## *HTS-barcode-checker*: automated taxonomic identification of illegally- traded species in suspect mixtures

Youri Lammers[[1]](#footnote-1), Tamara Peelen2,, Rutger A. Vos1, Barbara Gravendeel 1,3

## Abstract

### Background

Mixtures of organic substances traded internationally often raise the suspicion of Customs offices that they may contain parts of species protected by the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES). This may lead to confiscation, particularly in the case of traditional Chinese medicine. High-throughput sequencing (HTS) of DNA barcoding markers obtained from such samples currently provides a lot more insight into species composition in a much shorter timeframe as compared to a few years ago but manual verification of the results against the CITES appendices is still labor intensive. In addition, false positive hits can occur for DNA barcodes deposited in NCBI GenBank with incorrect taxonomic names. Lastly, incongruence between the taxonomies of CITES and NCBI GenBank can result in erroneous estimates of illegal trade.

### Results

*HTS-barcode-checker* is an application for automated processing of sets of next generation sequences to determine whether these contain DNA barcodes obtained from species listed on the CITES appendices. This analytical pipeline builds upon and extends existing open-source applications for read sorting and serial searching against the NCBI GenBank reference database. In a single operation, many reads are converted into taxonomic identifications and possible hits with names on the CITES appendices. By inclusion of a blacklist and ‘reconciled names’ database, *HTS-barcode-checker* mitigates against false positives and identifies taxonomic heterogeneity in the results.

### Conclusions

We show that *HTS-barcode-checker* can detect correctly identified DNA barcodes of CITES-protected species large volumes of reads obtained from TCM samples in reasonable time. As a result, our pipeline can aid in better monitoring of trade in endangered species to help prevent their extinction in the wild. *HTS-barcode-checker* is available at <http://github.com/naturalis/HTS-barcode-checker>.

## Keywords

Biodiversity informatics; High-throughput sequencing; Taxonomy; Wildlife Forensics.

## Background

The Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) was signed in 1973 to control and regulate trade in endangered species. The convention produces appendices on which species are listed in which trade is tightly controlled or prohibited under strict legal sanction. These lists can be found on the Internet as text (pdf and html) documents (www.cites.org/eng/app/index.php).

Control in trade of CITES-listed species is challenging when parts like antlers, horns, leaves, rhizomes, roots, stems or tails are used in mixtures such as traditional Chinese medicines (TCMs). During manufacturing of TCMs, parts are often processed (ground, heated, dried, or blended with other products), which makes species identification based on chemical, chromatographic or morphological methods challenging (Coghlan et al., 2012; Gathier et al. 2013). In addition, labels often do not provide sufficient warranties about the actual content of a product. New methods are therefore needed to protect both consumers and producers from fraud, and endangered species from over-exploitation.

DNA barcoding (Savolainen, Cowan, Vogler, Roderick, & Lane, 2005) is a powerful new tool for the emerging field of wildlife forensics. Species composition of mixtures can be retrieved by sequencing a variable marker of the genome. For animals, the mitochondrial *COI* marker is frequently used. For plants, next to the official Barcode of Life (BoLD) plastid *matK* and *rbcL* markers, the nuclear ribosomal Internal Transcribed Spacer (nrITS) regions are often sequenced. DNA barcoding sequences obtained are compared against reference databases such as NCBI GenBank or BoLD (Hebert, Cywinska, Ball, & deWaard, 2003). In both databases, reference DNA barcodes of animal and plant species have been submitted during the past 10 years or more. In the latter database, next to sequence data, images of reference specimens and additional sampling details can also be found. The availability of increasingly large reference databases has opened up the way for DNA barcoding as a standard tool applied by regulatory institutions such as Customs offices worldwide to control illegal trade in endangered species.

With high-throughput sequencing (HTS) techniques (Shendure & Ji, 2008) an increasingly large number of barcode sequences can be generated and analyzed at low cost, which leads to a greater identifying potential for complex species samples. The process of going through a set of identified sequences and manually comparing them to the CITES appendices is labor intensive and error prone for a variety of reasons. First of all, HTS continues to increase the volume of reads, which in turn increases the data processing time. Secondly, the CITES appendices are only available as text (pdf and html) documents on the Internet. Manual verification of sequencing results against the 46 pages of these appendices is labor intensive. In addition, false positive hits can occur for DNA barcodes deposited in NCBI GenBank with incorrect taxonomic names. Lastly, taxonomies of the CITES appendices and NCBI GenBank are not always congruent. This can lead to false conclusions about illegal trade in endangered species.

