

1 *Journal : Molecular Ecology Resources*

2 *Subject area: Computer Programs*

3 **SPART, a versatile and standardized data exchange format**

4 **for species partition information**

5

6 Aurélien Miralles^{1*}, Jacques Ducasse², Sophie Brouillet¹, Tomas Flouri³, Tomochika

7 Fujisawa⁴, Paschalia Kapli³, L. Lacey Knowles⁵, Sangeeta Kumari⁶, Alexandros

8 Stamatakis^{7,8}, Jeet Sukumaran⁹, Sarah Lutteropp⁷, Miguel Vences^{6*}, Nicolas Puillandre^{1*}

9

10 ¹Institut de Systématique, Évolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS,
11 Sorbonne Université, EPHE, Université des Antilles 57 rue Cuvier, CP 50, 75005 Paris, France

12 ²49 rue Eugène Carrière, 75018 Paris, France

13 ³Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Environment, University
14 College London, London WC1E 6BT, UK

15 ⁴Center for Data Science and Education and Research, Shiga University, 1-1-1 Banba, Hikone, 522-8522, Shiga,
16 Japan

17 ⁵Department of Ecology and Evolution, University of Michigan, Ann Arbor, MI 48109, USA

18 ⁶Braunschweig University of Technology, Zoological Institute, Mendelssohnstraße 4, 38106 Braunschweig,
19 Germany

20 ⁷Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies, Schloss-
21 Wolfsbrunnenweg 35, 69118 Heidelberg Germany

22 ⁸Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Am Fasanengarten 5, 76131 Karlsruhe,
23 Germany

24 ⁹Biology Department, LS 262, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182-4614,
25 USA

26

27 ***Corresponding authors**

28 **Contact:** miralles.skink@gmail.com, m.vences@tu-braunschweig.de, nicolaspuillandre@gmail.com

29

30 **Short running title:** SPART, a standardized format for species partition

31 Abstract

32

33 A wide range of data types can be used to delimit species and various computer-based tools
 34 dedicated to this task are now available. Although these formalized approaches have
 35 significantly contributed to increase the objectivity of SD under different assumptions, they
 36 are not routinely used by alpha-taxonomists. One obvious shortcoming is the lack of
 37 interoperability among the various independently developed SD programs. Given the frequent
 38 incongruences between species partitions inferred by different SD approaches, researchers
 39 applying these methods often seek to compare these alternative species partitions to evaluate
 40 the robustness of the species boundaries. This procedure is excessively time consuming at
 41 present, and the lack of a standard format for species partitions is a major obstacle. Here we
 42 propose a standardized format, SPART, to enable compatibility between different SD tools
 43 exporting or importing partitions. This format reports the partitions and describes, for each of
 44 them, the assignment of individuals to the “inferred species”. The syntax also allows to
 45 optionally report support values, as well as original trees and the full command lines used in
 46 the respective SD analyses. Two variants of this format are proposed, overall using the same
 47 terminology but presenting the data either optimized for human readability (matricial SPART)
 48 or in a format in which each partition forms a separate block (SPART.XML). ABGD,
 49 DELINEATE, GMYC, PTP and TR2 have already been adapted to output SPART files and a
 50 new version of LINES has been developed to import, export, merge and split them.

51

52 **Key words.** Species delimitation programs; SPART; Species partition format; Integrative
 53 taxonomy ; LINES v2.0

54

55 **Introduction**

56

57 Species delimitation (SD) is a burgeoning, fully fledged research field in systematic biology
 58 (Sites & Marshall 2003; Camargo & Sites 2013; Flot 2015, Ducasse et al. 2020). SD benefits
 59 from the interpretation of species as independent evolutionary lineages (De Queiroz 1998,
 60 2007) that can be distinguished from each other using a variety of operational SD criteria
 61 (Samadi & Barberousse 2006). In integrative taxonomy (Dayrat 2005; Padial et al. 2010),
 62 various lines of evidence and a wide range of data types can be used in formalised analytical
 63 workflows to propose species hypotheses, from DNA barcodes to phylogenomic data, discrete
 64 morphological characters, morphometric measurements, ecological traits, geographic
 65 occurrence, bioacoustic signals, metabolomic profiles, and others (Miralles et al. 2020).

66

67 If many, and among them the earliest, formalised SD procedures are mostly carried out
 68 manually, e.g. by comparing trees with the geographic occurrence of individuals, calculating
 69 correlations between geographic and genetic distances, assessing steepness of hybrid zones,
 70 or seeking for correlation between genetic distance and morphological characters (Good &
 71 Wake 1992, Wiens & Penkrot 2002, Vieites et al. 2009, Flot et al. 2010, Weisrock et al. 2010,
 72 Puillandre et al. 2012a, Miralles & Vences 2013, Derkarabetian & Hedin 2014, Dufresnes et
 73 al. 2015), a substantial number of computer-based tools has been developed to delimit
 74 species, often based on statistical criteria. These programs can analyse large datasets, with a
 75 strong focus on the use of sequence data (Table 1). These methods have significantly
 76 contributed to increase the objectivity, repeatability, and speed of species delimitation
 77 inferences under different mathematical models and assumptions (e.g. Multispecies
 78 coalescent model, DNA barcode gap, haplotype fields of recombination, cf. de Queiroz 1998,

79 2007, Knowles & Carstens 2007, Yang & Rannala 2010, Carstens et al. 2013, Leavitt et al.
80 2015, Rannala 2015).

81

82 Although the number and importance of SD tools is likely to sharply increase in the
83 immediate future, they are not yet routinely used in the majority of alpha-taxonomic studies
84 that result in the naming of over 15,000 new species of organisms every year (Miralles et al.
85 2020). One obvious shortcoming is the lack of interoperability among the various
86 independently developed SD programs, and the lack of comprehensive software suites that
87 offer various user-friendly features, such as those for data visualization and comparison of
88 results across methods. For instance, incongruent species partitions resulting from different
89 SD approaches applied to a given dataset are common. They can even be significant, if not
90 striking in some cases (such as excessive splitting or lumping leading to highly different
91 number of species delimited; Carstens et al. 2013, Miralles & Vences 2013, Dellicour & Flot
92 2015, Kapli et al. 2016, Postaire et al. 2016, Renner et al. 2017 for empirical cases; and
93 Sukumaran & Knowles 2017, Chan et al 2020, Luo et al. 2018, Mason et al. 2020 and Zhang
94 et al. 2011 for more methodological studies on SD limitations). Integrative taxonomists will
95 seek to compare these alternative species partitions across SD approaches (but see Rannala
96 2015), and eventually estimate their robustness by integrating other data sources
97 (morphological variation, geographic distribution, etc), in order to make an informed choice—
98 a procedure that is excessively time consuming at present, given the lack of a standard format
99 for species partitions.

