Pheromone Evolution, Reproductive Genes, and Comparative Transcriptomics in Mediterranean Earthworms (Annelida, Oligochaeta, Hormogastridae)

Marta Novo,*1,2 Ana Riesgo,1,2 Antoni Fernández-Guerra,2 and Gonzalo Giribet1
1Museum of Comparative Zoology, Department of Organismic and Evolutionary Biology, Harvard University
2Centro de Estudios Avanzados de Blanes, CSIC, Girona, Spain

*Corresponding author: E-mail: mnrodrig@fas.harvard.edu.
Associate editor: Barbara Holland

Abstract

Animals inhabiting cryptic environments are often subjected to morphological stasis due to the lack of obvious agents driving selection, and hence chemical cues may be important drivers of sexual selection and individual recognition. Here, we provide a comparative analysis of de novo-assembled transcriptomes in two Mediterranean earthworm species with the objective to detect pheromone proteins and other reproductive genes that could be involved in cryptic speciation processes, as recently characterized in other earthworm species. cDNA libraries of unspecific tissue of Hormogaster samnithica and three different tissues of H. elsa were sequenced in an Illumina Genome Analyzer II or Hi-Seq. Two pheromones, Attractin and Temptin were detected in all tissue samples and both species. Attractin resulted in a reliable marker for phylogenetic inference. Temptin contained multiple paralogs and was slightly overexpressed in the digestive tissue, suggesting that these pheromones could be released with the casts. Genes involved in sexual determination and fertilization were highly expressed in reproductive tissue. This is thus the first detailed analysis of the molecular machinery of sexual reproduction in earthworms.

Key words: Annelida, pheromones, reproductive genes, transcriptomes, earthworm.

Introduction

Cryptic milieu, such as soil environments, may drive chemical signaling to play a more important role than morphology in sexual selection (Lee and Frost 2002). For example, earthworms, one of the most paradigmatic soil inhabitants, show morphological stasis with high levels of cryptic speciation (e.g., King et al. 2008; Novo et al. 2009, 2010; James et al. 2010; Buckley et al. 2011) and homoplasy (e.g., Novo, Fernández, et al. 2012).

Chemical signals are an ancient form of communication, being present in a great variety of taxa, including insects (Roelofs et al. 2002; Saudan et al. 2002), molluscs (Susswein and Nagle 2004), annelids (Zeeck et al. 1998; Ram et al. 1999), fish (Sorensen 2004), amphibians and reptiles (Houck 2009), mammals (Brennan and Keverne 2004), and even protozoans (Luporini et al. 2005) or yeasts (Kodama et al. 2003). Pheromones—molecules involved in animal communication by inducing a behavioral reaction or developmental process among individuals of the same species (Cardé and Millar 2009)—are semiochemicals (chemicals involved in communication) that can either be detected by “sniffing” air or water, or by contact chemoreception (Wyatt 2003). Attractin was the first water-borne peptide sex pheromone ever characterized in invertebrates, and it was described in two species of Aplysia (Mollusca, Gastropoda; Painter et al. 1998). Some other sex pheromones were subsequently described, including Enticin, Temptin, and Seductin (Cummins et al. 2004, 2006). No sex pheromone has yet been described or characterized in earthworms, although it has been suggested that they can leave trails containing pheromones (Rosenkoetter and Boice 1975) and that they present chemoreceptors (Lavarrack 1960). Alarm pheromones have been detected in earthworms (Ressler et al. 1968), which can deter other members of the species but can act as a chemoattractant to other animals such as snakes (Jiang et al. 1990). Also, one hormone, Annetocin, which induces the egg-laying behavior in Eisenia fetida, has been described (Oumi et al. 1996). In annelids, Temptin has been described in Pomatoceros lamarkii (Takahashi et al. 2009), and the sperm-release pheromone cysteine-glutathione disulfide (“Nereithione”) has been described in Nereis succinea (Zeeck et al. 1998; Ram et al. 1999).

Not only sex pheromones play an essential role in the sexual reproduction of animals, but also many other proteins are required to generate gametes, and among them, germ line determination proteins are crucial to maintaining the totipotency of the gametes (Extablour 2007). In annelids, the proteins encoded by the genes vasa, PI.10, piivi, and nanos have been found to play a role in the embryonic determination of the germ line (Rebscher et al. 2007; Dill and Seaver 2008; Sugio et al. 2008; Giani et al. 2011). However, the remainder of the germ line machinery is poorly known in these animals. For sex determination, annelids appear to use double-sex and mab-3-related proteins (Sukuzi et al. 2005), as many other metazoans do (Volff et al. 2003). After the formation of gametess and
mating, proteins such as Fertilin and Acrosin play an essential role in fertilization (Vacquier 1998; Howes and Jones 2002).

Next-generation sequencing platforms (e.g., Illumina) have made genomic and transcriptomic data progressively more affordable for research groups working on nonmodel organisms. Illumina RNA-seq is becoming popular for de novo assembly of animal transcriptomes (Reich et al. 2010; Feldmeyer et al. 2011; Siebert et al. 2011; Smith et al. 2011; Hartmann et al. 2012; Protasio et al. 2012), and recently, Riesz et al. (2012) provided comparative characterization of transcriptomic data across multiple species throughout the animal phyla, including an earthworm.

Despite being key organisms for the correct functioning of soil systems, earthworms, which captured Darwin’s attention (Darwin 1881) and have been the target of applied research for some time (e.g., Lavelle and Spain 2001; Edwards 2004), have been featured in few articles focusing on transcriptome profiling. These articles have targeted lumbricids (Lee et al. 2005; Pirroznia et al. 2007; Owen et al. 2008; Gong et al. 2010) and megascoleids (Cho et al. 2009), and just one of these has used next-generation sequencing technologies (Gong et al. 2010). No earthworm genome has yet been released, although Lumbricus rubellus is in preparation (www.earthworms.org). Therefore, genetic resources for hormogastrids provide a useful complement to the already studied species because they differ in their life strategy, morphology, and phylogenetic position. Moreover, none of these studies focuses on genes involved in reproduction, centering instead on exposure to contaminants (Pirroznia et al. 2007; Owen et al. 2008), regeneration (Cho et al. 2009), midgut expression profiles (Lee et al. 2005), and oligo arrays design (Gong et al. 2010). Only Owen et al. (2008) prepared cDNA libraries including reproductive tissue, but they were not able to identify a substantial sample of transcripts associated with the biological process of sexual reproduction. Therefore, this is the first time that the molecular machinery of sexual reproduction in earthworms is detailed.

Hormogastrid earthworms, endogeic and endemic to the Mediterranean region (Cobolli-Sbordoni et al. 1992) are abundant in sandy dry soils, potentially unsuitable for most other earthworm species (Hernández et al. 2007) and have shown interesting biogeographical and evolutionary patterns, such as cryptic speciation or differences in substitution rates among sister clades (Novo et al. 2009, 2010, 2011; Novo, Almodóvar, et al. 2012). Identification of the sex pheromones postulated to drive cryptic speciation would be very useful given the morphological stasis found in this group.

