Chapter 19

DNA Barcode-Based Species Diagnosis with MolD

Alexander Fedosov, Nicolas Puillandre, Frank Fischell, Stefanos Patmanidis, Aurélien Miralles, and Miguel Vences

Abstract

Rapid biodiversity loss sets new requirements for taxonomic research, prompting updating some long-established practices to maximize timely documentation of species before they have gone extinct. One of the crucial procedures associated with the description of new taxa in Linnean taxonomy is assigning them a diagnosis, which is an account of the specific features of the taxon, differentiating it from already described species. Traditionally, diagnostic characters have been morphological, but especially in the case of morphologically cryptic species, molecular diagnoses become increasingly important. In this chapter, we provide detailed protocols for molecular taxon diagnosis with the bioinformatic tool MolD which is available as open-source Python code, command-line driven binary, GUI-driven executable for Windows and Mac, and Galaxy implementation. MolD identifies diagnostic combinations of nucleotides (DNCs) in addition to single (pure) diagnostic sites, enabling users to base DNA diagnoses on a minimal number of diagnostic sites necessary for reliable differentiation of taxa.

Key words Cryptic species, Molecular diagnosis, Diagnostic nucleotide combination, User-friendly tools

1 Introduction

Diagnosing species is at the heart of Linnean taxonomy [1]. The Codes of Nomenclature (Botany, Zoology, Microbiology) require researchers, when naming a new species-level taxon, to provide information on how the new species can be differentiated from other, already described species. The Code of Zoological Nomenclature [2] (fourth edition) stipulates that a newly named taxon must "be accompanied by a description or definition that states in words characters that are purported to differentiate the taxon" (Article 13.1.1). Similarly, the International Code of Nomenclature for Algae, Fungi, and Plants [3] specifies that "a name of a new taxon [. . .] must [. . .] be accompanied by a description or diagnosis of the taxon" [or by a reference to such a description/diagnosis], and defines a diagnosis as "a statement of that which in the opinion

Robert DeSalle (ed.), *DNA Barcoding: Methods and Protocols*, Methods in Molecular Biology, vol. 2744,
https://doi.org/10.1007/978-1-0716-3581-0_19,

of its author distinguishes the taxon from other taxa" (Article 38.1/38.2). In particular, the Zoological Code is therefore quite strict in requiring a diagnosis, and more specifically, one that describes differential characters in words.

A recurrent theme in contemporary taxonomy is, however, the discovery of cryptic species, that is, species that cannot or can only with great difficulty be distinguished by morphology [4, 5]. Because the Codes do not limit diagnosis to morphological characters, it is obvious that molecular diagnostics provide a useful alternative, especially with—but not limited to—cases of morphologically cryptic species. Molecular diagnoses from DNA sequences are more straightforward to produce by computer programs than morphological diagnoses, and therefore particularly suitable for approaches of turbo-taxonomy [6] where large numbers of new species are named along with type images and simplified descriptions in an effort to accelerate the taxonomic inventory of life [7]. In recent turbo-taxonomy studies, DNA barcode sequences themselves are used as descriptions and are interpreted as "characters in words," differentiating the taxon, thus supposedly satisfying the requirements of the Zoological Code. To be unambiguously Code-compliant, it is however recommendable, in a formal taxon diagnosis, to rely on diagnostic sites rather than sequences or genetic distances if no other (e.g., morphological) diagnostic characters are available [5]. Because diagnostic positions are always computed relative to the sequence alignment used, it is important to ensure repeatability, e.g., by making the alignment used in a publication also available [8, 9].

Various tools have been developed to identify diagnostic sites from multiple DNA sequence alignments, all applicable to DNA barcodes: CAOS [10], Fastachar [11], DeSignate [12], QUIDDICH [14], DNAdiagnoser [13], and MolD [9]. The wide applicability of these tools remains to be tested [16]; reproducibility, flexibility, and user-friendliness are among the main factors that will determine their success.

Here, we present and recommend using MolD [9] as a standalone menu-driven executable. MolD is tailored for seamlessly recovering DNA-based diagnoses from large DNA datasets and is capable of identifying diagnostic combinations of nucleotides (DNCs) in addition to single (pure) diagnostic sites, enabling users to base DNA diagnoses on a minimal number of diagnostic sites necessary for reliable differentiation of taxa, rather than providing a long list of such sites. The MolD algorithm assembles DNA diagnoses that fulfill predefined criteria of reliability, which is achieved by repeatedly scoring diagnostic nucleotide combinations against datasets of in silico mutated sequences. The new version of MolD presented herein has been updated to expand its functionality in two aspects. First, it integrates some features of DNA diagnoser [13], especially regarding the output, which now includes the

option to identify diagnostic sites for pairwise comparisons of taxa, and produces textual output that can directly be copy-pasted into the diagnosis section of taxonomic revisions. Second, the updated version provides more flexibility in selecting taxa to be diagnosed. Furthermore, the GUI of the stand-alone executable version of MolD has been completely reworked for user-friendliness.

