

## 1 iTaxoTools 0.1: Kickstarting a specimen-based software toolkit for taxonomists

2  
3 Miguel Vences<sup>1\*</sup>, Aurélien Miralles<sup>2</sup>, Sophie Brouillet<sup>2</sup>, Jacques Ducasse<sup>3</sup>, Alexander Fedosov<sup>4</sup>,  
4 Vladimir Kharchev<sup>1</sup>, Ivaylo Kostadinov<sup>5</sup>, Sangeeta Kumari<sup>1</sup>, Stefanos Patmanidis<sup>6</sup>, Mark D.  
5 Scherz<sup>7</sup>, Nicolas Puillandre<sup>2</sup>, Susanne S. Renner<sup>8</sup>

6  
7 1 Department of Evolutionary Biology, Zoological Institute, Technische Universität Braunschweig,  
8 Mendelssohnstraße 4, 38106 Braunschweig, Germany

9 2 Institut de Systématique, Évolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne  
10 Université, EPHE, Université des Antilles 57 rue Cuvier, CP 50, 75005 Paris, France

11 3 49 rue Eugène Carrière, 75018 Paris, France

12 4 A.N. Severtsov Institute of Ecology and Evolution, Russian Academy of Sciences, Leninsky prospect 33, 119071  
13 Moscow, Russian Federation

14 5 GFBio - Gesellschaft für Biologische Daten e.V., c/o Research II, Campus Ring 1, 28759 Bremen, Germany

15 6 School of Electrical and Computer Engineering, National Technical University of Athens, Iroon Polytechniou St 9,  
16 15780 Athens, Greece

17 7 Faculty of Mathematics and Natural Sciences, Institute for Biochemistry and Biology, University of Potsdam,  
18 Potsdam, Germany

19 8 Department of Biology, Washington University, 1 Brookings Drive, Saint Louis, MO 63130, USA

20 \* Corresponding author; m.vences@tu-braunschweig.de

### 21 22 23 **Abstract**

24  
25 While powerful and user-friendly software suites exist for phylogenetics, and an impressive  
26 cybertaxonomic infrastructure of online species databases has been set up in the past two decades,  
27 software specifically targeted at facilitating alpha-taxonomic work, i.e., delimiting and  
28 diagnosing species, is still in its infancy. Here we present a project to develop a bioinformatic  
29 toolkit for taxonomy, based on open-source Python code, including tools focusing on species  
30 delimitation and diagnosis and centered around specimen identifiers. At the core of iTaxoTools is  
31 user-friendliness, with numerous autocorrect options for data files and with intuitive graphical  
32 user interfaces. Assembled standalone executables for all tools or a suite of tools with a launcher  
33 window will be distributed for Windows, Linux, and Mac OS systems, and in the future also  
34 implemented on a web server. The alpha version (iTaxoTools 0.1) distributed with this paper  
35 contains GUI versions of six species delimitation programs (ABGD, ASAP, DELINEATE,  
36 GMYC, PTP, tr2) and a simple threshold-clustering delimitation tool. There are also new Python  
37 implementations of existing algorithms, including tools to compute pairwise DNA distances,  
38 ultrametric time trees based on non-parametric rate smoothing, species-diagnostic nucleotide  
39 positions, and standard morphometric analyses. Other utilities convert among different formats of  
40 molecular sequences, geographical coordinates, and units; merge, split and prune sequence files  
41 and tables; and perform simple statistical tests. As a future perspective, we envisage iTaxoTools  
42 to become part of a bioinformatic pipeline for next-generation taxonomy that accelerates the  
43 inventory of life while maintaining high-quality species hypotheses.

44  
45 **Key words:** integrative taxonomy, molecular diagnosis, species delimitation, ABGD, PTP,  
46 GMYC, TR2, DELINEATE, Limes.

## 50 **Introduction**

51

52 Bioinformatics has become the core of modern biology, especially in the context of high-  
53 throughput workflows that are becoming commonplace in many fields, in particular related to -  
54 omics approaches. The big data volumes obtained by these techniques require ever more efficient  
55 and sophisticated software, which is being developed and refined at a vigorous pace. In the field  
56 of systematics, powerful programs for phylogenetic analysis have been developed, and databases  
57 and data aggregators have been set up to deal with the massive globally-generated taxonomic  
58 dataset comprised of over one million species and many millions of specimen records. Yet, only  
59 few bioinformatic tools so far have been tailored to specifically fit the practical work of  
60 taxonomists, who diagnose and name some 15,000–20,000 new species of organisms per year, a  
61 task that still is largely performed by single or small teams of (professional and amateur)  
62 researchers (Miralles et al. 2020). Most existing tools are aimed at the construction of  
63 identification keys (e.g. Dallwitz, 1974; Clark 2003; Delgado Calvo-Flores et al. 2006; Zhang et  
64 al. 2006; MacLeod 2008; Vignes Lebbe 2015; Tofilski 2018), which in some groups help field  
65 identification. Only a handful of software packages (EDIT: cybertaxonomy.eu, TaxonWorks:  
66 taxonworks.org, Scratchpads: scratchpads.org) are tailored towards facilitating descriptive work  
67 itself, but none of these is so far widely used; furthermore, these programs do not include various  
68 important aspects of the alpha-taxonomic workflow, such as species delimitation or molecular  
69 diagnosis (Miralles et al. 2020) which can also be of high relevance for other fields such as  
70 molecular ecology.

71 Although most taxonomic studies are still relying on morphology only (as shown in a recent  
72 review; Miralles et al. 2020), taxonomy increasingly integrates diverse lines of evidence (Padial  
73 et al. 2010), a procedure called integrative taxonomy by Dayrat (2005). Discovering, delimiting,  
74 diagnosing, and naming new species requires taxonomists to examine voucher specimens and  
75 associated catalogues, field books and pictures; take, tabulate and statistically analyze  
76 morphometric measurements; define, tabulate and document phenotypic character states; estimate  
77 geographical ranges based on specimen provenances; align and analyze DNA sequences; and  
78 elaborate accurate specimen tables, species diagnoses and identification keys. Depending on the  
79 organism under study, it also may involve more specialized procedures such as comparing  
80 acoustic and visual signal repertoires of animals, or isolate and culture unicellular organisms. In  
81 addition, to fulfil standards of cybertaxonomy, data sets need to be archived in specialized  
82 repositories and new species names registered in online databases (Miralles et al. 2020). With  
83 rising best-practice standards, these many and varied tasks generally involve the use of different  
84 computer programs – and thus lead to an extra burden on taxonomists who may lack  
85 bioinformatic training.

86

87

## 88 **The concept of iTaxoTools**

89

90 We aim to develop a bioinformatic platform to facilitate the core work of taxonomists, that is,  
91 delimiting, diagnosing and describing species. Our initiative produced an integrative taxonomy  
92 toolkit – iTaxoTools (Fig. 1; Table 1). The concept of iTaxoTools rests on four pillars: (1) **fully**  
93 **open source** code; (2) a **diversified** set of stand-alone programs (‘modules’) that in future  
94 versions will become increasingly interconnected; (3) a **specimen-centered** architecture, where at  
95 present tables (tab-delimited text files) with specimen identifier columns serve as main input

96 format; and (4) a focus on **user-friendliness**, accessibility, and clear and transparent  
97 documentation.

98 All of the code developed by us is **fully open source** and available from a dedicated GitHub  
99 repository (<https://github.com/iTaxoTools>). In the case of tools programmed by other researchers,  
100 we make this information transparent, and the GUI we added specifies the original references and  
101 programmers. The current pre-release of compiled executables is available from  
102 <https://github.com/iTaxoTools/iTaxoTools-Executables>. See Table 2 for repositories of each  
103 single tool.

104 The toolkit is **diversified**, including simple format converters of molecular or geographic  
105 data, text and spreadsheet merging and pruning, simple statistical analyses e.g. of morphometric  
106 data, but especially focuses on two main aspects: species delimitation and diagnosis, based on  
107 multiple kinds of data.

108 The distribution of the tools is also diversified, including command-line tools for those users  
109 familiar/comfortable with Python; standalone GUI executables of each module for Windows,  
110 Linux, and Mac operating systems for those looking for a single functionality, e.g. a converter, to  
111 be called from a single and easily portable file – these tools will necessarily be ‘heavier’ and  
112 slightly slower than command-line executables; and a single software package containing all  
113 libraries (currently developed for Windows and Linux), from which each module can be launched  
114 (Fig. 2). In the future, the latter software package will also enable data transfer between different  
115 modules. The GUI software versions are designed to be stable over many different versions of the  
116 respective operating systems, e.g., from Windows 7 to Windows 10.

117 Alpha taxonomy is a primarily **specimen-centered** research field in which specimens –  
118 mostly single individual organisms or parts thereof, or cultured isolates composed of multiple  
119 individuals – are grouped into species. Consequently, iTaxoTools has implemented tab-delimited  
120 text as standard format for most tools, with one column indicating the specimen identifier. This  
121 will in subsequent versions allow the user to save the output of different tools for each specimen,  
122 and combine these results for further analysis. The tab-delimited format also allows easy editing  
123 of the data tables in spreadsheet editors. This specimen-based architecture needed for alpha-  
124 taxonomic programs remains valid whether specimens are represented by physical vouchers,  
125 images, or in the future maybe by full genome sequences.

126 Simplicity and **user-friendliness** are at the core of the toolkit we are developing. Because  
127 the majority of taxonomists is not familiar with programming languages, such as Python, all our  
128 tools are accessible via graphical user interfaces (GUI) – analyses can therefore be carried out  
129 with a few intuitive mouse clicks, under default or custom settings, without the need to enter  
130 commands in a command line. We also have added autocorrect routines to avoid the loss of time  
131 associated with the search for small misspellings or incorrect characters in input files that cause  
132 programs to fail. Furthermore, we will provide detailed manuals and wikis with screenshots,  
133 along with example files. We chose Python as the main programming language for our package,  
134 because it is characterized by its good readability and simple-to-learn syntax, and we documented  
135 newly written code extensively, to allow its re-use by other programmers. This comes at the cost  
136 of speed that would have been achieved by using the C programming language, but our toolkit in  
137 this early phase is not designed to cope with huge genomic datasets or analyses with tens of  
138 thousands of specimens. Currently iTaxoTools is designed to provide support for the most  
139 common taxonomic research projects that discover and name a limited number of species only  
140 (Miralles et al. 2020), but will be extended to large-scale projects in the future.