The main objectives of this study are thus: 1) to characterize and compare the transcriptomes of two closely related hormogastrid species (Hormogaster elisae, Alvarez [1977] and H. samnitica, Cognetti [1914]); 2) to identify the genes involved in reproductive and recognition processes, such as germ line determination, mating, and fertilization, with special emphasis on the sex pheromones; 3) to understand the evolution of these proteins across metazoans and hormogastrids and to test their suitability for phylogenetics; and 4) to compare the presence and levels of expression of these genes in different tissues of an individual, specifically on reproductive versus nonreproductive tissue.

Results

Sequence Assembly

Statistics for the assemblies are presented in tables 1–3. In tables 1 and 2, the data after different thinning values are shown for the CLC Genomics Workbench 4.6.1 (CLC) assemblies. Velvet/Oases (V/O) assemblies were subsequently done with the best read set (following the criteria of Riesz et al. 2012). We selected 0.05-thinned reads for the H. samnitica sample and 0.005-thinned reads for all the tissue samples from H. elisae. Comparison of the different final used assemblies is presented in table 3, and the number of contigs and their length are shown in figure 1. V/O provided longer contigs than CLC in both species, and in both assemblies there was a high proportion of contigs shorter than 500 bp.

Basic Local Alignment Search Tool and Functional Annotation

De novo assembled transcriptomes from the three earthworm species (H. samnitica, H. elisae, and L. rubellus) were BLASTed against a metazoan nonredundant (nr) database, and the results showed that a minimum of 17.23% and a maximum of 8.37% of the contigs recovered Basic Local Alignment Search Tool (BLAST) hits and from 1.15% to 12.03% contigs were annotated (fig. 2). In all cases, the contigs assembled with V/O presented more BLAST hits than those assembled with CLC (fig. 2). Interestingly, although the more stringent e value (1e-10) yielded a lower proportion of BLAST hits, as expected, the annotation was similar in the case of V/O contigs, whereas it was also lower in the case of CLC assemblies. BLAST results with specified contig size are shown in supplementary fig. S1, Supplementary Material online. Although contigs shorter than 500 bp were very abundant in the assemblies (fig. 1), they did not recover a high proportion of BLAST hits (supplementary fig. S1, Supplementary Material online).

BLAST analyses performed with CLC assemblies showed more gene uniqueness (unique hits) (fig. 3). The BLAST analysis of the assemblies generated by CLC against those generated by V/O show that CLC assemblies normally contained a much larger number of contigs, but the overlap between contigs from both programs is not very high, indicating that both assembly strategies recover many private contigs, as illustrated in figure 4. The overlap of the three earthworm species is also represented (fig. 4). CLC assemblies show a higher overlap among the species, not only between hormogastrids but also with L. rubellus, whose contigs were retrieved from another study (www.earthworms.org). Comparison among tissues also shows a high number of private contigs that may represent tissue-specific genes or variants (fig. 4).

Gene Ontology Terms

For H. elisae, we obtained gene ontology (GO) terms for 2,028 (for the higher stringency BLAST value, e value 1e-10) and 8,042 contigs for the less stringent value (1e-5).
Table 1. Assembly Parameters for All Trials in CLC Genomics Workbench that Helped to Decide the Thinning Limit to be Used in V/O.

<table>
<thead>
<tr>
<th>Species</th>
<th>Tissue</th>
<th>Sequencer</th>
<th>Size (bp)</th>
<th>Thin Limit</th>
<th>N Reads BT</th>
<th>N Reads AT</th>
<th>N Reads Trimmed</th>
<th>Avg. L AT</th>
<th>N Contigs</th>
<th>N Bases</th>
<th>Avg. L Contigs</th>
<th>SD</th>
<th>Max Contig L</th>
<th>N50 Contig Count</th>
<th>Avg. L of N50</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>H. samnitica</em></td>
<td>Posterior part</td>
<td>GAII 447</td>
<td>A</td>
<td>50,789,952</td>
<td>47,857,894</td>
<td>98.728</td>
<td>52.01</td>
<td>189,839</td>
<td>75,902,786</td>
<td>399.83</td>
<td>312.52</td>
<td>7,319</td>
<td>423</td>
<td>49,506</td>
<td>766.61</td>
<td>426.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B</td>
<td>50,789,952</td>
<td>45,094,358</td>
<td>93.77</td>
<td>83.77</td>
<td>169,050</td>
<td>68,625,148</td>
<td>407.13</td>
<td>320.15</td>
<td>5,661</td>
<td>436</td>
<td>43,377</td>
<td>793.34</td>
<td>432.37</td>
</tr>
<tr>
<td><em>H. elisae</em></td>
<td>REP Hi-seq</td>
<td>331</td>
<td>A</td>
<td>107,887,540</td>
<td>105,784,930</td>
<td>98.59</td>
<td>100.58</td>
<td>301,292</td>
<td>134,789,005</td>
<td>450.35</td>
<td>475.86</td>
<td>17,619</td>
<td>497</td>
<td>64,973</td>
<td>1,042.46</td>
<td>757.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B</td>
<td>107,887,540</td>
<td>103,123,804</td>
<td>98.41</td>
<td>100.44</td>
<td>299,301</td>
<td>134,789,005</td>
<td>450.35</td>
<td>475.86</td>
<td>17,619</td>
<td>497</td>
<td>64,973</td>
<td>1,042.46</td>
<td>757.65</td>
</tr>
<tr>
<td></td>
<td>DIG Hi-seq</td>
<td>307</td>
<td>A</td>
<td>36,873,140</td>
<td>35,998,970</td>
<td>97.09</td>
<td>100.74</td>
<td>120,501</td>
<td>63,002,546</td>
<td>522.84</td>
<td>510.76</td>
<td>13,265</td>
<td>650</td>
<td>25,233</td>
<td>1,248.43</td>
<td>725.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B</td>
<td>36,873,140</td>
<td>35,196,193</td>
<td>96.32</td>
<td>100.67</td>
<td>119,745</td>
<td>62,689,709</td>
<td>523.57</td>
<td>511.55</td>
<td>13,265</td>
<td>651</td>
<td>25,052</td>
<td>1,251.29</td>
<td>725.96</td>
</tr>
<tr>
<td></td>
<td>REST Hi-seq</td>
<td>268</td>
<td>A</td>
<td>56,019,876</td>
<td>54,758,116</td>
<td>97.23</td>
<td>100.77</td>
<td>113,820</td>
<td>51,708,606</td>
<td>455.51</td>
<td>393.59</td>
<td>14,792</td>
<td>530</td>
<td>27,371</td>
<td>944.63</td>
<td>551.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B</td>
<td>56,019,876</td>
<td>53,713,426</td>
<td>97.11</td>
<td>100.77</td>
<td>113,820</td>
<td>51,708,606</td>
<td>455.51</td>
<td>393.59</td>
<td>14,792</td>
<td>530</td>
<td>27,371</td>
<td>944.63</td>
<td>551.22</td>
</tr>
<tr>
<td></td>
<td>All together</td>
<td>a</td>
<td>a</td>
<td>192,033,423</td>
<td>189,839</td>
<td>97.23</td>
<td>100.77</td>
<td>351,596</td>
<td>157,992,055</td>
<td>449.36</td>
<td>475.92</td>
<td>17,620</td>
<td>499</td>
<td>75,491</td>
<td>1,046.43</td>
<td>760.69</td>
</tr>
</tbody>
</table>

