Cluster oligonucleotide signatures for rapid identification by sequencing

Manuel Zahariev²*, Wen Chen¹*, Cobus M. Visagie⁴ and C. André Lévesque³

Abstract

Background: Oligonucleotide signatures (signatures) have been widely used for studying microbial diversity and function in wet-lab settings, but using them for accurate in silico identification of organisms from high-throughput sequencing (HTS) data is only a proof of concept. Existing signature design programs for sequence signatures (signatures matching exactly one sequence) or clade signatures (signatures matching every sequence in a phylogenetic clade) are not able to identify all possible polymorphic sites for sequences with high similarity and perform poorly when handling large genome sequencing datasets.

Results: We introduce cluster signatures: subsequences that match perfectly and exclusively any group of sequences in a data set. Cluster signatures provide complete recall for primer/probe design and increased discrimination between sequences beyond that of clade signatures. Using cluster signatures for in silico identification of HTS targets achieves good precision/recall and running time performance. This method has been implemented into an open source tool, the Automated Oligonucleotide Design Pipeline (adop), included in supplementary material and available at: https://bitbucket.org/wenchen_aafc/aodp_v2.0_release.

Conclusions: Cluster signatures provide a rapid and universal analysis tool to identify all possible short diagnostic DNA markers and variants from any DNA sequencing dataset. They are particularly useful in discriminating genetic material from closely related organisms and in detecting deleterious mutations in highly or perfectly conserved genomic sites.

Keywords: DNA hybridization, Oligonucleotide signatures, Metabarcoding, Metagenomics, Regulated pathogens

Background

Biodiversity research and survey require accurate identification of organisms from the environment, especially those of public concerns, e.g. quarantine species and select agents monitored by national biosafety and biosecurity programs. Identifying the sequences, e.g. DNA markers or genome regions, of concern in ecosystems is the fundamental strategy [1], especially in the metagenomics era which requires high-throughput processing without compromising accuracy and sensitivity.

A widely used strategy for taxonomic assignment of shotgun metagenomes or metabarcodes is to bin [2, 3] or cluster [4, 5] sequencing reads followed by comparing with reference databases using string search algorithms, which, as reviewed previously [6] either depend on alignment-based phylogenetic distances (homology-search), such as BLAST [7] and HMMER [8–13] or k-mer frequency profile (composition) comparison, such as USEARCH [4, 14–20]. These algorithms are implemented in "off-the-shelf" suites for classification of HTS data, such as ShortBRED [21], PanPhlAn [22], MIDAS [23] and mOTU [24], developed for taxonomic classification or for identifying gene homologs from HTS data.

Aligners using BLAST to map reads are very precise, but with high computational cost, while composition-based programs and aligners using suffix-prefix tries are fast but can be imprecise, compounding errors present in most HTS techniques. For example, the average classification accuracy for all fragments of 16S rRNA genes longer than 100 bp was 70% using the Ribosomal Database...
Project (RDP) Classifier, a text-based Bayesian classifier [17]. A recent study using the same classifier could classify metabarcodes of the 16rRNA genes to family and genus levels with accuracy 75% or lower [25]. Sigma [6] and Pathoscope [10, 26] are systems developed for subspecies and strain-level inference of metagenomics data, but are not applicable to metabarcoding data, since DNA barcodes are known to lack discriminating power for many taxon lineages [27–30]. For instance, the internal transcribed spacer 1 (ITS1) of Tilletia indica, a quarantine pathogen in many countries, and T. walkeri which is not regulated by most countries except for South Korea, differ only by two bases.

The Minimum Entropy Decomposition (MED) algorithm implemented in OligoTyping [31, 32] identifies information-rich polymorphic sites and iteratively partitions a set of metabarcodes to homogeneous operational taxonomic units (OTUs), until the Shannon entropy profile of a given node is converged or below a given threshold.

Eren et al. [31] stated that MED was able to discriminate taxa with less than 1% sequence variance and is computationally efficient. While MED is excellent in identifying distinct subgroups of a taxon adapted to specific environmental niches, it works best on abundant OTUs/taxa observed across diverse ecosystems, while many pathogens present as rare taxa in the environment. In addition, alignment is required prior to MED when differences in sequence length do not represent biologically meaningful variation, which can be a main constraint on efficiency when processing HTS reads not of the same length, e.g. quality trimmed Illumina data or 454 pyrosequencing data. While the discriminating positions identified by MED have the potential for strain-typing microorganisms, MED does not directly extract oligonucleotide signatures associated with these positions that may be used as primers or probes for the development of molecular diagnostic assays.

Oligonucleotide signatures (signatures) are short sequence strings (λ-mers) of fixed length (signature length λ), normally 18 to 100 bp, that match exactly and exclusively one or more targeted sequence(s) (targets) in a given genetic data set, usually from the same region of the genomes of targeted taxa. Most existing approaches only design sequence signatures, i.e. signatures for single sequences [33–35] or a single group of sequences per run [36, 37] as reviewed previously [38–40].

A few applications were developed to design signatures for pre-defined groups of genomes [41, 42], gene families [43] or clade signatures2, i.e. signatures for a single phylogenetic clade3 [38]. However, these applications either suffer from memory and runtime issues, or are part of larger, special purpose systems [39].

Algorithm 1: Sequence matching algorithm using $T_{λ,ψ}$, a table of clusters from a given training set: signature of length $λ$ → cluster (set) of matching sequences.

| Data: $T_{λ}$ | // signature → cluster of sequences |
| Input: $Q$ | // query sequence |
| $Ω$ ← $∅$ | // list of matching clusters |
| for $q ∈ Q$ subsequence of length $|q| = λ$ do |
| $Ω$ ← $Ω ∪ T_{λ}(q)$ | // add cluster $T_{λ}(q)$ |
| $Ψ$ ← $Ψ$ | // kernel of matching sequences |
| for $ω ∈ Ω$ do |
| $η$ ← $0$ | // portion of $Q$ covered by $ω$ |
| for $q ∈ Q$ where $ω ∩ T_{λ}(q) ≠ ∅$ do |
| $ω$ ← $ω ∩ T_{λ}(q)$ |
| $η ← η + 1$ |
| if $η/|Q| ≥ γ$ then $Ψ ← Ψ ∪ ω$ | // constant $γ = 0.75$ |
| for $ψ ∈ Ψ$ do |
| $α(ψ, Q)$ ← align($ψ, Q$) | // pairwise alignment |
| Output: $Ω = \arg \max_{ψ∈Ψ} α(ψ, Q)$ | // all sequences $ψ ∈ Ψ$ that maximize $α(ψ, Q)$ |

In addition, phylogenetic clades and other a priori groupings can be very restrictive to the identification of viable signatures, which may be caused by conflicting phylogenetic signals among loci shared by different taxonomic domains [44] as found in our own studies [45, 46]. This restrictiveness is further compounded by additional experimental constraints, such as primer/probe melting temperature [47] or Kane’s conditions [48–50].

