RESOLVING BIOLOGY'S DARK MATTER: SPECIES RICHNESS, SPATIOTEMPORAL DISTRIBUTION, AND COMMUNITY COMPOSITION OF A DARK TAXON

Emily Hartop[1,2,3,4], Leshon Lee[5,6], Amrita Srivathsan[4,5], Mirkka Jones[7,8], Pablo de la Peña Aguilera[9], Otso Ovaskainen[10,7,11], Tomas Roslin[7,9], Rudolf Meier[4,5]

[1] Department of Natural History, NTNU University Museum, Norwegian University of Science and Technology, Trondheim, NO-7491, Norway
[2] Zoology Department, Stockholm University, SE-106 91, Stockholm, Sweden
[3] Station Linné, Ölands Skogsby 161, 386 93, Sweden
[4] Center for Integrative Biodiversity Discovery, Leibniz Institute for Evolution and Biodiversity Science, Museum für Naturkunde, Invalidenstraße 43, 10115 Berlin, Germany
[5] Department of Biological Sciences, National University of Singapore, Science Drive 4, Singapore 117558
[6] National University of Singapore, Lee Kong Chian Natural History Museum, 2 Conservatory Dr, Singapore 117377
[7] Research Center for Ecological Change, Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences, University of Helsinki, P.O. Box 65, Helsinki FI-00014, Finland
[8] Institute of Biotechnology, HILIFE Helsinki Institute of Life Science, University of Helsinki, P.O. Box 65, Helsinki FI-00014, Finland
[9] Swedish University of Agricultural Sciences (SLU), Department of Ecology, Ulls väg 18B, SE-75651 Uppsala, Sweden
[10] Department of Biological and Environmental Science, P.O. Box 35, FI-40014, University of Jyväskylä, Jyväskylä, Finland.
[11] Centre for Biodiversity Dynamics, Department of Biology, Norwegian University of Science and Technology, N-7491 Trondheim, Norway

ABSTRACT

Background: Biology´s dark matter is found in hyperdiverse, poorly known taxa that are numerically dominant and yet largely unstudied, even in temperate regions of the globe where charismatic taxa are already well understood. Biology's dark matter is everywhere, but identification impediments caused by high diversity, abundance, and small size have historically stymied its study. We here show how entomological dark matter can be converted from dark to reasonably well known by using high-throughput DNA barcoding ("megabarcoding"). We reveal the surprisingly high abundance and diversity of scuttle flies (Diptera: Phoridae) across Sweden by processing 31 800 specimens from 37 sites across four seasonal time-periods. We ask how many scuttle fly species are likely to occur in Sweden and what environmental factors are driving community changes between seasons and across the country.

Results: We find that Swedish scuttle fly diversity is much higher than previously known, with 549 mOTUs (putative species) detected, compared to only 374 species recorded previously. Hierarchical Modelling of Species Communities reveals that scuttle fly communities are highly structured by latitude, and strongly driven by climatic (including temporal) factors. The large dissimilarities between sites and seasons are largely driven by turnover rather than nestedness (i.e. local differences in species richness). We predict that climate changes are likely to greatly affect the 29% of the species that show a positive and the 18% of species that show a negative response to mean annual

temperature. Surprisingly, these results were robust regardless of whether we used haplotype diversity or species-proxies (mOTUs: molecular Operational Taxonomic Units) as response variables. Furthermore, species-level models of common taxa adequately predict overall species richness. **Conclusions:** Understanding the bulk of the diversity around us is imperative during an era of biodiversity loss and we here show that dark insect taxa can be efficiently characterised and surveyed with megabarcoding. Furthermore, we show that undersampling of rare taxa and the choice of operational taxonomic units do not change the main ecological inferences, so that the time appears rife for tackling biology's dark matter.

INTRODUCTION

If we go on the way we have, the fault is our greed and if we are not willing to change, we will disappear from the face of the globe, to be replaced by the insect. – Jacques Yves Cousteau

Biodiversity loss in the Anthropocene is driven by changes in climate, land use, but also species introductions (Díaz et al. 2006; Cardinale et al. 2012). Such loss can result in the concomitant decline in ecosystem services vital to society (Mace et al. 2018; Lu et al. 2020) and may ultimately disrupt global supply chains and food security (WEF, 2020). Accurate monitoring of biodiversity is therefore a global priority (Naeem et al. 2016).

A crucial first step to monitoring biodiversity is obtaining robust quantitative baseline data. Most biologists would consider this to be data on species diversity, abundance, and biomass for those taxa that contribute substantially to these quantitative metrics. However, most biodiversity studies cover only a few well-studied groups that are relatively easily identified and quantified (e.g., birds, mammals, amphibians, bees, and butterflies). Such charismatic taxa are then used as proxies for all taxa in a region (Ricklefs 2006; Wiens 2007; Svenning et al. 2008; Ollerton 2017; Garda et al. 2018; Bell et al. 2019; Bogoni et al. 2020; Orr et al. 2020) instead of basing our understanding of global biodiversity on a broad and unbiased representation of biodiversity covering a wide range of traits and responses to environmental change (Bar-On et al. 2018, Goodsell et al. 2024).

A key component of global biodiversity is "dark taxa", i.e., taxonomic groups for which less than 10% of the diversity is described and the species diversity is estimated to be upwards of 1000 (Hartop et al. 2022). Such poorly known groups do not only inhabit inaccessible realms like deep sea trenches (Rabone et al. 2023) but also the terrestrial habitats in which we live. A recent study (Srivathsan et al. 2023) revealed that 20 insect families (of which 10 belong to Diptera) account for >50% of local species diversity. Alarmingly, the very same families suffered from extreme taxonomic neglect and are therefore poorly represented in biodiversity surveys. Identifying and tackling the diversity of these dark taxa with scalable techniques thus emerges as an urgent priority for biodiversity science.

Large-scale studies of dark taxa have only become feasible in recent years due to advancements in sequencing technologies coupled with efficient single-specimen DNA barcoding workflows (Srivathsan et al. 2019a; Yeo et al. 2020; Srivathsan et al. 2021; Hartop et al. 2022; Meier et al. 2024; Chua et al., 2023: "megabarcoding"). Such workflows allow large numbers of specimens to be processed and sorted to putative species (mOTUs), while providing exact specimen counts, and vouchers for subsequent morphological, taxonomic, and biological work.

Scuttle flies in the family Phoridae (Diptera) have been considered the seventh most speciose and abundant insect family globally (Srivathsan et al. 2023). However, to date only ca. 4000 species have been described, although their actual diversity may be two orders of magnitude larger (Srivathsan et al. 2019a). In addition to their extreme species richness, scuttle flies occupy a wide range of ecological niches, from herbivores and predators to scavengers, parasitoids, and parasites (reviewed by Disney 1994). Nonetheless, previous ecological studies focusing on scuttle flies (Durska 1996, 2001, 2002, 2003, 2006, 2013, 2020; Durska et al. 2005) have been of limited scope due to time consuming morphological identification methods.

In this study, we use sorting with megabarcoding to generate the data for answering fundamental questions about this dark taxon. We ask how many species of scuttle flies occur in Sweden, how are they distributed across time and space, and what environmental variables drive their distribution. To test whether the choice of species and species delimitation method will affect the results, we carry out the same analyses using both mOTUs as species-proxies and haplotype diversity, and test whether rare species influence the overall inferences. We show how that "Dark Taxon Science" can quickly yield the answers urgently needed in the Anthropocene.

RESULTS

Diversity

We obtained a total 31 739 *COI* barcodes belonging to 2 697 haplotypes. At the species threshold (1.7%) we detected 549 mOTUs. Species accumulation curves revealed that scuttle fly species diversity was incompletely sampled overall and across sites, zones, and time periods (Fig. 2, Supp. Figs. S3, S4). Between 38 and 145 species were observed per site, with Chao1 richness estimates per site ranging from 83 to 244 species (Fig. 2b,c). Regional richness estimates varied between nearly 400 species estimated in the southernmost coastal zone (Zone 1), to fewer than 200 estimated for the alpine zone (Supp. Fig. S3). Midsummer and late-summer time-periods were both characterised by richness estimates of around 450 species, while the late spring and offseason time-periods showed lower species richness estimates of around 300 species (Supp. Fig. S4). Total species richness in Sweden was estimated at between 652 species (for Chao1, Fig. 2a) and 713 species (for the CNE estimator), suggesting that 100-160 species of scuttle flies remain to be sampled.

Ordinations of community composition

NMDS plots revealed a clear distinction between scuttle fly communities in the southern (zones 1-3) and northern (zones 4-alpine) plant hardiness zones (Supp. Fig. S5). ANOSIM supported significant differences between southern and northern zones at this threshold (R=0.58, p=0.001), and SIMPER revealed that North-South similarity was just 13.2%, as compared to similarities of 24.6% and 23.2%, respectively, among samples within the northern and southern zones (Supp. Tables S3-S4).

The separation between zones was consistent across clustering thresholds ranging from haplotypes-as-such to a threshold of 1.7% sequence similarity, with stress values around 0.21 (Supp. Fig. S5 top

row). Above a threshold of 3%, however, the patterns were increasingly blurred and the stress values were higher (0.25-0.27) (Supp. Fig. S5 bottom row). These patterns were also evident in ANOSIM analyses, where the sample statistic decreased from 0.54 for haplotypes to 0.27 for 5% mOTUs, indicating a decreasing relationship between scuttle fly community composition and plant hardiness zones at higher clustering thresholds (Supp. Table S2). Similarly, in SIMPER analyses, average similarity between zones increased from 20.2% for haplotypes to 41.4% for 5% mOTUs (Supp. Table S3). For species, scuttle fly communities were found to be distinct across most zones (R=0.51, p=0.001), with higher (average=28.5%) similarity within zones than between (average=19.2%) zones (Supp. Tables S2, 3).

