Application Article

Accuracy and consistency of grass pollen identification by human analysts using electron micrographs of surface ornamentation[1]

Luke Mander[2,6], Sarah J. Baker[2], Claire M. Belcher[2], Derek S. Haselhorst[3], Jacklyn Rodriguez[4], Jessica L. Thorn[2], Shivangi Tiwari[5], Dunia H. Urrego[2], Cassandra J. Wesseln[3], and Surangi W. Punyasena[4]

[2]College of Life and Environmental Sciences, University of Exeter, Prince of Wales Road, Exeter, Devon EX4 4PS, United Kingdom; [3]Program in Ecology, Evolution, and Conservation Biology, University of Illinois, 505 South Goodwin Avenue, Urbana, Illinois 61801 USA; [4]Department of Plant Biology, University of Illinois, 505 South Goodwin Avenue, Urbana, Illinois 61801 USA; and [5]Department of Earth Sciences, Indian Institute of Technology Roorkee, Roorkee 247667, Uttarakhand, India

- *Premise of the study:* Humans frequently identify pollen grains at a taxonomic rank above species. Grass pollen is a classic case of this situation, which has led to the development of computational methods for identifying grass pollen species. This paper aims to provide context for these computational methods by quantifying the accuracy and consistency of human identification.
- *Methods:* We measured the ability of nine human analysts to identify 12 species of grass pollen using scanning electron microscopy images. These are the same images that were used in computational identifications. We have measured the coverage, accuracy, and consistency of each analyst, and investigated their ability to recognize duplicate images.
- *Results:* Coverage ranged from 87.5% to 100%. Mean identification accuracy ranged from 46.67% to 87.5%. The identification consistency of each analyst ranged from 32.5% to 87.5%, and each of the nine analysts produced considerably different identification schemes. The proportion of duplicate image pairs that were missed ranged from 6.25% to 58.33%.
- *Discussion:* The identification errors made by each analyst, which result in a decline in accuracy and consistency, are likely related to psychological factors such as the limited capacity of human memory, fatigue and boredom, recency effects, and positivity bias.

Key words: automation; classification; expert analysis; identification; palynology.

Fossil pollen grains are a valuable empirical record of the history of plant life on Earth. They are used to investigate a broad range of questions in plant evolution and paleoecology (e.g., Birks and Birks, 1980), and are used by the hydrocarbon exploration industry to date and correlate sedimentary rocks (Traverse, 2007; Punyasena et al., 2012a). Fossil pollen grains are identified based on aspects of their morphology (e.g., Traverse, 2007; Punt et al., 2007), and to extract the maximum amount of evolutionary, paleoecological, or biostratigraphic information from an assemblage of fossil pollen grains, researchers generally aim to identify pollen grains at the species level. In many cases, however, species-level identification of pollen grains is not possible, and researchers default to identifications at relatively low taxonomic ranks such as the genus or family level to ensure that their identifications are reproducible by other workers (Punyasena et al., 2012b). In such situations, the fossil pollen record is said to suffer from low taxonomic resolution, which presents a major barrier to the accurate reconstruction of vegetation history (Birks and Birks, 2000; Jackson and Booth, 2007; Mander, 2011; Punyasena et al., 2011, 2012b; May and Lacourse, 2012; Mander et al., 2013).

Grass pollen is a classic case of low taxonomic resolution, and is seldom identified below the family level in routine palynological studies that use fossil pollen grains to reconstruct vegetation history (Strömberg, 2011). As a result, most of the fossil evidence for the evolutionary and ecological history of grasses (members of the Poaceae family) has been provided either by molecular phylogenetic methods (Edwards et al., 2010; Grass Phylogeny Working Group II, 2012) or from the fossil record of phytoliths (microscopic silica bodies that form in plant tissues), which can be used to identify grasses to a much finer taxonomic resolution than pollen grains (up to genus level; Piperno, 2006; Strömberg, 2011). Nevertheless, fossil grass pollen grains are a potentially rich source of information on the evolutionary and ecological history of grasses because of their wide dispersal, production in large numbers, and excellent preservation potential in most depositional settings apart from very oxidative environments. Consequently, researchers have made several attempts to increase the taxonomic resolution of

[1]Manuscript received 18 March 2014; revision accepted 3 June 2014.

The authors thank two anonymous referees and Kat Holt for helpful comments on this study. We acknowledge funding from the National Science Foundation (DBI-1052997 and DBI-1262561 to S.W.P.) and a Marie Curie International Incoming Fellowship within the 7th European Community Framework Program (PIIF-GA-2012-328245 to L.M.).

[6]Author for correspondence: luke.mander@gmail.com

doi:10.3732/apps.1400031

the grass pollen fossil record. These include morphometric approaches to identify taxa based on the size and shape of characters such as the entire pollen grain, the pore and the annulus (e.g., Andersen, 1979; Tweddle et al., 2005; Joly et al., 2007; Schüler and Behling, 2011a, b), phase-contrast microscopy to identify taxa based on aspects of the organization of the grass pollen exine (Fægri et al., 1992; Beug, 2004; Holst et al., 2007), and scanning electron microscopy (SEM) to identify taxa based on the patterns of surface ornamentation (Andersen and Bertelsen, 1972; Page, 1978; Peltre et al., 1987; Chaturvedi et al., 1998; Mander et al., 2013).

The most recent of these attempts employed a combination of high-resolution imaging (using SEM) and computational image analysis to identify 12 species of extant grass pollen based on the size and shape of sculptural elements on the pollen surface and the complexity of the ornamentation patterns they form (Mander et al., 2013). This approach differs from most routine palynological work in that it involves investigating and comparing detailed portions of the surface of individual pollen grains, rather than identifying pollen grains by viewing entire specimens using brightfield microscopy, and resulted in a species-level identification accuracy of 77.5% (Mander et al., 2013). By way of comparison, seven human analysts identified the same SEM images of grass pollen surface ornamentation with accuracies ranging from 68.33% to 81.67% (Mander et al., 2013). However, these seven analysts only analyzed one set of images, and as a result their self-consistency was not measured. This is problematic because low self-consistency, which is the degree to which an analyst makes identifications that are consistent with their own previous identifications (MacLeod et al., 2010), is cited as a primary reason to support the development of computational identification methods instead of manual identifications by human analysts (e.g., Culverhouse et al., 2003; Culverhouse, 2007; MacLeod et al., 2010).

In the present paper, we address this issue by testing the ability of nine human analysts to identify the pollen of 12 species of grass using SEM images of surface ornamentation. This study builds on the preliminary investigation of Mander et al. (2013) and has the following specific aims: (1) to measure the identification accuracy of the nine analysts; and (2) to measure the consistency of the identification produced by each analyst. An overarching goal of this work is to provide context for the errors produced by computational methods of identifying grass pollen and to explore whether identification of grass pollen by human analysts in future work may provide reliable records of ancient grass diversity.

MATERIALS AND METHODS

We used the image library of SEM images of the pollen of 12 grass species generated by Mander et al. (2013) as the raw material for our study (Fig. 1). This library contains SEM images of 20 specimens of each grass species. These images were acquired by mounting specimens of pollen from each species onto separate SEM stubs, coating them with gold-palladium using a sputter coater, and imaging them at 2000×, 6000×, and 12,000× magnification using a JEOL JSM-6060-LV SEM (JEOL USA, Peabody, Massachusetts, USA) at 15 kV (Mander et al., 2013). In this study, we have used 400 × 400-pixel windows that were manually cropped from 6000× SEM images (Fig. 1). These are the same images that were used to develop algorithmic identifications of grass pollen by Mander et al. (2013). From this image library, we generated a training set of five SEM images of each species that were correctly classified and labeled. We also generated two test sets each containing 120 unidentified SEM images of grass pollen that were then manually identified by nine human analysts. We have used nine analysts because this was the number of people who agreed to participate in this work. One of the analysts (L.M.; Analyst 7) also analyzed the data. This should be borne in mind when interpreting the results of this study because this analyst may have gained an advantage through greater familiarity with the images. However, in the context of the performance of all the analysts who participated in this study, any advantage is not immediately apparent in the identification accuracy and consistency of this analyst. In this paper, we follow the terminology of Sokal (1974), in which classification is defined as the ordering of objects into groups on the basis of their relationships, and identification is defined as the assignment of additional unidentified objects to the correct class.

The test sets contained 10 images of each of the 12 species. Both test sets contained the same images of the same species, and each image was engraved with a unique number (Fig. 1). Of the 240 SEM images in the two test sets combined, 48 were duplicate pairs. This arbitrary number of duplicate pairs was generated by randomly selecting two specimens of each species as duplicates. However, no identical images were present in both the training and the test sets, which ensured that the material identified by each analyst was independent of the material used for learning.

