Frequent origins of traumatic insemination involve convergent shifts in sperm and genital morphology

Jeremias N. Brand[1*], Gudrun Viktorin[1], R. Axel W. Wiberg[1], Christian Beisel[2], Luke J. Harmon[3] and Lukas Schärer[1]

[1] University of Basel, Department of Environmental Sciences, Zoological Institute, Vesalgasse 1, 4051 Basel, Switzerland

[2] Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland

[3] Department of Biological Sciences, University of Idaho, Moscow, USA

Short title:

Frequent origins of traumatic insemination

*Corresponding Author

University of Basel,

Department of Environmental Sciences, Zoological Institute,

Vesalgasse 1, 4051 Basel, Switzerland

Email: jeremias.br@gmail.com

Abbreviations

PC: principal component

pPCA: phylogenetically corrected principal component analysis

Abstract

Traumatic insemination is a mating behaviour during which the (sperm) donor uses a traumatic intromittent organ to inject an ejaculate through the epidermis of the (sperm) recipient, thereby frequently circumventing the female genitalia. It likely evolves due to sexual selection and sexual conflict, since it allows the donor to bypass pre- and postcopulatory choice and resistance mechanisms of the recipient. Several striking cases of traumatic insemination have been studied to date, but the frequency of its evolution, the intermediate stages via which it originates, and the morphological changes that such shifts involve remain poorly understood. Based on observations of reproductive traits in 145 species of the free-living flatworm genus *Macrostomum*—in combination with a robust phylogenomic phylogeny—we present comparative work on the evolution of traumatic insemination. We identify at least nine, but up to 13, independent origins of traumatic insemination, which involve convergent shifts in both sperm design and male and female genital morphology. We further find that traumatic insemination may originate via internal wounding. Finally, we highlight the large diversity of reproductive traits across the genus and show evidence for male-female coevolution. Our findings indicate that sexual selection and sexual conflict repeatably favour the evolution of behaviours—like traumatic insemination—that allow donors to bypass control mechanisms of recipients, leading to predictable shifts in both male and female reproductive traits.

Keywords: traumatic mating, hypodermic insemination, copulatory wounding, phylogenetics, evolution, female genitalia, correlated evolution, parallel evolution, sexually antagonistic coevolution

Introduction

The sexes frequently show a difference in mating propensity because male fertility (i.e. fertilized egg production) is often limited by the number of matings a male can achieve, while female fertility is often limited by the amount of resources a female can invest into eggs and offspring [1–3]. The resulting conflict over mating rate has far-reaching consequences, often resulting in "Darwinian sex roles" with choosy females and eager males [4]. Choice may be beneficial for females, since it allows them to preferentially reproduce with males based on genetic compatibility, genetic quality [5] and/or direct benefits (e.g. nuptial gifts [6]). Indeed, evidence for female choice is abundant and there are many species where females mate multiply, suggesting that polyandry may indeed result in such benefits [7]. However, females may also simply mate multiply as a result of male harassment, and while multiple mating could be costly to females, resisting male harassment might be even costlier [7, 8]. Costly harassment is expected to arise frequently, since female choice necessarily goes against the rejected males' interests [9], potentially leading to sexually antagonistic coevolution between male persistence and female resistance traits [8, 10, 11].

In polyandrous species, sexual selection and sexual conflict will continue after copulation through intricate interactions of the female genital tract with the male intromittent organs and the received ejaculate [12–14]. Female genitalia might exert postcopulatory control mechanisms through differential sperm storage, sperm ejection or sperm digestion, thus applying a selective filter on male genital and ejaculate traits. In analogy to the precopulatory conflict, it is then possible for traits in males to arise that attempt to bypass or influence the female choice and resistance mechanisms, again resulting in sexually antagonistic coevolution.

Sexually antagonistic coevolution can drive the emergence of male traits that inflict considerable harm on females [11, 15, 16]. A striking example that implicates such harm is traumatic insemination, which occurs in some internally fertilising species and involves the infliction of a wound to the female's integument through which the male then transfers its ejaculate [17]. Since traumatic insemination frequently occurs in both gonochoristic (separate-sexed) and hermaphroditic species

[17], we in the following use the more general terms (sperm) donor and (sperm) recipient to refer to the two sexual roles, with no loss of generality [18].

Although traumatic insemination often results in costs to the recipients [15, 17, 19–22], it has evolved repeatedly across various animal groups [17]. While natural selection might play a role in its evolution in some taxa—especially the endoparasitic Strepsiptera [23, 24]—it likely often evolves because it allows the donor to bypass pre- and postcopulatory choice and resistance mechanisms of the recipient. Specifically, traumatic insemination can enable donors to force copulation and thus minimise the control that the recipient could otherwise exert over mating [15]. Traumatic insemination usually also allows the donor to bypass the recipient's genitalia, by depositing sperm either closer to the site of fertilisation [24, 25] or even directly within the relevant tissue [15, 26], thus likely reducing the recipient's ability to control the fate of the received ejaculate [12, 17]. In this view, traumatic insemination allows the donor to bypass the influence of the recipient's sexually antagonistic choice and resistance mechanisms, temporarily gaining an advantage in the coevolutionary chase.

However, since conflicts persist under traumatic insemination, we expect selection to then act on traits that allow the recipient to regain control over mating and/or the fate of the received ejaculate. For example, some species of bed bugs have evolved what could be considered a secondary vagina, a structure that has been shown to reduce the costs incurred due to traumatic insemination [19, 27]. But even without the emergence of new organs, the recipients could evolve behavioural or physiological responses to avoid traumatic insemination (such as parrying strikes during penis fencing in polyclad flatworms [28]) or to manipulate and control the hypodermically received ejaculate (e.g. similar to sperm digestion in copulating species [29–31]).

Besides bypassing recipient choice and resistance mechanisms, traumatic insemination could also evolve due to sperm competition, since in many internally fertilising species sperm of unrelated donors compete within the female genital tract for fertilisation of the recipient's eggs [32]. In this context, traumatic insemination might allow donors to bias sperm competition in their favour and prevent competing donors from removing their previously donated sperm, potentially resulting in

paternity benefits [17]. Indeed, traumatic insemination seems to affect sperm competition in a family of spiders, where sperm precedence is biased towards the first male in a species with traumatic insemination, while it is biased towards the second male in its non-traumatically mating relatives [17, 23, 33]. In contrast, traumatic insemination is associated with last male precedence in one species of bed bug [26], so the effects of it on sperm competition might depend on a species' morphology and ecology.

Interestingly, since traumatic insemination likely affects not only how sperm interact, e.g. concerning sperm mixing, sperm removal and sperm density, but also the degree to which they are accessible to recipient control, it could influence the nature of both sperm competition and cryptic female choice. Cryptic female choice, sperm removal, and sperm displacement can introduce skews in the distribution of sperm competing in the female storage organ, which may reduce the intensity of sperm competition [34–38]. These same postcopulatory mechanisms can also favour sperm designs that increase the chance of being retained in storage. This could include adaptations to better anchor or position sperm, making them able to resist removal or displacement by either rival donors or the recipient [39]. Traumatic insemination could lead to less removal and more mixing, resulting in a more fair-raffle like sperm competition [40, 41], in which paternity success is more determined by the relative number of sperm donated, leading to strong selection on sperm number. Since sperm number is generally assumed to trade-off with sperm size [42–45], it is expected that traumatic insemination is associated with smaller (and potentially also more numerous) sperm [40, 41].

Traumatic insemination might evolve more frequently in hermaphrodites due to sexual conflict over mating roles [12, 18, 46–48]. In general, and in analogy to the situation outlined above for gonochorists [1], a hermaphrodite that already carries enough received sperm to fertilise its own eggs might gain little from additional matings as a recipient, while it might still gain additional fertilisations by acting as a donor [12]. It is thus likely that, on average, individual hermaphrodites show a preference for sperm donation [12, 18, 46–48] and this prediction is supported in several laboratory studies [47, 49, 50]. Traumatic insemination then potentially allows unilateral enforcement

of donation while avoiding receipt. Additionally, harmful traits are expected to evolve more readily in simultaneous hermaphrodites because fitness costs incurred by a recipient may be partially compensated by fitness benefits from the same individual acting as a donor. A hermaphrodite may thus also engage in matings that involve harmful effects, as long as the net fitness benefits of mating are positive [46, 51]. Therefore, hermaphrodites could engage in traumatic insemination even if this results in considerable harm to the recipient.

Indeed, traumatic insemination appears to arise more commonly in simultaneous hermaphrodites compared to gonochorists, since they comprise 11 out of the 23 well-supported independent origins listed in [17], which is a lot considering that hermaphrodites make up ~6% of animals [52]. Hermaphroditic animals are thus ideal study organisms for investigations of traumatic insemination, since—while it has been studied in some charismatic systems [15, 25, 28, 53–56]—we know little about how frequently it evolves or via which intermediate states a copulating species can transition to traumatic insemination [17, 20, 22].

Here we present comparative work on the evolution of traumatic insemination across the genus *Macrostomum*, a species-rich taxon of hermaphroditic free-living flatworms. In *Macrostomum*, traumatic insemination is called hypodermic insemination, since in several species the donor uses a needle-like stylet (**Fig 1**) to inject sperm through the epidermis of the mating partner and sperm then move through the body of the recipient to the site of fertilisation [41, 57, 58]. Injected sperm can often be observed inside the parenchymal tissues of these highly transparent animals [41, 57–59], making it feasible to screen a large number of species for the convergent evolution of hypodermic insemination. And while we here present evidence that not all traumatically mating *Macrostomum* species may inject sperm through the external epidermis, we nevertheless use the term hypodermic insemination for consistency with the earlier literature.

The genus consists of two phylogenetically well-separated clades, with one clade currently thought to only mate through hypodermic insemination (“hypodermic clade”, referred to as Clade 1 in [41]) and a second clade primarily containing reciprocally mating species (“reciprocal clade”, referred to

as Clade 2 in [41]). Within the reciprocal clade, hypodermic insemination has previously been shown to have convergently evolved in *M. hystrix* [41]. Reciprocally copulating *Macrostomum* species engage in a copulatory handshake, where two individuals insert their often relatively blunt stylet (Fig 1) via their partner's female genital opening into the female sperm storage organ, the female antrum (further only called antrum), so that both can donate and receive sperm in the same mating [60]. Many reciprocally copulating species then perform a postcopulatory behaviour, where worms place their mouth over their own female genital opening and suck [41, 61], presumably in an attempt to remove components of the received ejaculate from the antrum (called the suck behaviour, [39, 60]). This removal of ejaculate could target manipulative seminal fluids, since the ejaculate of the main *Macrostomum* model species, *M. lignano*, contains substances that influence the mating partners behaviour, including the propensity to perform the suck behaviour [62, 63]. The suck behaviour could also be aimed at reducing the number of stored sperm (e.g. to lower the risk of polyspermy), constitute a form of cryptic female choice (e.g. to favour donors of higher quality), and/or represent a resistance trait in sexual conflict over mating roles (i.e. to undo unwanted sperm receipt) [41].

If the suck behaviour is a recipient resistance trait, we might expect the evolution of donor persistence traits, potentially leading to antagonistic coevolution [8]. Indeed, the sperm of reciprocally copulating species generally have a thin anterior feeler and two stiff lateral bristles that could represent such persistence traits (**Fig 1**), serving to anchor the sperm in the antrum to prevent removal during the suck behaviour [39, 41]. In contrast, sperm of species with hypodermic insemination (i.e. the hypodermic clade and *M. hystrix*) lack lateral bristles and have a simplified morphology, presumably because they no longer need to resist the suck behaviour [39, 41] which has so far never been observed in species with hypodermic insemination. These sperm may instead be adapted to efficiently move through the partner's tissues (**Fig 1**), and one such adaptation could include a size reduction, as outlined above. Moreover, while species with reciprocal copulation have an antrum with a thickened epithelium, the species with hypodermic insemination have a simple antrum, presumably because it no longer interacts with the donor's stylet and sperm, and instead is used for egg-laying only [41].

Based on these findings, the observed adaptations to reciprocal copulation and hypodermic insemination have been described as the reciprocal and hypodermic mating syndrome, respectively, since they each constitute specific combinations of morphological (sperm, stylet and antrum) and behavioural traits [41].

If hypodermic insemination evolved as a resolution of sexual conflict over the mating roles, then we would expect it to occur frequently, but it is currently unclear whether hypodermic insemination has convergently arisen more than once within the reciprocal clade. It is also unclear if the transitions are reversible or if the emergence of hypodermic insemination alters the coevolutionary dynamics between donor and recipient, so that they cannot readily revert back to reciprocal copulation. Little molecular data is currently available for most species in the genus *Macrostomum* and the description of important structures, such as the sperm morphology and aspects of the mating behaviours, are often neglected in the taxonomic literature.

Here we collected molecular and morphological information on 145 *Macrostomum* species to identify additional independent origins of hypodermic insemination, to highlight species that represent possible transitional states in the evolution of hypodermic insemination, and to quantitatively assess convergent changes in both sperm design and in male and female genital morphology that accompany its evolution. Using ancestral state reconstruction, we further ask whether species can revert back to reciprocal copulation once hypodermic insemination has arisen. Moreover, if hypodermic insemination evolves in the context of sexually antagonistic coevolution, then signatures of male-female coevolution in the form of a diversity of resistance and persistence traits should be evident across the genus. Indeed it is predicted that donors might carry multiple persistence traits even if recipients resistance renders them ineffective [64]. We thus test for covariation between male and female genital morphology and survey the genus for novel resistance and persistence traits.

Results

Species collected and phylogenetics

We included 145 *Macrostomum* species in our analysis, and established a phylogeny inferred from a combination of newly generated transcriptome data (98 species, see Material and Methods) and a partial *28S rRNA* gene sequence (permitting to add 47 additional species). Except for 14 species, we collected and systematically documented all species ourselves. Based on an integrative approach—using detailed documentation of *in vivo* morphology and *28S rRNA* sequences—we identified 51 species as previously described and a striking 94 species as likely new to science (see Material and Methods for details on species delimitation; **Tab 1** for a summary of all collected species; **Tab S1** for the sample sizes of morphological measures across species; and also SI Taxonomy). Since our focus here was not on taxonomy, we delimited new species as operational taxonomic units. We deposited extensive image, video [65], geographic and molecular data for all specimens collected and their operational species assignment (see **Tab S2**), facilitating their future description (see http://macrostomorpha.info and [65]).

We inferred phylogenies based on two protein alignments generated from transcriptome data (see Material and Methods), one with many genes while having lower occupancy (L alignment with 8218 genes) and one with high occupancy but fewer genes (H alignment with 385 genes, **Tab 2**). For both alignments, we calculated phylogenies using maximum likelihood with IQ-TREE [66] (referred to as L-IQ-TREE and H-IQ-TREE) and summary methods with ASTRAL III [67] (L-ASTRAL and H-ASTRAL). We applied summary methods to account for incomplete lineage sorting and the gene tree – species tree conflict it can cause. Second, to include species for which we lacked transcriptomes, we added information from *28S rRNA* to the H alignment to construct a combined (C) maximum likelihood phylogeny (C-IQ-TREE, see Material and Methods). Third, we performed two types of Bayesian analyses on the H alignment, one that like the maximum-likelihood analysis operates on partitioned amino acids [68] (H-ExaBayes) and a second performed on the unpartitioned DNA coding

sequence alignment to better account for rate heterogeneity across sites [69] (H-PhyloBayes, see Material and Methods).

Across these phylogenies, all species represented by more than one transcriptome (i.e. 15 x 2, 3 x 3, and 1 x 9 transcriptomes) were monophyletic, with the notable exception of *M. lignano* (**Fig S1**). One of the *M. lignano* transcriptomes was from the DV1 inbred line from the type locality near Bibione, Italy [70] and the other from an outbred population from the Sithonia Peninsula, Greece [61, 71]. *M. lignano* was monophyletic in the L-IQ-TREE, H-PhyloBayes, and both ASTRAL phylogenies, but the Greek population was sister to *M. janickei* in the H-IQ-TREE and H-ExaBayes phylogenies, although node support for this split was low for H-IQ-TREE (**Fig S1**), and removal of cDNA synthesis primer sequences from the alignment rendered *M. lignano* monophyletic in H-IQ-TREE (see SI Primer removal and Material and Methods). A closer investigation of the morphology of the Greek vs. Italian population also revealed that the former had a considerably larger stylet (96.7 vs. 60.7 µm) and longer sperm (76.5 vs. 62.7 µm). We still consider *M. lignano* a proper species, but a closer comparison of the Greek and Italian populations would be interesting.

Next, we discuss other agreements and discrepancies between the different inferred phylogenies. The grouping of the major clades was mostly consistent across all phylogenetic approaches (**Fig 2** and **Fig S1**). All phylogenies recovered six large species groups (further called the hypodermic, spirale, lignano, finlandense, tuba, and tanganyika clade), two smaller species groups (the minutum and hamatum clade), and two consistent species pairs (*M*. sp. 45 + 46 and *M*. sp. 4 + 89). However, the backbone, and the positions of some species with long branches (*M*. sp. 37, *M*. sp. 39, *M*. sp. 90 and *M. curvituba)*, differed depending on the alignment and method used. Despite these discrepancies, the Robinson-Foulds distances [72] between the phylogenies were low (**Tab S3**), indicating good agreement between methods.

In all phylogenies, the hypodermic clade was deeply split from the reciprocal clade, but with large phylogenetic distances also appearing within the hypodermic clade (see also below). Also consistent was the grouping of the tanganyika, tuba and finlandense clades (**Fig 2** and **Fig S1**). The hamatum

and minutum clades, together with *M.* sp. 4 + 89, were always the closest relatives to the former three clades. However, the exact relatedness patterns were uncertain. In the L-IQ-TREE, H-IQ-TREE and H-ExaBayes phylogenies, *M.* sp. 4 + 89 were the closest relatives to these three clades, followed by the hamatum and minutum clades. In contrast, in the ASTRAL and the H-PhyloBayes phylogenies, the hamatum clade was more closely related to the minutum clade (with the latter nested within the former in case of H-PhyloBayes) and both were sister to the grouping of tanganyika, tuba and finlandense. Moreover, the exact branching order at the base of the reciprocal clade was not clearly resolved. The spirale clade split off first in the H-IQ-TREE and H-PhyloBayes phylogenies, while the lignano clade split off first in the other phylogenies. Consistent with the conflict between methods, the quartet support from the ASTRAL analysis indicated gene tree – species tree conflict at most nodes in the phylogeny's backbone (**Fig S2**). These internal nodes were separated by short branches suggestive of rapid speciation events, such as during adaptive radiation [73], where substantial incomplete lineage sorting is expected. This pattern is also consistent with ancient hybridisation, which is a distinct possibility, since there is evidence for hybridisation within the genus under laboratory conditions [74].

The topology of C-IQ-TREE was identical to H-IQ-TREE when we removed all the species added based on *28S rRNA* (**Fig 3**) and thus adding these species did not negatively influence the overall topology of the tree. Node support in C-IQ-TREE was somewhat lower, as could be expected given the placement of the additional species based solely on *28S rRNA* sequences. Nevertheless, the added inferential power obtained from a ~50% increase in species representation is highly worthwhile, and we therefore focus on results based on this combined phylogeny in the main text.

As already mentioned above, many species in the hypodermic clade were highly molecularly diverged, even though they are difficult to distinguish morphologically. Most of these species had stylets that consisted of a short proximal funnel that tapered to a curved and drawn-out asymmetrical needle-like thickening (**Fig 3**A, top). It was possible to distinguish some species based on general habitus. These differences are reflected by the four deeply split clades containing, *M. rubrocinctum*,

M. hystricinum, *M. gabriellae*, and *M. pusillum*, respectively (although the latter two clades were also quite similar in habitus). We have indicated substantial morphological similarity by appending a letter to the species name (e.g. *M. hystricinum a*, *c*, *d*). Investigations of species within the hypodermic clade without support from molecular data thus require considerable caution. Moreover, given these striking morphological similarities, it may often not be clear to which of these species the name-bearing type specimens belong to [41], so the species names in this clade should be considered tentative. Eventually, one may need to name these species afresh, applying more extensive molecular species delimitation, and either suppress the original names of species without detailed enough morphological descriptions and/or lacking adequate type material, or to define neotypes, including molecular voucher specimens [75]. While such a detailed taxonomical revision of the genus *Macrostomum* goes well beyond the scope of the current study, we feel that readers need to be aware of these caveats (see also SI Taxonomy).