141 Considering that powerful programs exist for phylogenetic, phylogenomic and DNA  
142 metabarcoding analyses, we did not attempt to include such functionalities in our toolkit.  
143 Similarly, we also did not focus on dedicated multiple sequence alignment programs or genome  
144 assemblers because (i) these bioinformatic tasks are more efficiently carried out by programs  
145 written in C language, (ii) GUI-driven programs and pipelines already exist for alignment and  
146 phylogeny (e.g., PAUP, MEGA, PAUP, BEAST: Swofford 2003; Kumar et al. 2018; Bouckaert  
147 et al. 2019), genomics, and DNA metabarcoding (e.g., Anslan et al. 2017) and (iii) there is an  
148 active community both of commercial companies and academic research teams constantly  
149 extending these kinds of programs. We are, however, adding graphical user interfaces and new  
150 functionalities to other existing tools that are important for analyses in the context of systematics  
151 and that are not yet optimized with user-friendly GUIs. For instance, we have updated the code of  
152 Partitionfinder (Lanfear et al. 2016) from Python v. 2 to v. 3, and aim to add a GUI also to the  
153 sequence alignment program MAFFT (Kato & Standley 2013). These developments will be  
154 added successively to iTaxoTools.

155  
156

### 157 **Functionalities implemented in iTaxoTools 0.1**

158

159 Our work on iTaxoTools is ongoing and will be intensified in the period 2021–2023 thanks to  
160 support by the DFG SPP 1991 TaxonOmics priority funding program. The current version,  
161 published along with the present article, already includes a series of functional tools that we  
162 predict will be useful in different steps of the alpha-taxonomic workflow, Data Preparation  
163 (mainly Conversion), Analysis, Delimitation, and Diagnosis.

164

#### 165 Data Preparation

166 Several tools convert among data formats, with the major modules being **dnaconvert** for  
167 converting among common DNA sequence formats, **latlonconverter** for converting among  
168 geographic coordinate systems (elaborated upon below), and **pyr8s** for converting non-  
169 ultrametric trees to ultrametric. A collection of simpler tools includes **fastmerge** and **fastsplit** for  
170 splitting and merging large fasta and fastq files, including advanced filtering options by sequence  
171 name or sequence motif; **specimentablemerger** and **specimentablepruner** for splitting and  
172 merging tab-delimited text files by specimen identifiers; **linebreaker** for converting among Linux  
173 and Windows line-break styles (often necessary when processing input files from other  
174 bioinformatic tools); **nodenamecorrector** for replacing all non-standard ASCII characters from  
175 Newick-format trees; and **unitconverter** for distance, time, volume, molarity, and other units.

176 **dnaconvert** is a versatile tool to convert DNA (and protein) sequence data among  
177 commonly used formats such as fasta, fastq, phylip, or nexus (Fig. 3). Compared to other  
178 sequence format converters, dnaconvert is particularly user-friendly in that it autocorrects  
179 numerous issues that usually create compatibility problems, e.g., by automatically replacing non-  
180 standard ASCII characters from sequence names or auto-renaming sequences in formats of  
181 limited sequence name length such as phylip. A main novelty is the support for tab-delimited files  
182 because in our experience, it is useful, for small to medium-sized taxonomy projects, to store and  
183 organize specimen-based DNA sequence information (DNA barcodes) in spreadsheet editors  
184 such as Microsoft Excel or its freeware equivalents Libre Office / Open Office Calc. From these  
185 spreadsheets, it is then easy to copy-paste the sequence, specimen-voucher, species and locality  
186 columns into dnaconvert and obtain a sequence file for analysis, e.g. in fasta format, with all  
187 respective information concatenated in the sequence name. The program also supports a format in

188 which these metadata are bracketed as required for uploading the sequence data along with  
189 metadata to the NCBI Genbank repository (i.e., via Submission Portal or BankIt). Lastly, the  
190 program also converts Genbank flatfiles into a tabular format, allowing the user to immediately  
191 have all relevant metadata associated with the Genbank record in separate columns in a  
192 spreadsheet.

193 **latlonconverter** allows batch conversion of geographic coordinates from a large number of  
194 different formats into standard decimal format as required by most geographical information  
195 system (GIS) programs. By performing a series of autocorrections of possible typos and then  
196 using a heuristic approach, latlonconverter is able to recognize and transform many idiosyncratic  
197 formats of geographical coordinates as they are commonly found in specimen databases  
198 containing geographical information taken by different researcher. With **spartmapper**,  
199 geographical coordinates in combination with a species partition file (spart; Miralles et al.  
200 submitted) can be previewed on a map, and then transformed in a kml file that plots all localities  
201 on Google Earth and visualizes the geographical distribution of respective species hypothesis  
202 (Fig. 4).

203 **pyr8s** is one of our flagship modules (Fig. 5). For many evolutionary analyses, but also for  
204 species delimitation (e.g. GMYC), ultrametric phylogenies are required where non-ultrametric  
205 trees are available. This conversion is rather complex and can be time-intensive. While numerous  
206 programs exist to calculate time trees (e.g., MCMCtree, BEAST, MEGAX: Yang & Rannala  
207 2006; Bouckaert et al. 2019; Kumar et al. 2018), they usually require DNA sequence information  
208 in addition to a previously inferred phylogenetic tree. For iTaxoTools, we opted to recuperate a  
209 vintage approach, non-parametric rate smoothing (NPRS), initially developed by Sanderson  
210 (1997) and later implemented as part of the program r8s (Sanderson 2003). This method only  
211 requires a phylogenetic tree as input, with the option to add one or more time calibration points.  
212 NPRS has previously been implemented in the R package ape (Paradis et al. 2004), but was  
213 removed from the latter and from the newest releases of r8s due to licensing issues. Specifically,  
214 the original version of r8s relied on a modified implementation of Powell's conjugate direction  
215 method which was incompatible with open-source licensing (Powell 1964; Gill et al. 1981; Press  
216 et al. 1992). In the GUI-driven tool **pyr8s**, the NPRS algorithm has been newly coded, making  
217 use of the open-source libraries DendroPy (Sukumaran & Holder 2010) and SciPy (Virtanen et al.  
218 2020), thus resolving the previous licensing issues. This new version provides a GUI for user-  
219 friendly setting of time constraints, exposes a Python interface for lower-level analysis and  
220 maintains support for r8s-formatted input files.

## 221 Analysis

222 We include several data analysis modules: **TaxI2** for calculation of pairwise distances among  
223 individuals, and **morphometricanalyzer** for basic morphometric analyses (elaborated upon  
224 below). For convenience, we also include **simplestatscalculator**, a utility for quick, basic  
225 statistical analyses of manually entered or pasted data.

226 **TaxI2** is a tool for pairwise sequence comparison. To analyze DNA barcoding data sets,  
227 Steinke et al. (2005) proposed the program TaxI, which performs pairwise alignments between  
228 sequences and calculates pairwise distances based on these alignments. Compared to a multiple  
229 sequence alignment (MSA) the authors argued that these distance calculations may be more  
230 accurate in the case of highly divergent markers including multiple insertions and deletions, such  
231 as stretches of mitochondrial ribosomal RNA genes. The pure-Python tool **TaxI2** performs  
232 similar calculations, with numerous added functionalities such as support for pre-MSA aligned  
233 data sets. The tool has two main analysis modes: First, following the original TaxI approach, it

235 can compare a set of sequences against a reference database, via pairwise alignments, identifies  
236 for each query the closest reference sequence, and calculates various genetic distances among the  
237 two. Second, it also can perform all-against all comparisons of a set of sequences. In this latter  
238 approach, sequences can be added in tab-delimited table format along with species name, and  
239 from these data the program calculates within-species, between-species, and between-genus  
240 distances. Various metrics and graphs defining the barcode gap in a given data set are also  
241 included in the output. The program furthermore performs a simple threshold-based clustering of  
242 DNA sequences into OTUs, following the approach previously implemented in TaxonDNA  
243 (<http://taxondna.sourceforge.net/>; Meier et al. 2006), and outputs the resulting species partition as  
244 SPART file (Miralles et al. 2021)

245 **morphometricanalyzer** is our tool for exploratory analysis of morphometric datasets.  
246 Integrative taxonomists do not only use molecular data. In many cases, a limited number of one-  
247 dimensional morphometric measurements such as body length and width (or leaf length and  
248 width in plants) are taken and compared among groups of individuals. For simple statistical  
249 analyses, we have included the tool **morphometricanalyzer** which performs a series of  
250 exploratory routine comparisons from morphometric data. It takes as input tab-delimited text files  
251 with species hypotheses and a series of other optional categories, and then performs automatically  
252 a series of statistical comparisons between species (and between other categories), such as  
253 calculations of means, medians, standard deviation, minimum and maximum values; pairwise  
254 Mann-Whitney U-tests and Student's t-tests between all pairs of species; a simple Principal  
255 Component analysis; and calculation of ratios among original values as a means to size-correct  
256 them, followed by statistical comparison of these size-corrected values. Finally, the program also  
257 outputs pre-formulated taxonomic diagnoses, with full-text sentences specifying by which  
258 morphometric value or ratio a species/population differs from other species/populations with  
259 statistical significance, or without value overlap. It would also be possible to explore non-  
260 morphological (e.g. bioacoustic) data with this tool, although it is primarily developed for  
261 morphometrics.

## 262 263 Delimitation

264 A special emphasis in the first development phase of iTaxoTools is species delimitation, a  
265 burgeoning field in systematics. The available species delimitation algorithms mostly use DNA  
266 sequence data and tend to overestimate the number of species in a data set (e.g., Miralles et al.  
267 2013); indeed, they may delimit populations rather than species (Sukumaran & Knowles 2017).  
268 Yet, such automated delimitation may play a role in formulating initial species hypotheses that  
269 can then be tested in an integrative taxonomy pipeline. In the first version of iTaxoTools, we have  
270 focused on tools already available in Python programming language. For these tools, we added  
271 user-friendly GUIs and slightly extended the functionality, for example by enabling them to  
272 output species partition information in the standardized "spart" format proposed by Miralles et al.  
273 (2021). The current version of iTaxoTools includes GUI-enhanced versions of **PTP** (Zhang et al.  
274 2013) (Fig. 6) and **GMYC** (Pons et al. 2006; Fujisawa & Barraclough 2013; Python version J.  
275 Zhang) which delimit species from single-locus trees; **tr2** (Fujisawa et al. 2016) and  
276 **DELINEATE** (Sukumaran et al. 2020) that use coalescence-based approaches on multiple gene  
277 trees; and **ABGD** (Puillandre et al. 2012) (Fig. 7) and **ASAP** (Puillandre et al. 2020) that are  
278 alignment-based and rely on calculations of genetic distances. For some of these tools (PTP,  
279 GMYC, tr2, DELINEATE) the current pre-release GUI versions are still basic and only run under  
280 default settings; options to change and refine parameters will be added to the first complete

281 release. iTaxoTools also includes **LIMES 2.0**, a program to handle and compare species  
282 partitions obtained by these various approaches (Ducasse et al. 2020, Miralles et al. 2021).

