Neglected dipterans in stream studies

Bernadett Boóz,[1]* Arnold Móra,[1] Márk Ficsór,[1,2] Petr Paři[1,3] Raúl Acosta,[4,5] Bea Bartalovics,[1] Thibault Datry,[6] José Maria Fernández-Calero,[4,7] Maxence Forcellini,[6] Marko Miliša,[8] Heikki Mykrä,[9] Bálint Pernecker,[1] Vendula Polášková,[3] Luka Polović,[3,8] Henna Snåre,[9] Zoltán Csabai[1,10,11]

[1]Department of Hydrobiology, Faculty of Sciences, University of Pécs, Hungary; [2]Department of Public Health, Laboratory for Environmental Protection, Government Office of Borsod-Abaúj-Zemplén County, Miskolc, Hungary; [3]Department of Botany and Zoology, Faculty of Science, Masaryk University, Brno, Czechia; [4]Department of Evolutionary Biology, Ecology and Environmental Sciences, Section of Ecology, FEHM-Lab, Faculty of Biology, University of Barcelona, Spain; [5]Institute of Environmental Assessment and Water Research (IDAEA-CSIC), Barcelona, Spain; [6]National Research Institute for Agriculture, Food and Environment (INRAE), UR RiverLy, Centre Lyon-Grenoble Auvergne-Rhône-Alpes, Villeurbanne, France; [7]Biodiversity Research Institute (IRBio), University of Barcelona, Spain; [8]Department of Biology, Faculty of Science, University of Zagreb, Croatia; [9]Finnish Environment Institute (SYKE), Freshwater Centre, Oulu, Finland; [10]HUN-REN Balaton Limnological Research Institute, Tihany, Hungary; [11]HUN-REN Centre for Ecological Research, Institute of Aquatic Ecology, Debrecen, Hungary

*Corresponding author: bb950828@gmail.com

Key words: aquatic Diptera; macroinvertebrates; literature meta-analysis; European ecoregions; quantitative sampling; family level identification.

Contributions: all the authors made a substantive intellectual contribution, read and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Conflict of interest: the authors declare that they have no competing interests, and all authors confirm accuracy.

Funding: field samplings were fully supported by the DRYvER project (H2020 grant agreement No 869226). ZC was also supported by Hungarian Research Fund (OTKA FK-135136). PP was supported by the Czech Science Foundation (GA23-05268S).

Data availability: summary data used for the literature meta-analysis are available as an electronic supplementary material attached to this paper. Case study data that support the findings are available from the authors, but temporary restrictions apply to the availability of these data based on the data policy of the DRYvER project, so are not publicly available during an embargo period until they will be published as a part of a bigger dataset. These data are, however, available from the authors upon reasonable request and with permission from the DRYvER coordinator.

Citation: Boóz B, Móra A, Ficsór M, et al. Neglected dipterans in stream studies. *J Limnol 2024;83:2191*.

Edited by: Diego Fontaneto, *National Research Council, Water Research Institute (CNR-IRSA), Verbania Pallanza, Italy*.

Received: 8 April 2024.
Accepted: 15 July 2024

J. Limnol., 2024; 83:2191
DOI: 10.4081/jlimnol.2024.2191

ABSTRACT

True flies comprise approximately one-tenth of all animal species on Earth, yet despite their prevalence and ecological significance in freshwater ecosystems, members of the insect order Diptera are frequently neglected in stream studies. This absence or inconsistency regarding Diptera in literature and taxonomic lists may leave readers with a sense of discrepancy. To illustrate this underrepresentation in quantitative ecological investigations, we conducted a targeted literature-based meta-analysis, assessing the average level of Diptera identification and the reported number of families. These findings were compared to data from 639 quantitative samples collected across six European ecoregions (Mediterranean, Alpine, Continental, Balkanic, Pannonian, Boreal) during six, bimonthly repeated sampling campaigns in 2021 and 2022. Our analysis revealed that, compared to other macroinvertebrate groups, Diptera were typically identified at a less detailed level, often only to the family level, thereby failing to fully represent Diptera diversity, especially regarding rare, less abundant families. In our review of literature studies, we identified references to a total of 40 families. Notably, Chironomidae, Ceratopogonidae, and Simuliidae were consistently represented across the majority of studies, whereas nearly half of the families were exclusively mentioned in one or two studies. No significant differences were found in the number of families across continents or various habitat types. In our case studies the number of families was significantly higher than in European stream studies, suggesting that several rare families occasionally completely neglected during sampling, sample sorting or identification. We explored potential connections among Diptera assemblages through correlation and coexistence analyses. Our results highlighted the significant influence of the more frequent Chironomidae, Ceratopogonidae, and Simuliidae on the presence or absence of other families. While correlations between Diptera families were identified, attempts to develop a predictive model for the diversity and occurrence of minor families based on the abundance of major ones proved inconclusive. For future quantitative studies on macroinvertebrate communities, it is essential to recognize, identify and incorporate less abundant Diptera families, even on family level, or in higher taxonomic resolution, if possible, to enhance understanding and prevent the loss of information concerning this compositionally and functionally uniquely diverse insect group, which represent a significant part of the entire community, and gain a better understanding on their interactions with other aquatic groups.

INTRODUCTION

With more than 159,000 species described worldwide the order of true flies (Diptera) comprise no less than one-tenth of all described animal species on Earth (Courtney *et al*., 2017), being by far the most diverse insect group along with Coleoptera, Hymenoptera and Lepidoptera (Zhang, 2011). In Europe alone, nearly 19,300 species of 126 families have been recorded (de Jong *et al*., 2014). Having successfully colonized all continents, including Antarctica (Usher and Edwards, 1984), true flies are presumably the most diverse insect order from an ecological point of view (Kitching *et al*., 2005). In aquatic ecosystems, where they inhabit all habitat types – including seas, oceans, shoreline saline pools, all kinds of stagnant waters (*e.g.*, lakes, ponds, and marshes), seepages and groundwater zones, plant-held waters (phytotelmata), cold and hot springs and the whole river continuum as well (Courtney and Cranston, 2015) – more than half of the recorded insect species belong to Diptera (Sundermann *et al*., 2007).

Besides its widespread distribution and exceptional diversity, in many cases Diptera is the most abundant taxon in freshwater ecosystems (Sundermann *et al*., 2007), thus playing a crucial role in ecosystem functioning. In food webs, true flies take part in decomposition and nutrient release by consuming large quantities of detritus, serve as essential food source for other freshwater organisms (Smith, 1989; Hövemeyer, 2000), and provide valuable functions as scavengers, predators, parasitoids, herbivores, and pollinators (Courtney and Cranston, 2015). Dipteran species show a wide range of tolerance: some of them live exclusively in pristine habitats, while others can tolerate various forms of environmental stress or perturbation, and thus survive in extremely degraded habitats of heavily polluted waterbodies (Lenat, 1993; Barbour *et al*., 1999; Courtney *et al*., 2017). Freshwater Diptera larvae are, therefore, frequently used as bio-indicators for water quality and bioassessment studies (Paine *et al*., 1956; Başören and Kazancı, 2020, and see also references in Sundermann *et al*., 2007). Families Chironomidae (Saether, 1979; Rosenberg, 1992; Timmermans *et al*., 1992) and Simuliidae (Feld *et al*., 2002; Lautenschläger and Kiel, 2005; Illéšová *et al*., 2008; Cuadrado *et al*., 2019) are traditionally the most widely used groups applied in classifying the extent of impacts on water bodies. The remarkable potential – provided by their high diversity and ecological variability – that could be used in ecological evaluation, is not adequately exploited, even in the case of these better-known families. Additionally, the family Ceratopogonidae is barely known compared to Chironomidae and Simuliidae, even though it is often diverse and abundant in freshwaters, including streams (Nakano and Nakamura, 2008; Thakur *et al*., 2022). Based on their dominant roles in streams, these three families (Chironomidae, Simuliidae, and Ceratopogonidae) are hereinafter referred to as major families, while the other Diptera families are referred to as minor families.

The most obvious explanation for the neglect of Diptera is that the identification of their larvae is mostly challenging, even achieving family level can be problematic (Smith, 1989; Oosterbroek, 2006; Dobson, 2013), and reaching lower taxonomic levels (genus or species) requires significant preparatory processes and extensive taxonomic experience (Sundermann *et al*., 2007). Difficulties may also arise from the large amount of unknown and undescribed species, the exceeding number of unknown larval forms, poorly known habitat preference of most of the species, and the deficiency or lack of identification keys for many families (Sundermann *et al*., 2007; Dobson, 2013).

As a result of the difficulties in the identification of their immature stages, ecological studies often underestimate the importance of Diptera, or even neglect the group partially or completely (Sundermann *et al*., 2007). There are numerous studies focusing on Diptera, for example, in relation to taxa that are vectors of human and/or animal diseases (Gerhardt and Lawrence, 2019), or restricted to those being beneficial to humans (*e.g.*, by plant pollination or water purification) (McLean, 2000). Aspects of their ecology, however, were barely investigated, compared to their significance in freshwater ecosystems (Omelková *et al*., 2013; Campos, 2015; Ivković *et al*., 2015; Polášková *et al*., 2020), and most of the ecological studies related to Diptera larvae are focused only on a single highlighted taxon (Cortelezzi *et al*., 2011; McCreadie and Adler, 2012; Cazorla and Campos, 2020). In quantitative ecological studies based on the whole aquatic macroinvertebrate community the detail of identification of Diptera larvae is generally not as high as that of other insect groups (Sarremejane *et al*., 2017; Miliša *et al*., 2022).

We hypothesize that Diptera are underrepresented in quantitative ecological studies. To explore this underrepresentation, we first examine how well stream studies describe dipteran assemblages and compare these with a dataset composed of 639 quantitative samples collected across six river networks spread in different ecoregions of Europe. We then intend to model the presence and abundances of rare and often neglected Diptera families based on occurrence and representation of the dominant and frequently studied ones.

