A novel approach using 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, 5008, 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)

Abstract

Gene 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 however limited by the intrinsic limitations of common sequence similarity searches and available databases. Thus, a subset of the complement of an organism consists of hidden orthologs, those with no apparent homology with common sequenced animal lineages –mistakenly considered new genes– but actually representing fast evolving orthologs of presumably lost proteins. Here, we describe ‘Leapfrog’, an automated pipeline that uses increased taxon sampling to overcome long evolutionary distances and identify hidden orthologs in large transcriptomic databases. As a case study, we used 35 transcriptomes of 29 flatworm lineages to recover 3,597 hidden orthologs. Unexpectedly, we do not observe a correlation between the number of hidden orthologs in a lineage and its ‘average’ evolutionary rate. Hidden orthologs do not show unusual sequence composition biases (e.g. GC content, average length, domain composition), but do appear to be more common in genes with binding or catalytic activity. 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, qualifying the impact of gene losses and gains in gene complement evolution.

Introduction

Changes in gene complement are a fundamental aspect of organismal evolution (Ohno 1970; Olson 1999; Long, et al. 2003; De Robertis 2008). Current genome analyses estimate that novel genes –the so-called 'taxonomically-restricted' genes (TRGs) or 'orphan' genes; those without a clear homolog in other taxa– represent around 10–20% of the gene complement of most animal genomes (Khalturin, et al. 2009; Tautz and Domazet-Loso 2011). Although reported in some cases as non-functional open reading frames (ORFs) (Clamp, et al. 2007), TRGs are likely essential for the biology and evolution of an organism (Loppin, et al. 2005; Khalturin, et al. 2009; Knowles and McLysaght 2009; Li, et al. 2010; Colbourne, et al. 2011; Warnefors and Eyre-Walker 2011; Martin-Duran, et al. 2013; Palmieri, et al. 2014). The continuous increase in gene content is, however, balanced by a high rate of depletion among newly evolved genes (Tautz and Domazet-Loso 2011; Palmieri, et al. 2014) and by losses within the conserved, more ancient gene complement of animals (Kortschak, et al. 2003; Krylov, et al. 2003; Edvardsen, et al. 2005; Technau, et al. 2005).

Understanding the dynamic evolution of gene repertoires is often hampered by the difficulties of confidently identifying gene losses and gains. Gene annotation pipelines and large-scale comparisons (e.g. phylostratigraphy methods) 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 insertions or deletions and/or exhibits similarity with other counterparts in only a small subset of residues (Moyers and Zhang 2015). These limitations can generate significant biases when studying the evolution of protein-coding gene families (Elhaik, et al. 2006; Moyers and Zhang 2015). Accordingly, a proportion of the gene complement of an organism will be represented by genes that lack obvious affinity with homologs in the gene sets of the best annotated genomes– thus mistakenly considered potential TRGs– but actually representing fast evolving orthologs that we call hidden orthologs. This systematic error can potentially be overcome by more sensitive, although computationally intense, detection methods (e.g. profile HMMs, PSI-BLAST) (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 (fig. 1A).

Platyhelminthes (flatworms) is a morphological and ecologically diverse animal group characterized by significantly high rates of molecular evolution (Edgecombe, et al. 2011; Struck, et al. 2014; Laumer, Bekkouche, et al. 2015). Accordingly, changes in gene complement seem to be important drivers of adaptive evolution in this group (Berriman, et al. 2009; Martin-Duran and Romero 2011; Riddiford and Olson 2011; Tsai, et al. 2013). For instance, parasitic forms (e.g. tapeworms and flukes) have many unidentifiable genes and are reported to be missing myriad genes, including important developmental genes that are highly conserved in most other animals (Riddiford and Olson 2011; Tsai, et al. 2013). The presumed loss of critical genes has led to the inference that these animals have either developed alternative ways to implement critical steps in conserved pathways or that these pathways are no longer active (Wang, et al. 2011; Tsai, et al. 2013). 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, Hejnol, et al. 2015). These important resources provide an ideal case study 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.

Here, we describe a tool, which we have called 'Leapfrog,' that we have used to identify thousands of hidden orthologs across 27 different flatworms species by using an intermediate 'slow-evolving' flatworm species as a 'bridge.' Counter-intuitively, we show that the number of hidden orthologs does not correlate with the 'average' evolutionary rate of each particular species and unusual sequence composition biases, such as GC content, transcript length and domain architecture that could affect BLAST searches. Instead, some hidden orthologs appear to be related to certain gene ontology classes, and thus to particular highly divergent biological features of flatworms. In this context, we identify tens of presumably lost centrosomal-related genes (Azimzadeh, et al. 2012) and recover several homeodomain classes previously reported as absent (Tsai, et al. 2013). Altogether, our findings demonstrate that a functionally relevant proportion of genes without clear homology are indeed hidden orthologs in flatworms, thus alleviating the previously believed extensive gene loss exhibited by Platyhelminthes

(Azimzadeh, et al. 2012; Tsai, et al. 2013). In a broader context, our study suggests that hidden orthologs likely comprise a significant proportion of the gene repertoire of every organism, improving our understanding of gene complement evolution in animals.

Results

The 'Leapfrog' pipeline

To identify hidden orthologs in large transcriptomic datasets we created 'Leapfrog', which automates a series of BLAST-centric processes (fig. 1B). We started with a set of well-annotated sequences –the human RefSeq protein dataset– as our main queries and conducted a TBLASTN search of these sequences against each of our target flatworm transcriptomes (supplementary table 1, Supplementary Material online). Any queries that had zero BLAST hits with E-values less than our cutoff (0.01) were considered candidate hidden orthologs. We then looked for reciprocal best TBLASTX hits between these candidates and the transcriptome of the polyclad flatworm *Prostheceraeus vittatus*, a lineage that evolves at a slower rate than most other flatworms with available sequence data (as evidenced by branch lengths in (Laumer, Hejnol, et al. 2015)). If there was a reciprocal best BLAST hit in our 'bridge' transcriptome, the 'bridge' transcript was used as query in a BLASTX search against the initial annotated human RefSeq protein dataset. If there was a human reciprocal hit, and the human sequence was the starting query, then we deemed the candidate a hidden ortholog.

Leapfrog identified hundreds of hidden orthologs in flatworm transcriptomes

To validate 'Leapfrog', 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. Under these conditions, ‘Leapfrog’ identified a total of 3,597 hidden orthologs, 1,243 of which were unique and 671 were species-specific (fig. 2A, B; supplementary table 2, Supplementary Material online). From the annotation of their human ortholog, the hidden orthologs represented a wide array of different proteins, from genes involved in signaling transduction (e.g. GFRA3, a *GDNF family receptor alpha-3*) to oncogenes (e.g. BRCA2, the *breast cancer type 2 susceptibility protein*) and cytoskeleton regulators (e.g. COBLL1 or *cordon-bleu*). 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 hidden orthologs recovered in each particular lineage ranged from 41 in the rhabdocoel *Provortex sphagnorum* to 198 in the planarian *S. mediterranea* (fig. 3). The number of hidden orthologs varied considerably between different species belonging to the same group of flatworms. Within Tricladida, for instance, we identified 125 hidden orthologs in the marine species *Bdelloura candida*, 183 in the continenticolan species *Dendrocoelum lacteum* and 198 in the model species *S. mediterranea*. However, we only recovered 71 hidden orthologs for *Dugesia tigrina*, a freshwater planarian related to *S. mediterranea*. We observed a similar issue in Macrostomorpha, Prorhynchida, and Rhabdocoela (fig. 3). Interestingly, the 'Leapfrog' pipeline also reported hidden orthologs in the outgroup taxa (*Austrognathia* sp., 63; *L. patella*, 21; and *L. squamata*, 35) and *Microstomum lineare* (71), a flatworm lineage that shows a slower rate of evolutionary change than *P. vittatus* (Laumer, Hejnol, et al. 2015).

