions represent the mean ± 1 standard deviation. The vertical
590 dashed lines show the main peaks and their interpretation (S = sporopollenin). The "1170" and "1034" correspond to peaks attributed to acetolysis

592

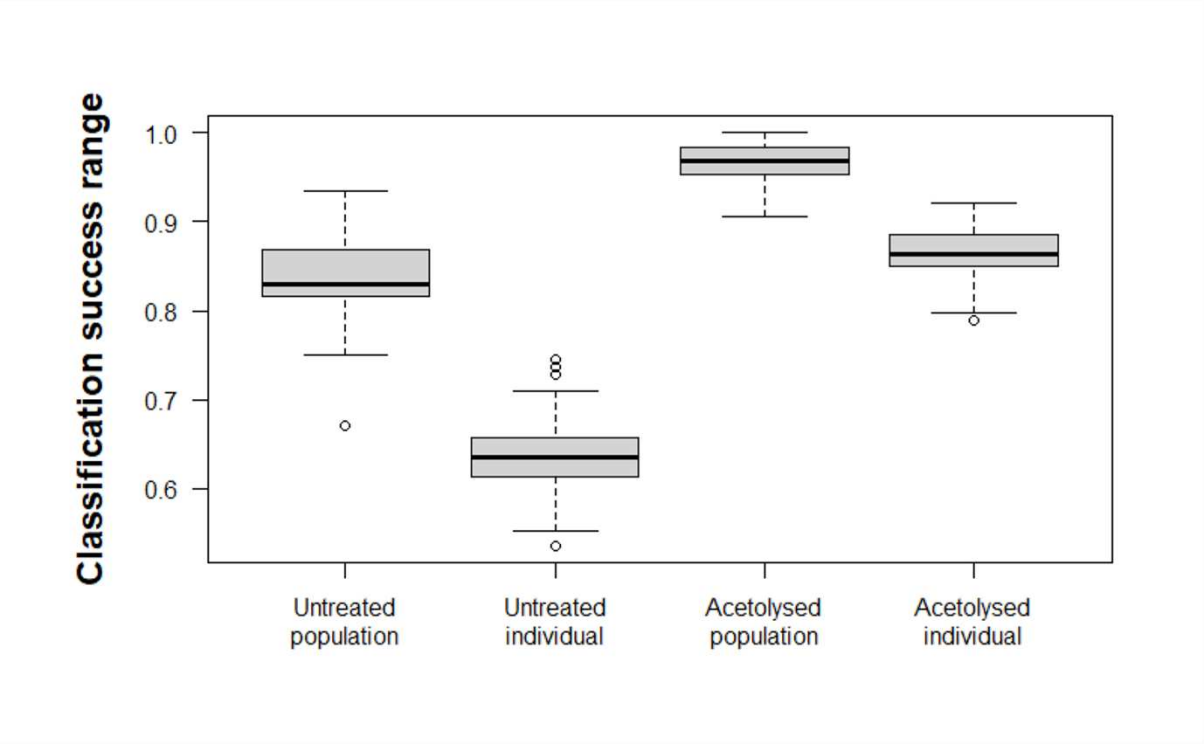


Fig. 3 Classification accuracy range of the test subset from untreated populations, untreated individual pollen grains, acetolysed populations and individual pollen grains after repeating RF models 100 times with different train subsets

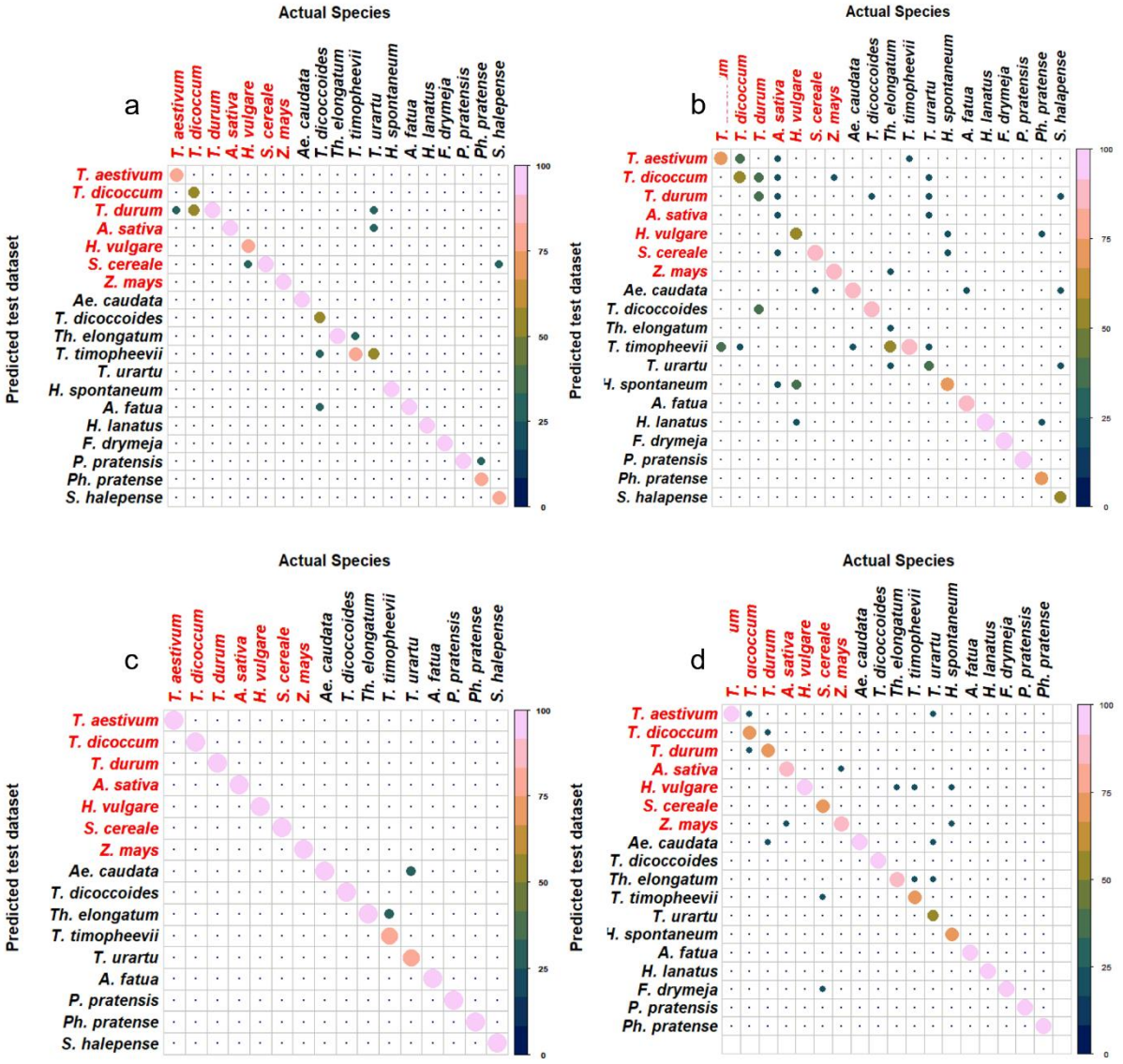
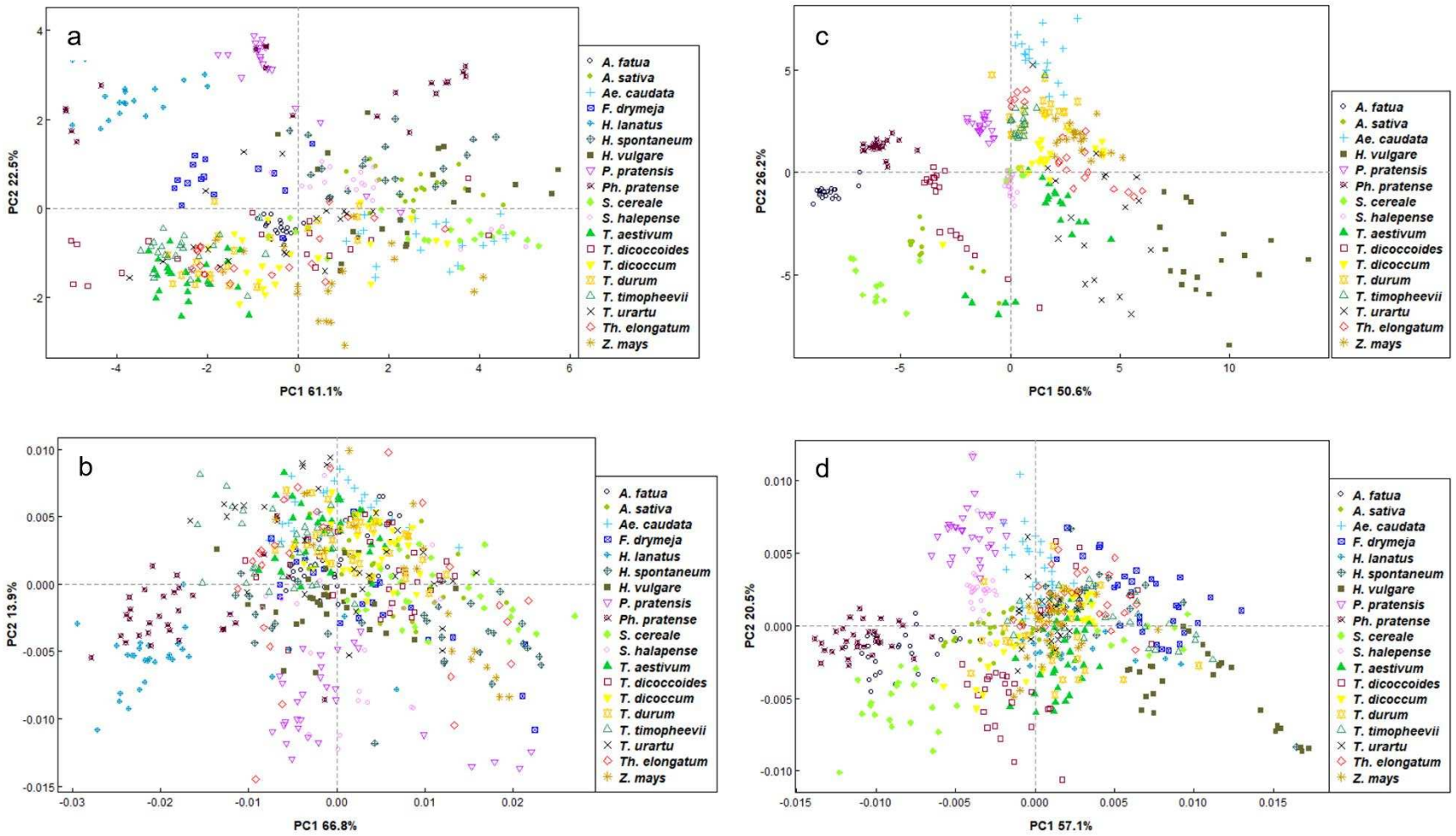