METHODS

Literature search and meta-analysis

We conducted comprehensive online searches for peer-reviewed papers in two search platforms: Web of Science and Google Scholar on November 4, 2022 (Tab. S1). We used the keywords “macroinvertebrate” AND “Diptera” AND “stream” for the first search in both platforms (in Web of Science with and without using quotation marks, searching in the content of ‘all fields’). For Web of Science all relevant hits were checked while in Google Scholar only the first three hundred hits were considered. Under the same conditions, an additional search was performed on Google Scholar covering the first 500 hits. To increase the representation of Europe, we performed a second search in Google Scholar based on “macroinvertebrate” AND “stream” AND “Diptera” AND “Europe” keywords, where the first 300 hits were considered. Only papers published between 2010 and 2022 were kept and listed, since taxonomical knowledge, identification methods, and quality of the habitats (pollution, restoration, intermittency) changed a lot in the last decades, so we felt older records are not comparable with recent studies and projects. In the first step, out of all hits, unsuitable records were excluded by title screening and duplicate records were removed, which led to 214 individual documents, supplemented by 104 documents from the additional search. They were further selected considering four criteria: i) written in English, ii) peer-reviewed research articles, iii) based on quantitative field studies

of aquatic macroinvertebrates, iv) must contain information about number of taxa and/or the number of individuals of all macroinvertebrates, and number of taxa and/or the number of individuals of Diptera, or the relative abundance of Diptera within the whole community. As a result, a total of 61 papers were finally included in the meta-analysis (Tab. S2). In addition, the level of identification (family, genera, species), if available, of all macroinvertebrates and of Diptera, was also extracted from the articles.

Field study area, sampling, and sample processing

To present the composition and quantitative relationships of stream-dwelling dipteran communities in European streams, we rely on data from the DRYvER project (Datry *et al*., 2021) in which all aquatic macroinvertebrate groups were sorted and identified with the same thoroughness by a small team of experienced specialists. The DRyVER project focuses on drying river networks, however, there are almost no watersheds in Europe where at least certain sections have not dried up for at least a short period of time, especially in the last 10 years. The studied area covers six river networks from six countries in different European ecoregions (Fig. 1): Mediterranean – Genal River, Spain (hereinafter SP); Alpine – Albarine River, France (FR); Continental – Velička River, Czechia (CZ); Balkanic – Butižnica River, Croatia (CR); Pannonian – Bükkösdi-víz River, Hungary (HU); and Boreal – Lepsämänjoki River, Finland (FI). Between 15 and 26 sites per study area added up to a total of 126 sampling sites. Sampling was carried out through a series of six, bi-monthly repeated sampling campaigns in 2021–2022. Due to the intermittence of some sites in certain seasons, a total of 639 samples were taken. Qualitative, close and harvest type multihabitat sampling was conducted at each sampling site from a selected reach of 50-150 meters, depending on a maximum mean wetted width of the riverbed. For a more detailed description of the case study areas and sampling procedure see Datry *et al*. (2021). After the sorting of all collected macroinvertebrates in the laboratory using stereomicroscopes, different groups were treated separately and specimens were identified to the lowest possible taxonomic level but at least to genera, while Diptera larvae were identified to family level according to the identification keys by Sundermann *et al*. (2007), Tachet *et al*. (2010), Dobson (2013), and Kriska (2013).

Data analysis

Literature-based meta-analysis: to visualize the processing level of the literature data (the distribution between families/genus/species) and to present the numbers of data within each family, we used a ternary plot created with the ggtern extension (Hamilton and Ferry, 2018) of the ggplot2 package (Wickham, 2016) in the R statistical environment (R

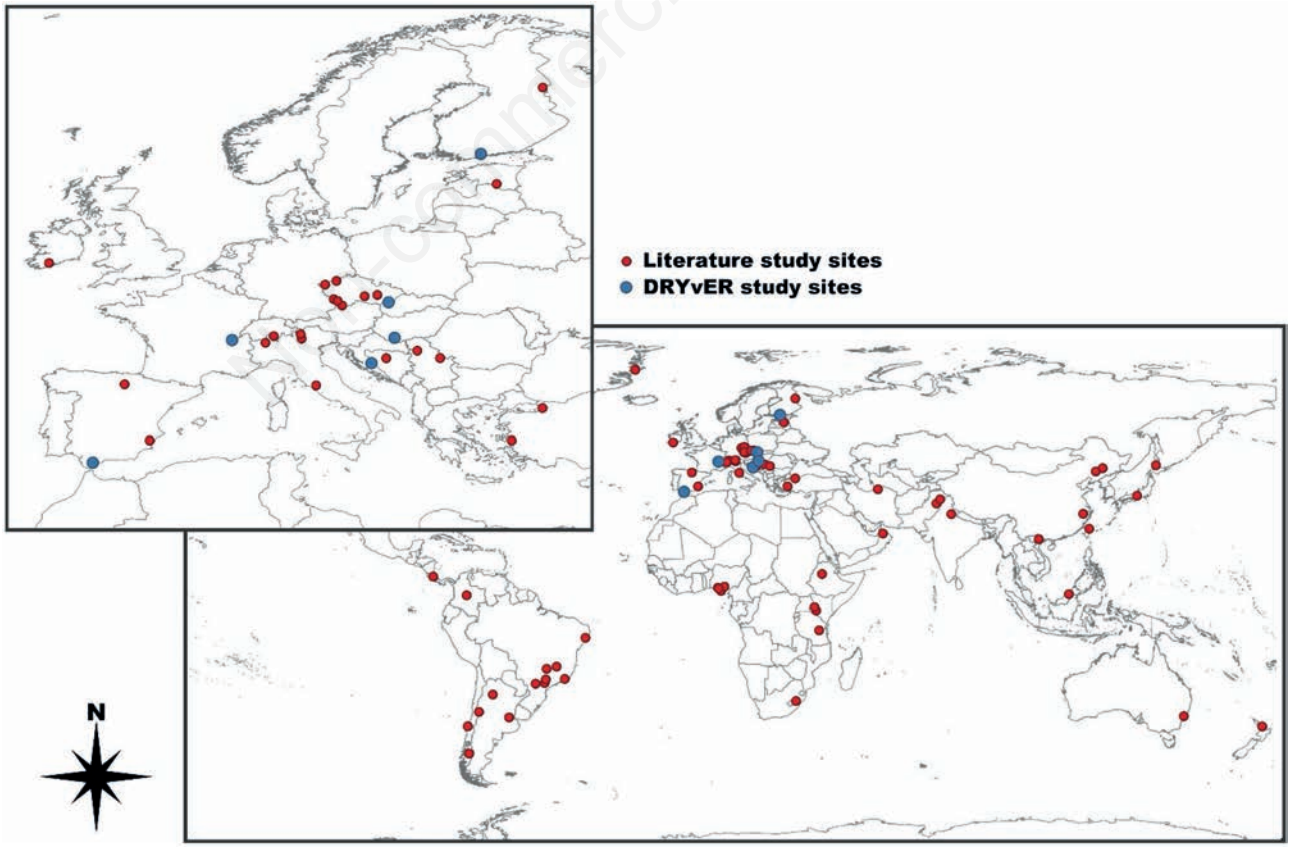

Fig. 1. Geographical location of the literature and field study sites included in the analyses over the world (bottom right) and zoomed to Europe (top left). North America was not represented in the literature studies except for one site in Greenland.

Core Team, 2023). The structure of data did not follow a normal distribution it could not be achieved even by transformation. Therefore, non-parametric methods were used for all analyses. Mann-Whitney U, Kruskal-Wallis and Dunn's tests were used to compare the number of Diptera families (N_{fam}) and the percentage distribution of Diptera within the whole aquatic macroinvertebrate community ($\%_{Dip}$) among geographical locations (continent, highlands/lowlands), watercourse types (small streams to larger rivers), and according to number of sampling locations and sampling periods. To ensure the most relevant comparison with our case study datasets, in addition to representing all the processed literature studies, we also made subsets of the European and non-European studies and a subset of exclusively small watercourse studies within Europe.

Case study analyses and modelling: the initial dataset contained Diptera abundance data (ind./m^2) for 632 of the 639 samples. The data were standardized based on the size of the sampled area (number of subunits), the number of sampling occasions, and then min-max normalization was performed to ensure comparability. To test whether the occurrence and abundance of selected major families can be used to predict the occurrence and diversity of minor families, we used three approaches.

First, using correlation and coexistence analyses, we examined which families' occurrences show a relationship with each other. Pairwise non-parametric correlation analyses (Spearman's r) were utilized to explore the relationships between the abundance data of Diptera families. Coexistence analyses were performed according to Schmera *et al*. (2007), Diptera families which occurred in less than six cases were excluded from the analysis. Altogether 10,000 random pseudo-assemblages were generated with constant species abundances in a sample. When an observed value of co-existence index falls in the upper marginal tail of the random distribution (2.5 percentile range) positive association can be observed between two taxa, but if its value falls in the lower marginal tail negative association is presumed. To map the relationship between the abundance of the major families, and the number of other Diptera ($N_{otherfam}$) families and the Shannon diversity of other families ($SH_{otherFam}$), we also used non-parametric pairwise correlation analyses (Spearman's r).

Second, pre-selecting predictor families, we tested their explanatory power for the occurrence of other families. From the initial dataset, data of families that were represented in less than 1/10 of the samples, thus considered rare, were excluded from the analysis (Bibionidae, Blephariceridae, Cecidomyiidae, Chaoboridae, Culicidae, Cylindrotomidae, Dolichopodidae, Ephydridae, Fanniidae, Rhagionidae, Syrphidae and Thaumaleidae). Families, whose abundance data could explain the presence, absence or abundance of others were selected consequently by using a series of 5-fold cross-validated Random Forest (RF) classification algorithms (Breiman, 2001). The RF algorithms – tuned for the best value of hyperparameter *mtry* based on the AUC metric – were trained to classify the presence and absence (decoded as binary outcome) of each family based on the abundance data of all the other families as predictor variables. Families, with an importance higher than 75% in each RF model were selected and retained for further analysis in case they had an above average number of cases being part of the selection. Retained families were used as explanatory variables, whose effects on the abundance of other families were studied using redundancy analysis (RDA, van den Wollenberg, 1977).

Finally, without prior ranking, we built models by including the families that explain the occurrence characteristics of each family significantly and to the greatest extent. The relationship between the abundances of families that were not considered rare were further analysed by generalized additive models (GAMs, Hastie and Tibshirani, 1986). In these, the relation between the abundance (log) of each non-rare families, as the response variable and the abundance (log) of all other non-rare families, as predictor variables were examined. Predictors with the significance of smooth terms' p-value <0.01 were retained for each family, and consequently used to predict their abundance. Predicted abundance was then plotted by the function of predictor family abundances. For the assessment of their effect on the abundance of the response group, partial dependence plots (PDPs) were created for each predictor families within a model, and descriptive statistics (*i.e.* means and ranges) of the marginal impact values on the Y-axes were compared.

