Supporting Information - Tackling discrepancies in freshwater invertebrate trait databases: Harmonising across continents and aggregating taxonomic resolution

Stefan Kunz[1], Ben J. Kefford[2], Astrid Schmidt-Kloiber[3], Christoph D. Matthaei[4], Philippe Usseglio-Polatera[5], Wolfram Graf[3], N. LeRoy Poff[6], Leon Metzeling[7], Laura Twardochleb[8], Charles P. Hawkins[9], and Ralf B. Schäfer[1]

[1]Institute for Environmental Sciences, University of Koblenz-Landau, Landau, Germany

[2]Centre for Applied Water Science, Institute for Applied Ecology, University of Canberra, Canberra, Australia

[3]Institute of Hydrobiology and Aquatic Ecosystem Management, University of Natural Resources and Life Sciences Vienna (BOKU), Vienna, Austria

[4]Department of Zoology, University of Otago, Dunedin, New Zealand

[5]University of Lorraine, CNRS, LIEC, Metz, France

[6]Department of Biology, Colorado State University, Fort Collins, USA

[7]Environment Protection Authority Victoria, Applied Sciences Division, Macleod, Australia

[8]Department of Fisheries and Wildlife, Michigan State University, East Lansing, USA
[9]Department of Watershed Sciences, National Aquatic Monitoring Center, and the Ecology Center, Utah State University, Logan, USA

Table S1: Comparison of trait definition differences between invertebrate trait databases. Only traits that are differently described across databases are listed. The definition is quoted if it enables differences to be identified, otherwise the differences are described. Bullet points indicate when several traits have been used to describe a morphological or behavioural property. The hyphen indicates a missing trait. Reproduction was captured in multiple grouping features per database. Hence, differences for reproduction have been described in the paper. Body form traits are not different between databases, except that the Vieira database contains the trait Bluff (blocky) which does not appear in the other databases.

Trait	Freshwater-ecology.info	Tachet	CONUS	Vieira	Australia	New Zealand

Feeding shredder	"Feed from fallen leaves, plant tissues, CPOM"	"Eat coarse detritus, plants or *animal material*"	• "Shred decomposing vascular plant tissue" • Trait herbivore includes among others insect that shred *living aquatic plants*	Shredder	• Detrivore† • Trait herbivore includes among others the trait shredder	Shredders

Feeding predator	"Eating from prey"	• Carvers, engulfers & swallowers • Piercers (plants & animals) are an additional trait	Engulfers ("ingest prey whole or in parts") & piercers ("prey tissues and suck fluids")	Predator	Piercer & engulfer	Predator
Feeding filter-feeder	• Active filter feeders • Passive filter feeders	No distinction between active and passive	No distinction between active and passive	No distinction between active and passive	No distinction between active and passive	No distinction between active and passive
Semivoltine	"One generation in two years"	"Life cycle lasts *at least* two years"	"< 1 generation per year"	"< 1 generation per year"	"< 1 generation per year"	"< 1 reproductive cycle per year"

Multi-voltine	"*Three* or more generations per year"‡	"Able to complete *at least* two successive generations per year"	"> 1 generations per year"	"> 1 generations per year"	• 1-2 generations per year • bi - or multi-voltine • up to 5 generations per year • up to 10 generations per year	"> 1 reproductive cycles per year"

	Locomotion swimming	• Passive movement like floating or drifting (trait swimming/scating) • Active movement (trait swimming/diving) .	• Surface swimmers (over and under the water surface) • Full water swimmers (e.g. Baetidae).	”Adapted for ”fishlike” swimming”	Swimmer	• Swimmer • Skater	Swimmers (water column)

Locomotion burrowing	"Burrowing in *soft* substrates or boring in *hard* substrates"	• Burrowing "within the first centimeters of the benthic fine sediment" • Interstitial (endobenthic)	"Inhabiting *fine* sediment of streams and lakes"	Burrower	"Moving deep into the substrate and thus avoiding flow"	Burrowers (infauna)
Locomotion sprawling & walking	"Sprawling or walking actively with legs, pseudopods or on a mucus"	-	Sprawling: "inhabiting the surface of floating leaves of vascular hydrophytes or fine sediments"	Sprawler	-	-

Locomotion crawling	-	”Crawling over the bottom substrate”	Defined as crawling on the surface of floating leaves or fine sediments on the bottom	-	• Crawler • Sprawler • Climber • Clinger	Crawlers (epibenthic)
Locomotion sessil	Does not distinguish temporarily and permanently attached	• Temporarily attached • Permanently attached	Does not distinguish temporarily and permanently attached	Does not distinguish temporarily and permanently attached	• Temporarily attached • Permanently attached	Does not distinguish temporarily and permanently attached

Respiration plastron & spiracle	Plastron and spiracle (aerial) are two separate traits	Definition includes respiration using air stores of aquatic plants	Plastron and spiracle combined into one trait	• Spiracular gills • Plastron • Atmosph. breathers • Plant breathers	• Plastron and spiracle (termed aerial) occur as separate and combined traits • Air (plants) • Atmospheric • Functional spiracles	• Plastro • Spiracle (termed aerial)
Body size small	-	Multiple size classifications[¶]	< 9 mm	< 9 mm	< 9 mm [†§]	Multiple size classifications
Body size medium	-		9 - 16 mm	9 - 16 mm	9 - 16 mm	
Body size large	-		> 16 mm	> 16 mm	> 16 mm	

† Traits from Botwe et al.

‡ Contains also bivoltine (two generations per year), trivoltine (three generations per year) and flexible.

§ Contains a size trait with numeric size values. Contains also traits classifying size like Tachet and like the North American trait databases.

¶ Size classifications: $<= 0.25\ cm$, $> 0.25 - 0.5\ cm$, $0.5 - 1\ cm$, $1 - 2\ cm$, $2 - 4\ cm$, $4 - 8\ cm$, $> 8\ cm$. No distinction into small, medium and large.

⋆ Size classifications: $> 0.25 - 0.5\ cm$, $0.5 - 1\ cm$, $1 - 2\ cm$, $2 - 4\ cm$, $4 - 8\ cm$. No distinction into small, medium and large.