Fig. 4 Confusion matrices showing the classification accuracy (%) of each species from: a) untreated populations, b) untreated individual pollen grains, c) acetolysed populations, and d) acetolysed individual pollen grains. The species labelled in red are domesticated cereal species and the species labelled in black are wild grasses. The confusion matrices present the median classification accuracies of the 100 random forest model runs for each dataset


a
PC2 22.5%
PC1 61.1%
b
PC2 13.9%
PC1 66.8%
c
PC2 26.2%
PC1 50.6%
d
PC2 20.5%
PC1 57.1%
A. fatua
A. sativa
Ae. caudata
F. drymeja
H. lanatus
H. spontaneum
H. vulgare
P. pratensis
Ph. pratense
S. cereale
S. halapense
S. halepense
T. aestivum
T. dicoccoides
T. dicoccum
T. durum
T. timopheevii
T. urartu
Th. elongatum
Z. mays

606

Fig. 5 Principal components analysis (PCA) plots for 1st-derivative spectra of: a) untreated populations, b) untreated individual pollen grains, c) acetolysed populations, and d) acetolysed individual pollen grains. Closed symbols were used for domesticated crops and open for wild grasses. In this PCA, only wavenumbers with variable importance above 70% (according to the variable importance of the random forest runs) were used. The percentage of variance explained by principal components 1 and 2 is indicated in the axes titles

612 **Tables:**

Table 1: List of species (alphabetical order) included in the analysis and their collection origin.

Family	Genus	Species	Common name	Place of collection	Number of plants sampled	Wild or domesticated
POACEAE	**Aegilops**	*Aegilops caudata*	wild wheat	Sutton Bonington (UK)	3	wild
POACEAE	**Avena**	*Avena fatua*	wild oat	Greece	3	wild
POACEAE	**Avena**	*Avena sativa*	oat	Germany, Greece	3	domesticated
POACEAE	**Festuca**	*Festuca drymeja*		Germany	3	wild
POACEAE	**Holus**	*Holus lanatus*	Yorkshire fog, tufted grass, and meadow soft grass	Sutton Bonington, Nottingham and Sheffield (UK)	3	wild
POACEAE	**Hordeum**	*Hordeum spontaneum*	wild barley	Greece and James Hutton Institute (UK)	4	wild
POACEAE	**Hordeum**	*Hordeum vulgare*	barley	Greece, Sutton Bonington and James Hutton Institute (UK)	3	domesticated

POACEAE	**Poa**	*Poa pratensis*	Kentucky bluegrass, smooth meadow-grass, or common meadow-grass	Allergon stock pollen		wild
POACEAE	**Phleum**	*Phleum pratense*	Timothy grass	Nottingham	3	wild
POACEAE	**Secale**	*Secale cereale*	rye	Greece, Germany, Sigma-Aldrich stock pollen	3+	domesticated
POACEAE	**Sorghum**	*Sorghum halepense*	Johnson grass or Johnsongrass	Sigma-Aldrich stock pollen		wild
POACEAE	**Triticum**	*Triticum aestivum*	common wheat or bread wheat	Germany, Greece, Sheffield (UK)	3	domesticated
POACEAE	**Triticum**	*Triticum dicoccoides*	wild emmer	Germany, Sheffield (UK)	3	wild
POACEAE	**Triticum**	*Triticum dicoccum*	emmer wheat	Germany, Greece, Sheffield (UK)	3	domesticated
POACEAE	**Triticum**	*Triticum durum*	pasta wheat or macaroni wheat	Germany, Greece, Sheffield (UK)	3	domesticated

POACEAE	**Triticum**	*Triticum timopheevii*	Timopheev's wheat or Zanduri wheat	Germany, Sutton Bonington and Sheffield (UK)	3	wild
POACEAE	**Triticum**	*Triticum urartu*	wild einkorn wheat	Germany, Sutton Bonington and Sheffield (UK)	3	wild
POACEAE	**Thinopyrum**	*Thinopyrum elongatum*	tall wheatgrass	Sutton Bonington (UK)	2	wild
POACEAE	**Zea**	*Zea mays*	maize	Germany	3	domesticated

10. SUPPLEMENTARY MATERIAL

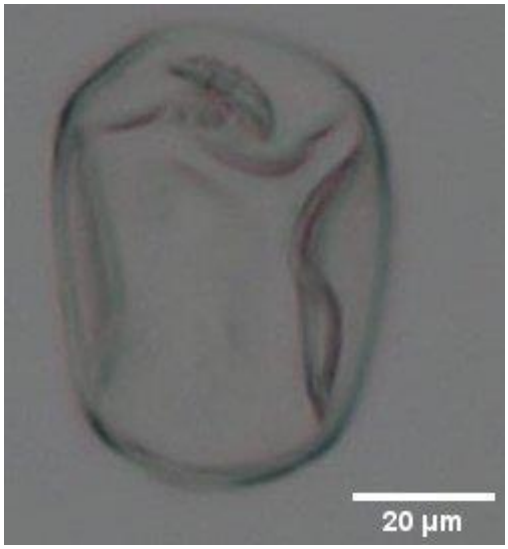
Tables

S2 Table with pooled standard variation values per species for untreated/acetolysed populations and individual pollen grains.

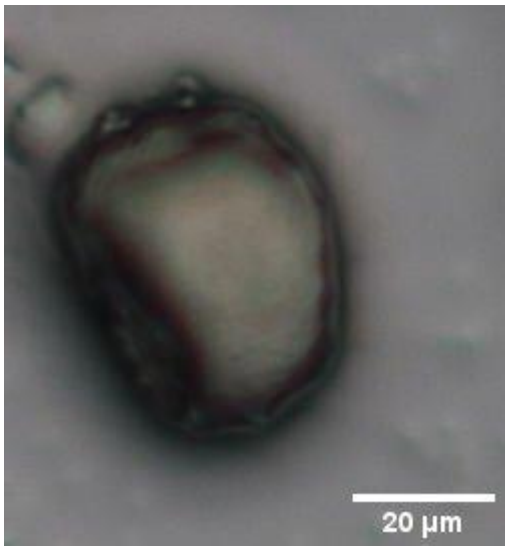
	Pooled SD values for untreated populations	Pooled SD for untreated individual pollen grains	Pooled SD for acetolysed populations	Pooled SD for acetolysed individual pollen grains
T. aestivum	0.003148579	0.005238022	0.009991912	0.005232399
T. dicoccum	0.004435576	0.006206741	0.010528546	0.006200078
T. durum	0.003678485	0.006442158	0.009601097	0.006432725
A. sativa	0.004925851	0.009631786	0.006239655	0.009626445
H. vulgare	0.007960686	0.009868383	0.010851349	0.009862911
S. cereale	0.005891131	0.008435695	0.008881108	0.008420651
Z. mays	0.005215815	0.01222964	0.009368466	0.012222383
Ae. caudata	0.003533944	0.005590626	0.010123889	0.005636001
T. dicoccoides	0.005634411	0.006508374	0.007379908	0.006516387
Th. elongatum	0.004075455	0.012709734	0.011909734	0.012695149
T. timopheevii	0.003431314	0.005749591	0.006538684	0.005769477
T. urartu	0.005484107	0.008425443	0.015188691	0.008389126
H. spontaneum	0.00492411	0.011691405	-	0.011618033
A. fatua	0.002585723	0.006420723	0.004855898	0.006433032
H. lanatus	0.004349624	0.01120298	-	0.011090831
F. drymeja	0.005774456	0.011051747	-	0.01110348
P. pratensis	0.00482031	0.011329483	0.005074327	0.011329483
Ph. pratense	0.006464326	0.010614296	0.004502205	0.010602902
S. halepense	0.004102461	0.010986219	0.005552432	0.010978108