283

### 284 Diagnosis

285 The diagnosis of new species – rather than its lengthy description – represents the most important  
286 part of the alpha-taxonomic process, and in all Nomenclatural Codes, diagnosis can be based on  
287 molecular, as well as morphological characters (Renner, 2016). Several software tools have been  
288 proposed to extract diagnostic nucleotide positions of clades and species, either phylogeny-based  
289 (caos; Sarkar et al. 2008) or primarily alignment-based (MolD, Fastachar, DeSignate: Fedosov et  
290 al. 2019; Merckelbach & Borges 2020; Hütter et al. 2020). In order to facilitate the use of such  
291 DNA characters in differential diagnoses of new species, we implemented a crucial new tool for  
292 DNA taxonomy named **dnadiagnoser**. Compared to other tools, dnadiagnoser has various  
293 functionalities to improve the use of DNA characters in species diagnosis. It takes as input tab-  
294 delimited text files in which one column specifies the unit for analysis (typically the species), and  
295 provides as output pre-formulated text sentences which specify (i) in a pairwise fashion, all the  
296 diagnostic sites of one species against all other species, and (ii) the unique diagnostic sites (if  
297 any) that differentiate a species against all other species. These text sentences can then directly be  
298 used in species diagnoses. As a further innovation dnadiagnoser interprets one of the sequences in  
299 the input alignment as reference sequence and outputs the diagnostic sites relative to this  
300 sequence. To facilitate such comparisons, the program also includes a series of standard reference  
301 sequences (such as the full *Homo sapiens* COI or *cox1* gene) and allows as input unaligned  
302 sequences, which are then pairwise aligned against the reference sequence to identify diagnostic  
303 positions and label them according to their position in the reference sequence, a procedure that  
304 works reliably in sets of sequences with no or only few insertions or deletions such as COI. In  
305 addition, we have also programmed a GUI for **MolD** (Fedosov et al. 2019), a program that is  
306 tailored for recovering DNA-based diagnoses in large DNA dataset, and is capable of identifying  
307 diagnostic combinations of nucleotides (DNCs) in addition to single (pure) diagnostic sites. The  
308 crucial and unique functionality of MolD allows assembling DNA diagnoses that fulfil pre-  
309 defined criteria of reliability, which is achieved by repeatedly scoring diagnostic nucleotide  
310 combinations against datasets of in-silico mutated sequences.

311

312

### 313 **Future extensions**

314

315 Our goal with this paper is to make the tools we have developed available to the community  
316 as soon as possible so they may be critically evaluated and improved. The next developments will  
317 be in three fields: (i) **Geography**: iTaxoTools will not compete with geographical information  
318 systems (GIS), but there are a number of recurrent and rather simple geographical analyses in  
319 alpha-taxonomy that can be facilitated by bioinformatic tools, in particular calculation of linear  
320 distances among sites and of the surface (minimum convex polygon) of a distribution range of a  
321 species, based on a set of georeferenced locality points, and most importantly, a simple graphical  
322 editor that outputs publication-ready distribution maps, with customizable colors and symbols for  
323 different species, from a set of georeferenced locality records. For more sophisticated analyses,  
324 connecting iTaxoTools (via data formats such as SPART) with dedicated toolboxes for analysis  
325 of spatial biodiversity data such as SDMTtoolbox (Brown et al. 2017) could allow e.g. for  
326 comparative niche modelling of alternative species partitions. (ii) **Extraction of diagnostic traits  
327 from specimen data**: Besides molecular diagnosis with MolD and dnadiagnoser, we plan to

328 develop a tool that automatically outputs diagnoses based on (specimen-based) categorical data  
329 sets of morphological characters. (iii) **Connection to other programs:** We also plan to explore  
330 options to connect iTaxoTools to the DELTA (DEscription Language for TAXonomy) software  
331 package (Coleman et al. 2010). DELTA is a format for coding descriptive taxonomic information  
332 that however is primarily species-based (not specimen-based as iTaxoTools), and a series of  
333 programs have been developed on this basis, spearheaded by M. Dallwitz at CSIRO (Canberra,  
334 Australia) (Dallwitz 1974, 1980). The new Free DELTA platform launched in 2000  
335 (<http://freedelta.sourceforge.net/>) includes options for editing and maintenance of data sets in  
336 DELTA formats, as well as utilities for data conversion, interactive identification of taxa,  
337 automated generation of diagnostic keys, and descriptions. Especially, information on species-  
338 specific molecular and morphological characters identified in iTaxoTools could be seamlessly  
339 coded in DELTA, making use of pydelta (<http://freedelta.sourceforge.net/pydelta/>).

340 The biggest gap in taxonomy software so far is the integrative aspect in the sense of Dayrat's  
341 (2005) concept of integrative taxonomy (see also Padial et al. 2010). That is, the many available  
342 species delimitation programs all output a species hypothesis based on one analytical approach –  
343 usually based on only a molecular data set, with a few exceptions such as iBPP, which can  
344 integrate morphometric and molecular data (Solís-Lemus et al. 2015). Approaches that combine  
345 information from different lines of evidence into species delimitation are exceedingly scarce. As  
346 an example, DELINEATE (Sukumaran et al. 2020) allows the user to fix a series of species  
347 hypotheses (i.e., firmly assign a series of specimens to species) while letting the other specimens  
348 "float" freely in the analysis and assign them to either one of the previously defined species, or to  
349 a new species. Such an option of "prior" species delimitation should be universally available to  
350 users as a manual option (e.g., if evidence comes from field or experimental data on  
351 hybridization, genomic information, or other data that are yet difficult to code or implement in  
352 species delimitation software), similar to what is implemented in DELINEATE. But ideally,  
353 automated proposals of firm *a priori* evidence for two specimens to either belong to two species,  
354 or to the same species, could also be elaborated by the software – for instance, using evidence  
355 such as sympatric geographical occurrence without gene flow, full concordance between genetic  
356 and morphological characters, or exceedingly high genetic distances. We plan to develop  
357 concepts for such analysis priors, and start implementing them in a iTaxoTools webserver  
358 pipeline, in the next years.

359 Importantly, our project is open for other developers to join, and for the taxonomic  
360 community as a whole to provide suggestions. We especially welcome proposals of additional  
361 tools that could help to streamline and accelerate the whole process of delimiting and naming  
362 species (whether it concerns the initial step of data acquisition, their treatment, their analyses, or  
363 their final submission to a dedicated repository). Only practicing taxonomists know which parts  
364 of the alpha-taxonomic workflow for their group of taxa is particularly time-consuming, and  
365 where time and effort is lost with repetitive, manual tasks that could be as well automatically  
366 performed by a computer program – and thereby formulate requirements for such dedicated  
367 programs.

368  
369

### 370 **Perspectives for iTaxoTools**

371

372 The different taxonomic tools made available here are performing analyses offline on a local  
373 computer (and in the future will also be available on a webserver), but without linking to external  
374 resources. True next-generation taxonomy will require linking specimen-based taxonomy



375 software with online resources and databases, and scaling the analyses to data of many thousands  
376 of specimens. On the one hand, this involves archiving newly acquired data in dedicated  
377 repositories (Miralles et al. 2020). But on the other hand, it means aggregating for each specimen  
378 identifier DNA sequences, morphological characters, images, and increasingly -omics data (e.g.,  
379 Lendemer et al. 2020), and then entering these large-scale cyberspecimen data into species  
380 delimitation, diagnosis and naming pipelines. The process could be coupled with machine-  
381 learning programs to automatically extract diagnostic traits e.g. from images, with data  
382 aggregators such as GBIF ([gbif.org](http://gbif.org)) and online tools such as Map of Life ([mol.org](http://mol.org)), or Timetree  
383 of Life ([timetree.org](http://timetree.org)) to obtain geographical and temporal context, and distribution models for  
384 alternative species hypotheses. These bioinformatic opportunities may gain power under a view  
385 of species as probabilistic hypotheses that may allow defining probability thresholds of  
386 integrative taxonomic analysis above which lineages can be confidently named as species by  
387 semi-automated pipelines. While the current version of iTaxoTools is far from this vision, it may  
388 represent a seed for developing the necessary environment, and a sandbox to test software tools  
389 with the potential to significantly accelerate the inventory of life.

390  
391

## 392 **Acknowledgments**

393

394 We are grateful to Mike Sanderson for information on the initial code of r8s. This study was  
395 supported by the Deutsche Forschungsgemeinschaft (grant VE247/16-1 – HO 3492/6-1 and RE  
396 603/29-1) in the framework of the ‘TaxonOmics’ priority program.

397  
398

## 399 **References**

400

- 401 Anslan, S., Bahram, M., Hiiesalu, I. & Tedersoo, L. (2017) PipeCraft: Flexible open-source  
402 toolkit for bioinformatics analysis of custom high-throughput amplicon sequencing data.  
403 *Molecular Ecology Resources*, 17, e234-e240.
- 404 Bik, H.M. (2017) Let's rise up to unite taxonomy and technology. *PLoS Biology*, 15, e2002231.
- 405 Bouckaert, R., Vaughan, T.G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A.,  
406 Heled, J., Jones, G., Kühnert, D., De Maio, N., Matschiner, M., Mendes, F.K., Müller, N.F.,  
407 Ogilvie, H.A., du Plessis, L., Poppinga, A., Rambaut, A., Rasmussen, D., Siveroni, I.,  
408 Suchard, M.A., Wu, C.H., Xie, D., Zhang, C., Stadler, T. & Drummond, A.J (2019)  
409 BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS*  
410 *Computational Biology*, 15, e1006650.
- 411 Brooke, M. de L. (2000) Why museums matter. *Trends in Ecology and Evolution*, 15, 136–137.
- 412 Brown, J.L., Bennett, J.R. & French, C.M. (2017) SDMtoolbox 2.0: the next generation Python-  
413 based GIS toolkit for landscape genetic, biogeographic and species distribution model  
414 analyses. *PeerJ*, 5, e4095.
- 415 Calvo-Flores, M.D., Contreras, W.F., Galindo, E.G., & Pérez-Pérez, R. (2006) XKey: A tool for  
416 the generation of identification keys. *Expert Systems with Applications*, 30, 337–351.
- 417 Clark, J.Y. (2003) Artificial neural networks for species identification by taxonomists.  
418 *Biosystems*, 72, 131–147.