The training set and the test sets were transmitted to the nine analysts electronically. The analysts were told that each test set contained 10 specimens of each species, and were instructed that each species should be represented by no more than 10 images in their identification scheme. The analysts were instructed not to guess at the taxonomic affinity of an image and to construct a list of images that were left unidentified. Identification was performed by comparing each image in the test set with the images in the training set, and listing the unique number engraved into each unknown image next to the appropriate taxon in a spreadsheet. Identification of images in the test sets was undertaken in two rounds. The second data set was transmitted to the analysts one month after the first, and after the analysts had completed the first identification round. The analysts did not receive any feedback on their performance after the first classification round. Analysts were instructed to record their reasons for each of their identifications in both the first and second identification rounds, and could use either technical (e.g., Punt et al., 2007) or nontechnical language to do so.

Each analyst was instructed to place themselves into one of four groups based on their level of experience identifying pollen grains or any other microscopic objects that involve identification based on morphology (Table 1). These groups were as follows: (i) *Novice* (analyst has up to one month of experience studying pollen grains or any other microscopic objects that involve identification based on morphology); (ii) *Intermediate* (analyst has between one month and one year of experience); *(*iii) *Expert* (analyst has over one year of experience, but does not yet hold a PhD in palynology, or a PhD that involves the identification of microscopic objects using morphological criteria); and (iv) *Professional* (analyst holds a PhD in palynology, or a PhD that involved the identification of microscopic objects using morphological criteria). The unequal distribution of analyst experience is a consequence of the small, available pool of participants.

We then examined the identification performance of the nine analysts by measuring the coverage, accuracy, and consistency of their identifications. Coverage was measured by calculating the proportion of images in each test set that each analyst attempted to identify (Kohavi and Provost, 1998), which provides a baseline measure of analyst confidence. Accuracy was measured by calculating the proportion of all images in each test set that were identified correctly, with images left unidentified treated as errors. Identification consistency was measured using two metrics. Metric one was generated by calculating the proportion of images that were identified as the same taxon in both identification rounds irrespective of whether the identification was correct or not. Metric two was generated by calculating the proportion of images that were correctly identified as the same taxon in both identification rounds. We also investigated the ability of each analyst to recognize duplicate images by measuring the proportion of duplicate image pairs that were split by misidentification in the two combined test sets. These metrics are summarized in Table 2.

RESULTS

Coverage ranged from 87.5% (analyst 1 in round one) to 100% (analyst 9 in round two) (Table 1). Five analysts increased their coverage from the first to the second round, two analysts decreased their coverage from the first to the second round, and the coverage of two analysts remained the same in both rounds (Table 1). Averaged across both identification rounds, all analysts attempted to identify at least 90% of the images presented to them (Table 1).

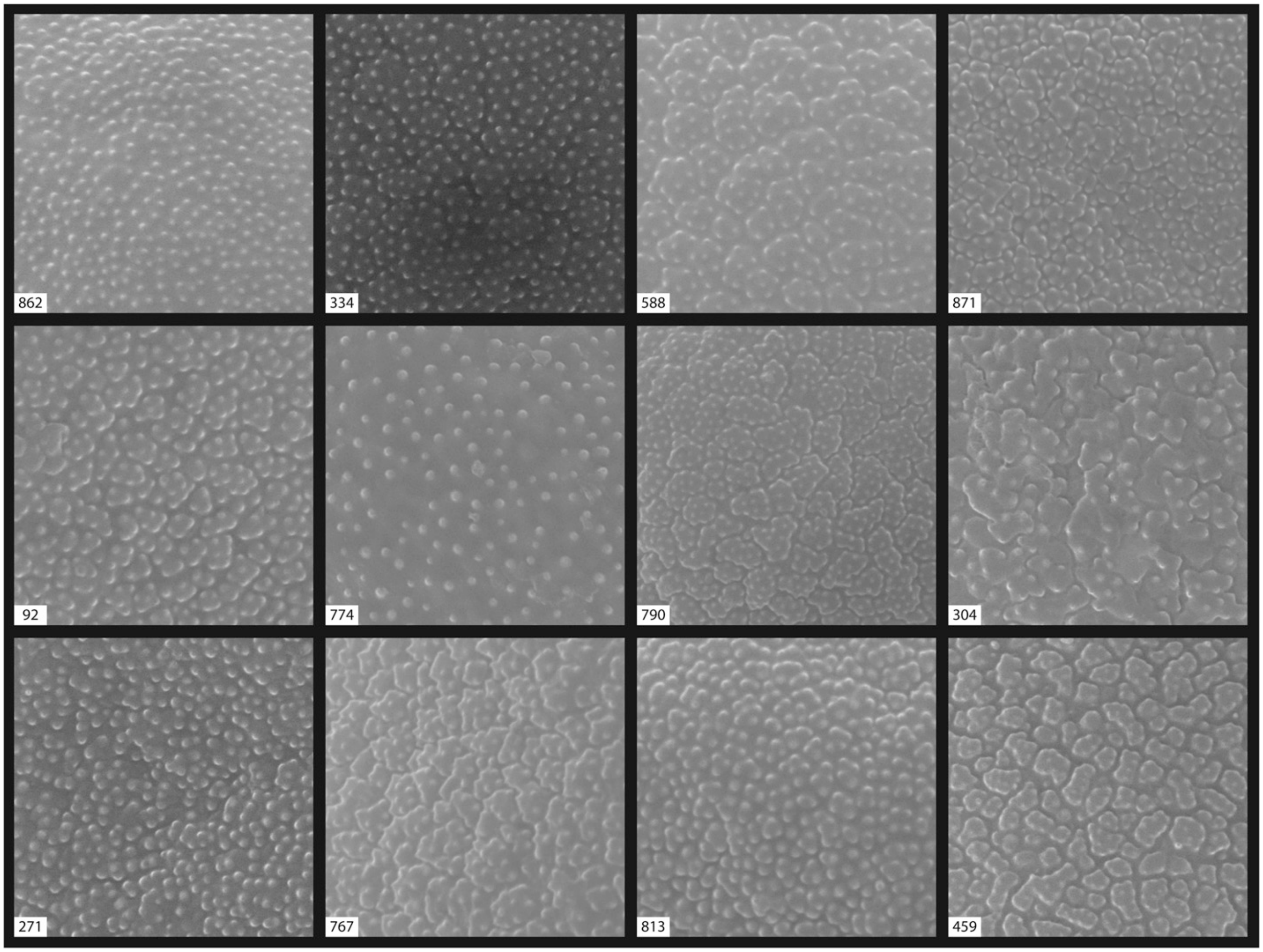


Fig. 1. Example SEM images showing a portion of the surface of a pollen grain from each grass species used in the identification experiment described in this paper. Species identified by the engraved unique number as follows: *Anthoxanthum odoratum* L. (862), *Dactylis glomerata* L. (92), *Phalaris arundinacea* L. (271), *Poa australis* R. Br. (334), *Stipa tenuifolia* Steud. (774), *Cynodon dactylon* (L.) Pers. (767), *Eragrostis mexicana* (Hornem.) Link (588), *Sporobolus pyramidalis* P. Beauv. (790), *Triodia basedowii* Pritz. (813), *Bothriochloa intermedia* (R. Br.) A. Camus (871), *Digitaria insularis* (L.) Fedde (304), *Oplismenus hirtellus* (L.) P. Beauv. (459).

Identification accuracy ranged from 36.67% (analyst 1 in round one) to 90% (analyst 9 in round two) (Table 1, Fig. 2A), and mean accuracy averaged over the two identification rounds ranged from 46.67% to 87.5% (Table 1, Fig. 2B). The identification accuracy of six analysts increased from round one to round two, and the accuracy of three analysts decreased (Table 1,

Table 1. Coverage, accuracy, and consistency (reported as percentages) of grass pollen identification by the nine analysts in this study.

		Coverage			Accuracy			Consistency		
Analyst[a]	Experience[b]	Round 1	Round 2	Mean	Round 1	Round 2	Mean	Measure 1	Measure 2	Duplicates split
1	Nov.	87.50	92.50	90.00	36.67	56.67	46.67	32.50	22.50	58.33
2	Nov.	99.17	99.17	99.17	61.67	59.17	60.42	51.67	41.67	37.50
3	Prof.	90.83	95.00	92.92	53.33	67.50	60.42	48.33	41.67	47.92
4	Prof.	90.00	91.67	90.83	78.33	71.67	75.00	65.83	61.67	35.42
5	Int.	98.33	95.83	97.08	73.33	77.50	75.42	65.83	60.83	29.17
6	Int.	96.67	98.33	97.50	79.17	80.00	79.58	73.33	68.33	29.17
7	Prof.	97.50	97.50	97.50	80.83	85.00	82.92	72.50	70.00	29.17
8	Ex.	95.83	95.00	95.42	87.50	84.17	85.83	81.67	77.50	20.83
9	Ex.	99.17	100.00	99.58	89.17	90.00	89.58	87.50	84.17	6.25

[a] Analysts ordered by their mean classification accuracy.
[b] Experience levels: Nov. = Novice; Int. = Intermediate; Ex. = Expert; Prof. = Professional.