Figures

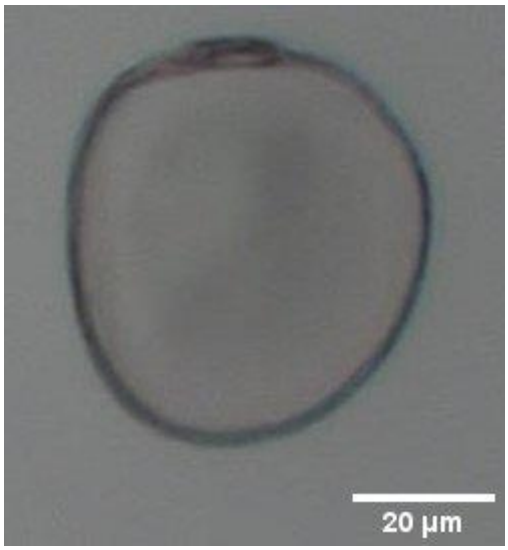
Aegilops caudata (untreated)



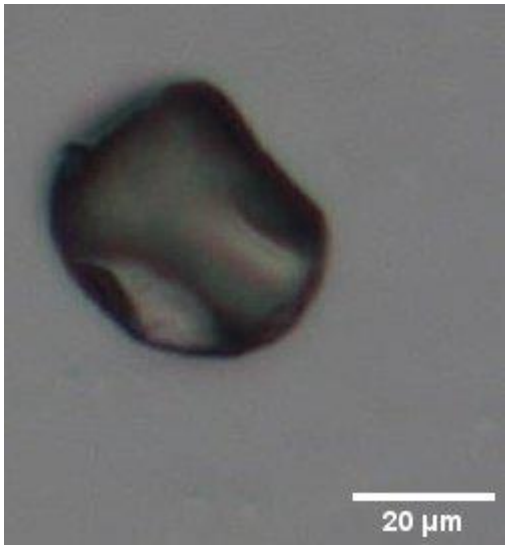
Aegilops caudata (acetolysed)



Avena fatua (untreated)

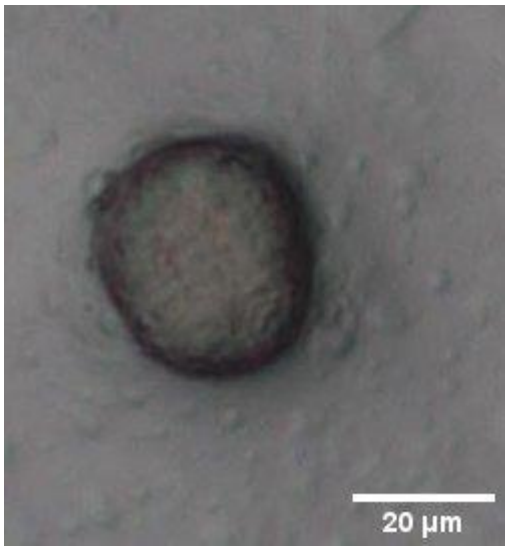


Avena fatua (acetolysed)



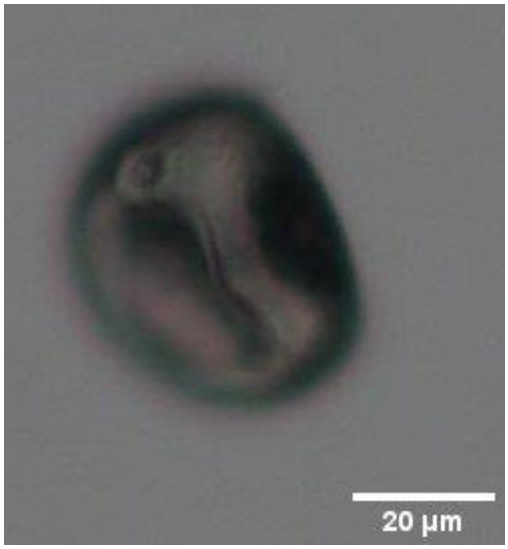
Avena sativa (untreated)

Avena sativa (acetolysed)

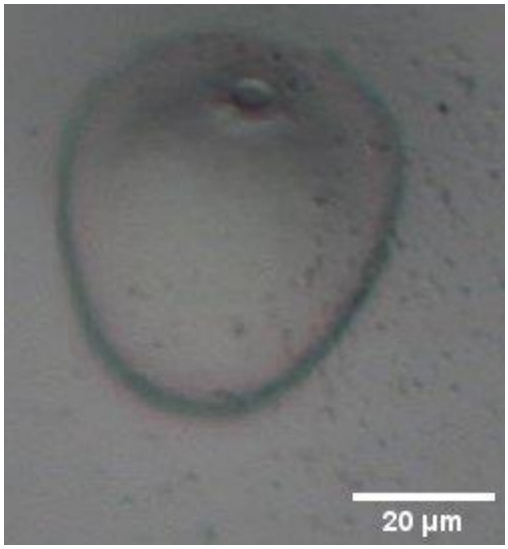


Festuca drymeja (untreated)

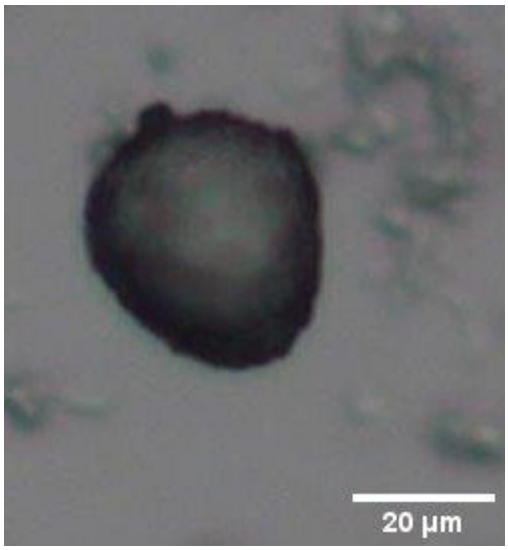
Festuca drymeja (acetolysed)



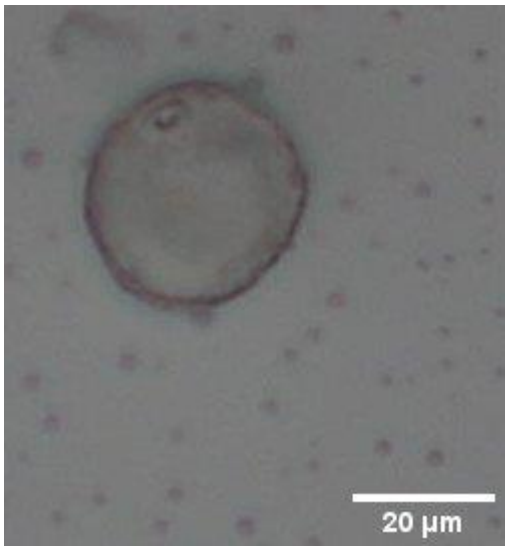
Holus lanatus (untreated)



Holus lanatus (acetolysed)

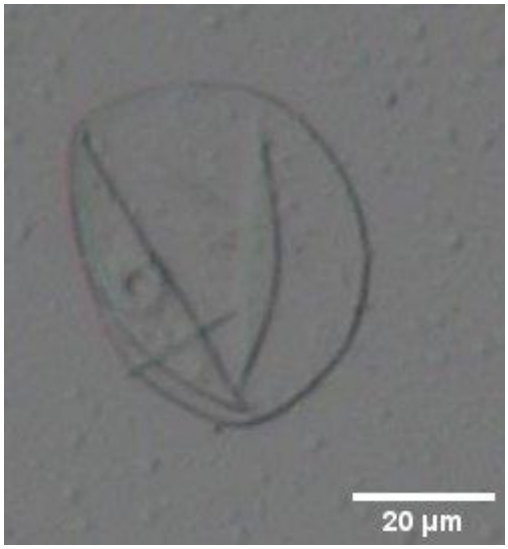


Hordeum spontaneum (untreated)

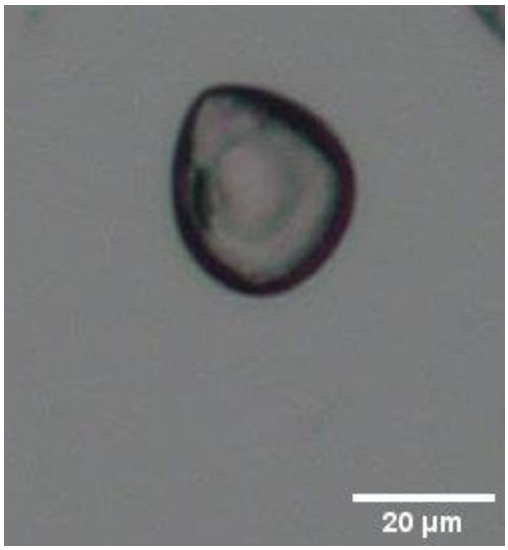


Hordeum spontaneum (acetolysed)

