Research Plan

Doctoral candidate: Maria Khan, 2215004

Doctoral Degree Programme: Biology

Faculty of Science, Ecology and Genetics Research Unit

Supervisors: Senior Curator Marko Mutanen,

Assistant Professor Stefan Prost

Professor Niklas Wahlberg (Lund University)

Follow-up group: Chair, Dr. Juan Galarza Pavia,

Professor Anna-Maria Pirttilä,

Postdoctoral Researcher Johanna Honka

Title of the Thesis: Characterizing species diversity through genomics: Envisioning global bio-literacy using megadiverse insects as a model

Type of thesis: Article-based

Start date of study right: 24.05.2023

Abstract: Biodiversity crises are a pressing global concern, with a significant portion of species facing extinction due to environmental changes. Despite this urgency, our understanding of biodiversity remains limited, with many species yet to be discovered and described. Traditional taxonomy struggles with species differentiation, hindering biomonitoring efforts and obscuring species distributions, variations, and conservation statuses. To address this knowledge gap, leveraging DNA-based approaches presents a promising solution. By harnessing cutting-edge genomics technologies, we can unlock vast amounts of genetic information from individuals, enabling efficient species delimitation and identification. This proposal advocates for the adoption of high-throughput genomics to revolutionize biodiversity research, offering standardized and quantifiable methods for species characterization. The project aims to demonstrate the benefits of DNA-based approaches across three key research questions: efficient species characterization under complex evolutionary circumstances, tackling diverse species groups, and integrating historical species descriptions with modern genomic techniques. Through international collaboration and state-of-the-art sequencing technologies, this research endeavors to advance our understanding of biodiversity and inform conservation efforts for a sustainable future.

1. Background

1.1. Aims and Significance of the research

Biodiversity crises have taken the world by storm. According to the Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services[1], 25% of species are threatened with extinction due to the environmental changes that we are facing nowadays. Yet, we remain largely illiterate concerning the biodiversity that surrounds us, with approximately 80% of the world's multicellular species still awaiting to be discovered and described[1]. Even for species that are already described, it is often difficult to tell them apart due to strikingly similar morphological faces, permitting only dedicated species experts to accomplish that, or not even so. Additionally, with many species groups taxonomic expertise being completely absent, biomonitoring is rendered entirely impossible. Other than that, their distributions, variation, conservation status, and interconnections to other species remain incompletely understood. All this can be attributed to the considerable deficit in our biodiversity knowledge[2]. For conservation efforts to be efficient, it is crucial for us to become bio-literate regarding the species diversity surrounding us, because species are central units to all biodiversity-related activities. To bridge this literacy gap we need to take a step forward from traditional practices that are time-consuming and laborious and may not provide a solution in most cases due to the very high number of species.

All taxonomically relevant information is encoded in an organism's DNA[3]. Basing biodiversity monitoring on DNA would provide multiple remarkable benefits, as state-of-the-art genomics techniques enable sequencing DNA at an unprecedented speed and accuracy. DNA provides feasible solutions to overcome the two main components of our present illiteracy regarding biodiversity. First, we can obtain vast amounts of genetic information from individuals and delimit species based on DNA sequence information. Second, we can benefit from high-throughput genomics by sequencing standard fragments of DNA (i.e., DNA barcodes) from a vast number of specimens simultaneously[4]. My proposal puts forward the idea to make use of high-throughput genomics to provide solutions to both aspects.

2. Research Questions, Objectives, and Methods

In the proposed project, I will explore and demonstrate the benefits of DNA-based approaches in biodiversity research. My ambition is to convince the taxonomic research community about the benefits of cutting-edge genomics technologies as compared to traditional largely manual and non-quantifiable practices. In each of my three work packages, I will apply a different, efficient DNA

technology, and address different yet crucially important problems of biodiversity characterization. More specifically, I will address the following research questions:

1. How could genomics approaches provide more efficient ways to characterize species under complicated evolutionary circumstances?