Signatures have wide applications in the biological field, such as being used as primers and probes in PCR and DNA-hybridization [40, 46, 51–53] or lab-on-a-chip detection methods, as well as in targeted enrichment methods for focused high-throughput sequencing (HTS) [54, 55]. Kallisto [56] uses signatures (of length $k$: kmers) from RNA-Seq reads to create $k$-compatibility classes, whose intersection represents the set of possible sequences matching a given read. Similarly, Salmon [57] builds equivalence classes over fragments of reads (in effect signatures), from which it infers statistically the relative abundance of transcripts. Kraken [58] infers the taxonomic classification for HTS reads by building a database of phylogenetic lowest common ancestors using clade signatures. We show further that a significant number of signatures cross clade boundaries. We also show that while signatures work very well on perfectly preserved reads, they are brittle to errors introduced by the HTS process.

Signatures have also been used to detect pathogenic microbes from metagenomics sequencing data. This theoretical approach, termed Electronic probe Diagnostic Nucleic acid Analysis (EDNA) [59, 60], shows promising research and diagnosis direction (75% precision on a mock database) using shotgun metagenomics data, but relies on a priori groupings of the training data set (reference genomes), and a priori differentiation against false positives identified using near neighbor comparisons in a
reference database. EDNA also depends on an external program for signature design, the Tool for Oligonucleotide Fingerprint Identification (TOFI) [42], which introduces runtime efficiency constraints. A system that can streamline this process would be ideal as a regulatory tool in pest detection and management.

We present here a research tool for unrestricted design of signatures that can be used for the detection of any kind of groups (mutants, species/subspecies, or any a priori groupings) in a wide range of molecular biology assays or DNA sequence data for in silico probing.

**Methods**

The usefulness of signatures is based on the low probability of accidental match between a signature and unrelated genetic material. The probability \( p \) of an accidental match (collision) between two 4-base nucleotide strings of length \( \lambda \) (\( 4^\lambda \) possibilities) in a data set of size \( N \) can be modeled by the birthday formula [61]:

\[
p = \frac{4^\lambda!}{4^N (4^\lambda - N)!} \approx 1 - e^{-\frac{N^2}{2\lambda^2}} \text{ for } N \ll 4^\lambda \quad (1)
\]

Accidental matches between portions of nucleotide strings occur in random genetic material when there is no taxonomic or functional relationship between the query and testing sequences. For example, it is likely to encounter the subsequence of length \( \lambda = 4 \) "ACGT" multiple times in different sequences in any given large reference database. The birthday formula quantifies the probability of such accidental matches.

Assuming an uniform distribution of nucleotides and signature length \( \lambda = 36 \), \( p < 10^{-4} \) applies for data sets with \( N < 2.94 \times 10^{10} \) nucleotides (approximately 274 GiB of unaligned FASTA files). In practice, data sets of taxonomically related DNA have a higher degree of similarity between sequences, which increases the probability that any two identical subsequences have a taxonomic or functional relationship between them and do not represent accidental matches. Unless explicitly specified, signatures of length \( \lambda = 36 \) are used for analyses in this study.

**Clusters**

We introduce an extension of clade signatures: for a given set of sequences, a **cluster** is a group of sequences for which at least one signature (**clade signature**) can be found, that matches all sequences in the group but does not match any sequences not in the group.

Notably, clusters are not required to represent the same groups as those in phylogenetic clades; they are any groups of sequences for which signatures can be found, as opposed to clade signatures for predefined phylogenetic groups. Any subsequence of length \( \lambda \) of any sequence is a signature for exactly one cluster, i.e. a cluster signature.

It is not obvious how to predict the number of clusters expected for a given data set of taxonomically related sequences. For example, \( S \) identical sequences of length \( L_i > \lambda \) where \( 1 \leq i \leq S \), will generate exactly 1 cluster. Since any subsequence of length \( \lambda \) from any sequence can be found in every other sequence, the single cluster will contain all sequences. By contrast, a data set with size \( N < 4^\lambda \) containing \( S \) sequences randomly generated using a uniform distribution of nucleotides will have \( S \) clusters of signatures of length \( \lambda \). Each such cluster will contain one sequence, since it is very likely that every subsequence of length \( \lambda \) in any sequence is a sequence signature: it will not be found anywhere else in the data set \( p < 10^{-4} \) (Eq. 1).

**The automated oligonucleotide design pipeline**

We have built an open source tool, **aoodp**, the Automated Oligonucleotide Design Pipeline (**aoodp** v. 2.5 is included in the supplementary material), which generates efficiently signatures for sequences, clades and clusters by enumerating all \( \lambda \)-mers (signatures) for each sequence in a given data set. The list of originator sequences is collected for each enumerated signature. All distinct sets of originator sequences for all signatures form the list of clusters for the data set. Facilities for enumerating and cross-referencing signatures and clusters are provided.

Furthermore, **aoodp** can be used to find the closest matching sequences from a training set to a query sequence, assumed to be an imperfectly recovered portion of an unknown sequence (such as an HTS read) by computing the union of all sequences in all clusters matching any portion of the HTS read and then heuristically eliminating all but sequences that explain the largest portion of the HTS read. All remaining sequences are then compared to the HTS read and only the ones with the highest overlap score are kept.

More formally, the matching algorithm 1 works as follows: first, compute a set of matching clusters \( \Omega \) for each query sequence \( Q \) (loop 1): we observe later (Table 3) that the set of all training sequences contained in all matching clusters \( \Omega \) has average size \( \bar{\Omega} < S \) smaller than the size of the training set; second, minimize a subset (kernel) \( \Psi \) of \( \Omega \) (loop 2): we observe that the average size of the kernel \( \bar{\Psi} \ll S \) is much smaller than the size of the training set; and, finally, compute the sequence similarity of each training sequence in the kernel \( \Psi \) against the query sequence (loop 4) using a global alignment algorithm [62].

The result (the set of mapped HTS reads) is the subset of sequences of \( \Psi \) which maximize the alignment score to the query sequence.