Note.—N bases represent the total contig length. bp, base pairs; N number; BT, before thinning; AT, after thinning; Avg., average; L, length; SD, standard deviation; REP, reproductive tissue; DIG, digestive tissue; REST, rest of the tissues not included in REP and DIG. Thinning was performed using 0.05 (A) and 0.005 (B) as the limit in CLC. Preferred assemblies and thin limits are shaded.

Table 2. Distribution of Contig Lengths after the Different Assemblies Performed in CLC Genomics Workbench That Helped to Select the Thin Limit Used.

<table>
<thead>
<tr>
<th>Species</th>
<th>Contig size (bp)</th>
<th>Homo*gaster samnitica</th>
<th>Homo*gaster elisae</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>A Percentage</td>
<td>B Percentage</td>
</tr>
<tr>
<td></td>
<td>&lt;300</td>
<td>98.728</td>
<td>52.01</td>
</tr>
<tr>
<td></td>
<td>100,000</td>
<td>50,789,952</td>
<td>47,857,894</td>
</tr>
<tr>
<td></td>
<td>1,000,000</td>
<td>107,887,540</td>
<td>105,784,930</td>
</tr>
<tr>
<td></td>
<td>20,000,000</td>
<td>36,873,140</td>
<td>35,998,970</td>
</tr>
<tr>
<td></td>
<td>40,000,000</td>
<td>56,019,876</td>
<td>54,758,116</td>
</tr>
<tr>
<td></td>
<td>60,000,000</td>
<td>59.728</td>
<td>52.01</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>189,839</td>
<td>189,839</td>
</tr>
</tbody>
</table>

Note.—Thinning was performed using 0.05 (A) and 0.005 (B) as the limit in CLC. Preferred assemblies and thin limits are shaded. REP, reproductive tissue; DIG, digestive tissue; REST, rest of the tissues not included in REP and DIG.
In *H. samnitica*, between 4,246 and 7,420 contigs had GO terms assigned (both values 1e-10 and 1e-5, respectively), whereas in *L. rubellus*, 2,570 and 7,676 contigs had GO terms assigned (also both values 1e-10 and 1e-5, respectively). No functional category of genes was lacking in any of the analyzed transcriptomes, as shown by the GO assignment, and there was no clear bias toward any particular category of terms (fig. 5).

The percentages of sequences mapped to given GO terms were highly similar for the three species and between CLC and V/O, although those differed slightly when treated as totals (supplementary table S1, Supplementary Material online), where V/O assemblies showed higher values and so did *H. elisae*. This species showed a higher number of GOs, particularly concerning molecular function. This may be simply related to the greater amount of initial contigs available from Metazoa

<table>
<thead>
<tr>
<th>Species</th>
<th>Assembler</th>
<th>N Contigs</th>
<th>Avg. L Contigs</th>
<th>Max Contig L</th>
<th>N50</th>
<th>N50 Contig Count</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>H. samnitica</em></td>
<td>CLC</td>
<td>189,839</td>
<td>399.83</td>
<td>7,319</td>
<td>423</td>
<td>49,506</td>
</tr>
<tr>
<td></td>
<td>V/O</td>
<td>59,853</td>
<td>677.10</td>
<td>9,307</td>
<td>1,156</td>
<td>56,674</td>
</tr>
<tr>
<td><em>H. elisae</em></td>
<td>CLC</td>
<td>351,596</td>
<td>449.36</td>
<td>17,620</td>
<td>499</td>
<td>75,491</td>
</tr>
<tr>
<td></td>
<td>V/O</td>
<td>202,194</td>
<td>1,234.30</td>
<td>16,498</td>
<td>2,414</td>
<td>187,360</td>
</tr>
</tbody>
</table>

Table 3. Characteristics of the Final Assemblies After the Selected Thinning Threshold.

Note.—Individual tissues were only assembled with CLC (see tables 1 and 2). N, number; Avg, average; L, length; V/O, Velvet/Oases.
When analyzing the GO terms of each tissue sample in *H. elisae*, we detected only significant differences in the GO complements for molecular function between the reproductive tissue (REP) and the normal tissue (REST) in the terms related to ion channel activity (fig. 6).

**Pheromones and Other Genes Involved in Reproduction**

From the four sex pheromone genes searched in the earthworms transcriptomes (*attractin*, *temptin*, *enticin*, and *seducitin*), only two were detected: *attractin* and *temptin*. Three paralogs were found for *temptin* in hormogastrids, all showing one or two domains of Copper type II ascorbate-dependent monooxygenase (fig. 7). The fact that three paralogs were found for this gene made it inappropriate for phylogenetic analyses and species differentiation (fig. 7). In addition, two isoforms were detected in the digestive tissue (DIG) for *temptins* 1 and 2. Two more *temptin*-like genes were found in *H. samnitica*, whereas only one was detected in *H. elisae*, all of them similar to dopamine beta-hydroxylase sequences (table 4). In turn, *attractin* showed phylogenetic signal (fig. 8), recovering key animal clades such as Bilateria, its main division in Deuterostomia and Protostomia, and a split of the latter into Ecdysozoa and Spiralia (see Edgecombe et al. 2011). It also finds monophyly of Annelida as well as that of Clitellata and Hormogastridae. In addition to the strong phylogenetic signal for deep metazoan relationships, within earthworms this protein could be useful to differentiate closely related species, as shown by the 11% uncorrected *p*-distance in the DNA sequence of the *attractin* pheromone gene between *H. elisae* and *H. samnitica*.

The evolution of amino acid sequence of Attractin in metazoans shows clear domain reorganization (fig. 8). Pheromones were found from protozoans to vertebrates, including some groups where they have never been previously reported, like sponges and filastereans. Regarding the protein structure, the sponge *Petrosia ficiformis* and the solitary anemone *Nemastostella vectensis* maintained the exact same number of domains and position, but in the limpet *Lottia gigantea*, a Kelch domain appears duplicated and one PSI is lost. In annelids, the three upstream domains are reorganized, epidermal growth factor (EGF) located before CUB. A duplication of PSI is observed in arthropods. In vertebrates, a CLECT domain was added before the terminal two EGF domains and the transmembrane domain (fig. 8). In the three earthworms analyzed, the protein structure is maintained (fig. 8). This is also the case for all the known arthropods.
FIG. 5. Percentage of contigs mapped to given GO terms for the three earthworm species. CLC assemblies are used for hormogastrids, and data retrieved from www.earthworms.org (Elsworth B, personal communication) are used for Lumbricus rubellus.