Characterizing and delimiting species is one of the main commitments of taxonomic research. It is also proven to be a very complicated procedure. These complications largely arise from the great variability of biological diversity and the high similarity of many species. Species are born through an evolutionary process called speciation, meaning that every species has a common ancestor with another species, sometimes not far in the past. As speciation is a process rather than an event, there is a “grey zone” during which species have undergone differentiation but have not yet reached full species status. This means that while many species have fully speciated and do not show any gene flow anymore, the situation may be much more complicated in some others. Basing species characterization on genomic data would enable standardizing and quantifying the process. Further, to enable efficient species delineation, all relevant information about organisms is encoded in their DNA. In this work package, I will focus on two cases of long-standing issues in the characterization of two closely related species (or populations). In the first study, I will focus on a pair of sibling species of Small Emperor moth *Saturnia pavonia* and *S. pavoniella*. The two species show a parapatric model of distribution, i.e., their distributions meet but do not overlap. They appear nearly identical in their external morphology, and they are shown to hybridize readily in the laboratory. This suggests that they have speciated but have not isolated ecologically. The second case focuses similarly on a pair of species with virtually identical appearance and problematic taxonomy: Burnished Brass *Diachrysia chrysitis* and *D. stenochrysis* and/or D. *tutti*. However, the ranges of the two putative species largely overlap, and their females are shown to release chemically different sexual pheromones to attract males. Individuals cannot be reliably separated by morphology. Using a target-enrichment approach[5] and Illumina NovaSeq sequencing[7], I will demonstrate that if the species truly are two distinct biological entities, their characterization and identification will be efficient when based on ca 2,000 standard benchmarked genetic markers. I will demonstrate that such a DNA taxonomy-based approach will provide a scientifically rigorous manner to circumscribe species under challenging evolutionary circumstances[5].

2. How can we tackle super diverse species groups representing “dark” diversity?

Only a fraction of all species has been scientifically described. Most described species also come from the ‘easiest’ groups, i.e., groups where species have large body sizes and that provide many

good features for species discrimination. Insect species have been estimated to be several million, and in many groups, the look-alikes are so many that their comprehensive characterization and identification based on their morphology is not a realistic scenario. In this work package, I will focus on parasitic wasps (Hymenoptera) more specifically on Braconidae that are known to be immensely diverse and largely remain in the dark for their species diversity[2]. The Braconidae family of parasitoid wasps is known to be highly diverse. According to recent research, the number of discovered species in this family is estimated to be around 17,000 recognized species, with many thousands more undescribed. One analysis estimated a total between 30,000 and 50,000, and another provided a narrower estimate between 42,000 and 43,000 species[11]. A more recent article mentions 19,801 described species belonging to 1071 genera, which represent nearly 20% of the total hymenopteran diversity[12]. Therefore, the exact number of species in the Braconidae family is still not fully determined, but it is clear that the family is highly speciose. I will assess their species diversity using DNA barcodes, i.e., short standard fragments of DNA, as a species proxy, and sequencing the barcodes from 9,025 specimens using the high-throughput Sequel sequencing platform and a PacBio SEQUEL platform pipeline already established by my supervisors and collaborators. I will demonstrate that DNA barcoding coupled with high-throughput sequencing will provide efficient means to sort out immensely species-rich groups into putative species. Furthermore, I will test the validity of these putative species by sequencing two nuclear markers for putative species using the Oxford Nanopore MinION platform and a pipeline that is already established in Oulu. This will provide efficient means to assess species diversity accurately based on large numbers of samples, and it will speed up species characterization and taxonomic procedure, including naming, in an unprecedented way.

3. How to consider the history of previously described but not molecularly characterized species?

When a species is described and named, a type specimen for it is always designated. The type specimen provides a link between the biological species and the name[8]. This is important because for example in case the species is divided into two, the type specimen determines for which one of them the old name should be applied. However, as taxonomy and species descriptions have been made for over 260 years, most of the type specimens lack genetic information, rendering it difficult to associate previously described species with those observed based on genetic analyses. Luckily, technologies to recover DNA from historical samples have developed fast in the last 15 years, making it feasible to obtain genetic information, such as DNA barcodes, from old collection specimens[6]. In this work package, I will demonstrate how modern genomic technologies enable the recovery of DNA information from old museum-preserved type specimens of small parasitic wasps,

i.e., the same group of insects the WP2 will focus on. To reach this target, I will analyze museum specimens using Illumina sequencing platform. The approach to be used is called Shotgun sequencing. My collaborator Prof. Wahlberg has developed a highly functional pipeline for sequencing of old samples with degraded DNA[9].