Table 2. Summary of the measures used to examine the identification performance of the nine analysts.

Measure	Explanation
Coverage	Proportion of images that each analyst attempted to identify
Accuracy	Proportion of all images in each test set that were identified correctly. Unidentified images are treated as errors.
Consistency metric 1	Proportion of images that were identified as the same taxon in both identification rounds irrespective of whether the identification was correct or not
Consistency metric 2	Proportion of images that were correctly identified as the same taxon in both identification rounds
Duplicate images	Proportion of duplicate image pairs that were split by misidentification

Fig. 2A). The largest increases in identification accuracy were by analysts 1 and 3, whose accuracy increased by 20% and 14.17%, respectively (Table 1). Analysts 1, 2, and 3 had markedly lower mean accuracies than the other analysts (Fig. 2B). Analysts 1 and 2 placed themselves into the Novice category, and analyst three placed themselves into the Professional category (Table 1).

Using metric one, the identification consistency of each analyst ranged from 32.5% to 87.5% (Table 1, Fig. 3A). Using metric two, the identification consistency of each analyst ranged from 22.5% to 84.17% (Table 1, Fig. 3A). There is a positive relationship between mean identification accuracy and identification consistency using both metric one (Fig. 4A) and metric two (Fig. 4B). The proportion of duplicate image pairs that were split by misidentification varied widely between analysts. For example, analyst 1 split 58.33% of the image pairs, but analyst 9 split just 6.25% of these images (Table 1, Fig. 3B). The proportion of duplicate image pairs split by the other seven analysts ranges from 20.83% to 47.92%, and three analysts each split 29.17% of these image pairs (Table 1, Fig. 3B). There is a negative relationship between mean identification accuracy and the proportion of duplicate image pairs split by misidentification (Fig. 4C).

Each of the nine analysts produced different identification schemes, and this is highlighted by error matrices showing the identification errors made by each analyst in identification round two (Figs. 5–7). The identifications of analysts 1, 2, and 3 are characterized by numerous and widely scattered errors that differ considerably from one another, and typically there is confusion between two and three species, and occasionally between four and six other species (Fig. 5). For example, of the 10 specimens that analyst 2 identified as *Eragrostis mexicana* (Hornem.) Link, five were correct, but the other five specimens were each confused with a different species (Fig. 5B). The identifications of analysts 7, 8, and 9 are characterized by far fewer errors, but each of these analysts makes different identification errors (Fig. 7). For example, of the 10 specimens assigned to *Triodia basedowii* Pritz. by analyst 7, one was actually *Bothriochloa intermedia* (R. Br.) A. Camus and one was actually *Phalaris arundinacea* L. (Fig. 7A), but of the 10 specimens assigned to *T. basedowii* by analyst 8, one was actually *B. intermedia* and two were *Dactylis glomerata* L. (Fig. 7B). Similarly, of the 10 species assigned to *T. basedowii* by analyst 9, one was actually *P. arundinacea* and one was actually *Anthoxanthum odoratum* L. (Fig. 7C).

However, there are also some areas of agreement among the analysts. For example, in identification round two all nine analysts correctly identified at least nine out of 10 images of *Stipa tenuifolia* Steud. (Figs. 5–7). This species is characterized by relatively simple surface ornamentation, consisting of regularly spaced granula, that is visually distinctive in the context of the 12 species investigated here (Fig. 1). Similarly, certain species appear relatively difficult for all analysts to identify. For example, just five out of 10 specimens of *E. mexicana* were identified correctly by analysts 2, 3, 4, and 6 (Figs. 5, 6). Analysts 1, 5, 7, and 8 identified six out of 10 specimens of this species correctly (Figs. 5–7), and analyst 9 identified eight out of 10 specimens of *E. mexicana* correctly (Fig. 7C).

DISCUSSION

A growing body of evidence indicates that human analysts are unable to identify microscopic natural objects such as pollen grains with 100% accuracy (e.g., Ginsburg, 1997; Culverhouse et al., 2003, 2014; Culverhouse, 2007; Mander et al., 2013). Although based on portions of individual specimens, the mean identification accuracy of the nine analysts investigated here supports this view (46.67–87.5%; Fig. 2B). There are several psychological factors that are thought to reduce the ability of human analysts to identify objects (Evans, 1987) and that have been invoked to partly explain why human analysts identifying marine dinoflagellates achieved accuracies between 84% and 95% (Culverhouse et al., 2003).

The first of these is the limited capacity of human memory. Classic work has shown that the human short-term memory has a general capacity of between five and nine items (Miller, 1956), and the visual information subsystem of the short-term memory can retain up to 16 individual features when they are distributed across four different objects (Luck and Vogel, 1997). Some of the identification errors made by each analyst in our study are likely to be related to this because the number of SEM images in each identification round (120) far exceeds the known capacity of human short-term memory. The second factor is fatigue and boredom. Several analysts reported that they suffered both fatigue and boredom during the course of this study, which may have prevented analysts from focusing adequately on the task, and may have led to identification errors. One analyst, however, reported that they felt no fatigue and boredom during the study, and instead described intense enjoyment of the activity and the challenge it posed. They felt that if they were to complete the task too quickly, which might happen if an analyst was aiming to avoid fatigue and boredom, then their accuracy would drop. The third factor is the recency effect, whereby more recent experiences are influential in judgments about present situations (Jones and Sieck, 2003). In the context of identification, recency effects mean that a new identification is biased toward those specimens in the set of most recently identified specimens (Culverhouse, 2007). The fourth is positivity bias, where an analyst's identification is biased by their expectations of the species likely to be present in the sample. Certainly the nine analysts in this were all subject to positivity bias because they were told that each test set contained 10 specimens of each species, and were instructed that each species should be represented by no more than 10 images in their identification scheme. However, although each of these

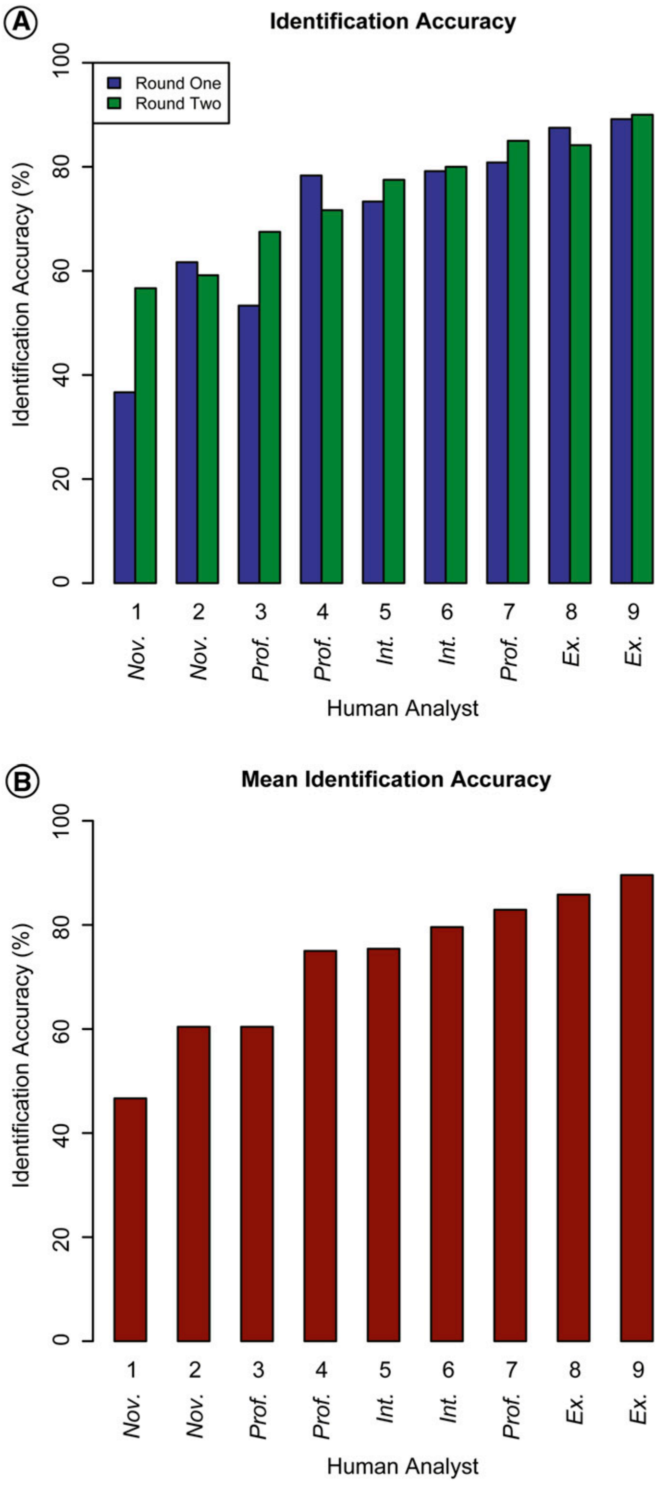


Fig. 2. Identification accuracy of the nine analysts reported for each of the two identification rounds separately (A), and as the mean of the two identification rounds (B). Abbreviations beneath each analyst number denote level of analyst experience (see Table 1).