Univariate tests were performed in JASP (ver. 0.16.4.0, JASP Team, 2022) and in PAST (ver. 4.11, Hammer *et al*., 2001) software environments. Coexistence analyses were run using a Microsoft Excel Macro, based on Schmera *et al*. (2007). RDA and GAM analyses and visualisation were performed in the R statistical environment (R Core Team, 2023) using the following packages: caret (Kuhn, 2008), DALEX (Biecek, 2018), ggplot2 (Wickham, 2016), ggpubr (Kassambara, 2023), ggrepel (Slowikowski, 2023), mgcv (Wood, 2011), Polychrome (Coombes *et al*., 2019), stringr (Wickham, 2022), vegan (Oksanen *et al*., 2020).

RESULTS

Diptera assemblages in literature-based studies

From the last 10 years a total of 61 publications fully met the selection criteria and search conditions, showing a heterogeneous picture in time and space. The studied watercourse sections ranged from mountainous headwater streams to small or medium-sized lowland rivers. The extent of the sampling area, the number of sampled sites and occasions also varied from a few sampling points on a single watercourse to a complex river network including more than 100 stream sections, or even to a large-scale study listing 271 sampling sites in several countries, scheduled only one time or with monthly, bimonthly, or seasonal sampling frequency. The included studies originated from all the inhabited continents, there was no remarkable difference in their number, apart from the fact that based on their area, North America, Africa, and Australia were underrepresented, while Europe was deliberately overrepresented due to the optimization of the database search (Africa 9, America 13, Asia 14, Australia and Oceania 2, Europe 23 studies; Fig. 1, Tab. S2.)

Breakdown of the order Diptera to family or lower level was done in 58 cases, but the exact taxonomic level to which the Diptera specimens were identified was clearly revealed in 55 studies (in three cases there are only vague references to this). Compared to other aquatic macroinvertebrates, where the proportion of family-only-level identification was low (~13% of all papers), the identification level of Diptera taxa was generally less detailed: in 14 of 55 cases (~25%) only family-level identification was done. In 13 studies, some Diptera families

(<30% of all found) were identified to genera at least, while in further 16 articles most of the families (>70%) were processed at a finer resolution, but in six of these cases the number of families was lower than eight. In the remaining 14 studies, genus level identification was made for approximately half of the families (30-70%).

In the 61 studies evaluated in detail we found mentions of 40 Diptera families altogether, where the most frequently included were Chironomidae (55 studies), Simuliidae (46) and Ceratopogonidae (45) (Fig. 2), while almost half of the families occurred in one (13 families) or two (4) studies (Tab. S2). The average number of Diptera families in the studies was ~10±4.33 (mean±SD), ranging from 2 to 23. A quarter of the families (9, ~22%) always remained at family level and seven (~17%) were always processed at genus level, although most of the families (16) in these two groups were found in only one or two studies. For the 22 families that appear more frequently in studies (≥4), species-level identification occurs in 16 cases, but in 14 of these cases it remains below 25% (Fig. 2). For each family, the mean proportion of studies using species-level identification is 7.1%±17.1, for the genus level it is 41.9%±34.8, while at the family level it is 50.8%±35.9.

The range of data (including raw data, summary data, metadata) that could be extracted from the involved studies and from the available electronic supplementary materials varied widely. While the number of macroinvertebrate taxa found was revealed in all cases, specific data on abundances could only be found in 39 cases. It was possible to calculate the proportion of the Diptera assemblage to the entire macroinvertebrate community in 26 cases, but specific abundance records were only found in 18 cases. Based on this, to be able to include relevant and at the same time representative (*i.e.*, can be retrieved from sufficient number of studies) characteristics in the comparisons, we excluded the raw abundance data and used i) the number of Diptera families detected in the given study (N_{fam}), and ii) the proportion of Diptera order to the whole macroinvertebrate communities ($\%_{Dip}$). Regarding these features, we found no differences between habitat types (from small streams to medium-sized rivers, Kruskal-Wallis, N_{fam}: χ^2=5.92, p=0.11; $\%_{Dip}$: χ^2=0.19, p=0.98), the number of examined sections/sites (N_{fam}: χ^2=4.08, p=0.13; $\%_{Dip}$: χ^2=0.23, p=0.89), geographical location (among continents, treating Africa, Australia, and South-America pooled due to the very low numbers of available data from the first two, N_{fam}: χ^2=5.19, p=0.07 $\%_{Dip}$: χ^2=0.92, p=0.63), or altitude (highland or also including lowland sections, Mann-Whitney-U, N_{fam}: U=117, p=0.62; $\%_{Dip}$: U=34, p=0.43). Therefore, we did not use

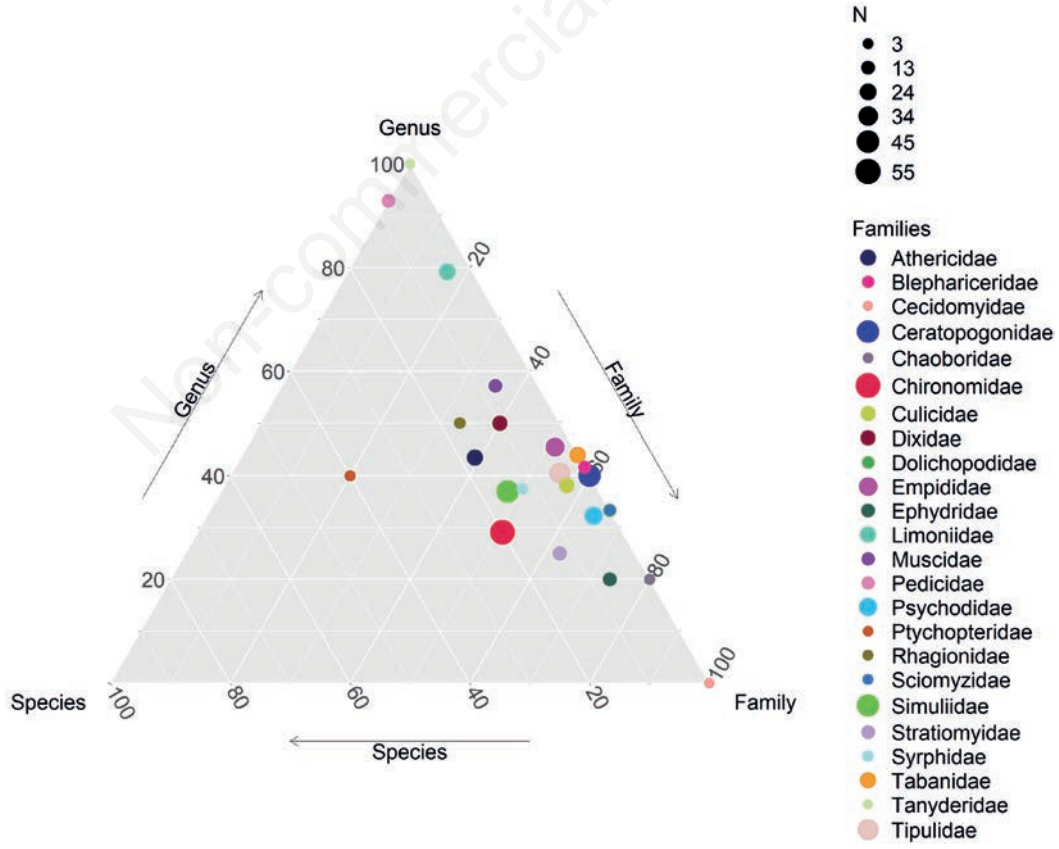

Fig. 2. Ternary plot showing the variance in share of identification level (family, genus, species) of each Diptera family with more than three mentions in papers emphasizes the very low proportion of species level processing and various but generally half-share between family and genus level among 46 studies. N, number of studies in which the given family was included. Families Ptychopteridae and Sciomyzidae are overlapping, shown under the sign of the latter.

such breakdowns in further analyses. In seasonal sampling frequency we did not detect differences in number of families but in $\%_{Dip}$ (1, 2-3, 4+ sampling campaigns, N_{fam}: χ^2=0.19, p=0.91; $\%_{Dip}$: χ^2=7.72, p=0.02). Since it is clearly reflecting the phenology of invertebrates, affecting only the literature studies, we also did not include it in the detailed analyses.

Diptera assemblages in the case studies

The N_{fam} was not significantly different among the case study catchments (Fig. 3B; χ^2=8.48, p=0.12). The average N_{fam} was ~16.83±1.94, altogether 26 families were found in the six case studies. The $\%_{Dip}$ showed much higher variance between (χ^2=20.42, p=0.001) and within the case study areas (Fig. 3D), the average was 31.67±22.77. Families with highest abundance were Chironomidae, Simuliidae, and Ceratopogonidae. In addition, families that occurred in all catchments, but in lower abundances, were Pediciidae, Psychodidae, Limoniidae, Muscidae, Rhagionidae, Tabanidae, Tipulidae, and Empididae. There were some families that only appeared in one catchment: Bibionidae (HU), Thaumaleidae (SP), Cylindrotomidae, and

