Increased taxon sampling reveals thousands of hidden orthologs in flatworms

José M. Martín-Durán[1§], Joseph F. Ryan[1,2§], Bruno C. Vellutini[1], Kevin Pang[1], Andreas Hejnol[1*]

[1]Sars International Centre for Marine Molecular Biology, University of Bergen, Thormøhlensgate 55, Bergen, 5006, Norway

[2]Whitney Laboratory for Marine Bioscience, University of Florida, 9505 Ocean Shore Blvd., St Augustine, FL, 32080, USA

[§]These authors contributed equally to this work

*Corresponding author: Andreas Hejnol (andreas.hejnol@uib.no)

Running title: Uncovered hidden orthology in flatworms

Keywords: Gene losses, orphan genes, genome evolution, BLAST, orthology, Platyhelminthes, *Schmidtea mediterranea*, centrosome.

Abstract

Gains and losses shape the gene complement of animal lineages and are a fundamental aspect of genomic evolution. Acquiring a comprehensive view of the evolution of gene repertoires is limited by the intrinsic limitations of common sequence similarity searches and available databases. Thus, a subset of the gene complement of an organism consists of hidden orthologs, i.e., those with no apparent homology to sequenced animal lineages –mistakenly considered new genes– but actually representing rapidly evolving orthologs or undetected paralogs. Here, we describe Leapfrog, a simple automated BLAST pipeline that leverages increased taxon sampling to overcome long evolutionary distances and identify putative hidden orthologs in large transcriptomic databases by transitive homology. As a case study, we used 35 transcriptomes of 29 flatworm lineages to recover 3,427 putative hidden orthologs, some of them not identified by OrthoFinder and HaMStR, two common orthogroup inference algorithms. Unexpectedly, we do not observe a correlation between the number of putative hidden orthologs in a lineage and its 'average' evolutionary rate. Hidden orthologs do not show unusual sequence composition biases that might account for systematic errors in sequence similarity searches. Instead, gene duplication with divergence of one paralog and weak positive selection appear to underlie hidden orthology in Platyhelminthes. By using Leapfrog, we identify key centrosome-related genes and homeodomain classes previously reported as absent in free-living flatworms, e.g. planarians. Altogether, our findings demonstrate that hidden orthologs comprise a significant proportion of the gene repertoire in flatworms, qualifying the impact of gene losses and gains in gene complement evolution.

Introduction

Understanding the dynamic evolution of gene repertoires is often hampered by the difficulties of confidently identifying orthologous genes, i.e. those genes in different species originated by speciation (Fitch 1970). Gene annotation pipelines and large-scale comparisons largely rely on sequence-similarity approaches for gene orthology assignment (Alba and Castresana 2007; Domazet-Loso et al. 2007; Tautz and Domazet-Loso 2011; Yandell and Ence 2012). These approaches depend on taxonomic coverage and the completeness of the gene databases used for comparisons. Although extremely useful in many contexts, sequence-similarity methods, such as Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990), can be confounded in situations in which a gene evolves fast, is short, has an abundance of indels and/or exhibits similarity with other gene family members in only a small subset of residues (Moyers and Zhang 2015). These limitations, as well as intrinsic systematic errors of most algorithms (Liebeskind et al. 2016), can generate significant biases when studying the evolution of protein-coding gene families (Elhaik et al. 2006; Moyers and Zhang 2015; Moyers and Zhang 2016). Accordingly, a proportion of the gene complement of an organism will be represented by genes that lack obvious affinity with orthologs in the gene sets of the best annotated genomes –thus mistakenly considered potential 'taxonomically-restricted', or 'orphan', genes (Khalturin et al. 2009; Tautz and Domazet-Loso 2011)– but actually representing fast evolving orthologs that we call hidden orthologs. This issue can potentially be overcome by more sensitive, although computationally intense, detection methods (e.g. profile HMMs, PSI-BLAST) (Altschul et al. 1997; Eddy 2011; Kuchibhatla et al. 2014), but also by increasing taxon sampling, which helps to bridge the long evolutionary gaps between hidden orthologs and their well-annotated, more

conservative counterparts, e.g. via transitive homology (Pearson 1996; Yona et al. 1998) (Fig. 1A).

Platyhelminthes (flatworms) exhibit significantly high rates of molecular evolution (Edgecombe et al. 2011; Struck et al. 2014; Laumer et al. 2015a). Gene loss and orphan genes have been attributed to evolutionary changes leading to Platyhelminthes morphology (Berriman et al. 2009; Martin-Duran and Romero 2011; Riddiford and Olson 2011; Tsai et al. 2013; Breugelmans et al. 2015). A prime example is the loss of centrosomes in planarian flatworms, where the apparent absence of genes critical to the functioning of animal centrosomes was used as evidence supporting the secondary loss of these organelles in Platyhelminthes (Azimzadeh et al. 2012). Recently, two phylogenomic analyses have provided an extensive transcriptomic dataset for most platyhelminth lineages, in particular for those uncommon and less studied taxa that otherwise occupy key positions in the internal relationships of this group (Egger et al. 2015; Laumer et al. 2015b). These important resources provide an ideal opportunity to address how increasing taxon sampling may improve the resolution of gene complement evolution in a fast evolving –and thus more prone to systematic error– animal group.

Results

The Leapfrog pipeline identifies hundreds of hidden orthologs in flatworms

To identify hidden orthologs in large transcriptomic datasets we created Leapfrog, a simple pipeline that automates a series of BLAST-centric processes (Fig. 1B; Methods). We assembled a dataset including 35 publicly available transcriptomes from 29 flatworm species, and incorporated the transcriptomes of the gastrotrich *Lepidodermella squamata*, the rotifer *Lepadella patella*, and the gnathostomulid *Austrognathia* sp. as

closely related outgroup taxa (Struck et al. 2014; Laumer et al. 2015a) (Supplemental Table 1). Using a single-isoform version of the human RefSeq protein dataset as initial queries, Leapfrog identified a total of 3,427 putative hidden orthologs, 1,217 of which were unique and 636 were species-specific (Fig. 2A, B; Supplemental Table 2). In 30 cases (less than 1% of the total recovered hidden orthologs), Leapfrog associated two or more human proteins, likely paralogs, ohnologs or inparalogs, with the same hidden ortholog and 'bridge' contig (Supplemental Table 3). Alignments of recovered hidden orthologs with their human and *P. vittatus* counterparts show that many amino acid positions that differ between the human and the hidden ortholog products are conserved between *P. vittatus* and one or the other sequences (e.g., Fig. 2C).

The number of putative hidden orthologs recovered in each particular lineage ranged from 41 in the rhabdocoel *Provortex sphagnorum* to 198 in the planarian *S. mediterranea*, and varied considerably between different species belonging to the same group of flatworms (Supplemental Fig. S1). However, completeness of each transcriptome and sequencing depth were also very irregular (Supplemental Fig. S1; Supplemental Table 1), suggesting that the number of putative hidden orthologs we recovered with Leapfrog is sensitive to the quality of the transcriptomes, perhaps influenced by the source tissue(s) used for transcriptome sequencing.

Accuracy of Leapfrog and impact of the 'bridge' transcriptome

In order to ensure that domain shuffling was not a source of false positives, we analyzed the domain content of 130 'bridge' proteins and their corresponding human sequences from the *S. mediterranea* analysis (Supplemental Table 4). In most cases (112; 86.15%), human and 'bridge' proteins had identical domain content. In the other 18 cases, the

'bridge' ortholog had a domain not reported in the human protein or vice versa. Nonetheless, many of these may be due to InterProScan 5 (Jones et al. 2014) sensitivity. For example, in the case of *BATF2*, InterProScan 5 identified the bZIP domain bZIP_1 (PF00170) in the human *BATF2* and bZIP_2 (PF07716) in the 'bridge' sequence; however, bZIP_1 and bZIP_2 are very similar and both domains are recovered in an identical region of *BATF2*. In general, there was agreement between the domain architecture of the human and bridge sequences, and we found no evidence suggesting that these results were inflated due to homologous over-extension (Gonzalez and Pearson 2010).

To recover even more putative hidden orthologs in our dataset, we implemented an iterative approach that employed each flatworm transcriptome with ≥ 85% CEGs as 'bridge' and 'target' in Leapfrog. This approach increased the number of recovered hidden orthologs 3.4 times (4,216 versus 1,240) with respect to the original Leapfrog with a constant, single 'bridge' lineage (Supplemental Table 5). Each hidden ortholog was identified on average by 2.11 different 'bridges', which suggests that each 'bridge' transcriptome is only suitable to retrieve by transitive homology a subset of the total amount of hidden orthologs present in a given 'target' lineage. Noticeably, the 'bridge' *M. lineare*, the flatworm species with the shortest evolutionary distance (Laumer et al. 2015b), recovered less hidden orthologs than *P. vittatus* used as 'bridge', suggesting that evolutionary rate is not necessarily the best criteria for choosing a 'bridge' lineage.

