

II. Taxonomy in the Electronic Age: An e-Monograph of the Papaya Family (Caricaceae) as an Example [§]

Fernanda Antunes Carvalho^{1,*}, Denis Filer², Susanne S. Renner¹

¹ Systematic Botany and Mycology, University of Munich (LMU), Menzinger Strasse 67,
80638 Munich, Germany

² Department of Plant Sciences, University of Oxford, South Parks Road Oxford,
OX1 3RB, United Kingdom

* Corresponding author: antunesfc@gmail.com

[§] published in: *Cladistics*, 31(3), 321–329, June 2015, doi: 10.1111/cla.12095

© The Willi Hennig Society 2014, reproduced here with kind permission of John Wiley & Sons, Inc.

Abstract

The need for taxonomists to take full advantage of biodiversity informatics has been clear for at least 10 years. Significant progress has been made in providing access to taxonomic resources online, including images of specimens, especially types; original species descriptions; and georeferenced collection data. However, in spite of persuasive calls for e-monography, there are few, if any, completed projects, even though monographic research is the only mechanism for reducing synonymous names, which are estimated to comprise 50% of all published names. Caricaceae is an economically important family of flowering plants from Africa and the Neotropics, best known for the fruit crop papaya. There is a large amount of information on the family, especially on chemistry, crop improvement, genomics, and the sex chromosomes of papaya, but up-to-date information on the 230 names and which species they might belong was not available. A dynamically updated e-monograph of the Caricaceae now brings together all information on this family, including keys, species descriptions, and specimen data relating the 230 names to 34 species and one hybrid. This may be the first taxonomic monograph of a plant family completely published online. The curated information will be continuously updated to improve the monograph's comprehensiveness and utility.

Introduction

The Plant List (2010) shows 1,040,426 published names for plants of which 29% are accepted, 25% of unclear status, and 46% considered synonymous with other species names. The problem of synonymous names arises because taxonomists inadvertently name the same species several times, usually because it is widespread and has been collected in far-apart regions and/or because widespread species often are morphologically variable, sometimes in correlation with their environment, making it difficult to assess species status until a dense collection series can be studied. In the flowering plants, there may be 3–4 synonyms for every accepted name (Scotland and Wortley, 2003;

Wortley and Scotland, 2004; Paton et al., 2008; The Plant List, 2010). The problem of synonymous names is by no means restricted to plants, although reliable estimates for all eukaryotes are difficult to obtain (Alroy, 2002; Mora et al., 2011). Synonymous names are not a harmless nuisance, and their rate seems to be increasing apace with the rate at which new species are described (Fig. 1). When it comes to conserving species or using species for medical or any other kind of purpose, synonymous names will result in two kinds of errors: they result in wrong, usually narrower, species range estimates than warranted because each name will be associated with its own “species” range; and they make it difficult to find material of, or published information on, a particular biological entity because users cannot know which names refer to which good species.

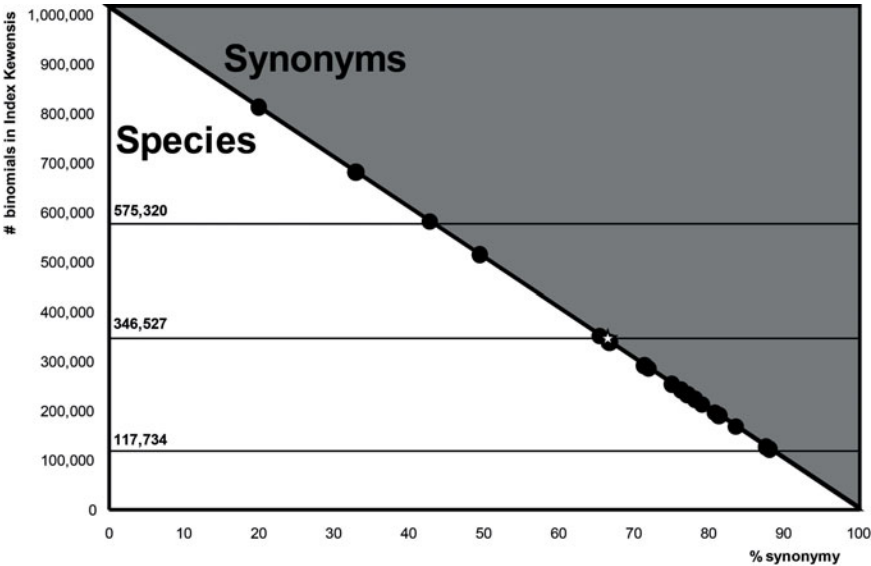


Fig. 1. Relationship between synonymy percentage and number of species from Wortley and Scotland (2004), reproduced with permission of the authors.