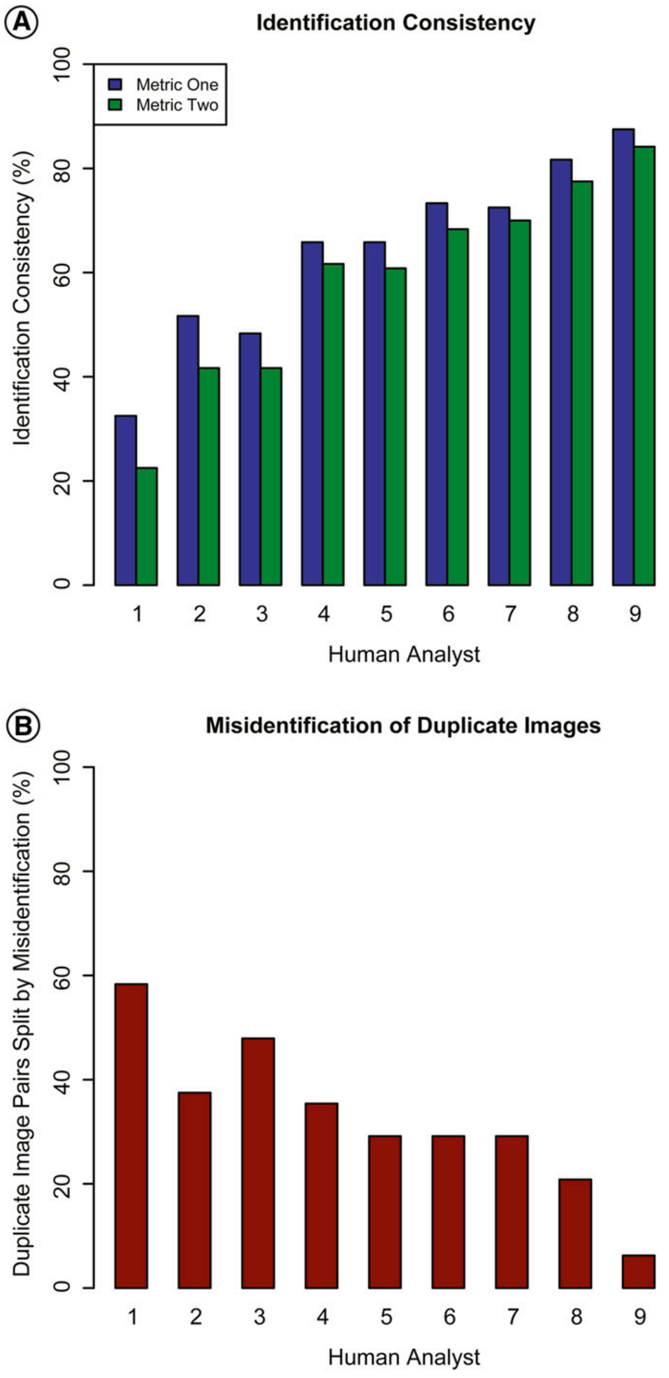


Fig. 3. Identification consistency of the nine analysts expressed using the two consistency metrics described in the main text (A), and as the percentage of duplicate image pairs that were split by misidentification (B).

four factors is a likely cause of misidentifications, it is not possible for us to convincingly tie specific identification errors to any one of these factors specifically. For example, of the 10 specimens identified as *Poa australis* R. Br. by analyst 8, two were actually *E. mexicana* (Fig. 7B), but we are unable to say conclusively whether these specific errors are the result of problems with short-term memory capacity, boredom, fatigue, recency effects, or positivity bias.

These factors are also likely to play a role in the identification consistency of the nine analysts in this study, who exhibited a greater range of self-consistency values (32.5–87.5% metric one, 22.5–84.17% metric two; Fig. 3A) than trained personnel asked to identify marine dinoflagellates in previous work (67–83%;