Leapfrog identifies orthologs not detected by OrthoFinder and HaMStR

In order to be certain that we were identifying orthologs that would not be detected in a typical analysis, we compared our pipeline with the commonly deployed orthogroup

inference algorithm OrthoFinder (Emms and Kelly 2015). To do this, we first performed an OrthoFinder analysis on human and the planarian *S. mediterranea*, and then evaluated the impact of adding the 'bridge' species *P. vittatus* to the calculation of orthogroups. Initially, OrthoFinder identified 5,638 orthogroups containing at least one sequence of *H. sapiens* and *S. mediterranea* (Supplemental Fig. S2). The inclusion of the translated transcripts of *P. vittatus* led to an increase in orthogroups that included both human and planarian sequences (5,816; Supplemental Fig. S2). However, OrthoFinder only recovered 82.7% (62/75) of the putative hidden orthologs identified by the single 'bridge' Leapfrog pipeline in the same *S. mediterranea* transcriptome (Brandl et al. 2016) with a similar E-value cutoff. Remarkably, the inclusion of the entire dataset in OrthoFinder reduced the number of orthogroups shared by human and *S. mediterranea* (5,213 orthogroups), and retrieved less putative hidden orthologs (56 out of 75; 74.67%).

We also performed a HaMStR analysis (Ebersberger et al. 2009) using *S. mediterranea* with default parameters, and evaluated how many of the Leapfrog putative hidden orthologs it was able to recover. In the initial HMM step only two of the 75 *S. mediterranea* Leapfrog hidden orthologs were present in the HMM outputs above the threshold, with an additional two below the threshold. Therefore, HaMStR is also not as sensitive to hidden orthologs as Leapfrog.

The number of hidden orthologs does not relate to the branch length of each lineage

To investigate the parameters that might influence the appearance and identification of putative hidden orthologs in our dataset, we performed a principal component analysis (PCA) including variables related to the quality and completeness of the transcriptome,

the mean base composition of the transcriptome and the evolutionary rate of each lineage (Fig. 3A; Supplemental Table 6). The first principal component (PC1) was strongly influenced by the quality of the transcriptome, while the second principal component (PC2) mostly estimated the balance between evolutionary change (branch lengths and hidden orthologs) and transcriptome complexity (GC content). The two first principal components explained 67% of the variance of the dataset, indicating that additional interactions between the variables exist (e.g. the GC content can affect sequencing performance (Dohm et al. 2008; Benjamini and Speed 2012), and thus transcriptome quality and assembly).

Despite the fact that the branch length of a given lineage and the number of putative hidden orthologs affected the dispersion of our data in a roughly similar manner, we did not detect a linear correlation ($R^2 = 0.131$, p-value = 0.053; Fig. 3B) between these two variables, not even when we considered those transcriptomes with similar completeness ($\geq$ 85% CEGs identified; $R^2 = 0.331$, p-value = 0.04). This result supported our previous observation that lineages with similar branch lengths could exhibit remarkably different sets of hidden orthologs (Supplemental Fig. S1).

Flatworm hidden orthologs do not show sequence composition biases

A recent report showed that very high GC content and long G/C stretches characterize genes mistakenly assigned as lost in bird genomes (Hron et al. 2015). In flatworms, however, hidden orthologs do not show a significantly different GC content and average length of G/C stretches than the majority of transcripts (Fig. 3C). We confirmed this observation for each particular transcriptome of our dataset (Supplemental Fig. S3).

Systematic error in sequence-similarity searches is also associated with the length of the sequence and the presence of short conserved stretches (i.e. protein domains with only a reduced number of conserved residues). Short protein lengths decrease BLAST sensitivity (Moyers and Zhang 2015). We thus expected hidden orthologs to consist of significantly shorter proteins, as is seen in *Drosophila* orphan genes (Palmieri et al. 2014). When analyzed together, the length of the flatworm hidden ortholog transcripts are not significantly different from that of the rest of the transcripts (Supplemental Table 7). However, the putative hidden orthologs are significantly longer than the rest of the transcripts when only high-coverage transcriptomes (≥ 85% CEGs identified) are considered (Fig. 3D).

We next addressed whether the 1,243 non-redundant human proteins homolog to the flatworm hidden orthologs were enriched in particular sequence motifs that could hamper their identification by common sequence similarity searches. We recovered a total of 1,180 unique PFAM annotations, almost all of them present only in one (1,016) or two (112) of the identified hidden orthologs (Supplemental Table 8). The most abundant PFAM domain (Table 1) was the pleckstrin homology (PH) domain (PFAM ID: PF00169), which occurs in a wide range of proteins involved in intracellular signaling and cytoskeleton (Scheffzek and Welti 2012). PH domains were present in 11 of the candidate hidden orthologs. Most other abundant domains were related to protein interactions, such as the F-box-like domain (Kipreos and Pagano 2000), the forkhead-associated domain (Durocher and Jackson 2002), and the zinc-finger of C2H2 type (Iuchi 2001). These more abundant domains vary significantly in average length and number of generally conserved sites (Table 1).

Lastly, we looked to see if there were any patterns of codon usage associated with the putative hidden orthologs. We did not observe a statistically significant difference between the codon adaptation index of hidden orthologs of the planarian species *B. candida*, *D. tigrina* and *S. mediterranea* and other open reading frames of these transcriptomes (Fig. 3E). Altogether, these analyses indicate that hidden orthologs do not show intrinsic properties that could cause systematic errors during homology searches.

The possible mechanisms driving hidden orthology in Platyhelminthes

To assess the contribution of duplication and divergence (Force et al. 1999) towards the generation of hidden orthologs, we looked for paralogs of the putative hidden orthologs in OrthoFinder orthogroups. Focusing on Tricladida (planarian flatworms), we observed that a putative hidden ortholog co-occurred in the same orthogroup with one or more sequences from the same species in 14–30% of the cases (depending on the species) (Fig. 4A). These hidden orthologs can be interpreted as fast-evolving paralogs. For those one-to-one hidden orthologs of *S. mediterranea*, we calculated the number of non-synonymous substitutions per non-synonymous sites (K_a) and the number of synonymous substitutions per synonymous sites (K_s) in pairwise comparisons with their respective ortholog in the 'bridge' transcriptome (Fig. 4B). Although for almost half of them the K_s value appeared to be saturated ($K_s > 2$), the K_a/K_s ratio for most of the rest was above or close to 0.5, which is often interpreted as a sign of weak positive selection or relaxed constraints (Nachman 2006).

A gene ontology (GO) analysis of the non-redundant hidden orthologs identified in all flatworm transcriptomes and the planarian *S. mediterranea* revealed a wide spectrum of

GO terms, with binding and catalytic activities being the most abundant (Supplemental Fig. S4). The statistical comparison of the GO categories of the putative *S. mediterranea* hidden orthologs revealed 248 significantly ($p < 0.05$) enriched GO terms (Supplemental Table 9). Interestingly, the putative hidden orthologs were enriched for biological processes and cellular compartments related to mitochondrial protein translation and the mitochondrial ribosome respectively. Indeed, ribosomal proteins were amongst the most common hidden orthologs recovered from our dataset (Supplemental Table 2). These findings suggest that mitochondrial genes show accelerated evolutionary rates (Solà et al. 2015), which might be causing nuclear-encoded proteins that are exported to the mitochondrion adapt to this change (Barreto and Burton 2013).

To test if hidden orthologs are a result of rapid compensatory evolution, we investigated whether there were a disproportionate number of interactions between putative hidden orthologs in *S. mediterranea*. We used the BioGRID database (Chatr-Aryamontri et al. 2015) to count the number of physical interactions between *S. mediterranea* hidden orthologs. We identified 71 such interactions in *S. mediterranea*, which is statistically significant when compared with 1,000 random sets of human genes (one-tailed Monte Carlo analysis with 100,000 iterations; p-value 1.9×10^{-4}; Fig. 4C). These findings suggest that compensatory mutations in binding partners and/or otherwise interconnected proteins are contributing to the origin of hidden orthologs.

The identified hidden orthologs fill out gaps in the flatworm gene complement

We used an expanded Leapfrog strategy to identify possible hidden orthologs for those centrosomal and cytoskeleton-related genes supposedly lost in the flatworms *M.*

lignano, *S. mediterranea*, and *S. mansoni* (Azimzadeh et al. 2012). First, we used a reciprocal best BLAST strategy to identify orthologs of the human centrosomal proteins in each of our transcriptomes under study, and thereafter we used Leapfrog to identify any hidden member of this original gene set. We recovered at least one reciprocal best BLAST hit for 56 of the 61 centrosomal genes, and identified fast-evolving putative orthologs in 19 of the 61 centrosomal genes (Fig. 5). In total, the number of hidden orthologs identified was 58 (counting only once those for the same gene in the different analyzed *S. mediterranea* transcriptomes). Most importantly, we found putative hidden orthologs for the genes *CEP192* and *SDCCAG8* in the planarian *S. mediterranea* (Fig. 5; Supplemental Fig. S5; Supplemental Fig. S6), which were two of the five key genes essential for centrosome assembly and duplication (Azimzadeh et al. 2012). Noticeably, the expression patterns of these two genes (Supplemental Fig. S7) differ considerably from those reported for other centrosomal genes in *S. mediterranea* (Azimzadeh et al. 2012), which suggests that they might not be co-expressed on those planarian cells that assemble centrosomes and thus might have evolved alternative functions.