Samples from the late spring, midsummer, and late-summer time-periods showed a clear progression along the ranked plant hardiness zones, while offseason samples appeared more randomly distributed (Supp. Fig. S6). Northern sites (IV-alpine) showed higher distinctness across seasons (R=0.59) than when considering all sites (R=0.27) (Supp. Table S4). The between-season similarity of all sites averaged 16.9% and within time-period similarity averaged 23.7% (Supp. Table S5, top). However, for northern sites only (IV-alpine), between time-period similarity was 21.3% and within time-period similarity was 37.2% (Supp. Table S5, bottom).

Hierarchical Modelling of Species Communities (HMSC)

Models of the four response matrices (species-occurrence, haplotypes-occurrence, species-abundance, haplotypes-abundance) showed relatively good MCMC convergence with potential scale reduction factors of the models' beta and omega parameters close to one (Supp. Fig. S7).

Explained variance averaged c. 30% for the occurrences and c. 60% for the abundances of both species and haplotypes, but there were large differences in model fit among taxa (Supp. Fig. S8), haplotype occurrence mean ± sd Tjur R^2=0.30 ± 0.15 (range 0.05 – 0.82); haplotype abundance R^2=0.61 ± 0.26 (range 0.00-1.00); species presence-absence Tjur R^2=0.30 ± 0.14 (range 0.05-0.72); species abundance R^2=0.59 ± 0.26 (range 0.02-1.00)).

Sampling time-period explained the largest fraction of variance, on average, in all four HMSC models (mean 10-11% in the occurrence models; mean 15-16% in the abundance models; Fig. 3, Supp. Fig. S8). Mean annual temperature explained almost as much variance as sampling time-period in the occurrence models (mean 7% and 8% in the species vs haplotype models), but clearly less than sampling season in the abundance models (mean 6% in the species models and 5% in the haplotype models) (Fig. 3, Supp. Fig. S8). All response variables showed strong climatic (including temporal) community structure and there was also strong spatial structure in species and haplotype occurrences linked with the latitudinal temperature gradient across Sweden (Fig. 3). Tree cover explained more variance in the abundance models (mean 6% for both species and haplotypes) than in the occurrence models (mean 1% for both species and haplotypes). Trapping effort also explained more variance on average in the abundance models (mean 7% and 8%) than in the occurrence models (mean 1.5% for both), as did the effect of having (vs not having) sequenced the full trap sample (2% mean for occurrence, 6% mean for abundance).

While the occurrences of taxa and haplotypes showed statistically supported responses to all model covariates (Fig. 4), their abundances showed responses less frequently, and responses with strong support were primarily with seasonal covariates (Fig. 4).

Most species and haplotypes with a statistically supported seasonal abundance trend peaked in the late spring. Compared to late spring, the midsummer fauna showed a reduction of 28% in species counts and 16% in haplotype counts, respectively, with a further reduction of 24% in both species counts and haplotype counts towards the late summer. Taxa and haplotypes were also usually more prevalent in late spring than in mid- or especially late summer. However, a minority of species and haplotypes (9% and 7%, respectively) showed the reverse pattern, being more prevalent in late summer than in the spring. Offseason captures in the late fall to early winter were consistently low, and the number of offseason samples included in the models was the smallest (n = 20 samples). Nonetheless, 33% of the species modelled and 29% of the haplotypes modelled were occasionally detected in offseason samples.

Taxon occurrences showed a mixture of positive (29% vs 30% for species and haplotypes, respectively) and negative (18% vs 23%) responses to the annual temperature gradient ("bio1") across Sweden (Fig. 4), reflecting the broad-scale compositional changes from southern towards northern Sweden seen in the NMDS ordinations. Where detected, the occurrence responses of taxa to forest or woodland cover were more often positive than negative (10% positive vs 4% negative for species; 8% vs 2% for haplotypes; Fig. 4). Longer trapping periods did not consistently result in higher detection probabilities of taxa, presumably due to seasonal differences in trapping duration (4% positive vs 5% negative responses for species; 8% positive vs 4% negative for haplotypes; Fig. 4). As expected, most taxa (81-82% in both the species and haplotype models) showed a negative occurrence response to the binary variable indicating whether all specimens in a sample were sequenced or not ("FullSample") (Fig. 4). The mean annual temperature gradient across Sweden was predicted to affect the prevalence of 38% of species, but does not appear to be a main driver of species abundance (Fig. 4). The predicted effect of the temperature gradient on species prevalence during late summer was positive in 22% of taxa and negative in 16% of taxa (Supp. Fig. S9). Beyond spatial patterns explained by these climatic predictors, there was evidence of localised spatial autocorrelation in species and haplotype site occupancies at scales of less than 40 km. Neither species nor haplotype abundances were spatially autocorrelated, nor did we detect statistical support for temporal autocorrelation in any model.

Residual correlations in the distributions of taxa were detected among sites and samples in both the species and especially the haplotype occurrence models (Supp. Fig. S10). Residual associations of taxa over time were also evident, but less frequently (Supp. Fig. S10). Residual associations between taxon/haplotype occurrences likely indicate that our models either lack or imperfectly represent some of the variables that structure the occurrences of scuttle fly taxa and haplotypes in space and time. No residual associations among taxa were evident among sites or over time in the species or haplotype abundance models, and very few were detected among samples (Supp. Fig. S10). Hence, covariance in the abundances of taxa and haplotypes across occupied samples was well modelled by the environmental and other covariates in these models.

The observed richness of the excluded rare vs modelled common species and haplotypes was strongly positively correlated (R = 0.59 for species and R = 0.80 for haplotypes). Compositional differences between samples in terms of the rare vs common species and haplotypes were also positively correlated (for taxon presence-absence Mantel R = 0.31 for species and Mantel R = 0.64 for haplotypes; for taxon abundance Mantel R = 0.30 for species and Mantel R = 0.60 for haplotypes).

Community Dissimilarity

Turnover accounted for the bulk of spatial, temporal and spatio-temporal variation in community dissimilarity (Fig. 5). In the spatial analyses, the mean turnover of species and haplotypes between sample pairs was 0.75 and 0.87, respectively, while the corresponding mean values of nestedness were 0.07 and 0.03. In temporal analyses, similarly, the mean values of turnover were 0.45 and 0.74 for species and haplotypes respectively, and the corresponding means of nestedness were 0.13 vs. 0.07. Finally, mean turnover values for the spatio-temporal analyses were 0.75 and 0.84 vs a mean nestedness of 0.07 and 0.04 for species and haplotypes, respectively.

Regardless of the time-period and operational taxonomic units used, we found a significant positive correlation between turnover and geographical distances, meaning that communities in closer proximity to each other are more similar in composition (Supp. Fig. S11). The mean spatial turnover of species communities was highest in the summer time periods (midsummer: 0.82, late summer: 0.84) and lower in late spring and offseason (0.78 and 0.72, respectively). However, we did not find any significant correlation between temporal distances (difference in mean week) and turnover (see Table S2). Patterns of nestedness showed no detectable correlation with any of the distances explored (except for the grouped temporal distance) (Supp. Table S2). Within each time-period, scuttle fly communities become more distinct from neighbouring communities from late spring to late summer (Supp. Table S2 and Supp. Fig. S11-13).

DISCUSSION

Our study marks the first country-wide examination of a dark taxon's diversity and distribution. It revealed more than 500 species based on processing ca. 31 800 specimens. The ecological analyses suggest that climate change may have profound (and quickly apparent) effects on communities of scuttle flies, that may serve as early indicators of future shifts in the environment. We demonstrate that armed with recent advancements in sequencing technologies, bioinformatics pipelines and molecular barcoding workflows (Meier et al. 2006; Wang et al. 2018, 2018a; Srivathsan et al. 2019b, 2021; Yeo et al. 2020; Hartop et al. 2021), we are now able to resolve patterns in the alpha diversity, spatial and temporal turnover, and species communities of challenging dark taxa. We show that for the scuttle flies in Sweden, a major part of all species remains undiscovered, that local communities show major turnover in space and time, and that ecological patterns are largely robust to the finer details of species delimitation. Below, we will address each of these findings in further detail – and the importance of these patterns in dark taxon biology to biodiversity science.

Diversity

Our sample only scratched the surface of the scuttle fly fauna of Sweden (Fig. 2, Supp. Figs. S3, S4). While the true Swedish fauna of scuttle flies is still hard to estimate, the 549 putative species found (and 652-713 predicted based on current sampling) greatly exceed the 374 previously documented from the country. Previous estimates for the scuttle fly fauna have ranged from 1 100 to nearly 2 000 species, suggesting that the true numbers may be even higher (Ronquist et al. 2020). This would not be surprising, given that our study included sampling from just 37 traps in a country of over 450 000 square kilometers; i.e., many habitats remained unsampled. However, the available data confirms that for dark taxa the basic pattern in ecology holds that most species are rare and that finding all species would require massive sampling (Callaghan et al. 2023).

Where the remaining species will be lurking is unknown. While Malaise traps are remarkably effective at capturing flies (Karlsson et al. 2020; Skvarla et al. 2020), other trapping methods would certainly reveal additional species. An "All Diptera Biodiversity Inventory" conducted in Costa Rica utilised a wide variety of methods (Borkent and Brown 2015), revealing that 59% of fly species were unique to a specific collection method (Borkent et al. 2018). Intriguingly, species-rich taxa with particularly interesting and often impactful biologies (e.g. parasitoids) may be underrepresented in Malaise trap samples. For example, Brown (1996) reported that nearly half of the species of ant-parasitising scuttle flies were caught exclusively over army ant raids. Similar results were obtained by Disney et al. (1982), who found some species of scuttle flies were uniquely or primarily collected in water traps rather than Malaise traps. Another possible frontier is the canopy. Ongoing research in upper levels of the canopy in the Amazon suggests that it contains many scuttle fly species not found in lower sampling elevations (Gilliland 2021), contrasting with previous data showing scuttle flies as a taxon primarily collected at ground-level (Kitching et al. 2005). In a limited sample (159 specimens) from canopy trap prototypes run concurrently with the Malaise trap sampling for this study we found a single unique mOTU not seen in the 31 794 scuttle flies from ground level, indicating that even in a temperate environment we may have more to find in the canopy.