Morphological diversity

Before we report on our formal analysis of convergent evolution and coevolution, we highlight several exciting morphologies, which considerably expand the known morphospace of *Macrostomum* (**Fig 3**). As mentioned, the hypodermic clade shows little variation in terms of stylet and sperm morphology. However, the stylet of *M.* sp. 93 differed clearly from that stereotypical form, by having a small proximal funnel extending into a straight and obliquely-cut tube (**Fig 3**A). This shape is similar to the stylets of several species in the reciprocal clade, namely *M. shenda*, *M.* sp. 34 and *M.* sp. 64 (as well as *M. orthostylum*, for which we have no phylogenetic placement). Since we observed hypodermic received sperm in both *M.* sp. 93 and *M.* sp. 64 (**Fig 3**), this shape appears adapted for hypodermic insemination. And while we did not observe hypodermic received sperm in *M. orthostylum* and *M.* sp. 34 (nor was such sperm reported for *M. shenda* by [76]), these species are nevertheless likely to also mate through hypodermic insemination, not least since stylets with similar shapes are also used for hypodermic insemination in related macrostomid flatworms [77].

In contrast to the largely canalized stylet and sperm of the hypodermic clade, we found remarkable variation in these structures within the reciprocal clade (**Fig 3**). For example, we documented five species in the spirale clade that had stylets with lateral protrusions close to the distal opening. The protrusion is shaped like a rod or spike in *M. evelinae*, *M.* sp. 29 and *M.* sp. 42, while it consists of two thin filaments in *M.* sp. 30 and *M.* sp. 43 (see also SI Taxonomy). These highly modified stylets may have coevolved with the complex antra in these species, which have two separate chambers connected via a ciliated muscular sphincter. The sperm of these species is also remarkable because they carry only a single bristle instead of the typical two. The sperm of *M*. sp. 30 carry only a single curved bristle, while the bristle of *M*. *evelinae*, *M*. sp. 29, and *M*. sp. 42 is additionally modified, being thicker and appearing flexible, frequently curving back towards the sperm body. A second bristle might also be absent in the closely related *M*. sp. 13 (indicated with a shaded second bristle in **Fig 3**), but the available material currently does not allow an unambiguous assessment.

Sperm are also highly variable across the entire reciprocal clade. Particularly striking are the sperm modifications of *M*. sp. 82, which give the anterior part of the sperm a translucent appearance under phase contrast (discussed in more detail below). We also document extraordinarily long sperm in several species in the tanganyika clade. While it is not entirely clear which part of the sperm is modified here, it appears that they have very long sperm feelers, with the bristles thus being located unusually far posterior. Finally, we observed numerous species that either had reduced (length <3.2 µm) or no sperm bristles. Such reduction and losses repeatedly occur across the whole reciprocal clade and appear to coincide with changes in the stylet and antrum morphology (see our formal tests below). Finally, a striking modification of sperm design occurs in *M. distinguendum* (finlandense clade), which appears to lack sperm bristles, but instead carries novel club-shaped lateral appendages, which in light microscopic images bear no resemblance to the usual sperm bristles.

The antrum morphology was also highly variable within the reciprocal clade, with several species having complex female antra with multiple chambers. Particularly striking are the two female genital openings present in four quite distantly related taxa (see SI Taxonomy). A second opening is present

in *M. spiriger*, *M. gieysztori* (and its three close relatives, *M.* sp. 16, *M.* sp. 17 and *M.* sp. 18), *M. paradoxum*, and *M.* sp. 82, suggesting multiple independent origins across the genus (see SI Female opening). The observed variation in male and female genital morphology, and sperm design could be important for sexual selection, as we outline in the discussion.

Frequent origins of hypodermic insemination

We inferred the number of convergent transitions to hypodermic insemination using ancestral state reconstruction (ASR) of three reproductive traits: received sperm location, sperm bristle state, and antrum state (see SI Morphology and Tab 3). Because received sperm location involves the direct observation of sperm within the recipient's tissue, it provides direct evidence for hypodermic insemination. However, observation of injected sperm in field-collected specimens can be challenging, especially in species with low investment into sperm production, thus lowering the sample size. Since tests of correlated evolution revealed strong associations of hypodermic sperm with absent/reduced sperm bristles and a thin antrum (see the next section), we also performed ASR using these states as proxies for hypodermic insemination. Finally, we performed ASR on the inferred mating syndrome, which represents a synthesis of all available information (see Material and Methods and Tab 3). We performed ASR, with all traits scored as binary and, where appropriate, also as trinary, to test if hypodermic insemination could evolve via an intermediate state (see Material and Methods). All reconstructions indicated frequent origins of hypodermic insemination (Tab 4 and Fig S3). In all analyses with trinary states, an ordered transition model without gains once traits have been lost (ORD-Dollo) was preferred, and in all analyses with binary states, a model without gains (Dollo) was preferred. However, other models, including some permitting gains, also received at least some support (Tab 4). ASR of trinary states inferred frequent transitions to the intermediate state, which were driven by the ordered model's requirements to transition through it. These transitions were often placed along internal branches of the phylogeny, primarily within the finlandense clade, which contains several species with reduced or absent states and, nested within them, two species with

353 present states (*M*. sp. 12 and *M*. sp. 44, with received sperm in the antrum, long bristles, and assigned
354 to the reciprocal mating syndrome; Fig S3 A, C, F).
355 We estimated a lower bound for the number of transitions by requiring an origin of the derived state
356 to be separated by other such origins via nodes with a >95% posterior probability of having the
357 ancestral state. Applying this rule to traits scored as binary, we find nine transitions to hypodermic
358 received sperm, 17 losses/reductions of sperm bristles, 13 simplifications of the antrum, and 13
359 transitions to the hypodermic or intermediate mating syndrome (see red stars and numbers in Fig S3).
360 Moreover, these lower-bound estimates were slightly lower for trinary states. Finally, we also
361 performed these analyses for the H-IQ-TREE and H-ExaBayes phylogenies, and found qualitatively
362 very similar results to those of the C-IQ-TREE phylogeny reported above, albeit, since they contain
363 fewer species, showing somewhat lower numbers of transitions (Tab S4).

364 ## Correlated evolution

365 We performed tests of correlated evolution to ask if the numerous convergent changes in received
366 sperm location, sperm bristle state and antrum state are evolutionarily dependent. We found strong
367 support for correlated evolution of received sperm location with both sperm bristle state and antrum
368 state (Fig 4A+B). This supports previous findings that hypodermic insemination is associated with
369 changes in sperm design and antrum simplification [41]. Therefore, when observations of received
370 sperm are missing, both sperm bristle state and antrum state are likely good proxies for the mating
371 syndrome. We expand on the earlier analyses by also providing evidence for the correlated evolution
372 between the sperm bristle state and antrum state (Fig 4C), which was implied in [41], but not formally
373 tested. We find substantially stronger support for correlated evolution than [41] across the board, with
374 Bayes factors that are 7.7 and 7 times larger for the correlated evolution of sperm bristle state or
375 antrum state with the inferred mating syndrome, respectively. Moreover, the analyses were robust,
376 with respect to the phylogeny and the priors used (see SI Correlated evolution).

Hypodermic insemination and convergence in morphospace

Next, we used phylogenetically corrected principal component analysis (pPCA) to investigate if these convergent transitions to hypodermic insemination also coincide with changes in a greater variety of reproductive traits (see SI Morphology). The first two principal components, PC1 and PC2, together captured nearly half of the variation in the analysed reproductive traits (Fig 5), followed by additional principal components with relatively small contributions (Tab S5). Specifically, PC1 captured a change in stylet phenotype, with larger values indicating species with longer, more curved stylets, that are distally more symmetric and less sharp (Fig 5). Larger values of PC1 also indicated both longer sperm and bristles, and an increased probability for the sperm to carry a brush. Finally, high values of PC1 indicated a thickened antrum with a more pronounced cellular valve, and a more complex internal structure. In comparison, PC2 had a less clear interpretation, with high values indicating larger species with larger proximal and distal stylet openings.

Species in the hypodermic clade (stippled outlines) had similar values in PC1 and mainly differed in PC2. Interestingly, species from the reciprocal clade that we had assigned to the hypodermic mating syndrome (solid and left yellow) grouped closely with the species in the hypodermic clade, indicating striking convergence in morphospace concerning stylet, sperm and antrum morphology (see also Fig S4). PC1 further separated species based on the received sperm location, with hypodermic received sperm (right yellow) only found in species with low PC1 values, indicating that PC1 captures a morphology necessary for hypodermic insemination. Almost all species with reduced (triangles) or absent (circles) sperm bristles grouped closely together in PC1, with the notable exception of *M*. sp. 68 and *M*. sp. 82 (black arrowheads), which cluster together with other species that we assigned to the reciprocal mating syndrome. We observed sperm in the antrum of both species (i.e. in 2 of 7 specimens in *M*. sp. 68 and 16 of 21 specimens in *M*. sp. 82) and the antrum is similar in both, with a long muscular duct that performs a 90° turn towards the anterior before it enters a second chamber that is strongly muscular (Fig 6). Moreover, both species have a similar L-shaped stylet with

a blunt tip, which makes it unlikely that they mate through hypodermic insemination (see also Discussion).

Pathways to hypodermic insemination

While 117 of the 145 studied species could be assigned to either the hypodermic or reciprocal mating syndrome, some were not easily identified as performing either type of mating behaviour (Tab 3), which may either have been due to lack of detailed observations, or more interestingly, due to potentially transitional patterns that deviated from these syndromes. This included two species that we categorised as intermediate because we observed sperm both in the antrum and embedded inside the recipient's tissues (light green triangles in Fig 5). Both species (*M*. sp. 3 and *M*. sp. 101) have a sharp stylet and sperm with reduced bristles, fitting with the hypodermic mating syndrome. But they also have a thickened antrum wall, which indicates the reciprocal mating syndrome, and these species indeed have intermediate values of PC1. Multiple observations of hypodermic sperm in *M*. sp. 3 show them to be embedded deeply in the anterior wall of the antrum and more deeply in the tissue lateral to the body axis, extending up to the ovaries, and in the tail plate (Fig 7). And while in *M.* sp. 101 we did not observe sperm as deeply in the recipient's tissues, some were fully embedded in the cellular valve and just anterior to it, and thus close to the developing eggs (Fig 8). One explanation for these findings would be that during mating, the stylet of both species pierces the recipient's antrum wall and sperm is traumatically injected into the body internally. Unfortunately, no copulations were observed in mating observations of *M*. sp. 101, but this species has been seen performing the suck behaviour, as expected if ejaculate is, at least partially, deposited in the antrum (pers. comm. P. Singh). We lack mating observations for *M*. sp. 3 and further investigations of the mating behaviour and the antrum histology of both species would be highly desirable.

Three additional species (*M*. sp. 14, *M*. sp. 51 and *M*. sp. 89) were also difficult to classify because, although their morphology indicates hypodermic insemination, we clearly observed received sperm in the antrum. *M*. sp. 51 and *M*. sp. 89 grouped with the hypodermically mating species in PC1 (red

arrowheads in Fig 5), while we did not include *M*. sp. 14 in this analysis due to missing data for sperm bristle length. We found sperm within the antrum in only 1 of 4 specimens in *M*. sp. 51 and 1 of 12 specimens in *M*. sp. 89, and it is thus possible that sperm is hypodermically injected and may later enter the antrum when an egg passes through the cellular valve into the antrum before egg laying. But these species could also represent an intermediate state between the mating syndromes. We observed sperm in the antrum in 3 of 5 specimens in *M*. sp. 14 and it therefore seems less likely that sperm entering during the transition of the egg into the antrum is the cause of its presence here as well. Instead, sperm is probably deposited in the antrum by the mating partner during copulation. However, based on its general morphology, we predict that closer investigations of this species will reveal hypodermic received sperm in a similar location as found in the other intermediate species (*M*. sp. 3 and *M*. sp. 101).

Finally, we were not able to assign *M*. sp. 10 to a mating syndrome, because—although we found received sperm in the antrum and its sperm carry long bristles—it also has a sharp stylet and a simple antrum. From our previous findings, we would expect this species to have a thickened antrum due to its interaction with sperm and the mating partner's stylet. This discrepancy could possibly be attributed to misclassification of the antrum morphology, since *M*. sp. 10 has very pronounced shell glands, making it difficult to see the anterior part of the antrum, possibly obscuring a thickening or cellular valve (see specimen IDs MTP LS 788 and MTP LS 801 for a possibly thin cellular valve).

Hypodermic insemination and sperm morphology

In addition to the changes in sperm design mentioned above, we tested whether hypodermic insemination is associated with a change in sperm length using phylogenetic least squares (PGLS) regression. We used received sperm location, sperm bristle state, antrum state, and the inferred mating syndrome as predictors and the $\log_{10}$ transformed sperm length as the response variable. In all cases, the states that indicate the reciprocal mating syndrome were associated with longer sperm, with the largest effect for antrum state, followed by the inferred mating syndrome (Fig 9). This seems

reasonable, since the bristle type falsely classified *M.* sp. 68 and *M.* sp. 82 as hypodermically mating, while the received sperm location and inferred mating syndrome analyses had lower samples size than that with antrum state. The predictive value of the PGLS models was high, indicating that a large proportion of the variation in sperm length is explained by the phylogeny and these indicators of the reciprocal mating syndrome (Fig 9, Tab S6). Note that despite these strong associations, there is considerable overlap in sperm length between the species exhibiting the different states, with some species with the reciprocal mating syndrome having short sperm (Fig 9, Tab S6) and an overall 6.7-fold variation in sperm length across all species (mean length ranging from 25.6 to 173.1 μm).

Coevolution of male and female genitalia

To investigate coevolution between male and female genital traits, we independently summarised five male and four female genital traits using pPCA. Stylet PC1 was positively loaded with stylet length and the width of the distal opening and negatively loaded with distal asymmetry. Therefore, high values of Stylet PC1 represent a more elongate stylet with a wider and less sharp distal opening (Fig 10A, Tab S7). Antrum PC1 was positively loaded with all input variables, meaning that large values represent more complex female genitalia (Fig 10B). A PGLS regression of Stylet PC1 on Antrum PC1 across all species revealed a significant positive relationship (Fig 10C). This relationship closely matches the loadings on PC1 in the pPCA analysis of all reproductive traits (see Fig 5) and could therefore be mainly driven by the simplification of the antrum in hypodermically mating species. Therefore, we repeated the analysis on the subset of species assigned to the reciprocal mating syndrome. We again found a positive relationship between Stylet PC1 and Antrum PC1, indicative of male-female coevolution of genital morphology among the reciprocally mating species (Fig 10C).

Discussion

Across the genus *Macrostomum*, hypodermic insemination has evolved independently at least 9 times when assessed based on the location of received sperm, and at least 13 times when our more inclusive inferred mating syndrome is considered. The frequent origins of this type of traumatic insemination is remarkable considering that according to Lange et al. [17] merely 12 and 11 such origins have to date been documented in gonochorists and simultaneous hermaphrodites, respectively (including the two cases previously documented in *Macrostomum*). Our detailed analysis of a single genus of free-living flatworms thus approximately doubles the number of documented origins of hypodermic insemination among hermaphrodites. Moreover, three additional origins of traumatic insemination have also recently been documented in the Macrostomorpha [77], the parent group of the genus *Macrostomum*, suggesting that the origin of traumatic insemination occurs commonly in the Macrostomorpha, and maybe other groups of free-living flatworms.

The majority of the collected species are likely undescribed, and a large proportion of the diversity in this genus is yet to be discovered. Since free-living flatworms remain understudied in many regions of the world [78], we have likely not yet documented all convergent origins of hypodermic insemination within *Macrostomum*. As a case and point, the deeply split placement of *M.* sp. 15 (from South Africa) and *M.* sp. 118 (from Southern France) suggests the existence of an additional clade, possibly hinting at large amounts of additional molecular diversity. While we have no observations on the sperm morphology of *M.* sp. 15 or *M.* sp. 118, the stylet morphology of *M.* sp. 15 is suggestive of hypodermic insemination, and closer investigations of this clade would be interesting since this could facilitate inference about the ancestral state leading to the reciprocal clade.

Frequent convergent evolution of this extreme type of mating bolsters the interpretation that it represents an adaptive resolution to sexual conflict over mating rate, mating role or both [12, 18, 28, 39, 41, 46]. Hypodermic insemination likely is an alternative strategy in an ongoing evolutionary chase between donor and recipient, with donor persistence traits, such as complex sperm with bristles

498 and manipulative seminal fluids, and recipient resistance traits, such as the suck behaviour and
499 complex female genitalia, engaged in constant antagonistic coevolution [12, 17, 20, 46]. Moreover,
500 this coevolution does not only seem to drive the frequent origin of hypodermic insemination, but also
501 appears to result in the emergence of multiple morphological innovations, such as modifications of
502 stylet, sperm and antrum morphology (see also below).

503 Interestingly, we find no clear evidence for reversals back to reciprocal mating once hypodermic
504 insemination has arisen. However, while a Dollo model was preferred in all our ancestral state
505 reconstructions, this evidence was not completely decisive, as alternative models also received at
506 least some support. Reciprocal copulation clearly is the ancestral state of the reciprocal clade, but the
507 state of the most recent common ancestor of the genus is less clear, allowing for either a gain or a
508 loss (Fig S3). Similarly, in the finlandense clade we either have two independent losses with *M*. sp. 12
509 and *M*. sp. 44 retaining the ancestral state, or a single loss with a gain in these two sibling species
510 (Fig S3).

511 Since hypodermic insemination fundamentally alters the nature of the mating interaction, it might be
512 difficult for a species to revert to copulation. Specifically, once copulation no longer involves
513 inserting the stylet into the antrum, a reversal back to copulation would presumably require both
514 mating partners to again engage in a reciprocal behaviour. Also, a simplification of the antrum could
515 further hinder reversals, since reciprocally copulating species have traits that presumably reduce the
516 risk of injury (e.g. a thickened antrum epithelium). If these traits have been lost and are missing in
517 hypodermically mating species, then occasional reciprocal copulations could result in high fitness
518 costs for both partners.

519 In contrast, hypodermic insemination can presumably be performed unilaterally and might not require
520 the cooperation of both partners, making an evolutionary transition in that direction more likely.
521 Consequently, the origin of hypodermic insemination likely is a one-way process that canalises taxa
522 into this mating behaviour. However, across macroevolutionary time recipients may evolve
523 secondary female genitalia to avoid costs and regain control over the received ejaculate [17]. Why

this has not occurred in *Macrostomum* is unclear, but it might imply that costs of hypodermic insemination are generally low (possibly due to the striking regeneration ability of these flatworms [79]) or that the location of insemination is too variable for the evolution of a localised novel organ.

The analysis of correlated evolution clearly shows a strong association between both the sperm bristle and antrum state with the received sperm location. While on first sight this might simply appear to represent a confirmation of previous findings [41], albeit with more species and consequently a larger Bayes factors, it actually represents a very important qualitative improvement of the evidence for these associations. This is due to how tests of correlated evolution generally, and the Pagel test implemented in the BayesTraits analysis specifically, are set up. Tests of correlated evolution evaluate evidence for convergence, by testing whether a model, in which the evolution of the state of one trait is dependent of the state of the other trait, is more likely than a model where such shifts occur independently. While these tests supposedly correct for phylogenetic dependencies, they, somewhat counterintuitively, can support the dependent model of evolution even with only a single (unreplicated) origin of the trait states in question, provided that these origins map to appropriate nodes in the phylogeny (as recently outlined: [80, 81]). However, descendants of a clade often share multiple traits, many of which may not be functionally linked (as in the example given by [80] of the joint presence of fur and middle ear bones in mammals). Naturally, one aim of such a correlation analysis is to explore whether there might be a causal relationship between the traits in question. Thus, while the previous findings with two independent origins of hypodermic insemination in *Macrostomum* could be considered evidence for correlated evolution [41], that evidence is not as decisive as the large Bayes factors may have suggested. By sampling many more convergent events, we here could remedy the limitations of these earlier results, giving us substantially more confidence that our findings indeed reveal a causal link between these traits and allowing us to evaluate the consequences of shifts to hypodermic insemination in a more quantitative context.