100

101 The main output of species delimitation, and therefore of any SD program, is a species
 102 partition. The term “partition” here follows the set theory concept: the organization of a set of
 103 *elements* into mutually-exclusive and jointly-comprehensive *subsets*, not including the empty
 104 subset (Hrbacek & Jech 1999). In an SD application, the *elements are individuals* (i.e.
 105 samples or specimens), and a specific species delimitation hypothesis is a particular
 106 assignment (i.e. a *partition*) of these individuals to subsets, where *each subset corresponds to*
 107 *a distinct inferred species*. Categories resulting from an SD analysis have been referred to by
 108 various terms, such as primary species hypothesis, operational taxonomic unit (OTUs),
 109 barcode index number (BINs; Ratnasingham & Hebert 2013), or even cluster (without any
 110 particular status (Fig. 1)), but all of them match the aforementioned definition of a subset.
 111 Furthermore, while some tools produce *de novo* species partitions (i.e. directly aggregating
 112 individuals into species hypotheses; exploratory methods), others statistically compare
 113 competing species hypotheses that have been defined *a priori* (hypothesis-testing methods),
 114 and these programs require a species partition as input. SD methods may also assign scores,
 115 either to the entire inferred partition (e.g., ASAP-score in the program ASAP; Puillandre et al.
 116 2021), to the distinctiveness of each subset from the others (e.g., posterior probabilities in the
 117 programs BPP and bPTP; Yang & Rannala 2010; Zhang et al. 2013), or to the presence of
 118 each individual in a given subset (e.g., probability of placement in calculation of BINs,
 119 Ratnasingham & Hebert 2013).

120

121 **A standardized Species PARTition format (SPART)**

122

123 Typically, each SD program exports the resulting species partitions in its own idiosyncratic
 124 format. Some, for instance, provide a table of assignments of individual specimens to the

125 subsets (e.g. GMYC) while others, conversely, list the different subsets with the included
 126 individuals (e.g. ABGD, PTP), whereas again others graphically report subsets on a tree
 127 topology (e.g. GMYC). These different formats may or may not include complementary data
 128 (e.g., scores, topologies, metadata, number of species delimited, etc.), and are not designed to
 129 be parsed by other tools for downstream analyses. Their manual conversion into a versatile
 130 and easily reusable plain text species partition (e.g., CSV) is not always straightforward. It
 131 can be particularly error prone and time consuming with large datasets, as species
 132 delimitations on several hundreds, or even thousands, of specimens are becoming common
 133 practice in molecular taxonomy (e.g., Ahrens et al. 2016, Renner et al 2017, Garcià-Melo et
 134 al. 2019, Hoffmann et al. 2019, Solihah et al. 2020, Christodoulou et al. 2020).

135

136 We here propose a standardized species partition format, SPART, to enable compatibility
 137 between different tools producing (export) or using (import) species partitions. Our format
 138 facilitates:

139 (1) statistical comparison of different alternative species partitions such as their overall
 140 congruence, similarity or resolving power, identification of the subsets that are congruently
 141 delimited (currently implemented in the program LIMES v2.0; Ducasse et al. 2020);

142 (2) assessment of multiple competing SD hypotheses, including those used as input in e.g.
 143 DELINEATE and BPP to evaluate them (Sukumaran et al. 2020, Yang & Rannala 2010);

144 (3) visualization and comparison of species partitions (e.g., DNA-based species partitions
 145 compared with manually-edited species partitions obtained from alternative methods and data

such as Principal Component Analysis of morphometry, haplotype networks, geographic distribution, habitat type, external phenetic similarity, or simply, current taxonomy);

(4) extraction, from original data files, of specific data for each subset under different species partition assumptions (e.g. lists of molecular and morphological diagnostic character states, descriptive statistics characterizing each of the inferred species, or ecological or distributional traits); and

(5) potential taxonomic reassignment of specimens in databases.

More generally, the SPART format is designed to be versatile and fully integrative in the sense that it can include any species partition descriptors, independently of the method or data-type used to generate the species partition (Fig. 2). SPART does not convey any interpretation on the quality of the species partition, nor on the pros and cons of the methods used to define them, but is simply a common format that seeks at homogenising the way species partitions are recorded. It can therefore be implemented in any method used to generate one or several species partitions as output. Likewise, any method using (analysing, comparing, automatically reassigning or graphically representing) multiple subsets of specimens might benefit from being able to import SPART files as input data.

Matricial and serial implementation of the SPART format

SPART files include information on one or multiple species partitions for a given set of elements (i.e. individuals) and use standardized terminology to denote the number of species partitions included in the file (“N_partitions”) and for each partition, the number of individuals (“N_individuals”), number of subsets (“N_subsets”), and the assignment of individuals to subsets (“Assignment”) (Supporting information 1). The syntax also allows to optionally include support values for species partitions, subsets, and the assignment of individuals to subsets, as well as original trees and the full command line used in the respective SD analyses, the program version number as well as comments and species partition comparison indices as calculated with LIMES 2.0, a new version of LIMES (Ducasse et al. 2020) recently published.

To account for the diversity of possible future applications, we propose two variants of the SPART format (for details see Supporting information 1). Both of these use largely the same terminology but represent the data differently:

The first SPART variant is optimized for human readability and its syntax has been designed to be compatible with Nexus (a widely used data format in phylogenetic inference software: Maddison et al. 1997). This allows to include SPART specifications as blocks in Nexus files if required by future applications. If information from multiple partitions is included, then it is combined into a single block, presenting the respective assignments and assignment scores per individual from different species partitions concatenated on a single line, separated by separator symbols. This enables easy manual transformation into a spreadsheet format if required. Due to the presentation of information from multiple partitions in one block as a concatenated matrix, we denote this variant as *matricial SPART* format, or simply SPART.