2 Materials

2.1 Computational Resources Needed, Program Code, and Software Implementation

1. All iTaxoTools programs are available as open-source code from Github [https://github.com/iTaxoTools]. This allows any user to compile them on their preferred platform, or extend, modify or improve the code in order to integrate the tools in their own noncommercial software. MolD is written in the computer language Python 3 and can also be directly run under Python in command-line mode. Command line versions are not further explained in this chapter; refer to the respective manuals of the software tools.
2. Stand-alone executables (binaries) with graphical user interfaces are provided in the framework of iTaxoTools (13; http://itaxotools.org/download.html), for Windows, Mac, and Linux computers. These are simple executables that do not need installation and run just by clicking on the respective file. The Mac executables are Universal 2 fat binaries that support both 64-bit Intel and Apple processors. They are compiled for MacOs Ventura 13.1 and are signed using an Apple developer code signing certificate. This ensures trouble-free execution on the most recent Macintosh operating systems and chip architectures as of 2022. The Windows executables are also signed with an open-source code signing certificate and will work reliably on Win 10 and 11 platforms (and some of them on Win 7 and Win 8; for others, legacy versions for Win 7 and Win 8 are available). Step-by-step protocols to use the stand-alone executables are provided below.
3. MolD has also been made available for the Galaxy platform [17]. Galaxy is an open-source platform for data analysis that enables users to access, via a standard web browser, tools from various domains through its graphical web interface. Galaxy installations have been set up for the public domain but can also be installed in restricted computational environments. MolD has been made available along with XML interface files in the Galaxy tool shed [15] and can be thus seamlessly integrated into any Galaxy environment. As an advantage for future development of an integrative taxonomy software environment, Galaxy allows the creation of the so-called "Workflows," where various tools perform a set of tasks in the form of an automated pipeline.

2.2 Input File Formats

MolD requires input data in specific file formats:

1. The program accepts DNA barcode sequences provided in the FASTA format. MolD requires sequences in the FASTA input file to be aligned, and sequence names to have a specific format, each comprising a sequence identifier and a taxon identifier separated by a pipe character (see specifications below).
2. MolD in the current command-line version also accepts input files formatted as a tab delimited text, with three columns corresponding to i) sequence ID, ii) taxon ID, and iii) DNA sequence (aligned across an input file). However, this functionality has not yet been implemented in the GUI-driven executable.
3. Refer to Fig. 1 of Chap. 18 for input file examples.

3 Methods

3.1 Terminology

1. Query taxon: the taxon being diagnosed.
2. Reference taxon (taxa): all taxa included in the analysis that are not query taxon.
3. DNP: diagnostic nucleotide position. A nucleotide site (position in a nucleotide alignment) that differentiates all DNA sequences of a query taxon from some or all DNA sequences of the reference taxa.
4. pDNP: pairwise diagnostic nucleotide position. Nucleotide positions that are diagnostic for either taxon in a pairwise comparison of DNA sequences of two taxa.
5. uDNP: uniquely/universally/unambiguously diagnostic nucleotide: A DNP that unambiguously differentiates all query taxon sequences from all the reference taxa sequences in a DNA alignment, therefore constituting a proper diagnosis.
6. DNC: diagnostic nucleotide combination. A combination of nucleotides at specified sites of the alignment, unique for a query taxon. MolD attempts to minimize the number of nucleotide positions in a DNC to allow for a concise (and accurate) molecular diagnosis.
7. Draft DNC: A DNC produced in the first step of the MolD algorithm, containing a maximum number of DNPs as specified by the Maxlen1 parameter, obtained by a stepwise procedure of adding new positions to reduce (to zero) the number of taxa sharing a DNC with the query taxon. A draft DNC will typically contain more nucleotide positions than the minimal necessary number.