FIG. 6. Differentially enriched GO terms in the reproductive tissue (REP) of Hormogaster elisae compared with the normal tissue (REST). (A) Hierarchical diagram of differentially enriched GO terms in the REP (shaded) and their relationships with related terms. (B) For the differentially enriched GO terms, the percentage of sequences assigned in both tissue types is represented.
In *H. elisae*, 11 genes were detected to be part of the germ line determination, four involved in sex differentiation, and eight in fertilization (table 4). For *H. samnitica*, 12 genes were found to belong to the germ line determination machinery and four to the sex differentiation process, whereas no genes involved in fertilization were found (table 4). In general, the sequences found in the *H. elisae* data set were longer than those in the *H. samnitica* data set (table 4). Interestingly, although two paralog genes were found for *nanos* in *H. samnitica*, only one copy was detected in *H. elisae* (table 4). The germ line marker *oskar* was not found in any of the data sets.

Expression Profiles

Two different heat maps were obtained, one showing the expression levels of the three tissues including all the contigs from *H. elisae* (supplementary fig. S2, Supplementary Material online) and one filtering those contigs longer than 1,000 bp (supplementary fig. S3, Supplementary Material online). Only the mapped reads were considered for this analysis. It appears that most genes that were highly expressed in the DIG were also overexpressed in the other two transcriptomes (supplementary fig. S2, Supplementary Material online). In turn, differential expression was detected for both REP and REST, with a higher number of genes uniquely expressed in the REP (supplementary fig. S2, Supplementary Material online). In the case of contigs longer than 1,000 bp, it was observed in the heat map that mostly REP genes were upregulated and the rest showed low expression rates (supplementary fig. S3, Supplementary Material online). The expression levels of these contigs ranged between 0 and 1,113.5 nRPKM (reads per kilobase of exon model per million mapped reads), and the maximum value was obtained in REST (supplementary fig. S4, Supplementary Material online). Most of the contigs are below 100 nRPKM, and therefore, those above this level in the REP were further analyzed. The list of genes highly expressed in the REP is presented in table 5.

Specific information on genes related to attraction (sex pheromones), sexual differentiation, and determination and fertilization is presented in table 4. These genes were searched for in hormogastrid species, and their expression levels were measured for different tissues of *H. elisae*. Although some of the genes involved in germ line determination (*PL10 1*, *PL10 2*, *piwi 1*, *tsunagi*, *Piwi 2*, *smaug*, *nanos*, and *mago nashi*) were only slightly more expressed in the REP than in DIG and REST (table 4), other genes, such as *vasa*, *germ cell-less*, and *piwi 1*, showed markedly higher values of nRPKM in the REP than in the other two transcriptomes (table 4). The gene *PL10 3* was not upregulated in the REP but was upregulated in the rest (table 4). All genes involved in sex determination, except for *sperm-associated antigen 7* (SAA-7), were upregulated in the REP and REST (table 4). As for the fertilization genes, only *acrosin 2* was upregulated in the REP, whereas the other genes seemed to have similar expression values (table 4). For the sex pheromone genes, two paralogs of the *temptin* (*temptin 1* and 2) and *temptin-like 1*, and *attractin* were slightly more expressed in the DIG. No *acrosin*, *fertilin*, or *annetocin-precursor* sequences were found in the transcriptome of *H. samnitica*, probably because of the lower coverage of this data set.

Discussion

This study accounts for the potential of transcriptomic data for multiple biological purposes and represents one of the very few studies of expressed sequence tags in earthworms (Lee et al. 2005; Pirooznia et al. 2007; Owen et al. 2008; Cho...
et al. 2009; Gong et al. 2010). It is also the first transcriptomic comparative study including multiple hormogastrid species. The only complete earthworm genome (L. rubellus; www.earthworms.org) is still unpublished, and therefore, we had no reference genome for assembly. It has been shown that even closely related species can present very different genomes, with low levels of conservation (Ewen-Campen et al. 2011) and indeed a high divergence even within lineages has been found in earthworms when looking at a few genes (e.g., Chang et al. 2009; Rougerie et al. 2009; Fernández et al. 2011; Novo et al. 2009, 2010; Novo, Almodóvar, et al. 2012).

### Sequence Assembly, BLAST, and Functional Annotation

The assembly results showed that V/O produces longer contigs than CLC, as when comparing V/O with other programs...

---

**Table 4. Selected Genes Involved in Reproduction and Sexual Differentiation Processes Identified in Hormogastrid Transcriptomes.**

<table>
<thead>
<tr>
<th>Genes</th>
<th><strong>Hormogaster elisae</strong></th>
<th><strong>H. samnitica</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Presence</td>
<td>Expression (nRPKM)</td>
</tr>
<tr>
<td>Pheromones</td>
<td></td>
<td>Digestive</td>
</tr>
<tr>
<td>Attractin</td>
<td>Yes</td>
<td>1,177</td>
</tr>
<tr>
<td>Temptin 1</td>
<td>Yes</td>
<td>296</td>
</tr>
<tr>
<td>Temptin 2</td>
<td>Yes</td>
<td>96</td>
</tr>
<tr>
<td>Temptin 3</td>
<td>Yes</td>
<td>120</td>
</tr>
<tr>
<td>Temptin-like 1</td>
<td>Yes</td>
<td>278</td>
</tr>
<tr>
<td>Temptin-like 2</td>
<td>No</td>
<td>—</td>
</tr>
<tr>
<td>Other hormones</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annetocin precursor</td>
<td>Yes</td>
<td>130</td>
</tr>
<tr>
<td>Annetocin receptor 1</td>
<td>Yes</td>
<td>140</td>
</tr>
<tr>
<td>Annetocin receptor 2</td>
<td>Yes</td>
<td>129</td>
</tr>
<tr>
<td>Annetocin receptor 3</td>
<td>Yes</td>
<td>132</td>
</tr>
<tr>
<td>Annetocin receptor 4</td>
<td>Yes</td>
<td>70</td>
</tr>
<tr>
<td>Sexual differentiation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DMRT3</td>
<td>Yes</td>
<td>340</td>
</tr>
<tr>
<td>SAA</td>
<td>Yes</td>
<td>198</td>
</tr>
<tr>
<td>SPATA2</td>
<td>Yes</td>
<td>384</td>
</tr>
<tr>
<td>SOX 3</td>
<td>Yes</td>
<td>296</td>
</tr>
<tr>
<td>Germ line determination</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vasa</td>
<td>Yes</td>
<td>153</td>
</tr>
<tr>
<td>PL10 1</td>
<td>Yes</td>
<td>435</td>
</tr>
<tr>
<td>PL10 2</td>
<td>Yes</td>
<td>359</td>
</tr>
<tr>
<td>PL10 3</td>
<td>Yes</td>
<td>294</td>
</tr>
<tr>
<td>Germ cell-less</td>
<td>Yes</td>
<td>414</td>
</tr>
<tr>
<td>Tsunagi</td>
<td>Yes</td>
<td>178</td>
</tr>
<tr>
<td>Piwi 1</td>
<td>Yes</td>
<td>906</td>
</tr>
<tr>
<td>Piwi 2</td>
<td>Yes</td>
<td>745</td>
</tr>
<tr>
<td>Smaug</td>
<td>Yes</td>
<td>597</td>
</tr>
<tr>
<td>Nanos 1</td>
<td>Yes</td>
<td>145</td>
</tr>
<tr>
<td>Nanos 2</td>
<td>No</td>
<td>—</td>
</tr>
<tr>
<td>Mago-nashi</td>
<td>Yes</td>
<td>149</td>
</tr>
<tr>
<td>Fertilization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fertilin 1</td>
<td>Yes</td>
<td>400</td>
</tr>
<tr>
<td>Fertilin 2</td>
<td>Yes</td>
<td>611</td>
</tr>
<tr>
<td>Fertilin 3</td>
<td>Yes</td>
<td>781</td>
</tr>
<tr>
<td>Fertilin 4</td>
<td>Yes</td>
<td>288</td>
</tr>
<tr>
<td>Fertilin 5</td>
<td>Yes</td>
<td>287</td>
</tr>
<tr>
<td>Acrosin 1</td>
<td>Yes</td>
<td>216</td>
</tr>
<tr>
<td>Acrosin 2</td>
<td>Yes</td>
<td>205</td>
</tr>
<tr>
<td>Acrosin 3</td>
<td>Yes</td>
<td>337</td>
</tr>
</tbody>
</table>

