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

60 short, has an abundance of insertions or deletions and/or exhibits similarity with other
61 counterparts in only a small subset of residues (Moyers and Zhang 2015). These
62 limitations can generate significant biases when studying the evolution of protein-
63 coding gene families (Elhaik, et al. 2006; Moyers and Zhang 2015). Accordingly, a
64 proportion of the gene complement of an organism will be represented by genes that
65 lack obvious affinity with homologs in the gene sets of the best annotated genomes–
66 thus mistakenly considered potential TRGs– but actually representing fast evolving
67 orthologs that we call hidden orthologs. This systematic error can potentially be
68 overcome by more sensitive, although computationally intense, detection methods (e.g.
69 profile HMMs, PSI-BLAST) (Kuchibhatla, et al. 2014), but also by increasing taxon
70 sampling, which helps to bridge the long evolutionary gaps between hidden orthologs
71 and their well-annotated, more conservative counterparts (fig. 1A).
72
73 Platyhelminthes (flatworms) is a morphological and ecologically diverse animal group
74 characterized by significantly high rates of molecular evolution (Edgecombe, et al.
75 2011; Struck, et al. 2014; Laumer, Bekkouche, et al. 2015). Accordingly, changes in
76 gene complement seem to be important drivers of adaptive evolution in this group
77 (Berriman, et al. 2009; Martin-Duran and Romero 2011; Riddiford and Olson 2011;
78 Tsai, et al. 2013). For instance, parasitic forms (e.g. tapeworms and flukes) have many
79 unidentifiable genes and are reported to be missing myriad genes, including important
80 developmental genes that are highly conserved in most other animals (Riddiford and
81 Olson 2011; Tsai, et al. 2013). The presumed loss of critical genes has led to the
82 inference that these animals have either developed alternative ways to implement
83 critical steps in conserved pathways or that these pathways are no longer active (Wang,
84 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

110 (Azimzadeh, et al. 2012; Tsai, et al. 2013). In a broader context, our study suggests that
111 hidden orthologs likely comprise a significant proportion of the gene repertoire of every
112 organism, improving our understanding of gene complement evolution in animals.
113
114 **Results**
115 ***The 'Leapfrog' pipeline***
116 To identify hidden orthologs in large transcriptomic datasets we created 'Leapfrog',
117 which automates a series of BLAST-centric processes (fig. 1B). We started with a set of
118 well-annotated sequences –the human RefSeq protein dataset– as our main queries and
119 conducted a TBLASTN search of these sequences against each of our target flatworm
120 transcriptomes (supplementary table 1, Supplementary Material online). Any queries
121 that had zero BLAST hits with E-values less than our cutoff (0.01) were considered
122 candidate hidden orthologs. We then looked for reciprocal best TBLASTX hits between
123 these candidates and the transcriptome of the polyclad flatworm *Prostheceraeus*
124 *vittatus*, a lineage that evolves at a slower rate than most other flatworms with available
125 sequence data (as evidenced by branch lengths in (Laumer, Hejnol, et al. 2015)). If there
126 was a reciprocal best BLAST hit in our 'bridge' transcriptome, the 'bridge' transcript
127 was used as query in a BLASTX search against the initial annotated human RefSeq
128 protein dataset. If there was a human reciprocal hit, and the human sequence was the
129 starting query, then we deemed the candidate a hidden ortholog.
130
131 ***Leapfrog identified hundreds of hidden orthologs in flatworm transcriptomes***
132 To validate 'Leapfrog', we assembled a dataset including 35 publicly available
133 transcriptomes from 29 flatworm species, and incorporated the transcriptomes of the
134 gastrotrich *Lepidodermella squamata*, the rotifer *Lepadella patella*, and the

135 gnathostomulid *Austrognathia* sp. as closely related outgroup taxa. Under these
136 conditions, 'Leapfrog' identified a total of 3,597 hidden orthologs, 1,243 of which were
137 unique and 671 were species-specific (fig. 2A, B; supplementary table 2,
138 Supplementary Material online). From the annotation of their human ortholog, the
139 hidden orthologs represented a wide array of different proteins, from genes involved in
140 signaling transduction (e.g. GFRA3, a *GDNF family receptor alpha-3*) to oncogenes
141 (e.g. BRCA2, the *breast cancer type 2 susceptibility protein*) and cytoskeleton
142 regulators (e.g. COBLL1 or *cordon-bleu*). Alignments of recovered hidden orthologs
143 with their human and *P. vittatus* counterparts show that many amino acid positions that
144 differ between the human and the hidden ortholog products are conserved between *P.*
145 *vittatus* and one or the other sequences (e.g., fig. 2C).
146
147 The number of hidden orthologs recovered in each particular lineage ranged from 41 in
148 the rhabdocoel *Provortex sphagnorum* to 198 in the planarian *S. mediterranea* (fig. 3).
149 The number of hidden orthologs varied considerably between different species
150 belonging to the same group of flatworms. Within Tricladida, for instance, we identified
151 125 hidden orthologs in the marine species *Bdelloura candida*, 183 in the
152 continenticolan species *Dendrocoelum lacteum* and 198 in the model species *S.*
153 *mediterranea*. However, we only recovered 71 hidden orthologs for *Dugesia tigrina*, a
154 freshwater planarian related to *S. mediterranea*. We observed a similar issue in
155 Macrostomorpha, Prorhynchida, and Rhabdocoela (fig. 3). Interestingly, the 'Leapfrog'
156 pipeline also reported hidden orthologs in the outgroup taxa (*Austrognathia* sp., 63; *L.*
157 *patella*, 21; and *L. squamata*, 35) and *Microstomum lineare* (71), a flatworm lineage
158 that shows a slower rate of evolutionary change than *P. vittatus* (Laumer, Hejnol, et al.
159 2015).

160

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

171

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

185

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

198

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

200 To investigate the parameters that might influence the evolutionary appearance and
201 methodological identification of hidden orthologs in our dataset, we first performed a
202 principal component analysis (PCA) including variables related to the quality and
203 completeness of the transcriptome (number of sequenced bases, number of assembled
204 contigs, mean contig length, and number of CEGs), the mean base composition of the
205 transcriptome (GC content) and the evolutionary rate of each lineage (branch length,
206 and number of identified hidden orthologs) (fig. 4A; supplementary table 3,
207 Supplementary Material online). We observed that the first principal component (PC1)
208 was strongly influenced by the quality of the transcriptome, while the second principal
209 component (PC2) mostly estimated the balance between evolutionary change (branch

210 lengths and hidden orthologs) and transcriptome complexity (GC content). The two first
211 principal components explained 67% of the variance of the dataset, indicating that
212 additional interactions between the variables exist (e.g. the GC content can affect
213 sequencing performance (Dohm, et al. 2008; Benjamini and Speed 2012), and thus
214 transcriptome quality and assembly).
215
216 Despite the fact that the branch length of a given lineage and the number of putative
217 hidden orthologs affected the dispersion of our data in a roughly similar manner, we did
218 not detect a strong linear correlation ($R^2 = 0.124$; fig. 4B) between these two variables,
219 even when we only considered those transcriptomes with similar completeness (≥ 85%
220 CEGs identified; $R^2 = 0.332$). This result supported our previous observation that
221 lineages with similar branch lengths could exhibit remarkably different sets of hidden
222 orthologs (fig. 3).
223
224 ***Flatworm hidden orthologs do not show sequence composition biases***
225 A recent report showed that very high GC content and long G/C stretches characterize
226 genes mistakenly assigned as lost in bird genomes (Hron, et al. 2015). To test whether a
227 similar case is observed in the flatworm hidden orthologs, we first plotted the GC
228 content and average length of the G/C stretches of all recovered hidden orthologs and
229 compared them with all flatworm transcripts (fig. 4C). Contrary to the situation
230 observed in birds, hidden orthologs in flatworms do not show a significantly different
231 GC content and average length of G/C stretches than the majority of transcripts. We
232 confirmed this observation for each particular transcriptome of our dataset (fig. 4C;
233 supplementary fig. 1, Supplementary Material online).
234