Here, we present a pipeline that automates both the identification and CITES listing verification step to scan large numbers of samples and sequences efficiently for the presence of DNA barcodes derived from protected species.

## The pipeline

### Overview

The *HTS-barcode-checker* pipeline verifies whether reads produced by HTS originate from CITES-protected species by comparing the DNA sequence with data in NCBI GenBank’s reference database (Benson, Karsch-Mizrachi, Lipman, Ostell, & Sayers, 2009) using BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990) searches. The NCBI taxonomic identifiers (taxon ID) of the resulting BLAST hits are compared to a list of taxon IDs that correspond to CITES-listed species. Any putative matches are reported back to the user including the immediately surrounding context of the CITES appendix text. The steps of the pipeline are shown in Figure 1 and will be explained in more detail below.

### Local CITES database

As an offline process, a local database is maintained containing the corresponding NCBI taxon IDs of CITES-listed species. To update this database, a copy of the CITES appendix is automatically downloaded from the Internet (www.cites.org/eng/app/index.php) (Figure 1 - step 1), which the pipeline scans to retrieve names of CITES-protected species and the appendix that contains these names. By default, the pipeline checks the CITES appendices every time it is run and performs a downloading and preprocessing step whenever contents have changed. This behavior can be influenced by forcing a download (with the --force\_download flag) or avoiding one (using --avoid\_download) regardless whether CITES has released new appendices.

For each entry in the CITES appendix the corresponding taxon ID is initially queried using approximate string searches in the NCBI taxonomic database (Figure 1 - step 2). Since an entire genus or family can be listed in the CITES appendix (for example: *Dendrobium* or Orchidaceae), returned higher taxa are expanded into all lower ranks to which GenBank sequences may be annotated. When no taxon ID can be obtained, a taxonomic name reconciliation web service (TNRS, <http://api.phylotastic.org/tnrs>, (Stoltzfus et al., n.d.)) is used (Figure 1 - step 3) to obtain a list of synonyms, based on which the pipeline retries to obtain terminal taxon IDs. The taxon IDs for CITES-protected species are locally stored, along with CITES appendix information and NCBI taxon names, in a comma separated value (CSV) file that can be read with standard spreadsheet software.

### Sequence identification

To identify putative CITES-listed species from DNA sequence data the pipeline takes a set of sequences (in FASTA format) and searches these using BLAST against the NCBI GenBank database (Figure 1 - step 4). Data obtained using a variety of HTS technologies can be used, such as from the Roche 454, Illumina and IonTorrent platforms. All GenBank databases and BLAST algorithms are supported, but by default the nucleotide database *nr* is used in combination with the *blastn* algorithm. The user can influence how results are filtered on match quality by optionally specifying the minimum percentage identity of returned hits (default is 97%), the minimum coverage in bases (default is 100) and the maximum e-value (default is 0.05).

As some HTS platforms produce large amounts of data the user is advised to consider reducing these amounts using any of the following pre-processing steps, which can be performed using a variety of commonly-used bioinformatics tools:

1. Only supply data obtained by targeted sequencing of barcoding loci, not e.g. whole genome shotgun sequencing
2. Filter reads by omitting those with poor quality scores
3. Remove duplicate reads
4. Cluster similar reads into Operational Clustered Taxonomic Units (e.g. using octupus, [http://octupus.sourceforge.net](http://octupus.sourceforge.net/)) and pick a representative (or consensus) from among each cluster

Note that remote BLAST searches through NCBI must be limited in volume. Our implementation of *HTS-barcode-checker* is based on the BioPython library’s BLAST client, which automatically respects these limits by introducing a 3 second waiting period between polling requests to the NCBI server.

As NCBI GenBank contains erroneous taxonomic names (Groenenberg, Neubert, & Gittenberger, 2011), the pipeline uses a user-editable blacklist of GenBank accession numbers for which taxonomic identification is known to be incorrect (Figure 2). For each BLAST hit that passes filtering through this blacklist (Figure 1 - step 5) the taxon ID is obtained from the sequence record. This taxon ID is then matched against the local CITES database (Figure 1 - step 6) to determine if the sequence originates from a CITES-protected species.

### Pipeline output

The final result is a CSV spreadsheet file containing the query sequences, BLAST hits and, in case a CITES-listed species was found, also the surrounding textual context from the relevant appendix. A condensed example of such output is shown in Table 1. As taxonomic heterogeneity (i.e. multiple hits from different species are returned for the same query) has the potential of causing both Type I and Type II errors such occurrences must be evaluated further by the user. To draw attention to this, the pipeline flags such issues with a critical warning message such as:

X out of a total of Y distinct taxa for “sequence Z” are CITES-listed

CITES appendices contain multiple exceptions for certain species, e.g. based on their geographic location, domestication status or the enforcement of trade quota. The pipeline is not capable of handling these exceptions, as they are not made available in a structured format. All results that match the names listed in the CITES appendices are therefore flagged and all surrounding context is reported to the user.