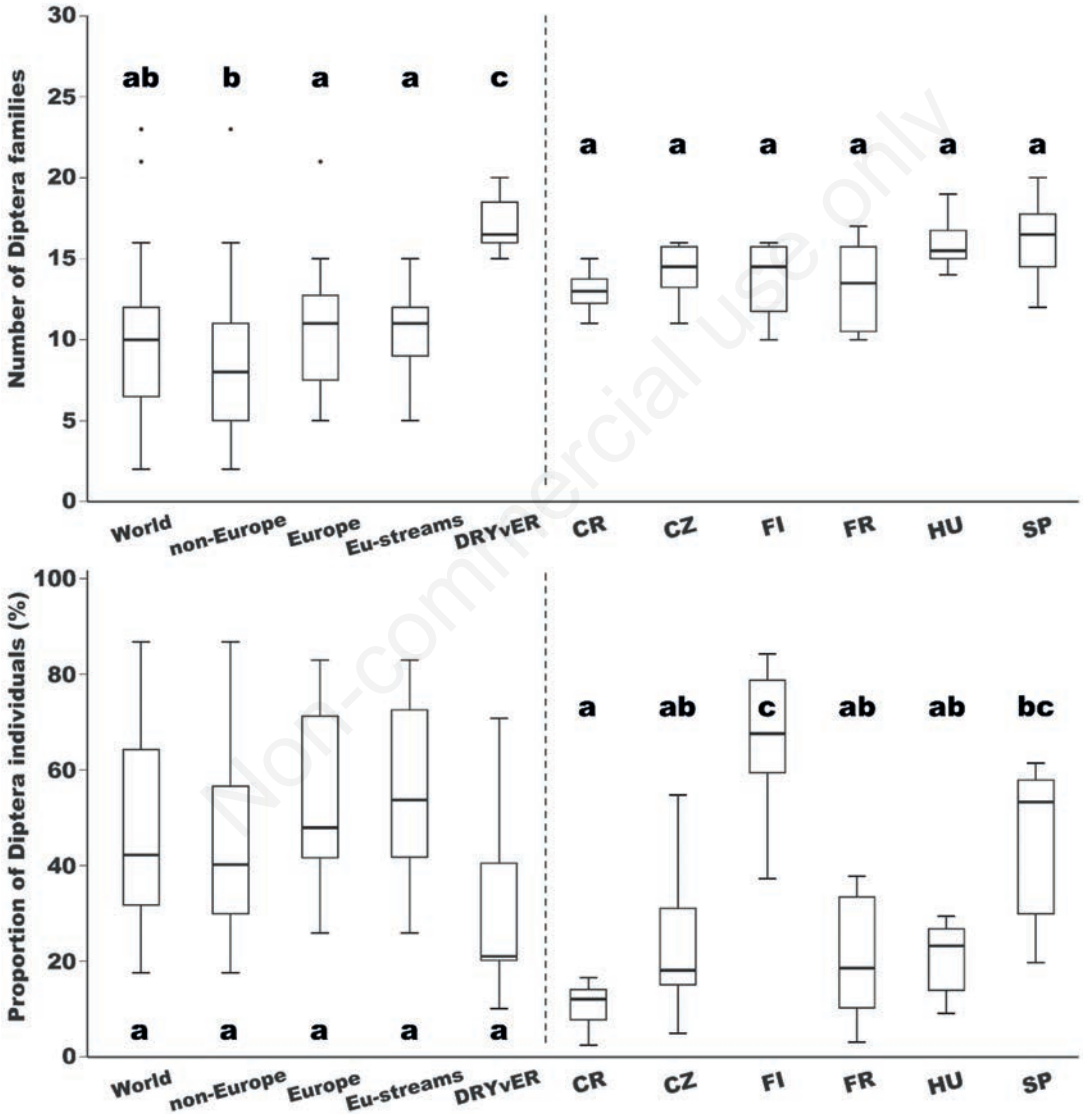

Fig. 3. Characteristics of Diptera assemblages are comparable among all type of studies, but literature studies are barely report lesser-known and rare families, as illustrated by the comparison of the number of Diptera families **(A,B)**, and the share of Diptera in the whole aquatic macroinvertebrate community **(C,D)** in different literature and field case study groups. World, studies from all involved papers [n=46 **(A)**, 22 **(C)**]; non-Europe, literature studies from all countries outside of Europe [n=26 **(A)**, 13 **(C)**]; Europe, literature studies from Europe [n=20 **(A)**, 9 **(C)**]; Eu-streams: literature studies including only low-order streams from Europe [n=15 **(A)**, 6 **(C)**]; DRYvER, field case studies from the DRYvER project [n=6 **(A,C)**]; CR, Croatia, Butižnica catchment; CZ, Czechia, Velička catchment; FI, Finland: Lepsämänjoki catchment; FR, France, Albarine catchment; HU, Hungary, Bükkösdi-víz catchment; SP, Spain, Genal catchment (n=6 for all catchments, pooled data per sampling campaign). Lowercase letters indicating grouping based on Kruskal-Wallis tests followed by pairwise Dunn-tests comparisons, where each panel **(A-D)** is to be interpreted separately.

Chaoboridae (FI) and some that were found in two catchment Syrphidae (CZ, HU) and Fanniidae (FR, HU). Families with few individuals but occurring in several places were Ephydridae, Cecidomyiidae, and Dolichopodidae.

Comparing literature to case study datasets

We found significant differences among the nested category groups of literature studies in N_{fam} (χ^2=21.46, p=0.001) but not in $\%_{Dip}$ (χ^2=5.95, p=0.28). The slightly higher values characterising European studies regarding N_{fam} (Fig. 3A). However, N_{fam} was significantly higher (χ^2=18.11, p=0.001) in DRYvER case studies compared to any literature study groups (Fig. 3A), even in the case of comparing the same habitat types (low order streams from Europe). These differences are evident even if we compare not only the pooled case study data but each case study individually to the literature study groups (Fig. 3 A,B). Although in the DRYvER case studies the average $\%_{Dip}$ was slightly lower than in the literature studies, but, regarding the pooled data, this deviation was not significant (χ^2=5.94, p=0.2) (Fig. 3C). Comparing the $\%_{Dip}$ of each case study individually to themselves and to the literature studies, there are case studies with significantly lower mean values (Fig. 3 C,D).

Modelling relationships within Diptera assemblages

More than 20% of the potential relationships based on correlations (Fig. 4A) in abundances between Diptera families proved to be significant, but the correlations were strong ($r>|0.7|$) only in two cases: between Chironomidae and Ceratopogonidae (positive), and between Ceratopogonidae and Athericidae (negative). Much more positive relations (57) were detected than negative (16). Coexistence analyses (Fig. 4B) revealed 123 significant associations between families based on their occurrence characteristics, 66 of them were negative and 57 were positive. The proportion between positive and negative association with other families varied widely. For example, the family Athericidae were almost exclusively negatively associated with most of the families (2 positive *versus* 15 negative), similarly to Simuliidae (4:11) or Dolichopodidae (2:9). Whereas the most abundant Chironomidae were more positively (8:2) associated with other families, and Ceratopogonidae, Psychodidae, Tabanidae and Empididae favoured or avoided almost equal numbers of families (9:7, 8:9, 9:8, 7:9).

The abundance of Chironomidae and Ceratopogonidae significantly positively correlated with the number of other families

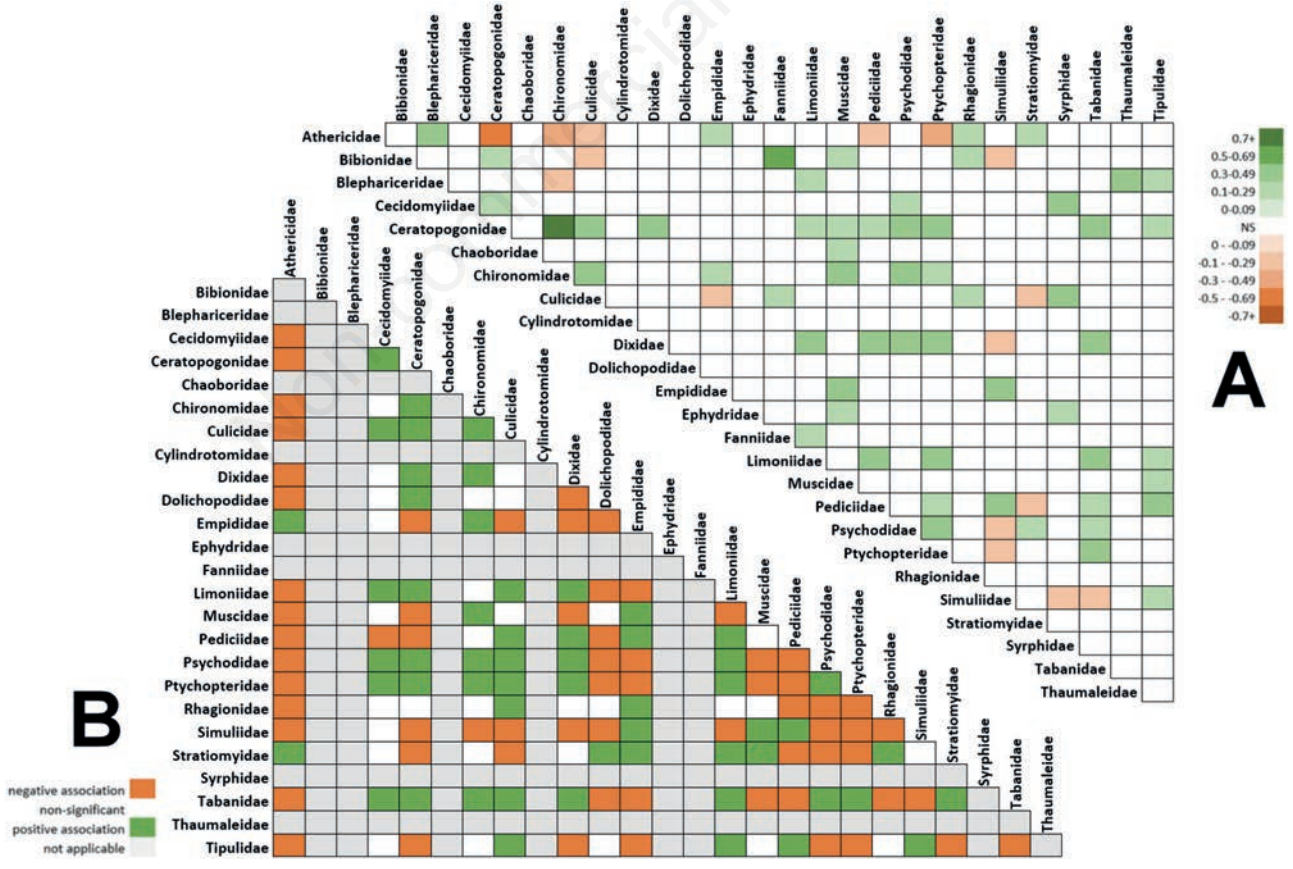

Fig. 4. The heatmaps of correlation **(A)** and coexistence **(B)** estimates based on abundance data clearly show the families occurring together or avoiding each other. Orange colour coding negative, green colour coding positive relationships. Only significant ($p<0.05$) relationships are included with colours. In the upper half of the figure, the colour shades refer to the value of the correlation coefficient (r) according to the legend on the right side.

(r_s=0.245, p=0.006; r_s=0.323, p=0.001, respectively), and negatively, although not always significantly, with their diversity expressed as Shannon-index (r_s=-0.153, p=0.09; r_s=-0.323, p=0.001, respectively). The abundance of Simuliidae not significantly negatively correlated with the number of other families (r_s=-0.008, p=0.96), and significantly negatively (r_s=-0.223, p=0.01) with their diversity.