To asses how the completeness of each transcriptome was influencing 'Leapfrog', we calculated the proportion of core eukaryotic genes (CEGs) (Parra, et al. 2007) present in each transcriptome. Consistent with the differences in sequencing depth (supplementary table 1, Supplementary Material online), we observed a broad range of CEG content between transcriptomes: from a reduced 8% in *P. sphagnorum* –the flatworm transcriptome with less recovered hidden orthologs– to an almost complete 99% of the polyclad *Stylochus ellipticus* and our "bridge" species *P. vittatus* (fig. 3). Importantly, our dataset included highly complete transcriptomes (with > 85% CEGs) for each major flatworm group (Macrostomorpha, Polycladida, Prorhynchida, Rhabdocoela, Proseriata, Adiaphanida, and Neodermata).

The comparison of these highly complete transcriptomes with the other representatives of their respective groups showed that the number of recovered hidden orthologs was in many cases species-dependent. For instance, we recovered 85 putative hidden orthologs in *Geocentrophora applanata* and 137 in *Prorhynchus* sp. I, despite both prorhynchids having highly complete transcriptomes (fig. 3). The opposite case can be seen in the Macrostomorpha, where 71 (five species-specific) and 75 (four species-specific) hidden orthologs were recovered in *Microstomum lineare* and *Macrostomum lignano* respectively, both of which have highly complete transcriptomes. However, we identified 129 hidden orthologs (34 species-specific) in the closely related macrostomorph *Macrostomum* cf. *ruebushi*, whose transcriptome showed only a 60% of CEGs (fig. 3). These results together suggest that the number of hidden orthologs we recovered with 'Leapfrog' is sensitive to the quality of the transcriptomes, but overall seems to be strongly restricted by species.

We evaluated whether the use of a different 'bridge' transcriptome –with comparable completeness as *P. vittatus*– could be used to recover even more hidden orthologs in our datasets. We used the transcriptome of *M. lineare* because this species had the shortest branch in a published phylogenomic study (Laumer, Hejnol, et al. 2015). Using *M. lineare* as a 'bridge' we predicted hidden orthologs in the transcriptome of *S. mediterranea*, the lineage with the most hidden orthologs identified using *P. vittatus* as a 'bridge.' Surprisingly, we only recovered 62 putative hidden orthologs under these conditions, as opposed to 198 when using *P. vittatus*, suggesting that evolutionary rate is not necessarily the best criteria for choosing a 'bridge' lineage. Noticeably, only 33 of the recovered 169 unique hidden orthologs overlapped between the two analyses, demonstrating the potential of using different transcriptomes as 'bridges' to identify additional hidden orthologs.

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

To investigate the parameters that might influence the evolutionary appearance and methodological identification of hidden orthologs in our dataset, we first performed a principal component analysis (PCA) including variables related to the quality and completeness of the transcriptome (number of sequenced bases, number of assembled contigs, mean contig length, and number of CEGs), the mean base composition of the transcriptome (GC content) and the evolutionary rate of each lineage (branch length, and number of identified hidden orthologs) (fig. 4A; supplementary table 3, Supplementary Material online). We observed that 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 strong linear correlation ($R^2 = 0.124$; fig. 4B) between these two variables, even when we only considered those transcriptomes with similar completeness (≥ 85% CEGs identified; $R^2 = 0.332$). This result supported our previous observation that lineages with similar branch lengths could exhibit remarkably different sets of hidden orthologs (fig. 3).

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). To test whether a similar case is observed in the flatworm hidden orthologs, we first plotted the GC content and average length of the G/C stretches of all recovered hidden orthologs and compared them with all flatworm transcripts (fig. 4C). Contrary to the situation observed in birds, hidden orthologs in flatworms do not show a significantly different GC content and average length of G/C stretches than the majority of transcripts. We confirmed this observation for each particular transcriptome of our dataset (fig. 4C; supplementary fig. 1, Supplementary Material online).

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). However, the length of the flatworm hidden transcripts are not significantly different from that of the rest of the transcripts (fig. 4D; supplementary table 4, Supplementary Material online).

We next performed a domain-composition analysis of the 1,243 non-redundant candidates, to address whether 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 (supplementary table 6, Supplementary Material online). 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 IPT/TIG domain (Aravind and Koonin 1999; Bork, et al. 1999), 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 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. 4E). Altogether, these analyses indicate that hidden orthologs do not show intrinsic properties that could cause systematic errors during homology searches.

Flatworm hidden orthologs include multiple GO categories

We next asked whether hidden orthologs were associated with particular biological traits of flatworm lineages. We thus performed a gene ontology (GO) analysis of the human orthologs for the 1,243 non-redundant hidden orthologs identified in our flatworm transcriptomes. We recovered a wide spectrum of GO terms describing biological processes (fig. 5A) and cellular components (fig. 5B), with no particular predominant GO category. In contrast, in the analyses of molecular function, binding and catalytic activities were more abundant among hidden ortholog GO categories (fig. 5C). A similar distribution of GO terms was observed with the 198 non-redundant candidate genes recovered from the planarian *S. mediterranea* (fig. 5D-F). The statistical comparison of the GO categories of the hidden orthologs identified in *S. mediterranea* with its whole annotated transcriptome revealed 248 significantly ($p < 0.05$) enriched GO terms, 145 of them corresponding to the biological process category, 70 to the cellular component category, and 33 to the molecular function (table 2; supplementary table 7, Supplementary Material online). Interestingly, hidden orthologs were enriched for biological processes and cellular compartments related to mitochondrial protein translation and the mitochondrial ribosome respectively, which might be a result of the changes in the mitochondrial genetic code observed in rhabditophoran flatworms (Telford, et al. 2000). Indeed, ribosomal proteins are amongst the most common hidden orthologs recovered from our dataset (supplementary table 2, Supplementary Material online). However, we also identified five mitochondrial ribosomal proteins (39S ribosomal proteins L50, L10 and L40, and 28S ribosomal proteins S30 and S27) and three mitochondrial-related proteins (PET117, ECSIT and ATP5I genes) as hidden orthologs in the catenulid *Stenostomum leucops*, suggesting that the sequence divergence of the mitochondrial components might be independent of the genetic code modifications.

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

A previous study suggested the loss of an important proportion of centrosomal and cytoskeleton-related genes in the flatworms *M. lignano*, *S. mediterranea*, and *S. mansoni* (Azimzadeh, et al. 2012). We thus used an expanded 'Leapfrog' strategy to identify possible hidden orthologs for that group of genes in our set of flatworm transcriptomes. 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. 6). 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 the hidden orthologs for the genes CCCAP (SDCCAG8) and CEP192 in the planarian *S. mediterranea* (fig. 6; supplementary fig. 2 and supplementary fig. 3, Supplementary Material online), which were two of the five key essential centrosomal genes thought to be missing and essential for centrosome assembly and duplication (Azimzadeh, et al. 2012).