Using as a 'bridge' the orthologs found in the more conservative rhabditophoran species *M. lignano* and *P. vittatus*, we found hidden orthologs for *gsc*, *dbx*, *vax*, *drgx*, *vsx* and *cmp* in the planarian *S. mediterranea* (Table 2; Supplemental Fig. S8 and Supplemental Fig. S9), which places the loss of these homeodomain classes most likely at the base of the last-common neodermatan ancestor. Importantly, the Hhex family was present in *P. vittatus*, but was not identified in *M. lignano* and *S. mediterranea*, and the Prrx and Shox families were present in *M. lignano*, but absent from *P. vittatus* and *S. mediterranea* transcriptomes. These observations suggest that many of the losses of

homeobox genes occurred in the ancestors to the Rhabitophora and Neodermata, with only a few losses of specific gene classes in particular lineages of free-living flatworms.

Discussion

Our study implements a simple automated BLAST pipeline that uses increased taxon sampling and transitive homology (Pearson 1996; Yona et al. 1998) to overcome large evolutionary distances and identify putative hidden orthologs (Fig. 2). However, our results on an extensive dataset of flatworm transcriptomes are likely an underestimation of the true number of hidden orthologs. First, we based our identification of hidden orthologs on reciprocal best BLAST hits, a valid and widely used approach (Tatusov et al. 1997; Overbeek et al. 1999; Wolf and Koonin 2012), but limited (Fulton et al. 2006; Dalquen and Dessimoz 2013). Second, an iterative Leapfrog recovers many more putative hidden orthologs (Supplemental Table 5). This indicates that there might be natural circumstances (e.g., presence of hidden orthologs and missing genes), even in more conservatively evolving lineages, which contribute to the suitability of a particular transcriptome to act as a 'bridge'. Furthermore, we demonstrate that using hidden orthologs themselves as 'bridge queries' on other lineages can help recover even more new hidden orthologs (Table 2). Finally, 16 out of the 35 analyzed transcriptomes contain less than 80% of core eukaryotic genes (Supplemental Fig. S1), and can be regarded as fairly incomplete (Parra et al. 2009).

In our dataset, putative hidden orthologs are not significantly shorter, and do not exhibit either particular sequence composition biases (Fig. 3) or protein domains (Table 1) that could account for the difficulties in being detected by standard homology searches. Instead, hidden orthologs represent restricted fast evolving orthologs, which have been

driven by either gene duplication events, weak positive selection and relaxed constraints, and/or by compensatory mutations between protein partners (Fig. 4). As observed in other organisms, whole or partial genomic duplications (Van de Peer et al. 2009a; Van de Peer et al. 2009b) and transposable elements (Feschotte 2008), which are common in flatworms (Benazzi and Benazzi-Lentati 1976; Garcia-Fernandez et al. 1995; Sperb et al. 2009), could also contribute to generate sequence diversity, and thus hidden orthology, in Platyhelminthes. In some cases, however, hidden orthologs are apparently associated with divergent biological features of Platyhelminthes (Fig. 5; Table 2). The fact that most of them are species-specific indicates that the gene complement of an organism is in fact heterogeneous, composed of genes evolving at different evolutionary rates (Wolfe 2004), sometimes much higher or much lower than the 'average' exhibited by that lineage.

Although orthology does not strictly imply functional conservation (Fitch 1970; Gabaldon and Koonin 2013), the 'orthology conjecture' states that orthologous genes show a higher functional similarity than paralogs (Koonin 2005; Dolinski and Botstein 2007; Studer and Robinson-Rechavi 2009; Kryuchkova-Mostacci and Robinson-Rechavi 2016). Since only 14–30% of the putative hidden orthologs identified in Tricladida are fast evolving paralogs (Fig. 4A), the orthology conjecture would imply that most of the putative hidden orthologs identified are indeed functionally conserved. In this regard, hidden orthologs are often protein partners (Fig. 4C), suggesting that compensatory mutations to maintain interaction, and thus probably functionality, could have promoted sequence divergence in some cases. However, a more detailed analysis of the two centrosomal hidden orthologs *SDCCAG8* and *CEP192* shows remarkably different expression patterns in the planarian *S. mediterranea*, making unlikely that they

still cooperate to assemble the centrosomes. This indicates that unraveling whether hidden genes are maintaining ancestral functions despite very high mutation rates or are abandoning highly conserved ancestral functions but continuing to contribute to the biology of the organism requires of detailed individual analyses of each of the putative hidden orthologs.

The recovered hidden orthologs have an immediate impact on our understanding of gene complement evolution in Platyhelminthes, and in particular on those lineages that are subject of intense research, such as the regenerative model *Schmidtea mediterranea* and parasitic flatworms (Berriman et al. 2009; Wang et al. 2011; Olson et al. 2012; Sánchez Alvarado 2012). The identification of fast-evolving orthologs for important centrosomal proteins in *S. mediterranea* and other flatworms lineages (Fig. 5) indicates that the evolutionary events leading to the loss of centrosomes are probably more complex than initially surmised (Azimzadeh et al. 2012). Similarly, the presence of presumably lost homeobox classes in *S. mediterranea* may affect our current view of gene loss and morphological evolution in flatworms (Tsai et al. 2013). The use of more molecularly conserved flatworm lineages, such as *P. vittatus*, can improve the identification of candidate genes, as well as help with the annotation of the increasingly abundant flatworm RNA-seq and genomic datasets (Berriman et al. 2009; Wang et al. 2011; Tsai et al. 2013; Robb et al. 2015; Wasik et al. 2015; Brandl et al. 2016). Therefore, we have now made available an assembled version of those highly complete non-planarian flatworm transcriptomes in PlanMine, an integrated web resource of transcriptomic data for planarian researchers (Brandl et al. 2016). Importantly, the Leapfrog pipeline can also be exported to any set of transcriptomes/predicted proteins.

Methods

Leapfrog Pipeline

We assembled a dataset including 35 transcriptomes of 29 flatworm species, three outgroup taxa and a single-isoform version of the human RefSeq proteome (see Supplemental Methods for further details). Transcriptome completeness and quality analyses were performed with CEGMA (Parra et al. 2007) and TransRate (Smith-Unna et al. 2015) (see Supplemental Methods). All BLAST searches were conducted using BLAST+ version 2.2.31 (Camacho et al. 2009) using multiple threads (from 2 to 10 per BLAST). We first ran a TBLASTN search using the proteome HumRef2015 as a query against the untranslated *Prostheceraeus vittatus* transcriptome (tblastn -query HumRef2015 -db Pvit -outfmt 6 -out Hs_v_Pv). We next ran a BLASTX search using the *Prostheceraeus vittatus* transcriptome (DNA) as a query against the HumRef2015 protein dataset (blastx -query Pvit -db HumRef2015 -outfmt 6 -out Pv_v_Hs). We ran a series of TBLASTX searches using the *Prostheceraeus vittatus* transcriptome (DNA) as a query against each of our target untranslated transcriptome database (e.g., tblastx -query "TRANSCRIPTOME" -db Pvit -outfmt 6 -out "TRANSCRIPTOME"_v_Pvit). Lastly, we ran a series of TBLASTX searches using our transcriptome databases (DNA) as queries against the untranslated *Prostheceraeus vittatus* transcriptome (e.g., tblastx -query Pvit -db Sman -out Pvit_v_Sman -outfmt 6). The tab-delimited BLAST outputs generated above were used as input to the Leapfrog program (Supplemental File 1; https://github.com/josephryan/leapfrog). The default E-Value cutoff (0.01) was used for all Leapfrog runs. The Leapfrog program identifies HumRef2015 proteins that fit the following criteria: (1) they have no hit to a target flatworm transcriptome, (2) they have a reciprocal best BLAST hit with a 'bridge' (e.g. *Prostheceraeus vittatus*) transcript, and (3) the 'bridge' transcript has a reciprocal best BLAST hit to the target flatworm

transcriptome. The output includes the HumRef2015 Gene ID, the *Prostheceraeus vittatus* transcript and the target flatworm transcript. The annotation of the hidden orthologs is provided in Supplemental Table 2.

OrthoFinder and HaMStR analyses

Single best candidate coding regions of all flatworm transcriptomes were predicted with TransDecoder v3.0.0 (Haas et al. 2013). The resulting flatworm proteomes, together with HumRef2015, were used to identify orthologous groups with OrthoFinder v0.7.1 (Emms and Kelly 2015). We ran HaMStR version 13.2.6 (Ebersberger et al. 2009) on *S. mediterranea* using the following command: (hamstr -sequence_file=Smed.dd.mod -taxon=Smed -hmmset=modelorganisms_hmmer3 -refspec=DROME -central). We used a Perl script to compare the list of putative hidden orthologs from Leapfrog to the HMM outputs of HaMStR (Supplemental File 1).