The assessment, and reassessment, of the status of a name as either a synonym or a good species is done during monographic research. Monographic research is based on bringing together the information pertaining to all names that have ever been published for some group, typically a genus or a family. This will include the publication in which a name was first proposed (the so-called protologue), all specimens to which the name has been applied (rightly or wrongly), the accepted names and their synonyms, morphological descriptions for each species, geographic coordinates of relevant collections, chromosome numbers, chemical traits, flowering or fruiting times, and DNA sequences from specimens given one or several of the names in question. A monographer will study the specimens, often do some phylogenetic work based on DNA sequences of a representative subset, and reach a conclusion about which names refer to which species. He/she next constructs a key to identify the accepted species and prepares an authoritative list of the accepted and synonymized names. Monography is the only known mechanism for achieving quality control in taxonomy and for reducing the number of synonymous names that clutter up databases and hinder progress in our knowledge of the World's biodiversity and its conservation status.

Because taxonomy is the portal to the entire information available about species, the need for taxonomic research to "move into the electronic age" has long been clear (Bisby et al., 2002; Godfray, 2002; Wilson, 2003; Kress, 2004; Wheeler, 2004; Scotland and Wood, 2012). Indeed species descriptions of animals and plants are now increasingly being published online (Blagoderov et al., 2010; Knapp, 2010; Penev et al., 2010; Knapp et al., 2011). Monography, however, has not followed suit, in spite of the availability of massive online databases of literature and digitized specimen, wikis, ever cheaper digital photography and microscopy (essential to the study of herbarium specimens), and dedicated platforms for taxonomy, such as the Botanical Research and Herbarium Management System (BRAHMS, <http://herbaria.plants.ox.ac.uk/bol/>) and Scratchpads (<http://scratchpads.eu/>). The new "cyber-taxonomy" or "e-taxonomy" (Zauner, 2009; Wheeler and Valdecasas, 2010) is reality only for species descriptions and lists of names but not yet for

monographic research (Scotland and Wood, 2012). Although there are several ongoing taxon-centered initiatives (Appendix 1), to our knowledge no revision or monograph of any large group has been completed. The advantages of online monography, such as the possibility of including near-unlimited color images and the option of up-dating information, have thus not been realized.

Overview of the Electronic Monograph of Caricaceae and its Underlying Database

Here we present a recently completed electronic monograph of a plant family (Caricaceae), the result of research that brought together the available collections with digital libraries, digitized specimen data, and other taxonomic and methodological tools available, including DNA sequencing for barcoding the recognized species (Carvalho and Renner, 2012, 2013).

Caricaceae is a small family of flowering plants from Africa and the Neotropics, best known for the fruit crop *Carica papaya*. The family's economic importance lies not only in the papaya fruit, but also in the production of papain, a cysteine proteinase widely used in food and pharmaceutical industries. A search for the topics 'papaya' and 'papain' in Web of Knowledge retrieves approximately 20,823 and 42,100 citations, respectively (ReutersISI, 2013). Several Caricaceae are considered as unexploited crops because of their nutritive fruits, high concentration of papain-like enzymes, and resistance to pathogens (Kyndt et al., 2007; Ramos-Martínez et al., 2012). Among these are the so-called highland papayas, species of *Vasconcellea*, a genus thought to be synonymous with *Carica* until Badillo (2000) cleared up their morphological distinctness (Badillo, 2000). Molecular data have revealed that the closest relative of papaya is a clade of four species in Mexico and Guatemala entirely neglected by ecologists and breeders (Carvalho and Renner, 2012). The lack of knowledge before 2012 on the true closest relatives of papaya resulted in the assumption that the highland papayas (*Vasconcellea* species) were the best group to use in papaya improvement (Scheldeman et al., 2011; Coppens d'Eeckenbrugge et al., 2014).

As required in a taxonomic monograph, the e-monograph of Caricaceae (<http://herbaria.plants.ox.ac.uk/bol/caricaceae>) allocates all names (here 230) to recognized species (here 34 and one hybrid), providing a comprehensive data infrastructure for scientists and nonscientists alike. The database is being developed, managed and published online using BRAHMS (<http://herbaria.plants.ox.ac.uk/bol>) developed at the University of Oxford. In carrying out this research on the Caricaceae, we added a range of new features to BRAHMS that facilitate cyber-monography emphasizing thus the importance of close collaborations among taxonomists and bioinformaticians (Stein, 2008).