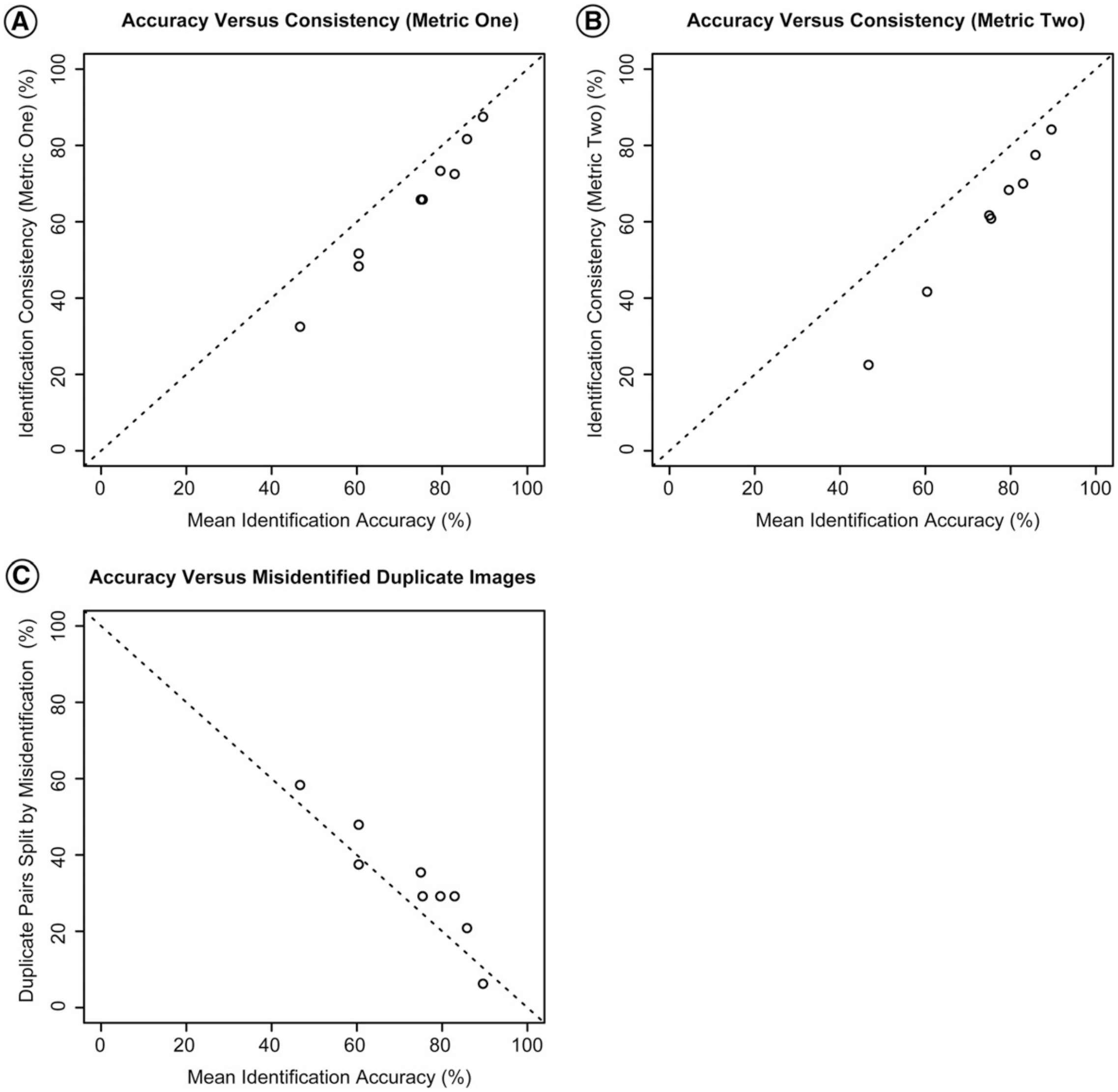
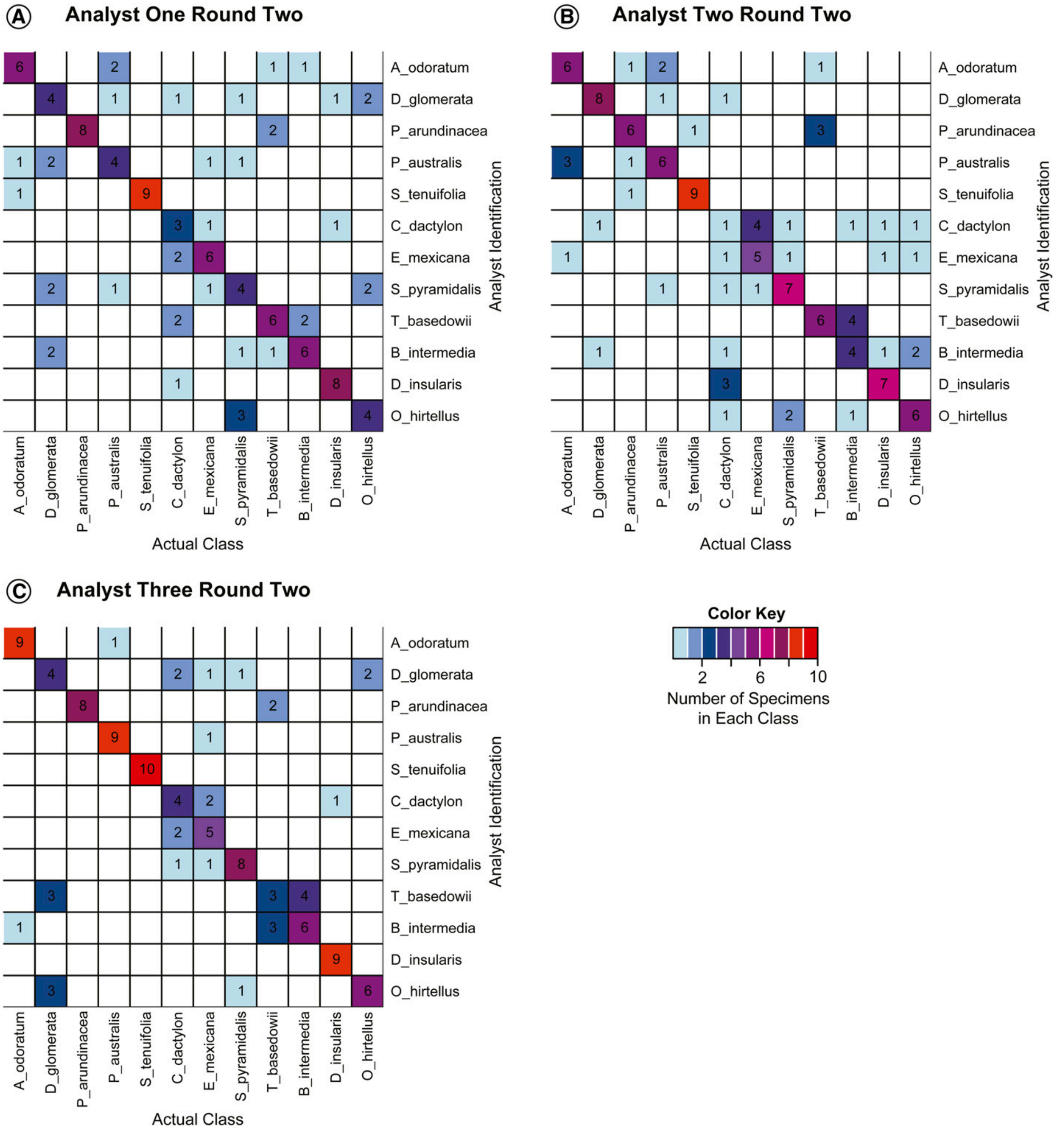


Fig. 4. Graphical comparisons of the identification accuracy and consistency of each of the nine analysts. Plot (A) shows mean identification accuracy against consistency metric one (described in the main text), plot (B) shows mean identification accuracy against consistency metric two (described in the main text), and plot (C) shows mean identification accuracy against the percentage of duplicate image pairs that were split by misidentification. Dashed diagonal line in each plot is a line of equality.

Culverhouse et al., 2003). The error matrices shown in Figs. 5–7 also highlight that each analyst produced a unique identification scheme. One of the roots of such inconsistency between workers is that human analysts are thought to create their own rules for identifying objects, so that the features used to identify an object by one analyst may not be the same as the features used to identify the same object by a different analyst (Sokal, 1974). The analysts in this study were instructed to complete the two identification rounds alone and without collaboration, and this allows us to look for evidence of such individualistic behavior.

In some cases, there is evidence that the analysts used different features to identify the species, and this is reflected in the reasons given by each analyst for their identifications. *Poa australis* (see Fig. 1), for example, was described as having “large, expansive areolae (exine islands) with high numbers of granulae; low contrast between islands and negative reticulum” by analyst 8, but analyst 9 stated that the “granulae appear brighter, islands better defined” and also that they used “intuition” as part of their identification of this species. Of the 10 specimens identified as *P. australis* by these two analysts, eight were correct and two were not (Fig. 7). However, in the case of analyst 8, these two misidentified specimens were actually *E. mexicana* (Fig. 7B), whereas in the case of analyst 9, these two misidentified specimens were actually *A. odoratum* (Fig. 7C). These two analysts used different features as the basis of their identifications of this species, with analyst 8 using the size of the areolae and the number of granula on the surface of the pollen grain. It is possible that this is an example of the individualistic behavior described by Sokal (1974), and may explain the lack of consensus between these two analysts on the identification of *P. australis* (Fig. 7B, 7C).

In most cases in this study, however, analysts appear to focus on the same features but use different vocabulary to describe them. The surface ornamentation of *Stipa tenuifolia* (see Fig. 1), for example, was described as follows: “small circular pustules



Ⓐ Analyst One Round Two

Analyst Identification	A_odoratum	D_glomerata	P_arundinacea	P_australis	S_tenuifolia	C_dactylon	E_mexicana	S_pyramidalis	T_basedowii	B_intermedia	D_insularis	O_hirtellus
A_odoratum	6			2					1	1		
D_glomerata		4		1		1		1			1	2
P_arundinacea			8						2			
P_australis	1	2		4			1	1				
S_tenuifolia	1				9							
C_dactylon						3	1				1	
E_mexicana						2	6					
S_pyramidalis		2		1			1	4				2
T_basedowii						2			6	2		
B_intermedia		2						1	1	6		
D_insularis						1					8	
O_hirtellus								3				4

Actual Class

Ⓑ Analyst Two Round Two

Analyst Identification	A_odoratum	D_glomerata	P_arundinacea	P_australis	S_tenuifolia	C_dactylon	E_mexicana	S_pyramidalis	T_basedowii	B_intermedia	D_insularis	O_hirtellus
A_odoratum	6		1	2					1			
D_glomerata		8		1		1						
P_arundinacea			6		1				3			
P_australis	3		1	6								
S_tenuifolia			1		9							
C_dactylon		1				1	4	1		1	1	1
E_mexicana	1					1	5	1			1	1
S_pyramidalis				1		1	1	7				
T_basedowii									6	4		
B_intermedia		1				1				4	1	2
D_insularis						3					7	
O_hirtellus						1		2		1		6

Actual Class

Ⓒ Analyst Three Round Two

Analyst Identification	A_odoratum	D_glomerata	P_arundinacea	P_australis	S_tenuifolia	C_dactylon	E_mexicana	S_pyramidalis	T_basedowii	B_intermedia	D_insularis	O_hirtellus
A_odoratum	9			1								
D_glomerata		4				2	1	1				2
P_arundinacea			8						2			
P_australis				9			1					
S_tenuifolia					10							
C_dactylon						4	2				1	
E_mexicana						2	5					
S_pyramidalis						1	1	8				
T_basedowii		3							3	4		
B_intermedia	1								3	6		
D_insularis											9	
O_hirtellus		3						1				6

Actual Class

Fig. 5. Error matrices highlighting the errors in the round two identifications produced by analysts 1 (A), 2 (B), and 3 (C). These analysts achieved between 47% and 61% mean identification accuracy (see Table 1). The actual class of each specimen is shown on the *x*-axis of each matrix. For example, analyst 3 (C) identified 10 images as *Phalaris arundinacea* L., but of those 10 images, two were actually *Triodia basedowii* Pritz. and eight were *P. arundinacea*. Specimens left unidentified by each analyst are not shown, and rows do not always sum to 10 as a result.