Comparing aggregation methods

Comparison of family-level aggregated traits with family-level assigned traits

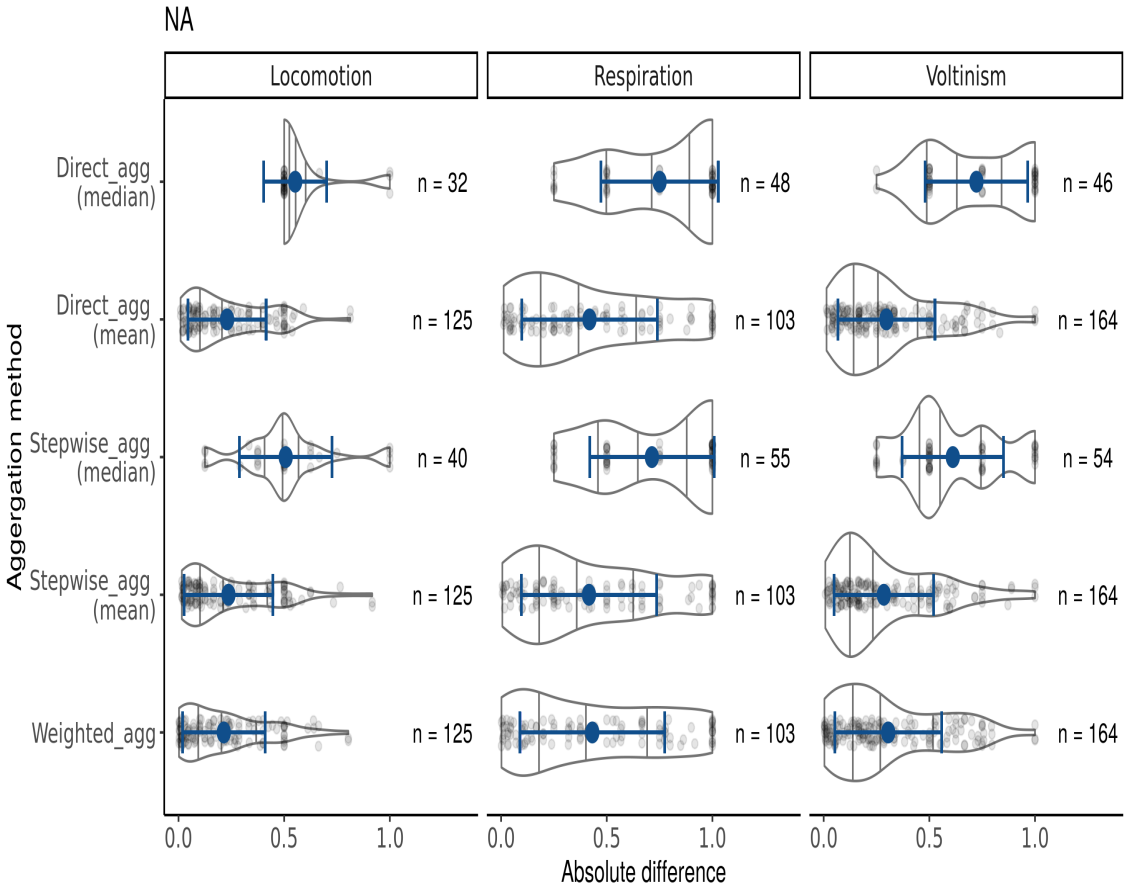

Figure S1: Cases (factor combination of investigated families and traits) where differences occurred between aggregated traits and expert assigned traits at family level for the North American dataset. Violin plots - mirrored density plots - show the density of the absolute trait affinity differences for the grouping features locomotion, respiration, and body size. For more details see Figure 2.

Effect of phylogeny and trait variability on aggregation outcomes

To examine the influence of phylogeny and trait variability on the outcomes of the different trait aggregation methods, we created three hypothetical scenarios. We simulated three different families, each containing 25 total species but with different phylogenetic structures. The three families consisted of (1) a family with an equal number of genera and species (five genera each with five species each), denoted as *sim_base*; (2) a family in which one genus had a much larger number of species than the other four genera (1 genus with 13 species, 4 genera with 3 species each), denoted as *sim_extreme*; (3) a family in which all genera had a different number of species (8, 2, 7, 3, 5), denoted as *sim_variation*. We assigned a hypothetical grouping feature with three traits (T1, T2, and T3) to each scenario. We then simulated the 25 affinities, one per species, for each trait, by sampling from a truncated normal distribution bound by 0 and 1 and with a mean value of 0.5. To simulate different levels of trait variability, we repeated the sampling 100 times for each of the 5 standard deviations (0.2, 0.4, 0.6, 0.8, and 1), resulting in 12,500 simulated trait affinities for each simulated trait (25 species per family $\times$ 5 levels of trait variability $\times$ 100 replicates). We converted simulated trait affinities to proportions, as was done during trait database processing, and assigned the 12,500 simulated affinities per trait to each of the three family scenarios. We then applied the five trait aggregation methods described above to each simulated dataset. We compared the resulting ranges of aggregated trait affinities between levels of trait variability and phylogenetic scenario as well as the differences in trait affinities obtained by each aggregation method.

Results: Comparison of aggregation methods with varying phylogenies and trait variability

The simulations showed that both phylogenetic structure and trait variability affected aggregated trait affinities, although only to a small degree, and that aggregation method mattered in terms of the ranges of trait affinities produced over simulation replicates. The effect of phylogenetic structure differed across aggregation

methods, but, as we expected, the range of trait affinities increased with increasing trait variability for all aggregation methods (Figure S2).

Phylogenetic structure appeared to influence the outcomes of the different aggregation methods. For the *sim_base* scenario (equal numbers of genera and species), the mean aggregation methods yielded similar ranges of aggregated trait affinities within each level of trait variability, and the median aggregation methods consistently produced greater ranges of aggregated trait affinities than the other methods. The largest ranges were produced by the *stepwise_agg* $_{median}$ method. For the more complex phylogenetic structures, *sim_extreme* (one genus with a much larger number of species than the other four) and *sim_variation* (all genera with a different number of species), the *stepwise_agg* $_{median}$ method still produced the largest ranges of trait affinities for most levels of trait variability (and $direct\_agg_{mean}$ produced the narrowest ranges for most levels of trait variability), but there was much less consistency in the ranges of trait affinities for all aggregation methods (i.e., the ranges were different among all aggregation methods).

Although trait aggregation methods were affected by phylogenetic structure and trait variability to some degree, in most simulated datasets, the different aggregation methods resulted in similar trait affinities. Only 1.4 %, or 213 out of 15.000 total comparisons (3 scenarios $\times$ 5 levels of trait variability $\times$ 10 unique comparisons of trait aggregation methods $\times$ 100 replicates) showed a difference equal or greater than an absolute trait affinity of 0.1. Most (83.5 %) of these differences occurred in the *sim_extreme* scenario and were found between the aggregation methods $direct\_agg_{mean}$ and $stepwise\_agg_{median}$, $direct\_agg_{median}$ and $stepwise\_agg_{median}$, and $stepwise\_agg_{median}$ and *weighted_agg* (Figure S4).

