

Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing

KEN KRAAIJEVELD,*† LETTY A. DE WEGER,‡ MARINA VENTAYOL GARCÍA,* HENK BUERMANS,* JEROEN FRANK,* PIETER S. HIEMSTRA‡ and JOHAN T. DEN DUNNEN*

*Human and Clinical Genetics, Leiden Genome Technology Center, Leiden University Medical Center, PO Box 9600, 2300RC Leiden, The Netherlands, †Bioinformatics, University of Applied Sciences Leiden, Zernikedreef 11, 2333CK Leiden, The Netherlands, ‡Department of Pulmonology, Leiden University Medical Center, PO Box 9600, 2300RC Leiden, The Netherlands

Abstract

Pollen monitoring is an important and widely used tool in allergy research and creation of awareness in pollen-allergic patients. Current pollen monitoring methods are microscope-based, labour intensive and cannot identify pollen to the genus level in some relevant allergenic plant groups. Therefore, a more efficient, cost-effective and sensitive method is needed. Here, we present a method for identification and quantification of airborne pollen using DNA sequencing. Pollen is collected from ambient air using standard techniques. DNA is extracted from the collected pollen, and a fragment of the chloroplast gene *trnL* is amplified using PCR. The PCR product is subsequently sequenced on a next-generation sequencing platform (Ion Torrent). Amplicon molecules are sequenced individually, allowing identification of different sequences from a mixed sample. We show that this method provides an accurate qualitative and quantitative view of the species composition of samples of airborne pollen grains. We also show that it correctly identifies the individual grass genera present in a mixed sample of grass pollen, which cannot be achieved using microscopic pollen identification. We conclude that our method is more efficient and sensitive than current pollen monitoring techniques and therefore has the potential to increase the throughput of pollen monitoring.

Keywords: DNA metabarcoding, molecular identification, next-generation sequencing, pollen allergy, pollen monitoring

Received 29 November 2013; revision received 19 May 2014; accepted 23 May 2014

Introduction

Respiratory allergies caused by plant pollen affect hundreds of millions of people worldwide (Bousquet *et al.* 2007). The severity of the symptoms differs extensively between plant species and thus flowering time and is influenced by meteorological variables such as wind, temperature and humidity and by air pollution (Galán *et al.* 1995; D'Amato *et al.* 2005). Accurate, detailed and up-to-date information on local levels of airborne pollen allows patients to take preventive measures. Pollen monitoring is widely used to provide pollen-allergic patients with daily information about previous pollen levels in the air (e.g. <https://www.lumc.nl/pollen>).

Ideally, pollen monitoring should be organized as a dense, continent-wide network of monitoring stations. Unfortunately, several aspects of the current pollen monitoring process hamper the establishment of such networks. Pollen monitoring currently involves pollen collection from ambient air, usually by immobilizing it

on a sticky tape, followed by microscopic identification and quantification under a microscope (Rittenour *et al.* 2012; Scheifinger *et al.* 2013). This procedure requires trained personnel able to recognize the various pollen types, is labour intensive and therefore costly. This prohibits more frequent sampling and establishment of multiple pollen monitoring stations, which would be highly beneficial for both research and provision of patient information. Another problem is that not all pollen can be identified to the genus level, reducing the sensitivity of the current pollen monitoring method. For example, the pollen of the hundreds of grass species is microscopically identical in morphology. Identification of the grass pollen to the genus level is highly relevant as the different genera may differ in the degree to which they cause allergic reactions (Hrabina *et al.* 2008; De Weger *et al.* 2011). Therefore, there is a clear need for a more efficient, cost-effective and sensitive method of monitoring airborne pollen.

Attempts to increase the throughput of pollen monitoring procedures focus on digital image recognition and light scattering principles (Kawashima *et al.* 2007; Landsmeer *et al.* 2009; Rittenour *et al.* 2012; Scheifinger *et al.*

Correspondence: Ken Kraaijeveld, Fax: +31 20 5987004; E-mail: ken@kenkraaijeveld.nl

2013). These methods have great difficulty discriminating pollen from dust particles, and the potential to identify pollen to genus level is limited. DNA-based pollen identification has been explored with some success (Longhi *et al.* 2009), but at that time, identification of only a small number of plant species was possible. Furthermore, previous attempts at DNA-based pollen monitoring were hampered by high costs and low throughput of the available technology at the time. Finally, immunological and spectrophotometric methods have been evaluated (Rittenour *et al.* 2012). To our knowledge, none of these approaches have thus far resulted in an operational method to replace the conventional counting method.

Recent developments in DNA sequencing technology, so-called next-generation sequencing (NGS), have opened new possibilities for DNA-based identification and quantification of environmental samples (Chariton *et al.* 2010; Fonseca *et al.* 2010; Hajibabaei *et al.* 2011; Epp *et al.* 2012; Pompanon *et al.* 2012; Thomsen *et al.* 2012; Yoccoz *et al.* 2012; Yu *et al.* 2012; Bellemain *et al.* 2013; Ji *et al.* 2013; Schmidt *et al.* 2013). Here, we show that such an approach is feasible and results in accurate identification of pollen from mixed samples, including species that cannot be distinguished microscopically. Furthermore, we show that the relative abundance of pollen from a sample taken from the air matches that recorded using the traditional method. We argue that with decreasing cost and increasing accuracy, NGS-based pollen monitoring will soon present a cost-effective superior alternative to microscopic pollen counting.

Methods

Pollen collection

We made use of the routine pollen monitoring facilities at the Leiden University Medical Center (LUMC), the Netherlands (De Weger *et al.* 2013). Pollen was collected from ambient air using Burkard traps (Burkard Manufacturing, Rickmansworth, UK) placed on the roof of the main LUMC building. The trap uses a vacuum pump to suck in air. Pollen and other particles are impacted on adhesive tape supported on a clockwork-driven drum. The drum rotates at a constant speed, so that a given section of tape corresponds to a known length of time. The tape was removed from the trap and cut into seven pieces from which microscope slides were prepared. In an area corresponding to a 1-day period (48 × 20 mm), the pollen grains were counted microscopically after staining with safranin (0.002%). A second trap, identical to and placed within 2 m of the first, was operated simultaneously at the same site. The tape from this trap was cut into 1-day sections and subsequently into two sections lengthwise (48 × 10 mm) (Roger's *et al.* 2009).