From 14 non-rare families Chironomidae, Ceratopogonidae, and Simuliidae were identified as having the highest influence on the presence and absence of other families. Although the RDA scatterplot based on these three families as predictors showed some relationships with them (Tabanidae, Psychodidae, and Ptychopteridae were positively, while Tipulidae, Pediciidae, and Athericidae were negatively correlated with Ceratopogonidae; Pediciidae, and Empididae were positively correlated with Simuliidae, and Athericidae was positioned far opposite to Chironomidae), the first two axes explained only a cumulative 2.93% of the total variance (Fig. S1).

The results of GAMs (Fig. 5) showed significant relationship between the abundance of several pairs of families. Al-

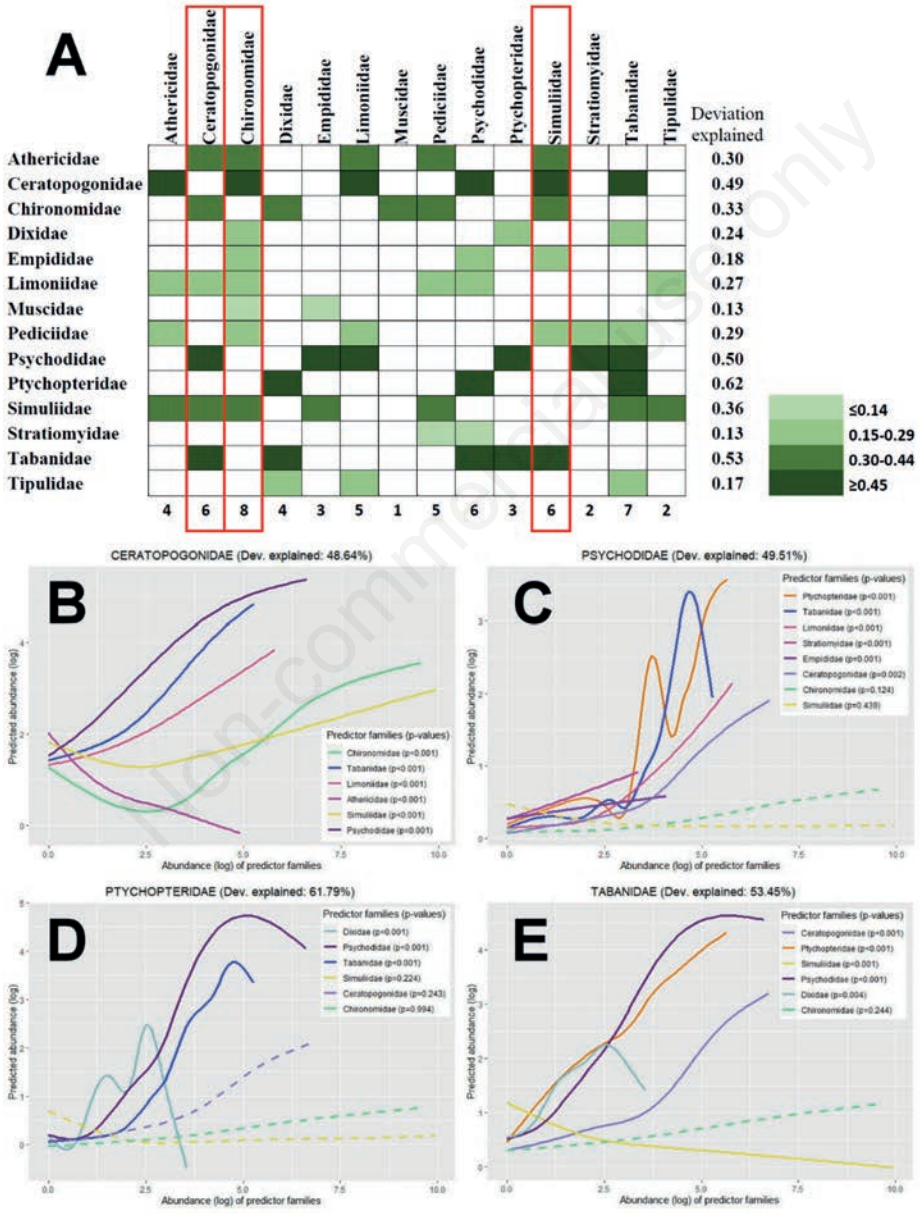

Fig. 5. Results of the Generalized Additive Models show high variation in the number of explanatory families and in the predictive power of models for each family **(A)**. There are no generalizable rules based on which abundances can be reliably predicted for all, but well-explaining models can be built for several families including various predictor families **(B-E)**. Continuous lines show significant predictors, dashed lines represent non-significant predictors, latter only showed in case of Chironomidae, Ceratopogonidae or Simuliidae as well-known and abundant families, which are marked by red rectangles in **(A)**.

though the explained deviation in these models were relatively low in many cases (Fig. S2), they were still an order magnitude higher than in RDA. In GAM models, different predictor families appeared to achieve the best explanatory power for each family (Fig. 5A). The number of predictor families varied between 2 and 8, and the degree of explained variance varied between 13-62%. The largest explained variance could be seen in cases of Ptychopteridae, Tabanidae, Psychodidae, and Ceratopogonidae (Fig. 5 B-E, 49-62%), while the smallest was given by the best model of Stratiomyidae, Tipulidae, Muscidae, and Empididae (Fig. S2, 13-18%). No family was found that had explanatory power in all cases (Fig. 5A). Most of the selected families showed significant explanatory power in only half of the cases, similarly to the better-known, most abundant ones (Chironomidae, Simuliidae, and Ceratopogonidae, 8, 6 and 6 cases, respectively, Fig. 5A). In models with the three highest values of explained variance, the latter three appeared as a significant predictor in only 0, 1 and 2 cases, respectively (Fig. 5 C-E), and at the same time, the Chironomidae and Simuliidae appeared as significant predictors in Ceratopogonidae's best model (Fig. 5B).

The abundance of predictor families, despite their statistical significance, mostly had minor effects on the predicted abundance of response groups/ones. The "effect sizes", defined as the mean and range of partial dependence values, showed variability both between and within models, and were generally higher in models with high explanatory power (Fig. 6).

DISCUSSION

The literature meta-analysis and its comparison with the thorough case study datasets clearly showed that members of certain, presumably rarer, smaller, and lesser-known families are less likely to be found in the literature studies. The numbers of families in literature studies were lower than in the case-studies, confirming our assumption that Diptera diversity is generally underrepresented in stream studies. It is also clear that the identification of some better-known families - which are being represented in large numbers in the streams and are found in almost every study - was done at a lower (genus or species) level in several cases. Based on these facts, we tested whether it is possible

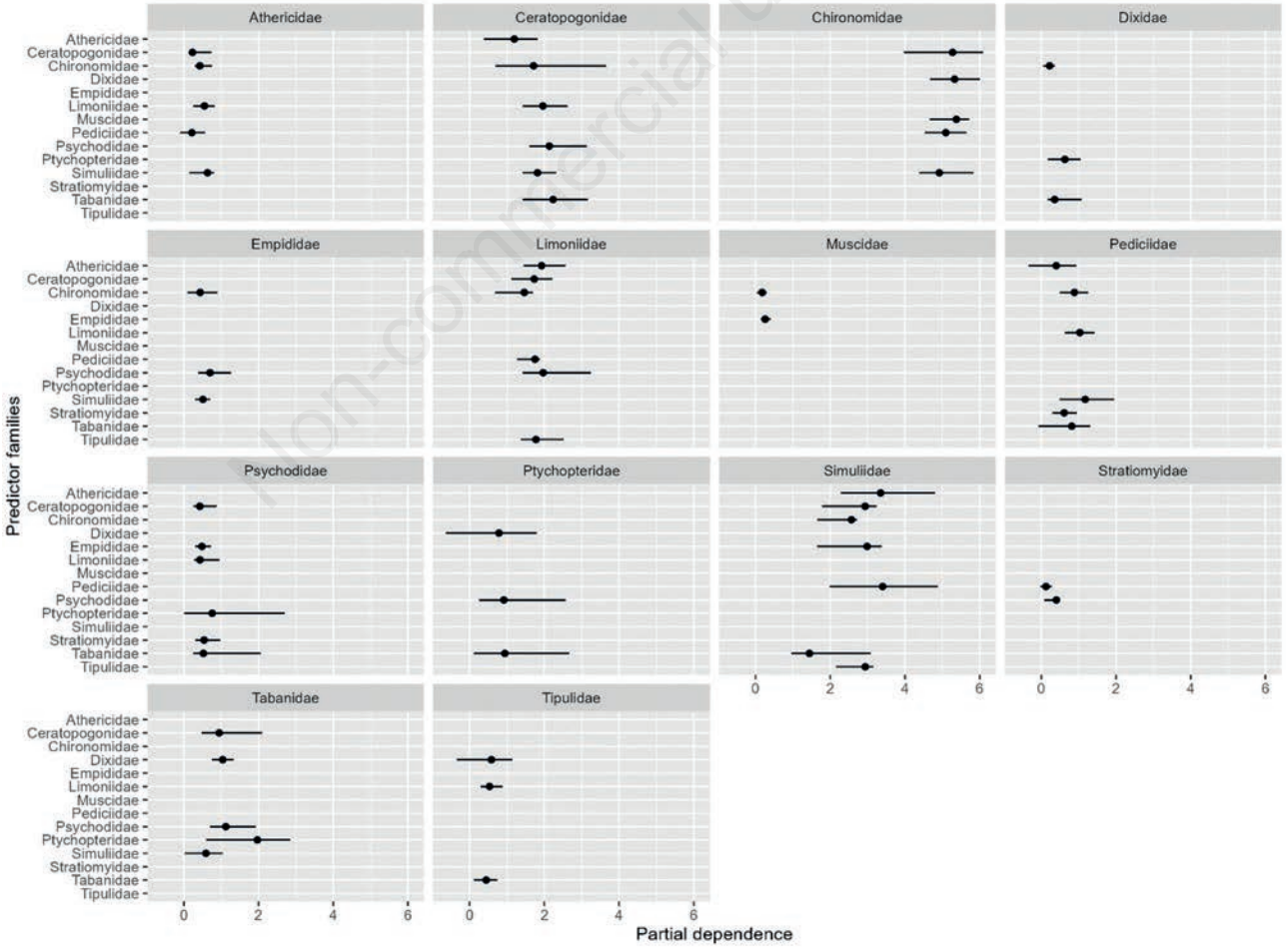

Fig. 6. Effect sizes of predictor families based on the partial dependence of the response families also indicated that predictor families in the final models vary widely, with their contribution to the predictions being significant but often weak.

to predict the characteristics of neglected families based on the occurrences and abundances of the frequently studied ones. Although we have identified many connections and co-occurrences, which can make certain families likely to be present in the presence of other families, and may be used for prediction of single families, our results also showed that it is not possible to model the expected diversity and the probability of the appearance of individual families based exclusively on common and well-known families. These families are not, or not the only ones, based on which we can make predictions with higher explanatory power.