GC content, sequence length, CAI index, interaction analyses and K_a/K_s values

Custom-made scripts were used to calculate the GC content of hidden orthologs and transcripts of our dataset, the average length of the G/C stretches of each sequence, and the length of hidden orthologs and other transcripts (Supplemental File 1). We used the Codon Usage Database (Nakamura et al. 2000) and CAIcal server (Puigbo et al. 2008) to calculate 'codon adaptation indexes' in hidden orthologs and three sets of transcripts of same size, randomly generated. Alignments of 53 one-to-one orthologs between *S. mediterranea* and *P. vittatus* were calculated with MAFFT v.5 (Katoh and Standley 2013) and PAL2NAL (Suyama et al. 2006) was used to infer codon alignments and trim gap regions and stop codons. K_a and K_s values were calculated with KaKs_calculator v2.0 (Wang et al. 2010) with the default model averaging (MA) method. A custom Perl

script identified physical interactions in *S. mediterranea* hidden orthologs. The same script built 1,000 random gene sets (sampling genes from HumRef2015 that were current in Entrez Gene as of August 2016) and determined the number of physical interactions in each of these. The complete analysis can be repeated by running the biogrid.pl script (Supplemental File 1). All values were plotted in R (R Core Team 2015) using the ggplot2 package (Wickham 2009).

GO and InterPro analyses

All GO analyses were performed using the free version of Blast2GO v3 (Conesa et al. 2005). Charts were done with a cutoff value of 30 GO nodes for the analyses of all hidden orthologs, and 10 GO nodes for the analyses of *S. mediterranea* hidden orthologs. Resulting charts were edited in Illustrator CS6 (Adobe). InterProScan 5 (Jones et al. 2014) was used to analyze protein domain architectures.

Orthology assignment and gene expression analyses

Protein alignments (available in Supplemental File 2) were performed with MAFFT v.5 (Katoh and Standley 2013) using the G-INS-i option, and spuriously aligned regions were removed with gblocks 3 (Talavera and Castresana 2007). Orthology assignments were performed with RAxML v8.2.6 (Stamatakis 2014) with the autoMRE option. The genes *SDCCAG8* and *CEP192* were cloned using gene-specific primers (Supplemental Methods) on cDNA obtained from adult planarians and mixed regenerative stages. Whole mount *in situ* hybridization was performed as described previously (Umesono et al. 1997; Agata et al. 1998). Representative specimens were cleared in 70% glycerol and imaged under a Discovery.V8 SteREO microscope equipped with an AxioCam MRc camera (Zeiss, Germany).

Data Access

Raw RNA-seq data produced in this study have been submitted to the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra/) under accession number SRX1343824.

Software availability

The Leapfrog pipeline and all scripts employed in this study can be found in Supplemental File 1. In addition, they are available in https://github.com/josephryan/leapfrog and https://github.com/josephryan/reduce_refseq.

Acknowledgements

We thank the members of the Hejnol's lab for support and discussions, and in particular Daniel Thiel and Anlaug Furu for taking care of the *M. lignano* and *S. mediterranea* cultures. We appreciate the advance access given to platyhelmith transcriptomes by Gonzalo Giribet and Chris Laumer at the beginning of the project. We thank Axios Review (axiosreview.org) editor Titus Brown, and the six anonymous referees of this manuscript for their efficient and insightful reviewing. We also thank Francesc Cebrià and Susanna Fraguas for kindly providing *S. mediterranea* animals and cDNA, as well as Marta Iglesias for helping with planarian fixation and *in situ* hybridization. This research was funded by the Sars Centre core budget and the European Research Council Community's Framework Program Horizon 2020 (2014–2020) ERC grant agreement 648861 to AH. JFR was supported by startup funds from the University of Florida DSP

Research Strategic Initiatives #00114464 and the University of Florida Office of the Provost Programs, JMMD was supported by Marie Curie IEF 329024 fellowship.

Author's contributions

JMMD and JFR designed the study. JMMD, AH, and KP collected material for the transcriptomes of *M. lignano, P. vittatus*, and *L. squammata*. JFR wrote the code of Leapfrog. JMMD, JFR and BCV performed the analyses. JFR, JMMD and AH wrote the manuscript. All authors read and approved the final manuscript.

Disclosure declaration

The authors declare that they have no competing interests.

References

Agata K, Soejima Y, Kato K, Kobayashi C, Umesono Y, Watanabe K. 1998. Structure of the planarian central nervous system (CNS) revealed by neuronal cell markers. *Zoolog Sci* **15**: 433-440.

Alba MM, Castresana J. 2007. On homology searches by protein Blast and the characterization of the age of genes. *BMC Evol Biol* **7**: 53.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403-410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-3402.

Azimzadeh J, Wong ML, Downhour DM, Sanchez Alvarado A, Marshall WF. 2012. Centrosome loss in the evolution of planarians. *Science* **335**: 461-463.

Barreto FS, Burton RS. 2013. Evidence for compensatory evolution of ribosomal proteins in response to rapid divergence of mitochondrial rRNA. *Mol Biol Evol* **30**: 310-314.

Benazzi M, Benazzi-Lentati G. 1976. *Animal cytogenetics*. Gebru der Borntraeger, Berlin.

Benjamini Y, Speed TP. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* **40**: e72.

Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC, Mashiyama ST, Al-Lazikani B, Andrade LF, Ashton PD et al. 2009. The genome of the blood fluke Schistosoma mansoni. *Nature* **460**: 352-358.

Brandl H, Moon H, Vila-Farre M, Liu SY, Henry I, Rink JC. 2016. PlanMine - a mineable resource of planarian biology and biodiversity. *Nucleic Acids Res* **44**: D764-773.

Breugelmans B, Ansell BR, Young ND, Amani P, Stroehlein AJ, Sternberg PW, Jex AR, Boag PR, Hofmann A, Gasser RB. 2015. Flatworms have lost the *ri*ght *o*pen reading frame kinase 3 gene during evolution. *Sci Rep* **5**: 9417.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L et al. 2015. The BioGRID interaction database: 2015 update. *Nucleic Acids Res* **43**: D470-478.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**: 3674-3676.

Dalquen DA, Dessimoz C. 2013. Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals. *Genome Biol Evol* **5**: 1800-1806.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36**: e105.

Dolinski K, Botstein D. 2007. Orthology and functional conservation in eukaryotes. *Annu Rev Genet* **41**: 465-507.

Domazet-Loso T, Brajkovic J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet* **23**: 533-539.

Durocher D, Jackson SP. 2002. The FHA domain. *FEBS Lett* **513**: 58-66.

Ebersberger I, Strauss S, von Haeseler A. 2009. HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evol Biol* **9**: 157.

Eddy SR. 2011. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**: e1002195.

Edgecombe GD, Giribet G, Dunn CW, Hejnol A, Kristensen RM, Neves RC, Rouse GW, Worsaae K, Sørensen MV. 2011. Higher-level metazoan relationships: recent progress and remaining questions. *Org Divers Evol* **11**: 151-172.

Egger B, Lapraz F, Tomiczek B, Muller S, Dessimoz C, Girstmair J, Skunca N, Rawlinson KA, Cameron CB, Beli E et al. 2015. A transcriptomic-phylogenomic analysis of the evolutionary relationships of flatworms. *Curr Biol* **25**: 1347-1353.

Elhaik E, Sabath N, Graur D. 2006. The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. *Mol Biol Evol* **23**: 1-3.

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

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**: 397-405.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. *Syst Zool* **19**: 99-113.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**: 1531-1545.

Fulton DL, Li YY, Laird MR, Horsman BG, Roche FM, Brinkman FS. 2006. Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics* **7**: 270.

Gabaldon T, Koonin EV. 2013. Functional and evolutionary implications of gene orthology. *Nat Rev Genet* **14**: 360-366.

Garcia-Fernandez J, Bayascas-Ramirez JR, Marfany G, Munoz-Marmol AM, Casali A, Baguna J, Salo E. 1995. High copy number of highly similar *mariner*-like transposons in planarian (Platyhelminthe): evidence for a trans-phyla horizontal transfer. *Mol Biol Evol* **12**: 421-431.

Gonzalez MW, Pearson WR. 2010. Homologous over-extension: a challenge for iterative similarity searches. *Nucleic Acids Res* **38**: 2177-2189.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M et al. 2013. *De novo* transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nat Protoc* **8**: 1494-1512.

Hron T, Pajer P, Pačes J, Bartunek P, Elleder D. 2015. Hidden genes in birds. *Genome Biol* **16**: 164.

Iuchi S. 2001. Three classes of C2H2 zinc finger proteins. *Cell Mol Life Sci* **58**: 625-635.

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**: 1236-1240.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772-780.

Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet* **25**: 404-413.

Kipreos ET, Pagano M. 2000. The F-box protein family. *Genome Biol* **1**: REVIEWS3002.

Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* **39**: 309-338.

Kryuchkova-Mostacci N, Robinson-Rechavi M. 2016. Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs. *PLoS Comput Biol* **12**: e1005274.