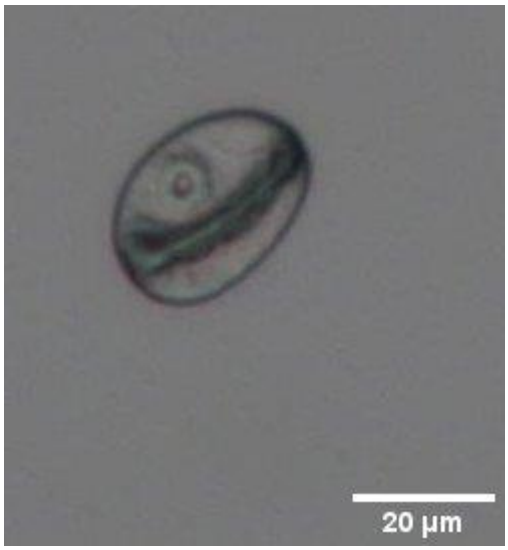
Hordeum vulgare (untreated)



Hordeum vulgare (acetolysed)

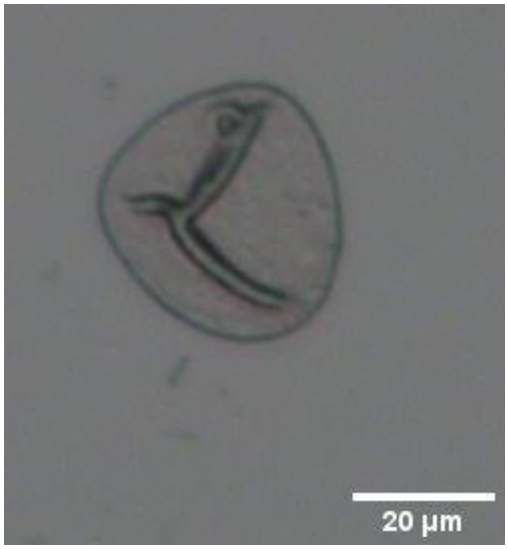


Poa pratensis (untreated)

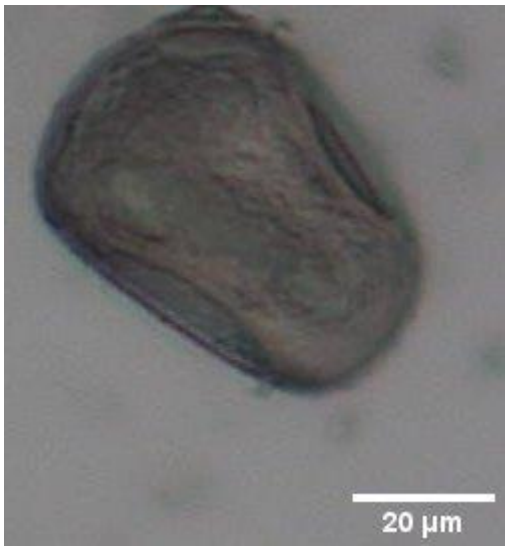


Poa pratensis (acetolysed)

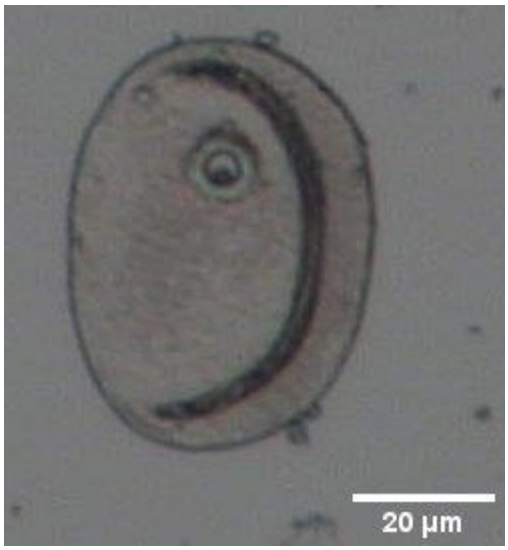
Phleum pratense (untreated)



Phleum pratense (acetolysed)

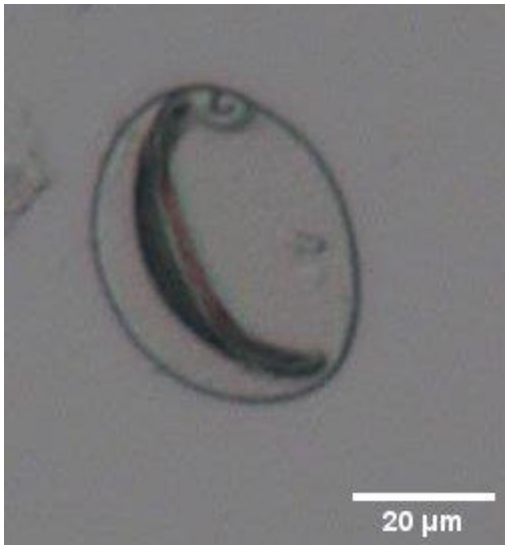


Secale cereale (untreated)



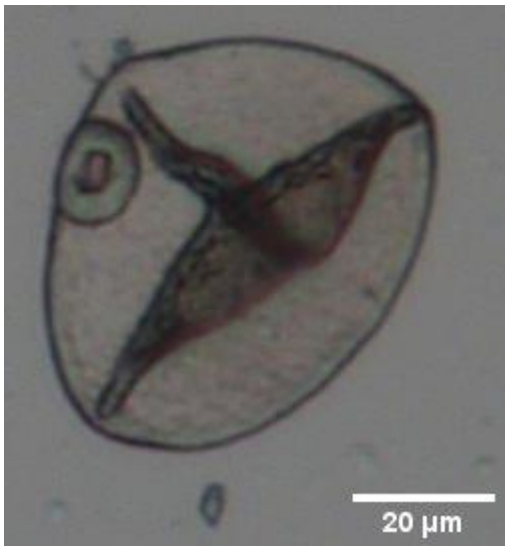
Secale cereale (acetolysed)

Sorghum halepense (untreated)

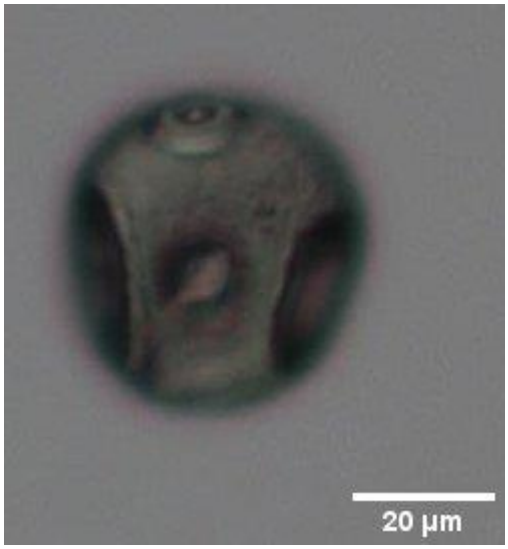


Sorghum halepense (acetolysed)

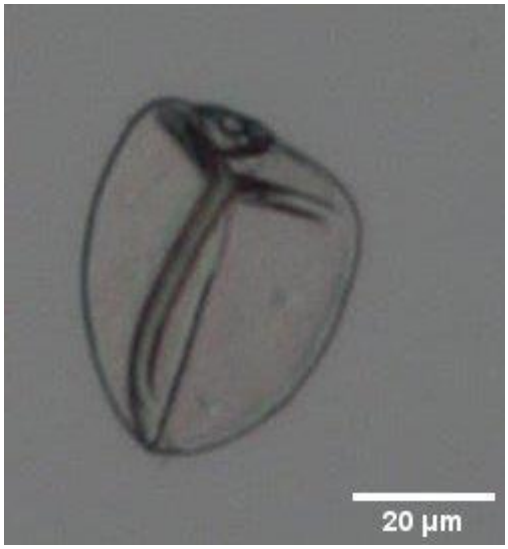
Triticum aestivum (untreated)



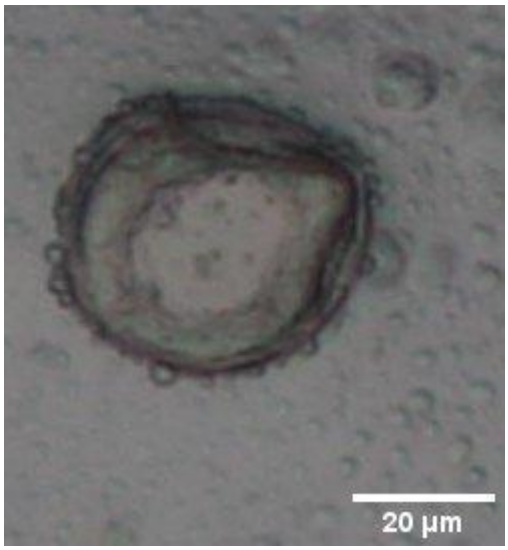
Triticum aestivum (acetolysed)