with low frequency, irregular distribution" (analyst 4); "no clustering, no islands, large spots, spots not dense" (analyst 6); "lack of areolae (exine islands) and very prominent, round granulae" (analyst 8); "Sculptural elements appear widely spaced. Lacks islands" (analyst 9). In these descriptions, the analysts have all described that this species lacks areolae, either by using this term (analyst 8) or by using the term "islands" instead (analysts 6 and 9), or by omitting this feature from the description altogether (analyst

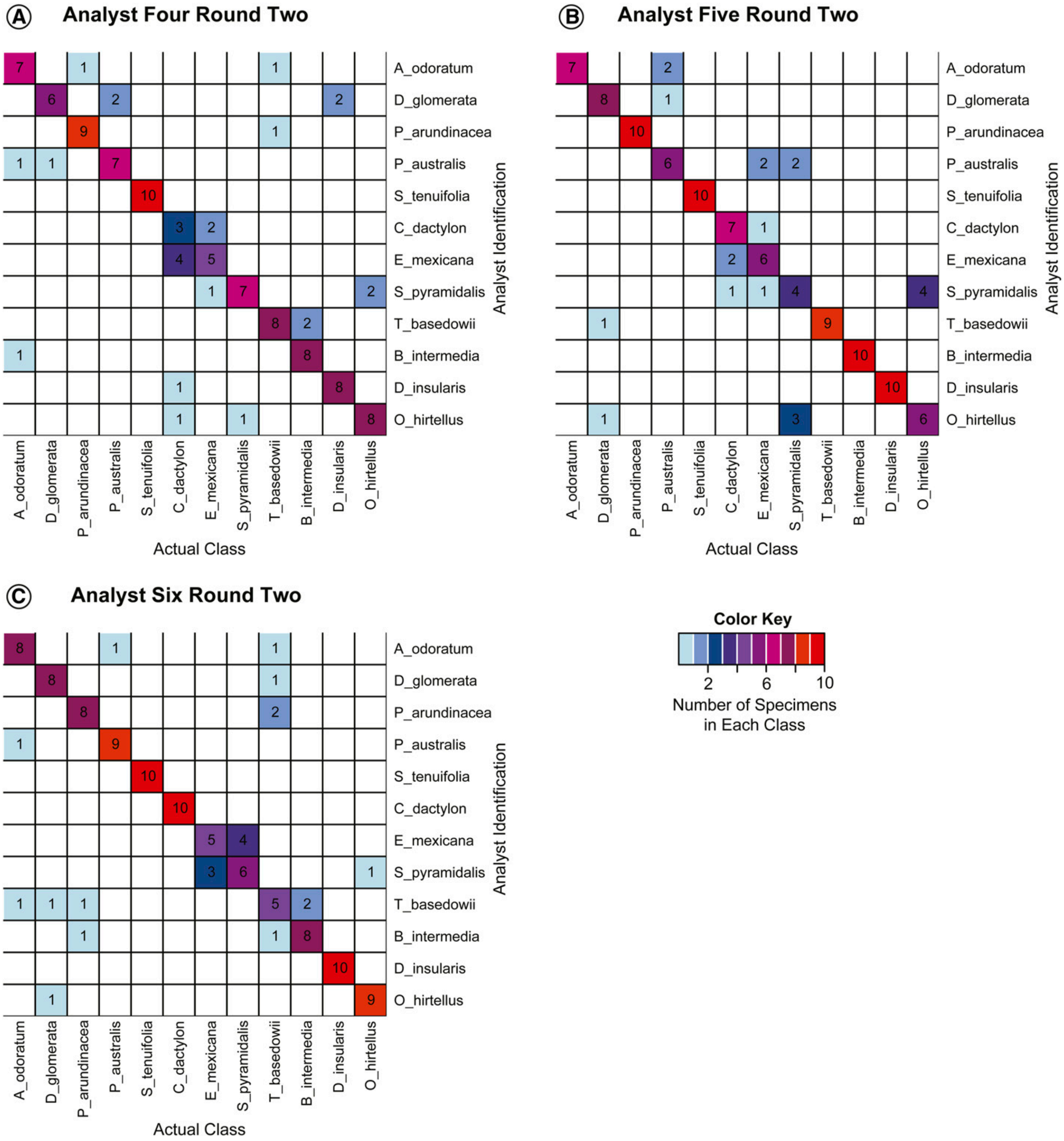
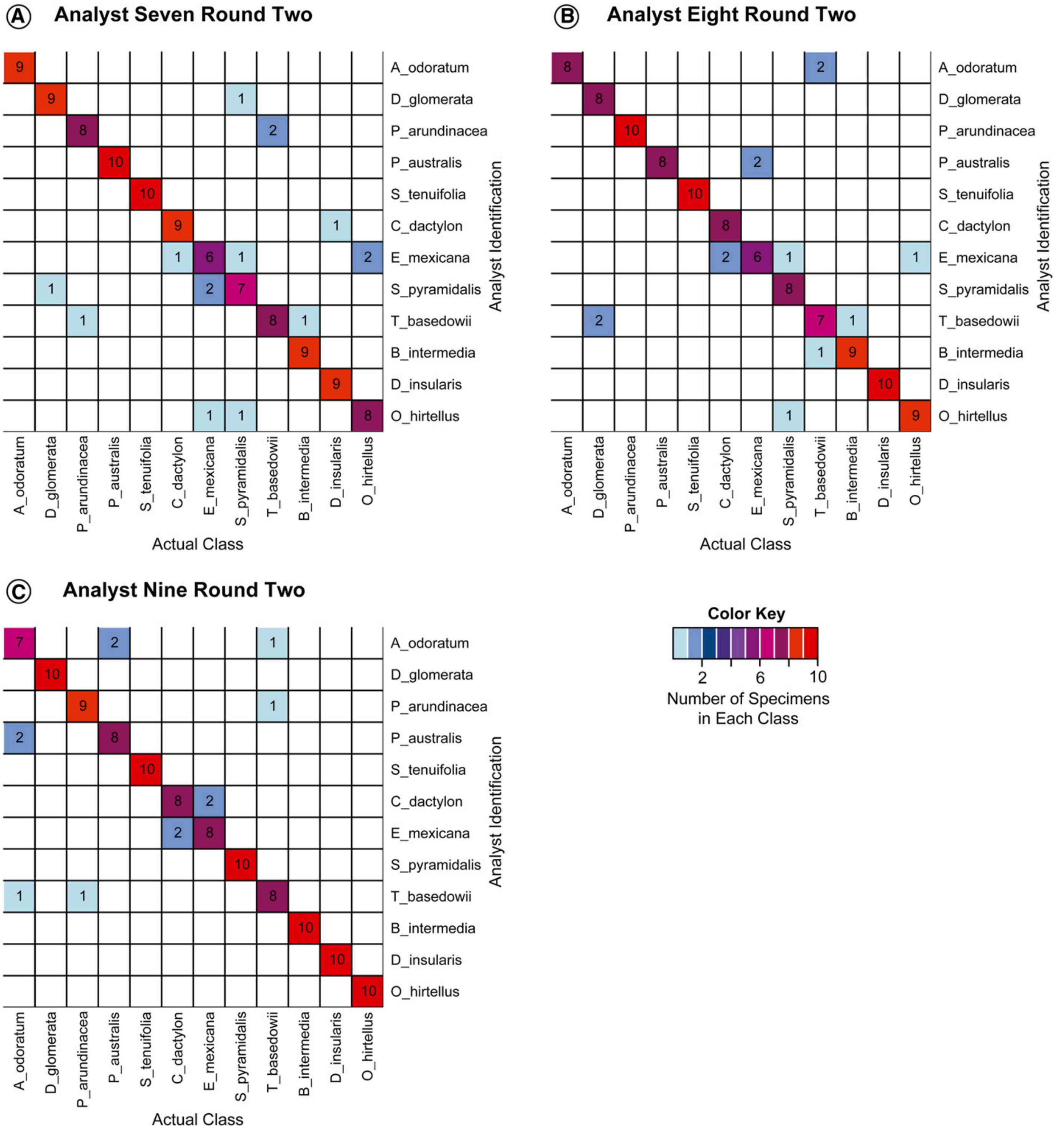


Fig. 6. Error matrices highlighting the errors in the round two identifications produced by analysts 4 (A), 5 (B), and 6 (C). These analysts achieved between 75% and 80% mean identification accuracy (see Table 1). Details as for Fig. 5.