189 The second SPART variant is optimized for machine readability, and relies on XML
190 (eXtensible Markup Language), a lightweight data-interchange format that can be easily
191 parsed and written by software tools, while it can still be read and written by humans as well.
192 When information from multiple partitions is included, each partition forms a separate block
193 containing information on the number of subsets, individual assignments and assignment
194 scores. We therefore denote this variant as *SPART.XML* format.

195

196 **Tools already implementing SPART and future perspectives**

197

198 The proposed format is already implemented in several widely-used SD programs. Both the
199 matricial SPART and SPART.XML output files are already generated by GUI-driven
200 standalone versions (<https://github.com/iTaxoTools>) of ABGD, ASAP, GMYC, PTP, mPTP,
201 TR2 and DELINEATE (Vences et al. submitted), by the native Python version of TR2, and in
202 the web versions of ABGD and ASAP; and in progress for the Python versions of GMYC and
203 PTP. Furthermore, the species partition comparison tool LIMES v2.0 has been expanded to
204 import, export and convert SPART files, in particular to (1) compare, by calculating indices
205 (e.g., *Ctax*, *Ratx*, *Match Ratio*, cf. Ducasse et al. 2020) for species partitions from SPART
206 files (including each one or several species partitions); (2) merge species partitions included
207 in different SPART files into one SPART file, (3) import species partition(s) table(s) from
208 spreadsheet editors such as Microsoft EXCEL and save it (them) into a single SPART file. A
209 new software tool named SPARTMAPPER has also been developed; it takes SPART files as
210 input along with a tab-delimited series of geographical coordinates linked to specimen names,
211 plots the distribution of alternative delimited species on a map, and exports a .kml file to
212 visualize this information in Google Earth.

213 [\[note to reviewers: all of these new software versions will be released before or upon publication of this](#)
 214 [manuscript, and the web based ABGD and ASAP are already freely available since February 2021](#)
 215 <https://bioinfo.mnhn.fr/abi/public/abgd/abgdweb.html>, <https://bioinfo.mnhn.fr/abi/public/asap/>);
 216 [for review, various preliminary Windows executables are available under this link:](#)
 217 [https://hidrive.ionos.com/share/ohaymwcgjd#\\$/Win%20executables](https://hidrive.ionos.com/share/ohaymwcgjd#$/Win%20executables)]

218 In the context of future work, we envisage the development of visualization tools to
 219 automatically illustrate information from species partitions along with support values and
 220 phylogenetic hypotheses (Fig. 1). There is still a long way to go before programs will be able
 221 to infer species based on combining evidence using different data sources such as genetics,
 222 morphology, ecology, behaviour, geographic distribution, etc. However, eventually, reliable
 223 computer-based, species delimitation procedures that mirror the procedures of integrative
 224 taxonomy will be at the core of next generation taxonomy (Vences 2020). Our SPART data
 225 exchange format would thus contribute to this next generation taxonomy, by simplifying
 226 computational approaches to completing the inventory of life on Earth.

227
 228

229 **Acknowledgments**

230

231 We are grateful to Susanne Renner who stimulated this work by leading the priority program
 232 SPP 1991 "Taxon-Omics" of the Deutsche Forschungsgemeinschaft (DFG), specifically in the
 233 context of a grant on taxonomic data integration and management (RE 603/29-1), and to
 234 many members of the Taxon-Omics consortium and Guillaume Achaz (MNHN) for fruitful
 235 discussion. MV and SK were supported by DFG grant VE247/20-1, NP by the European
 236 Research Council (ERC) under the European Union's Horizon 2020 research and innovation
 237 programme (grant agreement No. 865101), and AS and SL by the Klaus-Tschira foundation.

238

239

240

241

242

243

244 REFERENCES

245

246

247 Ahrens, D., Fujisawa, T., Krammer, H. J., Eberle, J., Fabrizi, S., & Vogler, A. P. (2016).

248 Rarity and incomplete sampling in DNA-based species delimitation. *Systematic Biology*,

249 65, 478–494. [doi:10.1093/sysbio/syw002](https://doi.org/10.1093/sysbio/syw002)

250 Camargo, A., & Sites, J. Jr. (2013). Species delimitation: a decade after the renaissance. In:

251 The Species Problem - Ongoing Issues (ed. I. Y. Pavlinov). IntechOpen.

252 Carstens, B. C., Pelletier, T. A., Reid, N. M., & Satler, J. D. (2013). How to fail at species

253 delimitation. *Molecular Ecology*, 22, 4369–4383. [doi:10.1111/mec.12413](https://doi.org/10.1111/mec.12413)

254 Chan, K. O., Hutter, C. R., Wood, P. L. Jr., Grismer, L. L., Das, I., & Brown, R. M. (2020).

255 Gene flow creates a mirage of cryptic species in a Southeast Asian spotted stream frog

256 complex. *Molecular Ecology*, 29(20), 3970–3987. [doi:10.1111/mec.15603](https://doi.org/10.1111/mec.15603)

257 Christodoulou, M., O'Hara, T., Hugall, A. F., Khodami, S., Rodrigues, C. F., Hilario, A.,

258 Vink, A., & Martinez Arbizu, P. (2020) Unexpected high abyssal ophiuroid diversity in

259 polymetallic nodule fields of the northeast Pacific Ocean and implications for

260 conservation, *Biogeosciences*, 17, 1845–1876. [doi:10.5194/bg-17-1845-2020](https://doi.org/10.5194/bg-17-1845-2020)

261 Dayrat, B. (2005). Toward integrative taxonomy. *Biological Journal of the Linnean Society*,

262 85, 407–415. [doi:10.1111/j.1095-8312.2005.00503.x](https://doi.org/10.1111/j.1095-8312.2005.00503.x)

263 Dellicour, S., & Flot J.-F. (2015). Delimiting species-poor data sets using single molecular

264 markers: a study of barcode gaps, haplowebs and GMYC. *Systematic Biology*, 64, 900–

265 908. [doi:10.1093/sysbio/syu130](https://doi.org/10.1093/sysbio/syu130)

266 de Queiroz, K. (1998). The general lineage concept of species, species criteria, and the

267 process of speciation. In: D.J. Howard & S.H. Berlocher, S.H. (Eds.), *Endless Forms:*

268 *Species and Speciation*. (pp. 57–75). New York: Oxford University Press.

