https://doi.org/10.35662/unine-thesis-2792

UNIVERSITÉ DE NEUCHÂTEL

Assessing wild bee diversity using next generation sequencing

By

Morgan Gueuning

PhD thesis submitted to the Institute of Biology,
University of Neuchâtel,
Switzerland

Thesis committee:
Dr. Christophe Praz (co-director)
Prof. Betty Benrey (co-director)
Dr. Jürg Frey (Supervisor)
Prof. Edward Mitchell
Prof. Nadir Alvarez

Defended on 18 October 2019

UNIVERSITÉ DE NEUCHÂTEL

Faculté des Sciences
Secrétariat-décanat de Faculté
Rue Emile-Argand 11
2000 Neuchâtel – Suisse
Tél : + 41 (0)32 718 21 00
E-mail : secretariat.sciences@unine.ch

IMPRIMATUR POUR THESE DE DOCTORAT

La Faculté des sciences de l'Université de Neuchâtel autorise l'impression de la présente thèse soutenue par

Monsieur Morgan GUEUNING

Titre:

"Assessing wild bee diversity using next generation sequencing"

sur le rapport des membres du jury composé comme suit:

- Dr Christophe Praz, co-directeur de thèse, Université de Neuchâtel, Suisse
- Prof. titulaire Betty Benrey, co-directrice de thèse, Université de Neuchâtel, Suisse
- Dr Jürg Frey, Agroscope, Wädenwil, Suisse
- Prof. Edward Mitchell, Université de Neuchâtel, Suisse
- Prof. Nadir Alvarez, Muséum d'histoire naturelle, Genève et Université de Lausanne, Suisse

Neuchâtel, le 5 novembre 2019

Le Doyen, Prof. P. Felber

P. Fel

- Remerciements -

Les travaux présentés dans cette thèse de doctorat ont été réalisés avec l'aide et le soutien de nombreuses personnes auxquelles je souhaiterais témoigner ma gratitude.

Je voudrais avant tout remercier mes deux superviseurs, Christophe Praz et Jürg Frey, qui m'ont donné cette chance de travailler à leurs côtés et avec lesquels j'ai pris beaucoup de plaisir à travailler. Ils m'ont offert dans leur guidance une liberté de pensée, de mouvement et d'action inespérée, faisant de ces trois années et demi une expérience incroyablement enrichissante. Je remercie particulièrement Jürg de m'avoir continuellement challengé et poussé à me remettre en question pour alimenter ma réflexion, et particulièrement Christophe pour m'avoir partagé son savoir et sa passion pour les abeilles.

Un grand merci à mes collègues pour leurs inspirations, leurs conseils critiques et constructifs, ainsi que de leur soutien moral. Je tiens à remercier spécialement Beatrice et Daniel pour leur précieuse aide en laboratoire, sans lesquels je n'aurais peut-être pas pu accomplir mon travail de recherche de façon aussi agréable.

J'adresse également mes sincères remerciements à Nadir Alvarez qui a été mon mentor et m'a fait l'honneur d'être présent dans le jury de ma thèse. J'en profite aussi pour remercier les autres membres du jury, Betty Benrey et Edward Mitchell, pour avoir accepté de faire partie du comité de thèse.

Je remercie également mes amis, qui m'ont soutenu durant cette aventure et auxquels j'adresse parallèlement mes excuses pour mes nombreuses absences aux différents événements sociaux avec lesquels le doctorat n'était malheureusement pas toujours compatible.

Enfin, je tiens à exprimer ma gratitude à ma famille de m'avoir offert l'opportunité d'en arriver jusque-là, et d'avoir toujours cru en moi, sans avoir à me dicter la direction à prendre.

Bien évidemment, mes plus tendres remerciements reviennent à ma partenaire de vie, mon étoile du berger qui m'a donné la force et le courage de toujours me dépasser, et de toujours retirer quelque chose de positif des moments les plus douloureux.

Enfin, je souhaiterais remercier l'Agroscope et l'Université de Neuchâtel pour leur encadrement, la dispense de cours de très haute qualité et la mise à disposition des outils nécessaires et infrastructures modernes ayant fortement contribué à la réussite de travail.

- Table of content -

- Thesis abstract -

Wild bees are essential pollinators and therefore play a key role in both natural and agricultural ecosystems. However, bees have often been neglected in conservation studies and policies worldwide, which is surprising given their ecological importance. As a result, little is known on the conservation status of the vast majority of wild bee species in Europe, and even less worldwide. Limited surveys suggest important declines in the abundance and diversity of most wild bee communities worldwide. It is therefore urgent to implement targeted measures for the conservation of these keystone species. Once implemented, the effectiveness of these measures must be evaluated using adequate monitoring programs. To date, wild bee surveys are entirely based on morphological identification, which is both labor intensive and time consuming. Consequently, an affordable, high-throughput identification method is needed to reduce costs and improve bee monitoring.

The objective of this thesis was to evaluate novel genetic techniques based on Next Generation Sequencing (NGS) methods for facilitating surveys of wild bees. NGS tools were mainly investigated for bridging two important impediments to wild bee conservation efforts, i.e., the cost of biodiversity assessment schemes and taxonomic incompleteness. With the development of NGS techniques, DNA barcoding has gained enormous momentum, enabling cost-effective, fast and accurate identifications. Before these methods can be routinely used in monitoring programs, there are however still important knowledge gaps to fill. These gaps mainly concern the detection of rare species and the acquisition of accurate quantitative data on species abundance; more generally the cost and labour effectiveness of these methods need to be evaluated. To provide a comprehensive presentation of the advantages and weaknesses of different NGS-based identification methods, we assessed three of the most promising ones, namely metabarcoding, mitogenomics and NGS barcoding. Using a regular monitoring data, we found that NGS barcoding performed best for both species' presence/absence and abundance data, producing only few false positives and no false negatives. The other methods investigated were less reliable in term of species detection and inference of abundance data, and partly led to erroneous ecological conclusions. In terms of workload and cost, we showed that NGS techniques were more expensive than morphological identification with our dataset, although these techniques would become slightly more economical in large-scale monitoring programs.

A second aim of this thesis was to provide an easy and robust genomic solution to alleviate taxonomical incompleteness, one of the major impediments to the effective conservation of many insect taxa. For conservation purposes, having stable and well-delimited species hypotheses is essential. Currently, most species are delimitated based on morphology and/or DNA barcoding. These methods are however associated with important limitations, and it is widely accepted that species delimitation should rely on multi-locus genomic markers. To overcome these limitations, ultraconserved elements (UCEs) were tested as a fast and robust approach using different species-complexes harbouring cryptic diversity, mitochondrial introgression, or mitochondrial paraphyly. Phylogenetic analyses of UCEs were highly conclusive and yielded meaningful species delimitation hypotheses in all cases. These results provide strong evidence for the potential of UCEs as a fast method for delimiting species even in cases of recently diverged lineages. Advantages and limitations of UCEs for shallow phylogenetic studies are further discussed.

- Résumé de la thèse -

Les abeilles sauvages sont des pollinisateurs essentiels et jouent donc un rôle clé dans les écosystèmes naturels et agricoles. Cependant, les abeilles ont souvent été négligées dans les études et les politiques de conservation dans le monde entier, ce qui est surprenant étant donné leur importance écologique. Par conséquent, nous en savons peu sur l'état de conservation de la grande majorité des espèces d'abeilles sauvages en Europe, et encore moins dans le monde. Des études limitées suggèrent un déclin important de l'abondance et de la diversité de la plupart des communautés d'abeilles sauvages dans le monde. Il est donc urgent de mettre en œuvre des mesures ciblées pour la conservation de ces espèces clés. Une fois mises en place, l'efficacité de ces mesures doit être évaluée au moyen de programmes de surveillance adéquats. Jusqu'à présent, les enquêtes sur les abeilles sauvages sont entièrement basées sur l'identification morphologique, représentant un travail à la fois long et exigeant. Par conséquent, une méthode d'identification abordable et à haut débit est nécessaire pour réduire les coûts et améliorer le suivi des abeilles.

L'objectif de cette thèse était d'évaluer de nouvelles techniques génétiques basées sur les méthodes de séquençage de nouvelle génération (Next Generation Sequencing - NGS) pour faciliter les enquêtes sur les abeilles sauvages. Les outils NGS ont été principalement étudiés pour surmonter deux obstacles importants aux efforts de conservation des abeilles sauvages, à savoir le coût des programmes d'évaluation de la biodiversité et le caractère incomplet de la taxonomie. Avec le développement des techniques NGS, le barcoding d'ADN a pris un énorme essor, permettant des identifications économiques, rapides et précises. Avant que ces méthodes puissent être utilisées de façon routinière dans les programmes de surveillance, il reste toutefois d'importantes lacunes à combler en matière de connaissances. Ces lacunes concernent principalement la détection des espèces rares et l'obtention de données quantitatives précises sur l'abondance des espèces; plus généralement, le coût et la rentabilité de ces méthodes doivent être évalués. Afin de fournir une présentation complète des avantages et des faiblesses des différentes méthodes d'identification basées sur les méthodes NGS, nous avons évalué trois méthodes des plus prometteuses, à savoir le metabarcoding, la mitogénomique et le NGS barcoding. À l'aide de données de surveillance régulières, nous avons constaté que le NGS barcoding donne les meilleurs résultats pour les données sur la présence/absence et l'abondance des espèces, ne produisant que quelques faux positifs et aucun faux négatif. Les autres méthodes étudiées étaient moins fiables en termes de détection des espèces et d'inférence des données d'abondance, et ont conduit en partie à des conclusions écologiques erronées. En termes de charge de travail et de coût, nous avons montré que les techniques NGS étaient plus coûteuses que l'identification morphologique avec notre ensemble de données, bien que ces techniques deviendraient légèrement plus économiques dans les programmes de surveillance à grande échelle.

Un deuxième objectif de cette thèse était de fournir une solution génomique facile et robuste pour alléger les lacunes taxonomiques, l'un des principaux obstacles à la conservation efficace de nombreux taxons d'insectes. Pour la conservation, il est essentiel de disposer d'hypothèses stables et bien délimitées sur les espèces. Actuellement, la plupart des espèces sont délimitées en fonction de leur morphologie et/ou de leur code-barres ADN. Ces méthodes sont toutefois associées à d'importantes limitations, et il est largement admis que la délimitation des espèces devrait reposer sur des marqueurs génomiques multi-locus. Pour surmonter ces limites, les éléments ultra-conservés (UCE) ont été testés comme approche rapide et robuste utilisant différents complexes d'espèces abritant une diversité cryptique, une introgression mitochondriale ou une paraphylie mitochondrial. Les analyses phylogénétiques des UCEs ont été très concluantes et ont donné lieu à des hypothèses de délimitation d'espèces significatives dans tous les cas. Ces résultats fournissent des preuves solides du potentiel des UCEs comme méthode rapide de délimitation des espèces, même dans le cas de lignées ayant récemment divergé. Les avantages et les limites des UCEs pour étudier des divergences phylogénétiques récentes sont discutés plus en détail.

– General introduction –

Understanding and controlling our surrounding nature and environment has always been of tremendous importance and has unquestionably played a vital role in the expansion (both geographical and demographical), and evolutionary success of mankind. For biologists, understanding nature entails, above all, identifying and naming biological species. Carl Linnaeus (1707 - 1778), the so-called "father of modern taxonomy", devoted his life to describing and classifying plants and animals. In total, Linnaeus named close to 4400 animal species and 5900 plant species (Müller-Wille, 2006). Less than three centuries later, 1.8 million species have been named according to Linnaeus' binomial nomenclature system (Roskov et al., 2019). These taxonomical efforts have however not completely bridged some important knowledge gaps, like for instance the overall number of species present on earth. This fundamental question has drawn considerable attention but still remains largely unresolved, with estimates varying between ~2 million (Costello, Wilson, & Houlding, 2012) and ~1 trillion species (Locey & Lennon, 2016). Nevertheless, estimates for insects seem to have stabilized around 5.5. million species (Stork, 2017). Approximately 1 million insect species have so far been described, among which numerous species names are suspected to be synonymous (Stork, 2017), demonstrating the huge knowledge gap for this animal phylum.

Luckily, the development of sequencing technologies has dramatically improved and even changed our perception of biodiversity. With technological advances, we entered a "new age of discovery" (Donoghue & Alverson, 2000) characterized by the discovery of a vast number of new species and ecological processes. Historically, species were solely described and classified based on morphological criteria. However, for many species, especially for microorganisms, morphological identification can be virtually impossible. DNA sequencing has changed this paradigm and revolutionised many aspects of biological research. For instance in microbiology, the description and classification of bacteria was until recently only possible for a small proportion of cultivable strains (Amann, Ludwig, & Schleifer, 1995; Hall, 2007; although see Martiny, 2019), but with the development of next generation sequencing (NGS) it is now possible to explore a vast variety of bacterial communities, even in the most extreme environments (Rothschild & Mancinelli, 2001). The accumulation of data for bacterial communities provided empirical proof of the suspected hyper-diversity of this group, which is estimated to account for more than 70% of the global diversity (Larsen, Miller, Rhodes, & Wiens, 2017).

For larger organisms, DNA sequencing also played a vital role in uncovering "hidden", or "cryptic" diversity. There are many well-documented examples in which DNA sequencing has helped discovering and disentangling unexpected diversity. For instance, in the well-studied group of mammals, more than 1000 new species have been described globally over the last decade (Burgin, Colella, Kahn, & Upham, 2018). Among the newly described species, some belong to the largest animals such as giraffes (Fennessy et al., 2016), elephants (Roca, Georgiadis, Pecon-Slattery, & O'Brien, 2001), dolphins (Charlton, 2007) and orangutans (Nater et al., 2017). These discoveries would not have been possible without DNA assisted species delimitations.

The molecular identification of organisms is mainly based on so-called "DNA barcoding". Analogous to digital barcodes found on commercialized products, DNA barcodes are standard DNA regions used for species identification. Ideally, DNA barcodes should be (Valentini, Pompanon, & Taberlet, 2009): species-specific, yet with minimal intra-specific variation, universal with highly conserved priming sites, phylogenetically informative, but short enough to accommodate technical sequencing limitations. Although the perfect DNA sequence probably does not exist, a portion of the mitochondrial gene cytochrome-oxidase subunit I (COI)

was proposed as a near-perfect candidate in the early 2000' (Brunner, Fleming, & Frey, 2002; Hebert, Cywinska, Ball, & DeWaard, 2003). Since then, this ~650 base pair (bp) long sequence has become the gold standard for animal DNA barcoding, and other taxon like protists (Jan Pawlowski et al., 2012). For plants, bacteria and fungi other barcode gene fragments were found to approach the above mentioned criteria better than COI (Kress, García-Robledo, Uriarte, & Erickson, 2015).

Initially, the application of DNA barcoding aimed for the fast and accurate identification of species (Hebert et al., 2003). With the notable exception of cryptic diversity described above, molecular identification is rarely needed for larger animals. In contrast, for smaller, morphologically uniform animals, classical morphological identification can be very challenging. Indeed, morphological identification requires expertise and morphological identification keys which do not always exist for understudied taxa or biogeographical regions. Furthermore, some taxa exhibit important phenotypic plasticity or sex-dimorphisms rendering identifications difficult. The presence of cryptic diversity, i.e. morphologically near-identical but genetically distinct species, may also obscure morphological identification. A text book example of unexpected diversity unravelled by DNA barcoding is provided by a clade of neotropical skipper butterfly *Astraptes fulgerator* (Hebert, Penton, Burns, Janzen, & Hallwachs, 2004). Based on morphological features at the adult stage, this clade was initially thought to be composed of a single species. However large divergence in DNA barcodes between populations across the distribution range suggested the presence of multiple lineages. Re-examination of the caterpillars' morphology and ecology showed that this taxon is indeed composed of several species. Besides complementing taxonomy, DNA barcoding is a unique tool for species identification in cases where no morphologically identifiable biological material is available. For instance, DNA barcoding is routinely used in the investigation of illegal trades of protected or endangered wildlife, species mislabelling or food industry frauds (Staats et al., 2016).

Given these advantages, DNA barcoding and classical taxonomy have become intimately associated. DNA barcoding-based species identification relies on reference databases containing sequences associated to morphologically identified specimens. The association between correctly identified specimens and their respective DNA barcode is thus vital. The main weakness of this concept is that the main reference databases (i.e. GenBank, EMBL, ITS2db or SILVA) currently contain an increasing number of erroneous sequences, resulting from sequencing errors, contaminations, or misidentifications. To cope with this problem, the Barcode of Life Data Systems database (BOLD, http://www.barcodinglife.org) was initiated. In contrast to GenBank, which acts solely as sequence repository, the philosophy of BOLD is to provide a curated database storing sequences together with associated metadata. To be labelled as "barcode-compliant" sequence in BOLD, sequences must be larger than 500 bp and need to be associated with a specimen deposited in a public institution, the precise collection records, the name of the person that performed the identification, the primer information, and the Sanger sequencing chromatogram. Despite these rather stringent requirements BOLD is not free of errors. A recent study compared the identification performance of BOLD and NCBI using a diverse set of insect taxa and found NCBI to outperform the curated database (Meiklejohn, Damaso, & Robertson, 2019).

Database errors are not the only error source in DNA barcoding. Because animal DNA barcodes are located on the mitochondrial genomes, disagreements between morphology and DNA barcoding may be induced by divergent evolution forces acting solely on the organelle (Funk & Omland, 2003). Such discrepancies are often mediated through adaptive introgression, *Wolbachia* infection or sex-biased asymmetries (i.e. male-biased

dispersal, mating behaviour or sex-biased offspring production) (Toews & Brelsford, 2012). Sequencing of nuclear genes is often required to clarify such disagreements between DNA barcoding and morphological identifications.

DNA barcoding has been revolutionised with the development of NGS. In contrast to Sanger sequencing, the high throughput of NGS platforms now allows capturing DNA barcodes of many organisms in bulk mixtures, a process called metabarcoding. Metabarcoding was used to answer a wide range of questions such as invasive species detection (e.g. Bohmann et al., 2014), gut content analysis (e.g. Leray et al., 2013), and assessment of diversity in difficult taxonomical groups such as bacteria (e.g. Caporaso et al., 2011), fungi (e.g. Amend, Seifert, & Bruns, 2010), plants (e.g. Hiiesalu et al., 2012) or invertebrates (e.g. Ji et al., 2013; Pawlowski, Lejzerowicz, Apotheloz-Perret-Gentil, Visco, & Esling, 2016). Lately, metabarcoding was even used to test hypotheses on the nature of a mystical creature in a well-known Scottish "loch" (Gemmel et al., in press).

The drastic reduction in sequencing cost, the development of easy-to-use bioinformatic tools and comprehensive DNA barcoding databases made NGS-based species identifications very attractive. It was for instance suggested as worthy cost-effective alternative to morphological identifications for routine monitoring of arthropods (Ji et al., 2013; Taberlet, Coissac, Pompanon, Brochmann, & Willerslev, 2012). Traditionally, monitoring programs have relied on morphological identification of species, which is extraordinarily demanding, requires expert knowledge of morphology and taxonomy, and is therefore a lengthy and expensive procedure. For instance for wild bees, Lebuhn et al. (2013) estimated that to be able to detect small annual declines (2-5%) in biodiversity, programs should encompass 200-250 sampling sites visited twice a year. At the identification pace of less than 2 minutes per sample, they estimated the cost of such a program to $2,000,000 over 5 years. In an era where cost-effectiveness has become central in biodiversity conservation policies, monitoring programs could benefit from high-throughput identification pipelines. With the recent advance of NGS, methods have been developed that enable an accurate description of the biodiversity of certain groups using genetic tools. An important advantage of NGS based methods is the potential for automation of the procedure, which can be rapidly carried out by laboratory technicians without specialist taxonomic knowledge. As with all new approaches, these new methods are currently being tested before they can be applied for routine biodiversity monitoring.

The high monitoring costs –majorly linked to morphological identifications– are an important reason why insects have largely been neglected in conservation studies and policies worldwide. In fact, funding for biodiversity projects is largely biased towards only few taxa, with for instance mammals receiving 1000 times more resources than insects (Cardoso, Erwin, Borges, & New, 2011). This lack of funding is also partially responsible for important scientific knowledge gaps which in turn are obstructing public and politic will (Simaika & Samways, 2018). In fact, among the seven impediments to invertebrate conservation cited by Cardoso et al. (2011), four where associated with knowledge gaps, namely: (1) taxonomical incompleteness; (2) unknown distribution of described species; (3) unknown abundance and changes in space and time; (4) unknown ecology and sensitivity to habitat change.

While some of these gaps can be bridged with increased monitoring, taxonomical incompleteness – the failure to recognize and detect all species present - requires more attention. In part, taxonomical incompleteness can either arise from incomplete surveys, which only uncover part of the total biodiversity. However, it can also arise form cryptic diversity. This last case if often considered as the worst-case scenario of taxonomic

incompleteness (Delić, Trontelj, Rendoš, & Fišer, 2017). Typically, cryptic species are observed but misidentified as other, more common species; thus, they remain imperceptible, their conservation cannot be assessed and they cannot be efficiently protected. Assessing geographical ranges and population sizes at the species-complex level instead of at species level may indeed severely bias conservation status (Delić et al., 2017; Funk, Caminer, & Ron, 2012; Niemiller et al., 2013; Theodoridis, Nogués-Bravo, & Conti, 2019). Moreover, cryptic species are most often ecologically differentiated (Fišer, Robinson, & Malard, 2018) which implied that conservation measures need to be specifically adapted (Ashrafi, Beck, Rutishauser, Arlettaz, & Bontadina, 2011; Bickford et al., 2007; Brown et al., 2007). In sum, for conservation policy the unbiased assessment of cryptic diversity is essential (Trontelj & Fier, 2009).

Overall, neglecting insects in biodiversity conservation policies has undoubtedly accelerated current losses in this ecologically vital group. A recent study found that more than 40% of insect species are declining and one third is endangered. Estimates of the extinction rates in insects are eight times higher than those of mammals, birds and reptiles (Figure 1; Sánchez-Bayo & Wyckhuys, 2019). These estimates in diversity losses are thought to be even more aggravated by the taxonomical incompleteness in insects. At current rates, it is predicted that many insect species might go extinct even before they are discovered (Lees & Pimm, 2015; Pimm et al., 2014; Pimm, Jenkins, Joppa, Roberts, & Russell, 2010; but also see Costello, May, & Stork, 2013).

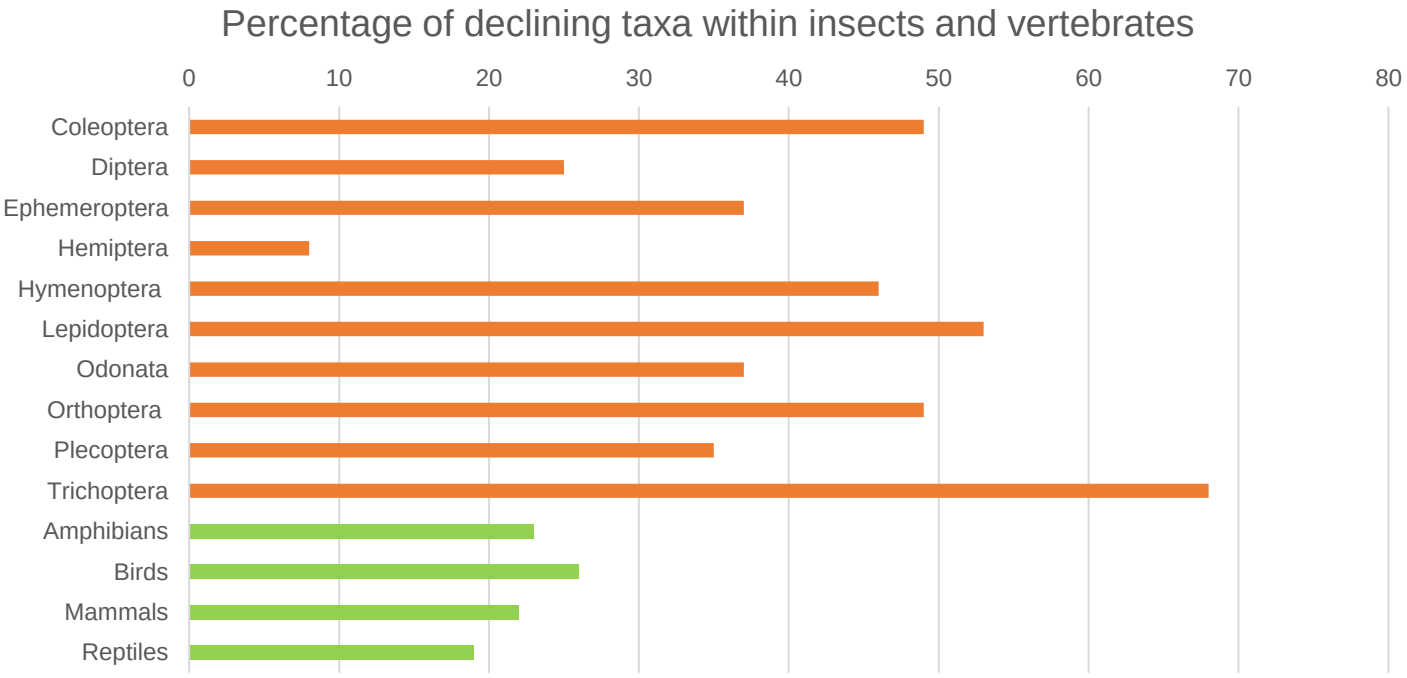


Figure 1 Proportion of declining species per taxon according to IUCN criteria. Data source: Sánchez-Bayo & Wyckhuys, 2019.

Particular case of a vital group of pollinators: wild bees

Pollinators are responsible for the fertilisation of more than 80% of flowering plants worldwide (Ollerton, Winfree, & Tarrant, 2011) and hence play a key role in the majority of terrestrial ecosystems. Often referred to as an ecosystem service, pollination is not only vital to natural habitats but also provides great direct and

indirect benefits to human societies (Fisher, Turner, & Morling, 2009). In 2005 it was estimated that 70% of global crops directly used for human consumption depended on animal pollination (Klein et al., 2007), a service amounting to €153 billion (Gallai, Salles, Settele, & Vaissière, 2009). Based on the trends of the last decades, the value of this global pollination is expected to further increase (Lautenbach, Seppelt, Liebscher, & Dormann, 2012).

While the contribution of wild pollinators to agriculture has long been underestimated, recent studies have demonstrated their direct and indirect importance in the agricultural landscape (Winfree, Williams, Gaines, Ascher, & Kremen, 2008). In many cases, wild pollinators are equal or even better pollinators than honeybees (Garibaldi et al., 2014) and essential to economically important crops such as coffee (Klein, Steffan-Dewenter, & Tscharntke, 2003), tomatoes (Greenleaf & Kremen, 2006a), sunflowers (Greenleaf & Kremen, 2006b), watermelons (Kremen, Williams, & Thorp, 2002; Winfree et al., 2008), blueberries (Cane, 1997) and canola (Morandin & Winston, 2005). They also provide a buffer against pollination shortage and indirectly complement the service provided by honeybees by enhancing in some cases the fruit set of plants visited by honeybees (Greenleaf and Kremen, 2006b).

During the last decades, there has been a worldwide collapse in honeybees with a loss of 25% of colonies in central Europe between 1985 and 2005 (Potts et al., 2010) and an even more pronounced decline (59%) in North America during the same period (van Engelsdorp, Hayes, Underwood, & Pettis, 2008). While this alarming loss is extensively monitored in managed honeybees, only little is known about the situation in wild bees. Based on sparse surveys of some common wild bee species conducted in Europe, the distribution range of various species has significantly shrunk throughout the continent, leading in some cases to local and global extinctions (Biesmeijer et al., 2006; Lebuhn et al., 2012). However, due to the lack of historical records, trends for the vast majority of European wild bee species (i.e., 79%) are unknown (Nieto et al., 2014). Among the remaining species (i.e., 21%), over one third is thought to be in decline.

As for honeybees, the current wild bee decline is associated to multiple factors (Goulson, Nicholls, Botías, & Rotheray, 2015). Some species or region-specific stressors were reported, but in general, habitat loss and fragmentation appear to be the main drivers of decline (Brown & Paxton, 2009; Potts et al., 2010; Zurbuchen & Müller, 2012). Like managed honeybees, wild bees may also suffer from pathogens and parasites (Graystock, Yates, Darvill, Goulson, & Hughes, 2013; Ravoet et al., 2014) but most importantly, bees have to contend with the interaction of multiple stressors. Some factors, when applied individually, may not have great direct impact on bee fitness but when combined with other stressors, such subliminal factors increase their overall stress level and, in addition, may enhance the effect of other factors (Sih, Bell, & Kerby, 2004).

Regardless of the numerous conservation activities performed so far, Switzerland is not spared from this decline in pollinators. With more than 50% of the Swiss wild bee fauna stated as “probably threatened” or “threatened” on the IUCN red list (Cordillot & Kraus, 2011), having effective conservation policies is urgently required to preserve these keystone species. Given the importance of bees and the lack of monitoring, the **overarching aim of my PhD thesis was to evaluate NGS methods as an important new tool for enhancing conservation efforts of wild bees**. We mainly focused on the potential of NGS tools for bridging two of the major impediments to wild bee conservation efforts, i.e., the cost of biodiversity assessment schemes, and taxonomic incompleteness.

First, we explored the potential of NGS identification methods as cost-effective alternatives to morphological identification. Despite the recent development of this research field, there have already been tremendous efforts to establish high throughout identification pipelines for biomonitoring (e.g. Liu et al., 2013; Papadopoulou, Taberlet, & Zinger, 2015; Taberlet & Coissac, 2012). To date, a variety of tools have been developed (Hajibabaei, Baird, Fahner, Beiko, & Golding, 2016), and even though most tools show great potential to fulfil the necessary criteria for such a pipeline, all current approaches are still facing important technical problems. The aim of the **first chapter** was to provide a comprehensive presentation of the advantages and weaknesses of the three most promising methods, namely metabarcoding, mitogenomics and NGS barcoding in biodiversity monitoring. Using a regular monitoring dataset, these methods were compared in terms of species detection and abundance estimates, but also in terms of cost and workload using morphology identification as reference point.

In the second and third chapters, we addressed the problem of taxonomic incompleteness. The development of molecular tools, and especially DNA barcoding, was proposed to provide a technical solution to the taxonomical incompleteness and shortfalls of morphology for several groups, such as insects. However, there is cumulative evidence showing that using the uniparentally inherited organelles (COI) as markers for animal species delimitation can be problematic. Because there are currently no adequate genomic methods for species delimitation, we examined the use of ultraconserved nuclear genetic elements (UCEs). UCEs are highly conserved genomic regions found throughout the tree of life and were therefore used to resolve deep phylogenetic divergences. However, the flanking regions of the UCEs harbour substantial sequence variation and thus UCEs have been suggested to also be a suitable marker for resolving recent divergences (Zarza et al., 2018). As a proof of concept, in the **second chapter**, we used UCEs to delimitate species in a long-debated species complex, the *Andrena-bicolor* group. This particular group was chosen because it harboured deeply divergent sympatric mitochondrial lineages within two species (i.e. *Andrena bicolor* and *A. amieti*). Using UCEs, we investigated the reproductive isolations of these mitochondrial lineages using species delimitation tests and population genetic analyses.

In the **third chapter**, UCEs were further tested to unravel five species complexes that have defeated taxonomists so far in spite of intensive morphological examination and several DNA barcoding studies. These complexes span three of the five bee families present in Europe, allowing us to test for the universality of this technique in bees.

- Chapter I -

Evaluating NGS methods for routine monitoring of wild bees: metabarcoding, mitogenomics or NGS barcoding

Morgan Gueuning, Dominik Ganser, Simon Blaser, Matthias Albrecht, Eva Knop
Christophe Praz & Juerg E. Frey

(Published in Molecular Ecology Resources, 2019)
DOI: 10.1111/1755-0998.13013

Abstract

Implementing cost-effective monitoring programs for wild bees remains challenging due to the high costs of sampling and specimen identification. To reduce costs, next generation sequencing (NGS)-based methods have lately been suggested as alternatives to morphology-based identifications. To provide a comprehensive presentation of the advantages and weaknesses of different NGS-based identification methods, we assessed three of the most promising ones, namely metabarcoding, mitogenomics and NGS barcoding. Using a regular monitoring dataset (723 specimens identified using morphology), we found that NGS barcoding performed best for both species' presence/absence and abundance data, producing only few false positives (3.4%) and no false negatives. In contrast, the proportion of false positives and false negatives was higher using metabarcoding and mitogenomics. Although strong correlations were found between biomass and read numbers, abundance estimates significantly skewed the communities' composition in these two techniques. NGS barcoding recovered the same ecological patterns as morphology. Ecological conclusions based on metabarcoding and mitogenomics were similar to those based on morphology when using presence/absence data, but different when using abundance data. In terms of workload and cost, we show that metabarcoding and NGS barcoding can compete with morphology, but not mitogenomics which was consistently more expensive. Based on these results, we advocate that NGS barcoding is currently the seemliest NGS method for monitoring of wild bees. Furthermore, this method has the advantage of potentially linking DNA sequences with preserved voucher specimens, which enable morphological re-examination and will thus produce verifiable records which can be fed into faunistic databases.

Keywords: survey, pollinators, insects, molecular identification, DNA barcoding, conservation biology

1. Introduction

During the last decades, insect pollinators, and especially bees, have declined in several regions of the world (Bartomeus, Stavert, Ward, & Aguado, 2019; Biesmeijer et al., 2006; Burkle, Marlin, & Knight, 2013; Imperatriz-Fonseca et al., 2016; Ollerton, Erenler, Edwards, & Crockett, 2014). While these losses are extensively monitored in managed honeybees (Potts et al., 2010; vanEngelsdorp & Meixner, 2010), less is known on the status, trends and stressors of wild bee populations, as they are more difficult to survey (Goulson, Nicholls, Botías, & Rotheray, 2015; Potts, Biesmeijer, Bommarco, Kleijn, & Scheper, 2015). Due to the lack of adequate cost-effective monitoring programs, trends for the vast majority of European bee species are unknown (Goulson et al., 2015; Imperatriz-Fonseca et al., 2016; Nieto et al., 2015; Potts et al., 2010). Therefore, there is an urgent need for developing and testing comprehensive, robust and systematic monitoring programs that deliver the information needed for policy makers to decide on the most appropriate conservation measures.

To date, most monitoring programs have relied on morphological identifications, which require a sound knowledge of taxonomy and careful analysis of each individual specimen, making it a lengthy and expensive procedure (Lebuhn et al., 2013). The recent advances of "Next Generation Sequencing" (NGS) techniques offer new opportunities for the assessment of biodiversity (e.g. Schnell et al., 2012; Taberlet, Bonin, Zinger, & Coissac, 2018). Molecular species identifications by DNA barcoding are particularly appealing when classical morphological identifications are not possible [e.g. eDNA, diet assessments; (Rodgers et al., 2017; Taberlet, Coissac, Pompanon, Brochmann, & Willerslev, 2012)], but DNA barcoding has also been suggested for the taxonomical assessment of morphologically identifiable taxa as a mean to reduce costs (Brunner et al., 2002; Hebert, Cywinska, Ball, & DeWaard, 2003).

Although DNA-based monitoring methods have emerged only recently, there have been numerous efforts to establish reliable molecular identification pipelines (e.g. Gibson et al., 2015; Ji et al., 2013). For the successful implementation of NGS-identification tools into monitoring programs, the approach should be reliable, reproducible, cost- and time-effective, easily applicable and, ideally, quantitative to enable assessing species abundance (Joseph, Field, Wilcox, & Possingham, 2006). To date, a variety of tools have been developed, and even though most tools have great potential, each is associated with limitations. Presently, most approaches have been assessed in terms of accuracy (species detection and abundance), but only few have been compared regarding costs and workload (e.g. Elbrecht, Vamos, Meissner, Aroviita, & Leese, 2017). Furthermore, substantial variation in terms of species detection rates and abundance estimates can be observed among studies applying the same molecular methods (although with slightly different parameters), casting doubt on their reproducibility (e.g. see Liu et al., 2013; Yu et al., 2012 for inter-study variation, or Brandon-Mong et al., 2015 for intra-study variation). There is thus an urgent need for a comprehensive and reliable benchmark study assessing the strengths and weaknesses of different methods in terms of species detection and abundance estimates, but also in terms of cost and workload. In this study, we assessed and compared three NGS approaches likely to be among the most suitable to be implemented in routine monitoring programs, namely metabarcoding (Taberlet et al., 2012; Yu et al., 2012), mitogenomics (Zhou et al., 2013), and NGS barcoding (Shokralla et al., 2014).

As in conventional barcoding, metabarcoding (MB) relies on the amplification of a taxonomically informative gene fragment (“barcode”). However, the DNA extraction used as template in MB comes from a bulk mixture

of specimens (Ji et al., 2013), rendering quantification of species abundance difficult. With NGS methods, abundance inference is generally based on the assumption that the number of output reads correlate to the initial amount of input DNA, a proxy for biomass. Thus, if the biomass of each species in the bulk mixture were known in advance, it should theoretically be possible to infer the number of specimens per operational taxonomical unit (OTU). Nevertheless, due to the very nature of the amplification steps involved in MB, this method can be subject to heavy bias, making quantifications doubtable (Dowle, Pochon, Banks, Shearer, & Wood, 2016; Elbrecht & Leese, 2015; Elbrecht, Taberlet, Dejean, Valentini, & Usseglio-polatera, 2016; Piñol, Mir, Gomez-Polo, & Agustí, 2015; Tang et al., 2015; Yu et al., 2012).

To cope with the current lack of solid quantitative output from MB techniques, a PCR-free approach has been suggested (Zhou et al., 2013): mitogenomics (MG), also referred to as mitochondrial metagenomics (Crampton-Platt et al., 2015) or mito-metagenomics (Tang et al., 2014), an ultra-deep sequencing approach using mitochondrial DNA as a "super-DNA-barcode" (Tang et al., 2015). Derived from bacterial metagenomics, it has been successfully applied for mitochondrial mining on arthropod communities (Choo, Crampton-Platt, & Vogler, 2017; Crampton-Platt et al., 2015; Gillett et al., 2014; Gomez-Rodriguez, Crampton-Platt, Timmermans, Baselga, & Vogler, 2015; Linard, Crampton-Platt, Gillett, Timmermans, & Vogler, 2015; Linard et al., 2018; Liu et al., 2016; Tang et al., 2015, 2014; Wilson, Brandon-Mong, Gan, & Sing, 2019; Zhou et al., 2013). Using total DNA extraction of bulk mixtures, shotgun sequencing on high-throughput NGS platforms is performed and raw data is bioinformatically assembled either de novo or mapped to reference databases. MG is not subject to an amplification bias, making it more suitable for quantitative inference (Gomez-Rodriguez et al., 2015; Tang et al., 2015; Zhou et al., 2013). However, even though estimates of species abundance are approaching morphology-based results, MG is still facing methodological limitations, mostly due to the low coverage of target sequences (Crampton-Platt, Yu, Zhou, & Vogler, 2016). Although mitochondria are found in vast copy numbers in animals, mitochondrial DNA (mtDNA) only accounts for a small fraction of the total DNA compared to nuclear sequences. Consequently, the vast majority of data (e.g. 99.47%, in Zhou et al., 2013) produced with MG is not informative, making this approach hardly cost-efficient. Furthermore, as initially presented, MG relies on databases containing full mitogenomes for all investigated species. Because only few full mitogenomes are currently available, this approach is not realistic at this point. To overcome this problem, Gomez-Rodriguez et al., (2015) compared results obtained using full mitogenomic databases with those obtained using only cytochrome oxidase I (COI) reference databases, and found only a slight decrease in species detection and abundance rates in the latter.

In the third method investigated here, NGS barcoding (NGSB), each specimen is processed separately from extraction to sequencing, unlike in MB and MG (Shokralla et al., 2014). Similar to MB, this method relies on the amplification of a genetic marker, but instead of amplifying from total bulk extracts, PCR amplifications are done individually. Because each specimen is uniquely tagged, this approach is quantitative by design and therefore independent of species biomass information. An additional advantage of this method is that each specimen can be preserved for subsequent identification verification or simply to be archived in natural history collections (Wang, Srivathsan, Foo, Yamane, & Meier, 2018). However, processing all specimens individually increases cost and workload related to the library preparation, which constitutes the main limitation of this approach.

To assess the suitability of these three methods for monitoring purposes, we used a dataset collected under regular monitoring conditions. The data was sampled to measure the effectiveness of three different types of

flower strips (FS) in promoting wild and managed bees, and the crop pollination services they provide, in Swiss agricultural landscapes. To answer this question, we compared bee species richness and abundance (relative and absolute) found across the three different types of FS. Additionally, we evaluated the influence of plant species richness on wild bee abundance and diversity.

This realistic monitoring dataset allowed us to assess the performance of each NGS method with respect to variation levels found among sampling sites under realistic conditions. The number of species and specimens characterizing a dataset have a large influence on the overall precision, cost and workload associated with the different NGS methods, which is why estimations of those metrics only make sense with a realistic dataset. Finally, using a realistic dataset allowed us to determine if the accuracy level (presence/absence, relative and absolute abundance) of the explored methods would allow us to detect ecological patterns and reach similar conclusions, and thus validate their use in monitoring programs.

Overall, in this study we compared 1) species detection rates (presence/absence data only), 2) relative and absolute species abundances, 3) ecological patterns and finally 4) costs and workload of the three different NGS-identification methods outlined above compared to morphological identification.

2. Material and Methods

2.1. Sampling

The dataset (sampling material) used in this study was collected in 2017 in agricultural landscapes of the central Swiss Midland. The sampling scheme was designed to identify the effectiveness of three types of sown FS for providing foraging resources to pollinators. In total, 20 different FS were sampled three times over the flower season (two FS were collected four times and one FS two times). FS were either sown in April 2013 (FS type 1, n = 8), April 2016 (FS type 2, n = 8), and September 2016 (FS type 3, n = 8). All three types of FS harboured unique floral mixtures, composed of species of annual (all three types) and perennial flowering plants (types 1 and 2), which were primarily selected due to their high pollen and nectar production.

To be able to obtain quantitative information on the number of pollinators present at each sampling round, a strict sweep-netting protocol was applied. During each sampling round, transects were slowly walked up while sweeping two times 25 sweeps with one minute pause in between. After 50 sweeps, the collected material was transferred into a plastic bag and directly stored at -20 °C in a portable freezer. Furthermore, during each sampling round, we monitored plant species richness, allowing us to additionally assess the importance of this parameter in promoting bees.

To determine the degree of variation within each FS, the exact same protocol was repeated within the same FS after five minutes (hereafter referred to as “transect I” and “transect II”). Transect II started from the end point of transect I. In total, the dataset encompasses 122 sampling points [hereafter referred to as “communities”: (17 FS × 3 sampling rounds × 2 transects) + (2 FS × 4 sampling rounds × 2 transects) + (1 FS × 2 sampling rounds × 2 transects)].

2.2. Identification methods

2.2.1. Morphological identification

In the laboratory, raw sampling material was sorted to isolate wild bees from plant material, other insects, as well as honeybee workers. Each specimen (n = 723) was then pinned, labelled, dried for at least 72h in a desiccator containing silica gel, and identified by an expert. Most specimens were identified to species-level, but in the following cases morphological identifications were performed to species-group level: *Bombus terrestris*-group for workers belonging to *B. terrestris*, *B. lucorum* and *B. cryptarum*; *Halictus simplex*-group for females of *H. simplex*, *H. langobardicus* and *H. eurygnathus*; and *Andrena ovatula*-group for females of *A. ovatula* and *A. wilkella*. Morphological identification was complemented by Sanger sequencing using COI barcoding for all specimens identified to species-group level and not to species level (n = 29) or left undetermined because of lack of intact morphological criteria (n = 11). For clarity, we still refer to this dataset as "morphological" even if for a limited number of specimens morphological identifications have been complemented using Sanger sequencing. Details of the Sanger sequencing protocol are given in Supplementary Information S1.

2.2.2. Metabarcoding

Bulk DNA extractions were performed on each community using a proteinase K solution and digested overnight at 56 °C. Volumes of proteinase K solutions were adapted according to the number of specimens per community so that all specimens were immersed into the solution. To reduce costs linked to commercial kits, we purified the extracts following the Canadian Center for DNA Barcoding (CCDB) DNA extraction protocol (Ivanova, Dewaard, & Hebert, 2006). For each community, to increase species detection rates and normalize template abundance, DNA purifications were performed in triplicates and immediately pooled after extraction. To reduce workflow and limit numbers of PCR reactions required during the library preparation, amplification was carried out using fusion primers. In addition to the priming sequence, fusion primers have overhangs composed of Illumina indexes and a unique tag of 8 base pairs (bp) designed using the software Barcode generator (Meyer & Kircher, 2010). The overhangs allow amplicons to be directly loaded onto the Illumina sequencer. To overcome the inherent limitation of Illumina platforms in sequencing low complexity libraries, we added a "heterogeneity spacer" between the labelled tag and the priming sequence, as recommended in Fadrosh et al. (2014). The PCR primer sequences of the fusion primers were those of mlCOIintF and of HCO2198 (Leray et al., 2013) and targeted a 313 bp region of the COI gene. Overall, forward and reverse primers were 95 bp long (± 3 bp). Per community, bulk amplification was performed in five different PCR-replicates, each harbouring a unique combination of forward and reverse tags. Further details on MB library preparation are given in Supplementary Information S2. Final library was sequenced on an Illumina Miseq using a v3 kit (2 x 300 bp) and spiked with 20% Phix.

The majority of bioinformatics analyses (detailed in Supplementary Information S3) were performed using QIIME1 (Caporaso et al., 2010). Briefly, raw data were trimmed based upon the FASTQC profile before joining paired-end reads. After demultiplexing, adaptors, spacers and primer sequences were trimmed. Chimeric sequences were identified de novo and removed using usearch61 (Edgar, 2010). Filtered sequences were then clustered using the uclust algorithm (Edgar, 2010) at the default similarity threshold of 97%. Taxonomical assignment of OTUs was performed using the same algorithm by fitting reads to reference sequences. To determine the impact of database quality on the species detection performance, OTU's were assigned using two separate COI databases. The first database ("uncurated") encompassed all available COI sequences of bee species (barcodes for ca. 2000 species) available on BOLD (Barcode of Life Database) and

Genbank (downloaded in June 2017). Additional verifications were made to ensure the presence of multiple barcodes ($n \geq 3$) for all species present in our dataset. The second database ("curated") was downloaded from BOLD and corresponds to sequences deposited by Schmidt and colleagues (2015) in their extensive barcoding study on western-Europeans bees (dx.doi.org/10.5883/DS-GBAPI). This dataset was initially missing barcodes of two species present in our dataset (i.e. *Andrena flavipes* and *Chelostoma florisomne*) and barcodes for these two species were downloaded from other projects on BOLD and manually added to the database. Similarly, to determine the best similarity threshold, the MB bioinformatic pipeline was ran several times using different similarities thresholds [from 90% (default) to 99%]. Corresponding community matrices were compared to the morphological community matrix and the threshold performing best was retained for downstream analyses. The same empirical approach was applied to determine the optimal cross-validation setting among replicates (i.e. minimal occurrence of a species among replicates to be validated).

2.2.3. Mitogenomics

Aliquots of the DNA extracts used for MB (prior to library preparation) were sheared using an ultrasonicator (Bioruptor). The mitogenomics (MG) library was built using a commercial Illumina 96 TruSeq DNA Nano kit following the manufactures recommendations. To reduce differences in sequencing depth, we homogenized sequencing depth on the number of specimens per community by applying the same correction factor as for MB (Supplementary Information S2). The library was sequenced on an Illumina Miseq using a v3 kit (2 x 300 bp) and spiked with 1% Phix.