4). There is some evidence of the analysts focusing on different features, with analysts 4 and 8 describing the shape of the individual granula on the pollen surface. This example shows that analysts can achieve consensus in terms of identification accuracy (each of these analysts identified *S. tenuifolia* with 100% accuracy in round two [Figs. 6–7]) despite using different terminology and, in the case of analysts 4 and 8, despite describing subtly different morphological features during the identification process.

It is difficult to make general statements about reasons for differences in the identification accuracy and consistency of the nine analysts. In this study, we have ranked each analyst in terms of their experience in classifying pollen grains or any other microscopic



Ⓐ Analyst Seven Round Two

	A_odoratum	D_glomerata	P_arundinacea	P_australis	S_tenuifolia	C_dactylon	E_mexicana	S_pyramidalis	T_basedowii	B_intermedia	D_insularis	O_hirtellus
A_odoratum	9											
D_glomerata		9						1				
P_arundinacea			8						2			
P_australis				10								
S_tenuifolia					10							
C_dactylon						9					1	
E_mexicana						1	6	1				2
S_pyramidalis		1					2	7				
T_basedowii			1						8	1		
B_intermedia										9		
D_insularis											9	
O_hirtellus							1	1				8

Ⓑ Analyst Eight Round Two

	A_odoratum	D_glomerata	P_arundinacea	P_australis	S_tenuifolia	C_dactylon	E_mexicana	S_pyramidalis	T_basedowii	B_intermedia	D_insularis	O_hirtellus
A_odoratum	8								2			
D_glomerata		8										
P_arundinacea			10									
P_australis				8			2					
S_tenuifolia					10							
C_dactylon						8						
E_mexicana						2	6	1				1
S_pyramidalis								8				
T_basedowii		2							7	1		
B_intermedia									1	9		
D_insularis											10	
O_hirtellus								1				9

Ⓒ Analyst Nine Round Two

	A_odoratum	D_glomerata	P_arundinacea	P_australis	S_tenuifolia	C_dactylon	E_mexicana	S_pyramidalis	T_basedowii	B_intermedia	D_insularis	O_hirtellus
A_odoratum	7			2					1			
D_glomerata		10										
P_arundinacea			9						1			
P_australis	2			8								
S_tenuifolia					10							
C_dactylon						8	2					
E_mexicana						2	8					
S_pyramidalis								10				
T_basedowii	1		1						8			
B_intermedia										10		
D_insularis											10	
O_hirtellus												10

Fig. 7. Error matrices highlighting the errors in the round two identifications produced by analysts 7 (A), 8 (B), and 9 (C). These analysts achieved between 83% and 90% mean identification accuracy (see Table 1). Details as for Fig. 5.

objects, such as charcoal, based on morphology. Using these categories, the level of analyst experience seems a poor predictor of classification accuracy as the two analysts with an intermediate level of experience achieved higher classification accuracy than two of the analysts with a professional level of experience, and the analysts with the highest classification accuracy have an expert level of experience (Table 1). Additionally, the mean classification accuracy of analysts 2 and 3 was identical, despite a wide gap in the level of experience of these two analysts (Table 1). These results may provide some support for the suggestion that the experience of an analyst measured in terms of "years on the job" is only weakly related to classification performance (Ericsson and Lehmann, 1996; Culverhouse, 2007). However, the scale on which we have ranked each analyst does not measure the intensity or quality of the hours that have been spent classifying objects using morphology. The two analysts with the highest classification accuracies are currently studying for PhDs in palynology and have been studying pollen morphology intensively recently for over a year. Perhaps these results should instead be interpreted as corroborating the idea that deliberate practice over a sustained period of time is crucial to the generation of expert levels of performance (Ericsson and Lehmann, 1996).

The high identification accuracy achieved by some of the analysts in this study is heartening from the perspective of using SEM images of fossil grass pollen grains to track changes in the diversity and composition of grasslands through time (e.g., Mander et al., 2013). The performance of analyst 9, who was also able to classify with quite high self-consistency and to also recognize most of the duplicate image pairs in this study (Table 1), is especially encouraging. Grass species can clearly be identified using SEM images of the surface ornamentation on their pollen grains (Andersen and Bertelsen, 1972; Page, 1978; Peltre et al., 1987; Chaturvedi et al., 1998; Mander et al., 2013) (Fig. 2). Some additional features of grass pollen morphology that also have taxonomic significance, such as the distribution of tectal columellae (Fægri et al., 1992; Beug, 2004; Holst et al., 2007), cannot be seen using the SEM because this instrument records information from the surface of individual specimens (Sivaguru et al., 2012). Similarly, the shape of the grass pollen pore (Schüler and Behling, 2011a, b) can be hidden from view because of the orientation of the specimen on the SEM stub, and the collapsing of grains on the SEM stub can prevent the overall size of pollen grains from being measured accurately (e.g., Moore et al., 1991). Nevertheless, where possible, the use of additional features such as pore shape and grain size (e.g., Andersen, 1979; Tweddle et al., 2005; Joly et al., 2007; Schüler and Behling, 2011a, b) will presumably increase the accuracy of grass pollen identification by human analysts.

This invites comparison of the accuracy of the nine analysts studied here and computational methods for identifying the same SEM images of grass pollen (e.g., Mander et al., 2013). Four of the analysts exceeded the accuracy of computational identifications of the same images based on quantifying the complexity of grass pollen surface ornamentation (Fig. 8), but only analyst 9 exceeded the accuracy of computational identifications based on descriptions of grass pollen surface ornamentation using histograms of local quantized image patches (Fig. 8). This shows that some human analysts are able to compete with current computational methods in terms of identification accuracy, albeit with half the number of specimens that were used for computational identifications based on quantifying the complexity of grass pollen surface ornamentation (Mander et al., 2013). These computational identifications took approximately eight hours.

However, we emphasize that high identification accuracy alone does not necessarily represent success. It is the low consistency of identifications by the same analyst (e.g., Fig. 3) and by different analysts (Figs. 5–7) that leads researchers to identify pollen grains at relatively low taxonomic ranks, such as the genus or family, in an attempt to ensure that their identifications are reproducible (Punyasena et al., 2012b). It is this drive for consistency and repeatability that underpins the need for computational approaches to identification (MacLeod et al., 2010; Punyasena et al., 2012b). There are additional concerns, which include the amount of time it takes for human analysts to undertake difficult classification tasks such as the one described in this paper. For example, although the identifications of analyst 9 surpass both computational methods and are also reasonably consistent, this analyst spent around eight hours completing identification round one, and around five and half hours completing identification round two. One view of this might be that the length of time taken on a particular task should be of little concern if the data collected are valuable, but an alternative view is that spending too much time on a particular focused task reduces scientific productivity.

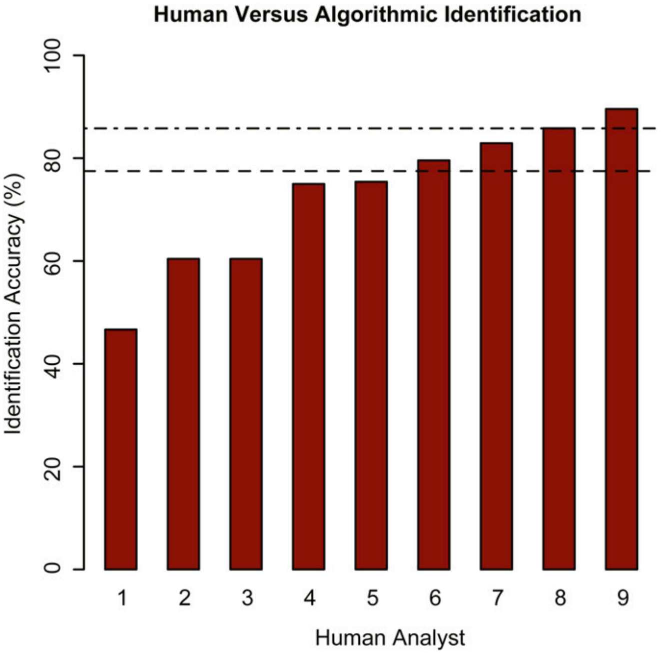


Fig. 8. Bar chart comparing the mean identification accuracy of the nine analysts in this study with the accuracy of computational methods of identifying grass pollen. The lower dashed line represents computational identifications based on quantifying the complexity of grass pollen surface ornamentation (77.5% accuracy; Mander et al., 2013). The upper dot-dash line represents computational identifications based on descriptions of grass pollen surface ornamentation using histograms of local quantized image patches (85.8% accuracy; Mander et al., 2013).