Thus, no single sampling method will suffice to reveal all taxa and additional trapping sites that target regions or habitats un- or under-represented in the current study will also be needed for an exhaustive inventory of scuttle flies. Some of this complementary sampling may have already been carried out by the Swedish Malaise Trap Project years ago (Karlsson et al. 2020). An excellent follow-up to this study would be sequencing scuttle fly samples from that sampling campaign, or from another insect campaign that is ongoing in Sweden – the Insect Biome Atlas project (www.insectbiomeatlas.com).

In addition to focusing future efforts on revealing more species, the analysis of other life stages may offer a more nuanced understanding of scuttle fly communities. Our study is based on a sample of adult flies. Patterns for the longer-living larval stages are at this point unknown, as they are not easily collected.

Lessons from Ecological Analyses

Our ordinations of community composition suggested strong structuring of scuttle fly communities by Swedish horticultural zones (Fig. 1) and season (Supp. Fig. S6). Clear north-south structuring of the Swedish insect fauna has also been observed in damselflies, mosquitos, and caddisflies (Gullefors 2008, Wellenreuther et al. 2012, Lundström et al. 2013), and strong phenological patterning is a hallmark of any high-latitude fauna (Wolda, 1988). Our ordinations also indicated that spatial structuring was largely independent of mOTU clustering threshold at or below the species proxy level (Supp. Fig. S5).

To address these findings more rigorously and without the constraints of pre-determined zones, we implemented HMSC. This confirmed that scuttle fly community variation is both highly seasonal and strongly tied to the latitudinal temperature gradient over Sweden, with all models strongly predicted by climatic covariates (Fig. 3, Supp. Fig. 8). The presence of significant spatial autocorrelation in the HMSC occurrence models reflects the gradation of scuttle fly distributions across space. Scuttle fly communities also showed clear compositional changes over time during the warm season, from late spring through mid and late summer, while offseason sampling was too inconsistent to reveal any patterns (Supp. Figs. S2, S6). Consistent with our findings that scuttle fly communities are driven by climatic covariates, we found more rapid spatial turnover in taxa at the species than at the haplotype level (Fig. 5). Recent work has predicted that when dispersal limitation is the dominant driver of species distributions, the rate of spatial turnover of biological communities will be similar at both the haplotype and species levels (Baselga et al. 2022). Conversely, when environmental conditions strongly constrain species ranges, community similarity is predicted to decay at different rates across genealogical scales.

Our results have several implications in the face of climate change. Individual taxa show a mixture of positive and negative responses to mean annual temperature, and to a lesser extent to seasonality. This implies that we may see a substantial number of both climate change "winners and losers" in the future, as species ranges and phenology expand, contract, or shift (Fig. 4, Supp. Fig. S9). Adult scuttle flies are ephemeral – with short lifespans and high turnover and mobility – they may respond rapidly and serve as indicators for future shifts of other taxa and in the environment more broadly. Our observation of steeply declining abundance from late spring through the summer into the offseason may partially reflect the extreme temperatures and drought in Europe in summer 2018 (Bakke et al. 2020, Bastos et al. 2020). While this suggests that our seasonal results may be atypical, they are perhaps also a sign of summers to come, as climate change increases the frequency and severity of these events (Hari et al. 2020).

With our initial results confirmed by HMSC, the predictive value of the simple plant hardiness map for scuttle fly distributions offers excellent news. It suggests that even such a relatively coarse-grained tool can be used as a reliable indicator of the regional compositions of scuttle flies. That this is the case is only intuitive: Many scuttle flies exist close to ground level and therefore, like plants, their distributions may be closely tied to soil temperature, acidity, moisture, or composition (all of which would be interesting covariates to explore in future modelling). Additionally, some species are known to have direct interactions with plants, fungi, and ground-dwelling insects – factors which may again tie them to the microclimate at ground level. Future studies might focus on these

microenvironmental variables, to address the unexplained variance in scuttle fly communities from this study.

Dark Taxon Science

Dark taxa have historically been largely ignored due to the complexities involved in studying groups of highly diverse and abundant organisms of small size. How, then, do we make the study of Dark Taxa efficient and start "dark taxon science" to respond to the need for quantitative data on biodiversity?

To address this, we assessed whether the taxa analysed need to exactly match taxonomically validated species for ecological hypotheses to be addressed. If not, we may avoid endless discussions about species delimitation and can proceed with ecological analyses. Promisingly, our current results suggest that ecological patterns are largely robust to the exact clustering thresholds used. Overall biodiversity patterns, patterns of community turnover and drivers of distribution were virtually identical using thresholds from 0% (haplotype data) to 1.7% (Figs. 3-5, Supp. Figs. S5, 8, 11-13). Haplotype data are convenient because they do not require any taxonomic decisions, which may please molecular ecologists, with the important caveat that exceeding an appropriate species proxy threshold will obscure patterns because too many species will be lumped together (Supp. Fig. S5, bottom row).

A second potential stumbling block in the study of dark taxa lies in the large numbers of rare species. If understanding basic ecological patterns is dependent on these rare species, dark taxon science will be difficult. Our results indicate that both richness and compositional differences between samples of rare versus common species and haplotypes were both positively correlated. This suggests that more common and rarer species respond similarly to the same drivers. Again, this is excellent news, since it suggests that ecological inferences regarding the drivers of species distribution and community composition can be based on the more common species – which are much easier to detect.

Conclusion

We here argue for a dark taxon science and illustrate that it is not a hopeless undertaking by targeting scuttle flies as a typical dark taxon in Sweden. We sample across the entire country, we estimate the species richness, resolve patterns of distribution across time and space, and pinpoint environmental features that drive these distributions. Our results suggest that such assessments will be insensitive to specific taxonomic cut-offs, and robust to undersampling of rare taxa. Overall, the study hopes to contribute to a more quantitative approach to biodiversity. In the future, advances in molecular workflows, bioinformatics, robotics and automation will make these groups increasingly efficient to study (Meier et al. 2024). We hope that our case study will serve to propel forward Dark Taxon Science, bringing the main part of diversity into the realm of biodiversity science.

MATERIALS AND METHODS

Target taxon: the scuttle flies of Sweden

Sweden has one of the best-known animal faunas in the world due to efforts dating from Carl Linnaeus to the Swedish Taxonomy Initiative and Malaise Trap Project (Linnaeus 1735; Miller 2005; Karlsson et al. 2020, but see Ronquist et al. 2020). While 374 species of scuttle flies have been documented in the country (SLU Artdatabanken 2021), this is an underestimate. For comparison, a single suburban garden in Cambridge, UK, yielded nearly 100 species of scuttle flies (Brown and Hartop 2017), while backyards in Los Angeles, California, can support up to 82 species (Brown and Hartop 2017). Previous estimates of scuttle fly diversity in Sweden have proposed that the true fauna may approach 2000 species (CNE estimate in Ronquist et al. 2020), but this may be an overestimate based on an error-prone process of morphological identification (Ronquist et al. 2020).

Sampling

To start resolving the species richness, spatiotemporal distribution, and community composition of Swedish scuttle flies, we sampled communities of flying insects at 37 locations across Sweden (Fig. 1, Supp. Fig. S1, Supp. Table S1). These samples were collected in ~80% ethanol using Townes-style Malaise traps (Townes 1972). Scuttle flies were sorted from the trap samples and preserved in ~80% ethanol at -20°C.

Sampling started in May 2018 and continued into the following year. Although trapping was continuous, the sample periods varied across sites due to their extensive distribution across Sweden. Four time periods that matched the summer phenology were therefore established in sequence for each site: late spring, midsummer, late summer, and offseason (Supp. Fig. S2, Supp. Table S1). Some offseason samples could not be retrieved in late 2018 and were collected in 2019. To ensure uniformity in the number of weeks per sampling period and avoid introducing biases, we attributed the latest sampling date for the offseason traps in 2018 to those samples collected in 2019. Additionally, late spring samples were only obtained from 25 sites, as 12 sites were not installed until later in the season (see Supp. Table S1 for sampling details). To compare richness estimates across seasons, we plotted accumulation curves for each time-period excluding the 12 traps that were not yet installed in late spring.

Sequencing and bioinformatics

A total of 136 samples were selected for analysis. Most samples (94) contained thousands of scuttle fly specimens and subsampling was needed. Individuals were randomly selected with a minimum of two 96-well microplates of scuttle flies processed per sample. This resulted in a total of at least 190 specimens extracted per site and time-period for most samples. 31 samples containing fewer than 190 specimens and were thus processed completely, and two samples were found to contain no scuttle flies. Additional plates of specimens from nine samples that had been processed in an earlier study (Hartop et al. 2022) were also included (Supp. Table S1). All sample information, including numbers of barcodes obtained per sample and whether a sample was fully processed, is found in Supp. Table S1.

DNA were obtained from the specimens using 10 µL of HotSHOT lysis buffer (Truett et al. 2000). The extraction was carried out in a thermocycler at 65 °C for 18 minutes, then for 2 minutes at 98 °C. Amplification was carried out on a 313-bp fragment of Cytochrome Oxidase 1 (*COI*) using primers m1COIintF: 5'-GGWACWGGWTGAACWGTWTAYCCYCC-3' (Leray et al. 2013) and modified jgHCO2198: 5'-TANACYTCNGGRTGNCCRAARAAYCA-3' (Geller et al. 2013). For a small subset of samples, *COI* was amplified using the primer pair jgHCO2198 and LCO1490 (Folmer et al. 1994; Geller et al. 2013). Amplifications were conducted with tagged primers following Wang et al. (2018) for Illumina and Srivathsan et al. (2019b, 2021) for MinION. PCR reactions contained 4 µl Mastermix from CWBio, 1 µl of 1 mg/ml BSA, 1 µl of 10 µM of each primer, and 1 µl of DNA. PCR conditions were a 5 min initial denaturation at 94 °C followed by 35 cycles of denaturation/annealing/extension (94 °C (1 min)/47 °C (2 min)/72 °C (1 min)), and a final extension at 72 °C (5 min). A subset of wells (N=8-12 with negative control) from each PCR plate was run on an agarose gel to check for plate-wide failure, before products were pooled and then purified using AMPure XP beads (Beckman Coulter Life Sciences, Indiana, USA). DNA concentration was quantified with a Qubit™ dsDNA HS Assay Kit (Invitrogen, California, USA).