269 de Queiroz, K. (2007). Species concepts and species delimitation, *Systematic Biology*, 56,

270 879–886. [doi:10.1080/10635150701701083](https://doi.org/10.1080/10635150701701083)

271 Derkarabetian, S., & Hedin, M. (2014). Integrative taxonomy and species delimitation in

272 harvestmen: a revision of the western North American genus *Sclerobunus* (Opiliones:

273 Laniatores: Travunioidea). *PLoS One*, 9, e104982. [doi:10.1371/journal.pone.0104982](https://doi.org/10.1371/journal.pone.0104982)

274 Ducasse, J., Ung, V., Lecointre, G., Miralles, A. (2020). LIMES : a tool for comparing

275 species partition. *Bioinformatics*, 2282–2283. [doi:10.1093/bioinformatics/btz911](https://doi.org/10.1093/bioinformatics/btz911)

276 Dufresnes, C., Brelsford, A., Crnobrnja-Isailović, J., Tzankov, N., Lymberakis, P., & Perrin,

277 N. (2015). Timeframe of speciation inferred from secondary contact zones in the European

278 tree frog radiation (*Hyla arborea* group). *BMC Evolutionary Biology* 15, 1-8. doi:

279 10.1186/s12862-015-0385-2

280 Ence, D.D., & Carstens, B.C. (2011). SpedeSTEM: A rapid and accurate method for species

281 delimitation. *Molecular Ecology Resources*, 11, 473–480. [doi:10.1111/j.1755-](https://doi.org/10.1111/j.1755-0998.2010.02947.x)

282 [0998.2010.02947.x](https://doi.org/10.1111/j.1755-0998.2010.02947.x)

283 Flot, J.-F., Couloux, A., & Tillier, S. (2010). Haplowebs as a graphical tool for delimiting

284 species: a revival of Doyle's "field for recombination" approach and its application to the

285 coral genus *Pocillopora* in Clipperton. *BMC Evolutionary Biology*, 10, 372.

286 [doi:10.1186/1471-2148-10-372](https://doi.org/10.1186/1471-2148-10-372)

Flot, J.-F. (2015). Species delimitation's coming of age, *Systematic Biology*, 64, 897–899.

Fontaneto, D., Herniou, E., Boschetti, C., Caprioli, M., Melone, G., Ricci, C., & Barraclough, T.G. (2007). Independently evolving species in asexual bdelloid rotifers. *PLoS Biology*, 5, e87. [doi:10.1371/journal.pbio.0050087](https://doi.org/10.1371/journal.pbio.0050087)

Fujisawa, T., Aswad, A., Barraclough, T. G. (2016). A rapid and scalable method for multilocus species delimitation using Bayesian model comparison and rooted triplets. *Systematic Biology*, 65(5), 759–771. [doi:10.1093/sysbio/syw028](https://doi.org/10.1093/sysbio/syw028)

García-Melo, J. E., Oliveira, C., Da Costa Silva, G. J., Ochoa-Orrego, L. E., Garcia Pereira, L. H., Maldonado-Ocampo, J. A. (2019). Species delimitation of Neotropical characins (Stevardiinae): Implications for taxonomy of complex groups. *PLoS ONE* 14(6), e0216786. [doi:10.1371/journal.pone.0216786](https://doi.org/10.1371/journal.pone.0216786)

Good, D. A., & Wake, D. B. (1992). Geographic variation and speciation in the torrent salamanders of the genus *Rhyacotriton* (Caudata: Rhyacotritonidae). *University of California Publications in Zoology*, 126, 1–91.

Hofmann, E. P., Nicholson, K. E., Luque-Montes, I. R., Köhler, G., Cerrato-Mendoza, C. A., Medina-Flores, M., Wilson, L. D., & Townsend, J. H. (2019). Cryptic diversity, but to what extent? Discordance between single-locus species delimitation methods within mainland anoles (Squamata: Dactyloidae) of Northern Central America. *Frontiers in Genetics*, 10, 11. [doi:10.3389/fgene.2019.00011](https://doi.org/10.3389/fgene.2019.00011)

Hrbacek, K. & Jech, T. (1999). Introduction to set theory, third edition, revised and expanded. Monographs and Textbooks in pure and applied mathematics, vol. 220, Marcel Dekker Inc. New York, Basel.

Jones, G. (2017). Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. *Journal of Mathematical Biology*, 74, 447. [doi:10.1007/s00285-016-1034-0](https://doi.org/10.1007/s00285-016-1034-0)

Jones, G., Aydin, Z., & Oxelman, B. (2014). DISSECT: An assignment free Bayesian discovery method for species delimitation under the multispecies coalescent. *Bioinformatics*, 31, 991–998. [doi:10.1093/bioinformatics/btu770](https://doi.org/10.1093/bioinformatics/btu770)

Kapli, P., Lutteropp, S., Zhang, J., Kobert, K., Pavlidis, P., Stamatakis, A., & Flouri, T. (2016). Multi-rate Poisson Tree Processes for single-locus species delimitation under Maximum Likelihood and Markov Chain Monte Carlo. *Bioinformatics*, 33, 1630–1638. [doi:10.1093/bioinformatics/btx025](https://doi.org/10.1093/bioinformatics/btx025)

Knowles, L. L., Carstens, B. C. (2007) Delimiting species without monophyletic gene trees. *Systematic Biology*, 56 (6), 887–895. [doi:10.1080/10635150701701091](https://doi.org/10.1080/10635150701701091)

Leavitt, S. D., Moreau, C. S., & Lumbsch, H. T. (2015). The dynamic discipline of species delimitation: Progress toward effectively recognizing species boundaries in natural populations. In *Recent Advances in Lichenology* (pp. 11–44). New Delhi: Springer. [doi:10.1007/978-81-322-2235-4_2](https://doi.org/10.1007/978-81-322-2235-4_2)

Luo, A., Ling, C., Ho, S. Y. W., Zhu, C.-D. (2018). Comparison of methods for molecular species delimitation across a range of speciation scenarios. *Systematic Biology*, 67(5), 830–846. [doi:10.1093/sysbio/syy011](https://doi.org/10.1093/sysbio/syy011)