In their earlier study [41] posited that specific combinations of adaptations were necessary for efficient hypodermic insemination, consisting of a needle-like stylet and sperm adapted for movement

through tissue. As a consequence of hypodermic insemination, the antrum should then become simpler, since it no longer interacts with the sperm and stylet of the partner, instead being only used for egg-laying. While the data of [41] was suggestive, the issue was that the second hypodermic origin consisted only of *M. hystrix*, which made quantitative analysis difficult. The principal component analysis we performed here shows that species with hypodermic insemination indeed have similar values of PC1, with such values corresponding tightly to the mating syndromes described by [41] (but note that we here slightly adjusted the definitions of the syndromes, since we lacked behavioural observations for most species). We show that hypodermic insemination is indeed associated with a distinct syndrome of reproductive traits that have convergently evolved many times in this genus. Therefore, these traits likely constitute predictable adaptations to this type of mating.

We observed several species with an unclear or intermediate mating syndrome, and these species suggest a possible route to hypodermic insemination via the initial evolution of a traumatic stylet in reciprocally mating species. A sharp stylet could provide anchorage during copulation, as potentially occurs in *M. spirale* (pers. obs.) and *M. hamatum* (pers. comm. P. Singh), but it may also serve to stimulate the partner during copulation, or aid in the destruction or removal of rival sperm already present in the partner's antrum. Finally, internal wounding by the stylet may help to embed sperm in the antrum wall and/or cellular valve to prevent their removal, either by rival mating partners or by the recipient during the suck behaviour. However, the fact that we observe many *Macrostomum* species with stylets with blunt distal thickenings (Fig 3) suggests that such internal wounding may not always be advantageous for the donor and that these structures may have evolved to avoid harm to the recipient [41]. Irrespective of the initial selective advantage that internal wounding may confer, it could then evolve further to complete internal traumatic insemination, and eventually complete avoidance of the female genitalia and hypodermic insemination via the epidermis.

Accidental sperm transfer due to copulatory wounding has generally been suggested as a possible route from copulation to traumatic insemination (e.g. in traumatically mating bedbugs) [17, 20]. In bedbugs, the attachment during mating is a two-step process, indicating that traumatic insemination

576 was evolutionarily preceded by traumatic penetration for attachment [17]. Similar transitions have
577 also been proposed for *Drosophila* species in the melanogaster group, where extragenital wounding
578 structures are typically used for anchorage during copulation. In traumatically inseminating
579 *Drosophila,* these structures are modified and pierce the mating partner's integument to inject sperm
580 into the genital tract [55, 82]. In this light our findings are remarkable, because previous examples
581 only compared species with and without traumatic insemination, whereas we potentially observe
582 species "in the act" of transitioning to traumatic insemination. These intermediate species should be
583 excellent targets for future studies of the costs and benefits of traumatic insemination, especially
584 because the two candidates represent independent evolutionary transitions.
585 The effects of traumatic insemination on sperm morphology are mostly unknown, and we are only
586 aware of the previous study on hypodermic insemination in *Macrostomum* [41]. We here present
587 quantitative evidence for a shift in sperm length under hypodermic insemination. The PGLS analysis
588 showed that, on average, the sperm of species with the hypodermic mating syndrome are significantly
589 shorter. There are several possible explanations for these findings. First, hypodermic insemination
590 avoids the recipient's genitalia, and these potentially allow both cryptic female choice (e.g. via the
591 suck behaviour), and sperm displacement or removal by competing donors. Hypodermic insemination
592 could therefore alter the mode of postcopulatory sexual selection and particularly sperm competition.
593 This is because postcopulatory sexual selection can introduce skews in sperm representation [35, 37,
594 38, 83], which likely results in lower sperm competition compared to a "fair-raffle" type sperm
595 competition, where sperm are expected to mix more freely [32, 43, 44]. Such skews seem less likely
596 to occur under hypodermic insemination. If sperm size trades-off with sperm number [42–44] then
597 more intense sperm competition under hypodermic insemination will likely favour the evolution of
598 smaller sperm [40, 41].
599 Second, sperm in *Macrostomum* is quite large compared to the size of the antrum, and it intimately
600 interacts with its epithelium, often being partially embedded in the cellular valve with the feeler [39,
601 41, 60, 84]. Sperm is also in close contact with rival sperm when animals mate multiply [85, 86].

Under such conditions of high sperm density, i.e. when sperm displacement is likely (e.g. [87–89]), sperm are predicted to be bigger compared to species in which the sperm storage organ is substantially larger than the sperm [45, 90]. While in species with hypodermic insemination the sperm will still intimately interact with the partner's tissue, the "storage organ" could now potentially include the whole body of the recipient and would then be considerably larger compared to the antrum. Third, sperm size could decrease due to natural selection, if small sperm are better able to move through the dense parenchymal tissue of the mating partner [41]. We know little about sperm movement within the recipient's tissues, but it seems analogous to the movement one can observe within the antrum, namely via undulation of the sperm body using cortical microtubules [91]. Presumably, smaller sperm need less energy to overcome tissue resistance and thus might move more efficiently.

Note, that these three explanations are not mutually exclusive, and their relative importance might depend on the physiology, morphology, and ecology of each species. Even though sperm morphology is exceptionally diverse, little is known about its functional significance [92]. Because traumatic insemination originates frequently, it offers an exciting opportunity to disentangle how natural and sexual selection shape sperm morphology. Such investigations can contribute to an integrative view, considering not only the sexual selection perspective, but also the direct functional constraints a sperm's ecology imposes (e.g. survival in sperm storage vs. movement through tissue) [93].

To disentangle mechanisms shaping sperm length evolution, we should ideally investigate the sperm morphology of other species with traumatic insemination and make use of natural variation in the location of sperm injection. For example, in bedbugs, the elaboration of the sperm receiving organ varies considerably from just being a slightly thickened epithelium to a complex spermalege [27, 94]. If movement efficiency is a crucial constraint, we might expect a negative correlation between sperm length and tissue transit time. Also of interest are comparative investigations of sperm length in taxa with species with traumatic insemination directly into the recipient's reproductive tract (e.g. the fly *Drosophila parabipectinata* [55] or the spider *Harpactea sadistica* [33]), because here movement through tissue is absent and presumably factors related to sexual selection dominate.

628 Besides changes in sperm length, we strengthen earlier findings that hypodermic insemination is
629 associated with the loss of sperm bristles [41]. Furthermore, we document hypodermic received sperm
630 in species with reduced bristles, indicating that hypodermic insemination likely precedes the complete
631 loss of bristles. The ancestral state reconstruction using three sperm bristle states found support for
632 an ordered evolutionary model, suggesting a transition via an intermediate state may be the rule. If
633 bristles only confer a selective advantage in reciprocally mating species, we could expect them to be
634 lost through drift once hypodermic insemination evolves, since these structures are no longer needed
635 to persist against the suck behaviour. Moreover, sperm bristles might also be selected against in
636 hypodermically mating species if they result in costs for the donor, such as a reduced rate of
637 spermatogenesis of complex sperm. Indeed, spermatogenesis of the complex sperm with bristles of
638 *M. lignano* takes about six days, which is two days longer than the development of the simpler sperm
639 in *M. pusillum* [95–97]. But note that this could also be due to sperm length differences between these
640 species, since longer sperm have been associated with longer development times [98, 99]. Moreover,
641 hypodermic sperm that carry bristles could also have reduced mobility [41].
642 The fact that there are hypodermically mating species with reduced bristles suggests that selection
643 might only be efficient at reducing bristle size until they no longer result in strong costs. Short bristles
644 might not hinder movement much, and once the bristles are reduced to such an extent that they do not
645 protrude much from the sperm, they could potentially be retained across longer evolutionary times.
646 The sperm of a member of the *M. pusillum* species-complex in the hypodermic clade contains
647 electron-dense bodies [100] that are similar to the bristle anchor structures identified in the
648 reciprocally mating *M. tuba* and *M. lignano* [91, 101]. If these structures are indeed remnants of
649 bristles, this would support the hypothesis (in agreement with our ancestral state reconstruction) that
650 bristles are symplesiomorphic in *Macrostomum*—with bristle loss as the derived condition—
651 suggesting that such greatly reduced bristles are not costly or detrimental to sperm function.
652 Moreover, sperm bristles have not been observed in three species of *Psammomacrostomum* (pers.
653 obs.), the sister taxon of *Macrostomum* in the Macrostomidae [77], nor in a presumably closely

associated genus (e.g. *Dunwichia* [102]). Sperm bristles thus appear to be a novel trait that is restricted to the genus *Macrostomum* (and some closely associated and taxonomically problematic genera that should be synonymised with *Macrostomum*; see SI Taxonomy). However, more detailed investigations of sperm ultrastructure of species in the hypodermic clade and other Macrostomorpha are needed to evaluate this hypothesis.

Our data further suggest that the correlation between reduced/absent bristles and the hypodermic mating syndrome is not perfect, because we document two species without sperm bristles, that very likely mate through reciprocal copulation (*M*. sp. 68 and *M*. sp. 82; black arrows in Fig 5). In both species, we have observed sperm in the antrum that in both is very muscular and complex (**Fig 6**). We speculate that sperm is deposited deeply inside the antrum, so that sperm bristles may no longer serve an anchoring function in these species. Additionally, the sperm of *M*. sp. 82 has a peculiar feeler (Fig 6), which could represent an adaptation to this complex antrum morphology. It is also not clear if these species perform the suck behaviour, as this was not observed in mating observations of *M*. sp. 82 (pers. comm. P. Singh), while we currently have no data at all on the mating behaviour of *M*. sp. 68. The loss of sperm bristles in reciprocally mating species should caution us against a too hasty conclusion that bristle loss is adaptive. Instead, bristles could be lost simply due to drift because they no longer serve a function.

We also find clear evidence for male-female genital coevolution, both across all species and when the analysis is restricted to species assigned to the reciprocal mating syndrome. This evidence is in line with previous findings of such coevolution in hermaphrodites (e.g. [53, 103]) and contributes to a growing body of evidence that male-female genital coevolution frequently occurs in both hermaphrodites and gonochorists [104–107]. Genital coevolution is expected both under scenarios of sexual selection and sexual conflict. Under the sexual selection perspective, we expect coevolution due to cryptic female choice, where the recipient will choose based on genital traits of the donor [108]. Donors are therefore selected to closely match their genital morphology to the selection criteria of the recipient. Similarly, we expect coevolution to result from sexual conflict, since persistence and

680 resistance traits of the donor and recipient will coevolve due to sexually antagonistic selection [8].
681 The number and nature of the traits involved in this coevolution, as well as their respective selective
682 optima, might potentially differ between species, driving trait diversification and speciation [109,
683 110].
684 We see signatures of evolutionary diversification across the genus, with the emergence of several
685 novel and clade-specific traits. Five related species in the spirale clade carry lateral stylet protrusions,
686 which might serve to anchor the donor during copulation or allow sperm removal, and these species
687 have complex antra and an unusual sperm design with only a single bristle. Moreover, several species
688 in the tanganyika clade have long sperm, which could constitute a novel persistence trait serving a
689 similar function as is hypothesised for the sperm bristles. It appears possible that these long structures,
690 which are probably greatly elongated sperm feelers, can embed themselves in the recipient's antrum.
691 They may thereby anchor the sperm and help to prevent removal, making bristles less important to
692 resist the suck behaviour. Indeed, several of these species have quite short bristles (including the
693 above-mentioned *M.* sp. 68 that lacks bristles entirely), in spite of showing the reciprocal mating
694 syndrome, thus supporting this interpretation. Other novel traits include the peculiar sperm feeler of
695 *M.* sp. 82, the novel sperm appendages in *M. distinguendum*, and the frequent emergence of a sperm
696 velum, the function of which is currently completely unknown.
697 While coevolution can drive trait diversification, it is also interesting to consider cases of
698 convergence. Repeated origins of a trait indicate that it is a readily available target of selection.
699 Besides hypodermic insemination, another example of convergent evolution in *Macrostomum* is the
700 origin of a second female genital opening. The phylogeny suggests that a second female genital
701 opening has evolved at least four times independently within the genus (see SI Female openings for
702 a more detailed discussion). In all species, the novel second opening is associated with a muscular
703 bursa that could allow cryptic female choice, since it might eject sperm through muscular contraction.
704 Such contraction is also at play during the suck behaviour, where, at least in *M. hamatum*, sperm can

be seen to be ejected from the antrum even before the worm places its mouth on the female genital opening (pers. comm. P. Singh).

In summary, our findings document a dynamic evolutionary history of male-female coevolution driving frequent innovations. Such coevolution does not only lead to trait diversification, but likely also drives the frequent origin of traumatic insemination, because it allows donors to overcome pre- and postcopulatory choice and/or resistance mechanisms of the recipient. The drastic and repeatable morphological changes associated with traumatic insemination indicate a shift in functional constraints on sperm morphology and function, which are presumably not unique to these flatworms, and which should therefore also be studied in gonochoristic models for traumatic insemination. Such studies will allow us to disentangle specific and general factors advancing our general understanding of sperm form and function.

Material and Methods

Field collections and documentation

Almost all specimens were collected from the field in freshwater, brackish or marine habitats, either from sediment samples or from water plants (details on all sampled specimens and their measurements in **Tab S2**). We documented specimens extensively with digital photomicrography, as previously described [41, 77, 84], using light microscopes (Leica DM2500, Olympus BH2, Leitz Diaplan, Zeiss Axioscope 5) with differential interference contrast (DIC) and equipped with digital cameras (Sony DFW-X700, Sony DFK41 and Ximea xiQ). We collected both images and videos at various magnifications (40x to 1000x), documenting the general habitus and details of internal organs, which is possible due to the small size and high transparency of these organisms. To document sperm morphology we amputated a worm's tail slightly anterior of the seminal vesicle, ruptured the seminal vesicle (as described in [111]), and documented the released sperm in a smash preparation using DIC (and sometimes also phase-contrast microscopy). When possible, we prepared whole-mount permanent preparations of these amputated tails to preserve a physical specimen of the male intromittent organ, the stylet [41, 77]. Finally, we preserved the entire animal, or its anterior portion when amputated, for molecular analysis, in either RNAlater solution (Sigma, stored at 4°C up to a few weeks and then at -80°C) or in absolute ethanol (stored cool for up to a few weeks and then at -20°C).

Morphological data

To characterise morphology, we collected both quantitative (Q) and categorical (C) data from the detailed images and videos of the collected specimens (or from the taxonomical descriptions of the few species we did not collect ourselves). Categorical data were determined on a per species basis, while quantitative data were taken per individual. We measured body size (Q) as the total body area and either measured or scored various aspects of the stylet (Q: length, curviness, width of the proximal

opening, width of the distal opening, and asymmetry of the distal thickening; C: sharpness of the distal thickening), the sperm (Q: total length, bristle length; C: sperm bristle state, presence of a brush, and presence of a velum) and antrum (Q: number of genital openings; C: antrum thickness, presence and thickness of anterior cellular valve, antrum chamber complexity, and an overall compound measure of antrum complexity). We refer the reader to the SI Morphology section for detailed explanations of these measures. Morphometric analyses were performed using the software ImageJ (version 1.51w) and the plugin ObjectJ (version 1.04r), which allows marking structures in the original images in a non-destructive manner. The pixel length of structures was converted into µm, by calibrating the different microscope setups using a stage micrometre. For comparative analysis we transformed body area ($\log_{10}$ of the square-root) and $\log_{10}$ transformed all linear measures (stylet length, width of the proximal opening, width of the distal opening, sperm length, and bristle length). The sample sizes for all quantitative measurements are given in **Tab S1** and species averages in Tab S9.

Sequence data generation

We extracted both DNA and RNA from the RNAlater samples using the Nucleospin XS Kit in combination with the Nucleospin RNA/DNA Buffer Set (Macherey-Nagel), and we extracted DNA from the ethanol samples using the DNeasy Blood and Tissue kit (Qiagen, Germany). Extracted DNA and RNA was stored at -80°C. We amplified a partial *28S rRNA* sequence from DNA samples using PCR primers ZX-1 and 1500R and, for some fragments, additional nested PCR using primers ZX-1 and 1200R, or 300F and 1500R, with polymerases and cycling conditions as previously described [41, 61]. We sequenced the resulting fragments from both sides using the PCR primers (Microsynth, Switzerland), assembled them in Geneious (v 11.1.5) with the built-in assembler (using the Highest sensitivity setting), with minor manual trimming and correction. For some sequences, we obtained additional sequences using internal primers 1090F, ECD2 (both [41]).

764 To generate the RNA-Seq library for *M. clavituba* we extracted RNA from 40 pooled animals using
765 Tri™ reagent (Sigma) and then prepared the library using the TruSeq® Stranded mRNA kit
766 (Illumina). And to generate the RNA-Seq libraries for all other selected RNA samples, we used
767 SMART-Seq v4 (Clontech Laboratories, Inc.) in combination with the Nextera XT DNA library
768 preparation kit (Illumina). When at least 5ng total RNA was available, we performed the SMART-
769 Seq v4 protocol with 12 preamplification steps and otherwise we used 1ng and performed 15
770 preamplification steps. Libraries were checked for quality using a Fragment Analyzer (Agilent) and
771 then sequenced as 101 paired-end reads on the HiSeq2500 platform (using the HiSeq® SBS Kit v4,
772 Illumina) at the Genomics Facility Basel of the University of Basel and the Department of Biosystems
773 Science and Engineering of the ETH Zürich.

774 ## Species delimitation

775 Most collected specimens could not be assigned to a taxonomically described *Macrostomum* species
776 (http://turbellaria.umaine.edu) [112] and are therefore likely new to science. We present
777 transcriptome-level information for most species but could not conduct RNA-Seq on all >1600
778 collected specimens. We, therefore, relied on morphology and the partial *28S rRNA* sequences for
779 species assignment since this fragment is widely used as a DNA barcode for flatworms (e.g. [77, 113–
780 115]. We constructed haplotype networks from the partial *28S rRNA* sequences using the TCS
781 algorithm [116] implemented in the TCS software [117] and chose to delimit species with >3
782 mutational differences in the network. Such a difference in this *28S rRNA* fragment indicates distinct
783 species among microturbellarians [115], but it is insufficient to clearly distinguish recently diverged
784 species within *Macrostomum* [61]. Additional markers like COI are frequently used to detect recent
785 divergence (e.g. [41, 77]), but despite considerable efforts, we were unable to develop universal
786 primers (a common issue in flatworms, see e.g. [118]) and individual primer optimisation, as in [61],
787 was not feasible here. Instead, we chose to also delimit species with ≤3 differences in this *28S rRNA*
788 fragment if they showed clear diagnostic differences in morphology. We thus err on the side of

lumping specimens, with species with shallow molecular and no diagnostic morphological divergence being assigned to the same operational species. We provide haplotype networks accompanied by drawings of the diagnostic features for all species (see SI Haplotype networks).

When the morphology of a species was diagnostic, we sequenced several specimens, when possible from different sample locations, to confirm that they were indeed molecularly similar and then assigned additional specimens based on morphology only. When no unambiguous diagnostic traits could be defined (as was the case for many species of the hypodermic clade) we sequenced all specimens collected for molecular assignment. We used 668 sequences generated for this study (604 sequences, Accessions: MT428556-MT429159) or from public databases (64 sequences), representing the available diversity across the genus. We also included 15 sequences from seven species of our chosen outgroup genus *Psammomacrostomum*. We aligned sequences using MAFFT („—genafpair –maxiterate 1000“) and generated haplotype networks for the recovered clades. We removed all columns that contained at least one gap before running TCS, leaving us with an alignment with 787 sites and 385 variable bases. We thus ignore indels to avoid the generation of cryptic species solely based on these, since scoring indels is non-trivial, and they are frequently treated as missing data [119].

Phylogenetics

We used 134 transcriptomes representing 105 species, including four distant outgroups, *Haplopharynx rostratum*, *Microstomum lineare*, *Myozonaria bistylifera*, and *Karlingia* sp. 1, three species from the sister genus, *Psammomacrostomum* (see also [77]), and 98 *Macrostomum* species in the phylogenomic analysis. This included the four publicly available high-quality transcriptomes of *M. hystrix*, *M. spirale*, *M. pusillum* [120], and *M. lignano* [121, 122], as well as 130 *de novo* assembled transcriptomes. Transcriptomes were assembled as previously described [120] and were mostly derived from whole single specimens (and a few from anterior fragments only, which may reduce the transcriptome repertoire), while nine were generated by combining RNA-Seq data sets

from several animals or pooling animals that were collected at the same location and assigned to the species based on our taxonomic expertise into one sample (for details on the transcriptomes used see **Tab S8**). We assessed transcriptome quality using TransRate (version 1.0.2, [123]), which maps the reads back to the assembly and calculates mapping metrics, and BUSCO (version 2.0, [124]), which searches for the presence of a curated set of core conserved genes. Specifically, we ran the BUSCO analysis with the metazoan dataset consisting of 978 genes (version uploaded 2016-11-01). These BUSCO scores were also used to select one representative transcriptome when multiple transcriptomes were available for a species (see below). We determined the empirical insert size of our libraries by mapping the reads to the assemblies using SNAP (version 1.0, [125]) and then extracting the mean insert size using Picard (version 2.20.2).

