SPART: A versatile and standardized data exchange format for species partition information

Aurélien Miralles, Jacques Ducasse, Sophie Brouillet, Tomas Flouri, Tomochika Fujisawa, Paschalia Kapli, L. Lacey Lacey Knowles, Sangeeta Kumari, Alexandros Stamatakis, Jeet Sukumaran, et al.

Journal : Molecular Ecology Resources

Subject area: Computer Programs

SPART, a versatile and standardized data exchange format for species partition information

Aurélien Miralles[1*], Jacques Ducasse[2], Sophie Brouillet[1], Tomas Flouri[3], Tomochika Fujisawa[4], Paschalia Kapli[3], L. Lacey Knowles[5], Sangeeta Kumari[6], Alexandros Stamatakis[7,8], Jeet Sukumaran[9], Sarah Lutteropp[7], Miguel Vences[6]*, Nicolas Puillandre[1]*

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

[2]49 rue Eugène Carrière, 75018 Paris, France

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

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

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

[6]Braunschweig University of Technology, Zoological Institute, Mendelssohnstraße 4, 38106 Braunschweig, Germany

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

[8]Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Am Fasanengarten 5, 76131 Karlsruhe, Germany

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

***Corresponding authors**

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

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

Abstract

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

Key words. Species delimitation programs; SPART; Species partition format; Integrative taxonomy ; LIMES v2.0

Introduction

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

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

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

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

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

A standardized Species PARTition format (SPART)

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

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

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

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

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

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

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

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

152 (5) potential taxonomic reassignment of specimens in databases.

153

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

163

164 **Matricial and serial implementation of the SPART format**

165

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_spartitions") 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.

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

Tools already implementing SPART and future perspectives

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

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

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

Acknowledgments

We are grateful to Susanne Renner who stimulated this work by leading the priority program SPP 1991 "Taxon-Omics" of the Deutsche Forschungsgemeinschaft (DFG), specifically in the context of a grant on taxonomic data integration and management (RE 603/29-1), and to many members of the Taxon-Omics consortium and Guillaume Achaz (MNHN) for fruitful discussion. MV and SK were supported by DFG grant VE247/20-1, NP by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation 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, 65*, 478–494. doi: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 Ecolology, 22,* 4369–4383. doi: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
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
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
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
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
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
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
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 Ecolology Resources, 11*, 473–480. doi:10.1111/j.1755-
282 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

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

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

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

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

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. Ney 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

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

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

Knowles, L. L., Carstens, B. C. (2007) Delimiting species without monophyletic gene trees. *Systematic Biology*, 56 (6), 887–895. doi :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

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

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

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

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

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

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

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

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

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

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

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

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

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

Rabiee, M., Mirarab, S. (2019). SODA: Multi-locus species delimitation using quartet frequencies. *bioRxiv*, 869396. doi: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

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

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

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

Sites, J. W., & Marshall J. C. (2003). Delimiting species: a Renaissance issue in systematic biology. *Trends in Ecology and Evolution, 18,* 462–470. doi:10.1016/S0169-5347(03)00184-8

Sholihah, A., Delrieu-Trottin, E., Sukmono, T., Dahruddin, H., Risdawati, R., Elvira, R., Wibowo, A., Kusno, K., Busson, F., Sauri, S., Nurhaman, U., Zein, M. S. A., Fitriana, Y., Utama, I., Muchlisin, Z. A., Agnèse, J. F., Hanner, R., Wowor, D., Steinke, D., Keith, P., Rüber, L., Hubert, N., (2020). Disentangling the taxonomy of the subfamily Rasborinae (Cypriniformes, Danionidae) in Sundaland through DNA barcodes**.** *Scientific Reports,* 10, 2818. doi:10.1038/s41598-020-59544-9

Solís-Lemus, C., Knowles, L.L., & Ané, C. (2015). Bayesian species delimitation combining multiple genes and traits in a unified framework. *Evolution, 69,* 492–507. doi:10.1111/evo.12582