Hidden orthologs obtained in particular lineages could also be used as a “bridge” to manually identify their counterparts in other flatworm groups. For instance, we used the GFRA3 sequence from the fecampiid *Kronborgia* cf. *amphipodicola* and the FHAD1 (*fork head-associated phosphopeptide binding domain 1*) sequence from the rhabdocoel *Lehardyia* sp. to identify their putative orthologs in the planarian *S. mediterranea*. Surprisingly, the ‘Leapfrog’ pipeline did not recover many developmental genes, albeit flatworm lineages have supposedly lost important components of many developmental signaling pathways (Olson 1999; Berriman, et al. 2009; Martin-Duran and Romero 2011; Riddiford and Olson 2011; Tsai, et al. 2013; Koziol, et al. 2016). To explore the possibilities of this approach, we tried to manually identify in the planarian *S. mediterranea* classes of homeodomain genes previously reported as missing in free-living flaworms (Tsai, et al. 2013), using as a ‘bridge’ the orthologs found in the more conservative rhabditophoran species *M. lignano* and *P. vittatus*. We found orthologs for *gsc*, *dbx*, *vax*, *arx*, *drgx*, *vsx* and *cmp* in all these species (table 3; supplementary fig. 4 and supplementary fig. 5, Supplementary Material online), which places the loss of these homeodomain classes most likely at the base of the last-common neodermatan ancestor. Importantly, most of the classes absent in the transcriptomes of *P. vittatus* and *M. lignano* were also missing in *S. mediterranea*. 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 reveals thousands of hidden orthologs in Platyhelminthes (fig. 2, 3), and thus illustrates the importance of a dense taxon sampling to confidently study gene losses and gains during gene complement evolution. Nevertheless, our approach is conservative and these results are likely an underestimation of the true number of hidden orthologs in these data.

Since our goal was to demonstrate how increased taxon sampling and the use of intermediate taxa with moderate evolutionary rates can help identify fast evolving orthologs, we based our automated pipeline on BLAST searches (fig. 1B), by far the most common methodology for quickly identifying putative orthologs. However, other methods (e.g. profile HMM, PSI-BLAST) are more sensitive than BLAST when dealing with divergent sequences (Altschul and Koonin 1998; Eddy 1998), and have been shown, for instance, to recover homology relationships for many potential TRGs in viruses (Kuchibhatla, et al. 2014). Second, 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 with some limitations (Dalquen and Dessimoz 2013). Third, different ‘bridge’ transcriptomes generate different sets of hidden orthologs. This is an important observation, as it indicates that overall conservative lineages may themselves have hidden orthologs. Therefore, an approach in which each transcriptome is used both as a ‘bridge’ and as a target will likely uncover even more hidden orthologs. Furthermore, we demonstrate that using hidden orthologs themselves as 'bridge queries' on other lineages can help recover even more new hidden orthologs (table 3). Finally, 16 out of the 35 analyzed transcriptomes contain less than 80% of core eukaryotic genes (fig. 3), and can be regarded as fairly incomplete (Parra, et al. 2009). All things considered, it is highly likely that the number of hidden orthologs in these flatworm lineages is far greater than what we are able to show in this study.

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 the centrosomal proteins CEP192 and SDCCAG8 in *S. mediterranea* (fig. 6), as well as other core components in other flatworms lineages, indicates that the evolutionary events leading to the loss of centrosomes are probably more complex, or at least different from previously thought (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). These two examples illustrate how our study and computational tools can serve the flatworm research community. The use of intermediate, conservatively evolving 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 RNAseq 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 *P. vittatus* 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, and is freely available on GitHub (see Materials and Methods).

In a broader context, our study may help clarify the composition of animal gene repertoires. Because they have diverged beyond the threshold of similarity searches, hidden orthologs can be simultaneously interpreted as false positive TRGs and false negative missing genes. From our conservative approach, we estimate that hidden orthologs comprise around 1% of the whole proteome of *S. mediterranea* (227/26,008; number of predicted unigenes in the sexual strain in SmedGD 2.0) (Robb, et al. 2015), but as discussed above, there are likely many more. Considering that TRGs often represent around 10-20% of the gene complement (Khalturin, et al. 2009), our study suggests that at least 5–10% of the presumed TRGs are indeed hidden orthologs (i.e. false positives).

In our dataset, hidden orthologs are not significantly shorter, and do not exhibit either particular sequence composition biases (fig. 4) or protein domains (table 1) that could account for the difficulties in being detected by standard homology searches. Instead, hidden orthologs seem to represent restricted fast evolving orthologs, in some cases associated with divergent biological features of Platyhelminthes (fig. 5, 6; table 3). 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.

Previous studies suggested that more sensitive methods would reveal the real estimate of TRGs in animal genomes (Tautz and Domazet-Loso 2011). However, these methodologies are often time consuming and computationally intense, and thus hard to scale when dealing with large transcriptomes in a broad phylogenetic context. Our study proves that an alternative way to partially overcome this issue is by relying on improved taxon sampling, which is feasible as sequencing prices drop and the use of high-throughput sequencing becomes even more common in non-model organisms. Therefore, we envision a combination of both improved methodologies and expanded taxon sampling as the path to follow in future studies of gene complement evolution in animals.

The natural next step is to figure out what percentage of these hidden orthologs are functionally conserved. If it is a large percentage, then how are these genes able to diverge to such extremes when they are so highly conserved in most other animal lineages? One hypothesis is that such “leaps” in sequence diversity may require simultaneous mutations in different parts of the gene, since function-maintaining mutational space available to one-at-a-time mutations is small. Another hypothesis supported by the preponderance of hidden orthologs involved in binding (fig. 5B,E) is that hidden orthologs are being produced by compensatory mutations in binding partners. In both of these cases, genomes experiencing very high mutation rates like Platyhelminthes are especially suited to explore this larger mutational space.

Altogether, our study uncovers a so-far neglected fraction of the gene repertoire of animal genomes (fig. 7). Overlooked by common similarity searches, hidden orthologs include genes of biological relevance that were thought missing from the transcriptome/genome of most flatworms. These hidden genes are either 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. Either way, these results suggest that the prevalence of missing genes and orphan genes is likely exaggerated, and that caution is necessary in interpreting gene loss and gain when analyzing genomes.

Materials and methods

Macrostomum lignano *transcriptome*

Adult and juveniles of *M. lignano* were kept under laboratory conditions as described elsewhere (Rieger, et al. 1988). Animals starved for four days were homogenized and used as source material to isolate total RNA with the TRI Reagent (Life Technologies) following the manufacturer's recommendations. A total of 1 μg was used for Illumina paired-end library preparation and sequencing in a HiSeq 2000 platform. Paired-end reads were assembled *de novo* with Trinity v.r20140717 using default settings (Grabherr, et al. 2011).