Two different bioinformatics approaches were compared [i.e. (i) de novo assembly and (ii) raw read mapping] and the approach recovering the highest number of species was retained for downstream analyses. (i) The de novo assembly approach mainly followed Crampton-Plat et al. (2015). Details are given in Supplementary Information S3; briefly, libraries were quality assessed using FASTQC and residual adaptors trimmed with Trimmomatic (Bolger, Lohse, & Usadel, 2014). Then, libraries were filtered to retaining only mitochondrial reads using blastn (Camacho et al., 2009) and a database containing all publically available (partial and full) mitogenomes of bee species (336 mitogenomes of 82 species; among which 18 present in our dataset). Putative mtDNA reads were then assembled using IDBA-UD (Peng, Leung, Yiu, & Chin, 2012) with a 98% similarity threshold. Contigs were mapped at a 98% similarity against a custom database using BBMap (Bushnell, 2015). Since only 18 reference mitogenomes were available for the investigated species, additional COI barcodes from the curated COI database (see above) were added to the mitogenome database. Finally, SAMtools (Li et al., 2009) was used to index and extract the number of reads that mapped reference sequences. (ii) The raw read mapping approach relied on BBMap (Bushnell, 2015) to map unfiltered reads against COI reference sequences. Because only a small fraction of sequences will match to the COI reference database, it is crucial for this approach that the database is not only comprehensive, but also well curated. The presence of uncurated sequences (e.g. numts) will have a major influence upon the outcome, much more than for amplicon-based approach where coverage-based filtering will in most cases obliterate errors originated form the database. Therefore, only the curated database was used in this approach. To further reduce false positives due mapping of reads in the flaking regions of COI, sequencing spanning over the classical 658 bp COI barcoding region were filtered out of the curated database. As in Tang et al. (2015), a high similarity threshold (99%) was used to reduce false positives and reads were mapped once. Mapped reads were indexed and extracted using SAMtools (Li et al., 2009).

2.2.4. NGS barcoding

Before performing bulk DNA extractions described above, a single leg of each specimen was taken for DNA extraction (one extraction per specimen) following the CCDB protocol. As for MB, fusion primers were used to amplify individually all extractions and PCRs were conducted following the same conditions as for MB. After amplification, each PCR product was examined on a 1.5% agarose gel and amplicons were pooled equimolarly as estimated based on their amplification intensity. Pooled PCR products were purified with NucleoFast 96 PCR clean-up kits (Macherey-Nagel) using 300 µl of PCR product per well and eluted in 100 µl ddH2O. Cleaned PCR products were sequenced on an Illumina Miseq using a v3 kit (300 bp x 2) spiked with 20% Phix.

Data processing of the NGSB library is similar to the MB procedure. The filtered reads were clustered using uclust at a similarity threshold of 99% and OTU's were taxonomically assigned using the same algorithm but with a default threshold parameter (90%). A lower taxonomical assignment threshold than for MB was used to decrease the number of unassigned OTUs since only the most abundant species assignment per specimen was retained in the final matrix. The number of false positives was therefore not affected by this lower threshold. As for MB, taxonomical assignments of OTU's was performed using the two different databases (curated and uncurated).

2.3. Data analyses

2.3.1. Species richness

For all NGS methods, we compared species richness with morphological species richness for each community and assessed species detection rates using the Jaccard similarity index (Jaccard, 1912). To determine variation between two transects collected five minutes apart within the same FS, we also computed the Jaccard index between the samples identified based on morphology.

2.3.2. Quantitative inference

In this study, species quantification (relative and absolute abundance) for both bulk methods (i.e. MB and MG) was defined as a measure of the species biomass, and not numbers of specimens per species. To assess quantification accuracy for MB and MG, we correlated the number of reads per species (ln-transformed) with the corresponding species biomass measurements. For solitary bees, dry weight can be accurately estimated by the following exponential relationship (Cane, 1987): $y = 0.77(x)0.405$, where y is the shortest linear distance between the wing plates (intertegular distance; mm) and x is the dry weight (mg). A photograph was taken of each specimen using a stereomicroscope-mounted camera (Leica M4000), and intertegular distance was measured, which enabled to measure biomass for each specimen. To compare quantitative data on the number of specimens per species among all methods, we transformed the morphological absolute abundance (number of specimens per species) into relative abundance of biomass.

2.3.3. Comparison of ecological patterns

To determine if the detected ecological patterns would be similar across our three NGS approaches as well as the classic morphological approach, we applied the same statistical analyses on presence/absence data and on relative and absolute abundance data. First, to explore how much of the observed variance in species composition across sampling sites was explained by the identification method, we performed a non-parametric multivariate analysis of variance using distance matrices [i.e., PERMANOVA; (Anderson, 2001)]. The same test was also performed on the morphological dataset to determine the biological variance found between the

two transects sampled five minutes apart. Theses PERMANOVA tests ("adonis" function in the R cran vegan package) were performed using the Jaccard dissimilarity index for presence/absence data and the Bray-Curtis distance dissimilarity index for both relative and absolute abundance data. All adonis analyses were run with 10,000 permutations. Second, to complement the adonis analyses, we performed non-metric multidimensional scaling (NMDS) to visualize and compare community compositions of FS among the identification methods. The goodness-of-fit between the superimposed shapes of NMDS plots was assessed by Procrustes tests computed with the "protest" function (vegan package). The NMDS analyses were performed with the "metaMDS" function implemented in the vegan package with the "noshare" function activated to use extended dissimilarities when sampling sites did not share species. "Spider" diagrams were added to connect communities sharing the same FS type. Third, to determine and compare the effectiveness of the three different types of FS in promoting wild bees, we ran linear mixed models (LMM) and generalized linear mixed models (GLMM) using the "lme4" package (Bates, Mächler, Bolker, & Walker, 2015). Species richness and species abundance (relative and absolute) were used as response variables (see details of models in Supplementary Information S12). Finally, to determine the importance of flower richness on promoting wild bees, we applied similar models with the predictor variable being the interaction between plant species richness and identification method. The relationship between plant species richness and bee richness or abundance were plotted using linear regressions with 95% confidence intervals.

2.4. Cost and Workload

Costs estimates are based upon suppliers' prices applied in 2018 in Switzerland and do not contain cost linked to workload. To compensate for the cost of wet lab consumables, overall costs were increased by 15%. For the morphological identifications, the workload includes mounting, labelling and data basing of the specimens and the cost corresponds to the identifications performed by the taxonomist. Regarding the workload estimate for NGS methods, only hands-on laboratory processes were recorded, leaving out time needed for over-night digestions, PCR amplifications, electrophoresis or other incubation times.

To predict the relationship between overall cost and total number of specimens, we divided the price per specimen into fixed (i.e. independent from the number of specimens) and variable costs (dependent on the number of specimens). For the three NGS methods, we thus subtracted the cost of the sequencing kit (variable cost) to the grand total and divided the result by the number of specimens (fixed cost). Cost estimates for morphological identifications only included fixed costs.

Finally, since Illumina platforms offers the possibility to run different kits harboring variable outputs, we estimated the overall cost and sequencing depth for all kits allowing to span our targeted read length (~450 bp; including tags and technical sequences) for MB and NGSB, namely the Miseq v3 (2 x 300 bp), Miseq v2 (2 x 250 bp) and the Miseq v2 Nano (2 x 250 bp) kit; and for MG, the Miseq v3 (2 x 300 bp), Hiseq 4000 (1 x 50 bp) and Hiseq 4000 (2 x 75) kit.

3. Results

3.1. Morphological identification

Wild bees were found in 83 of the 122 sampling points. After sorting wild bees from the honeybees (n = 1422 honeybees) and other arthropods (mainly aphids, dipterans and coleopterans) we counted 723 wild bee

specimens. A total of 683 specimens were identified morphologically to species level, 29 to species-group level (among which 20 were identified as workers from the Bombus terrestris-group), and 11 remained unidentified. Sanger sequencing, used as complement for the identification to the species level of the species-groups and undetermined specimens, was successful for 39 of 40 specimens. The one unidentified specimen for whom Sanger sequencing failed was classified as "unidentified".

The morphological dataset, complemented with Sanger sequencing, comprised 723 specimens and 58 species, of which 382 specimens belonged to the transects I, and 341 to transects II (Supplementary Information S4). The median number of specimens per community was 5 and the mean (± SD) number 8.71 (± 10.12), with a minimum of 1 and a maximum of 55 specimens.

3.2. Sequencing outputs

The Miseq runs produced 13.8, 17.5 and 9.0 million reads, respectively, for the MB, MG and NGSB libraries (Supplementary Information S5). After read merging, demultiplexing and data filtering, the MB and NGSB datasets encompassed respectively 4.5 and 3.4 million reads. Raw reads from the MG library were not filtered but directly mapped to the COI reference database. In total, 28.26%, 0.02% and 32.22% of reads mapped to the database, for MB, MG and NGSB respectively. To estimate the average coverage per specimen and community, the number of mapped reads was divided by either the number of specimens (n = 723) or the number of communities (n = 83). On average, the number of reads per specimen was 5450, 4 and 3959 for MB, MG and NGSB respectively, and 47'471, 38, 34'485 per community, respectively.

3.3. Impact of the quality of the COI reference databases in MB and NGSB

For both MB and NGSB, species detection rates were higher while using the uncurated COI reference database (Supplementary Information S6). The use of this database uncovered more true positives and decreased the number of false negatives. For NGSB, using the uncurated database however introduced one supplementary false positive. Based on these results, the uncurated database was used for all subsequent analyses.

3.4. Metabarcoding parameters

Similarity thresholds for the taxonomical assignment of OTUs considerably influenced the overall number of false positives and negatives (Supplementary Information S7.A). The similarity threshold providing the highest species detection rates (Jaccard similarity index) were 97% and 98%. Since species detection rates were similar for 98% and 97%, the more widely accepted threshold of 97% was favored and used in all subsequent analyses. At this threshold, the mean percentage of unassigned OTU's per community was 18.1% (Supplementary Information S8).

Cross-validation thresholds had a lesser effect and produced similar number of false positives and false negatives when validating species present in at least 1, 2, 3, 4 or 5 of out of 5 replicates (Supplementary Information S7.B). The less stringent thresholds (i.e. 1/5 and 2/5) introduced one additional false positive while the correlation between biomass and read numbers was slightly higher than for the more stringent thresholds. Because the higher correlation between biomass and read number did not reduce the overall difference found between the morphological and MB matrices, and because this less stringent threshold slightly increased the false positives rate, the more conservative threshold of 3 out of 5 replicates was favored and used for subsequent analyses.

3.5. Mitogenomics pipelines

The Jaccard index for the de novo assembly pipeline was considerably lower than for the raw mapping pipeline (Supplementary Information S9). The former pipeline uncovered 17 true positives whereas the latter 53 true positives. Based on these results, the raw mapping pipeline was favored for downstream analyses.

3.6. Species richness

The Jaccard similarity index between morphological and NGS datasets was highest for NGSB, followed by MB and MG (Table 1). For NGSB, all species present in the morphological dataset were recovered and only two additional species (false positives) were identified (Supplementary Information S4). The number of false negatives was similar for MB (n = 5) and MG (n = 5), although MG harboured substantially more false positives (n = 16) than MB (n = 4). There was no clear overlap in species identity between the false positives and negatives found in those two methods (Supplementary Information S4). The lowest Jaccard similarity index was found among transects of the morphological identification method (Jaccard index = 0.508). Jaccard indexes between transects of each NGS method were consistently close (Supplementary Information S10).

3.7. Quantitative inference

Individual species biomass (as computed based on morphological identifications and measured intertegular distances) was significantly correlated to the sequencing output for both MB and MG for relative and absolute abundance (Figure 1, Supplementary Information S11). For both NGS methods correlations were higher when using relative abundance than absolute abundance. MG displayed higher correlation coefficients than MB, especially for relative abundance (Figure 1).

Datasets	Transects	Species Richness	# Shared Species	False Positives	False Negatives	Jaccard Index
Between transects of Morpho	I	43	30 (30/43 = 69.8%)	-	-	0.508
	II	46	30 (30/46 = 65.2%)	-	-	0.508
	I & II	58	-	-	-	-
Between MB and Morpho	I & II	57	53 (53/57 = 93.0%)	4 (4/57 = 7.0%)	5 (5/57 = 8.8%)	0.855
Between MG and Morpho	I & II	69	53 (76.8%)	16 (23.2%)	5 (7.5%)	0.716
Between NGSB and Morpho	I & II	60	58 (96.7%)	2 (3.4%)	0 (0%)	0.967

Table 1: Jaccard similarity index between the global diversity of morphological (Morpho) and molecular (MB, MG and NGSB) datasets. Similarity indexes per transect for the molecular methods are given in Supplementary Information S10.

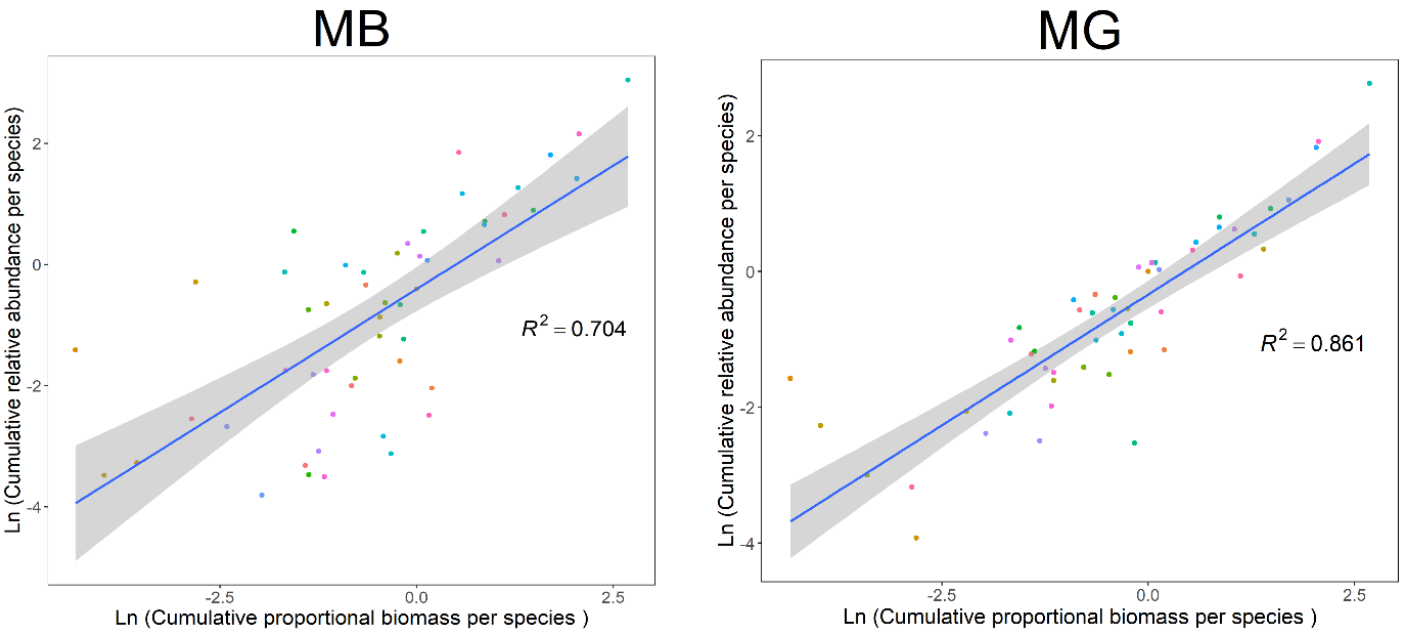


Figure 1: Correlation between the ln transformed relative read number per bee species and the ln-transformed estimate proportional biomass per species for metabarcoding and mitogenomics datasets. Grey areas represent the 95% confidence interval. Proportions were cumulated across all sampling sites. Each coloured dot represents a different species. Correlations were significant with *P-values* < 0.0001.

3.8. Ecological patterns

PERMANOVA tests, performed to analyze and quantify differences in community compositions between NGS and morphological datasets, revealed significant differences in the abundance data for both MB and MG, but not for NGSB (Table 2). With presence/absence data, the differences were significant only for MG datasets. Overall, the identification method explained 0.1%, 9.0% and 10.7% of the variance found compared to the morphological dataset for MPS, MB and MG, respectively.

The NMDS ordinations showed similarities in community composition across the morphological and the NGS methods (Figure 2, Supplementary Information S12). This was especially true for the NGSB datasets for whom the Procrustes tests revealed highly similar community compositions to the morphological one (Table 2). For the MB and MG datasets, Procrustes tests also depicted significant correlations with the morphological dataset in community composition, although with lower correlation coefficients. As in PERMANOVA analyses, the lowest correlation coefficient for MB and MG were found with absolute abundance data.

While testing for difference in bee species richness or abundance among the three different types of FS, the GLMM (presence/absence) and LMM (relative and absolute abundance) analyses depicted no statistical difference among FS types for all identification methods (Figure 4, Supplementary Information S13). Similarly, using the plant species richness as predictor, all identification methods showed comparable relationships between plant species richness and bee species richness (Supplementary Information S14-S15). However, the relationships between plant species richness and bee relative abundance were significantly different from the morphological dataset for MB or MG (Figure 3, Supplementary Information S15). Indeed, MB and MG showed a negative relationship between bee abundance and plant species richness whereas this relationship was positive for the morphological and NGSB datasets.

Furthermore, MG overall underestimated the bee relative abundance, whereas MB overestimated it for plots low in species abundance and underestimated it for species-rich abundant plots (Figure 3).

			Presence/absence				Relative Abundance				Absolute Abundance			
Method	Test	Levels	Df	F model	R^2	P value	Df	F model	R^2	P value	Df	F model	R^2	P value
Morpho	PERMANOVA	Transect	1	0.729	0.009	0.796	1	0.617	0.008	0.850	1	0.617	0.008	0.852
		Residuals	80		0.991		80		0.992		80		0.992	
		Total	81		1.000		81		1.000		81		1.000	
MB	PERMANOVA	Identification	1	0.760	0.005	0.727	1	3.614	0.022	**<0.001**	1	15.809	0.088	**<0.001**
		Residuals	164		0.995		164		0.978		164		0.912	
		Total	165		1.000		165		1.000		165		1.000	
	Procrustes				0.803	**0.001**			0.819	**0.001**			0.783	**0.001**
MG	PERMANOVA	Identification	1	19.044	0.106	**<0.001**	1	10.974	0.064	**<0.001**	1	15.625	0.089	**<0.001**
		Residuals	160		0.894		160		0.936		160		0.911	
		Total	161		1.000		161		1.000		161		1.000	
	Procrustes				0.543	**0.001**			0.651	**0.001**			0.350	**0.001**
NGSB	PERMANOVA	Identification	1	0.207	0.001	1.000	1	0.251	0.001	0.995	1	0.228	0.001	0.998
		Residuals	164		0.999		164		0.999		164		0.999	
		Total	165		1.000		165		1.000		165		1.000	
	Procrustes				0.934	**0.001**			0.854	**0.001**			0.900	**0.001**

Table 2: Non-parametric multivariate analysis of variance on distance matrices (PERMANOVA) using the *adonis* function and Procrustes test ("*protest*" function) between NMDS of molecular (MB, MG and NGSB) and morphological (Morpho) identifications. Jaccard dissimilarity index was used to transform the presence/absence datasets and the Bray-Curtis index for both abundance formats. For the morphological identified dataset, the PERMANOVA test was performed between transects. *P-values* under the 0.05 threshold are in bold.

3.9. Cost and workload

With respect to cost, morphological identification was approximately half the price of the cheapest NGS-identification method (MB) and approximately three times cheaper than the priciest one (MG) (Supplementary Information S16), when cost was estimated based on the number of specimens included in this study. For all investigated NGS methods, the sequencing kits used in this study represented the principal fraction of the overall cost. Since the sequencing kit cost is independent from the number of specimens sequenced (as long as the desired sequencing depth is reached), we calculated costs with increasing number of specimens. Based on this calculation, after approximately 1795 and 4639 specimens, MB and NGSB would become respectively more cost-efficient than morphology-based identifications for the Miseq v3 kits (Supplementary Information S17, see Supplementary Information S18 for cost details). Because of sequencing depth limitations, MG stayed largely costlier than morphological identification. Alternatively, instead of increasing specimen numbers, cost could be reduced by using smaller, less expensive sequencing kits. Based on the mean sequencing depth of MB and NGSB, we estimated the coverage and overall cost for two alternative kits (Miseq v2 [2 x 250 bp] and Miseq v2 Nano [2 x 250 bp]; Supplementary Information S19).

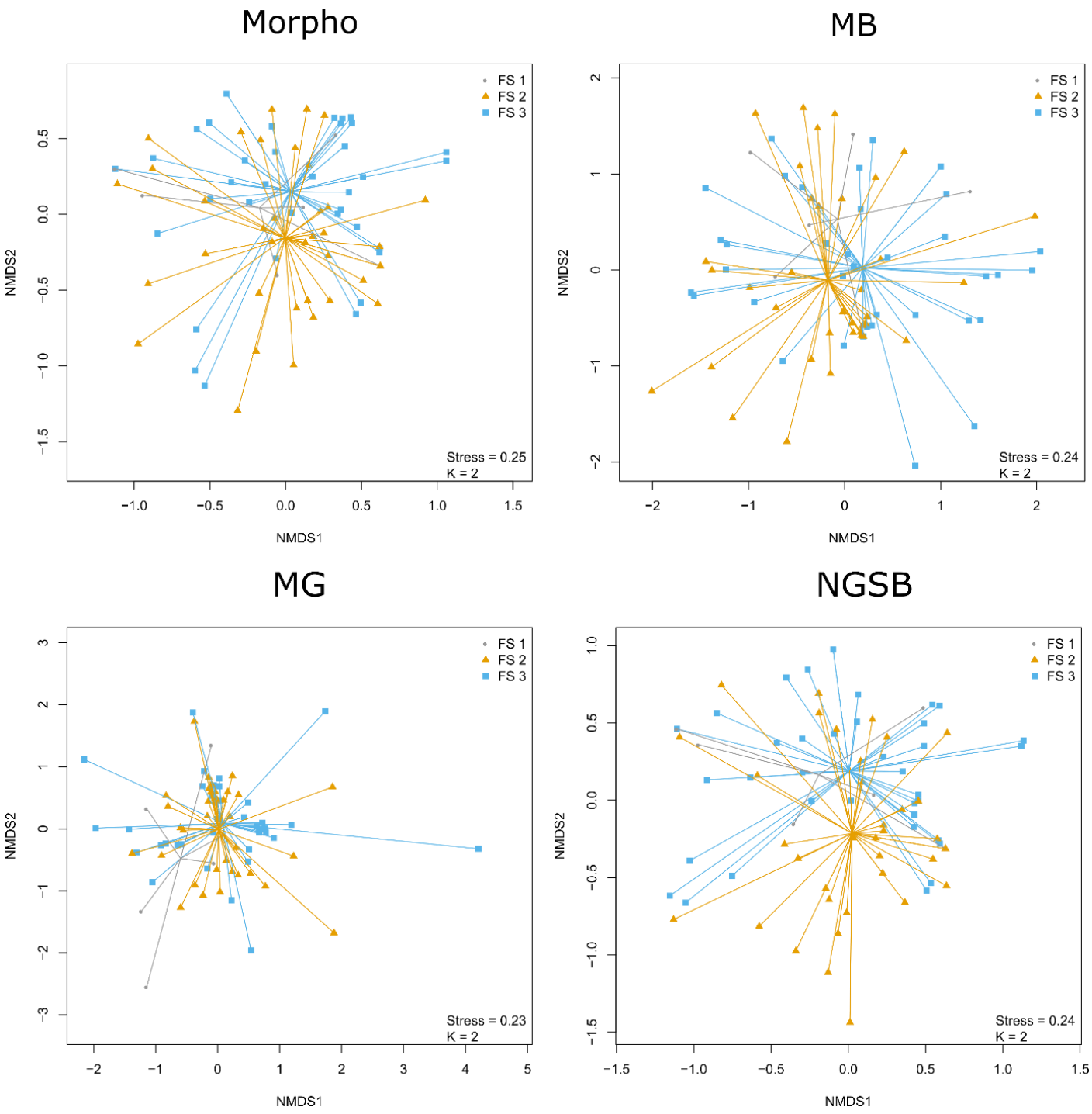


Figure 2: Non-metric multidimensional scaling (NMDS) of bees' relative abundance obtained by four different species identification methods. The NMDS analyses were performed using the Bray-Curtis index with the "*metaMDS*" function implemented in the *vegan* package. "Spider" diagrams connect communities sharing the same flower stripes (FS) type. Goodness-of-fit between the superimposed shapes of the molecular NMDS plots with the corresponding morphological NMDS plots were assessed using Procrustes tests, computed with the "*protest*" function (*vegan* package) (see Table 2). Note the close similarity between datasets based on morphology and NGBS.

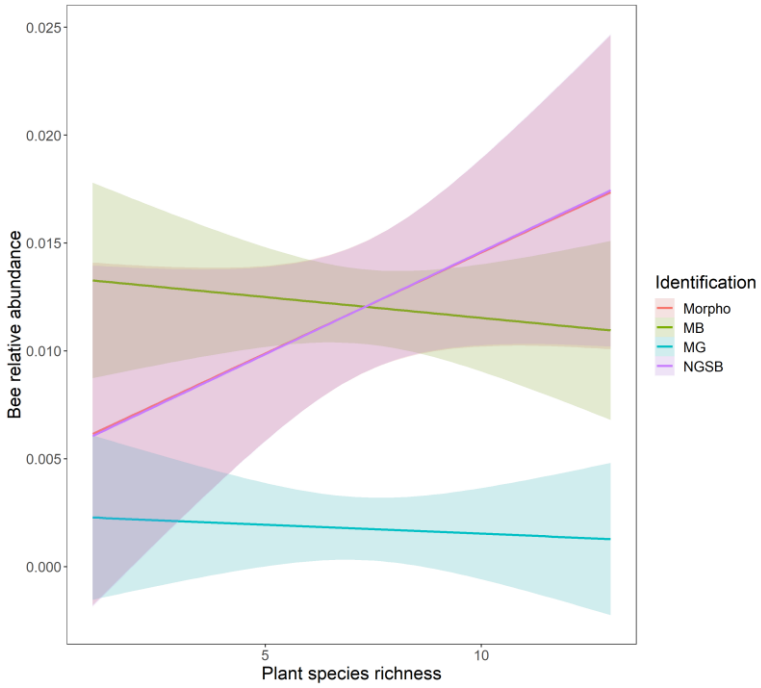


Figure 3: Relationship between plant species richness and the relative abundance of bees for different identification methods. Lines were computed by linear regressions as implemented in ggplot2. Coloured areas represent the 95% confidence interval. Statistical differences in relationships of the molecular identification method compared to the morphological identification method were assessed by linear mixed models. For bee species richness, no difference in relationship were found between the morphology and NGSB (regressions overlap), while MB and MG showed significant deviation compared to the relationship based on morphological identifications (See Supplementary Information S15 for LMM results).

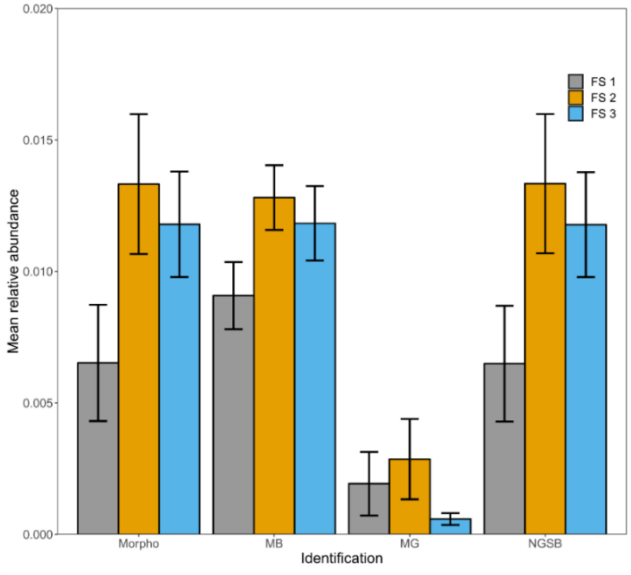


Figure 4: Mean relative abundance of bees for three different types of flowering strips (FS). Means were computed per identification methods and error bars correspond to the mean standard error. Statistical difference among means within each identification method was assessed with linear mixed models. No statistical difference among types of FS was found within method.

Although sequencing depth attained in this study for MB and NGSB were slightly under-optimal (Supplementary Information S4), coverage estimations based on these figures suggest sufficient sequencing depth, even for the smallest sequencing kits (Supplementary Information S19).

Regarding workload, MB was the identification method requiring less workload. Morphological and MG required similar workloads and NGSB moderately more (Supplementary Information S15).

4. Discussion

Overall, our results show that (I) NGS barcoding (NGSB) provided the dataset most similar to the morphological dataset, both in terms of species detection and abundance. (II) As predicted, the correlation between biomass and read numbers was stronger for mitogenomics (MG) than for metabarcoding (MB). Nevertheless, MG produced more false positives (23.2% against 7.0% for MB) and therefore considerably decreased similarities in community compositions compared to the morphological dataset. (III) For both MB and MG, species abundance estimates were better when using relative abundance than absolute abundance. (IV) Ecological patterns were similar across all identification methods when using presence/absence data. However, when using abundance data (both relative and absolute) the conclusions based on MB and MG identification, but not NGSB, differed from those based on morphology; (V) Finally, the overall cost of all three NGS methods were higher than morphological identifications. However, MB and NGSB become more cost-effective by either using smaller sequencing kits (e.g. Miseq v2 Nano kit) or by increasing specimen numbers. Hereafter, we summarize the advantages and weaknesses of each NGS method.

4.1. Metabarcoding

Since Taberlet et al., (2012) proposed MB as a modern tool for assessing biodiversity, MB has been widely accepted when alternative means of species identification are lacking (e.g. eDNA, diet analyses). However, for cases where morphological identification is possible (e.g. pollinators surveys), MB is still in a validation phase. To date, the vast majority of metabarcoding studies have been tested against laboratory-assembled communities of known composition (e.g. Elbrecht & Leese, 2015; Elbrecht et al., 2016; Piñol et al., 2015; Tang et al., 2015; Yu et al., 2012), and the reported detection rates are highly variable.

For instance, Tang et al. (2015) compared the accuracy of MB and MG on a dataset taxonomically similar to ours (33 wild bee species represented by 250 specimens) and found as many as 11 false negatives and 49 false positives, for 53 true positives. Based on these figures, the Jaccard similarity index between morphological and MB identification would be 0.47.

As illustrated in the study by Tang et al. (2015), MB detection rates are frequently obliterated by high numbers of false positives and negatives (Gentile Francesco Ficetola, Taberlet, & Coissac, 2016), a problem that strongly biases the overall interpretation of species detectability (Lahoz-Monfort, Guillera-Arroita, & Tingley, 2016). To overcome this limitation, replicates are crucial (Mata et al., 2019). Although it is possible to estimate the number of required replicates (Ficetola et al., 2015), the optimal replication level largely depends on the dataset. In our study, we empirically tested different settings and observed no major differences among them. Although detection rates may vary across studies, to our knowledge, all rates of species detection were under 100%. Because a perfect match between NGS and morphological identification is illusive, Ji et al. (2013)

investigated the effect of such discrepancies on policy making and management issues. To do so, they compared MB with standard morphology-based datasets and found that both exhibited similar alpha- and beta-diversities, leading to similar policy conclusions. Although insightful and pioneering, this study was conducted on a very large dataset (55,813 arthropods and bird specimens) in which small variations in species presence/absence would be unlikely to have a strong influence. Applying a similar approach to our much smaller dataset resulted in similar conclusions: morphological and MB datasets exhibited similar species composition (Table 2), revealing similar ecological patterns with (1) no differences in bee species richness among the three different types of FS and (2) similar positive relationships between plant species richness and bee species richness (Supplementary Information S14-S15).

Nevertheless, these conclusions are based on presence/absence data while the majority of monitoring programs rely on species abundance data, which gives a more precise picture of community composition (Joseph et al., 2006; MacKenzie, 2005). Therefore, there has been numerous efforts to foster the reliability of MB species count, and currently there is an equal number of studies claiming or disclaiming quantification reliability (see Piñol, Senar, & Symondson, 2018). A study investigating the variability in quantification recorded the level of variance in read numbers associated to individual nematodes between PCR/library replicates and found an overall very consistent read count per individual ($R^2 = 0.99$) (Porazinska, Sung, Giblin-Davis, & Thomas, 2010). However, their results also highlighted consistent variance in read numbers among species, even after correcting for their body size. In a similar attempt to uncover variation sources in read quantification at the inter-species level, Elbrecht & Leese (2015) sequenced libraries build with the exact same biomass of different species and found substantial differences in read abundance among species (up to 4 times higher or lower read abundances). These results underline an inherent problem linked to PCR-based techniques, i.e. the primers' species-specific efficiency. PCR amplification efficiency is primarily (73%) influenced by the number of template-primer mismatches (Piñol et al., 2015) and therefore the selection of primers will greatly influence the quantitative output (Piñol et al., 2019). While testing 15 common universal COI primer pairs, Piñol et al. (2018) found a significant relationship between DNA concentration pre- and post-PCR for the vast majority of primers (14/15) although R^2 values were variable. The primer pair used in our study performed relatively well, even though other primers performed better (e.g. ArF5 & ArR5, Gibson et al., 2014). The problems outlined above likely contribute to the large differences in quantification inference reported in the literature. Furthermore, bulk-based approaches might inform on the biomass, but not necessarily on specimen numbers because of intra-specific biomass variations (e.g. sex or "cast" polymorphism in social bees).

In our study, we found strong correlations between read numbers and estimated biomass, especially when using relative abundance data (up to $R^2 = 0.704$; Figure 1). The beta-diversity of MB and morphological datasets was also highly similar for relative abundance with only 2.2% variance explained by the identification methods alone (Table 2). Furthermore, the Procrustes test depicted a relatively high correlation between the NMDS shapes of the MB and morphological datasets ($R^2 = 0.819$, P-value < 0.001; Figure 2). Although these results are promising, we still found evidence of a bias introduced because of quantitative inference. Indeed, the LMM analysis depicted contrasting relationships between plant species richness and bee relative abundance depending on the identification method (Figure 3, Supplementary Information S15); while the relationship between these variables was positive for the Morphological dataset, it was slightly negative for the MB dataset (Figure 3). These results show that regardless of high correlations between estimated biomass

and inferred abundance in morphology and MB, the overall ecological patterns are skewed by a biased estimate of species abundance, ultimately leading to incorrect ecological conclusions.

4.2. Mitogenomics

As initially suggested by Zhou et al. (2013) and several follow-up studies (Gomez-Rodriguez et al., 2015; Tang et al., 2015), we corroborate that quantitative inference based on biomass is less biased with a PCR-free approach: regardless of the quantitative community format, Pearson's correlations were higher for MG (relative abundance: $R^2 = 0.861$; absolute abundance: $R^2 = 0.623$) than MB (relative abundance: $R^2 = 0.704$; absolute abundance: $R^2 = 0.549$) (Figure 1, Supplementary Information S11). Interestingly, the correlation coefficients found in our study are similar to those found in other MG analyses (Zhou et al., 2013: $R^2 = 0.64$; Gomes-Rodriguez et al., 2015: $R^2 = 0.69$; but see Tang et al., 2015: $R^2 = 0.25$).

While estimates of biomass appear to be more reliable and precise when using MG, the higher number of false positives and negatives (Table 1) skewed the overall species composition and introduced greater variance than with MB (Table 2). Although often claimed as less prone to false positives and negatives than PCR-based methods (Zhou et al., 2013, Tang et al., 2015), we nevertheless found in our study substantially more false positives (23.2%) than with MB (7.0%). We argue that these high rates could mainly be attributed to two factors: the reference database and the low coverage. First, the database used in our study features available sequences of much higher diversity (> 450 species) than present in our dataset (58 species). This approach was favoured to mimic monitoring conditions with limited a prior knowledge on species richness. To date, previous studies often opted for a more conservative approach and used the same DNA extracts for building the reference databases and the NGS library (e.g. Gomez-Rodriguez et al., 2015), which most likely increases the mapping success. Additionally, using a full mitogenomes reference database has been shown to slightly decrease the false negatives and positives rates (Gomez-Rodriguez et al., 2015), but is presently illusive for monitoring purposes due to the lack of published and annotated mitogenomes. In our study, the reduced number of false positives found with the de novo assembly approach (Supplementary Information S9) also indicates that an exhaustive database can considerably improve the outcome of MG. Second, higher coverage rates could help reducing false discovery rates by filtering out all mappings under a certain threshold, or by adding replicates to cross-validate species presence/absence as we did here on the MB dataset. In general, sequencing depth is a major limitation for MG as the vast majority of sequences produced with MG do not correspond to mitochondrial sequences and are therefore currently uninformative (although see Linard et al., 2015). In our study, approximately 0.02% of all reads mapped to the COI reference database for the raw read mapping pipeline (Supplementary Information S5). For the de novo assembly pipeline, approximately 5% of the reads where mapped to the mtDNA reference database. Using full-mitogenomes databases unsurprisingly increases the overall percentage of mapped reads, but in most cases, the mitochondrial fraction will nevertheless plateau around 1% (see review on mitogenomics by Crampton-Platt et al., 2016).

Despite these limitations, this PCR-free method has the advantage of not relying on taxon-specific primers and is therefore universally applicable to any group of animals, or even to plants, fungi or bacteria if other organelles or genes are targeted.

4.3. NGS Barcoding

In terms of species detection and abundance, NGSB performed best by far. Indeed, we found highly similar community compositions compared to the morphological identification data (Tables 2, Figures 2-4,

Supplementary Information S4-S10). Noteworthy, in transect II, two specimens belonging two *Halictus simplex* (as determined by Sanger sequencing) were most probably miss-identified as *H. langobardicus* by NGSB, a species for which barcoding is often challenged due to the co-amplification of nuclear copies of mitochondrial genes (i.e. numts; unpublished data C.Praz). For most western European bee fauna COI barcoding is reliable and provides enough resolution to discriminate at the species level, however there are some known cases of barcode sharing. In our dataset, only one problematic case of barcoding sharing species was sampled (i.e. *Andrena dorsata*, which shares barcodes with *A. propinqua*). After verification, this species was correctly identified for two out of the three methods (i.e. MB and NGSB). For MG, *A. dorsata* was not identified however neither was its sister species (i.e. *A. propinqua*). Therefore, potential biases due to barcoding sharing can be excluded in our study. The PERMANOVA and Procrustes tests on relative or absolute abundance data also indicate high similarity between this method and morphology in terms of species abundance and ecological patterns (Figures 2-3). The level of accuracy found in this study is in the range of previous studies. For instance, Shokralla et al. (2015) applied NGSB to a diverse dataset of arthropods (11 orders) and obtained an overall recovery rate of 97.3% (n = 1010), and 96.5% for Hymenoptera alone (n = 226). Likewise, Wang et al. (2018) sequenced over 4000 ants using NGSB and obtained 95% of correspondence between taxonomy and morphology.

Besides high accuracy, NGSB holds several other advantages over bulk-based approaches (i.e. MB and MG). First, individual DNA extractions and the preservation of associated specimens provide the possibility of verifying unexpected records (e.g. rare species or species outside their known range) through morphology since exoskeletons remain mostly unaltered after proteinase-K digestions. Alternatively, DNA extractions can be performed on single legs as done in our study, and reference specimens could be kept nearly intact, although at the cost of additional workload. The preservation of reference specimens provides a valuable back-up and therefore NGSB data is more likely to be considered for national or international databases, which can be used for purposes other than monitoring (for example, compiling red lists or more generally for conservation biology). Second, DNA barcodes generated using NGSB can be fed into existing DNA databases since a link to the specimen is maintained. Third, DNA extractions can further be used for population genetic or phylogenetic studies. Finally, contrary to MB, NGSB does not require PCR-replicates. Thereby, the sequencing runs of NGSB can encompass larger datasets and provide higher coverages and thus further reduce costs.

Dealing at the specimen instead of community level has, however, a major drawback. Individual extractions and PCRs considerably increase cost and workload linked to the library preparation. This additional workload and cost difference with bulk-based approaches will however largely depend on the number of specimens sampled per community.

4.4. Cost and workload effectiveness

One of the main arguments brought forward for promoting NGS-identification tools in monitoring programs is the potential cost reduction in identifications. Although often stated as more cost-efficient than morphological identification, only few studies have systematically assessed the financial advantages of NGS tools over morphology using "real" monitoring datasets. Overall, we found that all investigated NGS-identification methods were costlier than morphological identification (Supplementary Information S16). For MB and NGSB, sequencing kits constituted the largest fraction of the total cost (Supplementary Information S18). To reduce the overall cost for both methods it is possible to either use smaller sequencing kits or to

increase the number of specimens by sequencing run. Based upon estimations, the smallest Miseq sequencing kit able to span our targeted fragment would considerably decrease costs without compromising sequencing depths (Supplementary Information S19). Although the output of a Miseq v2 Nano kit (2 x 250 bp) corresponds to approximately 1/30 of a Miseq v3 kit, the estimated coverage will remain high, with over 100 mapped reads per specimens. Higher coverages can be expected if clustering optima during sequencing runs are reached. Using the Miseq v2 Nano sequencing kit, the overall cost of MB and NGSB are largely reduced and drop in the range of morphological identification (Supplementary Information S19). Alternatively, with the same sequencing kit used in this study, we estimated that MB and NGSB become more cost-efficient than morphology after 1675 and 4434 specimens, respectively. Noteworthy, several steps of our pipeline could be optimized to even further reduce cost and labor time. For instance, one could reduce hands-on time required for DNA extraction to only a few minutes by using quick DNA extraction kits such as QuickExtract DNA Extraction kit (Lucigen; see Kranzfelder, Ekrem, & Stur, 2016). Studies reducing as much as possible laboratory costs report that sequencing can be performed for approximately 0.50$ per specimen (Wang et al., 2018). Nevertheless, such cost reduction often implies fine-tuning protocols for the targeted taxon, mainly because DNA is amplified through direct-PCR (Wong et al., 2014). Additionally, such price optimization requires running libraries on partial kits/lanes, which is not always possible or proposed by sequencing suppliers. For MG, our cost estimations on the Illumina Miseq platform show that this method will hardly overpass morphology in terms of cost-efficiency. Indeed, sequencing depth is a main bottleneck for this method since only a minor fraction of the data is informative. Therefore, we would recommend sequencing MG libraries on more appropriated platforms, such as Hiseq 4000, Hiseq X or even Novaseq 6000. Although there are indications that the ability to sequence shorter fragments negatively affects the overall mitochondrial proportion, and therefore the fraction of reads corresponding to mitochondrial DNA may be reduced on a Hiseq sequencer (Crampton-Platt et al., 2016; Maddock et al., 2016), using larger scale sequencing platforms will drastically reduce costs and increase species detection rates.

In terms of workload, MB was the least labor-intensive method with approximately 27% less hands-on work than morphological identification. NGSB is unsurprisingly the method requiring most workload, although it is in a close range to MG and morphological identification. Compared to MB, NGSB relies on individual DNA extraction, which is a time demanding procedure, especially since extractions were performed on single legs. With a well-organized protocol, the sorting and DNA extractions required for NGSB may considerably be reduced, potentially to a similar level than MB. Indeed bulk-based approaches like MB or MG also require pre-sorting of raw sampling material to isolate bees from plant material, from numerous honeybees (n = 1422, thus nearly twice as many wild bees in our dataset) and other insects. If none-targeted taxa, and especially honeybees are not removed, the sequencing depth, and therefore detection rates and biomass estimations would largely be affected.

4.5. Conclusions

For routine monitoring of wild bees using molecular identification methods, we recommend NGSB. The reliability and accuracy levels of this method are hardly attainable with bulk-based approaches, especially for species abundance estimation. Furthermore, this approach provides a valuable supplementary security since specimens can be re-examined morphologically if required. NGSB is thus more likely to yield occurrence data that can be validated and integrated into national faunistic databases and thus used by bee experts and by conservation practitioners. Feeding national faunistic databases is an important by-product of monitoring programs (e.g., in Switzerland: http://www.biodiversitymonitoring.ch/en/home.html).

Acknowledgement

We would like to gratefully acknowledge Beatrice Frey and Daniel Frei for their valuable support and advice during the library preparations, as well as Melanie Schirrmann and Ernest Hennig for the fruitful discussions and advices on figure layouts and statistics. We are very thankful the taxonomist Andreas Müller for having accepted to identify our dataset. We also would like to thank all farmers involved in this study for allowing us to sample on their properties. Finally, we are grateful to Marc Matter and Jeannette Regan for having edited several sections of this manuscript. This study was funded by the Swiss Federal Office for Agriculture (FOAG) ("Wild bee metabarcoding project").

Data Accessibility

Absolute abundance matrix of all identification methods is available at Dryad (https://doi.org/10.5061/dryad.gh830j7). This repertory also contains raw sequencing files and associated metadata.

Authors contribution

Sampling scheme was designed and conducted by DG, MA and EK. Laboratory protocols were designed and performed by MG and SB. MG executed the bioinformatics steps and MG, DG, MA and EK performed the statistical analyses. A first draft of the manuscript was written by MG, CP and JF. All authors contributed to the writing of the final version of this paper.

References

Anderson, M. J. (2001). Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian Journal of Fisheries and Aquatic Sciences*, *58*(3), 626–639. doi:10.1139/cjfas-58-3-626

Bartomeus, I., Stavert, J. R., Ward, D., & Aguado, O. (2019). Historical collections as a tool for assessing the global pollination crisis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1763), 20170389. doi:10.1098/rstb.2017.0389

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). doi:10.18637/jss.v067.i01

Biesmeijer, J. C., Roberts, S. P. M., Reemer, M., Ohlemüller, R., Edwards, M., Peeters, T., … Kunin, W. E. (2006). Parallel Declines in Pollinators and Insect-Pollinated Plants in Britain and the Netherlands. *Science*, *313*(5785), 351. doi:10.1126/science.1127863

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. doi:10.1093/bioinformatics/btu170

Brandon-Mong, G.-J., Gan, H.-M., Sing, K.-W., Lee, P.-S., Lim, P.-E., & Wilson, J.-J. (2015). DNA metabarcoding of insects and allies: an evaluation of primers and pipelines. *Bulletin of Entomological Research*, 1–11. doi:10.1017/S0007485315000681

Brunner, P. C., Fleming, C., & Frey, J. E. (2002). A molecular identification key for economically important thrips species (Thysanoptera: Thripidae) using direct sequencing and a PCR-RFLP-based approach. *Agricultural and Forest Entomology*, *4*(2), 127–136. doi:10.1046/j.1461-9563.2002.00132.x

Burkle, L. A., Marlin, J. C., & Knight, T. M. (2013). Plant-Pollinator Interactions over 120 Years: Loss of Species, Co-Occurrence, and Function. *Science*, *339*(6127), 1611–1615. doi:10.1126/science.1232728

Bushnell, B. (2015). BBMap (version 35.14) [Software]. *Available at Https://Sourceforge.Net/Projects/Bbmap/*.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009).