Maddison, D. R., Swofford, D. L., Maddison, W. P. (1997). NEXUS: An extensible file format for systematic information. *Systematic Biology*, 46, 590–621. [doi:10.1093/sysbio/46.4.590](https://doi.org/10.1093/sysbio/46.4.590)

Mason, N. A., Fletcher, N. K., Gill, B. A., Funk, C., Zamudio, K. R. (2020). Coalescent-based species delimitation is sensitive to geographic sampling and isolation by distance. *Systematics and Biodiversity*, 18(3), 269–280. [doi:10.1080/14772000.2020.1730475](https://doi.org/10.1080/14772000.2020.1730475)

Masters, B. C., Fan, V., & Ross, H. A. (2011). Species Delimitation - a Geneious plugin for the exploration of species boundaries. *Molecular Ecology Resources*, 11, 154–157. [doi:10.1111/j.1755-0998.2010.02896.x](https://doi.org/10.1111/j.1755-0998.2010.02896.x)

Miralles, A. & Vences, M. (2013). New metrics for comparison of taxonomies reveal striking discrepancies among species delimitation methods in *Madascincus* lizards. *PlosONE*, 8, e68242. [doi:10.1371/journal.pone.0068242](https://doi.org/10.1371/journal.pone.0068242)

Miralles, A., Bruy, T., Wolcott, K., Scherz, M. D., Begerow, D., Beszteri, B., Bonkowski, M., Felden, J., Gemeinholzer, B., Glaw, F., Glöckner, F. O., Hawlitschek, O., Kostadinov, I., Nattkemper, T. W., Printzen, C., Renz, J., Rybalka, N., Stadler, M., Weibulat, T., Wilke, T., Renner, S., Vences, M. (2020). Repositories for taxonomic data: where we are and what is missing. *Systematic Biology*, 69, 1231–1253. [doi:10.1093/sysbio/syaa026](https://doi.org/10.1093/sysbio/syaa026)

Monaghan, M. T., Wild, R., Elliot, M., Fujisawa, T., Balke, M., Inward, D. J., Lees, D. C., Ranaivosolo R., Eggleton, P., Barraclough, T.G., & Vogler, A.P. (2009). Accelerated species inventory on Madagascar using coalescent-based models of species delineation. *Systematic Biology*, 58, 298–311. [doi:10.1093/sysbio/syp027](https://doi.org/10.1093/sysbio/syp027)

Padial, J.M., Miralles, A., De la Riva, I., & Vences, M. (2010). The integrative future of taxonomy. *Frontiers in Zoology*, 7, 16. [doi:10.1186/1742-9994-7-16](https://doi.org/10.1186/1742-9994-7-16)

Pons, J., Barraclough, T. G., Gomez-Zurita, J., Cardoso, A., Duran, D. P., Hazell, S., Kamoun, S., Sumlin, W. D., & Vogler, A. P. (2006). Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology*, 55, 595–609. [doi:10.1080/10635150600852011](https://doi.org/10.1080/10635150600852011)

Postaire B., Magalon H., Bourmaud C. A., & Bruggemann J. H. (2016). Molecular species delimitation methods and population genetics data reveal extensive lineage diversity and cryptic species in Aglaopheniidae (Hydrozoa). *Molecular Phylogenetics and Evolution*, 105, 36–49. [doi:10.1016/j.ympev.2016.08.013](https://doi.org/10.1016/j.ympev.2016.08.013)

Puillandre, N., Lambert, A., Brouillet, S., & Achaz, G. (2012b). ABGD, Automatic Barcode Gap Discovery for primary species delimitation, *Molecular Ecology*, 21, 1864–1877. [doi:10.1111/j.1365-294X.2011.05239.x](https://doi.org/10.1111/j.1365-294X.2011.05239.x)

Puillandre, N., Modica, M. C., Zhang, Y., Sirovich, L., Boisselier, M. C., Cruaud, C., Holford, M., Samadi, S. (2012a). Large-scale species delimitation method for hyperdiverse groups. *Molecular Ecology*, 11, 2671–3691. [doi:10.1111/j.1365-294X.2012.05559.x](https://doi.org/10.1111/j.1365-294X.2012.05559.x)

Puillandre, N., Brouillet, S., Achaz, G. (2021). ASAP: Assemble Species by Automatic Partitioning. *Molecular Ecology Resources*, 21(2), 609–620. [doi:10.1111/1755-0998.13281](https://doi.org/10.1111/1755-0998.13281)

Rabiee, M., Mirarab, S. (2019). SODA: Multi-locus species delimitation using quartet frequencies. *bioRxiv*, 869396. [doi:10.1101/869396](https://doi.org/10.1101/869396)

Rannala, B. (2015). The art and science of species delimitation. *Current Zoology*, 61, 846–853. [doi:10.1093/czoolo/61.5.846](https://doi.org/10.1093/czoolo/61.5.846)

Ratnasingham, S., & Hebert P.D.N. (2013). A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. *PLoS ONE*, 8: e66213. [doi:10.1371/journal.pone.0066213](https://doi.org/10.1371/journal.pone.0066213)

Renner, M.A., Heslewood, M.M., Patzak, S.D., Schäfer-Verwimp, A., & Heinrichs J. (2017). By how much do we underestimate species diversity of liverworts using morphological evidence? An example from Australasian *Plagiochila* (Plagiochilaceae: Jungermanniopsida). *Molecular Phylogenetics and Evolution*, 107, 576–593. [doi:10.1016/j.ympev.2016.12.018](https://doi.org/10.1016/j.ympev.2016.12.018)

Samadi, S., & Barberousse, A. (2006). The tree, the network, and the species. *Biological Journal of the Linnean Society*, 89(3), 509–521. [doi:10.1111/j.1095-8312.2006.00689.x](https://doi.org/10.1111/j.1095-8312.2006.00689.x)