Materials and protocols for WP1-WP3 and sequencing costs:

The study specimens for the WPs 1-2 have already been collected in the connection of previous and ongoing research projects by my supervisors. The sequencing analyses for the WP1 will be funded by Marko Mutanen's (MM) academy project. An efficient laboratory and bioinformatics pipeline for target enrichment approach is established by my supervisors. For WP2, the material has been collected using Malaise traps in the connection of MM's BIOMON project BIOLITERACY, with all required permissions to the protected sampling areas having been granted to my supervisor. The wasp specimens to be used in my study are preserved in 100% alcohol. The sequencing costs of these samples will be covered by the BIOLITERACY project. The SEQUEL analyses will be conducted at the Centre for Biodiversity Genomics (UGuelph, Canada) and an efficient MinION pipeline has been established at Oulu. The type specimens for WP3 have not been assembled yet, but I am confident that with the non-destructive protocols to be used, obtaining tissues of type specimens will not be an issue. The sequencing of those samples will be covered by the research available funds of the supervisors and collaborators. The laboratory work will be done at the Ecology and Genetics Research Unit labs at Oulu, and partially at Lund university. All the required equipment will be available in these institutes. Demanding computational analyses of massive genomic-scale datasets will be carried out with CSC supercomputers.

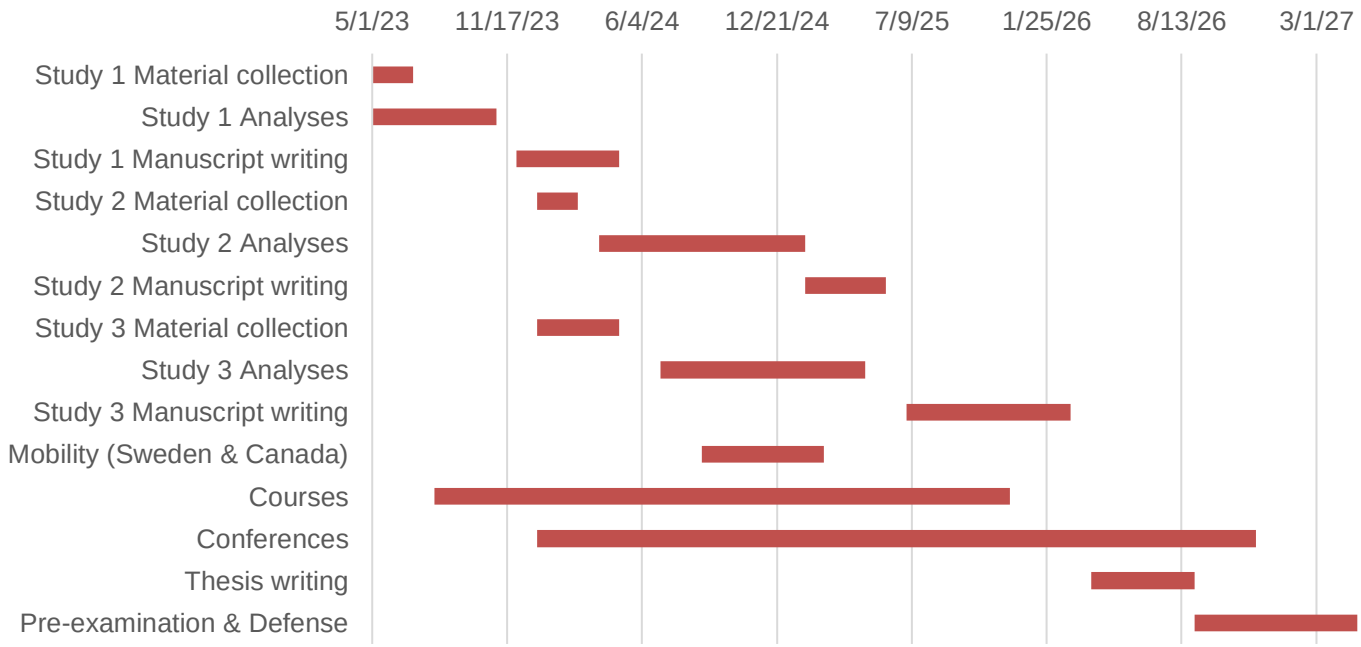
3. Work Plan and Schedule

3.1. Research Environment and Mobility Plan

The research will be led and carried out at the Ecology and Genetics Research Unit, University of Oulu, Finland. My research will be conducted in international collaboration. The main supervision will be at Oulu and by Marko Mutanen and Stefan Prost, but studies will be conducted in collaboration with Lund University and Professor Niklas Wahlberg. He is globally among the leading researchers of the field of molecular systematics and phylogenetics. He also has been a pioneering researcher in "museomics", a field that develops genetic approaches to recover DNA information from old specimens with degraded DNA. It is therefore particularly the WP3 that is to be conducted in close collaboration with Wahlberg. I will make a two-year visit (September 2024 to 2026) to the Lund University and Wahlberg lab. During this stay, I will learn methods of shotgun sequencing and associated bioinformatics. Equally importantly, this visit is a very important opportunity for me also to learn laboratory and working practices in another group and meet colleagues of my field and share ideas with them. I also plan to make a shorter visit (ca 1 month) to the Centre

for Biodiversity Genomics, University of Guelph, Canada, where MM's group has close links and collaboration especially with DNA barcoding and PacBIO SEQUEL sequencing platform and related bioinformatics. This visit will take approximately one month. Naturally, I will be actively participating and disseminating my results in the major conferences of the field.

3.2. Preliminary Timeline



3.2. Funding Plan:

I have qualified for the I4World Doctoral program which will fund my PhD for the next 4 years (2023-2027). This includes my salary until 14th May 2027, travel grant of upto 5000€ and publishing costs including open access.

4. Applicant and Supervisors Details

4.1. Applicant Details and Project relevant merits