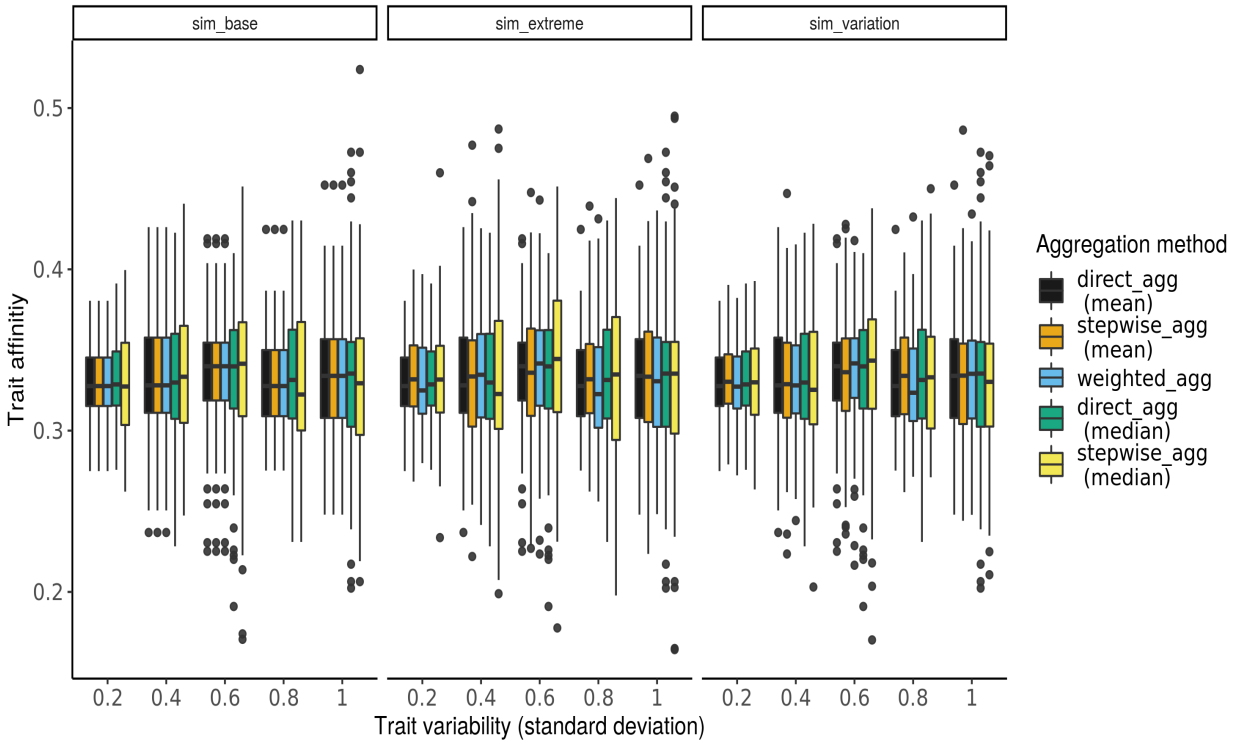

Figure S2: Ranges of aggregated trait affinities for the three examples of phylogenetic structures and simulated levels of trait variability. Shown are the results only for one simulated trait (T1). Similar results were obtained for the other simulated traits (Figure S3). Boxplots depict results for 100 replicated simulations of each trait aggregation method. The boxplot depicts the median and encompasses the 25^{th} and 75^{th} percentile. Horizontal black lines depict the median. Whiskers extend to the largest and smallest value respectively no further than 1.5 $\times$ the inter-quartile range. Outliers beyond the end of the whiskers are plotted as grey dots. Trait aggregation methods are in order of least to greatest produced ranges to improve visual inspection.

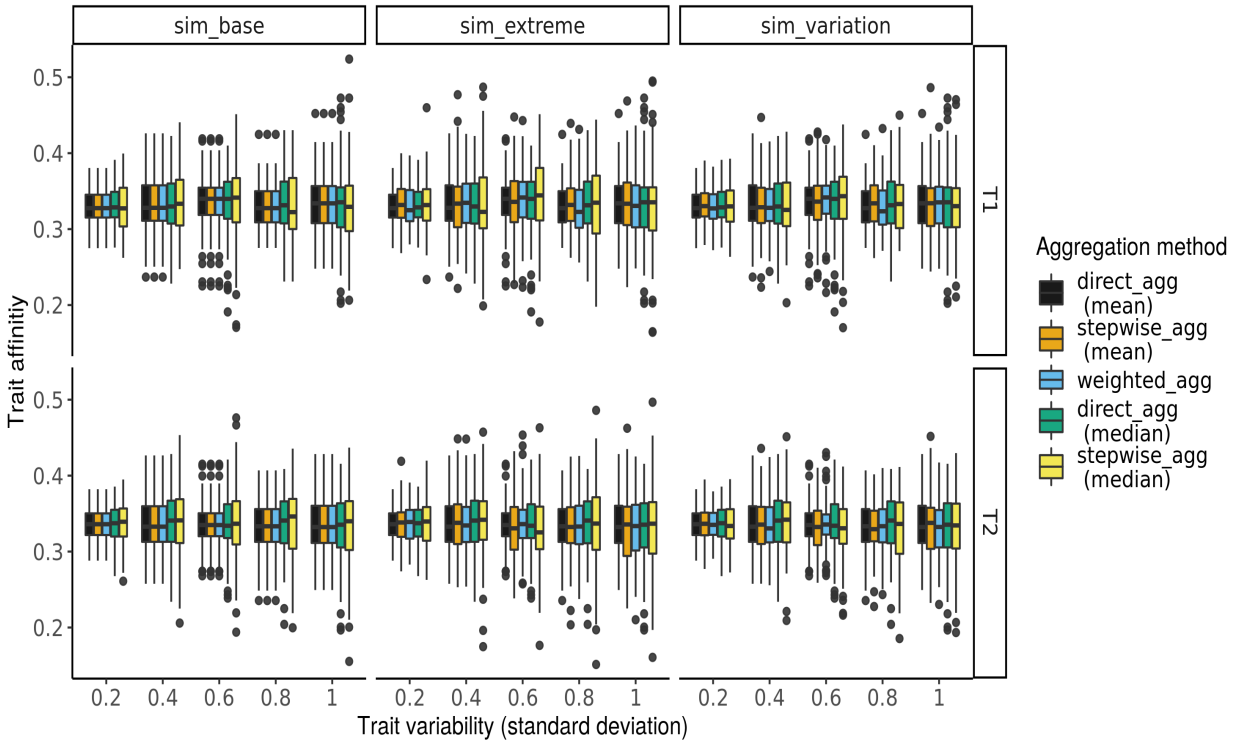

Figure S3: Ranges of aggregated trait affinities for the three examples of taxonomic hierarchies and simulated levels of trait variability. Shown are the results for the simulated traits T2 and T3. Boxplots depict results for 100 replicated simulations of each trait aggregation method. Trait aggregation methods are in order of least to greatest produced ranges to improve visual inspection. For more details see Figure 3.