- 381 Sites, J. W., & Marshall J. C. (2003). Delimiting species: a Renaissance issue in systematic
382 biology. *Trends in Ecology and Evolution*, 18, 462–470. [doi:10.1016/S0169-](https://doi.org/10.1016/S0169-5347(03)00184-8)
383 [5347\(03\)00184-8](https://doi.org/10.1016/S0169-5347(03)00184-8)
- 384 Sholihah, A., Delrieu-Trottin, E., Sukmono, T., Dahruddin, H., Risdawati, R., Elvira, R.,
385 Wibowo, A., Kusno, K., Busson, F., Sauri, S., Nurhaman, U., Zein, M. S. A., Fitriana, Y.,
386 Utama, I., Muchlisin, Z. A., Agnès, J. F., Hanner, R., Wowor, D., Steinke, D., Keith, P.,
387 Rüber, L., Hubert, N., (2020). Disentangling the taxonomy of the subfamily Rasborinae
388 (Cypriniformes, Danionidae) in Sundaland through DNA barcodes. *Scientific Reports*, 10,
389 2818. [doi:10.1038/s41598-020-59544-9](https://doi.org/10.1038/s41598-020-59544-9)
- 390 Solís-Lemus, C., Knowles, L.L., & Ané, C. (2015). Bayesian species delimitation combining
391 multiple genes and traits in a unified framework. *Evolution*, 69, 492–507.
392 [doi:10.1111/evo.12582](https://doi.org/10.1111/evo.12582)
- 393 Spöri, Y., & Flot, J. F. (2020). HaplowebMaker and CoMa: two web tools to delimit species
394 using haplowebs and conspecificity matrices. *Methods in Ecology and Evolution*, 11(11),
395 1434–1438. [doi:10.1111/2041-210X.13454](https://doi.org/10.1111/2041-210X.13454)
- 396 Sukumaran, J., & Knowles, L. (2017). Multispecies coalescent delimits structure, not species.
397 *Proceedings of the National Academy of Sciences USA*, 114(7), 1607–1612.
398 [doi:10.1073/pnas.1607921114](https://doi.org/10.1073/pnas.1607921114)
- 399 Sukumaran, J., Holder, T. M., Knowles, L. L. (2020). Incorporating the speciation process
400 into species delimitation. <https://github.com/jeetsukumaran/delineate>.
- 401 Vences, M. (2020). The promise of next-generation taxonomy. *Megataxa*, 1, 35–38.
402 [doi:10.11646/megataxa.1.1.6](https://doi.org/10.11646/megataxa.1.1.6)
- 403 Vences, M., Miralles, A., Brouillet, S., Ducasse, J., Fedosov, A., Kharchev, V., Kumari, S,
404 Patmanidis, S., Puillandre, N., Scherz, M. D., Kostadinov, I., Renner, S. S. (submitted).
405 iTaxoTools 0.1: Kickstarting a specimen-based software toolkit for taxonomists. **Submitted**
406 **manuscript will become publicly available on BioRxiv on 26 March 2021.**
- 407 Vieites, D. R., Wollenberg, K. C., Andreone, F., Köhler, J., Glaw, F., & Vences M. (2009).
408 Vast underestimation of Madagascar’s biodiversity evidenced by an integrative amphibian
409 inventory. *Proceedings of the National Academy of Sciences U. S. A.*, 106, 8267–8272.
410 [doi:10.1073/pnas.0810821106](https://doi.org/10.1073/pnas.0810821106)
- 411 Weisrock, D. W., Rasoloarison R. M., Fiorentino, I., Ralison, J. M., Goodman, S. M.,
412 Kappeler, P. M., & Yoder, A.D. (2010). Delimiting species without nuclear monophyly in
413 Madagascar’s mouse lemurs. *PLoS ONE*, 5, e9883. [doi:10.1371/journal.pone.0009883](https://doi.org/10.1371/journal.pone.0009883)
- 414 Wiens, J. J., & Penkrot, T. A. (2002). Delimiting species using DNA and morphological
415 variation and discordant species limits in spiny lizards (*Sceloporus*). *Systematic Biology*,
416 51, 69–91. [doi:10.1080/106351502753475880](https://doi.org/10.1080/106351502753475880)
- 417 Yang, Z., & Rannala, B. (2010). Bayesian species delimitation using multilocus sequence
418 data. *Proceedings of the National Academy of Sciences U. S. A.*, 107, 9264–9269.
419 [doi:10.1073/pnas.0913022107](https://doi.org/10.1073/pnas.0913022107)
- 420 Yang, Z., & Rannala, B. (2014). Unguided species delimitation using DNA sequence data
421 from multiple loci. *Molecular Biology and Evolution*, 31, 3125–3135.
422 [doi:10.1093/molbev/msu279](https://doi.org/10.1093/molbev/msu279)
- 423 Zhang, J., Kapli, P., Pavlidis, P., & Stamatakis, A. (2013). A general species delimitation
424 method with applications to phylogenetic placements. *Bioinformatics*, 29, 2869–2876.
425 [doi:10.1093/bioinformatics/btt499](https://doi.org/10.1093/bioinformatics/btt499)
- 426 Zhang, C., Zhang, D. X., Zhu, T., Yang, Z. (2011). Evaluation of Bayesian coalescent method
427 of species delimitation. *Systematic Biology*, 60(6), 747–761. [doi:10.1093/sysbio/syr071](https://doi.org/10.1093/sysbio/syr071)
428