BLAST+: architecture and applications. *BMC Bioinformatics*, *10*(1), 421. doi:10.1186/1471-2105-10-421

Cane, J. H. (1987). Estimation of Bee Size Using Intertegular Span (Apoidea). *Journal of the Kansas Entomological Society*, *60*(1), 145–147. Retrieved from http://www.jstor.org/stable/25084877

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336. doi:10.1038/nmeth.f.303

Choo, L. Q., Crampton-Platt, A., & Vogler, A. P. (2017). Shotgun mitogenomics across body size classes in a local assemblage of tropical Diptera: Phylogeny, species diversity and mitochondrial abundance spectrum. *Molecular Ecology*, *26*(19), 5086–5098. doi:10.1111/mec.14258

Crampton-Platt, A., Timmermans, M. J. T. N., Gimmel, M. L., Kutty, S. N., Cockerill, T. D., Khen, C. V., & Vogler, A. P. (2015). Soup to tree: The phylogeny of beetles inferred by mitochondrial metagenomics of a bornean rainforest sample. *Molecular Biology and Evolution*, *32*(9), 2302–2316. doi:10.1093/molbev/msv111

Crampton-Platt, A., Yu, D. W., Zhou, X., & Vogler, A. P. (2016). Mitochondrial metagenomics: letting the genes out of the bottle. *GigaScience*, *5*(1), 15. doi:10.1186/s13742-016-0120-y

Dowle, E. J., Pochon, X., Banks, J. C., Shearer, K., & Wood, S. A. (2016). Targeted gene enrichment and high-throughput sequencing for environmental biomonitoring: A case study using freshwater macroinvertebrates. *Molecular Ecology Resources*, *16*, 1240–1254. doi:10.1111/1755-0998.12488

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461. doi:10.1093/bioinformatics/btq461

Elbrecht, V., & Leese, F. (2015). Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol. *PLoS ONE*, *10*(7), 1–16. doi:10.1371/journal.pone.0130324

Elbrecht, V., Taberlet, P., Dejean, T., Valentini, A., & Usseglio-polatera, P. (2016). Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. *PeerJ*, 1–12. doi:10.7717/peerj.1966

Elbrecht, V., Vamos, E. E., Meissner, K., Aroviita, J., & Leese, F. (2017). Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. *Methods in Ecology and Evolution*, *8*(10), 1265–1275. doi:10.1111/2041-210X.12789

Fadrosh, D. W., Ma, B., Gajer, P., Sengamalay, N., Ott, S., Brotman, R. M., & Ravel, J. (2014). An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome*, *2*(1), 6. doi:10.1186/2049-2618-2-6

Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., … Taberlet, P. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, *5*(3), 543–556. doi:10.1111/1755-0998.12338

Ficetola, G. F., Taberlet, P., & Coissac, E. (2016). How to limit false positives in environmental DNA and metabarcoding? *Molecular Ecology Resources*, *16*(3), 604–607. doi:10.1111/1755-0998.12508

Gibson, J. F., Stein, E. D., Baird, D. J., Max, F. C., Zhang, X., & Hajibabaei, M. (2015). Wetland Ecogenomics – The Next Generation of Wetland Biodiversity and Functional Assessment. *Wetland Science and Practice*. doi:10.1007/978-3-540-70962-6_5

Gibson, J., Shokralla, S., Porter, T. M., King, I., van Konynenburg, S., Janzen, D. H., … Hajibabaei, M. (2014). Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(22), 8007–12. doi:10.1073/pnas.1406468111

Gillett, C. P. D. T., Crampton-Platt, A., Timmermans, M. J. T. N., Jordal, B. H., Emerson, B. C., & Vogler, A. P. (2014). Bulk de novo mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea). *Molecular Biology and Evolution*, *31*(8), 2223–2237. doi:10.1093/molbev/msu154

Gomez-Rodriguez, C., Crampton-Platt, A., Timmermans, M. J. T. N., Baselga, A., & Vogler, A. P. (2015). Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. *Methods in Ecology and Evolution*, 883–894. doi:10.1111/2041-210X.12376

Goulson, D., Nicholls, E., Botías, C., & Rotheray, E. L. (2015). Bee declines driven by combined stress from parasites, pesticides, and lack of flowers. *Science*, *347*(6229). doi:10.1126/science.1255957

Hebert, P. D. N., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, *270*(1512), 313–321. doi:10.1098/rspb.2002.2218

Imperatriz-Fonseca, V. L., Potts, S., Andreas Baste, I., Apau Oteng Yeboah, A., Alfredo Joly, C., Bartuska, A., … Kamaljit, B. (2016). *The assessment report of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services on pollinators, pollination and food production*.

Ivanova, N. V., Dewaard, J. R., & Hebert, P. D. N. (2006). An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Molecular Ecology Notes*, *6*(4), 998–1002. doi:10.1111/j.1471-8286.2006.01428.x

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, *11*(2), 37–50. doi:10.1111/j.1469-8137.1912.tb05611.x

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., … Yu, D. W. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, *16*(10), 1245–1257. doi:10.1111/ele.12162

Joseph, L. N., Field, S. A., Wilcox, C., & Possingham, H. P. (2006). Presence-absence versus abundance data for monitoring threatened species. *Conservation Biology*, *20*(6), 1679–1687. doi:10.1111/j.1523-1739.2006.00529.x

Kranzfelder, P., Ekrem, T., & Stur, E. (2016). Trace DNA from insect skins: a comparison of five extraction protocols and direct PCR on chironomid pupal exuviae. *Molecular Ecology Resources*, *16*(1), 353–363. doi:10.1111/1755-0998.12446

Lahoz-Monfort, J. J., Guillera-Arroita, G., & Tingley, R. (2016). Statistical approaches to account for false-positive errors in environmental DNA samples. *Molecular Ecology Resources*, *16*(3), 673–685. doi:10.1111/1755-0998.12486

Lebuhn, G., Droege, S. A. M., Connor, E. F., Gemmill-Herren, B., Potts, S. G., Minckley, R. L., … Parker, F. (2013). Detecting Insect Pollinator Declines on Regional and Global Scales. *Conservation Biology*, *27*(1), 113–120. doi:10.1111/j.1523-1739.2012.01962.x

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., … Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, *10*(1), 34. doi:10.1186/1742-9994-10-34

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. doi:10.1093/bioinformatics/btp352

Linard, B., Crampton-Platt, A., Gillett, C. P. D. T., Timmermans, M. J. T. N., & Vogler, A. P. (2015). Metagenome skimming of insect specimen pools: potential for comparative genomics. *Genome Biology*

and Evolution, *7*(6), 1474–1489. doi:10.1093/gbe/evv086

Linard, B., Crampton-Platt, A., Moriniere, J., Timmermans, M. J. T. N., Andújar, C., Arribas, P., … Vogler, A. P. (2018). The contribution of mitochondrial metagenomics to large-scale data mining and phylogenetic analysis of Coleoptera. *Molecular Phylogenetics and Evolution*, *128*, 1–11. doi:10.1016/j.ympev.2018.07.008

Liu, S., Li, Y., Lu, J., Su, X., Tang, M., Zhang, R., … Zhou, X. (2013). SOAPBarcode: Revealing arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons. *Methods in Ecology and Evolution*, *4*(12), 1142–1150. doi:10.1111/2041-210X.12120

Liu, S., Wang, X., Xie, L., Tan, M., Li, Z., Su, X., … Zhou, X. (2016). Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis. *Molecular Ecology Resources*, *16*(2), 470–479. doi:10.1111/1755-0998.12472

MacKenzie, D. I. (2005). What Are the Issues with Presence-Absence Data for Wildlife Managers? *The Journal of Wildlife Management*, *69*(3), 849–860. Retrieved from http://www.jstor.org/stable/3803327

Maddock, S. T., Briscoe, A. G., Wilkinson, M., Waeschenbach, A., San Mauro, D., Day, J. J., … Gower, D. J. (2016). Next-Generation Mitogenomics: A Comparison of Approaches Applied to Caecilian Amphibian Phylogeny. *Plos One*, *11*(6), e0156757. doi:10.1371/journal.pone.0156757

Mata, V. A., Rebelo, H., Amorim, F., McCracken, G. F., Jarman, S., & Beja, P. (2019). How much is enough? Effects of technical and biological replication on metabarcoding dietary analysis. *Molecular Ecology*, *28*(2), 165–175. doi:10.1111/mec.14779

Meyer, M., & Kircher, M. (2010). Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harbor Protocols*. doi:10.1101/pdb.prot5448

Nieto, A., Roberts, S. P. M., Kemp, J., Rasmont, P., Kuhlmann, M., Criado, M. G., … Michez, D. (2015). *European Red List of Bees*. doi:10.2779/77003

Ollerton, J., Erenler, H., Edwards, M., & Crockett, R. (2014). Extinctions of aculeate pollinators in Britain and the role of large-scale agricultural changes. *Science*, *346*(6215), 1360–1362. doi:10.1126/science.1257259

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, *28*(11), 1420–1428. doi:10.1093/bioinformatics/bts174

Piñol, J., Mir, G., Gomez-Polo, P., & Agustí, N. (2015). Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular Ecology Resources*, *15*, 819–830. doi:10.1111/1755-0998.12355

Piñol, J., Senar, M. A., & Symondson, W. O. C. (2019). The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. *Molecular Ecology*, *28*(2), 407–419. doi:10.1111/mec.14776

Porazinska, D. L., Sung, W., Giblin-Davis, R. M., & Thomas, W. K. (2010). Reproducibility of read numbers in high-throughput sequencing analysis of nematode community composition and structure. *Molecular Ecology Resources*, *10*(4), 666–676. doi:10.1111/j.1755-0998.2009.02819.x

Potts, S., Biesmeijer, K., Bommarco, R., Kleijn, D., & Scheper, J. A. (2015). *Status and trends of European pollinators. Key findings of the STEP project*. 70, , : Pensoft Publishers. Retrieved from http://edepot.wur.nl/389377

Potts, S. G., Biesmeijer, J. C., Kremen, C., Neumann, P., Schweiger, O., & Kunin, W. E. (2010). Global pollinator declines: Trends, impacts and drivers. *Trends in Ecology and Evolution*, *25*(6), 345–353. doi:10.1016/j.tree.2010.01.007

Rodgers, T. W., Xu, C. C. Y., Giacalone, J., Kapheim, K. M., Saltonstall, K., Vargas, M., … Jansen, P. A. (2017). Carrion fly-derived DNA metabarcoding is an effective tool for mammal surveys: Evidence from a known tropical mammal community. *Molecular Ecology Resources*, *17*(6), 133–145. doi:10.1111/1755-0998.12701

Schnell, I. B., Thomsen, P. F., Wilkinson, N., Rasmussen, M., Jensen, L. R. D., Willerslev, E., … Gilbert, M. T. P. (2012). Screening mammal biodiversity using DNA from leeches. *Current Biology*, *22*(8), 262–263. doi:10.1016/j.cub.2012.02.058

Shokralla, S., Gibson, J. F., Nikbakht, H., Janzen, D. H., Hallwachs, W., & Hajibabaei, M. (2014). Next-generation DNA barcoding: Using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Molecular Ecology Resources*, *14*(5), 892–901. doi:10.1111/1755-0998.12236

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). *Environmental DNA: For Biodiversity Research and Monitoring*. Oxford: Oxford University Press. doi:10.1093/oso/9780198767220.001.0001

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*, 2045–2050. doi:10.1111/j.1365-294X.2012.05470.x

Tang, M., Hardman, C. J., Ji, Y., Meng, G., Liu, S., Tan, M., … Yu, D. W. (2015). High-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods in Ecology and Evolution*, *6*(9), 1034–1043. doi:10.1111/2041-210X.12416

Tang, M., Tan, M., Meng, G., Yang, S., Su, X., Liu, S., … Zhou, X. (2014). Multiplex sequencing of pooled mitochondrial genomes - A crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research*, *42*(22), 1–13. doi:10.1093/nar/gku917

vanEngelsdorp, D., & Meixner, M. D. (2010). A historical review of managed honey bee populations in Europe and the United States and the factors that may affect them. *Journal of Invertebrate Pathology*, *103*, 80–95. doi:10.1016/J.JIP.2009.06.011

Wang, W. Y., Srivathsan, A., Foo, M., Yamane, S. K., & Meier, R. (2018). Sorting specimen-rich invertebrate samples with cost-effective NGS barcodes: Validating a reverse workflow for specimen processing. *Molecular Ecology Resources*, *18*(3), 490–501. doi:10.1111/1755-0998.12751

Wilson, J.-J., Brandon-Mong, G.-J., Gan, H.-M., & Sing, K.-W. (2019). High-throughput terrestrial biodiversity assessments: mitochondrial metabarcoding, metagenomics or metatranscriptomics? *Mitochondrial DNA Part A*, *30*(1), 60–67. doi:10.1080/24701394.2018.1455189

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., & Ding, Z. (2012). Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, *3*(4), 613–623. doi:10.1111/j.2041-210X.2012.00198.x

Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L., … Huang, Q. (2013). Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience*, *2*(1), 4. doi:10.1186/2047-217X-2-

Supplementary Information

S1: Sanger sequencing protocol.

A fragment of the COI gene was amplified by Polymerase Chain Reaction (PCR) using the primers LCO1490 and HCO2198 (Folmer, Black, Hoeh, Lutz, & Vrijenhoek, 1994). After PCR-cleanup, linear amplification was done using Big Dye Terminator v3.1 chemistry (Applied Biosystems) and reaction cleanup was performed using DyeEX 96 kits (Qiagen). Sequences obtained from a 3130xl Genetic Analyzer (Applied Biosystems) were assembled and taxonomically identified by BLAST queries on the NCBI GenBank database using Geneious v10.2.3. Samples for which amplification did not work were re-amplified using the set of primers mlCOIintF / HCO2198 (Leray et al., 2013). See details of reaction below.

1.PCR amplification

Reagents	Volume/sample (ul)
HotStarTaq DNA polymerase Mix (2x) (Qiagen)	10
Primer F (20uM)	0.4
Primer R (20uM)	0.4
H2O	8.2
DNA	1
Final	20

PCR conditions:
Lep F/ Lep R
15min 95°C // **35** cycles of 94°C for 1 min, **45**°C for 1min and 72°C for 90 s; 72 for 10 min; hold 4°C

mlCOIintF/HCO
15min 95°C // **35** cycles of 94°C for 1 min, **45**°C for 1min and 72°C for 90 s; 72 for 10 min; hold 4°C

2. Electrophoresis
Agarose gel 1.5%

3.Purification
Marchery-Nagel purification plates (as recommended by supplier)

4. Linear amplification

Reagents	Volume/sample (ul)
BigDye Terminator Mix (ThermoFisher)	2
Primer F (20uM)	0.2
H2O	5.3
DNA	[0.5;5.8]
Final	8

5.Purification
DyeEx 96 plates (Qiagen) (as recommended by supplier)

S2: Metabarcoding library preparation details.

PCR amplification was performed in a total volume of 20 µl with 1 µl each of 10 µM of forward and reverse primer, 10 µl of HotStartTaq DNA polymerase 5 U/µl mix (Qiagen), 7 µl of ddH2O and 1 µl of bulk DNA.

PCR conditions were as follows: initial activation of 15 min at 95 °C, 40 cycles of denaturation for 40 sec at 95 °C (1 °C/sec ramping), annealing for 15 sec at 45 °C (1 °C/sec ramping), one minute ramping to 60 °C (0.250 °C/sec ramping) and extension for 2 min at 72 °C (1 °C/sec ramping), followed by a final extension at 72 °C for 7 min (1 °C/sec ramping) (Frey et al., 2013). Success of PCR amplifications was verified on a 1.5% agarose gel after pooling all barcoded PCR-replicates per community. Pooled PCR products were purified using NucleoFast 96 PCR clean-up kits (Marcherey-Nagel) and eluted in 100 µl ddH2O. Purified PCR products were then quantified using a Qubit v4 (Thermofisher Scientific). To optimize sequencing depth and have higher coverages for specimens-rich communities than for specimens-poor communities, the communities' PCR products were pooled into a final library based upon specimen-richness (number of specimens) following this correction factor:

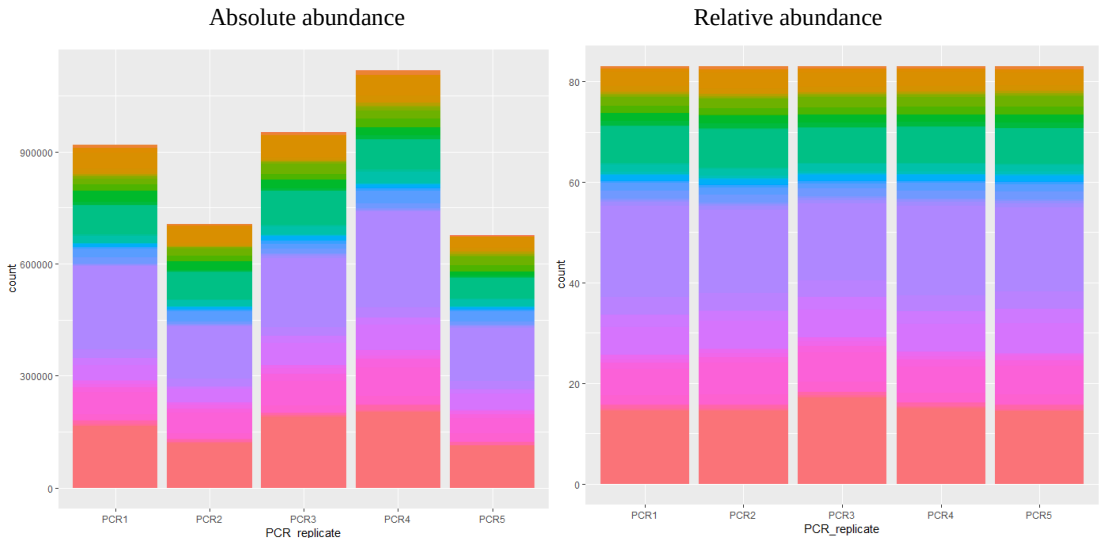
$$Pooling\ volume\ per\ community = ln(\frac{200}{[concentration]} * nb\ individuals\ within\ community)$$

Based upon the correction factor, communities were pooled and sequenced on an Illumina Miseq using a v3 kit (300 bp x 2) spiked with 20% Phix.

S3: Bioinformatic pipelines for metabarcoding, next generation sequencing barcoding and mitogenomics.

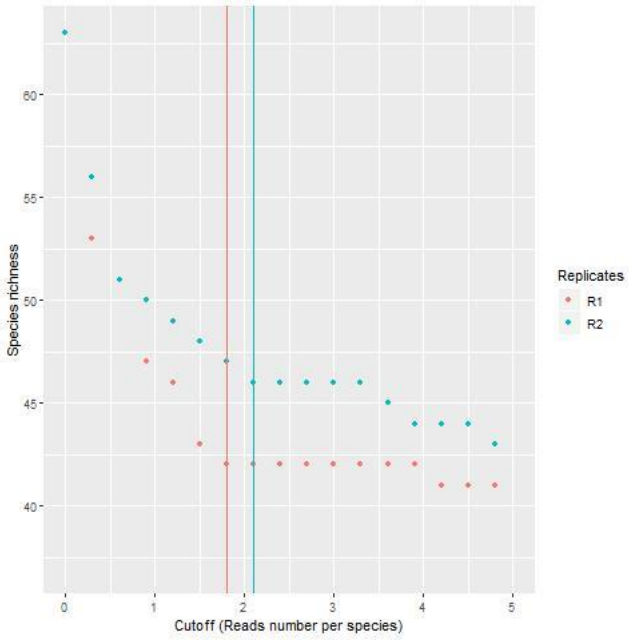
Metabarcoding

Non-assigned OTUs, likely corresponding to contaminants, nuclear pseudogenes or chimeric sequences, were filtered out. Differences in species composition, species relative abundance and species absolute abundances (read numbers) between the five PCR-replicates were visually inspected by stacked barplots:



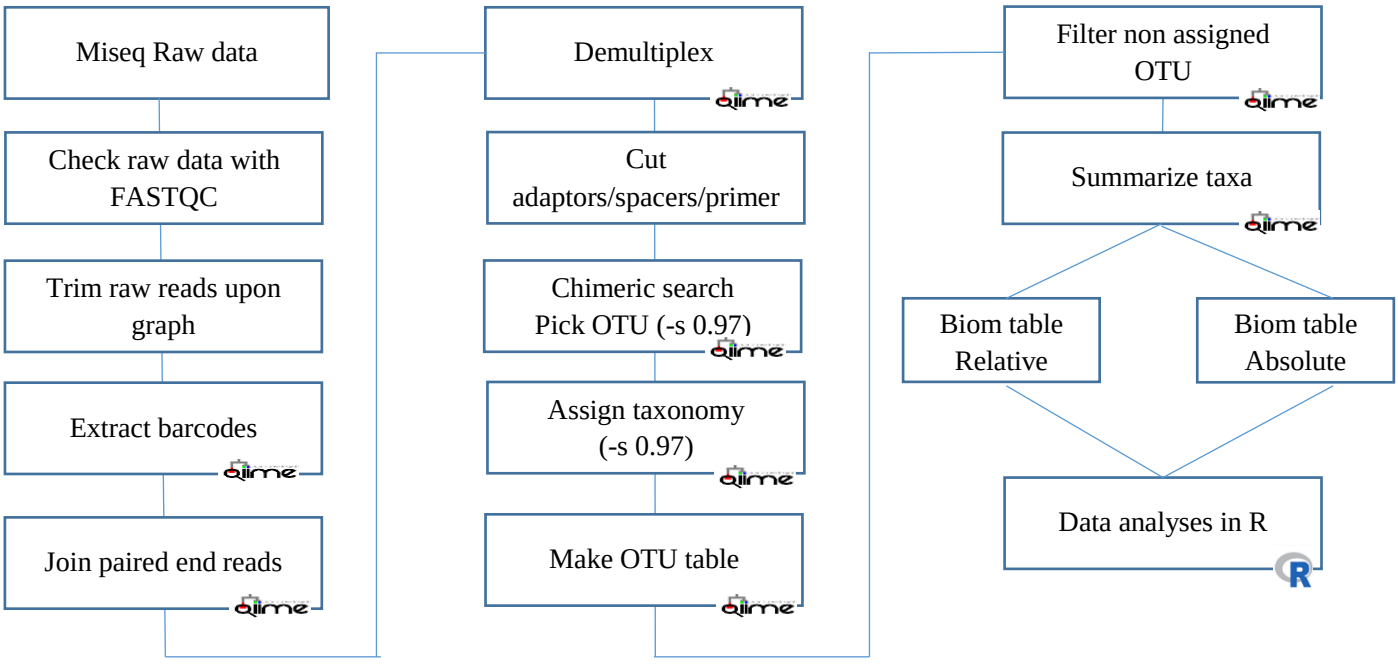
Metabarcoding abundance and relative abundance comparison in species composition between technical replicates. Each colour represents a different species.

To reduce the number of false positives, a supplementary filter was applied to only retain taxonomical assignments that were shared among 3 different PCR-replicates per community. Read numbers were multiplied by the exponential of their pooling volume to obtain uncorrected raw abundances (see equation Supplementary Information S2). Finally, to further reduce the number of false positives, a quality filtering based on OTU abundance was applied (Bokulich et al., 2013). To find the optimal filtering threshold, we plotted decreasing numbers of species according to the filtering threshold applied:

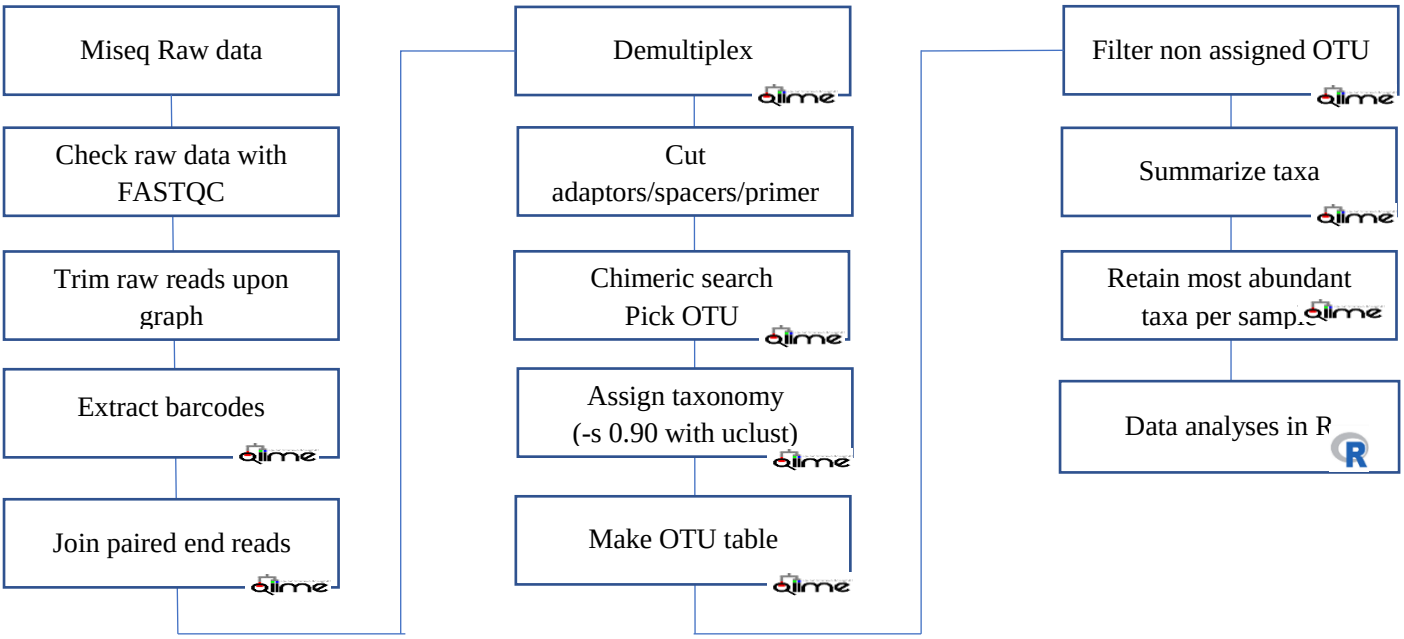


Plot of species richness per transect according to filtering thresholds based upon read numbers.

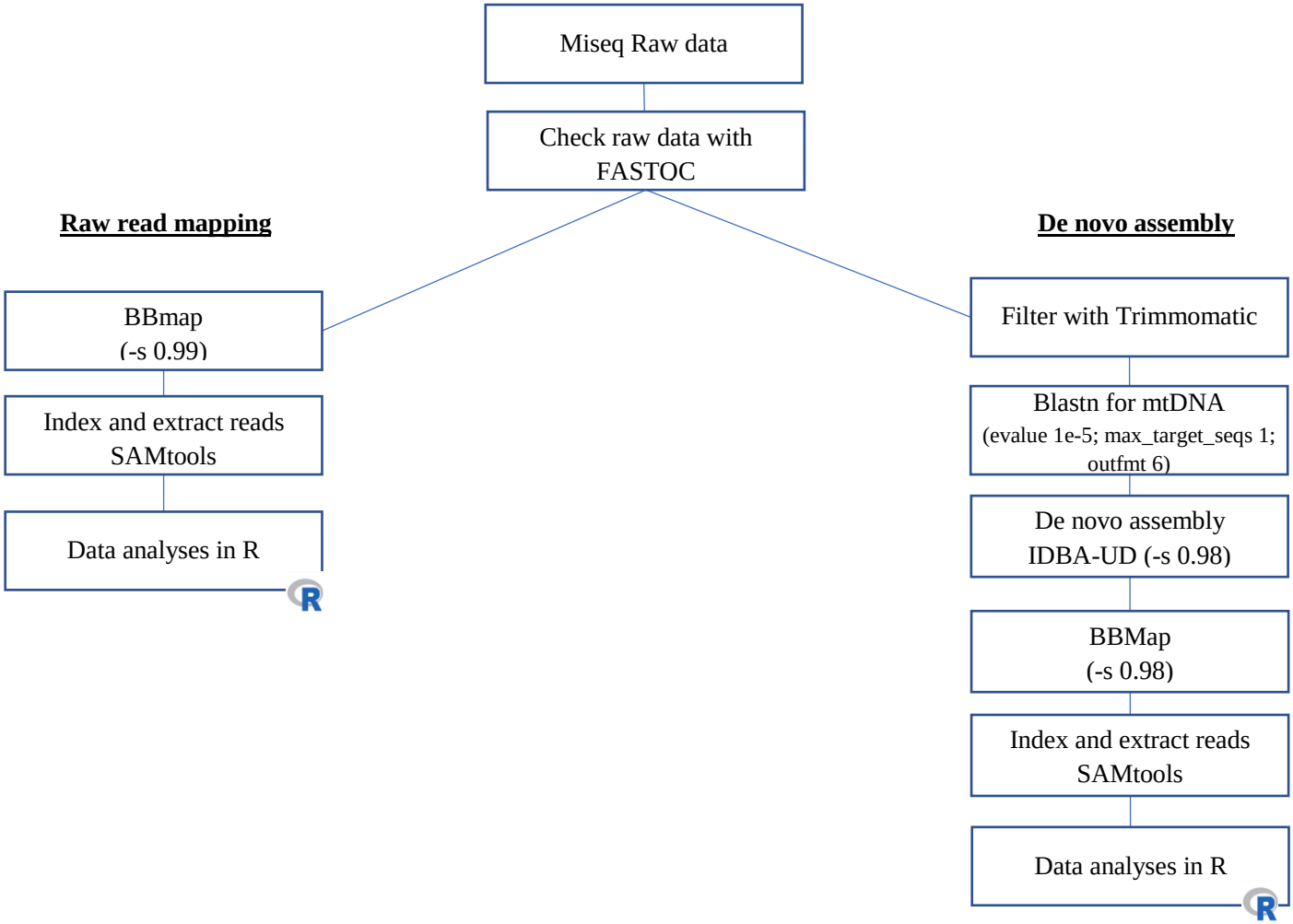
The optimal filtering threshold was then visually determined, similar to the process of the "elbow" method used to determine the optimal number of clusters in portioning clustering methods (e.g. K-means). Once all filtering steps were performed, the obtained absolute abundance data was transformed into relative abundance (per community) and presence/absence. Schematic workflow of bioinformatics steps is given hereunder.



Next Generation Sequencing Barcoding



Mitogenomics



S4: Summary table of species composition for all identification methods. Table depicts per transect the number of specimens for the morphological (Morpho) and next generation sequencing barcoding (NGSB) methods, as well as the relative abundance of read numbers for the metabarcoding (MB) and mitogenomics (MG) methods and/or the relative biomass for Morpho. To enable comparisons, the biomass of each specimen was summed and transformed into relative abundance for the morphological dataset. False negatives are highlighted in blue and false positives in red.

	Transect I					Transect II				
	Nb specimens		Relative biomass			Nb specimens		Relative biomass		
Species	Morpho	NGSB	Morpho	MB	MG	Morpho	NGSB	Morpho	MB	MG
Andrena dorsata	-	-	-	-	0.0001	1	1	0.0031	0.0200	-
Andrena flavipes	12	2	0.0502	-	0.0006	10	1	0.0478	-	<0.0001
Andrena fulvago	-	-	-	-	-	1	1	0.0042	0.0146	<0.0001
Andrena lagopus	3	3	0.0111	0.0006	-	-	-	-	-	-
Andrena minutula	2	2	0.0022	0.0101	0.0005	-	-	-	-	-
Andrena ovatula	6	7	0.0160	0.0669	0.0009	4	4	0.0193	0.0127	<0.0001
Andrena pandellei	-	-	-	-	-	1	1	0.0045	0.0547	0.0021
Andrena subopaca	-	-	-	-	-	1	1	0.0021	0.0034	<0.0001
Andrena wilkella	-	-	-	-	-	1	1	0.0063	0.0018	<0.0001
Bombus lucorum	2	2	0.0331	0.0420	0.0192	2	2	0.0581	0.0142	0.0007
Bombus subterraneus	-	-	-	-	-	1	1	0.0266	0.0003	<0.0001
Bombus barbutellus	-	-	-	-	-	1	1	0.0375	0.0076	0.0012
Bombus bohemicus	1	1	0.0155	0.0004	0.0004	-	-	-	-	-
Bombus hortorum	1	1	0.0189	0.0030	<0.0001	2	3	0.0608	0.0031	0.0005
Bombus hypnorum	-	-	-	-	-	1	1	0.0297	0.0096	0.0004
Bombus lapidarius	3	3	0.0426	0.0204	0.0075	2	2	0.0369	0.0412	0.0010
Bombus norvegicus	1	1	0.0303	0.0014	0.0006	-	-	-	-	-
Bombus pascuorum	5	4	0.0740	0.0431	0.2925	1	1	0.0149	0.0007	0.0003
Bombus pratorum	3	3	0.0371	0.0362	0.1828	-	-	-	-	-
Bombus rupestris	1	1	0.0184	0.0005	0.0003	-	-	-	-	-
Bombus sylvarum	-	-	-	-	0.0254	1	1	0.0141	0.0040	0.0003
Bombus terrestris	16	16	0.2747	0.1685	0.0591	3	3	0.0917	0.1048	0.7834
Ceratina cyanea	-	-	-	-	-	1	1	0.0011	0.0024	<0.0001
Chelostoma florisomne	-	-	-	-	-	1	1	0.0037	0.0088	-
Chelostoma rapunculi	5	5	0.0079	0.0025	<0.0001	-	-	-	-	-
Colletes similis	3	3	0.0086	0.0342	0.0059	-	-	-	-	-
Eucera nigrescens	-	-	-	-	-	1	1	0.0179	0.0011	0.0002
Halictus langobardicus	2	5	0.0043	0.0321	<0.0001	-	2	-	0.0073	-
Halictus maculatus	1	2	0.0014	0.0006	<0.0001	1	1	0.0020	0.0110	0.0001
Halictus scabiosae	3	3	0.0110	0.0067	0.0239	13	13	0.0723	0.0143	0.0010
Halictus simplex	3	2	0.0068	0.0076	0.0017	3	1	0.0120	0.0189	0.0013
Halictus subauratus	5	5	0.0063	0.0123	<0.0001	1	2	0.0019	0.0037	-
Halictus tumulorum	6	7	0.0064	0.0173	<0.0001	12	12	0.0205	0.1245	0.0006
Heriades truncorum	4	4	0.0046	0.0006	0.0007	3	3	0.0074	0.0005	<0.0001
Hylaeus communis	3	2	0.0049	-	<0.0001	3	3	0.0033	0.0037	<0.0001
Hylaeus difformis	1	1	0.0013	0.0050	0.0763	-	-	-	-	-
Hylaeus gredleri	2	1	0.0008	-	-	-	-	-	-	-
Lasioglossum calceatum	7	7	0.0155	0.0029	0.0154	4	4	0.0092	0.0003	<0.0001
Lasioglossum fulvicorne	3	3	0.0046	0.0097	0.0001	2	3	0.0033	0.0257	<0.0001
Lasioglossum glabriusculum	7	7	0.0035	0.0005	<0.0001	10	11	0.0058	0.0010	<0.0001

Lasioglossum interruptum	2	2	0.0029	0.0035	<0.0001	1	1	0.0015	0.0039	0.0003
Lasioglossum laticeps	17	16	0.0218	0.0180	0.0018	22	24	0.0361	0.0236	0.0016
Lasioglossum leucozonium	2	2	0.0066	0.0032	-	4	4	0.0129	0.0020	<0.0001
Lasioglossum lineare	4	4	0.0033	0.0049	0.0008	1	1	0.0012	0.0005	-
Lasioglossum malachurum	113	121	0.1413	0.2392	0.1069	81	87	0.1631	0.2355	0.0042
Lasioglossum morio	16	16	0.0104	0.0928	0.0018	17	17	0.0166	0.0450	0.0003
Lasioglossum nigripes	7	7	0.0175	-	<0.0001	5	4	0.0189	-	<0.0001
Lasioglossum pauxillum	43	39	0.0292	0.0166	0.0083	57	52	0.0601	0.0382	0.0005
Lasioglossum politum	38	40	0.0180	0.0295	0.0001	32	34	0.0223	0.1160	0.0003
Lasioglossum puncticolle	1	3	0.0016	0.0019	-	-	-	-	-	-
Lasioglossum villosulum	23	24	0.0238	0.0415	0.0002	22	21	0.0329	0.0154	<0.0001
Lasioglossum zonulum	1	1	0.0018	0.0009	<0.0001	2	2	0.0056	0.0016	<0.0001
Osmia caerulescens	2	2	0.0081	0.0216	0.0002	-	-	-	-	-
Sphecodes crassus	-	-	-	-	-	1	1	0.0011	0.0003	-
Sphecodes ephippius	1	1	0.0007	-	<0.0001	1	1	0.0027	-	-
Sphecodes ferruginatus	-	-	-	-	-	1	1	0.0016	0.0003	<0.0001
Sphecodes geoffrellus	-	-	-	-	-	1	1	0.0013	-	<0.0001
Sphecodes puncticeps	1	1	0.0009	0.0004	<0.0001	4	4	0.0033	-	-
Undetermined	-	-	-	-	-	1	-	0.0008	-	-
Andrena falsifica	-	-	-	-	-	-	-	-	0.0005	-
Andrena gravida	-	-	-	-	0.0001	-	-	-	-	-
Andrena humilis	-	-	-	-	0.0318	-	-	-	-	-
Andrena pilipes	-	-	-	-	0.0010	-	-	-	-	<0.0001
Apis mellifera	-	1	-	-	0.1203	-	4	-	-	0.1431
Bombus argillaceus	-	-	-	-	0.0021	-	-	-	-	-
Bombus cryptarum	-	-	-	-	0.0048	-	-	-	-	0.0281
Bombus wurflenii	-	-	-	-	0.0053	-	-	-	-	0.0280
Dasypoda suripes	-	-	-	-	-	-	-	-	-	<0.0001
Eucera interrupta	-	-	-	-	-	-	-	-	-	0.0001
Halictus compressus	-	-	-	0.0003	-	-	-	-	-	-
Halictus gavarnicus	-	-	-	-	-	-	-	-	0.0010	-
Hylaeus intermedius	-	1	-	-	-	-	-	-	-	-
Hylaeus kahri	-	-	-	-	<0.0001	-	-	-	-	-
Hylaeus moricei	-	-	-	-	<0.0001	-	-	-	-	<0.0001
Hylaeus punctulatissimus	-	-	-	-	<0.0001	-	-	-	-	-
Lasioglossum albipes	-	-	-	-	-	-	-	-	-	<0.0001
Lasioglossum clypeare	-	-	-	-	-	-	-	-	-	<0.0001
Lasioglossum crenicornis	-	-	-	0.0005	-	-	-	-	0.0004	-
Seladonia gavarnica	-	-	-	-	-	-	-	-	-	<0.0001
Sphecodes niger	-	-	-	-	0.0004	-	-	-	-	-
Total nb Specimens	382					341				

S5: Sequencing output and read mapping information for metabarcoding (MB), mitogenomics (MG) and next generation sequencing barcoding (NGSB) libraries. The average species and community coverage were computed by dividing the number of matching reads by the total number of specimens (n = 723) and communities (n = 83). For MG, raw reads were directly mapped to the reference database without filtering.

	Total read number (per direction)	Read number after filtering	Read number matching the Apoidea database	Average specimen coverage	Average community coverage
MB	13,828,724	4,530,461	3,940,098 (28%)	5450	47,471
MG	17,483,469	NA	3126 (0.018%)	4	38
NGSB	9,055,239	3,449,357	2,862,255 (32%)	3959	34,485

S6: Results of two different COI reference datasets on species detection for MB and NGSB. For both methods, two different reference datasets were used two assign OTU's taxonomy. The uncurated dataset encompassed all available COI sequences of Apoidea members (barcodes for ca. 2000 species) presently available on BOLD and Genbank (downloaded in June 2017). The curated dataset encompassed sequences deposited on BOLD by Schmidt and colleagues (2015) in their barcoding study on western-Europeans bees (dx.doi.org/10.5883/DS-GBAPI). Both datasets were verified to harbor barcodes for all investigated species. In case of lacking sequences, additional barcodes were download form BOLD and added manually to the databases.
The Jaccard similarity index were computed between the global diversity of the molecular and morphological methods.

Method	Database	Species Richness	# Shared species	False Positives	False Negatives	Jaccard Index
MB	Uncurated	57	53	4	5	0.855
	Curated	57	52	5	6	0.825
NGSB	Uncurated	60	58	2	0	0.967
	Curated	57	55	2	3	0.917

S7: Result for different bioinformatic parameters for metabarcoding identifications. Jaccard indexes was computed based on the number of shared species between the morphological and molecular identification method. The Adonis results were computed by non-parametric multivariate analysis of variance (PERMANOVA; *Adonis* function) between morphological and metabacoding community matrixes and represent the proportion (R^2) of variance explained by the solo effect of the identification method for presence/absence (PA), absolute abundance (AB) and relative abundance (RA) data. Asterisk (*) represent values that significantly explain de variance between both morphological and MB matrices (Significance codes: 0 '*' 0.001 '**' 0.01 '*'). Parameters that were retained for downstream analyses are marked in bold. Table A depict result for different similarity thresholds in uclust between the OTU's representative sequence and the reference database. Table B depicts results for different cross-validation settings between replicates (i.e. minimal time a species has to occurrence between replicates to be validated) using the best similarity threshold found in table A.**

A.

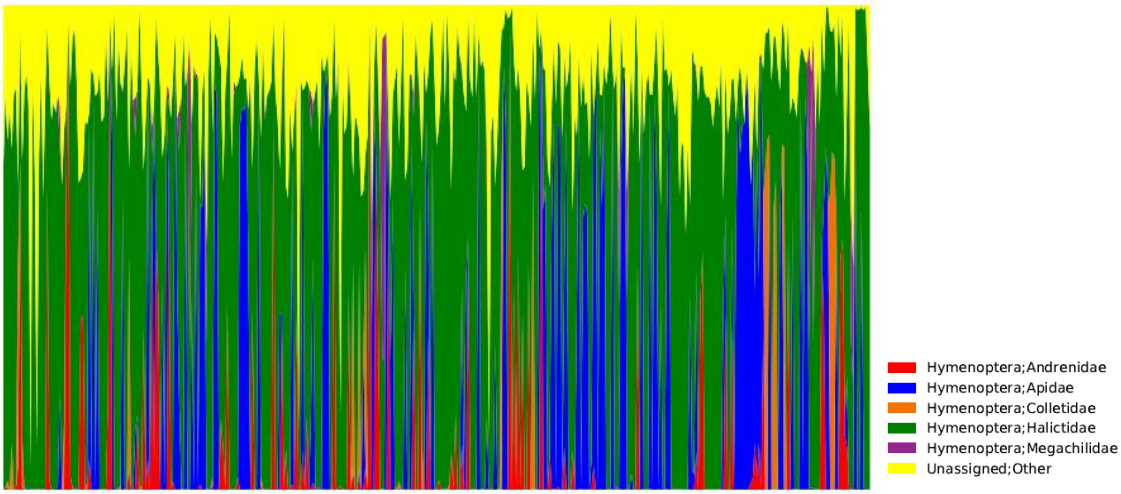
uclust %	Sp Richness	# Shared Sp	False Positives	False Negatives	Jaccard Index	Adonis PA	Adonis AB	Adonis RA
90%	64	50	14	8	0.694	0.010*	0.091***	0.023***
95%	66	52	14	6	0.722	0.018***	0.089***	0.026***
96%	59	51	8	7	0.773	0.007	0.089***	0.023***
97%	**57**	**53**	**4**	**5**	**0.855**	**0.005**	**0.088*** **	**0.022*** **
98%	57	53	4	5	0.855	0.008	0.088***	0.022***
99%	31	30	1	28	0.508	0.041***	0.064***	0.045***

B.

uclust %	Replicate Filtering	Sp Richness	# Shared Sp	False Positives	False Negatives	Jaccard Index	Adonis PA	Adonis AB	Adonis RA	Cor_AB	Cor_RA
97%	1/5	58	53	5	5	0.841	0.005	0.088***	0.022***	0.582	0.745
97%	2/5	58	53	5	5	0.841	0.005	0.088***	0.022***	0.582	0.745
97%	**3/5**	**57**	**53**	**4**	**5**	**0.855**	**0.005**	**0.088*** **	**0.022*** **	**0.549**	**0.704**
97%	4/5	57	53	4	5	0.855	0.005	0.088***	0.022***	0.549	0.704
97%	5/5	57	53	4	5	0.855	0.005	0.088***	0.022***	0.534	0.674

S8: Information on number of unassigned OTU's per community. (A.) Stacked barplot of the taxonomical assignment of OTU's per family. Yellow proportions correspond to the unassigned OTU's. The plot was drawn using the QIMME v1 (Caporaso et al., 2010) script "plot_taxa_summary.py" (B.) Summary statistics on the number of unassigned OTU's per community.

A.



B.

Min unassigned OTU's	1st Qu unassigned OTU's	Median unassigned OTU's	Mean unassigned OTU's	3rd Qu unassigned OTU's	Max. unassigned OTU's
0.002485	0.110934	0.163157	0.181541	0.231936	0.956276

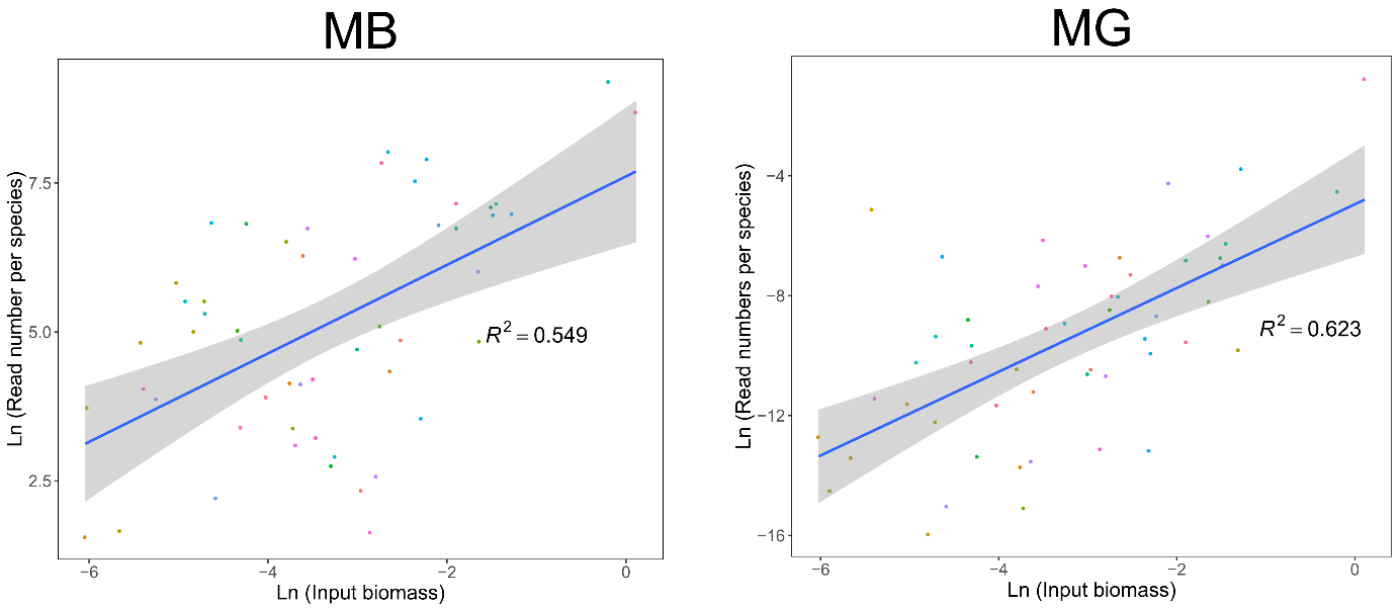
S9: Jaccard similarity index between the global diversity of morphological and two different bioinformatic pipelines for the mitogenomics dataset.

Datasets	Methods	Transects	Species Richness	# Shared Species	False Positives	False Negatives	Jaccard Index
Between MG and Morpho	Raw read mapping	I & II	69	53	16	5	0.716
	De novo assembly	I & II	35	17	18	41	0.224

S10: Jaccard similarity index between the global diversity of morphological and molecular datasets by transects. Percentages were computed upon the morphological species richness of the respective transect.