Spöri, Y., & Flot, J. F. (2020). HaplowebMaker and CoMa: two web tools to delimit species using haplowebs and conspecificity matrices. *Methods in Ecology and Evolution,* 11(11), 1434–1438. doi:10.1111/2041-210X.13454

Sukumaran, J., & Knowles, L. (2017). Multispecies coalescent delimits structure, not species. *Proceedings of the National Academy of Sciences USA,* 114(7), 1607–1612. doi:10.1073/pnas.1607921114

Sukumaran, J., Holder, T. M., Knowles, L. L. (2020). Incorporating the speciation process into species delimitation. https://github.com/jeetsukumaran/delineate.

Vences, M. (2020). The promise of next-generation taxonomy. *Megataxa, 1,* 35–38. doi:10.11646/megataxa.1.1.6

Vences, M., Miralles, A., Brouillet, S., Ducasse, J., Fedosov, A., Kharchev, V., Kumari, S, Patmanidis, S., Puillandre, N., Scherz, M. D., Kostadinov, I., Renner, S. S. (submitted). iTaxoTools 0.1: Kickstarting a specimen-based software toolkit for taxonomists. Submitted manuscript will become publicly available on BioRxiv on 26 March 2021.

Vieites, D. R., Wollenberg, K. C., Andreone, F., Köhler, J., Glaw, F., & Vences M. (2009). Vast underestimation of Madagascar's biodiversity evidenced by an integrative amphibian inventory. *Proceedings of the National Academy of Sciences U. S. A., 106,* 8267–8272. doi:10.1073/pnas.0810821106

Weisrock, D. W., Rasoloarison R. M., Fiorentino, I., Ralison, J. M., Goodman, S. M., Kappeler, P. M., & Yoder, A.D. (2010). Delimiting species without nuclear monophyly in Madagascar's mouse lemurs. *PLoS ONE, 5,* e9883. doi:10.1371/journal.pone.0009883

Wiens, J. J., & Penkrot, T. A. (2002). Delimiting species using DNA and morphological variation and discordant species limits in spiny lizards (*Sceloporus*). *Systematic Biology, 51,* 69–91. doi:10.1080/106351502753475880

Yang, Z., & Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences U. S. A., 107,* 9264–9269. doi:10.1073/pnas.0913022107

Yang, Z., & Rannala, B. (2014). Unguided species delimitation using DNA sequence data from multiple loci. *Molecular Biology and Evolution,* 31, 3125–3135. doi:10.1093/molbev/msu279

Zhang, J., Kapli, P., Pavlidis, P., & Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics, 29,* 2869–2876. doi:10.1093/bioinformatics/btt499

Zhang, C., Zhang, D. X., Zhu, T., Yang, Z. (2011). Evaluation of Bayesian coalescent method of species delimitation. *Systematic Biology,* 60(6), 747–761. doi:10.1093/sysbio/syr071