The main objective of the algorithm is to minimize the number of computationally expensive global alignments (step 5). The complexity has no direct dependency on the size of the training set: loops 1 and 3 have complexity
$O(|Q|)$ linear in the size of the query sequence and loop 2 has complexity $O(|Ω|)$ linear in the number of clusters matching the query sequence, which can be further reduced at the implementation level through the elimination of repeated set operations in loop 3.

A general limitation of algorithms for matching HTS reads (including our method) is that metabarcoding regions used in HTS do not always have sufficient discriminating power to differentiate very closely related species [28] represented by clades in a training dataset containing almost identical sequences, which, however, belong to multiple valid species. In this case, all matched reference sequences are given to a query sequence, and it should be the users’ decision if alternative DNA markers or wet lab molecular diagnostic assays should be used to confirm or validate the existence of targeted taxa of interest.

**Data sets**

Data sets for four important mycotoxin genera (*Alternaria*, *Aspergillus*, *Claviceps* and *Penicillium*) were built using the following methodology: internal transcribed spacer rDNA region (ITS) data sets were compiled from GenBank [63] using ex-type sequences of 16S rRNA genes of a bacterium (*Pectobacterium*) were downloaded from GenBank. The ITS dataset *Phytophthora* was obtained from [46]. The sequences for each dataset were aligned in MAFFT v. 7.305b [69], using the G-INS-i algorithm and trimmed manually in Geneious v. 8.1.8. Neighbour-joining trees were calculated in PAUP* v. 4.0b10 [70].

Reference ITS sequences for fungi (*Anisogramma*, *Ceratobasidium*, *Ceratocystis*, *Colletotrichum*, *Coniella*, *Diaporthe*, *Fusarium*, *Elsinoe*, *Talaromyces*, *Tilletia*), oomycetes (*Peronospora*), as well as the 16S rRNA genes of a bacterium (*Pectobacterium*) were downloaded from GenBank. The ITS dataset *Phytophthora* was obtained from [46]. The sequences for each dataset were aligned using the G-INS-i algorithm in MAFFT [69], and trimmed manually in BioEdit v.7.2.5 [71]. The approximate maximum likelihood trees were reconstructed using FastTree v.2.1.8 [72].

Each data set contains DNA sequences and a phylogenetic tree with the sequences as leaf nodes. The data sets were combined into a sequence database *17DataSets*, provided as supplementary material. Sequences with more than five ambiguous bases were removed from each data set. The characteristics of each data set are summarized in Table 1.

The distribution of cluster size and number of cluster signatures was also studied on a much larger dataset (*Unite*; included in the supplementary material) of 271,017 sequences fully identified down to the species level and which include an authoritative Latin binomial name for each species. The data set was extracted from the UNITE+INSD database released by the User-friendly Norwegian ITS Ectomycorrhiza Database (UNITE, version 7.1) [73]. A phylogenetic tree was automatically built from the Unite taxonomy using tax2newk, a companion utility of acdαp.

The sequence matching functionality was evaluated using a training set of 1,338 mycotoxin sequences (4Mycotoxins; included in the supplementary material) by combining the data sets *Alternaria*, *Aspergillus*, *Claviceps* and *Penicillium*. Sequences from each data set not classified to the principal genus of the data set and/or with more than five ambiguous bases were eliminated.

The precision and recall of the matching algorithm were evaluated using a testing set *4MycotoxinsBootstrap* bootstrapped from *4Mycotoxins*: subsequences of exactly $|Q| = 100$ bp starting at a random position were extracted from each sequence. In each subsequence, each nucleotide was modified to another nucleotide or a gap. Individual modifications were made at one of six error rates: $ε ∈ \{0.00, 0.01, 0.02, 0.03, 0.04, 0.05\}$. For each sequence and each error rate, 10 subsequences were generated. All random choices were drawn from uniform distributions driven by a Mersenne twister [74], seeded with a high resolution timestamp.

The efficiency of the matching algorithm was evaluated on a testing set *97AerobiotaSamples* containing 4,713,791 sequences (sequence length $|Q| ≈ 436$ bp ± 55; only sequences at least 325 bp long are selected) from a data set deposited in the Sequence Read Archive (SRA) under project accession number PRJNA358221. The error rate assigned to the data set was $ε = 0.01$ [75].

**Comparisons with other algorithms**

We have compared the computational efficiency of our matching algorithm with BLAST+ v.2.6.0 [7] and USEARCH v10.0.240_i86linux32 [4] testing on the *97AerobiotaSamples* data set and using the *4Mycotoxins* reference data set. All test runs were conducted on a system with Intel Core i7-3632QM CPU 2.20GHz ×8 running Ubuntu 16.04.

The following parameters were used for BLAST:

```
-word_size 11 -outfmt 6 -num_threads 8 -evalue 10 -max_target_seqs 100
```

The following parameters were used for USEARCH:

```
-usearch_global -strand plus -id 0.98 -maxaccepts 256 -maxrejects 1024 -wordlength 8 -blast6out
```

In all instances, output was ignored (redirected to /dev/null) in order to eliminate I/O contention. Separately, we compared precision and recall (Eqs. 2 and 3) of our matching algorithm with USEARCH, on the *4MycotoxinsBootstrap* dataset using the *4Mycotoxins* reference set.