Data set preparation

We downloaded the Human RefSeq FASTA file from the NCBI FTP site last updated on March 25, 2015 (ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/H_sapiens/protein/protein.fa.gz). We also downloaded the gene2accession data file from NCBI, which was last updated on July 3, 2015 (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2accession.gz). We then used the reduce_refseq script (available at https://github.com/josephryan/reduce_refseq) to generate a non-redundant Human RefSeq FASTA file with the following command:

(reduce_refseq --fasta=protein.fa.gz --gene2accession=gene2accession.gz > HumRef2015.fa). This script prints only the first isoform for each Gene ID in the RefSeq FASTA file. The resulting file (available from the reduce_refseq repository) will be hereafter referred to as HumRef2015. Additionally, we downloaded the 28 RNA-Seq *de novo* assemblies from (Laumer, Hejnol, et al. 2015) and 6 additional *S. mediterranea* datasets from PlanMine v1.0 (Brandl, et al. 2016) on May 29, 2015. On July 14, 2015 we downloaded *Schistosoma mansoni, Hymenolepis microstoma,* and *Girardia tigrina* gene models from the Sanger FTP site. Further details on datasets are available in supplementary table 1 (supplementary Material online).

Leapfrog Pipeline

All BLASTs were conducted using BLAST+ version 2.2.31 using multiple threads (from 2 to 10 per BLAST). We first ran a TBLASTN search using HumRef2015 as a query against the *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 as a query against the HumRef2015 dataset (blastx -query Pvit -db HumRef2015 -outfmt 6 -out Pv_v_Hs). We ran a series of TBLASTX searches using the *Prostheceraeus vittatus* transcriptome as a query against each of our target transcriptome database (e.g., tblastx -query “TRANSCRIPTME” -db Pvit -outfmt 6 -out “TRANSCRIPTME”_v_Pvit). Lastly, we ran a series of TBLASTX searches using our transcriptome databases as queries against the *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 (available from 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 *Prostheceraeus vittatus* transcript, and (3) the *Prostheceraeus vittatus* 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. All leapfrog output files are provided as supplementary data.

CEGMA analysis, transcriptome quality assessment, and statistics

Transcriptome completeness was evaluated with CEGMA (Parra, et al. 2007; Parra, et al. 2009). We could not run the CEGMA pipeline in the transcriptomes of *G. tigrina*, *Microdalyellia* sp. and *H. microstoma* due to an untraceable error. We calculated the contig metrics for each transcriptome assembly with TransRate (Smith-Unna, et al. 2015). Principal component analysis was performed in R and plotted using the ggplot2 package.

GC content analyses, sequence length and CAI index

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. All scripts are available upon request. The codon usage matrices for *B. candida*, *D. tigrina* and *S. mediterranea* available at the Codon Usage Database (Nakamura, et al. 2000) were used as reference to calculate the 'codon adaptation index' with CAIcal server (Puigbo, et al. 2008). For each species, hidden orthologs were compared with three sets of transcripts generated by randomly choosing the same number of sequences than the number of hidden orthologs from the complete set of CDS sequences. All values were plotted in R using the ggplot2 package.

GO and InterPro analyses

GO analyses were performed with the human ortholog sequences from HumRef2015, using the free version of Blast2GO v3. 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). GO enrichment analysis of *S. mediterranea* hidden orthologs was performed with Blast2GO v3 comparing the GO annotations of the hidden orthologs against the GO annotations of the whole *S. mediterranea* transcriptome. InterProScan 5 was used to analyze the domain architecture of the recovered hidden orthologs using the human ortholog sequence.

Multiple sequence alignments and orthology assignment

Full-length protein sequences of the human and *P. vittatus* SDCCAG8 gene were aligned to the SDCCAG8 cryptic ortholog recovered for *S. mediterranea*. Alignment was performed with MAFFT v.5 (Katoh and Standley 2013) using the G-INS-i option. Resulting alignment was trimmed between positions 319 and 494 of the human protein and edited with Illustrator CS6 (Adobe) to show the conserved residues between the three species. Multiple sequence protein alignments were constructed with MAFFT v.5 and spuriously aligned regions were removed with gblocks 3 (Talavera and Castresana 2007). Alignments are available upon request. Orthology assignments were performed with RAxML v8.2.6 (Stamatakis 2014) with the autoMRE option. The models of protein evolution (CEP192: RtRev+I+G+F; CCCAP: JTT+G+F; Homeodomains:

LG+G) were calculated with ProtTest (Abascal, et al. 2005). Resulting trees were edited with FigTree and Illustrator CS6 (Adobe).

Competing interests

The authors declare that they have no competing interests.

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.

Acknowledgements

We thank the members of the Hejnol's lab for support and discussions, and in particular Daniel Thiel and Annie Boddington for taking care of the *M. lignano* cultures. We appreciate the advance access given to platyhelmith transcriptomes by Gonzalo Giribet and Chris Laumer at the beginning of the project. 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.

References

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104-2105.

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, Koonin EV. 1998. Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. Trends Biochem Sci 23:444-447.

Aravind L, Koonin EV. 1999. Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. J Mol Biol 287:1023-1040.

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

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.

Bork P, Doerks T, Springer TA, Snel B. 1999. Domains in plexins: links to integrins and transcription factors. Trends Biochem Sci 24:261-263.

Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, Amores A, Desvignes T, Batzel P, Catchen J, et al. 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nat Genet 48:427-437.

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.

Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES. 2007. Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci U S A 104:19428-19433.

Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, et al. 2011. The ecoresponsive genome of *Daphnia pulex*. Science 331:555-561.

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.

De Robertis EM. 2008. Evo-devo: variations on ancestral themes. Cell 132:185-195.

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.

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.

Eddy SR. 1998. Profile hidden Markov models. Bioinformatics 14:755-763.

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.

Edvardsen RB, Seo HC, Jensen MF, Mialon A, Mikhaleva J, Bjordal M, Cartry J, Reinhardt R, Weissenbach J, Wincker P, et al. 2005. Remodelling of the homeobox gene complement in the tunicate *Oikopleura dioica*. Curr Biol 15:R12-13.

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.

Grabherr M, Haas BJ, Yassour M, Levin J, Thompson D, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol 29:644-652.

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.

Kao D, Felix D, Aboobaker A. 2013. The planarian regeneration transcriptome reveals a shared but temporally shifted regulatory program between opposing head and tail scenarios. BMC Genomics 14:797.

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.

Knowles DG, McLysaght A. 2009. Recent *de novo* origin of human protein-coding genes. Genome Res 19:1752-1759.

Kortschak RD, Samuel G, Saint R, Miller DJ. 2003. EST analysis of the cnidarian *Acropora millepora* reveals extensive gene loss and rapid sequence divergence in the model invertebrates. Curr Biol 13:2190-2195.

Koziol U, Jarero F, Olson PD, Brehm K. 2016. Comparative analysis of Wnt expression identifies a highly conserved developmental transition in flatworms. BMC Biol 14:10.

Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res 13:2229-2235.

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. 2015. Spiralian phylogeny informs the evolution of microscopic lineages. Curr Biol 25:2000-2006.

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

Li CY, Zhang Y, Wang Z, Zhang Y, Cao C, Zhang PW, Lu SJ, Li XM, Yu Q, Zheng X, et al. 2010. A human-specific *de novo* protein-coding gene associated with human brain functions. PLoS Comput Biol 6:e1000734.

Long M, Betran E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. Nat Rev Genet 4:865-875.

Loppin B, Lepetit D, Dorus S, Couble P, Karr TL. 2005. Origin and neofunctionalization of a *Drosophila* paternal effect gene essential for zygote viability. Curr Biol 15:87-93.