Illumina and MinION sequencing were used to sequence the amplicons for this study. Illumina libraries were prepared using a TruSeq DNA PCR-free kit to obtain 250bp PE sequences using Illumina HiSeq 2500; the sequencing was outsourced. Nanopore sequencing using MinION was conducted in house following Srivathsan et al. (2019b). Library preparation was conducted using either SQK-LSK109 or SQK-LSK110 ligation sequencing kits (Oxford Nanopore Technologies, Oxford, UK) with 200-ng pooled and purified PCR products. The manufacturer's instructions were followed except for use of 1× Ampure beads instead of 0.4× as suggested in the instructions because the minibarcode amplicons in our experiments were short (~391 bp with primers and tags), and a modified protocol for end-repair as described in Srivathsan et al. (2021). The sequencing was carried out using a MinION sequencer with either R9.4.1 or R10.3 flow cells for a maximum of 72 hours. Basecalling was conducted using Guppy versions 2.3.5+53a111f and 4.2.3+f90bd04. FastQ files were demultiplexed to obtain individual *COI* sequences identifiable back to individual specimens with unique specimen codes. This was done following Srivathsan et al. (2019a, 2021) for MinION and Wang et al. (2018) for Illumina.

To exclude contaminants and potential mis-sorts, *COI* sequences were matched using BLAST to NCBI GenBank's nucleotide database and sequences with >97% similarity to non-scuttle fly taxa were removed. Sequences were then aligned using MAFFT v7 (Katoh and Standley 2013) and clustered into molecular Operational Taxonomic Units (mOTUs) using Objective Clustering (part of "TaxonDNA", see Meier et al. 2006), a distance-based method that groups sequences based on a user-defined threshold for minimum interspecific uncorrected p-distance.

A species-proxy clustering threshold of 1.7% was used for the primary analyses as this was the distance that maximized cluster congruence to species-level morphology for a subset of 18 000 scuttle flies from this dataset (Hartop et al. 2022). Hereafter, references to "species" are to mOTUs at this threshold. To assess the impact of the selected threshold on distribution patterns, we examined other thresholds ranging from 0-5% and ran models using both species and haplotype data.

Diversity

To characterise local and national species richness, we used accumulation curves and Chao1 richness, as implemented in EstimateS (Colwell 2013) and R (R Development Core Team) package iNEXT (Chao et al. 2014) for total diversity, diversity per zone, and diversity per site. Following the approach of Ronquist et al. (2020), a combined non-parametric estimator (CNE) was used to obtain an alternative estimate of total species richness.

Characterisation of community variation

To visualise scuttle fly community composition in the context of geographic and climatic variation, we used Non-metric Multi-Dimensional Scaling (NMDS) plotted with R packages ggplot2 (Wickham, 2016) and Vegan v2.5-4 (Oksanen et al., 2019) using both abundance-based (Bray-Curtis) and incidence-based (Jaccard) dissimilarity indices. Since both metrics yielded nearly identical results, the results in the paper are based on Bray-Curtis indices. Sites were classified according to the nine plant hardiness zones of the Swedish Horticultural Society that synthesize climatic variables of particular importance to horticultural plants into horticultural zones: strong and rapid temperature changes, very low temperatures sustained for long periods, evaporation occurring when the sun is shining but the ground is still frozen, and temporal considerations of climatic and environmental conditions (Fig. 1, Riksförbundet Svensk Trädgård 2018, used with permission). As these zones have proven useful in identifying the survival and growth of horticultural plants across Sweden, we hypothesized that they may offer a relevant description of the environment for any organism sensitive to similar climatic variables – including scuttle flies. They also offer a more regional approach than the broad biogeographic classification previously used for spatial analysis of scuttle flies in Sweden (Ronquist et al. 2020). We therefore tested whether the zones could predict scuttle fly richness and distributions (Disney 1994; Kitching et al. 2005; McGlynn et al. 2019).

To quantify differences in scuttle fly community composition between plant hardiness zones and to assess the significance of those differences, we used Analysis of Similarities (ANOSIM) and Similarity Percentage analysis (SIMPER). These tests were run with PRIMER v7 (Clarke and Gorley 2006).

We excluded samples that contained fewer than 100 specimens from the ordination analyses, as small sample sizes can artificially generate large distances in the NMDS plots, thus obscuring structured variation in community composition arising from responses to the environment.

Hierarchical Modelling of Species Communities

To relate variation in community composition to continuous environmental drivers without the assumptions of zones used in our NMDS visualisations, we used Hierarchical Modelling of Species Communities (HMSC; (Ovaskainen et al. 2017; Ovaskainen and Abrego 2020), a type of joint species distribution model (Warton et al. 2015)). Due to the zero-inflated nature of the data, we fitted hurdle models, i.e., one model for the occurrence of taxa across samples (probit regression), and a second model for their abundance conditional on presence (linear regression for log-transformed count data, with zeros masked as missing data), henceforth referred to simply as an “abundance model”.

To evaluate the impact of the criterion used in species delimitation, we used four different response variables: species presence-absence, species abundance (where present), haplotype occurrence and haplotype abundance (where present), thus resulting in four HMSC models being fit. As these analyses are uninformative for taxa with very sparse data, we included only haplotypes or species with at least five occurrences (n = 391 haplotypes and 273 species, out of the 2 697 haplotypes and 549 species observed; see Results). To be able to relate each sample to specific climatic conditions, we also excluded 15 samples for which the trapping duration exceeded approximately a month (>34 days).

We included five predictor variables, coded as fixed effects, in our HMSC models. To test for seasonal changes in scuttle fly communities among the sampling time-periods, we included the four-category variable "Time.Period". To assess the effect of the spatial gradient in mean climatic conditions across Sweden on scuttle fly communities, we included the continuous variable mean annual temperature ("bio1" from the Worldclim database, Fick and Hijmans 2017). The importance of tree cover for scuttle fly distributions was assessed by including a continuous variable describing the percentage of forest or woodland cover within a 50m buffer ("ForestWood") derived from the Swedish National Land Cover Database (https://www.naturvardsverket.se/en/services-and-permits/maps-and-map-services/national-land-cover-database/). Site values for "bio1" and "ForestWood" were extracted using the R package raster (Hijmans 2024). To account for the effect of sampling effort on taxon detection, we further included a continuous variable quantifying the total number of trapping days per sample ("TrapDays") and a binary variable indicating whether all specimens in a sample were sequenced or not ("FullSample"). The latter, binary variable defines whether the sample was fully sequenced or not, and was included to account for the higher likelihood of encountering taxa in samples that had very high individual abundances (i.e., samples that could not be fully sequenced). We moreover included a spatially explicit random effect based on sample site coordinates ("Random: site") and a temporally explicit random effect based on median sampling date ("Random: time") to model any spatial or temporal autocorrelation in the scuttle fly dataset, and a categorical random effect representing sample identity ("Random: sample").
We fitted the model with the R-package HMSC (Tikhonov et al. 2020) assuming the default prior distributions. We sampled the posterior distribution of four MCMC chains, each of which was run for 37 500 iterations, of which the first 12 500 were removed as burn-in. The iterations were thinned by 100 to yield 250 posterior samples per chain, and thus 1 000 posterior samples in total. To explore the rate of Markov chain Monte Carlo (MCMC) convergence, we also fitted otherwise identical models but with 375 iterations (burn-in 125, thin 1) and 3,750 iterations (burn-in 1 250, thin 10). Convergence was assessed by examining the potential scale reduction factor (PSRF) distribution over the fixed effect (β) and random effect (Ω) parameters, equivalent to the Gelman–Rubin statistic (Gelman et al. 2014).

We examined the explanatory power of the models through species-specific coefficients of discrimination (Tjur R^2) for the occurrence models, which measure how well the model discriminates those samples in which a taxon occurs from those in which it does not occur, and R^2 for the abundance models. Tjur R^2 is defined as the difference in average model-predicted probability of occurrence for samples in which the species is present vs absent (Tjur 2009).

To assess whether the distributions of rare taxa are likely to be driven by similar factors to the common ones, we calculated Pearson and Mantel correlations, respectively, between the observed species richness and community composition (Bray-Curtis dissimilarity) of the modelled more common taxa and the excluded rare taxa (those with <5 occurrences).

Community dissimilarity

To further partition variation in scuttle fly community composition in time (between seasons) and space (across Sweden), we dissected overall community dissimilarity (β-diversity) into its turnover (i.e. species replacement) and nestedness (differences of species richness between sites) components (Baselga, 2010).

Our analyses encompassed three aspects: 1) spatial differences; differences in community composition between each sampling site considering their geographical distance, 2) temporal differences; differences in community composition between each sampling period, accounting for their temporal distance in weeks, and 3) spatio-temporal differences; community differences between each sampling site and sampling period including their joint effect. We computed pairwise Jaccard community dissimilarity values using both the species and haplotypes datasets for comparison. Given the observed geographical distance between sampling sites, these distances were rescaled to a mean of zero and a standard deviation of one before analyses. To characterise the temporal difference between samples in each sample pair, we used the difference in the mean week of sampling (for sample-specific details, see Supp. Table S1). Values of total β-diversity, turnover, and nestedness were calculated for each pairwise comparisons of sampling sites, sampling period and sample pairs. We excluded self-pairs and included data for each pair only once. To test the correlation between each beta diversity component and the distances in space and time we used Mantel tests based on Pearson moment correlations. All calculations were implemented in R package `betapart´ (Baselga & Orme, 2012).