To infer a set of orthologous genes (orthologs) we predicted open reading frames (ORFs) for each transcriptome using TransDecoder (version 5.3.0 [126]) with Pfam searches (version 32.0) to retain transcripts with predicted proteins and kept only one ORF per transcript using the "--single_best_only" option. We then clustered predicted proteins with at least 99.5% sequence identity using the CD-HIT clustering algorithm (version 4.7, [127]). The amino acid sequences where then processed with OrthoFinder (version 2.2.6, [128]) using the „-os" option to perform a length-adjusted reciprocal BLAST searches followed by MCL clustering. We processed the resulting set of homologous genes (homogroups) using modified scripts from the phylogenomic dataset construction workflow [129]. We aligned all homogroups that contained at least 10 species using MAFFT (version 7.310, „–genafpair –maxiterate 1000" [130]), inferred the best substitution model with ModelFinder [131], and the gene tree using IQ-TREE (version 1.5.5, [66], command: "-mset DCMut, JTTDCMut, LG, mtZOA, PMB, VT, WAG –mfreq FU,F –mrate G"). Then we trimmed the gene trees using "trim_tips.py" to remove tip branches that were longer than two substitutions per site and "mask_tips_by_taxonID_transcripts.py" to remove monophyletic paralogs by choosing the sequence with the best representation in the alignment. We split off subtrees with long internal branches using "cut_long_internal_branches.py" and inferred orthologs using the rooted outgroup method in

"prune_paralogs_RT.py". This method uses known outgroup taxa to root the phylogenies, which then allows to infer the history of speciation and duplication and extract the most inclusive set of orthologs. We defined *Haplopharynx rostratum*, *Microstomum lineare*, *Myozonaria bistylifera*, and *Karlingia sp. 1* as the outgroup and all *Macrostomum* and *Psammomacrostomum* as an ingroup (following the phylogeny of [77]), since the algorithm does not include the defined outgroup in the output, and we then used *Psammomacrostomum* to root our final ortholog trees. Next, we realigned the orthogroups using MAFFT and trimmed the alignment with ZORRO [132], discarding any columns in the alignment with a score of less than five and filtering alignments that were shorter than 50 amino acids after trimming. Finally, we inferred ortholog gene trees with 100 non-parametric bootstraps with IQ-TREE, by inferring the best fitting model „-mset DCMut, JTTDCMut, LG, mtZOA, PMB, VT, WAG -mfreq FU,F -mrate E,I,G,I+G“. These best fitting substitution models were later also used for the partitioned maximum-likelihood analysis.

We generated two gene matrices, one matrix containing many genes but a relatively moderate occupancy (called L for low occupancy) and one with a lower gene number but a higher occupancy (called H for high occupancy, **Tab 2**). We conducted most analyses on the H alignment for computational tractability and since missing data can have a negative influence on tree inference, particularly when using Bayesian methods ([133], but see [134]), and only the less computationally demanding summary and maximum-likelihood methods were also applied to the L alignment. So for both alignments, we calculated a maximum likelihood species phylogeny using IQ-TREE ("L-IQ-TREE" and "H-IQ-TREE") with a partition for each gene and using the best substitution model for each gene (see above). We calculated 1000 ultrafast bootstraps and conducted an approximate-likelihood ratio test to assess branch support. We inferred Bayesian phylogenies using ExaBayes (version 1.5, [68]) for the H alignment ("H-ExaBayes") with partitions for each gene, equal prior probability on all available amino acid substitution models, and with gamma models for all partitions. We ran four independent chains retaining every 500th iteration and discarding the first 365,000 iterations as burn-in. We terminated the analysis after 1.46 million generations since the average

deviation of split frequencies between all the chains was below 1%, indicating convergence. We further assessed convergence using Gelman's R of the likelihoods with the R package coda [136], which showed three chains had converged, while one appeared stuck on a local peak. Since all chains quickly converged on the same topology, the local peak probably occurs due to differences in the substitution models applied, with the HIVB model being present at a low rate in the three converged chains, but being absent in the fourth. We combined the three converged chains, discarded the fourth and calculated a consensus tree using quadratic path distance optimisation (using the ls.consensus function in the R package phytools).

To account for potential issues caused by model misspecification, we also performed an unpartitioned analysis using the CAT model implemented in PhyloBayes (version 1.5, [69]). Because a full GTR model of the amino acids was too parameter rich, we ran the tool on the DNA coding sequence of the H alignment and fit the CAT+GTR model ("H-PhyloBayes"). Two chains were run in parallel on 400 CPUs for two weeks. Due to the high cost of the analysis, we terminated the run after 23,311 iterations, at which point the chains showed an identical topology (after removal of 10,000 iterations as a burn-in). At this point the realised difference in the likelihoods between the two H-PhyloBayes chains was high (0.21 with an ESS of 858) suggesting the chains had not converged. However, Gelman diagnostics for the likelihoods suggested convergence. Our chains could potentially be at a local peak where the hypodermic clade is paraphyletic, possibly due to poor mixing, potentially biasing the results. Therefore, we discuss the H-PhyloBayes phylogeny, but exclude it from the comparative analysis.

To construct phylogenies while accounting for potential incomplete lineage sorting, we ran the quartet based method implemented in ASTRAL-III (version: 5.6.1, [67]) on both alignments ("L-ASTRAL" and "H-ASTRAL"). We assessed the level of gene tree – species tree conflict at each node of the phylogeny using the quartet support score, which gives the proportion of quartet trees induced by the gene trees that are supporting the species tree topology as opposed to the two possible alternatives [67]. Strong support for the species tree partition can be interpreted as little disagreement between

892 the gene trees, while support for the alternate topologies indicates strong gene tree – species tree
893 conflict.

894 Furthermore, we chose representative *28S rRNA* sequences from the haplotype network analysis to
895 infer phylogenetic placement for species without a transcriptome. For species with a transcriptome,
896 we mostly chose the *28S rRNA* sequence derived from the same specimen (for exceptions see
897 **Tab S8**). We aligned the sequences using MAFFT ("--maxiterate 1000 –globalpair"), trimmed the
898 start and end using trimAl ("–nogaps –terminalonly" [137]) and determined the best substitution
899 model using ModelFinder with the BIC criterion. We then combined this alignment with the H amino
900 acid alignment, and analysed it with the best fitting substitution model using IQ-TREE as described
901 above (referred to as "C-IQ-TREE", called C for combined). For comparative analysis, we
902 transformed the phylogenies (C-IQ-TREE, H-IQ-TREE and H-ExaBayes) to be ultrametric and with
903 a root depth of 1 using the penalized marginal likelihood approach [138] implemented in the software
904 TreePL [139].

905 We performed all follow-up analyses on the H-IQ-TREE and the H-ExaBayes phylogenies to explore
906 possible effects of phylogenetic uncertainty and included the C-IQ-TREE phylogeny to incorporate
907 more species. We pruned the phylogenies to include one representative tip per species, choosing the
908 transcriptome with the best BUSCO score. We present the results obtained with all three phylogenies,
909 but primarily discuss the ones obtained with the C-IQ-TREE phylogeny, since all results are
910 quantitatively similar and qualitatively identical.

911 After the above analyses were completed, we discovered that some transcripts of the RNA-Seq
912 libraries constructed with the SMART-Seq v4 cDNA kit contained cDNA synthesis primer
913 sequences. We did not remove these primers before transcriptome assembly since i) the manufacturer
914 specifically states they should not occur in RNA-Seq reads if used in combination with the Nextera
915 XT DNA library preparation kit (Appendix C in SMART-Seq v4 Ultra Low Input RNA Kit for
916 Sequencing User Manual, Takara Bio Inc. available at: https://www.takarabio.com/assets/a/114825)
917 and ii) because the relevant primer sequence is proprietary (both confirmed by Takara EU tech

support in November 2020), so that we initially did not have access to its sequence. We only discovered the offending primer sequence due to ongoing work on two *Macrostomum* genome projects. Because the analyses presented here had at this point used approximately 650,000 CPU hours of computation, and since the effect of the primer was likely small, we elected to perform some follow-up analyses to determine how robust our results were to its removal, rather than opting for a complete re-assembly and re-analysis of all the data. As we outline in SI Primer removal, the removal of the cDNA synthesis primer sequences had little effect on the topology and branch lengths of L-IQ-TREE, H-IQ-TREE and C-IQ-TREE and, therefore, we present the results based on the original alignments.

Frequent origins of hypodermic insemination

The original definitions of the mating syndromes integrated both morphological and behavioural traits [41]. Because we here lacked behavioural data for most species, we adapted the definitions, relying instead on several morphological traits and the observed received sperm location only to derive the inferred mating syndrome (Tab 3). We assigned species to the hypodermic mating syndrome if we exclusively found hypodermic received sperm, since this represents strong evidence for hypodermic insemination, as opposed to species where we observed both hypodermic sperm and received sperm in the female antrum, which we classify as intermediate (Tab 3). Moreover, because hypodermic sperm can be difficult to observe, especially in species with low investment into sperm production, we also assigned species that lacked received sperm observations to the hypodermic mating syndrome based on their morphology alone, namely when they had a simple antrum, a sharp stylet, and absent or reduced sperm bristles (**Tab 3**). And while observing received sperm in the female antrum may not exclude occasional hypodermic insemination, it is a strong indication of the reciprocal mating syndrome, especially when it occurs in a species with a blunt stylet. We, therefore, assigned all species with received sperm in the antrum and a blunt stylet to the reciprocal mating syndrome (**Tab 3**). And since some reciprocally mating species also have a sharp stylet (e.g. *M. spirale*), which could possibly

943 wound the partner internally during mating (pers. obs.), we also assigned these species to the
944 reciprocal mating syndrome, provided that we observed received sperm in the antrum, and that they
945 had sperm with bristles (**Tab 3**). These assignments based on morphology alone are supported by our
946 analysis of correlated evolution, showing a strong association between the received sperm location
947 and both sperm bristle state and antrum type, respectively (see Results). The inferred mating
948 syndrome is therefore a more inclusive classification of hypodermic insemination compared to an
949 assignment based on received sperm location alone.
950 We estimated ancestral states of the mating syndrome and three reproductive traits linked to it, namely
951 received sperm location, sperm bristle state, and antrum state. First, we used binary scorings (see SI
952 Morphology), equivalent to how they were used in the tests for correlated evolution (see below).
953 However, since we predicted that losses/reductions of the trait would transition via an intermediate
954 state, we also reconstructed ancestral states with the inferred mating syndrome, received sperm
955 location and sperm bristle state scored as trinary states. We estimate the history of transitions using
956 stochastic character mapping [140] with the R package phytools [141]. We determined the
957 appropriate transition matrix for reconstruction by fitting MK-models with either all state transitions
958 with equal rates (ER), with symmetric rates (SYM), with all rates different (ARD), and with a model
959 without the possibility of gains once the trait is lost (Dollo). For traits with trinary states, we
960 additionally fit an ordered model, where transitions have to go via the intermediate state (ORD) and
961 an ordered Dollo model without the possibility of gains once the trait is completely lost, but allowing
962 transitions back from the intermediate state (ORD-Dollo). We reconstructed ancestral states for all
963 models with a corrected AIC weight >0.15 (Table 3). We used the fully Bayesian implementation of
964 stochastic character mapping with a gamma prior on each transition ($\alpha = 1$, $\beta = 1$, this results in a low
965 prior on the number of transitions) and reconstructed 1000 character histories (10.000 iterations burn-
966 in followed by 10,000 iterations and retaining every 10^{th} character history). We summarised the
967 number of transitions as the average number of changes as well as the 95% credible interval.

Correlated evolution

We performed a number of tests of correlated evolution to ask if the numerous convergent changes in received sperm location, sperm bristle state and antrum state were evolutionarily dependent. But since we do not have direct observations of received sperm in all species, we first conducted a correlation test between sperm bristle state and received sperm location, and then tested for correlated evolution between both of these variables and the antrum type. We scored all traits as binary and applied Pagel's correlation test [142] as implemented in BayesTraits3 (available at http://www.evolution.rdg.ac.uk/BayesTraitsV3.0.2/BayesTraitsV3.0.2.html). This test compares the marginal likelihood of a model where the transition probability of the first trait depends on the state of the second trait with a model where the two traits evolve independently. For all analysis we ran four independent MCMC chains for 510 million iterations, discarding the first 10 million iterations as burn-in and retaining every 1000^{th} iteration. We calculated the marginal likelihood of the models using the stepping stone sampler of BayesTraits3 with 1000 power posteriors estimated with 10,000 iterations each. We assessed convergence of the chains by calculating Gelman's R using the coda R package [136] and upon confirming convergence merged the chains for further analysis. To evaluate sensitivity to the prior, we ran all analyses with a uniform prior (U 0 100), an exponential prior (exp 10) and a reversible-jump hyperprior with a gamma distribution between 0 and 1 for both the rate prior and the hyperprior (rjhp gamma 0 1 0 1). Models were compared with Bayes factors using the marginal likelihoods calculated by the stepping stone sampler (i.e. $BF=2(\log LH_{dependent} - \log LH_{independent})$). To assess the influence of the phylogeny we conducted these tests on three different phylogenies. This setup resulted in 54 model runs consisting of 216 MCMC chains (three trait pairs x three priors x three topologies x four chains for the independent and the dependent model; see also the Correlated evolution section in the Supporting Information). The number of species included was largest for the correlation between sperm bristle state and antrum type (C-IQ-TREE: 124; H-IQ-TREE & H-ExaBayes: 95), and similar for received sperm location and antrum type (C-IQ-TREE: 100; H-

993 IQ-TREE & H-ExaBayes: 77) and received sperm location and sperm bristle state (C-IQ-TREE: 101;
994 H-IQ-TREE & H-ExaBayes: 76).

995 Hypodermic insemination and convergence in morphospace

996 We conducted a multivariate analysis to investigate whether the convergent evolution of hypodermic
997 insemination is associated with changes in a variety of reproductive traits (see SI Morphology, and
998 Tables A1-2 and Figures A1-3 therein). For this we summarized all our data on stylet, sperm and
999 antrum morphology (including both quantitative and categorical data) using principal component
1000 analysis. Since regular principal component analysis assumes independence of observations, an
1001 assumption violated by the phylogenetic relationships of species [143], we calculated a
1002 phylogenetically corrected principal component (pPCA), using the phyPCA function in phytools with
1003 the lambda model. Since we combined data with different scales, we used the correlation matrix for
1004 all calculations. When discussing loadings of principal components we apply an aggressive threshold
1005 of ±0.5, since although this results in erosion of power, it keeps false-positive rate within expectations
1006 [144].

1007 Hypodermic insemination and sperm morphology

1008 To test the influence of hypodermic insemination on sperm length, we performed phylogenetically
1009 corrected ordinary least squared regression (PGLS) with the *gls* function in the R package *nlme*
1010 (version 3.1). We used *gls* because it allowed us to simultaneously incorporate phylogenetic signal in
1011 the residuals and account for variation in the number of measured specimens by using the sample size
1012 of the response as weights. We determined the best fitting evolutionary model for the covariance in
1013 the residuals by comparing corrected AIC of PGLS fitted with Brownian motion, lambda or Ornstein-
1014 Uhlenbeck models. We assessed if the assumptions of the PGLS were met by checking the
1015 distributions of the phylogeny-corrected residuals for normality and profiled the likelihood of the
1016 parameter of the correlation structure (i.e. lambda or alpha). Since R-squared values are problematic

for PGLS models [145] we calculated R_{pred} [146] to show model fits. As predictors, we used the binary traits included in the test of correlated evolution (received sperm location, sperm bristle state and antrum state) since they all are strong indicators of hypodermic insemination. Moreover, we also included the inferred mating syndrome as a predictor, but coded it as binary (hypodermic and reciprocal), and excluding the intermediate syndrome due to the low sample size of this group. Analysing all four of these predictors is somewhat redundant, but while received sperm location is the most direct evidence for hypodermic insemination it results in a reduced sample size (see above) and the other predictors allow the inclusion of more species.

Coevolution of male and female genitalia

If male and female genitalia coevolve, we expect correlations in aspects of their morphology across species. Ideally, this is detected by identifying homologous male and female structures in all species and correlating their shape evolution (e.g. [107]). However, the stylets within the reciprocal clade are highly variable, making the identification and placement of consistent landmarks challenging. Similarly, capturing the morphology of the antrum is challenging because these soft internal structures are highly transparent in live specimens. And while the antrum can be studied in histological sections, it can appear distorted and contracted due to the fixation, making it difficult to accurately measure length or volume. With these caveats in mind, we independently summarised male and female genital complexity using pPCAs (see above) and then performed a PGLS regression (see above) between the first principal components of each pPCA. We used five stylet traits (stylet length, curviness, width of proximal opening, width of distal opening, and distal asymmetry) and four antrum traits (thickness, cellular valve, chamber complexity and number of genital openings). We performed the PGLS regression both across all species and as well as only including species categorised as showing the reciprocal mating syndrome, since the species with the hypodermic mating syndrome were nearly invariant.

Acknowledgment

We thank the numerous people that have helped with field work. Especially, we are grateful for the help of, in no particular order, Werner Armonies, Benny Glasgow, Mohamed Charni, Edith Zemp, Bernhard Egger, Peter Ladurner, Gregor Schulte, Floriano Papi, Kazuya Kobayashi, Christopher Laumer, Wim Willems, Tom Artois, Christian Lott, Miriam Weber, Ana-Maria Leal-Zanchet, Kaja Wasik, Mariana Adami, Walter Salzburger, Adrian Indermaur, Bernd Egger, Fabrizia Ronco, Heinz Büscher, Victoria Huwiler, Philipp Kaufmann, Michaela Zwyer, Stefanie von Fumetti, Joe Ryan, Mark Q. Martindale, Marta Chiodin, John Evans, Leigh Simmons, Mauro Tognon, Piero Tognon, Cristiano Tognon, Pragya Singh, Nikolas Vellnow, Christian Felber, Ulf Jondelius, Sarah Atherton, Tim Janicke, Georgina Rivera-Ingraham, Ben Byrne, Yvonne Gilbert, Rod Watson, Jochen Rink, Miquel Vila-Fare, Helena Bilandžija, and Sasho Trajanovski. We thank Katja Eschbach of the Genomics Facility Basel for preparing and sequencing RNA-Seq libraries. We thank Peter Fields and Lukas Zimmermann for IT advice. We thank Jürgen Hottinger, Daniel Lüscher and Yasmin Picton for administrative and technical support. We thank Dita Vizoso for the use of sperm and stylet drawings for some of the previously described species and for providing inspiration for the new drawings. We thank Yu Zhang and his collaborators for kindly providing specimens of *M. baoanensis*. We thank Nick Goldman and Ziheng Yang for organising the summer school on Computational Molecular Evolution, which has helped and inspired the first author to explore phylogenetics. Calculations were performed, in part, at sciCORE (http://scicore.unibas.ch/) scientific computing centre at the University of Basel.

Funding

This work was supported by Swiss National Science Foundation (SNSF) research grants 31003A_162543 and 310030_184916 to LS.

Competing Interests

All Authors declare that they have no competing interests.

Data availability

The raw sequencing data generated for this study are available in the NCBI Sequence Read Archive repository with the following accession: PRJNA635941. The partial *28S rRNA* sequence are available on NCBI with the following accessions: MT428556-MT429159. Extensive image and video material of all documented specimens are deposited on Zenodo at: 10.5281/zenodo.4482135. Transcriptome assemblies, gene alignments and phylogenetic trees are deposited on Zenodo at: 10.5281/zenodo.4543289.

References

1. Bateman AJ. Intra-sexual selection in *Drosophila*. Heredity. 1948;2:349–68.

2. Arnold SJ. Bateman's principles and the measurement of sexual selection in plants and animals. Am Nat. 1994;144:S126–49.

3. Janicke T, Häderer IK, Lajeunesse MJ, Anthes N. Darwinian sex roles confirmed across the animal kingdom. Sci Adv. 2016;2:e1500983–e1500983.

4. Parker GA. The sexual cascade and the rise of pre-ejaculatory (Darwinian) sexual selection, sex roles, and sexual conflict. Cold Spring Harb Perspect Biol. 2014;6:a017509–a017509.