Kuchibhatla DB, Sherman WA, Chung BY, Cook S, Schneider G, Eisenhaber B, Karlin DG. 2014. Powerful sequence similarity search methods and in-depth manual analyses can identify remote homologs in many apparently "orphan" viral proteins. *J Virol* **88**: 10-20.

Laumer CE, Bekkouche N, Kerbl A, Goetz F, Neves RC, Sorensen MV, Kristensen RM, Hejnol A, Dunn CW, Giribet G et al. 2015a. Spiralian phylogeny informs the evolution of microscopic lineages. *Curr Biol* **25**: 2000-2006.

Laumer CE, Hejnol A, Giribet G. 2015b. Nuclear genomic signals of the 'microturbellarian' roots of platyhelminth evolutionary innovation. *Elife* **4**.

Liebeskind BJ, McWhite CD, Marcotte EM. 2016. Towards Consensus Gene Ages. *Genome Biol Evol* **8**: 1812-1823.

Martin-Duran JM, Romero R. 2011. Evolutionary implications of morphogenesis and molecular patterning of the blind gut in the planarian *Schmidtea polychroa*. *Dev Biol* **352**: 164-176.

Moyers BA, Zhang J. 2015. Phylostratigraphic bias creates spurious patterns of genome evolution. *Mol Biol Evol* **32**: 258-267.

Moyers BA, Zhang J. 2016. Evaluating Phylostratigraphic Evidence for Widespread De Novo Gene Birth in Genome Evolution. *Mol Biol Evol* **33**: 1245-1256.

Nachman M. 2006. Detecting selection at the molecular level. In *Evolutionary Genetics: Concepts and Case Studies*, (ed. CW Fox, JB Wolf). Oxford University Press, New York.

Nakamura Y, Gojobori T, Ikemura T. 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res* **28**: 292.

Olson PD, Zarowiecki M, Kiss F, Brehm K. 2012. Cestode genomics - progress and prospects for advancing basic and applied aspects of flatworm biology. *Parasite Immunol* **34**: 130-150.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. 1999. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96**: 2896-2901.

Palmieri N, Kosiol C, Schlotterer C. 2014. The life cycle of *Drosophila* orphan genes. *Elife* **3**: e01311.

Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**: 1061-1067.

Parra G, Bradnam K, Ning Z, Keane T, Korf I. 2009. Assessing the gene space in draft genomes. *Nucleic Acids Res* **37**: 289-297.

Pearson WR. 1996. Effective protein sequence comparison. *Methods Enzymol* **266**: 227-258.

Puigbo P, Bravo IG, Garcia-Vallve S. 2008. CAIcal: a combined set of tools to assess codon usage adaptation. *Biol Direct* **3**: 38.

R Core Team. 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Riddiford N, Olson PD. 2011. Wnt gene loss in flatworms. *Dev Genes Evol* **221**: 187-197.

Robb SM, Gotting K, Ross E, Sanchez Alvarado A. 2015. SmedGD 2.0: The *Schmidtea mediterranea* genome database. *Genesis* **53**: 535-546.

Sánchez Alvarado A. 2012. Q&A: What is regeneration, and why look to planarians for answers? *BMC Biol* **10**: 88.

Scheffzek K, Welti S. 2012. Pleckstrin homology (PH) like domains - versatile modules in protein-protein interaction platforms. *FEBS Lett* **586**: 2662-2673.

Smith-Unna RD, Boursnell C, Patro R, Hibberd JM, Kelly S. 2015. TransRate: reference free quality assessment of de-novo transcriptome assemblies. *BioRxiv* **021626**.

Solà E, Álvarez-Presas M, Frías-López C, Littlewood DT, Rozas J, Riutort M. 2015. Evolutionary analysis of mitogenomes from parasitic and free-living flatworms. *PLoS One* **10**: e0120081.

Sonnhammer EL, Östlund G. 2015. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res* **43**: D234-239.

Sperb F, Schuck DC, Rodrigues JJ. 2009. Occurrence and abundance of a *mariner*-like element in freshwater and terrestrial planarians (Platyhelminthes, Tricladida) from southern Brazil. *Genet Mol Biol* **32**: 731-739.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312-1313.

Struck TH, Wey-Fabrizius AR, Golombek A, Hering L, Weigert A, Bleidorn C, Klebow S, Iakovenko N, Hausdorf B, Petersen M et al. 2014. Platyzoan paraphyly based on phylogenomic data supports a noncoelomate ancestry of Spiralia. *Mol Biol Evol* **31**: 1833-1849.

Studer RA, Robinson-Rechavi M. 2009. How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet* **25**: 210-216.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**: W609-612.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**: 564-577.

Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. *Science* **278**: 631-637.

Tautz D, Domazet-Loso T. 2011. The evolutionary origin of orphan genes. *Nat Rev Genet* **12**: 692-702.

Tsai IJ, Zarowiecki M, Holroyd N, Garciarrubio A, Sanchez-Flores A, Brooks KL, Tracey A, Bobes RJ, Fragoso G, Sciutto E et al. 2013. The genomes of four tapeworm species reveal adaptations to parasitism. *Nature* **496**: 57-63.

Umesono Y, Watanabe K, Agata K. 1997. A planarian *orthopedia* homolog is specifically expressed in the branch region of both the mature and regenerating brain. *Dev Growth Differ* **39**: 723-727.

Van de Peer Y, Fawcett JA, Proost S, Sterck L, Vandepoele K. 2009a. The flowering world: a tale of duplications. *Trends Plant Sci* **14**: 680-688.

Van de Peer Y, Maere S, Meyer A. 2009b. The evolutionary significance of ancient genome duplications. *Nat Rev Genet* **10**: 725-732.

Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. 2010. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics* **8**: 77-80.

Wang X, Chen W, Huang Y, Sun J, Men J, Liu H, Luo F, Guo L, Lv X, Deng C et al. 2011. The draft genome of the carcinogenic human liver fluke Clonorchis sinensis. *Genome Biol* **12**: R107.

Wasik K, Gurtowski J, Zhou X, Ramos OM, Delas MJ, Battistoni G, El Demerdash O, Falciatori I, Vizoso DB, Smith AD et al. 2015. Genome and transcriptome of the regeneration-competent flatworm, *Macrostomum lignano*. *Proc Natl Acad Sci U S A* **112**: 12462-12467.

Wickham H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.

Wolf YI, Koonin EV. 2012. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biol Evol* **4**: 1286-1294.

Wolfe K. 2004. Evolutionary genomics: yeasts accelerate beyond BLAST. *Curr Biol* **14**: R392-394.

Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* **13**: 329-342.

Yona G, Linial N, Tishby N, Linial M. 1998. A map of the protein space--an automatic hierarchical classification of all protein sequences. *Proc Int Conf Intell Syst Mol Biol* **6**: 212-221.

Figure Legends

Figure 1. Hidden orthologs and the Leapfrog pipeline. (**A**) Taxonomically-restricted genes (TRGs) are genes with no clear orthology relationship (dashed line and question mark) to other known genes (e.g. orthology group of red dots). Improved sensitivity in the detection methods and/or improved taxon sampling can help uncover hidden orthology relationships, thus referring to these former TRGs as hidden orthologs. (**B**) The Leapfrog pipeline performs a series of reciprocal BLAST searches between an initial well-annotated dataset (e.g. human RefSeq), and a target and a 'bridge' transcriptomes. First, Leapfrog blasts the human RefSeq against the target (1) and the 'bridge' transcriptome (2), and identifies reciprocal best-hit orthologs between the 'bridge' and the human RefSeq proteins (3). These annotated genes of the 'bridge' are then used to find orthologs in the target transcriptomes by reciprocal best BLAST hits (4 and 5). If these two pairs of reciprocal best BLAST hit searches are consistent between them, the gene in the target transcriptome is deemed a hidden ortholog. Colored shapes within green boxes represent different sequences of each dataset.