For USEARCH we used the following parameters:

```
-usearch_global -strand plus -wordlength 8 -blast6out
```

```
Additionally, the “-id” parameter was set to 1 - 2ε to correspond to the error rate of the data set, “maxaccepts” χ was varied for different runs χ ∈ {4, 16, 64, 256, 1024} and “maxrejects” was set to 32 × χ.

For both USEARCH and aocdp, the match between a query sequence and a training sequence was considered correct if it is returned by the tool, and it has the highest percentage overlap compared to all other matching training sequences.

### Results

Large scale dependencies for the number of clusters were measured on the data set Unite. Most clusters have a relatively small number of sequences (Fig. 1): 85% have less than 100 sequences, 50% have less than 10 sequences and approximately 15% have one sequence (signable sequences). Clusters have a relatively small number of signatures (Fig. 2): 65% have less than 10 signatures and almost 30% have exactly one signature.

Other dependencies for the number of clusters are measured on the 17DataSets database. The number of clusters c is found to be comparable with the number of phylogenetic clades n = S + l in each of the data sets (0.7 ≤ c/n ≤ 3.6). Power law dependencies on the size of the data set N for the number of clusters c and number of signable clades n* are indicated by a log-log plot (Fig. 3).

A power law dependency of the number of clusters c on the number of signable clades n* is indicated by regression lines.

The dependency of the number of clusters and signable clades on signature size 12 ≤ λ ≤ 252 (increments of 4 nucleotides) is measured for the data set Penicillium (Fig. 4). The number of clusters c decreases rapidly with the signature length λ, because of the further reduction of the number of signatures in each cluster. The number of signable clades is relatively stable (slow initial increase).

### Clusters for probe design

Characteristics of clusters, signable clades and signable sequences were calculated in aggregate for all data sets and reported in Table 1. An incidence matrix for sequences (vertical axis) against clusters (horizontal axis) for the Ceratohiza data set is shown in Fig. 5. Signed sequences and internal signable are grouped in regions to the left of the figure.

The number of signable clades is smaller than the number of clades n* < n, in some cases substantially, for example n*/n = 17% for Alternaria.

The number of sequences s0 that are not contained in any signable clades can be substantial. This

### Table 1 Data sets included in database 17DataSets

| Data set       | N  | S  | i  | n  | L ± σ(L) | n* | c/n | n*/n | s0  | s1  | s2  | s3  | α  |
|----------------|----|----|----|----|----------|----|-----|-----|-----|-----|-----|-----|----|
| Anisogramma    | 15248 | 28 | 26 | 54 | 545 94 | 33 | 139 | 2.6 | 61% | 1 4% | 24 | 86% | 27 | 96% | 28 | 100% | 4% |
| Pectobacterium | 72624 | 37 | 42 | 79 | 1671 290 | 43 | 258 | 3.3 | 54% | - - | 25 | 68% | 28 | 76% | 35 | 95% | 19% |
| Ceratohiza     | 24645 | 37 | 35 | 72 | 647 60 | 36 | 137 | 1.9 | 50% | 7 19% | 24 | 65% | 25 | 68% | 34 | 92% | 24% |
| Coniella       | 23708 | 48 | 46 | 94 | 481 64 | 45 | 143 | 1.5 | 48% | 7 15% | 32 | 67% | 37 | 77% | 48 | 100% | 23% |
| Talaromyces    | 54964 | 88 | 86 | 174 | 625 220 | 126 | 626 | 3.6 | 72% | - - | 87 | 99% | 88 | 100% | 88 | 100% | - |
| Elsinoe        | 79740 | 132 | 63 | 195 | 586 146 | 54 | 199 | 1.0 | 28% | 1 1% | 52 | 27% | 74 | 38% | 82 | 42% | 149 | 77% | 35% |
| Claviceps      | 77453 | 140 | 139 | 279 | 553 45 | 92 | 376 | 1.3 | 33% | 16 | 11% | 58 | 41% | 63 | 45% | 82 | 59% | 14% |
| Ceratocystis   | 112291 | 193 | 179 | 372 | 582 205 | 115 | 631 | 1.7 | 31% | 52 | 27% | 74 | 38% | 82 | 42% | 149 | 77% | 35% |
| Phytophthora    | 201815 | 253 | 238 | 491 | 798 24 | 319 | 1103 | 2.2 | 65% | - - | 149 | 59% | 166 | 66% | 184 | 73% | 7% |
| Diaporthe      | 213202 | 399 | 338 | 737 | 530 99 | 196 | 1008 | 1.4 | 27% | 149 | 37% | 140 | 35% | 150 | 38% | 266 | 67% | 29% |
| Peronospora    | 428994 | 513 | 400 | 913 | 824 377 | 349 | 1984 | 2.2 | 38% | 64 | 12% | 200 | 39% | 222 | 43% | 310 | 60% | 17% |
| Alternaria     | 280418 | 551 | 550 | 1101 | 509 11 | 187 | 734 | 0.7 | 17% | - - | 78 | 14% | 86 | 16% | 101 | 18% | 3% |
| Aspergillus    | 547127 | 1032 | 1032 | 2064 | 530 39 | 591 | 2331 | 1.1 | 29% | 19 | 2% | 285 | 28% | 313 | 30% | 414 | 40% | 10% |
| Colletotrichum | 591867 | 1198 | 918 | 2116 | 576 297 | 477 | 2101 | 0.9 | 23% | 562 | 47% | 379 | 32% | 397 | 33% | 667 | 56% | 23% |
| Tilletia       | 743335 | 1200 | 915 | 2115 | 618 259 | 574 | 2666 | 1.3 | 27% | 394 | 33% | 376 | 31% | 403 | 34% | 649 | 54% | 20% |
| Penicillium    | 743954 | 1438 | 1437 | 2875 | 517 12 | 597 | 2675 | 0.9 | 21% | 57 | 4% | 310 | 22% | 325 | 23% | 413 | 29% | 6% |
| Fusarium       | 1604775 | 2946 | 2261 | 5207 | 533 133 | 1165 | 4417 | 0.8 | 22% | 1492 | 51% | 969 | 33% | 1001 | 34% | 1778 | 60% | 26% |

N: size of data set (nucleotides), S: number of sequences (other than sequences with more than 5 ambiguous bases), i: number of internal clades in the phylogenetic tree, n: total number of phylogenetic clades n = S + l, L: average length of sequences in the data set (rounded to closest integer), σ(L): corrected sample standard deviation for the sequence length (rounded to closest integer), n*: number of signable clades, c: number of clusters (λ=36) identified by aocdp, c/n: ratio between clusters and phylogenetic clades, n*/n: ratio between signable clades and phylogenetic clades, sc: number of sequences that are not included in any signable clades, ss: unique signable sequences (also unique signable sequence patterns), sc: unique signable cladepatterns, s: unique cluster patterns, δ: s = sc − ss: discrimination increase attributable to clusters (difference between unique cluster patterns and unique signable cladepatterns)
Fig. 1 Distribution of number of sequences per cluster (cumulative percentage), data set Unite, λ = 36

Fig. 2 Distribution of number of signatures per cluster (cumulative percentage), data set Unite, λ = 36
likely indicates a data set with high degree of similarity between sequences in different phylogenetic clades.

For example, 50% of sequences in the *Fusarium* data set are not contained in any signable clade. For the *Ceratocystis* data set, seven sequences (0, 4, 6, 8, 14, 16 and 17) have no signable clades (18% of the total).

Since every subsequence of length \( \lambda \) of every sequence is a cluster signature, every sequence is a member of at least one cluster. In other words, clusters provide signatures for every sequence of a data set (complete recall).

Each sequence has an associated cluster pattern (fingerprint) in the sequence-to-cluster incidence matrix (Fig. 5). This pattern may be unique for the sequence or can be shared with other sequences. For example, sequence 36 has a unique cluster pattern, but sequences 22, 23 and 24 have identical patterns. We call the number \( s_c \) of unique row patterns in the incidence matrix, unique cluster patterns.