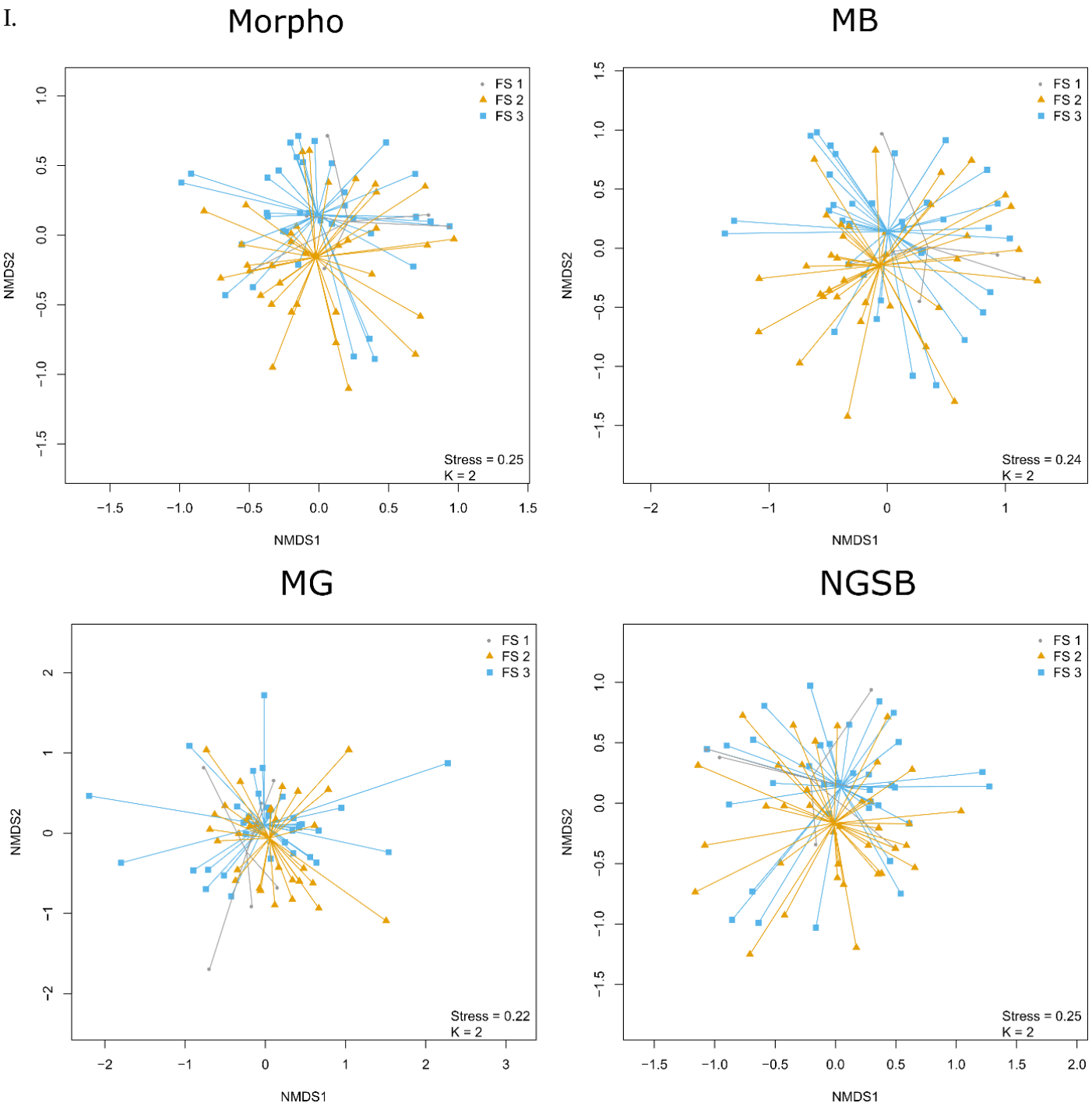
Datasets	Transects	Species Richness	# Shared Species	False Positives	False Negatives	Jaccard Index
Between transects of Morpho	I	43	30 (30/43 = 69.8%)	NA	NA	0.508
	II	46	30 (30/46 = 65.2%)	NA	NA	0.508
Between MB and Morpho	I	40	38 (38/40 = 95%)	2(2/40 = 5%)	5 (5/40 = 12.5%)	0.844
	II	44	40 (40/44 = 90.9%)	4(4/44 = 9.1%)	6 (6/44 = 13.6%)	0.800
Between MG and Morpho	I	53	39 (73.6%)	14 (26.4%)	4 (7.5%)	0.684
	II	48	38 (79.2%)	10 (20.8%)	8 (16.7%)	0.678
Between NGSB and Morpho	I	45	43 (95.6%)	2 (4.4%)	0 (0%)	0.956
	II	47	45 (95.7%)	2 (4.2%)	1 (2.1%)	0.937

S11: Correlation between the ln transformed absolute read numbers per species and the ln transformed absolute estimate of biomass per species for the metabarcoding and mitogenomics datasets.

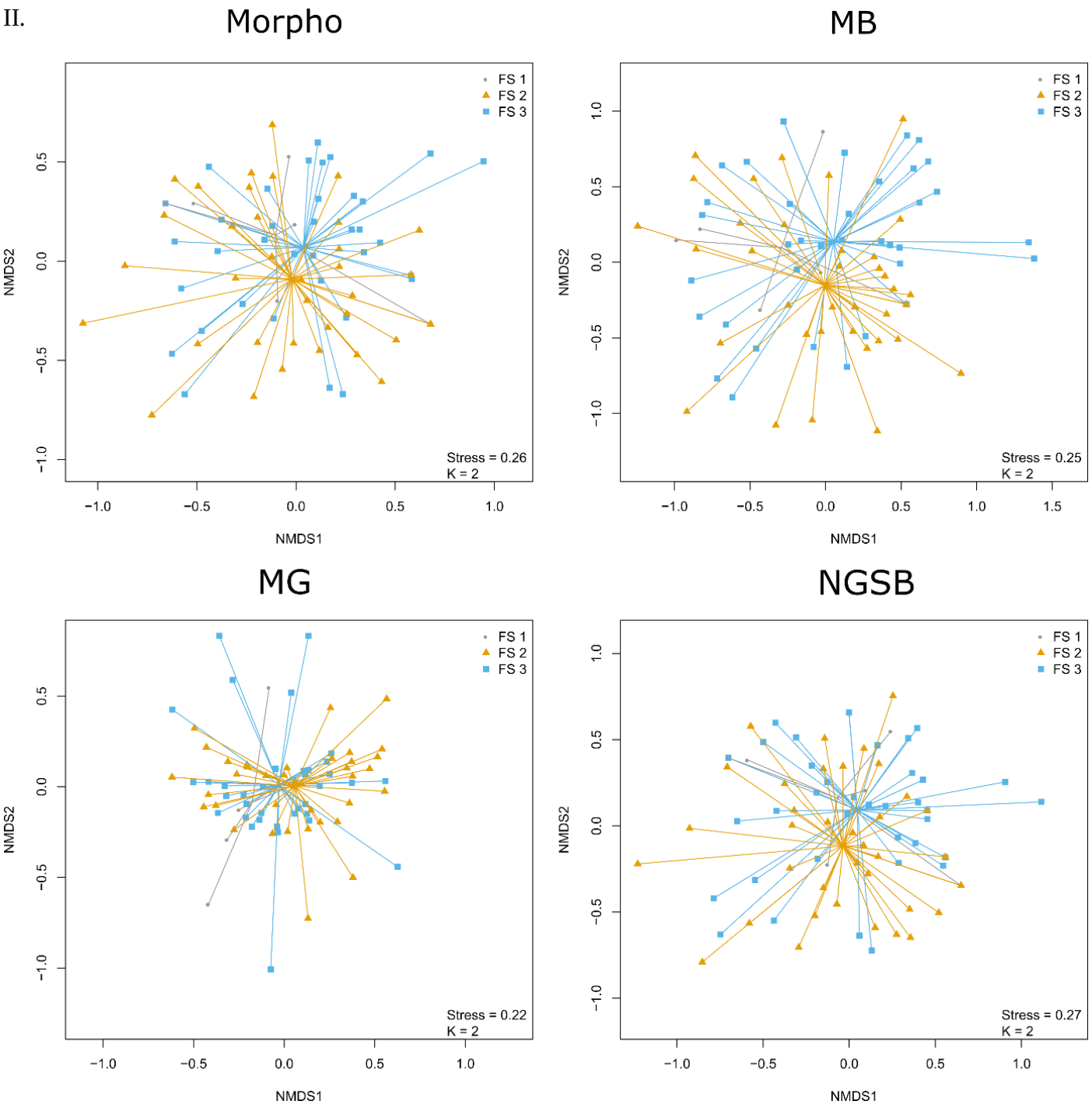


S12: Non-metric multidimensional scaling (NMDS) on (I) occurrence dissimilarity matrices computed with the Jaccard index and (II) absolute abundance using dissimilarity matrices computed with the Bray-Curtis index. The NMDS analyses were performed with the "*metaMDS*" function implemented in the *vegan* package. "Spider" diagrams connect communities sharing the same type of flower strip (FS). Goodness-of-fit between the superimposed shapes of the molecular NMDS plots with the corresponding morphological NMDS plots were assessed using Procrustes tests computed with the "*protest*" function (*vegan* package). Results of the Procrustes tests are given in Table 2.

I.



II.

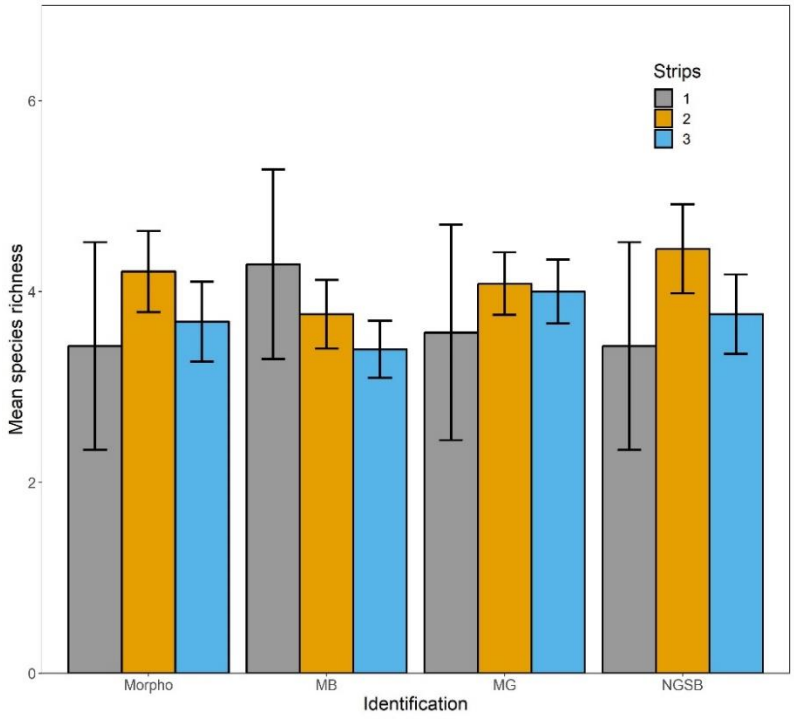

Morpho
MB
MG
NGSB
FS 1
FS 2
FS 3
NMDS1
NMDS2
Stress = 0.26
K = 2
Stress = 0.25
K = 2
Stress = 0.22
K = 2
Stress = 0.27
K = 2

S13: Mean bee (I) species richness and (II) absolute abundance for 3 different types of flowering strips (FS). Means were computed per identification methods and error bars correspond to the mean standard error. Statistical difference between means within each identification method was assessed by generalized linear mixed models (for species richness) or linear mixed models (for relative and absolute abundance). No statistical difference between types of FS was found within method. See manuscript for details on models.

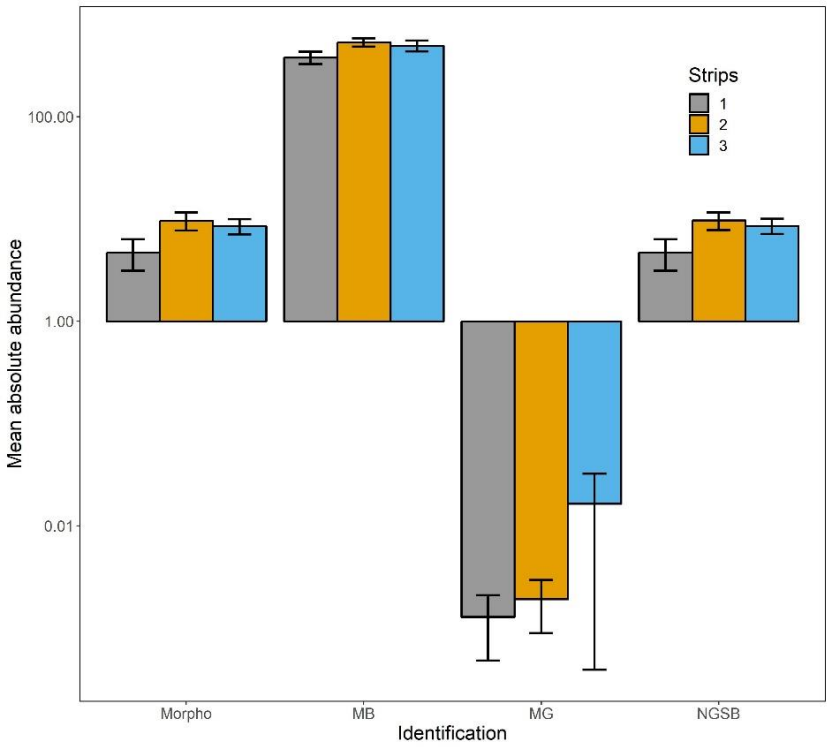
Model details:

Because we were interested in determined and comparing bee species richness and abundance between the different types of FS within each method, the predictor variable was set as the interaction between FS type and identification methods. The transects within sampling sites were used as a random factor. For species richness, we ran a GLMM using the "glmer" function and the poisson family. For abundance data, we ran LMM using the "lmer" function on both absolute and relative abundance data using the same model as above. Relative abundance was computed by dividing the sum of species (Morphological dataset and NGSB) or reads (MB and MG) for each site by the total number of species or reads per identification method. Difference in mean (± SE) species richness or abundance between FS types were graphically displayed using ggplot2 (Wilkinson, 2011).

I.

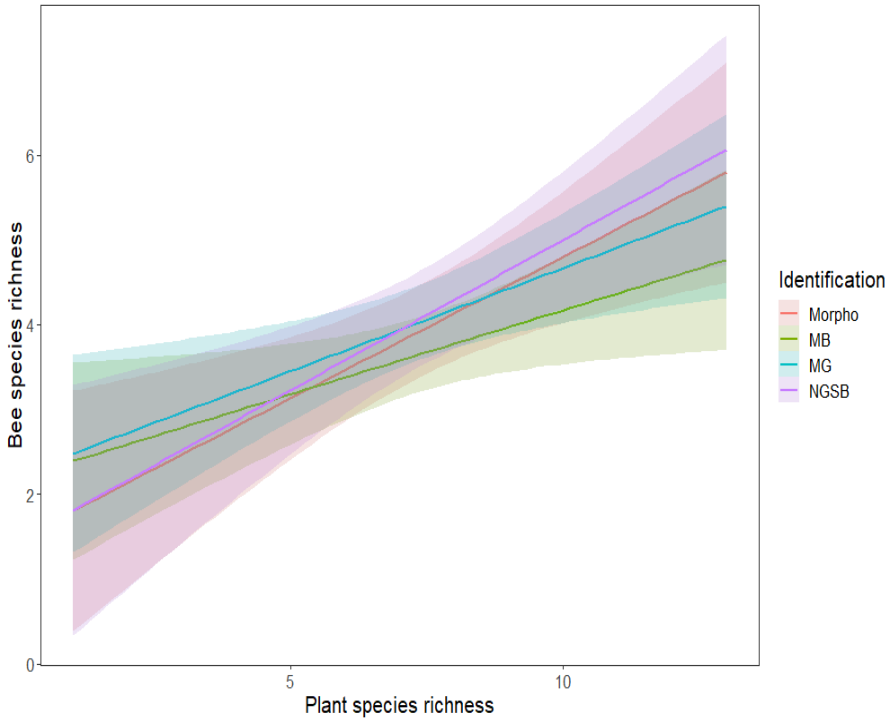


II.

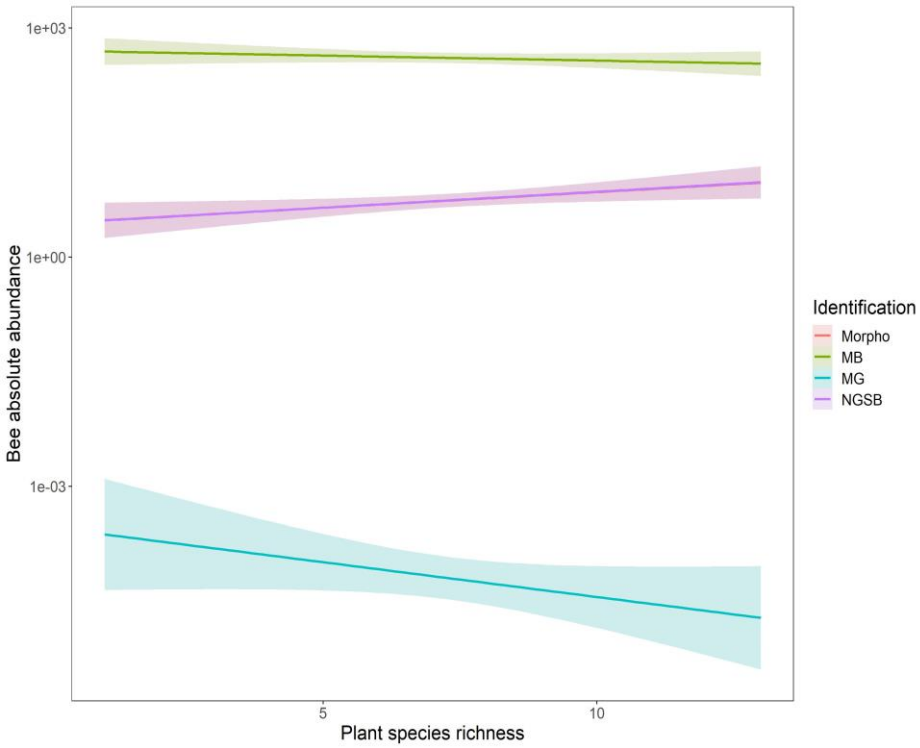

Strips
1
2
3
100.00
1.00
0.01
Mean absolute abundance
Morpho
MB
MG
NGSB
Identification

S14: Relationship between plant species richness and (I) bee species richness or (II) bee absolute abundance different identification methods. Lines were computed by linear regressions as implemented in ggplot2. The grey areas represent the 95% confidence interval. Statistical difference between the relationship of the molecular identification method compared to the morphological identification were assessed by generalized linear mixed models (for species richness) or linear mixed models (for relative and absolute abundance). For bee species richness, no difference in relationship was found between the morphological and molecular identifications. For bee absolute abundance, MB and MG showed significant deviation to the morphological relationship towards plant species richness.

I.



II.



S15: Linear mixed models (LMM) and generalized linear mixed models (GLMM) models and output between the plant species richness (plant_sr) and the bee species richness and species abundance (relative and absolute). Figures in brackets correspond to the standard error.

Bee species richness: *m1 <-glmer(bee_sr ~ plant_sr*Identification + (1|Site/Transect), data=PA_matrix, family = poisson(link = "log"),control=glmerControl(optimizer="bobyqa", optCtrl=list(maxfun=100000)))*

Bee absolute abundance: *m2 <-lmer((log(Abs_Abundance+1)) ~ plant_sr*Identification +(1 |Site/Transect), data= AB_matrix)*

Bee relative abundance: *m3 <-lmer((sqrt(Rel_Abundance)) ~ plant_sr*Identification +(1 |Site/Transect), data= RB_matrix)*

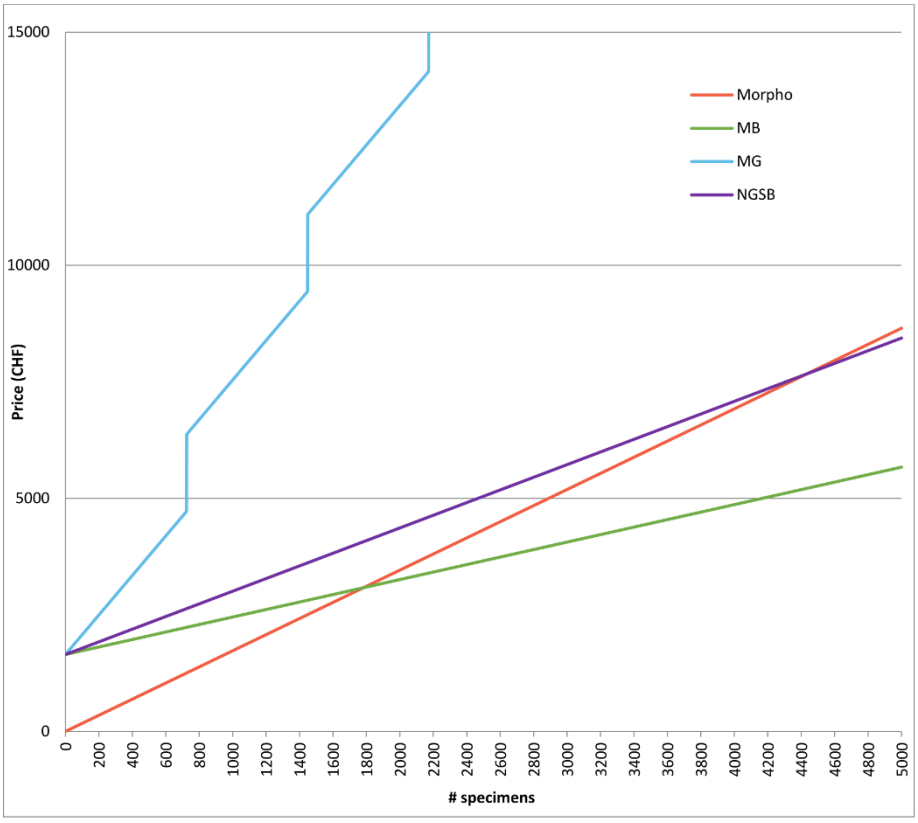
	Dependent variable (estimates ± SE)		
	Presence/Absence	**Log(Absolute abundance+1)**	**sqrt(Relative abundance)**
Plant species richness (plant_sr)	0.084** (0.029)	0.076** (0.028)	0.004* (0.002)
MB	0.167 (0.248)	4.843*** (0.298)	0.045* (0.019)
MG	0.225 (0.247)	-1.386*** (0.301)	-0.033 (0.020)
NGSB	0.018 (0.245)	-0.007 (0.295)	-0.001 (0.019)
plant_sr:MB	-0.030 (0.030)	-0.105** (0.038)	-0.005* (0.002)
plant_sr:MG	-0.024 (0.030)	-0.075 (0.039)	-0.006* (0.003)
plant_sr:NGSB	0.002 (0.03)	0.002 (0.038)	0.0001 (0.002)
Constant	0.633** (0.233)	1.377*** (0.215)	0.067*** (0.014)
Observations	328	327	327
Log Likelihood	-636.048	-336.180	533.372
Akaike Inf.Crit.	1292.096	694.360	-1044.745
Bayesian Inf.Crit.	1330.026	736.050	-1003.055

Note: *p<0.05;**p<0.01;***p<0.001

S16: Cost and workload estimates. All prices are estimated upon suppliers' prices applied for Switzerland in 2018 and are given in Swiss francs (1 CHF ~ 1 USD). Prices are exclusive of additional costs related to disposables (e.g. tips, PCR plates, insect pins and storage boxes). Estimation of total workload are based upon the hand-on time required for all laboratory and bioinformatics analyses. For the morphological identification method, workload does not encompass the taxonomist working hours, which are included into the grand total price.

Method	Grand total (CHF)	Price/Specimen (n = 723)		Price/Community (n = 83)		Workload (hand-on)
		Excl. consumables	Incl. consumables	Excl. consumables	Incl. consumables	
Morpho	1250.-	1.73.-	1.80.-	15.06.-	15.66.-	22:17:00
MB	2237.-	3.09.-	3.55.-	26.95.-	30.99.-	16:12:00
MG	4340.-	5.99.-	6.89.-	52.29.-	60.13.-	21:32:00
NGSB	2645.-	3.65.-	4.20.-	31.86.-	36.64.-	24:32:00

S17: Cost estimations per total number of specimens. Morphological identification is assumed to have a fixed price per specimen. For the three molecular identification methods, price per individual is divided by fixed (extraction, PCR, purification, etc.) and variables costs (sequencing kit). For MG, sequencing depth limit was fixed at 723 specimens and therefore each 723 specimens the cost a new sequencing kit was added. Prices are given in Swiss francs (1CHF ~ 1USD).



S18: Cost details of all laboratory steps required for the library preparation and sequencing of metabarcoding (MB), mitogenomics (MG) and next generation sequencing barcoding (NGSB). Table A depicts details of bulk purchases used to compute the price per reaction. Table B shows the number of reactions used by method and the overall cost of each library step. Grand totals are given at the end of table B. All prices are given in CHF (1 CHF ~ 1 USD) and are based on the supplier's prices for the year 2018.

A.

Laboratory Step	Item	Company	Price (CHF)	Units	Volume per reaction	Price/reaction
DNA extraction	AcroPrep 96 3.0 um	PALL	250.-	5*96 well-plates	NA	0.521.-/ext.
	Proteinase K (Lyophilized)	Promega	116.-	5000 ul	Variable	0.023.-/ul
PCR	Hot StarTaq Master Mix Kit	QIAGEN	1760.-	25 ml (2500U)	10 ul	0.70.-/PCR
	Genomic scale, desalted, dried primers	Microsynth	90.-	1700ul (100uM)	0.08 ul/primer (100 uM)	0.008.-/PCR
Electrophoresis	Agarose NEEO ultra-quality	ROTH	606.6.-	1 kg	2.5 g/96 well-plate	0.016.-/well
	Ethidium Bromide Solution (10 mg/ml)	BIO-RAD	121.-	10 ml	25 ul/96 well-plate	0.003.-/well
PCR clean-up	NucleoFast 96 PCR	Machery-Nagel	2704.-	50*96 well-plates	NA	0.563.-/well
Commercial library-prep	96 TruSeq DNA Nano	Illumina	2828.-	96 samples	NA	29.458.-/sample
Sequencing	Miseq Reagents Kit v3 (600 cycles)	Illumina	1650.-	NA	NA	1650.-/run

B.

Laboratory steps		MB		MG		NGSB	
Item	**Price/reaction**	**# reactions**	**Price**	**# reactions**	**Price**	**# reactions**	**Price**
AcroPrep 96 3.0 um	0.521.-/extraction	249 extr. (83 comm.*3 repl.)	129.7.-	249 extr. (83 comm.*3 repl.)	129.7.-	723 specimens	376.7.-
Proteinase K (Lyophilized)	0.023.-/ul	5000ul	115.-	5000ul	115.-	3765.625 ul (5.2 ul/sample)	86.6.-
Hot StarTaq Master Mix Kit	0.70.-/PCR	415 (83 comm.* 5 repl.)	290.5.-	NA	-	723 PCRs	506.1.-
Genomic scale, desalted, dried primers	0.008.-/PCR	415 (83 comm.* 5 repl.)	3.3.-	NA	-	723 PCRs	5.8.-
Agarose NEEO ultra-quality	0.016.-/well	83 samples	1.3.-	NA	-	723 samples	11.6.-
Ethidium Bromide Solution (10 mg/ml)	0.003.-/well	83 samples	0.2.-	NA	-	723 samples	2.2.-
NucleoFast 96 PCR	0.563.-/well	83 samples	46.8-	NA	-	10 wells (300 ul/well)	5.6.-
96 TruSeq DNA Nano	29.458.-/sample	NA	-	83	2445.-	NA	-
Miseq Reagents Kit v3 (600 cycles)	1650.-/run	1	1650.-	1	1650.-	1	1650.-
Grand total (CHF) without consumables			**2237.-**		**4340.-**		**2645.-**
Grand total (CHF) with consumables[1]			**2572.-**		**4991.-**		**3042.-**

[1] Consumables were estimated to increase cost by 15%

S19: Estimation of cost and sequencing coverages for different Illumina sequencing kits/platforms. The output of each kit is based upon the manufacture's figures. Mean coverage per specimens and community were compute upon the number of mapped reads using a Miseq v3 kit. Prices are given in Swiss francs (1CHF ~ 1USD), including consumables.

Kit	Read length	Output[1]	Price kit[3]	MB				MG				NGSB			
				Sp. coverage	Com. coverage	Overall cost[2]	Cost/ specimen	Sp. coverage	Com. coverage	Overall cost[2]	Cost/ specimen	Sp. coverage	Com. coverage	Overall cost[2]	Cost/ specimen
Miseq v3	2 x 300 bp	13.2-15 Gb	1650.-	5450	47471	2572.-	3.56	4	38	4991.-	6.90	3959	34485	3042.-	4.21
Miseq v2	2 x 250 bp	7.5-8.5 Gb	1259.-	3095	26960	2181	3.02	-	-	-	-	2250	19585	2651	3.67
Miseq v2 Nano	2 x 250 bp	500 Mb	359.-	160	1425	1281	1.77	-	-	-	-	118	1034	1751	2.42
Hiseq 4000	1 x 50 bp	105-125 Gb	817.-	-	-	-	-	32	304	4158	5.75	-	-	-	-
Hiseq 4000	2 x 75 bp	325-375 Gb	1193.-	-	-	-	-	100	950	4534	6.27	-	-	-	-

[1] Claimed output by Illumina.
[2] Cost were inflated by 15% to compensate for consumable prices
[3] Price for Miseq kits are based upon Illumina online shop; Price for Hiseq are based upon prices applied by NYU lagone health sequencing center (for CI member prices) and converted into Swiss francs.

References

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., … Caporaso, J. G. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods, 10*(1), 57–59. doi:10.1038/nmeth.2276

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods, 7*(5), 335–336. doi:10.1038/nmeth.f.303

Folmer, O., BLACK, M., HOEH, W., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*. doi:10.1371/journal.pone.0013102

Frey, J. E., Guillén, L., Frey, B., Samietz, J., Rull, J., & Aluja, M. (2013). Developing diagnostic SNP panels for the identification of true fruit flies (Diptera: Tephritidae) within the limits of COI-based species delimitation. *BMC Evolutionary Biology*. doi:10.1186/1471-2148-13-106

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., … Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, *10*(1), 34. doi:10.1186/1742-9994-10-34

Wilkinson, L. (2011). ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H. *Biometrics*. doi:10.1111/j.1541-0420.2011.01616.

- Chapter II -

Ultraconserved genetic elements uncover cryptic diversity and reveal patterns of mitochondrial-nuclear discordance within bees of the *Andrena-bicolor* complex (Hymenoptera, Andrenidae)

Morgan Gueuning, Juerg E. Frey & Christophe Praz

(Publication in preparation)

Abstract

The rapidly growing number of discovered cryptic species in various groups of organisms challenges traditional species concepts and provides evidence on the universality and ubiquity of this phenomenon. While most cryptic diversity is being discovered in relatively poorly investigated taxa, cases of cryptic species are still being identified in some of the most broadly studied taxa. Because cryptic species cannot be resolved on morphological bases, DNA barcoding is often the first indicator for the presence of cryptic diversity. Recently, DNA barcoding analyses conducted on the controversial *Andrena-bicolor* species complex suggested the presence of at least four species, and additionally revealed the presence of deeply divergent sympatric mitochondrial lineages within two species (i.e. *Andrena bicolor* and *A. amieti*). Using ultraconserved genetic elements (UCEs), we further investigated whether these divergent mitochondrial lineages represented reproductively isolated lineages, or putative cryptic species. We compared a wide range of analytical methods (phylogenetics, structure and admixture analyses, PCA and coalescence-based delimitation analyses) to propose species delimitation hypotheses within this complex. Within *A. bicolor*, the nuclear and mitochondrial patterns were congruent and most analyses suggested reproductive isolation with no admixture between lineages within the investigated region, providing evidence of overlooked cryptic diversity in one of the most widely distributed wild bee species complex in Europe. For *A. amieti*, we found patterns of mitochondrial-nuclear discordance with no reproductive isolation between sympatric mitochondrial lineages. The mitochondrial-nuclear discordance found in *A. amieti* strongly supports the view that both nuclear and mitochondrial markers should be used for correctly assessing cryptic diversity. Overall, our study provides a convincing demonstration of the utility of UCEs for species delimitation.

1. Introduction

In view of the current losses in biodiversity (Sánchez-Bayo & Wyckhuys, 2019), it is an important yet delicate task to adjust estimates of global diversity. Although species richness estimates are still uncertain for numerous taxa, (Costello, 2015; Troudet, Grandcolas, Blin, Vignes-Lebbe, & Legendre, 2017), the accumulation of data for many terrestrial taxa seems to have reach a consensus on species richness estimates (Stork, 2017). However, these numbers are often thought to be underestimates because of an unknown proportion of cryptic species (i.e. group of species classified as a single species due to lack of identifiable morphological features) (Adams, Raadik, Burridge, & Georges, 2014; Adis, 1990; Leasi et al., 2018). The concept of cryptic species has been recognized long ago (Winker, 2006), however this "hidden" diversity has only been unlocked recently with the development of molecular identification tools (Struck et al., 2018). During the last two decades, the number of identified cryptic species has exponentially increased, providing evidence that cryptic diversity is present in most taxa and biomes (Pérez-Ponce de León & Poulin, 2016; Pfenninger & Schwenk, 2007). This cumulative proof of ubiquitous cryptic diversity is raising questions on the actual proportion of cryptic diversity, a critical information towards fostering our comprehension on the evolutionary mechanisms and ecological functionalities of this hidden biodiversity (Bickford et al., 2007; Fišer, Robinson, & Malard, 2018). While assessing the proportion of cryptic diversity in Australian freshwater fishes, a study found 15 times more species than initially identified morphologically (Adams et al., 2014). Although this might be a dramatic example of exceptional levels of cryptic diversity, cryptic species are still being discovered even within thoroughly studied taxa [i.e. European butterflies (Dincă, Lukhtanov, Talavera, & Vila, 2011; Mutanen et al., 2016), beetles (Hendrich et al., 2015) or bees (Praz, Müller, & Genoud, 2019)].

Currently, most studies have identified cryptic species using one or two DNA markers (Fišer et al., 2018; Struck et al., 2018). For animals, very often the standard cytochrome oxidase subunit I (COI) barcode is combined with a nuclear marker (e.g. 18S and 28S rDNA, ITS or EF1). Although this approach has been successfully applied for identifying cryptic species (e.g. Hebert, Penton, Burns, Janzen, & Hallwachs, 2004; Magnacca & Brown, 2012; Murray, Fitzpatrick, Brown, & Paxton, 2008; Williams, Brown, et al., 2012), the usage of the uniparental inherited organelle as marker can be problematic. Indeed, there are numerous examples of mitochondrial-nuclear discordances (reviewed in Toews & Brelsford, 2012) with for instance patterns of deeply divergent sympatric mitochondrial lineages within species (e.g. Hinojosa et al., 2019) or patterns of lack of lineage sorting among related species (e.g. Funk & Omland, 2003; McKay & Zink, 2010). These patterns are often mediated through mtDNA introgression, demographic disparities or sex-biased asymmetries (Toews & Brelsford, 2012). Because mitochondrial markers do not necessarily reflect species history, the assessment of cryptic diversity should be investigated using nuclear, multi-locus approaches. Recently, few studies have used ultraconserved elements (UCEs) for identifying cryptic diversity (Smith, Harvey, Faircloth, Glenn, & Brumfield, 2014; Zarza et al., 2018). Although this targeted-capture based method was initially developed to resolve deep and complicated phylogenies (e.g. Branstetter, Longino, Ward, & Faircloth, 2017; Crawford et al., 2012; Zarza et al., 2018), UCEs can provide information at different evolutionary time scale thanks to an increasing variability from the core region (Faircloth et al., 2012). The use of UCE to uncover species diversity has never been investigated in bees.

In this study we investigated species boundaries and tested cases of overlooked cryptic diversity in a problematic group of bees, the taxa related to *Andrena bicolor* (Gusenleitner & Schwarz, 2002; Scheuchl & Willner, 2016). Species delimitation within this group has long remained controversial, especially in the alpine

region where two species co-occur with *A. bicolor*: *A. allosa* and *A. montana*. In a revision of the Swiss *Andrena* fauna, Amiet et al. (2010) raised doubts on the validity of these two taxa because of the presence of intermediate forms or populations. Phylogenetic analyses on a mitochondrial (COI) and a nuclear gene (LW-rhodopsin) confirmed the validity of both *A. allosa* and *A. montana*, but also revealed the presence of an additional cryptic species undescribed until then [i.e. *A. amieti*; (Praz et al., 2019)]. While solving some issues in this group, these analyses also raised new questions. Indeed, these phylogenetic analyses on the mitochondrial marker also revealed two sympatric clades for *A. bicolor* (also reported in Schmidt, Schmid-Egger, Morinière, Haszprunar, & Hebert, 2015) and two sympatric clades for *A. amieti* (Praz et al., 2019). Genetic distances between these sympatric clades were comparatively high (approximately 3.7% and 2.4% for *A. bicolor* and *A. amieti*, respectively), thus comparable to distances among valid species in this group. In addition, the two sympatric clades within *A. amieti* formed a paraphyletic unit from which a distinct alpine species arose, *A. allosa*. It remains thus unclear if these intra-specific deep mitochondrial divergences unmask further cryptic diversity or are solely the fruit of other evolutionary processes.

To address this question and further unravel the phylogeny of this species-complex, we used UCEs for 91 specimens representing all these lineages in Switzerland. More specifically, we: 1. evaluate species monophyly in all species of the *Andrena bicolor* group; 2. investigate concordance between mitochondrial and nuclear patterns in *A. bicolor* and *A. amieti*; 3. test for additional cryptic diversity within this group.

2. Material and Methods

2.1. Sampling

Bees were initially acquired from a collection composed of specimens collected across Switzerland between 2011 to 2018 within a project aiming at updating the Swiss bee red list. To further extent the dataset, sites known to harbour large populations or several species/clades in sympatry were additionally sampled in spring and summer 2018. Bees collected within the red list surveys were killed in ethyl acetate, pinned and preserved dry. Samples collected in 2018 were preserved in 70% EtOH at 4°C to ensure good DNA preservation. All bees were morphologically identified by one of us (C. Praz).

The cytochrome oxidase subunit I (COI) mitochondrial marker of approximately 300 specimens was sequenced either by Sanger or NGS-barcoding sequencing using the primer pairs LepF/LepR (Hebert et al., 2004) or mlCOIntF/HCO (Leray et al., 2013) (following protocols in Gueuning et al., 2019). The sequences were aligned and a phylogenetic tree was produced (See details of phylogenetic analysis below) and used as support for selecting 96 specimens for the UCE library. The sub-selection was performed based on locality, phylogenetic clustering, and DNA quality.

Our sampling revealed the two clusters previously found within *A. amieti* (*A. amieti* groups 1 and 2 in Praz et al. 2019, referred to as *amieti* lineage I and *amieti* lineage II here). *A. amieti* lineage I is known from Southern Italy (Calabria) -where *A. amieti* lineage II has not been found- and from one site in the Swiss Alps (Kandersteg) where *A. amieti* lineage II predominates (Praz et al., 2019). To increase the number of specimens for *A. amieti* lineage I, 96 samples from Kandersteg were sampled and analysed. Despite this extra sampling effort, only three specimens from this lineage were found. All available specimens from *A. amieti* lineage I were selected for the UCE library.

Within *A. bicolor*, two clades have been reported (Praz et al., 2019; Schmidt et al., 2015), referred to here as lineages I and II (corresponding to *A. bicolor* clades I and II in Praz et al., 2019).

Based on available sequences, both lineages are widely distributed in Europe from Greece to UK. In Switzerland, *A. bicolor* lineage I is restricted to the Valais canton, while lineage II is found throughout the country. We sampled similar numbers of specimen from lineage I (n = 30) and lineage II (n = 28). The divergent mitochondrial lineages within both *A. amieti* and *A. bicolor* are unlikely to represent nuclear inserts of mitochondrial DNA fragments (NUMTs), for the following two reasons. First, these sequences do not contain stop codons and all chromatograms were clean, without double peaks; second, identical and clean sequences have been obtained for the same specimens using universal and specific primers (Praz et al., 2019). Lastly, two specimens collected in Greece and probably representing an undescribed species were included; this species is referred to as *Andrena* sp3, as in Praz et al. (2019). A single specimen of *A. montana* was used to root the phylogenetic trees.

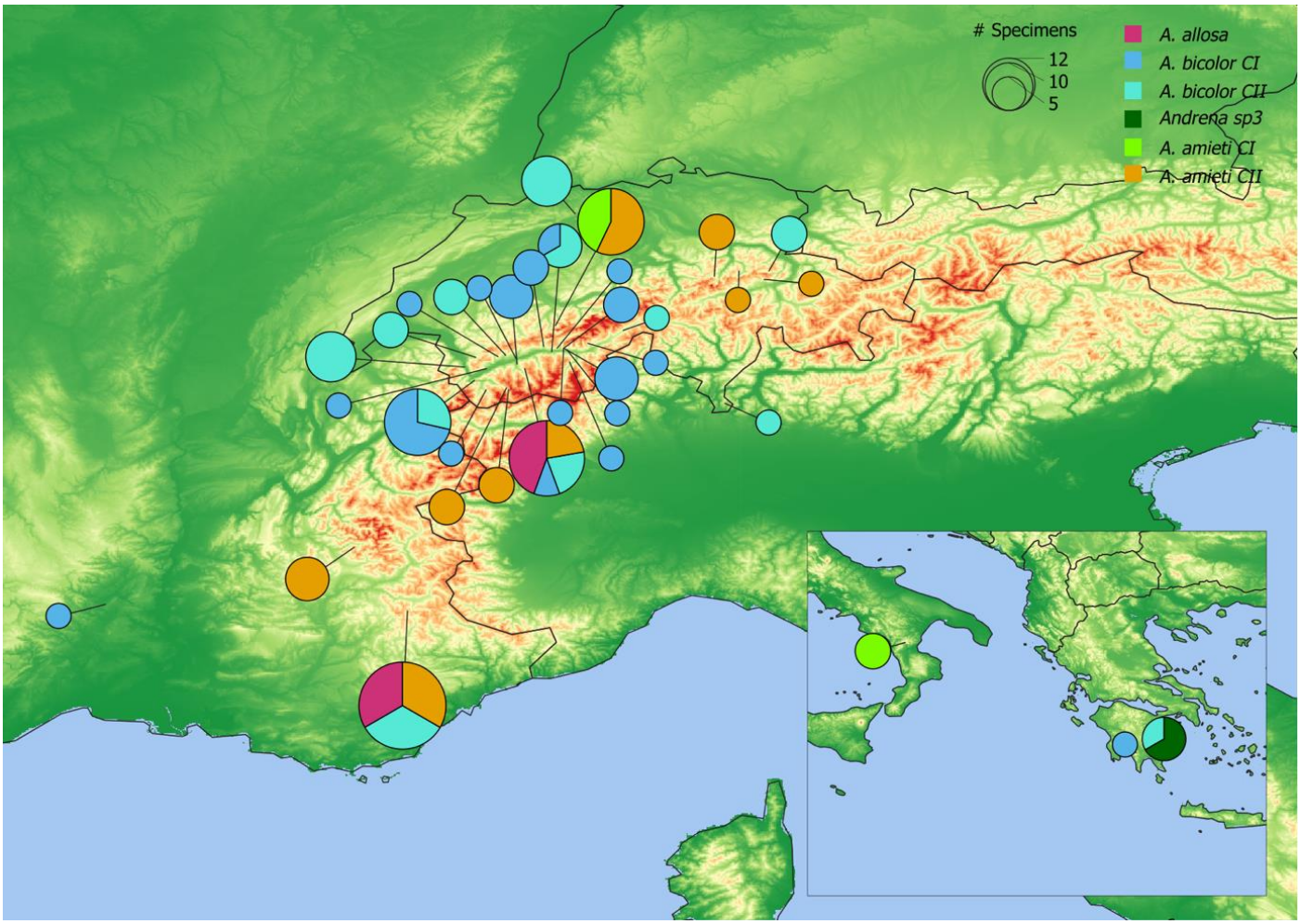


Figure 1: Sampling localities of 91 specimens selected for UCE sequencing and passing bioinformatic filtering. Specimens were morphologically assigned to the species level and assigned to mitochondrial lineage based on the sequencing of the COI marker.

2.2. UCE library preparation and NGS-sequencing

Whole body DNA extractions were performed overnight in a proteinase K buffer at 56°C and purified using a Qiagen Biosprint 96 extraction robot following the manufacturer's protocol. DNA extracts were quantified using Qubit v4 (Thermofisher Scientific) and 50 ng per sample were sheared to approximately 500 bp using a Bioruptor ultrasonicator (Diagenode). Sheared DNA was then processed into an Illumina compatible library using a Kapa Hyper prep kit (Roche). All library preparation steps was carried in ¼ of the manufacture's recommended volumes except for the library amplification PCR for which the recommended volumes were used as described in Branstetter et al. (2017).

Libraries were quantified (Qubit v4) and 41.6 ng per library were pooled 12 by 12. On the 96 samples, we obtained thus a total of 8 equimolar pools (96 specimens /12 pools) at 500 ng. Pools were enriched by hybridization using the Hymenopteran v2 kit (UCE Hymenoptera 2.5Kv2 Principal/Full, myBaits, Arborbiosci) targeting 2500 UCEs (Branstetter et al., 2017). Enrichments were performed using the manufactures v4 protocol with a hybridization step of 24 h at 65°C. Enriched pools were amplified using 14 cycles and after amplification, pools were purified using AMPure XP beads (1.5X). Enrichments were verified using qPCR and pooled in equimolar ratios. The final pool was sequenced using two Miseq v3 kits (2 x 300 bp).

2.3. Bioinformatic processing of UCE data

Demultiplexed data from both runs were merged and processed using *PHYLUCE* tools (Faircloth, 2016). Raw data were cleaned with *illumprocessor* (Faircloth, 2016), a tool wrapped around *trimmomatic* (Bolger, Lohse, & Usadel, 2014). Clean reads were assembled with *velvet* (K = 55) (Zerbino & Birney, 2008). Obtained contigs were mapped against the corresponding UCE reference file using *Lastz* (Harris, 2007) and matching reads were extracted and aligned between specimens using *MAFFT* (Katoh & Standley, 2013). Because taxa are thought to be closely related (< 30-50 MYA), alignments were edge-trimmed (Faircloth, 2016). Locus shared by less than 75% of the maximum number of specimens sharing a locus were filtered out. Remaining alignments were concatenated into a single fasta file. An additional filtering step was applied to remove specimens with more than 90% of missing data.

2.4. COI analyses

To determine which specimens to select for the UCE library, a phylogenetic analysis on over 300 specimens was performed on COI using *FastTree* v2.1.5 (Price, Dehal, & Arkin, 2010) in *Geneious* (v11.0.5) with the GTR model. To facilitate comparison between phylogenetic trees between COI and UCE datasets, additional COI trees were built only with specimens passing the UCE bioinformatic pipeline. One tree for *A. amieti* complex and one for complex *A. bicolor*. These trees were build using *RAxML* v8.2.11 (Stamatakis, 2014) with the GTR gamma model using 100 bootstrap replicates. After bootstrapping, trees harbouring the highest likelihood were retained. Since sequences were obtained using different methods (i.e. Sanger, NGS) and different primers, all sequences were truncated to the same length (n = 269 bp). Additionally, a principal component analysis (PCA) was performed using the *adegenet* package (Jombart, 2008). The three first principal components were plotted using the *scatterplot3d* package (Ligges & Mächler, 2003). The translation and skimming of sequences did not revealed the presence of stop codons.