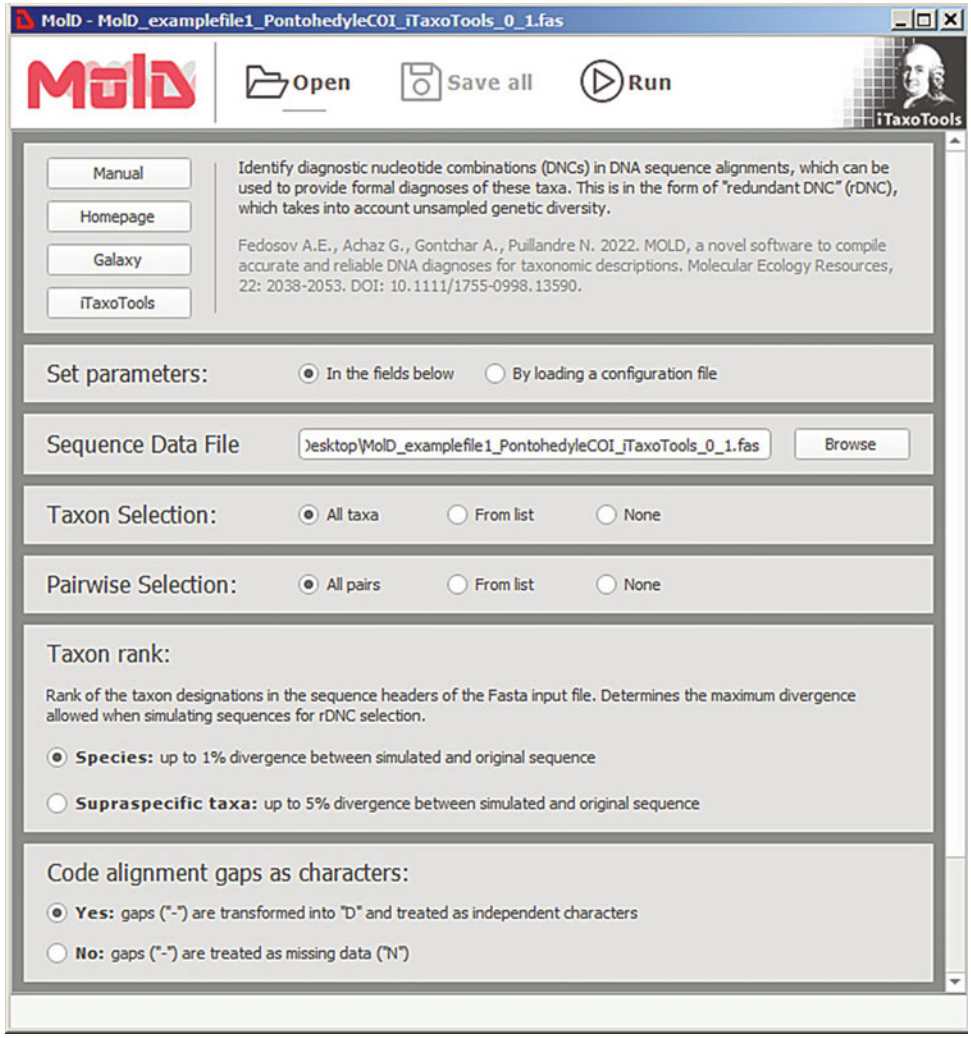

Fig. 1 Screenshot of the starting window of the MolD stand-alone executable developed in the iTaxoTools project

8. mDNC (= minimal DNC): a DNC, comprising a minimal number of DNPs sufficient to differentiate a query taxon from all reference taxa in an alignment. Because it comprises a minimal number of nucleotide sites necessary for diagnosis, any mutation in the mDNC in a single specimen of a query taxon will automatically disqualify it as a diagnostic combination. In MolD, mDNCs are calculated from draft DNCs by the removal of redundant nucleotide positions.
9. Single-nucleotide mDNC: an mDNC consisting of only a single DNP that differentiates a taxon from all other taxa in the analysis, i.e., a term largely synonymous to uDNP as used herein.
10. Independent mDNC: two or more mDNCs are independent if they constitute nonoverlapping sets of DNPs.

11. rDNC (=robust/redundant DNC): a DNC, which comprises more than a minimal necessary number of diagnostic sites, and therefore is robust to single nucleotide replacements. Even if a mutation arises in one of the rDNC sites, the remaining ones will (with high probability) remain sufficient to diagnose the query taxon. In MolD, an rDNC is computed from a list of mDNCs by testing the robustness of diagnostic nucleotide combinations using simulated datasets (in silico randomly mutated alignment replicates).

3.2 Defining Research Questions and Preparing the Data

1. Decide on the taxonomic level at which you wish to diagnose taxa against each other. Typically, this will be at the species level: i.e., you want to find nucleotide sites that can robustly distinguish one species from all others. However, you may also want to diagnose units at other levels: subspecies, species groups, subgenera, genera, or unnamed OTUs or geographical population-level clusters. For each such unit, you need to define a unique name which in the FASTA file will be used as a taxon identifier (see point 4 below).
2. Decide which taxon you want to diagnose against which reference taxa. Note that the more taxa you include in the analysis, the longer it will take. For example, if you plan to use a molecular diagnosis in the context of naming a new species, it may be sufficient to diagnose the new taxon against all other nominal species in the genus or in the species group, but not all of the nominal species against each other. You can specify the comparisons to be carried out either in the menu of the standalone executable or in a configuration file (=parameter file).
3. Decide if you prefer running the analysis by setting parameters in the GUI or by using a separate configuration file (= parameter file). For routine analyses using mostly default parameters, a parameter file is not necessary. For more advanced applications where numerous parameters deviate from the defaults, it is advantageous to use a parameter file; an added value is reproducibility, as you can easily run the same analysis again with exactly the same settings. Therefore, it is recommended to archive the parameter file along with the output files from the analysis. Please refer to the manual of MolD for the detailed format of a parameter file.
4. Prepare a FASTA file in the format required by MolD: each entry starts with the identifier line and one or more lines of nucleotide sequence. The identifier line starts with ">"and must contain two parts, separated by a pipe ("|") symbol. The first part is a free-style *sequence identifier*. The names of the taxa to be diagnosed correspond to the second element, i.e., the *taxon identifier*: >SequenceID####|taxonID. The taxon identifier corresponds to the name of the species, genus, etc., to

which you assign this sequence. There is no need to specify at this time which taxon is the query that should be diagnosed from the reference taxa: query and references are defined later in the settings of the program, and it is also possible to diagnose all against all taxa.