Literature-based meta-analysis

Despite the availability of several macroinvertebrate-based ecological research and publications on the topic, a relatively low number of articles, 61 out of 214+104 met all our search criteria. The main reason for this is that a complete list of taxa is not available in many articles, and neither the final identification level of the taxa, nor the exact number of individuals belonging to a given taxa could be extracted. Our study therefore points to the importance of OPEN DATA, as well as to the crucial accessibility of basic and/or main meta-data, even though we ourselves are also not authorized to make our data freely accessible publicly until the end of an embargo period of approximately a year.

It is obvious that in most studies of the reviewed literature, identification was done at the family or genus level (Fig. 2). One reason for this could be the complicated and often unclear larval taxonomy of Diptera. In some cases, even the family-level identification is problematic. For example, the superfamily Tipuloidea (former family Tipulidae) now includes the individual families Tipulidae, Limoniidae, Pediciidae, and Cylindrotomidae that are morphologically very similar to each other (Oosterbroek and Theowald, 1991). In other cases, the high morphological heterogeneity within a family combined with remarkably similar appearance of larvae of different families due to convergent adaptation to similar environmental conditions makes the identification very hard (Smith and Ferrar, 2000). Especially if these families also include terrestrial forms that are not keyed out in identification guides for aquatic Diptera (*e.g.*, Ephydridae, Muscidae, Psychodidae, Scathophagidae, Syrphidae, Tipulidae). In addition, reaching lower taxonomic level (*i.e.*, genus or species) often requires significant preparatory processes in case of many taxa, which cause the Diptera identification more difficult (Sundermann *et al*., 2007). Another reason for using lower taxonomic resolution (*i.e.*, family level) could be that it is suggested to be sufficient in certain cases, *e.g.*, when using robust bioassessment methods, or in large scale monitoring (de Oliveira *et al*., 2020; Pires et al., 2021), despite the controversial assessment of its efficacy (see Pires *et al*., 2021 and references cited in). Excessive numbers of individuals to be processed, even from the families that are easier to identify, could be a relevant, although professionally less acceptable cause of lower-level identification. This exactly happened to be the case in our own case studies, where >400k individuals were processed on a morphological basis. Although species level information is crucial for basic and applied ecological studies (Heino, 2014; de Oliveira *et al*., 2020), the above-mentioned facts may lead to a disproportionately frequent use of family-level identification.

Meta-analysis and case studies

Based on the literature data, Chironomidae, Ceratopogonidae, and Simuliidae, as major families, are the most common and most abundant stream dwelling Diptera and the most information is available about them, while only much more limited ecological information is available about other, less frequent, minor families. In accordance with the literature, our results show the prevalence of Chironomidae, Simuliidae and Ceratopogonidae too, as they were the most frequent and abundant families occurring in all studied areas with the largest numbers of individuals.

The number of Diptera families (N_{fam}) and the proportion of Diptera individuals in the whole community ($\%_{Dip}$) were similar in literature studies. In contrast, the N_{fam} was significantly higher in our case studies than found in the literature (Tab. S3). As our stream types were in concordance with published studies, we only suspect that the details of sampling protocols could be responsible for the observed inconsistences in this parameter. First, during the sample processing in our case studies, a highly detailed sample sorting was performed under stereoscopes, allowing to find small individuals and rare taxa with higher probability. In contrast, quantitative samples are often pre-sorted in the field and sub-sampled in the lab, which can result in the more likely loss of small individuals and rare taxa (Friberg *et al*., 2006), which could lead to a great loss of information. Second, in most cases no up-to-date keys were used for Diptera, and taxonomic changes often were overlooked in the literature. For example, in older keys, used in many papers included in our analysis, the families Tipulidae, Limoniidae and Pediciidae were treated as one family, which might contribute to a lower number of families. However, the number of families was so much higher in our case studies than in the literature, that adding one or two families to the literature checklists hardly can have effects on our results. Third, even the family-level identification could be complicated due to the inadequate identification literature and comprehensive keys, caused by, beyond the unclear taxonomy, the flexible definition for aquatic Diptera (Dobson, 2013). Some Diptera families consist of taxa with merely aquatic larvae, while others contain aquatic, semi-aquatic and terrestrial representatives as well (Nilsson, 1997; Hövemeyer, 2000). In some cases, identification keys do not completely overlap in types. For example, Chaoboridae, Culicidae, Dixidae, Ptychopteridae, Simuliidae, and Thaumaleidae are the exclusively aquatic families according to Nilsson (1997), while Dobson (2013) lists the family Blephariceridae as exclusively aquatic too. Moreover, especially in the case of mainly terrestrial families that contain only a few semi-aquatic species (*e.g.*, Fanniidae, Scatopsidae, Cecidomyiidae, *etc.*), different keys include different families, such as in those used in our study (Sundermann *et al*., 2007; Tachet *et al*., 2010; Dobson, 2013; Kriska, 2013). Accordingly, families to be considered as aquatic might depend on the identification keys used by researchers, and that can not only contribute to differences in the taxa-lists, but might also lead to information loss (*e.g.*, when a questionable family occurs in a sample, but excluded from further analyses because it is not mentioned or classified as aquatic in the key used). However, some newly published identification keys include a recent revision of Diptera families which can help classify larvae as aquatic or semi-aquatic (Faasch, 2015; Fusari *et al*., 2018; Lencioni *et al*., 2023). In our case studies we took all families included in at least one of the used keys into consideration. The more accurate identification combined with detailed sorting might lead to the higher number of families. This result highlights

that sorting process without sub-sampling and using multiple determination keys are essential to avoid information loss despite being time and energy consuming.

Although the N_{fam} was higher, the $\%_{Dip}$ showed lower values in the case studies than in the literature studies. For the first assumption, the higher number of families would lead to higher number of individuals, but since rare taxa were also included, the increased number of families did not significantly influence the number of individuals. Since samplings in the literature studies were performed in different seasons and did not intend to cover the entire vegetation period, phenology can also be in the background of lower $\%_{Dip}$ values. In case of summer samplings (Gao *et al*., 2014), lower numbers of early swarming species from other insect orders, like Ephemeroptera, Plecoptera, and Trichoptera would be collected, causing higher relative abundance of Diptera. Another reason for lower $\%_{Dip}$ values could be that different invertebrate groups were also involved in each study found in literature.

The generally fewer macroinvertebrate groups included in other studies, especially the lack of occasionally abundant groups, like Oligochaeta (Docile *et al*., 2016), may result in higher $\%_{Dip}$ values than in our own case studies. However, it is difficult to determine whether an invertebrate group that was not included in the taxa list did really not occur in the sample or was ignored in the study for any other reason showing false information on the relative abundances. Crustacea, for example, was not included at all in many studies (Narangarvuu *et al*., 2014; Docile *et al*., 2016; Alemneh *et al*., 2017; Bartošová *et al*., 2019; Aazami *et al*., 2020; Debiasi *et al*., 2022). Different studies defined different invertebrate groups as 'aquatic macroinvertebrates': in some cases, for example Cnidaria (Ono *et al*., 2020), Acarina (Marchamalo *et al*., 2018), Lepidoptera (Korte, 2010; Docile *et al*., 2016) or Collembola (Souto *et al*., 2011; Docherty *et al*., 2018; Marrochi *et al*., 2021) were included, while in other cases these were not. Although these groups are usually not represented in high numbers, collected specimens may affect the total abundance and the $\%_{Dip}$. Data untraceability makes it even more difficult to figure out these facts, highlighting the importance of the availability of basic data or main metadata for a better understanding of aquatic communities.

Modelling relationships within Diptera assemblages

Our literature meta-analysis also revealed that no studies have focused on the relationships between presence and abundance of stream dwelling Diptera, so we provide here the first information on their co-occurrence and correlation patterns. As we defined Chironomidae, Ceratopogonidae, and Simuliidae as major families, we examined the occurrence of other families in connection with them as well. It should be mentioned, however, that although Chironomidae and Simuliidae are among the most studied and relatively better-known families (Armitage *et al*., 1995; Currie and Adler, 2008), Ceratopogonidae is still a poorly understood one, while others are even more insufficiently known (Wagner *et al*., 2008), making the further evaluation of our results merely provisory.

Our results showed a strong positive relationship between Ceratopogonidae and Chironomidae, and negative relationships between both these families and Simuliidae. It is not surprising, since Ceratopogonidae and Chironomidae can be found in large numbers in pools and shallow, slow-flowing parts of streams (Armitage *et al*., 1995; Szadziewski *et al*., 1997), while Simuliidae is exclusively connected to running waters (Currie and Adler, 2008). Furthermore, we can distinguish a Chironomidae/Ceratopogonidae-related assemblage (including Culicidae, Dixidae, Psychodidae, Ptychopteridae, Tabanidae). In contrast, although there were other families that are typically connected to running waters (*e.g.*, Athericidae, Pediciidae), these did not form a distinguishable assemblage along with Simuliidae as they were negatively correlated with it, or positively correlated with Chironomidae and/or Ceratopogonidae too. Other minor families also did not show clear connection with the three major families. The diversity of species' autecology and habitat preference can be very high within a single family (Nilsson, 1997; Oosterbroek, 2006), and family-level identification might blur the species-specific ecological differences, resulting in the unclear co-occurrence patterns observed. Our results suggest that identification to taxonomic levels more precise than family is essential for better understanding the underlying processes behind organisation of Diptera assemblages (Heino, 2014; de Oliveira *et al*., 2020).