- 419 Coleman, C.O., Lowry, J.K. & Macfarlane, T. (2010) DELTA for Beginners: An introduction  
420 into the taxonomy software package DELTA. *ZooKeys*, 45, 1–75.
- 421 Crous, P.W., Gams, W., Stalpers, J.A., Robert, V. & Stegehuis G. (2004) MycoBank: an online  
422 initiative to launch mycology into the 21st century. *Studies in Mycology*, 50, 19–22.
- 423 Dallwitz, M.J. (1974) A flexible computer program for generating identification keys. *Systematic*  
424 *Zoology*, 23, 50–57.
- 425 Dallwitz, M.J. (1980) A general system for coding taxonomic descriptions. *Taxon*, 29, 41–46.
- 426 Dayrat, B. (2005) Toward integrative taxonomy. *Biological Journal of the Linnean Society*, 85,  
427 407–415.
- 428 De Queiroz, K. (2007) Species concepts and species delimitation. *Systematic Biology*, 56, 879–  
429 886.
- 430 De Mauro, A., Greco, M. & Grimaldi, M. (2016) A formal definition of Big Data based on its  
431 essential features. *Library Review*, 65, 122–135
- 432 Ducasse, J., Ung, V., Lecointre, G. & Miralles, A. (2020). LIMES: a tool for comparing species  
433 partition. *Bioinformatics*, 36, 2282–2283.
- 434 Fedosov, A., Achaz, G. & Puillandre, N. (2019) Revisiting use of DNA characters in taxonomy  
435 with MolD - a tree independent algorithm to retrieve diagnostic nucleotide characters from  
436 monolocus datasets. *bioRxiv*, 838151; doi: <https://doi.org/10.1101/838151>
- 437 Fujisawa, T., Aswad, A. & Barraclough, T.G. (2016) A rapid and scalable method for multilocus  
438 species delimitation using Bayesian model comparison and rooted triplets. *Systematic*  
439 *Biology*, 65, 759–771
- 440 Fujisawa, T. & Barraclough, T.G. (2013) Delimiting species using single-locus data and the  
441 Generalized Mixed Yule Coalescent approach: a revised method and evaluation on  
442 simulated data sets. *Systematic Biology*, 62, 707–724.
- 443 Güntsch, A., Groom, Q., Hyam, R., Chagnoux, S., Röpert, D., Berendsohn, W., Casino, A.,  
444 Droege, G., Gerritsen, W., Holetschek, J., Marhold, K., Mergen, P., Rainer, H., Smith, V. &  
445 Triebel, D. (2018) Standardised globally unique specimen identifiers. *Biodiversity*  
446 *Information Standards*, 2, e26658.
- 447 Gill, P. E., Murray, W. & Wright, M.H. (1981) *Practical optimization*. Academic Press, New  
448 York.
- 449 Heerlien, M., Van Leusen, J., Schnörr, S., De Jong-Kole S, Raes, N. & Van Hulsen, K. (2015)  
450 The natural history production line: An industrial approach to the digitization of scientific  
451 collections. *ACM Journal on Computing and Cultural Heritage*, 8, 3.
- 452 Hütter, T., Ganser, M.H., Kocher, M., Halkic, M., Agatha, S., Augsten, N. (2020) DeSignate:  
453 detecting signature characters in gene sequence alignments for taxon diagnoses. *BMC*  
454 *Bioinformatics*, 21, 151.
- 455 IISE (2011) State of Observed Species. Tempe, AZ. International Institute for Species  
456 Exploration. Accessed 15 March 2019. Available from: <http://species.asu.edu/SOS>

- 457 Katoh, K., Standley, D.M. (2013) MAFFT Multiple Sequence Alignment Software Version 7:  
458 improvements in performance and usability. *Molecular Biology and Evolution*, 30, 772–  
459 780.
- 460 Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. (2018) MEGA X: Molecular  
461 Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and*  
462 *Evolution*, 35, 1547–1549.
- 463 Lanfear, R., Frandsen, P.B., Wright, A.M., Senfeld, T. & Calcott, B. (2016) PartitionFinder 2:  
464 new methods for selecting partitioned models of evolution for molecular and morphological  
465 phylogenetic analyses. *Molecular Biology and Evolution*, 34, 772–773.
- 466 Larsen, B.B., Miller, E.C., Rhodes, M.K. & Wiens, J.J. (2017) Inordinate fondness multiplied and  
467 redistributed: the number of species on Earth and the new pie of life. *Quarterly Review of*  
468 *Biology*, 92, 229–265.
- 469 Lendemer, J., Thiers, B., Monfils, A.K., Zaspel, J., Ellwood, E.R., Bentley, A., LeVan, K., Bates,  
470 J., Jennings, D., Contreras, D., Lagomarsino, L., Mabee, P., Ford, L.S., Guralnick, R.,  
471 Gropp, R.E., Revelez, M., Cobb, N., Seltmann, K. & Aime, M.C. (2020) The extended  
472 specimen network: a strategy to enhance US biodiversity collections, promote research and  
473 education. *BioScience*, 70, 23–30.
- 474 MacLeod, N. (2008, ed.) *Automated Taxon Identification in Systematics: Theory, Approaches*  
475 *and Applications*. CRC Press, Boca Raton FL, USA. ISBN-13:978-0-8493-8205-5. 350 pp.
- 476 Merckelbach, L.M. & Borges, L.M.S. (2020) Make every species count: fastachar software for  
477 rapid determination of molecular diagnostic characters to describe species. *Molecular*  
478 *Ecology Resources*, 20, 1761–1768.
- 479 Meier, R., Kwong, S., Vaidya, G. & Ng, P.K.L. (2006) DNA Barcoding and taxonomy in  
480 Diptera: a tale of high intraspecific variability and low identification success. *Systematic*  
481 *Biology*, 55, 715–728.
- 482 Miralles, A., Bruy, T., Wolcott, K., Scherz, M.D., Begerow, D., Beszteri, B., Bonkowski, B.,  
483 Felden, J., Gemeinholzer, B., Glaw, F., Glöckner, F.O., Hawlitschek, O., Kostadinov, I.,  
484 Nattkemper, T.W., Printzen, C., Renz, J., Rybalka, N., Stadler, M., Weibulat, T., Wilke, T.,  
485 Renner, S.S., Vences, M. (2020) Repositories for taxonomic data: Where we are and what  
486 is missing. *Systematic Biology*, 69, 1231–1253.
- 487 Miralles, A. & Vences, M. (2013) New metrics for comparison of taxonomies reveal striking  
488 discrepancies among species delimitation methods in *Madascincus* lizards. *PLoS ONE*, 8,  
489 e68242.
- 490 Miralles, A., Ducasse, J., Brouillet, S., Flouri, T., Fujisawa, T., Kapli, P., Knowles, L.L., Kumari,  
491 S., Stamatakis, A., Sukumaran, J., Lutteropp, S., Vences, M. & Puillandre, N. (2021)  
492 SPART, a versatile and standardized data exchange format for species partition  
493 information. BioRxiv, doi: <https://doi.org/10.1101/2021.03.22.435428>
- 494 Nelson, G. & Ellis, S. (2018) The history and impact of digitization and digital data mobilization  
495 on biodiversity research. *Philosophical Transactions of the Royal Society B*, 374,  
496 20170391.
- 497 Padial, J.M., Miralles, A., De la Riva, I. & Vences, M. (2010) The integrative future of  
498 taxonomy. *Frontiers in Zoology*, 7, e16.

- 499 Paradis, E., Claude, J. & Strimmer, K. (2004) APE: Analyses of Phylogenetics and Evolution in  
500 R language, *Bioinformatics*, 20, 289–290.
- 501 Patterson, D.J., Cooper, J., Kirk, P.M., Pyle, R.L. & Remsen, D.P. (2010) Names are key to the  
502 big new biology. *Trends in Ecology and Evolution*, 25, 686–691.
- 503 Pons, J., Barraclough, T.G., Gomez-Zurita, J., Cardoso, A., Duran, D.P., Hazell, S., Kamoun, S.,  
504 Sumlin, W.D. & Vogler, A.P. (2006) Sequence-based species delimitation for the DNA  
505 taxonomy of undescribed insects. *Systematic Biology*, 55, 595–609.
- 506 Powell, M.J.D. (1964) An efficient method for finding the minimum of a function of several  
507 variables without calculating derivatives. *The Computer Journal*, 7, 155–162.
- 508 Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. (1992) *Numerical Recipes in C*.  
509 Cambridge University Press, New York. 2nd ed.
- 510 Puillandre, N., Brouillet, S. & Achaz, G. (2020) ASAP: assemble species by automatic  
511 partitioning. *Molecular Ecology Resources*. <https://doi.org/10.1111/1755-0998.13281>
- 512 Puillandre, N., Lambert, A., Brouillet, S. & Achaz, G. (2012) ABGD, Automatic Barcode Gap  
513 Discovery for primary species delimitation. *Molecular Ecology*, 21, 1864–1877.
- 514 Renner, S.S. (2016) A return to Linnaeus’s focus on diagnosis, not description: The use of DNA  
515 characters in the formal naming of species. *Systematic Biology*, 65, 1085–1095.
- 516 Riedel, A., Sagata, K., Surbakti, S., Tänzler, R. & Balke, M. (2013) One hundred and one new  
517 species of *Trigonopterus* weevils from New Guinea. *Zookeys*, 280, 1–150.
- 518 Roskov, Y., Ower, G., Orrell, T., Nicolson, D., Bailly, N., Kirk, P.M., Bourgoin, T., DeWalt,  
519 R.E., Decock, W., Nieukerken, E. van, Zarucchi, J. & Penev, L. (2019, eds.) Species 2000  
520 & ITIS Catalogue of Life, 26th February 2019. Digital resource at  
521 [www.catalogueoflife.org/col](http://www.catalogueoflife.org/col). Species 2000: Naturalis, Leiden, the Netherlands. ISSN 2405-  
522 8858.
- 523 Rupp, K. (2018) 42 Years of Microprocessor Trend Data. Website accessed 13 March 2019.  
524 Available from: <https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/>.
- 525 Ratnasingham, S. & Hebert, P.D. (2013) A DNA-based registry for all animal species: the  
526 barcode index number (BIN) system. *PLoS ONE*, 8, e66213.
- 527 Sanderson, M.J. (1997) A non-parametric approach to estimating divergence times in the absence  
528 of rate constancy. *Molecular Biology and Evolution*, 14, 1218–1231.
- 529 Sanderson, M.J. (2003) r8s: inferring absolute rates of molecular evolution and divergence times  
530 in the absence of a molecular clock. *Bioinformatics*, 19, 301–302.
- 531 Sarkar, I.N., Planet, P.J., Desalle, R. (2008) caos software for use in character-based DNA  
532 barcoding. *Molecular Ecology Resources*, 8, 1256–1259.
- 533 Sharkey, M.J., Janzen, D.H., Hallwachs, W., Chapman, E.G., Smith, M.A., Dapkey, T., Brown,  
534 A., Ratnasingham, S., Naik, S., Manjunath, R., Perez, K., Milton, M., Hebert, P., Shaw,  
535 S.R., Kittel, R.N., Solis, M.A., Metz, M.A., Goldstein, P.Z., Brown, J.W., Quicke, D.L.J.,  
536 van Achterberg, C., Brown, B.V. & Burns, J.M. (2021) Minimalist revision and description  
537 of 403 new species in 11 subfamilies of Costa Rican braconid parasitoid wasps, including  
538 host records for 219 species. *ZooKeys*, 1013, 1–665.