5. Please check that (i) Each identifier line has only one pipe symbol. (ii) No spaces are present in the sequence and taxon identifiers (we strongly recommend only using standard alphanumerical characters and underscores). (iii) All taxa identifiers are provided, correct, and spelled out consistently for all sequences of the same taxon. (iv) Sequence lines only contain valid nucleotides ("A", "C", "G", "T"), gaps ("-"), and ambiguous nucleotides ("N", "K", "M", "R", "S", "W", "Y").
6. You can use DNAconvert of the iTaxoTools set of programs to prepare a MolD-formatted FASTA file from a tab-delimited sequence file where "specimen_voucher" denotes the sequence identifier and "species" or "organism" denotes the taxon name. DNAconvert will automatically replace all nonalphanumeric characters with underscores and optionally add numbers to sequence identifiers to make sure they are short and unique, but it still is recommended to use concise sequence and taxon identifiers composed of alphanumeric characters only.
7. You may consider including in the FASTA file as the first sequence a length reference which ideally is a full sequence of the target gene from a model organism, for instance, the full COI sequence of mouse, zebrafish, or fruit fly in the case of classical animal DNA barcodes. All diagnostic nucleotide positions will then be indexed by the program relative to this sequence, which will make indexing of the DNC sites more reproducible. Note that the nucleotide positions will only be consistent with the length reference if the other sequences in the alignment do not include insertions.
8. Align the DNA barcode FASTA file using an alignment program of your choice. Among iTaxoTools programs you may use MAFFTpy or Concatenator [18] which both implement the FFT-NS-1 and G-INS-i alignment strategies of MAFFT [19].

3.3 Basic Steps and Parameters of a MolD Analysis

1. Parameters for a MolD analysis can be specified either by changing them in the interactive menu of the GUI or within a parameter file that can then be uploaded to the GUI version or used in the command line version.
2. Specify input and output files. If using a parameter file, specify the complete path and name for both the input file and the output file therein.

3. Specify the query taxa (qTAXA in the parameter file). Note that this parameter must be specified because there is no default setting. If all taxa in the dataset are to be diagnosed, use the setting "All" (case insensitive). If you wish all taxa with more than N sequences available to be diagnosed, use ">N." You can also explicitly enter the taxa you wish to diagnose by providing their taxon identifiers separated by comma ("Taxon1, Taxon2,…"). When specifying a selection of taxa for diagnosis, please make sure that all taxa identifiers are spelled correctly in the respective line of the parameter file or in the respective field in the GUI (identifiers must agree with the input alignment file).
4. Advanced options of query taxa input include the following: (i) combined (merged) taxa ("Taxon1 + Taxon2") allowing all sequences of two or more explicitly defined taxa (e.g., Taxon1 and Taxon2) to be diagnosed as a single taxon, (ii) calling taxa by a shared pattern (i.e., certain string of characters) in their taxon identifiers ("P: pattern") to have each of them diagnosed, or (iii) combined taxa obtained by merging all taxa with a certain pattern in their name ("P+:pattern"). It is also possible to enter specific combinations of taxa that should be diagnosed against each other ("Taxon1VSTaxon2," "Taxon1VSALL," "ALLVSALL"); output with diagnostic nucleotide sites differentiating the specified taxa from one another is then written to a separate file "*Basename_pairwise.out*" derived from the provided name of the output file. Also, multiple input options involving taxa names or "ALL" can be run together, such as Taxon1, Taxon2, Taxon1 + Taxon2, and Taxon1VSTaxon2; or ALL, ALLVSALL. Such complex settings are best passed through the parameter file.
5. Specify the taxon rank of your query taxa. In most cases, these will correspond to species or species-level OTUs. If your specified taxa correspond to supraspecific units such as genera or species groups, it is useful to change the setting accordingly to "supraspecific taxa" (in the parameter file: 1 for species, 2 for supraspecific taxa). The specified taxon rank will influence the maximum divergence allowed when simulating sequences in the last step of a MolD run, i.e., when calculating rDNCs. With the species level option set, sequences in this step are artificially mutated up to a maximum divergence of 1% from the original sequence, mirroring the usual divergence encountered within species and with the goal to simulate additional, unsampled haplotypes existing in this species. When the "supraspecific taxon" level is set, divergences up to 3% are allowed. Note that in some taxa with high amounts of intraspecific sequence variation such as many amphibians, it might be appropriate to use the "supraspecific" option also to obtain reliable rDNCs at

the species level. Alternatively, you can directly set the Pdiff parameter to a desired maximum sequence percent-divergence.

6. Specify if gaps should be coded as characters or as missing data. If specifying "Yes," dashes ("-") in the alignment are transformed into "D" characters and these are treated as independent character states. If specifying "No," dashes are treated as missing data ("N") and ignored. Note that MolD at present does not distinguish terminal vs. nonterminal dashes; if some of your sequences have missing data at the beginning or end, and you wish to code indels (dashes) as characters, make sure your missing data at the beginning and end are coded as "N" and not as dashes.
7. Advanced users: If desired you can adjust additional parameters for mDNC recovery. Cutoff (default 100): denotes the number of DNPs to be considered for inclusion into a mDNC; can be a natural number or a number preceded by ">." NumberN: number of ambiguously called nucleotides allowed in a sequence, natural number (default 5). Number_of_iterations (default 10,000). MaxLen1: maximum length of draft DNCs (default 12). MaxLen2: maximum length of mDNCs (default 7). For more detailed explanations see Fedosov et al. [9].
8. Advanced users: To score each candidate rDNC, 100 simulated test datasets are created. For this part of the analysis, the following parameters can be modified: Pdiff: Percent difference between original and modified sequences, natural number or a floating point (default 1 for species-level taxa, 3 for supraspecific taxa). NmaxSeq: maximum number of sequences per taxon to modify, natural number (default 10). Scoring: If an rDNC remains valid in a test dataset, the program adds 1 to the score. As 100 test datasets are created at each step, the lowest possible score is 0 and the highest is 100. If two consecutive scores are above the threshold value defined by a keyword argument (default is moderate), then the rDNC is sent to output. Arguments are "lousy" (=66), "moderate" (=75), "stringent" (=90), "very_stringent" (=95).