5. Puurtinen M, Ketola T, Kotiaho JS. The Good-Genes and Compatible-Genes Benefits of Mate Choice. Am Nat. 2009;174:741–52.

6. Arnqvist G, Nilsson T. The evolution of polyandry: multiple mating and female fitness in insects. Anim Behav. 2000;60:145–64.

7. Hosken DJ, Stockley P. Benefits of Polyandry: A Life History Perspective. In: Macintyre RJ, Clegg MT, editors. Evolutionary Biology. Boston, MA: Springer US; 2003. p. 173–94. doi:10.1007/978-1-4757-5190-1_4.

8. Arnqvist G, Rowe L. Sexual conflict. Princeton, N.J: Princeton University Press; 2005.

9. Parker GA. Sexual conflict over mating and fertilization: An overview. Philos Trans R Soc B Biol Sci. 2006;361:235–59.

10. Rice WR. Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. Nature. 1996;381:232–4.

11. Arnqvist G, Rowe L. Antagonistic coevolution between the sexes in a group of insects. Nature. 2002;415:787–9.

12. Charnov EL. Simultaneous hermaphroditism and sexual selection. Proc Natl Acad Sci U S A. 1979;76:2480–4.

13. Birkhead TR, Pizzari T. Evolution of sex: postcopulatory sexual selection. Nat Rev Genet. 2002;3:262–73.

14. Wedell N, Hosken DJ. The evolution of male and female internal reproductive organs in insects. In: Leonard JL, Córdoba-Aguilar A, editors. The evolution of primary sexual characters in animals. Oxford ; New York: Oxford University Press; 2010.

15. Morrow EH, Arnqvist G. Costly traumatic insemination and a female counter-adaptation in bed bugs. Proc R Soc B Biol Sci. 2003;270:2377–81.

16. Morrow EH, Arnqvist G, Pitnick S. Adaptation versus pleiotropy: why do males harm their mates? Behav Ecol. 2003;14:802–6.

17. Lange R, Reinhardt K, Michiels NK, Anthes N. Functions, diversity, and evolution of traumatic mating: function and evolution of traumatic mating. Biol Rev. 2013;88:585–601.

18. Schärer L, Janicke T, Ramm SA. Sexual conflict in hermaphrodites. Cold Spring Harb Perspect Biol. 2015;7:a017673.

19. Reinhardt K, Naylor R, Siva-Jothy MT. Reducing a cost of traumatic insemination: female bedbugs evolve a unique organ. Proc R Soc B Biol Sci. 2003;270:2371–5.

20. Reinhardt K, Anthes N, Lange R. Copulatory wounding and traumatic insemination. Cold Spring Harb Perspect Biol. 2015;7:a017582.

21. Benoit JB, Jajack AJ, Yoder JA. Multiple traumatic insemination events reduce the ability of bed bug females to maintain water balance. J Comp Physiol B. 2012;182:189–98.

22. Tatarnic NJ. Traumatic Insemination and Copulatory Wounding. In: Reference Module in Life Sciences. Elsevier; 2018.

23. Tatarnic NJ, Cassis G, Siva-Jothy MT. Traumatic insemination in terrestrial arthropods. Annu Rev Entomol. 2014;59:245–61.

24. Kathirithamby J, Hrabar M, Delgado JA, Collantes F, Dötterl S, Windsor D, et al. We do not select, nor are we choosy: reproductive biology of Strepsiptera (Insecta). Biol J Linn Soc. 2015;116:221–38.

25. Peinert M, Wipfler B, Jetschke G, Kleinteich T, Gorb SN, Beutel RG, et al. Traumatic insemination and female counter-adaptation in Strepsiptera (Insecta). Sci Rep. 2016;6:25052.

26. Stutt AD, Siva-Jothy MT. Traumatic insemination and sexual conflict in the bed bug *Cimex lectularius*. Proc Natl Acad Sci. 2001;98:5683–7.

27. Siva-Jothy M t. Trauma, disease and collateral damage: conflict in cimicids. Philos Trans R Soc B Biol Sci. 2006;361:269–75.

28. Michiels NK, Newman LJ. Sex and violence in hermaphrodites. Nature. 1998;391:647–647.

29. Sluys R. Sperm resorption in triclads (Platyhelminthes, Tricladida). Invertebr Reprod Dev. 1989;15:89–95.

30. Koene JM. Tales of two snails: sexual selection and sexual conflict in Lymnaea stagnalis and Helix aspersa. Integr Comp Biol. 2006;46:419–29.

31. Koene JM, Montagne-Wajer K, Roelofs D, Ter Maat A. The fate of received sperm in the reproductive tract of a hermaphroditic snail and its implications for fertilisation. Evol Ecol. 2009;23:533–43.

32. Parker GA. Sperm competition and the evolution of ejaculates: towards a theory base. In: Sperm competition and sexual selection. T.R. Birkhead and A.P. Møller. San Diego: Academic Press; 1998. p. 3–54.

33. Řezáč Milan. The spider *Harpactea sadistica*: Co-evolution of traumatic insemination and complex female genital morphology in spiders. Proc R Soc B Biol Sci. 2009;276:2697–701.

34. Parker GA. Sperm competition games: raffles and roles. Proc R Soc Lond B Biol Sci. 1990;242:120–6.

1144 35. Charnov EL. Sperm competition and sex allocation in simultaneous hermaphrodites. Evol Ecol.
1145 1996;10:457–62.

1146 36. Greeff JM, Parker GA. Spermicide by females: what should males do? Proc R Soc B Biol Sci.
1147 2000;267:1759–63.

1148 37. van Velzen E, Scharer L, Pen I. The effect of cryptic female choice on sex allocation in
1149 simultaneous hermaphrodites. Proc R Soc B Biol Sci. 2009;276:3123–31.

1150 38. Schärer L, Pen I. Sex allocation and investment into pre- and post-copulatory traits in
1151 simultaneous hermaphrodites: the role of polyandry and local sperm competition. Philos Trans R Soc
1152 B Biol Sci. 2013;368:20120052–20120052.

1153 39. Vizoso DB, Gunde Rieger, Lukas Schärer. Goings-on inside a worm: functional hypotheses
1154 derived from sexual conflict thinking. Biol J Linn Soc. 2010;99:370–83.

1155 40. Schärer L, Janicke T. Sex allocation and sexual conflict in simultaneously hermaphroditic
1156 animals. Biol Lett. 2009;5:705–8.

1157 41. Schärer L, Littlewood DTJ, Waeschenbach A, Yoshida W, Vizoso DB. Mating behavior and the
1158 evolution of sperm design. Proc Natl Acad Sci. 2011;108:1490–5.

1159 42. Parker GA. Selection on non-random fusion of gametes during the evolution of anisogamy. J
1160 Theor Biol. 1978;73:1–28.

1161 43. Parker GA. Why are there so many tiny sperm? Sperm competition and the maintenance of two
1162 sexes. J Theor Biol. 1982;:281–94.

1163 44. Parker GA. Sperm competition games: sperm size and sperm number under adult control. Proc R
1164 Soc Lond B Biol Sci. 1993;253:245–54.

1165 45. Immler S, Pitnick S, Parker GA, Durrant KL, Lüpold S, Calhim S, et al. Resolving variation in
1166 the reproductive tradeoff between sperm size and number. Proc Natl Acad Sci. 2011;108:5325–30.

1167 46. Michiels NK. Mating conflicts and sperm competition in simultaneous hermaphrodites. In: Sperm
1168 Competition and Sexual Selection. T.R. Birkhead and A.P. Møller. Elsevier; 1998. p. 219–54.

1169 47. Anthes N, Putz A, Michiels NK. Sex role preferences, gender conflict and sperm trading in
1170 simultaneous hermaphrodites: a new framework. Anim Behav. 2006;72:1–12.

1171 48. Anthes N. Mate choice and reproductive conflict in simultaneous hermaphrodites. In: Kappeler
1172 P, editor. Animal Behaviour: Evolution and Mechanisms. Berlin, Heidelberg: Springer Berlin
1173 Heidelberg; 2010. p. 329–57.

1174 49. Anthes N, David P, Auld JR, Hoffer JNA, Jarne P, Koene JM, et al. Bateman gradients in
1175 hermaphrodites: an extended approach to quantify sexual selection. Am Nat. 2010;176:249–63.

1176 50. Pélissié B, Jarne P, David P. Sexual selection without sexual dimorphism: Bateman gradients in
1177 a simultaneous hermaphrodite. Evolution. 2012;66:66–81.

1178 51. Michiels NK, Koene JM. Sexual selection favors harmful mating in hermaphrodites more than in
1179 gonochorists. Integr Comp Biol. 2006;46:473–80.

1180 52. Jarne P, Auld JR. Animals mix it up too: the distribution of self-fertilization among
1181 hermaphroditic animals. Evolution. 2006;60:1816–24.

1182 53. Koene JM, Schulenburg H. Shooting darts: co-evolution and counter-adaptation in
1183 hermaphroditic snails. BMC Evol Biol. 2005;5:25.

1184 54. Koene JM, Pförtner T, Michiels NK. Piercing the partner's skin influences sperm uptake in the
1185 earthworm Lumbricus terrestris. Behav Ecol Sociobiol. 2005;59:243.

1186 55. Kamimura Y. Twin intromittent organs of *Drosophila* for traumatic insemination. Biol Lett.
1187 2007;3:401–4.

1188 56. Tatarnic NJ, Cassis G. Surviving in sympatry: Paragenital divergence and sexual mimicry
1189 between a pair of traumatically inseminating plant bugs. Am Nat. 2013;182:542–51.

1190 57. Ramm SA, Vizoso DB, Schärer L. Occurrence, costs and heritability of delayed selfing in a free-
1191 living flatworm. J Evol Biol. 2012;25:2559–68.

1192 58. Ramm SA, Schlatter A, Poirier M, Schärer L. Hypodermic self-insemination as a reproductive
1193 assurance strategy. Proc R Soc B Biol Sci. 2015;282:20150660.

1194 59. Winkler L, Ramm SA. Experimental evidence for reduced male allocation under selfing in a
1195 simultaneously hermaphroditic animal. Biol Lett. 2018;14:20180570.

1196 60. Schärer L, Joss G, Sandner P. Mating behaviour of the marine turbellarian *Macrostomum* sp.:
1197 these worms suck. Mar Biol. 2004;145:373–80.

1198 61. Schärer L, Brand JN, Singh P, Zadesenets KS, Stelzer C-P, Viktorin G. A phylogenetically
1199 informed search for an alternative *Macrostomum* model species, with notes on taxonomy, mating
1200 behavior, karyology, and genome size. J Zool Syst Evol Res. 2020;58:41–65.

1201 62. Patlar B, Weber M, Temizyürek T, Ramm SA. Seminal fluid-mediated manipulation of post-
1202 mating behavior in a simultaneous hermaphrodite. Curr Biol. 2020;30:143–9.

1203 63. Weber M, Patlar B, Ramm SA. Effects of two seminal fluid transcripts on post-mating behaviour
1204 in the simultaneously hermaphroditic flatworm *Macrostomum lignano*. J Evol Biol. 2020;:jeb.13606.

1205 64. Holland B, Rice WR. Perspective: chase-away sexual selection: antagonistic seduction versus
1206 resistance. Evolution. 1998;52:1–7.

1207 65. Brand JN. Specimens from Brand et al. 2020, "Frequent origins of traumatic insemination involve
1208 convergent shifts in sperm and genital morphology". Zenodo.
1209 http://doi.org/10.5281/zenodo.XXXXX. 2020.

1210 66. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic
1211 algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–74.

1212 67. Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: polynomial time species tree
1213 reconstruction from partially resolved gene trees. BMC Bioinformatics. 2018;19 Suppl 6.

1214 68. Aberer AJ, Kobert K, Stamatakis A. Exabayes: Massively parallel bayesian tree inference for the
1215 whole-genome era. Mol Biol Evol. 2014;31:2553–6.

69. Lartillot N, Rodrigue N, Stubbs D, Richer J. PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. Syst Biol. 2013;62:611–5.

70. Vellnow N, Vizoso DB, Viktorin G, Schärer L. No evidence for strong cytonuclear conflict over sex allocation in a simultaneously hermaphroditic flatworm. BMC Evol Biol. 2017;17:103.

71. Zadesenets KS, Vizoso DB, Schlatter A, Konopatskaia ID, Berezikov E, Schärer L, et al. Evidence for karyotype polymorphism in the free-living flatworm, *Macrostomum lignano*, a model organism for evolutionary and developmental biology. Plos One. 2016;11:e0164915.

72. Robinson DF, Foulds LR. Comparison of phylogenetic trees. Math Biosci. 1981;53:131–47.

73. Irisarri I, Singh P, Koblmüller S, Torres-Dowdall J, Henning F, Franchini P, et al. Phylogenomics uncovers early hybridization and adaptive loci shaping the radiation of Lake Tanganyika cichlid fishes. Nat Commun. 2018;9:3159.

74. Singh P, Ballmer DN, Laubscher M, Schärer L. Successful mating and hybridisation in two closely related flatworm species despite significant differences in reproductive morphology and behaviour. Sci Rep. 2020;10:12830.

75. Pleijel F, Jondelius U, Norlinder E, Nygren A, Oxelman B, Schander C, et al. Phylogenies without roots? A plea for the use of vouchers in molecular phylogenetic studies. Mol Phylogenet Evol. 2008;48:369–71.

76. Xin F, Zhang S-Y, Shi Y-S, Wang L, Zhang Y, Wang A-T. *Macrostomum shenda* and *M. spiriger*, two new brackish-water species of *Macrostomum* (Platyhelminthes: Macrostomorpha) from China. Zootaxa. 2019;4603:105.

77. Janssen T, Vizoso DB, Schulte G, Littlewood DTJ, Waeschenbach A, Schärer L. The first multi-gene phylogeny of the Macrostomorpha sheds light on the evolution of sexual and asexual reproduction in basal Platyhelminthes. Mol Phylogenet Evol. 2015;92:82–107.

78. Leasi F, Sevigny JL, Laflamme EM, Artois T, Curini-Galletti M, de Jesus Navarrete A, et al. Biodiversity estimates and ecological interpretations of meiofaunal communities are biased by the taxonomic approach. Commun Biol. 2018;1:1–12.

79. Egger B, Ladurner P, Nimeth K, Gschwentner R, Rieger R. The regeneration capacity of the flatworm *Macrostomum lignano*—on repeated regeneration, rejuvenation, and the minimal size needed for regeneration. Dev Genes Evol. 2006;216:565–77.

80. Maddison WP, FitzJohn RG. The unsolved challenge to phylogenetic correlation tests for categorical characters. Syst Biol. 2015;64:127–36.

81. Uyeda JC, Zenil-Ferguson R, Pennell MW. Rethinking phylogenetic comparative methods. Syst Biol. 2018;67:1091–109.

82. Kamimura Y. Copulation anatomy of *Drosophila melanogaster* (Diptera: Drosophilidae): wound-making organs and their possible roles. Zoomorphology. 2010;129:163–74.

83. Schärer L. Tests of sex allocation theory in simultaneously hermaphroditic animals. Evolution. 2009;63:1377–405.

84. Ladurner P, Schärer L, Salvenmoser W, Rieger RM. A new model organism among the lower Bilateria and the use of digital microscopy in taxonomy of meiobenthic Platyhelminthes:

1255 *Macrostomum lignano*, n. sp. (Rhabditophora, Macrostomorpha). J Zool Syst Evol Res.
1256 2005;43:114–26.

1257 85. Janicke T, Marie-Orleach L, De Mulder K, Berezikov E, Ladurner P, Vizoso DB, et al. Sex
1258 allocation adjustment to mating group size in a simultaneous hermaphrodite. Evolution.
1259 2013;67:3233–42.

1260 86. Marie-Orleach L, Janicke T, Vizoso DB, David P, Schärer L. Quantifying episodes of sexual
1261 selection: insights from a transparent worm with fluorescent sperm. Evolution. 2016;70:314–28.

1262 87. Miller GT, Pitnick S. Sperm-female coevolution in *Drosophila*. Science. 2002;298:1230–3.

1263 88. Lüpold S, Manier MK, Berben KS, Smith KJ, Daley BD, Buckley SH, et al. How multivariate
1264 ejaculate traits determine competitive fertilization success in *Drosophila melanogaster*. Curr Biol.
1265 2012;22:1667–72.

1266 89. Manier MK, Lüpold S, Belote JM, Starmer WT, Berben KS, Ala-Honkola O, et al. Postcopulatory
1267 sexual selection generates speciation phenotypes in *Drosophila*. Curr Biol. 2013;23:1853–62.

1268 90. Parker GA, Immler S, Pitnick S, Birkhead TR. Sperm competition games: Sperm size (mass) and
1269 number under raffle and displacement, and the evolution of P2. J Theor Biol. 2010;264:1003–23.

1270 91. Willems M, Leroux F, Claeys M, Boone M, Mouton S, Artois T, et al. Ontogeny of the complex
1271 sperm in the macrostomid flatworm *Macrostomum lignano* (Macrostomorpha, Rhabditophora). J
1272 Morphol. 2009;270:162–74.

1273 92. Birkhead TR, Hosken DJ, Pitnick S. Sperm Biology. 2009.

1274 93. Reinhardt K, Dobler R, Abbott J. An Ecology of Sperm: Sperm Diversification by Natural
1275 Selection. Annu Rev Ecol Evol Syst. 2015;46:435–59.

1276 94. Hoogstraal H, Usinger RL. Monograph of Cimicidae (Hemiptera-Heteroptera). J Parasitol.
1277 1967;53:222.

1278 95. Schärer L, Ladurner P, Seifarth C, Salvenmoser W, Zaubzer J. Tracking sperm of a donor in a
1279 recipient: an immunocytochemical approach. Anim Biol. 2007;57:121–36.

1280 96. Giannakara A, Schärer L, Ramm SA. Sperm competition-induced plasticity in the speed of
1281 spermatogenesis. BMC Evol Biol. 2016;16. doi:10.1186/s12862-016-0629-9.

1282 97. Giannakara A, Ramm SA. Self-fertilization, sex allocation and spermatogenesis kinetics in the
1283 hypodermically inseminating flatworm *Macrostomum pusillum*. J Exp Biol. 2017;220:1568–77.

1284 98. Pitnick S, Markow TA, Spicer GS. Delayed male maturity is a cost of producing large sperm in
1285 Drosophila. Proc Natl Acad Sci. 1995;92:10614–8.

1286 99. Pitnick S. Investment in Testes and the Cost of Making Long Sperm in Drosophila. Am Nat.
1287 1996;148:57–80.

1288 100. Rohde K, Faubel A. Spermatogenesis of *Macrostomum pusillum* (Platyhelminthes,
1289 Macrostomida). Invertebr Reprod Dev. 1997;32:209–15.

1290 101. Rohde K, Watson N. Ultrastructure of spermatogenesis and sperm of *Macrostomum tuba*. J
1291 Submicrosc Cytol Pathol. 1991;23:23–32.

1292 102. Faubel A, Blome D, Cannon LRG. Sandy beach meiofauna of eastern Australia (southern
1293 Queensland and New South Wales). I. Introduction and macrostomida (Platyhelminthes). Invertebr
1294 Syst. 1994;8:899–1007.

1295 103. Anthes N, Schulenburg H, Michiels NK. Evolutionary links between reproductive morphology,
1296 ecology and mating behavior in opisthobranch gastropods. Evolution. 2008;62:900–16.

1297 104. Brennan PLR, Prum RO, McCracken KG, Sorenson MD, Wilson RE, Birkhead TR. Coevolution
1298 of Male and Female Genital Morphology in Waterfowl. PLOS ONE. 2007;2:e418.

1299 105. Arnqvist G, Rowe L. Correlated evolution of male and female morphologles in water striders.
1300 Evol Int J Org Evol. 2002;56:936–47.

1301 106. McPeek MA, Shen L, Farid H. The Correlated Evolution of Three-Dimensional Reproductive
1302 Structures Between Male and Female Damselflies. Evolution. 2009;63:73–83.

1303 107. Simmons LW, Fitzpatrick JL. Female genitalia can evolve more rapidly and divergently than
1304 male genitalia. Nat Commun. 2019;10:1312.

1305 108. Eberhard W. Female control: sexual selection by cryptic female choice. Princeton: Princeton
1306 University Press; 1996.
1307 http://www.degruyter.com/view/books/9780691207209/9780691207209/9780691207209.xml.
1308 Accessed 16 May 2020.

1309 109. Arnqvist G, Edvardsson M, Friberg U, Nilsson T. Sexual conflict promotes speciation in insects.
1310 Proc Natl Acad Sci. 2000;97:10460–4.

