# Development of a Bioinformatical Pipeline for Integrative Species Delimitation in *Leucanthemum* Mill. (Compositae, Anthemideae)

**DISSERTATION**

ZUR ERLANGUNG DES DOKTORGRADES DER NATURWISSENSCHAFTEN (DR. RER. NAT.) DER FAKULTÄT FÜR BIOLOGIE UND VORKLINISCHE MEDIZIN DER UNIVERSITÄT REGENSBURG


vorgelegt von

Tankred Hans Michael Ott

aus

Deggendorf


im Jahr

2022

# Development of a Bioinformatical Pipeline for Integrative Species Delimitation in *Leucanthemum* Mill. (Compositae, Anthemideae)

**DISSERTATION**

ZUR ERLANGUNG DES DOKTORGRADES DER NATURWISSENSCHAFTEN (DR. RER. NAT.) DER FAKULTÄT FÜR BIOLOGIE UND VORKLINISCHE MEDIZIN DER UNIVERSITÄT REGENSBURG

vorgelegt von

Tankred Hans Michael Ott

aus

Deggendorf

im Jahr

2022

Das Promotionsgesuch wurde eingereicht am: 10.06.2022

Die Arbeit wurde angeleitet von: Prof. Dr. Christoph Oberprieler

Unterschrift: ……

Tankred Ott

# Abstract

The description and delimitation of species, the fundamental units of biodiversity, is one of the main research topics of systematics and taxonomy. Traditionally, species were usually described based on differences in morphological characters. In the course of the last century, ecological niches, geographical distributions, and the degree of interbreeding became additional criteria for species delineation. More recently, the availability of genetic marker systems allowed researchers to additionally incorporate the underlying genealogical information. The combination of those sources of evidence for taxonomic studies is called integrative taxonomy and is the foundation of modern species delimitation. The present thesis studies integrative species delimitation of the diploid and tetraploid taxa of the genus *Leucanthemum* Mill. (Compositae, Anthemideae), a young, hybridizing, and closely-knit polyploid complex. I propose a bioinformatical pipeline, a set of methods, tools, and analyses, comprising the collection and analysis of data, and the joint evaluation of the observed taxonomical patterns. In the framework of integrative taxonomy, genealogical, ecological, geographical, and morphological evidence is compiled, represented by restriction associated DNA sequencing (RADseq) data, abiotic niches, distribution areas, and leaf shape, respectively. The first two studies of the thesis describe the implementation of GINJINN, a deep learning software tool for the automatic detection and extraction of structures from digital images and show its successful application by, among other examples, retrieving leaf silhouettes from herbarium specimens of *Leucanthemum*. The third study focuses on the integrative species delimitation of the diploid *Leucanthemum* taxa based on the previously mentioned four data sources. In this context, novel methods and tools for consensus clustering, reconstructing leaf silhouettes extracted from herbarium specimens, and for approximating distribution areas from sparse collection data are implemented. The integration of the four data sources leads to the down-ranking of three diploid *Leucanthemum* species to the subspecies level. The last study applies the previously described methodological pipeline to delimit eight tetraploid *Leucanthemum* taxa and introduces a in this context new clustering method for supporting genealogical species delimitation, leading to the proposal of a new taxonomic rank for four taxa.

## References of Published, Submitted, and Prepared Manuscripts

The present thesis is composed of the following published, submitted, or prepared manuscripts:

A. Ott, T., Palm, C., Vogt, R., Oberprieler, C. (2020). GinJinn: An object-detection pipeline for automated feature extraction from herbarium specimens. *Applications in Plant Sciences*, 8(6), e11351.

B. Ott, T., Lautenschlager, U. (2022). GinJinn2: Object detection and segmentation for ecology and evolution. *Methods in Ecology and Evolution*, 13(3), 603-610.

C. Ott, T., Schall, M., Vogt, R., Oberprieler, C. (under review). The warps and wefts of a polyploidy complex: integrative species delimitation of the diploid *Leucanthemum* (Compositae, Anthemideae) representatives. *Annals of Botany*.

D. Ott, T., Vogt, R., Oberprieler, C. (in preparation). Untangling the weave: species delimitation of the tetraploid *Leucanthemum* Mill. (Compositae, Anthemideae) representatives.

In addition, I contributed to following publications, which are not part of this thesis:

E. Wagner, F., Ott, T., Schall, M., Lautenschlager, U., Vogt, R., Oberprieler, C. (2020). Taming the Red Bastards: Hybridization and species delimitation in the *Rhodanthemum arundanum*-group (Compositae-Anthemideae). *Molecular Phylogenetics and Evolution*, 144, 106702.

F. Oberprieler, C., Conti, F., Dorfner, M., Eder, S. M., Heischneider, A., Ott, T., Scheunert, A., Vogt, R. (2022). The taxonomy of *Leucanthemum ircutianum* (Asteraceae, Anthemideae) in the Apennine Peninsula based on AFLP fingerprinting, plastid DNA sequence variation and eco-climatological niche reconstruction. *Botanical Journal of the Linnean Society*, boac003.

G. Dorfner, M., Ott, T., Ott, P., Oberprieler, C. (2022). Long-read genotyping with SLANG (Simple-Long-read loci Assembly of Nanopore data for Genotyping). *Applications in Plant Sciences*, 10(3), e11484.

Personal Contributions:

Publication A

Tankred Ott (TO), Christoph Palm (CP)., and Christoph Oberprieler (CO) conceived the present study. Robert Vogt (RV) produced the scans of the preserved specimens. TO programmed the software. CP consulted the exemplary data analysis. A first draft of the paper was written by TO. with input from CP., RV., and CO.

Publication B

TO and Ulrich Lautenschlager (UL) envisioned the present work, implemented the software, carried out the analyses and wrote the manuscript. Both authors approved the final version of the manuscript. We further note that UL and TO contributed equally to this work. The order of their names in the author list was decided by coin toss.

Publication C

TO and CO conceived the present study. RV produced scans of the preserved specimens and supplied the collection locations. Maximilian Schall (MS) annotated the leaves. The methods for leaf processing and approximating geographical overlap were conceived and implemented by TO. TO conducted all analyses. A first draft of the manuscript was written by TO. The final manuscript was written by TO, CO, and RV.

Publication D

The study was conceived by TO and CO. RV provided scans of preserved specimens and supplied the collection locations. TO conducted the analyses. The first draft of the manuscript was written by TO. The final manuscript was written by TO, CO, and RV.

# Contents

# Electronic Supplementary

# Chapter 1: General Introduction

## Species Delimitation and Integrative Taxonomy

The delineation and description of species, the fundamental units of biodiversity, are the focal points of taxonomic research. While an interesting research topic, the delimitation of species is not only an end itself, but is of great importance for scientific communication among biologists, for the evaluation of biodiversity in nature conservation studies, and can even influence economic and politic decisions (Zachos 2016: 163-174).

Soundly delimiting species, however, is often difficult, as is evident by the plethora of species concepts, i.e., definitions of what species are, that have emerged over the years. Today, there are more than 30 different and partially overlapping species concepts (Zachos, 2016), and depending on the applied concept, the resulting species delimitation can be completely different. For example, Mayr's "biological species concept"(Mayr, 1970) requires the presence of reproductive barriers between entities to rank them as distinct species, while the "phylogenetic species concept" (Cracraft, 1983) demands monophyletic sister groups, to only name two. In practice, however, many of the currently described species were simply delimited based on morphological differences.

Over the last century, ecological and geographical variation, and reproductive barriers became more important for ranking entities on the species level, and in recent decades, the analysis of DNA sequences, allowed scientists to draw evidence from a completely new source of data. Researchers tried to combine those data sources to improve and solidify their taxonomic decisions from the beginning. As early as 1898, von Wettstein (1898) proposed a surprisingly modern, evolutionary approach to taxonomy, incorporating ecological, geographical, and morphological evidence for delimiting species from sub-species, depending on the degree of differentiation. Almost a century later, Wiley (1978) took a similar approach, proposing that species should be lineages "with their own evolutionary tendencies", while stressing the importance of ecological differentiation and geographical patters. Only relatively recently the term "integrative taxonomy" (Dayrat, 2005) was coined for such a combination of evidence. Integrative taxonomy is characterized by the simultaneous observation and the joint evaluation of multiple sources and modes of data for the ranking of biological entities. Even though most researchers will agree that it is better to draw conclusions from a wide range of perspectives, the exact implementation of this concept is still debated (Daglio & Dawson, 2019). There have been

proposals for methodological pipelines for the application of this concept (e.g., Schlick-Steiner et al., 2010) for single study groups, but a general consensus is not yet established. Nevertheless, many recent taxonomic studies implicitly or explicitly apply integrative species delimitation, rendering it the de-facto standard for modern species delimitation.

## Data Sources for Species Delimitation in Plants

Species delimitation can be approached as sequential procedure consisting of a) grouping and b) ranking of studied entities as described by Stuessy et al. (2014) or Reydon & Kunz (2019). In the framework of integrative taxonomy, the grouping step should search for discontinuities in a wide range of datasets comprising information on the morphology, ecology, geography, reproduction, and genealogy of the entities under study, while the latter evaluates those patterns and draws taxonomic conclusions. Collecting data for the grouping step in-situ or growing plants in botanical gardens can be prohibitively expensive and labor-intensive. Fortunately, today, herbaria all over the world house a wealth of preserved, and often digitized specimens for many plant groups. These collections can be tapped into for extracting morphological characters, reconstructing ecological niches and geographical distributions based on specimen sampling locations, and even for retrieving genetic information.

While the geographical sampling locations of specimens can be easily retrieved from most those collections, the extraction of morphological characters from the specimens themselves or from digital images of them is still a labor-intensive process. Many studies have dealt with simplifying and automating the retrieval of morphological information from herbarium specimens (e.g., Corney et al., 2012a; Unger et al., 2016; Carranza-Rojas et al., 2017; see chapters 2 and 3). The strategy for the extraction of data from digitized specimens, be it morphology or text, shifted from classic image processing techniques (e.g., Unger et al., 2016) to machine learning-based solutions (e.g., Weaver et al., 2020), such as those presented in chapters 2 and 3 within the present thesis. Automating the retrieval of morphological characters, such as, for example, leaf shapes or flower colors from herbarium specimens, allows scientists to tap into natural-history collections for species delimitation.

The geographical distribution and abiotic niches of studied entities are best determined based on comprehensive species distribution rasters determining presence or absence of the entities under study, which are often unavailable for non-model organism.

As a result, methods have been developed for reconstructing ecological niches based on incomplete distribution rasters, e.g., MaxEnt (Phillips et al., 2021), or for approximating true distributions from patchy sampling, such as convex or alpha hulls (Burgman & Fox, 2003; Capinha & Pateiro-López, 2014; Meyer et al., 2017), or kernel density estimates (Burgman & Fox, 2003; Cardillo & Warren, 2016). Using those methods, and collection locations, as well as additional data retrieved from natural collections, species delimitation of non-model organisms can be enhanced with geographical and ecological information.

The degree of reproductive isolation (RI) is another criterion that can be approached for species delineation in a study group. While directly measuring reproductive isolation by means of crossing experiments is extremely labor- and time-intensive, and is not as relevant in context of plants as even taxa from different ploidy levels can be crossed (e.g., Greiner & Oberprieler, 2012), RI can be approximated by observing genetic variation among entities. With the advent of next (NGS) and more recently third generation sequencing techniques, compiling comprehensive genetic datasets became increasingly feasible. Arguably, sequencing the whole genome for all entities under study is the optimal way to assess the genetic variation, but unfortunately problematic and pricey for non-model organisms without a published reference genome. As an alternative, restriction site-associated DNA sequencing (RADseq; Poland et al., 2012) became a popular method for population genetic and systematic studies. RADseq produces hundreds to ten-thousands of short (100 bp – 300 bp) and anonymous loci which are (mostly) randomly distributed over the whole genome, without requiring a reference. Another advantageous property of RADseq is that it can be successfully applied to degraded DNA that is often found in herbarium specimens (see Chapter 4). Observing variation within those anonymous loci among studied entities, can be seen as an approximation of the whole genomic variation among them, which in turn allows scientists to discover genealogy discontinuities indicating the presence of separately evolving units, i.e., species.

The four sources of data, namely morphology, abiotic niches, geographical distribution, and genetic variation, introduced in this section are only a fraction of potential evidence that may be used for species delimitation, which may, for example, include flower ecology or biotic niches. This fraction, however, represents a set of readily available data that can be extracted from preserved specimens, opening the possibility of taxonomic studies on large scale, for example the assessment of species delimitation of whole genera, without the need of collecting all data de-novo. This data can subsequently, in a species

delimitation pipeline, be used for grouping entities before the eventual ranking of these entities.

## The Need for a Species Delimitation Pipeline

The delimitation of species in the context of integrative taxonomy requires the availability of a wide range of data modes and should be based on an objective operational framework for collecting and analyzing those sources of evidence, and eventually drawing taxonomic conclusion from them. Even though such frameworks have been proposed (e.g., Schlick-Steiner et al., 2010), there is still no consensus on how to implement this integration and what kind of evidence is to be preferred for species delimitation studies. What is clear, is that scientists should strive for a taxonomy that is as objective as possible and well-supported (Schlick-Steiner et al., 2014; Daglio & Dawson, 2019). Schlick-Steiner et al. (2010), for example, proposed a semi-formal species delimitation protocol in which different "disciplines" (sources of evidence) are individually studied, and subsequently manually integrated while resolving disagreements among the disciplines according to evolutionary explanations.

Developing an objective framework or pipeline for species delimitation, applicable for a wide range of taxa, even within the plant kingdom, would be, arguably, infeasible due to the highly diverse nature of speciation and reproductive modes, including hybridization and the formation if polyploids (cf. Hörandl, 2022). Implementing an operational sequence for delineating species within a predefined group, e.g., a genus, on the other hand, should be a worthwhile endeavor as such a pipeline can give researchers a toolkit for collecting and analyzing evidence for integrative species delimitation (e.g., Schlick-Steiner et al., 2010). Finally, a bioinformatical pipeline will also increase the reproducibility and traceability of species delimitation studies, by providing a common methodological and technical framework.

## The Genus *Leucanthemum* Mill.

*Leucanthemum* Mill. (Compositae, Anthemideae; "Ox-eye daisy") is a flowering plant genus comprising 38 (see chapters 4 and 5) to 42 species (*The Euro+Med Plantbase Project*) forming a closely-knit polyploid complex with ploidy levels ranging from diploid ($2x$) to docosaploid ($22x$). Its distribution range covers the whole European continent, with some species reaching North-East Asia (*L. ircutianum* DC.) and the temperate regions of

the Northern and Southern hemisphere (Meusel & Jäger, 1992). Its center of species diversity are the Iberian and Apennine peninsulas, as well as central Europe (Marchi, 1982; Vogt, 1991, 2019). *Leucanthemum* is a strongly hybridizing (Wagner et al., 2017, 2019) and relatively recently diversified (1.14-2.94 Ma; Wagner et al., 2019) genus. The species description within the genus is traditionally based on morphological and karyological features, and to a lesser degree on the geographical distribution of the taxa (Vogt, 1991). The leaf shape, in particular, is one of the most important criteria for delineating *Leucanthemum* taxa (Vogt, 1991, 2019).

Recent studies based on single-locus or AFLP markers revealed a genetic substructure within the diploid *Leucanthemum* species, separating an older clade comprising genetically well-separated taxa and a younger clade with a reticulate genetic structure (Konowalik et al., 2015; Wagner et al., 2019). In addition, molecular genetic analyses in morphologically opaque groups within the genus provided new insights into the phylogeny and allowed to resolve taxonomic uncertainties (Greiner et al., 2013; Oberprieler et al., 2014; Wagner et al., 2017), and suggested an allopolyploid origin of several tetraploid taxa (Oberprieler et al., 2014, 2022). *Leucanthemum* have proven to provide an interesting system for studying hybridization and reticulate evolution, and polyploid speciation in recent and potentially ongoing diversification (Greiner & Oberprieler, 2012; Greiner et al., 2013; Oberprieler et al., 2011, 2014, 2018; Konowalik et al., 2015).

## Thesis Outline

The present study proposes a set of tools and methods, a bioinformatical pipeline, to collect and analyze data from several different sources of evidence, i.e. genealogy, morphology, geographical distribution, and abiotic niches, for delimiting the diploid and tetraploid *Leucanthemum* species in the framework of integrative taxonomy. Chapter 2 deals with the implementation of a deep learning-based object detection tool for extracting morphologically relevant structures, e.g., leaves, from digitized herbarium specimens. Chapter 3 describes a new version of this tool, providing additional features for data pre- and postprocessing and extending it for more general object detection and segmentation use cases. Chapter 4 proposes a framework for integrative species delimitation in the diploid *Leucanthemum* representatives based on RADseq, leaf morphometry, abiotic niche modeling, and geographical distribution. Chapter 5 refines this framework for delimiting the tetraploid *Leucanthemum* taxa.

The first study, chapter 2, describes GINJINN, a software tool for detecting and extracting structures, e.g., leaves or flowers, from digital images of preserved plant specimens. GinJinn uses deep learning-based object detection techniques for bounding-box detection, allowing it to be adapted to a variety of different datasets and detection objectives by providing the corresponding training data. Such a bounding-box detection is a first step for automating the collection of morphological data from herbarium specimens for morphometric studies. The tool's performance is demonstrated by successfully extracting intact leaves, which in turn may be used for species delimitation studies, from digitized herbarium specimens of the diploid *L. vulgare* and the tetraploid *L. ircutianum*.

Chapter 3 describes a new, improved version of GINJINN, namely GINJINN2. GINJINN2 provides new data pre- and postprocessing, as well as visualization und evaluation functionality. In addition to the bounding box detection of GinJinn, this new version also features deep learning-based segmentation capabilities. While the first version was meant to be used with preserved herbarium specimens, this new version is applicable to a wider range of objectives, such as the herbivore detection, seed mixture quality evaluation, and stomata detection, which are presented in this chapter.

The last two chapters present the integrative species delimitation of the diploid and tetraploid *Leucanthemum* taxa, respectively. Chapter 4 lays the foundation of a methodological pipeline for species delineation based on genealogical, morphological, ecological, and geographical evidence. In this context, new methods for the reconstruction of damaged leaves that were extracted from herbarium specimens, and for approximating the true geographic range from incomplete sampling data are proposed and implemented. The novel methods are combined with well-established techniques for sequence-based species delimitation, clustering, and ecological niche modeling, and are evaluated in the context of von Wettstein's (1898) and De Queiroz' (2007) species concepts. Chapter 5, building on the framework proposed in chapter 4 and extending it with the novel application of a semi-supervised clustering method for genealogic species delimitation, describes the delineation of the tetraploid *Leucanthemum* species.

# Chapter 2: GinJinn

## GinJinn: An object-detection pipeline for automated feature extraction from herbarium specimens

Tankred Ott, Christoph Palm, Robert Vogt, and Christoph Oberprieler

Published in Applications in Plant Sciences, 2020, 8(6)

### Abstract

**Premise**: The generation of morphological data in evolutionary, taxonomic, and ecological studies of plants using herbarium material has traditionally been a labor-intensive task. Recent progress in machine learning using deep artificial neural networks (deep learning) for image classification and object detection has facilitated the establishment of a pipeline for the automatic recognition and extraction of relevant structures in images of herbarium specimens.

**Methods and Results**: We implemented an extendable pipeline based on state-of-the-art deep-learning object-detection methods to collect leaf images from herbarium specimens of two species of the genus *Leucanthemum*. Using 183 specimens as the training data set, our pipeline extracted one or more intact leaves in 95% of the 61 test images.

**Conclusions**: We establish GINJINN as a deep-learning object-detection tool for the automatic recognition and extraction of individual leaves or other structures from herbarium specimens. Our pipeline offers greater flexibility and a lower entrance barrier than previous image-processing approaches based on hand-crafted features.

## Introduction

Herbarium collections represent a rich treasure of plant specimens from around the world, providing the raw material for evolutionary, taxonomic, and ecological research. The increasing digitization of these natural history collections and their free availability allow scientists to tap into this treasure for systematic, historical, and phenological studies. The Global Biodiversity Information Facility (https://www.gbif.org) alone references herbaria containing over 30 million digitized plant specimens. Until recently, this source of data remained largely untouched due to the amount of manual labor required for the analysis of herbarium photographs.

Modern image-processing methods, however, allow scientists to automate the analysis of digitized herbarium specimens (Corney et al., 2012a, b). In the past few years, progress in machine learning, especially the development of convolutional neural networks (CNNs), has made it possible to automatically identify the genus or species of herbarium specimens (Unger et al., 2016; Carranza-Rojas et al., 2017), or even automatically extract qualitative information like leaf arrangement, form, and structure from digital images of preserved plants (Younis et al., 2018). Recently, Lorieul et al. (2019) showed that machine learning–based image classification can be used to detect the phenological state of herbarium specimens. An area of machine learning that has only very recently gained traction within the plant science community is the explicit object detection of plant structures such as leaves, flowers, or fruits in preserved specimens (Goëau et al., 2020; White et al., 2020).

Here, we introduce GINJINN, an object-detection pipeline based on the TensorFlow (Abadi et al., 2016) object-detection application programming interface (API) designed to make supervised deep-learning object detection accessible for plant scientists. Its name relates to the "magical" [Jinn] extraction of herb(arium specimen)s [Gin] to detect morphological features/structures. GINJINN streamlines the process of moving from annotated images to a trained object-detection model that can be exported and used for the automatic extraction of relevant structures of interest from newly acquired images of a particular study group. Thus, GINJINN allows scientists with little or no prior knowledge of machine learning to apply modern visual-recognition tools and to incorporate object detection into their workflow by automatizing data-mining processes that were previously largely manually performed.

We provide an automatic setup of projects for 47 different bounding-box-based object-detection architectures together with the automatic download of pretrained models for 17 of them. While simplifying the process of model training and deployment, GINJINN still exposes the raw TensorFlow object-detection API configuration files, which gives advanced users full access to all the architectural, preprocessing, and augmentation options provided by TensorFlow.

To show the efficacy of our pipeline, we used GINJINN to train and evaluate a model for the extraction of intact leaves from digitized herbarium specimens. From a technical point of view, the automatic extraction of leaves is an interesting problem, as software is already available for the automatic morphometry of isolated leaves (Corney et al., 2012b; Bonhomme et al., 2014; Chuanromanee et al., 2019) but the process of isolating the leaves themselves is not yet fully automatized (Corney et al., 2012a), especially not using modern machine learning techniques. From a biological point of view, leaf morphometry is an important tool for species delimitation and recognition, as well as for the reconstruction of historical climate conditions (Royer et al., 2005, 2008).

Here, we use two closely related *Leucanthemum* Mill. (Compositae, Anthemideae) species with different ploidy levels, namely the diploid *L. vulgare* Lam. and the tetraploid *L. ircutianum* DC., to demonstrate the application of leaf detection and extraction in a herbaceous plant group.

## Methods and Results

### Software

GINJINN was originally developed as an internal tool for rapid iteration through deep-learning model architectures to find adequate neural network models for the detection and extraction of intact leaves in digital images of herbarium specimens for subsequent morphometric analyses. It has since evolved into a general object-detection pipeline for the setup, training, evaluation, and deployment of bounding-box-based object-detection models with a focus on providing easy access to a high number of different model architectures with little manual work for the user, including the automated download of pretrained models if available. With GINJINN, we provide plant scientists a tool for applying modern machine learning–based visual recognition to their own data sets without requiring a thorough theoretical background in machine learning and proficiency in programming, which is generally necessary to apply and deploy deep-learning object detection.

GINJINN is a Python 3 command-line application for the management, training, and application of object-detection models. In addition to the pipeline application, GINJINN contains several helper scripts that can be used separately from the main command line tool. GINJINN makes use of the free, open-source deep-learning framework TensorFlow (Abadi et al., 2016). Specifically, we are using the TensorFlow object-detection API to access highly optimized training and evaluation pipelines and modern neural network architectures. The object-detection models supported by GINJINN are bounding-box prediction models; segmentation models are not yet implemented. This means that, based on sufficient training data where representative instances of the objects of interest are annotated with encompassing bounding boxes, the CNN learns to recreate those bounding boxes on the training data, and is also able to transfer the learned image-to-bounding-box transformation to newly acquired, similar data (Girshick et al., 2014; Liu et al., 2016; see O'Shea & Nash [2015] for an introduction to CNNs). This allows the automatic recognition of structures of interest after the training of the neural network. In the context of herbarium specimens, those structures might be, for example, fruits, flowers, leaves, buds, or herbivore damage patterns. Structures extracted by GINJINN may be subsequently subjected to different downstream analyses aiming to quantify their shape, color, or texture; count different structure classes (number of buds vs. number of flowers vs. number of fruits in phenological studies); or quantify their positions relative to each other on the surveyed herbarium specimen (coordinates of members of a predefined structure class).

The two different meta-architectures of bounding-box prediction models that are supported by GINJINN are Regions with CNNs (R-CNNs; Girshick et al., 2014) and Single Shot Multibox Detectors (SSDs; Liu et al., 2016). R-CNNs basically employ a two-step procedure of first predicting regions of interest, so-called region proposals, and subsequently classifying the regions of interest (Girshick et al., 2014; Girshick, 2015; Ren et al., 2015). In contrast, SSDs combine both steps in a single neural network architecture (Liu et al., 2016). While SSDs are more modern and allow faster prediction of bounding boxes, a recent benchmarking study by Zhao et al., (2019) showed that R-CNNs achieve similar or better accuracies. The even more recently developed class of bounding-box prediction models, You Only Look Once (YOLO) (Redmon et al., 2016), is intentionally not supported by GINJINN, because these models focus on prediction speed by sacrificing accuracy (Zhao et al., 2019), which is not necessary for the extraction of structures from static images of preserved plants.

Although GINJINN makes heavy use of the TensorFlow object- detection API, it is not merely a wrapper to ease the use of the API. GINJINN provides additional tools for data preprocessing, setting up a standardized project structure, downloading pretrained models (if available), simple model exporting, and using the trained network for the extraction of bounding boxes from newly acquired data, which is a functionality not supported out-of-the-box by the TensorFlow object-detection API. While providing this additional functionality, we ensured that the intermediary and output files were kept compatible with TensorFlow to allow advanced users to seamlessly access the more advanced features of the TensorFlow object-detection API without having to leave the framework provided by GINJINN. Hence, for users who are new to the field of machine learning–based object detection, the pipeline can act as a gentle introduction and allow them to iteratively try out more advanced functionalities of modern deep-learning object detection. Additionally, the interoperability with TensorFlow allows GINJINN users to monitor the training and evaluation of their models live with the TensorBoard (Abadi et al., 2016) tool.

The GINJINN pipeline consists of six steps (Figure 2.1): (1) The generation of a project directory including a project configuration template file. The project configuration file is the place for the user to set data paths and select parameters for the subsequent pipeline steps. (2) The conversion of the data to an internal format and splitting of the data into training and test data sets. (3) Model preparation, which includes the setup of the model and the automatic download of pretrained models (if desired and available). Additionally, this step generates the TensorFlow model configuration file. Advanced users can modify this file to influence image preprocessing, as well as the training and evaluation of the model. (4) Simultaneous model training and evaluation. Model checkpoints are automatically saved. During this step, progress can be monitored via TensorBoard or the console output. (5) Model export, in which the user can select one saved model checkpoint for export. (6) The use of exported models for the extraction of structures from newly acquired images via an additional GINJINN command.

GINJINN accepts JPEG (.jpg, .jpeg) and PNG (.png) images with corresponding annotations in PASCAL Visual Object Classes Challenge (VOC; Everingham et al., 2010) XML format or VGG Image Annotator (Dutta & Zisserman, 2019) JSON format for the training and evaluation of the CNNs. PNG images can be supplied only without the alpha channel (transparency). Like the exported models, the intermediary outputs are also compatible with standard TensorFlow object-detection workflows. The prediction of bounding boxes on newly obtained image data can be performed based on both JPEG and

PNG formats. The output of the prediction is available as class-wise images with bounding boxes for visual inspection, cropped bounding boxes, or bounding-box coordinates in CSV (.csv) format. The output image formats are PNG or JPEG, depending on the format of the respective input images.

We have tested GINJINN on Windows 10 (Microsoft Corporation, Redmond, Washington, USA), Debian (https://www.debian.org/), and Ubuntu (https://ubuntu.com/). The pipeline requires an installation of Python 3.6 (Van Rossum & Drake, 2009) and a corresponding TensorFlow or TensorFlow–graphics processing unit (GPU) version. The latter version is recommended due to the speedup in training, evaluation, and inference time compared to the CPU version, but requires the installation of proprietary NVIDIA GPU drivers and toolkits. Apart from the computation time, both versions are equivalent. Detailed installation instructions can be found in the manual. GINJINN has been released open source under the MIT license. The source code, including the installation instructions, is available at https://github.com/AGOberprieler/ginjinn.

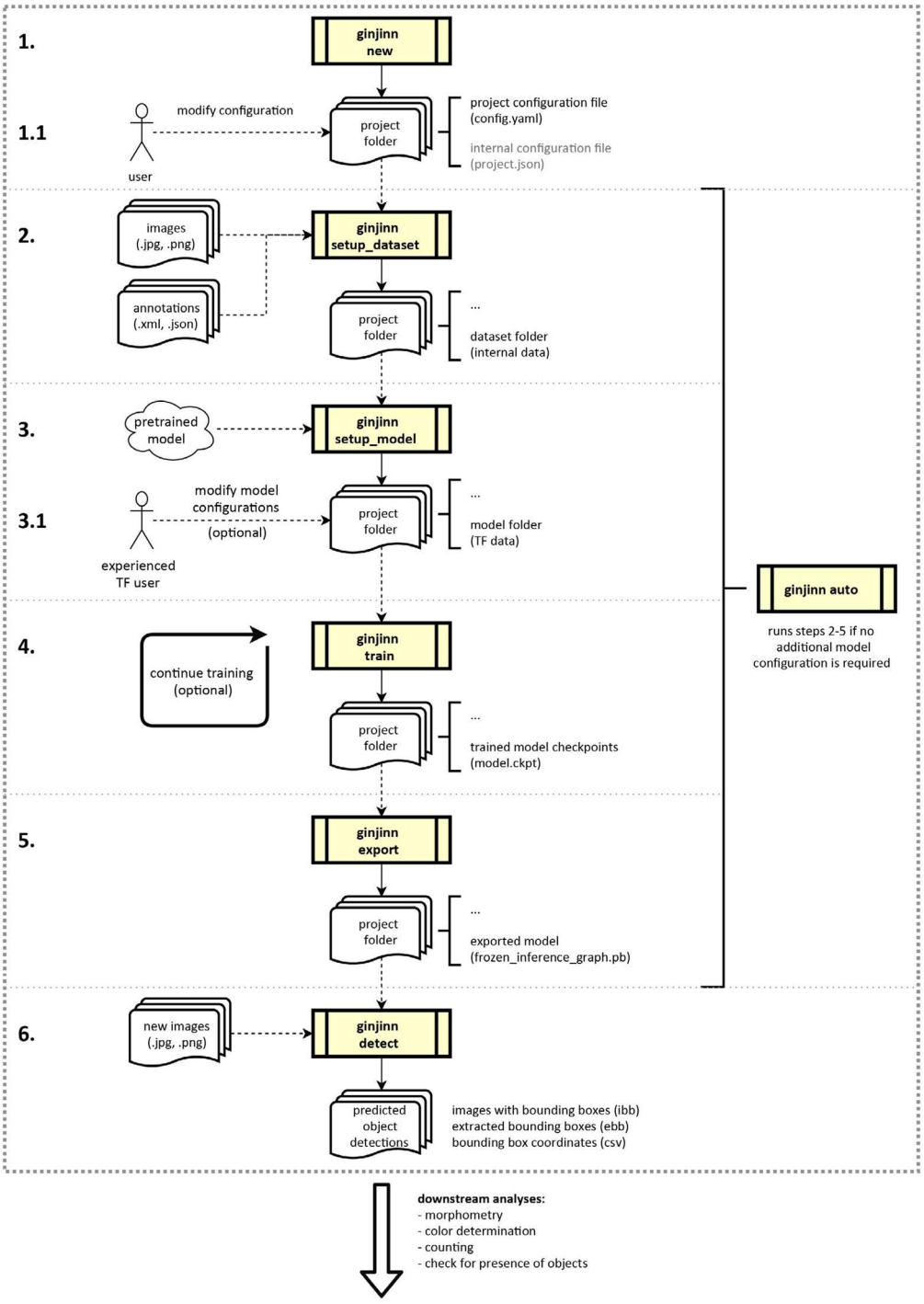


**Figure 2.1**. Flow diagram of the six GinJinn pipeline steps. A project folder is generated using ginjinn new (1) and the configuration file is modified depending on the user's needs (1.1). The preparation (2), processing (3), training (4), and export (5) steps are executed sequentially with specific GinJinn commands (setup_dataset, setup_model, train, and export, respectively), or alternatively at once with the single ginjinn auto command. When not using ginjinn auto, the user can modify intermediary TensorFlow configuration files (3.1) for additional control over the model parameters and augmentation options. The trained and exported model can be used for inference of bounding boxes on new data using ginjinn detect. GinJinn commands are indicated by the yellow process boxes. Data inputs and outputs are illustrated with solid and dashed arrows, respectively. After bounding box detection, the extracted structures of interest can be supplied to other tools for downstream analyses.

**Example application: *Leucanthemum* leaves**

As an example of the application of GINJINN, we present the recognition and extraction of intact leaves from digital images of preserved herbarium specimens of two species of *Leucanthemum* (ox-eye daisies; Compositae, Anthemideae) with different ploidy levels, namely the diploid *L. vulgare* and the tetraploid *L. ircutianum*. One important morphological character for the differentiation of those two species is the shape of the basal and middle cauline leaves (Wagenitz, 1977; Vogt, 1991).

The automated recognition of intact leaves on herbarium specimens, especially for plants with a high variability in leaf shape—as is the case for *Leucanthemum*—can be considered a complex task, because the occurrence of intact leaves in relation to the occurrence of non-intact leaves is rare. This high abundance of damaged leaves is caused by factors such as herbivore damage, shearing while handling the vouchers, and most prominently by the dry pressing process, where leaves are often unintendedly folded. Additionally, the difference between damaged and intact leaves can be very small, making it hard to clearly differentiate between the two cases, even for human observers. We defined as intact those leaves that were completely visible, non-overlapping, non-folded, and not damaged by herbivores. For cauline leaves, special care was taken to ensure that the leaf base was visible, as this is an important character for the distinction between the two *Leucanthemum* species (Wagenitz, 1977; Vogt, 1991). Damaged, overlapping, or folded leaves were not annotated. Accordingly, the detection of intact leaves in this study is posed as a single-class bounding-box detection problem. We have refrained from the subsequent downstream analyses of extracted structures (here: leaves) because these analyses of taxonomically relevant features such as the leaf outline, degree of dissection, color, or texture could be easily accomplished with existing software such as MASS (Chuanromanee et al., 2019) or Momocs (Bonhomme et al., 2014).

For the present example application of GINJINN, we used a data set consisting of 286 JPEG images of preserved plant herbarium specimens provided by the herbarium of the Botanic Garden and Botanical Museum Berlin-Dahlem (B), Berlin, Germany. The images were annotated using the free open-source tool LabelImg version 1.8.1 (https://github.com/tzuta lin/labelImg), resulting in a total of 889 annotated intact leaves in 243 images of herbarium specimens. For the 43 remaining images, no intact leaves were present. GINJINN was used to split the data into training and test data sets for model evaluation by randomly sampling 25% of the images into the test data set.

A model architecture consisting of a Faster R-CNN (Ren et al., 2015) meta-architecture and Inception-ResNet version 2 (Szegedy et al., 2016) as the feature extractor was selected. To speed up the training process, we applied so-called transfer learning by starting the training from a model that was pretrained on the Common Objects in Context (COCO) data set (T.-Y. Lin et al., 2014) provided by the TensorFlow object-detection API. The model was trained for 12,000 generations with a batch size of 1.