**Note.**—The presence/absence of the genes is indicated as well as the expression level in the tissues of *H. elisae*. nRPKM, reads per kilobase of exon model per million mapped reads (normalized values). Accession numbers for these sequences in GenBank database are included.
However, redundancy in V/O assemblies is higher, being the unique hits of CLC more abundant. It has been previously shown that different assemblers produce different contigs, which most of the times represent different genes, overlapping only in a small proportion as we detected in our Venn diagrams (e.g., Feldmeyer et al. 2011; Schulz et al. 2012) and therefore each providing different valuable information. Oases performs normally better with highly expressed transcripts, assuming that the ones that are lowly expressed represent sequencing errors, thus collapsing them with the highly expressed transcripts (Schulz et al. 2012). Therefore, this program could be eliminating paralogs or isoforms and could be the reason why the uniqueness is higher for the CLC assemblies. This makes V/O a good option for phylogenomic purposes, but CLC seems to fit better for gene hunting studies because it shows better all the variance of the expressed genome. One could think that sequencing errors could be producing the high number of variants found by CLC. However, Illumina platforms Hi-seq and GaII have shown to have a low error rate (Quail et al. 2012), which justifies the use of CLC assemblers and suggests that V/O is probably collapsing splicing variants into single contigs. Moreover, it is a good sign that CLC assemblies contain a much higher number of contigs shared with the earthworm L. rubellus (whose list of contigs was retrieved from www.earthworms.org), meaning that sequencing error does not seem probable. Most of the analyses performed subsequently with the different assemblies produced similar results, but CLC presented higher gene variability. GO annotation was similar after V/O and CLC assemblies and suggests that transcripts were broadly sampled.

We only found differences between the GO term complement of the category molecular function between the normal tissues (REST) and the reproductive tissues (REP) in nodes related to ion channel activity. Given that ionic fluxes play a key role in the activation of respiration and motility, and in chemotaxis of spermatozoa, the enriched ion channel complement in REP might be related to the occurrence of specific Ca$^{2+}$ channels of the sperm, as it occurs in the sperm of sea urchins (Darszon et al. 1994).

**Pheromones**

Our study provides the first report of sex pheromones in earthworms. For these animals, sex pheromones are probably important because they live in an environment where chemical signaling may play a crucial role in attracting a partner.

The gene sequences for the pheromones attractin and temptin, but not enticin or seductin, were identified in the hormogastrid transcriptomes. Cummins et al. (2006) suggested that binary blends of the pheromones Attractin with Enticin, Tempin, or Seductin stimulate mate attraction. Also, known insect pheromones are typically mixtures of multiple components (Kaissling 1996), but normally only two pheromone components are necessary to serve as an attractant (Christensen and Hildebrand 1994; Heinbockel et al. 2004). In the case of hormogastrids, it seems that the blend of Attractin with Tempin is what induces attraction. Their
union enhances the effectiveness of the mate attraction due to the role that Temptin plays in pheromone detection by organizing the interaction of Attractin with the cell surface receptor (Cummins et al. 2007). Particularly in Aplysia, it has been suggested that Temptin has a similar role in the pheromone complex to that of Fibrill in the extracellular matrix and mediates binding of Attractin to sensory cells in the chemosensory rhinophores (Cummins et al. 2007). Laverack (1960) suggested the presence of chemoreceptors in earthworms, being most sense organs that react to chemical stimuli located in the prostomium or buccal epithelium. The gene temptin seems to have a sequence homology to the EGF-like domains family, and therefore, it has been suggested that different isoforms could play similar roles of chemical communication in different tissues not related to pheromone function (Cummins et al. 2007). Indeed, we found different paralog genes for temptin in the analyzed transcriptomes, making this protein an unsuitable marker for species differentiation and phylogenetic inference. However, attractin is an informative phylogenetic marker ready to be used for species differentiation. The nucleotide sequence of attractin has an 11% divergence (uncorrected p distances) among studied conspecific hormogastrids. In Aplysia, the gene attractin is nearly identical for different species, making this pheromone a promiscuous signal, with different species of Aplysia found in the same egg-laying and mating aggregations (Cummins et al. 2006). However, in this case, it provides defense from predators, not required in the case of the endogeic Hormogaster. Experiments with hormogastrid earthworms would be necessary to unravel the exact working mechanism of this pheromone.

### Genes Involved in Reproduction

We have found the most complete machinery for germ line determination, sex differentiation, and fertilization reported to date in any annelid. Even though the genes vasa, Pl10, piwi, and nanos were found in polychaetes and clitellates, including hirudineans (Kang et al. 2002; Rebscher et al. 2007; Dill and Seaver 2008; Sugio et al. 2008; Giani et al. 2011), the rest of the germ line machinery (such as tsunagi, smaug, mago nashi, and germ cell-less) was unknown in these animals. Both tsunagi and mago nashi are involved in the germ line determination and oocyte differentiation of Drosophila (Parma et al. 2007).

---

### Table 5. List of Upregulated Genes within the REP of Hormogaster elisae, Number of the Contig That Contains the Sequence, and Accession Number of the Sequence That Matched in NCBI.