One of these tape fragments was used for DNA extraction and Ion Torrent sequencing as described below. Three samples were collected to represent different parts of the pollen season and thus to contain pollen from different plant species: 4 May 2013, 26 May 2012 and 17 July 2012. Slight modifications were made between the processing of the sample from 26 May 2012 and the other two as described below.

To assess whether our method could accurately distinguish between pollen from different species of grass that are indistinguishable using microscopic methods, we collected pollen from six species of grass (*Dactylis glomerata*, *Phleum pratense*, *Hordeum murinum*, *Holcus lanatus*, *Lolium perenne* and *Arrhenaterum elatius*) during the summer of 2013 in the vicinity of the LUMC. To provide unambiguous reference sequences, DNA was extracted from pollen of each species and sequenced for an approximately 200 bp region of the *trnL* gene using standard Sanger sequencing. Subsequently, these six pollen species were mixed in unstandardized proportions. DNA was extracted from the mixed sample, PCR-amplified and sequenced on the Ion Torrent PGM sequencing platform (Life Technology) as described below.

Lysis efficiency

To test the efficiency of our lysis procedure, we diluted 1 mg of two pollen samples (*Betula pendula* and *Dactylis glomerata*) in 400 µL water. Eight microlitre of each sample was transferred to a hemocytometer, and pollen grains were counted under a microscope. The remaining 392 µL of each sample was disrupted in a bullet blender as described below. The sample was then transferred to the hemocytometer, and the remaining intact pollen grains were counted under a microscope. For each pollen sample, we counted 10 microscope fields before and after lysis.

DNA extraction

The half-day section of tape containing pollen was spiralled into a 1.5-ml Safe-Lock Eppendorf tube with 400 µL of DNA isolation buffer (155 mM NaCl, 125 mM Tris-HCl pH 8.0, 15 mM EDTA, 1.25% polyvinylpyrrolidone, 0.6% SDS, 3 mM 1,10-phenanthroline monohydrate, 0.1% betamercaptoethanol) and five 3.2-mm stainless steel balls (Roger's *et al.* 2009). The sample was disrupted on a Next Advance bullet blender for 3 minutes at speed 8. Subsequently, 100 µL of 5% SDS was added to the sample, which was then incubated at 65°C for 30 min. DNA was extracted using the QIAamp DNA Mini kit according to the manufacturer's protocol for crude cell lysates. The DNA concentrations were mostly

below detection level using Qubit (Life Technologies), and we used 1 μ l in subsequent PCRs.

PCR diagnostic locus

We considered several loci that are widely used in molecular identification of plant material: *ITS*, *trnL*, *rbcL* and *matK*; the latter two have been recommended as a standard (Hollingsworth 2009; Hollingsworth *et al.* 2011). Following preliminary experiments, *trnL* was selected because it provided the most discriminatory power, allowing identification down to genus level in most cases.

TrnL primers with Ion Torrent tails were used for the PCR amplification. The forward primer was a combination of the Ion Torrent A adapter with *trnL-c* (Taberlet *et al.* 1991, 2007), and the reverse primer the Ion Torrent P1 adapter along with *trnL-h* (Taberlet *et al.* 2007; Table 1). The length of the target sequence varied between species, but averaged about 200 bp. For the May 26 sample, PCR was performed in a total volume of 25 μ l containing 1 $\times$ Phusion High Fidelity PCR Master Mix (New England Biolabs), 250 nM of each primer and 1 μ l of DNA. PCR conditions were 30 s at 98 °C, 25 cycles of 10 s at 98 °C, 30 s at 61 °C and 30 s at 72 °C, followed by 5 min at 72 °C. The other two samples were processed in the same way, except that the 1 $\times$ Phusion High Fidelity PCR Master Mix was substituted by 1 $\times$ Pim Mix (45 mM Tris pH 8.3, 11 mM (NH₄)₂SO₄, 4.5 mM MgCl₂, 6.7 mM β -mercaptoethanol, 0.5 mM dNTP's, 0.11 μ g/ μ l BSA fraction V, 4.4 μ M EDTA) and 0.5 U of Phusion Polymerase.

Ion Torrent library preparation and sequencing

Libraries were prepared for the Ion Torrent PGM according to the manufacturer's protocol. For the 26 May 2012 sample, the one-touch v1 300 bp template and sequencing kits were used, while for the other samples, the one-touch v2 200 bp template and sequencing kits v2 were used. Emulsion PCR was performed with the one-touch v1. Each library was run separately on an Ion Torrent chip. The 26 May 2012 sample was run on a 314 chip and the other two samples on a 316 chip, which yields more reads. Sequencing yields are summarized in Table 2.

Table 1 Primer sequences

Primer	Components	Sequence
Forward	A_barcode- <i>trnLc</i>	CCATCTCATCCCTGCGTGTCTCCGACTCAGCTTGTCCAATCGAAATCGGTAGACGCTACG
Reverse	P1- <i>trnLh</i>	CCTCTCTATGGGCAGTCGGTGATCCATTGAGTCTCTGCACCTATC

Barcode sequence underlined, locus-specific primer sequences in bold.

Table 2 Sequence output of the Ion Torrent runs

Sample	Initial read number	Reads with both primers present
Airborne samples		
4 May 2013	1 869 348	984 457
26 May 2012	732 279	351 029
17 July 2012	1 733 341	893 175
Known content		
Grassmix	1 739 802	574 437

Mixed amplicon sequencing

Several studies have recommended running multiple PCRs for environmental DNA samples and pooling these for sequencing (Thomsen *et al.* 2012; Yu *et al.* 2012; Leray *et al.* 2013; Schmidt *et al.* 2013). To test whether this approach would improve our method, we ran eight additional PCRs with the 26 May 2012 sample as template. Five reactions used the same conditions as above, while the other three reactions were run at different annealing temperatures (57, 59 and 63 °C). For each of these reactions, the primers were fitted with different adapter sequences. The samples were pooled and sequenced in one Ion Torrent run as described above (316 chip). The adapter sequences allowed the data for each PCR to be analysed either separately or in aggregate.