2.5. UCE analyses

Phylogenetic analyses were conducted on the concatenated alignments for all specimens passing the 90% missing data filtering step. Maximum likelihood trees were produced using *RAxML* v8.2.11 with the same parameters as for the COI *RAxML* trees. Genetic diversity within and between putative species was investigated using multivariate analyses. The analyses were conducted in *R* mainly using the *adegenet* package (Jombart, 2008). Sequences were first imported into *R* using the *"fasta2genlight"* function which reads aligned sequences and extracts binary SNPs before converting files into a genlight object. After conversion, the data was screened for significant departure from Hardy-Weinberg's equilibrium using the *dartR* package (Gruber, Unmack, Berry, & Georges, 2018). As for the COI dataset, a PCA was performed and results were plotted using the three first components. To further identify and describe genetic clusters, a discriminant analysis of principal components (DAPC) was ran. A first approach was used to verify the group's membership using a prior knowledge on the species assignments. For this analysis, *A. bicolor* and *A. amieti* were divided into two distinct groups (e.g. lineage I and II). To identify the optimal number of PCs to retain we used both the plotted cumulative variance of the eigenvalues and a cross-validation method implemented in the *"xvalDapc"* function. Results of posterior membership probabilities for each specimen were plotted using *ggplot2* (Wilkinson, 2011). In a second approach, we ran a DAPC by grouping specimens into genetic clusters without species a priori knowledge. The function *"find.clusters"* was used to determine the optimal number of genetic clusters.

To verify patterns of isolation by distance (IBD) for *A. bicolor* and *A. amieti*, we plotted per species the correlation between genetic and geographical distance using the "*gl.ibd"* function (*dartR* package) with 1000 permutations. Pairwise fixation indexes (Fst) were computed between each putative species (with *A. bicolor* treated as two lineages, and *A. amieti* as one lineage) using the *dartR* package with 10,000 permutations. Levels of observed genetic differences between lineages were tested using analyses of molecular variance (AMOVAs). To account for a potential bias due to geographical distance between populations, hierarchical AMOVAs were performed with populations niched into lineages. Analyses were conducted in R using the *pegas* (Paradis, 2010) implementation of AMOVA in the *poppr* package (Kamvar, Tabima, & Grünwald, 2014). Statistical significance was assessed with 10,000 permutations.

Finally, we used three independent analyses for testing species delimitation. (I) First, we performed a Generalized Mixed Yule Coalescent model (*GMYC*) on an ultrametric tree containing all specimens. Trees were built with *BEAST2* v2.5.2 (Bouckaert et al., 2014) using the JC69 substitution model and a strict molecular clock with a fixed rate of 1.0. The tree priors followed a yule model with a uniform distribution for "birthRate". MCMC ran for 250 million generations with sampling every 1000 generations. Chain convergence was assessed using the software *TRACER* v1.6 (Rambaut, Drummond, Xie, Baele, & Suchard, 2018). For computational purposes, trees were resampled to a total of 154 trees after 20% burn-in using the *logCombiner* software. The ultrametric trees were then imported into *R* using the *ape* package (Paradis & Schliep, 2019). *GMYC* was performed on the first tree using the *splits* package (Ezard, Fujisawa, & Barraclough, 2009). Interval of species number was set between 0 and 10 and the analysis was run using the single-threshold version. (II) Second, results from the first analyses were cross-validated using a Bayesian implementation of the GMYC model (*bGMYC*) (Reid & Carstens, 2012). The MCMC was set to 100,000 generations (8000 generations burnin) and sampled every 100 generations ("thinning"). (III) Third, we performed an analysis on the concatenated sequences using the Bayesian Phylogenetics and phylogeography model (*BPP*) (Yang & Rannala, 2010). *BPP* analyses was ran using the A11 model (unguided species

delimitation analysis) (Yang & Rannala, 2014) on the nexus files (each corresponding to one UCE) obtained after the 75% threshold filtering step. The population file was designed so that specimens were assigned to their species, with *A. bicolor* divided into two distinct lineages and *A. amieti* in one. Alpha and beta parameters of the inverted gamma distribution of the theta prior (average proportion of different sites between two sequences) were set to 3 and 0.004, respectively. For the tau prior, alpha and beta were set to 3 and 0.002. The analyses was run twice with a MCMC of 500,000 generations and a 10% burn-in period.

3. Results

3.1. Dataset after subselection

The UCE sequencing was performed on eight *A. allosa*, six *A. amieti* CI, 22 *A. amieti* CII, two *Andrena sp3*, 30 *A. bicolor* CI, 28 *A. bicolor* CII and one *A. montana*. Although 96 samples (excluding *A. montana* which was only used to root the phylogenetic trees) were sequenced, five specimens did not pass bioinformatic filtering. Therefore, downstream analyses for both COI and UCEs dataset were performed on 91 specimens (excluding *A. montana*). Metadata on sampling locations are given in Supplementary Information (S1).

3.2. UCE library output

Combined, both Miseq runs produced a total of 60.6 million reads with a median read number of 632,073 per specimen (min = 25,700; max = 2.06×10^6). The median number of identified UCEs contigs per specimen was 1174 with a minimum of 181 and a maximum of 1545 contigs. The final matrix with 75% of completeness contained 354 loci, which after concatenation and alignment produced 112,045 bp long reads with on average 27.4% missing data (min = 12.3%; max = 82.4%).

3.3. COI and UCE RAxML trees

The phylogenetic trees performed on the COI and UCEs datasets provided very similar topologies for *A. bicolor* (Figure 2). Both trees showed distinct monophyletic clades with strong bootstrapping support for the UCE tree. One specimen (i.e. "Pleoux1") sampled in Southern France was misplaced in the UCE tree compared to its position in the mitochondrial trees. For this specimen, as well as for two specimens sampled in Greece ("s871", "s876"), there is a strong effect of isolation by distance in the UCE dataset (Figure 1, Supplementary Information S4). For *A. amieti*, all specimens formed one monophyletic clade in the UCE tree, contrasting with the two paraphyletic clades in the COI tree. Furthermore, there was no apparent structuring in the UCEs tree among mitochondrial lineages. Specimens with large amounts of missing data (≥ 65%) exhibited longer branches in the UCE tree (e.g. "Kosmas1", "Pollino1"). There was again significant isolation by distance in the UCE dataset (Supplemental information S4), with specimens from Southern Italy forming a distinct monophyletic clade that was sister to all alpine specimens.

3.4. Multivariate analyses, AMOVA, Fst and IBD

For the COI and the concatenated UCE datasets, *adegenet* identified 37 and 8991 single nucleotide polymorphisms (SNPs) with on average 0.15% and 27.83% missing data, for COI and UCE respectively. No significant departure from Hardy-Weinberg equilibrium was observed in either dataset. Overall, PCAs from COI and UCEs showed highly similar patterns, with *A. bicolor* clearly separated into two clusters (Figure 2). *A. amieti* and *A. allosa* clustered together by the two first components but were separate by the third for UCEs (Figure 2). The two *Andrena sp3* specimens were located between the four clusters.

The plotted cumulative variance of the eigenvalues as well as the "xvalDapc" function suggested to retain the first five principal components (conserving 42.5% of the total variance) for the DAPC. The DAPC with a priori knowledge on the species identification correctly reassigned membership for the majority of specimens. All specimens were correctly reassigned with a 100% membership probability for 3 lineages (i.e. *A. bicolor* lineage II, *Andrena sp3* and *A. allosa*), suggesting no or very low levels of admixture.

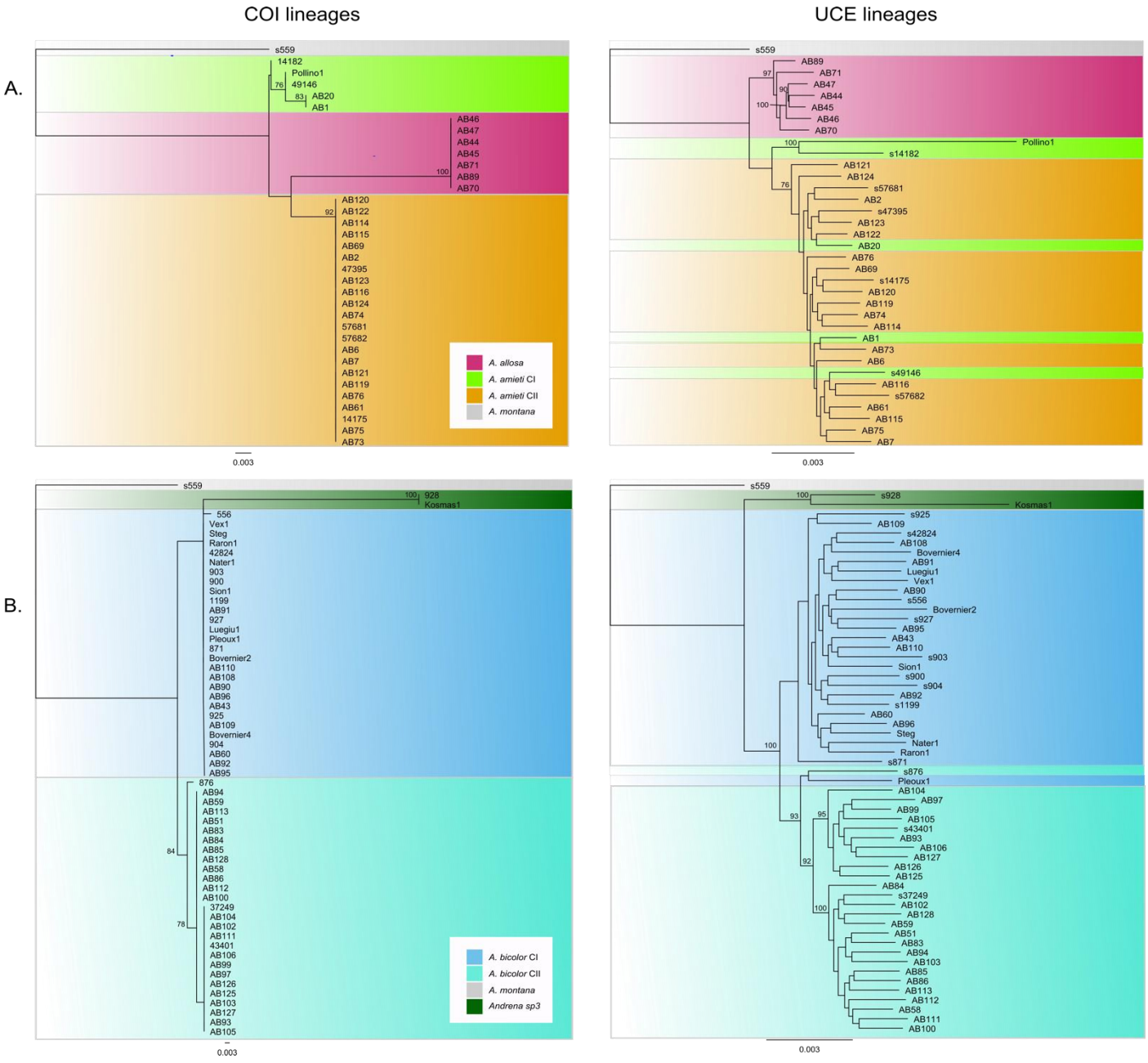


Figure 2: Maximum likelihood phylogenetic trees obtained for COI and the concatenated UCEs dataset. Trees were built in *RAxML* v8.2.11 with the GTR gamma model using 100 bootstrap replicates. For both trees, only bootstrap probability higher than 70% are shown. The colour code used for the COI and UCE tree correspond to the mitochondrial lineages.

The lack of clear cut between mitochondrial lineages for the alpine specimens is suggesting considerable levels of admixture. The optimal cluster number for the DAPC without identification a priori was between 2

and 4, with K = 4 harbouring the lowest BIC (Supplementary Information S2). With K = 4, all specimens were assigned to a single cluster with a posterior membership probability of 1 (Supplementary Information S3). As expected based on morphology and phenology (Praz et al., 2019), *A. amieti* and *A. allosa* formed distinct clades. For *A. bicolor*, all specimens beside "Pleoux1" were correctly assigned to their mitochondrial lineage. The two Greek specimens affiliated to *Andrena sp3* were grouped within *A. bicolor* lineage I, and grouped in their own separate lineage for K = 5 (Supplementary Information S3).

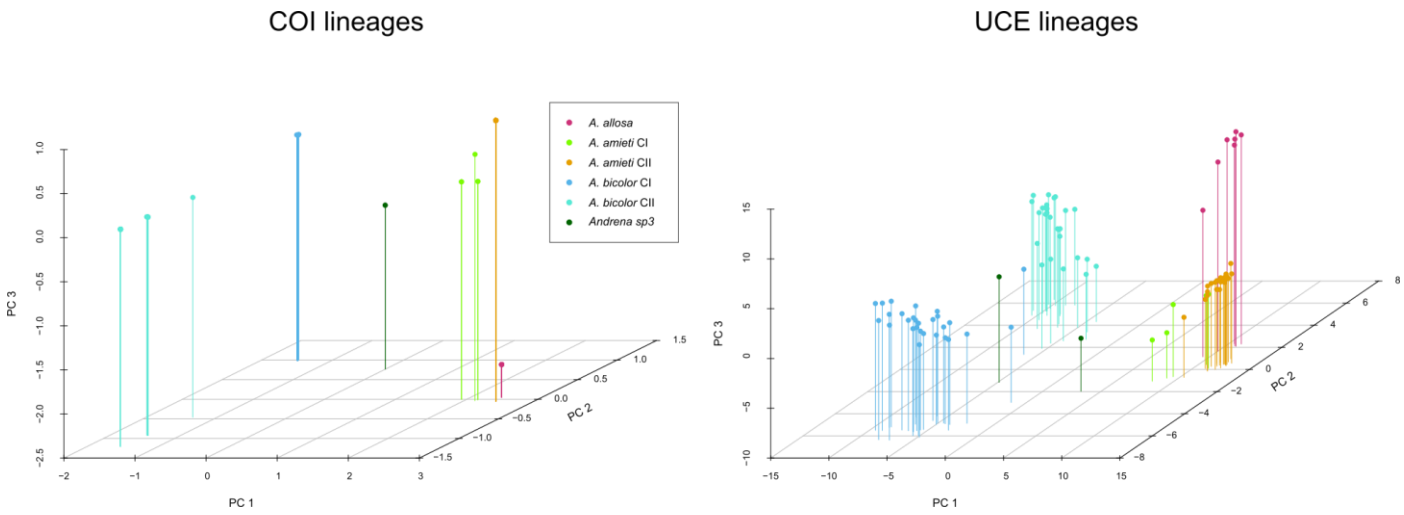


Figure 3: Principal component analyses performed on COI and UCE datasets. Both datasets harbour the same specimens however the COI dataset contains several identical sequences and therefore the same specimens fall into the same space.

The AMOVA (Table 1) depicted a moderate level (25.5%) of genetic diversity between the two *A. bicolor* lineages but no significant difference between the two *A. amieti* lineages. For the later, there was however slight difference (6.77%) between geographically isolated populations (i.e., South Italy and Alps) which is not surprising since we detected strong effects of IBD (Supplementary Information S4). The lowest, yet significant fixation index (Table 2) was obtained between both lineages of *A. bicolor* (Fst = 0.22), followed by *A. amieti* / *A. allosa* (Fst = 0.32) and *A. bicolor* lineage I / *Andrena sp3* (Fst = 0.33). The fixation index computed between clades of *A. amieti* was not significant, yet negative which resulted from a sampling bias between *A. amieti* CI (n = 5) and CII (n = 22).

3.5. Species delimitation

GMYC analysis computed on all specimens identified 4 clusters (Figure 4A): two clusters corresponding two the mitochondrial lineages found within *A. bicolor*, one cluster with *Andrena sp3* and one cluster with *A. amieti* and *A. allosa* merged together. The bGMYC analyses identified 5 clades with high posterior probabilities (p=0.95-1; Figure 4B). Clade delimitation between GMYC and bGMYC was identical except for *Andrena sp3* which was split into two separate clades, one for each specimen. The two parallel *BPP* analyses converged and were highly congruent. Both runs depicted: (i) one tree model [((*A. allosa*, *A. amieti* CI+CII), ((*A. bicolor* CI, *A. bicolor* CII), *Andrena sp3*))] with a posterior probability of 1; (ii) 5 delimited species (i.e. *Andrena sp3*, *A. bicolor* CI, *A. bicolor* CII, *A. allosa*, *A. amieti* CI +CII), all with a posterior probability of 1; (iii) and a posterior probability of 1 for having 5 species present in the dataset. Results of both *BPP* runs are given in Supplementary Tables S3.

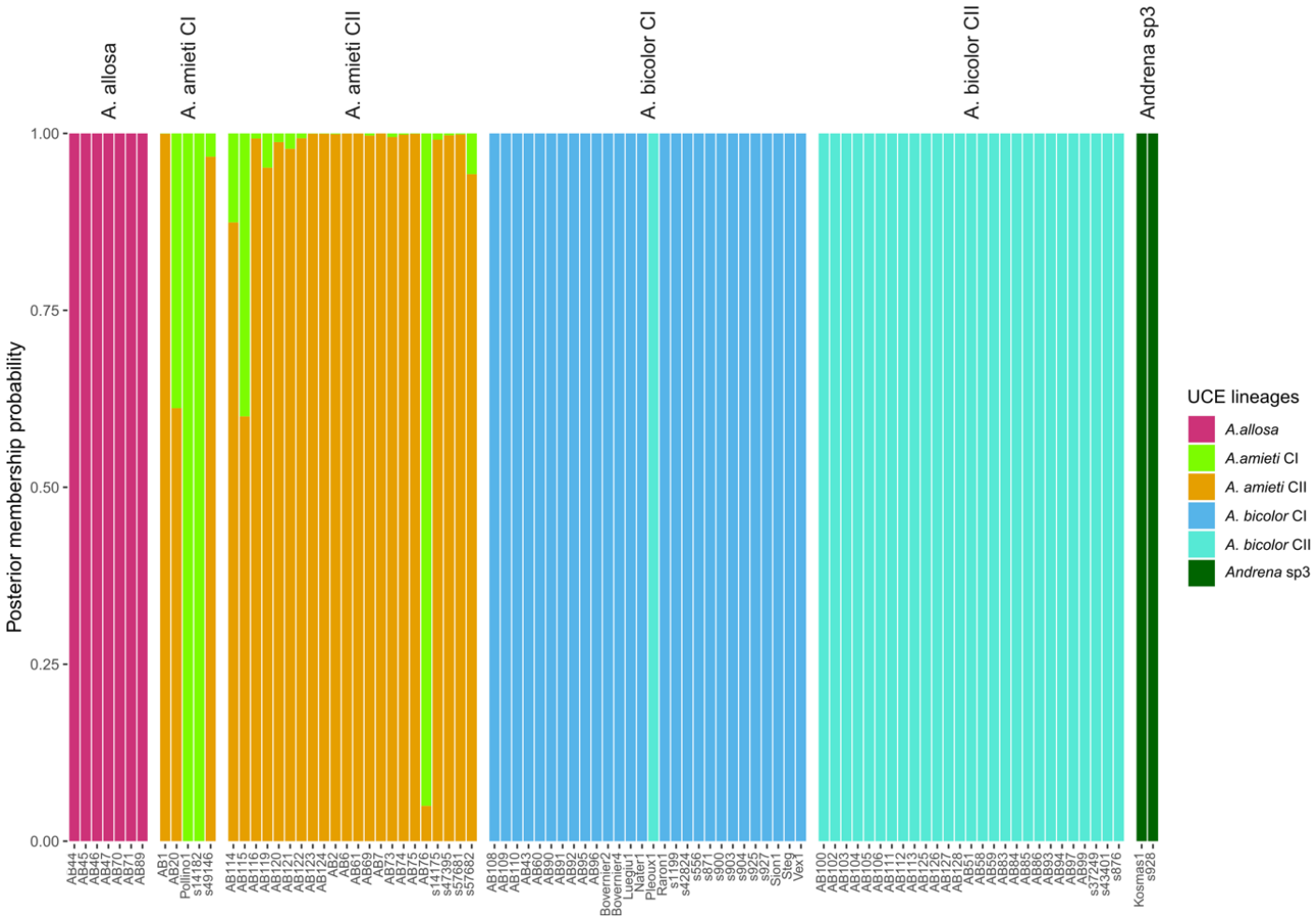


Figure 4: Discriminant analysis of principal components (DAPC) performed on the UCE concatenated dataset. Specimens are grouped by mitochondrial lineages with the corresponding group affiliation above each block and colours depict the nuclear lineages. For *A. amieti*, there is no clear clustering at the nuclear level between individuals from both mitochondrial lineages. Only two specimens (i.e. from southern Italy; "Pollino1", "s14182") from *A. amieti* clade I shown no sign of admixture with lineages II. For *A. bicolor*, all specimens beside "Pleoux1" were assigned with a 100% probability to the correct species/lineage, suggesting no or very low levels of admixture.

	Variance source	df	Sum of squares	Variance	% variance	P-value
A. amieti	**Between lineages**	1	355.76	19.14	9.74%	0.126
	Between pop within lineages	**9**	**1759.39**	**13.29**	**6.77%**	**0.001**
	Within Samples	16	2624.21	164.01	83.49%	
	Total	26	4739.36	196.45	100%	
A. bicolor	**Between lineages**	**1**	**2091.29**	**68.23**	**25.50%**	**<0.0001**
	Between pop within lineages	28	5815.32	10.61	3.96%	0.129
	Within Samples	25	4719.20	188.77	70.54%	
	Total	54	12625.80	267.61	100%	

Table 1: Analysis of molecular variance (AMOVA) between the phylogenetic clades observed using the COI dataset. The AMOVA was performed on the UCEs dataset using the poppr and pegas R packages. Significance was assessed through 10,000 permutations.

	A. allosa	***A. amieti* CI**	***A. amieti* CII**	***A. bicolor* CI**	***A. bicolor* CII**
A. allosa	-	-	-	-	-
***A. amieti* CI**	0.14*	-	-	-	-
***A. amieti* CII**	0.35*	**-0.25**	-	-	-
***A. bicolor* CI**	0.60*	0.52*	0.63*	-	-
***A. bicolor* CII**	0.66*	0.60*	0.67*	0.22*	-
***Andrena* sp3**	0.66*	0.51*	0.71*	0.33*	0.48*

Table 2: Pairwise Fst between all putative species were computed on the UCEs dataset. Statistical significance ($p \leq 0.05$) were assessed through 10,000 bootstraps and are depicted by asterisks (*). Only the Fst computed between both mitochondrial lineages of *A. amieti* was not significant (highlighted in bold). Tests were run in R using the dartR package. The negative Fst value found between both *A. amieti* CI and CII results from a sampling bias between both mitochondrial lineages (5 vs 22 specimens).

3.5. Species delimitation

GMYC analysis computed on all specimens identified 4 clusters (Figure 4A): two clusters corresponding two the mitochondrial lineages found within *A. bicolor*, one cluster with *Andrena sp3* and one cluster with *A. amieti* and *A. allosa* merged together. The bGMYC analyses identified 5 clades with high posterior probabilities (p=0.95-1; Figure 4B). Clade delimitation between GMYC and bGMYC was identical except for *Andrena sp3* which was split into two separate clades, one for each specimen. The two parallel *BPP* analyses converged and were highly congruent. Both runs depicted: (i) one tree model [((*A. allosa*, *A. amieti* CI+CII), ((*A. bicolor* CI, *A. bicolor* CII), *Andrena sp3*))] with a posterior probability of 1; (ii) 5 delimited species (i.e. *Andrena sp3*, *A. bicolor* CI, *A. bicolor* CII, *A. allosa*, *A. amieti* CI +CII), all with a posterior probability of 1; (iii) and a posterior probability of 1 for having 5 species present in the dataset. Results of both *BPP* runs are given in Supplementary Tables S3.

4. Discussion

In this study, we investigated the presence of cryptic diversity within a controversial group of closely related bee species allied with the widespread species *Andrena bicolor*. Comparing the relationships inferred using mitochondrial and multi-locus nuclear DNA sequences reveals that the deeply divergent mitochondrial lineages found within *A. bicolor* and *A. amieti* likely represents different evolutionary events.

4.1. Cryptic species uncovered within A. bicolor

The presence of two distinct mitochondrial lineages in the widespread *A. bicolor* was first described by Schmidt and colleagues (2015). The relatively high genetic distance between lineages, corresponding to values observed among distinct species, raised suspicion on the presence of cryptic diversity within *A. bicolor*. The same two mitochondrial lineages were confirmed later by Praz et al. (2019), although the status of these lineages remained unresolved with the analyses of one nuclear gene.

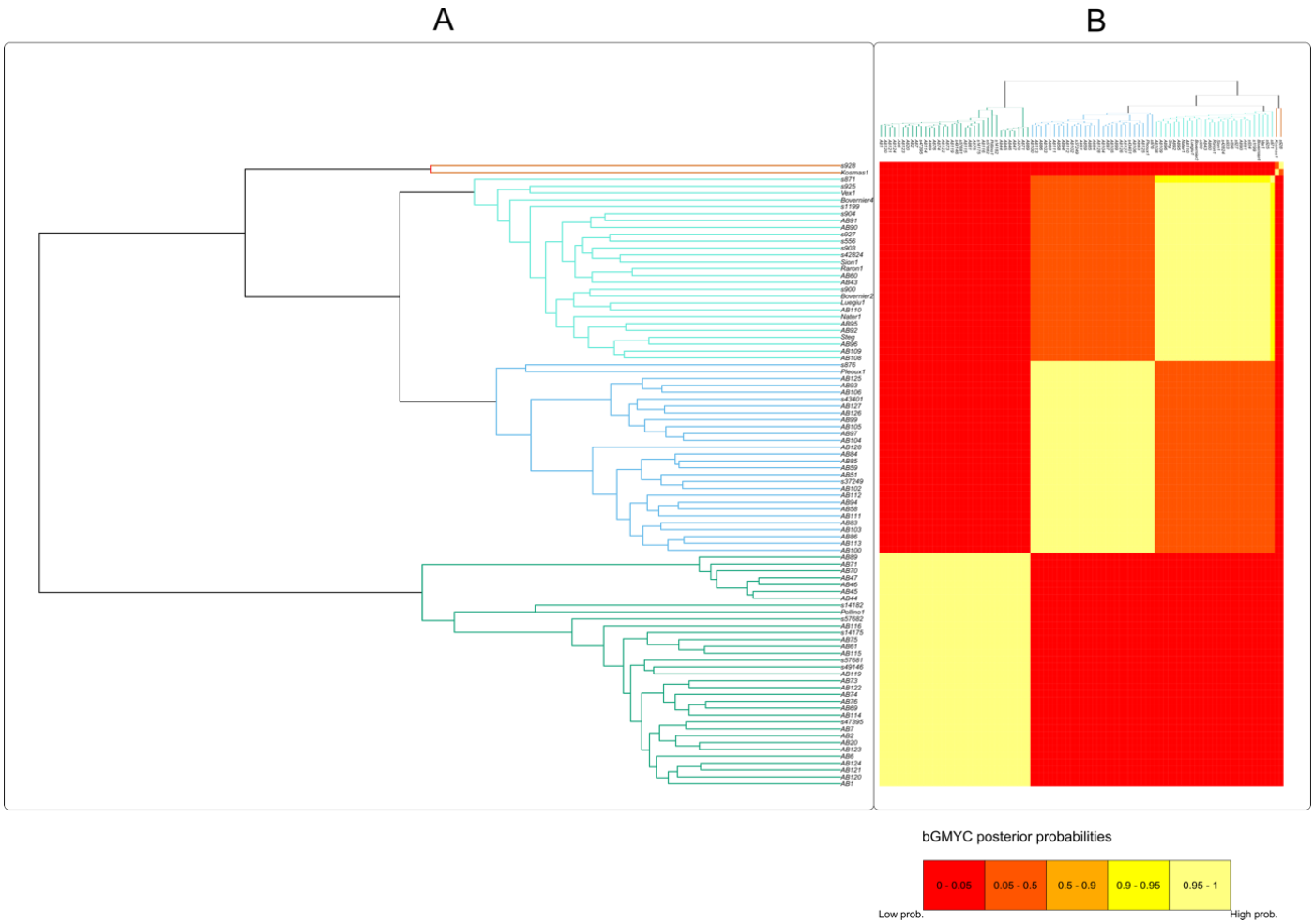


Figure 5: Results from Generalized Mixed Yule Coalescent model (*GMYC*) analyses for species delimitation. (A) Ultra-metric tree with colour code for the maximum likelihood clusters identified by the *GMYC* analysis. (B) The head map of conspecificity probabilities between specimens with associated probabilities depicted in legend under the map. Clear-yellow clusters harbour the highest posterior probabilities (prob. = 0.95-1). *bGMYC* was run with 100,000 generations (8000 generations burnin) and sampled every 100 generations.

While enriching for hundreds of nuclear homologues loci using UCEs, we found strong evidence for the presence of two cryptic sympatric lineages within *A. bicolor* in Switzerland. Indeed, our results showed (I) quasi-identical clades composition between the mitochondrial and nuclear lineages, with one single exception, a female from southern France (i.e. "Pleoux1"), (II) clear separation between both clades with little or no admixture; (III) statistical support for distinct lineages using three different species delimitation tests. With the exception of that specimen from southern France, there is little doubt that two distinct cryptic lineages coexist in Switzerland. Moreover, our sampling included one Greek specimen from each lineage, and although our data suggest some isolation by distance, both specimens clustered with the same clade in both mitochondrial and nuclear trees. Taken together, these results suggest that both lineages coexist in sympatry over 2000 km with little or no gene flow.

The pattern of mitochondrial-nuclear discordance observed for the lone specimen of southern France is difficult to explain with the available data. This specimen may alternatively be a hybrid between both lineages,

or may originate from a population showing some degree of introgression. It is unclear how hybrids would be detected with a UCE dataset. The specimen was a female and thus diploid; since polymorphisms were kept in the analysis, we feel that our UCE results clearly reject the hypothesis that this specimen was a first generation hybrid between both lineages. In addition, there was no clear sign of nuclear introgression since the specimens clearly cluster with one lineage in the PCA or in the phylogenetic trees, and since the DAPC assigned with 100% the specimen to a cluster. Possibly, this specimen represents a case of mitochondrial introgression with limited introgression in the nuclear genomes. Hybridization can be asymmetrical between populations and geographical areas (Bisconti, Porretta, Arduino, Nascetti, & Canestrelli, 2018; Johnson, White, Phillips, & Zamudio, 2015; Mastrantonio, Porretta, Urbanelli, Crasta, & Nascetti, 2016; Wielstra & Arntzen, 2012). Moreover, a few studies provided evidence of differential introgression rates between mitochondrial and nuclear genomes with cases of complete mitogenomes replacement without clear evidence of nuclear introgression (Good, Vanderpool, Keeble, & Bi, 2015; Pons, Sonsthagen, Dove, & Crochet, 2014; Zieliński et al., 2013). A more thorough sampling across the distribution area of *A. bicolor* would be required to draw a clear distribution map of the mitochondrial lineages and increase our sampling for UCE sequencing. A clear advantage of UCEs over other methods (e. g., RADseq or ddRADseq) is that our dataset can easily be complemented with additional specimens.

Cryptic diversity can be the outcome of different mechanisms, such as recent divergence, parallelism, convergence or stasis (Struck et al., 2018). Based on the phylogenetic and the ultra-metric trees (Figure 2, 5) both lineages from *A. bicolor* are closely related and seem to have diverged recently. This recent divergence time may explain the lack of morphological differentiation in the *Andrena bicolor* group, which is morphologically uniform with no conspicuous characters, e.g., in the male genitalias (Praz et al., 2019). In many cases, however, cryptic species harbour some differences in their ecology, for example in phenology, thermal specialisation and plant foraging, as documented in cryptic bumblebees of the *Bombus lucorum*-group (Scriven, Whitehorn, Goulson, & Tinsley, 2016). Exploiting slightly different ecological niches may allow closely related species to partition resources and therefore avoid competitive exclusion (Brunner & Frey, 2010; Fišer et al., 2018). So far, little is known on the ecology of both lineages revealed within *Andrena bicolor*. Both are bivoltine with a spring and a summer generation; based on our field observations, the first generation is polylectic while the second shows a preference for Campanulaceae in both lineages, as in the closely related *Andrena amieti*. More data are needed to shed light on the ecology of these two lineages in Europe.

4.2. Two mitochondrial barcodes, mitochondrial paraphyly, but one species in A. amieti

For *A. amieti*, we found mitochondrial and nuclear discordance with: (I) two mitochondrial lineages forming a paraphyletic unit with respect to another species, *A. allosa*; and (II) one monophyletic nuclear lineage. The mitochondrial lineages were composed of one "rare" mitochondrial lineage (*A. amieti* CI) with only five specimens and one larger mitochondrial lineage (*A. amieti* CII) including 22 specimens. Specimens from the "rare" lineage were sampled from two populations located in the Swiss Alps ("Kandersteg", n = 3) and in southern Italy ("Mt Pollino", n = 2). From the latter location, two additional specimens have been analysed by Praz et al. (2019), and they also belonged to the rare lineage. Taken together, barcode data for four specimens suggest that this rare mitochondrial lineage is likely dominant in southern Italy. For the alpine region, the three specimens belonging to the "rare" mitochondrial lineage represent only a small proportion (~ 3%) of all specimens collected (n = 96) in Kandersteg, and were sampled from one locality with numerous other specimens belonging to *A. amieti* CII. In phylogenetic tree based on the UCEs, specimens collected in southern Italy formed a well-supported monophyletic clade sister to all other specimens. In striking contrast,

the three alpine specimens from the rare *A. amieti* CI were distributed in a clade containing all other alpine specimens. Based on the discordance between the mitochondrial and nuclear lineages and the lack of relatedness in nuclear phylogenies between alpine specimens of lineage I, the existence of an isolated, cryptic lineage within *A. amieti* can be discarded.

Mitochondrial-nuclear discordances are not rare and are often linked to incomplete lineage sorting, mitochondrial introgressions, demographic disparities, *Wolbachia* infections or sex-biased asymmetries (i.e. male-biased dispersal, mating behaviour or sex-biased offspring production) (Toews & Brelsford, 2012). Interestingly, the mitochondrial-nuclear discordance was also accompanied by paraphyly between the two mitochondrial lineages. Species-level paraphyly and/or polyphyly with DNA barcoding can also be triggered by incomplete lineage sorting, introgression, or hybridization (Funk & Omland, 2003). Most often these events occur in recently diverged species and are not necessarily mutually exclusive (Mutanen et al., 2016). In this study, the low number of specimens collected in southern Italy renders the investigation on the underlying mechanism tedious. A more complete sampling across the entire distribution of *A. amieti* would be necessary to separate incomplete lineage sorting from the other mechanisms. In most cases, incomplete lineage is not associated with any biogeographical pattern (Funk & Omland, 2003; Toews & Brelsford, 2012). Rather, mitochondrial lineages are expected to be homogeneously distributed across nuclear lineages in the case of incomplete lineage sorting. In contrast, events such as hybridization/introgression often leave biogeographical footprints because they are unidirectional, which implies that the gene flow is directed from the native taxon towards the colonized taxon (Currat, Ruedi, Petit, & Excoffier, 2008; Nevado, Fazalova, Backeljau, Hanssens, & Verheyen, 2011; Pons et al., 2014). Therefore, introgression levels are highest at the hybridization zone and fade away over the colonized distribution zone (Toews & Brelsford, 2012). A more throughout sampling of *A. amieti* in Southern Italy, or between Southern Italy and the Alps, could potential provide an answer and rule out the incomplete lineage sorting theory if no specimens harbouring the "Alpine" mitochondrial allele are found.

4.3. UCEs for detecting cryptic diversity

The development of DNA sequencing methods has been a critical step towards the discovery of cryptic diversity. With the current increase in published sequences, DNA barcoding is an extremely useful tool for exploring and exposing cryptic diversity (Janzen et al., 2017). Nevertheless, as highlighted in this study, discordance between mitochondrial and nuclear DNA can be misleading. Therefore, delimitation of species should not be assessed only with DNA barcodes but rather with multi-locus nuclear markers. UCEs were initially developed and used to provide deep phylogenetic signals, and only few studies used UCEs for narrow evolutionary timescales. The major advantages of UCEs over other methods such as RADseq is that there is an overlap in orthologues loci between datasets of different studies and therefore datasets can theoretically be compiled and used for instance to enrich the tree of life. Also, library preparation of UCE is relatively straightforward and does not require high quality samples. With over 10 different bait sets currently available -covering major tetrapod, fish and insect taxa- UCEs can be applied to a large variety of taxa without requiring considerable changes in laboratory protocols. One of the concerns when using UCE is the presence of paralogous loci. Although UCEs exhibit low rates of paralogy (Derti et al., 2006), it is hard to exclude this possibility completely. During the bioinformatics process of UCE reads, paralogs are theoretically removed by aligning the assembled contigs to the target enrichment baits (Faircloth, 2016). The presence of paralogous loci will thus majorly depend on how bait sets were designed (Gustafson et al., 2019).

References

Adams, M., Raadik, T. A., Burridge, C. P., & Georges, A. (2014). Global Biodiversity Assessment and Hyper-Cryptic Species Complexes: More Than One Species of Elephant in the Room? *Systematic Biology*, *63*(4), 518–533. doi:10.1093/sysbio/syu017

Adis, J. (1990). Thirty million arthropod species – too many or too few? *Journal of Tropical Ecology*, *6*(1), 115–118. doi:10.1017/S0266467400004107

Bickford, D., Lohman, D. J., Sodhi, N. S., Ng, P. K. L., Meier, R., Winker, K., … Das, I. (2007). Cryptic species as a window on diversity and conservation. *Trends in Ecology and Evolution*, *22*(3), 148–155. doi:10.1016/j.tree.2006.11.004

Bisconti, R., Porretta, D., Arduino, P., Nascetti, G., & Canestrelli, D. (2018). Hybridization and extensive mitochondrial introgression among fire salamanders in peninsular Italy. *Scientific Reports*, *8*(1), 1–10. doi:10.1038/s41598-018-31535-x

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. doi:10.1093/bioinformatics/btu170

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

Branstetter, M. G., Longino, J. T., Ward, P. S., & Faircloth, B. C. (2017). Enriching the ant tree of life: enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. *Methods in Ecology and Evolution*, *8*(6), 768–776. doi:10.1111/2041-210X.12742

Brunner, P. C., & Frey, J. E. (2010). Habitat-specific population structure in native western flower thrips Frankliniella occidentalis (Insecta, Thysanoptera). *Journal of Evolutionary Biology*, *23*(4), 797–804. doi:10.1111/j.1420-9101.2010.01946.x

Costello, M. J. (2015). Biodiversity: The Known, Unknown, and Rates of Extinction. *Current Biology*, *25*(9), R368–R371. doi:10.1016/j.cub.2015.03.051

Crawford, N. G., Faircloth, B. C., McCormack, J. E., Brumfield, R. T., Winker, K., & Glenn, T. C. (2012). More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology Letters*, *8*(5), 783–786. doi:10.1098/rsbl.2012.0331

Currat, M., Ruedi, M., Petit, R. J., & Excoffier, L. (2008). The hidden side of invasions: Massive introgression by local genes. *Evolution*, *62*(8), 1908–1920. doi:10.1111/j.1558-5646.2008.00413.x

Dincă, V., Lukhtanov, V. A., Talavera, G., & Vila, R. (2011). Unexpected layers of cryptic diversity in wood white Leptidea butterflies. *Nature Communications*, *2*(1). doi:10.1038/ncomms1329

Ezard, T., Fujisawa, T., & Barraclough, T. G. (2009). Splits: species' limits by threshold statistics. *R Package Version*.

Faircloth, B. C. (2016). PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*, *32*(5), 786–788. doi:10.1093/bioinformatics/btv646

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, *61*(5), 717–26. doi:10.1093/sysbio/sys004

Fišer, C., Robinson, C. T., & Malard, F. (2018). Cryptic species as a window into the paradigm shift of the species concept. *Molecular Ecology*, *27*(3), 613–635. doi:10.1111/mec.14486

Funk, D. J., & Omland, K. E. (2003). Species-Level Paraphyly and Polyphyly: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics*, *34*(1), 397–423. doi:10.1146/annurev.ecolsys.34.011802.132421

Good, J. M., Vanderpool, D., Keeble, S., & Bi, K. (2015). Negligible nuclear introgression despite complete mitochondrial capture between two species of chipmunks. *Evolution*, *69*(8), 1961–1972. doi:10.1111/evo.12712

Gruber, B., Unmack, P. J., Berry, O. F., & Georges, A. (2018). dartr: An r package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Molecular Ecology Resources*. doi:10.1111/1755-0998.12745

Gueuning, M., Ganser, D., Blaser, S., Albrecht, M., Knop, E., Praz, C., & Frey, J. E. (2019). Evaluating next-generation sequencing (NGS) methods for routine monitoring of wild bees: Metabarcoding, mitogenomics or NGS barcoding. *Molecular Ecology Resources*, *19*(4), 847–862. doi:10.1111/1755-0998.13013

Gusenleitner, F., & Schwarz, M. (2002). Weltweite Checkliste der Bienengattung Andrena mit Bemerkungen und Ergänzungen zu paläarktischen Arten (Hymenoptera, Apidae, Andreninae, Andrena). *Entomofauna Supplement*, *12*, 1–1280.

Gustafson, G. T., Alexander, A., Sproul, J. S., Pflug, J. M., Maddison, D. R., & Short, A. E. Z. (2019). Ultraconserved element (UCE) probe set design: Base genome and initial design parameters critical for optimization. *Ecology and Evolution*, *9*(12), 6933–6948. doi:10.1002/ece3.5260

Harris, R. S. (2007). *Improved Pairwise Alignment of Genomic DNA. PhD thesis, Pennsylvania State University. PhD thesis, Pennsylvania State Univ. (2007)*. doi:/10.1016/j.brainres.2008.03.070

Hebert, P. D. N., Penton, E. H., Burns, J. M., Janzen, D. H., & Hallwachs, W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(41), 14812–14817. doi:10.1073/pnas.0406166101

Hendrich, L., Morinière, J., Haszprunar, G., Hebert, P. D. N., Hausmann, A., Köhler, F., & Balke, M. (2015). A comprehensive DNA barcode database for Central European beetles with a focus on Germany: Adding more than 3500 identified species to BOLD. *Molecular Ecology Resources*, *15*(4), 795–818. doi:10.1111/1755-0998.12354

Hinojosa, J. C., Koubínová, D., Szenteczki, M. A., Pitteloud, C., Dincă, V., Alvarez, N., & Vila, R. (2019). A mirage of cryptic species: Genomics uncover striking mitonuclear discordance in the butterfly Thymelicus sylvestris. *Molecular Ecology*, mec.15153. doi:10.1111/mec.15153

Janzen, D. H., Burns, J. M., Cong, Q., Hallwachs, W., Dapkey, T., Manjunath, R., … Grishin, N. V. (2017). Nuclear genomes distinguish cryptic species suggested by their DNA barcodes and ecology. *Proceedings of the National Academy of Sciences*, *114*(31), 8313–8318. doi:10.1073/pnas.1621504114

Johnson, B. B., White, T. A., Phillips, C. A., & Zamudio, K. R. (2015). Asymmetric Introgression in a Spotted Salamander Hybrid Zone. *Journal of Heredity*, *106*(5), 608–617. doi:10.1093/jhered/esv042

Jombart, T. (2008). Adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*. doi:10.1093/bioinformatics/btn129

Kamvar, Z. N., Tabima, J. F., & Grünwald, N. J. (2014). Poppr : an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, *2*, e281. doi:10.7717/peerj.281

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*. doi:10.1093/molbev/mst010

Leasi, F., Sevigny, J. L., Laflamme, E. M., Artois, T., Curini-Galletti, M., de Jesus Navarrete, A., … Thomas, W. K. (2018). Biodiversity estimates and ecological interpretations of meiofaunal communities are biased by the taxonomic approach. *Communications Biology*, *1*(1), 112. doi:10.1038/s42003-018-0119-2

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., … Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, *10*(1), 34. doi:10.1186/1742-9994-10-34

Ligges, U., & Mächler, M. (2003). scatterplot3d - An R Package for Visualizing Multivariate Data. *Journal of Statistical Software*, *8*(11), 1–36. doi:10.18637/jss.v008.i11

Magnacca, K. N., & Brown, M. J. F. (2012). DNA barcoding a regional fauna: Irish solitary bees. *Molecular Ecology Resources*, *12*(6), 990–998. doi:10.1111/1755-0998.12001

Mastrantonio, V., Porretta, D., Urbanelli, S., Crasta, G., & Nascetti, G. (2016). Dynamics of mtDNA introgression during species range expansion: insights from an experimental longitudinal study. *Scientific Reports*, *6*(1), 30355. doi:10.1038/srep30355

McKay, B. D., & Zink, R. M. (2010). The causes of mitochondrial DNA gene tree paraphyly in birds. *Molecular Phylogenetics and Evolution*, *54*(2), 647–650. doi:10.1016/j.ympev.2009.08.024

Murray, T. E., Fitzpatrick, Ú., Brown, M. J. F., & Paxton, R. J. (2008). Cryptic species diversity in a widespread bumble bee complex revealed using mitochondrial DNA RFLPs. *Conservation Genetics*, *9*(3), 653–666. doi:10.1007/s10592-007-9394-z

Mutanen, M., Kivelä, S. M., Vos, R. A., Doorenweerd, C., Ratnasingham, S., Hausmann, A., … Godfray, H. C. J. (2016). Species-Level Para- and Polyphyly in DNA Barcode Gene Trees: Strong Operational Bias in European Lepidoptera. *Systematic Biology*, *65*(6), 1024–1040. doi:10.1093/sysbio/syw044

Nevado, B., Fazalova, V., Backeljau, T., Hanssens, M., & Verheyen, E. (2011). Repeated unidirectional introgression of nuclear and mitochondrial dna between four congeneric tanganyikan cichlids. *Molecular Biology and Evolution*. doi:10.1093/molbev/msr043

Paradis, E. (2010). pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics*, *26*(3), 419–420. doi:10.1093/bioinformatics/btp696

Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*. doi:10.1093/bioinformatics/bty633

Pérez-Ponce de León, G., & Poulin, R. (2016). Taxonomic distribution of cryptic diversity among metazoans: not so homogeneous after all. *Biology Letters*, *12*(8), 20160371. doi:10.1098/rsbl.2016.0371

Pfenninger, M., & Schwenk, K. (2007). Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. *BMC Evolutionary Biology*, *7*, 1–6. doi:10.1186/1471-2148-7-121

Pons, J.-M., Sonsthagen, S., Dove, C., & Crochet, P.-A. (2014). Extensive mitochondrial introgression in North American Great Black-backed Gulls (Larus marinus) from the American Herring Gull (Larus smithsonianus) with little nuclear DNA impact. *Heredity*, *112*(3), 226–239. doi:10.1038/hdy.2013.98

Praz, C., Müller, A., & Genoud, D. (2019). Hidden diversity in European bees: Andrena amieti sp. n., a new Alpine bee species related to Andrena bicolor (Fabricius, 1775) (Hymenoptera, Apoidea, Andrenidae). *Alpine Entomology*, *3*, 11–38. doi:10.3897/alpento.3.29675

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, *5*(3), e9490. doi:10.1371/journal.pone.0009490

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology*, *67*(5), 901–904. doi:10.1093/sysbio/syy032

Reid, N. M., & Carstens, B. C. (2012). Phylogenetic estimation error can decrease the accuracy of species delimitation: A Bayesian implementation of the general mixed Yule-coalescent model. *BMC Evolutionary Biology*. doi:10.1186/1471-2148-12-196

Sánchez-Bayo, F., & Wyckhuys, K. A. G. (2019). Worldwide decline of the entomofauna: A review of its drivers. *Biological Conservation*, *232*(September 2018), 8–27. doi:10.1016/j.biocon.2019.01.020

Scheuchl, E., & Willner, W. (2016). *Taschenlexikon der Wildbienen Mitteleuropas*. Quelle & Meyer.