<table>
<thead>
<tr>
<th>Contig Number</th>
<th>Protein Name in NCBI</th>
<th>Species</th>
<th>E Value</th>
<th>Accession Number</th>
<th>Digestive nRPKM</th>
<th>Rest nRPKM</th>
<th>Reproductive nRPKM</th>
</tr>
</thead>
<tbody>
<tr>
<td>254294</td>
<td>Cytochrome P450 2J6</td>
<td>Crassostrea gigas</td>
<td>2.00E-65</td>
<td>EKC41577.1</td>
<td>0.1</td>
<td>0.3</td>
<td>169.4</td>
</tr>
<tr>
<td>241136</td>
<td>hypothetical protein Smp_086420</td>
<td>Schistosomes mansoni</td>
<td>8.60E+00</td>
<td>CCB40869.1</td>
<td>0.4</td>
<td>1.2</td>
<td>151.4</td>
</tr>
<tr>
<td>325591</td>
<td>EF-hand domain-containing protein 1</td>
<td>Crassostrea gigas</td>
<td>0.00E+00</td>
<td>EKC30173.1</td>
<td>0.4</td>
<td>1.2</td>
<td>146.4</td>
</tr>
<tr>
<td>255355</td>
<td>Ankyrin repeat and protein kinase domain-containing protein 1</td>
<td>Crassostrea gigas</td>
<td>8.00E-10</td>
<td>EKC18036.1</td>
<td>0.3</td>
<td>2.1</td>
<td>142.4</td>
</tr>
</tbody>
</table>

**Note:** The expression level is presented in nRPKM (Reads Per Kilobase of exon model per Million mapped reads, normalized values).
and interact with oskar (not found in our data sets) to establish the polarity of the embryo (Mohr et al. 2001). The gene germ cell-less is one of the genes acting early in germ line specification, and it is required to establish the transcription quiescence needed for germ cell determination in Drosophila (Leatherman et al. 2002) and might have a similar function in the mouse (Kimura et al. 1999). The roles of the complete germ line genetic machinery found in our study are still unknown in earthworms, but the genetic resources provided by this study will be a powerful tool to unravel the function of these genes during reproduction and help to unveil the evolution of the germ line determination in bilaterians.

For sex determination, earthworms might be using DM proteins, because we found the genes for double-sex and mab-3-related protein (DMRT3), as it is the case for other annelids (Suzuki et al. 2005), and many other metazoans do (Vollff et al. 2003). The protein SPATA2 is involved in the meiotic progress of male and female gametes in vertebrates, but it is absent in Drosophila and Caenorhabditis (Suzuki et al. 2011). The fact that we found the gene for SPATA2 in both earthworm species might indicate that the molecular interactions required early in meiotic prophase in both male and female germ cells in mice could also be present in oligochaetes. We have also found five paralog sequences in H. elisae of the gene fertilin, which has an important role during sperm–egg fusion in vertebrates (Vacquier 1998), and it has not been reported in any invertebrate other than molluscs (Cummins et al. 2006). Fertilins, which are sperm surface heterodimers, are thought to have evolved from pheromonal signaling mechanisms (Cummins et al. 2006). Moreover, three paralogs of acrosin were present in the transcriptome of H. elisae. Although Acrosin has long been considered as a zona lysin, some authors propose that it is a multifunctional protein that also plays a role in the secondary binding for retaining acrosome-reacted sperm on the zona surface (Howes and Jones 2002).

Additionally, the SOX3 gene, which is expressed in developing gonads and in the brain in humans, appears to be necessary for gonadal function (oocyte development, and male testis differentiation and gametogenesis) and not sex determination (Weiss et al. 2003). As a functional class, reproduction-specific genes evolve more rapidly than other functional gene classes, and as Grassa and Kulathinal (2011) found among vertebrates, there is a significantly higher protein divergence in gonadal genes (particularly in male-specific proteins, such as sperm development regulators) compared with nonreproductive genes. Therefore, these genes, involved in reproductive isolation, are prone to differentiate among cryptic species before other markers do. Grassa and Kulathinal (2011) conclude that sexual selection may be an important driver of evolutionary change and extends sexual selection theory to the level of molecules such as those found in gametogenesis and fertilization. In Drosophila, for example, there is evidence for adaptive evolution of seminal fluid proteins (Aguadé et al. 1992; Mueller et al. 2005), and it has been shown that a relatively high proportion of sex- and reproduction-related genes had experienced accelerated divergence throughout the genus Drosophila (Haerty et al. 2007). The study of more homogastrid species and other earthworm families would help to understand the evolution of reproduction-related proteins in these soil organisms.

Expression Profiles

All published gene expression studies conducted on earthworms so far have used microarrays (Bundey et al. 2008; Li et al. 2010; García-Reyero et al. 2011), this being the first RNA-seq analysis of multiple tissues of an earthworm species and the first study addressing the identification and determination of expression values of reproductive genes in earthworms. Among the genes highly expressed in the REP, we found proteins involved in the metabolism of glycogen such as glycogen phosphorylase, glycogen synthase, and phosphoglucomutase, and proteins involved in the formation and regulation of cytoskeleton, such as the EF-hand domain containing protein 1, and some which may enhance sperm motility such as Tektin-3 and ankyrin repeat and protein kinase domain-containing protein 1. Also, in the analysis of the differential gene expression, we found proteins involved in the synthesis of hormones such as Cytochrome P450 2J6, signaling peptides such as CUB and EGF-like domain-containing protein 2, stress-related proteins such as DNAJ-18, and proteins involved in apoptosis such as VDAC2 and poly ADP-ribose polymerase 14.

Pheromone genes in H. elisae (attractin and temptin) appeared to be slightly more expressed in the DIG, except for temptin 3, which was more expressed in the REP. temptins 1 and 2 were the genes showing the highest differences between expression levels in the DIG and the remaining tissues. It has been conjectured that the digestive tract might have a role in secreting pheromones in dipterans (Lu and Teal 2001), opening a question about the possibility of such explanation for the upregulation of pheromone precursors in the DIG of H. elisae. In earthworms, the release of the pheromones in conjunction with casts could be advantageous because it is a manner of leaving a trail of chemical signals for conspecifics. In this way, two individuals increase the possibility of encountering in the soil. Caro et al. (2012) already suggested the existence of chemical cues in galleries when they found that the presence of those accelerated the dispersion of an anec earthworm species. In this sense, these attractants could be directed not only toward reproduction but also toward the search of suitable environment or food.