We found that the higher the number of individuals of Chironomidae and Ceratopogonidae, the higher the number of co-occurring families, but the lower their Shannon diversity (*i.e.*, the assemblage of a few common and less frequent taxa). It shows that the habitats suitable for the major families are suitable for many minor families too, but these families appear as rare with higher probability. The diversity of the assemblage of the minor Diptera families is lower with the high abundance of Chironomidae and Ceratopogonidae, but we did not find any connection with the abundance of Simuliidae. Although the underlying processes are not clear, it suggests that Chironomidae and Ceratopogonidae might be more usable than Simuliidae in predicting the occurrence of minor Diptera families.

Our most crucial question, which promises practical implications, was whether it is possible to model the occurrence and abundance of rare, often lesser-known, and hardly recognized minor families in the stream-dwelling community based on those of common, abundant, and often better-known major families. If this was possible, in the case of any community, it would be possible to specify, either in advance or in retrospect, which previously neglected families can be expected to occur. The RDA clearly showed that by considering the three most abundant families together, only a tiny fraction (<3%) of the variance of the entire assemblage can be explained, meaning that it is not possible to reliably estimate abundance characteristics for any families based exclusively on their co-occurrence. Our other idea was whether there is a general combination of families that describes the abundances of minor ones with an acceptable probability and a sufficiently high explanatory power, and if there is, what role the major families play in it. When the explanatory variables (families) were chosen objectively for GAMs, the best models had higher explanatory powers (13-62%), and the relationship between certain families could be strongly assumed. At the same time, the characteristics of each family was always explained by a unique combination of different families. In other words, a general model with the same combination of predictor families cannot be set up, and the role of major families in these models are comparable with that of the other families. It means that in order to reveal and understand complex ecological processes in streams, it is necessary to involve the minor families, too.

Regrettably, acquiring the requisite expertise for a comprehensive identification of Diptera remains an elusive goal. A more

realistic solution lies in the establishment of robust DNA Barcode Reference Libraries (Morinière *et al*., 2019; Weigand *et al*., 2019). Using molecular tools (eDNA, metabarcoding), these databases can facilitate the analysis of DNA extracted from sorted samples or preservative alcohol, enabling a more accurate characterization of the hidden diversity, and holding potential significance for bioindication purposes. A notable technical challenge is applying these molecular tools to past samples already catalogued and archived in collections, but at the same time, this would enable the re-evaluation of historical communities by incorporating taxonomic groups, such as Diptera, that were previously underestimated.

CONCLUSIONS

Based on the result of a literature meta-analysis and its comparison with focused case studies, the underrepresentation of true flies, Diptera within macroinvertebrates became evident in stream studies, both in terms of diversity and abundance. This is especially true for the rare, less abundant, minor families, but also affects the major ones. Mapping the relationship between individual families may provide intimation for modelling the occurrence of certain families, however, a general model that can reliably predict the presence and abundance of minor families based on a few well-known major taxa cannot be established. Therefore, in quantitative or semi-quantitative studies of entire macroinvertebrate communities, it is still necessary to implement a very thorough sorting of samples taking all dipterans, especially the minor families with the lowest possible level of identification, into account to avoid significant loss of information in the abundance and biomass of this compositionally and functionally very diverse group that forms a significant part of the entire community. In the near future, DNA barcoding based on almost full-coverage reference libraries may make the identification of Diptera much easier, although the determination of their abundance will remain a critical and necessary step for a long time.

ACKNOWLEDGEMENTS

The Authors thank the six DRYvER field sampling teams for conducting the field work, and Éva Horváth-Tihanyi, József Balázs Berta, Anita Szloboda, Dorottya Hárságyi, Khouloud Sebteoui, Áron Zenke, Patrik Kis, and Zsolt Kovács (University of Pécs) for their enormous contribution in the laboratory processes (sample management and sorting).

REFERENCES

Aazami J, Maghsodlo H, Mira SS, Valikhani H, 2020. Health evaluation of riverine ecosystems using aquatic macroinvertebrates: a case study of the Mohammad-Abad River, Iran. Int J Environ Sci Te 17:2637-2644.

Alemneh T, Ambelu A, Bahrndorff S, Mereta ST, Pertoldi, Zaitchik BF, 2017. Modeling the impact of highland settlements on ecological disturbance of streams in Choke Mountain Catchment: Macroinvertebrate assemblages and water quality. Ecol Indic 73:452-459.

Armitage PD, Cranston PS, Pinder LCV, 1995. The Chironomidae: biology and ecology of non-biting midges. Chapman & Hall, London: 572 pp.

Barbour M, Gerritsen TJ, Snyder BD, Stribling JB, 1999. Rapid bioassessment protocols for use in streams and wadable rivers: periphyton, benthic macroinvertebrates and fish. United States Environmental Protection Agency, Office of Water, Washington: 202 pp.

Bartošová M, Schenková Polášková V, Bojková J, Šorfová V, Horsák M, 2019. Macroinvertebrate assemblages of the post-mining calcareous stream habitats: Are they similar to those inhabiting the natural calcareous springs? Ecol Eng 136:38-45.

Başören Ö, Kazancı N, 2020. Distribution of aquatic Diptera larvae of Yeşilırmak River (Turkey) and ecological characteristics. Ege J Fish Aquat Sci. 37:397-407.

Biecek P, 2018. DALEX: Explainers for Complex Predictive Models in R. J Mach Learn Res 19:1-5.

Breiman L, 2001. Random forests. Mach Learn 45:5-32.

Campos RE, 2015. Aquatic Diptera assemblages in four sympatric Eryngium (Apiaceae) phytotelmata in flowering and senescent times. J Nat Hist 50:1-17.

Cazorla CG, Campos RE, 2020. Ceratopogonidae (Diptera) communities in a protected area threatened by urbanization. Neotrop Entomol 49:361-368.

Coombes KR, Brock G, Abrams ZB, Abruzzo LV, 2019. Polychrome: creating and assessing qualitative palettes with many colors. J Stat Softw 90:1-23.

Cortelezzi A, Paggi AC, Rodríguez M, Rodrigues Capítulo A, 2011. Taxonomic and nontaxonomic responses to ecological changes in an urban lowland stream through the use of Chironomidae (Diptera) larvae. Sci Total Environ 409:1344-1350.

Courtney GW, Cranston PS, 2015. Order Diptera, p. 1043-1058. In: J.H. Thorp and D.C. Rogers eds.), Thorp and Covich's Freshwater Invertebrates. Academic Press.

Courtney GW, Pape T, Skevington JH, Sinclair BJ, 2017. Biodiversity of Diptera, p. 229-278. In: R. Foottit and P. Adler (eds.), Insect biodiversity: science and society. Wiley Blackwell.

Cuadrado LA, Moncada LI, Pinilla GA, Larrañaga A, Sotelo AI, Adler PH, 2019. Black fly (Diptera: Simuliidae) Assemblages of high Andean rivers respond to environmental and pollution gradients. Environ Entomol 48:815-825.

Currie DC, Adler PH, 2008. Global diversity of black flies (Diptera: Simuliidae) in freshwater. Hydrobiologia 595:469-475.

Datry T, Allen D, Argelich R, Barquin J, Bonada N, Boulton A, et al., 2021. Securing biodiversity, functional integrity, and ecosystem services in drying river networks (DRYvER). Res Ideas Outcomes 7:e77750.

Debiasi D, Franceschini A, Paoli F, Lencioni V, 2022. How do macroinvertebrate communities respond to declining glacial influence in the Southern Alps? Limnetica 41:121-137.

de Oliveira SSJr, Ortega JCG, dos Santos Ribas LG, Lopes VG, Bini LM, 2020. Higher taxa are sufficient to represent biodiversity patterns. Ecol Indic 111:105994.

Dobson M, 2013. Family□level keys to freshwater fly (Diptera) larvae: A brief review and a key to European families avoiding use of mouthpart characters. Freshwater Rev 6:1-32.

Docherty CL, Hannah DM, Riis T, Lund M, Abermann J, Milner AM, 2018. Spatio-temporal dynamics of macroinvertebrate communities in northeast Greenlandic snowmelt streams. Ecohydrology 11:e1982.

Docile TN, Figueiró R, Portela C, Nessimian JL, 2016. Macroinvertebrate diversity loss in urban streams from tropical forests. Environ Monit Assess 188:237.

Faasch H, 2015. Identification guide to aquatic and semi-aquatic Diptera larvae. DGL, Dt. Ges. für Limnologie. Hardegsen, Essen: 179 pp.

Feld CK, Kiel K, Lautenschkager M, 2002. The indication of morphological degradation of streams and rivers using Simuliidae. Limnologica 32:273-288.

Friberg N, Sandin L, Furse MT, Larsen SE, Clarke RT, Haase P, 2006. Comparison of macroinvertebrate sampling methods in Europe. Hydrobiologia 566:365-378.

Fusari LM, Dantas GPS, Pinho LC, 2018. Order Diptera, p. 607-623. In: N. Hamada, J.H. Thorp and D.C. Rogers (eds.), Thorp and Covich's Freshwater Invertebrates: Keys to Neotropical Hexapoda. Elsevier.

Gao X, Niu C, Chen Y, Yin X, 2014. Spatial heterogeneity of stream environmental conditions and macroinvertebrates community in an agriculture dominated watershed and management implications for a large river (the Liao River, China) basin. Environ Monit Assess 186:2375-2391.

Gerhardt RR, Lawrence JH, 2019. Flies (Diptera), p. 171-190. In: G.R. Mullen and L.A. Durden (eds.), Medical and Veterinary Entomology. Academic Press.

Hamilton NE, Ferry M, 2018. "ggtern: Ternary Diagrams Using ggplot2." J Stat Softw 87:1-17.

Hammer Ø, Harper DAT, Ryan PD, 2001. PAST: Paleontological Statistics Software Package for Education and Data Analysis. Palaeontol Electron 4:4.

Hastie TJ, Tibshirani R, 1986. Generalized additive models. Stat Sci 1:297-310.

Heino J, 2014. Taxonomic surrogacy, numerical resolution and responses of stream macroinvertebrate communities to ecological gradients: Are the inferences transferable among regions? Ecol Indic 36:186-194.

Hövemeyer K, 2000. Ecology of Diptera, p. 437-489. In: L. Papp and B. Darvas (eds.), Contribution to a manual of Palaearctic Diptera (with special references of flies of economic importance). Science Herald.

Illéšová D, Halgoš J, Krno I, 2008. Blackfly assemblages (Diptera, Simuliidae) of the Carpathian river: habitat characteristics, longitudinal zonation and eutrophication. Hydrobiologia 598:163-174.