The evaluation was performed according to the PASCAL VOC challenge evaluation metrics (Everingham et al., 2010). The model achieved a mean average precision (mAP) of 0.49 at an intersection over union (IoU) of 50%. We were able to successfully detect the presence of one or more intact leaves in 95% of the 61 test images, for which the presence of an intact leaf was manually determined a priori. Figure 2.2A shows the resulting predicted bounding boxes on a test image, with Figure 2.2B and Figure 2.2C depicting true positive and false positive leaf detection, respectively. The results shown in Figure 2.2B could subsequently be used for morphometric analyses with tools such as MASS (Chuanromanee et al., 2019) or Momocs (Bonhomme et al., 2014), for example. Our results indicate the applicability of training a deep-learning model for the detection of objects in preserved plant specimens that can potentially assist or even automatize the extraction of leaves from herbarium images or assist further annotations, even with a relatively small data set of only 243 images. The software manual hosted at the GINJINN GitHub repository contains a dedicated section for the reproduction of these results. This section should also be considered a tutorial for new users.


**Figure 2.2**. GINJINN outputs. (**A**) Output type 'ibb' (image with bounding boxes) showing class-wise predicted bounding boxes of leaves with a score of 0.5 or higher drawn on the original image of an herbarium specimen. The score can be interpreted as a probability that the content of the bounding box belongs to a certain object class (in this case, a leaf). (**B**) Output type 'ebb' (extracted bounding boxes with a padding of 25 pixels) for selected true positive examples of the detected leaves shown in A. (**C**) Output type 'ebb' for selected false positive examples of the leaves shown in A.

## Conclusions

With GINJINN, we are introducing a new software tool that allows plant scientists to tap into modern deep-learning-based visual recognition for the exploration and exploitation of the rich treasure that digitized herbarium specimens in collections all over the globe represent. Here, we have shown that our pipeline is able to automatically extract intact leaves from herbarium specimen images for subsequent downstream analyses. This provides the potential to speed up and automatize previously work-intensive manual workflows, and can grant scientists access to huge amounts of morphological data for morphometric and phenological studies using herbarium specimens.

Previous work in the area of visual recognition of preserved plant materials used hand-crafted features for trait and structure extraction (Corney et al., 2012a; Henries &

Tashakkori, 2012; Unger et al., 2016), focused on classification instead of object detection (Jin et al., 2015; Munisami et al., 2015; Carranza-Rojas et al., 2017; Younis et al., 2018; Lorieul et al., 2019), or tried to directly extract characters from images without using explicit object-detection techniques (Ubbens & Stavness, 2017; Younis et al., 2018). GINJINN, in contrast, is a tool specifically developed for the extraction of structures such as leaves, flowers, buds, or fruits from digitized herbarium specimens. As such, GINJINN can be used to generate inputs for downstream analyses with existing tools, for example, the recently released MASS (Chuanromanee et al., 2019) software for morphometric analyses. A tool somewhat similar to GINJINN, LeafMachine (Weaver et al., 2020), is also newly available for the extraction of leaves from digital images of preserved plants. Whereas LeafMachine is designed to extract leaves via semantic segmentation, our pipeline can be used to extract instances of any kind of structure that it is trained for via bounding-box object detection. Furthermore, all dependencies of GINJINN are free and open source, while LEAFMACHINE depends on the proprietary MATLAB (MathWorks, Natick, Massachusetts, USA) environment. However, if a pixel-wise segmentation is required instead of cropped leaves, LeafMachine might be the better- suited tool. Another possibility would be to use both tools: GINJINN to first reduce the complexity of the problem via bounding-box cropping and a subsequent pixel-perfect extraction of leaf silhouettes based on the cropped leaves using LEAFMACHINE.

When compared to the usage of the TensorFlow object-detection API directly, our pipeline adds the additional features of project setup, data preparation, automatic download of pretrained models, and an easy-to-use inference routine with outputs fitted to the plant science community. Additionally, GINJINN can be used by scientists without proficiency in Python programming and generally does not require any knowledge about the architecture of TensorFlow and the TensorFlow object-detection API.

By designing the pipeline with ease of use in mind, it was necessary to reduce the feature set that is presented to the user when compared to TensorFlow. This drawback is partly mitigated by exposing the raw TensorFlow configuration and the run and export scripts in the project folders generated by GINJINN in such a way that advanced users can modify those files directly without leaving the framework of the pipeline.

As a machine learning–based tool, the performance of object- detection models trained using GINJINN is highly dependent on the quality of the available training data. This limitation, however, applies to all machine learning–based modeling. Care must be taken to ensure the data used for training the object-detection models resemble the expected test

data. A problem, for example, is strong differences in lighting conditions or the angle from which images were taken between the training and test data. This limitation is partially mitigated using the built-in augmentation options, which introduce small perturbations into the training images to make the model more resistant against that type of variability.

A temporary technical restriction is that only bounding-box prediction models are available, even though models for semantic and instance segmentation are also supported by the TensorFlow object-detection API. However, in future versions, those segmentation models will be made available through the GINJINN interface. Another goal for the next version of the application is to provide the configuration of additional data augmentation options. The long-term aim is the integration of the PyTorch (Paszke et al., 2017) framework as an alternative to TensorFlow, which would introduce a higher amount of available architectures and an easier setup of GPU acceleration for GINJINN.

We present GINJINN as a deep-learning object-detection tool for the automatic recognition and extraction of structures such as leaves or flowers from herbarium specimens. We showed that GINJINN can be applied to successfully extract intact leaves from images of herbarized *Leucanthemum* individuals, while offering greater flexibility and a lower barrier to entry compared with previous image-processing approaches based on handcrafted features.

## Acknowledgements

This work benefited from the collaboration of C.O. (OB 155/13-1) and R.V. (VO 1595/3-1) within the Deutsche Forschungsgemeinschaft (DFG) priority program SPP 1991 "Taxon-omics – New Approaches for Discovering and Naming Biodiversity" through the project "Setting-up a methodological pipeline for species delimitation and species network reconstruction in polyploid complexes." The comments of two reviewers improved an earlier version of this contribution considerably and are thankfully acknowledged.

## Author Contributions

T.O., C.P., and C.O. conceived the present study. R.V. produced the scans of the preserved specimens. T.O. programmed the software. C.P. consulted the exemplary data analysis. A first draft of the paper was written by T.O. with input from C.P., R.V., and C.O.

## Data Availability

The source code and manual are hosted at https://github.com/AGOberprieler/ginjinn. The annotated image data that were used for evaluation of the method are hosted by the German Federation for Biological Data (GFBio; https://data.bgbm.org/datas et/gfbio/0033/).

# Chapter 3: GinJinn2

## GinJinn2: Object detection and segmentation for ecology and evolution

Tankred Ott, Ulrich Lautenschlager

Published in Methods in Ecology and Evolution, 2022, 13(3)

### Abstract

1. Collection and preparation of empirical data still represent one of the most important, but also expensive steps in ecological and evolutionary/systematic research. Modern machine learning approaches, however, have the potential to automate a variety of tasks, which until recently could only be performed manually. Unfortunately, the application of such methods by researchers outside the field is hampered by technical difficulties.
2. Here, we present GINJINN2, a user-friendly toolbox for deep learning-based object detection and instance segmentation on image data. Besides providing a convenient command-line interface to existing software libraries, it comprises several additional tools for data handling, pre- and postprocessing, and building advanced analysis pipelines.
3. We demonstrate the application of GINJINN2 for biological purposes using four exemplary analyses, namely the evaluation of seed mixtures, detection of insects on glue traps, segmentation of stomata and extraction of leaf silhouettes from herbarium specimens.
4. GINJINN2, by providing a coding-free environment, will enable users with a primary background in biology to apply deep learning-based methods for object detection and segmentation in order to automate feature extraction from image data.

## Introduction

Leveraging image data for ecological and evolutionary/systematic research typically requires substantial effort for data collection and preparation. The ability to automate time-consuming steps of this process, possibly along with further downstream analyses, for example, using programming languages like Python or R, can not only increase productivity, but also allow otherwise infeasible large-scale analyses. Recent advances in machine learning (ML), both on the soft- and hardware side, make it even possible to automate tasks that are difficult to solve by means of classically designed algorithms. Computer vision, in particular, has largely profited from deep learning, which increasingly influences even the more traditional branches of organismic biology. Species identification tools running on smartphone devices (for an overview, see Jones, 2020; Wäldchen & Mäder, 2018) are prominent examples for this trend. Beyond pure classification tasks, a technically even more challenging problem consists in localizing objects like cells, organs, or individuals on images. Specialized tools address this problem for various areas of application, such as crop or weed detection (e.g. Buddha et al., 2019; Afonso et al., 2020), detection of leaves and other plant organs on herbarium specimens (e.g. Ott et al., 2020; Weaver et al., 2020; Younis et al., 2020), stomata counting using microscopic leaf images (e.g. Fetter et al., 2019), animal counting using camera traps (Norouzzadeh et al., 2021) and many more. Moreover, DEEPIMAGEJ (Gómez-de-Mariscal et al., 2021), an optional plugin for the popular IMAGEJ program (Schneider et al., 2012; Schroeder et al., 2021), provides easy access to a number of trained deep learning models for pre-defined tasks via a graphical user interface.

Despite the availability of increasingly convenient frameworks, adapting well-established ML methods to new areas of application typically requires an amount of technical knowledge that may discourage potential users. GINJINN2, whose core functionality is based on Detectron2 (Wu et al., 2019), aims at lowering this hurdle by providing an easy-to-use command-line interface to the latter, augmented by a number of utility functions, designed to help the user with building custom analysis pipelines. While GINJINN (Ott et al., 2020) focused on extracting leaves from digitized herbarium specimens, GinJinn2 aims at a wider scope of application. Unlike the former, which was based on the Tensorflow object detection API, it is not restricted to bounding-box object detection, but also incorporates functionality for instance segmentation, that is, pixel-precise detection and classification of individual objects.

In the present contribution, a number of example analyses demonstrate how ecological, agricultural or evolutionary/systematic studies may benefit from GINJINN2. Those include pest monitoring using yellow glue traps, leaf shape extraction from herbarium specimens, stomata segmentation and the evaluation of seed mixtures. We hope to encourage interested researchers to consider deep learning-based object detection or segmentation when faced with similar tasks. Using GINJINN2 together with pretrained models from Detectron2's model zoo, new applications can be explored with a minimum of invested time and effort, which makes it a potentially useful tool for both beginners and advanced users.

## Software

### Overview

GINJINN2 is a toolbox for deep learning-based bounding-box object detection and instance segmentation. As such, it provides functionality for model training, evaluation and application based on the Detectron2 framework, segmentation refinement based on CascadePSP (Cheng et al., 2020), a set of data pre- and postprocessing tools for handling annotated image datasets, and capabilities for data insight and visualization. GINJINN2 is not meant to be a replacement for existing frameworks like Detectron2 or the Tensorflow Object Detection API (Huang et al., 2017), but rather a toolkit enabling code-free access to deep learning-based object detection technologies. All of GINJINN2's functionality is accessible via an easy-to-use command-line interface (CLI).

### Dataset splitting

Besides the data used to train the model, it is generally advisable to use a so-called validation dataset in order to detect overfitting and to optimize model choice and training parameters. Using a separate dataset for those purposes is necessary because the model's fit to the training data does not provide information about its generalization capability. In other words, a trained model may accurately reproduce the training data, but perform poorly on images that have not been presented to it before. However, as soon as any optimizing decision has been made based on the validation data (e.g., when to stop the training process), the model may again show overly optimistic performance for this particular dataset. To obtain an unbiased evaluation of the final model, it is therefore necessary to provide an additional test dataset, which should not have been used for any other task

beforehand. The *ginjinn split* command partitions an input dataset in such a way that each image along with its annotated objects is assigned to one of the resulting subsets. To be representative for the original dataset, each of the latter should comprise similar proportions of objects from each category. Aiming at a high level of homogeneity, the proposed splits are generated by a greedy optimization algorithm (see Supporting Information S3.1). Despite being a relatively rough heuristic, this approach is often sufficient to create acceptable splits and can even be applied to large datasets.

**Object detection and instance segmentation**

GINJINN2, by leveraging Detectron2's model zoo, offers several Faster R-CNN (Ren et al., 2015) and Mask R-CNN (He et al., 2017) models for bounding-box detection and instance segmentation respectively. These are used in a supervised manner, that is, before being able to predict objects on new images in a meaningful way, their parameters ('weights') have to be fitted to images with known object occurrences ('training'). While training such models de novo can be highly GPU intensive, this process can be considerably abbreviated by starting from pretrained rather than randomly initialized weights ('transfer learning'). Accordingly, all available Detectron2 models have already been trained on a large image dataset. Using those pretrained networks reduces the training time for new, custom datasets as well.

Once the user has prepared datasets for training, and, optionally, validation and test (see Dataset splitting), a GINJINN2 project can be initialized using *ginjinn new*. Training models using *ginjinn train* constitutes the computationally most demanding part of a typical GINJINN2 pipeline. This process consists of a prespecified number of iterations, at each of which one or multiple images from the training dataset are presented to the model. The objects predicted by the latter are then compared to the known annotations and the model weights are adjusted to reduce deviations ('loss') from the desired output. While minimizing the loss with respect to the training dataset, at some point, the model's generalization capability may begin to degrade. This so-called overfitting can be recognized by an increasing loss for the validation dataset. The latter is therefore evaluated at predefined intervals. To enable a better assessment of the learning progress, COCO (T.-Y. Lin et al., 2014) evaluation metrics (AP, AP50, AP75, APs, APm and APl; for details see https://cocodataset.org) for the validation dataset are calculated as well. Since the model weights are stored periodically, in case of overfitting, the user can go back to an earlier

checkpoint without having to discard the complete training. Since GINJINN2 is using Detectron2 as modelling backend, all models that are trained in the context of a GinJinn2 project can be used with Detectron2's Python interface without modification.

The quality of the final, trained model is best assessed based on a hitherto unused dataset with known object occurrences. This can be done using *ginjinn evaluate*, which calculates COCO evaluation metrics for the specified test dataset.

The *ginjinn predict* command allows applying a trained model to predict object occurrences for arbitrary images. Instance segmentations can optionally be refined using CascadePSP (Cheng et al., 2020); while slowing down the predictions, this may considerably improve the quality of the object outlines, especially in case of clear object boundaries. To facilitate the further use of the predictions, GINJINN2 provides various output options: (a) visualization of the predictions on the original images, (b) writing a new COCO annotation file and (c) saving a cropped image and, if applicable, segmentation mask for each predicted object.

**Further functionality**

As a counterpart to the already described splitting command (*ginjinn split*), datasets can also be merged (*ginjinn utils merge*), which is particularly useful when using COCO's annotation format. In doing so, the input datasets are also checked for duplicated images.

Object annotations can be filtered by either category or size using *ginjinn utils filter_cat* or *ginjinn utils filter_size* respectively. The latter command is also capable of removing only small disjunct fragments from existing objects.

To simplify existing data, nested image directories can be summarized, making them compatible with GINJINN2 and other tools. *ginjinn utils flatten* recursively collects all images from a given directory and its sub-directories, renames and copies them into a single directory, and modifies associated annotations accordingly.

Due to the limited spatial resolution of common object detection models, detecting or segmenting objects that are small in relation to the image size can be difficult. To mitigate this problem, a sliding window approach can be used to split the original images into smaller sub-images (*ginjinn utils sw_split*), preserving annotated objects, if available. Conversely, predictions based on such fragmented images can be merged again (*ginjinn utils sw_merge*) in order to generate an annotation of the original image.

The *ginjinn utils crop* command creates an annotated sub-image for each annotated object from a given dataset. Similar to the sliding window approach, this can be utilized to increase objects sizes relative to the images. Specifically, performing instance segmentation based on previously cropped bounding boxes may lead to improved results.

The contents of a dataset can be briefly summarized using *ginjinn info*. More detailed information is provided by *ginjinn utils count*, which lists object occurrences individually for each image in a given dataset. Object annotations can be visualized with *ginjinn visualize*, which produces images overlaid by bounding boxes and, if available, segmentation polygons. Moreover, GINJINN2 allows to generate artificial datasets for testing purposes (*ginjinn simulate*).

### Installation and usage

GINJINN2 is implemented in Python3 and can be installed using the CONDA package manager, which also takes care of most of its dependencies. *ginjinn* and all its subcommands provide a help option to list available parameters along with a short description. Further guidelines regarding installation and usage, along with an introductory tutorial and exemplary applications, are provided at https://ginjinn2.readthedocs.io.

## Example Analyses

### Seed counting

In this section, we demonstrate how GINJINN2 can be applied for seed mixture analysis, an illustrative use case for bounding-box detection with subsequent counting. This approach could, for instance, be used to examine commercial seed mixtures or be applied to ecological samples (e.g. from seed traps). The presented analysis is based on a dataset consisting of 284 microscopic images of sand-contaminated seed mixtures of the two plant genera *Sedum* L. and *Arabidopsis* (DC.) Heynh.

For all images, intact seeds were annotated with bounding boxes using the Computer Vision Annotation Tool (CVAT, https://github.com/openvinotoolkit/cvat), resulting in 6,732 and 1,964 annotated seeds for *Arabidopsis* and *Sedum* respectively. The annotated images were exported as COCO dataset, which was then flattened (g*injinn utils flatten*), and split into sub-datasets for training, validation and testing. A Faster R-CNN model was simultaneously trained and validated (Figure 3.1a). The quality of the fit model was assessed using COCO evaluation metrics for bounding-box detection. In addition,

instances predicted for the test dataset were counted (*ginjinn utils count*) and compared with the manually obtained counts.

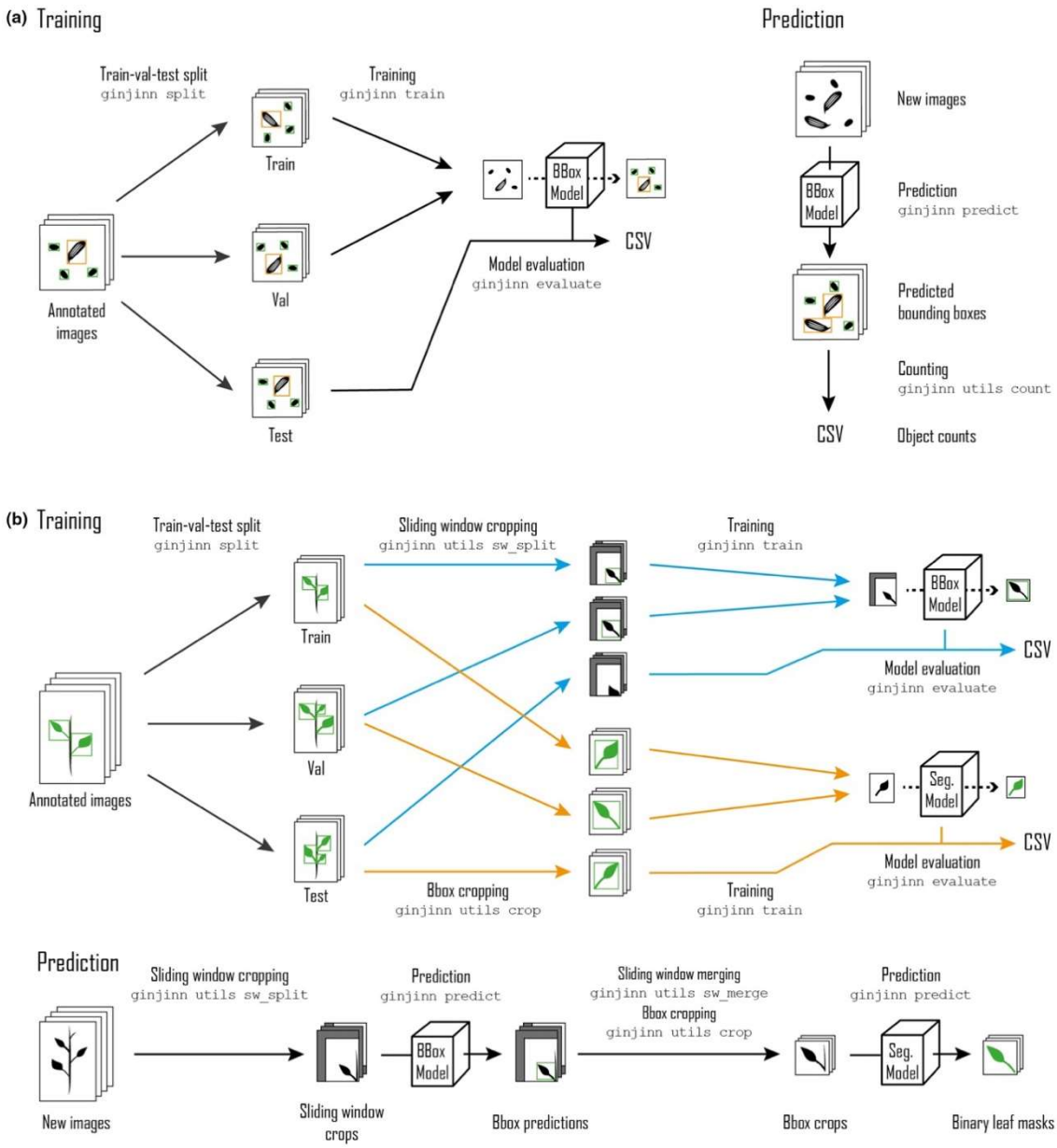


**Figure 3.1**. Seeds (a) and Leucanthemum (b) analysis workflows. The seeds dataset is split into training, validation and test datasets, which are used to train and evaluate a bounding-box model (a, Training). The trained model is applied to new data for seed counting (a, Prediction). The Leucanthemum dataset is also split into training, validation, and test datasets, but the workflow comprises training and evaluation of two separate models (b, Training). The blue branch refers to a bounding-box model for the detection of leaves on sliding window crops of the split dataset. The orange branch depicts the training and evaluation of an instance segmentation model on padded bounding boxes cropped from the split datasets. Leaf segmentations for new data are predicted by combining both models (b, Prediction).

After training, the AP50 was 98.6 and 98.9 for the validation and test dataset, respectively, which indicates that no overfitting occurred. The mean absolute error (MAE) of the class counts for the training dataset was 0.77 for *Arabidopsis* and 0.58 for *Sedum*, meaning that on average, less than a single object per image was misclassified, missed or falsely detected. The MAE of the seed proportions was 0.01, that is, only 1% deviation from the true seed proportions. Exemplary predictions are shown in Figure 3.3a (Supporting Information S3.2).

**Yellow-sticky-traps insect detection and counting**

As an example project for counting small, low-contrast objects on large images, the yellow-sticky-traps dataset (Nieuwenhuizen et al., 2018) was analyzed. This dataset consists of images of yellow glue traps that were placed in greenhouses to monitor insect abundance. Three categories of insects (true bugs) were annotated with bounding boxes: Whitefly (WF), *Macrolophus* (MR) and *Nesidiocoris* (NC).

After removing redundant images and correcting erroneous or missing annotations using CVAT, a cleaned sub-dataset comprising 120 images along with 4,913 bounding-box annotations (WF: 3,660, MR: 1,069, NC: 184) was exported in COCO format. In contrast to the seeds dataset, these bounding-box annotations are of considerably lower quality, often enclosing the insects only loosely.

The cleaned dataset was split into training, validation and test datasets using *ginjinn split*. Since the insects are relatively small compared to the total image size, a sliding window approach was applied (*ginjinn utils sw_split*) to crop sub-images along with corresponding object (sub-)annotations. The cropped datasets were used to train and evaluate a Faster R-CNN model for bounding-box detection. Finally, object instances predicted for the test dataset were counted (*ginjinn untils count*) and compared with true object counts.

The trained model achieved a validation and test AP50 of 90.12 and 92.4 respectively. The mean absolute error (MAE) of the instance counts was 1.67 for WF, 0.21 for NC and 0.79 for MR at an average of 27.1, 1.67 and 7.41 annotated instances per image for the respective object categories. The former amounts to a relative counting error of 6% for WF, 12.5% for NC and 10.6% for MR (weighted average: 7.24%). Exemplary predictions are illustrated in Figure 3.3 (Supporting Information S3.2).

**Stomata segmentation**

To demonstrate basic instance segmentation with the aim of detecting stomata, we applied GINJINN2 to microscopic images of epidermal plant material, retrieved from the Cuticle Database Project (Barclay et al., 2012). Results of such a segmentation can be used in downstream analyses for counting, measuring density or examining size and shape of the stomata.

Using CVAT, 147 images were annotated with 2,314 polygons, each enclosing the guard cells of a stoma. The annotated images were exported as COCO dataset and split into training, validation and test datasets used to train and evaluate a Mask R-CNN model.

The trained model achieved an AP of 49.46 and 51.32 for the validation and test dataset respectively. The mean absolute counting error amounts to 2.34 at an average of 14.69 stomata per image. An exemplary prediction is shown in Figure 3.2a.

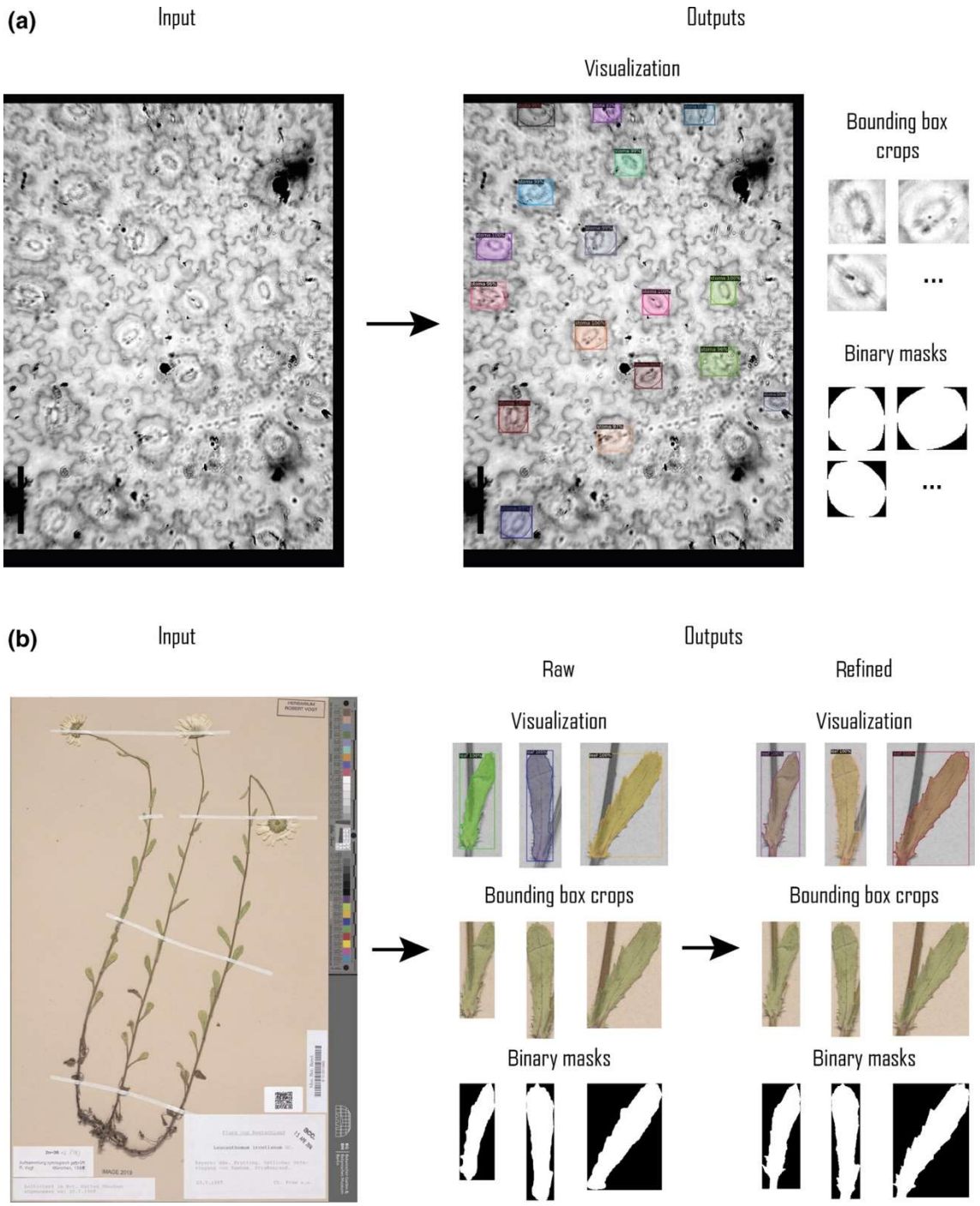


**Figure 3.2**. Exemplary outputs from the Stomata (**a**) and Leucanthemum (**b**) analyses. (**a**) depicts a single input image along with corresponding predictions by the stomata model, showing different output formats. Similarly, (**b**) shows an input image and corresponding predictions for the Leucanthemum pipeline, before and after segmentation refinement.

### *Leucanthemum* leaf segmentation

Morphometric studies often rely on outline data of specific animal or plant organs like, for example, leaves in the latter organism group. A common workflow to generate such data is to manually remove leaves from a living or herborized plant, fixate them on a

contrasting surface, capture digital images and finally apply semi-automatic thresholding methods (e.g. OTSU-thresholding) to construct binary segmentation masks. In this exemplary application of GINJINN2, we show an alternative way to segment individual leaves from digitized herbarium specimens based on a two-step approach involving separate models for bounding-box detection and segmentation.

For this purpose, the Botanic Garden and Botanical Museum Berlin provided us with 303 digitized herbarium specimens from 12 different *Leucanthemum* Mill. (ox-eye daisy) species. Using CVAT, the specimen images were annotated with polygons of the single object category 'leaf'. This category represents largely intact leaves, which are a prerequisite for reliable morphometric analyses. The annotated images, comprising 950 'leaf' instances, were exported from CVAT as COCO dataset, flattened (*ginjinn utils flatten*) and split into training, validation and test datasets.

A two-step pipeline (Figure 2.1b) was applied, consisting of (a) a Faster R-CNN bounding-box detection model that allows to extract individual leaves, and (b) a Mask R-CNN model to segment the leaves on those image parts. The Faster R-CNN was trained and evaluated on sliding window crops (*ginjinn utils sw_split*) of the three datasets. For the Mask R-CNN, sub-images (*ginjinn utils crop*) were cropped from the original annotated images, each containing a single annotated leaf. Based on those cropped datasets, the Mask R-CNN was trained and evaluated. In addition, segmentation refinement was applied to the predictions for the test dataset.

After training, the Faster R-CNN achieved an AP of 30.57 and 25.85 for the validation and test dataset respectively. The Mask R-CNN's AP scores were 76.44 and 74.54. Figure 3.2b illustrates an exemplary prediction. For new image data, the complete prediction process also involves sliding window merging as illustrated in Figure 2.1b in order to remove duplicated objects.

## Discussion

The GINJINN2 toolkit advances the original GinJinn by reimplementing its ideas on the basis of Detectron2, while also introducing new features like segmentation models including mask refinement, as well as several data pre- and postprocessing capabilities.

Based on four exemplary datasets we have shown applications of varying complexity. The seeds and yellow-sticky-traps analyses address multi-category object counting problems using bounding-box detection. We were able to predict the seed ratios

with an absolute error of only 1%, proving the potential of our software for the automation of such counting tasks. Considering the similar problem of counting insects on yellow glue traps, with an error of 7.2%, the accuracy of the trained model may appear less convincing. There are two likely causes for this difference in accuracy: (a) low contrast between objects (insects) and background (glue trap) and (b) low quality of annotations. The latter could easily be solved by a more careful annotation scheme. Nevertheless, the achieved accuracy might be sufficient for practical applications, for example, to measure the response to insecticide treatments or released beneficials in greenhouses.

The stomata analysis serves as a basic example of instance segmentation. Despite several previous works on the automated examination of stomata (Toda et al., 2018; Fetter et al., 2019; Li et al., 2019; Carrasco et al., 2020; Casado-García et al., 2020; Meeus et al., 2020; Song et al., 2020) this contribution, to our knowledge, is the first trying to automatically segment whole stomata (represented by their guard cells) using deep learning. With only 88 highly variable training images, our model achieved an AP of 51.32. Depending on the intended downstream analyses, this precision may already be acceptable if, for instance, only few high-quality object instances are required. Undoubtedly, a model trained on a larger dataset will achieve substantially higher predictive power.

Finally, the *Leucanthemum* analysis illustrates how to construct a pipeline consisting of sliding window-based bounding-box detection and subsequent segmentation for the extraction of high-quality leaf silhouettes from herbarium specimens. Here, the Faster R-CNN achieved an AP of 25.85. For potential morphometric analyses, we are not interested in extracting all leaves, but only largely intact ones, even at the cost of discarding viable instances. Therefore, the relatively low AP is sufficient. The Mask R-CNN, with an AP of 74.54 before refinement, was very successful at segmenting the leaves inside the bounding boxes. This pipeline already allows to generate leaf outlines for downstream analyses like Elliptic Fourier Analysis or Leaf Dissection Index calculation (for an overview of such methods, see McLellan & Endler, 1998) with little manual effort.

With the presented exemplary analyses, we hope to provide guidance for the application of GINJINN2 for automatic data collection and feature extraction. Despite GINJINN2's progress compared to its predecessor, there is still room for further improvements. At the moment, GINJINN2 is only available for Unix-like operating systems with access to an NVidia GPU while Windows support may become available with forthcoming updates to the Windows Subsystem for Linux (WSL). Moreover, there is only one meta-architecture for each of the two detection tasks available, namely Faster R-CNN

and Mask R-CNN. These, however, are among the most successful architectures for general-purpose object detection and segmentation. The integration of additional model architectures may be part of future versions.

We are confident that GINJINN2 will enable users, even those without programming experience, to apply deep learning-based methods for object detection and segmentation as part of their analysis pipelines. Advanced users may utilize GINJINN2 as a tool for rapid prototyping.

## Acknowledgements

First and foremost, we thank Christoph Oberprieler (Regensburg), who acquired the funding, for enabling this work and his comments on the manuscript. We thank Robert Vogt (Berlin) and Sergey Rosbakh (Regensburg) for providing digital images of *Leucanthemum* specimens and seed mixtures respectively. We also thank David Dilcher (Bloomington) for granting us permission to use the microscopic image shown in Figure 3.2a. The help of Maximilian Schall and Sebastian Segieth (both Regensburg), who annotated many of the *Leucanthemum* and seeds images, is much appreciated. We thank Tanja Wenzel for her support in designing the workflow diagrams. We also thank Agnes Scheunert (Regensburg) for her comments on the manuscript. The comments of Sam Church and two anonymous reviewers, who helped to significantly improve this contribution, are much appreciated. This work was supported by a Grant (OB 155/13-1) of the German Research Foundation (DFG) in the frame of the Priority Programme SPP 1991 'Taxon-omics—New Approaches for Discovering and Naming Biodiversity' to Christoph Oberprieler. Open Access funding enabled and organized by Projekt DEAL.

## Conflict of Interest

The authors declare no conflict of interest.

## Supporting Information

### S3.1: Greedy heuristic for dataset splitting

Sub-datasets for training and, optionally, validation and test are composed in a greedy manner. Starting with empty sub-datasets, at each iteration of the algorithm, one image from the input dataset is assigned to one of the sub-datasets to be generated until every image has found its place. The selection process works as follows: At the beginning of

iteration $i$, i.e., after having assigned $i-1$ images, the $j^{\text{th}}$ sub-dataset should include approximately $(i-1)p_j$ images, where $p_j$ is the proportion of images that we wish in this dataset (by default, 0.6, 0.2 and 0.2 for training, validation and test, respectively). The dataset whose current number of images differs the most from $ip_j$ (we will refer to this dataset as $j'$) is then allowed to select the next image from a number of candidates, randomly drawn from the remaining images. Limiting the number of candidates serves to avoid quadratic runtime and causes randomized results. Ideally, after selecting an appropriate image, dataset $j'$ should comprise approx. $n_k N_{j'}^{(i)}/N$ objects from each category $k$. Here, $n_k$ denotes the total number of objects from category $k$, $N$ denotes the total number of images, while $N_j^{(i)}$ is the number of images in the $j^{\text{th}}$ sub-dataset after completing iteration $i$. Among the available candidates, the image whose object counts minimize the sum of squared deviances from $n_k N_{j'}^{(i)}/N$ across each category $k$ is assigned to the current sub-dataset $j'$.