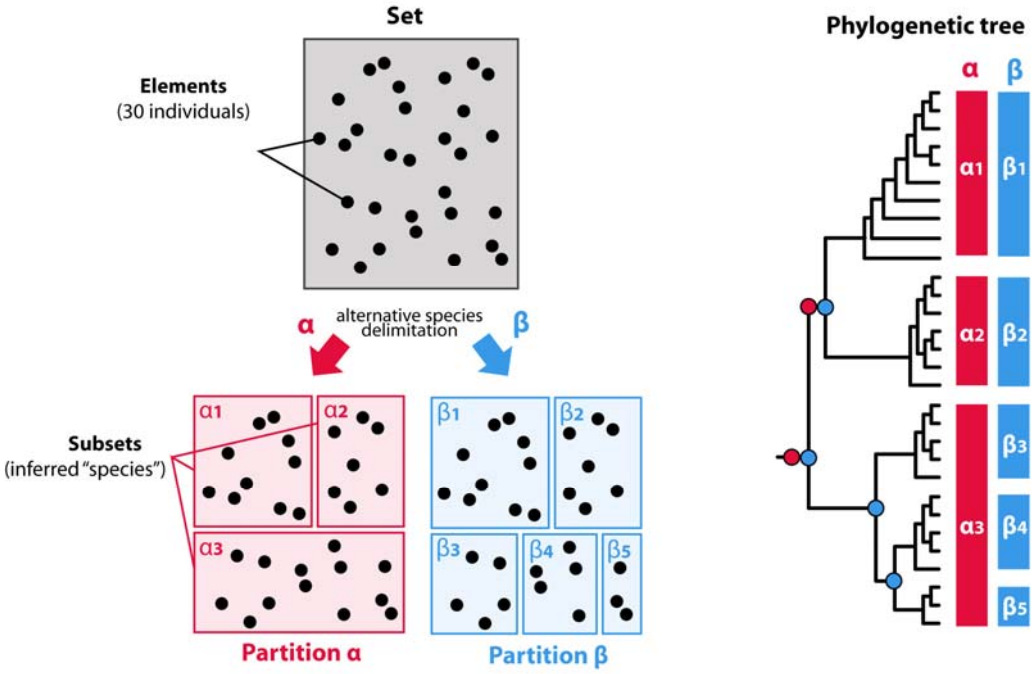
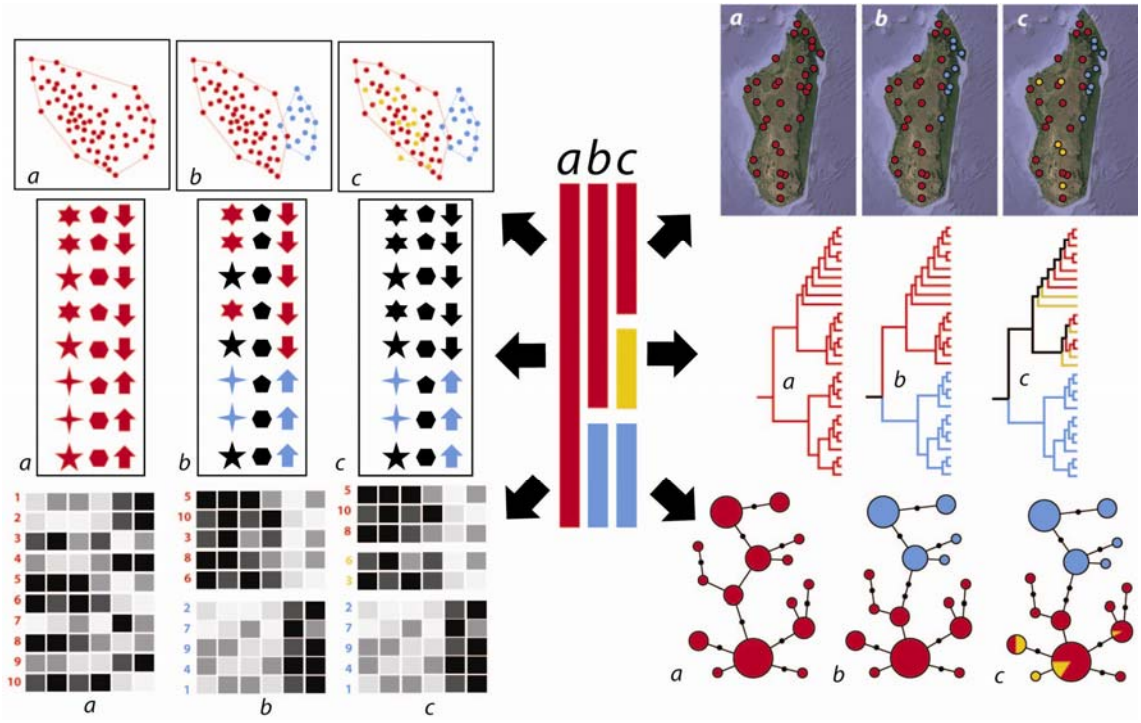