## Usage examples

In its simplest form, the pipeline is run on an input FASTA file, producing an output CSV file, like so:

HTS-barcode-checker -i <infile> -o <outfile>

A user-specified blacklist file (which is a text file containing one GenBank accession number on each line; Figure 2) can be specified, like so:

HTS-barcode-checker -i <infile> -o <outfile> -bl <blacklist>

To account for synonyms (Boyle et al. 2013) a local database of reconciled CITES names (see Figure 3) can be specified, like so:

HTS-barcode-checker -i <infile> -o <outfile> -cd <database>

To force or avoid updating the local database, add --force\_update or --avoid\_update, respectively. In addition, numerous command line arguments to control BLAST search behavior can be provided. Sensible defaults have been defined, and their usage can be recalled by issuing the --help argument.

## Results and discussion

### Caveats

### Performance evaluation

To our knowledge, the *HTS-barcode-checker* pipeline is the first tool for automated searches for DNA barcodes of CITES-protected species in HTS data. On the CITES website, several other online tools are available such as databases that can be queried for information about trade, management systems, export quotas, publications, identification manuals and photographs. None of these tools can be used to search for hits in HTS datasets yet.

To compare speed, we submitted a single TCM HTS dataset (... Mb) to 5 different students and let them search for CITES-listed species by visual comparison of the identifications obtained with the CITES Appendices. Processing time varied between 5 and 10 minutes and 1 species was missed by 1 of the 5 participants whereas the *HTS-barcode-checker* pipeline processed the data in less than 3 minutes and successfully retrieved all CITES-listed species.

### Future directions

Although the pipeline presented here is ready-made, several modifications are possible that would increase usability and impact. For example, although incorrect taxonomic identifications of GenBank records have previously been noted (Groenenberg et al., 2011), no community project yet exists to record and track such errors (Pennisi, 2008). The blacklist used by the  *HTS-barcode-checker* checker could be used for communal record keeping, especially as our usage of *git* as a decentralized revision control system provides the ideal infrastructure for this. Conversely, should an alternative community wide blacklist of NCBI GenBank come into existence,  *HTS-barcode-checker* could be modified to make use of it with the specific purpose of eliminating errors in taxonomic names of CITES protected species. Another possible modification is the addition of a web-based graphical user interface, which would make the pipeline accessible to non-expert users such as Customs officers as it would remove the need for local installations. In addition, this web application could be configured to update the local database of reconciled names and the blacklist at frequent intervals, thereby guaranteeing that the user always operates on state-of-the-art knowledge. Lastly, DNA barcodes of CITES protected species collected from well-identified specimens could be uploaded to BoLD where names can be changed by others when needed. The number of CITES protected species is currently 826 for mammals, ... for birds, ... for reptiles, ... for amphibians, ... for sharks, .... for fishes, 1 for sea cucumbers, 25 for scorpions and spiders, ... for insects, 2 for leeches, 37 for clams and mussels, 10 for snails and conches, ... for corals and sea anemones, ... for sea ferns, fire corals and stinging medusae, and 26234 for plants (estimates based on ... and Mabberley, 2008). Of this total of ... CITES protected species, roughly 16830 (... %) are covered in NCBI GenBank with DNA barcodes, leaving... to be sequenced.

## Conclusions

Increased production of HTS data leads to improved identification potential of endangered species traded in TCMs by DNA barcoding. Unfortunately, due to taxonomic errors in reference databases such as NCBI GeBank and incongruences in taxonomies of the CITES appendices and reference databases, incorrect conclusions can be drawn about illegal trade based on DNA barcoding.

We have developed the *HTS-barcode-checker* pipeline, which is a tool for automated identification of illegally traded species in suspect mixtures with applications for correcting and standardizing taxonomic names to overcome the risks mentioned above. Our pipeline provides a laborsaving and repeatable tool for assessing whether CITES-protected species are used in for example TCMs by analysis of HTS data. Tests carried out demonstrate the potential of the *HTS-barcode-checker* for reducing taxonomic errors and increasing integration between the NCBI GenBank reference database and the CITES appendices.

## Implementation

*HTS-barcode-checker* is written in python and uses the *bio-python*, *beautiful-soup* and *requests* packages to handle FASTA sequences and communicate with the various APIs and web services used.