## S3.2: Exemplary prediction from the Seeds and Yellow-sticky-traps analyses

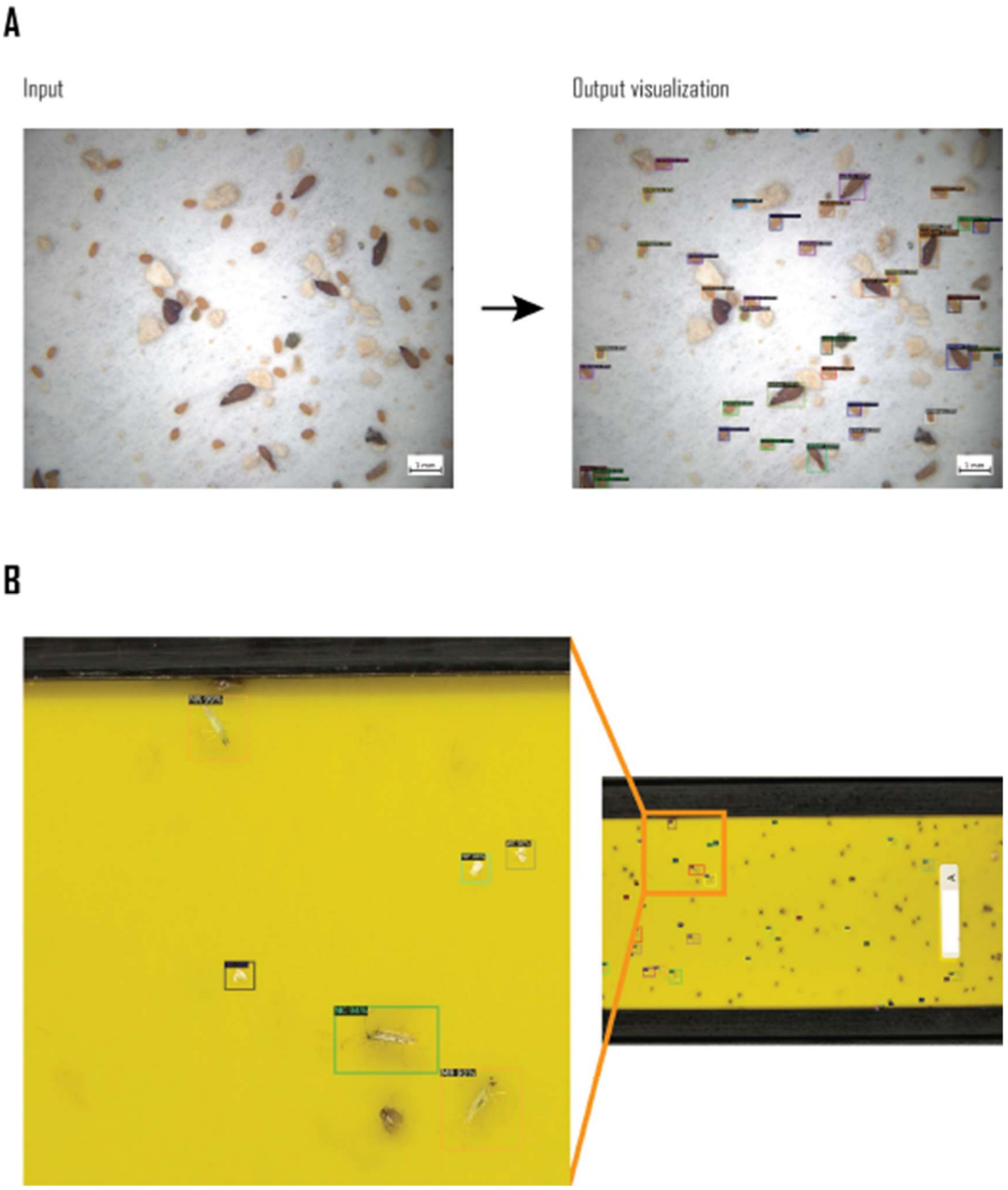


**Figure 3.3**. Exemplary predictions from the Seeds (**A**) and Yellow-sticky-traps (**B**) analyses. **A** shows a single input single input image along with corresponding predictions by the Seeds model. The image contains a mixture of seeds from *Sedum* and *Arabidopsis*, contaminated by sand. **B** illustrates an input image and predictions for the Yellow-sticky-traps pipeline, which makes use of the sliding window approach implemented by GINJINN2. The enlarged section on the left side comprises insects from all three categories, i.e., Whitefly (WF), *Macrolophus* (MR) and *Nesidiocoris* (NC).

# Chapter 4: The Warps and Wefts of a Polyploidy Complex

## The warps and wefts of a polyploidy complex: integrative species delimitation of the diploid *Leucanthemum* (Compositae, Anthemideae) representatives

Tankred Ott, Maximilian Schall, Robert Vogt, Christoph Oberprieler

A modified version of this chapter is published in Plants, 2022, 11(14).

### Abstract

**Background and Aims**: Owing to the paramount role of the species rank in evolutionary, ecological and nature conservation studies, the objective delimitation of species is an essential contribution of taxonomy to biodiversity research. In an 'integrative taxonomy' approach to the delimitation among species on the diploid level we search for the evolutionary significant units (the warps and wefts) that gave rise to the polyploid complex of European ox-eye daisies (*Leucanthemum*; Compositae-Anthemideae).

**Methods**: Species discovery and validation methods based on genetic, ecological, geographical, and morphometric datasets were applied to test the currently accepted morpho-species in diploid *Leucanthemum.* Novel approaches were applied in the analyses of RADseq data (consensus clustering), in morphometrics of reconstructed leaf silhouettes from digitized herbarium specimens, and in quantification species-distribution overlaps.

**Key Results**: We show that 14 of the 20 *Leucanthemum* morpho-species are supported by genetic evidence. The taxonomic rank of the remaining six morpho-species is resolved by combining genealogic, ecologic, geographic, and morphologic data in the framework of von Wettstein´s morpho-geographical species concept.

**Conclusions**: We herewith provide a methodological pipeline for the delimitation of species in an 'integrative taxonomy' fashion using sources of evidence from genealogical, morphological, ecological, and geographical data in the philosophy of De Queiroz´ "Unified Species Concept".

## Introduction

Delimiting species, the fundamental rank of taxonomy, is necessary to facilitate ecological, evolutionary, or nature conservation studies. Of the many species concepts proposed and used in taxonomic studies (Zachos, 2016), however, the majority is not applicable throughout the realm of organismic diversity. One of the major reasons for this is that speciation is a continuous process (Stankowski & Ravinet, 2021) and that the relative importance of the different criteria stressed in the different species concepts are variable throughout the tree of life. As a convincing solution to this problem, de Queiroz (2007) proposed his "Unified Species Concept" circumscribing species as independently evolving metapopulation lineages and that the properties used in other species concepts to define those entities may only be used as indicators for these independently evolving metapopulation lineages. As a consequence, these properties (morphology, physiology, ecology, reproductive isolation, geography, etc.) are not helpful in species conceptualization but can be used in species delimitation, and in their conjoint realization add trustworthiness to a delimitation hypothesis.

This conceptual breakthrough paved the way for revitalizing the "biosystematics" approaches to species delimitation from the 1960s and 70s as "integrative taxonomy" by Dayrat (2005) and Will et al. (2005) who questioned the solely DNA-based approaches to species delimitation championed by others (DNA barcoding; Hebert & Gregory, 2005). Subsequent researchers have conceptionalized this multisource approach to species delimitation and proposed procedural protocols for the joint and/or sequential application of different sources of evidence (Schlick-Steiner et al., 2010, Padial et al., 2010; but see Doyen & Slobodchikoff, 1974 for a comparable approach from the "biosystematics" era) and even tools for the computationally combined analysis of different lineage properties. As examples for the latter, the joint analyses of morphology, genetics, and geography (Geneland; Guillot et al., 2012), morphology and geography ("multivariate normal mixtures and tolerance regions" analysis; Zapata & Jiménez, 2012, Vásquez-Cruz et al., 2017), genealogy and morphology (iBPP; Solís-Lemus et al., 2015), or genetics and geography (regression analysis; Hausdorf & Hennig, 2020) are of interest.

A general feature of most of the mentioned approaches is that congruent datasets are necessary. This requirement is in conflict with the more prevalent situations encountered by taxonomists, in which data for the respective fields of evidence are fragmentary and only partially overlapping. Specimens sampled and genetically analyzed

usually represent a subset of specimens from natural collections available for morphological analyses and/or distribution mapping and ecological niche modelling. As a consequence, an integrative species-delimitation procedure should be based on a methodological pipeline, in which all fields of evidence are represented by datasets as comprehensive as possible, and which follows a reproducible protocol that allows for the integration of all results for a taxonomical decision. A flexible approach based on the procedures proposed by Schlick-Steiner et al. (2010) and Padial et al. (2010) will also allow for easier integration of novel or alternative analytical methods from the different fields of evidence, compared to a strict, computationally fixed pipeline. Here, we will apply such an approach to species delimitation questions in a plant group with a complex taxonomical history, namely the diploid representatives of *Leucanthemum* Mill.

The genus *Leucanthemum* Mill. (Compositae, Anthemideae) is a large polyploid complex comprising 42 species (*The Euro+Med Plantbase Project*) with ploidy levels ranging from diploid (2*x*) to dodecaploid (12*x*), and one species [*L. lacustre* (Brot.) Samp.] even showing a chromosome number of $2n = 22x = 198$. The genus is distributed all over the European continent, with one species (the tetraploid *L. ircutianum* DC.) reaching Siberia and some species introduced to many temperate regions of the northern and southern hemisphere (Meusel & Jäger, 1992). Most of the species are delimited based on differences in morphology, especially leaf morphology, geographical distribution, and ploidy level (e.g., Vogt, 1991, 2019). Phylogenetic relationships among the diploids of the genus – the fundamental layer or the "warps and wefts" of the polyploidy complex –were addressed in previous studies based on molecular markers (i.e., nrDNA ETS, AFLP fingerprinting, single-copy nuclear markers; Oberprieler et al., 2014, Konowalik et al., 2015, Wagner et al., 2017, 2019). Wagner et al. (2019) presented a multi-locus phylogenetic reconstruction of the subtribe Leucantheminae, in which the diversification among diploid *Leucanthemum* species was dated to the last 1.93 (1.14-2.94) Ma, arguing for the strong influence of Pleistocene oscillations on species formation.

Subjects of the present study are the 20 diploid *Leucanthemum* species of Central and Mediterranean Europe. The current species delimitation in this group is largely based on morphological characters, especially leaf shape, and ecology (Vogt, 1991, 2019). Instead of delimiting species de-novo, we are testing the plausibility of the current species delimitation, drawing evidence from genetic, morphometric, ecologic, and geographic analyses in the framework of integrative taxonomy. In this context, we present the application of modern methods for the analysis of genetic evidence in the form of

biologically informed species delimitation methods and Machine Learning techniques, morphological evidence by applying methods of leaf reconstruction and geometric morphometry, ecological evidence via ecological niche modeling and niche comparisons, and geographic patterns. By resolving the species boundaries in the diploid *Leucanthemum* representatives, we are laying the foundation for future research on the origin and phylogeny of polyploid representatives of the genus.

## Material and Methods

### RADseq Assembly

For the Double Digest RADseq (ddRADseq; Poland et al., 2012) procedure, a set of 54 individuals representing the 20 diploid *Leucanthemum* morpho-species was selected (see Table 4.1), comprising at least two individuals per taxon. The ddRAD data were generated in two batches, the first one comprising 52 silica-dried leaf samples from all morpho-species except *L. eliasii*, the second one two samples of *L. eliasii* from herbarium specimens. Genomic DNA was extracted according to the CTAB DNA extraction protocol of Doyle & Dickson (1987). Two replicates were generated by repeating the extraction process for *L. ageratifolium* (M60-01, M60-011) and *L. halleri* (162-03, 162-031). The DNA extracts were forwarded to LGC Genomics (Berlin, Germany) for ddRAD (2x150 bp) Illumina sequencing on an Illumina NextSeq 500 instrument (Illumina Inc., San Diego, CA, USA) with the restriction enzyme combination *Pst1* and *ApeK1*.

The raw reads were demultiplexed and adapter-clipped by LGC Genomics. Since there is no reference genome available for *Leucanthemum*, de-novo assembly of the preprocessed reads of the first ddRAD batch was performed using IPYRAD v0.9.54 (Eaton & Overcast, 2020). The pipeline parameters determining the clustering threshold (clust_threshold, *ct*) and the minimum number of samples required for a single locus to be retained (min_samples_locus, *msl*) were optimized, while the remaining parameters were kept at the default values. For the optimization, a cost function accounting for the assembly error of in-vitro replicates and the amount of missing data in the assembly was applied. The cost function is described by the formula $(1 - E_L) * (1 - E_S) * g(M_S, \mu = 50.0, \sigma = 50.0)$, where $E_L$ is the (mean) locus error rate, and $E_S$ the (mean) single nucleotide polymorphism (SNP) error rate considering the in-vitro replicates, $M_S$ is the percentage of missing data in the SNP data matrix returned by IPYRAD, and $g(M_S, \mu, \sigma)$ is the Gaussian transformation of $M_S$ with mean $\mu$ and standard deviation $\sigma$. The replicate-based error

measures $E_L$ and $E_S$ were calculated as described in Mastretta-Yanes et al. ( 2015). The idea behind this cost function is to combine the replicate-based error measures $E_L$ and $E_S$ with a rough target for the allowed amount of missing data in the assembly $g(M_S, \mu, \sigma)$. Assemblies with all parameter combinations of *msl* ranging from 4 to 52 in increments of 4 and *ct* ranging from 85 to 95 in increments of 2 were constructed. The assembly scoring best according to the previously described cost function was selected for constructing a ddRAD pseudoreference.

The aligned locus sequences were used to calculate a majority consensus sequence for each locus. The locus consensi were used as pseudoreference for a subsequent reference-based assembly using iPyrad of all 54 *Leucanthemum* accessions, keeping all parameters at default, except for *msl*, which was set to the optimal value according to the optimization procedure.

**Table 4.1**. List of samples used for the ddRAD analysis with information on voucher specimens (in B) and collection localities, coordinates, and collectors.

| Sample | Taxon | Voucher Specimens | Locality | Coordinates (latitude, longitude) | Collector |
|---|---|---|---|---|---|
| 135-05 | *L. ageratifolium* Pau | B100386712 | F, Occitania, Pyrénées-Orientales, 410 m | 42.5038, 2.9603 | *Konowalik KK42 & Ogrodowczyk* |
| M60-01 | *L. ageratifolium* Pau | B100345012, B100345013 | ES, Castile-La Mancha, Cuenca, 1157 m | 40.1019, -1.521 | *Cordel 60* |
| M02-01 | *L. ageratifolium* Pau | B100297950 | ES, Aragón, Huesca, 755 m | 42.525, -0.669 | *Cordel 2* |
| 90-01 | *L. burnatii* Briq. & Cavill. | B100464678 | F, Provence-Alpes-Côte d'Azur, Alpes-Maritimes, 1235 m | 43.7607, 6.9165 | *Vogt 16615 et al.* |
| 92-02 | *L. burnatii* Briq. & Cavill. | B100464676, B100464675 | F, Provence-Alpes-Côte d'Azur, Bouches-du-Rhône, 650 m | 43.545, 5.6626 | *Vogt 16618 et al.* |
| 60-01 | *L. cacuminis* Vogt & al. | B100413746 | ES, Galicia, Os Ancares, 1530 m | 42.8315, -6.8569 | *Höβl 60* |
| 62-01 | *L. cacuminis* Vogt & al. | B100413744 | ES, Galicia, Lugo, 750 m | 42.9249, -6.8657 | *Höβl 62* |
| 68-04 | *L. cacuminis* Vogt & al. | B100413738 | ES, Cantabria, Cantabria, 1770 m | 43.1538, -4.8053 | *Höβl 68 & Himmelreich* |
| L1008 | *L. eliasii* (Sennen & Pau) Sennen & Pau | B100484003 | ES, Castile and León, Burgos, 880 m | 42.503, -3.706 | *Lopéz 2537 & al.* |
| L1009 | *L. eliasii* (Sennen & Pau) Sennen & Pau | B100484004 | ES, Castile and León, Burgos, 920 m | 42.507, -3.705 | *Galán Cela 576 & Martín* |
| 159-11 | *L. gallaecicum* Rodr.Oubiña & S.Ortiz | B100386789, B100420775, B100464989 | ES, Galicia, Pontevedra, 375 m | 42.8498, -7.9878 | *Konowalik KK67 & Ogrodowczyk* |
| 161-03 | *L. gallaecicum* Rodr.Oubiña & S.Ortiz | No voucher | ES, Galicia, Corunna, 380 m | 42.8533, -7.9994 | *Konowalik s.n. & al.* |
| 58-02 | *L. gallaecicum* Rodr.Oubiña & S.Ortiz | B100413748 | ES, Galicia, Lugo, 490 m | 42.8205, -7.9504 | *Höβl 58* |
| 209-01 | *L. gaudinii* Dalla Torre | B100386664 | CH, Bern, Interlaken-Oberhasli, 2260 m | 46.5781, 7.97 | *Tomasello TS88* |
| 270-01 | *L. gaudinii* Dalla Torre | B100413007 | AT, Carinthia, Spittal an der Drau, 2200 m | 47.0025, 13.5275 | *Oberprieler 10859* |
| 276-01 | *L. gaudinii* Dalla Torre | B100413015 | AT, Carinthia, Feldkirchen, 2270 m | 46.8603, 13.8172 | *Oberprieler 10866* |
| 451-01 | *L. gaudinii* Dalla Torre | No voucher | PL, Lesser Poland, Giewont, 1860 m | 49.2505, 19.9343 | *Konowalik 20160909-01* |
| 84-10 | *L. gracilicaule* (Dufour) Pau | B100386704 | ES, Valencian Community, Alicante, 300 m | 38.8379, -0.1853 | *Konowalik KK20 & Ogrodowczyk* |
| 85-04 | *L. gracilicaule* (Dufour) Pau | B100386702 | ES, Valencian Community, Valencia, 340 m | 39.3135, -0.681 | *Konowalik KK25 & Ogrodowczyk* |

| Sample | Taxon | Voucher Specimens | Locality | Coordinates (latitude, longitude) | Collector |
|---|---|---|---|---|---|
| 116-01 | *L. graminifolium* (L.) Lam. | B100464684, B100464683 | F, Occitania, Hérault, 800 m | 43.7761, 3.2386 | *Vogt 16693 & al.* |
| 96-03 | *L. graminifolium* (L.) Lam. | B100464663 | F, Occitania, Aude, 600 m | 43.1494, 2.6294 | *Vogt 16656 & al.* |
| 162-03 | *L. halleri* (Vitman) Ducommun | B100386798 | D, Bavaria, Landkreis Garmisch-Partenkirchen, 2340 m | 47.4134, 11.1277 | *Konowalik KK67 & Tomasello* |
| 208-01 | *L. halleri* (Vitman) Ducommun | B100386672 | F, Valais, Sion, 2320 m | 46.3308, 7.2911 | *Tomasello TS65* |
| 280-01 | *L. laciniatum* Huter & al. | B100464203 | I, Calabria, Cosenza, 1580 m | 39.902, 16.1144 | *Tomasello 420* |
| 280-02 | *L. laciniatum* Huter & al. | B100464203 | I, Calabria, Cosenza, 1580 m | 39.902, 16.1144 | *Tomasello 420* |
| 366-01 | *L. legraeanum* (Rouy) B.Bock & J.-M.Tison | B100486634, B100486635, B100486636, B100486637, B100486638 | F, Provence-Alpes-Cote d'Azur, Var, 410 m | 43.1986, 6.3151 | *Vogt 17189* |
| 369-01 | *L. legraeanum* (Rouy) B.Bock & J.-M.Tison | B100486648, B100486649 | F, Provence-Alpes-Cote d'Azur, Var, 210 m | 43.2444, 6.3377 | *Vogt 17192* |
| 384-01 | *L. legraeanum* (Rouy) B.Bock & J.-M.Tison | B100627809, B100627810 | F, Provence-Alpes-Cote d'Azur, Var, 410 m | 43.1988, 6.3151 | *Vogt 17434 & al.* |
| 406-01 | *L. ligusticum* Marchetti & al. | B100627838, B100627839 | I, Liguria, La Spezia, 210 m | 44.247, 9.7728 | *Vogt 17460 & al.* |
| 412-01 | *L. ligusticum* Marchetti & al. | B100627849, B100627850, B100627851 | I, Liguria, Genova, 700 m | 44.3603, 9.5105 | *Vogt 17468 & al.* |
| 416-01 | *L. ligusticum* Marchetti & al. | B100627855, B100627856 | I, Liguria, Genova, 250 m | 44.3458, 9.4588 | *Vogt 17471 & al.* |
| 273-02 | *L. lithopolitanicum* | B100413012 | SL, Central Slovenia, 2100 m | 46.3633, 14.5715 | *Oberprieler 10862* |
| 274-02 | *L. lithopolitanicum* | B100413013 | SL, Savinja, 2000 m | 46.375, 14.5663 | *Oberprieler 10864* |

| Sample | Taxon | Voucher Specimens | Locality | Coordinates (latitude, longitude) | Collector |
|---|---|---|---|---|---|
| 128-01 | *L. monspeliense* (E.Mayer) Polatschek | B100464618 | F, Occitania, Gard, 750 m | 44.0888, 3.5786 | *Vogt 16712 & al.* |
| 131-02 | *L. monspeliense* (E.Mayer) Polatschek | B100464615 | F, Occitania, Gard, 380 m | 44.1412, 3.7316 | *Vogt 16716 & al.* |
| 340-01 | *L. monspeliense* (E.Mayer) Polatschek | B100486666, B100486667 | F, Occitania, Aveyron, 180 m | 44.5822, 2.184 | *Vogt 17156 & al.* |
| 40-09 | *L. pluriflorum* Pau | B100413758 | ES, Galicia, Corunna, 100 m | 42.8838, -9.2726 | *Höβl 40* |
| 42-04 | *L. pluriflorum* Pau | No voucher | ES, Galicia, Corunna, 150 m | 43.3069, -8.6186 | *Höβl 42* |
| 55-01 | *L. pluriflorum* Pau | B100413749 | ES, Galicia, Lugo, 10 m | 43.6309, -7.333 | *Höβl 55* |
| 266-01 | *L. pyrenaicum* Vogt & al. | B100464208 | ES, Aragon, Huesca, 1650 m | 42.7806, -0.2467 | *Tomasello TS382* |
| 266-02 | *L. pyrenaicum* Vogt & al. | B100464208 | ES, Aragon, Huesca, 1650 m | 42.7806, -0.2467 | *Tomasello TS382* |
| 267-03 | *L. pyrenaicum* Vogt & al. | B100464210 | ES, Aragon, Huesca, 2000 m | 42.6327, 0.453 | *Tomasello TS392* |
| 446-01 | *L. rotundifolium* (Willd.) DC. | No voucher | PL, Podkarpackie, Bieszczady, 920 m | 49.11905, 22.57755 | *Konowalik 20180622-02-01* |
| 447-01 | *L. rotundifolium* (Willd.) DC. | No voucher | RO, Bihor, Bihor, 1230 m | 46.51887, 22.66133 | *Konowalik 20180713-03-01* |
| 448-01 | *L. rotundifolium* (Willd.) DC. | No voucher | BH, Central Bosnia Canton, 1860 m | 43.95782, 17.74027 | *Konowalik 20180714-03-01* |
| 449-01 | *L. rotundifolium* (Willd.) DC. | No voucher | RO, Hunedoara, Râu de Mori, 1140 m | 45.31588, 22.77045 | *Konowalik 20180807-03-01* |
| 450-01 | *L. rotundifolium* (Willd.) DC. | No voucher | PL, Lesser Poland Voivodeship, Sucha County, 1100 m | 49.58787, 19.55152 | *Konowalik 20170920-01* |
| 278-01 | *L. tridactylites* (A.Kern. & Huter) Huter & al. | B100464207 | I, Abruzzo, Pescara, 2080 m | 42.1384, 14.1101 | *Tomasello 417* |
| 278-05 | *L. tridactylites* (A.Kern. & Huter) Huter & al. | B100464207 | I, Abruzzo, Pescara, 2080 m | 42.1384, 14.1101 | *Tomasello 417* |
| 225-02 | *L. virgatum* (Desr.) Clos | B100411746 | F, Provence-Alpes-Côte d'Azur, Alpes-Maritimes, 430 m | 43.9538, 7.2961 | *Vogt 16892 & Oberprieler* |
| 250-01 | *L. virgatum* (Desr.) Clos | B100350169, B100350172 | I, Liguria, Savona, 220 m | 44.0596, 8.0583 | *Vogt 16932 & Oberprieler* |
| 184-01 | *L. vulgare* Lam. | B100346626 | BH, Republic of Srpska, Nevesinje, 930 m | 43.2403, 18.3364 | *Vogt 16806 & Prem-Vogt* |
| L046-07 | *L. vulgare* Lam. | B100550249 | D, Bavaria, Deuerling, 450 m | 49.0333, 11.8833 | *Eder & Oberprieler s.n.* |
| 94-01 | *L. vulgare* Lam. | B100464674 | F, Occitania, Aude, 160 m | 43.1294, 2.6073 | *Vogt 16641 & al.* |

| Sample | Taxon | Voucher Specimens | Locality | Coordinates (latitude, longitude) | Collector |
|---|---|---|---|---|---|
| R016-12 | *R. arundanum* (Boiss.) B.H.Wilcox & al. | B100704793, B100704792 | MA, Fez-Meknes, cercle d'Outat El Haj, 2900 m | 33.6178, -3.8279 | *Vogt 17673 & al.* |
| R013-01 | *R. gayanum* (Coss. & Durieu) B.H.Wilcox & al. | B100704723 | MA, Fes-Boulemane, Boulemane, 2180 m | 33.6393, -4.1626 | *Vogt 17660 & al.* |
| R021-03 | *R. redieri* (Maire) B.H.Wilcox & al. | B100704774 | MA, Fez-Meknes, cercle d'Azrou, 2370 m | 33.2554, -4.9663 | *Vogt 17699 & al.* |

**Network Analysis**

As a first exploratory analysis, NeighborNets were calculated using SPLITSTREE4 v4.15.1 (Huson & Bryant, 2006) based on Kimura-2-parameter (K2P) distances of the concatenated SNPs (IPYRAD's .snps.phy output) and concatenated loci, and on SNP-based Nei distances calculated from the variant information (.vcf output). K2P distances with pairwise-deletion were calculated using the R package ape v5.4 (Paradis & Schliep, 2019); for calculation of the Nei distances a custom Python and C implementation of the Nei distance on the SNP level (https://github.com/TankredO/nei_vcf) as described in the POFAD manual (Joly & Bruneau, 2006; https://github.com/simjoly/pofad) was employed.

**Hybrid Detection**

Hybrid detection in the reduced dataset was performed using Patterson's D-statistics ("ABBA-BABA tests"; Green et al., 2010, Durand et al., 2011) implemented in IPYRAD, as described in Wagner et al. (2020). For each possible pair of morpho-species, all possible pairs of individuals from those species were subjected to Bonferroni corrected ABBA-BABA tests.

**Consensus Clustering**

Principal coordinate analysis (PCoA) was applied to transform pairwise K2P distances of the concatenated SNPs into a principal-coordinates matrix using the R package ape v5.4 (Paradis & Schliep, 2019). This was done for the whole dataset and for a reduced dataset comprising the samples from eleven morpho-species that were found being less clearly delimited according to the previous NeighborNet analysis. Based on the principal-coordinates matrices, consensus k-Means (CKM; Monti et al., 2003) clustering was conducted as described in Wagner et al. (2020). The number of clusters *k* was varied from 2 to 20 and 2 to 11, for the complete and reduced dataset, respectively. For both, CKMs were calculated with and without hybrid individuals. The feature and observation portions were both set to 0.8, meaning that the single k-Means runs (replicates) within a CKM use a random subset of the data matrix containing 80% of the samples and features. The number of replicates was set to 5,000. For both CKM analyses, the best *k* was determined using the Davies-Bouldin Index and the Silhouette Index. The analyses were conducted using the Python package pyckmeans v0.9.2 (https://github.com/TankredO/pyckmeans).

**Coalescent-based Species Delimitation**

Multispecies coalescent (MSC) species delimitation was performed using the BEAST2 (Bouckaert et al., 2014) package SNAPP v1.4.2 (Bryant et al., 2012; Leaché et al., 2014) based on the reduced dataset without putative hybrid individuals. Ten different species membership scenarios (S1-S10, see Figure 4.4) were surveyed by grouping taxa into clusters that were reasonable according to the current taxonomy and to previous analyses. SNAPP input files were constructed using BEAUTI v2.5.2 (Bouckaert et al., 2014). Since SNAPP analyses are computationally expensive, it was not possible to use the full SNP alignment (.phy.snps). Instead, IPYRAD's ".phy.usnps" output, comprising one randomly drawn SNP per locus, was used. The prior for the Yule model (lambda) was set to a gamma distribution with *alpha* = 2 and *beta* = 200. The population size prior was set to a gamma distribution with *alpha* = 1 and *beta* = 250. These parameter choices were motivated by the tutorial of Leaché & Bouckaert (2018). MCMC sampling (500,000 iterations, 500 sampling rate, 25% burn-in, alpha = 0.3) for the 10 scenarios with 48 path-sampling steps each, was conducted on the Athene HPC-cluster at the University of Regensburg.

**Ecological Niche Modeling**

Ecological niche modeling (ENM) was used to reconstruct potential distribution areas and compare the eco-climatological and edaphic niches of the six *Leucanthemum* morpho-species (*L. cacuminis*, *L. gallaecicum, L. gaudinii*, *L. pluriflorum*, *L. pyrenaicum*, *L. vulgare*), which could not be reliably delimited using genetic data. 732 collection locations of individuals from these taxa were retrieved from several herbaria [ARAN, B, BC, BCC, BCN, COI, E, FR, G, GOET, JACA, JBAG, LEB, LOU, LY, M, MA, MAF, MGC, NEU, P, PAD, SALA, SANT, SEV, TSM, VAL, VIT, W, WU; abbreviations according to the Index Herbariorum (http://sweetgum.nybg.org/science/ih/); see Electronic Supplementary ES1].

Rasters of 19 bioclimatic variables and ten edaphic variables from three different soil depth levels (0-5 cm, 5-15 cm, 15-30 cm) were retrieved from Worldclim Bioclim (Fick & Hijmans, 2017) and SoilGrids (Hengl et al., 2017), respectively (see Electronic Supplementary ES2 for variable descriptions). The rasters were cropped to a bounding box enclosing central Europe (longitude: -10.0° - 26.0°; latitude: 35.9° - 52°) and scaled to the resolution of the Bioclim rasters using the R package raster v3.3 (https://CRAN.R-

project.org/package=raster). The three depth levels of each of the edaphic variable were summarized by calculating the mean cell values, resulting in ten edaphic rasters. Principal component analysis (PCA) with normalization implemented in the R package ENMtools v1.0.2 (Warren et al., 2021) was applied to the Bioclim and SoilGrids rasters for dimensionality reduction and decorrelation. For each of the datasets, the three principal component rasters explaining the largest part of variance in the dataset were selected for subsequent analyses.

Based on the six principal component rasters and 732 collection locations, potential distribution areas were reconstructed for each of the six morpho-species using MAXENT v3.4.4 (Phillips et al., 2021). The cross-validation option was set to 6-fold. Niche equivalency, implemented in the R package ENMTools (Warren et al., 2021), was inferred among the morpho-species using MAXENT as ENM, a species range of 50 km, 1,000 background points with a radius of 20 km, and 200 permutations.

**Morphological Analyses**

For the morphological analyses of six genealogically close morpho-species, 66 images of digitized herbarium specimens were provided by the herbarium of the Berlin Botanical Museum (B; see Electronic Supplementary ES3). Based on the digital images, largely intact leaves were manually segmented (annotated) with polygons, and the corresponding midveins were annotated with polylines, using the Computer Vision Annotation Tool (CVAT) software (https://github.com/openvinotoolkit/cvat). In total, 417 contour-midvein pairs were retrieved and annotated images were exported as CVAT image (XML) dataset.

Leaf deformations that very likely influence the results of downstream geometric morphometric analyses were compensated by straightening the leaves using a custom procedure implemented in Python and C based on the as-rigid-as-possible algorithm (Igarashi et al., 2005) and relying on functionality from the Python packages scikit-image (van der Walt et al., 2014), numpy (Harris et al., 2020), triangle (Shewchuk, 1996; https://github.com/drufat/triangle), and igl (Jacobson et al., 2018). The process comprises four steps: (1) Generate a triangle mesh from the leaf polygon and midvein polyline by applying a constrained Delaunay triangulation (Chew, 1989), (2) move the midvein vertices in such a way that they form a straight line while preserving distances between the vertices, (3) move all other vertices according to the as-rigid-as-possible algorithm(Igarashi et al.,

2005), and (4) map the leaf texture from the original triangle mesh to the transformed triangle mesh (Figure 4.1).

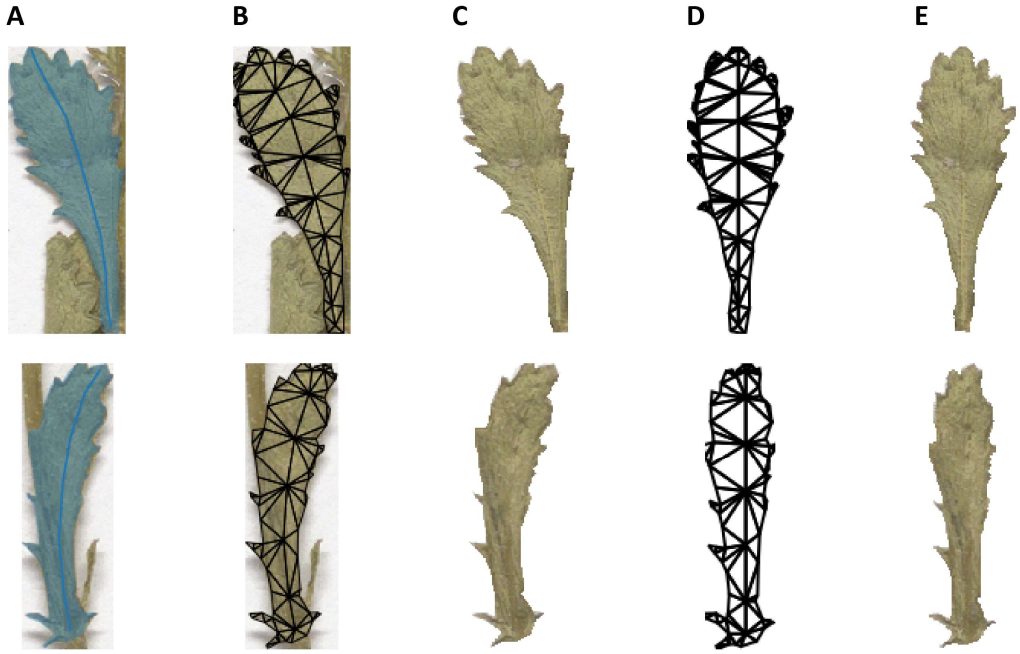


**Figure 4.1**. Leaf straightening process illustration. The leaf and midvein annotations (**A**) are used to construct a triangle mesh (**B**). Based on this mesh, the background is removed (**C**). The midvein points are arranged in a straight line, preserving distances between line vertices; the remaining vertices are updated according to the as-rigid-as-possible algorithm (**D**). Finally, the texture from the bent leaf mesh is mapped to the straightened leaf mesh (**E**).