Figure 1. In mathematics, a partition of a set is a grouping of its elements into non-empty subsets, in such a way that every element is included in exactly one such subset. The main output of a species delimitation inference therefore corresponds to a partition, independently of the theoretical context, 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

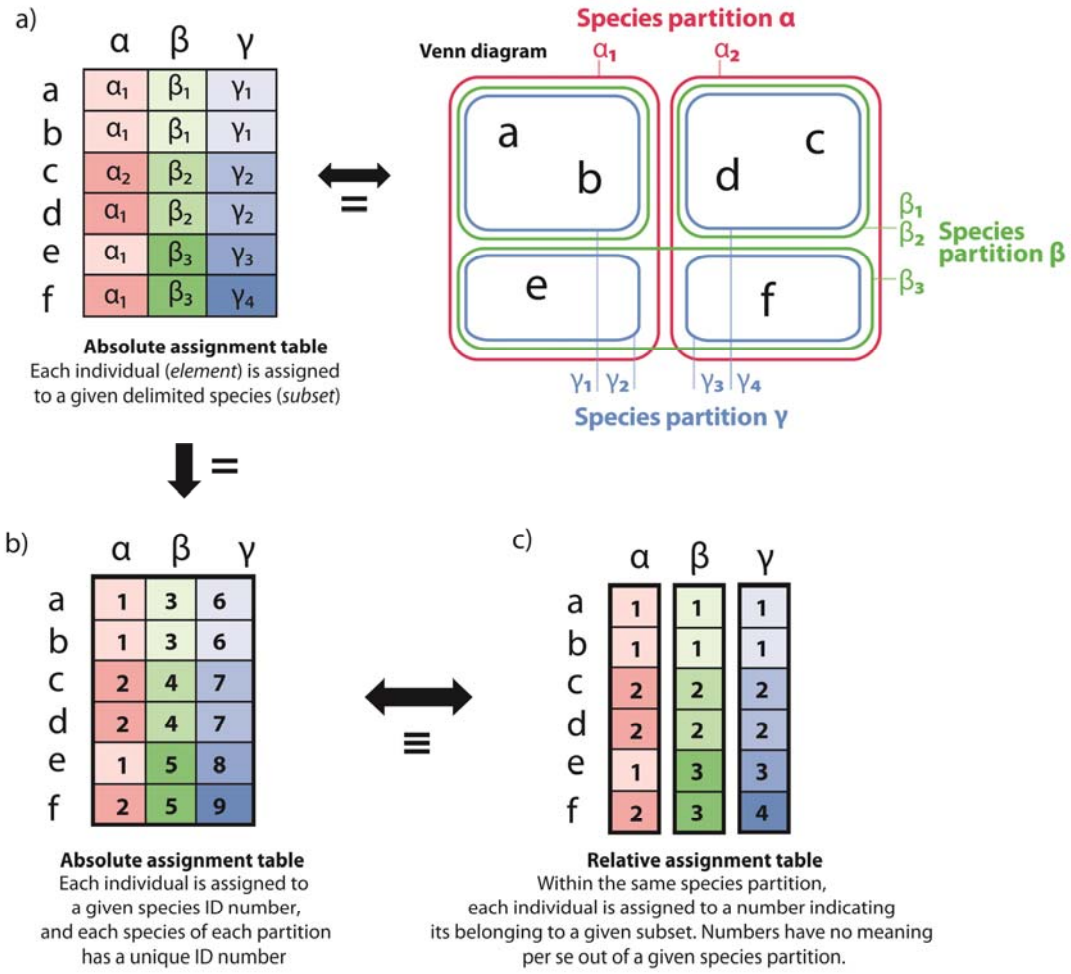


Figure 3. The SPART format can combine alternative species partitions of a same set of individuals (elements) into a unique multiple species partition file. (a) Example of set comprising six individuals split by three distinct SD analyses, resulting in three distinct species partitions (α, β and γ).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

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

Tools	General principle	Hypothetical partition needed as an input (a priori species assignement)	Optimal datasets and format	SPART impletementation	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)

478