Maria Khan ORCID: 0000-0003-2004-6391 is the PI on this proposal. I am a graduate of Biotechnology from the University of Karachi, Pakistan. As an undergraduate student, I have worked as a trainee with several labs and earned laboratory skills such as DNA extraction, purification, polymerase chain reaction (PCR), and other molecular biology skills. I also did my Master's in Ecology and Population Genetics at the University of Oulu, Finland, where I've studied for the past two years and learned to implement different Bioinformatic and Ecological method skills. I am practically experienced in using technical tools such as R studio and programming languages, including Python and Unix/Linux. I also learned to use Puhti supercomputers. All these experiences were useful during the analysis of the genomic data for my thesis, and I now believe that they will

be of great advantage to this research project. I have been a part of Biodiversity research since 2021 and did my master's thesis on the subject.

4.2. Project Supervisors

Marko Mutanen ORCID: 0000-0003-4464-6308 is the main supervisor on this proposal. He is a recognized expert in biodiversity genomics and holds a permanent position as Senior Curator at the University of Oulu. He leads the Insect Genomic Systematics research group whose research has been supported by two Finnish Academy grants (2014, 2018). He has successfully led the Finnish Barcode of Life project (FinBOL), a national node of iBOL, since 2010. He is now leading a 6-year and 8M-Euro 'Biodiverse Anthropocenes' PROFI6 programme at UOulu as well as two other funded research projects funded by UOulu/Finnish Ministry of Environment. He has published 112 peer-reviewed articles (including those in press), including many in leading journals such as *Systematic Biology*, *Proceedings B,* and *Ecology Letters*, and several non-peer-reviewed scientific articles. His research focus is on biodiversity and particularly bio literacy, i.e., how DNA-based tools could help us to better read, describe and map unexplored parts of it. He currently supervises six doctoral students and one postdoctoral researcher. His group focuses on species identification and species delimitation as well as phylogenetics by utilizing state-of-the-art molecular techniques, including high-throughput sequencing (HTS) technologies consistent with WP1-2. The Ecology and Genetics Research Unit also leads DNA-barcoding activities in Finland which is consistent with my WP2 and WP3 and will critically help me achieve my goal for the project.

Stefan Prost ORCID: 0000-0002-6229-3596 is a an Associate professor of UOulu and an expert of a variety of genomics tools and bioinformatics. His contribution will be particularly important in conducting Nanopore sequencing and bioinformatics consistent with WP2.

Niklas Wahlberg ORCID 0000-0002-1259-3363 is globally among the leading researchers of the field of molecular systematics and phylogenetics. He also has been a pioneering researcher in "museomics", a field that develops genetic approaches to recover DNA information from old specimens with degraded DNA. . The attention of his group is centered on the Evolutionary History of Lepidoptera and the processes such as genetic diversity, gene flow, admixture, and hybridization. To study the mentioned phenomenon, they utilize molecular systematic methods and are beginning to use Next Gen Sequencing technologies. Additionally, they utilize museum specimens in their work and as I mentioned in WP3 that I will be using museum-type specimens of small parasitic wasps then their expertise will aid in the success of the project.