ACKNOWLEDGEMENTS

We thank the Scuttle Fly Sorting Party crew at Station Linné – Carina Romero Ugarph, Harald Havnås, Johan Ennerfelt, Marianne "Mia" Blomqvist, Nino Pettersson, Robert Ennerfelt, and especially Dave Karlsson. We thank the members of the Evolutionary Biology Lab at the National University of Singapore for help with the many hours of wet lab work. We thank Darren Yeo for assisting us with visualisations. We thank Inger Ekrem at Riksförbundet Svensk Trädgård for helping with the plant hardiness zone map, Tomas Lagerström for further information on the map, and Eva Ronquist for first bringing our attention to the map. We thank the Swedish Taxonomy Initiative for the support to investigate the scuttle flies of Sweden (grant 2016-203 4.3 to EH), this study brings us one step closer to the ultimate goal of describing all of these species. TR and OO were funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ERC-synergy grant 856506—LIFEPLAN). OO was funded by Academy of Finland (grant no. 336212 and 345110). MJ was supported by the Academy of Finland's 'Thriving Nature' research profiling action.

DATA AVAILABILTY

Data and scripts are available on the project GitHub page at https://github.com/leshonlee/DocumentingPhorids. General HMSC pipeline scripts are available at https://www2.helsinki.fi/en/researchgroups/statistical-ecology/hmsc.

REFERENCES

Bakke, S. J., Ionita, M., & Tallaksen, L. M. 2020. The 2018 northern European hydrological drought and its drivers in a historical perspective. Hydrology and Earth System Sciences, 24(11), 5621–5653. https://doi.org/10.5194/hess-24-5621-2020

Bar-On, Y. M., Phillips, R., & Milo, R. 2018. The biomass distribution on Earth. Proceedings of the National Academy of Sciences, 115(25), 6506–6511. https://doi.org/10.1073/pnas.1711842115

Baselga, A. 2010. Partitioning the turnover and nestedness components of beta diversity. Global ecology and biogeography, 19(1), 134-143.

Baselga, A., Gómez-Rodríguez, C., Araújo, M. B., Castro-Insua, A., Arenas, M., Posada, D., & Vogler, A. P. 2022. Joint analysis of species and genetic variation to quantify the role of dispersal and environmental constraints in community turnover. Ecography, 2022(5), e05808. https://doi.org/10.1111/ecog.05808

Baselga, A., & Orme, C. D. L. 2012. betapart: an R package for the study of beta diversity. Methods in ecology and evolution, 3(5), 808-812.

Bastos, A., Ciais, P., Friedlingstein, P., Sitch, S., Pongratz, J., Fan, L., Wigneron, J. P., Weber, U., Reichstein, M., Fu, Z., Anthoni, P., Arneth, A., Haverd, V., Jain, A. K., Joetzjer, E., Knauer, J., Lienert, S., Loughran, T., McGuire, P. C., … Zaehle, S. 2020. Direct and seasonal legacy effects of the 2018 heat wave and drought on European ecosystem productivity. Science Advances, 6(24), eaba2724. https://doi.org/10.1126/sciadv.aba2724

Bell J.R., Botham M.S., Henrys P.A., Leech D.I., Pearce-Higgins J.W., Shortall C.R., Brereton T.M., Pickup J., Thackeray S.J. 2019. Spatial and habitat variation in aphid, butterfly, moth and bird phenologies over the last half century. Glob. Change Biol. 25:1982–1994.

Bogoni J.A., Peres C.A., Ferraz K.M.P.M.B. 2020. Extent, intensity and drivers of mammal defaunation: a continental-scale analysis across the Neotropics. Sci. Rep. 10:14750.

Borkent A., Brown B.V. 2015. How to inventory tropical flies (Diptera)-One of the megadiverse orders of insects. Zootaxa. 3949:301–322.

Borkent A.R.T., Brown B.V., Adler P.H., Amorim D.D.S., Barber K., Bickel D., Boucher S., Brooks S.E., Burger J., Burington Z.L., Capellari R.S., Costa D.N.R., Cumming J.M., Curler G., Dick C.W., Epler J.H., Fisher E., Gaimari S.D., Gelhaus J.O.N., Grimaldi D.A., Hash J., Hauser M., Hippa H., IbÁÑEz-Bernal S., Jaschhof M., Kameneva E.P., Kerr P.H., Korneyev V., Korytkowski C.A., Kung

G.-A., Kvifte G.M., Lonsdale O., Marshall S.A., Mathis W.N., Michelsen V., Naglis S., Norrbom A.L., Paiero S., Pape T., Pereira-Colavite A., Pollet M., Rochefort S., Rung A., Runyon J.B., Savage J., Silva V.C., Sinclair B.J., Skevington J.H., Stireman J.O. III, Swann J., Vilkamaa P., Wheeler T., Whitworth T., Wong M., Wood D.M., Woodley N., Yau T., Zavortink T.J., Zumbado M.A. 2018. Remarkable fly (Diptera) diversity in a patch of Costa Rican cloud forest: Why inventory is a vital science. Zootaxa. 4402.

Brown B. V., Hartop E.A. 2017. Big data from tiny flies: patterns revealed from over 42,000 phorid flies (Insecta: Diptera: Phoridae) collected over one year in Los Angeles, California, USA. Urban Ecosyst. 20:521–534.

Brown B.V. 1996. Phorid Newsletter 5. Phorid Newsl. 5.

Callaghan, C.T., Borda-de-Água, L., van Klink, R., Rozzi, R., Pereira, H.M. Unveiling global species abundance distributions. Nat Ecol Evol 7, 1600–1609 2023. https://doi.org/10.1038/s41559-023-02173-y

Cardinale B.J., Duffy J.E., Gonzalez A., Hooper D.U., Perrings C., Venail P., Narwani A., MacE G.M., Tilman D., Wardle D.A., Kinzig A.P., Daily G.C., Loreau M., Grace J.B., Larigauderie A., Srivastava D.S., Naeem S. 2012. Biodiversity loss and its impact on humanity. Nature. 486:59–67.

Chao A., Ellison A.M., Colwell R.K., Gotelli N.J., Sander E.L., Hsieh T.C., Ma K.H. 2014. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecol. Monogr. 84:45–67.

Chua PYS, Bourlat SJ, Ferguson C, Korlevic P, Zhao L, Ekrem T, Meier R, Lawniczak MKN. Future of DNA-based insect monitoring. Trends Genet. 2023 Jul;39(7):531-544. doi: 10.1016/j.tig.2023.02.012. Epub 2023 Mar 10. PMID: 36907721.

Clarke K., Gorley R.N. 2006. PRIMER v6: user manual/tutorial. PRIMER-E, Plymouth. 29:1060–1065.

Colwell R.K. 2013. EstimateS: Statistical estimation of species richness and shared species from samples. Version 9 and earlier. User's Guide and application. .

Díaz S., Fargione J., Chapin F.S., Tilman D. 2006. Biodiversity loss threatens human well-being. PLoS Biol. 4:1300–1305.

Disney R.H.L. 1994. Scuttle flies: the Phoridae. London: Chapman and Hall.

Disney R.H.L., Durska E. 2008. Conservation evaluation and the choice of faunal taxa to sample. Biodivers. Conserv. 17:449–451.

Disney R.H.L., Erzinclioglu Y.Z., D. J. de C. Henshaw D.H., Unwin D.M., Withers P., Woods A. 1982. Collecting methods and the adequacy of attempted fauna surveys, with reference to the Diptera. Field Stud. 5:607–621.

Durska E. 1996. The species composition and structure of scuttle fly communities (Diptera: Phoridae) in mature tree stands in pine forests at different stages of habitat degradation. Fragm. Faun. 39:267–285.

Durska E. 2001. Secondary succession of scuttle fly communities (Diptera: Phoridae) in moist pine forest in Bialowieza Forest. Fragm. Faun. 44:79–128.

Durska E. 2002. The phenology of dominant scuttle-fly (Diptera: Phoridae) species in the Bialowieza Forest. Entomol. Fenn. 13:123–127.

Durska E. 2003. The phenology of Triphleba Rondani species (Diptera: Phoridae) in moist pine forests in the Bialowieza Forest. Entomol. Fenn. 14:177–182.

Durska E. 2006. Diversity of scuttle fly (Diptera: Phoridae) communities in the plantations of moist pine forests of the Bialowieza Primeval Forest and the Tuchola Forest (Poland). Biodivers. Conserv. 15:385–393.

Durska E. 2013. Effects of disturbances on scuttle flies (Diptera: Phoridae) in Pine Forests. Biodivers. Conserv. 22:1991–2021.

Durska E. 2020. Preliminary data of the scuttle flies (Diptera: Phoridae) in the linden-oakhornbeam forest of the Wigry National Park, North East Poland. :6.

Durska E., Kaczorowska E., Disney R.H.L. 2005. Scuttle flies (Diptera: Phoridae) of saline habitats of the Gulf of Gdansk, Poland. Entomol. Fenn. 16:159–164.

Fick, S.E. and R.J. Hijmans, 2017. Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology 37: 4302-4315.

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. Mol. Mar. Biol. Biotechnol. 3:7.

Garda A.A., Stein M.G., Machado R.B., Lion M.B., Juncá F.A., Napoli M.F. 2018. Ecology, biogeography, and conservation of amphibians of the Caatinga. Caatinga: The Largest Tropical Dry Forest Region in South America. Springer International Publishing. p. 133–149.

Geller J., Meyer C., Parker M., Hawk H. 2013. Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. Mol. Ecol. Resour. 13:851–861.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B. 2014. Bayesian data analysis (3rd ed.). London: CRC Press.