## Availability and requirements

* Project name: HTS-barcode-checker
* Project home page: http://github.com/naturalis/HTS-barcode-checker
* Operating system(s): Platform independent
* Programming language: Python (version 2.7 or 3.0 and higher)
* Other requirements: the non-core Python packages bio-python, beautiful-soup, requests
* License: BSD-3
* No restrictions to use by non-academics

## List of abbreviations

* **API:** Application Programming Interface
* **BLAST**: Basic Local Alignment Search Tool
* **BoLD**: Barcoding of Life Database
* **CITES**: Convention on International Trade in Endangered Species of Wild Fauna and Flora
* **CSV**: Comma Separated Values
* **HTS**: High Throughput Sequencing
* **NCBI**:National Center for Biotechnology Information
* **TCM**:Traditional Chinese Medicine
* **TNRS**: Taxonomic Name Reconciliation Service

## Acknowledgements

The authors would like to thank Denise Rijkeboer and Elza Duijm for kindly providing us with Ion Torrent data on which to test the pipeline, Thomas Bolderink, Alex Hoogkamer and Roeben Vink for development of the first trial version, and Sigrid Beiboer for co-mentoring TB, AH and RV during the internship out of which this project evolved.

## Authors’ contributions

YL re-implemented a first prototype of *HTS-barcode-checker*. YL, RAV and BG contributed equally to the drafting of this manuscript. RAV oversaw software engineering, TP provided confiscated TCM samples for sequencing. All authors have reviewed and approved the final version of this manuscript.

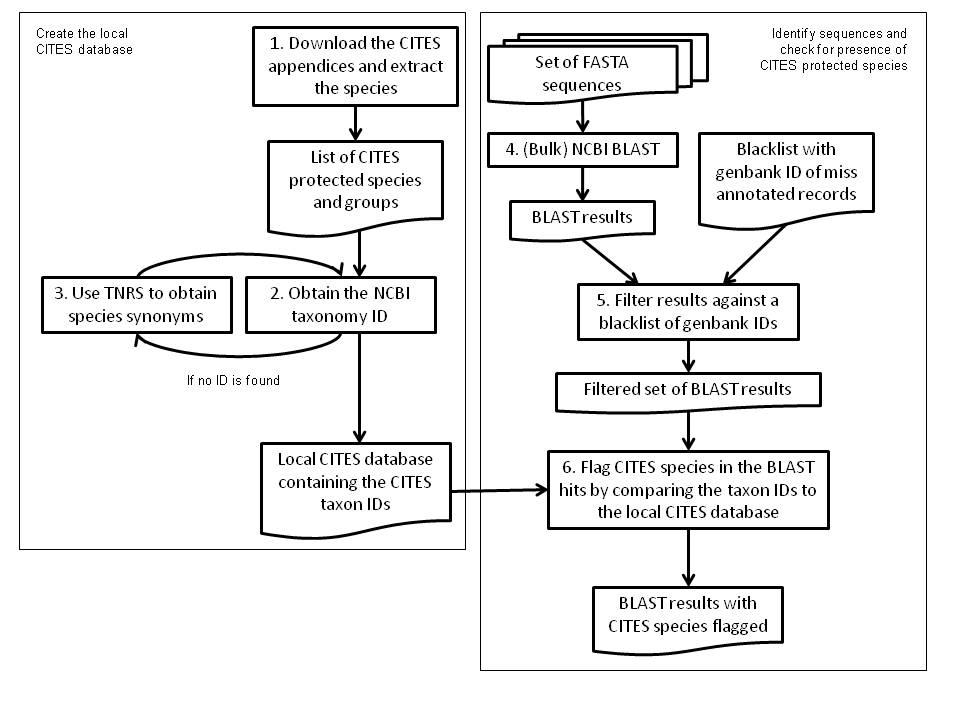
## Competing Interests

The authors declare that they have no competing interests.

## Figures

### Figure 1: Steps of the analysis pipeline

The panel on the left shows the offline process of updating and reconciling names of taxa listed on the CITES appendices with the NCBI taxonomy, the panel on the right shows the process of species identification from sequence data by matching against the reconciled database of CITES names. See text for details.



### 

### Figure 2: User-specified blacklist

Examples of NCBI GenBank accessions placed in our user-specified blacklist, erroneously listed on Genbank as belonging to *Gastrodia elata*, a highly endangered orchid (monocots) species listed on CITES appendix I, but containing nrITS sequences of the non-endangered eudicot Nyctaginaceae (first two) and fungal Agaromycetes (third) and Saccharomycetales (last three), instead.