235 Systematic error in sequence-similarity searches is also associated with the length of the
236 sequence and the presence of short conserved stretches (i.e. protein domains with only a
237 reduced number of conserved residues). Short protein lengths decrease BLAST
238 sensitivity (Moyers and Zhang 2015). We thus expected hidden orthologs to consist of
239 significantly shorter proteins, as is seen in *Drosophila* orphan genes (Palmieri, et al.
240 2014). However, the length of the flatworm hidden transcripts are not significantly
241 different from that of the rest of the transcripts (fig. 4D; supplementary table 4,
242 Supplementary Material online).
243
244 We next performed a domain-composition analysis of the 1,243 non-redundant
245 candidates, to address whether hidden orthologs were enriched in particular sequence
246 motifs that could hamper their identification by common sequence similarity searches.
247 We recovered a total of 1,180 unique PFAM annotations, almost all of them present
248 only in one (1,016) or two (112) of the identified hidden orthologs (supplementary table
249 6, Supplementary Material online). The most abundant PFAM domain (table 1) was the
250 pleckstrin homology (PH) domain (PFAM ID: PF00169), which occurs in a wide range
251 of proteins involved in intracellular signaling and cytoskeleton (Scheffzek and Welti
252 2012). PH domains were present in 11 of the candidate hidden orthologs. Most other
253 abundant domains were related to protein interactions, such as the F-box-like domain
254 (Kipreos and Pagano 2000), the IPT/TIG domain (Aravind and Koonin 1999; Bork, et
255 al. 1999), the forkhead-associated domain (Durocher and Jackson 2002), and the zinc-
256 finger of C2H2 type (Iuchi 2001). These more abundant domains vary significantly in
257 average length and number of generally conserved sites (table 1).
258

259 Lastly, we looked to see if there were any patterns of codon usage associated with
260 hidden orthologs. We did not observe a statistically significant difference between the
261 codon adaptation index of hidden orthologs of the planarian species *B. candida*, *D.*
262 *tigrina* and *S. mediterranea* and other open reading frames of these transcriptomes (fig.
263 4E). Altogether, these analyses indicate that hidden orthologs do not show intrinsic
264 properties that could cause systematic errors during homology searches.
265
266 ***Flatworm hidden orthologs include multiple GO categories***
267 We next asked whether hidden orthologs were associated with particular biological
268 traits of flatworm lineages. We thus performed a gene ontology (GO) analysis of the
269 human orthologs for the 1,243 non-redundant hidden orthologs identified in our
270 flatworm transcriptomes. We recovered a wide spectrum of GO terms describing
271 biological processes (fig. 5A) and cellular components (fig. 5B), with no particular
272 predominant GO category. In contrast, in the analyses of molecular function, binding
273 and catalytic activities were more abundant among hidden ortholog GO categories (fig.
274 5C). A similar distribution of GO terms was observed with the 198 non-redundant
275 candidate genes recovered from the planarian *S. mediterranea* (fig. 5D-F). The
276 statistical comparison of the GO categories of the hidden orthologs identified in *S.*
277 *mediterranea* with its whole annotated transcriptome revealed 248 significantly ($p <$
278 0.05) enriched GO terms, 145 of them corresponding to the biological process category,
279 70 to the cellular component category, and 33 to the molecular function (table 2;
280 supplementary table 7, Supplementary Material online). Interestingly, hidden orthologs
281 were enriched for biological processes and cellular compartments related to
282 mitochondrial protein translation and the mitochondrial ribosome respectively, which
283 might be a result of the changes in the mitochondrial genetic code observed in

284 rhabditophoran flatworms (Telford, et al. 2000). Indeed, ribosomal proteins are amongst
285 the most common hidden orthologs recovered from our dataset (supplementary table 2,
286 Supplementary Material online). However, we also identified five mitochondrial
287 ribosomal proteins (39S ribosomal proteins L50, L10 and L40, and 28S ribosomal
288 proteins S30 and S27) and three mitochondrial-related proteins (PET117, ECSIT and
289 ATP5I genes) as hidden orthologs in the catenulid *Stenostomum leucops*, suggesting
290 that the sequence divergence of the mitochondrial components might be independent of
291 the genetic code modifications.
292
293 ***The identified hidden orthologs fill out gaps in the flatworm gene complement***
294 A previous study suggested the loss of an important proportion of centrosomal and
295 cytoskeleton-related genes in the flatworms *M. lignano*, *S. mediterranea*, and *S.*
296 *mansoni* (Azimzadeh, et al. 2012). We thus used an expanded 'Leapfrog' strategy to
297 identify possible hidden orthologs for that group of genes in our set of flatworm
298 transcriptomes. First, we used a reciprocal best BLAST strategy to identify orthologs of
299 the human centrosomal proteins in each of our transcriptomes under study, and
300 thereafter we used 'Leapfrog' to identify any hidden member of this original gene set.
301 We recovered at least one reciprocal best BLAST hit for 56 of the 61 centrosomal
302 genes, and identified fast-evolving putative orthologs in 19 of the 61 centrosomal genes
303 (fig. 6). In total, the number of hidden orthologs identified was 58 (counting only once
304 those for the same gene in the different analyzed *S. mediterranea* transcriptomes). Most
305 importantly, we found the hidden orthologs for the genes CCCAP (SDCCAG8) and
306 CEP192 in the planarian *S. mediterranea* (fig. 6; supplementary fig. 2 and
307 supplementary fig. 3, Supplementary Material online), which were two of the five key

308 essential centrosomal genes thought to be missing and essential for centrosome
309 assembly and duplication (Azimzadeh, et al. 2012).
310
311 Hidden orthologs obtained in particular lineages could also be used as a "bridge" to
312 manually identify their counterparts in other flatworm groups. For instance, we used the
313 GFRA3 sequence from the fecampiid *Kronborgia* cf. *amphipodicola* and the FHAD1
314 (*fork head-associated phosphopeptide binding domain 1*) sequence from the rhabdocoel
315 *Lehardyia* sp. to identify their putative orthologs in the planarian *S. mediterranea*.
316 Surprisingly, the 'Leapfrog' pipeline did not recover many developmental genes, albeit
317 flatworm lineages have supposedly lost important components of many developmental
318 signaling pathways (Olson 1999; Berriman, et al. 2009; Martin-Duran and Romero
319 2011; Riddiford and Olson 2011; Tsai, et al. 2013; Koziol, et al. 2016). To explore the
320 possibilities of this approach, we tried to manually identify in the planarian *S.*
321 *mediterranea* classes of homeodomain genes previously reported as missing in free-
322 living flaworms (Tsai, et al. 2013), using as a 'bridge' the orthologs found in the more
323 conservative rhabditophoran species *M. lignano* and *P. vittatus*. We found orthologs for
324 *gsc*, *dbx*, *vax*, *arx*, *drgx*, *vsx* and *cmp* in all these species (table 3; supplementary fig. 4
325 and supplementary fig. 5, Supplementary Material online), which places the loss of
326 these homeodomain classes most likely at the base of the last-common neodermatan
327 ancestor. Importantly, most of the classes absent in the transcriptomes of *P. vittatus* and
328 *M. lignano* were also missing in *S. mediterranea*. The Hhex family was present in *P.*
329 *vittatus*, but was not identified in *M. lignano* and *S. mediterranea*, and the Prrx and
330 Shox families were present in *M. lignano*, but absent from *P. vittatus* and *S.*
331 *mediterranea* transcriptomes. These observations suggest that many of the losses of