Data analysis

From the Ion Torrent data, all reads that contained both the complete forward and reverse primer sequence, allowing for one mismatch for primer recognition, were selected using the TSSV software tool (Anvar *et al.* 2014). Primer sequences were clipped, and the remaining sequence was blasted against a local install of the NCBI database (downloaded on July 23rd, 2013). Blastn parameters were as follows: maximum number of aligned sequences to keep = 50, expectation value threshold = 1e-40, minimum per cent identity = 97% and penalty for a nucleotide mismatch = -5. The blast output was parsed to a tab-delimited text file for further processing. We selected hits of which the length of the aligned portion of the read was within 5 bp of the total

read length. We report the results for 98%, 99% and 100% identity separately. The hits to species within the same genus were pooled. When a sequence had hits of equal quality to multiple plant genera, we separated the genus names using '/'. We counted the number of sequences that had their top blast hits to each genus or group of genera. For the airborne pollen samples, we compared these counts to the results of standard microscopic counts from the same day using generalized linear models (GLM) with a Poisson error distribution and an empirically estimated scale parameter in R (R Development Core Team 2010). Significance was tested by dropping terms from the model and comparing the resulting change in deviance using an F test. The data from the mixed PCR amplicon experiment were compared with the single-PCR data set using Pearson's correlations. The number of sequences per genus for the eight PCRs for this experiment was compared using GLM as described above, and the number of genera in each PCR was compared using a chi-squared test.

Results

Lysis efficiency

Our lysis protocol resulted in almost complete lysis of both *Betula* and *Dactylis* pollen. We counted 14.1 ± 5.09 SD *Betula* pollen and 31.2 ± 20.43 SD *Dactylis* pollen per field before lysis. After disruption in the bullet blender, we counted 0.4 ± 0.52 SD remaining *Betula* pollen and 0.3 ± 0.48 SD *Dactylis* pollen per field. The lysis efficiency was thus 97% for *Betula* and 99% for *Dactylis*.

Analysis of airborne pollen samples

The DNA sequence data provided much greater resolution than the microscopic counts. In most cases, the blast results allowed identification of genera, while microscopic identification could only identify pollen down to the family level (e.g. *Cupressaceae* and *Taxus*, *Pinaceae*, *Poaceae*). To compare the DNA read counts to the microscopic counts, we lumped the DNA read counts for groups that could not be distinguished microscopically (indicated in Table S1, Supporting information). The number of sequence reads matching a given plant genus or family closely correlated with the number counted using standard microscopic pollen counts for each of the three airborne pollen samples (GLMs for 100% identity: 4 May 2012 $F_{1,46} = 13.53$, $P < 0.001$, 26 May 2013 $F_{1,46} = 38.16$, $P < 0.001$, 17 July 2012 $F_{1,46} = 36.83$, $P < 0.001$; P values for 99% and 98% lower; Fig. 1; Table S1, Supporting information). When the three samples were combined into one model, a significant interaction was observed between sampling date and

microscopic counts ($F_{2,138} = 19.37$, $P < 0.001$; Fig. S1, Supporting information). Therefore, the relationship between the sequence read counts and the microscopic counts differed between samples (Fig. S1, Supporting information), reflecting the fact that PCR obscures the absolute numbers of molecules of a given DNA sequence in each sample. Nevertheless, the relative abundance of sequence reads accurately reflected the relative abundance of pollen grains in the sample for a given species (Fig. 1).

The difference in resolution between the DNA read counts and the microscopic counts was most dramatic for grasses (*Poaceae*; Table S1, Supporting information). At least 34 different grass sequences were found, 20 of which could be identified to genus level. The relative abundance of these sequences differed between the three samples (Fig. 2). The 4 May sample contained few grass pollen. The 26 May sample was dominated by *Poa* and *Dactylis* and the 17 July sample by *Agrostis*-like sequences. This variation matched the flowering time of the different grasses. For example, *Dactylis glomerata* is an early flowering grass and peak numbers were found in the sample from 26 May. By contrast, sequence read numbers for the late-flowering *Agrostis* peaked in the 17 July sample.

Mixed grass pollen sample

Each of the *trnL* sequences of the six grasses in the mix showed a unique pattern of single nucleotide polymorphisms (SNPs) and insertion deletions (Fig. 3). Each of the *trnL* sequences showed a 100% match to the correct species in the NCBI database (and in most cases also to a number of related grass species). The only exception was *Holcus lanatus*, for which the reference sequence was truncated.

Four of the six grass species in the sample were identified unambiguously among the most abundant blast hits (*Holcus*, *Hordeum*, *Phleum* and *Dactylis*; Table 3). The remaining two species (*Lolium* and *Arrhenatherum*) were also among the most abundant hits, but could not be distinguished from *Festuca* and *Poa*, respectively (Table 3). Other abundant blast hits (*Puccinellia* and *Alopecurus*) are grasses with *trnL* sequences very similar to *Hordeum* and *Phleum*.

Mixed amplicons

The numbers of sequences for each plant genus in the pool of eight PCR amplicons correlated strongly to those in the single PCR (100% blast hits: $R = 0.94$, $P < 0.001$). The data for the pooled sample did not provide a better fit to the microscopic data than the single PCR (100% blast hits: $R = 0.73$ and $R = 0.49$, respectively), presumably because the large number of sequences in the