- 539 Solís-Lemus, C., Knowles, L.L. & Ané, C. (2015) Bayesian species delimitation combining  
540 multiple genes and traits in a unified framework. *Evolution*, 69, 492–507
- 541 Steinke, D., Salzburger, W., Vences, M. & Meyer, A. (2005) TaxI - A software tool for DNA  
542 barcoding using distance methods. – *Philosophical Transactions of the Royal Society*  
543 *London, Ser. B*, 360, 1975–1980.
- 544 Sukumaran, J. & Knowles, L.L. (2017) Multispecies coalescent delimits structure, not species.  
545 *Proceedings of the National Academy of the U.S.A.*, 114, 1607–1612.
- 546 Sukumaran, J., Holder, T.M. & Knowles, L.L. (2020) Incorporating the speciation process into  
547 species delimitation. <https://github.com/jeetsukumaran/delineate>.
- 548 Sukumaran, J. & Holder, M.T. (2010) DendroPy: A Python library for phylogenetic computing.  
549 *Bioinformatics*, 26, 1569–1571.
- 550 Swofford, D. L. (2003) PAUP\*. Phylogenetic Analysis Using Parsimony (\*and Other Methods).  
551 Version 4. Sinauer Associates, Sunderland, Massachusetts.
- 552 Tegelberg, R., Mononen, T. & Saarenmaa, H. (2014) High-performance digitization of natural  
553 history collections: Automated imaging lines for herbarium and insect specimens. *Taxon*,  
554 63, 1307–1313.
- 555 Tofilski, A. (2018). DKey software for editing and browsing dichotomous keys. *ZooKeys*, 735,  
556 131–140. <https://doi.org/10.3897/zookeys.735.21412>
- 557 Vignes Lebbe, R., Chesselet, P. & Diep Thi, M.H. (2015) Xper3: new tools for collaborating,  
558 training and transmitting knowledge on botanical phenotypes. In: Rakotoarisoa, N.R.,  
559 Blackmore, S., Riéra, B. (Eds) *Botanists of the 21st Century*. 11 pp.
- 560 Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski,  
561 E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J.,  
562 Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J.,  
563 Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R.,  
564 Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F.,  
565 van Mulbregt, P., SciPy 1.0 Contributors (2020) SciPy 1.0: fundamental algorithms for  
566 scientific computing in Python. *Nature Methods*, 17, 261–272.
- 567 Wheeler, Q.D., Knapp, S., Stevenson, D.W., Stevenson, J., Blum, S.D., Boom, B.M., Borisy,  
568 G.G., Buizer, J.L., De Carvalho, M.R., Cibrian, A., Donoghue, M.J., Doyle, V., Gerson,  
569 E.M., Graham, C.H., Graves, P., Graves, S.J., Guralnick, R.P., Hamilton, A.L., Hanken, J.,  
570 Law, W., Lipscomb, D.L., Lovejoy, T.E., Miller, H., Miller, J.S., Naeem, S., Novacek,  
571 M.J., Page, L.M., Platnick, N.I., Porter-Morgan, H., Raven, P.H., Solis, M.A., Valdecasas,  
572 A.G., Van Der Leeuw, S., Vasco, A., Vermeulen, N., Vogel, J., Walls, R.L., Wilson, E.O.  
573 & Woolley, J.B. (2012) Mapping the biosphere: exploring species to understand the origin,  
574 organization and sustainability of biodiversity. *Systematics and Biodiversity*, 10, 1–20.
- 575 Yang, Z. & Rannala, B. (2006) Bayesian estimation of species divergence times under a  
576 molecular clock using multiple fossil calibrations with soft bounds. *Molecular Biology and*  
577 *Evolution*, 23, 212–226.
- 578 Zhang J., Kapli P., Pavlidis P. & Stamatakis A. (2013) A general species delimitation method  
579 with applications to phylogenetic placements. *Bioinformatics*, 29, 2869–2876.

580 Zhang, X.-B., Chen, X.-X. & Cheng, J.-A. (2006) Lucid Phoenix: A tool for building and  
581        deploying interactive, multimedia keys through internet. *Entomotaxonomia*, 28, 231–234.  
582

583 **TABLE 1.** Overview of the software tools and functionalities currently included in the 0.1  
 584 version of iTaxoTools. The majority of the tools can be run (i) command-line driven in Python,  
 585 and is distributed as (ii) a standalone executable (.exe) file, (iii) as part of a full package with  
 586 launcher window (Fig. 1) in a single executable or part of a folder, and (iv) all tools will  
 587 furthermore be implemented on a webserver. Note that functionalities for pre-existing species  
 588 delimitation tools are explained in more detail in the original papers.  
 589  
 590

Tool	Purpose	Main functionalities
dnaconvert	Converts among DNA sequence formats	<ul style="list-style-type: none"> <li>- Supports typical sequence formats (fasta, fastq, phylip, nexus)</li> <li>- Autocorrects typical errors in sequence files such as non-standard characters in sequence names.</li> <li>- Reads GenBank flat files and converts from and to tab-delimited files to manage sequence in spreadsheet editors.</li> <li>- Single-file conversion, batch conversion and conversion of copy-pasted files</li> </ul>
latlonconverter	Converts among different geographic coordinate formats	<ul style="list-style-type: none"> <li>- Parses a large variety of formats of WGS84 geographical coordinates</li> <li>- Batch-conversion of coordinates in tables or copy-pasted lists which can contain coordinates in different formats (recognized by heuristic approach)</li> <li>- Main output in decimal degree format</li> </ul>
fastmerge	Merges DNA sequence files (fasta, fastq)	<ul style="list-style-type: none"> <li>- Can merge large /definition?/ files that usually cannot be opened in editors.</li> <li>- Works for any text file but includes additional features when processing fasta and fastq.</li> <li>- Allows for filtering sequences and sequence names with certain motifs and include/exclude them in the merged file</li> </ul>
fastsplit	Splits (large) DNA sequence files (fasta, fastq) into smaller files	<ul style="list-style-type: none"> <li>- Can split large files /definition?/ that usually cannot be opened in editors and split them into a series of equally sized smaller files</li> <li>- Designed for fasta and fastq, but works for any text file.</li> <li>- Allows for filtering sequences and sequence names with certain motifs and include/exclude them in the split files</li> </ul>
specimentablepruner	Removes rows from tables based on a list of values for the row "specimen"	<ul style="list-style-type: none"> <li>- Takes as input a tab-delimited file and a series of values of specimen identifiers</li> <li>- Removes all rows from the table where the column "specimen" (or other chosen column) agrees with any of the provided values</li> </ul>
specimentablemerger	Merges data from two tables based on values in the row "specimen"	<ul style="list-style-type: none"> <li>- Takes as input two or more tab-delimited files, compares values in column "specimen" (or other chosen column) and merges into one table, combining values for same specimen number in the same row</li> <li>- Automatically checks for duplicate values of the same variable and specimen and issues warnings</li> </ul>
linebreaker	Changes among line break formats (Unix, Windows, Mac)	<ul style="list-style-type: none"> <li>- Takes as input any text file and changes all line breaks to the specified format</li> </ul>
simplestatscalculator	Performs a series of basic statistical analyses based manually entered data	<ul style="list-style-type: none"> <li>- Values are typed or pasted into text boxes</li> <li>- Descriptive statistics (mean, median, standard deviation and others)</li> <li>- Pairwise comparisons (t-test, U-test)</li> <li>- Comparisons of distributions (Chi-square, normality, Fisher's)</li> <li>- Corrections for multiple testing</li> </ul>
unitconverter	Converts among different units	<ul style="list-style-type: none"> <li>- Values are typed into into one field, all other fields show converted values in real time</li> <li>- Separate tabs for conversion of distance, volume, time, molarity, and others</li> </ul>
spartmapper	Computes a kml file from geographical coordinates and spart file	<ul style="list-style-type: none"> <li>- Takes as input a text file with decimal geographical coordinates and specimen identifiers, and a species partition (spart) file</li> <li>- Outputs a kml file to show localities by species on Google Earth</li> </ul>
nodenamecorrector	Removes special characters from terminal node names in Newick-formatted trees	<ul style="list-style-type: none"> <li>- Takes as input a Newick treefile, identifies node names, searches for characters in node names that are not standard alphanumeric, and replaces them with underscores</li> </ul>
pyr8s	Calculates ultrametric timetrees	<ul style="list-style-type: none"> <li>- Takes as input treefiles with branch length (phylograms)</li> </ul>

	(chronograms) based on non-parametric rate-smoothing	<ul style="list-style-type: none"> <li>- Transforms into ultrametric using non-parametric rate smoothing, without the need to access original data (sequences)</li> <li>- User-friendly interface to set time constraints (calibrations) on nodes.</li> </ul>
TaxI2	Calculates inter- and intraspecific distances and the barcoding gap based on pairwise-aligning DNA sequences	<ul style="list-style-type: none"> <li>- Takes as input aligned or unaligned sequence files in fasta or tab-delimited text format</li> <li>- For unaligned sequences, pairwise alignment are performed</li> <li>- Calculates pairwise genetic distances among all sequences</li> <li>- If tab file contains row with species names, inter- and intra-species distances are calculated and summarized, and the barcoding gap as well as some summary statistics of the barcoding gap calculated</li> <li>- inter-species distances are separately calculated for species of the same genus vs. different genera</li> <li>- A histogram with illustrating the barcoding gap is produced in editable PDF format</li> </ul>
morphometricanalyzer	Calculates a series of basic statistical comparisons of species based on morphometric data	<ul style="list-style-type: none"> <li>- Takes as input tab-delimited files with morphometric measurements (continuous variables)</li> <li>- Allows specifying if analyses should be done by species, by sex/stage, or by species and sex/stage</li> <li>- Calculates summary statistics, pairwise comparisons (t-tests, U-tests), ANOVAs, PCA and DA</li> <li>- Size-corrects values by calculating ratio against a reference measurement such as body size</li> <li>- Outputs boxplots and scatterplots of PCA and DA factors, by species and/or sex/stage in editable PDF format</li> <li>- Writes text output summarizing diagnostic characters (scientifically different measurements between species, with and without overlap of ranges)</li> </ul>
dnadiagnoser	Computes diagnostic sites for species from DNA sequences	<ul style="list-style-type: none"> <li>- Takes as input aligned or unaligned sequence files in fasta or tab-delimited text format</li> <li>- Unaligned sequences are pairwise aligned to reference sequence and differences recorded relative to position in reference</li> <li>- Summarizes variation within species and outputs diagnostic sites among species</li> <li>- Outputs unique diagnostic sites for the whole data sets, as well as diagnostic sites in pairwise comparisons among species</li> <li>- Output is given in the form of tables but also as text which can be used for species diagnoses in taxonomic papers</li> </ul>
PTP	Species delimitation based on Poisson tree processes	<ul style="list-style-type: none"> <li>- Uses as input a non-ultrametric tree with branch lengths (phylogram) tree in Newick or Nexus format</li> <li>- Models speciation on branching events in terms of number of mutations (inferred from branch lengths)</li> <li>- Bayesian and ML versions of PTP are implemented</li> </ul>
GMYC	Species delimitation based on the Generalized Mixed Yule Coalescent	<ul style="list-style-type: none"> <li>- Uses as input an ultrametric tree in Newick or Nexus format</li> <li>- Uses a likelihood approach to analyse the timing of branching events, seeking for significant switches between a Yule (interspecific) and a coalescent (intraspecific) branching structure.</li> </ul>
tr2	Species delimitation using Bayesian model comparison and rooted triplets	<ul style="list-style-type: none"> <li>- Takes as input a set of gene trees, and optionally a guide species tree</li> <li>- Calculates posterior probability scores for user-specified delimitation hypotheses.</li> <li>- Alternatively, finds the best delimitation under a guide tree specifying a hierarchical structure of species grouping.</li> </ul>
DELINEATE	Species delimitation by integrating an explicit model of speciation into the multipopulation coalescent	<ul style="list-style-type: none"> <li>- Takes as input a rooted ultrametric tree from a multispecies coalescent analysis, in Nexus or Newick format</li> <li>- Second input file is a table assigning specimens to species, or flagging their species identity as unknown</li> <li>- Outputs various alternative species partitions, ranked by probability</li> </ul>
ABGD	Species delimitation by automatic barcoding gap discovery	<ul style="list-style-type: none"> <li>- Takes as input a set of aligned sequences and calculates pairwise distances</li> <li>- Uses a coalescent model to identify the position of the most likely barcode gap, based on a maximal genetic intraspecific divergence defined a priori by the user.</li> <li>- Uses the DNA barcoding gap to propose species partitions.</li> </ul>