The e-monograph of Caricaceae and its underlying database, store (and make available) data and images on collections, herbarium specimens, literature, and the revised nomenclature (including accepted names, synonyms, *nomina nuda*, illegitimate names, and excluded names). The monographic research resulted in updated circumscriptions of the recognized species, including detailed plates (Fig. 2), and precise geographic distribution of all relevant collections. Links to supportive literature and high-resolution images of type specimens are provided for each species as are cross-references to databases, such as The Plant List, TROPICOS, IPNI, and GBIF. General information on the family, including its ecology, sex chromosomes, and molecular phylogeny is provided, along with identification keys to all genera and species.

All these data are accessible through BRAHMS online and summarized in Table 1. Searches by taxon, collector, geographic place name, and map area (Fig. 3) generate tables that can also be shown in text format. Images can be grouped and filtered, and viewed at different resolutions. Maps are available using clustered Google Maps or Google Earth, both configurable with zoom features. A detailed description of the methods used to build the e-monograph is given in Appendix 2.

Discussion

Among the challenges for taxonomy today are to incorporate results and insights from molecular phylogenetic work and to tackle the problem of the 46–50% synonymous names already published (Scotland and Wortley, 2003;

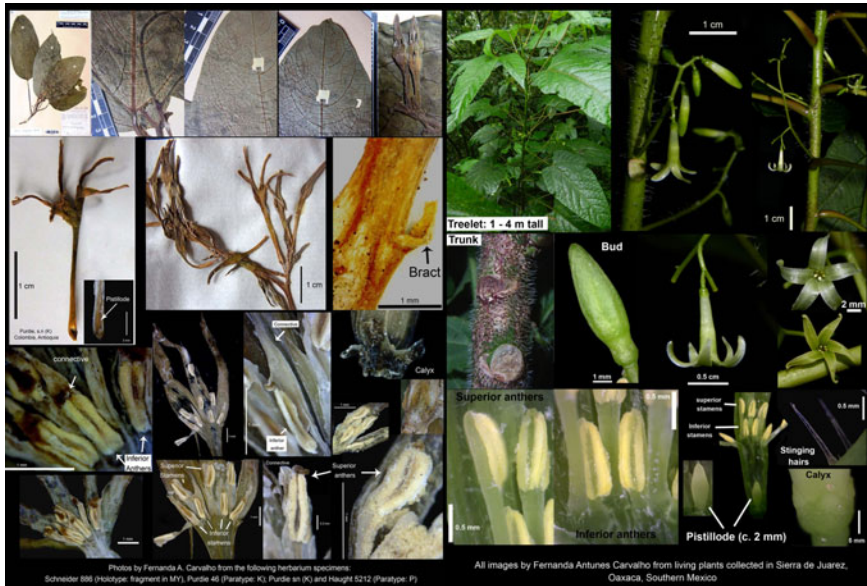


Fig. 2. Examples of species plates used to describe species in the website. To the left are images of details of male flowers and inflorescences based on herbarium specimens of *Vasconcellea longiflora*; to the right, images of living material of *Horovitzia cnidoscoloides*, one out of the four little known closest relatives of papaya.

Wortley and Scotland, 2004; The Plant List, 2010). Both challenges can only be addressed through monographic work in which species and genus circumscriptions are vetted and updated, based on the study of specimens and consideration of relevant phylogenetic results on relationships.

Reliably circumscribed and named species are also required to fulfill the promise of DNA barcodes, at least if that promise is finding names for unidentified specimens via matching of short DNA sequences (obviously, one can also match unnamed material to unnamed sequences). Simply increasing the rate of species discovery, while important, does not address either of these challenges because naming a newly discovered species does not require a complete assessment of all existing names that might apply (which would often take too much time). It is therefore likely that as the number of species descriptions increases (Costello et al., 2013), so does the number of newly created synonyms (Fig. 1).