1311 110. Ritchie MG. Sexual Selection and Speciation. Annu Rev Ecol Evol Syst. 2007;38:79–102.

1312 111. Janicke T, Schärer L. Sperm competition affects sex allocation but not sperm morphology in a
1313 flatworm. Behav Ecol Sociobiol. 2010;64:1367–75.

1314 112. Tyler S, Schilling S, Hooge M, Bush L F. Turbellarian taxonomic database. 2006.

1315 113. Chambrier A, Zehnder M, Vaucher C, Mariaux J. The evolution of the Proteocephalidea
1316 (Platyhelminthes, Eucestoda) based on an enlarged molecular phylogeny, with comments on their
1317 uterine development. Syst Parasitol. 2004;57:159–71.

1318 114. Chambrier A de, Waeschenbach A, Fisseha M, Scholz T, Mariaux J. A large 28S rDNA-based
1319 phylogeny confirms the limitations of established morphological characters for classification of
1320 proteocephalidean tapeworms (Platyhelminthes, Cestoda). ZooKeys. 2015;500:25–59.

1321 115. Scarpa F, Cossu P, Sanna D, Lai T, Norenburg JL, Curini-Galletti M, et al. An 18S and 28S-
1322 based clock calibration for marine Proseriata (Platyhelminthes). J Exp Mar Biol Ecol. 2015;463:22–
1323 31.

1324 116. Templeton AR, Crandall KA, Sing CF. A cladistic analysis of phenotypic associations with
1325 haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram
1326 estimation. Genetics. 1992;132:619–33.

1327 117. Clement M, Posada D, Crandall KA. TCS: a computer program to estimate gene genealogies.
1328 Mol Ecol. 2000;9:1657–9.

1329 118. Vanhove MPM, Tessens B, Schoelinck C, Jondelius U, Littlewood DTJ, Artois T, et al.
1330 Problematic barcoding in flatworms: A case-study on monogeneans and rhabdocoels
1331 (Platyhelminthes). ZooKeys. 2013;:355–79.

1332 119. Simmons MP, Müller K, Norton AP. The relative performance of indel-coding methods in
1333 simulations. Mol Phylogenet Evol. 2007;44:724–40.

1334 120. Brand JN, Wiberg RAW, Pjeta R, Bertemes P, Beisel C, Ladurner P, et al. RNA-Seq of three
1335 free-living flatworm species suggests rapid evolution of reproduction-related genes. BMC Genomics.
1336 2020;21:462.

1337 121. Wudarski J, Simanov D, Ustyantsev K, de Mulder K, Grelling M, Grudniewska M, et al.
1338 Efficient transgenesis and annotated genome sequence of the regenerative flatworm model
1339 *Macrostomum lignano*. Nat Commun. 2017;8:2120.

1340 122. Grudniewska M, Mouton S, Grelling M, Wolters AHG, Kuipers J, Giepmans BNG, et al. A
1341 novel flatworm-specific gene implicated in reproduction in *Macrostomum lignano*. Sci Rep.
1342 2018;8:3192.

1343 123. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate: reference-free quality
1344 assessment of de novo transcriptome assemblies. Genome Res. 2016;26:1134–44.

1345 124. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO
1346 applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 2017.
1347 https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msx319/4705839. Accessed 8
1348 Dec 2017.

1349 125. Zaharia M, Bolosky WJ, Curtis K, Fox A, Patterson D, Shenker S, et al. Faster and more accurate
1350 sequence alignment with SNAP. ArXiv Prepr ArXiv11115572. 2011.

1351 126. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. *De novo*
1352 transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation
1353 and analysis. Nat Protoc. 2013;8:1494–512.

1354 127. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation
1355 sequencing data. Bioinformatics. 2012;28:3150–2.

1356 128. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons
1357 dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:157.

1358 129. Yang Y, Smith SA. Orthology inference in nonmodel organisms using transcriptomes and low-
1359 coverage genomes: improving accuracy and matrix occupancy for phylogenomics. Mol Biol Evol.
1360 2014;31:3081–92.

1361 130. Nakamura T, Yamada KD, Tomii K, Katoh K, Hancock J. Parallelization of MAFFT for large-
1362 scale multiple sequence alignments. Bioinformatics. 2018;34:2490–2.

1363 131. Kalyaanamoorthy S, Minh BQ, Wong TKF, Haeseler A von, Jermiin LS. ModelFinder: fast
1364 model selection for accurate phylogenetic estimates. Nat Methods. 2017;14:587–9.

1365 132. Wu M, Chatterji S, Eisen JA. Accounting for alignment uncertainty in phylogenomics. PLOS
1366 ONE. 2012;7:e30288.

133. Roure B, Baurain D, Philippe H. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol Biol Evol. 2013;30:197–214.

134. Tan G, Muffato M, Ledergerber C, Herrero J, Goldman N, Gil M, et al. Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. Syst Biol. 2015;64:778–91.

135. Wang H-C, Minh BQ, Susko E, Roger AJ. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. Syst Biol. 2018;67:216–35.

136. Plummer M, Best N, Cowles K, Vines K. CODA: convergence diagnosis and output analysis for MCMC. R News. 2006;6:7–11.

137. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.

138. Sanderson MJ. Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. Mol Biol Evol. 2002;19:101–9.

139. Smith SA, O'Meara BC. treePL: divergence time estimation using penalized likelihood for large phylogenies. Bioinformatics. 2012;28:2689–90.

140. Bollback JP. SIMMAP: Stochastic character mapping of discrete traits on phylogenies. BMC Bioinformatics. 2006;7:88.

141. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things): phytools: R package. Methods Ecol Evol. 2012;3:217–23.

142. Pagel M. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc R Soc Lond B. 1994;255:37–45.

143. Revell LJ. Size-correction and principal components for interspecific comparative studies. Evolution. 2009;63:3258–68.

144. Peres-Neto PR, Jackson DA, Somers KM. Giving meaningful interpretation to ordination axes: assessing loading significance in principal component analysis. Ecology. 2003;84:2347–63.

145. Ives AR. R^2s for Correlated Data: Phylogenetic Models, LMMs, and GLMMs. Syst Biol. 2019;68:234–51.

146. Ives A, Li D. rr2: An R package to calculate R^2s for regression models. J Open Source Softw. 2018;3:1028.

147. Wang A-T. Three new species of the genus *Macrostomum* from China (Platyhelminthes, Macrostomida, Macrostomidae). Acta Zootaxonomica Sin. 2005;30:714–20.

148. Sun T, Zhang L, Wang A-T, Zhang Y. Three new species of freshwater *Macrostomum* (Platyhelminthes, Macrostomida) from southern China. Zootaxa. 2015;4012:120–34.

149. Lin Y, Zhou W, Xiao P, Zheng Y, Lu J, Li J, et al. Two new species of freshwater *Macrostomum* (Rhabditophora: Macrostomorpha) found in China. Zootaxa. 2017;4329:267.

150. Lin Y-T, Feng W-T, Xin F, Zhang L, Zhang Y, Wang A-T. Two new species and the molecular phylogeny of eight species of *Macrostomum* (Platyhelminthes: Macrostomorpha) from southern China. Zootaxa. 2017;4337:423.

151. Wang L, Xin F, Fang C-Y, Zhang Y, Wang A-T. Two new brackish-water species of *Macrostomum* (Platyhelminthes, Macrostomida) from mangrove wetland in southern China. Zootaxa. 2017;4276:107.

Figures

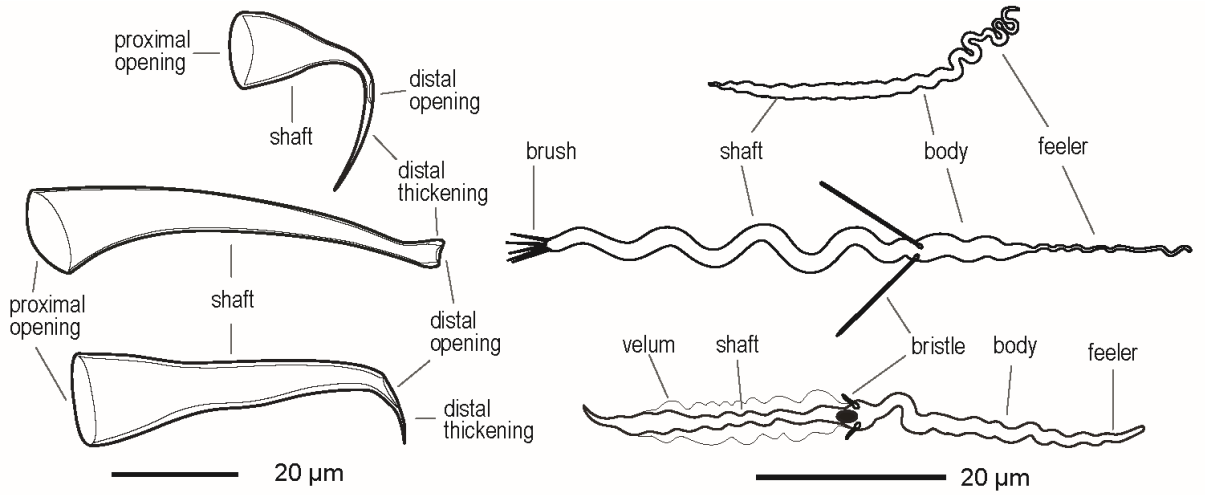


Fig 1. Representative drawings of the morphology of the stylet (male intromittent organ) (left) and the sperm (right) of three *Macrostomum* species.
Top: *M.* sp. 92, a hypodermically mating species from the hypodermic clade, with a typical needle-like stylet and a simple sperm morphology. Middle: The well-studied model organism *M. lignano* with the typical morphology for reciprocally mating species, showing a stylet with blunt distal thickenings and a complex sperm with an anterior feeler, two stiff lateral bristles, and a terminal brush. Bottom: *M.* sp. 9 representing one of the convergent origins of hypodermic insemination in the reciprocal clade, showing a stylet with a highly asymmetric and sharp distal thickening and sperm with reduced sperm bristles, no brush, but a thin velum along the shaft. Note that, given the striking diversity across the *Macrostomum* genus, it is not possible to clearly delimit all the sperm traits originally defined in *M. lignano* in some of the species.

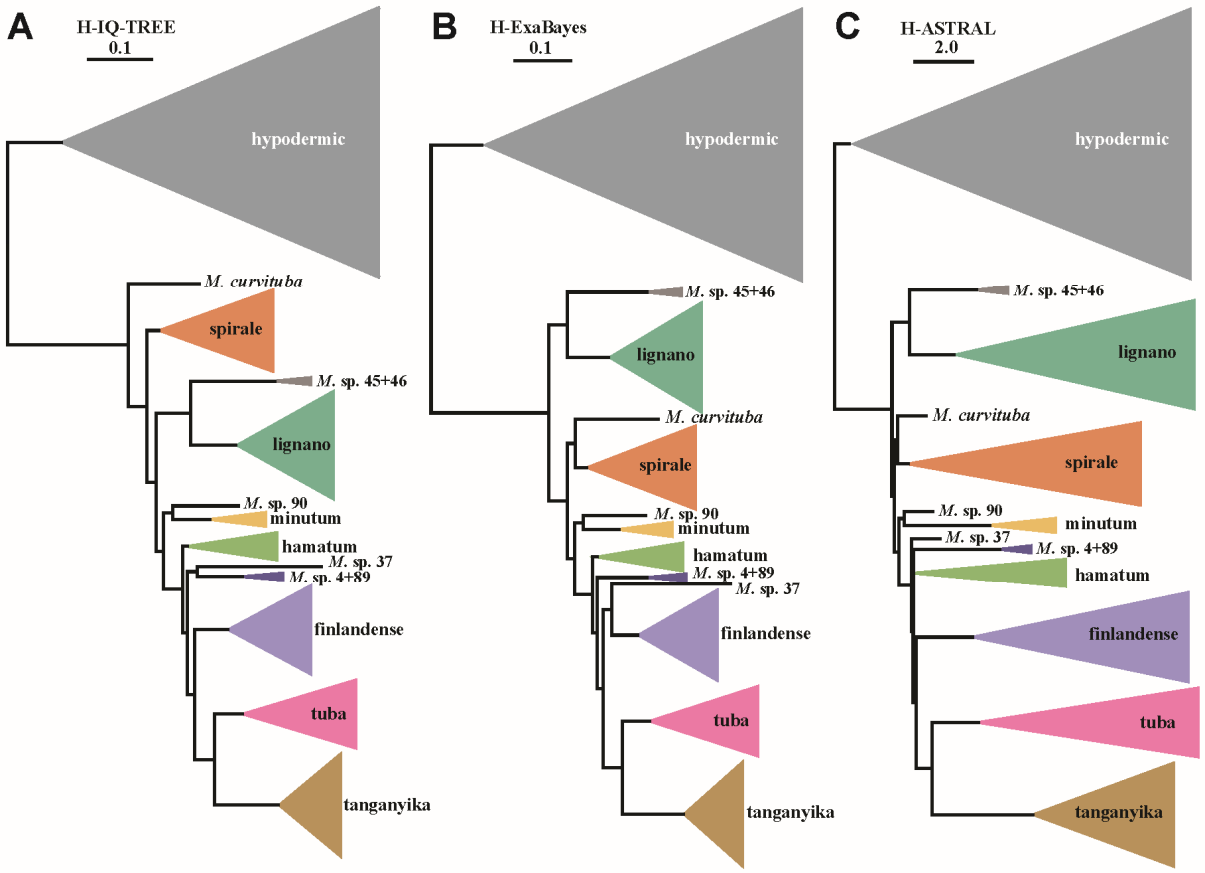
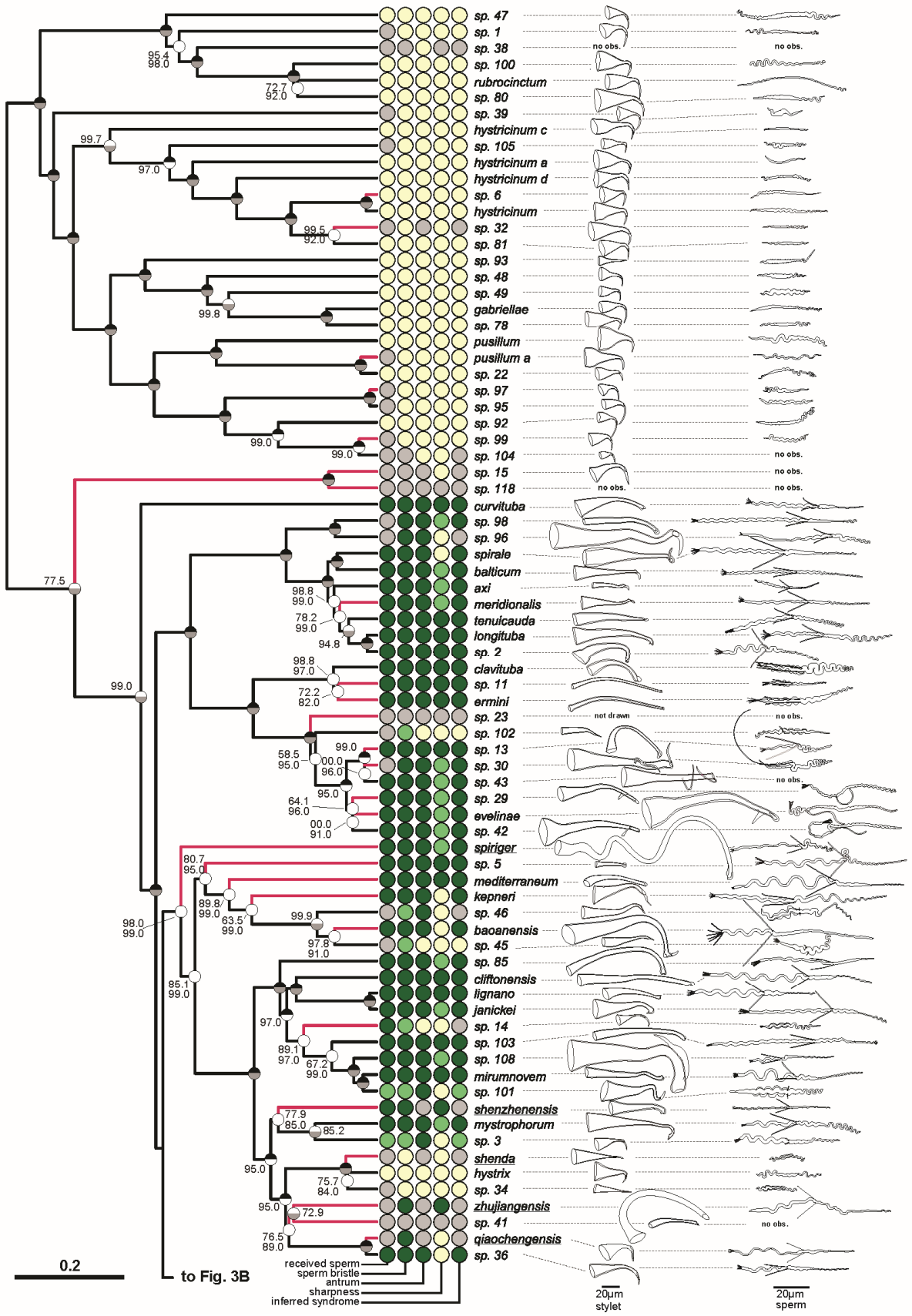
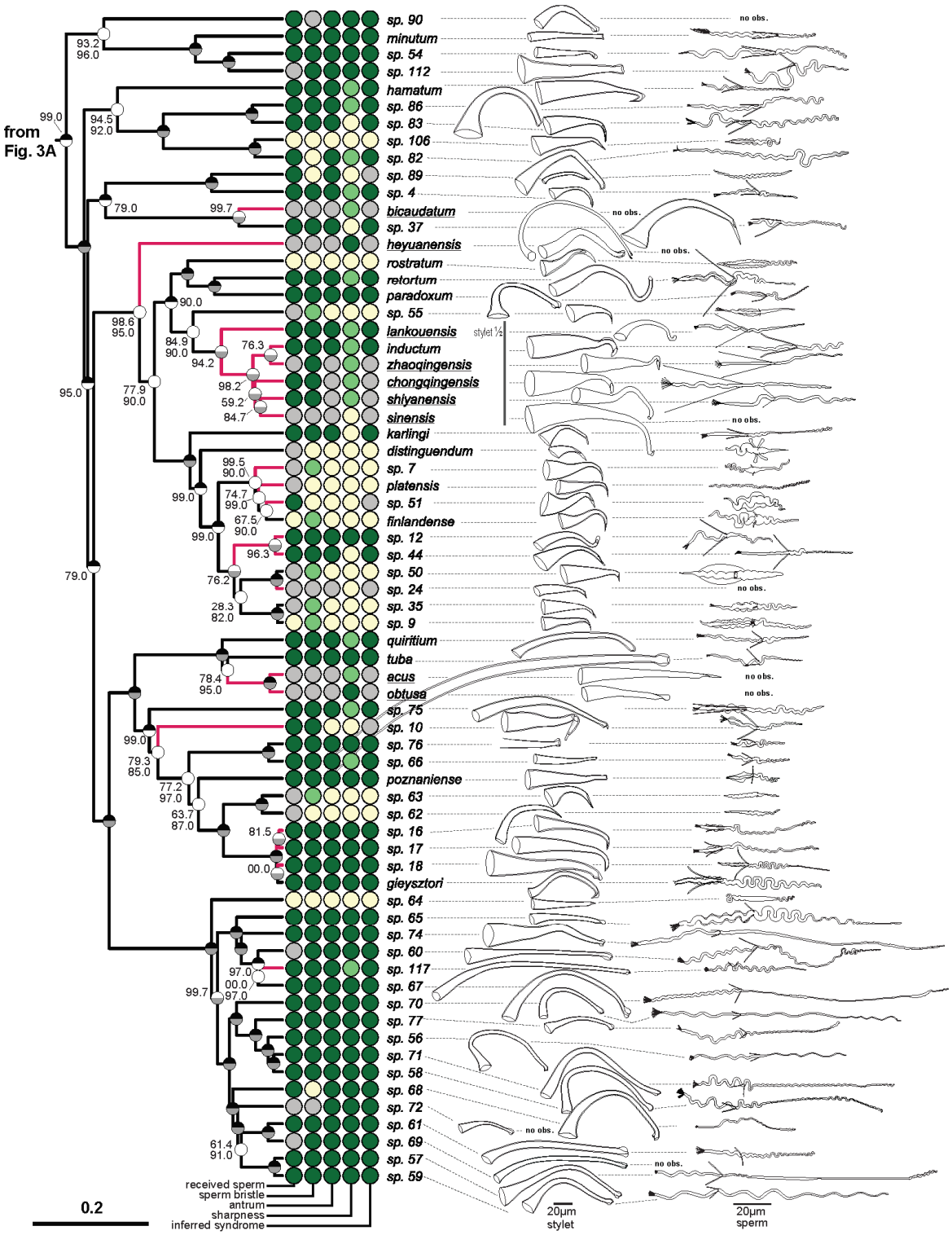


Fig 2. Simplified phylogenies of the genus *Macrostomum*.
Consistently recovered clades are collapsed into named triangles, with the height of the base proportional to the number of species, and single species names preceded by the genus abbreviation (in italics). Not shown here is the variable placement of *M.* sp. 39 within the hypodermic clade (see Figure S1). (A) Maximum-likelihood phylogeny (H-IQ-TREE) (B) Bayesian phylogeny (H-ExaBayes). (C) Summary method phylogeny (H-ASTRAL). Branch length represents substitutions per site in A & B and coalescent units in C. Support values for these groupings are generally high (for details see **Fig S1**).