If only sequence signatures are taken into account, all signable sequences can be uniquely and trivially identified by a sequence signature (a cluster of size 1). In other words, the number \( s_i \) of sequences that can be uniquely identified using sequence signatures is the number of signable sequences.

If taking into account signable clades and signable sequences, additional sequences may be uniquely identified from their signable clade pattern. For example, sequence 27 can be uniquely identified using its clade signature pattern, although it could not be identified using its sequence signature (it does not have one). We call the number \( s_n \) of such sequences, unique signable clade patterns.

For each data set, the quantities \( s_i, s_n \) and \( s_c \) were computed and the percent change \( \delta_c = s_c - s_n \), which we call discrimination attributable to clusters. For all data sets except *Talaromyces*, \( \delta_c > 0 \). Substantial gains can be seen, for example, for *Ceratocystis* (35%), *Diaporthe* (29%) and *Fusarium* (26%).

For the two data sets with the highest ratio less than 100% of unique signable clade patterns \( s_n \): *Anisogramma* (\( s_n =96\% \)) and *Coniella* (77%), the ratio of unique cluster patterns \( s_c \) increases to 100%.

Compared to using only phylogenetic clade signatures, where some sequences do not appear in any signable clade (\( s_0 > 0 \) in most cases) recall (Eq. 3) is always 100% when using clusters (since every sequence is a member of at least one cluster). Selectivity is also increased since more sequences can be differentiated through unique cluster patterns vs. unique signable clade patterns (\( \delta_c > 0 \) in most cases).
Consequently, the design of wet lab probes based on cluster signatures can improve recall and selectivity compared to designing only sequence or clade signature probes.

Clusters for high-throughput sequencing

Precision (Eq. 2), recall (Eq. 3) and the $F$-measure (Eq. 4) for the matching algorithm 1 were evaluated using the 4Mycotoxins training set and the bootstrapped 4MicotoxinsBootstrap testing set for signature lengths $\lambda \in \{8, 16, 24, 32, 40\}$. Identification was considered correct if for a query sequence its originator sequence from the training set is reported among the sequences with maximum similarity $\alpha$ over a threshold (Eq. 5) dependent on the error rate $\epsilon$.

\[
\text{precision} = \frac{\text{correctly identified}}{\text{total identified}} \quad (2)
\]

\[
\text{recall} = \frac{\text{correctly identified}}{\text{total number of terms}} \quad (3)
\]

\[
F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (4)
\]

\[
\alpha \geq 1 - 2\epsilon \quad (5)
\]

Precision was found to decrease with the increase of the error rate $\epsilon$ and signature length $\lambda$. Precision was consistently above 0.9 for $\lambda \geq 16$. Recall degraded below 0.5 for higher error rates.

Imperfect recall is due to two factors: “crowding” of defects in a query sequence to the point where there are no preserved subsequences of length $\lambda$, and to query sequences that have more errors, failing the similarity threshold (Eq. 5).

Imperfect precision is due to errors in the query sequence leading to accidental matches and higher similarity scores ($\alpha$) for sequences in the training set other than the originator sequence. This is likely to happen in training sets with high degree of similarity between sequences, for example, when an error may coincide with a single nucleotide polymorphism site.

Precision and recall for the bootstrapped test sets can be applied to only matching sequences in real test sets, where a substantial portion of the data may be unrelated to the training set.

Precision and recall are driven by the size and nature of the training set and the statistical properties of the error-introducing mechanism.

Precision and recall were compared on the same data sets with USEARCH by varying the $\text{maxaccepts}$ parameter $\chi \in \{4, 16, 64, 256, 1024\}$.

We notice that USEARCH outperforms aodp for the highest value of $\chi = 1024$ on the combined $F$ measure (Eq. 4), however aodp outperforms
USEARCH for smaller values of $\chi$ and at lower error rates $\epsilon \leq 0.03$. Moreover, the values of $\chi$ and the related USEARCH parameter “maxrejects” must be chosen a priori. The optimal value of this parameter likely depends on the degree of similarity of sequences within the training set. For example, for \textit{4Mycotoxins}, the optimal value is in range of the total number of sequences (1,338).

For aodp, the optimal value of $\lambda$ can be chosen based on the error rate of the testing set. The set of matching sequences self-calibrates to the size of the matching clusters.

The computational efficiency of the matching algorithm was measured on a realistic test set \textit{97AerobiotaSamples}, for different values of the signature length $\lambda \in \{16, 24, 32, 40\}$.

The number $\mu_{98}$ of matching query sequences (query sequences with similarity $\alpha \geq 1 - 2\epsilon = 0.98$ to at least one training sequence) is relatively stable for different values of $\lambda$. The matching kernel $\Psi \sim \Theta$ is a close approximation of the result set.

The set of all training sequences contained in all matching clusters $\Omega$ has average size $\overline{\Omega} < S = 1,338$ smaller than the size of the training set. The average size of the kernel $\overline{\Psi} \ll S$ (average number of alignments) is much smaller than the size of the training set, which shows that using clusters and reducing the matching sequence kernel are effective in reducing the number of alignments, and consequently running time compared to a brute force approach that would align every query sequence to every sequence in the training set.

Running time is dependent on the number of alignments (\overline{\Psi}), and to the set of training sequences in matching clusters ($\overline{\Omega}$): loop at line 2 in algorithm 1.

Running time degrades to impractical values for lower values of $\lambda$ (estimated at over 200h for $\lambda = 8$). This is due to very large values of $\overline{\Omega}$ (overfitting) from large number of

\textbf{Fig. 5} Sequence by cluster incidence matrix for the \textit{Ceratorhiza} data set ($\lambda=36$). Each row contains cluster matches associated with the sequence with numeric identifier on the y-axis (a fingerprint of the sequence). Each column represents sequences contained in a given cluster. Signable sequences and signable internal clades are at the left. The remaining (bare) clusters are at the right.
clusters in the training set at low signature length (Fig. 4) and high likelihood of accidental matches between short signatures and query sequences (Eq. 1).