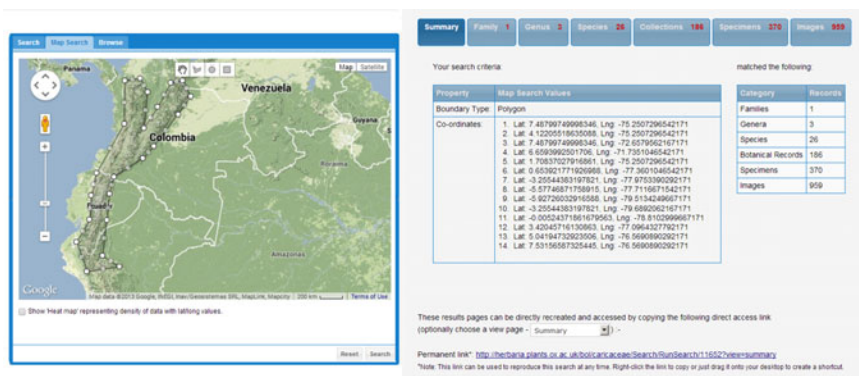


Fig. 3. Map search in BRAHMS. The left figure shows a polygon that can be drawn by the user to delimit the area of interest, in this case, the Andes from northern Peru to northern Colombia. To the right is a summary of the results, which includes number of genera, collections, specimens, and images available in the database. It also provides the coordinates of the polygons, which can be used to create a shape file.

A well-resolved, expert-vetted nomenclature and detailed information on the distribution of species are of great importance for many fields of research (Yesson et al., 2007; Bortolus, 2008; Patterson et al., 2010; Lis and Lis, 2011; Santos and Branco, 2012). However, high-quality data produced by taxonomists in revisions and monographs are of little use unless widely accessible (Kress, 2004). This is especially important for economically important groups, which often are also groups with a high rate of nomenclatural changes (as is the case for Caricaceae). Open-access information to this highly organized set of online data and images for the Caricaceae benefits the scientific community broadly as well as those working on the food and medicinal aspects of the family. This includes the community of herbarium curators, researchers focusing on papaya genomics (Fig. 4A), breeders, and the non-scientific public. In addition, georeferenced specimens are the basis for the growing field of bioclimatic modeling (Fig. 4B) and for a reliable baseline to document the effects of ongoing climatic changes on plant ranges.

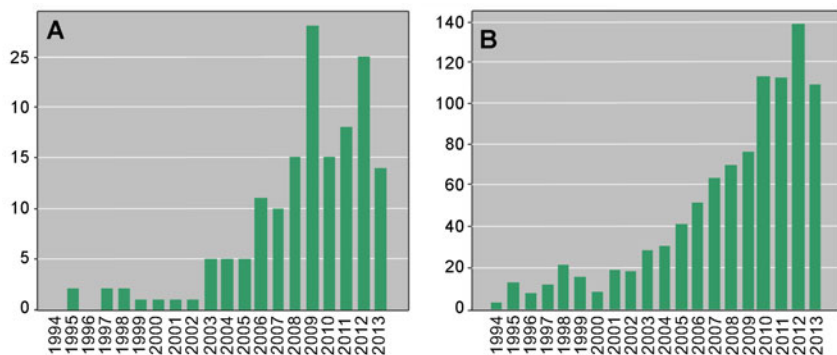


Fig. 4. (A) Number of published studies with the topic search fields “Caricaceae” and “genome”; a total of 168 records were found. **(B)** Number of published studies on Bioclimatic Modeling per year; in total 1,002 records. (Web of Knowledge accessed 18 November 2013).

In the case of the papaya family, the most recent taxonomic accounts were by Victor Manuel Badillo (1920–2008; <http://herbaria.plants.ox.ac.uk/bol/caricaceae#badillo>) a Venezuelan taxonomist who dealt with c. 200 names described in the family, 64 of these basionyms (meaning that the remainder result from changing generic concepts). The work of Badillo (1971, 1993, 2000) is poorly accessible, and since his last publication 13 years ago (Badillo, 2001) no further taxonomic work on the Caricaceae has been published. Meanwhile, molecular work on the family took off (Van Droogenbroeck et al., 2002; Kyndt, Romeijn-Peters, et al., 2005; Kyndt, Van Droogenbroeck, et al., 2005; Carvalho and Renner, 2012). The IUCN Red List of Threatened Species (IUCN, 2013) lists six endangered species of Caricaceae, none under the correct name; the new e-monograph available at <http://herbaria.plants.ox.ac.uk/bol/caricaceae>, now includes updated information on the vulnerability of species that together with the geographic and ecological information should help in conservation efforts.