Gilliland, H. C. 2021 Hundreds of new and unusual insects discovered in the Amazon's canopy. Available from https://www.nationalgeographic.com/magazine/article/hundreds-of-new-and-unusual-insects-discovered-in-the-amazon-canopy-feature.

Goodsell, R., Tack, A., Ronquist, F., van Dijk, L., Iwaszkiewicz-Eggebrecht, E., Miraldo, A., Roslin, T., Vanhatalo, J. 2024. The rarity of Invertebrates prevents reliable application of IUCN Red-List criteria. EcoEvoRxiv. https://doi.org/10.32942/X23G71

Gullefors B. 2008. Limes norrlandicus - a natural biogeographical border for caddisflies (Trichoptera) in Sweden. Ferrantia. 55:61-65.

Hari, V., Rakovec, O., Markonis, Y., Hanel, M., & Kumar, R. 2020. Increased future occurrences of the exceptional 2018–2019 Central European drought under global warming. Scientific Reports, 10(1), 12207. https://doi.org/10.1038/s41598-020-68872-9

Hartop E., Srivathsan A., Ronquist F., Meier R. 2021. Towards Large-Scale Integrative Taxonomy (LIT): Resolving the Data Conundrum for Dark Taxa, Systematic Biology, 2022;, syac033, https://doi.org/10.1093/sysbio/syac033

Hijmans R (2024). raster: Geographic Data Analysis and Modeling. R package version 3.6-27, https://rspatial.org/raster.

Karlsson D., Hartop E., Forshage M., Jaschhof M., Ronquist F. 2020. The Swedish Malaise Trap Project: A 15 Year Retrospective on a Countrywide Insect Inventory. Biodivers. Data J. 8:e47255.

Katoh K., Standley D.M. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30:772–780.

Kitching R.L., Bickel D.J., Boulter S. 2005. Guild Analyses of Dipteran Assemblages: A Rationale and Investigation of Seasonality and Stratification in Selected Rainforest Faunas. The Evolutionary Biology of Flies. New York: Columbia University Press. p. 388–415.

Leray M., Yang J.Y., Meyer C.P., Mills S.C., Agudelo N., Ranwez V., Boehm J.T., 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. Front. Zool. 10:34.

Linnaeus C. 1735. Systema naturæ, sive regna tria naturæ systematice proposita per classes, ordines, genera, & species. Leiden: Lugduni Batavorum.

Lu Y., Yang Y., Sun B., Yuan J., Yu M., Stenseth N.C., Bullock J.M., Obersteiner M. 2020. Spatial variation in biodiversity loss across China under multiple environmental stressors. Sci. Adv. 6:1–11.

Lundström, J. O., Schäfer, M. L., Hesson, J. C., Blomgren, E., Lindström, A., Wahlqvist, P., Halling, A., Hagelin, A., Ahlm, C., Evander, M., Broman, T., Forsman, M., & Vinnersten, T. Z. P. (2013). The geographic distribution of mosquito species in Sweden. Journal of the European Mosquito Control Association. 31.

Mace G.M., Barrett M., Burgess N.D., Cornell S.E., Freeman R., Grooten M., Purvis A. 2018. Aiming higher to bend the curve of biodiversity loss. Nat. Sustain. 1:448–451.

McGlynn T.P., Meineke E.K., Bahlai C.A., Li E., Hartop E.A., Adams B.J., Brown B.V. 2019. Temperature accounts for the biodiversity of a hyperdiverse group of insects in urban Los Angeles. Proc. R. Soc. B. 286:9.

Meier, R., Hartop, E., Pylatiuk, C., Srivathsan, A. 2024. Towards holistic insect monitoring: species discovery, description, identification, and traits for all insects. Phil. Trans. R. Soc. B. 10.1098/rstb.2023-0120

Meier R., Shiyang K., Vaidya G., Ng P.K. 2006. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Syst Biol. 55:715–28.

Miller G. 2005. Linnaeus's Legacy Carries On. Science. 307:1038–1039.

Naeem S., Chazdon R., Duffy J.E., Prager C., Worm B. 2016. Biodiversity and human well-being: an essential link for sustainable development. Proc. R. Soc. B Biol. Sci. 283:20162091.

Ollerton J. 2017. Pollinator Diversity: Distribution, Ecological Function, and Conservation. Annu Rev Ecol Evol Syst. 48:353–76.

Orr M.C., Hughes A.C., Chesters D., Pickering J., Zhu C.D., Ascher J.S. 2020. Global Patterns and Drivers of Bee Distribution. Curr. Biol.

Ovaskainen O., Abrego N. 2020. Joint Species Distribution Modelling. Cambridge University Press.

Ovaskainen O., Tikhonov G., Norberg A., Guillaume Blanchet F., Duan L., Dunson D., Roslin T., Abrego N. 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. Ecol. Lett. 20:561–576.

Page R. 2011. Dark taxa: GenBank in a post-taxonomic world. Available from https://iphylo.blogspot.com/2011/04/dark-taxa-genbank-in-post-taxonomic.html.

Page R.D. 2016. DNA barcoding and taxonomy: dark taxa and dark texts. Philos Trans R Soc Lond B Biol Sci. 371.

Rabone, M., Wiethase, J. H., Simon-Lledó, E., Emery, A. M., Jones, D. O. B., Dahlgren, T. G., Bribiesca-Contreras, G., Wiklund, H., Horton, T., & Glover, A. G. 2023. How many metazoan species live in the world's largest mineral exploration region? Current Biology, S0960982223005341.

Riksförbundet Svensk Trädgård. 2018. Zonkartan. Available from http://www.tradgard.org/svensk_tradgard/zonkartan.html.

Ronquist F., Forshage M., Häggqvist S., Karlsson D., Hovmöller R., Bergsten J., Holston K., Britton T., Abenius J., Andersson B., Buhl P.N., Coulianos C.-C., Fjellberg A., Gertsson C.-A., Hellqvist S., Jaschhof M., Kjærandsen J., Klopfstein S., Kobro S., Liston A., Meier R., Pollet M., Riedel M., Roháček J., Schuppenhauer M., Stigenberg J., Struwe I., Taeger A., Ulefors S.-O., Varga O., Withers P., Gärdenfors U. 2020. Completing Linnaeus's inventory of the Swedish insect fauna: Only 5,000 species left? PLOS ONE. 15:e0228561.

Skvarla M.J., Larson J.L., Fisher J.R., Dowling A.P.G. 2020. A Review of Terrestrial and Canopy Malaise Traps. Ann. Entomol. Soc. Am.:21.

SLU Artdatabanken. 2021. Famij: Phoridae - puckelflugor. Available from https://www.dyntaxa.se/Taxon/Info/2001326.

Srivathsan A, Ang Y, Heraty JM, Hwang WS, Jusoh WFA, Kutty SN, Puniamoorthy J, Yeo D, Roslin T, Meier R. Convergence of dominance and neglect in flying insect diversity. Nat Ecol Evol. 2023 Jul;7(7):1012-1021. doi: 10.1038/s41559-023-02066-0. Epub 2023 May 18. PMID: 37202502; PMCID: PMC10333119.

Srivathsan A., Hartop E., Puniamoorthy J., Lee W.T., Kutty S.N., Kurina O., Meier R. 2019a. Rapid, large-scale species discovery in hyperdiverse taxa using 1D MinION sequencing. BMC Biol. 17:96.

Srivathsan A., Hartop E.A., Puniamoorthy J., Lee W.T., Kutty S.N., Kurina O., Meier R. 2019b. Rapid, large-scale species discovery in hyperdiverse taxa using 1D MinION sequencing. BMC Biol. 17:96.

Srivathsan A., Lee L., Katoh K., Hartop E., Narayanan Kutty S., Wong J., Yeo D., Meier R. 2021. MinION barcodes: biodiversity discovery and identification by everyone, for everyone. BioRxiv Prepr.

Svenning J.C., Borchsenius F., Bjorholm S., Balslev H. 2008. High tropical net diversification drives the New World latitudinal gradient in palm (Arecaceae) species richness. J. Biogeogr. 35:394–406.

Tikhonov G., Opedal Ø.H., Abrego N., Lehikoinen A., Jonge M.M.J., Oksanen J., Ovaskainen O. 2020. Joint species distribution modelling with the R -package HMSC. Methods Ecol. Evol. 11:442–447.

Tjur, T. 2009. Coefficients of determination in logistic regression models - a new proposal: the coefficient of discrimination. *American Statistician.* 63: 366– 372.

Townes H. 1972. A light-weight Malaise trap. Entomol. News. 83:239–247.

Truett G.E., Heeger P., Mynatt R.L., Truett A.A., Walker J.A., Warman M.L. 2000. Preparation of PCR-Quality Mouse Genomic DNA with Hot Sodium Hydroxide and Tris (HotSHOT). BioTechniques. 29:52–54.

Wang W.Y., Srivathsan A., Foo M., Yamane S., Meier R. 2018. Sorting specimen-rich invertebrate samples with cost-effective NGS barcodes: validating a reverse workflow for specimen processing. Mol Ecol Resour.

Warton D.I., Blanchet F.G., O'Hara R.B., Ovaskainen O., Taskinen S., Walker S.C., Hui F.K.C. 2015. So Many Variables: Joint Modeling in Community Ecology. Trends Ecol. Evol. 30:766–779.

Wellenreuther M., Larson K.W., Svensson E.I. 2012. Climatic niche divergence or conservatism? Environmental niches and range limits in ecologically similar damselflies. Ecology. 93:1353–1366.

Wiens J.J. 2007. Global patterns of diversification and species richness in amphibians. Am. Nat. 170:S86–S106.

Wolda, H. (1988). Insect seasonality: why? Annual Review of Ecology and Systematics, 19(1), 1-18.

Yeo D., Srivathsan A., Meier R. 2020. Longer is Not Always Better: Optimizing Barcode Length for Large-Scale Species Discovery and Identification. Syst. Biol. 0:1–16.