Running time is also measured for BLAST and USEARCH on the same data sets. aodp outperforms BLAST by one order of magnitude. Running time is also faster, if comparable to USEARCH.

Discussion
The number of clusters is larger than the number of signable clades (Fig. 4), but comparable to the total number of clades. Within experimental constraints, it is feasible to design signatures for each cluster in a data set. Cluster signatures offer increased discrimination compared to sequences or clades signatures.

The number and composition of clusters is an objective property of a given data set. Conversely, phylogenies can be subjective when prepared by human taxonomists or inaccurate when automatically built using specific heuristics, in some cases with subjective parameters.

Most clusters have a small number of signatures (are brittle to additional experimental constraints) and a small number of sequences (have focused discrimination). To achieve optimal discrimination for clusters, signature length should be chosen as small as practical above the lower limit imposed by the birthday formula (Eq. 1).

Clusters provide signatures for every sequence in a data set (complete recall).

This makes it practical to design probes that identify DNA sequences from data sets with very closely related material, where some of the sequences may not be represented in any of the signable clades. Unique cluster patterns associated with sequences (Fig. 5) can help uniquely identify sequences from a data set, beyond the ability of unique signature clade patterns, in some cases for 100% of the sequences.

Cluster signatures can be used as clues for identifying partial, imperfectly copied query sequences (such as produced by HTS) against a training set of reference sequences. Combined with a global alignment algorithm for comparing candidate sequences from the matching sequence kernel of a set of matching clusters, a matching algorithm (algorithm 1) achieves good matching precision and recall for test sets of different quality (Table 2).

Using a set of matching clusters \( \Omega \) to the query sequence significantly reduces the number of pairwise comparisons (\( \Omega \)) compared to the brute force approach. Reducing to a kernel of matching sequences \( \Psi \) further decreases the number of alignments and provides good running time performance, with dependence on the number and size of clusters in the training database, but not on the actual size of the training database.

Increasing the signature length \( \lambda \) generally increases the precision and decreases the running time of the matching algorithm, but decreases the recall, even to unsatisfactory values (Table 2) for testing sets with high error rates \( \epsilon \). However, lower recall values (e.g. at or below 50%) may be acceptable when the assertion of existence of the target and not the accuracy in abundance was the objective of the investigation. Also, sequencing read accuracy at or above 98% (\( \epsilon \leq 0.02 \)) is provided by the majority of the algorithms, but decreases the recall, even to unsatisfactory values (Table 2) for testing sets with high error rates \( \epsilon \). However, lower recall values (e.g. at or below 50%) may be acceptable when the assertion of existence of the target and not the accuracy in abundance was the objective of the investigation. Also, sequencing read accuracy at or above 98% (\( \epsilon \leq 0.02 \)) is provided by the majority of

### Table 2 Precision and recall of our matching algorithm (aodp) and USEARCH using the 4Mycotoxins training set and the 4Mycotoxins testing set

| \( \lambda \) | aodp | USEARCH |
|---|---|---|
| 8 | 16 | 24 | 32 | 40 | 8 | 16 | 24 | 32 | 40 |
| \( \epsilon \) | Precision | Recall | Precision | Recall |
| 0.00 | 0.74 | 0.90 | 0.91 | 0.92 | 0.93 | 0.71 | 0.49 | 0.25 | 0.15 | 0.08 |
| 0.04 | 0.78 | 0.92 | 0.92 | 0.92 | 0.93 | 0.76 | 0.64 | 0.38 | 0.23 | 0.14 |
| 0.03 | 0.83 | 0.95 | 0.95 | 0.95 | 0.96 | 0.80 | 0.78 | 0.55 | 0.38 | 0.24 |
| 0.02 | 0.89 | 0.97 | 0.97 | 0.97 | 0.97 | 0.84 | 0.88 | 0.74 | 0.57 | 0.43 |
| 0.01 | 0.95 | 0.99 | 0.99 | 0.99 | 0.99 | 0.87 | 0.91 | 0.88 | 0.78 | 0.68 |
| 0.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

Rows have a given error rate \( \epsilon \) for aodp, columns have a given signature length \( \lambda \). For USEARCH, columns have given signature length \( \lambda \). For the “maxaccepts” parameter, Cells where USEARCH outperforms aodp on the \( F \) measure are in **bold**. Cells where aodp outperforms USEARCH on the \( F \) measure for \( \epsilon \leq 256 \) are also in **bold**.

### Table 3 Performance of the matching algorithm using the 4Mycotoxins training set (1,338 sequences) and the 97AerobiotaSamples testing set by signature length \( \lambda \)

| \( \lambda \) | \( \mu_{98} \) | \( \overline{\mu} / \overline{\psi} \) | \( \overline{\Psi} \) | \( \overline{\Omega} \) | \( t \) |
|---|---|---|---|---|---|
| 16 | 1352 | 0.93 | 0.317 | 17.41 | 17039 |
| 24 | 1353 | 0.94 | 0.311 | 13.27 | 9720 |
| 32 | 1342 | 0.95 | 0.299 | 11.83 | 6362 |
| 40 | 1325 | 0.94 | 0.298 | 11.06 | 3031 |

\( \mu_{98} \): number of matching query sequences with similarity \( \mu \geq 1 - 2 \epsilon = 0.98 \); \( \overline{\Omega} \): running time in seconds (system description in “Comparisons with other algorithms” section). Average values (algorithm 1) are reported for: \( \overline{\mu} / \overline{\Psi} \): number of sequences in all matching clusters \( \overline{\Omega} \); Ratio \( \overline{\mu} / \overline{\Psi} \); average size of the result set to the average size of the matching kernel. Running times are also reported for USEARCH and BLAST.
HTS techniques, although sometimes through building consensus [76].

Choosing very small values for the sequence length ($\lambda < 16$) leads to overfitting.

Additional thermodynamic constraints such as the elimination of homopolymer regions and filtering on melting temperature [47, 77] also apply to the design of signatures for assay development. Because of the variability of study objectives and experimental conditions, thermodynamic constraints have not been taken into account in analyses in this study, although support is built into aodp.

Results reported in the current study were drawn primarily from a wide variety of fungal groups, with a focus on plant pathogen and mycotoxin producers. We are confident that these can be generalized to include sequences from other organisms.

Since clusters do not rely on phylogenetic assumptions, but may only coincide with phylogenetic clades, there is no direct dependency of cluster parameters on a specific phylogeny. The phylogeny independent clusters can be particularly useful when it is important to follow some specific DNA sequences such as resistance point mutations or horizontally acquired genes.

A comprehensive study on a variety of data sets with different length distributions and systematically varied completeness and diversity may provide further insights (future research). Power law goodness-of-fit tests such as the Kolmogorov-Smirnoff statistic [78] for the dependencies of the number clusters and signable clades on number of sequences using a larger number of data sets may help quantify results.