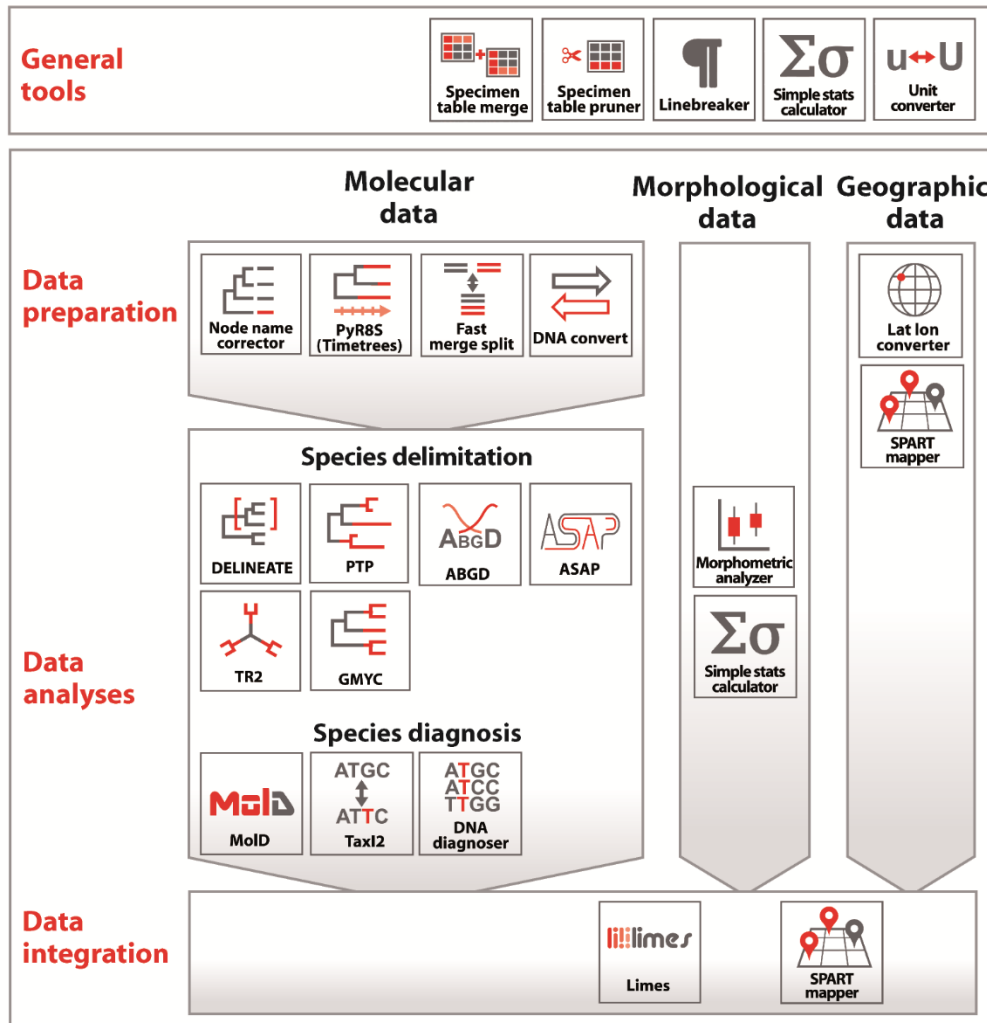
ASAP	Species delimitation from single-locus sequence data by the Assemble Species by Automatic Partitioning approach	<ul style="list-style-type: none"><li>- Takes as input a set of aligned sequences and calculates pairwise distances</li><li>- Proposes species partitions ranked by a new scoring system that uses no biological prior insight of intraspecific diversity.</li></ul>
LIMES 2.0	Compare species partitions by different indexes and parsing/merge/export spart files	<ul style="list-style-type: none"><li>- Reads species partition (spart) files, as well as species partition information in spreadsheet format</li><li>- Computes <math>C_{tax}</math>, <math>mC_{tax}</math>, <math>R_{tax}</math> and Match Ratio indexes</li><li>- Can merge, extract and export spart files</li></ul>
MolD	Recovers DNA-based diagnoses for taxa from DNA sequence alignments	<ul style="list-style-type: none"><li>- recovers diagnostic combinations of nucleotides (DNCs) for pre-defined groups of DNA sequences, corresponding to taxa</li><li>- Identifies pure diagnostic sites, minimal DNCs (mDNCs), and redundant DNCs (rDNCs), the latter fulfil predefined criteria of reliability</li></ul>

591  
592  
593  
594

595 **TABLE 2.** Repositories of the code of the tools included in the 0.1 version of iTaxoTools. The  
 596 table also lists the main programmers involved in the development of each tool or its graphical  
 597 user interface (GUI), and informs whether a tool was newly programmed for this project, adjusted  
 598 from existing code (by adding a GUI plus sometimes additional functionalities), or included as  
 599 original code and GUI without modification.  
 600

Tool	New / Adjusted / Original	Github repository (original / modified)	Main programmers (original program) / GUI
dnaconvert	New	<a href="https://github.com/iTaxoTools/DNAconvert">https://github.com/iTaxoTools/DNAconvert</a>	V. Kharchev
latlonconverter	New	<a href="https://github.com/iTaxoTools/latlon-converter">https://github.com/iTaxoTools/latlon-converter</a>	V. Kharchev
fastmerge	New	<a href="https://github.com/iTaxoTools/fastsplit-merge">https://github.com/iTaxoTools/fastsplit-merge</a>	V. Kharchev
fastsplit	New	<a href="https://github.com/iTaxoTools/fastsplit-merge">https://github.com/iTaxoTools/fastsplit-merge</a>	V. Kharchev
specimentablepruner	New	<a href="https://github.com/iTaxoTools/specimentablepruner">https://github.com/iTaxoTools/specimentablepruner</a>	V. Kharchev
specimentablemerger	New	<a href="https://github.com/iTaxoTools/specimentablemerger">https://github.com/iTaxoTools/specimentablemerger</a>	V. Kharchev
linebreaker	New	<a href="https://github.com/iTaxoTools/linebreaker">https://github.com/iTaxoTools/linebreaker</a>	S. Kumari
simplestatscalculator	New	<a href="https://github.com/iTaxoTools/simple_stat">https://github.com/iTaxoTools/simple_stat</a>	S. Kumari
unitconverter	New	<a href="https://github.com/iTaxoTools/UnitConverter">https://github.com/iTaxoTools/UnitConverter</a>	S. Kumari
spartmapper	New	<a href="https://github.com/iTaxoTools/linebreak_replacer">https://github.com/iTaxoTools/linebreak_replacer</a>	S. Kumari
nodenamecorrector	New	<a href="https://github.com/iTaxoTools/nodenamecorrector">https://github.com/iTaxoTools/nodenamecorrector</a>	V. Kharchev
pyr8s	New	<a href="https://github.com/iTaxoTools/pyr8s">https://github.com/iTaxoTools/pyr8s</a>	S. Patmanidis
TaxI2	New	<a href="https://github.com/iTaxoTools/TaxI2">https://github.com/iTaxoTools/TaxI2</a>	V. Kharchev
morphometricanalyzer	New	<a href="https://github.com/iTaxoTools/morphometricanalyzer">https://github.com/iTaxoTools/morphometricanalyzer</a>	V. Kharchev
dnadiagnoser	New	<a href="https://github.com/iTaxoTools/dnadiagnoser">https://github.com/iTaxoTools/dnadiagnoser</a>	V. Kharchev
PTP	Adjusted	<a href="https://github.com/zhangjiajie/PTP">https://github.com/zhangjiajie/PTP</a> <a href="https://github.com/iTaxoTools/PTP-PYQT5">https://github.com/iTaxoTools/PTP-PYQT5</a>	(J. Zhang) GUI: S. Kumari
GMYC	Adjusted	<a href="https://github.com/zhangjiajie/pGMYC">https://github.com/zhangjiajie/pGMYC</a> <a href="https://github.com/iTaxoTools/GMYC-PYQT5">https://github.com/iTaxoTools/GMYC-PYQT5</a>	(J. Zhang ) GUI: S. Kumari
tr2	Adjusted	<a href="https://github.com/TFujisawa/tr2-delimitation-git">https://github.com/TFujisawa/tr2-delimitation-git</a> <a href="https://github.com/iTaxoTools/pyqt5-tr2">https://github.com/iTaxoTools/pyqt5-tr2</a>	(T. Fujisawa) GUI: S. Kumari
DELINEATE	Adjusted	<a href="https://github.com/iTaxoTools/pyqt5-delineate">https://github.com/iTaxoTools/pyqt5-delineate</a>	(J. Sukumaran) GUI: S. Kumari
ABGD	Adjusted	<a href="https://bioinfo.mnhn.fr/abi/public/abgd/">https://bioinfo.mnhn.fr/abi/public/abgd/</a> <a href="https://github.com/iTaxoTools/ABGDpy">https://github.com/iTaxoTools/ABGDpy</a>	(S. Brouillet) GUI: S. Patmanidis
ASAP	Adjusted	<a href="https://bioinfo.mnhn.fr/abi/public/asap/">https://bioinfo.mnhn.fr/abi/public/asap/</a> <a href="https://github.com/iTaxoTools/ASAPy">https://github.com/iTaxoTools/ASAPy</a>	(S. Brouillet) GUI: S. Patmanidis
LIMES 2.0	Original	<a href="https://github.com/iTaxoTools/LIMES">https://github.com/iTaxoTools/LIMES</a>	J. Ducasse
MolD	Adjusted	<a href="https://github.com/SashaFedosov/MolD">https://github.com/SashaFedosov/MolD</a> <a href="https://github.com/iTaxoTools/MolD_pyqt5">https://github.com/iTaxoTools/MolD_pyqt5</a>	(A. Fedosov) GUI: S. Kumari