To extract morphological features from the straightened leaf images, elliptic Fourier analysis (EFA; Kuhl & Giardina, 1982) was conducted and leaf dissection indices (LDI; Kincaid & Schneider, 1983) were calculated. LDI calculation and EFA were performed with the Python packages scikit-image (van der Walt et al., 2014), numpy (Harris et al., 2020), scikit-learn (Pedregosa et al., 2011), and pyefd (https://github.com/hbldh/pyefd). For the EFA, the number of harmonics was set to 15 and the elliptic Fourier descriptors (EFD) were normalized. The non-constant EFDs were subjected to a principal component analysis (PCA) for decorrelation.

The first two principal component (PC) scores were used for permutation-based testing for significant differences of leaf shapes among morpho-species. For all possible pairs of morphospecies within genealogical clusters, the observed Euclidean distances in PC space were calculated. The observed distances were tested for significance against simulated null distributions generated by randomly swapping taxon labels and calculating

the corresponding distances 5,000 times. The *p*-values were Bonferroni corrected. Additionally, the LDI values were tested for significant differences using Bonferroni corrected Welch's tests.

**Geography**

Pair-wise geographic overlap between taxa is an important factor for species delimitation in our integrative taxonomical approach. Since there are no complete collection maps available for the six focal *Leucanthemum* morpho-species, it was necessary to approximate the true geographic distribution from the incomplete sampling data available. For this purpose, an ENM was fitted using MAXENT (Phillips et al., 2021), based on the BioClim, SoilGrid and, additionally, longitude and latitude rasters. For each morpho-species, the predicted suitability raster was subjected to a thresholded depth-first search (DFS), originating from those raster cells, for which collection data were available. A cell was only visited by the DFS, if the corresponding suitability score was above a defined threshold, in this case 0.25; cells that were not visited were set to 0.0. Finally, cells containing collection points were set to 1.0. The thresholded DFS removes areas that are disconnected from true sampling locations and additionally filters the suitability rasters according to the threshold. In the following, those filtered rasters were treated as approximate species-distribution maps.

Based on these, pairwise overlap between morpho-species was calculated by (1) multiplying the approximate distribution rasters, (2) thresholding the resulting raster by 0.25, (3) calculating the number of non-zero raster cells, and (4) dividing by the number of non-zero raster cells of the species with the smaller distribution area. To test for significant deviation from sympatry, a permutation-based approach was taken. For each comparison, 400 datasets were simulated by randomly swapping taxon labels, the distribution areas were approximated, and the overlap was calculated. Finally, the *p*-value was calculated by comparing the observed overlap with the simulated overlap. In this setup, a significant deviation from the null model in direction of a lower overlap means a deviation from sympatry that cannot be explained by chance alone.

## Results

### RADseq Assembly

The number of demultiplexed, adapter clipped, and restriction enzyme filtered reads was 85,612,720 (33.52 mean quality) and 6,936,593 (34.57 mean quality) for the first and second ddRAD batch, respectively. Considering the de-novo assembly, optimization of the proposed cost function led to an optimal parameter combination of *ct* = 95 and *msl* = 12 (Electronic Supplementary ES4). The corresponding assembly comprised a total of 10,246 RAD loci with 136,755 SNPs, 70,755 of which were parsimony informative. The mean locus coverage per sample was 5,035.8 (sd: 795.3). The reference mapping of all 54 *Leucanthemum* accessions against the pseudo-reference resulted in 9,248 RAD loci with 155,294 SNPs; 86,390 of which were parsimony informative, at a mean locus coverage of 5,415.1 (sd: 721.8) per accession.

### Network Analysis

We found that the NeighborNet analysis clusters members from 18 of the 20 *Leucanthemum* morpho-species (Figure 4.2). For the remaining two species (i.e., *L. cacuminis* and *L. ageratifolium*), a single accession lies outside the cluster containing the remaining individuals of the taxon. There are two larger groups visible in the NeighborNet, the left of them comprising nine morpho-species clusters that are separated from each other by relatively long branches when compared to the right group.

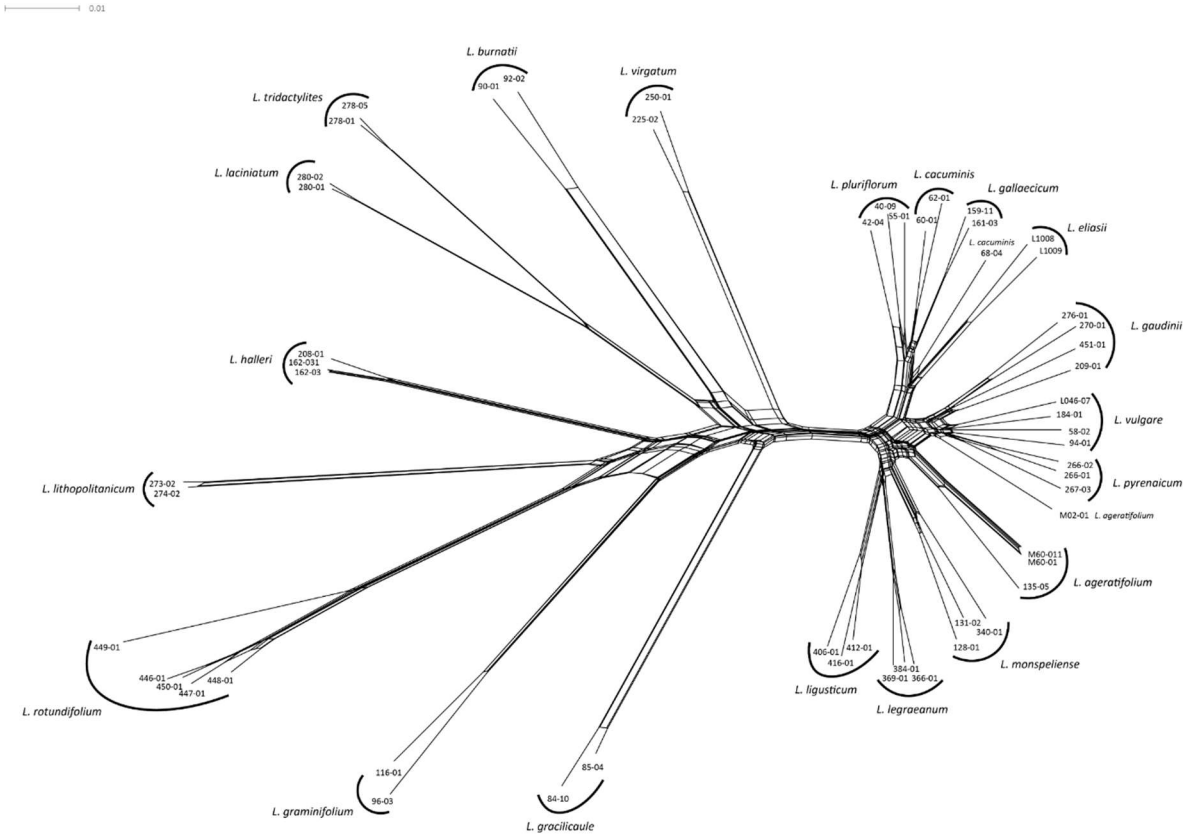


**Figure 4.2**. Neighbor net of the 20 Leucanthemum morpho-species, constructed using Kimura 2-parameter distances of the concatenated ddRAD loci from the IPYRAD pseudo-reference assembly. Species membership is indicated by enclosing curves, or text for single accessions. Accessions of 18 species form tight clusters, while accessions M02-01 (*L. ageratifolium*) and 68-04 (*L. cacuminis*) lie outside of the respective species clusters.

## Hybrid Detection

In total, 5,154 ABBA-BABA tests were conducted to detect hybrid individuals, 29 of which were significant (*alpha* = 0.01 after Bonferroni correction; see Electronic Supplementary ES6). The accession M02-01 *(L. ageratifolium)* was involved in 27 of those tests, indicating a strong hybridization signal. The remaining two significant tests involved the triplet consisting of accessions 60-01, 68-04, and 55-01. Of these only 68-04 *(L. cacuminis)* was already noticeable in previous analyses.

## Consensus Clustering

According to the Davies-Bouldin (DB) and silhouette index (SIL), the optimal number of clusters for the complete dataset was 15, both with and without hybrid individuals. The consensus clustering with $k$ = 15 delimits 13 of the 20 diploid *Leucanthemum* morpho-species (Figure 4.3). The remaining seven species form two

clusters, consisting of *L. vulgare*, *L. pyrenaicum*, and *L. gaudinii* on the one hand, and *L. pluriflorum*, *L. cacuminis*, *L. gallaecicum*, and *L. eliasii* on the other. The optimal number of clusters for the reduced dataset with and without hybrids was 11 and 8, respectively. It is noteworthy that even for the dataset including hybrids there is a local optimum in the DB and SIL measures at $k = 8$. The consensus clustering with $k = 8$ results in the merging of *L. vulgare* with *L. gaudinii* on the one side, and *L. pluriflorum*, *L. cacuminis*, and *L. gallaecicum* on the other. PCoA results are available in Electronic Supplementary ES5.

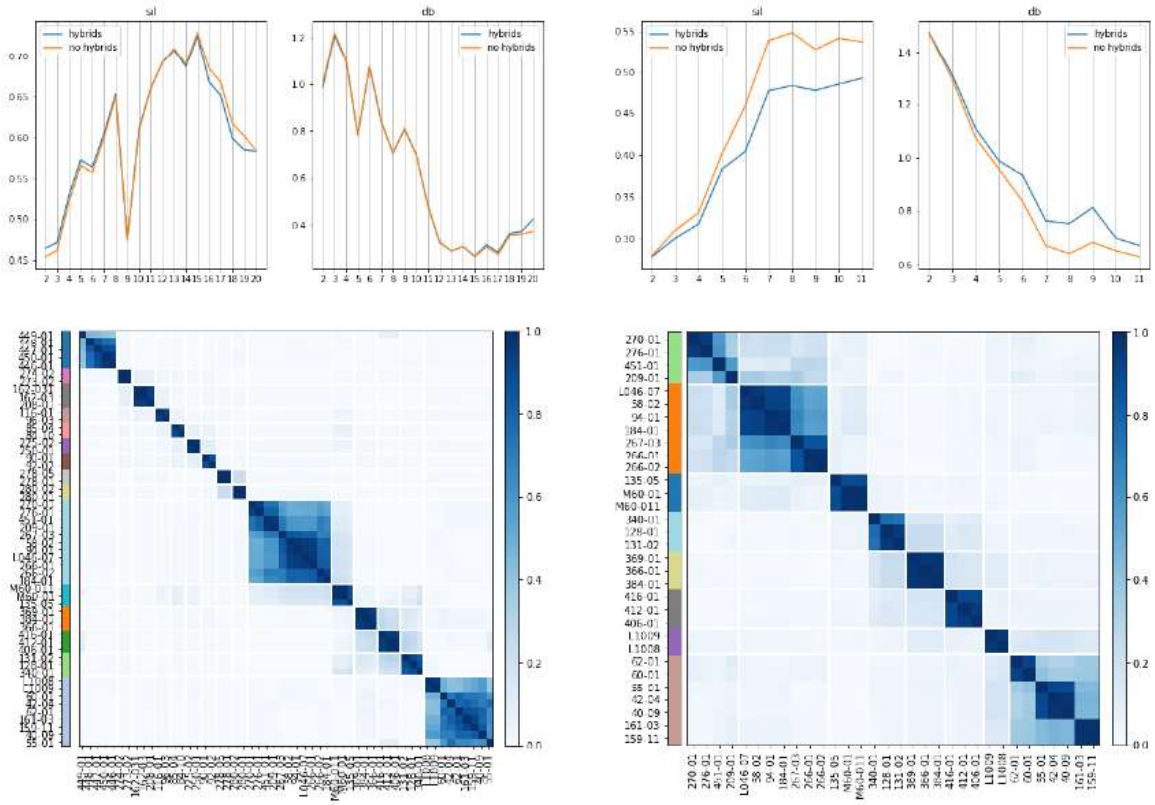


**Figure 4.3**. Consensus K-Means clustering results for the complete (left) and reduced (right) datasets. The line charts show the clustering quality metrics Silhouette Index (SIL; higher is better) and Davies-Bouldin Index (DB; lower is better) for datasets both with (blue) and without (orange) hybrid individuals (M02-01, 68-04). For the complete dataset, the optimal number of clusters is 15, independent of the presence of hybrid individuals. Considering the reduced dataset, the optimal k is either 11 or 8, with and without hybrids, respectively. The heat maps depict the consensus matrices for the optimal k of the complete (left) and reduced (right) datasets excluding hybrid individuals.

### Coalescent-based Species Delimitation

Marginal likelihood (ML) values were found to increase with the model complexity over the ten scenarios. The scenario with the highest ML was the one, in which each morpho-species formed its own cluster (scenario 10). As the tendency of MSC-based species delimitation methods to overestimate the number of species is known (e.g., Wagner

et al., 2020; Mason et al. 2020), we determined the knee of the ML-model complexity curve using the kneedle algorithm (Satopaa et al., 2011), in which we assumed model complexity being represented by species number. The knee (or elbow) is the point where the curvature of a function changes considerably; this can be interpreted as the point, where increasing the number of species only leads to a relatively low increase in model performance. We found a knee point at the model with seven groups (scenario 4), which merges the morpho-species *L. vulgare*, *L. gaudinii*, and *L. pyrenaicum*, and the species *L. pluriflorum*, *L. cacuminis*, and *L. gallaecicum* to a single group, respectively, while the remaining morpho-species are staying separate (Figure 4.4).

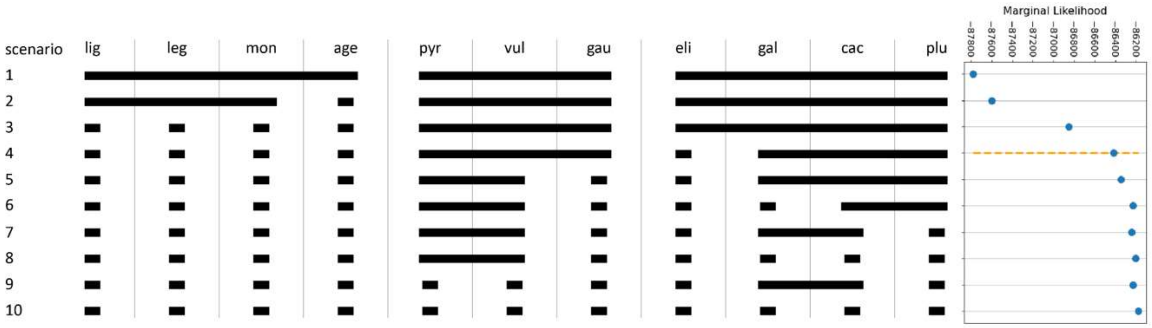


**Figure 4.4**. The ten species scenarios (scenario 1-10) resulting from merging morpho-species in ascending orders of complexity and the plot of the Marginal Likelihoods (ML) against species scenarios. The elbow point of the ML-complexity curve (orange dotted line) is located at scenario 4.

### Ecological Analysis

The first three principal components (PC) captured 82%, and 77% of the total variance considering the BioClim and SoilGrids raster PCAs, respectively. All niche reconstructions using MAXENT achieved cross-validation AUCs higher than 80%. All pairwise niche-equivalency tests, except for the pair *L. vulgare* - *L. gaudinii*, showed environmental niches being significantly different (Table 4.2).

### Morphological Analysis

Within the *L. vulgare*-cluster, the difference in PC space between *L. vulgare* and *L. pyrenaicum* is not significant (corr. $p = 5.02$), while the differences of the combinations *L. vulgare* - *L. gaudinii* (corr. $p = 0.00$) and *L. gaudinii* - *L. pyrenaicum* (corr. $p = 0.00$) are both significant. Within the *L. pluriflorum*-cluster, the test of *L. pluriflorum* against *L. cacuminis* is not significant (corr. $p = 0.28$), while *L. pluriflorum* vs. *L. gallaecicum* (corr. $p = 0.00$) and *L. gallaecicum* vs. *L. cacuminis* (corr. $p = 0.00$) are both significant (Table

2). Considering the LDI, the Welch's tests indicated significant differences (corr. $p < 0.01$) for the two comparisons *L. vulgare* vs. *L. gaudinii*, and *L. pyrenaicum* vs. *L. gaudinii*, while for the remaining comparisons no significant differences could be detected (Table 4.2).

### Geography

In the *L. vulgare*-cluster, the permutation-based tests for geographic overlap indicated a significant allopatry signal for the combination *L. pyrenaicum* vs. *L. gaudinii* (corr. $p = 0.00$), while *L. vulgare* vs. *L. pyrenaicum* (corr. $p = 6.00$) and *L. vulgare* vs. *L. gaudinii* (corr. $p = 2.22$) are not found being significantly allopatric. In the *L. pluriflorum*-cluster, all tests were significant (Table 4.2).

**Table 4.2**. Significance table for the pairwise comparisons in the L. pluriflorum- and L. vulgare-groups, with p-values for the environmental niche-overlap metric D, the permutation tests of the elliptic Fourier analysis (EFA), the Welch's tests of the leaf-dissection index (LDI), and for the permutation test of the geographical overlap (geo). In all cases, the p-values were Bonferroni-corrected.

| **Taxon A** | **Taxon B** | ***p(D)*** | ***p(EFA)*** | ***p(LDI)*** | ***p(geo)*** |
|---|---|---|---|---|---|
| *L. gaudinii* | *L. vulgare* | **0.03** | **0.00** | **0.02** | 2.22 |
| *L. gaudinii* | *L. pyrenaicum* | 0.06 | **0.00** | **0.00** | **0.00** |
| *L. vulgare* | *L. pyrenaicum* | **0.03** | 5.02 | 0.34 | 6.00 |
| *L. pluriflorum* | *L. gallaecicum* | **0.03** | **0.00** | 0.67 | **0.00** |
| *L. pluriflorum* | *L. cacuminis* | **0.03** | 0.28 | 3.65 | **0.00** |
| *L. gallaecicum* | *L. cacuminis* | **0.03** | **0.00** | 0.42 | **0.00** |

## Discussion

With the present study, we aimed at an objective delimitation of the diploid species of the genus *Leucanthemum* – the fundamental layer or the "warps and wefts" of the polyploidy complex this genus forms in central and southern Europe. In its taxonomic scope, the strategy of our analyses followed the operational sequence of (a) grouping and (b) ranking as described by (Stuessy, 1979; Stuessy et al., 2014) or Reydon & Kunz (2019). While the former procedure is engaged with finding discontinuities among representatives of a study group in datasets of different nature (e.g., morphology, genetics, ecology, etc.), the latter evaluates these taxonomic patterns and decides "at which level of the taxonomic hierarchy the taxa should be placed" (Stuessy, 1979: 623). Species delimitation (SD) frameworks are focused on the taxonomic rank of species and are considered comprising

both “discovery” and “validation” approaches (Carstens & Satler, 2013; Ence & Carstens, 2011). The “discovery” methods are actually discontinuity-inferring, pattern-recognition techniques that produce hypotheses on a taxonomic structure in the organism group under study, which are subsequently scrutinized by “validation” methods that incorporate models of variation exhibited at and below the species level and hence are based on underlying external assumptions (i.e., species concepts).

As in many modern taxonomic studies in areas of the world with a long tradition of (phyto-)taxonomical exploration, our present species-delimitation study in *Leucanthemum* diploids did not start from the scratch with an uninformed “grouping” step. In terms of an adequate sampling of the morphological and genetic diversity in the study area it was informed by hypotheses on taxonomic structure from former revisionary works based mainly on morphological information (i.e., Heywood, 1976, Vogt, 1991, 2019, *The Euro+Med Plantbase Project*). This may appear disadvantageous compared to a “total-ignorance” strategy that would prefer random sampling or sampling along transects or grids, but it represents the more frequent state-of-the-art in taxonomic studies nowadays. Though we have used this morpho-taxon concept for an informed sampling of our accessions in the multilocus RADseq analyses, we have subjected these genomic data both to “discovery” (consensus *k*-means clustering) and “validation” (coalescent-based species-delimitation with SNAPP) analyses. With this morpho-taxa-as-species-hypotheses approach, it was also possible to link the genealogical analyses with geographical, morphological, and ecological datasets based on a larger set of accessions for the final discussion on the ranking of the discovered entities.

Owing to the “integrative taxonomy” approach applied in the two most critical taxon-groups of the present study (the *L. vulgare*- and the *L. puriflorum*-group), we have used genealogical, morphological, and ecological “discovery” and/or “validation” methods, while for the ranking of the observed entities we additionally included geographical information on the entities´ distribution ranges. For addressing the genealogical, morphological, and geographical layers of evidence, novel methodological approaches were implemented and will be discussed in the following.

**The genealogical layer**

RADseq can be thought of an approximation of the total genomic variation and is thus well-suited as a source of genetic evidence in the absence of whole-genome

sequencing data. We implemented a pseudo-reference-based assembly approach for combining data from multiple ddRAD batches. This pseudo-reference assembly allows to efficiently combine ddRAD datasets while reducing the risk of dataset-specific noise influencing downstream analyses.

Species delimitation based on nucleotide data is known to be subject to the two opposing problems of over- and under-splitting (Sukumaran & Knowles, 2017). The former describes the problem of overestimating the number of species by wrongly recognizing (meta-)population structure as evidence for distinct species. This issue is known to occur especially in the context of species-delimitation tools, whose methodology is based on the multi-species-coalescent (MSC) model (Sukumaran & Knowles, 2017; see Wagner et al., 2020 for an example). Under-splitting, on the other hand, is the failure of detecting differences between species and consequently merging them to a single one. This problem can occur if hybridization is present, because MSC-based species-delimitation methods assume the biological species concept (BSC). Under-splitting is a problem that can be relatively easily mitigated by applying hybrid-detection methods to remove those intermediary individuals as we have done with the ABBA-BABA test. The problem of over-splitting, however, has not yet been solved when working with the MSC model. Here, we followed the suggestion of Wagner et al. (2020) by applying, in addition to a classical MSC-based method (i.e., SNAPP; Bryant et al., 2012, Leaché et al., 2014) a non-MSC method (consensus K-Means, CKM; Monti et al., 2003) for genetics-based SD. Interestingly, the results of both methods are largely concordant (Figure 4.3, Figure 4.4). Similar observations have been made by Wagner et al. (2020). It might be worthwhile, but out of scope for the present study, to investigate whether this is a general property of the non-parametric and significantly faster CKM to approximate MSC-based SD methods.

We find that 15 of the 20 currently described *Leucanthemum* morpho-species form clusters supported by CKM (Figure 4.3). Of the remaining five species, the groups consisting of *L. vulgare* and *L. pyrenaicum* (Figure 4.3, right, orange), and of *L. pluriflorum*, *L. cacuminis* and *L. gallaecicum* (Figure 4.3, right, brown) each form a cluster. The MSC-based SD additionally merges *L. gaudinii* with *L. vulgare* and *L. pyrenaicum* (Figure 4.4). Those two clusters, in the following called the *L. vulgare*-cluster and the *L. pluriflorum*-cluster, respectively, were the focal groups of the remaining analyses of ecology, geography and morphology.

**The morphological layer**

Leaf morphology is a key factor for the currently accepted species delimitation in *Leucanthemum*; especially the degree of leaf dissection is considered to be important for distinguishing species (Vogt, 1991, 2019). Leaf-dissection index (LDI, Kincaid & Schneider, 1983) and elliptic Fourier analysis (EFA; Kuhl & Giardina, 1982) are well-established methods for quantifying variations in leaf shape. While the LDI is invariant if leaves are deformed in such a way that the area and perimeter are conserved, the EFA is sensitive to such distortions. Unger et al. (2016) proposed a leaf normalization procedure that aligns pixels of leaf images so that the midvein forms a straight line. This method, however, introduces new distortions when the leaves are strongly bent. We picked up the idea of Unger et al. (2016) and combined it with a shape manipulation method, which minimizes shape distortions (Figure 4.1). Future research in this context could focus on the effect of straightening procedure on results of geometric morphometric analysis.

Based on the reconstructed leaves, we find that within the *L. vulgare*-cluster the pairs *L. vulgare* - *L. gaudinii* and *L. gaudinii – L. pyrenaicum* vary significantly in the degree of leaf dissection and also in the general leaf shape. Within the *L. pluriflorum*-cluster, we could not find any significant difference in the dissection of the leaves, but found that the general leaf shapes of the pairs *L. pluriflorum – L. gallaecicum,* and *L. gallaecicum – L. cacuminis* are significantly different.

**The ecological layer**

Both climatological and edaphic variables were incorporated into our analyses on the abiotic ecological niches of the morpho-species of the two *Leucanthemum* species-groups and followed traditional ecological niche modeling (ENM) as proposed by Raxworthy et al. (2007). For niche comparisons, we adopted the niche equivalency tests of Warren et al. (2021), whose efficacy for SD has been demonstrated in recent studies (e.g., Lin et al., 2021, Zheng et al., 2017). Our analyses show that the abiotic ecological niches are significantly different for all pairwise comparisons, except for the pair *L. vulgare – L. pyrenaicum*. These findings suggest that there is a signal of ecological differentiation even in the absence of clear genetic differentiation.

**The geographical layer**

Even if taxa occupy the same ecological niche, they can still be geographically separated. Conversely, taxa occupying different ecological niches can occur in spatial proximity. Indeed, those patterns have been long used to inform SD (von Wettstein, 1898). Determining the sympatry of species objectively can be problematic, especially in the absence of distribution maps, as it is the case for *Leucanthemum*. Due to this issue, we had to resort to retrieving historical collection data from herbaria. Since herbarium collection data is spatially fragmented and unequally sampled, it is necessary to approximate the true spatial distribution. Established methods for approximating the spatial distribution are convex hulls (Meyer et al., 2017), alpha hulls (Burgman & Fox, 2003; Capinha & Pateiro-López, 2014; Meyer et al., 2017), or kernel density estimates (Burgman & Fox, 2003; Cardillo & Warren, 2016) based on collection points. Those methods are very useful if the sampling is mostly complete or if the shape of the spatial distribution is following the corresponding hull function; they might fail, however, if the true distribution is only sparsely and spatially disjunctly sampled (Burgman & Fox, 2003).

Here, we described an ENM-based approach for approximating the true distribution raster from incomplete collection data. This method can be seen as an objective and reproducible variant of a distribution map that a specialist might draw according to available collection points and expert knowledge of the ecoclimatic variables at play. It might be worthwhile, but out of scope for this contribution, to study how this approximation behaves with other empirical or simulated datasets, and how performant it is at estimating the true distribution area. Based on the approximated distribution rasters, we tested the spatial overlap of the species, which we assumed to be a measure for sympatry and allopatry. We found that the currently delimited morpho-species in the *L. pluriflorum*-cluster all occur allopatrically and that within the *L. vulgare*-cluster only *L. pyrenaicum* and *L. gaudinii* were allopatrically distributed.

**Integration of layers and ranking of entities**

An early formal integration of morphology and geography in species-level taxonomy has been proposed by von Wettstein (1898). Representing an impressively modern evolutionary approach to taxonomy, this author considers allopatrically distributed, morphologically similar (i.e., closely-related) units being best acknowledged at subspecies level, while species-rank should be attributed to closely-related, but sympatrically

distributed entities. His argument is that only in the latter case ecological and/or reproductive differentiation between the units is sufficiently advanced to prevent merging of these lineages. This corresponds to the Evolutionary Species Concept of Wiley (1978) who considers species being “ancestor-descendant lineages that evolve separately from other such lineages and have their own evolutionary tendencies and historical fate” (Zachos, 2016: 83). It also stresses the importance of ecological differentiation among species for their geographical co-existence as independent evolutionary lineages as proposed by van Valen (1976) in his Ecological Species Concept.

We think that while morphology as a proxy for genetic similarity or genealogical relatedness is nowadays outperformed by genomic data, information on geographical separation and ecological differentiation as drivers, co-variants, and/or consequences of speciation processes definitively should be included in taxonomical decisions concerning ranks at or below the species-level. However, morphological information should not be disregarded completely, since discontinuities in this field of taxon properties may still correlate with genealogical differentiation and may either predate, coincide, or follow a speciation event [the Gray Zone of conflicting species concepts in de Queiroz' (2007: 882, Figure 1) argumentation scheme of speciation].

Owing to the pitfalls of an exclusively genealogy-based species-delimitation approach, which may lead to the misconception of population structure for species boundaries (Sukumaran & Knowles, 2017), we adopt an integration of genealogical, morphological, geographical, and ecological data that should cover patterns resulting from a broad array of speciation modes (i.e., allopatric, peripatric, parapatric, or sympatric speciation processes). In this conceptual framework based on von Wettstein's (1898) reasoning, species-level should be only accepted for two lineages that are genealogically distinct if they simultaneously are geographically overlapping and (but not necessarily) ecologically distinct. Significant morphological discontinuities between the two lineages may then only allow for the distinction between cryptic and phenetically appreciable species. In line with van Valen's (1976) Ecological Species Concept, allopatrically distributed and genealogically independent lineages should also deserve acknowledgement as species when their ecological niches are different enough to allow for expecting their reproductive isolation in potential sympatry.

**Taxonomical consequences**

With the described conceptual framework at hand, species delimitation in the two *Leucanthemum* clusters under study (the *L. vulgare*-cluster and the *L. pluriflorum*-cluster) could be put into effect as follows:

**(1)** The *Leucanthemum vulgare*-group: While *L. vulgare* is a taxon wide-spread in Europe, *L. gaudinii* and *L. pyrenaicum* are restricted to the Alps and the Pyrenees, respectively. In genealogical respects, *L. gaudinii* is on the verge of evolving as an independent lineage, with MSC-based SD showing no significant differentiation, but CKM clustering does, while the other two taxa are found belonging to the same metapopulation system. While both *L. gaudinii* and *L. pyrenaicum* are not significantly allopatric to *L. vulgare,* the former is morphologically different and the latter ecologically. In the case of *L. gaudinii,* its ecological and geographical overlap with *L. vulgare* combined with its genealogical (and morphological) distinctness argues for acknowledgement as an independently evolving lineage (i.e., species), while *L. pyrenaicum* represents only an ecologically deviating facies of *L. vulgare* – an ecotype that could be at best acknowledged taxonomically as subspecies of the widespread taxon.

**(a)** ***Leucanthemum gaudinii*** Dalla Torre in Sonklar & al., Anleit. Wiss. Beob. Alpenreisen 2: 244. 1882 ≡*Chrysanthemum gaudinii* (Dalla Torre) Dalla Torre & Sarnth., Fl. Tirol 6(3): 543. 1912 ≡*Chrysanthemum leucanthemum* var. *gaudinii* (Dalla Torre) Fiori, Nuov. Fl. Italia 2(4): 624. 1927 ≡*Chrysanthemum leucanthemum* ["f"] *gaudinii* (Dalla Torre) Fiori in Fiori & Paoletti, Fl. Italia 3(1): 239. 1903 – Neotypus [Gutermann, Phyton (Austria) 17: 37. 1975]: Kärnten: Nockgebiet, am Weg zur Falkerthütte – Bocksattel – S des Mallnock, 1850 m; mit *Calluna;* flachgründiger Boden über Silikat; leg. *A. Polatschek* P64/312; 2*n* = 18 (W! [W1965-0020139]).

**(b)** ***Leucanthemum vulgare*** Lam. [**subsp. v*ulgare***], Fl. Franç. 2: 137. 1779 ≡*Chrysanthemum leucanthemum* L., Sp. pl. 888. 1753 ≡*Tanacetum leucanthemum* (L.) Sch.Bip., Tanaceteen: 35. 1844 ≡*Matricaria leucanthemum* (L.) Desr. in Lam., Encycl. 3(2): 731. 1792 – Lectotypus [Böcher & Larsen, Watsonia 4: 15. 1957]: (BM-Hortus Cliffortianus).

**(c)** ***Leucanthemum vulgare*** **subsp.** ***barrelieri*** (Dufour ex DC.) O.Bolós & Vigo, Pl. Països Catalans 3: 816. 1996 "1995" ≡*Pyrethrum halleri* var. *barrelieri* Dufour ex DC., Prodr. 6: 55. 1838 (basionym) ≡*Leucanthemum gaudinii* subsp. *barrelieri* (Dufour ex DC.) Vogt in Ruizia 10: 89. 1991 ≡*Leucanthemum ceratophylloides* var. *barrelieri* (Dufour ex DC.)

Nyman, Consp. Fl. Eur.: 371. 1879 ≡*Pontia barrelieri* (Dufour ex DC.) Bubani, Fl. Pyr. 2: 219. 1899 ≡*Pyrethrum barrelieri* Dufour ex DC., Prodr. 6: 55. 1838, pro. syn., nom. inval. ≡*Leucanthemum vulgare* var. *pyrenaicum* Rouy, Fl. France 8: 272 & 274. 1903, nom. illegit. ≡*Leucanthemum pyrenaicum* Rouy, Fl. France 8: 274. 1903, pro syn., nom inval. [non *Leucanthemum barrelieri* Timb.-Lagr., Bull. Soc. Bot. France 13: 153. 1866 and in Rodet, Bot. Agric. Médic., ed. 2: 447. 1872] ≡*Leucanthemum pyrenaicum* Vogt et al. in Mol. Phylogenet. Evol. 92: 325. 2015. – Holotype: Pyren., près du Sommet de Monné, 1824 (G-DC! [G00450856]).

**(2)** The *Leucanthemum pluriflourm*-group: All three morpho-species of this group are allopatrically distributed: while *L. pluriflorum* is restricted to the coastline of Galicia in NW Spain (see Greiner et al., 2013; Vogt, 1991, 2019), *L. gallaecicum* is endemic on serpentine in central Galicia, and *L. cacuminis* (the former *L. gaudinii* subsp. *cantabricum;* Konowalik et al., 2015) is found in the mountainous regions of N Spain from the western Pyrenees in the east to the Picos de Europa in the west. Since no significant genealogical independence among the three taxa was found, acknowledgement on species-level is questionable, despite the significant non-overlapping in ecological respects of all three taxa and the morphological distinction of *L. gallaecicum* from the other two taxa. Therefore, our interpretation of the *L. pluriflorum*-group as a lineage of allopatrically and ecologically differentiating population group may be best represented taxonomically by ranking the three taxa as subspecies of a single species:

**(a)** ***Leucanthemum pluriflorum*** Pau [**subsp. p*luriflorum***] in Bol. Soc. Aragonesa Ci. Nat. 1: 31. 1902 – Holotype: San Ciprián, Galicia, *P. Merino S.J.* (MA! [MA128479]).

**(b)** ***Leucanthemum pluriflorum*** **subsp.** ***cantabricum*** (Font Quer & Guinea) T.Ott, Vogt & Oberpr., **comb. nov.** ≡*Leucanthemum vulgare* var. *cantabricum* Font Quer & Guinea in Guinea, Anales Jard. Bot. Madrid 7: 347-348. 1947 (basionym) ≡*Chrysanthemum leucanthemum* subsp. *cantabricum* (Font Quer & Guinea) Guinea, Catálogo florístico de Viscaya: 646. 1980 ≡*Leucanthemum gaudinii* subsp. *cantabricum* (Font Quer & Guinea) Vogt in Ruizia 10: 98. 1991 ≡*Chrysanthemum leucanthemum* var. *cacuminis* Font Quer & Guinea in Guinea, Bot. Santander: 327. 1953, nom. inval. ≡*Leucanthemum cacuminis* Vogt et al. in Mol. Phylogenet. Evol. 92: 325. 2015. – Holotype: Picos de Europa: in saxosis l. Vega de Liordes, ad 1890 m alt., 13.8.1944*, E. Guinea* (BC!).