332 homeobox genes occurred in the ancestors to the Rhabitophora and Neodermata, with
333 only a few losses of specific gene classes in particular lineages of free-living flatworms.
334
335 **Discussion**
336 Our study reveals thousands of hidden orthologs in Platyhelminthes (fig. 2, 3), and thus
337 illustrates the importance of a dense taxon sampling to confidently study gene losses
338 and gains during gene complement evolution. Nevertheless, our approach is
339 conservative and these results are likely an underestimation of the true number of
340 hidden orthologs in these data.
341
342 Since our goal was to demonstrate how increased taxon sampling and the use of
343 intermediate taxa with moderate evolutionary rates can help identify fast evolving
344 orthologs, we based our automated pipeline on BLAST searches (fig. 1B), by far the
345 most common methodology for quickly identifying putative orthologs. However, other
346 methods (e.g. profile HMM, PSI-BLAST) are more sensitive than BLAST when dealing
347 with divergent sequences (Altschul and Koonin 1998; Eddy 1998), and have been
348 shown, for instance, to recover homology relationships for many potential TRGs in
349 viruses (Kuchibhatla, et al. 2014). Second, we based our identification of hidden
350 orthologs on reciprocal best BLAST hits, a valid and widely used approach (Tatusov, et
351 al. 1997; Overbeek, et al. 1999; Wolf and Koonin 2012), but with some limitations
352 (Dalquen and Dessimoz 2013). Third, different 'bridge' transcriptomes generate
353 different sets of hidden orthologs. This is an important observation, as it indicates that
354 overall conservative lineages may themselves have hidden orthologs. Therefore, an
355 approach in which each transcriptome is used both as a 'bridge' and as a target will
356 likely uncover even more hidden orthologs. Furthermore, we demonstrate that using

357 hidden orthologs themselves as ‘bridge queries’ on other lineages can help recover even
358 more new hidden orthologs (table 3). Finally, 16 out of the 35 analyzed transcriptomes
359 contain less than 80% of core eukaryotic genes (fig. 3), and can be regarded as fairly
360 incomplete (Parra, et al. 2009). All things considered, it is highly likely that the number
361 of hidden orthologs in these flatworm lineages is far greater than what we are able to
362 show in this study.
363
364 The recovered hidden orthologs have an immediate impact on our understanding of
365 gene complement evolution in Platyhelminthes, and in particular on those lineages that
366 are subject of intense research, such as the regenerative model *Schmidtea mediterranea*
367 and parasitic flatworms (Berriman, et al. 2009; Wang, et al. 2011; Olson, et al. 2012;
368 Sánchez Alvarado 2012). The identification of fast-evolving orthologs for the
369 centrosomal proteins CEP192 and SDCCAG8 in *S. mediterranea* (fig. 6), as well as
370 other core components in other flatworms lineages, indicates that the evolutionary
371 events leading to the loss of centrosomes are probably more complex, or at least
372 different from previously thought (Azimzadeh, et al. 2012). Similarly, the presence of
373 presumably lost homeobox classes in *S. mediterranea* may affect our current view of
374 gene loss and morphological evolution in flatworms (Tsai, et al. 2013). These two
375 examples illustrate how our study and computational tools can serve the flatworm
376 research community. The use of intermediate, conservatively evolving flatworm
377 lineages, such as *P. vittatus*, can improve the identification of candidate genes, as well
378 as help with the annotation of the increasingly abundant flatworm RNAseq and genomic
379 datasets (Berriman, et al. 2009; Wang, et al. 2011; Tsai, et al. 2013; Robb, et al. 2015;
380 Wasik, et al. 2015; Brandl, et al. 2016). Therefore, we have now made available an
381 assembled version of *P. vittatus* in PlanMine, an integrated web resource of

382 transcriptomic data for planarian researchers (Brandl, et al. 2016). Importantly, the
383 'Leapfrog' pipeline can also be exported to any set of transcriptomes/predicted proteins,
384 and is freely available on GitHub (see Materials and Methods).
385
386 In a broader context, our study may help clarify the composition of animal gene
387 repertoires. Because they have diverged beyond the threshold of similarity searches,
388 hidden orthologs can be simultaneously interpreted as false positive TRGs and false
389 negative missing genes. From our conservative approach, we estimate that hidden
390 orthologs comprise around 1% of the whole proteome of *S. mediterranea* (227/26,008;
391 number of predicted unigenes in the sexual strain in SmedGD 2.0) (Robb, et al. 2015),
392 but as discussed above, there are likely many more. Considering that TRGs often
393 represent around 10-20% of the gene complement (Khalturin, et al. 2009), our study
394 suggests that at least 5–10% of the presumed TRGs are indeed hidden orthologs (i.e.
395 false positives).
396
397 In our dataset, hidden orthologs are not significantly shorter, and do not exhibit either
398 particular sequence composition biases (fig. 4) or protein domains (table 1) that could
399 account for the difficulties in being detected by standard homology searches. Instead,
400 hidden orthologs seem to represent restricted fast evolving orthologs, in some cases
401 associated with divergent biological features of Platyhelminthes (fig. 5, 6; table 3). The
402 fact that most of them are species-specific indicates that the gene complement of an
403 organism is in fact heterogeneous, composed of genes evolving at different evolutionary
404 rates (Wolfe 2004), sometimes much higher or much lower than the 'average' exhibited
405 by that lineage.
406

407 Previous studies suggested that more sensitive methods would reveal the real estimate
408 of TRGs in animal genomes (Tautz and Domazet-Loso 2011). However, these
409 methodologies are often time consuming and computationally intense, and thus hard to
410 scale when dealing with large transcriptomes in a broad phylogenetic context. Our study
411 proves that an alternative way to partially overcome this issue is by relying on improved
412 taxon sampling, which is feasible as sequencing prices drop and the use of high-
413 throughput sequencing becomes even more common in non-model organisms.
414 Therefore, we envision a combination of both improved methodologies and expanded
415 taxon sampling as the path to follow in future studies of gene complement evolution in
416 animals.
417
418 The natural next step is to figure out what percentage of these hidden orthologs are
419 functionally conserved. If it is a large percentage, then how are these genes able to
420 diverge to such extremes when they are so highly conserved in most other animal
421 lineages? One hypothesis is that such "leaps" in sequence diversity may require
422 simultaneous mutations in different parts of the gene, since function-maintaining
423 mutational space available to one-at-a-time mutations is small. Another hypothesis
424 supported by the preponderance of hidden orthologs involved in binding (fig. 5B,E) is
425 that hidden orthologs are being produced by compensatory mutations in binding
426 partners. In both of these cases, genomes experiencing very high mutation rates like
427 Platyhelminthes are especially suited to explore this larger mutational space.
428
429 Altogether, our study uncovers a so-far neglected fraction of the gene repertoire of
430 animal genomes (fig. 7). Overlooked by common similarity searches, hidden orthologs
431 include genes of biological relevance that were thought missing from the

432 transcriptome/genome of most flatworms. These hidden genes are either maintaining
433 ancestral functions despite very high mutation rates or are abandoning highly conserved
434 ancestral functions but continuing to contribute to the biology of the organism. Either
435 way, these results suggest that the prevalence of missing genes and orphan genes is
436 likely exaggerated, and that caution is necessary in interpreting gene loss and gain when
437 analyzing genomes.
438
439 **Materials and methods**
440 **Macrostomum lignano *transcriptome***
441 Adult and juveniles of *M. lignano* were kept under laboratory conditions as described
442 elsewhere (Rieger, et al. 1988). Animals starved for four days were homogenized and
443 used as source material to isolate total RNA with the TRI Reagent (Life Technologies)
444 following the manufacturer's recommendations. A total of 1 µg was used for Illumina
445 paired-end library preparation and sequencing in a HiSeq 2000 platform. Paired-end
446 reads were assembled *de novo* with Trinity v.r20140717 using default settings
447 (Grabherr, et al. 2011).
448
449 ***Data set preparation***
450 We downloaded the Human RefSeq FASTA file from the NCBI FTP site last updated
451 on March 25, 2015
452 (ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/H_sapiens/protein/protein.fa.gz). We also
453 downloaded the gene2accession data file from NCBI, which was last updated on July 3,
454 2015 (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2accession.gz). We then used the
455 reduce_refseq script (available at https://github.com/josephryan/reduce_refseq) to
456 generate a non-redundant Human RefSeq FASTA file with the following command:

457 (reduce_refseq --fasta=protein.fa.gz --gene2accession=gene2accession.gz >
458 HumRef2015.fa). This script prints only the first isoform for each Gene ID in the
459 RefSeq FASTA file. The resulting file (available from the reduce_refseq repository)
460 will be hereafter referred to as HumRef2015. Additionally, we downloaded the 28
461 RNA-Seq *de novo* assemblies from (Laumer, Hejnol, et al. 2015) and 6 additional *S.*
462 *mediterranea* datasets from PlanMine v1.0 (Brandl, et al. 2016) on May 29, 2015. On
463 July 14, 2015 we downloaded *Schistosoma mansoni, Hymenolepis microstoma,* and
464 *Girardia tigrina* gene models from the Sanger FTP site. Further details on datasets are
465 available in supplementary table 1 (supplementary Material online).
466
467 ***Leapfrog Pipeline***
468 All BLASTs were conducted using BLAST+ version 2.2.31 using multiple threads
469 (from 2 to 10 per BLAST). We first ran a TBLASTN search using HumRef2015 as a
470 query against the *Prostheceraeus vittatus* transcriptome (tblastn -query HumRef2015 -
471 db Pvit -outfmt 6 -out Hs_v_Pv). We next ran a BLASTX search using the
472 *Prostheceraeus vittatus* transcriptome as a query against the HumRef2015 dataset
473 (blastx -query Pvit -db HumRef2015 -outfmt 6 -out Pv_v_Hs). We ran a series of
474 TBLASTX searches using the *Prostheceraeus vittatus* transcriptome as a query against
475 each of our target transcriptome database (e.g., tblastx -query "TRANSCRIPTME" -db
476 Pvit -outfmt 6 -out "TRANSCRIPTME"_v_Pvit). Lastly, we ran a series of TBLASTX
477 searches using our transcriptome databases as queries against the *Prostheceraeus*
478 *vittatus* transcriptome (e.g., tblastx -query Pvit -db Sman -out Pvit_v_Sman -outfmt 6).
479 The tab-delimited BLAST outputs generated above were used as input to the 'Leapfrog'
480 program (available from https://github.com/josephryan/leapfrog). The default E-Value
481 cutoff (0.01) was used for all leapfrog runs. The leapfrog program identifies

482 HumRef2015 proteins that fit the following criteria: (1) they have no hit to a target
483 flatworm transcriptome, (2) they have a reciprocal best BLAST hit with a
484 *Prostheceraeus vittatus* transcript, and (3) the *Prostheceraeus vittatus* transcript has a
485 reciprocal best BLAST hit to the target flatworm transcriptome. The output includes the
486 HumRef2015 Gene ID, the *Prostheceraeus vittatus* transcript and the target flatworm
487 transcript. All leapfrog output files are provided as supplementary data.
488
489 ***CEGMA analysis, transcriptome quality assessment, and statistics***
490 Transcriptome completeness was evaluated with CEGMA (Parra, et al. 2007; Parra, et
491 al. 2009). We could not run the CEGMA pipeline in the transcriptomes of *G. tigrina*,
492 *Microdalyellia* sp. and *H. microstoma* due to an untraceable error. We calculated the
493 contig metrics for each transcriptome assembly with TransRate (Smith-Unna, et al.
494 2015). Principal component analysis was performed in R and plotted using the ggplot2
495 package.
496
497 ***GC content analyses, sequence length and CAI index***
498 Custom-made scripts were used to calculate the GC content of hidden orthologs and
499 transcripts of our dataset, the average length of the G/C stretches of each sequence, and
500 the length of hidden orthologs and other transcripts. All scripts are available upon
501 request. The codon usage matrices for *B. candida*, *D. tigrina* and *S. mediterranea*
502 available at the Codon Usage Database (Nakamura, et al. 2000) were used as reference
503 to calculate the 'codon adaptation index' with CAIcal server (Puigbo, et al. 2008). For
504 each species, hidden orthologs were compared with three sets of transcripts generated
505 by randomly choosing the same number of sequences than the number of hidden

506 orthologs from the complete set of CDS sequences. All values were plotted in R using
507 the ggplot2 package.
508
509 ***GO and InterPro analyses***
510 GO analyses were performed with the human ortholog sequences from HumRef2015,
511 using the free version of Blast2GO v3. Charts were done with a cutoff value of 30 GO
512 nodes for the analyses of all hidden orthologs, and 10 GO nodes for the analyses of *S.*
513 *mediterranea* hidden orthologs. Resulting charts were edited in Illustrator CS6 (Adobe).
514 GO enrichment analysis of *S. mediterranea* hidden orthologs was performed with
515 Blast2GO v3 comparing the GO annotations of the hidden orthologs against the GO
516 annotations of the whole *S. mediterranea* transcriptome. InterProScan 5 was used to
517 analyze the domain architecture of the recovered hidden orthologs using the human
518 ortholog sequence.
519
520 ***Multiple sequence alignments and orthology assignment***
521 Full-length protein sequences of the human and *P. vittatus* SDCCAG8 gene were
522 aligned to the SDCCAG8 cryptic ortholog recovered for *S. mediterranea*. Alignment
523 was performed with MAFFT v.5 (Katoh and Standley 2013) using the G-INS-i option.
524 Resulting alignment was trimmed between positions 319 and 494 of the human protein
525 and edited with Illustrator CS6 (Adobe) to show the conserved residues between the
526 three species. Multiple sequence protein alignments were constructed with MAFFT v.5
527 and spuriously aligned regions were removed with gblocks 3 (Talavera and Castresana
528 2007). Alignments are available upon request. Orthology assignments were performed
529 with RAxML v8.2.6 (Stamatakis 2014) with the autoMRE option. The models of
530 protein evolution (CEP192: RtRev+I+G+F; CCCAP: JTT+G+F; Homeodomains:

531 LG+G) were calculated with ProtTest (Abascal, et al. 2005). Resulting trees were edited
532 with FigTree and Illustrator CS6 (Adobe).
533
534 **Competing interests**
535 The authors declare that they have no competing interests.
536
537 **Author's contributions**
538 JMMD and JFR designed the study. JMMD, AH, and KP collected material for the
539 transcriptomes of *M. lignano, P. vittatus*, and *L. squammata*. JFR wrote the code of
540 'Leapfrog'. JMMD, JFR and BCV performed the analyses. JFR, JMMD and AH wrote
541 the manuscript. All authors read and approved the final manuscript.
542
543 **Acknowledgements**
544 We thank the members of the Hejnol's lab for support and discussions, and in particular
545 Daniel Thiel and Annie Boddington for taking care of the *M. lignano* cultures. We
546 appreciate the advance access given to platyhelmith transcriptomes by Gonzalo Giribet
547 and Chris Laumer at the beginning of the project. This research was funded by the Sars
548 Centre core budget and the European Research Council Community's Framework
549 Program Horizon 2020 (2014–2020) ERC grant agreement 648861 to AH. JFR was
550 supported by startup funds from the University of Florida DSP Research Strategic
551 Initiatives #00114464 and the University of Florida Office of the Provost Programs,
552 JMMD was supported by Marie Curie IEF 329024 fellowship.
553
554 **References**

555 Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein
556 evolution. Bioinformatics 21:2104-2105.
557 Alba MM, Castresana J. 2007. On homology searches by protein Blast and the
558 characterization of the age of genes. BMC Evol Biol 7:53.
559 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment
560 search tool. J Mol Biol 215:403-410.
561 Altschul SF, Koonin EV. 1998. Iterated profile searches with PSI-BLAST--a tool for
562 discovery in protein databases. Trends Biochem Sci 23:444-447.
563 Aravind L, Koonin EV. 1999. Gleaning non-trivial structural, functional and
564 evolutionary information about proteins by iterative database searches. J Mol Biol
565 287:1023-1040.
566 Azimzadeh J, Wong ML, Downhour DM, Sanchez Alvarado A, Marshall WF. 2012.
567 Centrosome loss in the evolution of planarians. Science 335:461-463.
568 Benjamini Y, Speed TP. 2012. Summarizing and correcting the GC content bias in
569 high-throughput sequencing. Nucleic Acids Res 40:e72.
570 Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC, Mashiyama
571 ST, Al-Lazikani B, Andrade LF, Ashton PD, et al. 2009. The genome of the blood fluke
572 Schistosoma mansoni. Nature 460:352-358.
573 Bork P, Doerks T, Springer TA, Snel B. 1999. Domains in plexins: links to integrins
574 and transcription factors. Trends Biochem Sci 24:261-263.
575 Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, Amores A,
576 Desvignes T, Batzel P, Catchen J, et al. 2016. The spotted gar genome illuminates
577 vertebrate evolution and facilitates human-teleost comparisons. Nat Genet 48:427-437.