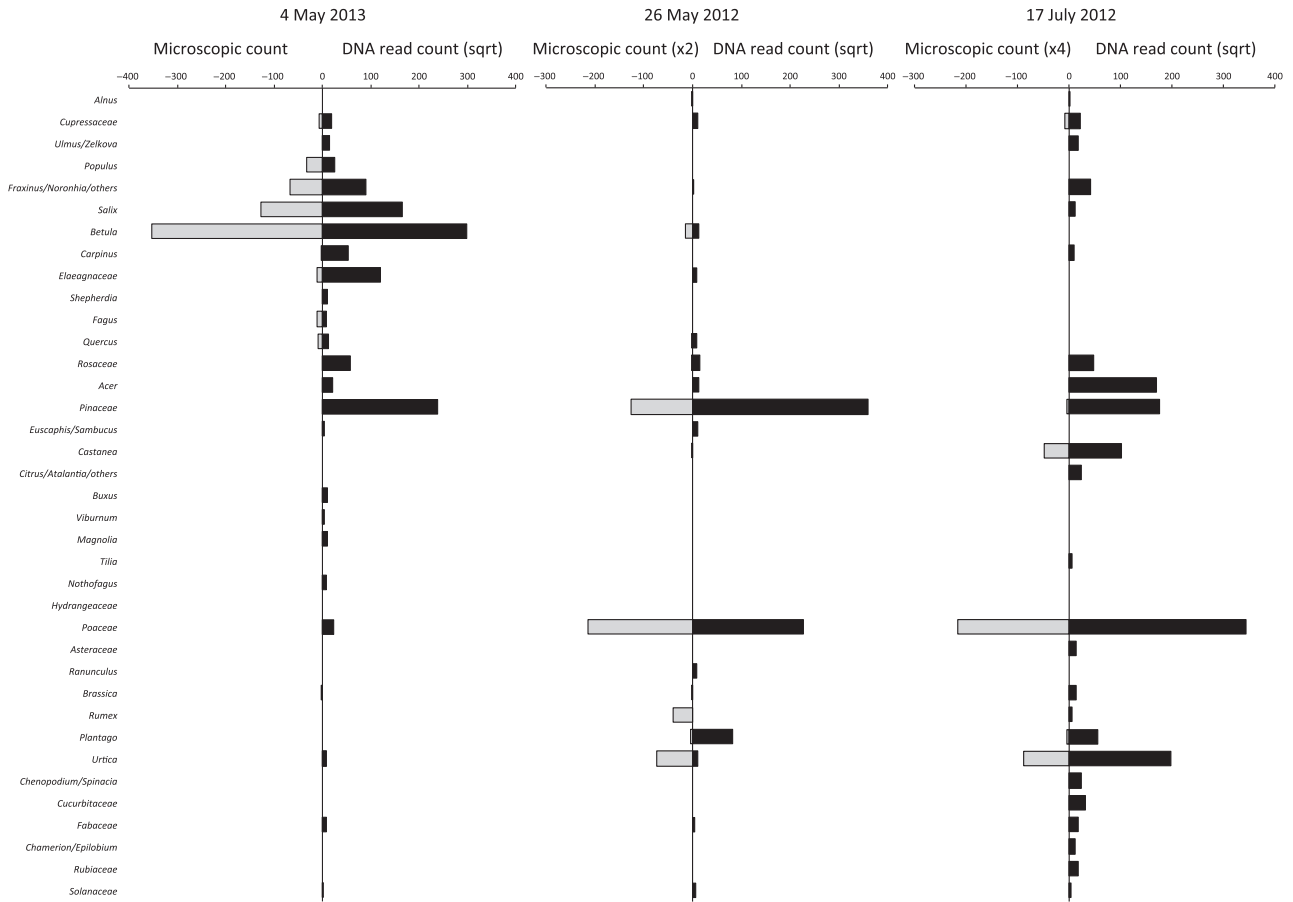


Fig. 1 Comparison between the number of pollen counted microscopically and the blast analysis of the Ion Torrent sequence reads for airborne pollen samples obtained on 3 days. DNA read counts were lumped for taxa that could not be distinguished microscopically (indicated in Table S1, Supporting information). For visualization purposes only, the number of pollen grains counted microscopically is multiplied by 0, 2 or 4 and turned negative, and the number of Ion Torrent sequence reads is square-root-transformed. Genera hits by less than 50 reads in any sample and not observed in any of the microscopic counts are omitted for this figure. For full data, see Table S1 (Supporting information).

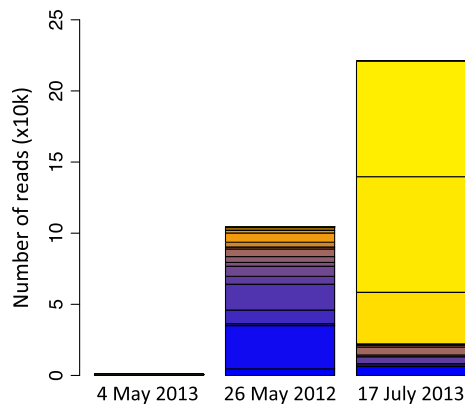


Fig. 2 Number of Ion Torrent reads from three samples of airborne pollen hitting different grass genera. For illustrative purposes, each colour represents a different grass genus. See main text for details and Table S1 (Supporting information) for full data.

pooled sample yielded even more genera that were not identified microscopically than the smaller single-PCR data set. There was no significant difference in the number of sequences per genus between the eight PCR amplicons (100% blast hits: GLM $F_{7,608} = 0.06$, $P = 0.99$). There was also no significant difference in the number of genera identified in each of the eight PCR amplicons ($\chi^2 = 1.37$, $P = 0.99$). Last, the actual genera identified in each of the eight PCR amplicons was very similar: every genus that was present with 100 sequences or more in at least one of the eight amplicons was also identified in each of the other reactions.

Discussion

We show that DNA sequence analysis can be used to provide a detailed and accurate assessment of the species composition of samples of airborne pollen obtained

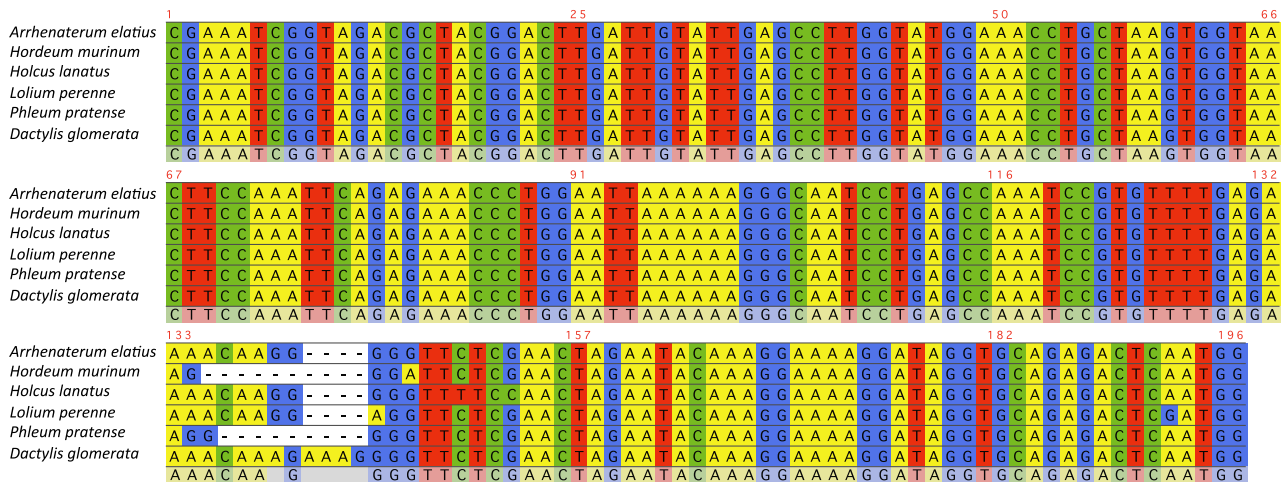


Fig. 3 Alignment of the reference sequences for the six grass species used in the grass mix. The sequences were generated using Sanger sequencing from the individual pollen samples before they were mixed.