The germ line genes (vasa, PL10 2, germ cell-less, piwi 1, and mago nashi), SPATA2, and acrosin were upregulated in the REP. This was expected given that these germ line genes play an essential role in gametogenesis, determining the precursor cells that will become gametes (Juliano and Wessel 2010). Also, the upregulation of the sex determination and fertilization genes SPATA2 and acrosin has been reported in gonads of other animals (Nayernia et al. 1994; La Salle et al. 2011). Marginal expression of germ line genes in other than REP is expected in our earthworms, because the expression of these genes has been previously reported from somatic tissues.
during embryogenesis (brain, mesodermal bands, and fore-gut) and in nongenital segments during the adulthood in other annelids (Oyama and Shimizu 2007; Dill and Seaver 2008) and other metazoans (Juliano and Wessel 2010). The fact that only vasa, PL10 2, germ cell-less, piwi 1, and mago nashi are upregulated in the REP may indicate a major role of these genes in germ line determination, whereas the other might be also implicated in maintaining totipotency, but not only in germ line cells.

**Conclusions**

We have used Illumina RNA-seq data for de novo assembling transcripts for two Mediterranean earthworm species, one of them segregated into three tissue sets. The two assembly approaches provided different contig sets, being longer the ones produced by V/O but more unique the ones provided by CLC. The pheromones Attractin and Temptin were found in these transcripts, the first showing potential to be used as a phylogenetic marker. It also shows interesting domain rearrangements during metazoan evolution. In turn, different paralogs and isoforms of temptin were detected questioning its validity for phylogenetics and species delimitation. These pheromones were overexpressed in the DIG, when compared with the others, opening the possibility of their release with casts to leave an attractant trail. We also describe the molecular machinery of sexual reproduction in these earthworms and found several genes involved in germ line determination, sexual differentiation, and fertilization.

**Materials and Methods**

**Sample Collection**

Two species of hormogastrid earthworms were collected by digging in appropriate soils: *H. elisae* from El Molar, Spain (40° 44’22.9”N, 3° 33’53.1”W) and *H. samnitica* from Gello, Italy (43° 19’49.0”N, 10° 42’30.2”E). Samples were preserved in RNAlater (Life Technologies) immediately after collection. For *H. samnitica*, pieces from the posterior of the animal were preserved (including tegument and digestive tissue, as well as nervous system, circulatory system, muscular septa, and nephridia), whereas for *H. elisae*, specific tissues were dissected out in RNAlater under the stereomicroscope. These resulted in: 1) reproductive tissue (REP, i.e., spermathecae containing sperm, seminal vesicles, seminal funnels, clitellum); 2) digestive tissue (DIG, i.e., gizzards, pharynx, eosophagus, stomach, intestine, typhlosole); and 3) remaining tissue (REST, i.e., nervous system, circulatory system, intestine, muscular septa, nephridia). Tissues were immersed in at least 10 volumes of RNAlater and stored in this buffer at –80 °C until RNA was extracted. Between 20 and 80 mg of tissue was placed in each eppendorf tube for subsequent processing. Tissue excisions were always performed with sterilized razor blades rinsed in RNaseZap (Ambion). All cleaning procedures were operated in an RNase-free and cold environment to avoid RNA degeneration.

**mRNA Extractions**

Total RNA was extracted, followed by mRNA purification. For total RNA extraction, we used a standard trizol-based method using TRI Reagent (Life Sciences) following the manufacturer’s protocol. Clean tissue pieces, previously stored in RNAlater, were flash frozen in liquid N₂ before tissue disruption, which was performed in flash frozen 500 µl of TRI Reagent using an RNase-free plastic pestle for grinding (with a drill). Another 500 µl of TRI Reagent was added, and after 5 min incubating at room temperature (RT), 100 µl of bromochloropropyl (BCP) was mixed by vortexing. After incubation at RT for 10 min, the samples were centrifuged at 16,000 rpm during 15 min at 4 °C. The upper aqueous layer was recovered, mixed with 500 µl of isopropanol, and incubated at –20 °C overnight. For the total RNA precipitation, the sample was centrifuged for 15 min at 16,000 rpm and 4 °C. Two washing steps of the pellet were performed by adding 1,000 µl of 75% EtOH and centrifuging first during 15 min at 16,000 rpm and 4 °C, and subsequently for 5 min at 7,600 rpm and 4 °C. The dried RNA pellet was eluted in 30 µl of Ambion RNA storage solution with 1 µl of ANTI-RNase (Life Technologies). Subsequent mRNA purification was done with the Dynabeads mRNA Purification Kit (Invitrogen) following manufacturer’s instructions.

**Next-Generation Sequencing**

cDNA library construction for *H. samnitica* was described in Riesgo et al. (2012); mRNA was used for random primed first-strand synthesis using SuperScript II Reverse Transcriptase (Life Technologies), followed by second strand synthesis with DNA Polymerase I and enzymatic fragmentation using the NEBNext dsDNA Fragmentase (New England Biolabs). End repair of the double-stranded cDNA (ds cDNA) was performed with NEBNext End Repair Module (New England Biolabs), and an additional dAmp was incorporated with the NEBNext dA-Tailing Module (New England Biolabs). ds cDNA was ligated to Illumina adapters using the NEBNext Quick Ligation Module (New England Biolabs). Size-selected cDNA fragments of around 350~450 bp were excised from a 2% agarose gel, purified and amplified using Illumina polymerase chain reaction (PCR) Primers for Paired-End reads (Illumina), and 18 cycles of the PCR program 98 °C for 30 s, 98 °C for 10 s, 65 °C for 30 s, 72 °C for 30 s, followed by an extension step of 5 min at 72 °C. For *H. elisae*, TruSeq for RNA Sample Preparation kit (Illumina) was used, following the manufacturer’s instructions and using a different index for each of the three types of tissue to be pooled into a single Illumina lane.

Concentration of the cDNA libraries was measured with a Qubit Fluorometer (Invitrogen) using the Qubit dsDNA High Sensitivity (HS) Assay Kit. Library quality and size selection were checked in an Agilent 2100 Bioanalyzer (Agilent Technologies) with the “HS DNA assay.” Fragment size was 447 bp for *H. samnitica* and 335, 307 bp and 268 bp for *H. elisae* samples (REP, DIG, and REST, respectively). The samples were run using the next-generation sequencing platform GAII (H. samnitica) and Illumina HiSeq (pooled libraries of
H. elisae) with paired-end reads of 150 bp and 101 bp, respectively, at the FAS Center for Systems Biology at Harvard University.

Sequence Assembly

Five different data sets were assembled: one for H. samnitica and four for H. elisae: one for each of the three individual tissue samples (REP, DIG, and REST) and one for the combination of the three. Thinning of the raw reads and removal of adapters and primer sequences were performed in CLC Genomics Workbench 4.6.1 (CLC bio, Aarhus, Denmark). Limit for thinning was set to 0.05 and 0.005 (based on Phred quality scores), and resulting quality of the thinned reads was visualized in FastQC (http://www.bioinformatics. bbsrc.ac.uk/projects/fastqc/). After thinning, reads showed good quality, deeming trimming unnecessary.