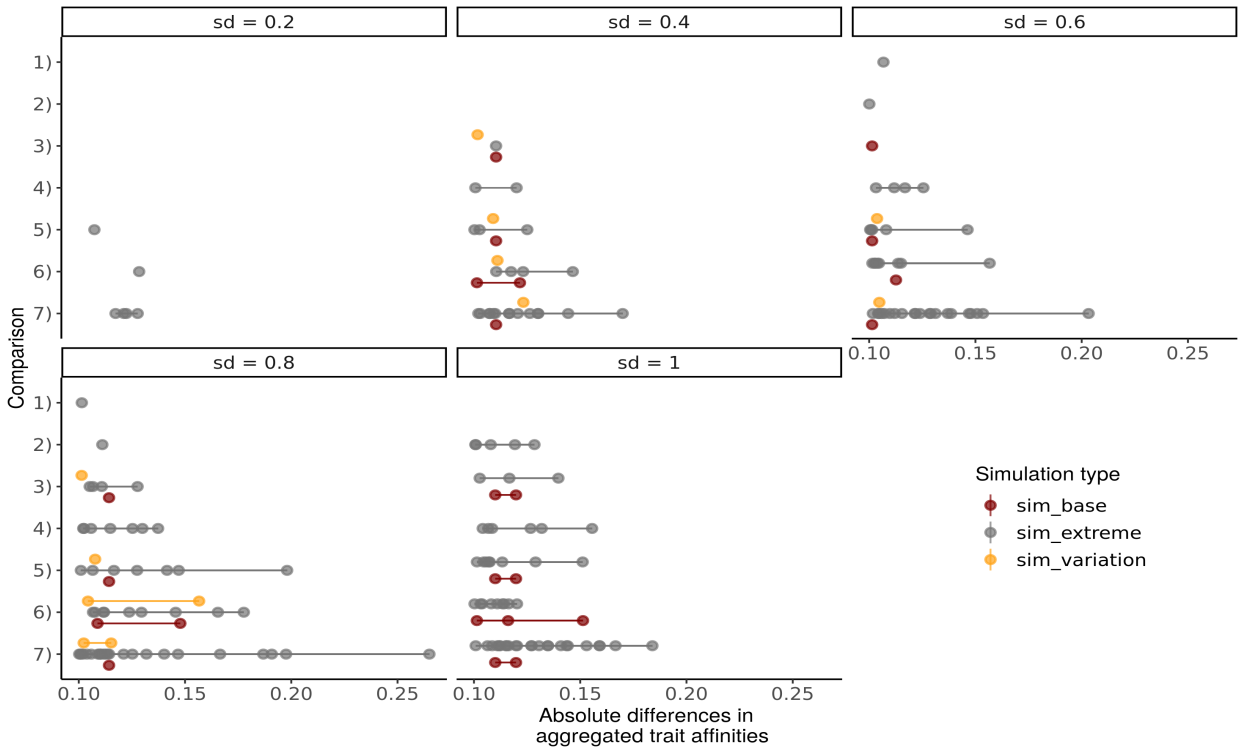

Figure S4: Comparison between the aggregated trait affinities produced by the different trait aggregation methods for every simulated dataset across all 3 simulated traits (5,000 comparisons per scenario). Results are shown for different levels of trait variability which is represented by the standard deviation (SD) of the simulated traits. Dots depict comparisons where absolute differences between aggregated trait affinities were greater than 0.1. Overall, there are 10 possible unique comparisons of which 7 produced absolute differences greater than 0.1.

Comparisons:

1) $direct\_agg_{median}$ - $stepwise\_agg_{mean}$
2) $direct\_agg_{median}$ - $weighted\_agg$
3) $stepwise\_agg_{mean}$ - $stepwise\_agg_{median}$
4) $stepwise\_agg_{mean}$ - $weighted\_agg$
5) $direct\_agg_{mean}$ - $stepwise\_agg_{median}$
6) $direct\_agg_{median}$ - $stepwise\_agg_{median}$
7) $stepwise\_agg_{median}$ - $weighted\_agg$

Discussion: Comparison of aggregation methods with varying phylogenies and trait variability

We expected that, in addition to averaging measures, different weighting approaches to aggregation (i.e., equal weight to each taxon, equal weight to each genus, or weighted by number of species) would affect affinities for families with varying phylogenetic structure. However, we found through simulations that trait variability had a greater effect on the range of affinities (i.e., greater ranges of affinities at higher levels of variance) and on the differences in affinities between weighting approaches. In the simulation, taxonomic structural unevenness appeared to produce some variation in affinities among weighting approaches, especially for a family where one genus had a much larger number of species than its other genera. The comparison between aggregated and assigned traits showed, that the number of differing cases and the mean absolute differences in trait affinities, as well as the distributions of absolute trait affinity differences to assigned traits, were similar across the mean aggregation and across the median aggregation methods, which suggests a small influence of the weighting approach on the aggregation outcomes. The minor impact of the weighting approaches on trait aggregation might be explained by the fact that a considerable portion of taxa had low numbers of genera or species. Of the taxa that were compared from the North American trait dataset, 14 % were identified at family and 62 % at genus level, 52 % comprised five or fewer genera, and 13 % contained just one genus (Figure S5). In the Australian dataset, 21 % of the compared taxa were identified at family, 40 % at genus, and 39 % at species level, 68 % of the taxa contained five or fewer genera, and 40 % just one genus (Figure S6). Hence, these results could change when more species-level trait information becomes available. Our findings show that the choice of aggregation method may matter less when phylogenetic structure is fairly even and traits are less variable, but the $stepwise\_agg_{median}$ method tended to produce the widest range in affinities, especially with high taxonomic unevenness and high trait variability.