Table 3 Blast hits of Ion Torrent sequence reads obtained from a mixed grass pollen sample of known content

Genus	blast_98	blast_99	blast_100
<i>Holcus</i>	145 316	99 740	0
<i>Festuca</i>	57 816	55 970	43 515
<i>Hordeum</i>	54 600	47 398	25 781
<i>Phleum</i>	43 904	40 662	25 781
<i>Dactylis</i>	42 358	38 537	27 833
<i>Puccinellia</i>	22 562	21 079	12 892
<i>Alopecurus</i>	21 885	20 326	12 892
<i>Arrhenatherum/Poa/others</i>	17 234	11 752	8782
<i>Festuca/Lolium/Dielsiochloa</i>	14 095	13 368	10 360
<i>Festuca/Lolium</i>	8285	8005	6216
<i>Avena</i>	5229	3652	27
<i>Helictotrichon</i>	1360	1035	32
<i>Poa</i>	714	357	262
<i>Elymus/Eremopyrum/others</i>	511	263	37
<i>Agrostis/Calamagrostis/others</i>	341	156	0
<i>Elymus/Leymus</i>	340	194	22
<i>Agropyron/Elymus/others</i>	283	102	11
<i>Pinus</i>	236	236	224
<i>Juglans/Pterocarya</i>	77	77	61
<i>Arctophila/Dupontia</i>	28	10	0
<i>Elymus</i>	21	15	2
<i>Phalaris/Poa</i>	17	1	0
<i>Plantago</i>	15	15	15
<i>Henrardia</i>	9	0	0
<i>Quercus</i>	9	9	9
<i>Phalaris</i>	6	0	0
Total	437 251	362 960	174 753

Shown are the numbers of reads hitting the genera in column 1 at 98%, 99% and 100% identity. Four species that were identified unambiguously are indicated in dark grey. Species that could not be distinguished from species present in the sample are indicated in light grey.

using routine techniques. Extracting DNA from pollen is notoriously difficult because the nucleated cells are surrounded by the resilient cellulose-rich intine and sporopollenin-containing exine (Baldi *et al.* 1987; Tanaka *et al.* 1987; Simel *et al.* 1997). However, we show that 97–99% of pollen is lysed using a bullet blender. We only sequenced about 200 bp of the *trnL* locus, which is located on the chloroplast DNA and is sufficiently variable to identify most plants to genus level (Taberlet *et al.* 2007). Our method relies on generic laboratory and bioinformatics techniques and does not require trained personnel able to identify pollen microscopically. Furthermore, it provides greater taxonomic detail than microscopic counting, which can only identify families in many cases. Although our method does not provide information on the actual allergens, we believe that these advantages over standard pollen monitoring will benefit the allergic population.

Allergenic symptoms are related to absolute, rather than relative allergen levels. It is therefore important to estimate the absolute number of each pollen type. However, the number of reads does not translate directly to the number of DNA molecules in the sample, because of the PCR amplification steps used (Deagle *et al.* 2013; Kelly *et al.* 2014). As a result, the slope of the relationship between the number of reads and the number of pollen counted microscopically was different for each of the three samples. PCR-free DNA-based taxon identification methods are being developed that may solve this issue (Zhou *et al.* 2013). For now, this problem is solved easily by correcting the number of reads by the total number of microscopically counted pollen. Thus, a simple total pollen count (without any species identification), followed by our NGS-based method, should give an accurate

estimate of the absolute number of each pollen type in the sample. Total pollen can be counted using image recognition, and therefore, this extra step could be automated.

A number of sources of bias could affect the accuracy of NGS-based pollen identification. First, sequence polymorphism in the primer sites could cause some plant species to be amplified less efficiently than others. This appears to be the case for *Urtica*, the *trnL* sequence of which showed a mismatch with our reverse primer. This could explain why this species was found in low numbers in the sequence data on 26 May 2012 when high numbers were counted microscopically. Using another primer site, or designing a primer with a degenerate nucleotide at the SNP position, could easily overcome this problem. Second, our method relies on the *trnL* sequence being present in the database. This appeared to pose few problems in our study, as plants from western Europe are well covered. However, incompleteness of the *trnL* database might hamper DNA-based pollen monitoring in other parts of the world. Still, some peculiarities in our results may be the result of database incompleteness. For example, the sequence data identified large numbers of *Fraxinus*-like pollen on both 4 May and 17 July. While the former is expected, the latter is highly unlikely based on the flowering time of *Fraxinus*. Closer inspection revealed that the majority (1167 of 1723) of reads assigned to *Fraxinus* in the 4 May sample had a stretch of 6 adenines near the 3' end of the sequence. By contrast, this stretch consisted of 7 adenines in 2520 of 2850 reads assigned to *Fraxinus* in the 17 July sample, indicating that the latter actually belonged to a different species that was not in the database. With massive sequencing projects going on, the database will soon be much richer in reference sequences (Erickson & Kress 2012). Also, it is straightforward to add relevant species to the database for a geographical region of interest, as we have done for grasses (see data accessibility section). Third, the number of chloroplasts may vary between pollen. Chloroplasts can be present in the vegetative or in both vegetative and generative cells of pollen grains, although they are degraded upon fertilization in most angiosperms, ensuring maternal inheritance (Sodmergen *et al.* 1992). We know of no estimates for chloroplast numbers per pollen grain. Detailed studies are needed to assess whether this is an important source of bias in the sequence data. Data provided by such a study could be used to calibrate the sequence data. Fourth, the DNA that we detect with our method can be derived from intact pollen but also from cytoplasmic debris and cell wall fragments from damaged pollen on the strip. As both the damaged pollen and intact pollen contain allergen or have released allergens (Grote 1999; Taylor *et al.* 2004), our DNA-based method reflects the amount of allergen

that can be produced. Fifth, stochasticity during PCR may affect the results. However, our mixed amplicon experiment demonstrates that the results are highly repeatable. Pooling amplicons for the same sample, as recommended for PCR of environmental samples using degenerate primers (Thomsen *et al.* 2012; Yu *et al.* 2012; Leray *et al.* 2013; Schmidt *et al.* 2013), did not improve the accuracy of the outcome. Last, reads with the same sequence may represent PCR duplicates, rather than true duplicates. PCR can bias the representation of input molecules in the final sequencing library, especially when many cycles are run. We are currently working on a modification to our laboratory protocol to resolve this issue.