Martin-Duran JM, de Mendoza A, Sebe-Pedros A, Ruiz-Trillo I, Hejnol A. 2013. A broad genomic survey reveals multiple origins and frequent losses in the evolution of respiratory hemerythrins and hemocyanins. Genome Biol Evol 5:1435-1442.

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.

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.

Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.

Olson MV. 1999. When less is more: gene loss as an engine of evolutionary change. Am J Hum Genet 64:18-23.

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.

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

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

Rieger R, Gehlen M, Haszprunar G, Holmlund M, Legniti A, Salvenmoser W, Tyler S. 1988. Laboratory cultures of marine Macrostomida (Turbellaria). Progr Zool 36:523.

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.

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.

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.

Technau U, Rudd S, Maxwell P, Gordon PM, Saina M, Grasso LC, Hayward DC, Sensen CW, Saint R, Holstein TW, et al. 2005. Maintenance of ancestral complexity and non-metazoan genes in two basal cnidarians. Trends Genet 21:633-639.

Telford MJ, Herniou EA, Russell RB, Littlewood DT. 2000. Changes in mitochondrial genetic codes as phylogenetic characters: two examples from the flatworms. Proc Natl Acad Sci U S A 97:11359-11364.

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.

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.

Warnefors M, Eyre-Walker A. 2011. The accumulation of gene regulation through time. Genome Biol Evol 3:667-673.

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.

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.

731 **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
PF01833	IPT/TIG domain	86.6	18%	EXOC2, PLXNA4, EBF3, EBF2, PLXNA1, EBF4
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
PF12763	Cytoskeletal-regulatory complex EF hand	95	31%	EHD2, EHD3, EHD4, EHD1
PF00536/ PF07647	SAM (Sterile alpha motif) domain	63.1/64.8	23%/20%	SAMD4A, SASH1, SAMD3, CNKSR3, SAMD10, SAMD15, SAMD15, SASH1

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

733 [b] Average values based on PFAM model

734

Table 2. Enriched GO categories in *S. mediterranea* hidden orthologs

GO term	Description	E-value
Biological process		
GO:0070124	Mitochondrial translational initiation	2.69E-12
GO:0070126	Mitochondrial translational termination	4.07E-12
GO:0070125	Mitochondrial translational elongation	6.05E-12
GO:0032543	Mitochondrial translation	4.76E-07
GO:0016064	Immunoglobulin mediated immune response	1.38E-06
GO:0019724	B cell mediated immunity	1.38E-06
Molecular function		
GO:0001056	RNA polymerase III activity	9.55E-03
GO:0005121	Toll binding	1.07E-02
GO:0000989	Transcription factor activity, transcription factor binding	4.05E-02
GO:0001635	Calcitonin gene-related peptide receptor activity	1.07E-02
GO:0043237	Laminin-1 binding	1.07E-02
GO:0005540	Hyaluronic acid binding	1.07E-02
Cellular compartment		
GO:0005761	Mitochondrial ribosome	3.87E-08
GO:0000313	Organellar ribosome	6.08E-08
GO:0005743	Mitochondrial inner membrane	9.09E-07
GO:0019866	Organelle inner membrane	6.39E-06
GO:0005762	Mitochondrial large ribosomal subunit	1.19E-05
GO:0031966	Mitochondrial membrane	4.54E-05

Table 3. 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]	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.

Figures

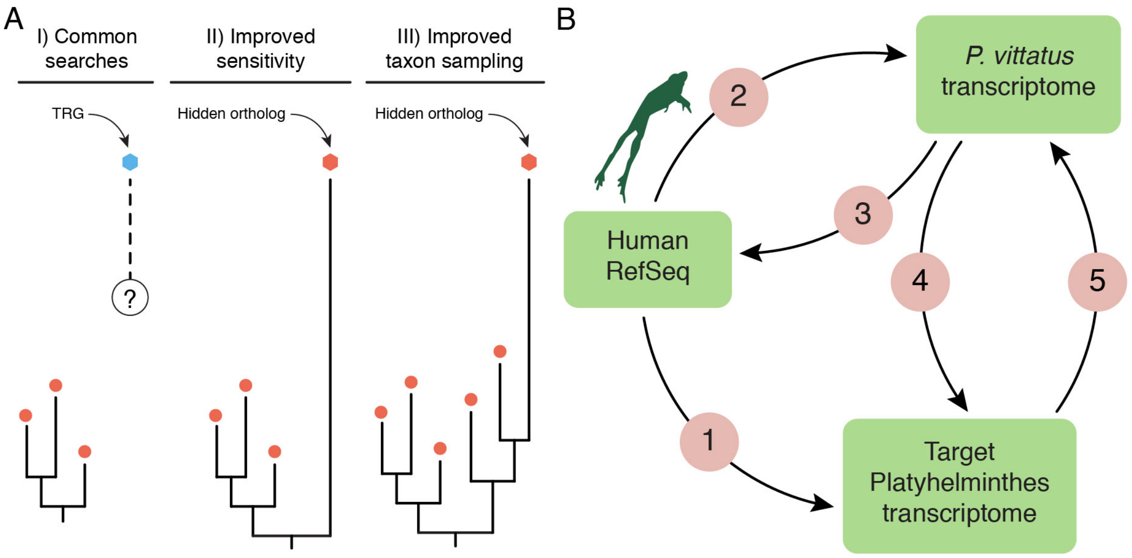

Figure 1. Hidden orthologs and the 'Leapfrog' pipeline. (**A**) Taxonomically-restricted genes (TRGs) are genes with no clear orthology relationship (dashed line and question mark) to other known genes (e.g. orthology group of red dots). Improved sensitivity in the detection methods and/or improved taxon 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 human RefSeq and the 'bridge' (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.