Taxonomic hierarchy in the trait datasets used for comparisons with assigned traits at family level

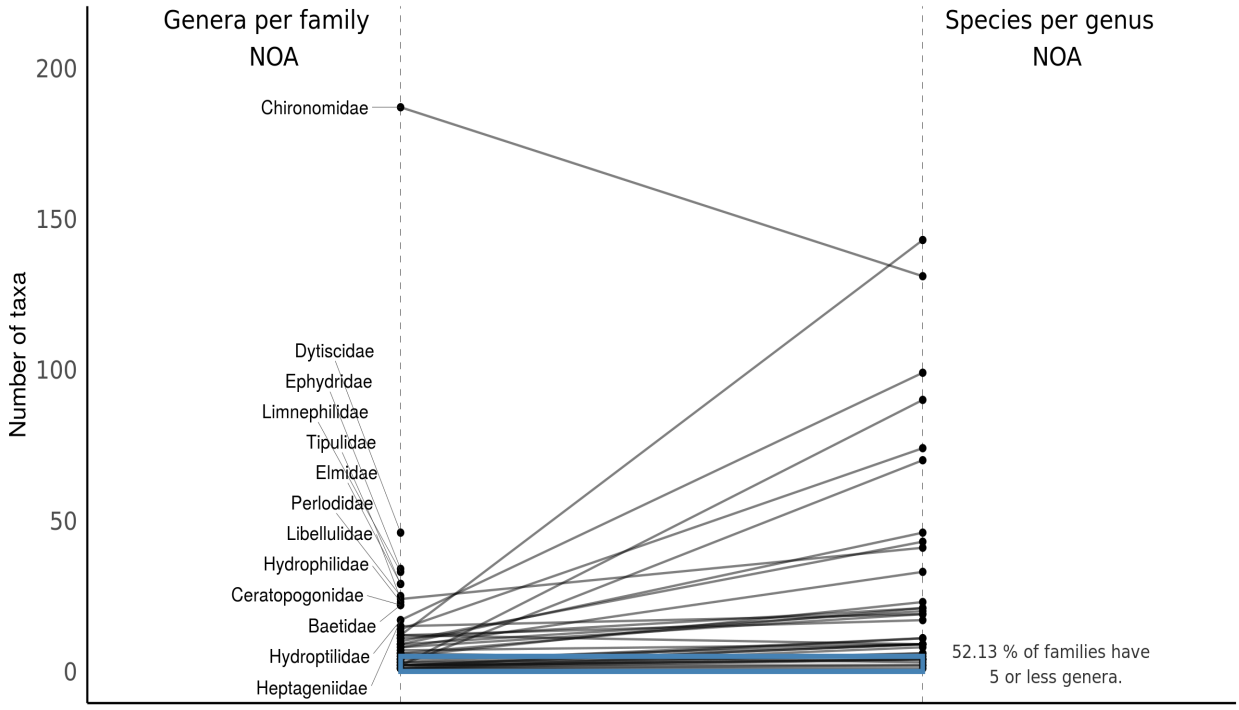

Figure S5: Number of genera per family and species per genus for those families of the North American trait dataset that have been compared with assigned traits at family level. For better visual display only families with more than 15 genera are displayed.

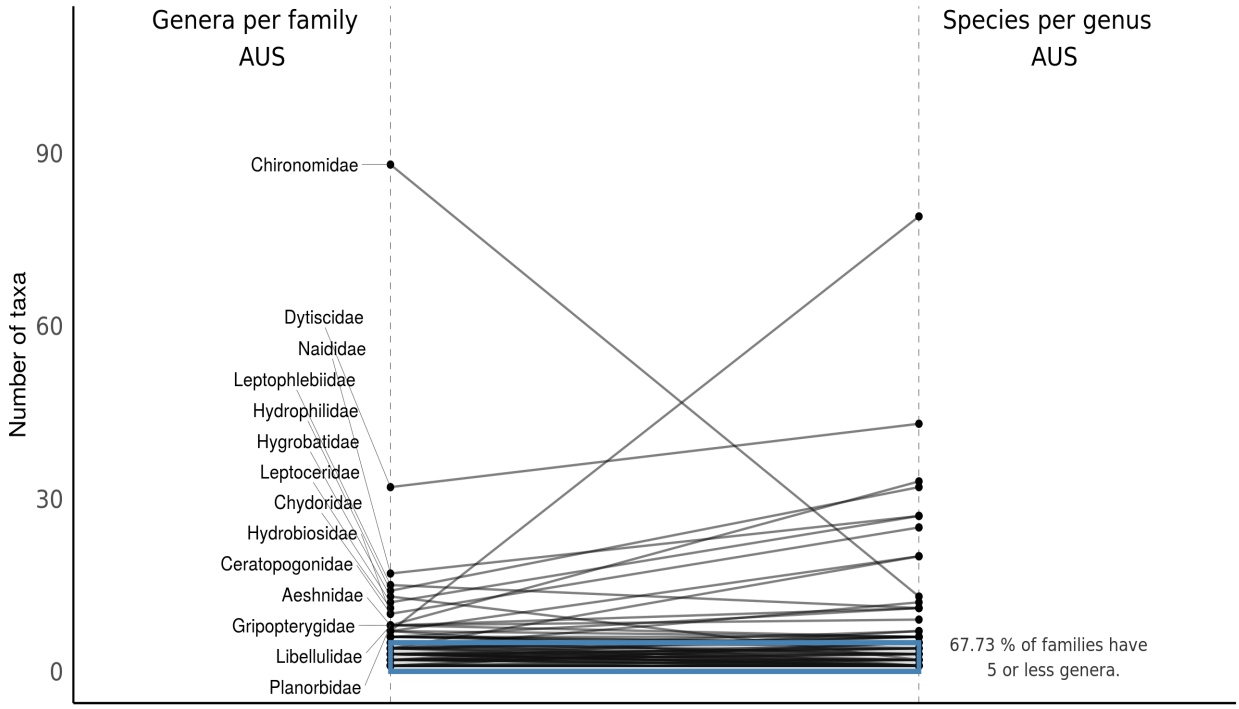

Figure S6: Number of genera per family and species per genus for the Australian trait dataset. For better visual display only families with more than 7 genera are displayed.

Table S2: Mean, median and standard deviation of the affinities of traits that were responsive to the salinity gradient in the original study but not in the re-analysis using the harmonised European trait dataset.