For identification, we compared each sequence read to the NCBI database using *blastn*. This is a computationally demanding process that requires the entire NCBI database to be installed locally. While this ensures that the search is as comprehensive as possible, it is not a feasible strategy for routine pollen monitoring, or for research groups that do not have easy access to high-power computing facilities. Numerous alternative strategies for taxonomy assignment exist (Coissac *et al.* 2012; Van Velzen *et al.* 2012; Tanabe & Toju 2013). A fast and light-weight strategy would be to create a database of *trnL* sequences and align the reads to that using dedicated alignment software.

We show that our method can reliably identify pollen that cannot be distinguished microscopically. This was especially striking for grasses, a large and diverse family that contributes many allergenic species (De Weger *et al.* 2011). While microscopic methods do not allow reliable distinction between pollen from different grass species, our NGS-based method unambiguously identified several genera. Four of the six species in our sample of mixed grass pollen were identified correctly. The short length of the DNA fragment we sequenced prevented distinction between, for example, *Festuca* and *Lolium*, a problem that should be solved when longer fragments are sequenced. Despite this shortcoming, the NGS-based method correctly identified several grasses that flower at different times of year in the airborne pollen samples.

Methods are being developed to measure allergen levels in ambient air. While only a small number of allergens can be measured at present, the measured allergen levels do not always correspond well to the pollen counts (Buters *et al.* 2012; Galan *et al.* 2013). Some of this variation may be due to different species or cultivars that cannot be identified by microscopic pollen counts (Cecchi 2013). For example, at least 11 groups of allergens have been isolated from grass pollen, which differ in the strength of the immune response they induce in patients (Andersson & Lidholm 2003). The prevalence of grass allergens varies over geographic regions (Barber *et al.* 2008), and grass species differ in their allergen

composition (Andersson & Lidholm 2003). The improved taxonomic identification of airborne pollen offered by our method will thus facilitate research on the link between, for example, grass-pollen-specific allergen and pollen levels. It needs to be noted that our DNA-based method does not provide insight into allergen gene expression levels.

We show that identification and quantification of airborne pollen using NGS is reliable and efficient. The ongoing technological developments are likely to allow further improvements to our method. For example, while we processed only three airborne pollen samples in this study, it will be easy to scale up the process. Many samples can be run in parallel, for example, in 96- or 384-well plates instead of in single tubes. Certain components or the process could also be automated, further increasing throughput. Our method used Ion Torrent for sequencing mixed DNA samples. At the time of our experiments, the read length was limited to about 300 bp. Further developments in NGS technology will lead to longer read lengths, which will allow longer *trnL* fragments to be sequenced, which will increase the discriminatory power of our method. Adding additional loci to the analysis could also increase the resolution. With the rapidly falling costs of NGS, our pollen monitoring method has the potential to facilitate considerable upscaling of pollen monitoring networks. Individual monitoring stations could avoid the initial investment costs by concentrating their DNA analyses in a centralized facility. Other applications are also possible. For example, the method could be used to (re-) analyse archived pollen samples or archaeological samples. The method could also easily be adapted to analyse other organisms including bacteria and fungal bioaerosols.

Acknowledgements

This work was supported by grants from the Centre for Medical Systems Biology and the Sequencing Analysis Support Core of the Leiden University Medical Center. Rolf Vossen helped with the molecular work and Yahya Anvar with the bioinformatics. We thank Isa Al Sawad, Merlina Kranenburg, Patrick Wijntjes, Nathalie van der Veen and Rieven Martis (students University of Applied Sciences Leiden) and Linda van der Steen for their help in the field, the laboratory and behind the computer.

References

Andersson K, Lidholm J (2003) Characteristics and immunobiology of grass pollen allergens. *International Archives of Allergy and Immunology*, **130**, 87–107.

Anvar SY, van der Gaag KJ, van der Heijden JWF *et al.* (2014) tssv: a tool for characterization of complex allelic variants in pure and mixed genomes. *Bioinformatics*, doi:10.1093/bioinformatics/btu068.

Baldi BG, Franceschi VR, Loewus FA (1987) Preparation and properties of pollen sporoplasts. *Protoplasma*, **141**, 47–55.

Barber D, de la Torre F, Feo F *et al.* (2008) Understanding patient sensitization profiles in complex pollen areas: a molecular epidemiological study. *Allergy*, **63**, 1550–1558.

Bellemain E, Davey ML, Kauserud H *et al.* (2013) Fungal palaeodiversity revealed using high-throughput metabarcoding of ancient DNA from arctic permafrost. *Environmental microbiology*, **15**, 1176–1189.

Bousquet J, Dahl R, Khaltayev N (2007) Global alliance against chronic respiratory diseases. *Allergy*, **62**, 216–223.

Buters JTM, Thibaudon M, Smith M *et al.* (2012) Release of Bet v 1 from birch pollen from 5 European countries. Results from the HIALINE study. *Atmospheric Environment*, **55**, 496–505.

Cecchi L (2013) From pollen count to pollen potency: the molecular era of aerobiology. *The European Respiratory Journal*, **42**, 898–900.

Chariton AA, Court LN, Hartley DM, Colloff MJ, Hardy CM (2010) Ecological assessment of estuarine sediments by pyrosequencing eukaryotic ribosomal DNA. *Frontiers in Ecology and the Environment*, **8**, 233–238.

Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular ecology*, **21**, 1834–1847.

D'Amato G, Liccardi G, D'Amato M, Holgate S (2005) Environmental risk factors and allergic bronchial asthma. *Clinical and Experimental Allergy: Journal of the British Society for Allergy and Clinical Immunology*, **35**, 1113–1124.

De Weger LA, Beerthuizen T, Gast-Strookman JM *et al.* (2011) Difference in symptom severity between early and late grass pollen season in patients with seasonal allergic rhinitis. *Clinical and Translational Allergy*, **1**, 18.