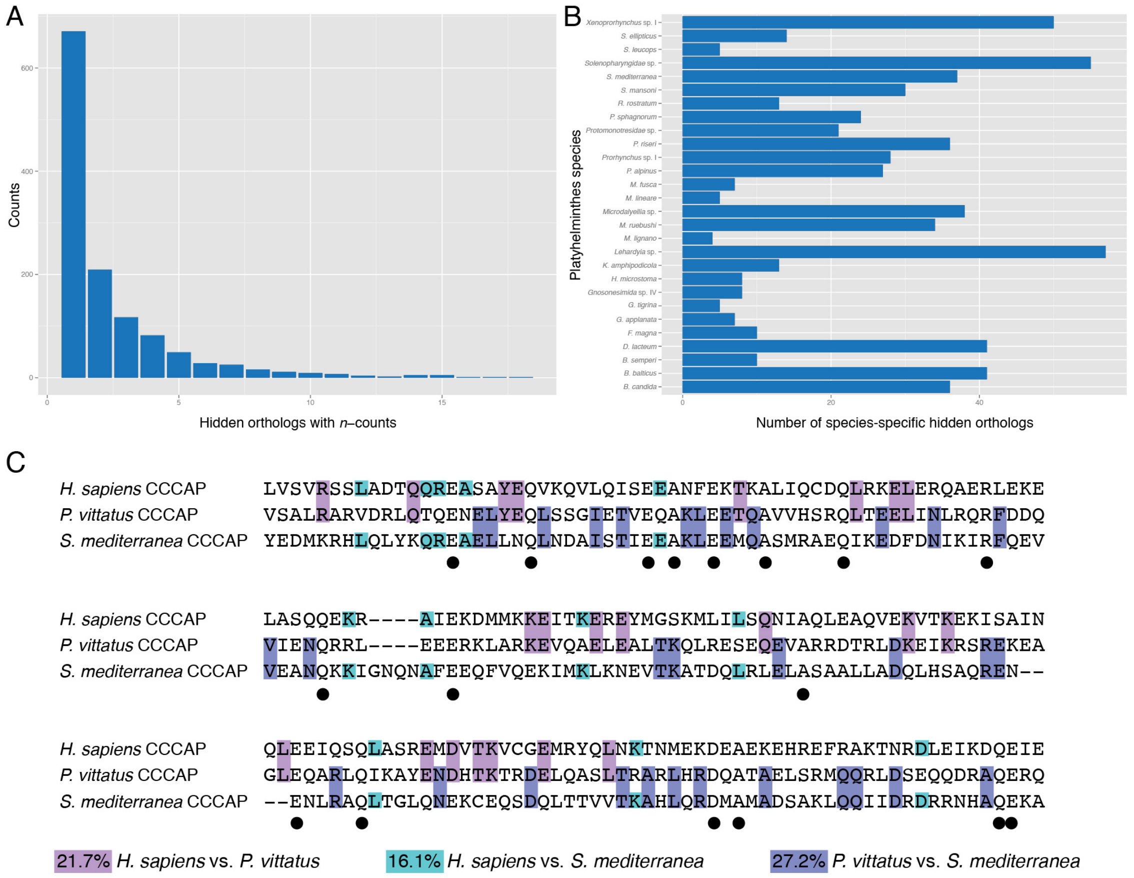

Figure 2. The Leapfrog pipeline recovers hundreds of hidden orthologs in Platyhelminthes. (**A**) Distribution of hidden orthologs according to their identification in one or more of the analyzed transcriptomes. Most of the hidden orthologs are unique of each lineage. (**B**) Distribution of species-specific hidden orthologs in each studied species. (**C**) Amino acid alignment of a fragment of the centrosomal protein CCCAP 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.

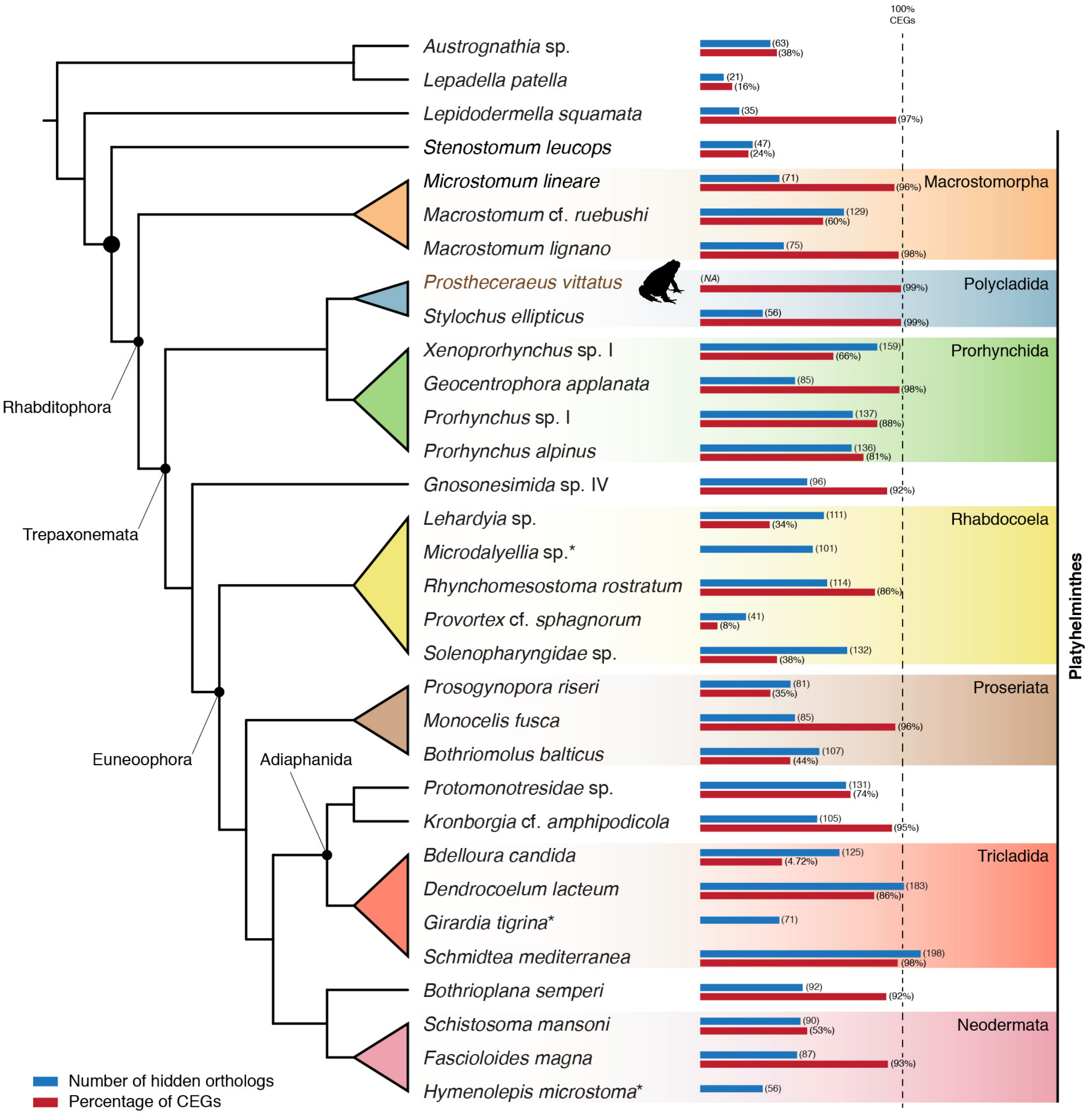

Figure 3. Distribution of hidden orthologs in the analyzed flatworm transcriptomes. The figure shows the total number of hidden orthologs in the analyzed transcriptomes in a phylogenetic context and with respect to their completeness (percentage of recovered core eukaryote genes, CEGs). The quality of the transcriptomes seems to be a limitation for the recovery of hidden orthologs in some flatworm lineages (e.g. *Provortex* cf. *sphagnorum*). However, the number of hidden orthologs is very species-specific.