Schmidt, S., Schmid-Egger, C., Morinière, J., Haszprunar, G., & Hebert, P. D. N. (2015). DNA barcoding largely supports 250 years of classical taxonomy: identifications for Central European bees (Hymenoptera, Apoidea partim). *Molecular Ecology Resources*, *15*(4), 985–1000. doi:10.1111/1755-0998.12363

Scriven, J. J., Whitehorn, P. R., Goulson, D., & Tinsley, M. C. (2016). Niche partitioning in a sympatric cryptic species complex. *Ecology and Evolution*, *6*(5), 1328–1339. doi:10.1002/ece3.1965

Smith, B. T., Harvey, M. G., Faircloth, B. C., Glenn, T. C., & Brumfield, R. T. (2014). Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. *Systematic Biology*. doi:10.1093/sysbio/syt061

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. doi:10.1093/bioinformatics/btu033

Stork, N. E. (2017). How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth? *Annual Review of Entomology*, *63*(1), 31–45. doi:10.1146/annurev-ento-020117-043348

Struck, T. H., Feder, J. L., Bendiksby, M., Birkeland, S., Cerca, J., Gusarov, V. I., … Dimitrov, D. (2018). Finding Evolutionary Processes Hidden in Cryptic Species. *Trends in Ecology & Evolution*, *33*(3), 153–163. doi:10.1016/j.tree.2017.11.007

Toews, D. P. L., & Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*, *21*(16), 3907–3930. doi:10.1111/j.1365-294X.2012.05664.x

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., & Legendre, F. (2017). Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports*, *7*(1), 9132. doi:10.1038/s41598-017-09084-6

Wielstra, B., & Arntzen, J. W. (2012). Postglacial species displacement in Triturus newts deduced from asymmetrically introgressed mitochondrial DNA and ecological niche models. *BMC Evolutionary Biology*, *12*(1), 161. doi:10.1186/1471-2148-12-161

Wilkinson, L. (2011). ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H. *Biometrics*. doi:10.1111/j.1541-0420.2011.01616.x

Williams, P. H., Brown, M. J. F., Carolan, J. C., An, J., Goulson, D., Aytekin, A. M., … Xie, Z. (2012). Unveiling cryptic species of the bumblebee subgenus Bombus s. str. worldwide with COI barcodes (Hymenoptera: Apidae). *Systematics and Biodiversity*, *10*(1), 21–56. doi:10.1080/14772000.2012.664574

Winker, K. (2006). Sibling species were first recognized by William Derham (1718). *The Auk*, *122*(2), 706. doi:10.1642/0004-8038(2005)122[0706:sswfrb]2.0.co;2

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

Yang, Ziheng, & Rannala, B. (2014). Unguided Species Delimitation Using DNA Sequence Data from Multiple Loci. *Molecular Biology and Evolution*, *31*(12), 3125–3135. doi:10.1093/molbev/msu279

Zarza, E., Connors, E. M., Maley, J. M., Tsai, W. L. E., Heimes, P., Kaplan, M., & McCormack, J. E. (2018). Combining ultraconserved elements and mtDNA data to uncover lineage diversity in a Mexican highland frog (Sarcohyla ; Hylidae). *PeerJ*, *6*, e6045. doi:10.7717/peerj.6045

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. doi:10.1101/gr.074492.107

Zieliński, P., Nadachowska-Brzyska, K., Wielstra, B., Szkotak, R., Covaciu-Marcov, S. D., Cogălniceanu, D., & Babik, W. (2013). No evidence for nuclear introgression despite complete mtDNA replacement in the Carpathian newt (Lissotriton montandoni). *Molecular Ecology*, *22*(7), 1884–1903. doi:10.1111/mec.12225

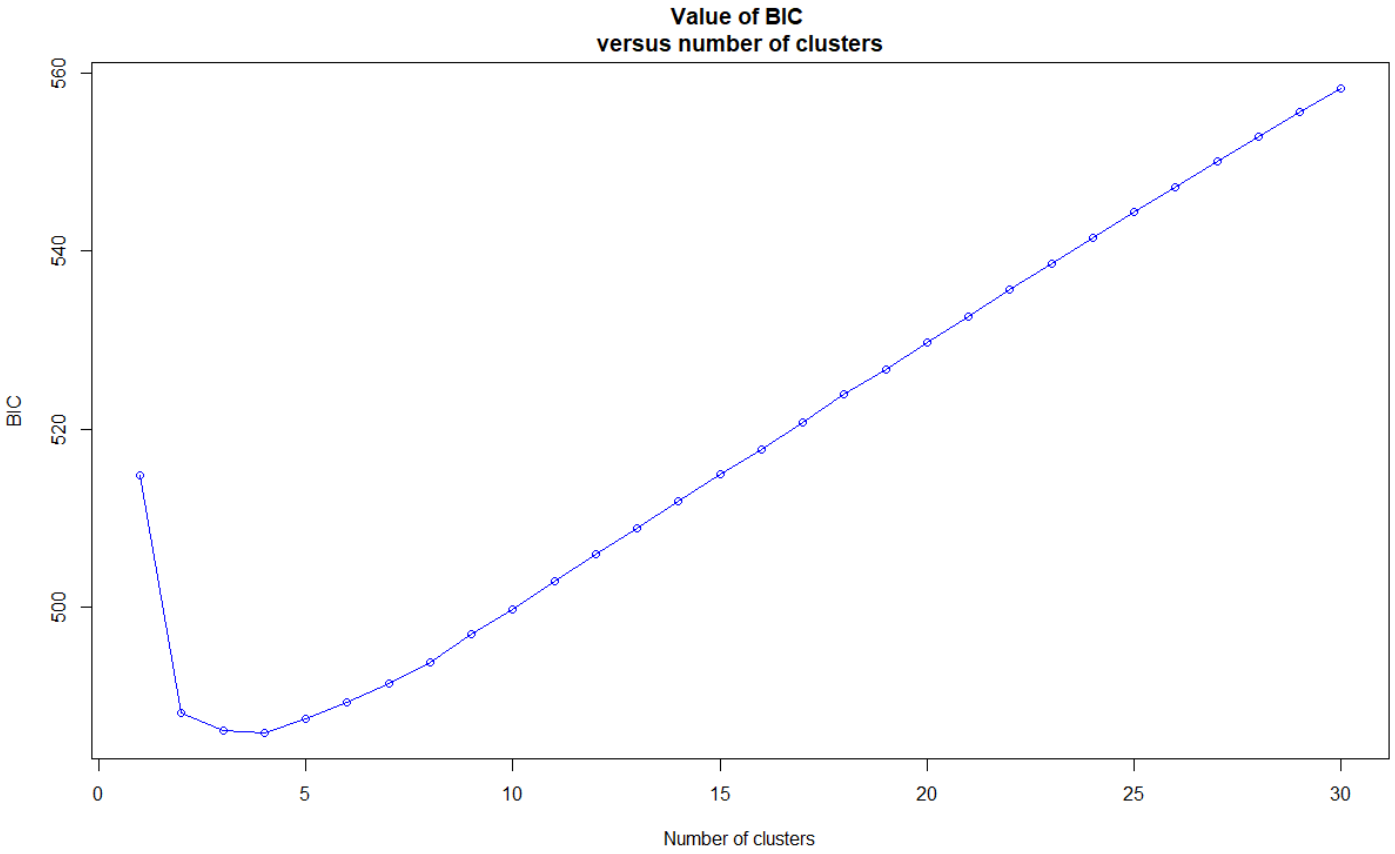
Supplementary Information

S1: Metadata

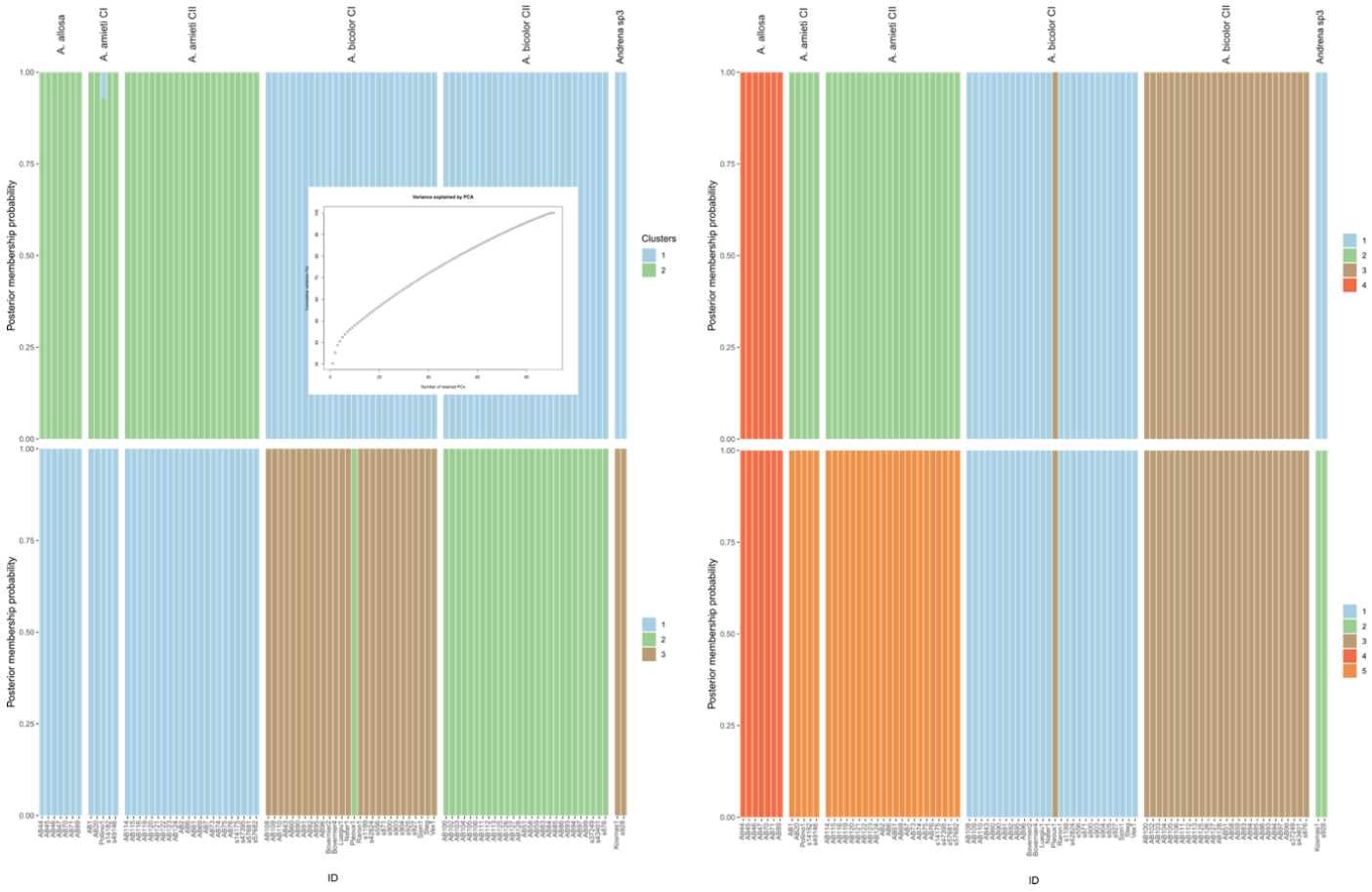
ID	Locality	Country	Day	Month	Year	Genus	Species	COI lineage	Sex	coordN	coordE
AB44	St-Martin	CH	11	5	2018	Andrena	allosa	NA	F	46.175	7.470
AB45	St-Martin	CH	11	5	2018	Andrena	allosa	NA	F	46.175	7.470
AB46	St-Martin	CH	11	5	2018	Andrena	allosa	NA	F	46.175	7.470
AB47	St-Martin	CH	11	5	2018	Andrena	allosa	NA	F	46.175	7.470
AB70	Allos	FR	5	6	2018	Andrena	allosa	NA	F	44.300	6.566
AB71	Allos	FR	5	6	2018	Andrena	allosa	NA	F	44.300	6.566
AB89	Allos	FR	5	6	2018	Andrena	allosa	NA	F	44.300	6.566
s57681	Madzeria	CH	2	7	2015	Andrena	amieti	CII	M	46.008	7.346
s57682	Madzeria	CH	2	7	2015	Andrena	amieti	CII	M	46.008	7.346
s47395	Bagnes	CH	3	7	2015	Andrena	amieti	CII	M	46.008	7.342
s14182	Monte Pollino	IT	4	7	2011	Andrena	amieti	CI	F	39.904	16.181
s14183	Monte Pollino	IT	4	7	2011	Andrena	amieti	CI	M	39.904	16.181
Pollino1	Monte Pollino	IT	4	7	2011	Andrena	amieti	CI	F	39.904	16.181
s49146	Kandersteg	CH	17	5	2017	Andrena	amieti	CI	F	46.503	7.707
AB119	St-Martin	CH	20	7	2018	Andrena	amieti	CII	F	46.169	7.472
AB120	St-Martin	CH	20	7	2018	Andrena	amieti	CII	F	46.169	7.472
s14175	Bagnes	CH	24	5	2017	Andrena	amieti	CII	F	46.021	7.329
AB121	Bargis	CH	25	7	2018	Andrena	amieti	CII	M	46.859	9.317
AB122	Bischofalp	CH	30	7	2018	Andrena	amieti	CII	F	46.923	9.122
AB114	Chapelle en valgaudemar	FR	31	7	2018	Andrena	amieti	CII	M	44.793	6.155
AB115	Chapelle en valgaudemar	FR	31	7	2018	Andrena	amieti	CII	M	44.793	6.155
AB116	Chapelle en valgaudemar	FR	31	7	2018	Andrena	amieti	CII	M	44.793	6.155
AB123	Fisetenpass	CH	31	7	2018	Andrena	amieti	CII	M	46.884	8.933
AB124	Fisetenpass	CH	31	7	2018	Andrena	amieti	CII	M	46.884	8.933
AB76	Allos	FR	5	6	2018	Andrena	amieti	CII	F	44.300	6.566
AB1	Kandersteg	CH	16	8	2018	Andrena	amieti	CI	M	46.501	7.715
AB2	Kandersteg	CH	16	8	2018	Andrena	amieti	CII	M	46.501	7.715
AB6	Kandersteg	CH	16	8	2018	Andrena	amieti	CII	M	46.501	7.715
AB7	Kandersteg	CH	16	8	2018	Andrena	amieti	CII	M	46.501	7.715
AB20	Kandersteg	CH	16	8	2018	Andrena	amieti	CI	M	46.501	7.715
AB61	Kandersteg	CH	16	8	2018	Andrena	amieti	CII	F	46.503	7.713
AB69	Allos	FR	5	6	2018	Andrena	amieti	CII	F	44.300	6.566
AB73	Allos	FR	5	6	2018	Andrena	amieti	CII	F	44.300	6.566
AB74	Allos	FR	5	6	2018	Andrena	amieti	CII	F	44.300	6.566
AB75	Allos	FR	5	6	2018	Andrena	amieti	CII	F	44.300	6.566
AB3	Kandersteg	CH	16	8	2018	Andrena	amieti	CII	M	46.501	7.715
s556	Emdt	CH	0	0	2013	Andrena	bicolor	CI	F	46.221	7.816
s925	Leuk	CH	1	4	2011	Andrena	bicolor	CI	F	46.303	7.678
s876	Kosmas	GR	2	6	2014	Andrena	bicolor	CII	F	37.107	22.728
s900	Leuk	CH	2	6	2011	Andrena	bicolor	CI	F	46.303	7.678
s5460	Locarno	IT	6	5	2016	Andrena	bicolor	CII	F	46.193	8.787
Sion1	Sion	CH	7	4	2017	Andrena	bicolor	CI	M	46.250	7.403
s871	Taygetus	GR	8	6	2014	Andrena	bicolor	CI	F	36.955	22.363
s1199	St-Martin	CH	8	7	2017	Andrena	bicolor	CI	F	46.170	7.472
s43401	Castel San Pietro	IT	10	6	2017	Andrena	bicolor	CII	M	45.902	9.015
AB90	Bovernier	CH	11	6	2018	Andrena	bicolor	CI	F	46.081	7.095
AB91	Bovernier	CH	11	6	2018	Andrena	bicolor	CI	M	46.081	7.095
AB92	Bovernier	CH	11	6	2018	Andrena	bicolor	CI	M	46.081	7.095
AB93	Bovernier	CH	11	6	2018	Andrena	bicolor	CII	M	46.081	7.095
AB94	Bovernier	CH	11	6	2018	Andrena	bicolor	CII	M	46.081	7.095
s42824	Liddes	CH	12	6	2017	Andrena	bicolor	CI	F	46.007	7.172

AB95	Inden	CH	12	6	2018	Andrena	bicolor	CI	F	46.344	7.616
AB96	Inden	CH	12	6	2018	Andrena	bicolor	CI	F	46.344	7.616
AB97	Saviese	CH	12	6	2018	Andrena	bicolor	CII	F	46.275	7.323
AB99	Saviese	CH	12	6	2018	Andrena	bicolor	CII	M	46.275	7.323
AB100	Les plans sur bex	CH	14	6	2018	Andrena	bicolor	CII	F	46.259	7.091
AB102	Les plans sur bex	CH	14	6	2018	Andrena	bicolor	CII	M	46.259	7.091
AB103	Evionnaz	CH	14	6	2018	Andrena	bicolor	CII	M	46.183	7.005
AB104	Evionnaz	CH	14	6	2018	Andrena	bicolor	CII	M	46.183	7.005
AB105	Evionnaz	CH	14	6	2018	Andrena	bicolor	CII	M	46.183	7.005
AB106	Evionnaz	CH	14	6	2018	Andrena	bicolor	CII	M	46.183	7.005
Nater1	Naters	CH	15	7	2014	Andrena	bicolor	CI	F	46.365	7.958
AB108	Steg	CH	15	6	2018	Andrena	bicolor	CI	F	46.329	7.770
AB109	Steg	CH	15	6	2018	Andrena	bicolor	CI	F	46.329	7.770
AB110	Steg	CH	15	6	2018	Andrena	bicolor	CI	F	46.329	7.770
AB113	Caumasle	CH	15	7	2018	Andrena	bicolor	CII	F	46.925	9.354
s927	Gampel-Bratsch	CH	16	7	2014	Andrena	bicolor	CI	F	46.329	7.720
s37249	Goms	CH	17	5	2017	Andrena	bicolor	CII	F	46.466	8.231
Steg	Steg-Hohtenn	CH	17	6	2014	Andrena	bicolor	CI	F	46.329	7.771
s903	Vex	CH	17	4	2012	Andrena	bicolor	CI	F	46.208	7.411
s904	Vex	CH	17	7	2012	Andrena	bicolor	CI	F	46.208	7.405
Luegiu1	Luegiu	CH	17	5	2017	Andrena	bicolor	CI	M	46.320	7.770
AB111	St-Martin	CH	20	7	2018	Andrena	bicolor	CII	F	46.169	7.472
AB125	Horn	CH	20	4	2018	Andrena	bicolor	CII	M	47.365	7.858
AB126	Horn	CH	20	4	2018	Andrena	bicolor	CII	M	47.365	7.858
AB127	Horn	CH	20	4	2018	Andrena	bicolor	CII	M	47.365	7.858
AB128	Horn	CH	20	4	2018	Andrena	bicolor	CII	M	47.365	7.858
AB112	Caumasle	CH	24	7	2018	Anderna	bicolor	CII	M	46.925	9.354
Pleoux1	Pleoux	FR	27	5	2017	Andrena	bicolor	CI	F	44.353	4.235
Raron1	Raron	CH	31	3	2014	Andrena	bicolor	CI	F	46.316	7.795
Bovernier1	Bovernier	CH	15	6	2012	Andrena	bicolor	CI	F	46.081	7.085
Bovernier2	Bovernier	CH	15	6	2012	Andrena	bicolor	CI	F	46.079	7.082
Bovernier4	Bovernier	CH	15	6	2012	Andrena	bicolor	CI	M	46.081	7.085
Bovernier3	Bovernier	CH	15	6	2012	Andrena	bicolor	CI	F	46.081	7.085
AB43	Conthey	CH	10	6	2018	Andrena	bicolor	CI	M	46.264	7.264
AB51	St-Martin	CH	11	5	2018	Andrena	bicolor	CII	F	46.175	7.470
AB58	Leuk	CH	10	6	2018	Andrena	bicolor	CII	F	46.335	7.657
AB59	Leuk	CH	10	6	2018	Andrena	bicolor	CII	F	46.335	7.657
AB60	Saillon	CH	10	6	2018	Andrena	bicolor	CI	F	46.171	7.176
AB83	Allos	FR	5	6	2018	Andrena	bicolor	CII	F	44.300	6.566
AB84	Allos	FR	5	6	2018	Andrena	bicolor	CII	F	44.300	6.566
AB85	Allos	FR	5	6	2018	Andrena	bicolor	CII	F	44.300	6.566
AB86	Allos	FR	5	6	2018	Andrena	bicolor	CII	F	44.300	6.566
Vex1	Vex	CH	18	6	2012	Andrena	bicolor	CI	F	46.208	7.401
s928	Kosmas	GR	2	6	2014	Andrena	Sp3	NA	F	37.107	22.728
Kosmas1	Kosmas	GR	2	6	2014	Andrena	Sp3	NA	F	37.107	22.728

S2: Number of clusters and their associated Bayesian Information Criterion (BIC) value identified on the UCE dataset without a prior on the specimens' identifications. Cluster with the lowest BIC value was K = 4.



S3: Discriminant analysis of principal components (DAPC) clustering specimens into genetic clusters without species a priori knowledge. Specimens are grouped by mitochondrial lineages with the corresponding group affiliation above each block. The analysis was run for K = 2 until K = 5, with K = 4 harbouring the lowest BIC value (see Supplementary Figure S2). A plot showing the cumulated variance explained by the eigenvalues of the PCA is niched in the first DAPC plot (k = 2). The optimal number of PCs to retained (i.e. 5 PCs conserving 42.5% of the total variance) was determined using both the plotted cumulated variance of the eigenvalues (niched in DAPC plot K = 2) and results from the xvalDapc function (adegenet package).



S4: BPP results of two consecutive runs using the A11 model. Specimens were assigned to their species, with A. bicolor divided into two distinct species. Alpha and beta parameters were set to 3 and 0.004 for inverted gamma distribution of the theta prior, and to 3 and 0.002 for the tau prior. The analyses were run with a MCMC of 500,000 generations and a 10% burn-in period.

(A) List of best models

Run	posterior probability	# species	SpeciesTree
1	0.99998	5	((A. allosa, A. amieti), ((A. bicolor CI, A. bicolor CII), Andrena sp3))
2	1	5	((A. allosa, A. amieti), ((A. bicolor CI, A. bicolor CII), Andrena sp3))

(B) Number of species delimitations

Run	# species delimited	posterior probability	# species
1	1	0.99998	5
2	1	1	5

(C) Species delimitation

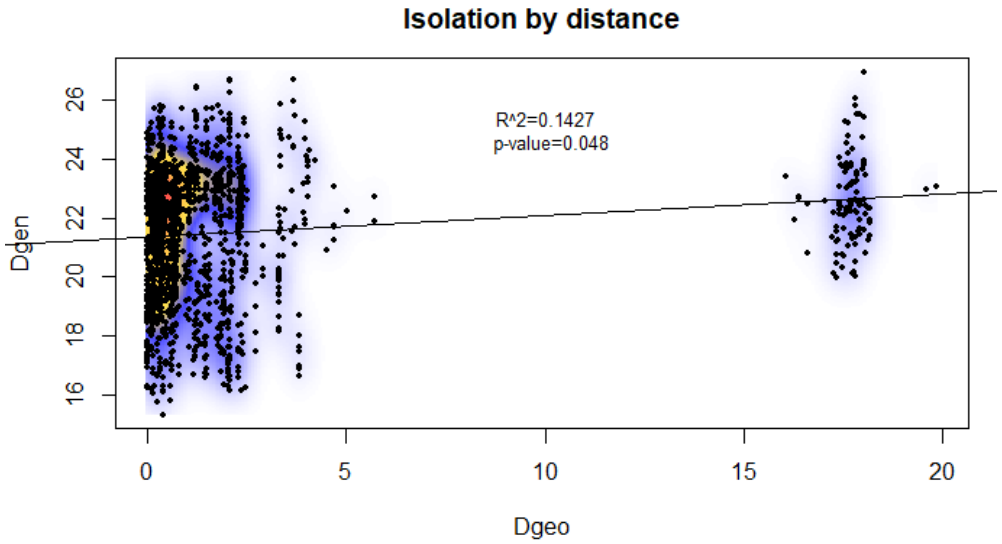
Run	posteriori probability	species name
1	0.99998	A. bicolor CII
	0.99998	A. amieti
	0.99998	A. bicolor CI
	0.99998	Andrena sp3
	0.99998	A. allosa
2	1	A. bicolor CII
	1	A. amieti
	1	A. bicolor CI
	1	Andrena sp3
	1	A. allosa

(D) Posterior probability for # of species

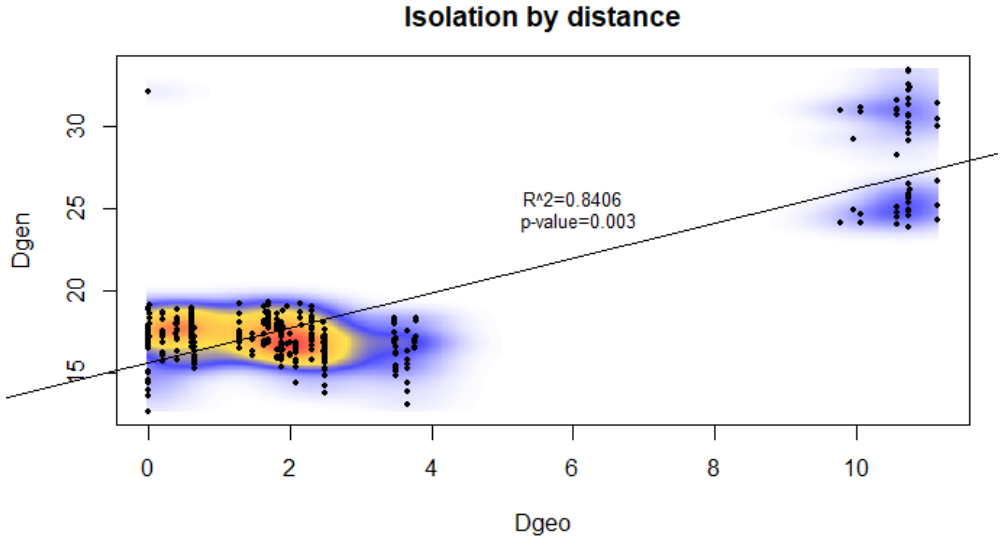
Run	Posterior probability	Prior
1	0	0.175
	0	0.175
	0	0.225
	0	0.25
	0.99998	0.175
2	0	0.175
	0	0.175
	0	0.225
	0	0.25
	1	0.175

S5: Isolation by distance analyses for *A. bicolor* (A) and *A. amieti (B)*. The analyses were run on the UCE dataset using the *gl.ibd* function (*dartR* package) with 1000 permutations. Plot shows the Euclidean geographical distance (i.e. Dgeo) against population based pairwise Fst/1-Fst (Rousseau's distance measure) (i.e. Dgen).

A



B



- Chapter III -

UCE overcome mitochondrial barcode limitations and provide a quick and robust genomic tool for species delimitation in Central European bees (Hymenoptera: Anthophila)

Morgan Gueuning, Juerg E. Frey & Christophe Praz

(Publication in preparation)

Abstract

Accurate and testable species delimitation hypotheses are essential for measuring, surveying and managing biodiversity. Today, taxonomists often rely on mitochondrial DNA barcodes to complement morphological species delimitations. However, there are strong limitations of mitochondrial DNA barcodes which can ultimately lead to erroneous signal, either for identifications or for species delimitation. First, what is regarded as one single species can be associated with two distinct DNA barcodes, which can either point to cryptic diversity, or simply to deep within-species mitochondrial divergences with no reproductive isolation. In extreme cases, these deep within-species divergences may be associated with mitochondrial paraphyly. Second, two distinct species can share barcodes, for instance due to mitochondrial introgression. These intrinsic limitations of mitochondrial DNA barcoding can only be addressed with nuclear genomic markers, which are expensive, labour intensive, often poorly repeatable, and often require high-quality DNA.

In this study, we examine the use of ultraconserved nuclear genetic elements (UCEs) as a quick and robust genomic approach for species delimitation. This genomic method was assessed using two cases of mitochondrial introgression among distinct, albeit closely related species; two cases of species paraphyly not associated with known morphological differences; and two additional cases where species delimitation hypotheses have been conflicting. In this study, UCEs recovered thousands of homologous nuclear loci and provided strong support for clear species delimitations in all investigated species-groups. Furthermore, this study also provided the first conclusive evidence for both mitochondrial introgression among distinct species, and mitochondrial paraphyly within a single species in bees.

1. Introduction

Human induced environmental changes are increasingly threatening vital ecosystem services and ultimately human well-being. Climate change is often cited as the most problematic environmental problem of the century, however another possibly connected crisis is silently taking place: loss in biodiversity (Ceballos, Ehrlich, & Dirzo, 2017). Over the past few decades, biodiversity decline has intensified so dramatically that according to some scientists, our planet is facing the onset of the sixth mass extinction (Barnosky et al., 2011; Pievani, 2014). However, targeted conservation efforts may enable slowing this dramatic decay of biodiversity. Yet such conservation efforts require accurate biodiversity databases as well as quick and robust monitoring techniques. For insects or invertebrates in general, important impediments are linked to incompleteness of their taxonomy and difficulties in species identification, rendering the survey of species distribution and population trends difficult (Cardoso, Erwin, Borges, & New, 2011). Providing the scientific community with accurate and robust identification tools and species hypotheses is thus a major task for taxonomists.

Traditionally, species were described by examining variation in morphological traits (Padial, Miralles, De la Riva, & Vences, 2010). Species were delimited in a way to minimize within-species variation and to maximise between-species variation in sets of variable characters. However, morphological delimitation is not always in full agreement with other species delimitation criteria. Indeed, morphological taxonomy is often challenged by morphological homogeneity i.e., the lack of variation between taxa, or conversely by sexual or generational polymorphisms within species, both of which will lead to an absence of a "morphological" gap between species. This underestimation of true levels of biodiversity results in so called "cryptic" diversity (Karanovic, Djurakic, & Eberhard, 2016). To complement morphology, DNA barcoding was introduced as a reliable, fast, and cheap identification method (Brunner, Fleming, & Frey, 2002; Hebert, Cywinska, Ball, & DeWaard, 2003), and has since been extensively used not only for specimen identification but also for species delimitation. For animals, the 5'-region of the cytochrome oxidase subunit I (COI) gene has quickly become the DNA barcode gold standard. The success of COI is mainly due to its universality, rapid evolutionary mutation rate and lack of recombination due to uniparental inheritance.

There are, nevertheless, numerous examples where COI-barcoding resulted in misleading taxon identification, contradicting morphological identifications. A number of possible reasons for such problematic barcoding results have recently emerged. For example, a growing body of literature is reporting that mitochondrial inheritance is more complicated than initially thought, with rare cases of paternal leakage, heteroplasmy or recombination (Ladoukakis & Zouros, 2017; White, Wolff, Pierson, & Gemmell, 2008). Furthermore, mitochondrial genomes can be subject to evolutionary forces acting solely at the organelle level [e.g. mitochondrial introgression, *Wolbachia* infection or sex-biased asymmetries; (Toews & Brelsford, 2012)]. Consequently, COI-based barcoding is subject to two types of errors. Error type I occurs when one biological species is associated with two distinct DNA barcodes, leading to the erroneous detection of two hypothetical species within a single biological species. This error is observed because of deep within-species divergences (in extreme cases associated with mitochondrial paraphyly), or artefacts such as nuclear insertions (Song, Buhay, Whiting, & Crandall, 2008). Type II errors occur when DNA barcoding fails to recognize two distinct species because of barcode sharing, most often linked to mitochondrial introgression, often mediated by *Wolbachia* in insects (Whitworth, Dawson, Magalon, & Baudry, 2007). Although these events are generally considered rare (but see Klopfstein, Kropf, & Baur, 2016; Neumeyer, Baur, Guex, & Praz, 2014; Nichols,

Jordan, Jamie, Cant, & Hoffman, 2012), they can considerably skew phylogenies or biodiversity estimates (Andriollo, Naciri, & Ruedi, 2015; Hinojosa et al., 2019; Mutanen et al., 2016). The consensus is that species delimitation should rely on multiple sources of information and, for molecular markers, species delimitation should use genes of both mitochondrial and nuclear origin.

Several different types of nuclear DNA markers can be used for species delimitation, yet the respective techniques are all associated with serious drawbacks. For instance, the fastest and most cost-effective solution is the amplicon sequencing of single-copied nuclear genes, such as elongation factor (EF) or of ribosomal DNA markers (Leneveu, Chichvarkhin, & Wahlberg, 2009; Martinet et al., 2018; Soltani, Bénon, Alvarez, & Praz, 2017; Williams, Lelej, & Thaochan, 2019). However the usefulness of these markers is often limited by the lack of phylogenetic resolution (Dellicour & Flot, 2018) or by within-genome variation (e.g. variation in the internal transcribed spacer [ITS]) rending sequencing complicated. For increased resolution, some studies have used population genetic markers such as microsatellites. Although microsatellites provide ample resolution for species delimitation (McKendrick et al., 2017), one major limitation is that loci are clade-specific and therefore require clade specific primers that have to be designed base on available genomic information. Alternative approaches that are slightly more universal and provide equal or higher resolution include genomic-reduction techniques such as RAD- or ddRAD sequencing (Lemopoulos et al., 2019). These methods are very powerful and can provide high-resolution, intraspecific information on population dynamics but again they suffer from some limitations. First, these methods rely on deep sequencing and are therefore costly and labor-intensive. Second, for standard applications, they generally require high-quality DNA, although protocols have been designed for ancient DNA (Schmid et al., 2017; Suchan et al., 2016). Finally, datasets obtained from different studies and/or taxa are hardly joinable due to the lack of repeatability. This last limitation of RAD-sequencing is a severe drawback for species delimitation, since taxonomic work essentially builds upon previous hypotheses, with new data continuously complementing earlier datasets.

Ideally, molecular species-delimitation markers should be: (i) nuclear, to reflect gene flow as a whole and not only organelle-based gene flow; (ii) genomic scale to cover numerous independent loci; (iii) sufficiently variable to capture recently diverged species; (iv) repeatable, so that datasets can be complemented once more material is available; (v) universal to the extent that datasets can complement each other.

In 2012, ultraconserved elements (UCEs) were introduced as a quick and essentially universal way to obtain “thousands of genetic markers spanning multiple evolutionary timescales” (Faircloth et al., 2012). UCEs appear to fulfil many of the above mentioned requirements. However, whether they harbor enough variation to capture divergence among recently diverged species remains an open question, since by definition they are highly conserved. In this study, we examine the use of UCEs for species delimitation in bees. For these taxa, as for the majority of animals, COI has been proven to have, in the majority of cases, sufficient sequence divergence to enable taxonomic differentiation at the species level (Gueuning et al., 2019; Sheffield, Hebert, Kevan, & Packer, 2009; Sheffield et al., 2017; Tang et al., 2015). Typically, a difference of 2% between barcode sequences corresponds to biologically different species (Ratnasingham & Hebert, 2013). However, several studies have reported cases of DNA barcoding errors. For instance, the large scale barcoding studies performed on the Apoidae fauna of Germany revealed 50 cases (9.7%) of DNA barcode sharing in 561 investigated species (Schmidt, Schmid-Egger, Morinière, Haszprunar, & Hebert, 2015). Furthermore, this same study found divergent barcodes within currently recognized species for 56 species (11%). A similar

study on the Canadian fauna reported barcode sharing for 3.2% of all 856 barcoded species (Sheffield et al., 2017).

Here we investigate UCEs as a delimitation tool for cases previously reported to be problematic in DNA barcoding. We include examples of both putative mitochondrial introgression and of multiple barcodes per species (Schmidt et al., 2015). More specifically, we focused on the following European species-complexes: *Andrena barbareae /cineraria; A. dorsata/propinqua; A. carantonica/trimmerana/rosae; Lasioglossum alpigenum/bavaricum/cupromicans; Nomada goodeniana/succincta.* Most of these cases are also controversial with respect to morphological delimitations (see below), so that current evidence based on the combined characteristics of morphology and COI-based DNA barcodes does not enable definite conclusions on the status of these species.

2. Material and methods

2.1. Species complexes

We selected five cases of known discrepancies between morphological and COI-based identifications in the Swiss bee fauna, mainly based on the comprehensive study of Schmidt et al. (2015). We tried to include all reported cases of mitochondrial introgression in Switzerland, with the exception of *Andrena nitida/A. limita* and *Colletes hederae* and *C. succinctus*, for which not enough material was available.

2.1.1. Andrena. barbareae/cineraria

A. barbareae and *A. cineraria* are sibling species, morphologically very close, although identifiable in most cases by a combination of characters in both genders (Amiet, Hermann, Müller, & Neumeyer, 2010). *A. cineraria* has a wider distribution than *A. barbareae*, which is mainly restricted to the Alps. Both species are polylectic but exhibit different phenologies with two generations for *A. barbareae* and one for *A. cineraria*. Because of their morphological, biogeographical and phenological differences, both species are generally considered as separated although this view was recently challenged because both taxa share identical barcodes (Schmidt et al., 2015).

2.1.2. Andrena carantonica/trimmerana/rosae

Taxonomical delimitation in this species-complex is a long-standing enigma with controversial species delimitation hypotheses due to morphologically divergent generations and unclear morphological differentiation. First, *Andrena rosae* and *A. stragulata* are considered by most authors to represent the summer and spring generations of the same species (Falk, 2016; Reemer, Groenenberg, Van Achterberg, & Peeters, 2008; Schmidt et al., 2015). Nevertheless, both taxa exhibit differences in terms of morphology, pollen collecting behaviour and nesting sites (Amiet et al., 2010; van der Meer, Reemer, Peeters, & Neve, 2006; Westrich, 2014). Second, morphological differentiation of *A. carantonica* and *A. trimmerana* is challenging. No clear morphological character allows to separate females; and while males of the first generation differ in the morphology of the mandible, those of the summer generation cannot be differentiated. Both taxa exhibit distinct phenologies with only one generation for *A. carantonica* (April to May) and two for *A. trimmerana* (March-April, June-July), although isolated late-summer specimens of *A. carantonica* are known (C. Praz and R. Paxton, unpublished data). Both taxa overlap in their distribution area, but *A. carantonica* is much more abundant than *A. trimmerana* for which only a few data points are available in Switzerland.

2.1.3. Andrena dorsata/propinqua

Depending on the author, *A. dorsata* and *A. propinqua* are considered as separate or conspecific taxa (Amiet, Hermann, Müller, & Neumeyer, 2010; Gusenleitner & Schwarz, 2002; Schmid-Egger & Scheuchl, 1997). Morphologically, the separation of both taxa is complicated and subject to errors, especially at the European scale. At the genetic level, both species were previously found to share COI barcodes, although identification errors could not be excluded (Schmidt et al., 2015).

2.1.4. Lasioglossum alpigenum/bavaricum/cupromicans

Species delimitation in this complex is generally accepted based on clear differences in male genital morphology (Amiet, Herrmann, Müller, & Neumeyer, 2001; Edmer, 1970). Identification of females is however challenging, and *L. bavaricum* and *L. cupromicans* were recently suggested to share the same COI barcode, although no male of *L. bavaricum* had been sequenced, rendering identifications for this taxon tentative.

2.1.5. Nomada goodeniana/succincta

The morphological separation between these two species are mainly relying on colour patterns and are therefore prone to identification errors, although both appear to differ in their hosts, phenology and possibly in the chemical composition of mandibular glandular secretions (Kuhlmann, 1997). In their barcoding study, Schmidt et al. (2015) found two divergent clusters for *A. succincta*: a northern European cluster containing specimens of *A. goodeniana* and a southern European cluster. A similar result was found in England (Creedy et al., 2019). As described above for *A. dorsata/propinqua*, the reported COI barcode sharing between *N. goodeniana* and *N. succincta* could potentially be due to misidentification, at least in Germany, where a previous study suggested that both species did not share the same COI barcode (Diestelhorst & Lunau, 2008).

2.2. Sampling

Most specimens were sampled across Switzerland in the frame of the "Red List of Swiss bee" project between 2008 and 2019. Two additional specimens were collected in France and one in northern Italy. Bees were in most cases collected in ether and pinned; morphological identifications were all verified by one of us (CP); for *A. carantonica/trimmerana*, phenology was used in addition to morphology for identification. In total, the dataset was composed of 97 specimens. For the *Andrea cineraria/barbareae* species-complex, we sequenced six specimens of each species and one specimen of *A. vaga* as outgroup. For the *A. carantonica/trimmerana* complex, we sequenced eleven *A. carantonica* and eight *A. trimmerana* (five from the summer generation and three from the spring generation, "*A. spinigera"*); as outgroup we used the species *A. rosae*, for which we included eight specimens (four of the summer generation, and four of the spring generation, "*A. stragulata*"). For the *A. dorsata/propinqua* complex we sequenced seven specimens from each species and one specimen of *A. congruens* as outgroup. For the *Lasioglossum alpigenum/bavaricum/cupromicans* complex, we included eight *L. alpigenum*, six *L. bavaricum*, five *L. cupromicans* and one *L. nitidulum* (outgroup); we favoured males whenever possible since females are challenging to identify. Finally, for the *Nomada goodeniana/succincta* complex, our dataset comprises nine *N. goodeniana*, eight *N. succincta* and one *N. bifisciata* as outgroup. Morphological identifications in all these groups are challenging.

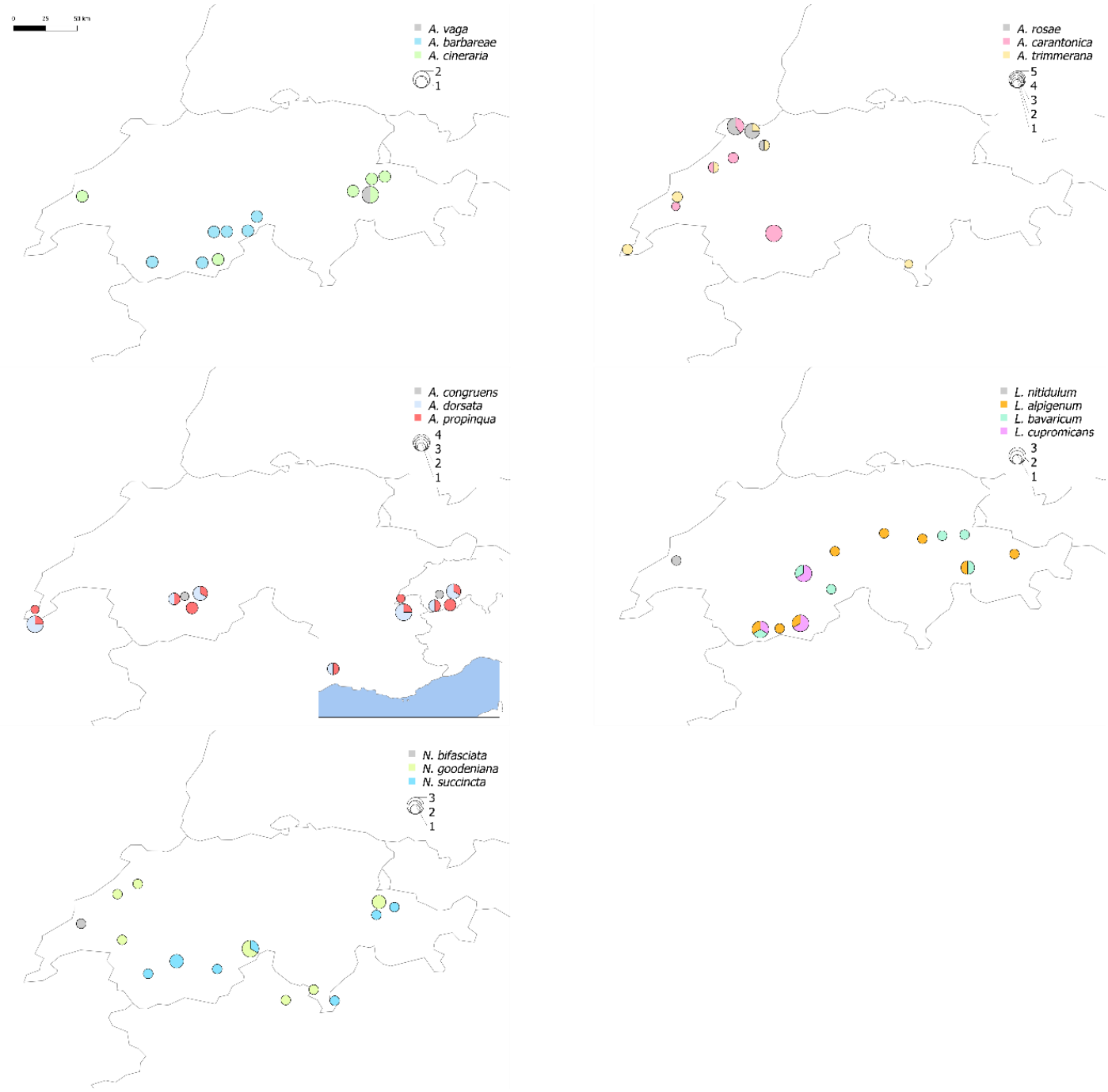


Figure 1: Sampling sites per species complex. Identification of specimens corresponds to the morphological identification. Number of specimens sampled per site is reflected by the pie charts size.

In the following cases, putative identification errors were detected and corrected: one male initially identified as *A. rosae* using the key of Amiet et al. (2010) clustered with *A. trimmerana* in both mitochondrial and nuclear analyses. Using the key of Falk (2016), this specimen was clearly identified as belonging to either *A. trimmerana* or *A. carantonica*, and the collecting date in early summer suggests *A. trimmerana*. Several specimens initially identified as *Nomada succincta* using the key of Amiet et al. (2007) (mostly the colour pattern of the scape and clypeus) clustered with specimens of *N. goodeniana* in both nuclear and mitochondrial phylogenetic trees. Using the criteria of Diestelhorst and Lunau (2008), in particular the colour pattern on the hind femora, identifications as *N. goodeniana* were confirmed. When possible, sites harbouring species in

sympatry were favoured and in all cases the perimeter around sampling localities for closely related taxa overlaps. Metadata can be found in Supplementary Table S1.

2.3. COI sequencing

The cytochrome oxidase unit I (COI) barcode was amplified for all 97 specimens. DNA extractions were performed over night at 65°C on whole bodies using a proteinase K solution. Extractions were automated on an extraction robot using the Qiagen BioSprint 96 DNA Blood Kit (Qiagen) following the manufacturer's protocol. A 313 bp region of the COI marker was amplified using the primer pair mlCOIintF/dgHCO2198 (Leray et al., 2013) and following PCR conditions described in Frey et al. (2013). Success of PCR amplifications was verified on a 1.5% agarose gel and PCR products were purified using NucleoFast 96 PCR clean-up kits (Macherey-Nagel). Linear amplification was performed using a BigDye terminator v1.1 kit (ThermoFisher) following the manufacturer's conditions. Sequencing reactions were purified using DyeEx 96 kits (Qiagen) and sequenced on a SeqStudio Genetic Analyzer (ThermoFisher).