578 Brandl H, Moon H, Vila-Farre M, Liu SY, Henry I, Rink JC. 2016. PlanMine - a
579 mineable resource of planarian biology and biodiversity. Nucleic Acids Res 44:D764-
580 773.
581 Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander
582 ES. 2007. Distinguishing protein-coding and noncoding genes in the human genome.
583 Proc Natl Acad Sci U S A 104:19428-19433.
584 Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita
585 S, Aerts A, Arnold GJ, Basu MK, et al. 2011. The ecoresponsive genome of *Daphnia*
586 *pulex*. Science 331:555-561.
587 Dalquen DA, Dessimoz C. 2013. Bidirectional best hits miss many orthologs in
588 duplication-rich clades such as plants and animals. Genome Biol Evol 5:1800-1806.
589 De Robertis EM. 2008. Evo-devo: variations on ancestral themes. Cell 132:185-195.
590 Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-
591 short read data sets from high-throughput DNA sequencing. Nucleic Acids Res
592 36:e105.
593 Domazet-Loso T, Brajkovic J, Tautz D. 2007. A phylostratigraphy approach to uncover
594 the genomic history of major adaptations in metazoan lineages. Trends Genet 23:533-
595 539.
596 Durocher D, Jackson SP. 2002. The FHA domain. FEBS Lett 513:58-66.
597 Eddy SR. 1998. Profile hidden Markov models. Bioinformatics 14:755-763.
598 Edgecombe GD, Giribet G, Dunn CW, Hejnol A, Kristensen RM, Neves RC, Rouse
599 GW, Worsaae K, Sørensen MV. 2011. Higher-level metazoan relationships: recent
600 progress and remaining questions. Org Divers Evol 11:151-172.

601 Edvardsen RB, Seo HC, Jensen MF, Mialon A, Mikhaleva J, Bjordal M, Cartry J,
602 Reinhardt R, Weissenbach J, Wincker P, et al. 2005. Remodelling of the homeobox
603 gene complement in the tunicate *Oikopleura dioica*. Curr Biol 15:R12-13.
604 Egger B, Lapraz F, Tomiczek B, Muller S, Dessimoz C, Girstmair J, Skunca N,
605 Rawlinson KA, Cameron CB, Beli E, et al. 2015. A transcriptomic-phylogenomic
606 analysis of the evolutionary relationships of flatworms. Curr Biol 25:1347-1353.
607 Elhaik E, Sabath N, Graur D. 2006. The "inverse relationship between evolutionary rate
608 and age of mammalian genes" is an artifact of increased genetic distance with rate of
609 evolution and time of divergence. Mol Biol Evol 23:1-3.
610 Grabherr M, Haas BJ, Yassour M, Levin J, Thompson D, Amit I, Adiconis X, Fan L,
611 Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-
612 seq data without a reference genome. Nat Biotechnol 29:644-652.
613 Hron T, Pajer P, Pačes J, Bartunek P, Elleder D. 2015. Hidden genes in birds. Genome
614 Biol 16:164.
615 Iuchi S. 2001. Three classes of C2H2 zinc finger proteins. Cell Mol Life Sci 58:625-
616 635.
617 Kao D, Felix D, Aboobaker A. 2013. The planarian regeneration transcriptome reveals a
618 shared but temporally shifted regulatory program between opposing head and tail
619 scenarios. BMC Genomics 14:797.
620 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
621 improvements in performance and usability. Mol Biol Evol 30:772-780.
622 Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC. 2009. More than just
623 orphans: are taxonomically-restricted genes important in evolution? Trends Genet
624 25:404-413.

625 Kipreos ET, Pagano M. 2000. The F-box protein family. Genome Biol
626 1:REVIEWS3002.
627 Knowles DG, McLysaght A. 2009. Recent *de novo* origin of human protein-coding
628 genes. Genome Res 19:1752-1759.
629 Kortschak RD, Samuel G, Saint R, Miller DJ. 2003. EST analysis of the cnidarian
630 *Acropora millepora* reveals extensive gene loss and rapid sequence divergence in the
631 model invertebrates. Curr Biol 13:2190-2195.
632 Koziol U, Jarero F, Olson PD, Brehm K. 2016. Comparative analysis of Wnt expression
633 identifies a highly conserved developmental transition in flatworms. BMC Biol 14:10.
634 Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003. Gene loss, protein sequence
635 divergence, gene dispensability, expression level, and interactivity are correlated in
636 eukaryotic evolution. Genome Res 13:2229-2235.
637 Kuchibhatla DB, Sherman WA, Chung BY, Cook S, Schneider G, Eisenhaber B, Karlin
638 DG. 2014. Powerful sequence similarity search methods and in-depth manual analyses
639 can identify remote homologs in many apparently "orphan" viral proteins. J Virol
640 88:10-20.
641 Laumer CE, Bekkouche N, Kerbl A, Goetz F, Neves RC, Sorensen MV, Kristensen
642 RM, Hejnol A, Dunn CW, Giribet G, et al. 2015. Spiralian phylogeny informs the
643 evolution of microscopic lineages. Curr Biol 25:2000-2006.
644 Laumer CE, Hejnol A, Giribet G. 2015. Nuclear genomic signals of the
645 'microturbellarian' roots of platyhelminth evolutionary innovation. Elife 4.
646 Li CY, Zhang Y, Wang Z, Zhang Y, Cao C, Zhang PW, Lu SJ, Li XM, Yu Q, Zheng X,
647 et al. 2010. A human-specific *de novo* protein-coding gene associated with human brain
648 functions. PLoS Comput Biol 6:e1000734.

649 Long M, Betran E, Thornton K, Wang W. 2003. The origin of new genes: glimpses
650 from the young and old. Nat Rev Genet 4:865-875.
651 Loppin B, Lepetit D, Dorus S, Couble P, Karr TL. 2005. Origin and
652 neofunctionalization of a *Drosophila* paternal effect gene essential for zygote viability.
653 Curr Biol 15:87-93.
654 Martin-Duran JM, de Mendoza A, Sebe-Pedros A, Ruiz-Trillo I, Hejnol A. 2013. A
655 broad genomic survey reveals multiple origins and frequent losses in the evolution of
656 respiratory hemerythrins and hemocyanins. Genome Biol Evol 5:1435-1442.
657 Martin-Duran JM, Romero R. 2011. Evolutionary implications of morphogenesis and
658 molecular patterning of the blind gut in the planarian *Schmidtea polychroa*. Dev Biol
659 352:164-176.
660 Moyers BA, Zhang J. 2015. Phylostratigraphic bias creates spurious patterns of genome
661 evolution. Mol Biol Evol 32:258-267.
662 Nakamura Y, Gojobori T, Ikemura T. 2000. Codon usage tabulated from international
663 DNA sequence databases: status for the year 2000. Nucleic Acids Res 28:292.
664 Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.
665 Olson MV. 1999. When less is more: gene loss as an engine of evolutionary change.
666 Am J Hum Genet 64:18-23.
667 Olson PD, Zarowiecki M, Kiss F, Brehm K. 2012. Cestode genomics - progress and
668 prospects for advancing basic and applied aspects of flatworm biology. Parasite
669 Immunol 34:130-150.
670 Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. 1999. The use of gene
671 clusters to infer functional coupling. Proc Natl Acad Sci U S A 96:2896-2901.
672 Palmieri N, Kosiol C, Schlotterer C. 2014. The life cycle of *Drosophila* orphan genes.
673 Elife 3:e01311.