Precision and recall were evaluated on a data set (4MycotoxinsBootstrap) of 100bp sequences with randomly-introduced errors at given error rates from a source data set (4Mycotoxins). While this does not account for variable read length generated by different HTS methods or for non-random defects, such as homopolymer errors or issues related to palindromic sequences, precision and recall targeted to specific methods can be modeled into the defect-introducing mechanism.

On precision and recall ($F$ measure), our matching algorithm outperforms USEARCH for lower values of “maxaccepts” ($\chi \leq 256$) and lower error rates ($\varepsilon \leq 0.03$). This likely happens in situations of closely related portions of training sequences, of which a large number (possibly overlapping a signature cluster) are equally similar to the training sequence. By imposing a limit on the size of the result set, the source sequence may or may not be included in the first $\chi$ USEARCH matches. Similar behaviour can be expected when varying the “max_target_seqs” in BLAST.

USEARCH outperforms our algorithm for higher values of $\chi$ (in range of the number of training sequences), but this parameter is dependent on the degree of similarity of the training set and must be chosen a priori. Always choosing large values may be impractical, since the size of the result sets increase dramatically with higher values of $\chi$.

Precision and recall were only evaluated on the bootstrapped data set, where introduced errors could be traced back to the originator sequence (ground truth) for comparison with the reported matches. Such a source of ground truth could not be easily derived for the larger 97AerobiotaSamples data set since the query sequences are, by definition, unknown and the number of matching terms may be too small to draw statistically sound conclusions: est. 1,300 vs. 80,280 terms for 4Mycotoxins-Bootstrap. Future research may look at comparisons on a real data set with a substantial number of matches with matches reported by BLAST as a source of ground truth. An additional complication for such a study may be the need to choose a high “max_target_seq” parameter to cover all possible matching training sequences, resulting in very high running times.

Our matching algorithm outperforms BLAST on running time by one order of magnitude on a realistic data set, with a "max_target_seqs" $\chi = 100$. BLAST performance degrades much further for higher values of $\chi$.

aodp also outperforms USEARCH on running time. This may be due to the nature of the testing set: approximately 1,300 matches in about 4.5 million reads, which may be reasonable for targeted studies of environmental samples, but may not hold for other types of investigations.

On a highly parallel system (80 hardware threads; results unreported), the difference increases to two orders of magnitude for BLAST and increases further for USEARCH because of good multithreading scalability of aodp.

The number and size of clusters are fixed parameters of a given database and represent the main drivers for the running time of the matching algorithm. In situations where short running time is essential without access to large computational resources, running time may be increased by increasing the signature length ($\lambda$) at the expense of recall, e.g. for a preliminary “quick” run, or the size of the source database can be reduced (which will lead to a reduction of the number and size of clusters).

Further improvement of the running time for aodp may be achieved by a more efficient implementation of the global alignment algorithm (step 5 in algorithm 1), such as using nucleotide k-mers, or alignment clues from the positioning of the matching signatures.

Conversely, cluster signatures could be used in a preprocessing step to quickly eliminate or identify candidate matching sequences, to be further validated using a matching algorithm with different objectives.

The study of the signability of other groupings such as gene function may be useful.
Another promising avenue of future research may be the study of cluster signatures for genetic variants in guiding the detection of mutations relevant to evolution, genetic diseases or rapid comparisons of genomes between tumors and healthy cells.

Conclusions
In this study, we evaluated the statistical properties of cluster signatures (oligonucleotide signatures for groups of sequences in data sets of DNA sequences) and their use for mass identification by sequencing.

Our method is universal as it can find oligonucleotide signatures for unique strains, species, higher level phylogenetic clades or mutations linked to genetic diseases or genetic abnormalities. Once diagnostic cluster signatures are known, rapid analysis tools for detection of high risk species, strains or mutations can be developed.

Our matching algorithm using signature clusters increases the efficiency of matching HTS reads against data sets of reference genetic material compared to string alignment methods (orders of magnitude faster than BLAST) and even outperforms high performance k-mer string search algorithms (such as USEARCH) for realistic environmental sample studies.

The matching algorithm also maintains good precision and recall compared to less sensitive string search methods and even outperforms USEARCH for reasonably high settings of the “maxaccepts” value on data sets with lower error rates ($\epsilon \leq 0.03$).

The matching algorithm does not rely on a-priori selection of a parameter limiting the result setlength, such as “max_target_seqs” for BLAST or “maxaccepts” for USEARCH, but self-calibrates to the size of the set of matching clusters of each query sequence.

Using cluster signatures improves recall and accuracy of existing in vitro methods of identification, especially for data sets containing closely related genetic material, without needing to rely on a priori hierarchical phylogenetic grouping.

Cluster signatures and the aodp utility can increase the sensitivity and accuracy of PCR-based and DNA hybridization-based experiments compared to traditional methods based on sequence or phylogenetic clade signatures. Cluster signatures can also be used for targeted enrichment-based HTS, developing accurate, sensitive and efficient diagnostic tools for in vivo or in silico detection of high-risk pathogens or mutation of genes linked to genetic disorders or tumors, using genomics, genetics and metagenomics sequencing data.

Endnotes
1 Signable sequence: sequence with at least one sequence signature.
2 Signable clade: phylogenetic clade with at least one clade signature.
3 Following phylogenetic tree terminology, we call clades with one sequence or with only identical sequences leaf clades and clades with more sequences internal clades.