1430



1431

Fig 3. Phylogeny of the genus *Macrostomum*, showing the striking diversity in stylet and sperm morphology across the genus.
The ultrametric phylogeny (C-IQ-TREE) includes all 145 studied species (with 77 species depicted in Fig 3A and 68 species in Fig 3B). Branch supports are ultrafast bootstraps (top, black if 100) and approximate likelihood ratio tests (bottom, grey if 100). Species without available transcriptomes that were added based on a *28S rRNA* fragment are indicated with red branches. Columns indicate the states of five reproductive traits from light to dark (i.e. yellow, light green and dark green for trinary states; or yellow and dark green for binary states; grey indicates missing data): received sperm location (hypodermic, both, in antrum), sperm bristle state (absent, reduced, present), antrum state (simple, thickened), sharpness of stylet (sharp, neutral, blunt), inferred mating syndrome (hypodermic, intermediate, reciprocal). Stylet and sperm are drawn based on our live observations, except for species with underlined names, which were redrawn based on the species description (*M. acus*, *M. obtusa* and *M. sinensis* from Wang 2005; *M. heyuanensis* and *M. bicaudatum* from Sun et al. 2015; *M. chongqingensis* and *M. zhaoqingensis* from Lin et al. 2017a; *M. shiyanensis* and *M. lankouensis* from Lin et al. 2017b; *M. shenzhenensis* and *M. qiaochengensis* from Wang et al. 2017; and *M. spiriger* and *M. shenda* from Xin et al. 2019). The stylet of *M*. sp. 15 is not drawn to scale, the stylets of some species are drawn at half size (stylet ½), and the stylet of *M*. sp. 23 is not drawn since it was incomplete (specimen ID MTP LS 913). Unobserved structures are marked as no observation (no obs.).

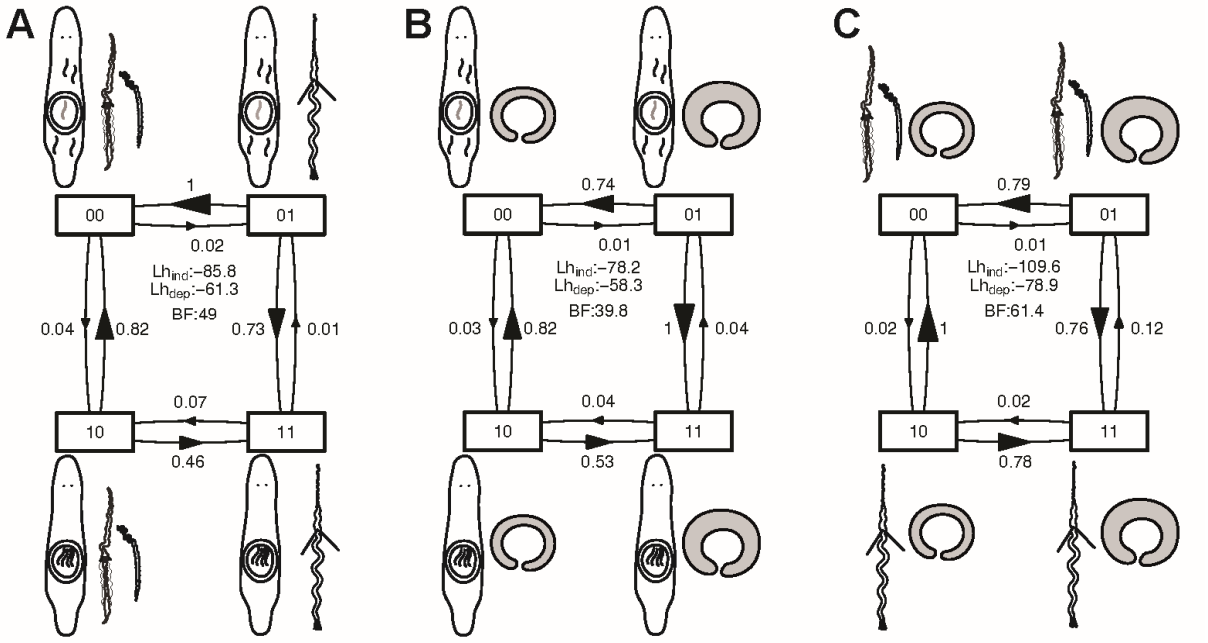


Fig 4. Results of correlated evolution analysis between (A) received sperm location and sperm bristle state, (B) received sperm location and antrum state, and (C) sperm bristle state and antrum state.
Shown are transition matrices for the dependent model from BayesTraits analysis, which was always preferred over the independent model. Transition rates are scaled so that the largest is unity (and arrow sizes are proportional). Also given are the likelihoods of the independent (Lh_{ind}) and dependent (Lh_{dep}) models, and the resulting BayesFactors (BF). An exponential prior and the C-IQ-TREE phylogeny was used for the results shown here. See SI Correlated evolution for runs with other priors (uniform and reversible-jump hyperprior) and other phylogenies (H-IQ-TREE and H-ExaBayes), which show qualitatively similar results.

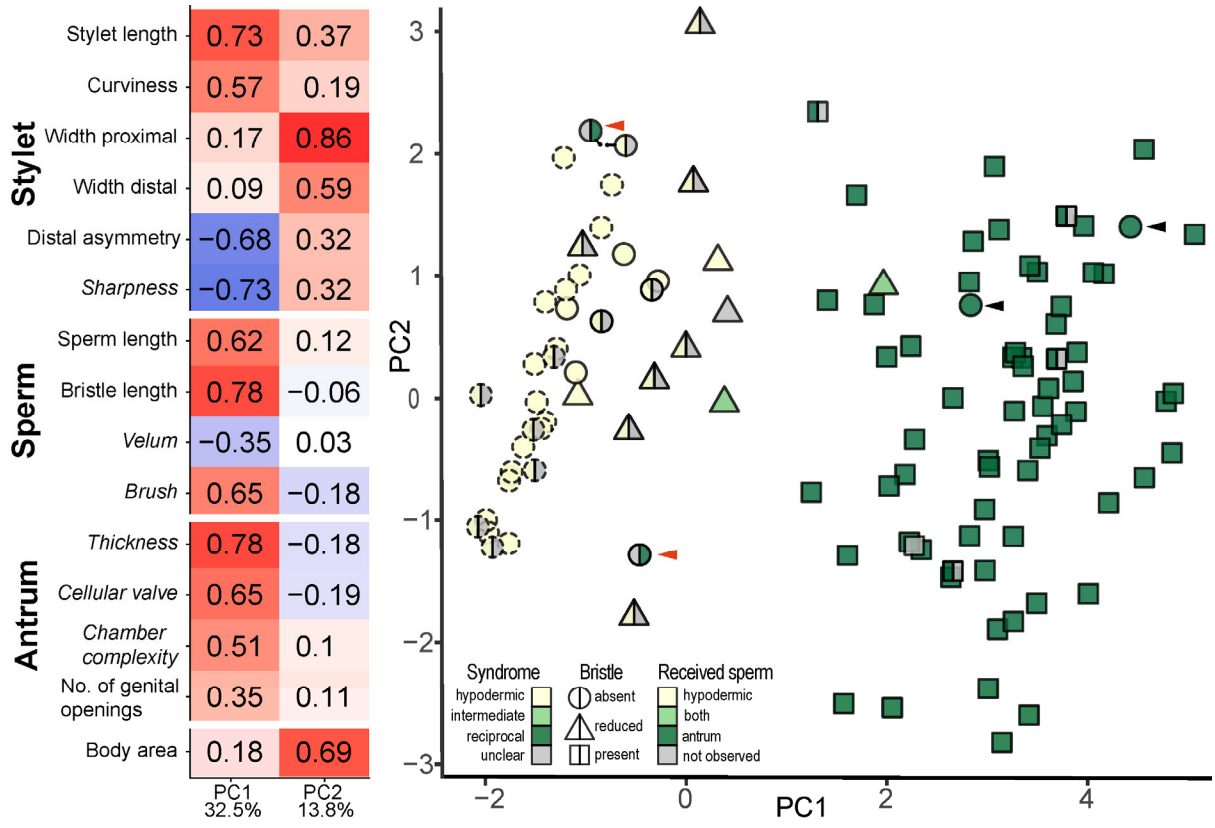


Fig 5. Results of a phylogenetically corrected principal component analysis of the measured quantitative and categorical (italics) reproductive traits.
Left: Loadings of PC1 and PC2, with the percentage of variance explained at the bottom. Right: 2D morphospace defined by PC1 and PC2. As indicated by the legend, the shape represents the sperm bristle state, while the colours represent the inferred mating syndrome (left side) and the received sperm location (right side). Species from the hypodermic clade are outlined with stippled lines. Red arrowheads indicate two species (*Macrostomum* sp. 51 and *M.* sp. 89) that cluster closely with species assigned to the hypodermic mating syndrome, but in which we observed received sperm in the antrum. Black arrowheads indicate two species (*M.* sp. 68 and *M.* sp. 82) assigned to the reciprocal syndrome, which have no discernible sperm bristles (see also Fig 6). The phylogenetic relationships of these species are represented as a phylomorphospace animation in Fig S4. Results shown here based on C-IQ-TREE, while detailed results including analyses with other phylogenies (H-IQ-TREE and H-ExaBayes) are in Tab S5**Fig S3**.

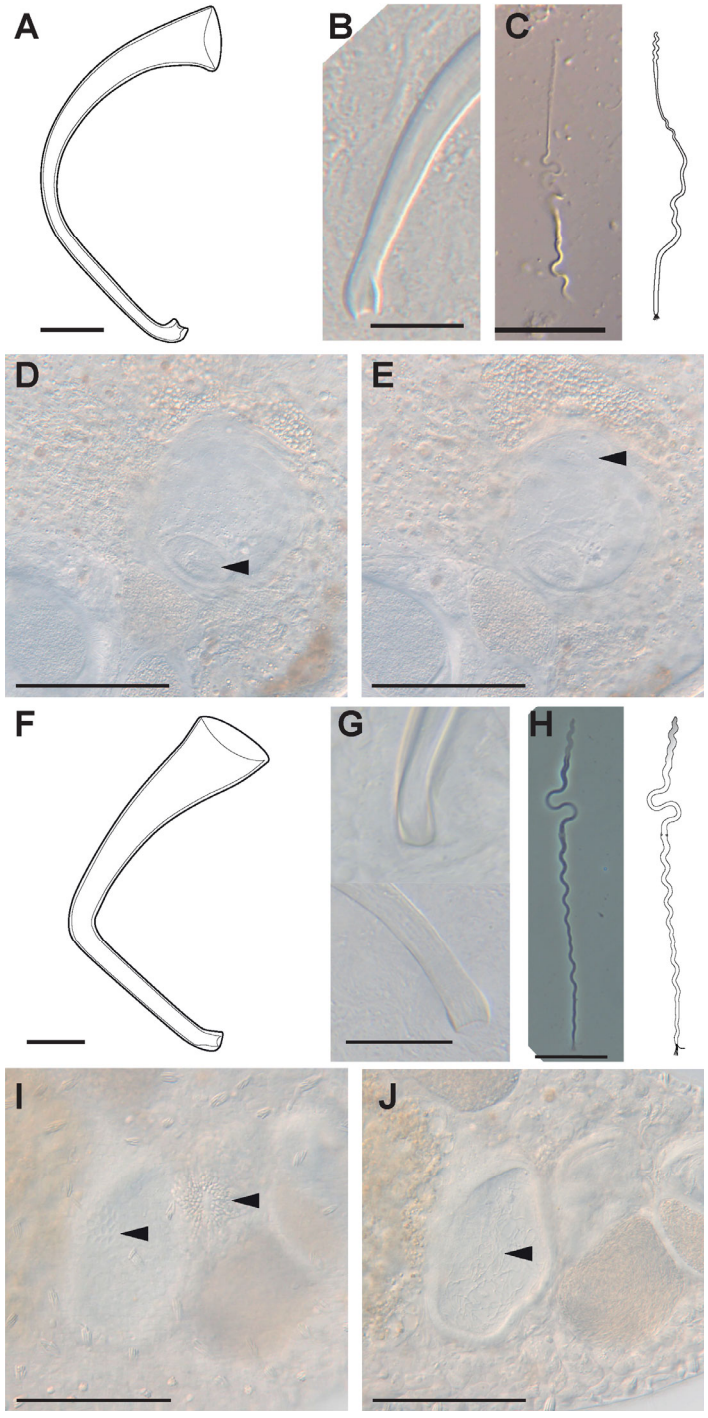


Fig 6. Details on the reproductive morphology of *Macrostomum* sp. 68 (A-E) and *M.* sp. 82 (F-J). *M.* sp. 68 (A) Stylet drawing showing the blunt distal thickenings; (B) distal stylet tip in a smash preparation (specimen ID MTP LS 2611). (C) Sperm image (MTP LS 2686) and drawing showing what seems to be a long feeler, but no apparent sperm bristles. (D-E). Details of the antrum (MTP LS 2562) indicating the muscular connection between the female genital opening and the antrum (arrowhead in D) and the anterior second chamber containing at least one received sperm (arrowhead in E). *M.* sp. 82 (F) Drawing of the stylet showing the slight blunt distal thickenings. (G) Distal stylet tip *in situ* (top, MTP LS 2845) and in a smash preparation (bottom, MTP LS 2846). (H) Sperm image (MTP LS 2877) and drawing indicating the modified anterior part of the sperm (shaded grey) and a less dense area approximately 1/3 along the sperm, which could be a vestigial bristle anchor location (arrowhead). (I-J) Details of the antrum (MTP LS 2848) indicating the anterior genital opening, the bursa pore (I, left arrowhead) next to the posterior genital opening, the gonopore (I, right arrowhead), both connecting into a large chamber containing many received sperm (J, arrowhead). Scale bars represent 100 µm in the antrum images and 20 µm otherwise.

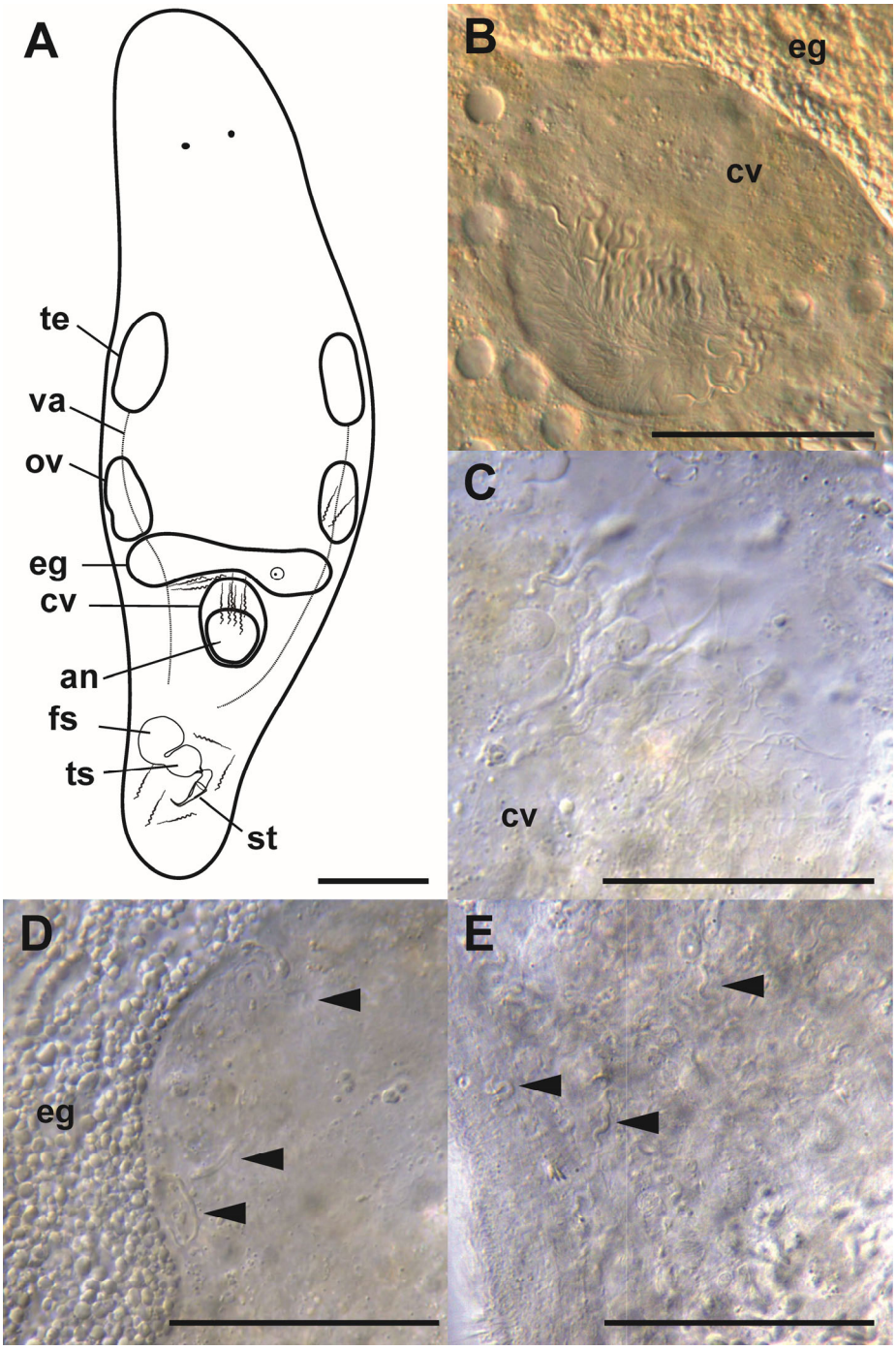


Fig 7. Detailed observations of received sperm location in *Macrostomum* sp. 3.
(A) Drawing of the sexual organs (te: testis, vd: vas deferens, ov: ovary, eg: developing egg, cv: cellular valve, an: antrum, fs: false seminal vesicle, ts: true seminal vesicle, st: stylet), with sperm drawn at locations where they were observed. (B-E) show sperm *in situ*. (B) sperm in the antrum (specimen ID MTP LS 3286) embedded in the cellular valve (anterior); (C) sperm in cellular valve with loose tissue (MTP LS 3314); (D) sperm close to developing egg (MTP LS 3314); (E) sperm close to the ovary (MTP LS 3317). Scale bars represent 100µm in (A) and 50 µm otherwise.