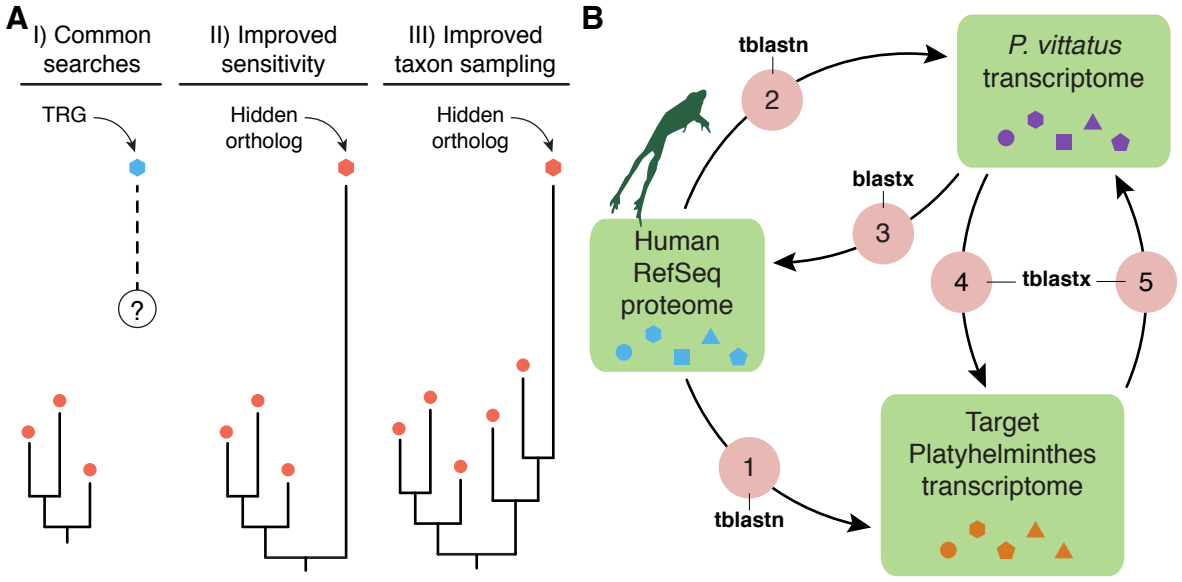
Figure 2. The Leapfrog pipeline recovers hundreds of hidden orthologs in Platyhelminthes. (**A**) Distribution of hidden orthologs according to their identification in one or more of the analyzed transcriptomes. Most of the hidden orthologs are unique of each lineage. (**B**) Distribution of species-specific hidden orthologs in each studied species. (**C**) Amino acid alignment of a fragment of the centrosomal protein SDCCAG8 of *H. sapiens*, *P. vittatus* and *S. mediterranea*, and pairwise comparison of conserved residues. Positions that differ between the human and the hidden ortholog products are conserved between *P. vittatus* and one or the other sequences. Black dots indicate residues conserved among the three species.

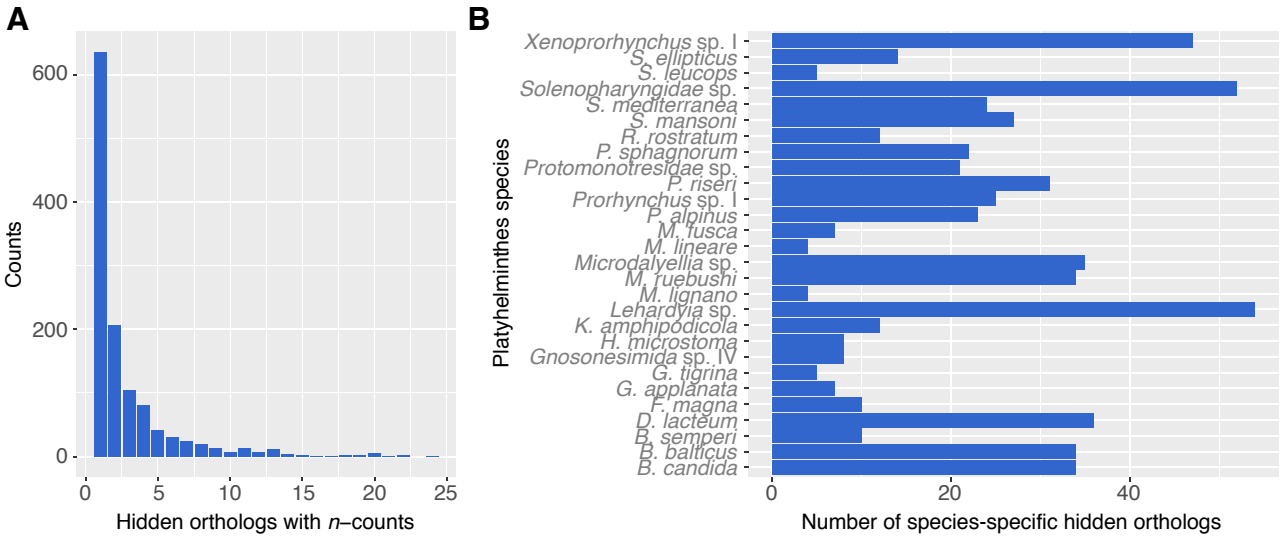
Figure 3. Hidden orthologs, evolutionary rates and sequence composition analyses. (**A**) Principal component analysis of the analyzed data showing the eigenvectors for each variable. The two first principal components (PC1, PC2) explain together 67.6% of the observed variability. (**B**) Number of hidden orthologs in relation to the branch length of each lineage (linear regression in blue; dots with external black line indicate the taxa with highly complete transcriptome). There is a low correlation between the two variables (R^2=0.124). (**C**) GC content of each transcript plotted against its average length of G/C stretches considering all studied flatworm transcriptomes (left) and only *S. mediterranea* (right). The transcripts corresponding to hidden orthologs are in blue. Hidden orthologs do not differentiate from the majority of transcripts. (**D**) Average length of hidden orthologs compared to the average length of the other genes in transcriptomes with ≥ 85% CEGs. Hidden orthologs are significantly longer than the rest (Mann-Whitney test; $p<0.05$). (**E**) Codon Adaptation Index (CAI) of the hidden orthologs of the planarian species *B. candida*, *D. tigrina* and *S. mediterranea* compared with non-hidden orthologs. CAI index in hidden orthologs does not significantly differ from the rest of transcripts (Mann-Whitney test; $p<0.05$).

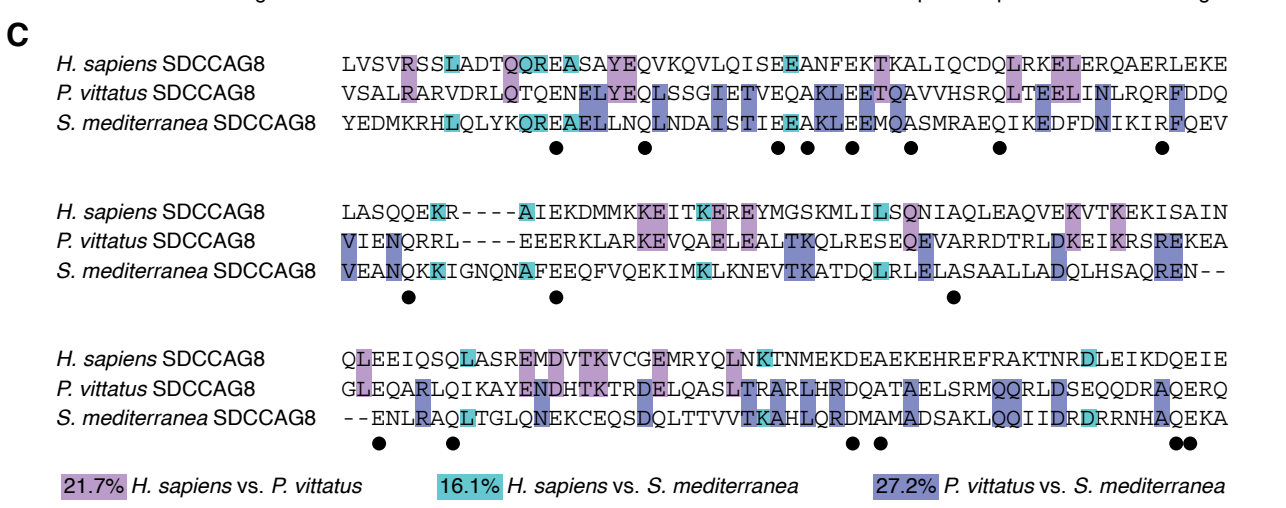
Figure 4. Level of paralogy and K_a/K_s values in triclad hidden orthologs. (**A**) Percentage of hidden orthologs identified by Leapfrog that are present in OrthoFinder and share orthogroup with other sequences of the same species. We dim these cases as probable fast evolving paralogs (hidden paralogy). (**B**) K_a and K_s values of 53 one-to-one hidden orthologs of *S. mediterranea* compared with their respective homologues in the 'bridge' species *P. vittatus*. Although in almost half of these hidden orthologs the K_s value suggested saturation ($K_s > 2$), for most of the rest the K_a/K_s value was above or around 0.5 (dotted line), which can be a sign of weak positive selection or relaxed