3.4 Step-by-Step Protocol for MolD-GUI (iTaxoTools)

MolD-GUI 1.0 is a stand-alone executable available for Windows, Mac, and Linux platforms that implements the code of MolD 1.4. The Windows and Mac versions are single files (with .exe extension in Windows), which can be executed by double-clicking. Alternatively, the command-line version of MolD 1.4 can be downloaded from its GitHub site (https://github.com/SashaFedosov/MolD) and executed in Python 3; refer to the MolD manual for instructions on how to run the command-line version.

1. Download the latest version of the MolD-GUI executable for your preferred computer platform, either from http://

itaxotools.org or from the iTaxoTools Github page (https://github.com/iTaxoTools).

2. Execute the file by double-clicking and wait until the window with the graphical user interface opens (Fig. 1). On Windows, you may have to relax your security settings to allow executing the .exe file.
3. Click the "Browse" button in the "Sequence Data File" field and specify the input file (aligned FASTA) prepared beforehand according to the specifications of MolD (see above). You can also upload the file using the "Open" button in the uppermost row of icons.
4. For advanced users: If you have prepared a parameter file with the settings for analysis, select the option by clicking the respective radio button in the field "Set parameters" to select "By loading a parameter file." Then press the "Browse" button and specify the parameter file. You can also upload the file using the "Open" button in the uppermost row of icons. After this, you can directly press the button "Run" in the upper row of icons to start the analysis. Note that settings in the parameter file will override all previously made settings in the graphical user interface; you can, however, adjust the settings in the interface after uploading the parameter file.
5. For routine analyses, we recommend adjusting settings in the respective fields of the GUI instead of uploading a parameter file. Note that if you leave default settings, the program will retrieve DNA diagnoses for all taxa in the input alignment, and among all possible pairs of taxa, which may take a long time with large datasets. See Sect. 3.3 for information on how various parameters can be set. Press the button "Run" in the upper row of icons to start the analysis. The progress of MolD execution will be output in the terminal-like window (Fig. 2).
6. After the analysis is completed, and depending on the parameter specifications, three files can be viewed or saved (Fig. 3): (i) the results of the molecular diagnosis based on the MolD algorithm (mDNCs and rDNCs), (ii) the results of a pairwise analysis that simply list all those nucleotides that are different between pairs of taxa (pDNPs), and (iii) a log file. For interpretation of the results, see Sect. 3.6.

3.5 Running MolD on the Galaxy Platform

1. MolD is available for any Galaxy installation on the Galaxy toolshed (https://toolshed.g2.bx.psu.edu/). Links to dedicated Galaxy webservers to run this and other programs of the iTaxoTools project are available on the project's website (http://itaxotools.org/).
2. In Galaxy (Fig. 4), first use the "Upload Data" button in the upper left to upload your files for analysis. Once your data files