Table 1. Summary of the Caricaceae e-monograph data available online as of 13 Feb. 2014. Invalid, Illegitimate, Excluded and Uncertain names are not included in this table

| Genera (6) | Species (34 + 1 hybrid) | Synonyms (160) | Collections examined (2950) | Specimens examined (4337) | Georeferenced collections (2204) | Images (10988) |
|--|---|----------------|-----------------------------|---------------------------|----------------------------------|----------------|
| <i>Cylicomorpha</i> | <i>C. parviflora</i> Urb. | 1 | 36 | 57 | 28 | 246 |
| | <i>C. solmsii</i> Urb. | 1 | 18 | 27 | 12 | 158 |
| <i>Carica</i> | <i>C. papaya</i> L. | 21 | 590 | 773 | 30 | 1911 |
| <i>Horovitzia</i> | <i>Horovitzia cnidoscoloides</i> | 1 | 97 | 26 | 19 | 68 |
| <i>Jarilla</i> | <i>Jarilla chocola</i> Standl. | 1 | 37 | 52 | 36 | 136 |
| | <i>Jarilla caudata</i> (Brandegee) Standl. | 4 | 50 | 62 | 48 | 159 |
| | <i>Jarilla heterophylla</i> (Cerv. ex La Llave) Rusby | 4 | 71 | 85 | 69 | 219 |
| <i>Jacaratia</i> (7 species) | <i>J. digitata</i> (Poepp. & Endl.) Solms-Laub. | 3 | 178 | 251 | 167 | 512 |
| | <i>J. spinosa</i> (Aubl.) A.DC. | 8 | 209 | 329 | 190 | 849 |
| | <i>J. chocoensis</i> A.H.Gentry & Forero | 0 | 15 | 21 | 15 | 31 |
| | <i>J. corumbensis</i> Kuntze | 3 | 41 | 86 | 34 | 281 |
| | <i>J. dolichaula</i> (Donn.Sm.) Woodson | 1 | 128 | 172 | 120 | 450 |
| | <i>J. mexicana</i> A. DC. | 7 | 142 | 158 | 132 | 500 |
| | <i>J. heptaphylla</i> (Vell.) A.DC. | 1 | 30 | 37 | 26 | 134 |
| <i>Vasconcellea</i> (21 species and 1 hybrid) | <i>V. candicans</i> (A.Gray) A.DC. | 3 | 26 | 38 | 19 | 141 |
| | <i>V. cauliflora</i> (Jacq.) A.DC. | 8 | 111 | 159 | 87 | 452 |
| | <i>V. crassipetala</i> (V.M.Badillo) V.M.Badillo | 1 | 6 | 16 | 5 | 55 |
| | <i>V. glandulosa</i> A.DC. | 9 | 80 | 153 | 71 | 419 |
| | <i>V. goudotiana</i> Triana & Planch. | 4 | 20 | 35 | 11 | 107 |
| | <i>V. horovitziana</i> (V.M.Badillo) V.M.Badillo | 1 | 13 | 34 | 3 | 125 |
| | <i>V. longiflora</i> (V.M.Badillo) V.M.Badillo | 1 | 6 | 10 | 2 | 26 |
| | <i>V. microcarpa</i> (Jacq.) A.DC. | 22 | 401 | 774 | 336 | 1508 |
| | <i>V. monoica</i> (Desf.) A.DC. | 7 | 32 | 70 | 14 | 156 |
| | <i>V. omnilingua</i> (V.M.Badillo) V.M.Badillo | 1 | 2 | 3 | 1 | 16 |
| <i>Vasconcellea</i> (21 species and 1 hybrid) | <i>V. palandensis</i> (V.M.Badillo, Van den Eynden & Van Damme) V.M.Badillo | 1 | 3 | 6 | 3 | 27 |
| | <i>V. parviflora</i> A.DC. | 5 | 61 | 109 | 46 | 323 |
| | <i>V. pubescens</i> A.DC. | 10 | 102 | 230 | 68 | 501 |
| | <i>V. pulchra</i> (V.M.Badillo) V.M.Badillo | 1 | 13 | 36 | 10 | 98 |
| | <i>V. quercifolia</i> A.St.-Hil. | 14 | 157 | 253 | 114 | 675 |

Molecular Phylogeny, Biogeography and an
e-Monograph of the Papaya Family (Caricaceae) as an
Example of Taxonomy in the Electronic Age

Antunes Carvalho, F.

2015, XIV, 147 p. 24 illus., Softcover

ISBN: 978-3-658-10266-1