Type	Trait	Mean	Median	SD	Responsive?
Stepw_median	Shredder	0.20	0.14	0.25	No
Stepw_mean	Shredder	0.18	0.12	0.22	No
Direct_median	Shredder	0.21	0.14	0.25	No
Direct_mean	Shredder	0.19	0.14	0.22	No
Weighted	Shredder	0.19	0.14	0.22	No
Harmonised; not_aggregated	Shredder	0.18	0.12	0.24	No
Original	Shredder	0.25	0.14	0.32	Yes
Stepw_median	Gills	0.30	0.27	0.32	Yes
Stepw_mean	Gills	0.29	0.22	0.32	Yes
Direct_median	Gills	0.30	0.30	0.32	Yes
Direct_mean	Gills	0.30	0.30	0.32	Yes
Weighted	Gills	0.30	0.30	0.32	Yes
Harmonised; not_aggregated	Gills	0.30	0.25	0.32	No
Original	Gills	0.28	0.00	0.33	Yes
Stepw_median	Short life cycle	0.64	0.75	0.39	No
Stepw_mean	Short life cycle	0.64	0.79	0.39	No
Direct_median	Short life cycle	0.67	0.75	0.37	Yes
Direct_mean	Short life cycle	0.67	0.79	0.38	Yes
Weighted	Short life cycle	0.67	0.79	0.38	Yes
Harmonised; not_aggregated	Short life cycle	0.64	0.75	0.40	Yes
Original	Short life cycle	0.64	0.75	0.40	Yes
Stepw_median	Long life cylce	0.36	0.25	0.39	No
Stepw_mean	Long life cylce	0.36	0.21	0.39	No
Direct_median	Long life cylce	0.33	0.25	0.37	Yes
Direct_mean	Long life cylce	0.33	0.21	0.38	Yes
Weighted	Long life cylce	0.33	0.21	0.38	Yes
Harmonised; not_aggregated	Long life cylce	0.36	0.25	0.40	Yes
Original	Long life cylce	0.36	0.25	0.40	Yes

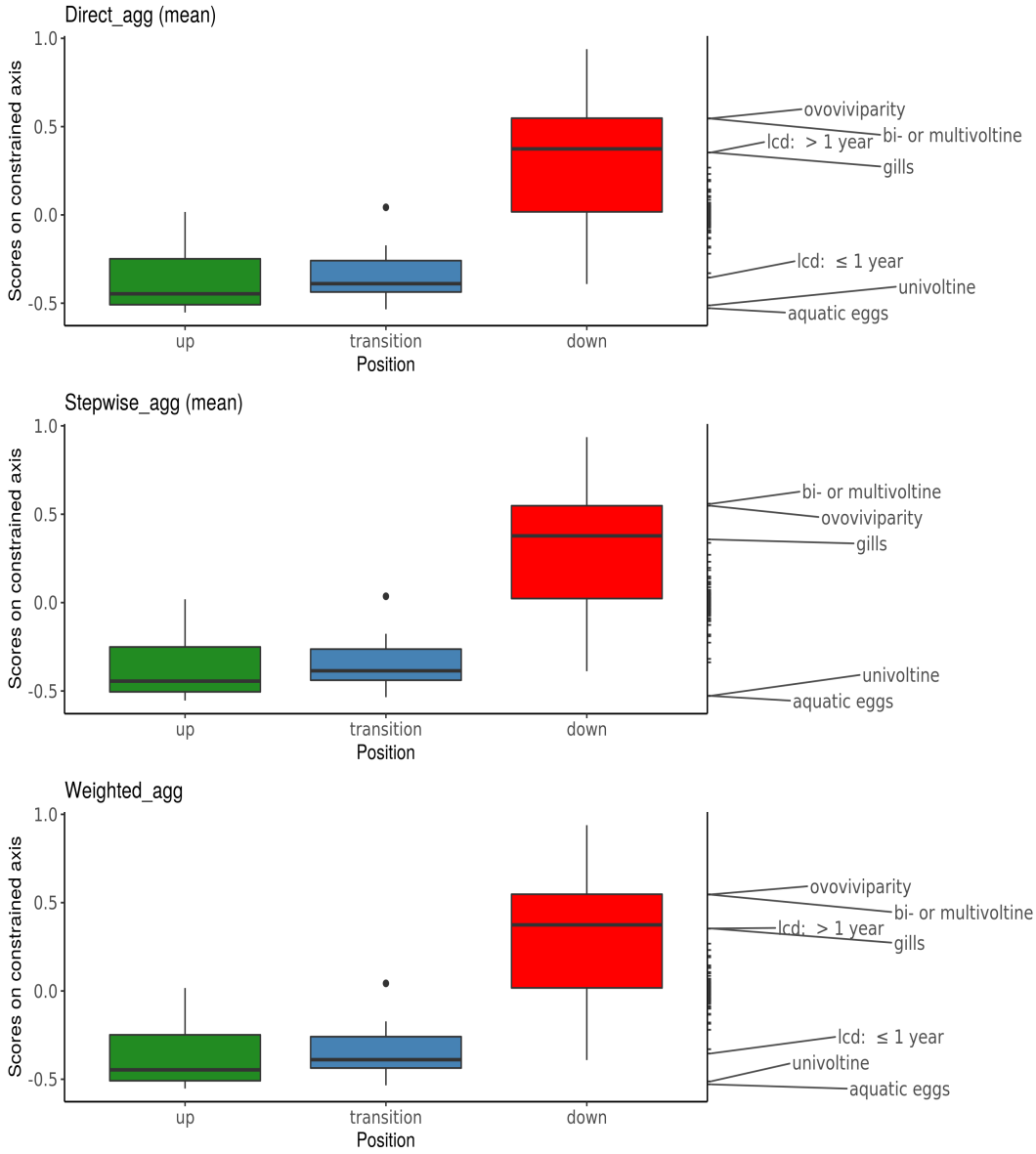

Figure S7: RDA of traits constrained by electric conductivity for the data aggregated with *direct_agg* $_{mean}$, *stepwise_agg* $_{mean}$, and *weighted_agg*. Shown are boxplots of the site scores along the conductivity axis. The rug on the right side of each plot indicates species scores of the traits on the conductivity axis. For more details see Figure 6. Abbreviations: lcd, life cycle duration; nr.cy, potential number of cycles per year.

Fourth corner analysis

We selected the ecologically relevant traits identified by Szöcs et al. (2014) that increased (shredder, life cycle duration > 1 year, gills, bi/multivoltine, ovoviviparity) or decreased (univoltine, eggs in clutches/cemented or fixed, life cycle duration <= 1 year) with increasing salinity and used these in the fourth corner analysis. The selected traits have also reacted to salinity in other studies (but see the discussion in Szöcs et al. (2014)). For some sites, environmental and abundance data have been sampled multiple times per year. Szöcs et al. (2014) used the full data and did not consider repeated sampling. We applied two analysis: one where we did not consider multiple samplings. And second where we selected, for the sites with multiple samplings, only those sampling events with the highest macroinvertebrate abundances per year and accordingly only used the conductivity data for these sampling events (this was not done for the RDA approach).Fourth corner analysis was carried out in R with the *fourthcorner()* function from the ade4 package. With the fourth corner analysis fewer trait-conductivity relationships than with the RDA approach were detected. With the original dataset significant relationships to conductivity were found for the traits shredder and ovoviviparity. In three out of the six aggregated and harmonised datasets ($direct\_agg_{mean}$, *weighted_agg*, harmonised (not aggregated)) four to five significant relationships (for the traits shredder, univoltinism, bi/multivoltine, ovoviviparity, and aquatic oviposition) were detected. In the remaining three datasets ($stepwise\_agg_{median}$, $stepwise\_agg_{mean}$, and $direct\_agg_{median}$) two significant relationships for the traits univoltinism and bi/multivoltinism were detected. Results were only slightly different when not considering multiple samplings (original dataset: one significant trait-conductivity relationship, others three to four). For the significant trait-conductivity relationships, the classification into characteristic traits for downstream and upstream sites was consistent with the results of the RDA analysis.