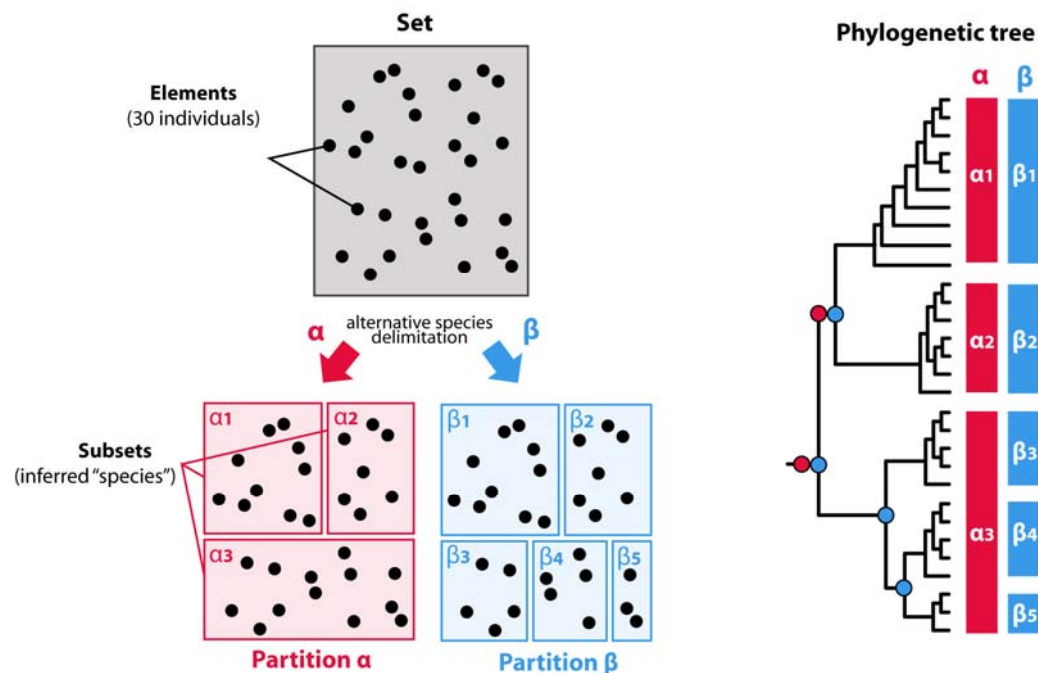
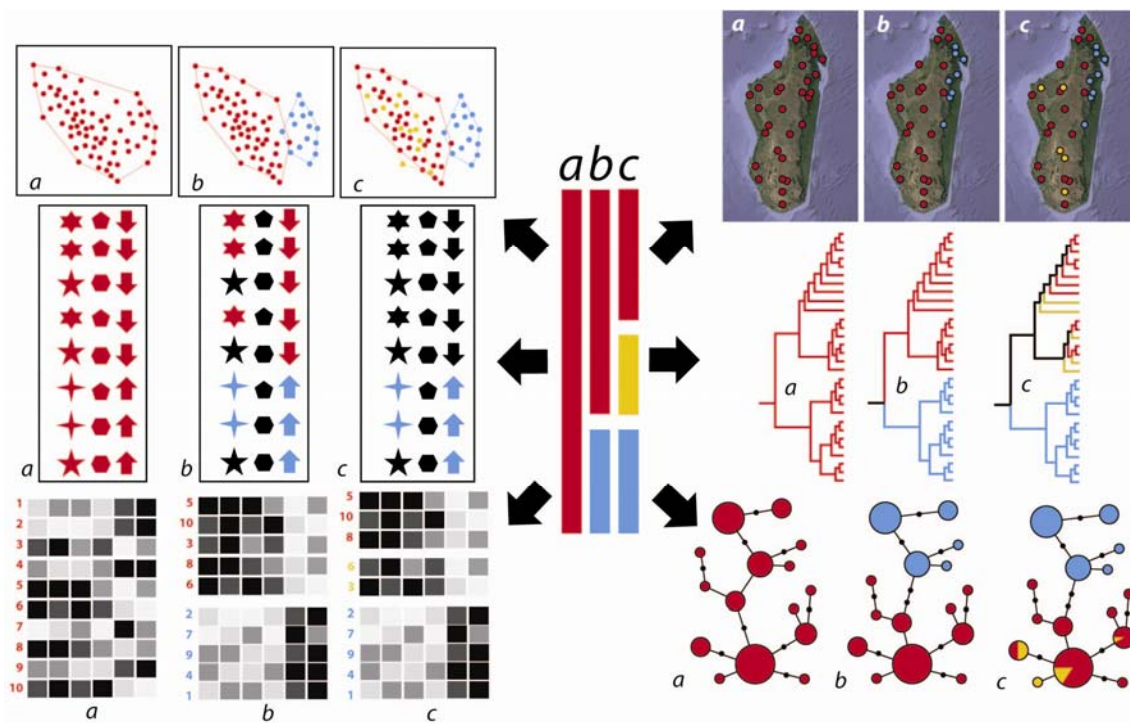


Figure 1. In mathematics, a partition of a set is a grouping of its elements into non-empty subsets, in such a way that every element is included in exactly one such subset. The main output of a species delimitation inference therefore corresponds to a partition, independently of the theoretical context, the biological input data, or the algorithms/models used. In our example, a set of 30 specimens is split by two different methods into two alternative partitions α and β , corresponding to 3 and 5 putative species (subsets), respectively. For the sake of clarity, these two alternative species partitions are represented as boxes reported next to each “species clade” in a phylogenetic tree, with hypothetical speciation events highlighted by circles via a corresponding color. Note that not all SD methods rely on a tree topology, and may therefore delimit non-monophyletic units (e.g., methods based on morphological or molecular divergence).