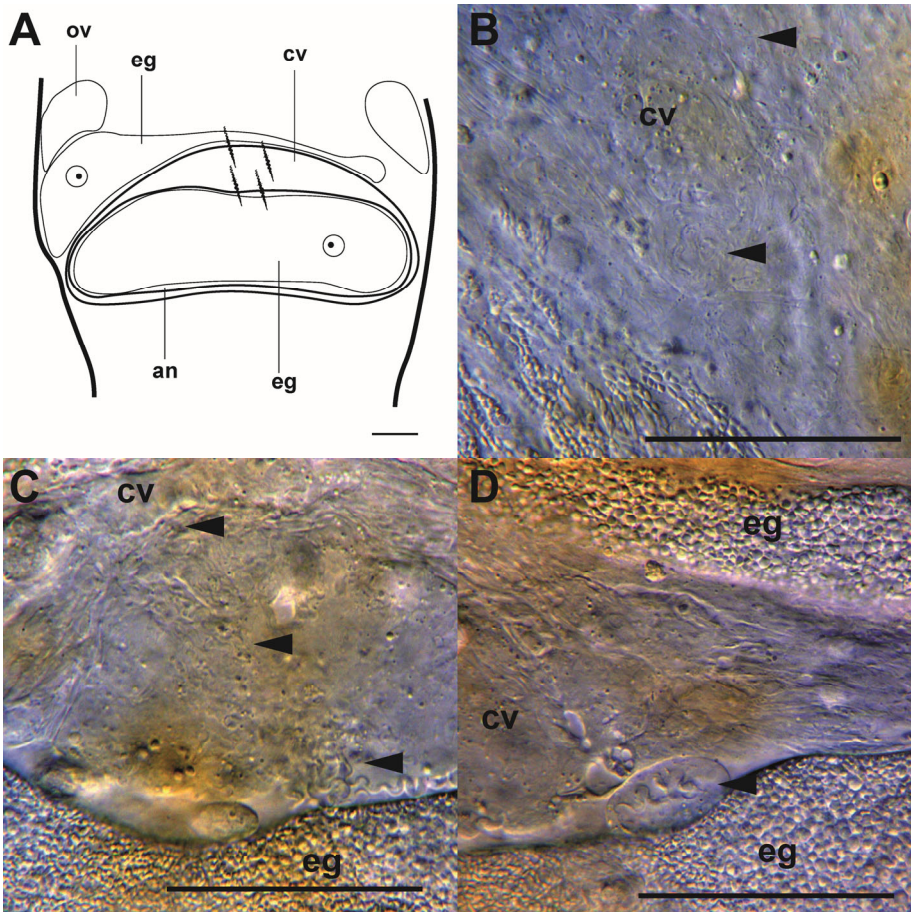
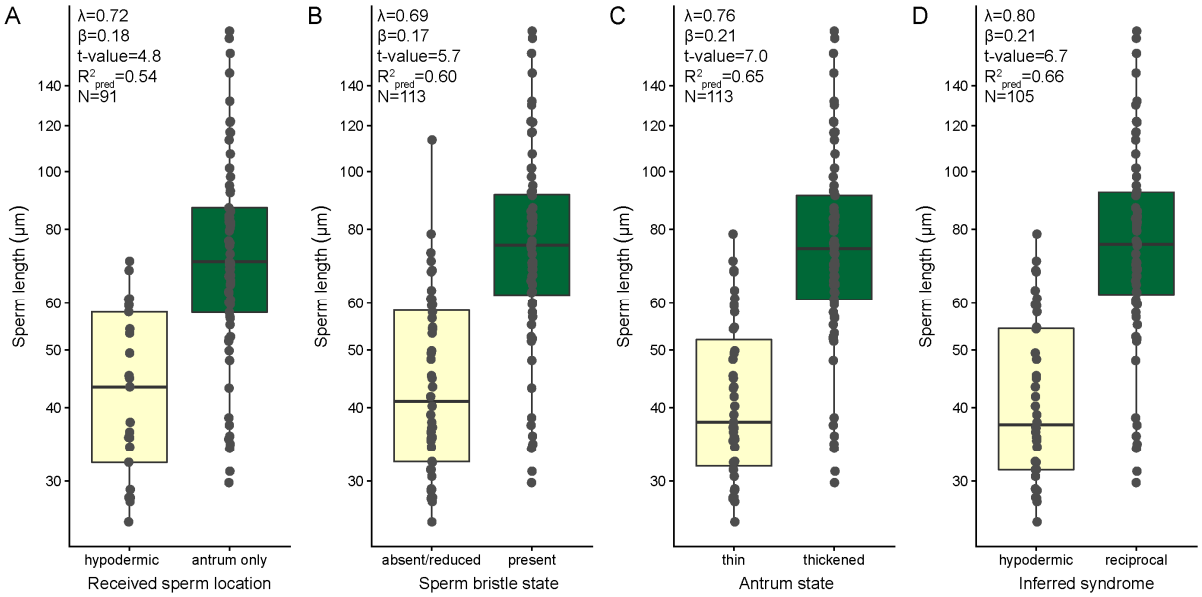


Fig 8. Detailed observations of received sperm location in *Macrostomum* sp. 101.
(A) Drawing of antrum region (ov: ovary, eg: developing egg, cv: cellular valve, an: antrum), with sperm drawn at locations where they were observed. (B-D) show sperm *in situ* with anterior on the top. (B) sperm (arrow) embedded in cellular valve (specimen ID MTP LS 3206); (C) sperm (arrow) in cellular valve and antrum (MTP LS 3127); (D) same specimen as C with dislodged sperm (arrow) surrounded by what appears to be cellular valve derived tissue. Scale bars represent 100µm in (A) and 50µm otherwise.

1500

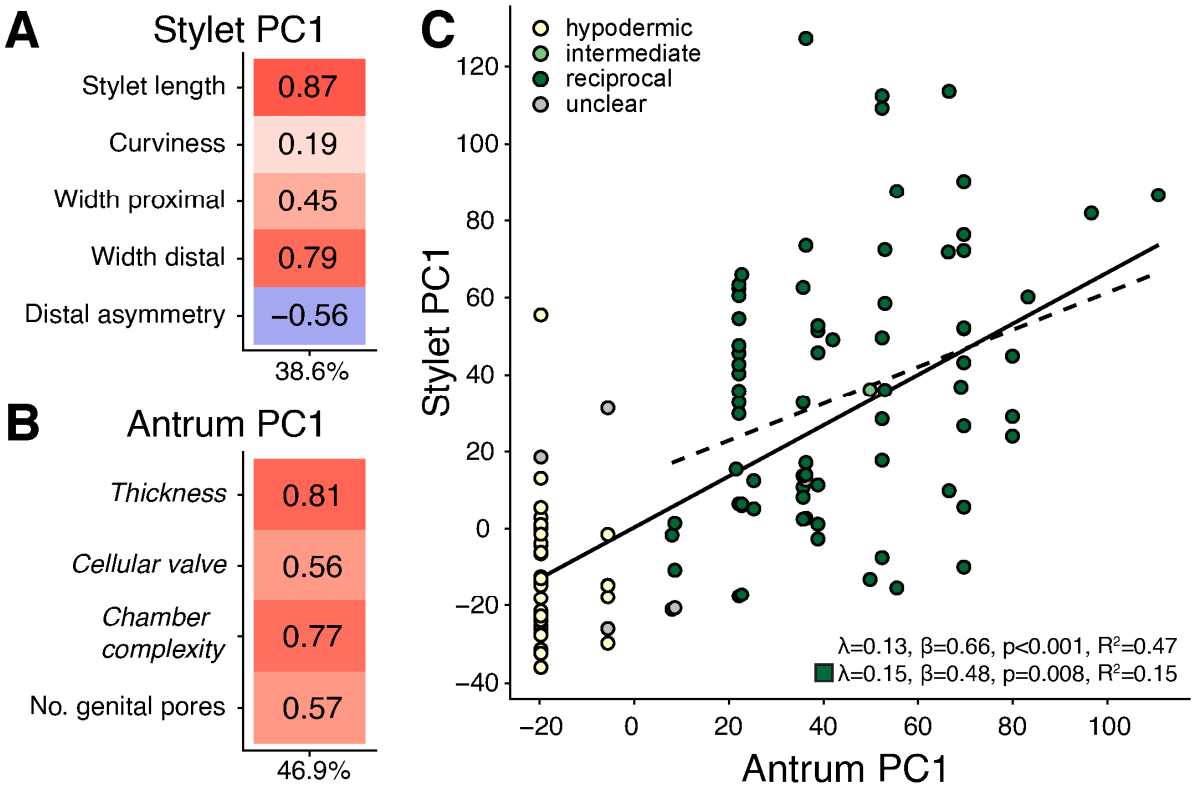


1501

1502 **Fig 9. Sperm length of species dependent on (A) sperm bristle state, (B) received sperm location (C) antrum state**
1503 **or (D) inferred mating syndrome.**
1504 Values are slightly jittered in the x direction, and the y-axis is on a log-scale. Within each panel the main results of
1505 PGLS analysis are given and in all tests the slopes were significant at $p < 0.001$. Detailed results including analyses with
1506 different phylogenies (H-IQ-TREE and H-ExaBayes) are given in Tab S6.

1507

1508



1509

1510 **Fig 10. Phylogenetically corrected principal component analyses of stylet (A) and antrum (B) traits, and evidence**
1511 **for male-female coevolution (C).**
1512 Loadings of Stylet and Antrum PC1, with the percentage of variance explained at the bottom, of the stylet (A) and
1513 antrum (B) traits, respectively. Categorical reproductive traits are in italics. (C) Shows the results from PGLS regression
1514 between Stylet PC1 and Antrum PC1 from (A-B). Regression was performed across all species (solid line, upper
1515 statistics) and restricted to species of the reciprocal mating syndrome (dashed line, lower statistics). The inferred mating
1516 syndrome of species is indicated by the colour of the dots. Results based on C-IQ-TREE, detailed results including
1517 analysis with other phylogenies (H-IQ-TREE and H-ExaBayes) are in Tab S7.

Tables

Tab 1. Summary of the taxonomic status of all the included *Macrostomum* species.
Species inferred based on single specimens (N=1) are listed separately.

Species in this study	N = 1	N > 1	Total
All	18	127	145
Described	5	46	51
Likely undescribed & immature	3	2	5
Likely undescribed, mature & no transcriptome	5	17	22
Likely undescribed, mature & transcriptome	5	62	67

Tab 2. Characteristics of the used protein alignments.
We used one alignment aimed at a high number of genes (L alignment) and one alignment aimed at high occupancy (H alignment). The alignments used were trimmed to only include regions with a high probability of being homologous using ZORRO, and the statistics are given for the alignments before and after trimming.

Alignment metrics	**L untrimmed**	**L**	**H untrimmed**	**H**
No. genes	8128	8128	385	385
No. amino acids AA	5,057,157	1,687,014	200,729	94,625
No. variable sites	3,263,955	1,157,689	135,887	74,175
No. parsimony informative sites	2,287,246	934,803	103,425	63,066
Missing data (%)	78.1	59.3	55.7	22.9

1531 **Tab 3. Assignment of the inferred mating syndrome based on different reproductive traits.**
1532 Species were assigned to an inferred mating syndrome based on the location of received sperm in the body (antrum, in
1533 the antrum only; hypodermic, hypodermic only; both, in the antrum and hypodermic; NA, no observation), the sperm
1534 bristle state (absent, reduced or present), the antrum state (simple or thickened), and the shape of the distal thickening of
1535 the stylet (sharp or blunt). 26 species with either not enough (22 species) or contradictory (four species) information
1536 were not assigned to a syndrome. Note, that all 24 species with only hypodermic sperm had the same morphological
1537 states, but this was not a condition for their assignment (hence the brackets). Similarly, all 69 species assigned to the
1538 reciprocal mating syndrome had a thickened antrum, but this was not a condition for their assignment.

	Received sperm location				Morphology			
Syndrome	**Antrum**	**Hypodermic**	**Both**	**NA**	**Sperm bristle**	**Antrum**	**Stylet**	**N**
Hypodermic		24			(Reduced/absent)	(Simple)	(Sharp)	24
Hypodermic				18	Reduced/absent	Simple	Sharp	18
Intermediate			2		Reduced	Thickened	Sharp	2
Reciprocal	61			6	Any state	(Thickened)	Blunt	67
Reciprocal	8				Present	(Thickened)	Sharp	8
Unclear	7			19	Other combinations			26

1539

1540

1541

Tab 4. Ancestral state reconstructions of reproductive traits, including received sperm location, sperm bristle state, antrum state, and inferred mating syndrome.
A range of MK-models were compared based on their AIC weights (ER: equal rate, SYM: symmetrical rate, ORD-Dollo: ordered model without gains once the trait is in state 0, Dollo: model without gains, ORD: ordered model, ARD: all rates different). For each trait the model with the highest AICc weight is shown in bold type, but we estimated the number of transitions between the states using stochastic character mapping with 1000 posterior samples for all models with an AICc weight >0.15. Given are the average number of transitions and the 2.5% and 97.5% quantiles in brackets. Results are based on the C-IQ-TREE phylogeny. For the quantitatively similar results with the H-IQ-TREE and H-ExaBayes phylogenies see Table S4.

Reproductive trait	model	dfs	log lik	AICc	ΔAICc	AICc weight	$N_{Species}$	$N_{Changes}$	0 →1	1→0	1→2	2→1	0→2	2→0
Received sperm location 0: hypodermic only 1: hypodermic & in antrum 2: in antrum only	ER	1	-43.3	88.6	10.1	0.004								
	SYM	3	-40.9	88.1	9.6	0.005								
	ORD-Dollo	**3**	**-36.1**	**78.5**	**0**	**0.579**	**102**	**21.2 (16, 32)**	**-**	**8.0 (7, 13)**	**4.0 (0, 11)**	**9.3 (7, 13)**	**-**	**-**
	Dollo	4	-36.1	80.6	2.2	0.196	102	10.8 (9, 17)	-	0.8 (0, 4)	0.9 (0, 4)	2.6 (2, 5)	-	6.5 (4, 8)
	ORD	4	-36.1	80.6	2.2	0.196	102	23.4 (16, 38)	0.7 (0, 4)	8.4 (6, 13)	4.8 (0, 15)	9.6 (7, 15)	-	-
	ARD	6	-36.1	85.1	6.6	0.021								
Received sperm location 0: hypodermic 1: in antrum only	ER	1	-36.8	75.6	3.1	0.135								
	Dollo	**1**	**-35.2**	**72.5**	**0**	**0.638**	**102**	**9.3 (9, 11)**	**-**	**9.3 (9, 11)**				
	ARD	2	-35.2	74.5	2.1	0.227	102	9.9 (9, 14)	0.9 (0, 4)	9.0 (8, 12)				
Sperm bristle state 0: absent 1: reduced 2: present	ER	1	-82.9	167.8	27.2	0								
	SYM	3	-81.7	169.5	28.9	0								
	ORD-Dollo	**3**	**-67.2**	**140.6**	**0**	**0.578**	**131**	**36.3 (29, 48)**	**-**	**12.2 (11, 17)**	**6.7 (1, 16)**	**17.5 (14, 23)**	**-**	**-**
	Dollo	4	-67.2	142.7	2.1	0.2	131	29.3 (22, 41)	-	7.1 (3, 13)	5.3 (1, 13)	12.4 (8, 18)	-	4.6 (0, 8)
	ORD	4	-67.2	142.7	2.1	0.199	131	37.1 (29, 50)	0.8 (0, 5)	13.0 (10, 20)	6.0 (1, 14)	17.3 (14, 22)	-	-
	ARD	6	-67.2	147.1	6.5	0.023								
Sperm bristle state 0: absent or reduced 1: present	ER	1	-57.6	117.3	5.7	0.038								
	Dollo	**1**	**-54.8**	**111.6**	**0**	**0.635**	**131**	**18.8 (18, 22)**	**-**	**18.8 (18, 22)**				
	ARD	2	-54.4	112.9	1.3	0.327	131	20.5 (17, 29)	2.7 (0, 9)	17.8 (15, 23)				
Antrum state 0: simple 1: thickened	ER	1	-48.1	98.2	3.8	0.086								
	Dollo	**1**	**-46.2**	**94.4**	**0**	**0.577**	**127**	**14.7 (14, 17)**	**-**	**14.7 (14, 17)**				
	ARD	2	-45.7	95.5	1.1	0.337	127	15.2 (13, 21)	2.1 (0, 6)	13.1 (11, 16)				
Inferred mating syndrome 0: hypodermic 1: intermediate 2: reciprocal	ER	1	-59.3	120.6	16.2	0								
	SYM	3	-54.4	114.9	10.4	0.003								
	ORD-Dollo	**3**	**-49.1**	**104.5**	**0**	**0.578**	**119**	**34.1 (27, 47)**	**-**	**13.7 (12, 18)**	**7.1 (2, 16)**	**13.2 (10, 19)**	**-**	**-**
	Dollo	4	-49.1	106.6	2.1	0.199	119	16.6 (14, 25)	-	0.7 (0, 5)	1.1 (0, 6)	2.8 (2, 6)	-	11.9 (9, 15)
	ORD	4	-49.1	106.6	2.1	0.198	119	35.4 (27, 50)	1.0 (0, 5)	14.3 (11, 20)	7.2 (2, 16)	12.9 (9, 18)	-	-
	ARD	6	-49.1	111	6.5	0.022								
Inferred mating syndrome 0: hypodermic & intermediate 1: reciprocal	ER	1	-49.4	100.9	14.675	0.001								
	Dollo	**1**	**-42.1**	**86.26**	**0**	**0.997**	**119**	**14.7 (14, 17)**	**-**	**14.7 (14, 17)**				
	ARD	2	-47.3	98.68	12.421	0.002								

Supp. Figures

Fig S1. Phylogenetic trees inferred using various methods on the L and H alignments. Clades are coloured consistently between the panels, with six large species groups (hypodermic, dark grey; spirale, orange; lignano, dark green; finlandense, purple; tuba, pink; tanganyika, brown), two smaller species groups (minutum, yellow; hamatum, light green), and two consistent species pairs (*M*. sp. 45 + 46, light grey; *M*. sp. 4 + 89, dark blue). Transcriptomes that are placed differently between approaches are highlighted in red. Species assignments are given as a short six letter abbreviation in the name of the transcriptomes (three letters each for the genus and species, respectively). See

Tab S8 for details.

see the file Fig_S1.pdf

Fig S2. The two ASTRAL phylogenies with quartet support shown for each node.
(A) L-ASTRAL phylogeny, and (B) H-ASTRAL phylogeny. The pie charts at the nodes represent the quartet support for the three possible partitions at that node. Blue represents support for the species tree topology and orange and red represent support for one or the other of the two alternative partitions. Mostly blue pie charts indicate little conflict. This analysis indicates that there is substantial gene tree – species tree conflict throughout the genus.

see the file Fig_S2.pdf

Fig S3. Ancestral state reconstructions of reproductive traits using the C-IQ-TREE phylogeny.
The trait and type of scoring (binary/trinary) is indicated at the bottom of each panel. Stochastic character mapping is summarised with pie charts representing the proportion of stochastic maps with the respective state. Shown is the reconstruction of the best-fitting ordered model without losses. The average number of transitions is given in Tab 4, while the red stars and numbers indicate the lower-bound number of transitions that have likely occurred (i.e. separated by nodes with >95% posterior probability of the ancestral state), while acknowledging that the ancestral state of the genus is often unclear (hence the brackets). The colours indicate the six large species groups (hypodermic, dark grey; spirale, orange; lignano, dark green; finlandense, purple; tuba, pink; tanganyika, brown), two smaller species groups (minutum, yellow; hamatum, light green), and two consistent species pairs (*M*. sp. 45 + 46, light grey; *M*. sp. 4 + 89, dark blue).

see the file Fig_S3.pdf

Fig S4. Animation of the phylomorphospace represented by PC1 and PC2 of the species in the C-IQ-TREE phylogeny.
The animation initially shows a cladogram that then gradually transforms into the phylomorphospace, which was calculated using the phylomorphospace function in phytools.

see the file Fig_S4.gif

Supp. Tables

Tab S1. The number of specimens analysed per *Macrostomum* species for all the included quantitative traits.

see the file Tab_S1.xlsx

Tab S2. Details on all specimens included in this study.

see the file Tab_S2.xlsx

Tab S3. Robinson-Foulds distance matrix between the various phylogenetic inference methods.
The first row gives the alignment used for inference (L or H) and the second row the inference software used.

see file Tab_S3.xlsx

Tab S4. Ancestral state reconstruction using stochastic character mapping.

see file Tab_S4.xlsx

Tab S5. Scores and loadings from the phylogenetically corrected principal component analysis.

see file Tab_S5.xlsx

Tab S6. Results of PGLS analysis of states indicating reciprocal copulation versus hypodermic insemination on sperm length.
All predictors were binary, with the reference level being the state indicating hypodermic insemination.

see file Tab_S6.xlsx

Tab S7. Results from PGLS correlating the first principal components of a phylogenetically corrected principal component analysis (pPCA) analysis including five stylet traits with the first principal component of a pPCA analysis including four antrum traits.
Analysis was performed across all species and restricted to the reciprocal mating syndrome. Also given are pPCA loadings and results for all three phylogenies.

see file Tab_S7.xlsx

Tab S8. Details on the transcriptomes used in this study.

see file Tab_S8.xlsx

Tab S9. Mean species values for all morphological variables.

see file Tab_S9.xlsx