EF090607

EU135607

EF090905

EU135906

EU135908

EU135909

### Figure 3: Reconciled names database

Examples of names in our re**conciliation database** **listed as distinct** species in NCBI GenBank but considered as synonyms **on the** CITES **Appendices**.

Panax ginseng C.A. Mey.; Panax sinensis J. Wen ined.

## Tables

### Table 1: Example output

Condensed version of results obtained by running the pipeline on data generated by Next Gen (Ion Torrent) sequencing of a confiscated traditional Chinese medicine (TCM) sample. The results show the retrieval of *Dendrobium cruentum*, which is listed on CITES appendix I. This species cannot be traded without CITES permits under any circumstances. In addition, *Panax ginseng* was retrieved, which is listed on CITES appendix II. Trade in wild collected roots from the last remaining natural population of this species in the Russian Federation is only allowed with a CITES permit; farmed roots can be traded without a CITES permit unless they are produced in China. Next to these CITES protected species, two other plant taxa were retrieved: *Dioscorea polystachya* and *Debregeasia elliptica*. Both species are not listed on the CITES appendices which means trade is not restricted.For simplicity’s sake, the columns for BLAST hit results (i.e. accession number, % identity, hit length, e-value and bit score) were omitted.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Query sequence** | **Taxon ID** | **Species** | **CITES info** | **Appendix** |
| OTU\_1 | 55575 | *Dioscorea polystachya* |  |  |
| OTU\_2 | 906701 | *Dendrobium cruentum* | *Dendrobium cruentum* | I |
| OTU\_3 | 540248 | *Debregeasia elliptica* |  |  |
| OTU\_4 | 4054 | *Panax ginseng* | *Panax ginseng*, Only the population of the Russian Federation; no other population is included in the Appendices) | II |

## References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* 1990, **215**: 403–10.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. GenBank. *Nucleic Acids Res.* 2009, **37**: D26–31.

Boyle, B., Hopkins, N., Lu, Z., Garay, J.A.R., Mozzherin, D., Rees, T., Matasci, N., Narro, M.L., Piel, W.H., McKay, S.J., Lowry, S., Freeland, C., Peet, R.K., Enquist, B.J. The taxonomic name resolution service: an online tool for automated standardiation of plant names. *BMC Bioinformatics* 2013, **14(16)**.

Coghlan, M. L., Haile, J., Houston, J., Murray, D. C., White, N. E., Moolhuijzen, P., Bellgard, M. I., et al. Deep Sequencing of Plant and Animal DNA Contained within Traditional Chinese Medicines Reveals Legality Issues and Health Safety Concerns. *PLoS Genetics* 2012, **8(4)**.

Gathier, G., van der Niet, T., Peelen, T., van Vugt, R.R., Eurlings, M.C.M., & Gravendeel, B. Forensic Identification of CITES protected Slimming Cactus (Hoodia) using DNA barcoding. *J. Forensic Sci.* 2013, In Press.

Groenenberg, D. S. J., Neubert, E., & Gittenberger, E. Reappraisal of the “‘Molecular phylogeny of Western Palaearctic Helicidae s.l. (Gastropoda: Stylommatophora)’”: When poor science meets GenBank. *Mol. Phylogenet. Evol.* 2011, **61**: 914–23.

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. Biological identifications through DNA barcodes. *P. Roy. Soc. B-Biol. Sci.* 2003, **270(1512)**: 313–21. Mabberley, D.J. Mabberley's Plant-Book. A portable dictionary of plants, their classification and uses. Third Edition. Cambridge University Press, 2008, Cambridge, U.K.

Pennisi, E. Proposal to “Wikify” GenBank Meets Stiff Resistance. *Science* 2008, **319(5870)**: 1598–9.

Savolainen, V., Cowan, R. S., Vogler, A. P., Roderick, G. K., & Lane, R. Towards writing the encyclopaedia of life : an introduction to DNA barcoding. *Roy. Soc. B-Biol. Sci.* 2005, **360:** 1805–11.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.* 2008, **26(10)**: 1135–45. Stoltzfus, A., Lapp, H., Matasci, N., Deus, H., Sidlauskas, B., Zmasek, C. M., Vaidya, G., et al. Phylotastic! Making Tree-of-Life Knowledge Accessible, Re-usable and Convenient. *BMC Bioinformatics*, (In Review).

1. Naturalis Biodiversity Center, Sylviusweg 72, 2333 BE, Leiden, The Netherlands

   2 Dutch Customs Laboratory, Kingsfordweg 1, 1043 GN , Amsterdam, The Netherlands

   3 University of Applied Science Leiden, Zernikedreef 11, 2333 CK, Leiden, The Netherlands [↑](#footnote-ref-1)