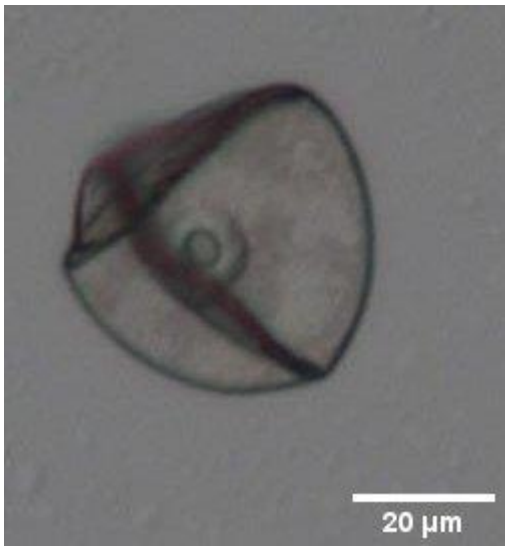
Triticum dicoccoides (untreated)



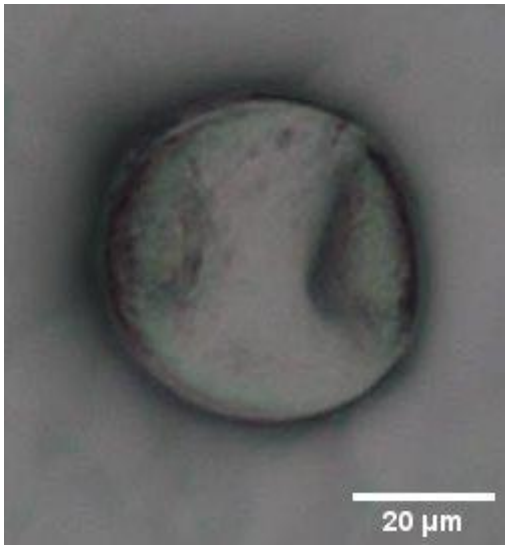
Triticum dicoccoides (acetolysed)



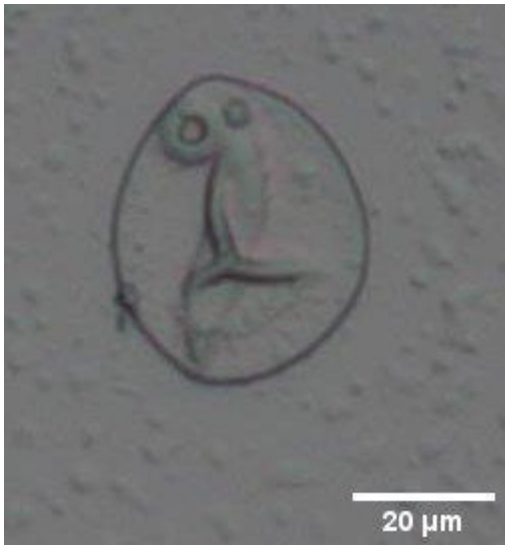
Triticum dicoccum (untreated)



Triticum dicoccum (acetolysed)

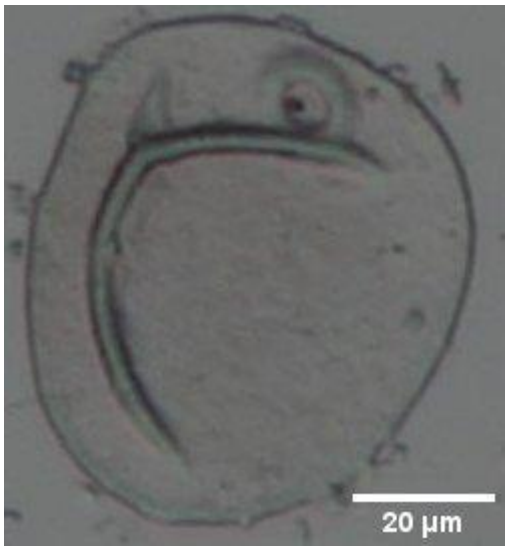


Triticum durum (untreated)

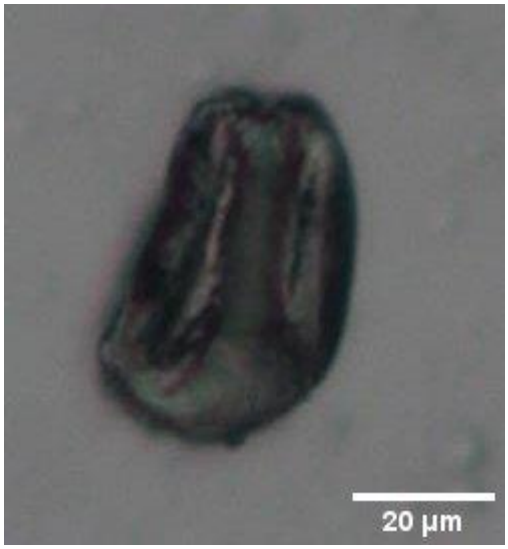


Triticum durum (acetolysed)

Triticum timopheevii (untreated)

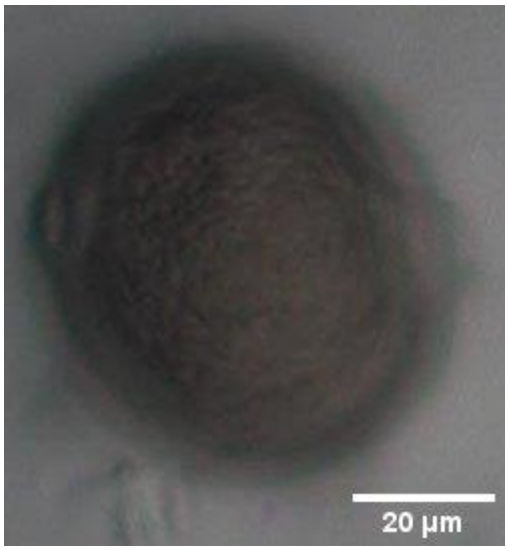


Triticum timopheevii (acetolysed)

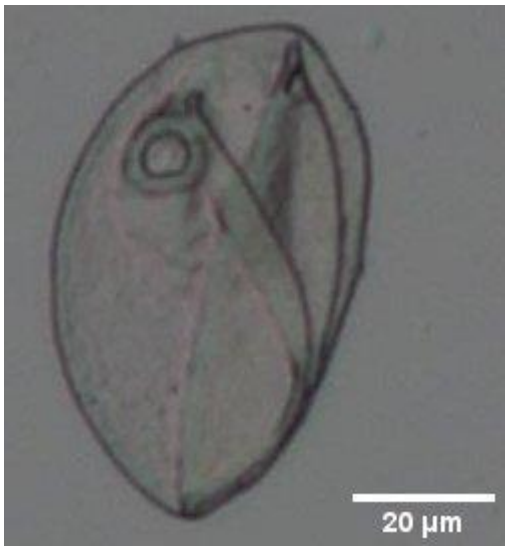


Triticum urartu (untreated)

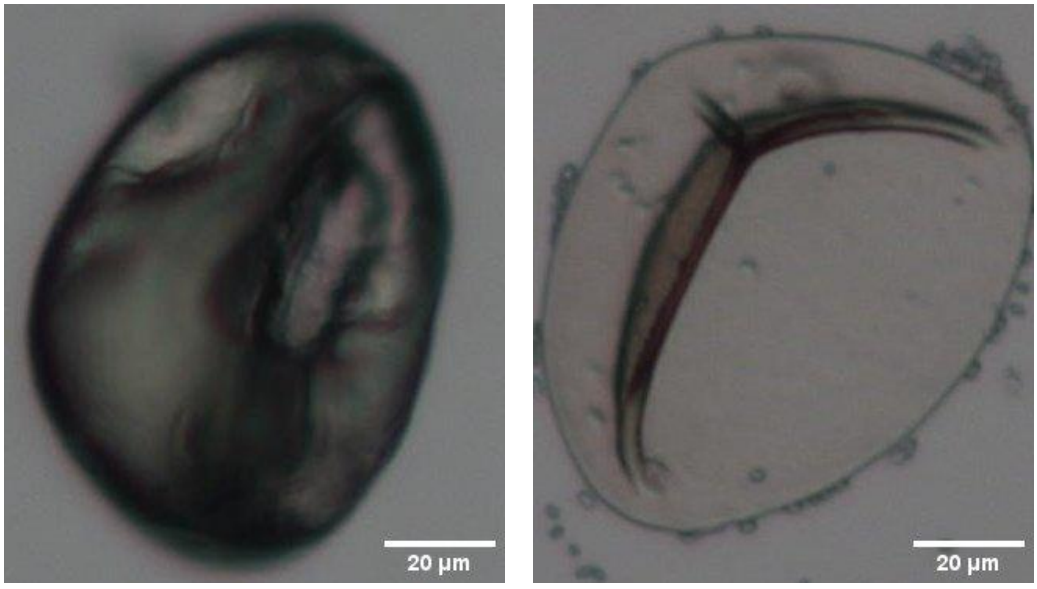
Triticum urartu (acetolysed)



Thinopyrum elongatum (untreated)