674 Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core
675 genes in eukaryotic genomes. Bioinformatics 23:1061-1067.
676 Parra G, Bradnam K, Ning Z, Keane T, Korf I. 2009. Assessing the gene space in draft
677 genomes. Nucleic Acids Res 37:289-297.
678 Puigbo P, Bravo IG, Garcia-Vallve S. 2008. CAIcal: a combined set of tools to assess
679 codon usage adaptation. Biol Direct 3:38.
680 Riddiford N, Olson PD. 2011. Wnt gene loss in flatworms. Dev Genes Evol 221:187-
681 197.
682 Rieger R, Gehlen M, Haszprunar G, Holmlund M, Legniti A, Salvenmoser W, Tyler S.
683 1988. Laboratory cultures of marine Macrostomida (Turbellaria). Progr Zool 36:523.
684 Robb SM, Gotting K, Ross E, Sanchez Alvarado A. 2015. SmedGD 2.0: The *Schmidtea*
685 *mediterranea* genome database. Genesis 53:535-546.
686 Sánchez Alvarado A. 2012. Q&A: What is regeneration, and why look to planarians for
687 answers? BMC Biol 10:88.
688 Scheffzek K, Welti S. 2012. Pleckstrin homology (PH) like domains - versatile modules
689 in protein-protein interaction platforms. FEBS Lett 586:2662-2673.
690 Smith-Unna RD, Boursnell C, Patro R, Hibberd JM, Kelly S. 2015. TransRate:
691 reference free quality assessment of de-novo transcriptome assemblies. BioRxiv
692 021626.
693 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-
694 analysis of large phylogenies. Bioinformatics 30:1312-1313.
695 Struck TH, Wey-Fabrizius AR, Golombek A, Hering L, Weigert A, Bleidorn C, Klebow
696 S, Iakovenko N, Hausdorf B, Petersen M, et al. 2014. Platyzoan paraphyly based on
697 phylogenomic data supports a noncoelomate ancestry of Spiralia. Mol Biol Evol
698 31:1833-1849.

699 Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent
700 and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564-
701 577.
702 Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families.
703 Science 278:631-637.
704 Tautz D, Domazet-Loso T. 2011. The evolutionary origin of orphan genes. Nat Rev
705 Genet 12:692-702.
706 Technau U, Rudd S, Maxwell P, Gordon PM, Saina M, Grasso LC, Hayward DC,
707 Sensen CW, Saint R, Holstein TW, et al. 2005. Maintenance of ancestral complexity
708 and non-metazoan genes in two basal cnidarians. Trends Genet 21:633-639.
709 Telford MJ, Herniou EA, Russell RB, Littlewood DT. 2000. Changes in mitochondrial
710 genetic codes as phylogenetic characters: two examples from the flatworms. Proc Natl
711 Acad Sci U S A 97:11359-11364.
712 Tsai IJ, Zarowiecki M, Holroyd N, Garciarrubio A, Sanchez-Flores A, Brooks KL,
713 Tracey A, Bobes RJ, Fragoso G, Sciutto E, et al. 2013. The genomes of four tapeworm
714 species reveal adaptations to parasitism. Nature 496:57-63.
715 Wang X, Chen W, Huang Y, Sun J, Men J, Liu H, Luo F, Guo L, Lv X, Deng C, et al.
716 2011. The draft genome of the carcinogenic human liver fluke Clonorchis sinensis.
717 Genome Biol 12:R107.
718 Warnefors M, Eyre-Walker A. 2011. The accumulation of gene regulation through time.
719 Genome Biol Evol 3:667-673.
720 Wasik K, Gurtowski J, Zhou X, Ramos OM, Delas MJ, Battistoni G, El Demerdash O,
721 Falciatori I, Vizoso DB, Smith AD, et al. 2015. Genome and transcriptome of the
722 regeneration-competent flatworm, *Macrostomum lignano*. Proc Natl Acad Sci U S A
723 112:12462-12467.

724 Wolf YI, Koonin EV. 2012. A tight link between orthologs and bidirectional best hits in
725 bacterial and archaeal genomes. Genome Biol Evol 4:1286-1294.
726 Wolfe K. 2004. Evolutionary genomics: yeasts accelerate beyond BLAST. Curr Biol
727 14:R392-394.
728 Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. Nat
729 Rev Genet 13:329-342.
730

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

735 **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

736

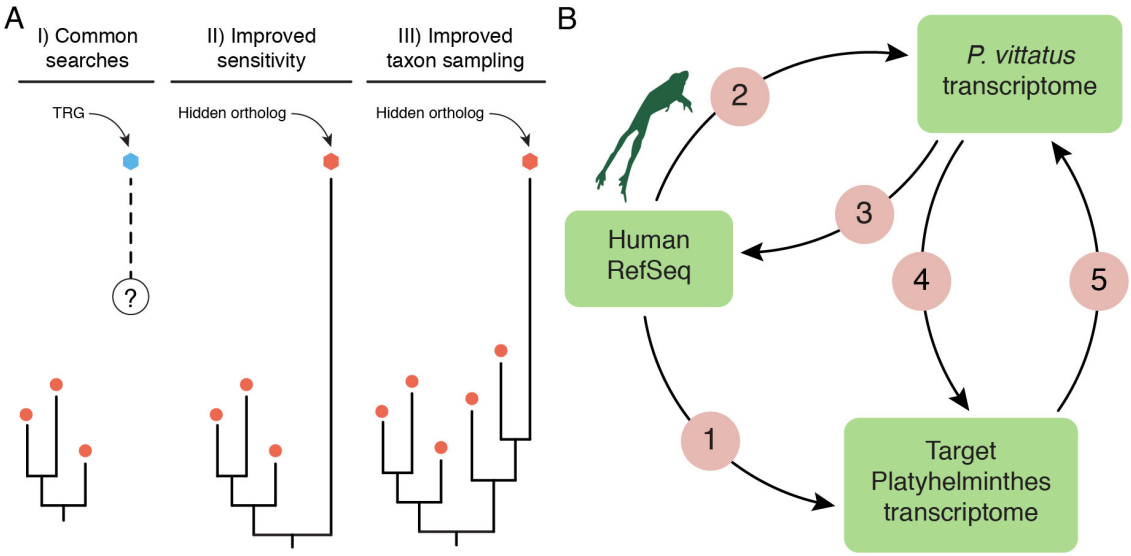
737 **Table 3. Presence/absence of hidden homeodomain genes in flatworms**

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	–	–	–

738 [1]gene present in the sister species *S. polychroa* (Martin-Duran and Romero 2011)
739 [2]Orthology based on BBH, not well supported by phylogenetic relationships.