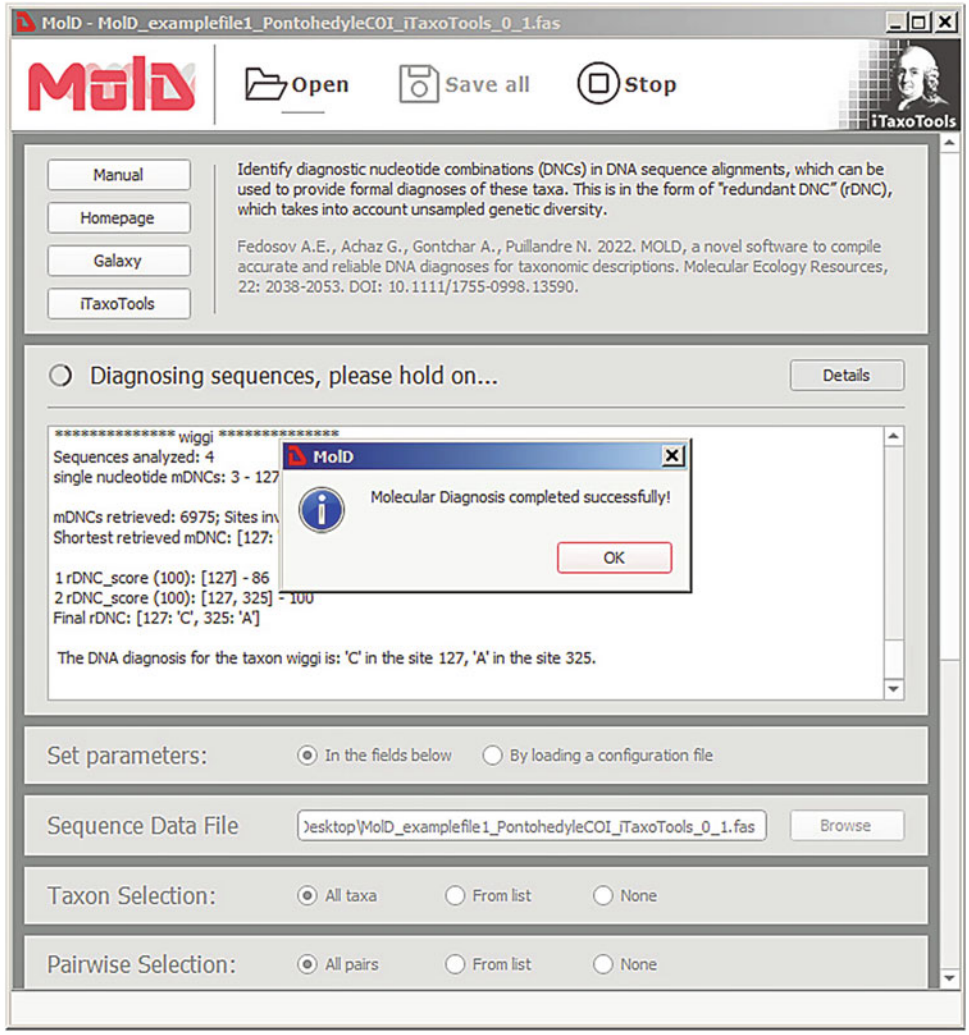

Fig. 2 Screenshot of the MolD stand-alone executable at the end of a successful run. The "terminal window" provides information on the progress of the analysis during the run

have been successfully uploaded, they will be shown in the right "history" bar in green color—wait until this process is completed before proceeding to the next step.

3. Select "MolD" in the left bar with installed programs. Note that the program is only available on those Galaxy servers that have previously included MolD among their installed software tools.
4. In MolD, select the previously uploaded FASTA file as the input file.
5. Click on "See taxa for selection" to open a menu with a list of all taxa/species in the input file. Select those taxa which you wish to diagnose against each other. You can also select "All"

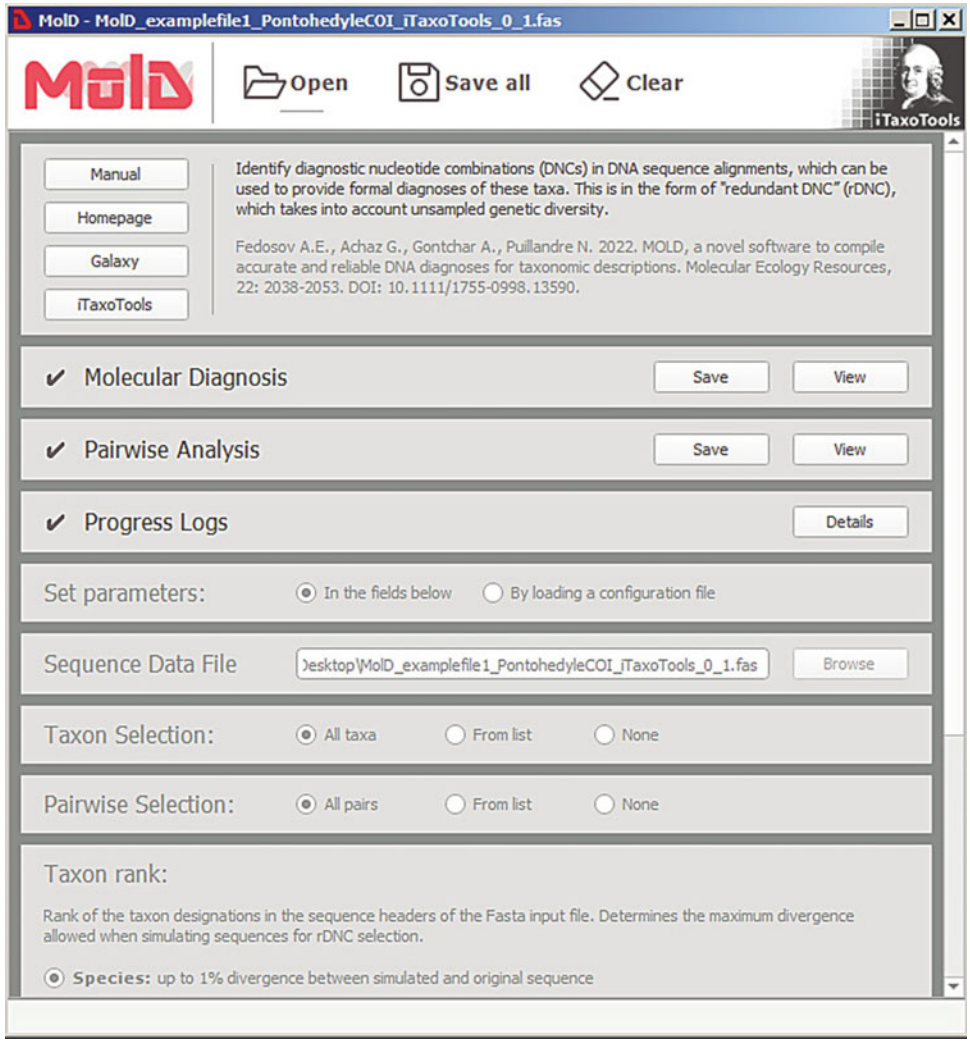

Fig. 3 Screenshot of the MolD stand-alone executable after a successfully completed run. Results for Molecular Diagnosis and Pairwise Analysis, as well as a progress logfile, can be viewed or saved using the respective buttons

but be aware that with large data files, computation time can be substantial.