4 Since $\lambda$ is generally much smaller than the size of each sequence, the total number of $\lambda$-mers can be approximated by the number of nucleotides in the database.

Acknowledgements
We thank Christine Lowe for testing aodp, Chunfang Zheng for reviewing the manuscript and Anatoly Belov for his help in reviewing existing HTS tools and in evaluating USEARCH and BLAST. Our deepest gratitude goes to the anonymous reviewers for their thoughtful and insightful comments.

Funding
This work is financially supported by Agriculture & Agri-Food Canada (AAFC) funded project #97 and Canadian Safety and Security Program (CSSP) funded project CRTI 09-462RD/CSSP 30v+01.

Availability of data and materials
The Automated Oligonucleotide Design Pipeline (aodp) used in analyses in this study is included in the supplementary material and is also available at: https://bitbucket.org/wenchen_aafc/aodp_v2.0.release. The 97AerobiotaSamples data set is available as samples in the Sequence Read Archive, https://www.ncbi.nlm.nih.gov/bioproject/PRJNA358221; all other datasets are available in the supplementary material.

Authors’ contributions
MZ, WC and AL contributed the concept. MZ designed the algorithms and provided the data analysis. WC designed the workflow of aodp and supervised all testing. WC, CV, AL and MZ provided the evaluation datasets. All authors provided input in the interpretation and presentation of data analyses, edited and reviewed the manuscript. All authors read and approved the final version of this manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
References

1. National Research Council. Sequence-based Classification of Select Agents: A Brighter Line. Washington: National Academies Press, 2010.

2. Sedlar K, Kupleva K, Provaznik I. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. Comput Struct Biotechnol J. 2017;15:48–55.

3. Lin H-H, Liao Y-C. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. Sci Rep. 2016;6:24175.

4. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460–1.

5. Fu L, Niu B, Zhu Z, Wu W. Sitao and Li: CD-HIT: accelerated for clustering high-confidence taxonomic assignments from metagenomic data. Bioinformatics. 2009;25(18):2329–30.

6. Ahn T-H, Choi J, Pan C. Sigma: strain-level inference of genomes from metagenomic analysis for biosurveillance. Bioinformatics. 2015;31(2):170–7.

7. Alschtschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.

8. Eddy SR. Multiple alignment using hidden Markov models. In: ISMB, vol. 3; 1995. p. 114–20.

9. Poulsen TM, Frith M. Variable-order sequence modeling improves bacterial strain discrimination for ion torrent dna reads. BMC Bioinformatics. 2017;18(1):299.

10. Hong C, Manimaran S, Shen Y, Perez-Rogers JF, Byrd AL, Castro-Nellar E, Crandall KA, Johnson WE. Pathoscope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. Microbiome. 2014;2(1):33

11. Haque M, Ghosh TS, Komanduri D, Mande SS. Sort-items: Sequence classification tool webserver for taxonomic classification of metagenomic sequences. Bioinformatics. 2010;26(11):132.

12. Lan Y, Wang Q, Cole JR, Rosen GL. Using the rdp classifier to predict taxonomic novelty and reduce the search space for finding novel organisms. PLoS ONE. 2012;7(3):32491.

13. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007;73(16):5261–7.

14. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Von Horn DJ, Weber CF. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7571–7.

15. Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. PLoS Comput Biol. 2010(6):e1553–7358.

16. MacDonald NJ, Parks DH, Beiko RG. Rapid identification of high-confidence taxonomic assignments for metagenomic data. Nucleic Acids Res. 2012;40(1):1362–4962.

17. Caminska J, Gibson MK, Franzosa EA, Segata N, Dantas G, Huttererhofer C. High-sensitivity targeted functional profiling in microbial communities with ShortBREAD. PLoS Comput Biol. 2015;11(12):1–22.

18. Scholz M, Ward DV, Pasolli E, Tiloi T, Zollo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods. 2016;13:435–8.

19. Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. PLoS Comput Biol. 2010;6(12):e1553–7358.

20. MacDonald NJ, Parks DH, Beiko RG. Rapid identification of high-confidence taxonomic assignments for metagenomic data. Nucleic Acids Res. 2012;40(14):1362–4962.

21. Kaminski J, Gibson MK, Franzosa EA, Segata N, Dantas G, Hutterhofer C. High-sensitivity targeted functional profiling in microbial communities with ShortBREAD. PLoS Comput Biol. 2015;11(12):1–22.

22. Scholz M, Ward DV, Pasolli E, Tiloi T, Zollo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods. 2016;13:435–8.
45. Zahariev M, Dahl V, Chen W, Lévesque CA. Efficient algorithms for the discovery of DNA oligonucleotide barcodes from sequence databases. Mol Ecol Resour. 2009;9(1):58–64.

46. Chen W, Djamæ ZR, Coffey MD, Martin FN, Bilodeau GJ, Radmer L, Denton G, Lévesque CA. Membrane-based oligonucleotide array developed from multiple markers for the detection of many Phytophthora species. Phytopathology. 2013;103(1):43–54.

47. SantaLucia Jr. J, Hicks D. The thermodynamics of DNA structural motifs. Annu Rev Biophys Biomol Struct. 2004;33:415–40.

48. Kane MD, Jatko TA, Stumpf CR, Lu J, Thomas JD, Madore SJ. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Res. 2000;28(22):4552–7.

49. Ille L, Ille S, Khoshrasta S, Bigvand AM. Seeds for effective oligonucleotide design. BMC Genomics. 2011;12(1):280.

50. Ilie L, Mohamadi H, Goldberg GB, Smyth WF. Bond: Basic oligonucleotide classification using exact alignments. Genome Biol. 2007;8(2):R6.

51. Hall TA. Bioedit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/nt. Nucleic Acids Symp Ser. 1999;41:97–9.

52. Price MN, Dehal PS, Arkin AP. FastTree 2 approximately maximum-likelihood trees for large alignments. PLOS ONE. 2010;5(3):9490.

53. Kjällåg U, Nilsson R, Abarenkov K, Tedersoo L, Taylor A, Bahram M, Bates S, Bruns T, Bengtsson-Palme J, Callaghan T, Douglass B, Drenkhan T, Eberhardt U, Dueñas M, Grebenc T, Griffith G, Hartmann M, Kirk P, Kohout P, Larsson E, Lindahl B, Lucking R, Martin M, Matheny P, Nguyen N, Niskanen T, Oja J, Peay K, Peintner U, Peterson M, Pöldmaa K, Saag L, Saar I, Schüller A, Scott J, Sené C, Smith M, Suija A, Taylor D, Telleria M, Weiß M, Larsson K-H. Towards a unified paradigm for sequence-based identification of fungi. Mol Ecol. 2013;22:5271–7.

54. Matsumoto M, Nishimura T. Mersenne twister: a 623-dimensionally equidistributed uniform random number generator. ACM Trans Model Comput Simul. 1998;8:3–30.

55. Giles A, Meglécz E, Pech N, Ferreira S, Malausa T, Martin J-F. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics. 2011;12(1):245.

56. Telleria M, Weiß M, Larsson K-H. Towards a unified paradigm for sequence-based identification of fungi. Mol Ecol. 2013;22:5271–7.

57. Clauset A, Shalizi CR, Newman MEJ. Power-law distributions in empirical data. SIAM Rev. 2009;51(4):661–703.

58. Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

Learn more at biomedcentral.com/submissions