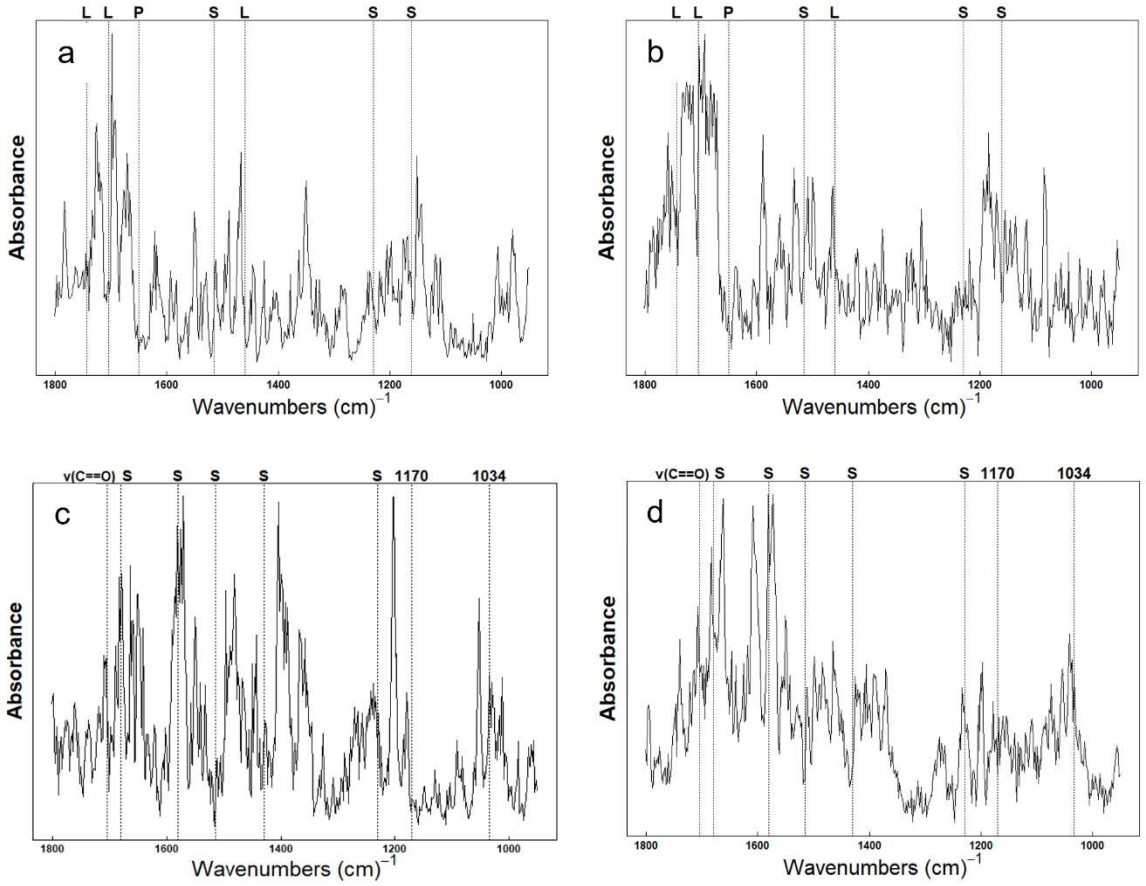


Thinopyrum elongatum (acetolysed)

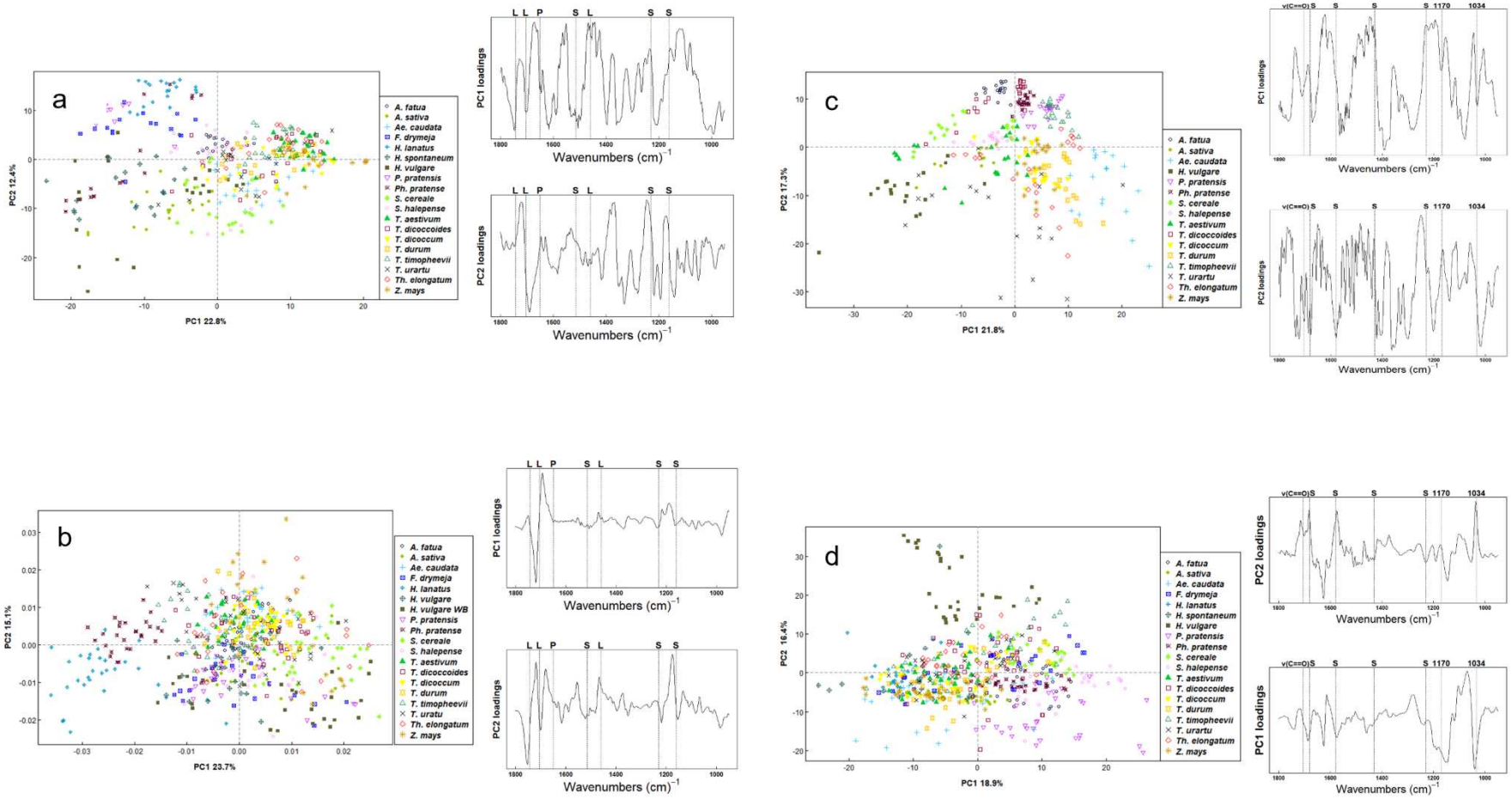


Zea mays (untreated) *Zea mays* (acetolysed)

S1 Light microscope images from Poaceae pollen used in this study. The left column includes the images of untreated pollen grains and the right column the acetolysed grains of the same species.



628 **S3** Wavenumbers' importance for RF classification of a) untreated populations, b)
untreated individual pollen grains, c) acetolysed populations and d) acetolysed
630 individual pollen grains.


a
PC2 12.4%
PC1 22.8%
A. fatua
A. sativa
Ae. caudata
F. drymeja
H. lanatus
H. spontaneum
H. vulgare
P. pratensis
Ph. pratense
S. cereale
S. halepense
T. aestivum
T. dicoccoides
T. dicoccum
T. durum
T. timopheevii
T. urartu
Th. elongatum
Z. mays
L L P S L S S
PC1 loadings
PC2 loadings
Wavenumbers (cm)$^{-1}$
b
PC2 15.1%
PC1 23.7%
H. vulgare WB
T. uratu
c
PC2 17.3%
PC1 21.8%
d
PC2 16.4%
PC1 18.9%
v(C=O) S S S S 1170 1034

632

S4 Principal components analysis (PCA) plots for first derivatives spectra from: a) untreated populations, b) untreated individual pollen grains, c) acetolysed populations, d) acetolysed individual pollen grains. For PCA analysis all wavenumbers of the fingerprint region (1800 cm^{-1} to 950 cm^{-1}) were used. The diagrams show PC1, PC2 and the percentage of variance explained by each principal component.

11. REFERENCES

638 Altieri, M.A., Nicholls, C.I., Henao, A., Lana, M.A., 2015. Agroecology and the design of climate change-resilient farming systems. Agron. Sustain. Dev. 35, 869–
640 890. https://doi.org/10.1007/s13593-015-0285-2

Andersen, S.T., 1979. Identification of wild grass and cereal pollen [fossil pollen,
642 Annulus diameter, surface sculpturing]. Aarbog. Danmarks Geologiske Undersoegelse (Denmark).

644 Andersen, T.S., Bertelsen, F., 1972. Scanning Electron Microscope Studies of Pollen of Cereals and other Grasses. Grana 12, 79–86.
646 https://doi.org/10.1080/00173137209428830

Anderson, T.W. (Theodore W., 2003. An introduction to multivariate statistical
648 analysis / T.W. Anderson., 3rd ed. ed, An introduction to multivariate statistical analysis, Wiley series in probability and mathematical statistics. Wiley-
650 Interscience, Hoboken, N.J.