6. MolD for Galaxy does not yet implement all of the options of the current version of MolD. The available options can be seen in the menu and under "Advanced parameters for pDNC recovery."
7. Click on "Execute" to run the program.
8. The output files can be viewed or downloaded from the "history" bar links. Note that MolD for Galaxy at present does not

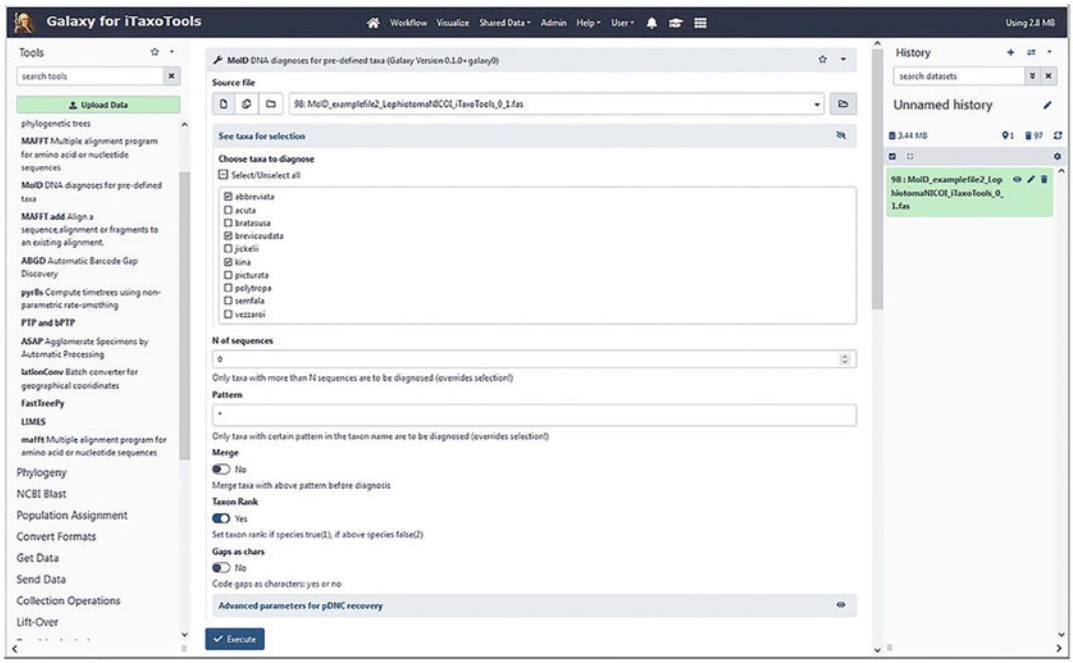

Fig. 4 Screenshot of the MolD starting page in Galaxy. The MolD example file has been uploaded and three taxa selected to be diagnosed against each other

support the option to output all pairwise diagnostic sites between taxa.

3.6 Interpreting the MolD Output

1. The two output files (molecular diagnosis and pairwise molecular diagnosis) are saved with .html extension and are best viewed in web browsers. However, it is also possible to open and view them in any text editor or to inspect them directly in the MolD-GUI viewer.
2. The first output file (molecular diagnosis) summarizes the main output of the MolD algorithm. After recapitulating the main parameter settings, it provides for each of the taxa selected for the analysis, basic information such as number of sequences analyzed, numbers of mDNCs and independent mDNCs recovered, rDNC scores, and a list of single-nucleotide mDNCs (=uDNPs) identified for the taxon (if any). Most importantly, the file provides for each taxon a final rDNC as well as a diagnosis text that can be copy-pasted into the respective section of a research paper for the purpose of diagnosing a taxon. Note that the rDNC is a diagnostic combination of positions, i.e., only the combination of the listed positions will provide a robust diagnosis (while each position on its own might or might not be universally diagnostic). Please,

also note that MolD may fail to recover a diagnosis for a selected query taxon (see point 4 below).

3. The second output file (pairwise) provides simple lists of nucleotide positions that differentiate two taxa (pDNPs), in a textual format, first listing the differences of the first vs. the second and then of the second vs. the first taxon. Each of these pDNPs allows diagnosing the two taxa against each other, but none of them has been checked for diagnostic use against other taxa. Also, similar to mDNCs, they are not robust against single mutations. If required for a pairwise differential diagnosis, the respective text can be directly copy-pasted into a research paper.
4. Troubleshooting: If no rDNCs are identified for a query taxon or the number of identified mDNCs is too small (< 10), then this may be caused by one of the following three issues: (i) sequence assignment to a wrong taxon will make both taxa (the one to which the sequence in fact belong, and the one, to which it is mistakenly assigned) undiagnosable; (ii) the query taxon is too poorly genetically differentiated—in this case even if some mDNCs are recovered, an attempt to identify an rDNC will likely fail; and (iii) too stringent parameters are selected that preclude recovery of an rDNC. Consider reducing values of NmaxSeq, and/or Pdiff, and/or opting for a more relaxed scoring. See the manual for more details on the possible settings.