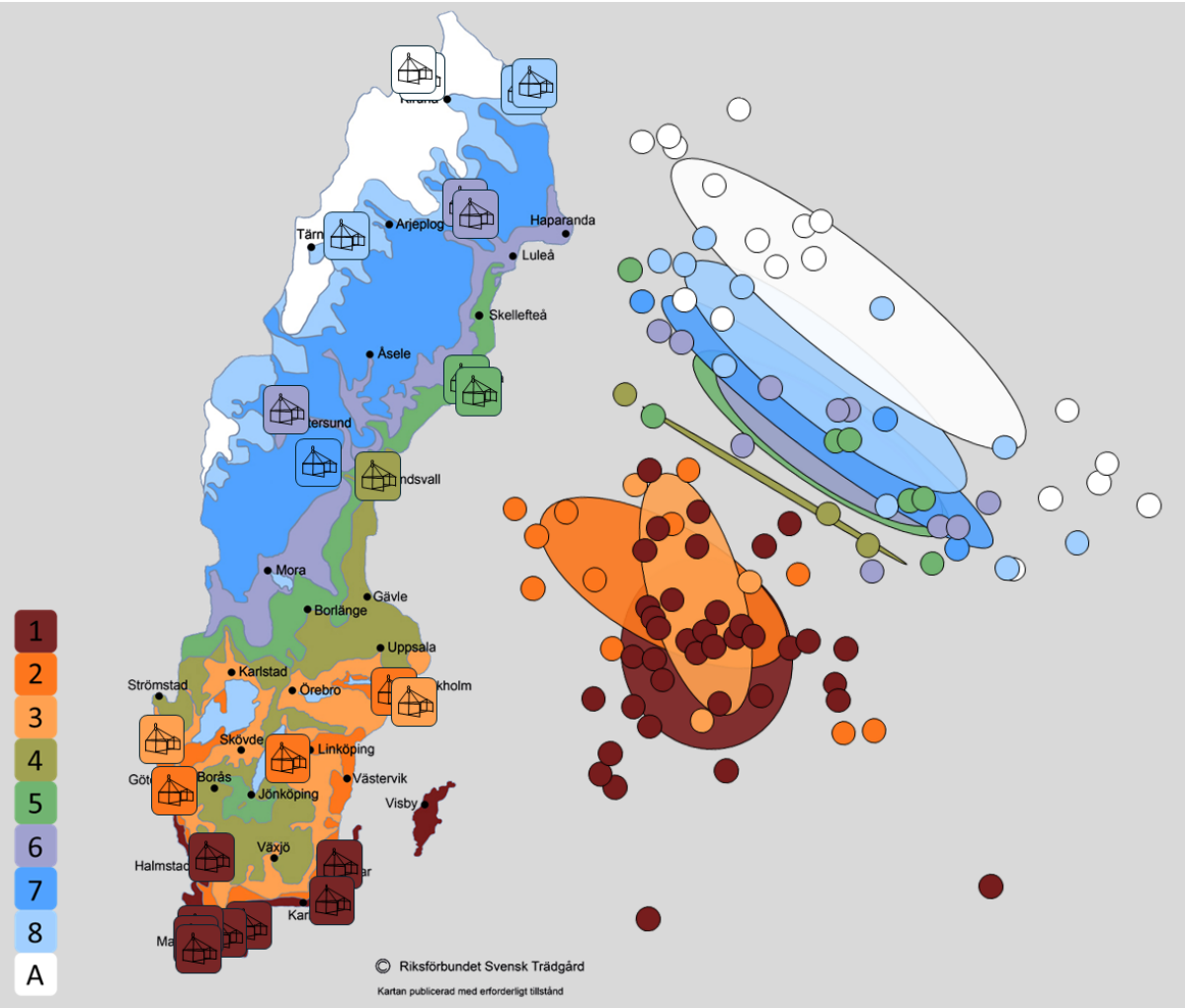

Figure 1. The location of 37 study sites colour-coded according to the plant hardiness zones (odlingszoner, 1-8 and Alpine) of the Swedish Horticultural Society (Riksförbundet Svensk Trädgård) (map used with permission) next to an NMDS plot of study samples colour-coded with the same zones.

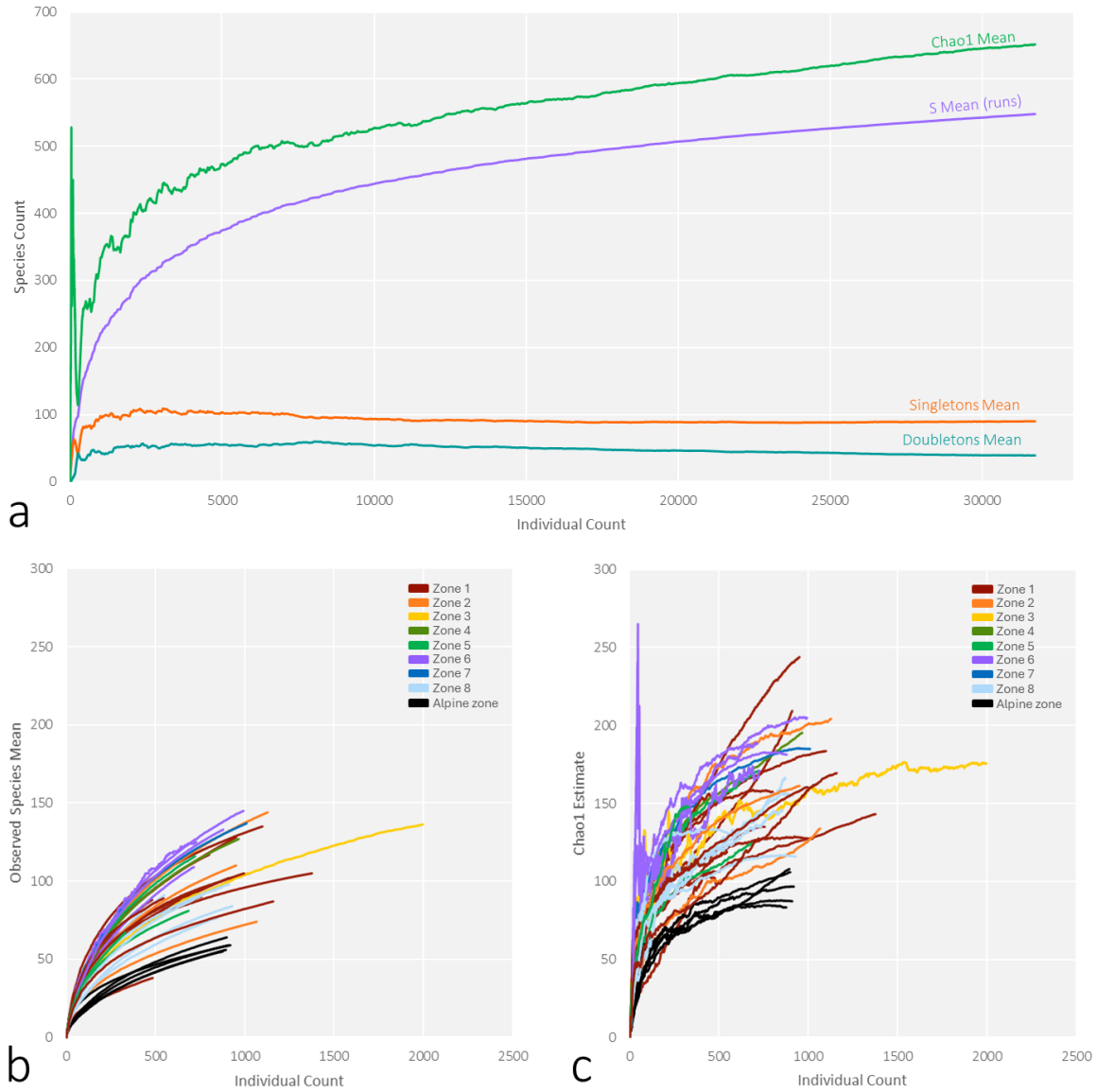

Figure 2. (a) Species accumulation and Chao1 estimate curves for the scuttle fly dataset across Sweden. Notably, current sampling is far from exhaustive, and numbers of singletons and doubletons in our dataset are high regardless of sample size, (b) Species accumulation curves by sampling sites, colour-coded by zone, and (c) Chao1 estimate curves by sampling sites, colour-coded by zone. For a map of the zones using the same colour codes, see Fig. 1.