740 **Figures**

741



742

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

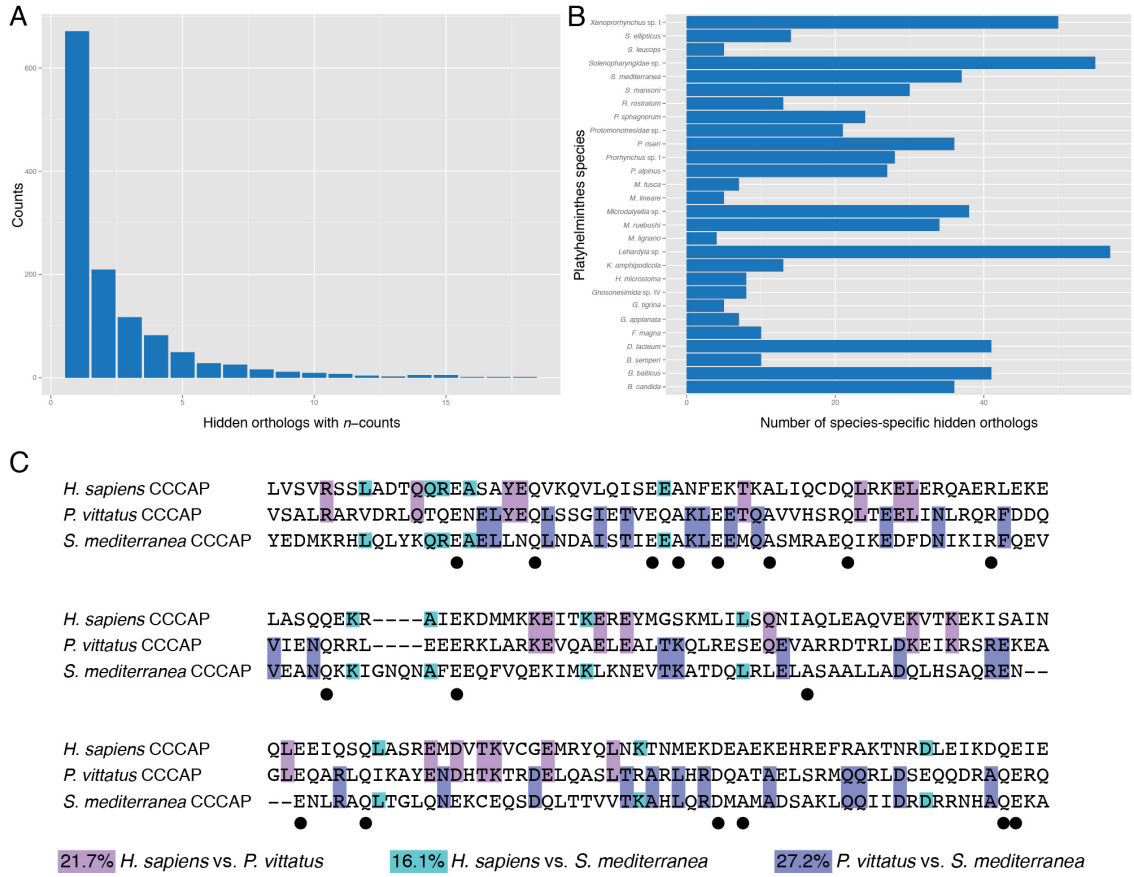


Figure 2. The Leapfrog pipeline recovers hundreds of hidden orthologs in Platyhelminthes. (**A**) Distribution of hidden orthologs according to their identification in one or more of the analyzed transcriptomes. Most of the hidden orthologs are unique of each lineage. (**B**) Distribution of species-specific 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.

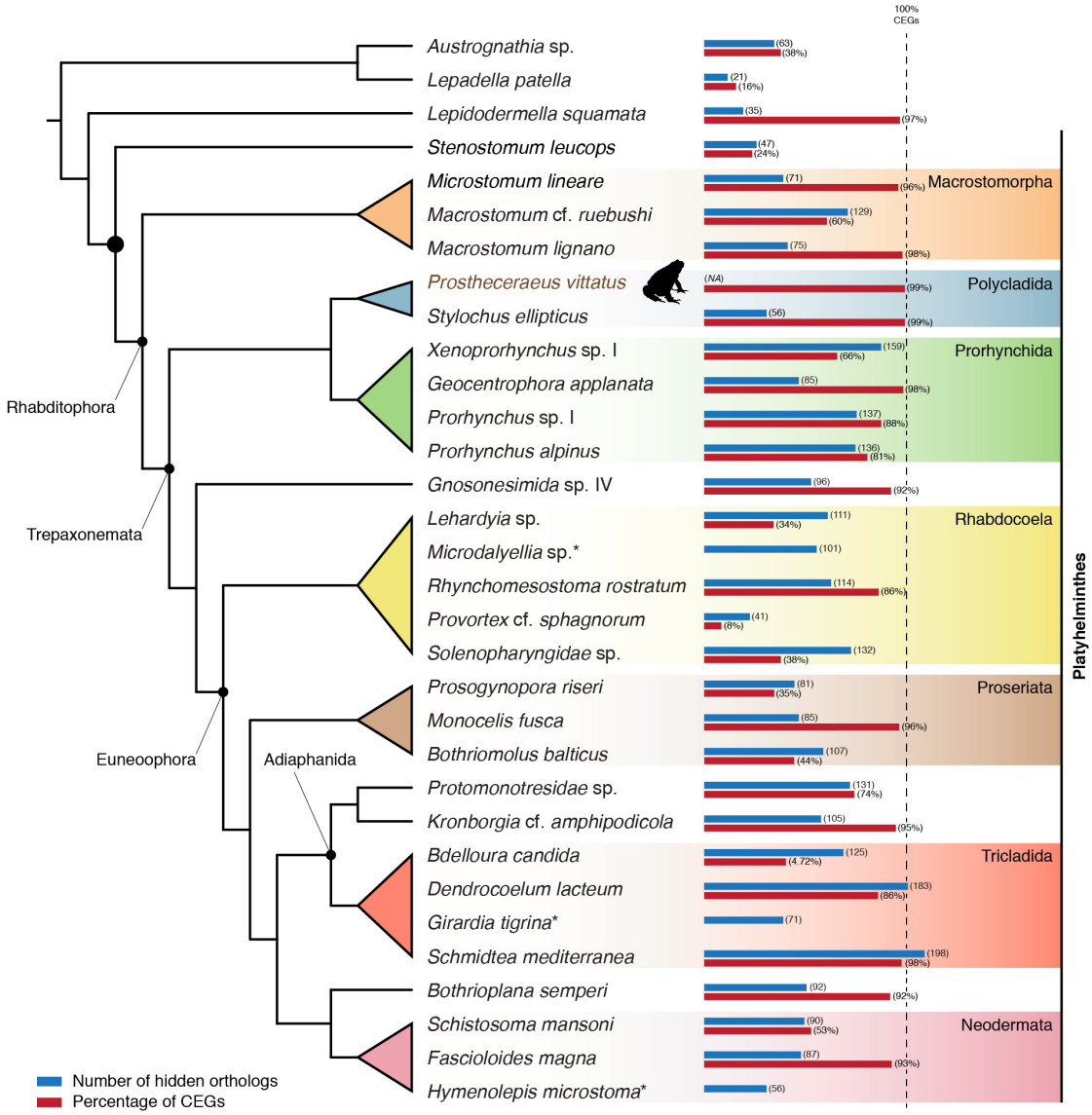


Figure 3. Distribution of hidden orthologs in the analyzed flatworm transcriptomes. The figure shows the total number of hidden orthologs in the analyzed transcriptomes in a phylogenetic context and with respect to their completeness (percentage of recovered core eukaryote genes, CEGs). The quality of 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.