De Weger LA, Beerthuizen T, Hiemstra PS, Sont JK (2013) Development and validation of a 5-day-ahead hay fever forecast for patients with grass-pollen-induced allergic rhinitis. *International journal of biometeorology*, DOI: 10.1007/s00484-013-0692-5, [Epub ahead of print]

Deagle BE, Thomas AC, Shaffer AK, Trites AW, Jarman SN (2013) Quantifying sequence proportions in a DNA-based diet study using Ion Torrent amplicon sequencing: which counts count? *Molecular Ecology Resources*, **13**, 620–633.

Epp LS, Boessenkool S, Bellemain EP *et al.* (2012) New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. *Molecular ecology*, **21**, 1821–1833.

Erickson DL, Kress WJ (2012) Future directions. *Methods in Molecular Biology*, **858**, 459–465.

Fonseca VG, Carvalho GR, Sung W *et al.* (2010) Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature Communications*, **1**, 98.

Galan C, Antunes C, Brandao R *et al.* (2013) Airborne olive pollen counts are not representative of exposure to the major olive allergen Ole e 1. *Allergy*, **68**, 809–812.

Galán C, Emberlin J, Domínguez E, Bryant RH, Villamandos F (1995) A comparative analysis of daily variations in the Gramineae pollen counts at Córdoba, Spain and A comparative analysis of daily variations in the Gramineae pollen counts at Cordoba, Spain and London, UK. *Grana*, **34**, 189–198.

Grote M (1999) In situ localization of pollen allergens by immunogold electron microscopy: allergens at unexpected sites. *International archives of allergy and immunology*, **118**, 1–6.

Hajibabaei M, Shokralla S, Zhou X, Singer GAC, Baird DJ (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE*, **6**, e17497.

Hollingsworth PM (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences, USA*, **106**, 12794–12797.

Hollingsworth PM, Graham SW, Little DP (2011) Choosing and using a plant DNA barcode. *PLoS ONE*, **6**, e19254.

Hrabina MÅ, Peltre G, Van Ree R, Moingeon PÅ (2008) Grass pollen allergens. *Clinical and Experimental Allergy Reviews*, **8**, 7–11.

Ji Y, Ashton L, Pedley SM *et al.* (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology letters*, **16**, 1245–1257.

Kawashima S, Clot B, Fujita T, Takahashi Y, Nakamura K (2007) An algorithm and a device for counting airborne pollen automatically using laser optics. *Atmospheric Environment*, **41**, 7987–7993.

- Kelly RP, Port JA, Yamahara KM, Crowder LB (2014) Using Environmental DNA to Census Marine Fishes in a Large Mesocosm (GE Hofmann, Ed.). *PLoS ONE*, **9**, e86175.
- Landsmeer SH, Hendriks EA, de Weger LA, Reiber JHC, Stoel BC (2009) Detection of pollen grains in multifocal optical microscopy images of air samples. *Microscopy research and technique*, **72**, 424–430.
- Leray M, Yang JY, Meyer CP *et al.* (2013) A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, **10**, 34.
- Longhi S, Cristofori A, Gatto P *et al.* (2009) Biomolecular identification of allergenic pollen: a new perspective for aerobiological monitoring? *Annals of allergy, asthma & immunology*, **103**, 508–514.
- Pompanon F, Deagle BE, Symondson WOC *et al.* (2012) Who is eating what: diet assessment using next generation sequencing. *Molecular ecology*, **21**, 1931–1950.
- R Development Core Team (2010) *R: A Language and Environment For Statistical Computing*. R Development Core Team, Austria, Vienna.
- Rittenour WR, Hamilton RC, Beezhold DH, Green BJ (2012) Immunologic, spectrophotometric and nucleic acid based methods for the detection and quantification of airborne pollen. *Journal of immunological methods*, **383**, 47–53.
- Rogers SL, Atkins SD, West JS (2009) Detection and quantification of airborne inoculum of *Sclerotinia sclerotiorum* using quantitative PCR. *Plant Pathology*, **58**, 324–331.
- Scheifinger H, Belmonte J, Buters J *et al.* (2013) Monitoring, modelling and forecasting of the pollen season. In: *Allergenic Pollen: A Review of the Production, Release, Distribution and Health Impacts* (eds Sofiev M & Bergmann K-C), pp. 71–126. Springer, Netherlands, Dordrecht.
- Schmidt P, Bálint M, Greshake B *et al.* (2013) Illumina metabarcoding of a soil fungal community. *Soil Biology & Biochemistry*, **65**, 128–132.
- Simel EJ, Saidak LR, Tuskan GA (1997) Method of extracting genomic DNA from non-germinated gymnosperm and angiosperm pollen. *Bio-Techniques*, **22**, 390–394.
- Sodmergen, Suzuki T, Kawano S *et al.* (1992) Behavior of organelle nuclei (nucleoids) in generative and vegetative cells during maturation of pollen in *Lilium longiflorum* and *Pelargonium zonale*. *Protoplasma*, **168**, 73–82.
- Taberlet P, Gielly L, Pautou G, Bouvet J (1991) Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant molecular biology*, **17**, 1105–1109.
- Taberlet P, Coissac E, Pompanon F *et al.* (2007) Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. *Nucleic acids research*, **35**, e14.
- Tanabe AS, Toju H (2013) Two new computational methods for universal DNA barcoding: a benchmark using barcode sequences of bacteria, archaea, animals, fungi, and land plants. *PLoS ONE*, **8**, e76910.
- Tanaka I, Kitazuma C, Ito M (1987) The isolation and culture of lily pollen protoplasts. *Plant Science*, **50**, 205–211.
- Taylor PE, Flagan RC, Miguel AG, Valenta R, Glovsky MM (2004) Birch pollen rupture and the release of aerosols of respirable allergens. *Clinical and Experimental Allergy: Journal of the British Society for Allergy and Clinical Immunology*, **34**, 1591–1596.
- Thomsen PF, Kielgast J, Iversen LL *et al.* (2012) Detection of a diverse marine fish fauna using environmental DNA from seawater samples. *PLoS ONE*, **7**, e41732.
- Van Velzen R, Weitschek E, Felici G, Bakker FT (2012) DNA barcoding of recently diverged species: relative performance of matching methods. *PLoS ONE*, **7**, e30490.
- Yoccoz NG, Bråthen K, a, Gielly L, *et al.* (2012) DNA from soil mirrors plant taxonomic and growth form diversity. *Molecular ecology*, **21**, 3647–3655.
- Yu DW, Ji Y, Emerson BC *et al.* (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.
- Zhou X, Li Y, Liu S *et al.* (2013) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *Gigascience*, **2**, 4.