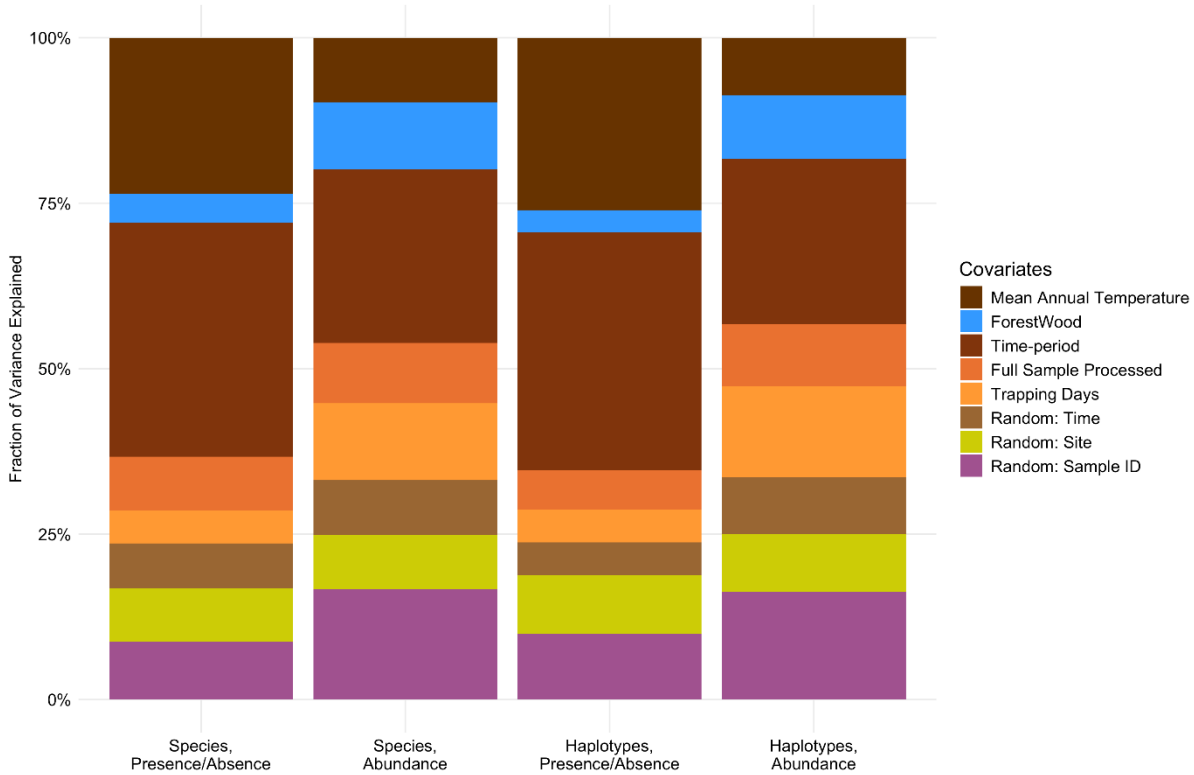

Figure 3. Summary of explained variance in the occurrences and abundances of scuttle fly haplotypes and species across samples (bar height corresponds to meanTjur's R^2 across the modelled taxa in the occurrence models and to mean R^2 in the abundance models). All four models show strong climatic structure (in shades of brown) on scuttle fly communities, as captured by fixed variables describing sampling season and mean annual temperature and a random effect based on median sampling date. The spatial fraction of explained variance (lime green) reflects community structural differences among sites that were not captured by the fixed effect covariates. Differences in sampling effort, i.e. whether or not trapped flies were all sequenced or not and the number of field trapping days per sample, also affect the predictability of community structure (shades of orange). The abundance of species and haplotypes, and to a lesser extent their occurrence, was also strongly structured by habitat type as described by forest and woodland cover (blue). Finally, we included a categorical random effect representing sample identity (purple).

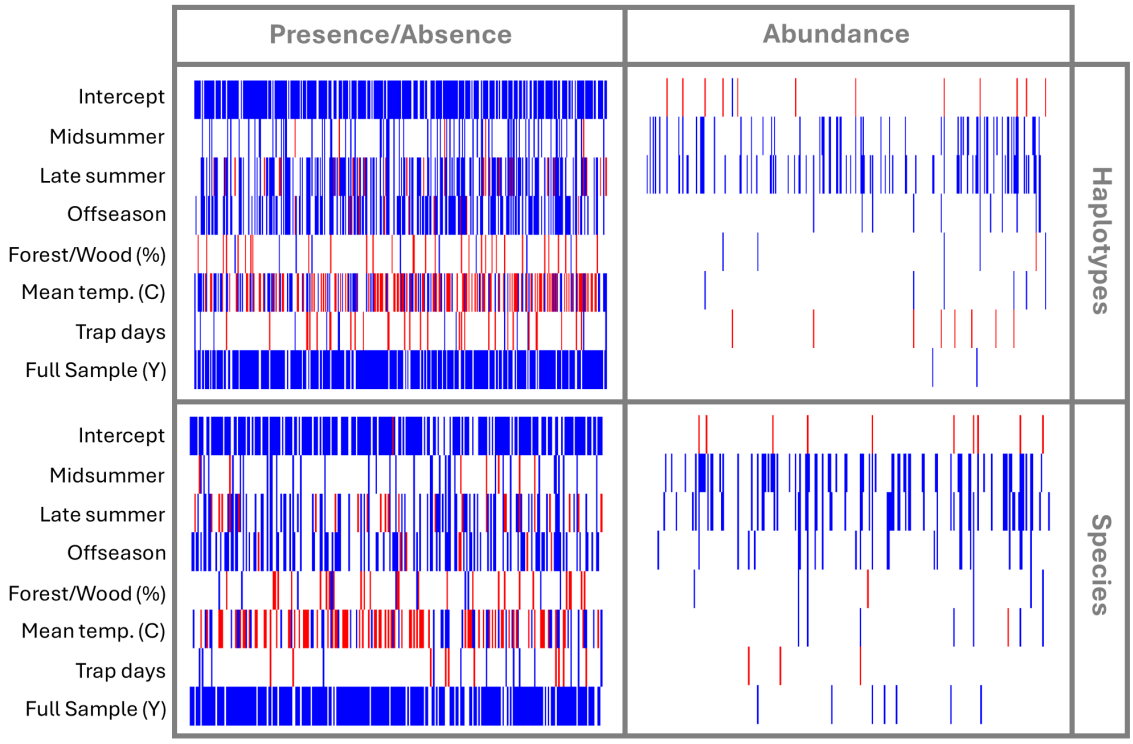

Figure 4. Predicted occurrence and abundance responses of 193 scuttle fly haplotypes and 162 mOTUs ("Species") to HMSC model covariates. Positive (red) and negative (blue) estimated responses with a posterior probability of 0.95 are illustrated. The three rows below the intercept illustrate the estimated effects of three levels of a categorical variable representing sampling time-period (midsummer, late summer, offseason) relative to the baseline (late spring). The subsequent three rows represent responses to % forest or woodland cover and mean annual temperature ("bio1") at sampling sites. The final two rows represent the effects of two sampling-related differences among samples: the number of trapping days and whether all specimens in the trap sample were sequenced, and hence available for HMSC analysis, or not.

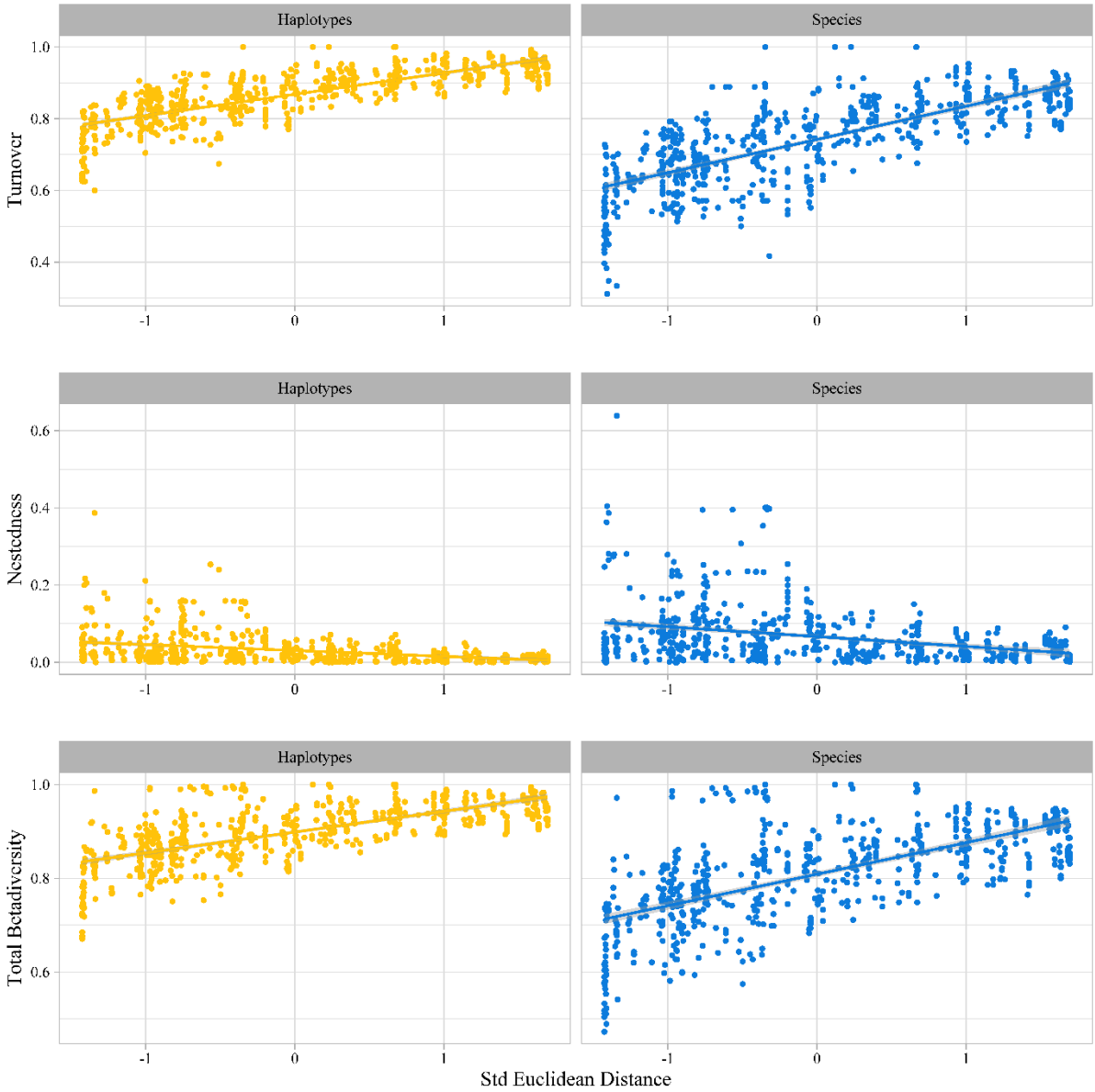

Figure 5. Pairwise community dissimilarity among scuttle fly species and haplotype samples, Following Baselga & Orme (2012), we partitioned overall dissimilarity into its turnover and nestedness components and illustrate these as well as overall dissimilarity as a function of distance (standardized Euclidean distance calculated from sample site coordinates).