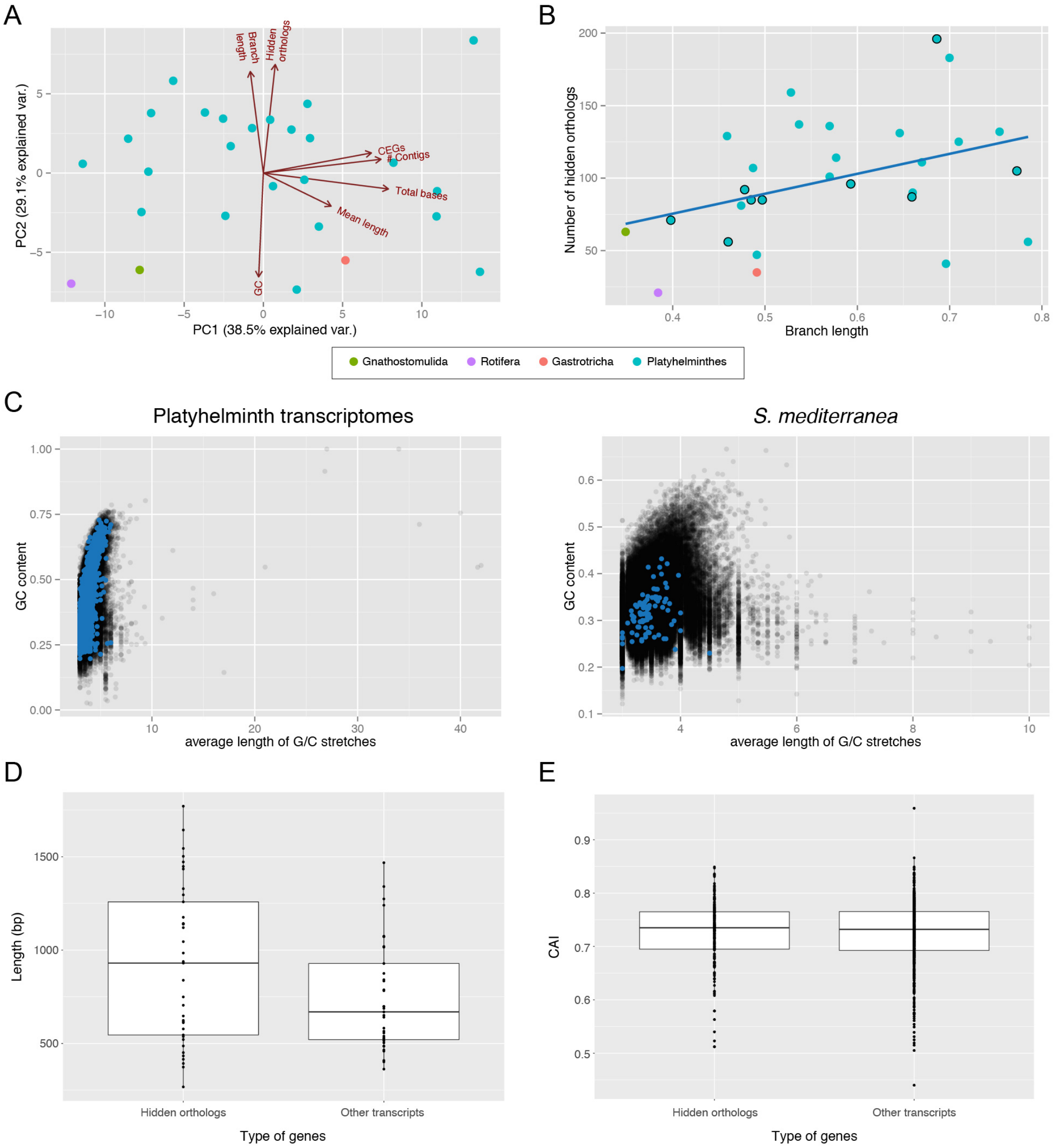

Figure 4. Hidden orthologs, evolutionary rates and sequence composition analyses. (**A**) Principal component analysis of the analyzed data showing the eigenvectors for each variable. The two first principal components (PC1, PC2) explain together 67.6% of the observed variability. (**B**) Number of hidden 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 of the transcriptome. Hidden orthologs are not 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; $p<0.05$).

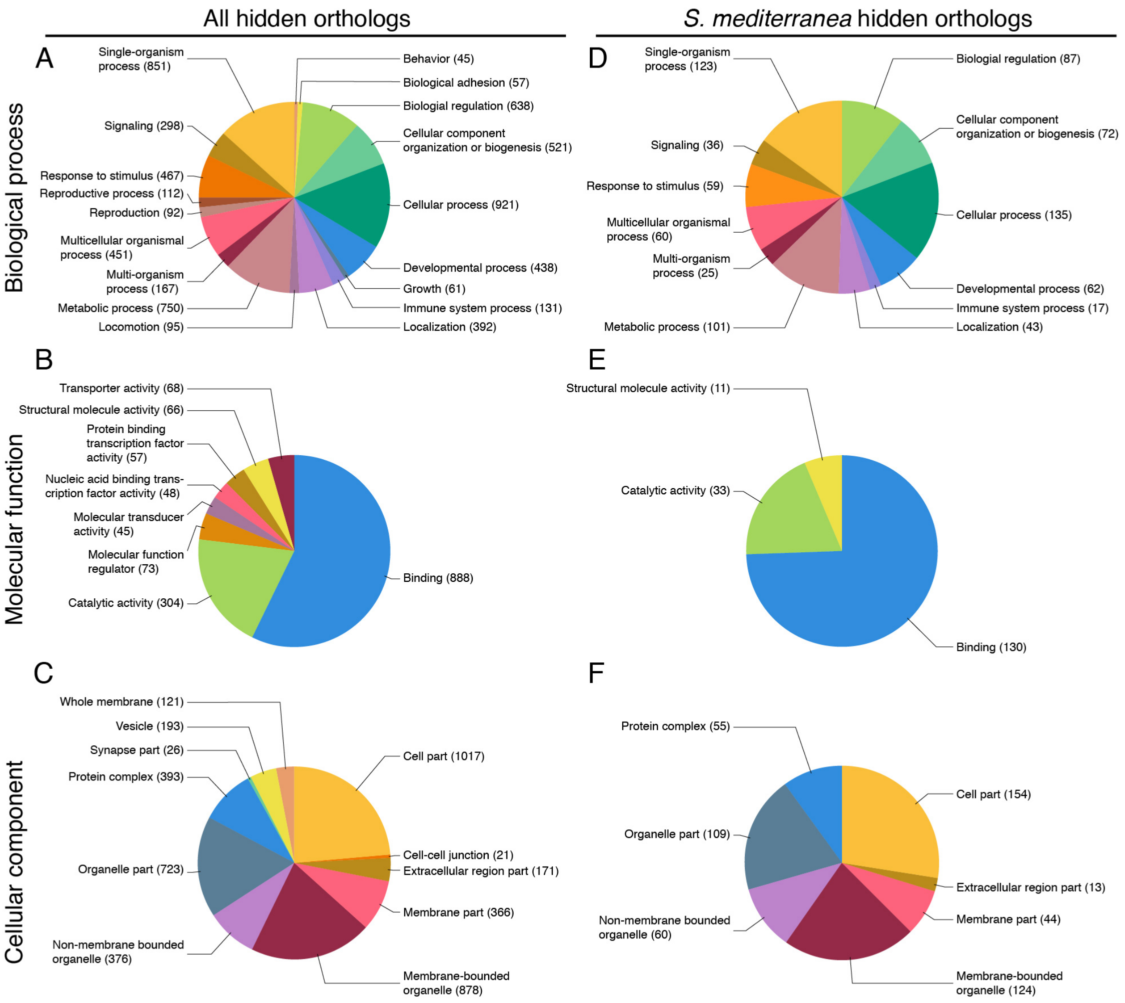

Figure 5. Gene Ontology (GO) characterization of hidden orthologs. Distribution of GO terms for all recovered hidden orthologs (**A**–**C**) and for the hidden orthologs identified in *S. mediterranea* (**D**–**F**). Hidden orthologs include a great diversity of GO categories, with a big proportion of binding and catalytic activity. The number of GO nodes in each category is indicated in parentheses.