Effect of harmonisation and trait aggregation on functional diversity metrics

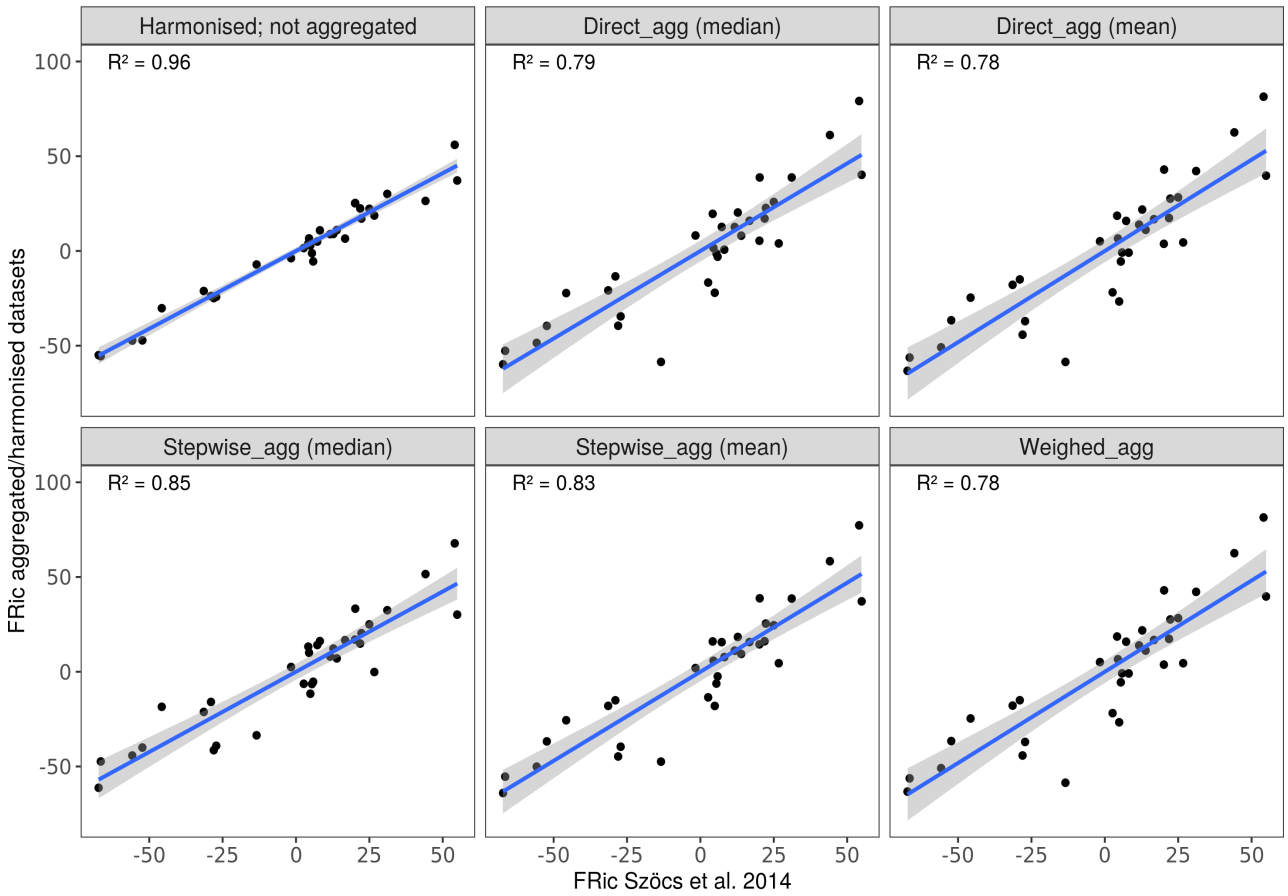

Figure S8: Fitted regressions between residuals of FRic with traits used in the original study and with harmonised and aggregated traits. Dots refer to each site-year combination. Sites have been sampled maximum three times in three years. For further details please refer to sections "Effects of harmonisation and trait aggregation on inferences regarding trait-environment relationships" and the original study.