Bağcıoğlu, M., Kohler, A., Seifert, S., Kneipp, J., Zimmermann, B., 2017. Monitoring
652 of plant–environment interactions by high-throughput FTIR spectroscopy of pollen. Methods in Ecology and Evolution 8, 870–880.
654 https://doi.org/10.1111/2041-210X.12697

Bağcıoğlu, M., Zimmermann, B., Kohler, A., 2015. A Multiscale Vibrational
656 Spectroscopic Approach for Identification and Biochemical Characterization of Pollen. PLOS ONE 10, e0137899.

Bassan, P., Byrne, H.J., Bonnier, F., Lee, J., Dumas, P., Gardner, P., 2009. Resonant Mie scattering in infrared spectroscopy of biological materials – understanding the 'dispersion artefact.' Analyst 134, 1586–1593. https://doi.org/10.1039/B904808A

Beug, H. J. 1961. *Leitfaden Der Pollenbestimmung*. Stuttgart: Gustav Fischer Verlag.

Bottema, S., 1992. Prehistoric cereal gathering and farming in the Near East: the pollen evidence. Review of Palaeobotany and Palynology, Festschrift For Professor Van Zeist 73, 21–33. https://doi.org/10.1016/0034-6667(92)90042-F

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

de Vareilles, A., Pelling, R., Woodbridge, J., Fyfe, R., 2021. Archaeology and agriculture: plants, people, and past land-use. Trends in Ecology & Evolution 36, 943–954. https://doi.org/10.1016/j.tree.2021.06.003

Dell'Anna, R., Lazzeri, P., Frisanco, M., Monti, F., Malvezzi Campeggi, F., Gottardini, E., Bersani, M., 2009. Pollen discrimination and classification by Fourier transform infrared (FT-IR) microspectroscopy and machine learning. Analytical and Bioanalytical Chemistry 394, 1443–1452.

Dickson, C., 1988. Distinguishing cereal from wild grass pollen: some limitations. Circaea 5, 67–71.

Diehn, S., Zimmermann, B., Tafintseva, V., Bağcıoğlu, M., Kohler, A., Ohlson, M., Fjellheim, S., Kneipp, J., 2020. Discrimination of grass pollen of different

species by FTIR spectroscopy of individual pollen grains. Anal Bioanal Chem. https://doi.org/10.1007/s00216-020-02628-2

Dodge, Y., 2008. Pooled Variance, in: The Concise Encyclopedia of Statistics. Springer New York, New York, NY, pp. 427–428. https://doi.org/10.1007/978-0-387-32833-1_323

Domínguez, E., Mercado, J.A., Quesada, M.A., Heredia, A., 1998. Isolation of intact pollen exine using anhydrous hydrogen fluoride. Grana 37, 93–96. https://doi.org/10.1080/00173139809362649

Eastwood, W.J., Fairbairn, A., Stroud, E., Roberts, N., Lamb, H., Yigitbas¸ ıoglu, H., S¸ enkul, Ç., Moss, A., Turner, R., Boyer, P., 2018. Comparing pollen and archaeobotanical data for Chalcolithic cereal agriculture at Çatalhoyük, Turkey. Quaternary Science Reviews 202, 4–18.

Fægri, K., Iversen, J., 1989. Textbook of pollen analysis, 4th ed. ed. Wiley, Chichester.

Fraser, W.T., Scott, A.C., Forbes, A.E.S., Glasspool, I.J., Plotnick, R.E., Kenig, F., Lomax, B.H., 2012. Evolutionary stasis of sporopollenin biochemistry revealed by unaltered Pennsylvanian spores. New Phytologist 196, 397–401. https://doi.org/10.1111/j.1469-8137.2012.04301.x

Fuller, D.Q., 2007. Contrasting Patterns in Crop Domestication and Domestication Rates: Recent Archaeobotanical Insights from the Old World. Annals of Botany 100, 903–924.

Fuller, D.Q., Lucas, L., 2014. Archaeobotany, in: Smith, C. (Ed.), Encyclopedia of Global Archaeology. Springer, New York, NY, pp. 305–310. https://doi.org/10.1007/978-1-4419-0465-2_2273

Hapsari, K.A., Ballauff, J., 2022. Distinguishing pollen grains of cereal from wild grasses in the Sundaland region using size separation. Review of Palaeobotany and Palynology 301, 104648. https://doi.org/10.1016/j.revpalbo.2022.104648

Heslop-Harrison, J. 1979. Aspects of the Structure, Cytochemistry and Germination of the Pollen of Rye. Annals of Botany, 44 (Supplement 1), Oxford University Press

Grohne, U. (1957) Die Bedeutung des Phasenkontrastverfahrens für die Pollenanalyse am Beispiel der Gramineenpollen vom Getreidetyp. Photograph. Forsch. 7 (8): 237-248

Jardine, P.E., Fraser, W.T., Lomax, B.H., Gosling, W.D., 2015. The impact of oxidation on spore and pollen chemistry. Journal of Micropalaeontology 34, 139–149. https://doi.org/10.1144/jmpaleo2014-022

Jardine, P.E., Fraser, W.T., Lomax, B.H., Sephton, M.A., Shanahan, T.M., Miller, C.S., Gosling, W.D., 2016. Pollen and spores as biological recorders of past ultraviolet irradiance. Sci Rep 6, 39269. https://doi.org/10.1038/srep39269

Jardine, P.E., Gosling, W.D., Lomax, B.H., Julier, A.C.M., Fraser, W.T., 2019. Chemotaxonomy of domesticated grasses: a pathway to understanding the origins of agriculture. Journal of Micropalaeontology 38, 83–95.

722 Jardine, P.E., Hoorn, C., Beer, M.A.M., Barbolini, N., Woutersen, A., Bogota-Angel, G., Gosling, W.D., Fraser, W.T., Lomax, B.H., Huang, H., Sciumbata, M., He,
724 H., Dupont-Nivet, G., 2021. Sporopollenin chemistry and its durability in the geological record: an integration of extant and fossil chemical data across the
726 seed plants. Palaeontology 64, 285–305. https://doi.org/10.1111/pala.12523

Joly, C., Barillé, L., Barreau, M., Mancheron, A., Visset, L., 2007. Grain and annulus
728 diameter as criteria for distinguishing pollen grains of cereals from wild grasses. Review of Palaeobotany and Palynology 146, 221–233.
730 https://doi.org/10.1016/j.revpalbo.2007.04.003

Jones, M.K., 1985. Archaeobotany beyond subsistence reconstruction, in: Baker,
732 G.W., Gamble, C. (Eds.), Beyond Domestication in Prehistoric Europe. Academic Press, New York, pp. 107–128.

734 Julier, A.C.M., Jardine, P.E., Coe, A.L., Gosling, W.D., Lomax, B.H., Fraser, W.T., 2016. Chemotaxonomy as a tool for interpreting the cryptic diversity of
736 Poaceae pollen. Review of Palaeobotany and Palynology 235, 140–147.

Köhler, E., Lange, E., 1979. A contribution to distinguishing cereal from wild grass
738 pollen grains by LM and SEM. Grana 18, 133–140. https://doi.org/10.1080/00173137909424973

740 Kuhn, M., 2019. caret: Classification and Regression Training.

Lee, L.C., Liong, C.-Y., Jemain, A.A., 2018. Partial least squares-discriminant
742 analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. Analyst 143, 3526–
744 3539. https://doi.org/10.1039/C8AN00599K

Li , F.S., Phyo, P., Jacobowitz, J., Hong, M. and Weng, J. K. 2019. The molecular structure of plant sporopollenin. Nature Plants, 5, 41–46.

Li, Y., Zhou, L., Cui, H., 2008. Pollen indicators of human activity. Chin. Sci. Bull. 53, 1281–1293. https://doi.org/10.1007/s11434-008-0181-0

Liland, K.H., 2017. EMSC: Extended Multiplicative Signal Correction.

Lutzke, A., Morey, K.J., Medford, J.I., Kipper, M.J., 2020. Detailed characterization of Pinus ponderosa sporopollenin by infrared spectroscopy. Phytochemistry 170, 112195. https://doi.org/10.1016/j.phytochem.2019.112195

Mander, L., Li, M., Mio, W., Fowlkes, C.C., Punyasena, S.W., 2013. Classification of grass pollen through the quantitative analysis of surface ornamentation and texture. Proceedings of the Royal Society B: Biological Sciences 280, 20131905. https://doi.org/10.1098/rspb.2013.1905

Marquer, L., Gaillard, M.-J., Sugita, S., Poska, A., Trodman, A.-K., Mazier, F., Nielsen, A.B., Fyfe, R., Jönsson, A.M., Smith, B., Kaplan, J.O., Alenius, T., Birks, J.H., Bjune, A.E., Christiansen, J., Dodson, J., Edwards, K.J., Giesecke, T., Herzschuh, U., Kangur, M., Seppä, H., 2017. Quantifying the effects of land use and climate on Holocene vegetation in Europe. Quaternary Science Reviews 171, 20–37.

Marston, J.M., 2021. Archaeological Approaches to Agricultural Economies. Journal of Archaeological Research, https://doi.org/10.1007/s10814-020-09150-0 29, 327–385.