4 Notes

1. Never forget to check your raw data as uncritical bioinformatic processing conveys the risk of overlooking flaws in the raw data. Critically explore the results of your analysis and correct the raw data if necessary, for instance, by inspecting a tree calculated from a set of sequences, or checking intra- and interspecific distances with a tool such as TaxI2 (Vences et al. Chap. 18).
2. Take into account biological phenomena such as different evolutionary ages of taxa and hybridization and introgression that can lead to the failure of taxo-informatic tools to identify reliable diagnostic sites.

References

1. Renner SS (2016) A return to Linnaeus's focus on diagnosis, not description: the use of DNA characters in the formal naming of species. Syst Biol 65:1085–1095. https://doi.org/10.1093/sysbio/syw032
2. ICZN (ed) (1999) International code of zoological nomenclature, 4th edn. International Trust for Zoological Nomenclature, London. http://www.iczn.org/iczn/index.jsp. Accessed 10 Jan 2023
3. Turland NJ, Wiersema JH, Barrie FR, Greuter W, Hawksworth DL, Herendeen PS, Knapp S, Kusber W-H, Li D-Z, Marhold K, May TW, Mc Neill J, Monro AM, Prado J,

Price MJ, Smith GF (eds) (2018) International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017. Regnum Vegetabile 159. Koeltz Botanical Books, Glashütten. https://doi.org/10.12705/Code.2018

4. Bickford D, Lohman DJ, Sodhi NS, Ng PK, Meier R, Winker K, Ingram KK, Das I (2007) Cryptic species as a window on diversity and conservation. Trends Ecol Evol 22:148–155. https://doi.org/10.1016/j.tree.2006.11.004
5. Fišer C, Robinson CT, Malard F (2018) Cryptic species as a window into the paradigm shift of the species concept. Mol Ecol 27:613–635. https://doi.org/10.1111/mec.14486
6. Sharkey MJ, Janzen DH, Hallwachs W, Chapman EG, Smith MA, Dapkey T, Brown A, Ratnasingham S, Naik S, Manjunath R, Perez K, Milton M, Hebert P, Shaw SR, Kittel RN, Solis MA, Metz MA, Goldstein PZ, Brown JW, Quicke DLJ, van Achterberg C, Brown BV, Burns JM (2021) Minimalist revision and description of 403 new species in 11 subfamilies of Costa Rican braconid parasitoid wasps, including host records for 219 species. ZooKeys 1013:1–665. https://doi.org/10.3897/zookeys.1013.55600
7. Vences M (2020) The promise of next-generation taxonomy. Megataxa 1:35–38. https://doi.org/10.11646/megataxa.1.1.6
8. Jörger KM, Schrödl M (2013) How to describe a cryptic species? Practical challenges of molecular taxonomy. Front Zool 10:59. https://doi.org/10.1186/1742-9994-10-59
9. Fedosov A, Achaz G, Gontchar A, Puillandre N (2022) Mold, a novel software to compile accurate and reliable DNA diagnoses for taxonomic descriptions. Mol Ecol Resour 22:2038–2053. https://doi.org/10.1111/1755-0998.13590
10. Sarkar IN, Planet PJ, Desalle R (2008) CAOS software for use in character-based DNA barcoding. Mol Ecol Resour 8:1256–1259. https://doi.org/10.1111/j.1755-0998.2008.02235.x
11. Merckelbach LM, Borges LMS (2020) Make every species count: fastachar software for rapid determination of molecular diagnostic characters to describe species. Mol Ecol Resour 20:1761–1768. https://doi.org/10.1111/1755-0998.13222
12. Hütter T, Ganser MH, Kocher M, Halkic M, Agatha S, Augsten N (2020) DeSignate: detecting signature characters in gene sequence alignments for taxon diagnoses. BMC Bioinform 21:151. https://doi.org/10.1186/s12859-020-3498-6
13. Vences M, Miralles A, Brouillet S, Ducasse J, Fedosov A, Kharchev V, Kostadinov I, Kumari S, Patmanidis S, Scherz MD, Puillandre N, Renner SS (2021) iTaxoTools 0.1: Kickstarting a specimen-based software toolkit for taxonomists. Megataxa 6:77–92. https://doi.org/10.11646/megataxa.6.2.1
14. Kühn AL, Haase M (2020) QUIDDICH: QUick IDentification of DIagnostic CHaracters. J Zool Syst Evol Res 58:22–26. https://doi.org/10.1111/jzs.12347
15. Blankenberg D, Von Kuster G, Bouvier E, Baker D, Afgan E, Stoler N, Galaxy Team, Taylor J, Nekrutenko A (2014) Dissemination of scientific software with Galaxy ToolShed. Genome Biol 15:403. https://doi.org/10.1186/gb4161
16. Ganser MH, Santoferrara LF, Agatha S (2022) Molecular signature characters complement taxonomic diagnoses: a bioinformatic approach exemplified by ciliated protists (Ciliophora, Oligotrichea). Mol Phylogenet Evol 170:107433. https://doi.org/10.1016/j.ympev.2022.107433
17. Galaxy Community (2022) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. Nucleic Acids Res 50:W345–W351. https://doi.org/10.1093/nar/gkac247
18. Vences M, Patmanidis S, Kharchev V, Renner SS (2022) Concatenator, a user-friendly program to concatenate DNA sequences, implementing graphical user interfaces for MAFFT and FastTree. Bioinform Adv 2:vbac050. https://doi.org/10.1093/bioadv/vbac050
19. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780