**(c) *Leucanthemum pluriflorum* subsp. *gallaecicum*** (Rodr.Oubiña & S.Ortiz) T.Ott, Vogt & Oberpr., **comb. et stat. nov.** ≡*Leucanthemum gallaecicum* Rodr.Oubiña & S.Ortiz in Anales Jard. Bot. Madrid 47: 498. 1990 (basionym). – Holotype: La Coruña: Toques, Paradela, 20-IX-1987, *J. Rodríguez Oubiña & S. Ortiz* (SANT; isotype: MA! [MA478919]).


## Acknowledgements

We thank Anja Heuschneider for preparing the samples for ddRADseq sequencing. The suggestions of Dr Agnes Scheunert and Ulrich Lautenschlager considering the analyses and the manuscript were always welcome and are very much appreciated.

## Funding Information

This work was funded by the grant of the German Research Foundation (DFG) in the frame of the SPP 1991 "Taxon-omics − New Approaches for Discovering and Naming Biodiversity" to C.O. (OB 155/13-1) and R.V. (VO 1595/3-1).


## Conflict of Interest

All authors declare no conflict of interest.

## Additional Analyses and Results

In the course of the study described above, phylogenetic reconstructions based on the RADseq data were conducted, but were ultimately not included in the manuscript. In this short section, the analyses and their results will be summarized.

### Methods

Gene trees of the concatenated ddRAD locus sequences with three *Rhodanthemum* accessions as outgroup were calculated using IQ-TREE (Minh et al., 2020). IQ-TREE reconstructions were run for 2,000 iterations (-nm 2000) using the general-time-reversible sequence model (-m GTR), with 1,000 bootstrap replicates (-bb 1000). Species-tree reconstruction of the *Leucanthemum* accessions was performed using ASTRAL-III (Zhang et al., 2018). For this purpose, locus-wise gene trees were reconstructed using IQ-Tree (-nm 4000, -bb 1000, -m GTR). Individuals were grouped according to the 20 morpho-taxa.

**Results**

Similar to the Neighbor-Net and PCoA results, the accessions cluster according to their morpho-species membership, the only exceptions being, accessions M02-01 and 68-04 (detected as hybrids by ABBA-BABA tests). Figure 4.5 depicts the maximum likelihood tree calculated based on the concatenated RAD loci, including the outgroup (*Rhodanthemum*). Here, *L. graminifolium* is the first diverging taxon of *Leucanthemum*. The species tree reconstruction using ASTRAL-III revealed the same species topology, the only exception being the placement of *L. ageratifolium* as sister-group to *L. vulgare*, *L. pyrenaicum*, and *L.* gaudinii (Figure 4.6). Only two splits have support values less than 0.95: the split between *L. virgatum* and the remaining ten species comprising the reduced dataset, and the split between *L. burnatii* and *L. tridactylites* + *L. laciniatum*

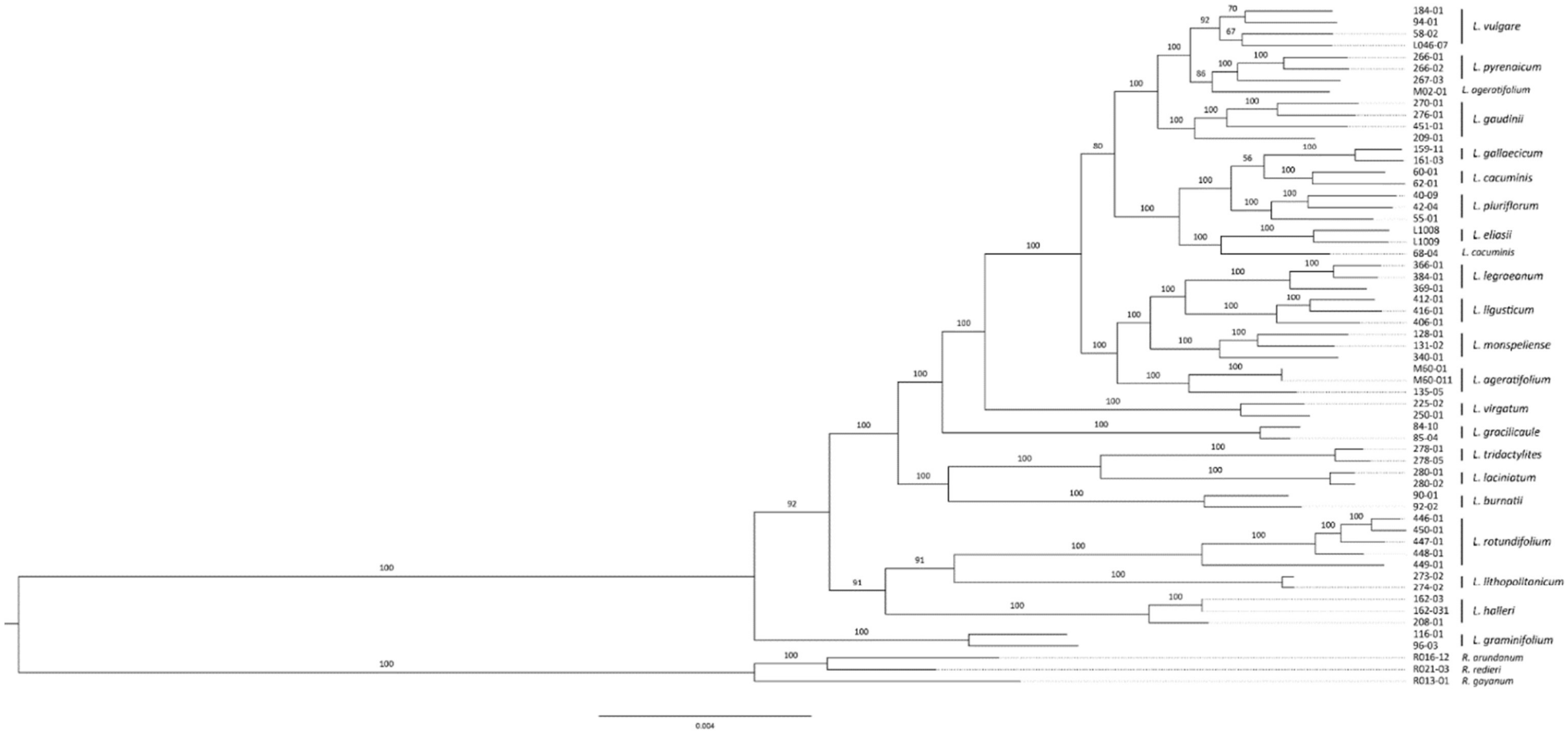


**Figure 4.5**. Maximum Likelihood tree of the concatenated RAD loci of the complete dataset including three *Rhodanthemum* accessions as outgroup, calculated using IQ-TREE. The branch labels represent bootstrap values.

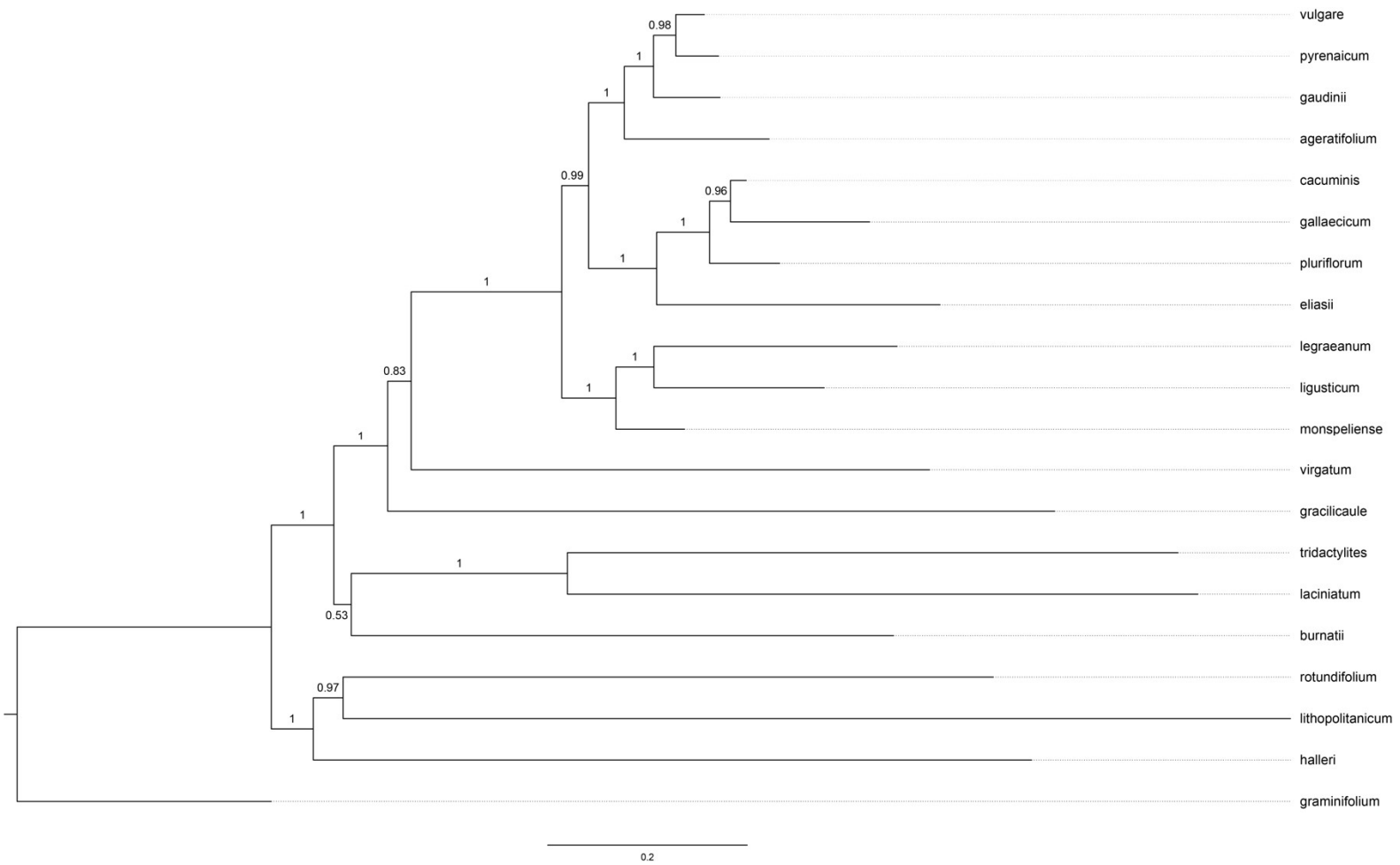


**Figure 4.6**. Species tree reconstructed using ASTRAL-III based on maximum likelihood trees of all RADseq loci. Branch labels represent support scores as returned by ASTRAL.

# Chapter 5: Untangling the Weave

## Untangling the weave: integrative species delimitation of the tetraploid *Leucanthemum* Mill. (Compositae, Anthemideae) representatives

Tankred Ott, Robert Vogt, Christoph Oberprieler

In preparation for publication.

### Abstract

*Leucanthemum* Mill. (Compositae, Anthemideae) is polyploidy complex comprising around 40 plant species, most of them polyploid, with chromosome numbers ranging from 2x (diploid) to 22x (docosaploid). Due to its young age and the high degree of hybridization and polyploidization within the genus, it presents an interesting but at the same time challenging group for studying evolution and species delimitation. Recently, the diploid species, the "warps and wefts", of this polyploidy complex were successfully delimited, but the complex "weave", the taxonomy of the polyploids, has not yet been resolved. Delimiting species, however, especially those of higher ploidy, is a difficult task, but of paramount importance for scientific discourse, nature conservation, and also as fundamental question of systematic research. Modern integrative taxonomy incorporates evidence from a wide range of data sources and methodologies to soundly and objectively delineating taxonomic units. Here, we are building on a methodological pipeline that was proposed and successfully applied for the delimitation of the diploid *Leucanthemum* species and extend it for resolving the taxonomic ranks within the tetraploid representatives of the genus. For this purpose, we combine genealogical, ecological, geographical, and morphological evidence, by means of genetic clustering based on RADseq data, ecological niche modeling and niche equivalency tests, measuring sympatry, and assessing differentiation of leaf shape and dissection based on preserved specimens, respectively. Additionally, we use SNP patterns to detect potential diploid parent species for the delineated tetraploid taxa. Our results suggest that *L. ircutianum* subsp. *leucolepis* and *L. delarbrei* subsp. *ruscinonense* should be raised to the species rank, and that *L. cantabricum* and *L. meridionale* should be treated as subspecies of *L. ircutianum* and *L. delarbrei*, respectively.

## Introduction

The delimitation of species, the fundamental units of biodiversity, is one of the major aims of taxonomy and systematics. An objective species delimitation, while especially difficult in presence of hybridization and polyploidization, is important for studies of ecology, nature conservation and even for law makers. In the present study we are delimiting the tetraploid representatives of the polyploidy complex-forming genus *Leucanthemum* Mill.

The genus *Leucanthemum* Mill. (Compositae, Anthemideae; "Ox-eye daisy") comprises 39 (Chapter 4) to 42 species (*The Euro+Med Plantbase Project*) with ploidy levels ranging from diploid ($2x$) to dodecaploid ($12x$), and one species [*L. lacustre* (Brot.) Samp.) from Portugal] even showing a chromosome number of $2n = 22x = 198$ (docosaploid level). *Leucanthemum* is distributed over the whole European continent and into Northern Asia (the tetraploid *L. ircutianum* DC. in Siberia), while some species were also introduced to the northern and southern hemisphere (Meusel & Jäger, 1992). The current species delimitation of the genus is mostly based on differences in morphology, especially general leaf shape and degree dissection since flower characters are relatively invariant, and ploidy level, as well as geographical distribution (Vogt, 1991, Vogt, 2019). More recently, genealogical characters were used as additional evidence for discriminating taxa (Chapter 4; Wagner et al., 2017).

Previous studies addressed species delimitation and phylogeny of the diploid *Leucanthemum* representatives based on multi-copy nuclear markers (nrDNA ETS), AFLP fingerprints, single-copy nuclear markers (Oberprieler et al., 2014, Oberprieler et al., 2014, Wagner et al., 2017, Wagner et al., 2019). Cloning of nrDNA ETS amplicons revealed that some diploid taxa exclusively possess a plesiomorphic ETS ribotype cluster closely related to ETS ribotypes of the outgroup and others are characterized by the exclusive possession of an apomorphic ETS ribotype cluster, while a third group of taxa exhibit an additive pattern of the two types (Oberprieler et al., 2014). This finding was supported by Konowalik et al. (2015) using AFLP fingerprinting and multilocus species-tree reconstructions based on single-copy markers of the nuclear genome, and it was demonstrated that the species with the plesiomorphic ETS ribotypes form an early-diverging paraphyletic grade, in which the monophyletic group of taxa with the apomorphic ETS ribotypes are nested. Furthermore, the authors applied coalescent-based simulations to distinguish between hybridization and incomplete lineage sorting (ILS), revealing that

for most of the diploid taxa involved (and especially for those of the second group) incongruence among gene trees could not be explained by ILS alone and recent hybridization or even homoploid hybrid speciation events must be assumed. Consequently, some infraspecific taxa were raised to species level to account for their assumed independent formation through homoploid hybrid speciation (i.e., *L. cacuminis*, *L. eliasii*, and *L. pyrenaicum*). Species delimitation in a morphologically closely-knit group of taxa in the clade with the apomorphic nrDNA ETS ribotypes was subsequently carried out by F. Wagner et al. (2017) who used AFLP fingerprinting, sequence information from plastid and nuclear single-copy markers, and coalescent-based Bayesian delimitation methods to infer species boundaries in the *L. ageratifolium*-group, despite the frequent presence of hybrid individuals in this group. Finally, Wagner et al. (2019) presented a multi-locus phylogenetic reconstruction of the subtribe Leucantheminae, in which the diversification among diploid *Leucanthemum* species was dated to the last 1.93 (1.14-2.94) Ma, arguing for the strong influence of Pleistocene oscillations on species formation in this genus. Most recently restriction site-associated DNA sequencing (RADseq), ecological niche modeling (ENM), geographical patterns, and morphometrics were combined for integrative species delimitation of the diploid *Leucanthemum* representatives, resulting in the proposal of a new classification for five *Leucanthemum* (Chapter 4).

In contrast to the extensive investigation of the diploids, the phylogeny and evolution of only a few tetraploid representatives has been subject of previous studies. Oberprieler et al. (2011) found indications for an allopolyploid origin of the tetraploid *L. ircutianum* subsp. *ircutianum* based on AFLP markers. One year later, Greiner & Oberprieler (2012) showed that the tetraploid *L. pseudosylvaticum* is able to form fertile offspring when crossed with its next diploid relative, *L. pluriflorum*, suggesting a relatively recent diversification. Another study of the *L. pluriflorum*-group by Greiner et al. (2013) gave evidence for the allopolyploid and autopolyploid origin of *L. pseudosylvaticum* and *L. corunnense*, respectively, and raised *L. pseudosylvaticum*, which was formerly a subspecies of *L. ircutianum* to the species rank. Finally, the most recent contribution to the taxonomy of the tetraploid *Leucanthemum* representatives by Oberprieler et al. (2022) affirmed the infraspecific rank of *L. ircutianum* subsp. *ircutianum* and *L. ircutianum* subsp. *leucolepis* based on a set of few nuclear and plastid DNA, and AFLP markers.

Building on the previous research of taxonomy and phylogeny of the diploid and tetraploid Leucanthemum representatives, the present study focuses on the species

delimitation of eight of the nine tetraploid *Leucanthemum* taxa, which we will each introduce briefly:

*Leucanthemum ircutianum* DC. subsp. *ircutianum*

Besides the diploid *L. vulgare* Lam., the most widely distributed species of the genus (Vogt, 1991); present in nearly all countries of the Euro + Med region (*The Euro+Med Plantbase Project*), but also introduced to all continents except Antarctica. Habitats are anthropogenetically influenced and include meadows and roadsides. *L. ircutianum* is morphologically very similar to *L. ircutianum* subsp. *leucolepis* but can, at least to a degree (Oberprieler et al., 2022), be differentiated by the dark-brown hyaline margins of involucral bracts.

*Leucanthemum ircutianum* subsp. *leucolepis* (Briq. & Cavill.) Vogt & Greuter [≡*L. leucolepis* (Briq. & Cavill.) Gajić]

A taxon with an amphi-Adriatic distribution with populations throughout Italy and along the Adriatic coast of the Balkan Peninsula. The taxon was described as a subspecies of the diploid *L. vulgare* but was subsequently considered a subspecies of the hexaploid *L. pallens* due to its white hyaline margins of involucral bracts or seen as an independent species. The observation of an allopatric distribution with *L. ircutianum* subsp. *ircutianum* as more Mediterranean facies of the species, together with the observation of hybrid individuals and hybrid swarms in areas of joint occurrence of the two tetraploid taxa in the Apennine and Balkan Peninsulas by Oberprieler et al. (2022) argued for a treatment as a subspecies of *L. ircutianum*. In contrast to *L. ircutianum* subsp. *ircutianum*, it has paler hyaline margins of the involucral bracts.

*Leucanthemum cantabricum* Sennen

Distributed between the western foothills of the Pyrenees (Spain and France), along the northern slopes of the Cantabrian Mts to Asturias and Galicia in the west. Habitats include road and meadow margins, meadows, and pastures; the elevational distribution ranges from sea level to 800 m. *Leucanthemum cantabricum* is characterized by dissected leaves (pinnatisect to pinnatipartite), non-succulent leaves, and dark-brown hyaline margins of involucral bracts. It was described as a species by Sennen (1936), but reduced

to subspecific rank under *L. ircutianum* by Vogt (1991) before being re-acknowledged at species rank by Vogt (2019).

*Leucanthemum crassifolium* (Lange) Lange

Endemic to the Cantabrian coast between Asturias and the Basque region (NE Spain and SW France). Habitats are coastal rocks and salt-influenced coastal slopes from sea level to 20 m. *Leucanthemum crassifolium* is characterized by succulent leaves and involucral bracts with dark-brown hyaline margins. It was first described as a variety of *L. pallens* by Lange (1861) and subsequently raised to subspecific (Vogt, 1991) and specific rank (Vogt, 2019; F. Wagner et al., 2019).

*Leucanthemum pseudosylvaticum* (Vogt) Vogt & Oberpr.

Distributed throughout the western part of the Iberian Peninsula (Portugal and Spain). Its habitats comprise road margins and slopes, margins of creeks and ditches, at an elevational range from 100 m to 1600 m. *Leucanthemum pseudosylvaticum* is characterized by leaves with proximally entire margins and involucral bracts with pallid to light-brown hyaline margins. It was described by (Vogt, 1991) as a subspecies of *L. ircutianum* but shown to merit species rank due to its evolutionary independence from the latter species by Greiner et al. (2012, 2013).

*Leucanthemum delarbrei* Timb.-Lagr. subsp. *delarbrei*

Endemic to the siliceous rocks of the old volcano cones of the central Massif Central in France (Monts Dore, Monts du Cantal); its habitats are siliceous rocks and meadows between 1550 m and 1750 m. *Leucanthemum delarbrei* subsp. *delarbrei* is characterized by strongly dissected leaves, from pinnatifid to bipinnatisect, and dark- to light-brown hyaline margins.

*Leucanthemum delarbrei* subsp. *ruscinonense* (Jeanb. & Timb.-Lagr.) Vogt, Florian Wagner & Oberpr.

Endemic to the eastern Pyrenees (Spain and France) and the SW parts of the Massif Central (Haut Languedoc, Montagne Noire). Habitats are comprising road margins, creeks and stony slopes, between 250 m and 1900 m. It is characterized by strongly dissected leaves, from pinnatifid to bipinnatisect, and dark- to light-brown hyaline margins. A high variability in terms of leaf dissection is speculated to be the result of potential hybridization.

The taxon was, for a long time, considered being part of *L. monspeliense* L. (e.g., Vogt, 1991), which is also distributed in the Massif Central (Cevennes), but has a diploid chromosome number.

*Leucanthemum meridionale* Legrand (1881).

Endemic to the Puy de Wolf (France, Aveyron, Rodez, Decaseville, Firmi), a serpentine mountain in the western Massif Central. It is found in dry and open, South-facing grassland on serpentine soil between 400 m and 600 m. *Leucanthemum meridionale* is characterized by its lanky habitus resembling the diploid *L. vulgare* or a member of the genus *Leucanthemopsis* with quite narrow leaves and cuneate leaf-bases.

The previously described eight species are part of the *Leucanthemum* polyploidy complex, whose taxonomy on the diploid level was already resolved in chapter 4; a similarly thorough study was not yet done for higher ploidy levels. Attaining an equally sound resolution of taxonomic relationships within the tetraploid taxa will facilitate future studies on the phylogeny and origin of the polyploid representatives of the genus. Delimiting polyploid species, however, is “notoriously difficult” due to the ”complexity of processes and different evolutionary scenarios that do not fit with classical species concepts” (Hörandl, 2022). A tetraploid species can originate from a single diploid species (autopolyploidization) or may from after a hybridization event of two or more diploid species (allopolyploidization). Tetraploid species can also form by homoploid speciation with or without hybridization on the tetraploid level. Of course, those processes can also happen conjointly or multiple times independently, rendering species delimitation particularly complex (see Hörandl, 2022).

Here, we will unravel the tetraploid *Leucanthemum* taxa, employing genealogical, ecological, geographical and morphometrical analyses by means of RADseq (Poland et al., 2012), ecological niche modeling, analysis of spatial distribution patterns, and geometric morphometry, respectively. For the genealogical layer, we will apply Weighted Ensemble of Random (k)k-Means (Lai et al., 2019), a semi-supervised clustering method, for finding genetic units, use SNiPLOID (Peralta et al., 2013) for searching SNP patterns that might help to unravel the diploid origin of the tetraploid taxa. Ecological niches will be compared using niche equivalency tests (Cardillo & Warren, 2016), and the overlap of spatial distribution areas will be analyzecx based on approximated distribution areas as proposed in chapter 4. For the morphological layer, the general leaf shape, and the degree of leaf

dissection, two important features for the current species delimitation within *Leucanthemum*, will be compared based on elliptic Fourier analyses (Kuhl & Giardina, 1982) and leaf dissection indices (Kincaid & Schneider, 1983). We will integrate all four sources of evidence to inform the taxonomic treatment of the taxa under study as objective as possible, laying the foundation for research on origin and evolution of the polyploid *Leucanthemum* representatives.

## Material and Methods

### RADseq Assembly

Double Digest RADseq (ddRADseq; Cardillo & Warren, 2016) was conducted based on a dataset comprising 51 individuals from all 17 diploid *Leucanthemum* species and 35 individuals from 8 different tetraploid *Leucanthemum* taxa (see Table 5.1), the latter being the focal group of the present study. ddRADseq reads of the diploid samples were collected from chapter 4. For the tetraploid individuals, genomic DNA was extracted from silica-dried specimens according to the CTAB DNA extraction protocol of Doyle & Dickson (1987). For assembly quality assessment, replicates were generated by repeating the extraction for two samples (*L. ircutianum* subsp. *leucolepis*: 170-02, 170-021; *L. ircutianum* DC. subsp. *lrcutianum*: L055-03, L055-031). ddRAD Illumina sequencing (2x150bp; NextSeq 500, Illumina Inc., San Diego, CA, USA) using the restriction enzymes *Pst1* and *ApeK1*, including demultiplexing and adapter clipping, was conducted by LGC Genomics (Berlin, Germany).

Demultiplexed and adapter clipped reads from the diploid and tetraploid samples were assembled using IPYRAD v0.9.81(Eaton & Overcast, 2020) against the reference constructed in Chapter 4; the minimum number of samples per locus (min_samples_locus) was set to 20, while the remaining parameters were kept at default values. For detection of assembly errors, locus, allele, and SNP error rates were calculated (Mastretta-Yanes et al., 2015; see chapter 4).

### Network Analysis

Neighbor nets based on SNP-level Nei distances (Joly & Bruneau, 2006; https://github.com/simjoly/pofad) were calculated using SPLITSTREE4 v4.15.1 (Huson & Bryant, 2006) for the full dataset, the diploid samples within the *L. vulgare*-group (see chapter 4) and all tetraploid samples, and for the tetraploids only. Distances were calculated

based on the variant output (VCF file) of IPYRAD using a custom Python and C tool (https://github.com/TankredO/nei_vcf). With SNP-level Nei distances, differences in variant (SNP) frequencies are directly included in the distance calculation, rendering this kind of distance metric particularly useful for comparing diploid and polyploid samples.

**Table 5.1**. List of samples used for the ddRAD analysis with information on voucher specimens in the botanical museum Berlin and ploidy, collection localities, coordinates, and collectors.

| Sample | Taxon | Ploidy | Voucher Specimens | Locality | Coordinates (latitude, longitude) | Collector |
|---|---|---|---|---|---|---|
| 135-05 | *L. ageratifolium* Pau | 2x | B100386712 | FR, Occitania, Pyrénées-Orientales, 410 m | 42.5038, 2.9603 | *Konowalik KK42 & Ogrodowczyk* |
| M60-01 | *L. ageratifolium* Pau | 2x | B100345012, B100345013 | ES, Castile-La Mancha, Cuenca, 1157 m | 40.1019, -1.521 | *Cordel 60* |
| M60-011 | *L. ageratifolium* Pau | 2x | B100345012, B100345013 | ES, Castile-La Mancha, Cuenca, 1157 m | 40.1019, -1.521 | *Cordel 60* |
| 90-01 | *L. burnatii* Briq. & Cavill. | 2x | B100464678 | FR, Provence-Alpes-Côte d'Azur, Alpes-Maritimes, 1235 m | 43.7607, 6.9165 | *Vogt 16615 et al.* |
| 92-02 | *L. burnatii* Briq. & Cavill. | 2x | B100464676, B100464675 | FR, Provence-Alpes-Côte d'Azur, Bouches-du-Rhône, 650 m | 43.5450, 5.6626 | *Vogt 16618 et al.* |
| 60-01 | *L. pluriflorum* subsp. *cantabricum* (Font Quer & Guinea) T.Ott | 2x | B100413746 | ES, Galicia, Os Ancares, 1530 m | 42.8315, -6.8569 | *Höβl 60* |
| 62-01 | *L. pluriflorum* subsp. *cantabricum* (Font Quer & Guinea) T.Ott | 2x | B100413744 | ES, Galicia, Lugo, 750 m | 42.9249, -6.8657 | *Höβl 62* |
| 68-04 | *L. pluriflorum* subsp. *cantabricum* (Font Quer & Guinea) T.Ott | 2x | B100413738 | ES, Cantabria, Cantabria, 1770 m | 43.1538, -4.8053 | *Höβl 68 & Himmelreich* |
| L1008 | *L. eliasii* (Sennen & Pau) Sennen & Pau | 2x | B100484003 | ES, Castile and León, Burgos, 880 m | 42.5030, -3.7060 | *Lopéz 2537 & al.* |
| L1009 | *L. eliasii* (Sennen & Pau) Sennen & Pau | 2x | B100484004 | ES, Castile and León, Burgos, 920 m | 42.5070, -3.7050 | *Galán Cela 576 & Martín* |
| 159-11 | *L. pluriflorum* subsp. *gallaecicum* (Rodr.Oubiña & S.Ortiz) T.Ott, Vogt & Oberpr. | 2x | B100386789, B100420775, B100464989 | ES, Galicia, Pontevedra, 375 m | 42.8498, -7.9878 | *Konowalik KK67 & Ogrodowczyk* |
| 161-03 | *L. pluriflorum* subsp. *gallaecicum* (Rodr.Oubiña & S.Ortiz) T.Ott, Vogt & Oberpr. | 2x | No voucher | ES, Galicia, Corunna, 380 m | 42.8533, -7.9994 | *Konowalik s.n. & al.* |
| 209-01 | *L. gaudinii* Dalla Torre | 2x | B100386664 | CH, Bern, Interlaken-Oberhasll, 2260 m | 46.5781, 7.9700 | *Tomasello TS88* |
| 270-01 | *L. gaudinii* Dalla Torre | 2x | B100413007 | AT, Carinthia, Spittal an der Drau, 2200 m | 47.0025, 13.5275 | *Oberprieler 10859* |
| 276-01 | *L. gaudinii* Dalla Torre | 2x | B100413015 | AT, Carinthia, Feldkirchen, 2270 m | 46.8603, 13.8172 | *Oberprieler 10866* |
| 451-01 | *L. gaudinii* Dalla Torre | 2x | No voucher | PL, Lesser PolanDE, Giewont, 1860 m | 49.2505, 19.9343 | *Konowalik 20160909-01* |
| 84-10 | *L. gracilicaule* (Dufour) Pau | 2x | B100386704 | ES, Valencian Community, Alicante, 300 m | 38.8379, -0.1853 | *Konowalik KK20 & Ogrodowczyk* |
| 85-04 | *L. gracilicaule* (Dufour) Pau | 2x | B100386702 | ES, Valencian Community, Valencia, 340 m | 39.3135, -0.6810 | *Konowalik KK25 & Ogrodowczyk* |

| Sample | Taxon | Ploidy | Voucher Specimens | Locality | Coordinates (latitude, longitude) | Collector |
|---|---|---|---|---|---|---|
| 116-01 | *L. graminifolium* (L.) Lam. | 2x | B100464684, B100464683 | FR, Occitania, Hérault, 800 m | 43.7761, 3.2386 | *Vogt 16693 & al.* |
| 96-03 | *L. graminifolium* (L.) Lam. | 2x | B100464663 | FR, Occitania, Aude, 600 m | 43.1494, 2.6294 | *Vogt 16656 & al.* |
| 162-03 | *L. halleri* (Vitman) Ducommun | 2x | B100386798 | DE, Bavaria, Landkreis Garmisch-Partenkirchen, 2340 m | 47.4134, 11.1277 | *Konowalik KK67 & Tomasello* |
| 162-031 | *L. halleri* (Vitman) Ducommun | 2x | B100386798 | DE, Bavaria, Landkreis Garmisch-Partenkirchen, 2340 m | 47.4134, 11.1277 | *Konowalik KK67 & Tomasello* |
| 208-01 | *L. halleri* (Vitman) Ducommun | 2x | B100386672 | FR, Valais, Sion, 2320 m | 46.3308, 7.2911 | *Tomasello TS65* |
| 280-01 | *L. laciniatum* Huter & al. | 2x | B100464203 | IT, Calabria, Cosenza, 1580 m | 39.9020, 16.1144 | *Tomasello 420* |
| 280-02 | *L. laciniatum* Huter & al. | 2x | B100464203 | IT, Calabria, Cosenza, 1580 m | 39.9020, 16.1144 | *Tomasello 420* |
| 366-01 | *L. legraeanum* (Rouy) B.Bock & J.-M.Tison | 2x | B100486634, B100486635, B100486636, B100486637, B100486638 | FR, Provence-Alpes-Cote d'Azur, Var, 410 m | 43.1986, 6.3151 | *Vogt 17189* |
| 369-01 | *L. legraeanum* (Rouy) B.Bock & J.-M.Tison | 2x | B100486648, B100486649 | FR, Provence-Alpes-Cote d'Azur, Var, 210 m | 43.2444, 6.3377 | *Vogt 17192* |
| 384-01 | *L. legraeanum* (Rouy) B.Bock & J.-M.Tison | 2x | B100627809, B100627810 | FR, Provence-Alpes-Cote d'Azur, Var, 410 m | 43.1988, 6.3151 | *Vogt 17434 & al.* |
| 406-01 | *L. ligusticum* Marchetti & al. | 2x | B100627838, B100627839 | IT, Liguria, La Spezia, 210 m | 44.2470, 9.7728 | *Vogt 17460 & al.* |
| 412-01 | *L. ligusticum* Marchetti & al. | 2x | B100627849, B100627850, B100627851 | IT, Liguria, Genova, 700 m | 44.3603, 9.5105 | *Vogt 17468 & al.* |
| 416-01 | *L. ligusticum* Marchetti & al. | 2x | B100627855, B100627856 | IT, Liguria, Genova, 250 m | 44.3458, 9.4588 | *Vogt 17471 & al.* |
| 273-02 | *L. lithopolitanicum* | 2x | B100413012 | SL, Central Slovenia, 2100 m | 46.3633, 14.5715 | *Oberprieler 10862* |
| 274-02 | *L. lithopolitanicum* | 2x | B100413013 | SL, Savinja, 2000 m | 46.375, 14.5663 | *Oberprieler 10864* |