2.4. COI phylogeny

Raw sequences were imported into Geneious v11.0.5 and consensus sequences between forward and reverse sequences were constructed for each specimen using the Geneious assembler. Consensus sequences were aligned per species-complex using *MAFFT* v7.308 (Katoh & Standley, 2013). Phylogenetic trees were built with *RAxML* v8.2.11 (Stamatakis, 2014) using the GTR GAMMA model and 100 bootstrap replicates.

2.5. UCE library preparation

Whole body DNA extracts (see above) were quantified using Qubit v4 (Thermofisher Scientific) and 50 ng DNA per specimen were sonicated to 500 bp fragment length using a Bioruptor ultrasonicator (Diagenode). The dual indexed Illumina compatible libraries were constructed using a Kapa Hyper prep kit (Roche) using one fourth of the manufacturer's recommended volumes (as described in Branstetter, Longino, Ward, & Faircloth, 2017). PCR amplifications were performed in the recommended volumes. Individual sample libraries were quantified on the Qubit and 12 specimens were equimolarly pooled. Libraries were UCE enriched using the Hymenopteran v2 hybridization kit (UCE Hymenoptera 2.5Kv2 Principal/Full, myBaits, Arborbiosci). Each enrichment was performed on a single pool of 12 individuals using 500 ng. The enrichment protocol followed the manufacturer's recommendations with a hybridization step of 24 h at 65°C, followed by a PCR amplification of 14 steps. Pools were sequenced on a Miseq (Illumina) using the Illumina v3 kit (2 x 300 bp; Illumina, location, Switzerland).

2.6. Bioinformatic processing of UCE data

Demultiplexed data from three Miseq runs were merged and mainly processed using *PHYLUCE* tools (Faircloth, 2016). Bioinformatic steps followed the process described in chapter 2 (Gueuning, Frey and Praz, in preparation) but using *SPAdes* (Nurk et al., 2013) instead of *Velvet* (Zerbino & Birney, 2008) for assembly. Briefly, raw data were cleaned using *illumiprocessor* (Faircloth, 2016) and assembled with *SPAdes* 3.8.0 with the single cell and careful options activated. Contigs were mapped with *Lastz* (Harris, 2007) against the Hymenopteran v2 UCE reference file and matching reads were extracted and aligned using *MAFFT* (Katoh & Standley, 2013). After edge-trimming aligned sequences (Faircloth, 2016), alignments with less than 75% of the maximum number of specimens sharing a locus were filtered out. Alignments passing filters were concatenated and additionally filtered out if harbouring more than 90% missing data.

2.7. UCE phylogeny

Concatenated contigs were imported into Geneious v11.0.5 and phylogenetic trees were constructed using the same parameters as for the COI datasets.

2.8. Principal component analyses

Principal component analyses (PCAs) were performed on the concatenated UCE reads separately for each species-complex. Concatenated reads were imported into R using the *adegenet* package (Jombart, 2008) and screened for binary SNPs which were extracted and converted into a compatible format (i.e. "genlight"). After conversion, departure from the Hardy-Weinberg equilibrium was assessed using the *dartR* package (Gruber, Unmack, Berry, & Georges, 2018). PCA were performed using the *dudi.pca* function (*ade4* package; Dray & Dufour, 2007) without scaling or centering the data. PCA results were plotted with *ggplot2* (Wilkinson, 2011) using the two first components.

3. Results

3.1. UCE sequencing

The Miseq runs produced in total 116.6 million reads (per direction) with on average 251,579 (SD ± 354,102) reads per specimen. After assembling we recovered a mean of 7492 (SD ± 6335) loci per specimen. After filtering none UCE loci and filtering for a 75% matrix completeness, the number of retrieved loci varied between 686 and 1761 across the five species-complexes (Supplementary table S2). The resulting concatenated reads were on average 787,783 bp long; the shortest reads were obtained for *N. goodeniana/succincta* (i.e. 429,820 bp) and the longest for *A. barbareae/cineraria* (i.e. 1,103,816 bp). In total, seven specimens did not pass the 90% missing data filter (Supplementary Information S1) and were therefore excluded. No significant departure form Hardy-Weinberg equilibrium was observed. The biallelic SNPs screening recovered on average 59,769 biallelic SNPs (SD ± 78,870) across all species-complexes (Supplementary Table S2).

3.2. Mitochondrial-nuclear phylogenies

3.2.1. Andrena barbareae/cineraria

Mitochondrial and nuclear phylogenies were discordant with no clear separation between both species in the COI tree and two well supported monophyletic clades corresponding to the two morphological species in the UCE tree (100% bootstrap support values, hereafter BS values; Figure 2). Results from the UCE phylogenetic tree were corroborated by the PCA in which both species were clearly separated (Figure 3).

3.2.2. Andrena carantonica/trimmerana/rosae

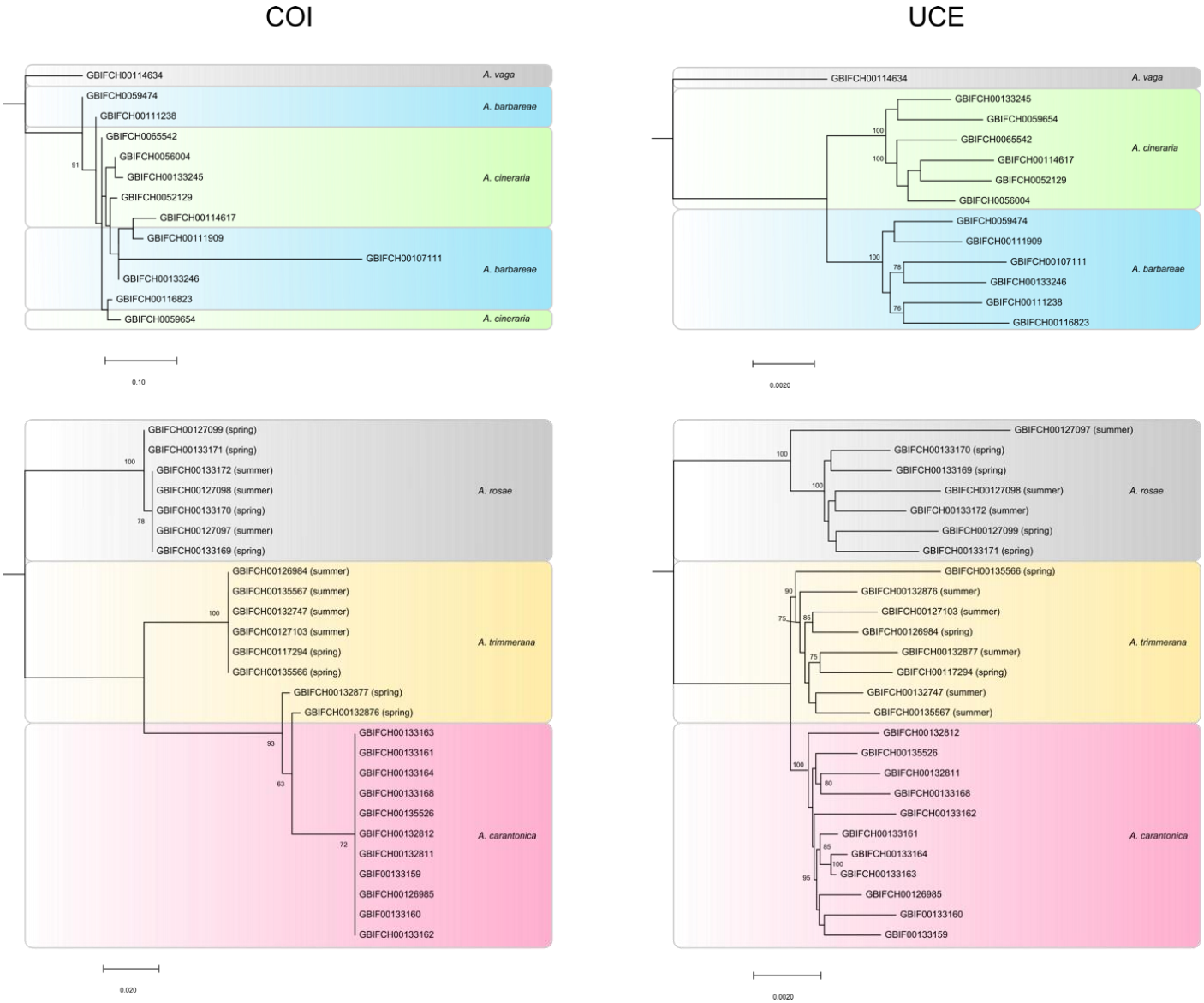
Mitochondrial phylogenies recovered well-supported (BS 72-100%) monophyletic clades for *A. carantonica* and *A. rosae*, but not for *A. trimmerana*, which was composed of two clades forming a paraphyletic unit from which *A. carantonica* arose; two specimens of *A. trimmerana* sampled in western Switzerland formed a clade that was sister to a clade containing all specimens of *A. carantonica*; support for this sister relationship was high (BS 93%). In contrast, all three species appeared as strongly supported monophyletic clades in the UCE tree (≥ 90% BS; Figure 2). Spring and summer generations of *A. rosae* and of *A. trimmerana* were intermixed in both mitochondrial and UCE trees, supporting the view that *A. stragulata* and *A. spinigera* constitute the morphologically differentiated spring generation of *A. rosae* and *A. trimmerana*, respectively.

3.2.3. Andrena dorsata/propinqua

Strong mito-nuclear discordances were detected within this species-complex. In mitochondrial trees, Swiss specimens formed two clusters corresponding to morphological identifications (Fig. 2), but one specimen of *A. propinqua* (GBIFCH00133244) collected in southern France was sister to a well-supported clade containing all other specimens of *A. dorsata* and of *A. propinqua* (Fig. 2). Our sampling also included one specimen of *A. dorsata* from this French site (GBIFCH00133243). Phylogenetic trees and PCAs based on UCEs recovered both species as separated clusters (Fig. 2, 3); the French specimens were not particularly divergent. Taken together, these results indicate that *A. dorsata* and *A. propinqua* are valid species.

3.2.4. Lasioglossum alpigenum/bavaricum/cupromicans

Comparison of mitochondrial and nuclear trees revealed mitochondrial-nuclear discordance for *L. bavaricum* and *L. cupromicans* (Figure 2). Both taxa were well delimited with highly supported monophyletic clades (95% bootstrap value) in the UCE tree but appeared to share the same COI barcode. All *L. alpigenum* specimens clustered in a single monophyletic clade sister to both other taxa in both trees. Our UCE results indicate that all three species are distinct as previously postulated based on morphology.



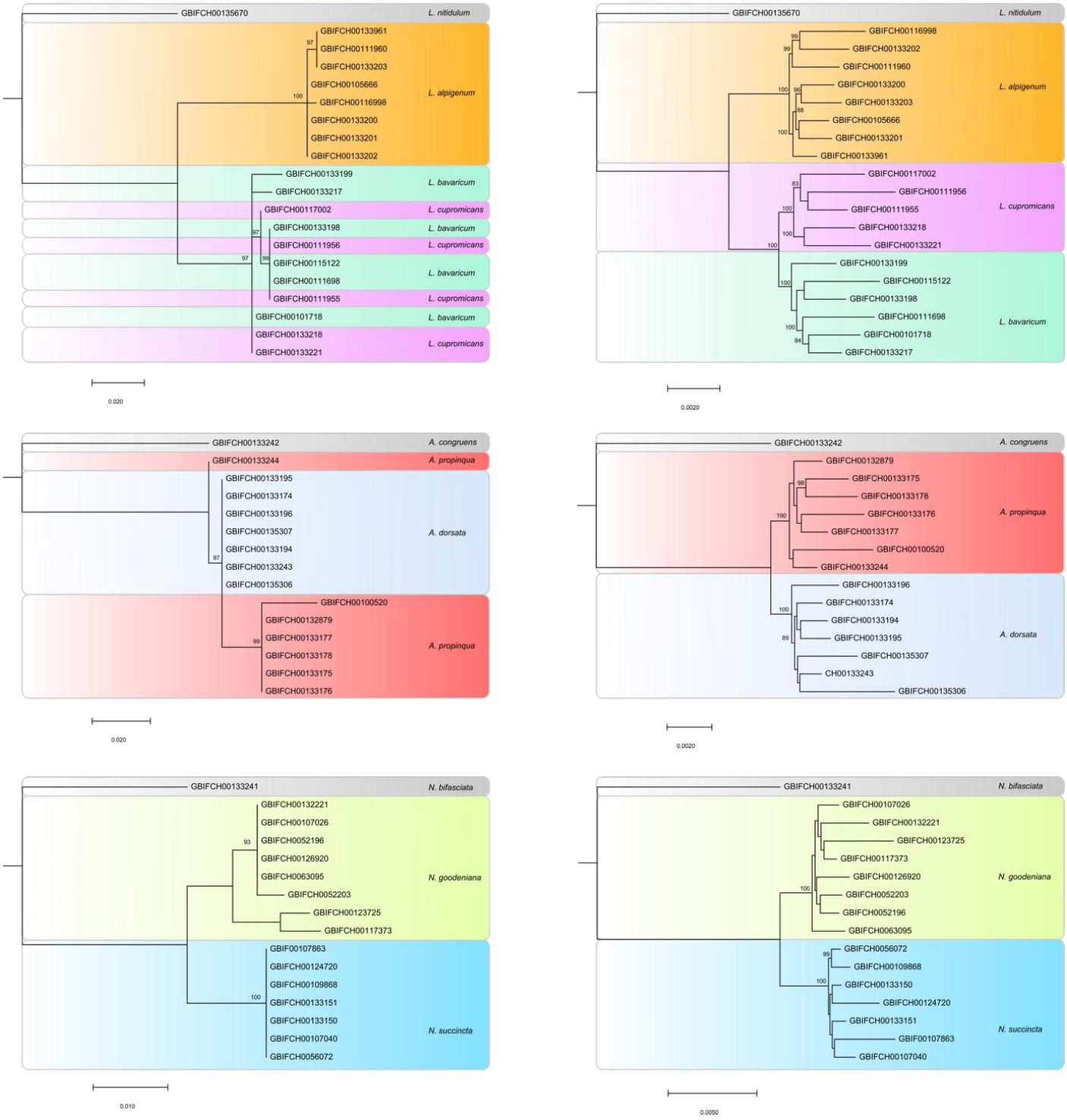


Figure 2: Phylogenetic trees for COI and UCE datasets per species-complex. Trees were built using *RAxML* with the GTR GAMMA model and 100 bootstrap replicates. Only bootstrap probabilities higher than 70% are shown. Colour codes are consistent across sampling maps and principal component analyses.

4. Discussion

4.1. Ultraconserved elements overcome mitochondrial barcode limitations

In all examined cases, mitochondrial barcodes limitations render conclusions on species delimitation difficult if used alone. To cope with identification errors and more generally with mitochondrial-morphological discordance, many studies have underlined the necessity to cross-validate COI-barcoding with nuclear genes (e.g. Dupuis, Roe, & Sperling, 2012). In this study we explored the potential of ultraconserved elements (UCEs) to provide a robust tool for species delimitation. Recently, a few studies have used UCEs for identifying cryptic diversity. Although this targeted-capture based method was initially developed to resolve deep and difficult phylogenies (Branstetter et al., 2017; Crawford et al., 2012), UCEs can also provide answers at shallower time scale (Harvey, Smith, Glenn, Faircloth, & Brumfield, 2016; Zarza et al., 2018). Indeed, because the variability within UCEs increased with increasing physical distance from the core, UCE can be applied to different time scales (Faircloth et al., 2012). In this study, UCEs recovered thousands of homologous nuclear loci and provided clear genetic signals for all investigated species-groups. Because of the universality of the probes, species belonging to three major bee families could be processed simultaneously using the same protocol, which would be nearly impossible with most other multigenic nuclear DNA markers. Our study provides strong evidence that all examined species are valid species in spite of difficult morphological identifications or mitochondrial barcode sharing.

4.2. Barcode sharing

Barcode sharing is a well-known issue in molecular diagnostics although it remains relatively rare among insects (Hebert, Dewaard, & Landry, 2010; but see Nicholls, Challis, Mutun, & Stone, 2012). Several large scale studies have reported barcode sharing rates mostly under 3% [i.e. lepidopterans (Ashfaq, Akhtar, Rafi, Mansoor, & Hebert, 2017; Hausmann et al., 2013; Huemer, Mutanen, Sefc, & Hebert, 2014); coleopterans (Hendrich et al., 2015; Pentinsaari, Hebert, & Mutanen, 2014)], a number slightly lower than that reported for the Apoidea fauna (see above). This difference in rates could possibly be linked to the tediousness of morphological identification of certain taxa, a common error source in barcoding studies (Mutanen et al., 2016). Barcode sharing is often linked to mitochondrial introgression or incomplete lineage sorting. Introgression events are a result from inter-specific hybridization followed by the fixation of foreign DNA. Because mitochondrial DNA are only transmitted by females and thus experience less gene flow than nuclear DNA, introgression levels are typically higher than for nuclear DNA (Currat, Ruedi, Petit, & Excoffier, 2008). Furthermore, the effect of mitochondrial introgression on phylogenies is exacerbated by the reduced level of recombination in maternally inherited organelles (Barr, Neiman, & Taylor, 2005; Funk & Omland, 2003). In most extreme cases, no or very little nuclear introgression can be observed despite a complete mitochondrial exchange (Good, Vanderpool, Keeble, & Bi, 2015; Zieliński et al., 2013). The second biological cause for barcode sharing is incomplete lineage sorting among recently isolated species.

Distinguishing both forces in the case of barcode sharing is trivial. Indeed, since mitogenomes are haploid and most often uniparentally inherited (Ladoukakis & Zouros, 2017), they have a four times smaller effective population size then nuclear genomes. The nuclear genomes will thus take four times longer to reflect lineage limits and history (Barrowclough & Zink, 2009; McKay & Zink, 2010; Zink & Barrowclough, 2008). Incomplete lineage sorting is thus expected to be more pronounced in nuclear than mitochondrial phylogenies. Consequently, in the case of barcode sharing due to incomplete lineage sorting, no phylogenetic structure at the nuclear level should be observed. In contrast, for the species-complexes *A. barbareae*/*cineraria* and *L.*

bavaricum/*cupromicans*, the clear phylogenetic separation at the nuclear level between species pairs appear to represent clear cases of mitochondrial introgression.

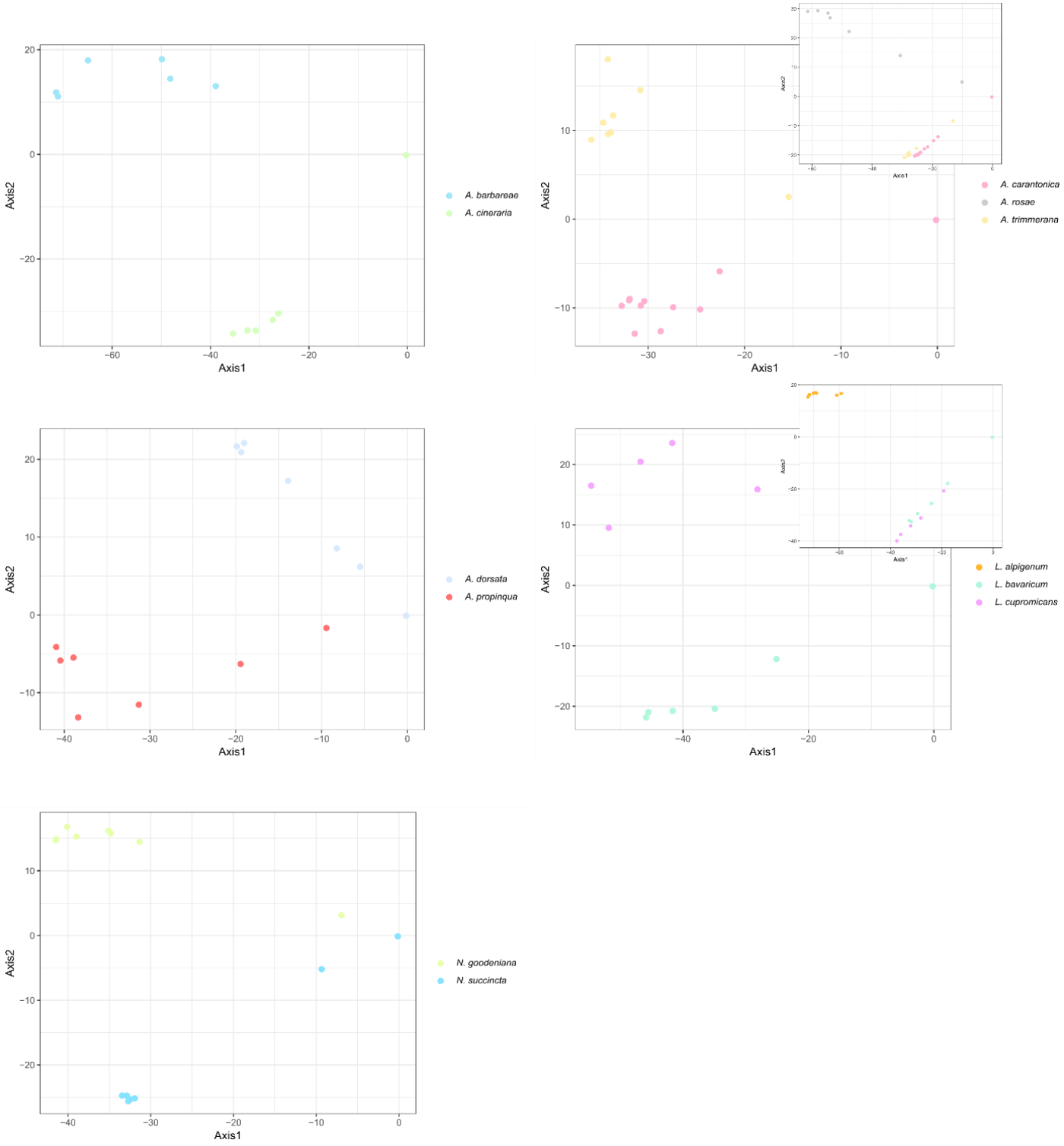


Figure 3: Principal component analyses (PCA) per species-complex. PCAs were computed based on the UCE datasets using the *adegenet* and *ade4* R package. Analyses were performed without outgroup. For species-groups composed of three species, the analysis was run twice: once with all three species, and once without the most divergent species.

4.3. Mitochondrial paraphyly

Our study additionally reveals two new and remarkable cases of mitochondrial paraphyly within a single biological species. In *A. trimmerana*, the two divergent mitochondrial sequences were found in sympatry, while in *A. propinqua*, the divergent mitochondrial haplotype was geographically separated. Such cases represent challenges that cannot be overcome with COI-based DNA species delimitations; in morphologically difficult groups, such as those examined here, morphology is also highly unlikely to resolve such cases, and having multi-locus nuclear markers is probably the only possibility to shed light on the status of these mitochondrial lineages.

Previous studies have suggested that mitochondrial paraphyly often originates from artefacts. First, putative paraphyly within a recognized species may in fact be due to the presence of more than one species. A first source of error is simply misidentification of some of the specimens (Mutanen et al., 2016). Second, the divergent mitochondrial lineages may in fact point to a hitherto unrecognized cryptic species (Dincă, Lukhtanov, Talavera, & Vila, 2011). In our case, both these situations appear unlikely, as in both cases the divergent specimens in mitochondrial trees were nested within presumed conspecific specimens in UCE trees.

One possible limitation of UCEs is their low rate of evolution. Could then the divergent mitochondrial haplotypes within *A. trimmerana* and *A. propinqua* represent cryptic species that diverged too recently to be picked up by the UCEs? In our study, UCEs were successful in recognizing species previously identified using morphological criteria, but largely failed to recover population-level divergences, such as isolation-by-distance. However, in *Andrena amieti* (Chapter 2), similar mito-nuclear discordance and mitochondrial paraphyly were uncovered; in this species, where the number of specimens and populations sampled was higher, isolation by distance was detected, leading us to strongly reject the presence of cryptic diversity in alpine populations of this taxon. Based on these results, we tentatively reject the hypothesis of additional cryptic diversity within *A. trimmerana* and *A. propinqua*.

4.4. Advantages and disadvantages of UCEs for species delimitation

Our study provides a convincing example of the utility of UCEs for species delimitation. We provide strong evidence that all species examined represent valid species, solving long-debated taxonomic questions. In all cases, mitochondrial barcodes provided useful data, but additionally exhibited distinct limitations, mainly caused by mitochondrial introgression and mitochondrial paraphyly. The former would also lead to erroneous barcode-based identifications, while the latter would still enable accurate specimen identifications in spite of the paraphyly.

The advantages of using UCEs for species delimitation can be summarized as follows. First, specimens with varying DNA quality can be used. Some species examined here are very rarely collected, such as males of the *Lasioglossum* taxa, or some *Andrena*, and we used all available specimens collected in Switzerland in the last 10 years. For these taxa, we mostly used pinned specimens, with high success. Obtaining ethanol-collected specimens for these, or other problematic bee taxa, would be very difficult. Second, existing datasets can be complemented progressively. In fact, sequencing was performed in two steps in our study, and specimen sampling was complemented during the second step based on the initial results obtained, which is exactly how taxonomy usually works. We are confident that our datasets will be complemented in the future by adding specimens from different regions. Last, the universality of the UCEs made it possible to process bees

belonging to three major bee families simultaneously, not only in the lab but also for bioinformatic analyses, which would be impossible with other methods.

An important question regarding the use of UCEs for delimitating closely related species is whether the captured variation will be sufficient to differentiate very recently separated species. We cannot exclude that all cases documented here may represent species that have diverged long ago, yet remained controversial because of difficult morphologies. Given the large number of open taxonomic questions in European bees, our study clearly demonstrates the huge potential UCEs for complementing COI-based DNA barcoding and morphological studies.

References

Amiet, F., Hermann, M., Müller, A., & Neumeyer, R. (2010). *Apidae 6, Andrena, Melliturga, Panurginus, Panurgus. Fauna Helvetica.* CSCF & SEG.

Amiet, F., Herrmann, M., Müller, A., & Neumeyer, R. (2001). *Apidae 3: Halictus, Lasioglossum. Fauna Helvetica, 208 pp.* CSCF & SEG.

Amiet, F., Herrmann, M., Müller, A., & Neumeyer, R. (2007). *Apidae 5: Ammobates, Ammobatoides, Anthophora, Biastes, Ceratina, Dasypoda, Epeoloides, Epeolus, Eucera, Macropis, Melecta, Melitta, Nomada, Pasites, Tetralonia, Thyreus, Xylocopa. Fauna Helvetica, 356 pp.* CSCF & SEG.

Andriollo, T., Naciri, Y., & Ruedi, M. (2015). Two mitochondrial barcodes for one biological species: The case of European Kuhl's pipistrelles (chiroptera). *PLoS ONE*. doi:10.1371/journal.pone.0134881

Ashfaq, M., Akhtar, S., Rafi, M. A., Mansoor, S., & Hebert, P. D. N. (2017). Mapping global biodiversity connections with DNA barcodes: Lepidoptera of Pakistan. *PLoS ONE*, *12*(3), 1–13. doi:10.1371/journal.pone.0174749

Barnosky, A. D., Matzke, N., Tomiya, S., Wogan, G. O. U., Swartz, B., Quental, T. B., … Ferrer, E. A. (2011). Has the Earth's sixth mass extinction already arrived? *Nature*, *471*(7336), 51–57. doi:10.1038/nature09678

Barr, C. M., Neiman, M., & Taylor, D. R. (2005). Inheritance and recombination of mitochondrial genomes in plants, fungi and animals. *New Phytologist*, *168*(1), 39–50. doi:10.1111/j.1469-8137.2005.01492.x

Barrowclough, G. F., & Zink, R. M. (2009). Funds enough, and time: mtDNA, nuDNA and the discovery of divergence. *Molecular Ecology*, *18*(14), 2934–2936. doi:10.1111/j.1365-294X.2009.04271.x

Branstetter, M. G., Longino, J. T., Ward, P. S., & Faircloth, B. C. (2017). Enriching the ant tree of life: enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. *Methods in Ecology and Evolution*, *8*(6), 768–776. doi:10.1111/2041-210X.12742

Brunner, P. C., Fleming, C., & Frey, J. E. (2002). A molecular identification key for economically important thrips species (Thysanoptera: Thripidae) using direct sequencing and a PCR-RFLP-based approach. *Agricultural and Forest Entomology*, *4*(2), 127–136. doi:10.1046/j.1461-9563.2002.00132.x

Cardoso, P., Erwin, T. L., Borges, P. A. V., & New, T. R. (2011). The seven impediments in invertebrate conservation and how to overcome them. *Biological Conservation*, *144*(11), 2647–2655. doi:10.1016/j.biocon.2011.07.024

Ceballos, G., Ehrlich, P. R., & Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the National Academy of Sciences*, *114*(30), E6089–E6096. doi:10.1073/pnas.1704949114

Crawford, N. G., Faircloth, B. C., McCormack, J. E., Brumfield, R. T., Winker, K., & Glenn, T. C. (2012). More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology Letters*, *8*(5), 783–786. doi:10.1098/rsbl.2012.0331

Creedy, T. J., Norman, H., Tang, C. Q., Qing Chin, K., Andujar, C., Arribas, P., … Vogler, A. P. (2019). *A validated workflow for rapid taxonomic assignment and monitoring of a national fauna of bees (Apiformes) using high throughput DNA barcoding . Molecular Ecology Resources*. doi:10.1111/1755-0998.13056

Currat, M., Ruedi, M., Petit, R. J., & Excoffier, L. (2008). The hidden side of invasions: Massive introgression by local genes. *Evolution*, *62*(8), 1908–1920. doi:10.1111/j.1558-5646.2008.00413.x

Dellicour, S., & Flot, J. F. (2018). The hitchhiker's guide to single-locus species delimitation. *Molecular Ecology Resources*. doi:10.1111/1755-0998.12908

Diestelhorst, O., & Lunau, K. (2008). Contribution to the clarification of the species status of Nomada

goodeniana. *Entomologie Heute*, 165–171.

Dincă, V., Lukhtanov, V. A., Talavera, G., & Vila, R. (2011). Unexpected layers of cryptic diversity in wood white Leptidea butterflies. *Nature Communications*, *2*(1). doi:10.1038/ncomms1329

Dray, S., & Dufour, A. B. (2007). The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software*. doi:10.18637/jss.v022.i04

Dupuis, J. R., Roe, A. D., & Sperling, F. A. H. (2012). Multi-locus species delimitation in closely related animals and fungi: One marker is not enough. *Molecular Ecology*, *21*(18), 4422–4436. doi:10.1111/j.1365-294X.2012.05642.x

Edmer, A. W. (1970). Die Bienen des Genus Halictus LATR. S.L. im Großraum von Linz (Hymenoptera, Apidae) Teil II. *Naturk.Jb.Stadt Linz*, 19–82.

Faircloth, B. C. (2016). PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*, *32*(5), 786–788. doi:10.1093/bioinformatics/btv646

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, *61*(5), 717–26. doi:10.1093/sysbio/sys004

Falk, S. (2016). Field Guide to the Bees of Great Britain and Ireland. *Bee World*, *93*, 85. doi:10.1080/0005772X.2016.1257474

Frey, J., Guillén, L., Frey, B., Samietz, J., Rull, J., & Aluja, M. (2013). Developing diagnostic SNP panels for the identification of true fruit flies (Diptera: Tephritidae) within the limits of COI-based species delimitation. *Bmc Evolutionary Biology*, *13*(1), 1–19. doi:10.1186/1471-2148-13-106

Funk, D. J., & Omland, K. E. (2003). Species-Level Paraphyly and Polyphyly: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics*, *34*(1), 397–423. doi:10.1146/annurev.ecolsys.34.011802.132421

Good, J. M., Vanderpool, D., Keeble, S., & Bi, K. (2015). Negligible nuclear introgression despite complete mitochondrial capture between two species of chipmunks. *Evolution*, *69*(8), 1961–1972. doi:10.1111/evo.12712

Gruber, B., Unmack, P. J., Berry, O. F., & Georges, A. (2018). dartr: An r package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Molecular Ecology Resources*. doi:10.1111/1755-0998.12745

Gueuning, M., Ganser, D., Blaser, S., Albrecht, M., Knop, E., Praz, C., & Frey, J. E. (2019). Evaluating next-generation sequencing (NGS) methods for routine monitoring of wild bees: Metabarcoding, mitogenomics or NGS barcoding. *Molecular Ecology Resources*, *19*(4), 847–862. doi:10.1111/1755-0998.13013

Gusenleitner, F., & Schwarz, M. (2002). Weltweite Checkliste der Bienengattung Andrena mit Bemerkungen und Ergänzungen zu paläarktischen Arten (Hymenoptera, Apidae, Andreninae, Andrena). *Entomofauna Supplement*, *12*, 1–1280.

Harris, R. S. (2007). *Improved Pairwise Alignment of Genomic DNA. PhD thesis, Pennsylvania State University. PhD thesis, Pennsylvania State Univ. (2007)*. doi:/10.1016/j.brainres.2008.03.070

Harvey, M. G., Smith, B. T., Glenn, T. C., Faircloth, B. C., & Brumfield, R. T. (2016). Sequence Capture versus Restriction Site Associated DNA Sequencing for Shallow Systematics. *Systematic Biology*. doi:10.1093/sysbio/syw036

Hausmann, A., Charles, H., Godfray, J., Huemer, P., Mutanen, M., Rougerie, R., … Hebert, P. D. N. (2013). Genetic patterns in European geometrid moths revealed by the Barcode Index Number (BIN) system. *PLoS ONE*, *8*(12), 1–11. doi:10.1371/journal.pone.0084518

Hebert, P. D. N., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, *270*(1512), 313–321. doi:10.1098/rspb.2002.2218

Hebert, P. D. N., Dewaard, J. R., & Landry, J. F. (2010). DNA barcodes for 1/1000 of the animal Kingdom. *Biology Letters*, *6*(3), 359–362. doi:10.1098/rsbl.2009.0848

Hendrich, L., Morinière, J., Haszprunar, G., Hebert, P. D. N., Hausmann, A., Köhler, F., & Balke, M. (2015). A comprehensive DNA barcode database for Central European beetles with a focus on Germany: Adding more than 3500 identified species to BOLD. *Molecular Ecology Resources*, *15*(4), 795–818. doi:10.1111/1755-0998.12354

Hinojosa, J. C., Koubínová, D., Szenteczki, M. A., Pitteloud, C., Dincă, V., Alvarez, N., & Vila, R. (2019). A mirage of cryptic species: Genomics uncover striking mitonuclear discordance in the butterfly Thymelicus sylvestris. *Molecular Ecology*, mec.15153. doi:10.1111/mec.15153

Huemer, P., Mutanen, M., Sefc, K. M., & Hebert, P. D. N. (2014). Testing DNA barcode performance in 1000 species of European Lepidoptera: Large geographic distances have small genetic impacts. *PLoS ONE*, *9*(12), 1–21. doi:10.1371/journal.pone.0115774

Jombart, T. (2008). Adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*. doi:10.1093/bioinformatics/btn129

Karanovic, T., Djurakic, M., & Eberhard, S. M. (2016). Cryptic Species or Inadequate Taxonomy? Implementation of 2D Geometric Morphometrics Based on Integumental Organs as Landmarks for Delimitation and Description of Copepod Taxa. *Systematic Biology*. doi:10.1093/sysbio/syv088

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*. doi:10.1093/molbev/mst010

Klopfstein, S., Kropf, C., & Baur, H. (2016). Wolbachia endosymbionts distort DNA barcoding in the parasitoid wasp genus Diplazon (Hymenoptera: Ichneumonidae). *Zoological Journal of the Linnean Society*, *177*(3), 541–557. doi:10.1111/zoj.12380

Kuhlmann, M. (1997). Zum taxonomischen Status von Nomada goodeniana (KlRBY, 1802) und Nomada succincta PANZER, 1798 (Hymenoptera, Apidae). *Entomofauna*, *18*(32), 521–528.

Ladoukakis, E. D., & Zouros, E. (2017). Evolution and inheritance of animal mitochondrial DNA: rules and exceptions. *Journal of Biological Research-Thessaloniki*, *24*(1), 2. doi:10.1186/s40709-017-0060-4

Lemopoulos, A., Prokkola, J. M., Uusi-Heikkilä, S., Vasemägi, A., Huusko, A., Hyvärinen, P., … Vainikka, A. (2019). Comparing RADseq and microsatellites for estimating genetic diversity and relatedness — Implications for brown trout conservation. *Ecology and Evolution*. doi:10.1002/ece3.4905

Leneveu, J., Chichvarkhin, A., & Wahlberg, N. (2009). Varying rates of diversification in the genus Melitaea (Lepidoptera: Nymphalidae) during the past 20 million years. *Biological Journal of the Linnean Society*. doi:10.1111/j.1095-8312.2009.01208.x

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., … Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, *10*(1), 34. doi:10.1186/1742-9994-10-34

Martinet, B., Lecocq, T., Brasero, N., Biella, P., UrbanovÁ, K., ValterovÁ, I., … Rasmont, P. (2018). Following the cold: geographical differentiation between interglacial refugia and speciation in the arcto-alpine species complex Bombus monticola (Hymenoptera: Apidae). *Systematic Entomology*. doi:10.1111/syen.12268

McKay, B. D., & Zink, R. M. (2010). The causes of mitochondrial DNA gene tree paraphyly in birds.

Molecular Phylogenetics and Evolution, *54*(2), 647–650. doi:10.1016/j.ympev.2009.08.024

McKendrick, L., Provan, J., Fitzpatrick, Ú., Brown, M. J. F., Murray, T. E., Stolle, E., & Paxton, R. J. (2017). Microsatellite analysis supports the existence of three cryptic species within the bumble bee Bombus lucorum sensu lato. *Conservation Genetics*. doi:10.1007/s10592-017-0965-3

Mutanen, M., Kivelä, S. M., Vos, R. A., Doorenweerd, C., Ratnasingham, S., Hausmann, A., … Godfray, H. C. J. (2016). Species-Level Para- and Polyphyly in DNA Barcode Gene Trees: Strong Operational Bias in European Lepidoptera. *Systematic Biology*, *65*(6), 1024–1040. doi:10.1093/sysbio/syw044

Neumeyer, R., Baur, H., Guex, G. D., & Praz, C. (2014). A new species of the paper wasp genus Polistes (Hymenoptera, Vespidae, Polistinae) in Europe revealed by morphometrics and molecular analyses. *ZooKeys*. doi:10.3897/zookeys.400.6611

Nicholls, J. A., Challis, R. J., Mutun, S., & Stone, G. N. (2012). Mitochondrial barcodes are diagnostic of shared refugia but not species in hybridizing oak gallwasps. *Molecular Ecology*. doi:10.1111/j.1365-294X.2012.05683.x

Nichols, H. J., Jordan, N. R., Jamie, G. A., Cant, M. A., & Hoffman, J. I. (2012). Fine-scale spatiotemporal patterns of genetic variation reflect budding dispersal coupled with strong natal philopatry in a cooperatively breeding mammal. *Molecular Ecology*, *21*(21), 5348–5362. doi:10.1111/mec.12015

Nurk, S., Bankevich, A., Antipov, D., Gurevich, A., Korobeynikov, A., Lapidus, A., … Pevzner, P. A. (2013). Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads BT - Research in Computational Molecular Biology. In M. Deng, R. Jiang, F. Sun, & X. Zhang (Eds.) (pp. 158–170). Berlin, Heidelberg: Springer Berlin Heidelberg.

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

Pentinsaari, M., Hebert, P. D. N., & Mutanen, M. (2014). Barcoding beetles: A regional survey of 1872 species reveals high identification success and unusually deep interspecific divergences. *PLoS ONE*, *9*(9). doi:10.1371/journal.pone.0108651

Pievani, T. (2014). The sixth mass extinction: Anthropocene and the human impact on biodiversity. *Rendiconti Lincei*, *25*(1), 85–93. doi:10.1007/s12210-013-0258-9

Ratnasingham, S., & Hebert, P. D. N. (2013). A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. *PLoS ONE*, *8*(7). doi:10.1371/journal.pone.0066213

Reemer, M., Groenenberg, D. S. J., Van Achterberg, C., & Peeters, T. M. J. (2008). Taxonomic assessment of Andrena rosae and a stragulata by DNA-sequencing (Hymenoptera: Apoidea: Andrenidae). *Entomologia Generalis*.

Schmid-Egger, C., & Scheuchl, E. (1997). llustrierte Bestimmungstabellen der Wildbienen Deutschlands und Österreichs unter Berücksichtigung der Arten der Schweiz. *Monografien Entomologie Hymenoptera*, *0030*, 1–180.

Schmid, S., Genevest, R., Gobet, E., Suchan, T., Sperisen, C., Tinner, W., & Alvarez, N. (2017). HyRAD-X, a versatile method combining exome capture and RAD sequencing to extract genomic information from ancient DNA. *Methods in Ecology and Evolution*. doi:10.1111/2041-210X.12785

Schmidt, S., Schmid-Egger, C., Morinière, J., Haszprunar, G., & Hebert, P. D. N. (2015). DNA barcoding largely supports 250 years of classical taxonomy: identifications for Central European bees (Hymenoptera, Apoidea partim). *Molecular Ecology Resources*, *15*(4), 985–1000. doi:10.1111/1755-0998.12363

Sheffield, C. S., Hebert, P. D. N., Kevan, P. G., & Packer, L. (2009). DNA barcoding a regional bee (Hymenoptera: Apoidea) fauna and its potential for ecological studies. *Molecular Ecology Resources*,

9(SUPPL. 1), 196–207. doi:10.1111/j.1755-0998.2009.02645.x

Sheffield, C. S., Heron, J., Gibbs, J., Onuferko, T. M., Oram, R., Best, L., … Rowe, G. (2017). Contribution of DNA barcoding to the study of the bees (Hymenoptera: Apoidea) of Canada: Progress to date. *Canadian Entomologist*, *149*(6), 736–754. doi:10.4039/tce.2017.49

Soltani, G. G., Bénon, D., Alvarez, N., & Praz, C. J. (2017). When different contact zones tell different stories: Putative ring species in the Megachile concinna species complex (Hymenoptera: Megachilidae). *Biological Journal of the Linnean Society*. doi:10.1093/biolinnean/blx023

Song, H., Buhay, J. E., Whiting, M. F., & Crandall, K. a. (2008). Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(36), 13486–13491. doi:10.1073/pnas.0803076105

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. doi:10.1093/bioinformatics/btu033

Suchan, T., Pitteloud, C., Gerasimova, N. S., Kostikova, A., Schmid, S., Arrigo, N., … Alvarez, N. (2016). Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on collection specimens. *PLoS ONE*. doi:10.1371/journal.pone.0151651

Tang, M., Hardman, C. J., Ji, Y., Meng, G., Liu, S., Tan, M., … Yu, D. W. (2015). High-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods in Ecology and Evolution*, *6*(9), 1034–1043. doi:10.1111/2041-210X.12416

Toews, D. P. L., & Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*, *21*(16), 3907–3930. doi:10.1111/j.1365-294X.2012.05664.x

van der Meer, F., Reemer, M., Peeters, T. M. J., & Neve, A. (2006). De roodzandbij Andrena rosae in de Zuid-Hollandse Biesbosch (Hymenoptera: apoidea: andrenidae). *Nederlandse Faunistische Mededelingen / Nationaal Natuurhistorisch Museum*.

Westrich, P. (2014). Beitrag zur Diskussion über den taxonomischen Status von Andrena rosae Panzer 1801 (Hymenoptera, Apidae. *Eucera*.

White, D. J., Wolff, N. J., Pierson, M., & Gemmell, N. J. (2008). Revealing the hidden complexities of mtDNA inheritance. *Molecular Ecology*, *17*(23), 4925–4942. doi:10.1111/j.1365-294X.2008.03982.x

Whitworth, T. L., Dawson, R. D., Magalon, H., & Baudry, E. (2007). DNA barcoding cannot reliably identify species of the blowfly genus Protocalliphora (Diptera: Calliphoridae). *Proceedings of the Royal Society B: Biological Sciences*. doi:10.1098/rspb.2007.0062

Wilkinson, L. (2011). ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H. *Biometrics*. doi:10.1111/j.1541-0420.2011.01616.x

Williams, K. A., Lelej, A. S., & Thaochan, N. (2019). New species of Myrmosinae (Hymenoptera: Mutillidae) from Southeastern Asia. *Zootaxa*, *4656*(3), 525–534. doi:10.11646/zootaxa.4656.3.9

Zarza, E., Connors, E. M., Maley, J. M., Tsai, W. L. E., Heimes, P., Kaplan, M., & McCormack, J. E. (2018). Combining ultraconserved elements and mtDNA data to uncover lineage diversity in a Mexican highland frog (Sarcohyla ; Hylidae). *PeerJ*, *6*, e6045. doi:10.7717/peerj.6045

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. doi:10.1101/gr.074492.107

Zieliński, P., Nadachowska-Brzyska, K., Wielstra, B., Szkotak, R., Covaciu-Marcov, S. D., Cogălniceanu, D., & Babik, W. (2013). No evidence for nuclear introgression despite complete mtDNA replacement in the Carpathian newt (Lissotriton montandoni). *Molecular Ecology*, *22*(7), 1884–1903. doi:10.1111/mec.12225

Zink, R. M., & Barrowclough, G. F. (2008). Mitochondrial DNA under siege in avian phylogeography. *Molecular Ecology*, *17*(9), 2107–2121. doi:10.1111/j.1365-294X.2008.

Supplementary Information

S1: Metada for selected specimens. Outgroups for each species-complex are marked with an asterisk (*) next to the species name. Specimens that did not pass bioinformatic filters are highlighted in bold.