Ivković M, Miliša M, Baranov V, Mihaljević Z, 2015. Environmental drivers of biotic traits and phenology patterns of Diptera assemblages in karst springs: The role of canopy uncovered. Limnologica 54:44-57.

JASP Team, 2022. JASP (Version 0.16.4.0). Available from: https://jasp-stats.org/

de Jong Y, Verbeek M, Michelsen V, Bjørn Pde P, Los W, Steeman F, et al., 2014. Fauna Europaea - all European animal species on the web. Biodivers Data J 17:e4034.

Kassambara A, 2023. ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.6.0. Available from: https://CRAN.R-project.org/package=ggpubr

Kitching RL, Bickel DJ, Boulter S, 2005. Guild analyses of dipteran assemblages, a rationale and investigation of seasonality and stratification in selected rainforest faunas, p. 388-415. In: D.K. Yeates and B.M. Wiegmann(eds.), The evolutionary biology of flies. Columbia University Press.

Korte T, 2010. Current and substrate preferences of benthic invertebrates in the rivers of the Hindu Kush-Himalayan region as indicators of hydromorphological degradation. Hydrobiologia 651:77-91.

Kriska Gy, 2013. Freshwater invertebrates in Central Europe. A field guide. Springer, Wien: 411 pp.

Kuhn M, 2008. Building Predictive Models in R Using the caret Package. J Stat Soft 28:1-26.

Lautenschläger M, Kiel E, 2005. Assessing morphological degradation in running waters using Blackfly communities (Diptera, Simuliidae): Can habitat quality be predicted from land use? Limnologica 35:262-273.

Lenat DR, 1993. A biotic index for the southeastern United States: derivation and list of tolerance values, with criteria for assigning water-quality ratings. J N Am Benthol Soc 12:279-290.

Lencioni V, Adler PH, Courtney GW, 2023. Order Diptera, p. 503-535. In: A. Maasri and J.H. Thorp (eds), Identification and ecology of freshwater arthropods in the Mediterranean Basin. Elsevier.

Marchamalo M, Springer M, Acosta R, González-Rodrigo B, Vásquez D, 2018. Responses of aquatic macroinvertebrates to human pressure in a tropical highland volcanic basin: Birrís River, Irazú Volcano (Costa Rica). Hidrobiológica 28:179-190.

Marrochi MN, Hunt L, Solis M, Scalise AM, Fanelli SL, Bonetto C, Mugni H, 2021. Land-use impacts on benthic macroinvertebrate assemblages in pampean streams (Argentina). J Environ Manage 279:111608.

McCreadie JW, Adler PH, 2012. The roles of abiotic factors, dispersal, and species interactions in structuring stream assemblages of black flies (Diptera: Simuliidae). Aquat Biosyst 8:14.

McLean IFG, 2000. 1.12. Beneficial Diptera and their role in decomposition, p. 491–517. In: L. Papp and B. Darvas (eds.), Contribution to a Manual of Palaearctic Diptera (with special references of flies of economic importance). Volume 1, General and Applied Dipterology. Science Herald.

Miliša M, Stubbington R, Datry T, Cid N, Bonada N, Šumanović M, Milošević D, 2022. Taxon-specific sensitivities to flow intermittence reveal macroinvertebrates as potential bioindicators of intermittent rivers and streams. Sci Total Environ 804:150022.

Morinière J, Balke M, Doczkal D, Geiger MF, Hardulak LA, Haszprunar G, et al., 2019. A DNA barcode library for 5,200 German flies and midges (Insecta: Diptera) and its implications for metabarcoding-based biomonitoring. Mol Ecol Resour 19:900-928.

Nakano D, Nakamura F, 2008. The significance of meandering channel morphology on the diversity and abundance of macroinvertebrates in a lowland river in Japan. Aquat Conserv 18:780-798.

Narangarvuu D, Hsu C-B, Shieh S-H, Wu F-C, Yang P-S, 2014. Macroinvertebrate assemblage patterns as indicators of water quality in the Xindian watershed, Taiwan. J Asia-Pac Entomol 17:505-513.

Nilsson AN, 1997. Aquatic insects of North Europe. A taxonomic handbook. Vol. 2. Odonata, Diptera. Apollo Books, Stenstrup: 440 pp.

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al., 2020: vegan: Community Ecology Package. R package, version 2.5-7.

Omelková M, Syrovátka V, Křoupalová V, Rádková V, Bojková J, Horsák M, et al., 2013. Dipteran assemblages of spring fens closely follow the gradient of groundwater mineral richness. Can J Fish Aquat Sci 70:689-700.

Ono ER, Manoel PS, Melo ALU, Uieda VS, 2020. Effects of

riparian vegetation removal on the functional feeding group structure of benthic macroinvertebrate assemblages. Commun Ecol 21:145-157.
Oosterbroek P, Theowald B, 1991. Phylogeny of the Tipuloidea based on characters of larvae and pupae (Diptera, Nematocera) with an index to the literature except Tipulidae. Tijdschr Entomol 134:211-267.
Oosterbroek P, 2006. The European families of the Diptera. Identification, diagnosis, biology. KNNV: Utrecht, Netherlands: 204 pp.
Paine GH, And JR, Gaufin AR, Taft RA, 1956. Aquatic Diptera as indicators of pollution in a Midwestern Stream. Ohio J Sci 56:291-304.
Pires MM, Grech MG, Stenert C, Maltchik L, Epele LB, McLean KI, et al., 2021. Does taxonomic and numerical resolution affect the assessment of invertebrate community structure in New World freshwater wetlands? Ecol Indic 125:107437.
Polášková V, Schenková J, Bílková M, Poláková M, Šorfová V, Polášek M, Schlaghamerský J, Horsák M, 2020. Drivers of small-scale Diptera distribution in aquatic-terrestrial transition zones of spring fens. Wetlands 40:235-247.
R Core Team, 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.
Rosenberg D, 1992. Freshwater biomonitoring and Chironomidae. Neth J Aquat Ecol 26:101-122.
Saether OA, 1979. Chironomid communities as water quality indicators. Ecography 2:65-74.
Sarremejane R, Mykrä H, Bonada N, Aroviita J, Muotka T, 2017. Habitat connectivity and dispersal ability drive the assembly mechanisms of macroinvertebrate communities in river networks. Freshwater Biol 62:1073-1082.
Schmera D, Erős T, Greenwood MT, 2007. Spatial organization of a shredder guild of caddisflies (Trichoptera) in a riffle - Searching for the effect of competition. Limnologica 37:129-136.
Slowikowski K, 2023. ggrepel: automatically position non-overlapping text labels with 'ggplot2'. R package version 0.9.4. Available from: https://CRAN.R-project.org/package=ggrepel
Smith KGV, 1989. An introduction to the immature stages of British flies. Handbooks for the identification of British insects, part 14, vol. 10. Royal Entomological Society of London, London: 280 pp.
Smith KGV, Ferrar P, 2000. Key to families – larvae, p. 201-239. In: L. Papp and B. Darvas (eds.), Contribution to a manual of Palaearctic Diptera (with special references of flies of economic importance). Volume 1, General and Applied Dipterology. Science Herald.
Souto RDMG, Facure KG, Pavanin LA, Jacobucci GB, 2011. Influence of environmental factors on benthic macroinvertebrate communities of urban streams in Vereda habitats, Central Brazil. Acta Limnol Bras 23:293-306.
Sundermann A, Lohse S, Beck LA, Haase P, 2007. Key to the larval stages of aquatic true flies (Diptera), based on the operational taxa list for running waters in Germany. Ann Limnol 43:61-74.
Szadziewski R, Krzywiński J, Giłka W, 1997. Diptera Ceratopogonidae, biting midges, p. 243-263. In: A.N. Nilsson (ed.) Aquatic insects of North Europe. A taxonomic handbook. Vol. 2. Odonata, Diptera. Apollo Books, Stenstrup.
Tachet H, Richoux P, Bournaud M, Usseglio-Polatera P, 2010. [Invertébrés d'eau douce. Systématique, biologie, écologie].[Book in French]. CNRS Éditions, Paris: 607 pp.
Thakur Y, Grover A, Sinha R, 2022. Differential distribution of macroinvertebrate associated with water quality. World Water Pol 9:84-112. https://doi.org/10.1002/wwp2.12089
Timmermans KR, Peeters W, Tonkes M, 1992. Cadmium, zinc, lead and copper in Chironomus riparius (Meigen) larvae (Diptera, Chironomidae): uptake and effects. Hydrobiologia 241:119-134.
Usher MB, Edwards M, 1984. A dipteran from south of the Antarctic Circle: *Belgica antarctica* (Chironomidae), with a description of its larva. Biol J Linn Soc 23:19-31.
van den Wollenberg AL, 1977. Redundancy analysis: An alternative for canonical correlation analysis. Psychometrika 42:207-219.
Wagner R, Barták M, Borkent A, Courtney G, Goddeeris B, Haenni J-P, et al., 2008. Global diversity of dipteran families (Insecta Diptera) in freshwater (excluding Simulidae, Culicidae, Chironomidae, Tipulidae and Tabanidae). Hydrobiologia 595:489-519.
Weigand H, Beermann AJ, Čiampor F, Costa FO, Csabai Z, Duarte S, et al., 2019. DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. Sci Total Environ 678:499–524.
Westheide W, Rieger R, 1996. [Spezielle Zoologie, Teil 1: Einzeller und wirbellose Tiere].[Book in German]. Gustav Fischer, Stuttgart.
Wickham H, 2016. ggplot2: Elegant graphics for data analysis. Available from: https://ggplot2.tidyverse.org
Wickham H, 2022. stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.5.0. Available from: https://CRAN.R-project.org/package=stringr
Wood SN, 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J R Stat Soc B 73:3-36.
Zhang ZQ, 2011. Animal biodiversity: An introduction to higher-level classification and taxonomic richness. Zootaxa 3148:7-12.

Online supplementary material:

Tab. S1. The number of studies found in the two databases through the searches and the number of hits matching the inclusion criteria.

Tab. S2. The listed criteria and reasons for inclusion or exclusion of publications in the analysis.

Tab. S3. List of processed papers with their citation, year of publication, DOI, and basic data derived from them.

Fig. S1. Redundancy analysis scatterplot reveal very minor impact of the predictor families on abundances of several but not all families, indicating the lack of a general rule.

Fig. S2. Results of the Generalized Additive Models for families where best models show lower explanatory power.