We close this paper with a discussion of the role of human analysts in the present era of computational classification and identification. One of the primary reasons that is cited in support of computational approaches to the identification of natural objects such as pollen grains is to “free [analysts] from the drudgery of routine identifications” (MacLeod et al., 2010, p. 155). The Classifynder automated pollen-counting system, for example, has been explicitly “designed to dramatically reduce the time that the palynologist must spend at the microscope” (Holt et al., 2011, p. 175). However, we urge palynologists not to dismiss all routine work as drudgery to be passed entirely on to automated pollen-counting systems. This is because expert levels of performance are achieved only after spending about 10 years in intense preparation with deliberate and structured practice lasting at most four hours per day (Ericsson and Lehmann, 1996), and we suggest that daily routine work is the time when palynologists can practice their core identification skills and attain expert levels of performance. Viewed in this light, routine palynological work could be seen as analogous to the scales and rudiments that are practiced by expert musicians. Of course, continual practice of routine identification can create large-scale recency effects and positivity bias (Culverhouse, 2007), which may reduce the ability of analysts to recognize a new species in the middle of a routine investigation.

However, if the time a palynologist spends undertaking routine identifications was structured to mimic the regimens of careful training and practice that lead to expert and exceptional performance in fields such as chess, music, and athletics (see Ericsson and Lehmann, 1996), then the performance of individual workers and the whole discipline would be raised considerably. Such improvement could be vital in the near future because some current automated identification systems, again using the Classifynder instrument as an example, require that "the classified images [be] then presented to the palynologist for checking" (Holt et al., 2011, p. 175). Clearly it is desirable to have any computationally generated identification system checked by an expert, particularly when tackling difficult palynological problems such as hyperdiverse tropical systems. Therefore, identification of pollen by human experts will remain a necessity in palynology. As automated systems become more common, palynologists must remain mindful that a certain level of regular exposure to pollen morphology is required to maintain their levels of expertise.

LITERATURE CITED

Andersen, S. T. 1979. Identification of wild grass and cereal pollen. *Danmarks Geologiske Undersøgelse Årbok* 1978: 69–92.

Andersen, S. T., and F. Bertelsen. 1972. Scanning electron microscope studies of pollen of cereals and other grasses. *Grana* 12: 79–86.

Beug, H.-J. 2004. Leitfaden de Pollenbestimmung für Mitteleuropa und angrenzende Gebiete. Verlag Dr. Friedrich Pfiel, Munich, Germany.

Birks, H. J. B., and H. H. Birks. 1980. Quaternary palaeoecology. Edward Arnold, London, United Kingdom.

Birks, H. H., and H. J. B. Birks. 2000. Future uses of pollen analysis must include plant macrofossils. *Journal of Biogeography* 27: 31–35.

Chaturvedi, M., K. Datta, and P. K. K. Nair. 1998. Pollen morphology of *Oryza* (Poaceae). *Grana* 37: 79–86.

Culverhouse, P. F. 2007. Natural object classification: Man versus machine. *In* N. MacLeod [ed.], Automated taxon identification in systematics: Theory, approaches and applications, 25–45. CRC Press, Boca Raton, Florida, USA.

Culverhouse, P. F., R. Williams, B. Reguera, V. Herry, and S. González-Gil. 2003. Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. *Marine Ecology Progress Series* 247: 17–25.

Culverhouse, P. F., N. MacLeod, R. Williams, M. C. Benfield, R. M. Lopes, and M. Picheral. 2014. An empirical assessment of the consistency of taxonomic identifications. *Marine Biology Research* 10: 73–84.

Edwards, E. J., C. P. Osborne, C. A. E. Strömberg, S. A. Smith, and the C_4 Grasses Consortium. 2010. The origins of C_4 grasslands: Integrating evolutionary and ecosystem science. *Science* 328: 587–591.

Ericsson, K. A., and A. C. Lehmann. 1996. Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology* 47: 273–305.

Evans, J. St. B. T. 1987. Bias in human reasoning: Causes and consequences. Laurence Erlbuam Associates, Hove, United Kingdom.

Fægri, K., P. E. Kaland, and K. Krzywinski. 1992. Textbook of pollen analysis. Wiley, Chichester, United Kingdom.

Ginsburg, N. 1997. Perspectives on the blind test. *Marine Micropaleontology* 29: 101–103.

Grass Phylogeny Working Group II. 2012. New grass phylogeny resolves deep evolutionary relationships and discovers C_4 origins. *New Phytologist* 193: 304–312.

Holst, I., J. E. Moreno, and D. R. Piperno. 2007. Identification of teosinte, maize, and *Tripsacum* in Mesoamerica by using pollen, starch grains, and phytoliths. *Proceedings of the National Academy of Sciences, USA* 104: 17608–17613.

Holt, K., G. Allen, R. Hodgson, S. Marsland, and J. Flenley. 2011. Progress towards an automated trainable pollen location and classifier system for use in the palynology laboratory. *Review of Palaeobotany and Palynology* 167: 175–183.

Jackson, S. T., and R. T. Booth. 2007. Validation of pollen studies. *In* S. A. Elias [ed.], Encyclopaedia of Quaternary sciences, 2413–2422. Elsevier Scientific Publishing, Amsterdam, The Netherlands.

Joly, C., L. Barillé, M. Barreau, A. Mancheron, and L. Visset. 2007. Grain and annulus diameter as criteria for distinguishing pollen grains of cereals from wild grasses. *Review of Palaeobotany and Palynology* 146: 221–233.

Jones, M., and W. R. Sieck. 2003. Learning myopia: An adaptive recency effect in category learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition* 29: 626–640.

Kohavi, R., and F. Provost. 1998. Glossary of terms. *Machine Learning* 30: 271–274.

Luck, S. J., and E. K. Vogel. 1997. The capacity of visual working memory for features and conjunctions. *Nature* 390: 279–281.

MacLeod, N., M. C. Benfield, and P. F. Culverhouse. 2010. Time to automate identification. *Nature* 467: 154–155.

Mander, L. 2011. Taxonomic resolution of the Triassic–Jurassic sporomorph record in East Greenland. *Journal of Micropalaeontology* 30: 107–118.

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

May, L., and T. Lacourse. 2012. Morphological differentiation of *Alnus* (alder) pollen from western North America. *Review of Palaeobotany and Palynology* 180: 15–24.

Miller, G. A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63: 81–97.

Moore, P. D., J. A. Webb, and M. E. Collinson. 1991. Pollen analysis. Blackwell Scientific, London, United Kingdom.

Page, J. S. 1978. A scanning electron microscope survey of grass pollen. *Kew Bulletin* 32: 313–319.

Peltre, G., M. T. Cerceau-Larrival, M. Hideux, M. Abadie, and B. David. 1987. Scanning and transmission electron microscopy related to immunochemical analysis of grass pollen. *Grana* 26: 158–170.

Piperno, D. R. 2006. Phytoliths: A comprehensive guide for archaeologists and paleoecologists. AltaMira Press, New York, New York, USA.

Punt, W., P. P. Hoen, S. Blackmore, S. Nilsson, and A. Le Thomas. 2007. Glossary of pollen and spore terminology. *Review of Palaeobotany and Palynology* 143: 1–81.

Punyasena, S. W., J. W. Dalling, C. Jaramillo, and B. L. Turner. 2011. Comment on "The response of vegetation on the Andean Flank in Western Amazonia to Pleistocene Climate Change." *Science* 333: 1825.

Punyasena, S. W., C. Jaramillo, F. de la Parra, and Y. Du. 2012a. Probabilistic correlation of single stratigraphic samples: A generalized approach. *AAPG Bulletin* 96: 235–244.

Punyasena, S. W., D. K. Tcheng, C. Wesseln, and P. G. Mueller. 2012b. Classifying black and white spruce using layered machine learning. *New Phytologist* 196: 937–944.

Schüler, L., and H. Behling. 2011a. 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.

Schüler, L., and H. Behling. 2011b. Characteristics of Poaceae pollen grains as a tool to assess palaeoecological grassland dynamics in South America. *Vegetation History and Archaeobotany* 20: 97–108.

Sivaguru, M., L. Mander, G. Fried, and S. W. Punyasena. 2012. Capturing the shape and surface texture of pollen: A comparison of microscopy techniques. *PLoS ONE* 7: e39129.

Sokal, R. R. 1974. Classification: Purposes, principles, progress, prospects. *Science* 185: 1115–1123.

Strömberg, C. A. E. 2011. Evolution of grasses and grassland ecosystems. *Annual Review of Earth and Planetary Sciences* 39: 517–544.

Traverse, A. 2007. Paleopalynology, 2nd ed. Springer, Dordrecht, The Netherlands.

Tweddle, J. C., K. J. Edwards, and N. R. J. Fieller. 2005. Multivariate statistical and other approaches for the separation of cereal from wild Poaceae pollen using a large Holocene dataset. *Vegetation History and Archaeobotany* 14: 15–30.