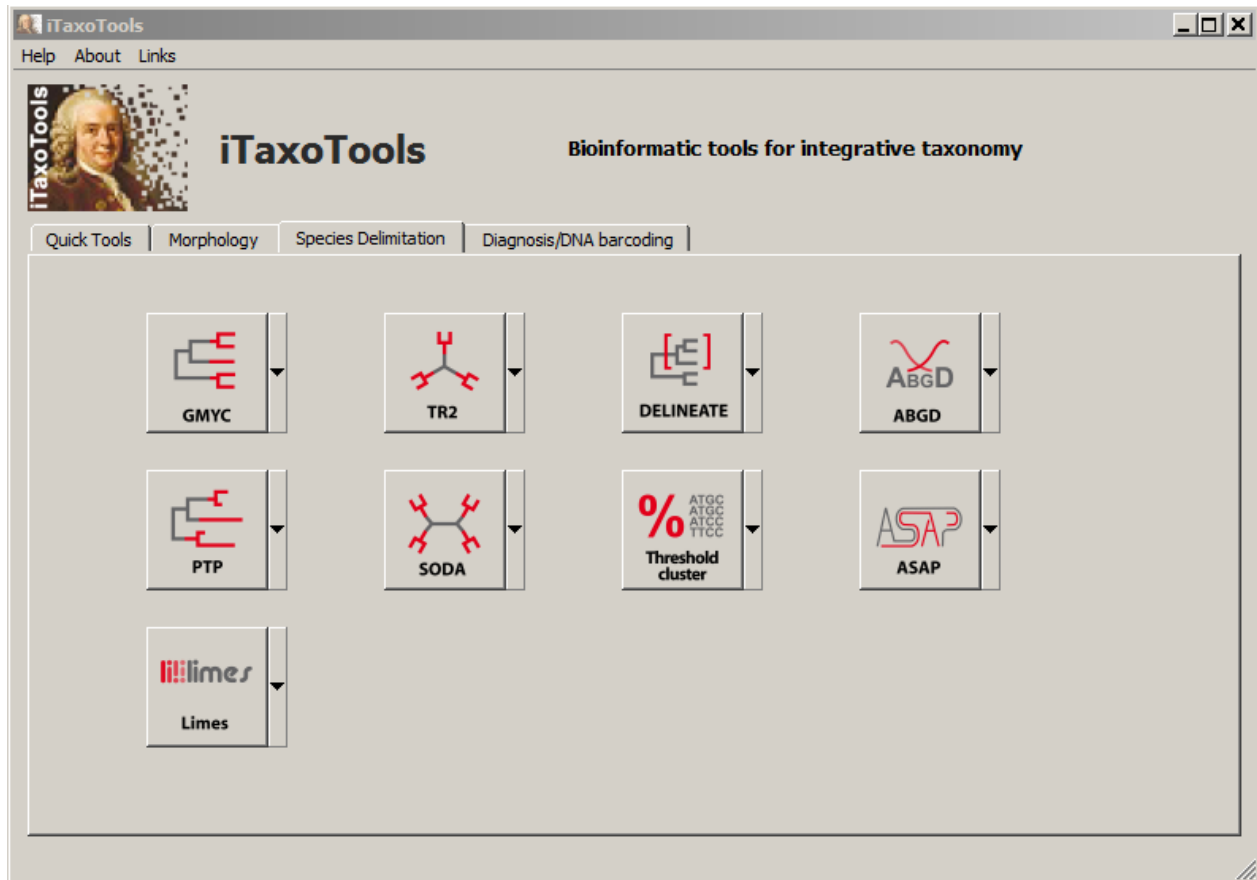
601  
 602  
 603  
 604  
 605  
 606



607  
608  
609  
610  
611  
612  
613  
614  
615  
616

**FIGURE 1.** Overview of the various tools implemented in iTaxoTools, and their scope. In the present version a focus is on molecular data analysis, but more functionalities to analyze and visualize morphological and geographic data will be implemented in the next future, while data integration remains the main focus for long-term implementation.

617



618

619

620

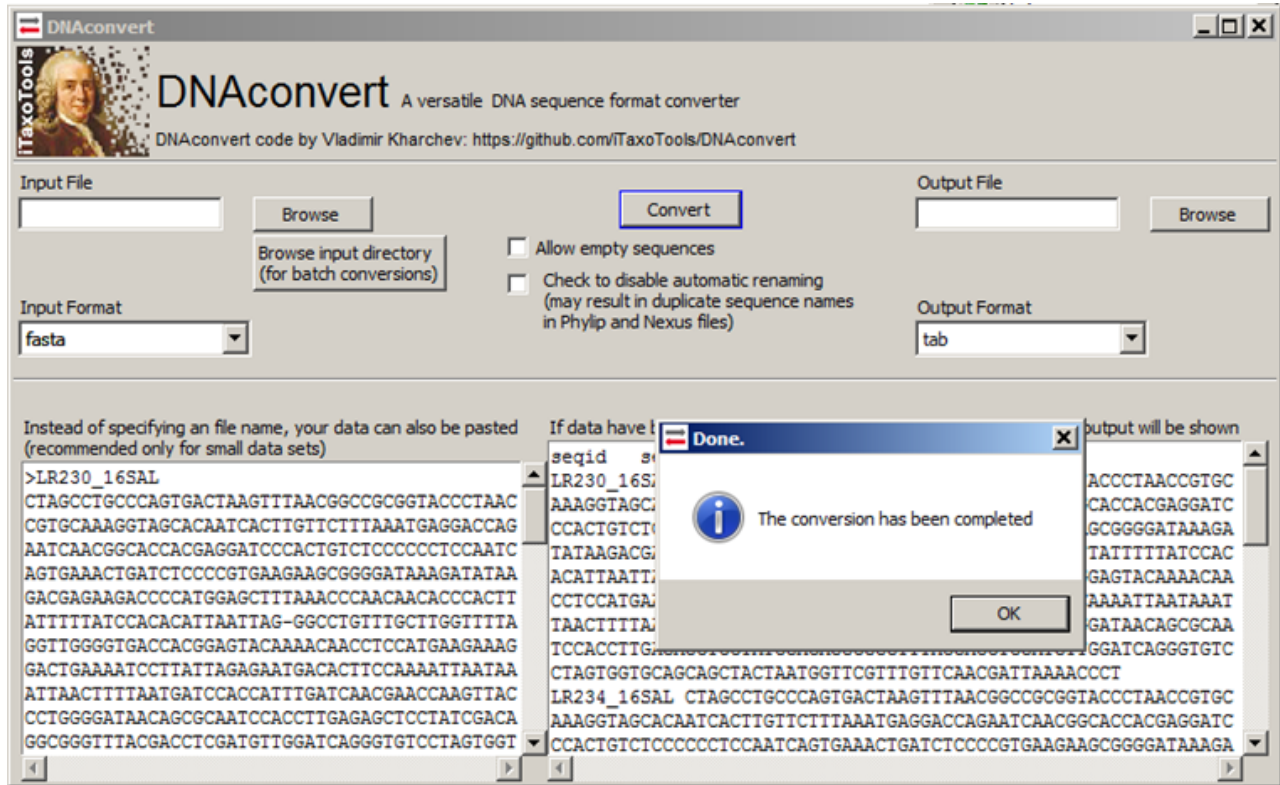
621

622

623

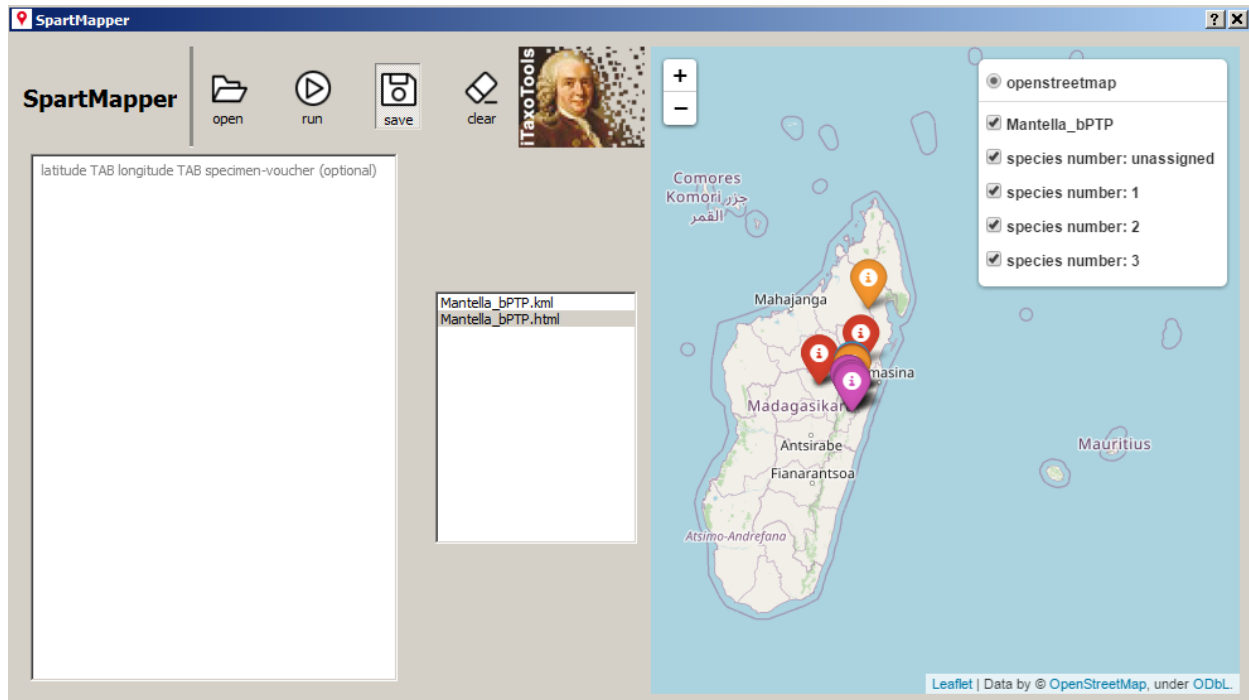
**FIGURE 2.** Main launcher window of iTaxoTools 0.1 with the option to start various species delimitation tools (additional tools can be started from the other tabs).