| Sample | Taxon | Ploidy | Voucher Specimens | Locality | Coordinates (latitude, longitude) | Collector |
|---|---|---|---|---|---|---|
| 128-01 | *L. monspeliense* (E.Mayer) Polatschek | 2x | B100464618 | FR, Occitania, GarDE, 750 m | 44.0888, 3.5786 | *Vogt 16712 & al.* |
| 131-02 | *L. monspeliense* (E.Mayer) Polatschek | 2x | B100464615 | FR, Occitania, GarDE, 380 m | 44.1412, 3.7316 | *Vogt 16716 & al.* |
| 340-01 | *L. monspeliense* (E.Mayer) Polatschek | 2x | B100486666, B100486667 | FR, Occitania, Aveyron, 180 m | 44.5822, 2.1840 | *Vogt 17156 & al.* |
| 40-09 | *L. pluriflorum* Pau | 2x | B100413758 | ES, Galicia, Corunna, 100 m | 42.8838, -9.2726 | *Höβl 40* |
| 42-04 | *L. pluriflorum* Pau | 2x | No voucher | ES, Galicia, Corunna, 150 m | 43.3069, -8.6186 | *Höβl 42* |
| 55-01 | *L. pluriflorum* Pau | 2x | B100413749 | ES, Galicia, Lugo, 10 m | 43.6309, -7.3330 | *Höβl 55* |
| M02-01 | *L. pyrenaicum Vogt & al.* | 2x | B100297950 | ES, Aragón, Huesca, 755 m | 42.5250, -0.6690 | *Cordel 2* |
| 446-01 | *L. rotundifolium* (Willd.) DC. | 2x | No voucher | PL, Podkarpackie, Bieszczady, 920 m | 49.1191, 22.5776 | *Konowalik 20180622-02-01* |
| 447-01 | *L. rotundifolium* (Willd.) DC. | 2x | No voucher | RO, Bihor, Bihor, 1230 m | 46.5189, 22.6613 | *Konowalik 20180713-03-01* |
| 448-01 | *L. rotundifolium* (Willd.) DC. | 2x | No voucher | BH, Central Bosnia Canton, 1860 m | 43.95782, 17.7403 | *Konowalik 20180714-03-01* |
| 449-01 | *L. rotundifolium* (Willd.) DC. | 2x | No voucher | RO, Hunedoara, Râu de MorIT, 1140 m | 45.3159, 22.7705 | *Konowalik 20180807-03-01* |
| 450-01 | *L. rotundifolium* (Willd.) DC. | 2x | No voucher | PL, Lesser Poland Voivodeship, Sucha County, 1100 m | 49.5879, 19.5515 | *Konowalik 20170920-01* |
| 278-01 | *L. tridactylites* (A.Kern. & Huter) Huter & al. | 2x | B100464207 | IT, Abruzzo, Pescara, 2080 m | 42.1384, 14.1101 | *Tomasello 417* |
| 278-05 | *L. tridactylites* (A.Kern. & Huter) Huter & al. | 2x | B100464207 | IT, Abruzzo, Pescara, 2080 m | 42.1384, 14.1101 | *Tomasello 417* |
| 225-02 | *L. virgatum* (Desr.) Clos | 2x | B100411746 | FR, Provence-Alpes-Côte d'Azur, Alpes-Maritimes, 430 m | 43.9538, 7.2961 | *Vogt 16892 & Oberprieler* |
| 250-01 | *L. virgatum* (Desr.) Clos | 2x | B100350169, B100350172 | IT, Liguria, Savona, 220 m | 44.0596, 8.0583 | *Vogt 16932 & Oberprieler* |
| 184-01 | *L. vulgare* Lam. | 2x | B100346626 | BH, Republic of Srpska, Nevesinje, 930 m | 43.2403, 18.3364 | *Vogt 16806 & Prem-Vogt* |
| 58-02 | *L. vulgare* Lam. | 2x | B100413748 | ES, Galicia, Lugo, 490 m | 42.8205, -7.9504 | *Höβl 58* |
| 94-01 | *L. vulgare* Lam. | 2x | B100464674 | FR, Occitania, Aude, 160 m | 43.1294, 2.6073 | *Vogt 16641 & al.* |
| L046-07 | *L. vulgare* Lam. | 2x | B100550249 | DE, Bavaria, Deuerling, 450 m | 49.0333, 11.8833 | *Eder & Oberprieler s.n.* |
| 266-01 | *L. vulgare* subsp. *barrelieri* (Dufour ex DC.) O.Bolós & Vigo | 2x | B100464208 | ES, Aragon, Huesca, 1650 m | 42.7806, -0.2467 | *Tomasello TS382* |
| 266-02 | *L. vulgare* subsp. *barrelieri* (Dufour ex DC.) O.Bolós & Vigo | 2x | B100464208 | ES, Aragon, Huesca, 1650 m | 42.7806, -0.2467 | *Tomasello TS382* |
| 267-03 | *L. vulgare* subsp. *barrelieri* (Dufour ex DC.) O.Bolós & Vigo | 2x | B100464210 | ES, Aragon, Huesca, 2000 m | 42.6327, 0.4530 | *Tomasello TS392* |

| Sample | Taxon | Ploidy | Voucher Specimens | Locality | Coordinates (latitude, longitude) | Collector |
|---|---|---|---|---|---|---|
| 45-01 | *L. cantabricum* Sennen | 4x | B100413756 | ES, Galicia, O Eume, 350 m | 43.3972, -8.0959 | *Hößl 45* |
| 50-01 | *L. cantabricum* Sennen | 4x | B100413752 | ES, Galicia, Ortegal, 10 m | 43.6749, -7.9022 | *Hößl 50* |
| 50-02 | *L. cantabricum* Sennen | 4x | B100413752 | ES, Galicia, Ortegal, 10 m | 43.6749, -7.9022 | *Hößl 50* |
| 50-03 | *L. cantabricum* Sennen | 4x | B100413752 | ES, Galicia, Ortegal, 10 m | 43.6749, -7.9022 | *Hößl 50* |
| 66-01 | *L. crassifolium* (Lange) Lange | 4x | B100413740 | ES, Asturias, Asturias, 60 m | 43.6565, -5.8510 | *Hößl 66 & Himmelreich* |
| 66-02 | *L. crassifolium* (Lange) Lange | 4x | B100413740 | ES, Asturias, Asturias, 60 m | 43.6565, -5.8510 | *Hößl 66 & Himmelreich* |
| 69-01 | *L. crassifolium* (Lange) Lange | 4x | B100413737 | ES, Cantabria, Cantabria, 50 m | 43.3943, -4.3571 | *Hößl 69 & Himmelreich* |
| 69-02 | *L. crassifolium* (Lange) Lange | 4x | B100413737 | ES, Cantabria, Cantabria, 50 m | 43.3943, -4.3571 | *Hößl 69 & Himmelreich* |
| 101-01 | *L. delarbrei* subsp. *ruscinonense* (Jean. & Timb.-Lagr.) Vogt | 4x | B100464658, B100464657 | FR, Occitania, Aude, 193 m | 43.3279, 2.3835 | *Vogt 16670 & al.* |
| 110-01 | *L. delarbrei* subsp. *ruscinonense* (Jean. & Timb.-Lagr.) Vogt | 4x | B100464637 | FR, Occitania, Hérault, 214 m | 43.5786, 3.0212 | *Vogt 16685 & al.* |
| 139-01 | *L. delarbrei* subsp. *ruscinonense* (Jean. & Timb.-Lagr.) Vogt | 4x | B100386786 | ES, Catalonia, Girona, 1015 m | 42.412, 2.7504 | *Konowalik KK46 & Ogrodowczyk* |
| 349-01 | *L. delarbrei* subsp. *ruscinonense* (Jean. & Timb.-Lagr.) Vogt | 4x | B100486690, B100486691, B100486692, B100486693 | FR, Occitania, Tarn, 502 m | 43.6236, 2.4984 | *Vogt 17171 & al.* |
| 355-03 | *L. delarbrei* subsp. *ruscinonense* (Jean. & Timb.-Lagr.) Vogt | 4x | B100486699, B100486700 | FR, Occitania, Hérault, 182 m | 43.7384, 3.3283 | *Vogt 17177 & al.* |
| 424-01 | *L. delarbrei* Timb.-Lagr. subsp. *delarbrei* | 4x | B100605498, B100605497, B100605489 | FR, Auvergne-Rhone-Alpes, Cantal, 1688 m | 45.0569, 2.7461 | *Vogt 17487 & Prem-Vogt* |
| 425-01 | *L. delarbrei* Timb.-Lagr. subsp. *delarbrei* | 4x | B100605488, B100605487, B100605486, B100605501 | FR, Auvergne-Rhône-Alpes, Cantal, 1553 m | 45.1050, 2.6847 | *Vogt 17489 & Prem-Vogt* |
| 427-01 | *L. delarbrei* Timb.-Lagr. subsp. *delarbrei* | 4x | B100605494 | FR, Auvergne-Rhone-Alpes, Puy-de-Dôme, 1818 m | 45.5286, 2.8128 | *Vogt 17494 & Prem-Vogt* |

| Sample | Taxon | Ploidy | Voucher Specimens | Locality | Coordinates (latitude, longitude) | Collector |
|---|---|---|---|---|---|---|
| 428-01 | *L. delarbrei* Timb.-Lagr. subsp. *delarbrei* | 4x | B100605492, B100603419 | FR, Auvergne-Rhone-Alpes, Puy-de-Dôme, 1457 m | 45.5364, 2.8125 | *Vogt 17495 & Prem-Vogt* |
| 106-01 | *L. ircutianum* DC. subsp. *ircutianum* | 4x | B100464641 | FR, Occitania, Tarn, 410 m | 43.4816, 2.3717 | *Vogt 16678 & al.* |
| 177-01 | *L. ircutianum* DC. subsp. *ircutianum* | 4x | B100346630 | ME, Old Royal Capital Cetinje, 920 m | 42.3681, 18.8903 | *Vogt 16794 & Prem-Vogt* |
| 343-01 | *L. ircutianum* DC. subsp. *ircutianum* | 4x | B100486672, B100486673 | FR, Occitania, Aveyron, 470 m | 44.5589, 2.3114 | *Vogt 17159 & al.* |
| 437-01 | *L. ircutianum* DC. subsp. *ircutianum* | 4x | B101003356 | FR, Corsica, Corse-du-SuDE, 920 m | 41.9727, 9.1843 | *Vogt 17865 & al.* |
| 87-01 | *L. ircutianum* DC. subsp. *ircutianum* | 4x | B100464680 | IT, Piemont, Cuneo, 670 m | 44.2982, 7.5052 | *Vogt 16611 & al.* |
| L052-02 | *L. ircutianum* DC. subsp. *ircutianum* | 4x | No voucher | DE, Bavaria, Regensburg, 385 m | 48.9833, 12.0833 | *Eder s.n.* |
| L055-03 | *L. ircutianum* DC. subsp. *ircutianum* | 4x | B100627297 | DE, Bavaria, Landkreis Garmisch-Partenkirchen, 1430 m | 47.5833, 11.1167 | *Eder & Oberprieler 10298* |
| L055-031 | *L. ircutianum* DC. subsp. *ircutianum* | 4x | B100627297 | DE, Bavaria, Landkreis Garmisch-Partenkirchen, 1430 m | 47.5833, 11.1167 | *Eder & Oberprieler 10298* |
| L062-04 | *L. ircutianum* DC. subsp. *ircutianum* | 4x | B100416530, B100627403 | AT, Vorarlberg, Bregenz, 1380 m | 47.3000, 10.1000 | *Eder, Oberprieler 10304a & Vogt 16222* |
| 170-02 | *L. ircutianum* subsp. *leucolepis* (Briq & Cavill.) Vogt & Greuter | 4x | B100346645 | ME, Herceg Novi Municipality, 34 m | 42.4742, 18.4731 | *Vogt 16724 & Prem-Vogt* |
| 170-021 | *L. ircutianum* subsp. *leucolepis* (Briq & Cavill.) Vogt & Greuter | 4x | B100346645 | ME, Herceg Novi Municipality, 34 m | 42.4742, 18.4731 | *Vogt 16724 & Prem-Vogt* |
| 174-01 | *L. ircutianum* subsp. *leucolepis* (Briq & Cavill.) Vogt & Greuter | 4x | B100346612, B100346613 | ME, Old Royal Capital Cetinje, 900 m | 42.4389, 18.8319 | *Vogt 16790 & Prem-Vogt* |
| 329-02 | *L. ircutianum* subsp. *leucolepis* (Briq & Cavill.) Vogt & Greuter | 4x | B100464093, B100464094 | IT, Abruzzo, Pescara, 125 m | 42.2447, 13.9659 | *Oberprieler 10870* |
| 371-01 | *L. ircutianum* subsp. *leucolepis* (Briq & Cavill.) Vogt & Greuter | 4x | B100486717 | HR, , Primorje-Gorski Kotar County, 770 m | 45.2772, 14.1944 | *Vogt 17198 & Prem-Vogt* |
| 323-03 | *L. meridionale* Legrand | 4x | B100464096, B100458572 | FR, Occitania, Aveyron, 388 m | 44.5514, 2.3044 | *Oberprieler 10867* |
| 323-04 | *L. meridionale* Legrand | 4x | B100464096, B100458572 | FR, Occitania, Aveyron, 388 m | 44.5514, 2.3044 | *Oberprieler 10867* |
| 323-05 | *L. meridionale* Legrand | 4x | B100464096, B100458572 | FR, Occitania, Aveyron, 388 m | 44.5514, 2.3044 | *Oberprieler 10867* |

| Sample | Taxon | Ploidy | Voucher Specimens | Locality | Coordinates (latitude, longitude) | Collector |
|---|---|---|---|---|---|---|
| 7-02 | *L. pseudosylvaticum* (Vogt) Vogt & Oberpr. | 4x | B100413784 | PT, Bragança District, 350 m | 41.5582, -6.6754 | *Hößl 7 &Hutschenreuther* |
| 1-05 | *L. pseudosylvaticum* (Vogt) Vogt & Oberpr. | 4x | B100413790 | ES, Galicia, Lugo, 1100 m | 42.7267, -7.0258 | *Hößl 1 & Hutschenreuther* |
| 3-08 | *L. pseudosylvaticum* (Vogt) Vogt & Oberpr. | 4x | B100413788 | ES, Galicia, Ourense, 470 m | 42.4634, -6.8950 | *Hößl 3 & Hutschenreuther* |

**Consensus Clustering**

Weighted Ensemble of Random (k)k-Means (WECR; Lai et al., 2019) was applied to detect clusters of genetically similar individuals within the tetraploid *Leucanthemum* representatives. WECR is a semi-supervised consensus (ensemble) clustering technique, allowing to a-priori define fuzzy pair-wise must-link and must-not-link constraints. Roughly, WECR fits a set number of k-Means clusterings (e.g., 1000), each on a random subset of features ("variables") and data points ("samples"), with a random number of clusters *k*. For each k-Means run, clustering-level and cluster-level consistencies are calculated, which measure deviations from the fuzzy constraints concerning the whole clustering and single clusters, respectively. In addition, the mean silhouette index is calculated as internal clustering quality metric. The three values are non-linearly combined to obtain a clustering weight, i.e., a value determining the clustering quality. The clustering co-association matrix, the matrix storing which samples were clustered together (this can be thought of as a similarity matrix with only 1 and 0 entries; 1 if the samples were placed in the same cluster and 0 otherwise), is multiplied by the clustering weight. All weighted co-association matrices are summed (and optionally scaled) to obtain a consensus (co-association) matrix. Finally, a consensus clustering is calculated based on this consensus matrix using hierarchical or spectral clustering.

For clustering of the tetraploid *Leucanthemum* representatives, the previously calculated SNP-level Nei distances (see Network Analysis) were subjected to a principal coordinate analysis (PCoA) with Lingoes' (1971) negative eigenvalue correction and the first principal coordinates (PCo) explaining at least 50% of the variance were selected. Based on the PCo scores, WKM was applied with must-link constraints (MLC) for the individuals of each of the taxa *L. cantabricum*, *L. crassifolium*, *L. delarbrei* subsp. *delarbrei*, *L. delarbrei* subsp. *ruscinonense*, *L. ircutianum* subsp. *leucolepis*, and L. *meridionale*, and must-not-link constraints (MNLC) among individuals from *L. delarbrei* subsp. *delarbrei* and *L. meridionale*. The MLC were set so that individuals of the same taxon are more likely to be clustered together, while the MNLC was used to encourage the algorithm to try to cluster the two genetically close taxa *L. delarbrei* subsp. *delarbrei* and *L. meridionale* with other taxa. The fraction of random features and samples was set to 0.60 and 0.80, respectively. The number of k-Means runs was set to 5,000, and *k* was allowed to take values from two to 14, including the lower and upper bounds. Finally, the consensus matrix was subjected to average-linkage hierarchical clustering for each of the *k* values,

and the clustering scoring best according to the Bayesian Information Criterion (BIC) and the Calinski-Harabasz criterion (CH) was selected. All analyses were performed using pyckmeans v0.9.4 (https://github.com/TankredO/pyckmeans).

**Ecological Niche Modeling**

Eco-climatological and edaphic niches of the tetraploid *Leucanthemum* taxa were reconstructed using ecological niche modeling (ENM) and compared using permutation-based statistical tests. For this purpose, we retrieved collection locations of 1,470 individuals from the database of the Botanical Garden and Botanic Museum Berlin (see Electronic Supplementary ES7), and obtained rasters of 19 bioclimatic and 10 edaphic variables at depth levels of 0-5 cm, 5-15 cm, and 15-30 cm (see Electronic Supplementary ES8) from Worldclim Bioclim (Fick & Hijmans, 2017) and SoilGrids (Hengl et al., 2017), respectively. The rasters were cropped to central Europe (longitude: -10.0° - 26.0°; latitude: 35.9° - 52°) and scaled to a resolution of 2.5 minutes using the R package raster v3.5.15 (https://github.com/rspatial/raster). In addition to the recent climate rasters, we retrieved paleoclimate rasters for the last glacial maximum (LGM; CCSM4) and the last interglacial (LIG; lig_30s) from Worldclim Bioclim and subjected them to the same preprocessing, but with additional recoding of the temperature, since LGM and LIG datasets used Bioclim 1 temperature encoding. For the edaphic variables, we averaged the raster values of the depth levels to obtain a single raster for each soil variable. Since it was computationally not tractable to apply ENM using all 29 rasters, we applied principal component analysis (PCA) with standardization as implemented in ENMtools v1.0.6 (Warren et al., 2021) for feature extraction to the recent Bioclim and SoilGrids rasters, separately. The first three Principal Component (PC) rasters for each of the datasets were selected for ENM.

To compare ecological niches, we reconstructed potential distribution areas for all tetraploid taxa except *L. meridionale* using MAXENT v3.4.4 (Phillips et al., 2021) with 5000 iterations and 6-fold cross-validation, and applied niche equivalency tests implemented in ENMTools (Warren et al., 2021) for all combinations of tetraploid taxa with MAXENT as ENM, 200 replicates, a species range of 50 km, and 1000 background points with a range of 20 km. *Leucanthemum meridionale* was excluded because this species is endemic to a very limited area, causing all collection points to fall in the same raster cell and thus rendering ENM unsuitable. Potential niches at LGM and LIG were also reconstructed using MAXENT with same parameters as for the PC rasters, but this time with LGM and LIG as

projection rasters and including collection data of the diploid species that was already used in chapter 4.

**Geographical Distribution**

The similarity in geographic distribution of the tetraploid taxa, excluding *L. meridionale* (only one sampling location, see Ecological Niche Modeling) was assessed by evaluating the overlap of spatial distributions using the same data as for the ENM analyses. Unfortunately, there are no comprehensive distribution rasters available for the tetraploid *Leucanthemum* representatives, which is why we had to resort to approximating the geographic distribution from the available sampling points. We applied the method proposed in chapter 4, which approximates the true geographical distribution by reconstructing the potential area using an ENM, in this case MAXENT v3.4.4 (Phillips et al., 2021), and subsequently removing unconnected regions, i.e. regions that are separated from collection points by areas of low probability.

To test for sympatry, we applied the permutation approach presented in chapter 4 with 400 simulated datasets and a threshold of 0.25; the resulting *p*-values were corrected for multiple testing using Bonferroni's method.

**Morphological Analyses**

Differences in leaf morphology, an important delimitation criterion for the currently accepted *Leucanthemum* taxa, were assessed by measuring the general leaf shape and the degree of dissection of the leaves by applying elliptic Fourier analysis (EFA; Kuhl & Giardina, 1982) and calculating leaf dissection indices (LDI), respectively. For this purpose, 285 images of digitized herbarium specimens were provided by the herbarium of the Berlin Botanical Museum (B; see Electronic Supplementary ES9). Using those images, we manually annotated 1,338 intact leaves with polygons around the leaves' outlines and polylines along the leaves' main veins using the Computer Vision Annotation Tool (CVAT; https://github.com/openvinotoolkit/cvat). According to procedure described in chapter 4, we straightened the leaves to reduce the influence of deformations introduced by the drying process or unequal growth, and extracted the leaf contours as binary masks.

Using the binary masks, we conducted EFA and calculated the LDI using the Python packages scikit-image (van der Walt et al., 2014), numpy (Harris et al., 2020), scikit-learn (Pedregosa et al., 2011), and pyefd (https://github.com/hbldh/pyefd). For the

EFA, we used 20 harmonics, normalized the descriptors, and applied principal component analysis (PCA) for decorrelation and feature extraction to the non-constant descriptors (normalization causes the three descriptors A1, B1 and C1 to be constant).

To find differences in the general leaf shape, we used the first 15 PCs, explaining at least 50% of the total variance with two different testing strategies: 1) the permutation-based test for differences of the Euclidean distance in PC space as described in chapter 4 and 2) Non-parametric multivariate analysis of variance (NPMANOVA; Anderson, 2001) implemented in the R package vegan 2.6 (function "adonis"). The number of permutations was set to 5,000 and 100, for the first and latter one, respectively. To assess differences in leaf dissection, we subjected the LDI values to Welch's tests. All tests were corrected using Bonferroni's method.

**Detection of Potential Diploid Parents**

To find the potential diploid parental species, we applied a custom version of the SNiPLOID method (Peralta et al., 2013). Very similar to the original SNiPLOID algorithm, we are comparing two parent individuals (i.e., diploids) with one child individual (i.e., a tetraploid) on a per SNP basis, distinguishing and counting the frequency of five different SNP categories: categories 1 and 2 are so-called inter-specific SNPs, where exactly one parent SNP is identical to the child SNP. Categories 3 and 4 are derived SNPs, meaning that a SNP is unique to the child individual, while the parents are homozygous with the same variant; patterns three and four cannot be distinguished based on unphased SNP data. Category five are homeo-SNPs, where the child is heterozygous, comprising both parents' homologous alleles. For all categories, only SNPs where both parents are homozygous were considered.

We applied SNiPLOID to look for signs of allopolyploid speciation, as proposed by Wagner et al. (2020): recent allotetraploids are expected to express a high number of category 5 (i.e., homeo-) SNPs, while established species should express a higher frequency of derived (i.e., category 3 and 4) SNPs. Finally, category 1 and 2 SNPs can indicate geneflow to the putative parents, or can alternatively be a sign for recent polyploidization. Here, for each tetraploid taxon and each possible pair of diploid parent species, we applied a custom Python implementation of the SNiPLOID algorithm to the concatenated SNP output returned by iPyrad (.snps ouput). For each triplet, we tested all individuals, and calculated the mean SNP patterns.

## Results

### RADseq Assembly

Of 216,551,342 raw reads, 81,517,011 demultiplexed, adapter clipped and restriction enzyme filtered reads mapped against the reference that was constructed in chapter 4. The resulting assembly comprised 7,342 loci with a mean length of 178.05 bp (sd: 67.40 bp), and 156,057 SNPs, of which 83,375 were parsimony informative. The amount of missing data for sequence and SNP matrices was 36.93% and 34.52%, respectively.

### Network Analysis

The network reconstruction based on SNP-level Nei distances of the full dataset (all diploid and tetraploid individuals) placed all tetraploid taxa into the so-called *L. vulgare*-group (Figure 5.1, A, right); the diploid samples are resolved according to the species membership. The tetraploid individuals are mostly clustering according to the taxon membership, apart from *L. ircutianum* subsp. *ircutianum* (Figure 5.1, A), which is divided into two clusters and a single individual, namely 437-01. When inspecting the tetraploid individuals in isolation, sample 437-01 is grouped with one of the *L. ircutianum* subsp. *ircutianum* clusters (Figure 5.1, B). The network indicates that there is only a small genetic distance between *L. ircutianum* subsp. *ircutianum* and *L. meridionale*. *L. pseudosylvaticum*, *L. crassifolium*, and the group consisting of *L. delarbrei* subsp. *delarbrei* and *L. meridionale* have, compared to the remaining clusters, relatively long basal branches (Figure 5.1, B).

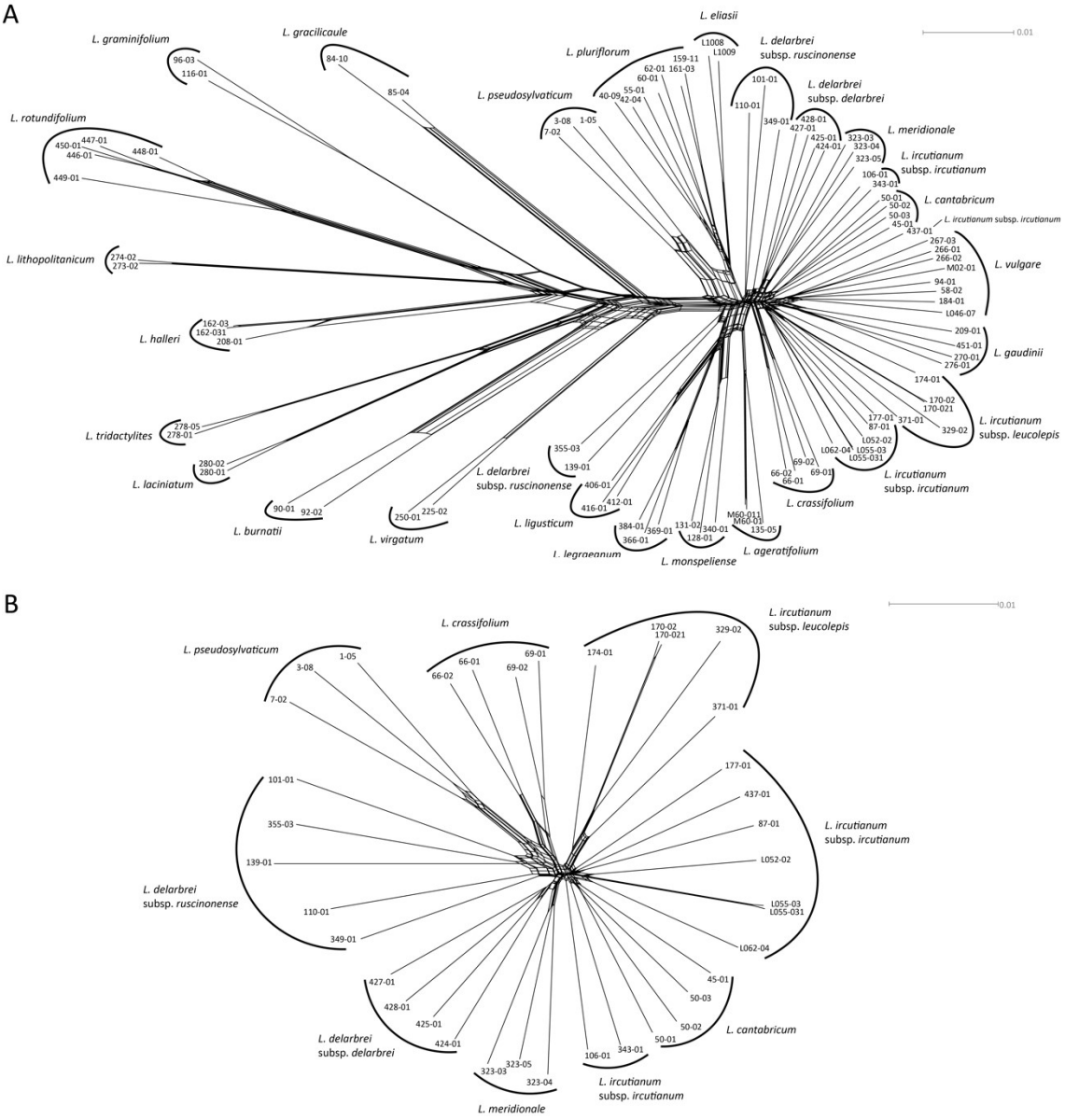


**Figure 5.1**. Neighbor net reconstructions based on SNP-level Nei distances of the full dataset (A), comprising individuals from all 17 diploid species and 8 tetraploid taxa, and of the tetraploids only (B). For the full dataset, all tetraploid taxa are found within the so-called *L. vulgare*-group (right). The *L. vulgare* group is separated from the remaining diploid species by set of relatively long branches (center). All diploid samples are clustered according to the species membership (A). The same is the case for the tetraploid taxa, except for *L. ircutianum* subsp. *ircutianum*, which is separated into two groups (B, lower right).

## Consensus Clustering

The optimal number of clusters according to the Bayesian Information Criterion (BIC; lower is better) was six (Figure 5.2, B, BIC), while the Calinksi-Harabasz (CH; higher is better) score was optimal for $k$=2, but also has a local optimum at six clusters (Figure 5.2, B, CH), suggesting six genetic clusters. The consensus clustering with the co-

association matrix at *k*=6 merged the taxa *L. meridionale* and *L. delarbrei* subsp. *delarbrei* (Figure 5.2, A, pink), and *L. ircutianum* subsp. *ircutianum* and *L. cantabricum* each to a single cluster (Figure 5.2, A, yellow). The remaining samples formed pure clusters, containing only individuals of the respective taxa.

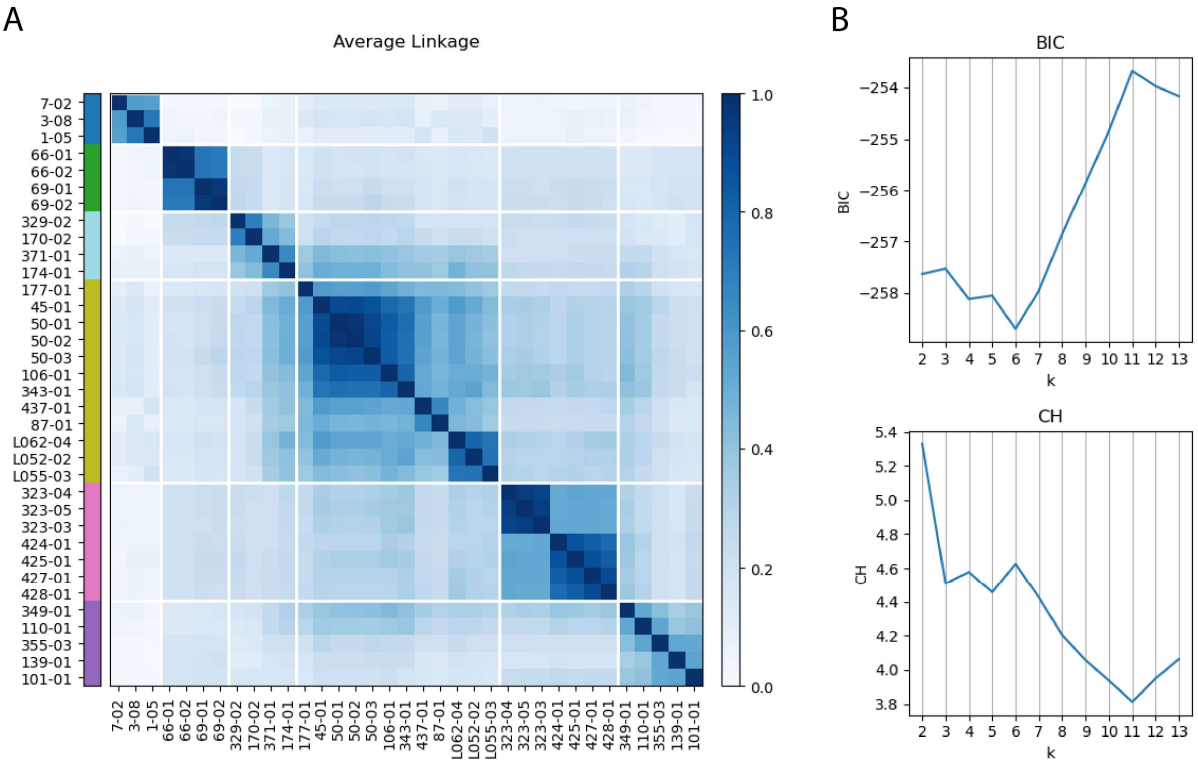


**Figure 5.2**. Consensus clustering results. A) Weighted co-association matrix of the tetraploid *Leucanthemum* taxa with consensus clusters for *k*=6 with color bars depicting consensus cluster membership. B) BIC and CH metrics for number of clusters ranging from 2 to 13. BIC and CH are optimal at *k*=6 and *k*=2, respectively; additionally, the CH has a local optimum at *k*=6 (B).

## Ecological Niche Modeling

All pairwise niche equivalency tests except for the two taxon combinations *L. cantabricum – L. crassifolium* and *L. cantabricum – L. delarbrei* subsp. *ruscinonense* were significant at an alpha level of 0.01 (Table 5.2). Note that the *p* values are exactly 0 if there is no overlap at all. Maps depicting predicted potential areas of the diploids and tetraploids based on recent, LGM, and LIG bioclimatic variables are provided in the supporting information (Electronic Supplementary E10, ES11).

**Table 5.2**. Bonferroni-corrected *p*-values for pairwise niche equivalency tests based on the two test statistics D (upper triangle) and I (lower triangle).

| | *L. cantabricum* | *L. crassifolium* | *L. delarbrei* subsp. *ruscinonense* | *L. delarbrei* subsp. *delarbrei* | *L. ircutianum* subsp. *ircutianum* | *L. ircutianum* subsp. *leucolepis* | *L. pseudosylvaticum* |
|---|---|---|---|---|---|---|---|
| *L. cantabricum* | | **0.52** | **0.31** | 0 | 0 | 0 | 0 |
| *L. crassifolium* | **0.31** | | 0 | 0 | 0 | 0 | 0 |
| *L. delarbrei* subsp. *ruscinonense* | **0.52** | 0 | | 0 | 0 | 0 | 0 |
| *L. delarbrei* subsp. *delarbrei* | 0 | 0 | 0 | | 0 | 0 | 0 |
| *L. ircutianum* subsp. *ircutianum* | 0 | 0 | 0 | 0 | | 0 | 0 |
| *L. ircutianum* subsp. *leucolepis* | 0 | 0 | 0 | 0 | 0 | | 0 |
| *L. pseudosylvaticum* | 0 | 0 | 0 | 0 | 0 | 0 | |

## Geographical Distribution

All taxon combinations except for *L. cantabricum* – *L. crassifolium* and *L. delarbrei* subsp. *ruscinonense* – *L. ircutianum* subsp. *ircutianum* were significant, suggesting a deviation from sympatric distribution for most of the tetraploid taxa (Table 5.3).

**Table 5.3**. Bonferroni-corrected *p*-values for pairwise tests of geographical overlap, i.e., sympatry. Corrected *p*-values are truncated to 1.

| | *L. crassifolium* | *L. delarbrei* subsp. *ruscinonense* | *L. delarbrei* subsp. *delarbrei* | *L. ircutianum* subsp. *ircutianum* | *L. ircutianum* subsp. *leucolepis* | *L. pseudosylvaticum* |
|---|---|---|---|---|---|---|
| *L. cantabricum* | **1.00** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *L. crassifolium* | | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *L. delarbrei* subsp. *ruscinonense* | | | 0.00 | 0.05 | 0.00 | 0.00 |
| *L. delarbrei* subsp. *delarbrei* | | | | 0.00 | 0.00 | 0.00 |
| *L. ircutianum* subsp. *ircutianum* | | | | | 0.00 | 0.00 |
| *L. ircutianum* subsp. *leucolepis* | | | | | | 0.00 |

## Morphology

The degree of leaf dissection varied significantly at an alpha level of 0.01 for all but four taxon combinations (Table 5.4, upper triangle). Differences of the general leaf shape, measured as Euclidean distance among PC transformed Fourier descriptors, were significant for 8 of the 28 taxon pairs (Table 5.4, lower triangle, left). When applying NPMANOVA to the PC scores, six additional taxon combinations showed significant differences in the leaf shape (Table 5.4, lower triangle, right).