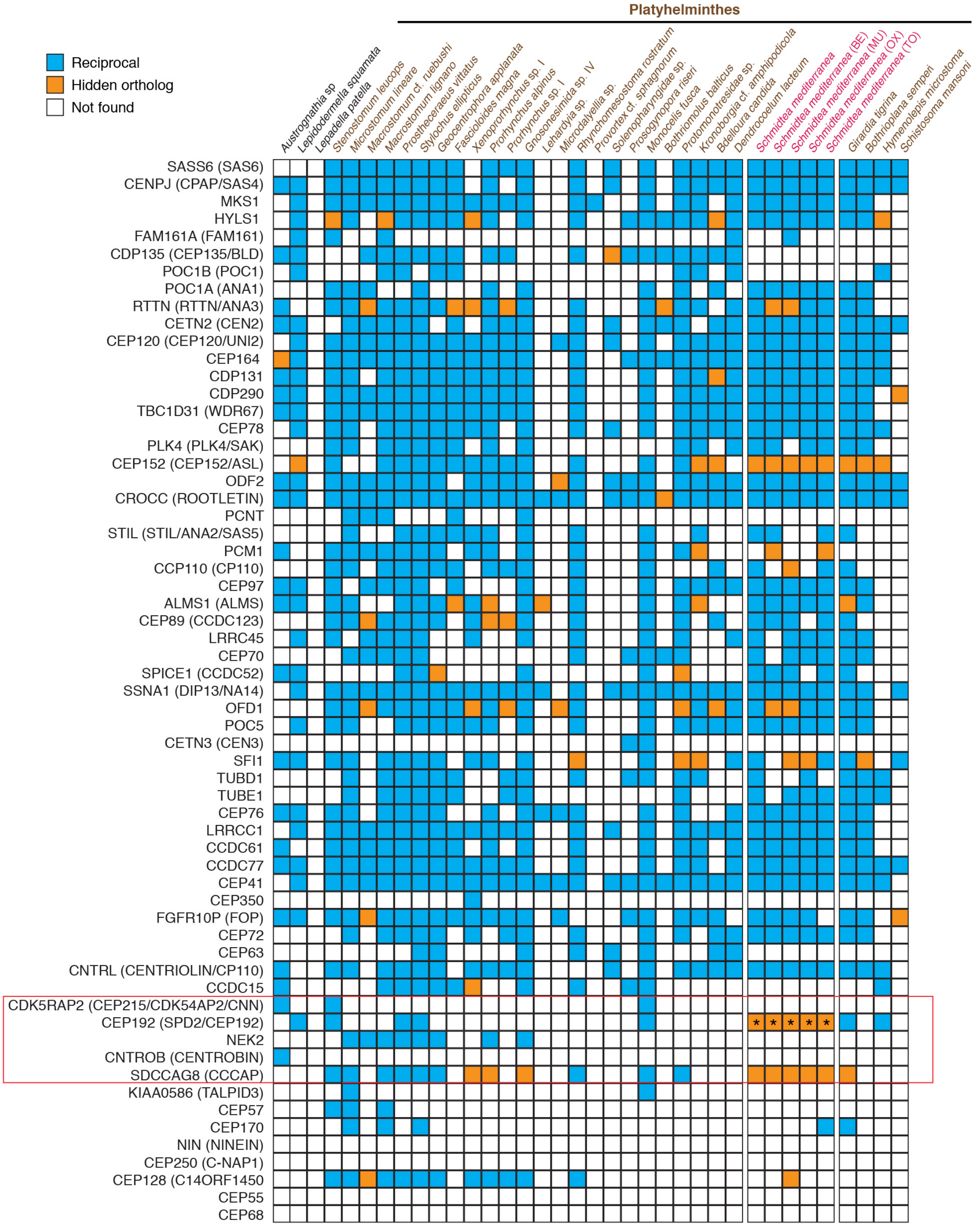

Figure 6. Hidden orthologs in the core set of centrosomal-related proteins. Presence (colored boxes) and absence (empty boxes) of the core set of centrosomal proteins (Azimzadeh, et al. 2012) in all the analyzed flatworm transcriptomes. Orthologs identified by direct reciprocal best BLAST hit are in blue boxes, and hidden orthologs are in orange. The CEP192 protein in the *S. mediterranea* transcriptomes (pink color

code) is indicated by asterisks. These proteins were manually identified with the *G. tigrina* CEP192 as ‘bridge’ by reciprocal best BLAST hit. The five proteins essential for centrosomal replication are squared in red.

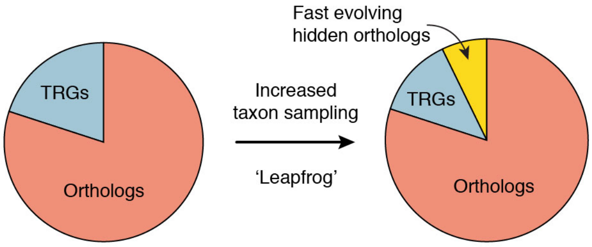

Figure 7. Increased taxon sampling uncovers fast-evolving hidden orthologs. Taxonomically restricted genes (TRGs) usually comprise 10-20% of the gene repertoire (left). Increasing taxon sampling in the group of study and the use of a 'slow evolving' intermediate species (i.e. 'Leapfrog' strategy) helps identify part of the TRGs of a given lineage as fast-evolving hidden orthologs, thus diminishing both the number of TRGs and inferred gene losses. The proportion of TRGs, common orthologs and hidden orthologs are not to scale.

Supplementary Material

Supplementary Figure 1. GC content in flatworm transcriptomes. GC content of each transcript plotted against its average length of G/C stretches for each flatworm species under study. The transcripts corresponding to hidden orthologs are in blue. Hidden orthologs do not differentiate from the majority of transcripts.

Supplementary Figure 2. Orthology analysis of the centrosomal CEP192 protein. CEP192 proteins do not contain any identifiable protein domain, and there is no known related protein that can help root the tree. Flatworm sequences are highlighted in red.

Supplementary Figure 3. Orthology analysis of the centrosomal CCCAP protein. CCCAP proteins contain a CCCAP domain (PFAM: PF15964), which is exclusive of these proteins. The domain is clearly recognizable in all flatworm sequences except *P. alpinus* (fragment too short) and the triclads *G. tigrina* and *S. mediterranea* (too divergent). Flatworm sequences are highlighted in red.

Supplementary Figure 4. Orthology analysis of the ANTP homeodomain class. The newly identify sequences in the macrostomid *M. lignano*, the polyclad *P. vittatus* and the triclad *S. mediterranea* are highlighted in red.

Supplementary Figure 5. Orthology analysis of the CUT homeodomain class. The newly identify sequences in the macrostomid *M. lignano*, the polyclad *P. vittatus* and the triclad *S. mediterranea* are highlighted in red.

Supplementary Table 1. Transcriptomes analyzed in this study.

Supplementary Table 2. Recovered hidden orthologs. Hidden orthologs (as in human RefSeq) recovered in each transcriptome after running ‘Leapfrog’ with the transcriptome of the polyclad *P. vittatus* used as the ‘bridge’.

Supplementary Table 3. Data set used for principal component analysis.

Supplementary Table 4. Length of hidden orthologs and ORFs in flatworm transcriptomes.

Supplementary Table 5. PFAM domains identified in the hidden orthologs.

Supplementary Table 6. Significantly enriched GO terms in the hidden orthologs recovered in *S. mediterranea*.