ID	Species	Date	Locality	Country	Sex	Lat	Long
GBIFCH00114634	Andrena vaga*	3/4/2016	Safiental	CH	F	46.8062	9.3270
GBIFCH00107111	Andrena barbareae	13/7/2017	Ernen	CH	F	46.3833	8.1357
GBIFCH00111238	Andrena barbareae	5/7/2017	Goms	CH	F	46.4660	8.2246
GBIFCH00111909	Andrena barbareae	6/7/2017	Zermatt	CH	F	46.0052	7.6791
GBIFCH00116823	Andrena barbareae	12/6/2017	Liddes	CH	F	46.0123	7.1788
GBIFCH00133246	Andrena barbareae	13/6/2014	Naters	CH	F	46.3161	7.9239
GBIFCH0059474	Andrena barbareae	17/5/2016	Eggerberg	CH	F	46.3126	7.8979
GBIFCH00114617	Andrena cineraria	3/4/2016	Safiental	CH	M	46.8061	9.3302
GBIFCH00133245	Andrena cineraria	8/4/2018	Ferreyres	CH	F	46.6671	6.4798
GBIFCH0052129	Andrena cineraria	14/4/2013	Haldenstein	CH	F	46.8639	9.5028
GBIFCH0056004	Andrena cineraria	9/4/2015	Tamins	CH	M	46.8393	9.4292
GBIFCH0059654	Andrena cineraria	24/6/2016	Täsch	CH	F	46.0362	7.8374
GBIFCH0065542	Andrena cineraria	25/4/2014	Lumnezia	CH	F	46.7192	9.1836
GBIFCH00107127	**Andrena rosae (summer)***	**19/7/2017**	**Wangen**	**CH**	**F**	**47.2016**	**8.8921**
GBIFCH00127097	Andrena rosae (summer)*	3/7/2015	Haute-Sorne	CH	F	47.3032	7.2072
GBIFCH00127098	Andrena rosae (summer)*	3/7/2015	Haute-Sorne	CH	M	47.3032	7.2072
GBIFCH00133172	Andrena rosae (summer)*	25/7/2016	Pieterien	CH	F	47.1685	7.3077
GBIFCH00127099	Andrena rosae (spring)*	10/4/2015	Haute-Sorne	CH	F	47.3032	7.2116
GBIFCH00133169	Andrena rosae (spring)*	20/4/2015	Glovelier	CH	F	47.3408	7.1990
GBIFCH00133170	Andrena rosae (spring)*	20/4/2015	Glovelier	CH	F	47.3408	7.1990
GBIFCH00133171	Andrena rosae (spring)*	20/1/2015	Glovelier	CH	F	47.3408	7.1990
GBIFCH00126985	Andrena carantonica	20/4/2016	Milvignes	CH	M	46.9619	6.8425
GBIFCH00132811	Andrena carantonica	6/5/2018	Cressier	CH	F	47.0567	7.0381
GBIFCH00132812	Andrena carantonica	6/5/2018	Cressier	CH	M	47.0567	7.0381
GBIFCH00133162	Andrena carantonica	20/4/2015	Glovelier	CH	F	47.3408	7.1990
GBIFCH00135525	**Andrena carantonica**	**6/5/2018**	**Ferreyres**	**CH**	**F**	**46.6648**	**6.4733**
GBIFCH00135526	Andrena carantonica	6/5/2018	La_Sarraz	CH	M	46.6670	6.4766
GBIF00133159	Andrena carantonica	30/4/2019	Ayent	CH	F	46.2730	7.4143
GBIF00133160	Andrena carantonica	30/4/2019	Ayent	CH	F	46.2730	7.4143
GBIFCH00133161	Andrena carantonica	30/4/2019	Ayent	CH	F	46.2730	7.4143
GBIFCH00133163	Andrena carantonica	30/4/2019	Ayent	CH	M	46.2708	7.4176
GBIFCH00133164	Andrena carantonica	29/3/2019	Ayent	CH	M	46.2730	7.4176
GBIFCH00133168	Andrena carantonica	20/4/2015	Glovelier	CH	M	47.3408	7.1990
GBIFCH00117294	Andrena trimmerana (spring)	10/4/2016	Astano	CH	M	46.0107	8.8023
GBIFCH00135566	Andrena trimmerana (spring)	8/4/2018	Ferreyres	CH	F	46.6626	6.4799
GBIFCH00132876	Andrena trimmerana (spring)	30/3/2019	Chancy	CH	F	46.1445	5.9842
GBIFCH00132877	Andrena trimmerana (spring)	30/3/2019	Chancy	CH	M	46.1445	5.9842
GBIFCH00126984	Andrena trimmerana (summer)	8/7/2016	Milvignes	CH	F	46.9642	6.8457
GBIFCH00127103	Andrena trimmerana (summer)	3/7/2015	Haute-Sorne	CH	F	47.3032	7.2116
GBIFCH00132747	Andrena trimmerana (summer)	18/7/2016	Pieterlen	CH	M	47.1808	7.3494
GBIFCH00135567	Andrena trimmerana (summer)	2/7/2018	Ferreyres	CH	F	46.6626	6.4799
GBIFCH00133242	Andrena congruens*	19/4/2018	Mollens	CH	M	46.3163	7.5209
GBIFCH00133174	Andrena dorsata	29/3/2019	Ayent	CH	M	46.2665	7.4192

GBIFCH00133194	Andrena dorsata	30/3/2019	Chancy	CH	M	46.1445	5.9842
GBIFCH00133195	Andrena dorsata	30/3/2019	Chancy	CH	M	46.1422	5.9778
GBIFCH00133196	Andrena dorsata	30/3/2019	Chancy	CH	M	46.1422	5.9810
GBIFCH00133243	Andrena dorsata	5/4/2015	Notre-Dame-De-Londres	FR	F	43.8400	3.7700
GBIFCH00135306	Andrena dorsata	1/4/2011	Leuk	CH	F	46.3161	7.6463
GBIFCH00135307	Andrena dorsata	11/7/2011	Leuk	CH	M	46.3161	7.6463
GBIFCH00100520	Andrena propinqua	25/6/2014	Russin	CH	F	46.1855	6.0252
GBIFCH00132879	Andrena propinqua	30/3/2019	Chancy	CH	F	46.1445	5.9842
GBIFCH00133175	Andrena propinqua	30/4/2019	Ayent	CH	M	46.2665	7.4192
GBIFCH00133176	Andrena propinqua	8/5/2015	Leuk	CH	F	46.3161	7.6463
GBIFCH00133177	Andrena propinqua	15/5/2018	Salgesch	CH	F	46.3027	7.5774
GBIFCH00133178	Andrena propinqua	15/5/2018	Salgesch	CH	F	46.3027	7.5774
GBIFCH00133244	Andrena propinqua	5/4/2015	Notre-Dame-De-Londres	FR	F	43.8400	3.7700
GBIFCH00135670	Lasioglossum nitidulum*	2/7/2018	La_Sarraz	CH	F	46.6671	6.4798
GBIFCH00105666	Lasioglossum alpigenum	12/8/2015	Emmetten	CH	M	46.9398	8.5498
GBIFCH00111960	Lasioglossum alpigenum	17/8/2017	Zermatt	CH	M	46.0007	7.6823
GBIFCH00116998	Lasioglossum alpigenum	5/8/2017	Bagnes	CH	M	45.9447	7.3629
GBIFCH00133200	Lasioglossum alpigenum	2/7/18	Fisetenpass	CH	F	46.8840	8.9327
GBIFCH00133201	Lasioglossum alpigenum	22/6/2017	Sufers	CH	F	46.5761	9.3646
GBIFCH00133202	Lasioglossum alpigenum	7/7/2016	Arolla	CH	F	45.9905	7.5075
GBIFCH00133203	Lasioglossum alpigenum	20/5/2018	Schwanden	CH	F	46.7594	8.0587
GBIFCH00133961	Lasioglossum alpigenum	17/8/2018	Davos	CH	M	46.7333	9.8512
GBIFCH00101718	Lasioglossum bavaricum	5/6/2015	Bagnes	CH	F	46.0133	7.3412
GBIFCH00111698	Lasioglossum bavaricum	27/7/2015	Riederalp	CH	M	46.3814	8.0214
GBIFCH00115122	Lasioglossum bavaricum	31/7/2017	Glarus_Süd	CH	M	46.9150	9.1305
GBIFCH00133198	Lasioglossum bavaricum	27/6/2018	Calfeisen	CH	F	46.9249	9.3541
GBIFCH00133199	Lasioglossum bavaricum	22/6/2017	Sufers	CH	F	46.5761	9.3646
GBIFCH00133217	Lasioglossum bavaricum	16/8/2018	Kandersteg	CH	M	46.5028	7.7096
GBIFCH00111955	Lasioglossum cupromicans	17/8/2017	Zermatt	CH	M	46.0052	7.6758
GBIFCH00111956	Lasioglossum cupromicans	17/8/2017	Zermatt	CH	M	46.0052	7.6758
GBIFCH00117002	Lasioglossum cupromicans	5/8/2017	Bagnes	CH	M	45.9469	7.3661
GBIFCH00133218	Lasioglossum cupromicans	10/7/2018	Kandersteg	CH	M	46.5028	7.7096
GBIFCH00133221	Lasioglossum cupromicans	10/7/2018	Kandersteg	CH	M	46.5028	7.7096
GBIFCH00133222	**Lasioglossum cupromicans**	**10/7/2018**	**Kandersteg**	**CH**	**M**	**46.5028**	**7.7096**
GBIFCH00133223	**Lasioglossum cupromicans**	**10/7/2018**	**Kandersteg**	**CH**	**M**	**46.5028**	**7.7096**
GBIFCH00133241	Nomada bifasciata*	8/4/2018	Ferreyres	CH	F	46.6603	6.4734
GBIFCH00107026	Nomada goodeniana	15/5/2017	Ernen	CH	F	46.3834	8.1292
GBIFCH00107027	Nomada goodeniana	15/5/2017	Ernen	CH	M	46.3834	8.1292
GBIFCH00117373	Nomada goodeniana	10/4/2016	Astano	CH	F	46.0107	8.7991
GBIFCH00123725	Nomada goodeniana	9/6/2017	Bedretto	CH	F	45.9058	8.5227
GBIFCH00126920	Nomada goodeniana	20/4/2016	Milvignes	CH	F	46.9619	6.8457
GBIFCH00132221	Nomada goodeniana	10/4/2017	Corsier-Sur-Vevey	CH	F	46.5056	6.8898
GBIFCH0052196	Nomada goodeniana	8/5/2013	Haldenstein	CH	F	46.8683	9.5095
GBIFCH0052203	Nomada goodeniana	8/5/2013	Haldenstein	CH	F	46.8660	9.5095
GBIFCH0063095	Nomada goodeniana	2/4/2014	Le_Landeron	CH	M	47.0640	7.0454
GBIF00107863	Nomada succincta	11/5/2018	Fully	CH	M	46.1692	7.1489
GBIFCH00107040	Nomada succincta	15/5/2017	Ernen	CH	M	46.3856	8.1325
GBIFCH00109868	Nomada succincta	11/5/2015	Lüen	CH	M	46.8331	9.6082
GBIFCH00117457	Nomada succincta	9/5/2017	Mendrisio	CH	F	45.9000	9.0122
GBIFCH00124720	Nomada succincta	30/4/2017	Embd	CH	F	46.2161	7.8387
GBIFCH00133150	Nomada succincta	30/4/2019	Ayent	CH	F	46.2685	7.4208

GBIFCH00133151	Nomada succincta	30/4/2019	Ayent	CH	F	46.2708	7.4176
GBIFCH0056072	Nomada succincta	11/5/2015	Tamins	CH	F	46.8365	9.4204
GBIFCH0074496	**Nomada succincta**	**21/5/2014**	**Osco**	**CH**	**M**	**46.4972**	**8.7621**

S2: UCE sequencing output

Species-complex	Total #Sp.	# Sp. ≥ 90% NA	# locus at 75% Compl.*	Length concat. locus	Mean locus length	Mean #Sp./locus	# SNPs
A. barbareae/cineraria	13	-	1671	1,103,816	617.78	10.6	30,249
A. carantonica/trimmerana/rosae	28	2	993	611,032	529.76	17.7	20,676
A. dorsata/proqinqua	15	-	1084	693,068	567.6	10.28	199,711
L. alpigenum/barbareae/cupromicans	22	2	1761	1,101,179	577.52	16.03	37,425
N. goodeniana/succincata	19	3	686	429,820	533.91	10.62	10,787

*Matrix completeness was determined upon the maximum number of specimen/locus and not total number of specimens.

– General discussion –

In spite of the importance of bees for ecosystems and humankind, currently very little is known on the population trends and conservation status for the majority of wild bee species. The main objective of this thesis was to evaluate different NGS methods as solutions for cost-effective monitoring (chapter I) and taxonomical impediments (chapter II and III). In chapter one, some NGS techniques were shown to outperform morphology for identifications in term of accuracy. Chapters 2 and 3 illustrate the huge step forward that NGS techniques represent for evaluating species boundaries in difficult species complexes. Together, our data illustrate how new techniques can assist taxonomists and ecologists in the field of conservation.

Benchmarking NGS methods for routine monitoring of insects

Presently, most monitoring schemes are still relying on morphological identification of individual specimens. This process is both demanding, as it requires expert taxonomic skills, and slow, making biomonitoring very costly and dramatically reducing the potential scale of coverage that could be reached with more cost-effective methods (Lebuhn et al., 2012). To overcome these limitations, several studies have recently contemplated NGS technologies to identify cost-effectively, rapidly and semi-automatically large numbers of specimens (Piper et al., 2019). However, despite serious efforts to improve NGS-based species identification methods, there are still open questions and major pitfalls to overcome.

In chapter I, we filled some important knowledge gaps concerning "next generation" monitoring, providing stakeholders and ecologists with a comprehensive overlook on the advantages and shortages of each method. To do so, we compared three of the most promising NGS methods (i.e. metabarcoding, mitogenomics and NGS barcoding) with respect to cost, workload and data-quality. Although numerous studies had already compared one or the other method, the novelty of this study lies in the comparison of three methods not only based on absolute mismatch values between NGS- and morphology-based identifications but also based on ecological patterns. Because NGS-based and morphology-based diversity assessments will always to some degree deviate from each other, it is essential to investigate the effect of these differences on the overall ecological patterns and conclusions.

In terms of data accuracy (i.e. species presence/absence and abundance), results showed that at least one NGS method (i.e. NGS barcoding) could provide accurate species presence/absence and abundance estimates. In terms of workload and cost, we found that gain in time and cost of using NGS methods was limited in our study. It is however important to stress that the dataset was collected in agricultural sites where the diversity of wild bees is well known and relatively low, which enables fast morphological identification. Studies focusing on more species-rich areas will likely face increased cost per specimen for morphological identifications. Similarly, studies targeting different taxonomic groups will require the input of several specialists. Although the morphological identification cost may not greatly vary among taxonomists or countries, the intervention of different taxonomists could substantially slow down the overall performance of monitoring programs. Likewise, for some speciose insect groups (e.g. Diptera, Coleoptera), morphological identifications are more difficult than for wild bees and may therefore be less reliable and costlier. Consequently, the cost for morphological identification in many monitoring projects should be higher than the one calculated for this study, which would make NGS-based monitoring schemes generally competitive.

Next generation monitoring

Based on the results of chapter I, we would encourage stakeholders and ecologists to take the turn and adopt NGS barcoding identifications into monitoring schemes. Indeed, this method provided excellent results in

terms of presence/absence and absolute abundance. Furthermore, NGS barcoding has the advantage of potentially linking DNA sequences with preserved voucher specimens, which enable morphological re-examination and will thus produce verifiable records. This method offers thus the opportunity to make a "soft" transition from classical morphology-based identification to NGS-based identification. Although NGS barcoding does not drastically reduce identification cost (see comment above), we predict that the reduction in sequencing cost will render the NGS method more cost-effective than the morphology method, for which the cost will remain fairly constant.

Currently, long-term surveys for the majority of insect taxa are lacking in most European countries. Changes in abundance and distribution of insects are generally estimated based on indicator groups (i.e. species reflecting a specific environmental condition). In Switzerland, for instance, the national Broad-based long-term survey (BDM) focuses on surveying well-known taxa like vascular plants, butterflies and breeding birds. For insects, butterflies are often the only group for which accurate measures of change can be obtained, mainly because they are well-known (Thomas, 2005). Furthermore, monitoring butterflies is cost-effective because data are based upon standardized transect counts of adults. For bees, as for many other taxa, long-terms monitoring studies will have to rely on more laborious sampling designs, like pan trapping. Pan traps have the advantage to provide quantitative and reproducible data necessary to assess population trends. They are therefore the most robust, replicable and standardized method for monitoring wild bees in European agriculture and grassland habitats (Lebuhn et al., 2012; Westphal et al., 2008). Nevertheless, pan trapping is labour intensive because it requires visiting sites twice in a short interval (to place and collect the traps). Pan trapping also renders the morphological identification more tedious and time demanding, as specimens are mingled and left in the soapy water for at least 24 hours which may jeopardize morphological integrity. Furthermore, specimens need to be sorted and prepared before specialists can perform the identifications, which is much more laborious than for specimens collected by sweep netting. For such datasets, NGS barcoding could greatly simplify identifications.

Besides, more and more programs start incorporating measurements of genetic diversity. For instance, in the Swiss BDM it is stated that: "*To attain the overriding goal of biodiversity conservation, efforts must begin below the species level. Genetic diversity may be lost long before a species becomes extinct. For financial reasons, BDM only monitors genetic diversity in crop plants and livestock […]. Representative surveys, such as those conducted by the programme conducts on the diversity of widespread and common species […], cannot be financed for genetic diversity"* (Swiss Biodiversity Monitoring BDM, 2014). Including NGS barcoding in monitoring schemes will not directly provide the necessary genetic information for measurements of genetic diversity, but will furnish the raw material (identified insects) and DNA extractions necessary for such surveys.

Although NGS barcoding is currently the best candidate for monitoring programs, this method, along with morphological identifications, could quickly become limited for very large datasets encompassing thousands or even millions of specimens. In such case, a more viable approach would be to couple metabarcoding with raw biomass measurements. Metabarcoding could provide estimates on the community composition (presence/absence) whereas biomass measurements could evaluate species abundance declines (Hallmann et al., 2017). Although this approach does not inform on the species-specific population trends, it would currently be the only viable option for assessing communities using very large datasets.

More broadly, incorporating NGS in monitoring programs provides the opportunity to combine the classical species assessment and ecological network analyses. Until now, biomonitoring was merely descriptive and assessed changes in species and communities in space and time. While these data are useful for assessing population trends or species conservation status, they do not provide information on underlying mechanisms behind ecosystem functions. A more powerful and holistic approach is provided by ecological network analyses, which describe species interactions, community structures and functionality, and resilience of ecosystems (Derocles et al., 2018). In the particular case of pollinators, building an ecological network would for instance involve identifying the pollen found on specimens by means of metabarcoding. Pollen is regularly found on: (1) pollen-collecting structures of female bees; and (2) on the body of any flower visitor. The combination of these data would allow building plant-pollinator networks which are extremely informative and can provide valuable insights, for instance in how environmental changes are affecting ecosystems (Benstead et al., 2010; Weiner, Werner, Linsenmair, & Blüthgen, 2014) or for assessing certain conservation policies at different ecological layers (Kaiser-Bunbury et al., 2017). Future studies should examine whether full body extraction may recover enough DNA to perform pollen metabarcoding alongside with NGS barcoding.

Importance of accuracy in biodiversity assessments

Biodiversity monitoring is essential to document changes over time and space, as well as to evaluate the effectiveness of conservation policies. However, monitoring efforts and conservation policies can only be effective if biodiversity is properly assessed. As mentioned in the introduction, biodiversity can be measured at different levels but the most fundamental and often-used level is the "species" level. Indeed, having clear and stable species delimitations is vital for many research fields such as taxonomy, ecology, and biogeography.

There are numerous examples illustrating how correctly delimitating lineages within cryptic species complexes can affect conservation status. For instance, molecular delimitation of the subterranean amphipod *Niphargus stygius* revealed the presence of 15 parapatric and sympatric cryptic lineages. After correct delimitation, these lineages were found to be much more endangered than estimated before, with records of single-site endemism (Delić et al., 2017). Conversely, after molecular delimitation of the cryptic species complex of the amblyopsid cavefish (*Typhlichthys subterraneus*), the conservation status was suggested to be downgraded from ''Vulnerable'' to ''Near Threatened'' on IUCN Red List (Niemiller et al., 2013).

Traditionally, species were mainly described and delimited based on the examination of variable morphological traits (Padial, Miralles, De la Riva, & Vences, 2010). Currently, the vast majority of described species have been delimitated using this morphological species concept. As outlined in chapter II, this species-concept has severe limitations (i.e. choice of morphological criteria, phenotypic variation, sex-dimorphism, holometabolous associated metamorphosis, cryptic species). To cope with the shortfalls of morphology, many taxonomists started complementing species delimitations with molecular data, especially DNA barcodes. Although DNA barcoding has been successfully applied at many occasions for unravelling cryptic diversity or to uncover morphologically complex taxa (reviewed in Struck et al., 2018), using uniparentally inherited markers (i.e. COI) does not necessarily provide accurate information on ongoing gene flow and reproductive isolation (e.g. Hinojosa et al., 2019). Events such as introgression, incomplete lineage sorting or hybridization can severely bias COI-based phylogenies and lead to discordance between morphology and COI barcoding.

Currently, there are few convenient molecular methods allowing to delimit closely related lineages and to shed light on the origin of such mitochondrial-nuclear discordances. To overcome this methodological shortfall, in Chapter II and III, we tested ultraconserved elements (UCEs) as a fast and robust approach to recover thousands of genomic markers.

As a proof of concept, we first applied UCEs on a species-complex suspected to harbour cryptic diversity due to the presence of deeply divergent sympatric mitochondrial lineages within two taxa (i.e. *Andrena bicolor* and *A. amieti*). The results from different species delimitation tests and phylogenetic analyses clearly unravelled the presence of two cryptic lineages within *A. bicolor*. For *A. amieti*, we found patterns of mitochondrial-nuclear discordance with no reproductive isolation between sympatric mitochondrial lineages. This species-complex provides an ideal case study to validate the UCEs for delimiting closely related cryptic species, as we found, in the same complex, patterns of mitochondrial-nuclear concordance as well as mitochondrial-nuclear discordance. This allowed us to conclude that the UCEs were variable enough to separate closely related lineages and that all mitochondrial genes were properly filtered out in the UCE dataset and therefore did not influence the nuclear phylogeny.

As a follow up study, UCEs were used in Chapter II to unravel some long-debated taxonomical enigmas. More specifically, we investigated two cases of mitochondrial introgression among distinct but closely related species; two cases of species paraphyly not associated with known morphological differences; and two additional cases where species delimitation hypotheses have been conflicting. For all species-complexes, UCEs provided clear and well-supported phylogenies allowing to draw clear conclusions on species delimitations.

UCEs also provided the first conclusive evidence of mitochondrial introgression in wild bees. Several examples have been suggested before (Schmidt, Schmid-Egger, Morinière, Haszprunar, & Hebert, 2015). However, these examples were based on closely related species and no evidence from nuclear DNA marker was presented to confirm introgression. Introgression occurs when alleles or entire organelle genomes from one species penetrate the gene pool of another species through interspecific mating (Funk & Omland, 2003). Formerly, interspecific hybridization, and therefore introgression, was thought to be mainly confined to plants. However, there is accumulating evidence that interspecific hybridization also occurs in at least 10% of animals species (see review Mallet, 2005) with over 4200 reported cases in insects (Schwenk, Brede, & Streit, 2008). Hybridization events do however not always develop into introgression. For introgression to occur, hybrids need to backcross with the parental taxa. Although introgression can occur at the mitochondrial and/or nuclear level, rates are often biased to one or the other genome. In haplodiploid systems, mitochondrial introgression should theoretically occur more often than nuclear introgression because of the particular genetic transmission (Patten, Carioscia, & Linnen, 2015). Nevertheless, the introgression events documented in chapter II between the species complexes *Andrena barbareae*/*cineraria* and *Lasioglossum bavaricum*/*cupromicans* are among the first records in wild bees; these cases were reported before (Schmidt et al., 2015), but without analyses of nuclear markers, shedding doubts on the distinctiveness of the taxa.

Introgression events can be problematic for barcoding and NGS-based identifications: introgression can affect biodiversity indexes since two or more lineages will share the same barcode. For wild bees, barcode sharing rates oscillating between 3% (Sheffield et al., 2017) and 10% (Schmidt et al., 2015) are reported. For NGS-based monitoring programs however, these cases of barcode sharing may be problematic. If ignored, not only

the overall biodiversity indexes will be underestimated, but more importantly, certain species will remain undetected while abundance and distribution of others will be inflated. To overcome this barcoding limitation, the ideal workflow for NGS-based monitoring programs would require the morphological reassessment of all specimens matching known cases of shared barcodes. For bulk based method such as metabarcoding or mitogenomics this is evidently not possible since all specimens are mangled. However the more pragmatic approach proposed by NGS barcoding is perfectly suited for the morphological re-examination of specimens.

Overall, the results from chapter II and III provide strong evidence on the usefulness of UCEs for detecting cryptic diversity and for overcoming COI barcoding limitations. Besides, being a rapid and a quasi "universal" approach, the main advantage of using UCEs lies in its compatibility between datasets. This feature could ultimately allow complementing the tree of life for a vast majority of taxa and foster our understanding on many open taxonomical questions.

For pragmatic reasons, we mainly focused on the Swiss Apoidae fauna. As mentioned previously, bee populations in Switzerland are under high pressure (Cordillot & Kraus, 2011), and regardless of being historically well studied, they harbour many taxonomical enigmas. However, compared to countries with higher species richness, population trends and taxonomy are relatively well defined in Switzerland. As a result, at the European scale, the conservation status of most species (≥ 55%) remains unknown because of lack of data (Nieto et al., 2015). Given the huge taxonomic incompleteness in European bees, we suggest that UCEs and NGS barcoding have a high potential to improve the tools at the disposal of biologists and policy makers for the conservation of wild bees.

Conclusion and perspectives

Science, technology and innovation are highly interdependent and drive one another. In the context of this thesis, it was proposed to use technological advances to overcome important impediments to the effective survey of wild bees, namely taxonomical incompleteness and cost-effective monitoring. In Chapter I we show that at least one NGS-based method (i.e. NGS barcoding) would be a solid candidate for routine monitoring program. NGS barcoding also provide the security to re-examine specimens morphologically. Furthermore, this method also provide the opportunity to assess biodiversity at two levels, the species level and the genetic level. Although there is an increasing number of monitoring programs including some measures of genetic biodiversity, it is still unclear how these measurements will influence conservation status. Our results therefore open a new path for examining changes in species diversity and genetic diversity in complex taxa, such as insects. In chapter II and III we provide strong evidence on the usefulness of UCEs delimiting species. This method has great potential for overcoming the current taxonomical incompleteness in the European Apoidae fauna. Indeed, this method is a fast and robust way to obtain thousands of nuclear loci providing deep phylogenetic signals for uncovering reproductively isolated lineages. Although UCEs uncover all investigated taxonomical enigmas, it remains uncertain if this capture method would provide shallow enough signals to capture ongoing speciation, or cases of speciation with ongoing gene flow. Furthermore, one of the main advantages of this method is linked to its compatibility among different datasets. However, there are currently no studies that have taken advantage of this feature. It remains thus unclear how many loci will be shared between different datasets, performed by different laboratories and sequenced on different platforms.

References

Amann, R. I., Ludwig, W., & Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews*, *59*(1), 143–169.

Amend, A. S., Seifert, K. A., & Bruns, T. D. (2010). Quantifying microbial communities with 454 pyrosequencing: Does read abundance count? *Molecular Ecology*, *19*(24), 5555–5565. doi:10.1111/j.1365-294X.2010.04898.x

Ashrafi, S., Beck, A., Rutishauser, M., Arlettaz, R., & Bontadina, F. (2011). Trophic niche partitioning of cryptic species of long-eared bats in Switzerland: implications for conservation. *European Journal of Wildlife Research*, *57*(4), 843–849. doi:10.1007/s10344-011-0496-z

Benstead, J. P., Beveridge, O. S., Blanchard, J., Brey, T., Brown, L. E., Cross, W. F., … Yvon-Durocher, G. (2010). Ecological Networks in a Changing Climate. *Advances in Ecological Research*, *42*, 71–138. doi:10.1016/B978-0-12-381363-3.00002-2

Bickford, D., Lohman, D. J., Sodhi, N. S., Ng, P. K. L., Meier, R., Winker, K., … Das, I. (2007). Cryptic species as a window on diversity and conservation. *Trends in Ecology and Evolution*, *22*(3), 148–155. doi:10.1016/j.tree.2006.11.004

Biesmeijer, J. C., Roberts, S. P. M., Reemer, M., Ohlemüller, R., Edwards, M., Peeters, T., … Kunin, W. E. (2006). Parallel declines in pollinators and insect-pollinated plants in Britain and the Netherlands. *Science*. doi:10.1126/science.1127863

Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., … de Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology and Evolution*. doi:10.1016/j.tree.2014.04.003

Brown, D. M., Brenneman, R. A., Koepfli, K. P., Pollinger, J. P., Milá, B., Georgiadis, N. J., … Wayne, R. K. (2007). Extensive population genetic structure in the giraffe. *BMC Biology*, *5*(1), 57. doi:10.1186/1741-7007-5-57

Brown, M. J. F., & Paxton, R. J. (2009). The conservation of bees: a global perspective. *Apidologie*, *40*(3), 410–416. doi:10.1051/apido/2009019

Brunner, P. C., Fleming, C., & Frey, J. E. (2002). A molecular identification key for economically important thrips species (Thysanoptera: Thripidae) using direct sequencing and a PCR-RFLP-based approach. *Agricultural and Forest Entomology*, *4*(2), 127–136. doi:10.1046/j.1461-9563.2002.00132.x

Burgin, C. J., Colella, J. P., Kahn, P. L., & Upham, N. S. (2018). How many species of mammals are there? *Journal of Mammalogy*. doi:10.1093/jmammal/gyx147

Cane, J. H. (1997). Lifetime monetary value of individual pollinators: The bee habropoda l4boriosa at rabbiteye blueberry (vaccinium ashei reade). In *Acta Horticulturae*. doi:10.17660/ActaHortic.1997.446.8

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336. doi:10.1038/nmeth.f.303

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., … Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America*, *108 Suppl*(Supplement_1), 4516–22. doi:10.1073/pnas.1000080107

Cardoso, P., Erwin, T. L., Borges, P. A. V., & New, T. R. (2011). The seven impediments in invertebrate conservation and how to overcome them. *Biological Conservation*, *144*(11), 2647–2655. doi:10.1016/j.biocon.2011.07.024

Charlton, K. (2007). A note on divergent mtDNA lineages of bottlenose dolphins from coastal waters of

southern Australia. *J. Cetaceans Res. Manage.*

Cordillot, F., & Kraus, G. (2011). Espèces menacées en Suisse. *Etat de l'Environnement, Listes Rouges - Biodiversité*, 1–111.

Costello, M. J., May, R. M., & Stork, N. E. (2013). Can We Name Earth's Species Before They Go Extinct? *Science*, *339*(6118), 413–416. doi:10.1126/science.1230318

Costello, M. J., Wilson, S., & Houlding, B. (2012). Predicting total global species richness using rates of species description and estimates of taxonomic effort. *Systematic Biology*, *61*(5), 871–883. doi:10.1093/sysbio/syr080

Delić, T., Trontelj, P., Rendoš, M., & Fišer, C. (2017). The importance of naming cryptic species and the conservation of endemic subterranean amphipods. *Scientific Reports*, *7*(1), 3391. doi:10.1038/s41598-017-02938-z

Derocles, S. A. P., Bohan, D. A., Dumbrell, A. J., Kitson, J. J. N., Massol, F., Pauvert, C., … Evans, D. M. (2018). Biomonitoring for the 21st Century: Integrating Next-Generation Sequencing Into Ecological Network Analysis. *Advances in Ecological Research*, *58*, 1–62. doi:10.1016/bs.aecr.2017.12.001

Donoghue, M. J., & Alverson, W. S. (2000). A New Age of Discovery. *Annals of the Missouri Botanical Garden*. doi:10.2307/2666212

Fennessy, J., Bidon, T., Reuss, F., Kumar, V., Elkan, P., Nilsson, M. A., … Janke, A. (2016). Multi-locus Analyses Reveal Four Giraffe Species Instead of One. *Current Biology*. doi:10.1016/j.cub.2016.07.036

Fišer, C., Robinson, C. T., & Malard, F. (2018). Cryptic species as a window into the paradigm shift of the species concept. *Molecular Ecology*, *27*(3), 613–635. doi:10.1111/mec.14486

Fisher, B., Turner, R. K., & Morling, P. (2009). Defining and classifying ecosystem services for decision making. *Ecological Economics*, *68*(3), 643–653. doi:10.1016/j.ecolecon.2008.09.014

Funk, C., Caminer, M., & Ron, S. (2012). High levels of cryptic species diversity uncovered in Amazonian frogs. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1734), 1806–1814. doi:10.1098/rspb.2011.1653

Funk, D. J., & Omland, K. E. (2003). Species-Level Paraphyly and Polyphyly: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics*, *34*(1), 397–423. doi:10.1146/annurev.ecolsys.34.011802.132421

Gallai, N., Salles, J. M., Settele, J., & Vaissière, B. E. (2009). Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. *Ecological Economics*, *68*(3), 810–821. doi:10.1016/j.ecolecon.2008.06.014

Garibaldi, L. A., Steffan-dewenter, I., Winfree, R., Aizen, M. A., Bommarco, R., Cunningham, S. A., … Carvalheiro, L. G. (2014). Honey Bee Abundance. *Science*, *339*(May), 1608–1611.

Goulson, D., Nicholls, E., Botías, C., & Rotheray, E. L. (2015). Bee declines driven by combined stress from parasites, pesticides, and lack of flowers. *Science*, *347*(6229). doi:10.1126/science.1255957

Graystock, P., Yates, K., Darvill, B., Goulson, D., & Hughes, W. O. H. (2013). Emerging dangers: Deadly effects of an emergent parasite in a new pollinator host. *Journal of Invertebrate Pathology*, *114*(2), 114–119. doi:10.1016/j.jip.2013.06.005

Greenleaf, S. S., & Kremen, C. (2006a). Wild bee species increase tomato production and respond differently to surrounding land use in Northern California. *Biological Conservation*. doi:10.1016/j.biocon.2006.05.025

Greenleaf, S. S., & Kremen, C. (2006b). Wild bees enhance honey bees' pollination of hybrid sunflower. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(37), 13890–13895. doi:10.1073/pnas.0600929103

Hajibabaei, M., Baird, D. J., Fahner, N. A., Beiko, R., & Golding, G. B. (2016). A new way to contemplate Darwin's tangled bank: how DNA barcodes are reconnecting biodiversity science and biomonitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1702), 20150330. doi:10.1098/rstb.2015.0330

Hall, N. (2007). Advanced sequencing technologies and their wider impact in microbiology. *Journal of Experimental Biology*, *210*(9), 1518–1525. doi:10.1242/jeb.001370

Hallmann, C. A., Sorg, M., Jongejans, E., Siepel, H., Hofland, N., Schwan, H., … de Kroon, H. (2017). More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PLOS ONE*, *12*(10), e0185809. doi:10.1371/journal.pone.0185809

Hebert, P. D. N., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, *270*(1512), 313–321. doi:10.1098/rspb.2002.2218

Hebert, P. D. N., Penton, E. H., Burns, J. M., Janzen, D. H., & Hallwachs, W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. *Proceedings of the National Academy of Sciences*, *101*(41), 14812–14817. doi:10.1073/pnas.0406166101

Hiiesalu, I., Opik, M., Metsis, M., Lilje, L., Davison, J., Vasar, M., … Pärtel, M. (2012). Plant species richness belowground: higher richness and new patterns revealed by next-generation sequencing. *Molecular Ecology*, *21*(8), 2004–16. doi:10.1111/j.1365-294X.2011.05390.x

Hinojosa, J. C., Koubínová, D., Szenteczki, M. A., Pitteloud, C., Dincă, V., Alvarez, N., & Vila, R. (2019). A mirage of cryptic species: Genomics uncover striking mitonuclear discordance in the butterfly Thymelicus sylvestris. *Molecular Ecology*, mec.15153. doi:10.1111/mec.15153

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., … Yu, D. W. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, *16*(10), 1245–1257. doi:10.1111/ele.12162

Kaiser-Bunbury, C. N., Mougal, J., Whittington, A. E., Valentin, T., Gabriel, R., Olesen, J. M., & Blüthgen, N. (2017). Ecosystem restoration strengthens pollination network resilience and function. *Nature*, *542*(7640), 223–227. doi:10.1038/nature21071

Klein, A.-M., Steffan-Dewenter, I., & Tscharntke, T. (2003). Pollination of Coffea canephora in relation to local and regional agroforestry management. *Journal of Applied Ecology*, *40*(5), 837–845. doi:10.1046/j.1365-2664.2003.00847.x

Klein, A. M., Vaissière, B. E., Cane, J. H., Steffan-Dewenter, I., Cunningham, S. A., Kremen, C., & Tscharntke, T. (2007). Importance of pollinators in changing landscapes for world crops. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1608), 303–313. doi:10.1098/rspb.2006.3721

Kremen, C., Williams, N. M., & Thorp, R. W. (2002). Crop pollination from native bees at risk from agricultural intensification. *Proceedings of the National Academy of Sciences of the United States of America*. doi:10.1073/pnas.262413599

Kress, W. J., García-Robledo, C., Uriarte, M., & Erickson, D. L. (2015). DNA barcodes for ecology, evolution, and conservation. *Trends in Ecology & Evolution*, *30*(1), 25–35. doi:10.1016/j.tree.2014.10.008

Larsen, B. B., Miller, E. C., Rhodes, M. K., & Wiens, J. J. (2017). Inordinate fondness multiplied and redistributed: The number of species on earth and the new pie of life. *Quarterly Review of Biology*, *92*(3), 229–265. doi:10.1086/693564

Lautenbach, S., Seppelt, R., Liebscher, J., & Dormann, C. F. (2012). Spatial and temporal trends of global pollination benefit. *PLoS ONE*. doi:10.1371/journal.pone.0035954

Lebuhn, G., Droege, S., Connor, E. F., Gemmill-Herren, B., Potts, S. G., Minckley, R. L., … Parker, F. (2012). Detecting insect pollinator declines on regional and global scales. *Conservation Biology : The Journal of the Society for Conservation Biology*, *27*(1), 113–20. doi:10.1111/j.1523-1739.2012.01962.x

Lees, A. C., & Pimm, S. L. (2015). Species, extinct before we know them? *Current Biology*, *25*(5), R177–R180. doi:10.1016/j.cub.2014.12.017

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., … Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, *10*(1), 34. doi:10.1186/1742-9994-10-34

Liu, S., Li, Y., Lu, J., Su, X., Tang, M., Zhang, R., … Zhou, X. (2013). SOAPBarcode: Revealing arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons. *Methods in Ecology and Evolution*, *4*(12), 1142–1150. doi:10.1111/2041-210X.12120

Locey, K. J., & Lennon, J. T. (2016). Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(21), 5970–5975. doi:10.1073/pnas.1521291113

Martiny, A. C. (2019). High proportions of bacteria are culturable across major biomes. *ISME Journal*, 3–6. doi:10.1038/s41396-019-0410-3

Meiklejohn, K. A., Damaso, N., & Robertson, J. M. (2019). Assessment of BOLD and GenBank – Their accuracy and reliability for the identification of biological materials. *PLoS ONE*, *14*(6), 1–14. doi:10.1371/journal.pone.0217084

Morandin, L. A., & Winston, M. L. (2005). Wild bee abundance and seed production in conventional, organic, and genetically modified canola. *Ecological Applications*. doi:10.1890/03-5271

Müller-Wille, S. (2006). Linnaeus' herbarium cabinet: a piece of furniture and its function. *Endeavour*, *30*(2), 60–64. doi:10.1016/j.endeavour.2006.03.001

Nater, A., Mattle-Greminger, M. P., Nurcahyo, A., Nowak, M. G., de Manuel, M., Desai, T., … Krützen, M. (2017). Morphometric, Behavioral, and Genomic Evidence for a New Orangutan Species. *Current Biology*. doi:10.1016/j.cub.2017.09.047

Niemiller, M. L., Graening, G. O., Fenolio, D. B., Godwin, J. C., Cooley, J. R., Pearson, W. D., … Near, T. J. (2013). Doomed before they are described? The need for conservation assessments of cryptic species complexes using an amblyopsid cavefish (Amblyopsidae: Typhlichthys) as a case study. *Biodiversity and Conservation*, *22*(8), 1799–1820. doi:10.1007/s10531-013-0514-4

Nieto, A., Roberts, S. P. M., Kemp, J., Rasmont, P., Kuhlmann, M., Criado, M. G., … Michez, D. (2014). *European Red List of Bees*. *European Red List of bees. Luxembourg*. doi:10.2779/77003

Nieto, A., Roberts, S. P. M., Kemp, J., Rasmont, P., Kuhlmann, M., Criado, M. G., … Michez, D. (2015). *European Red List of Bees*. *Luxembourg: Publication Office of the European Union*. doi:10.2779/77003

Ollerton, J., Winfree, R., & Tarrant, S. (2011). How many flowering plants are pollinated by animals? *Oikos*, *120*(3), 321–326. doi:10.1111/j.1600-0706.2010.18644.x

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

Papadopoulou, A., Taberlet, P., & Zinger, L. (2015). Metagenome skimming for phylogenetic community ecology: a new era in biodiversity research. *Molecular Ecology*, *24*(14), 3515–3517. doi:10.1111/mec.13263

Patten, M. M., Carioscia, S. A., & Linnen, C. R. (2015). Biased introgression of mitochondrial and nuclear genes: A comparison of diploid and haplodiploid systems. *Molecular Ecology*, *24*(20), 5200–5210.

doi:10.1111/mec.13318

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., … de Vargas, C. (2012). CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms. *PLoS Biology*. doi:10.1371/journal.pbio.1001419

Pawlowski, J., Lejzerowicz, F., Apotheloz-Perret-Gentil, L., Visco, J., & Esling, P. (2016). Protist metabarcoding and environmental biomonitoring: Time for change. *European Journal of Protistology*, *55*, 12–25. doi:10.1016/j.ejop.2016.02.003

Pimm, S. L., Jenkins, C. N., Abell, R., Brooks, T. M., Gittleman, J. L., Joppa, L. N., … Sexton, J. O. (2014). The biodiversity of species and their rates of extinction, distribution, and protection. *Science*, *344*(6187), 1246752–1246752. doi:10.1126/science.1246752

Pimm, S. L., Jenkins, C. N., Joppa, L. N., Roberts, D. L., & Russell, G. J. (2010). How Many Endangered Species Remain to be Discovered in Brazil? *Natureza & Conservação*, *08*(01), 71–77. doi:10.4322/natcon.00801011

Piper, A. M., Batovska, J., Cogan, N. O. I., Weiss, J., Cunningham, J. P., Rodoni, B. C., & Blacket, M. J. (2019). Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. *GigaScience*, *8*(8), 1–22. doi:10.1093/gigascience/giz092

Potts, S. G., Biesmeijer, J. C., Kremen, C., Neumann, P., Schweiger, O., & Kunin, W. E. (2010). Global pollinator declines: Trends, impacts and drivers. *Trends in Ecology and Evolution*, *25*(6), 345–353. doi:10.1016/j.tree.2010.01.007

Ravoet, J., De Smet, L., Meeus, I., Smagghe, G., Wenseleers, T., & de Graaf, D. C. (2014). Widespread occurrence of honey bee pathogens in solitary bees. *Journal of Invertebrate Pathology*, *122*, 55–58. doi:10.1016/j.jip.2014.08.007

Roca, A. L., Georgiadis, N., Pecon-Slattery, J., & O'Brien, S. J. (2001). Genetic evidence for two species of elephant in Africa. *Science*. doi:10.1126/science.1059936

Roskov, Y., Abucay, L., Orrell, T., Nicolson, D., Bailly, N., Kirk, P. M., … Penev, L. (2019). Species 2000 & ITIS Catalogue of Life, 2018 Annual Checklist, 30 March 2019.

Rothschild, L. J., & Mancinelli, R. L. (2001). Life in extreme environments. *Nature*, *409*(6823), 1092–1101. doi:10.1038/35059215

Sánchez-Bayo, F., & Wyckhuys, K. A. G. (2019). Worldwide decline of the entomofauna: A review of its drivers. *Biological Conservation*, *232*(September 2018), 8–27. doi:10.1016/j.biocon.2019.01.020

Schmidt, S., Schmid-Egger, C., Morinière, J., Haszprunar, G., & Hebert, P. D. N. (2015). DNA barcoding largely supports 250 years of classical taxonomy: identifications for Central European bees (Hymenoptera, Apoidea partim). *Molecular Ecology Resources*, *15*(4), 985–1000. doi:10.1111/1755-0998.12363

Schwenk, K., Brede, N., & Streit, B. (2008). Introduction. Extent, processes and evolutionary impact of interspecific hybridization in animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1505), 2805–2811. doi:10.1098/rstb.2008.0055

Sheffield, C. S., Heron, J., Gibbs, J., Onuferko, T. M., Oram, R., Best, L., … Rowe, G. (2017). Contribution of DNA barcoding to the study of the bees (Hymenoptera: Apoidea) of Canada: Progress to date. *Canadian Entomologist*, *149*(6), 736–754. doi:10.4039/tce.2017.49

Sih, A., Bell, A. M., & Kerby, J. L. (2004). Two stressors are far deadlier than one. *Trends in Ecology and Evolution*, *19*(6), 274–276. doi:10.1016/j.tree.2004.02.010

Simaika, J. P., & Samways, M. J. (2018). Insect conservation psychology. *Journal of Insect Conservation*, *22*(3–4), 635–642. doi:10.1007/s10841-018-0047-y

Staats, M., Arulandhu, A. J., Gravendeel, B., Holst-Jensen, A., Scholtens, I., Peelen, T., … Kok, E. (2016). Advances in DNA metabarcoding for food and wildlife forensic species identification. *Analytical and Bioanalytical Chemistry*. doi:10.1007/s00216-016-9595-8

Stork, N. E. (2017). How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth? *Annual Review of Entomology*, *63*(1), 31–45. doi:10.1146/annurev-ento-020117-043348

Struck, T. H., Feder, J. L., Bendiksby, M., Birkeland, S., Cerca, J., Gusarov, V. I., … Dimitrov, D. (2018). Finding Evolutionary Processes Hidden in Cryptic Species. *Trends in Ecology & Evolution*, *33*(3), 153–163. doi:10.1016/j.tree.2017.11.007

Taberlet, P., & Coissac, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding, *21*, 2045–2050.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*, 2045–2050. doi:10.1111/j.1365-294X.2012.05470.x

Theodoridis, S., Nogués-Bravo, D., & Conti, E. (2019). The role of cryptic diversity and its environmental correlates in global conservation status assessments: Insights from the threatened bird's-eye primrose (Primula farinosa L.). *Diversity and Distributions*, *25*(9), 1457–1471. doi:10.1111/ddi.12953

Thomas, J. . (2005). Monitoring change in the abundance and distribution of insects using butterflies and other indicator groups. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1454), 339–357. doi:10.1098/rstb.2004.1585

Toews, D. P. L., & Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*, *21*(16), 3907–3930. doi:10.1111/j.1365-294X.2012.05664.x

Trontelj, P., & Fier, C. (2009). Cryptic species diversity should not be trivialised. *Systematics and Biodiversity*, *7*(1), 1–3. doi:10.1017/S1477200008002909

Valentini, A., Pompanon, F., & Taberlet, P. (2009). DNA barcoding for ecologists. *Trends in Ecology and Evolution*, *24*(2), 110–117. doi:10.1016/j.tree.2008.09.011

van Engelsdorp, D., Hayes, J., Underwood, R. M., & Pettis, J. (2008). A survey of honey bee colony losses in the U.S., Fall 2007 to Spring 2008. *PLoS ONE*. doi:10.1371/journal.pone.0004071

Weiner, C. N., Werner, M., Linsenmair, K. E., & Blüthgen, N. (2014). Land-use impacts on plant–pollinator networks: interaction strength and specialization predict pollinator declines. *Ecology*, *95*(2), 466–474. doi:10.1890/13-0436.1

Westphal, C., Bommarco, R., Carré, G., Lamborn, E., Morison, N., Petanidou, T., … Steffan-Dewenter, I. (2008). Measuring bee diversity in different European habitats and biogeographical regions. *Ecological Monographs*, *78*(4), 653–671. doi:10.1890/07-1292.1

Winfree, R., Williams, N. M., Gaines, H., Ascher, J. S., & Kremen, C. (2008). Wild bee pollinators provide the majority of crop visitation across land-use gradients in New Jersey and Pennsylvania, USA. *Journal of Applied Ecology*. doi:10.1111/j.1365-2664.2007.01418.x

Zarza, E., Connors, E. M., Maley, J. M., Tsai, W. L. E., Heimes, P., Kaplan, M., & McCormack, J. E. (2018). Combining ultraconserved elements and mtDNA data to uncover lineage diversity in a Mexican highland frog (Sarcohyla ; Hylidae). *PeerJ*, *6*, e6045. doi:10.7717/peerj.6045

Zurbuchen, A., & Müller, A. (2012). *Wildbienenschutz - von der Wissenschaft zur Praxis*. *Bristol-Schriftenreihe*.