K.K. supervised the study, conducted part of the data analysis and wrote the manuscript; L.A.dW. supervised the study, conducted the microscopic pollen counting and provided the samples; M.V.G. conducted the laboratory work, H.B. supervised the laboratory work and conducted part of the data analysis; J.F. conducted part of the data analysis; and P.S.H. and J.T.dD. conceived and supervised the study.

Data Accessibility

DNA sequences: GenBank Accession no. KF876179.

Ion Torrent read sets: sequence read archive (SRA) Accession no. SUB526939.

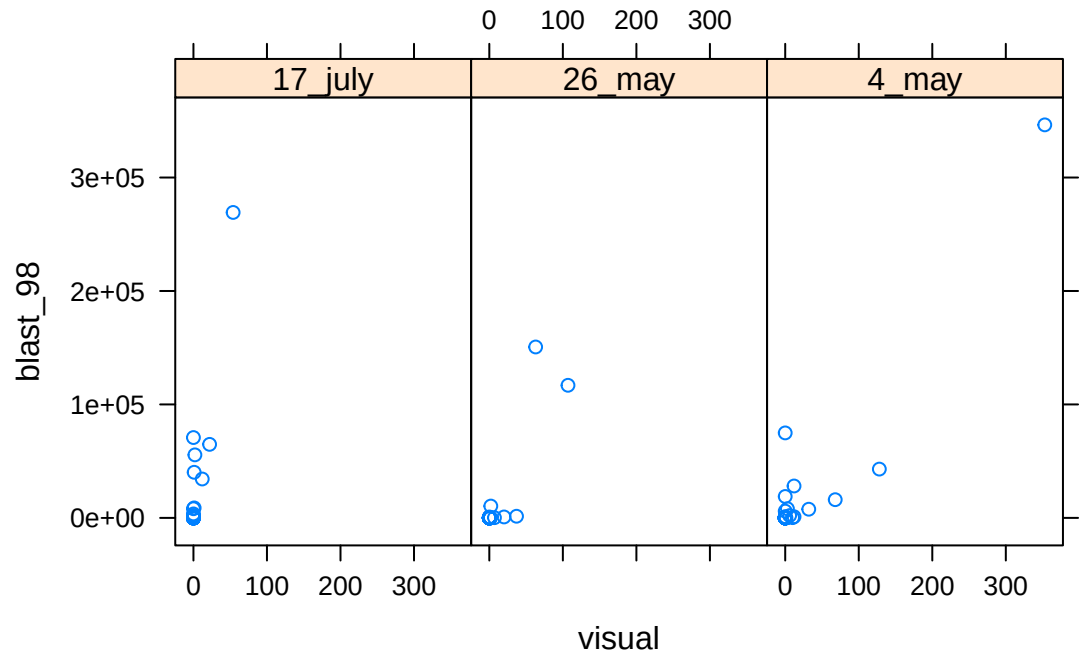
Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1 Scatterplots of the correlation between DNA read counts and microscopic counts for the three test strips.

Fig. S2 Alignment of the reference sequences for the six grass species used in the grass mix. This file was used to construct figure 3.

Table S1 Complete overview of the blast hits and microscopic pollen counts for the three test strips.



CLUSTAL W (1.83) multiple sequence alignment

```

Arrhenaterum_el CGAAATCGGTAGACGCTACGGACTTGATTGTATTGAGCCTTGGTATGGAAACCTGCTAAG
Hordeum_murinum CGAAATCGGTAGACGCTACGGACTTGATTGTATTGAGCCTTGGTATGGAAACCTGCTAAG
Holcus_lanatus   CGAAATCGGTAGACGCTACGGACTTGATTGTATTGAGCCTTGGTATGGAAACCTGCTAAG
Lolium_perenne   CGAAATCGGTAGACGCTACGGACTTGATTGTATTGAGCCTTGGTATGGAAACCTGCTAAG
Phleum_pratense  CGAAATCGGTAGACGCTACGGACTTGATTGTATTGAGCCTTGGTATGGAAACCTGCTAAG
Dactylis_glomer  CGAAATCGGTAGACGCTACGGACTTGATTGTATTGAGCCTTGGTATGGAAACCTGCTAAG

```

```

Arrhenaterum_el TGGTAACTTCCAAATTCAGAGAAACCCTGGAATTAATAAGGGCAATCCTGAGCCAAATC
Hordeum_murinum TGGTAACTTCCAAATTCAGAGAAACCCTGGAATTAATAAGGGCAATCCTGAGCCAAATC
Holcus_lanatus   TGGTAACTTCCAAATTCAGAGAAACCCTGGAATTAATAAGGGCAATCCTGAGCCAAATC
Lolium_perenne   TGGTAACTTCCAAATTCAGAGAAACCCTGGAATTAATAAGGGCAATCCTGAGCCAAATC
Phleum_pratense  TGGTAACTTCCAAATTCAGAGAAACCCTGGAATTAATAAGGGCAATCCTGAGCCAAATC
Dactylis_glomer  TGGTAACTTCCAAATTCAGAGAAACCCTGGAATTAATAAGGGCAATCCTGAGCCAAATC

```

```

Arrhenaterum_el CGTGTTTTGAGAAAACAAGG----GGGTTCTCGAACTAGAATACAAAGGAAAAGGATAGG
Hordeum_murinum CGTGTTTTGAGAAG-----GGATTCTCGAACTAGAATACAAAGGAAAAGGATAGG
Holcus_lanatus   CGTGTTTTGAGAAAACAAGG----GGGTTTTCCAAGCTAGAATACAAAGGAAAAGGATAGG
Lolium_perenne   CGTGTTTTGAGAAAACAAGG----AGGTTCTCGAACTAGAATACAAAGGAAAAGGATAGG
Phleum_pratense  CGTGTTTTGAGAAG-----GGGTTCTCGAACTAGAATACAAAGGAAAAGGATAGG
Dactylis_glomer  CGTGTTTTGAGAAAACAAGGAAAAGGGGTTCTCGAACTAGAATACAAAGGAAAAGGATAGG

```

```

Arrhenaterum_el TGCAGAGACTCAATGG
Hordeum_murinum TGCAGAGACTCAATGG
Holcus_lanatus   TGCAGAGACTCAATGG
Lolium_perenne   TGCAGAGACTCGATGG
Phleum_pratense  TGCAGAGACTCAATGG
Dactylis_glomer  TGCAGAGACTCAATGG

```