**Table 5.4**. Bonferroni-corrected *p*-values for pairwise tests of morphological similarity. The upper triangle represents *p*-values of the LDI-based Welch's tests, determining divergence in the

dissection of the leaves. The lower triangle comprises *p*-values of the permutation tests of Euclidean distances in PC space (left) and NPMANOVA results (right), quantifying differences in the general leaf shape. Corrected *p*-values are truncated to 1.

| | *L. cantabricum* | *L. crassifolium* | *L. delarbrei* subsp. *delarbrei* | *L. delarbrei* subsp. *ruscinonense* | *L. ircutianum* subsp. *ircutianum* | *L. ircutianum* subsp. *leucolepis* | *L. meridionale* | *L. pseudosylvaticum* |
|---|---|---|---|---|---|---|---|---|
| *L. cantabricum* | | 0.00 | **0.21** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *L. crassifolium* | **1.00**, **1.00** | | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *L. delarbrei* subsp. *delarbrei* | **1.00**, **1.00** | **0.44**, **1.00** | | 0.00 | 0.00 | 0.00 | **1.00** | 0.00 |
| *L. delarbrei* subsp. *ruscinonense* | **0.04**, 0.00 | **0.02**, 0.00 | **1.00**, **1.00** | | 0.00 | 0.00 | 0.00 | 0.00 |
| *L. ircutianum* subsp. *ircutianum* | **0.01**, 0.00 | **0.17**, 0.00 | **0.02**, **0.83** | 0.00, 0.00 | | **1.00** | 0.00 | 0.00 |
| *L. ircutianum* subsp. *leucolepis* | 0.00, 0.00 | 0.00, 0.00 | **0.01**, 0.00 | 0.00, 0.00 | **1.00**, **1.00** | | 0.00 | **0.01** |
| *L. meridionale* | **1.00**, **1.00** | **1.00**, **1.00** | **1.00**, **0.83** | **1.00**, **1.00** | **1.00**, **1.00** | **0.38**, 0.00 | | 0.00 |
| *L. pseudosylvaticum* | 0.00, 0.00 | 0.00, 0.00 | **0.06**, 0.00 | **0.01**, 0.00 | 0.00, 0.00 | 0.00, 0.00 | **1.00**, **1.00** | |

## Detection of Potential Diploid Parents

We found that SNP categories 1 and 2 were increasing with the genetic distance of the corresponding parental species from the child taxon. The sum of categories 3 and 4 was generally decreasing with the genetic distance. For the analyzed taxa, we could not find any irregularity in this pattern. The frequency of category 5 is relatively constant, but also slightly decreases with the genetic distance of the parent species. There seemed to be slight shift in categories frequencies, when at least one parent outside of the "*L. vulgare*-group" (*L. vulgare*, *L. gaudinii*, *L. pluriflorum*, *L. ageratifolium*, *L. monspeliense*, *L. legraenum*, *L. ligusticum*; see chapter 4) was included as parental species. Plots of the category proportion for all triplets are provided in the supporting information (Electronic Supplementary ES12).

## Discussion

The tetraploid *Leucanthemum* representatives comprise a taxonomically challenging group with relatively recent divergence times (Wagner et al., 2019) and a high potential for hybridization (Oberprieler et al., 2014). While there has been research on the evolutionary history of a subset of the tetraploid taxa of the genus and on the species delimitation based on morphological characters, especially the leaf shape (Vogt, 1991, 2019), the taxonomy on the species level is still rather uncertain. Our goal was, building on the current taxonomy, to delimit the observed entities as objective as possible by integrating genealogical, ecological, geographical, and morphological evidence. This integrative approach has already been applied to the closely-knit group of diploid *Leucanthemum* representatives (see chapter 4) and has proven its worth for untangling the taxonomy in this young and hybridizing plant group. Here, we employed the strategy of first grouping (i.e.,

finding discontinuities in the four sources of data), and then assigning the taxonomic rank of the corresponding group, as proposed by Stuessy et al. (2014) or Reydon & Kunz (2019). We also tried to find patterns in the genealogy and bioclimatological niche preferences that might give a first hint at the potential history and origin of the tetraploids. Owing to the integrative methodology, the grouping step is based on genealogical, ecological, and morphological evidence. The application, results, and interpretation of each of those data modes will be discussed in the following.

**Genealogical Evidence**

RADseq, by sampling a mostly random set of short and anonymous loci distributed over the whole genome, allows to approximate the genomic variation even in the absence of a reference genome. In chapter 4, we have shown that a pseudo reference-based assembly approach can be used to combine multiple RADseq dataset without the need for a reference genome. Here, we used the consensus loci prepared in chapter 4 as a reference for both, the diploid RADseq data from that chapter, and new data generated in the context of the present study, enabling us to build an assembly comprising all diploid *Leucanthemum* species and eight tetraploid *Leucanthemum* taxa, which are of central interest to this contribution.

Based on the combined assembly, we found that all tetraploid taxa are clustering within the so-called *L. vulgare*-group (see Figure 5.1, A, right), comprising the eight diploid *Leucanthemum* lineages that were the focal group of chapter 4. This group was shown to have only relatively recently diverged from the remaining *Leucanthemum* species (ca. 1 Ma; Wagner et al., 2019), suggesting a similar situation for the tetraploids. We found that clustering patterns in the neighbor nets based on genetic distances mostly conform with the current, morphology-based taxon membership; the only exception being the broadly distributed *L. ircutianum* subsp. *ircutianum*, which is separated into two groups with *L. cantabricum* inserted in between (Figure 5.1, A). The network of the tetraploid individuals suggests varying genetic differentiation of the focal taxa (Figure 5.1, B): *L. pseudosylvaticum*, *L. crassifolium*, *L. ircutianum* subsp. *leucolepis*, as well as the group comprising *L. delarbrei* subsp. *delarbrei* and *L. meridionale* have relatively long basal branch segments compared to the remaining taxa. Contrary to our expectation, the genetic distance of *L. ircutianum* and *L. delarbrei* subspecies to their respective infraspecific relatives is similar to the distance the remaining taxa, suggesting that the current taxonomic rank is not supported by genealogic evidence.

Since *Leucanthemum* is a polyploidizing and hybridizing genus, and due to the large amount of genetic information (around 150,000 SNPs), none of the currently available model-based species delimitation tools matched our runtime and methodological requirements sufficiently. Following chapter 4, we decided to apply a consensus clustering approach to detect groups of genetically similar individuals, i.e., reproductively isolated entities. Consensus k-Means clustering (Monti et al., 2003), as proposed by Wagner et al. (2020) and applied in chapter 4, however, should not be applied when the density, in this case the genetic variability within the groups varies drastically. This problem can be reduced by applying a related method, the semi-supervised Weighted Ensemble of Random (k)k-Means (WECR; Lai et al., 2019), which allows to integrate a-priori knowledge about the groups. Here, we encouraged WECR to cluster individuals of the same taxon with each other. This prior causes WECR to effectively cluster taxa rather than individuals. The clustering of the tetraploid *Leucanthemum* representatives was optimal at 6 groups (Figure 5.2, B), merging *L. ircutianum* subsp. *ircutianum* with *L. cantabricum*, and *L. meridionale* with *L. delarbrei* subsp. *delarbrei* (Figure 5.2).

**Ecological Evidence**

Based on ecoclimatologic and edaphic variables, the abiotic niches of seven tetraploid taxa were reconstructed by applying traditional ecological niche modeling (ENM) as proposed by Raxworthy et al. (2007). The eighths taxon, *L. meridionale*, is endemic to a single locality, the Puy de Wolf (France, Rodez, Decaseville, Firmi), causing all available samples of this taxon to fall in the same raster cell and thus rendering ENM infeasible. The remaining taxa with broader spatial distributions, whose niches could be reconstructed, have been compared using niche equivalency tests (Warren et al., 2021). We found the abiotic niches to be significantly different (Table 5.2) for all but one taxon combination, namely the pair *L. crassifolium* - *L. cantabricum*. This result might seem counter-intuitive at first, since *L. crassifolium* is described as inhabiting mainly coastal areas, while *L. cantabricum* is found at the foothills of the Pyrenees from see level up to an elevation of 800 m. It is possible that the relatively rough resolution of the ecoclimatological and edaphic rasters could not capture the differences in ecological niches on a small spatial scale. An alternative explanation may be that specific edaphic and climatological variables indicating proximity to coastal areas (e.g., soil salinity) are

missing. Interestingly, this pair was found to form genealogically distinct groups in the neighbor net and clustering analyses.

**Geographical Evidence.**

Joint spatial distribution (i.e., sympatry) is an important criterion for species delimitation in plants (e.g., von Wettstein, 1898). Determining sympatry, however, requires the availability of comprehensive distribution maps for the tested entities. Unfortunately, there are no such maps available for the tetraploid *Leucanthemum* representatives, and thus we had to approximate the true distribution of the focal taxa. Here, we applied the method proposed in chapter 4 for generating distribution rasters based on incomplete distribution maps by including abiotic niche information via ENM. Other methods such as convex hulls, alpha hulls, and kernel density estimated were considered, but ultimately discarded, due to bad performance with the available data. We found that only the two pairs *L. ircutianum* subsp. *ircutianum* – *L. delarbrei* subsp. *ruscinonense* and *L. cantabricum* – *L. crassifolium* did not significantly deviate from a sympatric distribution (Table 5.3).

**Morphological Evidence**

The current taxonomy of the *Leucanthemum* representatives is largely based on morphological features; leaf shape and dissection in particular are very important criteria for species delimitation (Vogt, 1991, 2019). In the present study, we analyzed leaf morphology on two levels: 1) the general leaf shape using elliptic Fourier analysis (EFA; Kuhl & Giardina, 1982) and 2) the division of the leaves using the leaf dissection index (LDI; Kincaid & Schneider, 1983). As expected, we found that almost all taxon pairs show significant differences in the dissection of the leaves, with the exception of the two *L. ircutianum* subspecies, and the pair *L. delarbrei* subsp. *delarbrei* – *L. meridionale* (Table 5.4, upper triangle). Variation in the general leaf shape is, according to the permutation tests based on Fourier descriptors, much less pronounced (Table 5.4, lower triangle).

**Potential Diploid Parents**

In addition to the grouping and ranking of the taxa, we wanted to get a first view into the evolutionary history of the taxa under study. In particular, we wanted to find out, whether they are of autopolyploid or allopolyploid origin. As proposed by Wagner et al. (2020), we employed SNiPLOID (Peralta et al., 2013) to detect genealogic patterns that

might give a hint on the diploid parent or parents of the tetraploid *Leucanthemum* representatives. Using this method, however, we could not find any signal indicating either allopolyploid or autopolyploid diversification (Electronic Supplementary ES12). Instead, we found a general tendency of the frequency of SNiPLOID patterns 1 and 2 to increase with the respective genetic distance of the diploid species to the tetraploid taxon. Pattern 3/4 is generally high when the parent species are genetically similar to the tetraploid and low when the distance between parents and tetraploid increases. The frequency of pattern 5, which according to Wagner et al. (2020) indicates allopolyploid speciation, was rather stable, and did only slightly decrease with increasing distance between diploids and the corresponding tetraploid taxon. A possible explanation might be that the young age of the genus renders the detection of ancestral patterns difficult. Nevertheless, this finding is surprising, since previous works studying *Leucanthemum* suggested that at least some of the tetraploid taxa might be of allopolyploid origin (e.g., Oberprieler et al., 2014, 2022). Future research on the phylogeny and evolutionary history of *Leucanthemum*, potentially based on phased long-read markers as, for example, produced by nanopore sequencing might help to unravel the origin of the tetraploid taxa (e.g., Dorfner et al., 2022).

**Integration of Evidence**

Recent taxonomic studies are more and more applying some form of integrative species delimitation by drawing evidence from diverse sources of data, such as genealogy, morphology, physiology, ecology, and geography. While most scientist agree that this combination of data modes is a better approach than basing conclusions on a singular source of evidence, there is still no generally accepted method for a formal integration. Rather, semi-formal procedures have been proposed (e.g., Schlick-Steiner et al., 2010, Padial et al., 2010), which allow researchers to flexibly add new data modes or alternative analytical methods. Here, we will apply the methodological framework described in chapter 4, which was already used for species delimitation of the diploid *Leucanthemum* representatives.

In this framework, influenced by De Queiroz' (2007) unified species concept, and the works of von Wettstein (1898), Wiley (1978) and van Valen (1976), the most important requirement for the species rank is genealogical differentiation: two entities can only be distinct species, if they differ significantly in their genetic composition, i.e., if they are sufficiently reproductively isolated, to assume “separately evolving metapopulations” (De

Queiroz, 2007). Genealogically distinct entities can be distributed sympatrically or allopatrically. In the first case, genetic differentiation in sympatry is a clear indicator for assigning the species rank, while ecology and morphometry can help in disentangling the infraspecific ranking. In the latter case (i.e., genetic differentiation in allopatry), additional ecologic differentiation is required to assume two distinct species.

**Taxonomic Implications**

Following the previously introduced framework, we propose the following species delimitation:

(1) *Leucanthemum ircutianum* and *Leucanthemum cantabricum*: While *L. ircutianum* subsp. *ircutianum* is broadly distributed in Europe and into east Asia, *L. ircutianum* subsp. *leucolepis* is spatially limited to the Apennine and Balkan Peninsula. Both taxa were only recently considered as subspecies of *L. ircutianum* (Oberprieler et al., 2022). We found that the subspecies form two distinct genetic cluster in allopatric distribution occupying different abiotic niches, with a very similar leaf shape and degree of dissection (Figure 5.3). In the chosen species delimitation framework, this is a clear indication for elevating the taxa to the species rank. Regarding *L. ircutianum* subsp. *ircutianum* and *L. cantabricum*, we could not find sufficient genetic differentiation that would justify the separation into two species. At the same time the two taxa are allopatrically distributed and occupy different abiotic niches, with a mostly distinct leaf morphology (Figure 5.3).

Due to the presence of genetical, ecological, and morphological differentiation from *L. ircutianum* subsp. *ircutianum* we propose to raise the subspecies *L. ircutianum* subsp. *leucolepis* to the species rank. Further, we suggest that *L. cantabricum* should be considered a subspecies of *L. ircutianum* instead of a distinct species, even so they are found in different habitats and are morphologically distinct, since those taxa form a genealogical unit.

(2) *Leucanthemum delarbrei* and *Leucanthemum meridionale*: The two *L. delarbrei* subspecies, *L. delarbrei* subsp. *delarbrei* and *L. delarbrei* subsp. *ruscinonense*, were found to form distinct genetic clusters occupying significantly different abiotic niches and having significantly different leaf morphology, while at the same time being allopatrically distributed (Figure 5.3). Indeed, *L. delarbrei* subsp. *delarbrei* is genealogically more similar to *L. meridionale* than to its intraspecific sister taxon. *Leucanthemum meridionale*

is endemic to a single location on serpentine soil in the western Massif Central (Puy de Wolf, France), which did not allow us to calculate geographic overlap and compare abiotic niches with this taxon. We will thus assume *L. meridionale* to be allopatrically distributed with *L. delarbrei* subsp. *delarbrei*, since the location of endemicity of *L. meridionale* lies outside the distribution range of *L. delarbrei* subsp. *delarbrei*. For the abiotic niche, we assumed that both taxa do not share the same niche due to the limited distribution of *L. meridionale* on serpentine soil.

The pronounced differences in genealogy, ecology, spatial distribution, and morphology lead us to propose the ranking of *L. delarbrei* subsp. *delarbrei* and *L. delarbrei* subsp. *ruscinonense* as two distinct species. *L. meridionale* should be considered a subspecies of *L. delarbrei* due to the absence of sufficient genetic differentiation of *L. delarbrei* subsp. *delarbrei*.

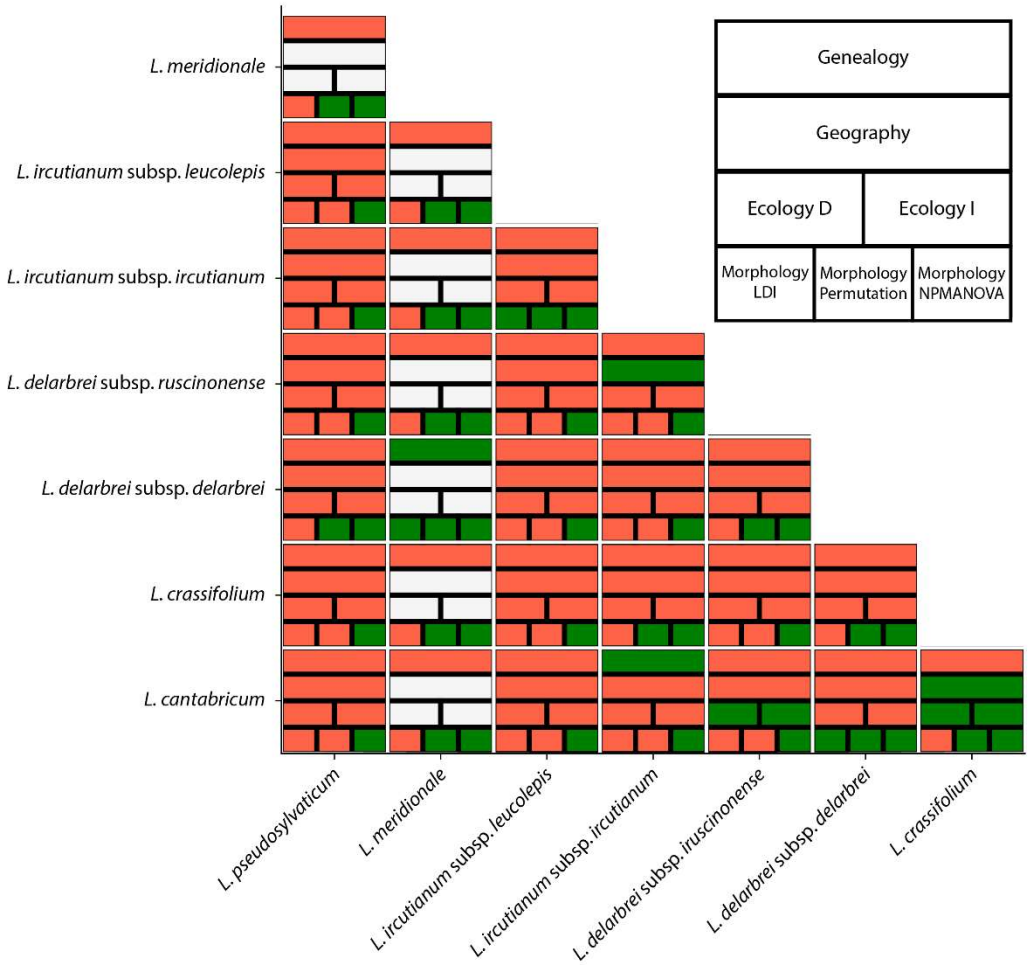


**Figure 5.3**. Taxon pair-wise differences in genealogy (first box row), geography (second box row), ecology (third box row; D and I), and morphology (fourth box row; LDI Welch's test, EFD permutation test, and EFD NPMANOVA). Each box represents a taxon pair and is subdivided into

fields for each source of evidence. A red field signifies divergence, a green field no sufficient divergence, and a white field missing data, regarding the corresponding data source.

## Funding Information

This work was funded by the grant of the German Research Foundation (DFG) in the frame of the SPP 1991 “Taxon-omics − New Approaches for Discovering and Naming Biodiversity” to C.O. (OB 155/13-1) and R.V. (VO 1595/3-1).

## Conflict of Interest

The authors declare no conflict of interest.

# Chapter 6: Summary and Discussion

## Summary

The present thesis proposes a set of tools, methods, and operational guidelines, a bioinformatical pipeline, for integrative species delimitation and explores its application to the young, hybridizing, and polyploid complex-forming genus *Leucanthemum*. Here, chapters 2 and 3 describe the implementation of a software for the collection of morphological data from preserved plant specimens, while chapters 4 and 5 focus on practically applying methods, some of them novel in their use in this context, for integrative species delimitation of the diploid and tetraploid *Leucanthemum* taxa.

The study presented in chapter 2 proposes a deep learning-based solution, namely GINJINN, for the detection of structures in digital images of preserved plant specimens. Traditionally, feature extraction from herbarium specimens in the context of plant sciences was conducted either manually using the physical voucher, or alternatively by applying classical, algorithmic object detection methods using hand-crafted features (e.g., Unger et al., 2016). Draw backs of those methods are a large amount of manual work, and the requirement of expert knowledge in computer vision, respectively. Modern deep learning models, like those leveraged by GINJINN, can circumvent both of those problems by learning a mapping from input images to the corresponding object localizations, provided that sufficient training data, i.e., images with object annotations, is available. With our tools, we were able to successfully automate the extraction of intact leaves from digitized herbarium specimens in 95% of 61 tested images, using a Faster R-CNN (Nieuwenhuizen et al., 2018) trained on 183 specimen images, and pretrained on the COCO (Lin et al., 2014) dataset. While its efficacy could be shown, GINJINN was limited to bounding box-detection functionality, i.e., it can only extract the box enclosing objects of interest.

The next study (chapter 3) describes the evolution of GINJINN to GINJINN2. GINJINN2 extends the functionality of GINJINN by not only providing bounding box-detection, but also segmentation capabilities. Segmentation is the pixel perfect detection of objects of interest. For leaves, for example, the tool's output is the exact leaf silhouette, provided the model is sufficiently trained. In addition, GINJINN2 provides several utilities for data pre- and postprocessing, visualization, data insight, and segmentation refinement using CascadePSP (Cheng et al., 2020). We successfully demonstrated the software's application on a set of diverse problems, ranging from herbivore detection, over seed mixture evaluation, and stomata extraction from microscopic images, to leaf silhouette

extraction from digital images of herbarium specimens. This range of applications shows that a deep leaning-based solution for data extraction from digital images can be a very useful tool in the context of ecology and evolution, in particular the extraction of plant and animal structures from preserved specimens may provide a new, easily accessible source of data for morphological or integrative species delimitation.

Chapter 4, the third study, describes the integrative species delimitation of all diploid *Leucanthemum* representatives using genealogical, morphological, ecological, and geographical evidence. Using RADseq (Poland et al., 2012) data, species were genealogically delimited using the multispecies coalescent-based SNAPP (Bryant et al., 2012) software, and the unsupervised consensus k-means clustering (Monti et al., 2003; Wagner et al., 2020). For the ecological layer, abiotic niches were reconstructed and tested for equivalency (Cardillo & Warren, 2016) among the morphotaxa. Geographical overlap was calculated based on approximated distribution maps. In this context, a novel method for reconstructing distribution areas using ecological niche modeling and a thresholded depth-first search was proposed. Differences among the taxa in the general leaf shape and the degree of dissection of the leaves, by means of elliptic Fourier analysis (Kuhl & Giardina, 1982) and leaf dissection index (Kincaid & Schneider, 1983), respectively, were determined based on leaf silhouettes extracted from digital images of preserved plant specimens. Since shape measurements can be influenced by a bend of the leaves, which can be caused due to the drying process (Unger et al., 2016) and due to unequal growth, a novel method for correcting this deformation based on the as-rigid-as-possible (Igarashi et al., 2005) algorithm was proposed and successfully applied. The integration of the sources of evidence inspired by De Queiroz' (2007) and von Wettstein's (1898) species concepts, lead to the recombination of *L. pyrenaicum* as subspecies of *L. vulgare*, and of *L. gallaecicum* and *L. cacuminis* as subspecies of *L. pluriflorum*.

The last study of the present thesis (chapter 5) describes the extension and adaption of the methodological pipeline built in chapter 4 for species delimitation of *Leucanthemum* at the tetraploid level. Again, ecological, geographical, and morphological evidence is integrated according to the methods proposed in chapter 4. For the genealogical layer, a in this context novel approach was taken by applying the semi-supervised WECR (k)k-means (Lai et al., 2019) method for detecting genetic clusters. In addition, the SNiPloid (Peralta et al., 2013) software was used as described by (Wagner et al., 2020) to detect SNP patterns that might indicate an allopolyploid origin and the respective diploid parent species. Surprisingly, and in contrast to previous studies (e.g., Oberprieler et al., 2014, 2022), no

evidence for or against allopolyploid speciation could be found based on RADseq data. Finally, the integration of evidence, as described in chapter 4, suggested that *L. ircutianum* subsp. *leucolepis* and *L. delarbrei* subsp. *ruscinonense* should be raised to the species rank, while *L. cantabricum* and *L. meridionale* are to be considered subspecies of *L. ircutianum* and *L. delarbrei*, respectively.

## An Integrative Species Delimitation Pipeline for *Leucanthemum*

In their review of integrative species description, Pante et al. (2015) argue that "most, if not all, taxonomists agree that objectively evaluating several lines of evidence within a formalized framework is the most efficient and theoretically grounded approach to defining robust species hypotheses". While several such frameworks were proposed (e.g., De Queiroz, 2007, Schlick-Steiner et al., 2010) a general applicable solution to the species delimitation problem has not yet been found, since "the number of arguments, data and criteria […] needed for defining new species and their hierarchy can vary depending on the taxon considered or the approach applied" (Pante et al., 2015). This problem is thought to be especially prevalent in study groups with ongoing or recent speciation, i.e. clades comprising closely related species, within so-called "gray zones" (De Queiroz, 2007), where different delimitation methods may arrive at different taxonomic conclusions. Besides recent speciation, other factors that complicate species delimitation may include the presence of hybridization, incomplete lineage sorting, phylogeographic structure, and demography (see Naciri & Linder, 2015).

De Queiroz' (2007) "unified species concept" tries to combine ideas from many precursing species concepts by requiring species to form "separately evolving metapopulation lineages". In the context of this concept, the separately evolving nature of taxa can be determined based on any kind of evidence reflecting disparate evolution. Such evidence might, for example, include crossing barriers, morphological discontinuities, ecological differentiation, or genealogical variance, where the latter might be assumed to be especially fitting for this purpose, since differentiation among lineages must be genetically detectable. Species delimitation using genealogic evidence alone, however, is known to be plagued by problems of over- and under-splitting (Sukumaran & Knowles, 2017), especially in the context of modern multispecies coalescent-based delimitation and in the presence of geneflow (Sukumaran & Knowles, 2017; e.g., Wagner et al., 2020). Depending on the availability of genealogical data, it is possible that a signal is not visible

in the selected genetic marker system and would only be unraveled by consulting whole-genome analyses, which is not feasible for many non-model organisms. This problem necessitates the incorporation of additional sources of evidence from a variety of data modes in the spirit of integrative taxonomy. The development of a general operational sequence, a pipeline, for species delimitation using an integrative approach to data collection and analysis, taxon description, and ranking is an idea that was pursued since the calls of Dayrat (2005) and Will et al. (2005) for integrative taxonomy (Padial et al., 2010; Schlick-Steiner et al., 2010; but also Yeates et al., 2011). The present thesis proposes such a pipeline, with chapters 2 and 3 developing a deep learning-based solution for morphological data collection, and chapters 4 and 5 implementing an operational sequence for species delimitation of the diploid and tetraploid representatives of *Leucanthemum*, respectively.

The ever-progressing digitalization of herbaria all over the world provides a rich treasure of readily available preserved plant specimens but the manual extraction of structures like, for example, leaves or flowers, for analyses, e.g., species delimitation, is a labor-intensive process. As a consequence, methods based on classical computer vision methods were developed to automatically extract structures or features from images of preserved specimens (e.g., Corney, Tang, et al., 2012, Henries & Tashakkori, 2012, Unger et al., 2016). Those techniques, while ingenious and performant at their time, lacked the capacity for generalization, i.e., the transfer to new types of data or species, and often required specialists for fine tuning. Fortunately, within the last decade, the popularization and evolution of deep learning-based frameworks for image classification and object detection provided a new set of tools alleviating those problems, and unleashing a wave of publications dealing with the automatic analysis of herbarium specimens (e.g., Carranza-Rojas et al., 2017, Goëau et al., 2020, Lorieul et al., 2019, Weaver et al., 2020, Younis et al., 2018, 2020). Chapters 2 and 3 are describing the development of GINJINN and GINJINN2, a tool for object detection and extraction using deep learning. While GINJINN was developed for the extraction of plant structures, it has since evolved to a general-purpose object detection tool (see chapter 3).

In contrast to classical computer vision algorithms, supervised machine learning methods, like GINJINN, require a substantial amount of so-called training data, i.e., images where to objects of interest are annotated, and can be difficult to interpret. Concerning the first issue, the famous COCO dataset (Lin et al., 2014), for example, comprises 83,000 training images for training a detection model with 91 different “common object” classes

(e.g., bicycle, car, person, etc.). In a scientific context, this is, arguably, an infeasible amount of required data. Fortunately, a model that was fit on one (potential large) dataset, e.g., COCO, can be used as a starting point for training a model for detecting completely different classes. This technique, called transfer learning, reduces the amount of data that is required to an adequate model fit drastically. Regarding the problem of model opacity, there is still no general solution available, however, the relatively recent movement of so-called explainable machine learning (e.g., Roscher et al., 2020, Belle & Papantonis, 2021) provides ideas on how to make deep learning models more transparent and interpretable. With the benefits greatly outweighing the drawbacks, deep learning-based image analyses are very likely here to stay, as the need for automatic data extraction from natural collections can be assumed to be increasing as they are growing. It can be assumed that it will become more convenient to apply machine learning methods due to the evolving ergonomics of available frameworks, development of new methods, and due to the accumulation of annotated datasets that can be used for transfer learning. Even today, in absence of those expected advancements, many recent publications (see Hussein et al., 2022 for a review), such as chapters 2 and 3 of this thesis, show that deep learning is well-suited for automatic data collection in the context of plant sciences.

Extracting morphologic information from herbarium specimens is hampered by degradation of the plants, and deformation or damaging of structures due to drying and storage. As a result, it may be necessary to restore the original state of the structures of interest, i.e., in the context of this thesis, leaves. Recent studies have used machine learning-based methods for removing herbivore damage from leaf silhouettes (Hussein et al., 2021; Silva Vieira et al., 2021), and some years prior, Unger et al. (2016) proposed a method for straightening leaves that were bent as a result of the drying process or because of unequal growth. For the *Leucanthemum* specimens used in the present thesis, only the latter is relevant. In chapter 4 (Figure 4.1), we proposed a novel leaf straightening procedure based on the as-rigid-as-possible algorithm (Igarashi et al., 2005), which allows to apply deformation-sensitive methods like the elliptic Fourier analysis (Kuhl & Giardina, 1982) to leaves extracted from herbarium specimens. This method requires the annotation of a median line along the leaves to be straightened. Future improvements of this methods might be the automatic detection of this median line without the need for manual annotation, or alternatively a machine learning-based implementation, potentially similar to the one used by Hussein et al. (2021) for herbivore damage removal.

Besides data collection, an integrative species delimitation pipeline should also address data analysis and taxonomic consequences. Chapters 4 and 5 describe the analysis of genealogical, ecological, geographical, and morphological evidence and their joint evaluation for species delineation. RADseq data formed the foundation of the genealogical layer, comprising multispecies coalescent (MSC)-based species delimitation using SNAPP (Bryant et al., 2012), and unsupervised and semi-supervised consensus clustering, by means of consensus k-means (CKM; Monti et al., 2003) and weighted ensemble of random (k)k-means (WECR; Lai et al., 2019), respectively. Species delimitation in context of the MSC is known to often result in the oversplitting of taxa (Sukumaran & Knowles, 2017), i.e., in the interpretation of population structure as evidence for species. Clustering without an explicit biological model (CKM and WECR) is, unfortunately, often looked down upon as "uncladistic", i.e., missing a phylogenetic foundation, even when basing the clustering on biologically informed distance metrics, i.e., models of sequence evolution. Interestingly, in chapter 4, we found that such an allegedly oversimplistic clustering procedure, namely consensus k-means, arrived at the same grouping of taxa as the highly complex MSC-based method but required a fraction of the computation time (~4 months on a HPC cluster vs. ~10 minutes on a desktop PC). A similar result was found by Wagner et al. (2020) in the sister genus *Rhodanthemum* B. H. Wilcox & al. It might be an interesting topic for future research to determine, whether this correlation can be traced back to an equivalency of the methods under certain circumstances. Finding approximate solutions to taxon grouping instead of explicitly modeling biological processes might be particularly important in the context of growing data set sizes and especially when studying polyploid taxa due to over-linearly increasing computational costs (see Lautenschlager et al., 2020, Yan et al., 2022).

The analysis of ecological, geographical, and morphological patterns has long been important for species delimitation, especially before genetic data was available. Chapter 4 and 5 describe the application of well-established statistical methods for detecting discontinuities among *Leucanthemum* taxa based on the equivalency of reconstructed ecological niches (Cardillo & Warren, 2016; Warren et al., 2021), and by detecting differences in the dissection of leaves (Kincaid & Schneider, 1983) and in their general shape (Kuhl & Giardina, 1982). To evaluate the geographical layer, i.e., the overlap of distribution areas, comprehensive distribution maps are required. Such maps, however, are not available for most non-model organism, including *Leucanthemum*. While there are many methods approximating distribution areas (convex hulls: see Meyer et al., 2017; alpha hulls: Burgman & Fox, 2003, Capinha & Pateiro-López, 2014, see Meyer et al., 2017;

kernel methods: Burgman & Fox, 2003, Cardillo & L. Warren, 2016), none of them were adequate for application in *Leucanthemum*. Chapter 4 proposes an ecological niche modeling-based approach to solve this problem. The resulting distribution maps should be treated like approximated distribution maps that a scientist might manually produce based on known ecological preferences of a species and sampling locations but are objective and reproducible. While the proposed method is working as intended in context of the present thesis, its general features and limitations should be tested on simulated and empirical datasets in comparison to other methods.

In their publications, Stuessy et al. (2014), and Reydon & Kunz (2019) describe species delimitation as a sequence of a) grouping, i.e., the detection of discontinuities, and b) ranking of the resulting clusters (Stuessy et al., 2014; Reydon & Kunz, 2019). In the framework of integrative taxonomy, the ranking step comprises the joint evaluation of a set of analyses based on diverse data modes for drawing taxonomic consequences. This evaluation is traditionally a descriptive process, comparing evidence from different data sources with each other and trying to reach a conclusion that is well-supported. In contrast to this descriptive approach, efforts have been made to formalize species delimitation algorithms (e.g., Solís-Lemus et al., 2015) or guidelines (e.g., Schlick-Steiner et al., 2010) but those did not take hold as single source of truth in the taxonomic community (see Daglio & Dawson, 2019), probably due to their limited applicability to single study groups. As an example, a species delimitation algorithm that can be used to delineate reproductively isolated animal species might not be adequate to delimit apomictic plant taxa, or even unicellular organism. In chapters 4, we thus took a rather descriptive approach to species delimitation in the diploid *Leucanthemum* representatives, factoring the species concepts of von Wettstein (1898), Wiley (1978), and De Queiroz (2007) into the taxonomic decision. This ranking method was finally applied to delimit the tetraploid *Leucanthemum* species. While the resulting taxonomy for *Leucanthemum* based on this approach is justifiable and reasonable compared to the prior taxonomy in this groups, it might be worthwhile to include even more sources of evidence, e.g., biotic niches or flower morphology, and to test mathematical methods for integrating different sources of data (e.g., multi-view clustering: see Fu et al., 2020) to reach a potentially even better supported taxonomical conclusion.

## Outlook

Integrative taxonomy is the current gold-standard and most likely also the future of research in the context of species delimitation (e.g., Schlick-Steiner et al., 2010, 2014, Daglio & Dawson, 2019). The present thesis' studies explored this integrative approach to species delineation in *Leucanthemum*, and proposed an operational sequence based on genealogy, ecology, geography, and morphology by means of RADseq data, abiotic niches, overlap of distribution areas, and differentiation of leaf shape, respectively. Those data sources, of course, only represent a fraction of all potentially available evidence that may be used for delimiting taxa. Zachos (2016), in his book on *Species Concepts in Biology*, compiled 32 different species concepts, listing alternative species conditions like complete reproductive isolation ("biological species concept"), monophyly ("phylogenetic species concept"), or divergent evolution ("evolutionary species concept"), to only name a few. In this context, it should be clear that integrative species delimitation must try to incorporate as much information as possible to arrive at taxonomic conclusions that are as objective as possible. Collecting and analyzing data for unraveling this wide range of processes and patterns will be a task for further research in this area, to make taxonomy even more integrative, and might, for example comprise research based whole genomic data, phylogenetic and evolutionary processes, and biotic niches and interactions.

For polyploid species, the origination from other species of lower or even higher ploidy level may be considered an important criterion for species delimitation, but also complicates the delineation of taxonomic significant units (Hörandl, 2022). In chapter 5, we tried to find SNP patterns indicating potential diploid parent taxa using RADseq data but were ultimately not able to find any. Phased, long read sequencing such as those produced by Nanopore sequencing might help to reconstruct the phylogenetic history of polyploids in relation to potential parents with a lower ploidy level. Based on this DNA sequencing data, it might be worthwhile to reconstruct diploid sub genomes (e.g., Lautenschlager et al., 2020) or to search for reticulation events explaining the currently observed polyploids (e.g., Yan et al., 2022). The problem of finding diploid parent species, however, is complicated by multiple formation of polyploid species from the same parental species, hybridization of parental species prior to polyploidization, hybridization of the polyploids after polyploidization, or potential chromosome reduction after polyploidization, to only name few. The species delimitation of the *Leucanthemum* taxa with higher ploidy levels (>4x) is still an open research question that hopefully might be

answered by applying an adapted version of the integrative species delimitation pipeline proposed in the present thesis.