624  
625  
626



627  
628  
629  
630  
631  
632  
633  
634  
635  
636

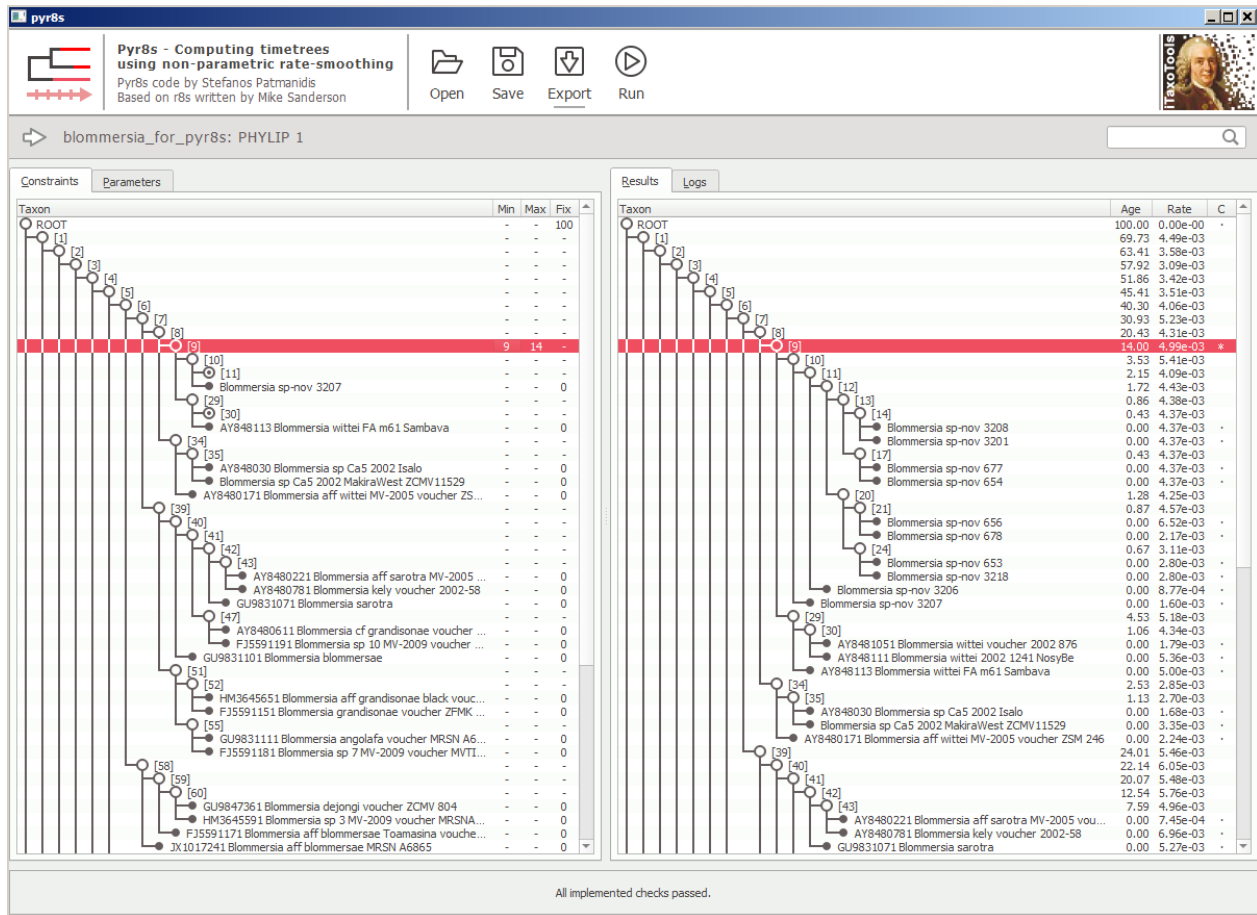
**FIGURE 3.** Screenshot of one of the newly programmed quick conversion tools, DNAconvert, which implements numerous autocorrect options to avoid sequence output files generating errors in downstream programs. DNAconvert also supports tab-delimited table input and its conversion to common sequence formats such as FASTA, NEXUS, or PHYLIP, to facilitate storage and management of sequences and sequence metadata in spreadsheet editors such as Microsoft Excel.



637  
638  
639  
640  
641  
642  
643  
644

**FIGURE 4.** Screenshot of spartmapper, a tool that plots distribution records from geographical coordinates on a map and categorizes the records based on a species hypothesis provided as SPART file (Miralles et al. 2021). The program allows live view and produces a kml file to visualize the records in Google Earth or Google Maps.

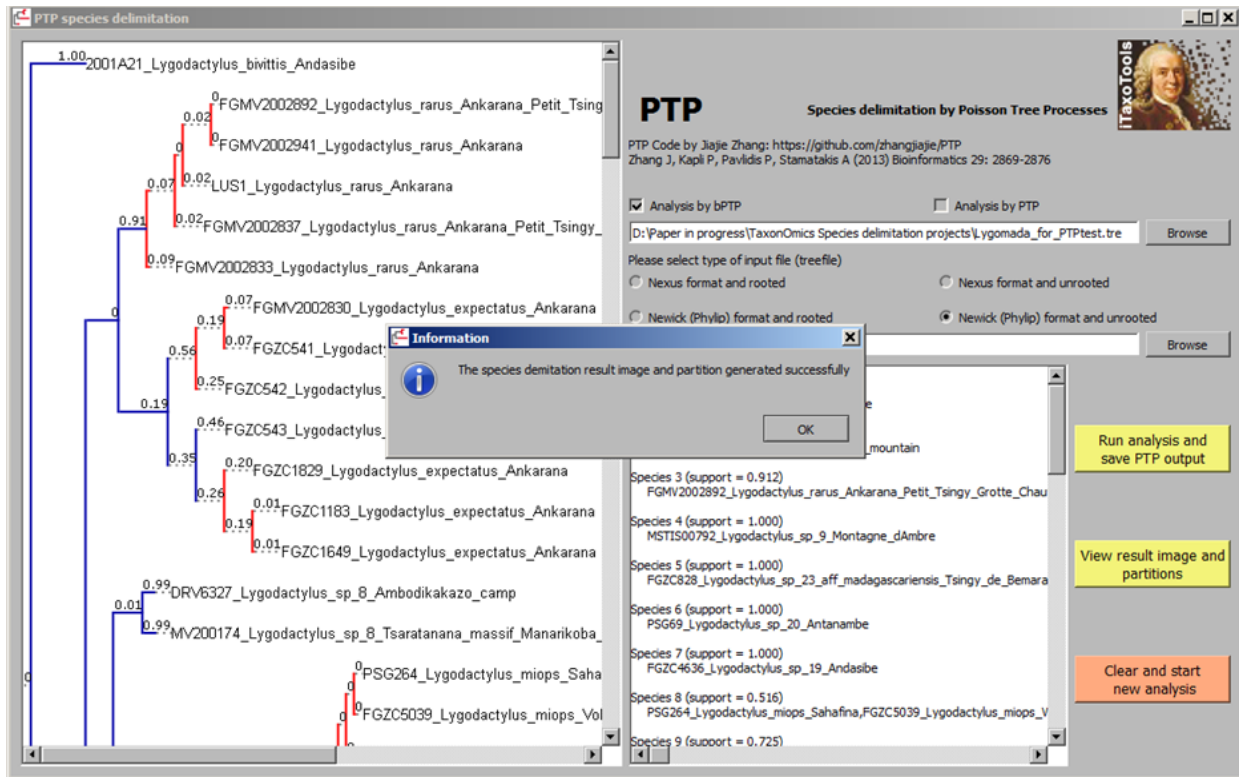
645



646  
647  
648  
649  
650  
651  
652  
653

**FIGURE 5.** Screenshot of pyr8s after running an analysis and converting a tree into an ultrametric timetree. The red-highlighted row marks a node for which age constraints had been set before the analysis.

654  
655  
656

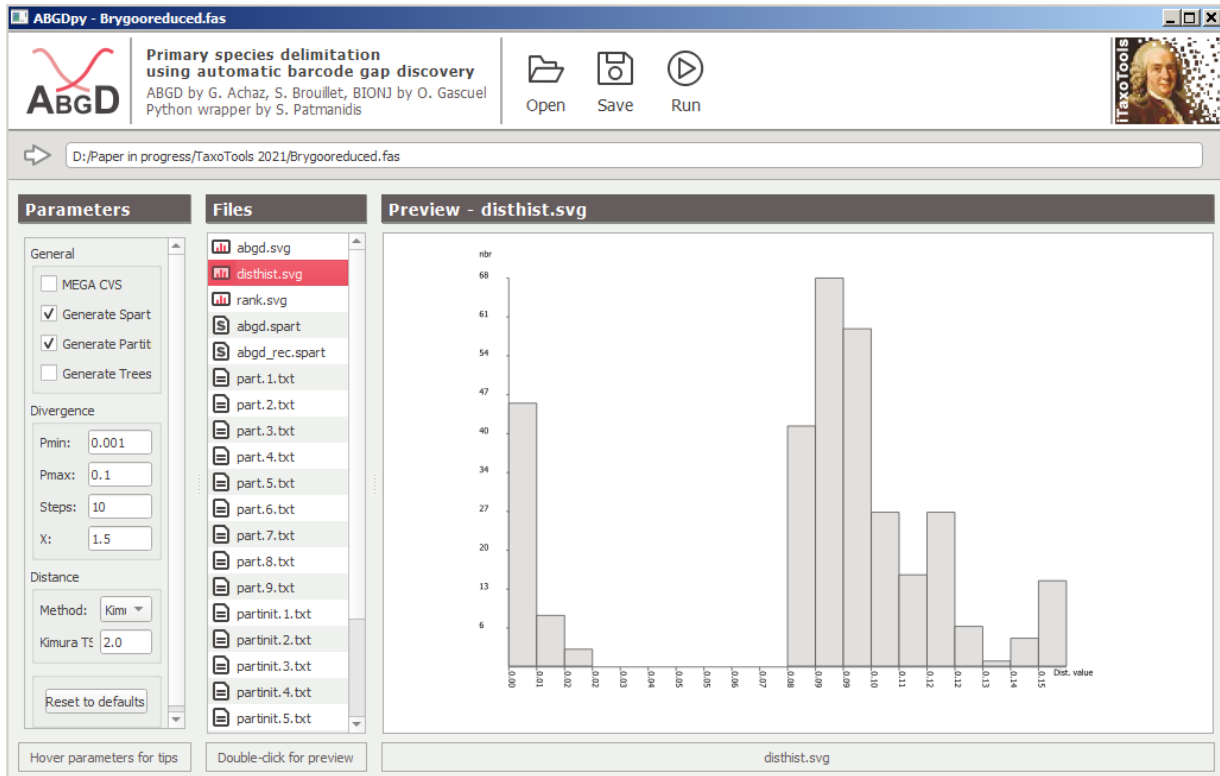


657  
658  
659  
660  
661  
662  
663  
664  
665

**FIGURE 6.** Screenshot of the GUI-based version of PTP, a program that delimits species from non-ultrametric trees. The original Python code of PTP was written by Zhang et al. (2013); iTaxoTools adds the GUI, as well as functionality to export species partition in the SPART format (Miralles et al. 2021).



666



667

668

669

670

671

672

673

674

675

676

**FIGURE 7.** Screenshot of the GUI-based version of ABGD, a program that delimits species by detecting the barcoding gap from pairwise single-locus sequence distances (Puillandre et al. 2012). For this tool, the original ABGD code written in C was wrapped with a Python GUI and compiled as standalone executable. The different output files produced by ABGD (text and graphs) can be selected and pre-viewed within the GUI.