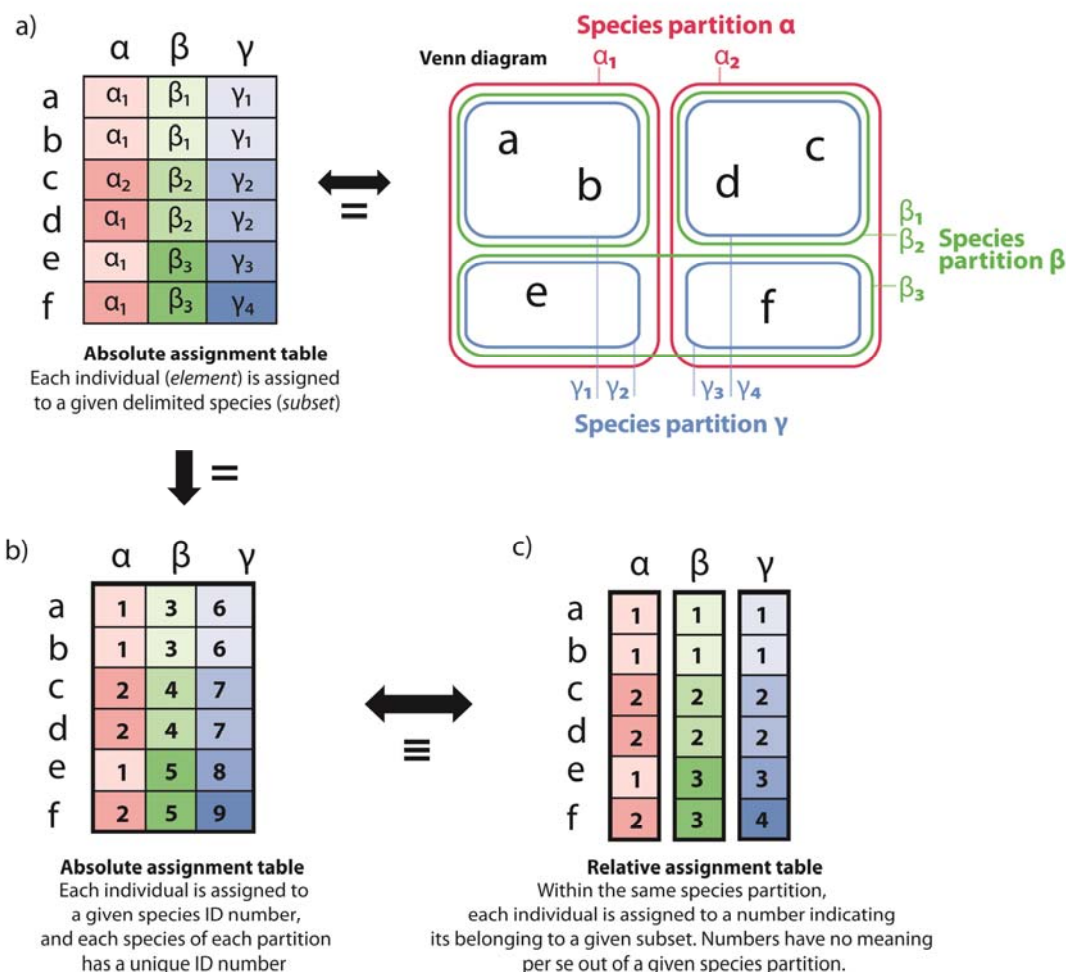
444



445

446 **Figure 2.** Illustration of exemplary potential applications of a species partition (SPART) file.
447 If it can be parsed by other programs, SPART might facilitate the exploration of taxonomic
448 datasets under various delimitation assumptions (such as morphometric Principal Component
449 Analysis, automated extraction of diagnostic traits, heatmap of meristic morphological traits,
450 distribution map, mitochondrial DNA-based phylogenetic tree, or haplotype network from
451 nuclear DNA). In the present example, the partition b represents the optimal delimitation from
452 a taxonomic perspective.
453

454



455

456 **Figure 3.** The SPART format can combine alternative species partitions of a same set of
 457 individuals (elements) into a unique multiple species partition file. (a) Example of set
 458 comprising six individuals split by three distinct SD analyses, resulting in three distinct
 459 species partitions (α , β and γ). All these species partitions are hierarchically compatible (i.e.
 460 they conform to the mathematical definition of nested sets), with the exception of the pair α -
 461 β (Venn diagram representing the alternative species partitions on the right, and
 462 corresponding assignment table on the left). These alternative species partitions can be coded
 463 in SPART either (b) by using a unique numbering for all the three species partitions (so that
 464 each species partition has its own set of species (subset) numbers) or (c) by using one
 465 numbering system per species partition. The latter representation allows combining different
 466 species partitions into a multiple species partition file without having to adjust each species or
 467 cluster number (subset). Both (b) and (c) are fully equivalent in SPART format, because the
 468 coding of each partition is independent from the others (subset assignment numbers have no
 469 meaning *per se*, they only indicate, within each partition, the common assignment to a
 470 specific subset).

471

472 **Table 1.** Automated tools dedicated to species delimitation. Abbreviations used: mtDNA, mitochondrial DNA; nDNA, nuclear DNA. Note that
 473 for programs marked with an asterisk (GMYC, PTP, SODA, DELINEATE) GUI-driven versions with SPART implementation have been
 474 prepared in the context of the iTaxoTools project but SPART output is not yet provided by all available versions. Other programs (ABGD,
 475 ASAP, TR2) already include native SPART output.

476

Tools	General principle	Hypothetical partition needed as an input (a priori species assignment)	Optimal datasets and format	SPART implementation	References
GMYC (mGMYC and bGMYC)	General mixed Yule-coalescent model	No	mtDNA – ultrametric gene tree	Yes *	Pons et al. (2006), Fontaneto et al. (2007), Monaghan et al. (2009)
BPP, iBPP	Multispecies coalescent model	Both options are possible	nDNA – multilocus alignments + (optionally in iBPP) matrix of morphological characters	In preparation	Yang & Rannala (2010, 2014), Solís-Lemus et al. (2015)
SPEDESTEM	Maximum likelihood and information theory	Yes	nDNA – ultrametric gene trees from multiple loci (nwk)	No	Ence and Carstens (2011)
ABGD	DNA barcode gap detection	No	mtDNA – sequence alignment or distance matrix	Yes	Puillandre et al. (2012)
SPECIES DELIMITATION	Coalescence / tree based approach	Yes	Topology (ultrametric tree)	No	Masters et al. (2011)
BINs	DNA barcode distance threshold + Markov clustering.	No	mtDNA – sequence alignment	No	Ratnasingham & Hebert (2013)
PTP (mPTP and bPTP)	Multi-rate Poisson ree processes model	No	Non ultrametric tree (nwk or NEXUS tree)	Yes (mPTP and bPTP) *	Zhang et al. (2013), Kapli et al. (2016)
DISSECT	Multispecies coalescent model	No	nDNA – multilocus alignments	No	Jones et al. (2014)
TR2	Multispecies coalescent model	No	nDNA – rooted gene trees from multiple loci (nwk)	Yes	Fujisawa et al. (2016)
STACEY	Multispecies coalescent model	No	nDNA – multilocus alignments	No	Jones (2017)
SODA	Quartet frequencies, based on coalescent model	No	Multiple gene tree topologies	Yes *	Rabiee & Mirarab (2019)
HaplowebMaker / CoMa	Mutual allelic exclusivity	No	nDNA – multilocus alignments	No	Spöri & Flot (2020)
ASAP	Distance-based partitions + coalescent-based scoring	No	mtDNA – sequence alignment or distance matrix	Yes	Puillandre et al. (2021)
DELINEATE	Multispecies coalescent model	Yes	Rooted ultrametric tree (nwk or NEXUS)	Yes *	Sukumaran et al. (2020)

477