766 Morrison, K.D., Hammer, E., Popova, L., Madella, M., Whitehouse, N., Gaillard, M.-J., LandCover6k Land-Use Group Members, 2018. Global-scale comparisons
768 of human land use: developing shared terminology for land-use practices for global change | PAGES. Past Global Change Magazine 26, 8–9.

770 Murphy, K.P., 2012. Machine Learning: A probabilistic perspective. MIT Press, London, UNITED KINGDOM.

772 Muthreich, F., Zimmermann, B., Birks, H.J.B., Vila-Viçosa, C.M., Seddon, A.W.R., 2020. Chemical variations in Quercus pollen as a tool for taxonomic
774 identification: Implications for long-term ecological and biogeographical research. Journal of Biogeography 47, 1298–1309.
776 https://doi.org/10.1111/jbi.13817

Nierop, K. G. J., Versteegh, G. J. M., Filley, T. R. and de Leeuw, J. W. 2019.
778 Quantitative analysis of diverse sporomorph-derived sporopollenins. Phytochemistry, 162, 207–215.

780 Pappas, C.S., Tarantilis, P.A., Harizanis, C., Polissiou, M.G., 2003. New Method for Pollen Identification by FT-IR Spectroscopy. Applied Spectroscopy 57.

782 Pedersen, T.L., Crameri, F., 2020. scico: Colour Palettes Based on the Scientific Colour-Maps.

784 Piperno, D.R., 2011. The Origins of Plant Cultivation and Domestication in the New World Tropics: Patterns, Process, and New Developments. Current
786 Anthropology 52, S453–S470. https://doi.org/10.1086/659998

Radziwill, N.M., 2017. Statistics (the easier way) with R: An informal text on applied statistics and data science, 2nd ed. Lapis Lucera.

Riehl, S., Asouti, E., Karakaya, D., Starkovich, B.M., Zeidi, M., Conard, N.J., 2015. Resilience at the Transition to Agriculture: The Long-Term Landscape and Resource Development at the Aceramic Neolithic Tell Site of Chogha Golan (Iran). BioMed Research International 2015, 532481. https://doi.org/10.1155/2015/532481

Riehl, S., Pustovoytov, K.E., Weippert, H., Klett, S., Hole, F., 2014. Drought stress variability in ancient Near Eastern agricultural systems evidenced by δ13C in barley grain. Proceedings of the National Academy of Sciences 111, 12348–12353. https://doi.org/10.1073/pnas.1409516111

Rinnan, Å., Berg, F. van den, Engelsen, S.B., 2009. Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends in Analytical Chemistry 28, 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007

Roberts, N., 2015. Revisiting the Beyşehir Occupation Phase: Land-Cover Change and the Rural Economy in the Eastern Mediterranean During the First Millennium AD. Late Antique Archaeology 11, 53–68. https://doi.org/10.1163/22134522-12340052

Roberts, N., 2002. Did prehistoric landscape management retard the post-glacial spread of woodland in Southwest Asia? Antiquity 76, 1002–1010. https://doi.org/10.1017/S0003598X0009181X

Rowley, J.R., 1960. The Exine Structure of “Cereal” and “Wild” Type Grass Pollen. Grana Palynologica 2, 9–15. https://doi.org/10.1080/00173136009429441

Salih, A., Jones, A.S., Bass, D., Cox, G., 1997. Confocal imaging of exine as a tool for grass pollen analysis. Grana 36, 215–224. https://doi.org/10.1080/00173139709362610

Schüler, L., Behling, H., 2011. Poaceae pollen grain size as a tool to distinguish past grasslands in South America: a new methodological approach. Vegetation History and Archaeobotany 20, 83–96. https://doi.org/10.1007/s00334-010-0265-z

Singh, A., Thakur, N., Sharma, A., 2016. A review of supervised machine learning algorithms. Presented at the 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, pp. 1310–1315.

Sobol, M.K., Finkelstein, S.A., 2018. Predictive pollen-based biome modeling using machine learning. PLoS One 13, e0202214. https://doi.org/10.1371/journal.pone.0202214

Taiyun, W., Viliam, S., 2017. R package “corrplot”: Visualization of a Correlation Matrix.

Trondman, A.-K., Gaillard, M.-J., Mazier, F., Sugita, S., Fyfe, R., Nielsen, A.B., Twiddle, C., Barratt, P., Birks, H.J.B., Bjune, A.E., Björkman, L., Broström, A., Caseldine, C., David, R., Dodson, J., Dörfler, W., Fischer, E., van Geel, B., Giesecke, T., Hultberg, T., Kalnina, L., Kangur, M., van der Knaap, P., Koff, T., Kuneš, P., Lagerås, P., Latałowa, M., Lechterbeck, J., Leroyer, C., Leydet,

M., Lindbladh, M., Marquer, L., Mitchell, F.J.G., Odgaard, B.V., Peglar, S.M., Persson, T., Poska, A., Rösch, M., Seppä, H., Veski, S., Wick, L., 2015. Pollen-based quantitative reconstructions of Holocene regional vegetation cover (plant-functional types and land-cover types) in Europe suitable for climate modelling. Global Change Biology 21, 676–697. https://doi.org/10.1111/gcb.12737

Wang, T., Bell, B.A., Fletcher, W.J., Ryan, P.A., Wogelius, R.A., 2023. Influence of common palynological extraction treatments on ultraviolet absorbing compounds (UACs) in sub-fossil pollen and spores observed in FTIR spectra. Frontiers in Ecology and Evolution 11, 1096099. https://doi.org/10.3389/fevo.2023.1096099

Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., Venables, B., 2020. gplots: Various R Programming Tools for Plotting Data.

Watson, J.S., Septhon, M.A., Sephton, S.V., Self, S., Fraser, W.T., Lomax, B.H., Gilmour, I., Wellman, C.H. And Beerling, D.J. 2007. Rapid determination of spore chemistry using thermochemolysis gas chromatography-mass spectrometry and micro-Fourier transform infrared spectroscopy. Photochemical & Photobiological Sciences, 6, 689–694.

Wei, C.-X., Jardine, P. E., Mao, L.-M., Mander, L., Li, M., Gosling, W. D., & Hoorn, C. (2023). Grass pollen surface ornamentation is diverse across the phylogeny: Evidence from northern South America and the global literature. Journal of Systematics and Evolution, n/a(n/a), Advance online publication. https://doi.org/10.1111/jse.13021

856 Williams, J.W., Grimm, E.C., Blois, J.L., Charles, D.F., Davis, E.B., Goring, S.J., Graham, R.W., Smith, A.J., Anderson, M., Arroyo-Cabrales, J., Ashworth,
858 A.C., Betancourt, J.L., Bills, B.W., Booth, R.K., Buckland, P.I., Curry, B.B., Giesecke, T., Jackson, S.T., Latorre, C., Nichols, J., Purdum, T., Roth, R.E.,
860 Stryker, M., Takahara, H., 2018. The Neotoma Paleoecology Database, a multiproxy, international, community-curated data resource. Quaternary
862 Research 89, 156–177. https://doi.org/10.1017/qua.2017.105

Woutersen, A., Jardine, P.E., Bogotá-Angel, R.G., Zhang, H.-X., Silvestro, D.,
864 Antonelli, A., Gogna, E., Erkens, R.H.J., Gosling, W.D., Dupont-Nivet, G., Hoorn, C., 2018. A novel approach to study the morphology and chemistry of
866 pollen in a phylogenetic context, applied to the halophytic taxon L.(Nitrariaceae). PeerJ 6, e5055. https://doi.org/10.7717/peerj.5055

868 Ziegler, A., König, I.R., 2014. Mining data with random forests: current options for real-world applications. WIREs Data Mining and Knowledge Discovery 4, 55–
870 63. https://doi.org/10.1002/widm.1114

Zimmermann, B., 2018. Chemical characterization and identification of Pinaceae
872 pollen by infrared microspectroscopy. Planta. An International Journal of Plant Biology 247, 171–180.

874 Zimmermann, B., 2010. Characterization of Pollen by Vibrational Spectroscopy. Appl Spectrosc 64, 1364–1373. https://doi.org/10.1366/000370210793561664

876 Zimmermann, B., Bağcıoğlu, M., Sandt, C., Kohler, A., 2015. Vibrational microspectroscopy enables chemical characterization of single pollen grains
878 as well as comparative analysis of plant species based on pollen

ultrastructure. Planta 242, 1237–1250. https://doi.org/10.1007/s00425-015-2380-7

Zimmermann, B., Bağcıoğlu, M., Tafinstseva, V., Kohler, A., Ohlson, M., Fjellheim, S., 2017. A high-throughput FTIR spectroscopy approach to assess adaptive variation in the chemical composition of pollen. Ecology and Evolution 7, 10839–10849. https://doi.org/10.1002/ece3.3619

Zimmermann, B., Kohler, A., 2013. Optimizing Savitzky–Golay Parameters for Improving Spectral Resolution and Quantification in Infrared Spectroscopy: Applied Spectroscopy. https://doi.org/10.1366/12-06723

Zimmermann, B., Tafinsteva, V., Bağcıoğlu, M., Berdahl, M.H., Kohler, A., 2016. Analysis of allergenic pollen by FTIR Microspectroscopy. Analytical Chemistry 88, 803–811.