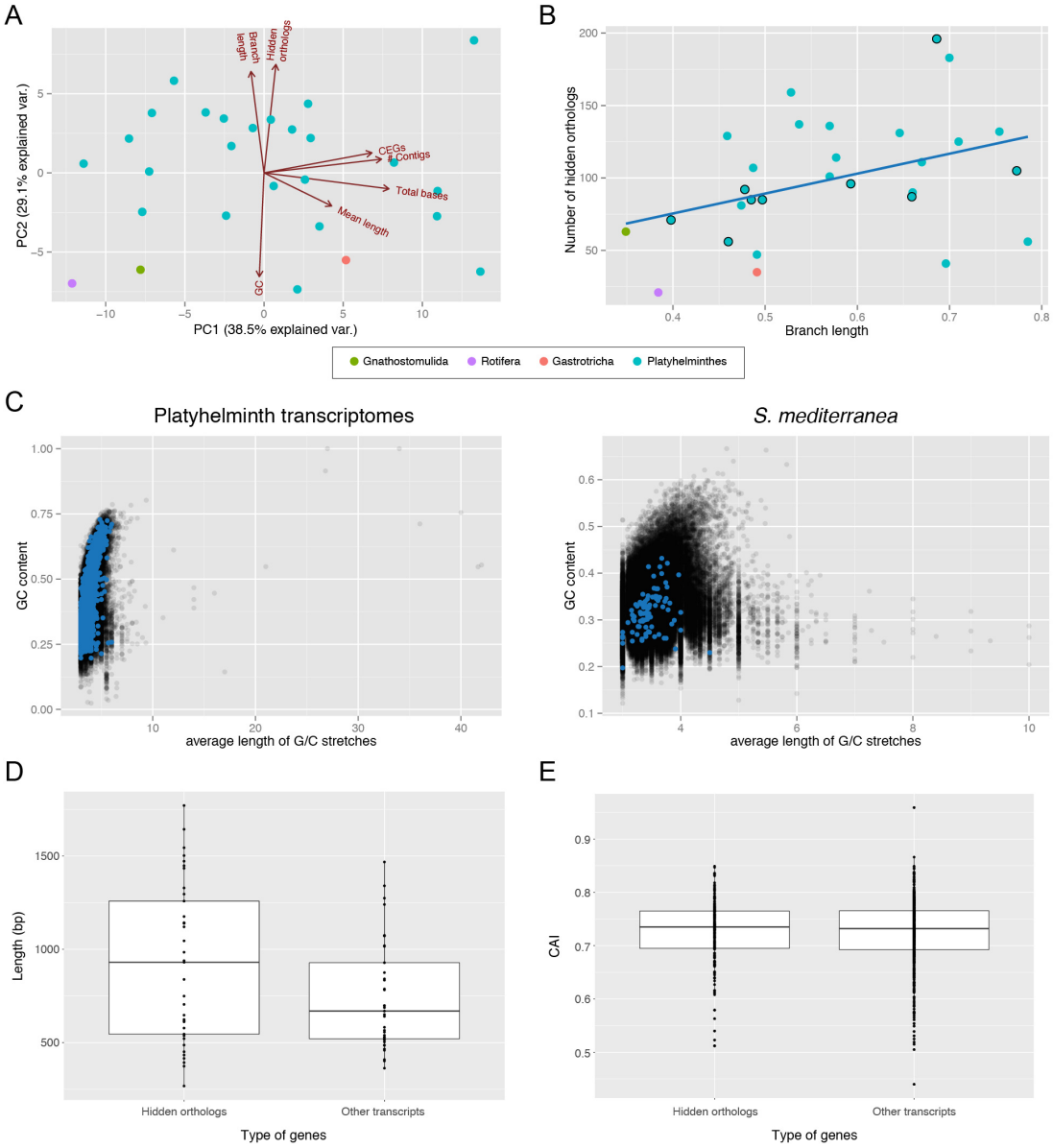


Figure 4. Hidden orthologs, evolutionary rates and sequence composition analyses. (**A**) Principal component analysis of the analyzed data showing the eigenvectors for each variable. The two first principal components (PC1, PC2) explain together 67.6% of the observed variability. (**B**) Number of hidden 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

783 *S. mediterranea* (right). The transcripts corresponding to hidden orthologs are in blue.
784 Hidden orthologs do not differentiate from the majority of transcripts. (**D**) Average
785 length of hidden orthologs compared to the average length of the other genes of the
786 transcriptome. Hidden orthologs are not significantly longer than the rest (Mann-
787 Whitney test; $p<0.05$). (**E**) Codon Adaptation Index (CAI) of the hidden orthologs of
788 the planarian species *B. candida*, *D. tigrina* and *S. mediterranea* compared with non-
789 hidden orthologs. CAI index in hidden orthologs does not significantly differ from the
790 rest of transcripts (Mann-Whitney test; $p<0.05$).

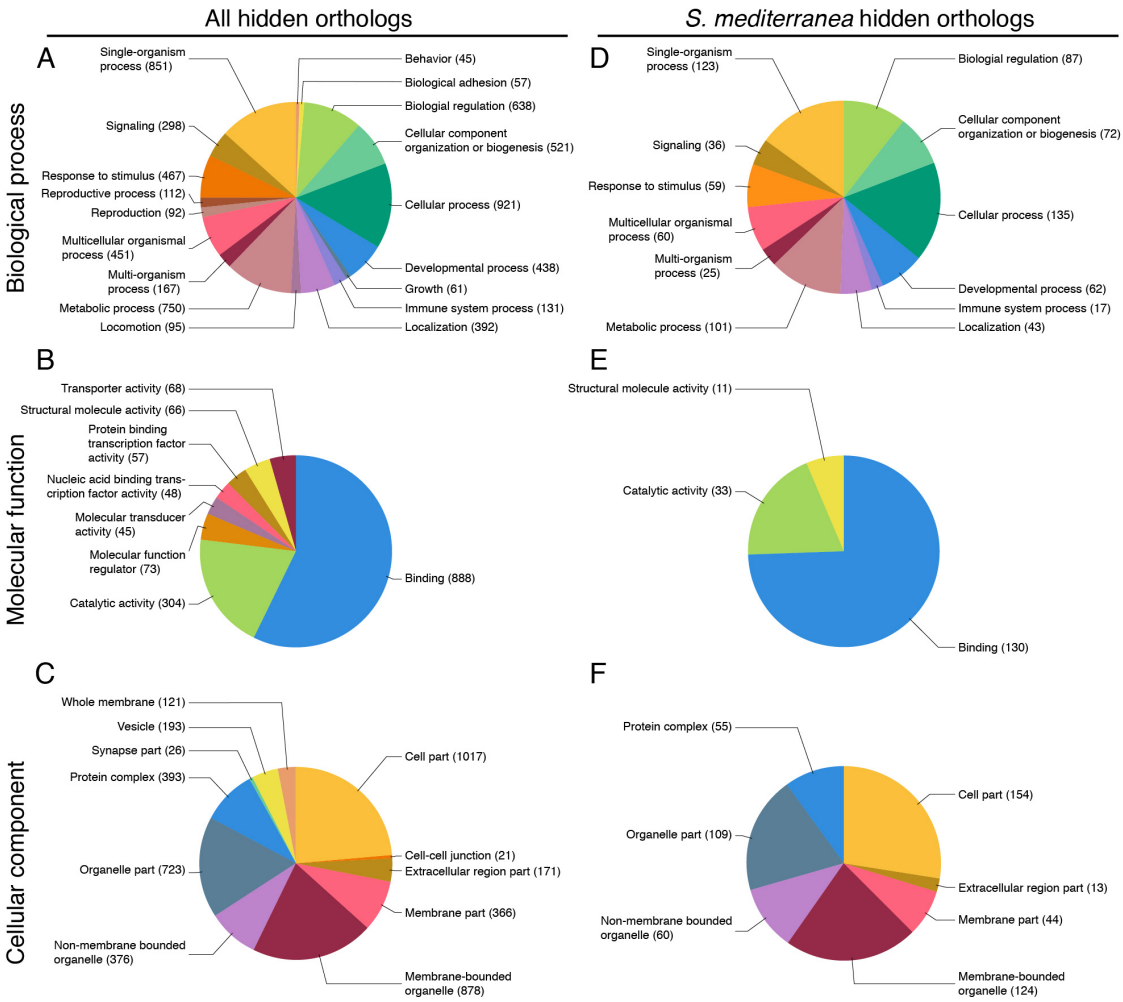


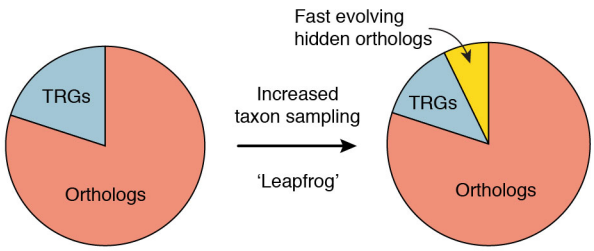
Figure 5. Gene Ontology (GO) characterization of hidden orthologs. Distribution of GO terms for all recovered hidden orthologs (**A**–**C**) and for the hidden orthologs identified in *S. mediterranea* (**D**–**F**). Hidden orthologs include a great diversity of GO categories, with a big proportion of 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

803 code) is indicated by asterisks. These proteins were manually identified with the *G.*

804 *tigrina* CEP192 as ‘bridge’ by reciprocal best BLAST hit. The five proteins essential for

805 centrosomal replication are squared in red.



806

807

808 **Figure 7. Increased taxon sampling uncovers fast-evolving hidden orthologs.**

809 Taxonomically restricted genes (TRGs) usually comprise 10-20% of the gene repertoire

810 (left). Increasing taxon sampling in the group of study and the use of a 'slow evolving'

811 intermediate species (i.e. 'Leapfrog' strategy) helps identify part of the TRGs of a given

812 lineage as fast-evolving hidden orthologs, thus diminishing both the number of TRGs

813 and inferred gene losses. The proportion of TRGs, common orthologs and hidden

814 orthologs are not to scale.

815 **Supplementary Material**

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

820

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

824

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

830

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

834

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

838

839 **Supplementary Table 1. Transcriptomes analyzed in this study.**

840

841 **Supplementary Table 2. Recovered hidden orthologs.** Hidden orthologs (as in human

842 RefSeq) recovered in each transcriptome after running 'Leapfrog' with the

843 transcriptome of the polyclad *P. vittatus* used as the 'bridge'.

844

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

846

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

848 **transcriptomes.**

849

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

851

852 **Supplementary Table 6. Significantly enriched GO terms in the hidden orthologs**

853 **recovered in *S. mediterranea*.**