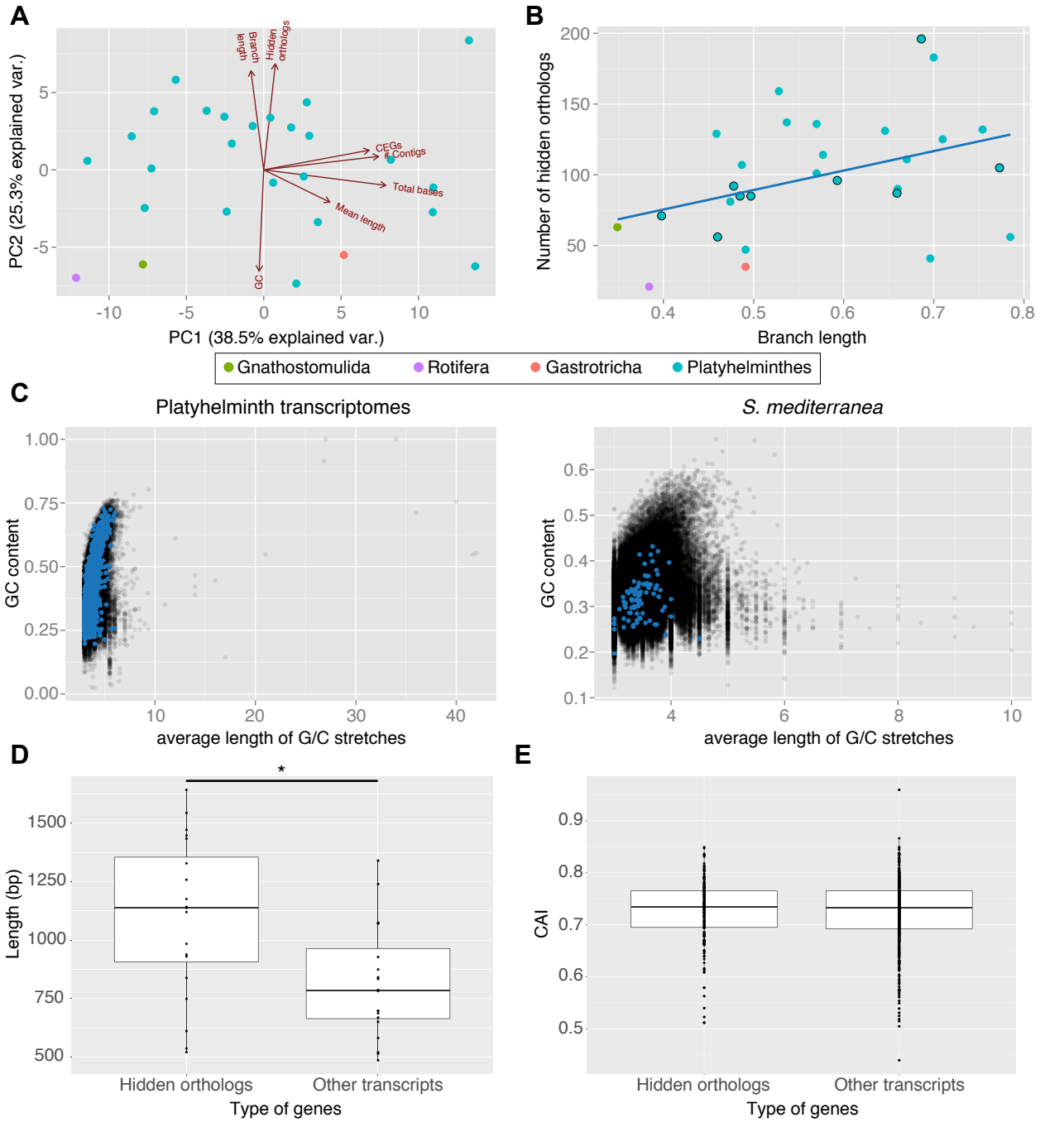
constraints. (**C**) Number of predicted protein-protein interactions in *S. mediterranea* hidden orthologs (red dot) compared with a distribution of interactions observed in 1,000 random samples of similar size (grey bars). Hidden orthologs show a significantly higher number of interactions, suggesting that complementary mutations between protein partners might drive hidden orthology in flatworms.

Figure 5. Hidden orthologs in the core set of centrosomal-related proteins. Presence (colored boxes) and absence (empty boxes) of the core set of centrosomal proteins (Azimzadeh et al. 2012) in all the analyzed flatworm transcriptomes. Orthologs identified by direct reciprocal best BLAST hit are in blue boxes, and hidden orthologs are in orange. The asterisks indicate the CEP192 protein in the *S. mediterranea* transcriptomes (pink color code). These proteins were manually identified with the *G. tigrina CEP192* sequence as 'bridge' by reciprocal best BLAST hit. The five proteins essential for centrosomal replication are squared in red.

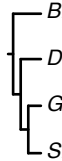

A
I) Common searches
II) Improved sensitivity
III) Improved taxon sampling
TRG
?
Hidden ortholog
Hidden ortholog
B
tblastn
2
P. vittatus transcriptome
blastx
3
Human RefSeq proteome
4
tblastx
5
1
tblastn
Target Platyhelminthes transcriptome


A
600
400
200
0
Counts
0
5
10
15
20
25
Hidden orthologs with n–counts
B
Xenoprorhynchus sp. I
S. ellipticus
S. leucops
Solenopharyngidae sp.
S. mediterranea
S. mansoni
R. rostratum
P. sphagnorum
Protomonotresidae sp.
P. riseri
Prorhynchus sp. I
P. alpinus
M. fusca
M. lineare
Microdalyellia sp.
M. ruebushi
M. lignano
Lehardyia sp.
K. amphipodicola
H. microstoma
Gnosonesimida sp. IV
G. tigrina
G. applanata
F. magna
D. lacteum
B. semperi
B. balticus
B. candida
Platyhelminthes species
0
20
40
Number of species-specific hidden orthologs


C
H. sapiens SDCCAG8 LVSVRSSLADTQQREASAYEQVKQVLQISEEANFEKTKALIQCDQLRKELERQAERLEKE
P. vittatus SDCCAG8 VSALRARVDRLQTQENELYEQLSSGIETVEQAKLEETQAVVHSRQLTEELINLRQRFDDQ
S. mediterranea SDCCAG8 YEDMKRHLQLYKQREAELLNQLNDAISTIEEAKLEEMQASMRAEQIKEDFDNIKIRFQEV
H. sapiens SDCCAG8 LASQQEKR----AIEKDMMKKEITKEREYMGSKMLILSQNIAQLEAQVEKVTKEKISAIN
P. vittatus SDCCAG8 VIENQRRL----EEERKLARKEVQAELEALTKQLRESEQEVARRDTRLDKEIKRSREKEA
S. mediterranea SDCCAG8 VEANQKKIGNQNAFEEQFVQEKIMKLKNEVTKATDQLRLELASAALLADQLHSAQREN--
H. sapiens SDCCAG8 QLEEIQSQLASREMDVTKVCGEMRYQLNKTNMEKDEAEKEHREFRAKTNRDLEIKDQEIE
P. vittatus SDCCAG8 GLEQARLQIKAYENDHTKTRDELQASLTRARLHRDQATAELSRMQQRLDSEQQDRAQERQ
S. mediterranea SDCCAG8 --ENLRAQLTGLQNEKCEQSDQLTTVVTKAHLQRDMAMADSAKLQQIIDRDRRNHAQEKA
21.7% H. sapiens vs. P. vittatus
16.1% H. sapiens vs. S. mediterranea
27.2% P. vittatus vs. S. mediterranea

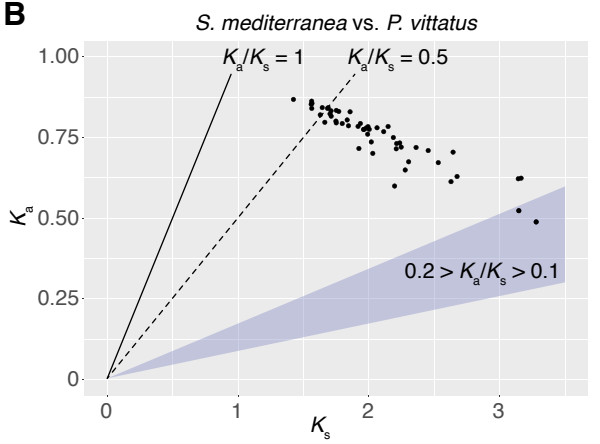

A
PC2 (25.3% explained var.)
PC1 (38.5% explained var.)
Branch length
Hidden orthologs
CEGs
Contigs
Total bases
Mean length
GC
B
Number of hidden orthologs
Branch length
Gnathostomulida
Rotifera
Gastrotricha
Platyhelminthes
C
Platyhelminth transcriptomes
S. mediterranea
GC content
average length of G/C stretches
D
Length (bp)
*
Hidden orthologs
Other transcripts
Type of genes
E
CAI
Hidden orthologs
Other transcripts
Type of genes

A

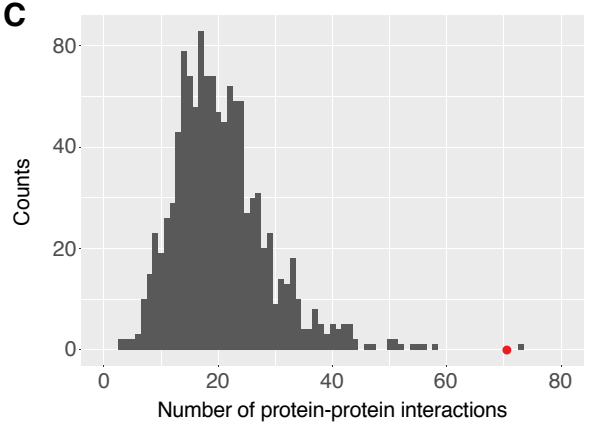


	# hidden orthologs	# hidden orthologs in OrthoFinder	% paralogy
B. candida	124	67	29.85% (20)
D. lacteum	179	58	13.79% (8)
G. tigrina	71	71	15.49% (11)
S. mediterranea	75	70	17.14% (12)