# References


Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *ArXiv:1603.04467 [Cs]*. http://arxiv.org/abs/1603.04467

Afonso, M., Fonteijn, H., Fiorentin, F. S., Lensink, D., Mooij, M., Faber, N., Polder, G., & Wehrens, R. (2020). Tomato Fruit Detection and Counting in Greenhouses Using Deep Learning. *Frontiers in Plant Science*, *11*. https://www.frontiersin.org/article/10.3389/fpls.2020.571299

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, *26*(1), 32–46. https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x

Barclay, R. S., Wilf, P., Dilcher, D. L., & McElwain, J. C. (2012). *The cuticle database project*. http://cuticledb.eesi.psu.edu

Belle, V., & Papantonis, I. (2021). Principles and Practice of Explainable Machine Learning. *Frontiers in Big Data*, *4*. https://www.frontiersin.org/article/10.3389/fdata.2021.688969

Bonhomme, V., Picq, S., Gaucherel, C., & Claude, J. (2014). Momocs: Outline Analysis Using R. *Journal of Statistical Software*, *56*(13), 1–24. https://doi.org/10.18637/jss.v056.i13

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., & Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLOS Computational Biology*, *10*(4), e1003537. https://doi.org/10.1371/journal.pcbi.1003537

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., & RoyChoudhury, A. (2012). Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis. *Molecular Biology and Evolution*, *29*(8), 1917–1932. https://doi.org/10.1093/molbev/mss086

Buddha, K., Nelson, H. J., Zermas, D., & Papanikolopoulos, N. (2019). Weed Detection and Classification in High Altitude Aerial Images for Robot-Based Precision Agriculture. *2019 27th Mediterranean Conference on Control and Automation (MED)*, 280–285. https://doi.org/10.1109/MED.2019.8798582

Burgman, M. A., & Fox, J. C. (2003). Bias in species range estimates from minimum convex polygons: Implications for conservation and options for improved planning. *Animal Conservation*, *6*(1), 19–28. https://doi.org/10.1017/S1367943003003044

Capinha, C., & Pateiro-López, B. (2014). Predicting species distributions in new areas or time periods with alpha-shapes. *Ecological Informatics*, *24*, 231–237. https://doi.org/10.1016/j.ecoinf.2014.06.001

Cardillo, M., & Warren, D. L. (2016). Analysing patterns of spatial and niche overlap among species at multiple resolutions. *Global Ecology and Biogeography*, *25*(8), 951–963. https://doi.org/10.1111/geb.12455

Carranza-Rojas, J., Goeau, H., Bonnet, P., Mata-Montero, E., & Joly, A. (2017). Going deeper in the automated identification of Herbarium specimens. *BMC Evolutionary Biology*, *17*(1), 181. https://doi.org/10.1186/s12862-017-1014-z

Carrasco, M., Toledo, P. A., Velázquez, R., & Bruno, O. M. (2020). Automatic Stomatal Segmentation Based on Delaunay-Rayleigh Frequency Distance. *Plants*, *9*(11), 1613. https://doi.org/10.3390/plants9111613

Carstens, B. C., & Satler, J. D. (2013). The carnivorous plant described as *Sarracenia alata* contains two cryptic species. *Biological Journal of the Linnean Society*, *109*(4), 737–746. https://doi.org/10.1111/bij.12093

Casado-García, A., del-Canto, A., Sanz-Saez, A., Pérez-López, U., Bilbao-Kareaga, A., Fritschi, F. B., Miranda-Apodaca, J., Muñoz-Rueda, A., Sillero-Martínez, A., Yoldi-Achalandabaso, A., Lacuesta, M., & Heras, J. (2020). LabelStoma: A tool for stomata detection based on the YOLO algorithm. *Computers and Electronics in Agriculture*, *178*, 105751. https://doi.org/10.1016/j.compag.2020.105751

Cheng, H. K., Chung, J., Tai, Y.-W., & Tang, C.-K. (2020). CascadePSP: Toward Class-Agnostic and Very High-Resolution Segmentation via Global and Local Refinement. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8887–8896. https://doi.org/10.1109/CVPR42600.2020.00891

Chew, P. L. (1989). Constrained delaunay triangulations. *Algorithmica*, *4*(1), 97–108. https://doi.org/10.1007/BF01553881

Chuanromanee, T. S., Cohen, J. I., & Ryan, G. L. (2019). Morphological Analysis of Size and Shape (MASS): An integrative software program for morphometric analyses of leaves. *Applications in Plant Sciences*, *7*(9), e11288. https://doi.org/10.1002/aps3.11288

Corney, D. P. A., Clark, J. Y., Tang, H., & Wilkin, P. (2012a). Automatic Extraction of Leaf Characters from Herbarium Specimens. *Taxon*, 61(1), 231-244. https://doi.org/10.1002/TAX.611016

Corney, D. P. A., Tang, H. L., Clark, J. Y., Hu, Y., & Jin, J. (2012b). Automating Digital Leaf Measurement: The Tooth, the Whole Tooth, and Nothing but the Tooth. *PLOS ONE*, *7*(8), e42112. https://doi.org/10.1371/journal.pone.0042112

Cracraft, J. (1983). Species Concepts and Speciation Analysis. In R. F. Johnston (Ed.), *Current Ornithology* (pp. 159–187). Springer US. https://doi.org/10.1007/978-1-4615-6781-3_6

Daglio, L. G., & Dawson, M. N. (2019). Integrative taxonomy: Ghosts of past, present and future. *Journal of the Marine Biological Association of the United Kingdom*, *99*(6), 1237–1246. https://doi.org/10.1017/S0025315419000201

Dayrat, B. (2005). Towards integrative taxonomy. *Biological Journal of the Linnean Society*, *85*(3), 407–417. https://doi.org/10.1111/j.1095-8312.2005.00503.x

De Queiroz, K. (2007). Species Concepts and Species Delimitation. *Systematic Biology*, *56*(6), 879–886. https://doi.org/10.1080/10635150701701083

Dorfner, M., Ott, T., Ott, P., & Oberprieler, C. (2022). Long-read genotyping with SLANG (Simple-Long-read loci Assembly of Nanopore data for Genotyping). *Applications in Plant Sciences*, *10*(3), e11484.

Doyen, J. T., & Slobodchikoff, C. N. (1974). An Operational Approach to Species Classification. *Systematic Biology*, *23*(2), 239–247. https://doi.org/10.1093/sysbio/23.2.239

Doyle, J. J., & Dickson, E. E. (1987). Preservation of Plant Samples for DNA Restriction Endonuclease Analysis. *Taxon*, *36*(4), 715–722. https://doi.org/10.2307/1221122

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for Ancient Admixture between Closely Related Populations. *Molecular Biology and Evolution*, *28*(8), 2239–2252. https://doi.org/10.1093/molbev/msr048

Dutta, A., & Zisserman, A. (2019). The VIA Annotation Software for Images, Audio and Video. *Proceedings of the 27th ACM International Conference on Multimedia*, 2276–2279. https://doi.org/10.1145/3343031.3350535

Eaton, D. A. R., & Overcast, I. (2020). ipyrad: Interactive assembly and analysis of RADseq datasets. *Bioinformatics*, *36*(8), 2592–2594. https://doi.org/10.1093/bioinformatics/btz966

Ence, D. D., & Carstens, B. C. (2011). SpedeSTEM: A rapid and accurate method for species delimitation. *Molecular Ecology Resources*, *11*(3), 473–480. https://doi.org/10.1111/j.1755-0998.2010.02947.x

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, *88*(2), 303–338. https://doi.org/10.1007/s11263-009-0275-4

Fetter, K. C., Eberhardt, S., Barclay, R. S., Wing, S., & Keller, S. R. (2019). StomataCounter: A neural network for automatic stomata identification and counting. *New Phytologist*, *223*(3), 1671–1681. https://doi.org/10.1111/nph.15892

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

Fu, L., Lin, P., Vasilakos, A. V., & Wang, S. (2020). An overview of recent multi-view clustering. *Neurocomputing*, *402*, 148–161. https://doi.org/10.1016/j.neucom.2020.02.104

Girshick, R. (2015). Fast R-CNN. *ArXiv:1504.08083 [Cs]*. http://arxiv.org/abs/1504.08083

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *ArXiv:1311.2524 [Cs]*. http://arxiv.org/abs/1311.2524

Goëau, H., Mora-Fallas, A., Champ, J., Love, N. L. R., Mazer, S. J., Mata-Montero, E., Joly, A., & Bonnet, P. (2020). A new fine-grained method for automated visual analysis of herbarium specimens: A case study for phenological data extraction. *Applications in Plant Sciences*, *8*(6), e11368. https://doi.org/10.1002/aps3.11368

Gómez-de-Mariscal, E., García-López-de-Haro, C., Ouyang, W., Donati, L., Lundberg, E., Unser, M., Muñoz-Barrutia, A., & Sage, D. (2021). DeepImageJ: A user-friendly environment to run deep learning models in ImageJ. *Nature Methods*, *18*(10), 1192–1195. https://doi.org/10.1038/s41592-021-01262-9

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., Hansen, N. F., Durand, E. Y., Malaspinas, A.-S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., … Pääbo, S. (2010). A Draft Sequence of the Neandertal Genome. *Science*, *328*(5979), 710–722. https://doi.org/10.1126/science.1188021

Greiner, R., & Oberprieler, C. (2012). The role of inter-ploidy block for reproductive isolation of the diploid *Leucanthemum pluriflorum* Pau (Compositae, Anthemideae)

and its tetra- and hexaploid relatives. *Flora - Morphology, Distribution, Functional Ecology of Plants*, *207*(9), 629–635. https://doi.org/10.1016/j.flora.2012.07.001

Greiner, R., Vogt, R., & Oberprieler, C. (2013). Evolution of the polyploid north-west Iberian *Leucanthemum pluriflorum* clan (Compositae, Anthemideae) based on plastid DNA sequence variation and AFLP fingerprinting. *Annals of Botany*, *111*(6), 1109–1123. https://doi.org/10.1093/aob/mct075

Guillot, G., Renaud, S., Ledevin, R., Michaux, J., & Claude, J. (2012). A Unifying Model for the Analysis of Phenotypic, Genetic, and Geographic Data. *Systematic Biology*, *61*(6), 897–911. https://doi.org/10.1093/sysbio/sys038

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hausdorf, B., & Hennig, C. (2020). Species delimitation and geography. *Molecular Ecology Resources*, *20*(4), 950–960. https://doi.org/10.1111/1755-0998.13184

He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2980–2988. https://doi.org/10.1109/ICCV.2017.322

Hebert, P. D. N., & Gregory, T. R. (2005). The Promise of DNA Barcoding for Taxonomy. *Systematic Biology*, *54*(5), 852–859. https://doi.org/10.1080/10635150500354886

Hengl, T., Jesus, J. M. de, Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE*, *12*(2), e0169748. https://doi.org/10.1371/journal.pone.0169748

Henries, D. G., & Tashakkori, R. (2012). Extraction of leaves from herbarium images. *2012 IEEE International Conference on Electro/Information Technology*, 1–6. https://doi.org/10.1109/EIT.2012.6220752

Heywood, V. H. (1976). *Leucanthemum* Miller. In *Flora Europaea* (Vol. 4).

Hörandl, E. (2022). Novel Approaches for Species Concepts and Delimitation in Polyploids and Hybrids. *Plants*, *11*(2), 204. https://doi.org/10.3390/plants11020204

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., & Murphy, K. (2017). Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3296–3297. https://doi.org/10.1109/CVPR.2017.351

Huson, D. H., & Bryant, D. (2006). Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution*, *23*(2), 254–267. https://doi.org/10.1093/molbev/msj030

Hussein, B. R., Malik, O. A., Ong, W.-H., & Slik, J. W. F. (2021). Reconstruction of damaged herbarium leaves using deep learning techniques for improving classification accuracy. *Ecological Informatics*, *61*, 101243. https://doi.org/10.1016/j.ecoinf.2021.101243

Hussein, B. R., Malik, O. A., Ong, W.-H., & Slik, J. W. F. (2022). Applications of computer vision and machine learning techniques for digitized herbarium specimens: A systematic literature review. *Ecological Informatics*, *69*, 101641. https://doi.org/10.1016/j.ecoinf.2022.101641

Igarashi, T., Moscovich, T., & Hughes, J. F. (2005). As-Rigid-as-Possible Shape Manipulation. *ACM Trans. Graph.*, *24*(3), 1134–1141. https://doi.org/10.1145/1073204.1073323

Jacobson, A., Panozzo, D., Schüller, C., Diamanti, O., Zhou, Q., Pietroni, N., & others. (2018). *libigl: A simple C++ geometry processing library*.

Jin, T., Hou, X., Li, P., & Zhou, F. (2015). A Novel Method of Automatic Plant Species Identification Using Sparse Representation of Leaf Tooth Features. *PLOS ONE*, *10*(10), e0139482. https://doi.org/10.1371/journal.pone.0139482

Joly, S., & Bruneau, A. (2006). Incorporating Allelic Variation for Reconstructing the Evolutionary History of Organisms from Multiple Genes: An Example from *Rosa* in North America. *Systematic Biology*, *55*(4), 623–636. https://doi.org/10.1080/10635150600863109

Jones, H. G. (2020). What plant is that? Tests of automated image recognition apps for plant identification on plants from the British flora. *AoB PLANTS*, *12*(6), plaa052. https://doi.org/10.1093/aobpla/plaa052

Kincaid, D. T., & Schneider, R. B. (1983). Quantification of leaf shape with a microcomputer and Fourier transform. *Canadian Journal of Botany*, *61*(9), 2333–2342. https://doi.org/10.1139/b83-256

Konowalik, K., Wagner, F., Tomasello, S., Vogt, R., & Oberprieler, C. (2015). Detecting reticulate relationships among diploid *Leucanthemum* Mill. (Compositae, Anthemideae) taxa using multilocus species tree reconstruction methods and AFLP fingerprinting. *Molecular Phylogenetics and Evolution*, *92*, 308–328. https://doi.org/10.1016/j.ympev.2015.06.003

Kuhl, F. P., & Giardina, C. R. (1982). Elliptic Fourier features of a closed contour. *Computer Graphics and Image Processing*, *18*(3), 236–258. https://doi.org/10.1016/0146-664X(82)90034-X

Lai, Y., He, S., Lin, Z., Yang, F., Zhou, Q., & Zhou, X. (2019). An Adaptive Robust Semi-Supervised Clustering Framework Using Weighted Consensus of Random $k$k-Means Ensemble. *IEEE Transactions on Knowledge and Data Engineering*, *33*(5), 1877–1890. https://doi.org/10.1109/TKDE.2019.2952596

Lautenschlager, U., Wagner, F., & Oberprieler, C. (2020). AllCoPol: Inferring allele co-ancestry in polyploids. *BMC Bioinformatics*, *21*(1), 441. https://doi.org/10.1186/s12859-020-03750-9

Leaché, A. D., & Bouckaert, R. R. (2018). Species trees and species delimitation with SNAPP: a tutorial and worked example. *Workshop on Population and Speciation Genomics, Český Krumlov*.

Leaché, A. D., Fujita, M. K., Minin, V. N., & Bouckaert, R. R. (2014). Species Delimitation using Genome-Wide SNP Data. *Systematic Biology*, *63*(4), 534–542. https://doi.org/10.1093/sysbio/syu018

Li, K., Huang, J., Song, W., Wang, J., Lv, S., & Wang, X. (2019). Automatic segmentation and measurement methods of living stomata of plants based on the CV model. *Plant Methods*, *15*(1), 67. https://doi.org/10.1186/s13007-019-0453-5

Lin, H.-Y., Gu, K.-J., Li, W.-H., & Zhao, Y.-P. (2021). Integrating coalescent-based species delimitation with ecological niche modeling delimited two species within the *Stewartia sinensis* complex (Theaceae). *Journal of Systematics and Evolution*, *n/a*(n/a). https://doi.org/10.1111/jse.12732

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014* (Vol. 8693, pp. 740–755). Springer International Publishing. https://doi.org/10.1007/978-3-319-10602-1_48

Lingoes, J. C. (1971). Some boundary conditions for a monotone analysis of symmetric matrices. *Psychometrika*, *36*(2), 195–203. https://doi.org/10.1007/BF02291398

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. *ArXiv:1512.02325 [Cs]*, *9905*, 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

Lorieul, T., Pearson, K. D., Ellwood, E. R., Goëau, H., Molino, J., Sweeney, P. W., Yost, J. M., Sachs, J., Mata-Montero, E., Nelson, G., Soltis, P. S., Bonnet, P., & Joly, A. (2019). Toward a large-scale and deep phenological stage annotation of herbarium specimens: Case studies from temperate, tropical, and equatorial floras. *Applications in Plant Sciences*, *7*(3), e01233. https://doi.org/10.1002/aps3.1233

Marchi, P. (1982). *Leucanthemum*—Margherita (9341). In *Flora d'Italia* (Vol. 3, p. 795). Edagricole.

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., & Emerson, B. C. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*, *15*(1), 28–41. https://doi.org/10.1111/1755-0998.12291

Mayr, E. (1970). *Populations, Species, and Evolution: An Abridgment of Animal Species and Evolution*. Harvard University Press.

McLellan, T., & Endler, J. A. (1998). The Relative Success of Some Methods for Measuring and Describing the Shape of Complex Objects. *Systematic Biology*, *47*(2), 264–281. https://doi.org/10.1080/106351598260914

Meeus, S., Van den Bulcke, J., & wyffels, F. (2020). From leaf to label: A robust automated workflow for stomata detection. *Ecology and Evolution*, *10*(17), 9178–9191. https://doi.org/10.1002/ece3.6571

Meusel, H., & Jäger, E. J. (1992). *Vergleichende Chorologie der Zentraleuropäischen Flora* (Vol. 3). Gustav Fischer Verlag.

Meyer, L., Diniz-Filho, J. A. F., & Lohmann, L. G. (2017). A comparison of hull methods for estimating species ranges and richness maps. *Plant Ecology & Diversity*, *10*(5–6), 389–401. https://doi.org/10.1080/17550874.2018.1425505

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning*, *52*(1), 91–118. https://doi.org/10.1023/A:1023949509487

Munisami, T., Ramsurn, M., Kishnah, S., & Pudaruth, S. (2015). Plant Leaf Recognition Using Shape Features and Colour Histogram with K-nearest Neighbour Classifiers. *Procedia Computer Science*, *58*, 740–747. https://doi.org/10.1016/j.procs.2015.08.095

Naciri, Y., & Linder, H. P. (2015). Species delimitation and relationships: The dance of the seven veils. *TAXON*, *64*(1), 3–16. https://doi.org/10.12705/641.24

Nieuwenhuizen, A. T., Hemming, J., & Suh, H. K. (2018). *Detection and classification of insects on stick-traps in a tomato crop using Faster R-CNN.* The Netherlands Conference on Computer Vision. https://edepot.wur.nl/463457

Norouzzadeh, M. S., Morris, D., Beery, S., Joshi, N., Jojic, N., & Clune, J. (2021). A deep active learning system for species identification and counting in camera trap images. *Methods in Ecology and Evolution*, *12*(1), 150–161. https://doi.org/10.1111/2041-210X.13504

Oberprieler, C., Conti, F., Dorfner, M., Eder, S.-M., Heuschneider, A., Ott, T., Scheunert, A., & Vogt, R. (2022). The taxonomy of *Leucanthemum ircutianum* (Asteraceae, Anthemideae) in the Apennine Peninsula based on AFLP fingerprinting, plastid DNA sequence variation and eco-climatological niche reconstruction. *Botanical Journal of the Linnean Society*, boac003. https://doi.org/10.1093/botlinnean/boac003

Oberprieler, C., Eder, C., Meister, J., & Vogt, R. (2011). AFLP fingerprinting suggests an allopolyploid origin of two members of the *Leucanthemum vulgare* aggregate (Compositae, Anthemideae) in central Europe. *Nordic Journal of Botany*, *29*(3), 370–377. https://doi.org/10.1111/j.1756-1051.2011.01025.x

Oberprieler, C., Greiner, R., Konowalik, K., & Vogt, R. (2014). The reticulate evolutionary history of the polyploid NW Iberian *Leucanthemum pluriflorum* clan (Compositae, Anthemideae) as inferred from nrDNA ETS sequence diversity and eco-climatological niche-modelling. *Molecular Phylogenetics and Evolution*, *70*, 478–491. https://doi.org/10.1016/j.ympev.2013.10.013

Oberprieler, C., Konowalik, K., Fackelmann, A., & Vogt, R. (2018). Polyploid speciation across a suture zone: Phylogeography and species delimitation in S French

*Leucanthemum* Mill. representatives (Compositae–Anthemideae). *Plant Systematics and Evolution*, *304*(9), 1141–1155. https://doi.org/10.1007/s00606-018-1537-9

O'Shea, K., & Nash, R. (2015). An Introduction to Convolutional Neural Networks. *ArXiv:1511.08458 [Cs]*. http://arxiv.org/abs/1511.08458

Ott, T., Palm, C., Vogt, R., & Oberprieler, C. (2020). GinJinn: An object-detection pipeline for automated feature extraction from herbarium specimens. *Applications in Plant Sciences*, *8*(6). https://doi.org/10.1002/aps3.11351

Padial, J. M., Miralles, A., De la Riva, I., & Vences, M. (2010). The integrative future of taxonomy. *Frontiers in Zoology*, *7*(1), 16. https://doi.org/10.1186/1742-9994-7-16

Pante, E., Schoelinck, C., & Puillandre, N. (2015). From Integrative Taxonomy to Species Description: One Step Beyond. *Systematic Biology*, *64*(1), 152–160. https://doi.org/10.1093/sysbio/syu083

Paradis, E., & Schliep, K. (2019). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). *Automatic differentiation in PyTorch*. https://openreview.net/forum?id=BJJsrmfCZ

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peralta, M., Combes, M.-C., Cenci, A., Lashermes, P., & Dereeper, A. (2013). SNiPloid: A Utility to Exploit High-Throughput SNP Data Derived from RNA-Seq in Allopolyploid Species. *International Journal of Plant Genomics*, *2013*, e890123. https://doi.org/10.1155/2013/890123

Phillips, S., Dudík, M., & Schapire, R. (2021). *Maxent software for modeling species niches and distributions* (3.4.4) [Computer software]. http://biodiversityinformatics.amnh.org/open_source/maxent/

Poland, J. A., Brown, P. J., Sorrells, M. E., & Jannink, J.-L. (2012). Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. *PLOS ONE*, *7*(2), e32253. https://doi.org/10.1371/journal.pone.0032253

Raxworthy, C. J., Ingram, C. M., Rabibisoa, N., & Pearson, R. G. (2007). Applications of Ecological Niche Modeling for Species Delimitation: A Review and Empirical Evaluation Using Day Geckos (*Phelsuma*) from Madagascar. *Systematic Biology*, *56*(6), 907–923. https://doi.org/10.1080/10635150701775111

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *ArXiv:1506.02640 [Cs]*. http://arxiv.org/abs/1506.02640

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems 28 (NIPS 2015)*. NeurIPS 2015. https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf

Reydon, T. A. C., & Kunz, W. (2019). Species as natural entities, instrumental units and ranked taxa: New perspectives on the grouping and ranking problems. *Biological Journal of the Linnean Society*, *126*(4), 623–636. https://doi.org/10.1093/biolinnean/blz013

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access*, *8*, 42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199

Royer, D. L., McElwain, J. C., Adams, J. M., & Wilf, P. (2008). Sensitivity of leaf size and shape to climate within *Acer rubrum* and *Quercus kelloggii*. *New Phytologist*, *179*(3), 808–817. https://doi.org/10.1111/j.1469-8137.2008.02496.x

Royer, D. L., Wilf, P., Janesko, D. A., Kowalski, E. A., & Dilcher, D. L. (2005). Correlations of climate and plant ecology to leaf size and shape: Potential proxies for the fossil record. *American Journal of Botany*, *92*(7), 1141–1151. https://doi.org/10.3732/ajb.92.7.1141

Schlick-Steiner, B. C., Arthofer, W., & Steiner, F. M. (2014). Take up the challenge! Opportunities for evolution research from resolving conflict in integrative taxonomy. *Molecular Ecology*, *23*(17), 4192–4194. https://doi.org/10.1111/mec.12868

Schlick-Steiner, B. C., Steiner, F. M., Seifert, B., Stauffer, C., Christian, E., & Crozier, R. H. (2010). Integrative Taxonomy: A Multisource Approach to Exploring Biodiversity. *Annual Review of Entomology*, *55*(1), 421–438. https://doi.org/10.1146/annurev-ento-112408-085432

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, *9*(7), 671–675. https://doi.org/10.1038/nmeth.2089

Schroeder, A. B., Dobson, E. T. A., Rueden, C. T., Tomancak, P., Jug, F., & Eliceiri, K. W. (2021). The ImageJ ecosystem: Open-source software for image visualization, processing, and analysis. *Protein Science*, *30*(1), 234–249. https://doi.org/10.1002/pro.3993

Sennen, F. (1936). *Diagnoses des nouveautés parues dans les e exsiccata Plantes d'Espagne et du Maroc de 1928 à 1935*.

Shewchuk, J. R. (1996). Triangle: Engineering a 2D quality mesh generator and Delaunay triangulator. In M. C. Lin & D. Manocha (Eds.), *Applied Computational Geometry Towards Geometric Engineering* (pp. 203–222). Springer. https://doi.org/10.1007/BFb0014497

Silva Vieira, G. D., Maria de Sousa, N., Rocha, B., Fonseca, A. U., & Soares, F. (2021). A Method for the Detection and Reconstruction of Foliar Damage caused by Predatory Insects. *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, 1502–1507. https://doi.org/10.1109/COMPSAC51774.2021.00223

Solís-Lemus, C., Knowles, L. L., & Ané, C. (2015). Bayesian species delimitation combining multiple genes and traits in a unified framework. *Evolution*, *69*(2), 492–507. https://doi.org/10.1111/evo.12582

Song, W., Li, J., Li, K., Chen, J., & Huang, J. (2020). An Automatic Method for Stomatal Pore Detection and Measurement in Microscope Images of Plant Leaf Based on a Convolutional Neural Network Model. *Forests*, *11*(9), 954. https://doi.org/10.3390/f11090954

Stankowski, S., & Ravinet, M. (2021). Defining the speciation continuum. *Evolution*, *75*(6), 1256–1273. https://doi.org/10.1111/evo.14215

Stuessy, T. F. (1979). Ultrastructural Data for the Practicing Plant Systematist. *American Zoologist*, *19*(2), 621–636. https://doi.org/10.1093/icb/19.2.621

Stuessy, T. F., Crawford, D. J., Soltis, D. E., & Soltis, P. S. (2014). *Plant Systematics: The Origin, Interpretation, and Ordering of Plant Biodiversity*. Koeltz Scientific Books.

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

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2016). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *ArXiv:1602.07261 [Cs]*. http://arxiv.org/abs/1602.07261

*The Euro+Med Plantbase Project*. Retrieved March 24, 2022, from http://ww2.bgbm.org/EuroPlusMed/query.asp

Toda, Y., Toh, S., Bourdais, G., Robatzek, S., Maclean, D., & Kinoshita, T. (2018). *DeepStomata: Facial Recognition Technology for Automated Stomatal Aperture Measurement* [Preprint]. Bioinformatics. https://doi.org/10.1101/365098

Ubbens, J. R., & Stavness, I. (2017). Deep Plant Phenomics: A Deep Learning Platform for Complex Plant Phenotyping Tasks. *Frontiers in Plant Science*, *8*. https://www.frontiersin.org/article/10.3389/fpls.2017.01190

Unger, J., Merhof, D., & Renner, S. (2016). Computer vision applied to herbarium specimens of German trees: Testing the future utility of the millions of herbarium specimen images for automated identification. *BMC Evolutionary Biology*, *16*(1), 248. https://doi.org/10.1186/s12862-016-0827-5

Valen, L. V. (1976). Ecological Species, Multispecies, and Oaks. *Taxon*, *25*(2/3), 233–239. https://doi.org/10.2307/1219444

van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., & Yu, T. (2014). scikit-image: Image processing in Python. *PeerJ*, *2*, e453. https://doi.org/10.7717/peerj.453

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

Vásquez-Cruz, M., Vovides, A. P., & Sosa, V. (2017). Disentangling Species Limits in the *Vauquelinia corymbosa* Complex (Pyreae, Rosaceae). *Systematic Botany*, *42*(4), 835–847. https://doi.org/10.1600/036364417X696519

Vogt, R. (1991). *Die Gattung Leucanthemum Mill. (Compositae—Anthemideae) auf der Iberischen Halbinsel*. Real Jardín Botánico, C.S.I.C. https://books.google.de/books?id=4WvdOr5P65QC

Vogt, R. (2019). *Leucanthemum* Mill. In *Flora Iberica* (3rd ed., Vol. 16, pp. 1848–1880). Real Jardín Botánico.

Von Wettstein, R. (1898). *Grundzüge der geographisch-morphologischen Methode der Pflanzensystematik*. Gustav Fischer Verlag.

Wagenitz, G. (1977). Zur Bestimmung der *Leucanthemum* Arten in Mitteleuropa nördlich der Alpen. *Gottinger Floristische Rundbriefe*, *10*, 80–85.

Wagner, F., Härtl, S., Vogt, R., & Oberprieler, C. (2017). “Fix Me Another Marguerite!”: Species delimitation in a group of intensively hybridizing lineages of ox-eye daisies (*Leucanthemum* Mill., Compositae-Anthemideae). *Molecular Ecology*, *26*(16), 4260–4283. https://doi.org/10.1111/mec.14180

Wagner, F., Ott, T., Schall, M., Lautenschlager, U., Vogt, R., & Oberprieler, C. (2020). Taming the Red Bastards: Hybridisation and species delimitation in the *Rhodanthemum arundanum*-group (Compositae, Anthemideae). *Molecular Phylogenetics and Evolution*, *144*, 106702. https://doi.org/10.1016/j.ympev.2019.106702

Wagner, F., Ott, T., Zimmer, C., Reichhart, V., Vogt, R., & Oberprieler, C. (2019). ‘At the crossroads towards polyploidy’: Genomic divergence and extent of homoploid hybridization are drivers for the formation of the ox-eye daisy polyploid complex (*Leucanthemum*, Compositae-Anthemideae). *New Phytologist*, *223*(4), 2039–2053. https://doi.org/10.1111/nph.15784

Wagner, N. D., He, L., & Hörandl, E. (2020). Phylogenomic Relationships and Evolution of Polyploid *Salix* Species Revealed by RAD Sequencing Data. *Frontiers in Plant Science*, *11*. https://www.frontiersin.org/article/10.3389/fpls.2020.01077

Wäldchen, J., & Mäder, P. (2018). Machine learning for image based species identification. *Methods in Ecology and Evolution*, *9*(11), 2216–2225. https://doi.org/10.1111/2041-210X.13075

Warren, D. L., Matzke, N. J., Cardillo, M., Baumgartner, J. B., Beaumont, L. J., Turelli, M., Glor, R. E., Huron, N. A., Simões, M., Iglesias, T. L., Piquet, J. C., & Dinnage, R. (2021). ENMTools 1.0: An R package for comparative ecological biogeography. *Ecography*, *44*(4), 504–511. https://doi.org/10.1111/ecog.05485

Weaver, W. N., Ng, J., & Laport, R. G. (2020). LeafMachine: Using machine learning to automate leaf trait extraction from digitized herbarium specimens. *Applications in Plant Sciences*, *8*(6), e11367. https://doi.org/10.1002/aps3.11367

White, A. E., Dikow, R. B., Baugh, M., Jenkins, A., & Frandsen, P. B. (2020). Generating segmentation masks of herbarium specimens and a data set for training segmentation models using deep learning. *Applications in Plant Sciences*, *8*(6), e11352. https://doi.org/10.1002/aps3.11352

Wiley, E. O. (1978). The Evolutionary Species Concept Reconsidered. *Systematic Biology*, *27*(1), 17–26. https://doi.org/10.2307/2412809

Will, K. W., Mishler, B. D., & Wheeler, Q. D. (2005). The Perils of DNA Barcoding and the Need for Integrative Taxonomy. *Systematic Biology*, *54*(5), 844–851. https://doi.org/10.1080/10635150500354878

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2019). *Detectron2*. https://github.com/facebookresearch/detectron2

Yan, Z., Cao, Z., Liu, Y., Ogilvie, H. A., & Nakhleh, L. (2022). Maximum Parsimony Inference of Phylogenetic Networks in the Presence of Polyploid Complexes. *Systematic Biology*, *71*(3), 706–720. https://doi.org/10.1093/sysbio/syab081

Yeates, D. K., Seago, A., Nelson, L., Cameron, S. L., Joseph, L., & Trueman, J. W. H. (2011). Integrative taxonomy, or iterative taxonomy? *Systematic Entomology*, *36*(2), 209–217. https://doi.org/10.1111/j.1365-3113.2010.00558.x

Younis, S., Schmidt, M., Weiland, C., Dressler, S., Seeger, B., & Hickler, T. (2020). Detection and annotation of plant organs from digitised herbarium scans using deep learning. *Biodiversity Data Journal*, *8*, e57090. https://doi.org/10.3897/BDJ.8.e57090

Younis, S., Weiland, C., Hoehndorf, R., Dressler, S., Hickler, T., Seeger, B., & Schmidt, M. (2018). Taxon and trait recognition from digitized herbarium specimens using deep convolutional neural networks. *Botany Letters*, *165*(3–4), 377–383. https://doi.org/10.1080/23818107.2018.1446357

Zachos, F. E. (2016). *Species Concepts in Biology*. Springer International Publishing. https://doi.org/10.1007/978-3-319-44966-1

Zapata, F., & Jiménez, I. (2012). Species Delimitation: Inferring Gaps in Morphology across Geography. *Systematic Biology*, *61*(2), 179. https://doi.org/10.1093/sysbio/syr084

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, *19*(6), 153. https://doi.org/10.1186/s12859-018-2129-y

Zhao, Z.-Q., Zheng, P., Xu, S.-T., & Wu, X. (2019). Object Detection With Deep Learning: A Review. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(11), 3212–3232. https://doi.org/10.1109/TNNLS.2018.2876865

Zheng, H., Fan, L., Milne, R. I., Zhang, L., Wang, Y., & Mao, K. (2017). Species Delimitation and Lineage Separation History of a Species Complex of Aspens in China. *Frontiers in Plant Science*, *8*, 375. https://doi.org/10.3389/fpls.2017.00375

## Acknowledgements

First, I want to thank my supervisor Prof. Dr. Christoph Oberprieler for igniting my interest in plant systematics, for guiding me during my thesis, and for always taking the time for exciting discussions on plant taxonomy, systematic methods, and especially species concepts and species delimitation.

I would like to thank Prof. Dr. Christoph Palm, my first mentor, for introducing me to computer vision and object detection, and for his very helpful consulting. I also want to thank my second mentor Dr. Robert Vogt for his valuable support on all things concerning the genus *Leucanthemum* and the formulation of taxonomic consequences.

For the great collaboration on GINJINN2, I want to sincerely thank my colleague Ulrich Lautenschlager. Likewise, I am grateful for all who directly or indirectly contributed to this thesis, namely Maximilian Schall, and Agnes Scheunert, Marco Dorfner, Conan Wolff and Daniel Sulas.

The participation of Prof. Dr. Florian Hartig, Prof. Dr. Joachim Ruther, and Prof. Dr. Christoph Schubart in the examination committee is very much appreciated.