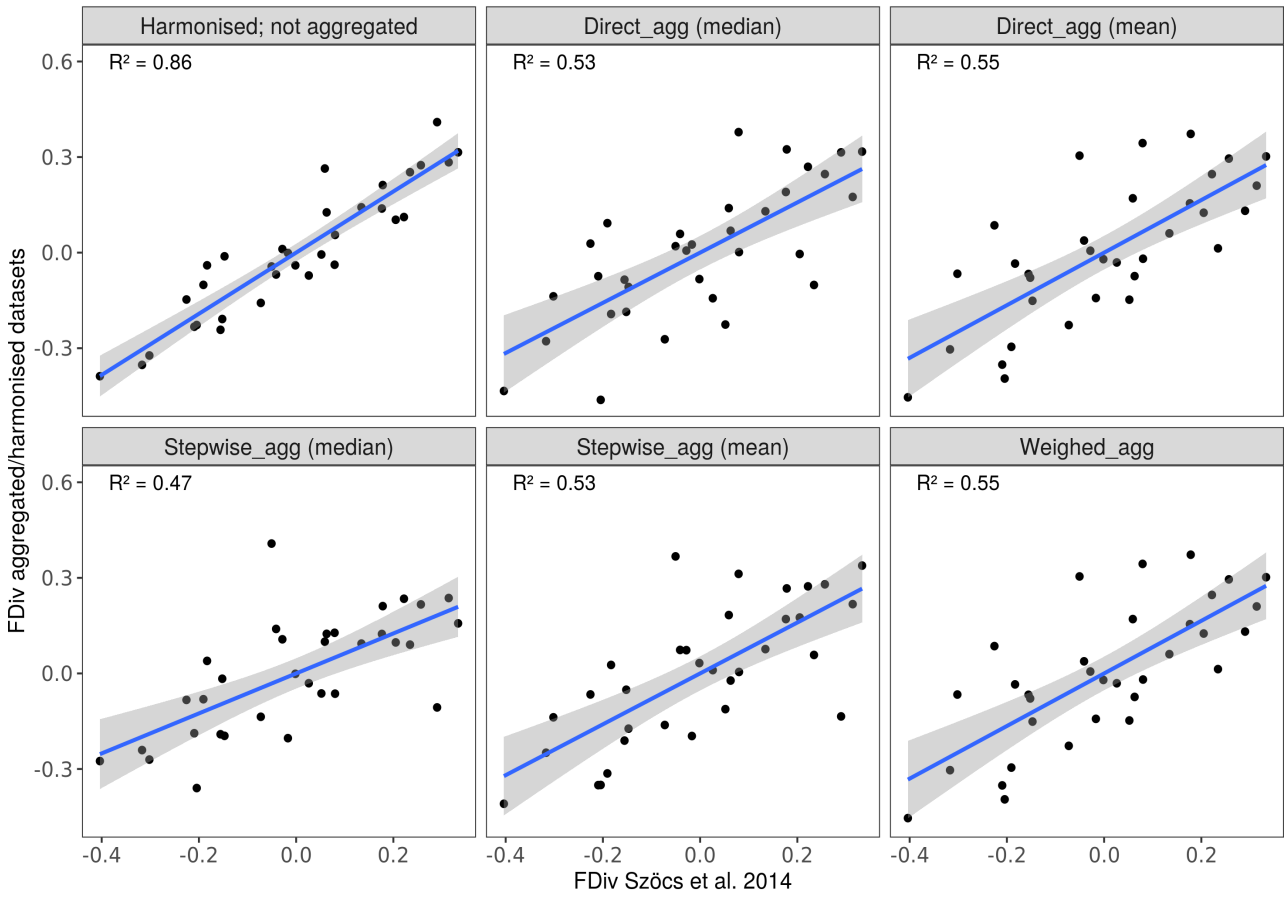

Figure S9: Fitted regressions between residuals of FDiv with traits used in the original study and with harmonised and aggregated traits. Dots refer to each site-year combination site-year combination. Sites have been sampled maximum three times in three years. Some sites were sampled multiple times within a year. In that case the highest sampling event with the highest total invertebrate abundance was taken (this explains the discrepancy to the original study, that did not consider multiple samplings). For further details please refer to sections "Effects of harmonisation and trait aggregation on inferences regarding trait-environment relationships" and the original study.

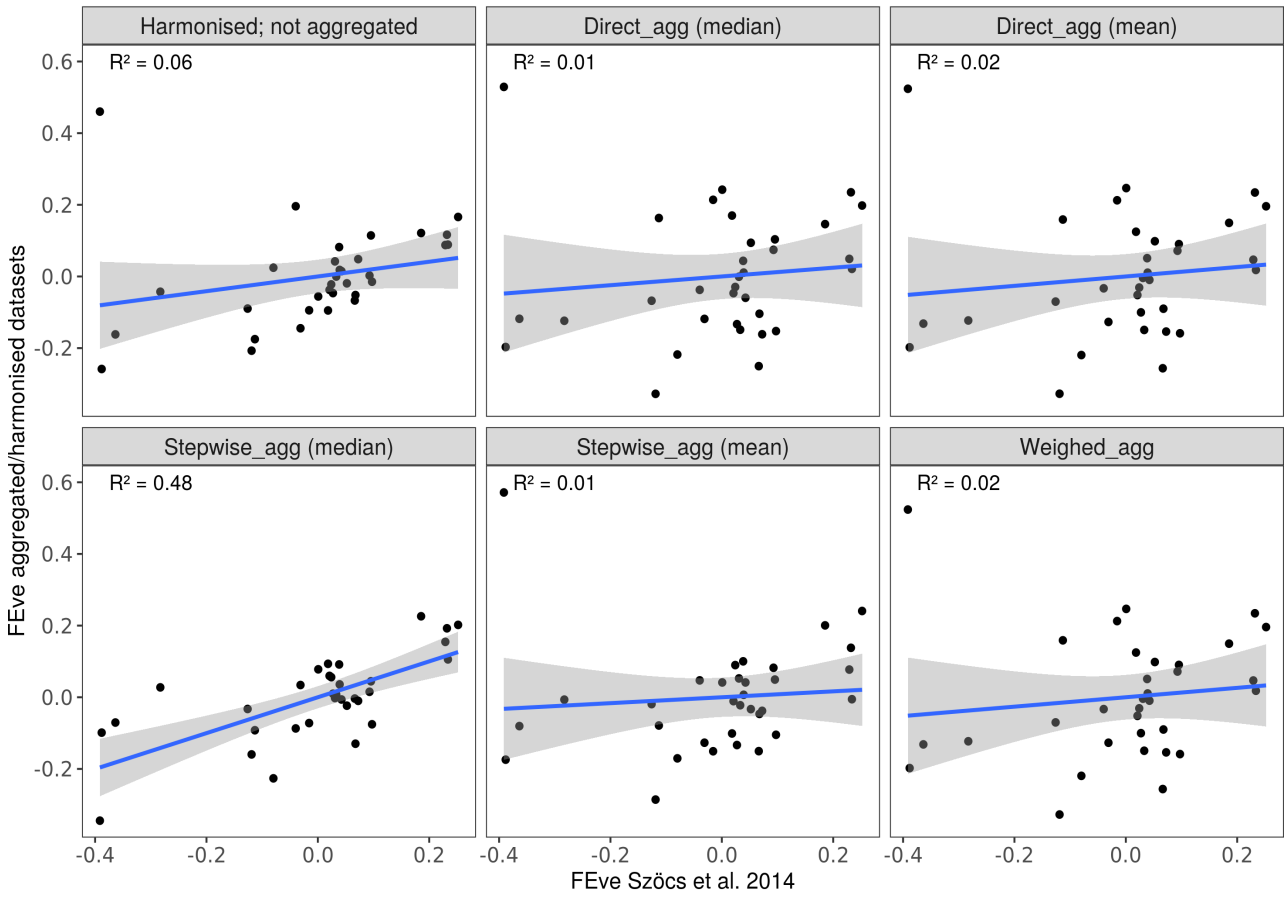

Figure S10: Fitted regressions between residuals of FEve with traits used in the original study and with harmonised and aggregated traits. Dots refer to each site-year combination site-year combination. Sites have been sampled maximum three times in three years. Some sites were sampled multiple times within a year. In that case the highest sampling event with the highest total invertebrate abundance was taken (this explains the discrepancy to the original study, that did not consider multiple samplings). For further details please refer to sections "Effects of harmonisation and trait aggregation on inferences regarding trait-environment relationships" and the original study.