B



C



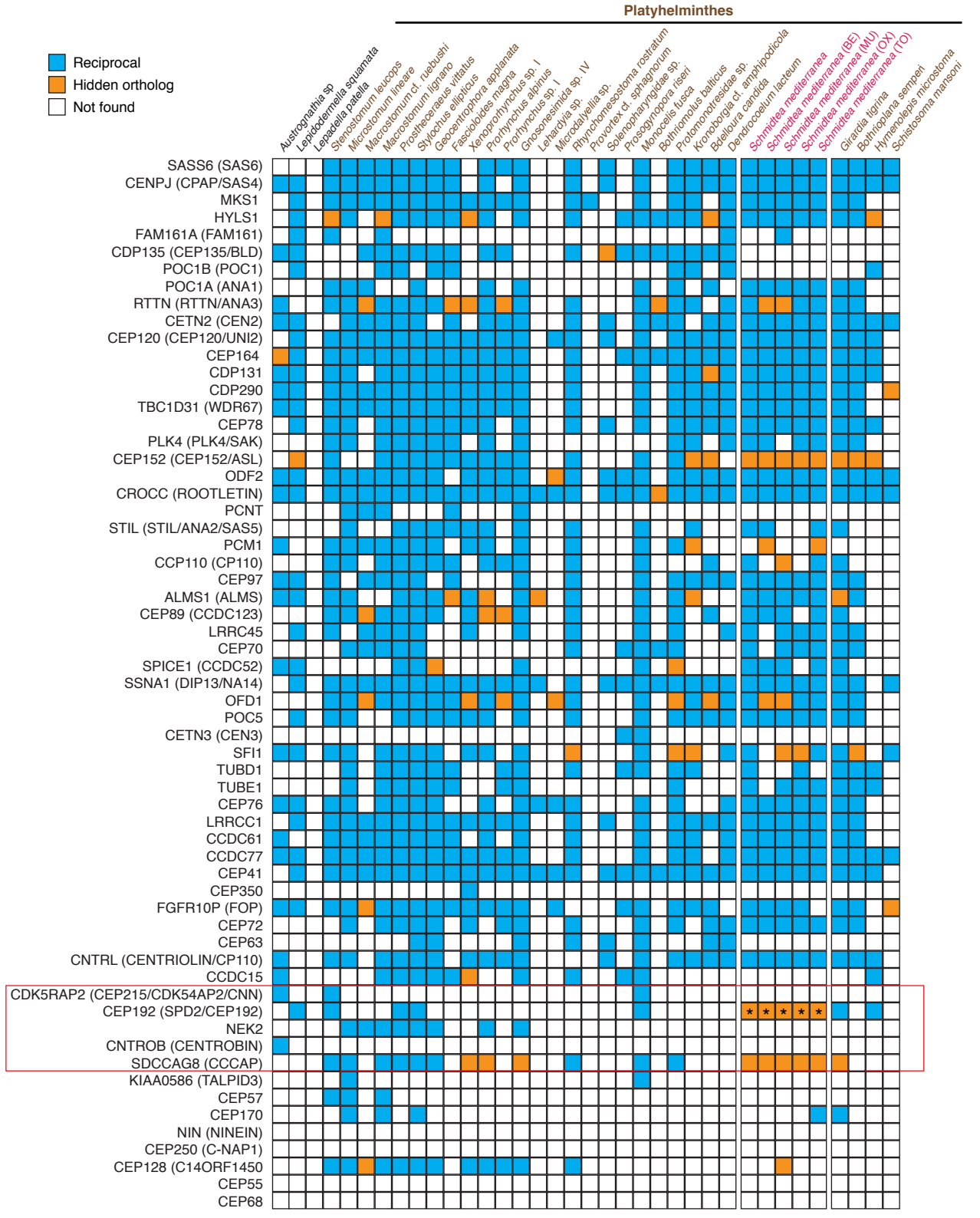

Platyhelminthes
Reciprocal
Hidden ortholog
Not found
Austrognathia sp
Lepidodermella squamata
Lepadella patella
Stenostomum leucops
Microstomum lineare
Macrostomum cf. ruebushi
Macrostomum tignano
Prostheceraeus vittatus
Stylochus ellipticus
Geocentrophora applanata
Fasciolopsis magna
Xenoprorhynchus sp.
Prorhynchus alpinus
Prorhynchus sp. I
Gnosonesimida sp. IV
Lehardyia sp.
Microdalyellia sp.
Rhynchomesostoma rostratum
Provortex cf. sphagnorum
Solenopharyngidae sp.
Prosogynopora riseri
Monocelis fusca
Bothrioplus balticus
Pratomonotresidae sp.
Kronoborgia cf. amphipodicola
Bdelloura candida
Dendrocoelum lacteum
Schmidtea mediterranea
Schmidtea mediterranea (BE)
Schmidtea mediterranea (MU)
Schmidtea mediterranea (OX)
Schmidtea mediterranea (TO)
Girardia tigrina
Bothrioplana semperi
Hymenolepis microstoma
Schistosoma mansoni
SASS6 (SAS6)
CENPJ (CPAP/SAS4)
MKS1
HYLS1
FAM161A (FAM161)
CDP135 (CEP135/BLD)
POC1B (POC1)
POC1A (ANA1)
RTTN (RTTN/ANA3)
CETN2 (CEN2)
CEP120 (CEP120/UNI2)
CEP164
CDP131
CDP290
TBC1D31 (WDR67)
CEP78
PLK4 (PLK4/SAK)
CEP152 (CEP152/ASL)
ODF2
CROCC (ROOTLETIN)
PCNT
STIL (STIL/ANA2/SAS5)
PCM1
CCP110 (CP110)
CEP97
ALMS1 (ALMS)
CEP89 (CCDC123)
LRRC45
CEP70
SPICE1 (CCDC52)
SSNA1 (DIP13/NA14)
OFD1
POC5
CETN3 (CEN3)
SFI1
TUBD1
TUBE1
CEP76
LRRCC1
CCDC61
CCDC77
CEP41
CEP350
FGFR10P (FOP)
CEP72
CEP63
CNTRL (CENTRIOLIN/CP110)
CCDC15
CDK5RAP2 (CEP215/CDK54AP2/CNN)
CEP192 (SPD2/CEP192)
NEK2
CNTROB (CENTROBIN)
SDCCAG8 (CCCAP)
KIAA0586 (TALPID3)
CEP57
CEP170
NIN (NINEIN)
CEP250 (C-NAP1)
CEP128 (C14ORF1450
CEP55
CEP68

Table 1. Most represented PFAM domains in flatworm hidden orthologs

PFAM	Description	Length[a]	Identity[b]	Hidden orthologs
PF00169	Pleckstrin homology domain	104.4	17%	APPL2, DOCK11, SH2B2, DOK1, PLEKHH1, ADAP1, PLEKHA3, DEF6, GAB1, RAPH1, PLEKHD1
PF00240	Ubiquitin family	70.7	36%	UBLCP1, TMUB2, TMUB1, HERPUD1, BAG1
PF00612	IQ calmodulin-binding motif	20.6	32%	IQGAP2, LRRIQ1, IQCE, RNF32, IQCD
PF07690	Major facilitator superfamily	311.2	12%	SLC46A3, SLC18B1, SLC22A18, MFSD3, KIAA1919
PF12874	Zinc-finger of C2H2 type	23.4	28%	SCAPER, ZMAT1, BNC2, ZNF385B, ZNF385D
PF12937	F-box-like	47.8	25%	FBXO18, FBXO7, FBXO33, FBXO15, FBXO39
PF00498	Forkhead-associated domain	72.4	24%	FHAD1, MDC1, NBN, MKI67
PF00536/ PF07647	SAM (Sterile alpha motif) domain	63.1/64.8	23%/20%	SAMD4A, SASH1, SAMD3, CNKSR3, SAMD10, SAMD15, SAMD15, SASH1

[a]in amino acids. Average values based on PFAM model.
[b]Average values based on PFAM model

Table 2. Presence/absence of hidden homeodomain genes in flatworms

Family	*M. lignano*	*P. vittatus*	*S. mediterranea*
Gsc	–	Present	–[1]
Pdx	–	–	–
Dbx	Present	Present	Present
Hhex	–	Present	–
Hlx	–	–	–
Noto	–	–	–
Ro	–	–	–
Vax	Present	Present	Present
Arx	Present[2]	Present[2]	–
Dmbx	–	–	–
Drgx	Present[2]	Present[2]	Present[2]
Prrx	Present	–	–
Shox	Present	–	–
Vsx	Present	Present	Present (Kao et al. 2013)
Pou1	–	–	–
Cmp	Present	Present	Present
Tgif	–	–	–

[1]gene present in the sister species *S. polychroa* (Martin-Duran and Romero 2011)
[2]Orthology based on BBH, not well supported by phylogenetic relationships.