5. Impact of the Research

Biodiversity loss is widely recognized as a major existential risk for humankind[10]. Species are central entities of biodiversity and serve as words of the biodiversity language. Since probably over 90% of species remain unknown to science, we also cannot communicate the bewildering diversity surrounding us efficiently. This also seriously complicates our attempts to conserve biodiversity efficiently, because this would require us to be able to map biodiversity and monitor changes in it. State-of-the-art molecular genetic tools have a high potential to provide an escape from this prevailing dead-end. University of Oulu is leading DNA barcoding activities nationwide and participates in global efforts to illuminate biodiversity through iBOL (https://ibol.org/) and researchers of University of Oulu do research in this area at the global forefront. I have a high desire to utilize new genetic tools to help us to overcome the biodiversity crisis and make humanity entirely bio-literate. I am convinced that genomics will revolutionize all biodiversity research, including species identification and delimitation, and I am keen to have an active role in this paradigmatic shift that I and many of my colleagues are foreseeing. The results that will be obtained from all the three work packages with the help of the cutting-edge technologies will aid in providing a solution to all the issues related to biodiversity crises mentioned above such as species delimitation, identification and eventually conservation. Also, it will serve as a template to further encourage other studies as well on the topic of biodiversity crises which needs immediate attention. We as scientists need to do important work that, for example, causes the government to effect change and make new policies regarding the biodiversity crises and environment in general.

References

[1]IPBES. (2019). Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (Version 1). (“Key Findings to Know from the IPBES Report on Biodiversity”) Zenodo. https://doi.org/10.5281/zenodo.6417333

[2]Rougerie, Rodolphe et al. “Molecular analysis of parasitoid linkages (MAPL): gut contents of adult parasitoid wasps reveal larval host.” *Molecular ecology* vol. 20,1 (2011): 179-86. doi:10.1111/j.1365-294X.2010.04918.x

[3]Khawaldeh, Saed et al. “Taxonomic Classification for Living Organisms Using Convolutional Neural Networks.” Genes vol. 8,11 326. 17 Nov. 2017, doi:10.3390/genes8110326

[4]Kennedy, Susan R et al. “High-throughput sequencing for community analysis: the promise of DNA barcoding to uncover diversity, relatedness, abundances and interactions in spider communities.” *Development genes and evolution* vol. 230,2 (2020): 185-201. doi:10.1007/s00427-020-00652-x

[5]Mayer, Christoph, et al. "Adding leaves to the Lepidoptera tree: capturing hundreds of nuclear genes from old museum specimens." *Systematic Entomology* 46.3 (2021): 649-71.

[6]Burrell, Andrew S, et al. "The use of museum specimens with high-throughput DNA sequencers." ("The use of museum specimens with high-throughput DNA sequencers") *Journal of human evolution* 79 (2015): 35-44.

[7]Ottenburghs, Jente, et al. "Highly differentiated loci resolve phylogenetic relationships in the Bean Goose complex." *BMC Ecology and Evolution* 23.1 (2023): 1-12.

[8]Sluys, Ronald. "Attaching names to biological species: the use and value of type specimens in systematic zoology and natural history collections." *Biological Theory* 16.1 (2021): 49-61.

[9]Twort, Victoria G., et al. "Museomics of a rare taxon: placing Whalleyanidae in the Lepidoptera Tree of Life." Systematic Entomology 46.4 (2021): 926-937.

[10]Torres, Phil. "Biodiversity loss: An existential risk comparable to climate change." The Bulletin of the Atomic Scientists. http://thebulletin. org/biodiversity-loss-existential-risk-comparable-climate-change9329#. VxRxznC6AAw. twitter. Accessed 11 (2016).

[11] JONES, O.R., PURVIS, A., BAUMGART, E. and QUICKE, D.L.J. (2009), Using taxonomic revision data to estimate the geographic and taxonomic distribution of undescribed species richness in the Braconidae (Hymenoptera: Ichneumonoidea). *Insect Conservation and Diversity*, 2: 204-212.

[12] Ahmad Z, Ghramh HA, Ansari A. Two new species of braconid wasps (Hymenoptera, Braconidae) from India. Zookeys. 2019 Nov 14;889:23-35.