De novo assemblies for each of the five data sets, each thinned with the two thinning limits (0.05 and 0.005) were performed in CLC. Global alignments for the de novo assemblies were always done using the following default parameters: mismatch cost = 2; insertion cost = 3; deletion cost = 3; length fraction = 0.5; similarity = 0.8; and randomly assigning the nonspecific matches. Best k-mer lengths were estimated by the software. The best assembly for each species/tissue was selected following criteria previously outlined by Riesgo et al. (2012).

After selecting the thinning value that produced the best assemblies for CLC (see below), the reads were assembled with Velvet/Oases (V/O) for the complete data sets (H. samnitica and H. elisae including reads from the three tissue sets) to compare both algorithms. A preliminary assembly was produced by Velvet v.1.2.03 (Zerbino and Birney 2008) and further improved with Oases v. 0.2.01 (Schulz et al. 2012). We examined the assemblies over a range of k-mer values from 41 to 69 with VelvetOpt, the latter resulting in the best for both species.

BLAST and Functional Annotation

Contigs from the assemblies for the two hormogastrid species (CLC and V/O based) were mapped against a selection of the nr National Center for Biotechnology Information (NCBI) GenBank database release 190 (June 2012) using the blastx program and including only proteins from Metazoa. All BLAST searches were conducted with BLAST + 2.2.23 (Altschul et al. 1990; Camacho et al. 2009) using an e-value threshold of 1e-5 and 1e-10. With the resulting files, we then used Blast2GO v2.5.0 (Conesa et al. 2005) to retrieve the GO terms (Harris et al. 2004) and their parents associated with the top BLAST hit for each sequence. We performed the same BLAST and annotations for the L. rubellus transcriptome retrieved from www.earthworms.org (Elsworth B, personal communication). Specific GO terms were searched for the categories “biological process,” “molecular function,” and “cellular component” following the criteria of Ewen-Campen et al. (2011) in the three species. GOs were calculated as total and percentage, and only those that were represented by more than 1% were included. We also performed a one-tailed Fisher’s exact test with multiple test correction by Benjamini–Hochberg false discovery rate (FDR) to analyze the differential GO term enrichment (P < 0.05) in each tissue.

Redundancy of the BLASTs was calculated to detect whether different contigs BLASTed against the same protein (i.e., unique hit understood as only one contig matching each protein and redundant hits as more than one contig matching the same protein). We performed BLASTs among the different assemblies to compare the hits of CLC and V/O assemblies as well as the three species (H. samnitica, H. elisae, and L. rubellus) and the three tissues of H. elisae. Venn diagrams were elaborated by calculating reciprocal BLAST hits among contig files.

Pheromones and Other Genes Involved in Reproduction

A major goal of this study was to identify candidate genes involved in attraction and reproduction. The gene sequences for the pheromones attractin, temptin, seduction, and enticin, previously described in Aplysia (Painter et al. 1998; Cummins et al. 2004, 2006), the germ line markers vasa, nanos, PL10, piwi, germ cell-less, tsunagi, mago-nashi, oskar, and smaug, and the reproductive genes DMRT, fertilin, acrosin, SAA7, SPATA2, and SOX3 were searched in the earthworms’ transcriptomes. We downloaded the sequences of at least three different orthologs of the selected protein targets from several invertebrate species (trying to find the closest available relatives of earthworms) and BLASTed them individually against our transcriptomes (using tblastn engines). The search was performed among the contigs longer than 500 bp to maximize the number of full-length proteins or at least with a length sufficient to recover full domains. We then selected only the hits with the maximum similarity and checked each open reading frame with ORF finder (http://www.ncbi.nlm.nih. gov/orf/orfig.cgi). Each predicted protein sequence was then re-BLASTed against the nr database in NCBI using the blastp program (http://blast.ncbi.nlm.nih.gov/), and the domain structure was rechecked in SMART v7 (http://smart.embl-heidelberg.de/) using the bundled HMMER to search for PFAM domains and internal repeats (Schultz et al. 1998; Letunic et al. 2012). The domain structure was plotted with the software DOG 2.0 (http://dog.biocuckoo. org/). If more than one sequence was found in the local BLAST searches for each gene, we used IsoSVM (Spitzer et al. 2006) to determine whether the sequences were paralogs or isoforms generated by alternative splicing.

Orthologs for the proteins Attractin and Temptin in metazoans and protozoans were downloaded from NCBI (see accession numbers in supplementary table S2, Supplementary Material online) and aligned with our translated sequences using the program MUSCLE in Seaview (Gouy et al. 2010). A phylogenetic analysis using amino acid sequences was performed with PhyML 3.0 (Guindon et al. 2010) using maximum likelihood with an LG model of amino acid replacement and estimated gamma-shape parameters obtained using Prottest 2 (Abascal et al. 2005). Bootstrap values were estimated in
PhyML with 1,000 replicates. The protozoan Capvaspora owczarzaki was used as outgroup for the Attractin tree, and the placozoan Trichoplax adhaerens was included as outgroup for the Temptin tree. The pheromone genes were also searched in the genome of the choanoflagellate Monosiga brevicollis without success.

Expression Profiles
Heat maps summarizing differential gene expression profiles were obtained with CLC Genomics Workbench 4.6.1 (CLC bio, Aarhus, Denmark) by comparing the expression levels among the three tissues of H. elisae. The 351,000 contigs were compared, and those with a size longer than 1,000 bp (N = 25,838) were represented in detail to assure full length. Expression was measured in RPKM, and scaling normalization was performed on the original expression values (nRPKM). Because no reference genome was available for the selected species, exons were not annotated, and in turn, the assembled contigs were assigned a complete exon. We plotted the nRPKM values against the contigs longer than 1,000 bp to analyze the distribution of the expression levels (see later) and to establish a threshold for the most striking differences between tissues. Because the main goal of this work was the study of the molecular machinery for earthworm reproduction, we focused on those contigs that showed differential expression among DIG and REST (<2 nRPKM) and REP (> 100 nRPKM) (N = 39). Finally, the expression levels of specific genes, related to the attraction (pheromones), sexual differentiation and determination, and fertilization (see list above) were compared among tissues of H. elisae and searched for their presence/absence in H. samnitica.

Data Availability
All sequenced data have been deposited at the NCBI Short Read Archive in the projects: PRJNA181254 (H. samnitica: accession no. SRS374608) and PRJNA196484 (H. elisae: REP accession no. SRS374609, DIG accession no. SRS374610, and REST accession no. SRS374611).

Supplementary Material
Supplementary tables S1 and S2 and figures S1–S4 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments
The authors are indebted to Rosa Fernández for field assistance, to Darío Díaz Cosín for his accurate dissections and for his kindness, to Ben Elsworth (member of the Earthworm Genome Sequencing Consortium, Blaxter Lab, University of Edinburgh, UK) for providing access to L. rubellus data, and to Daniel Navarro and Sónia Andrade for advice with data analyses. The FAS Center for Systems Biology at Harvard University provided technical support and the NGS infrastructure. M.N. was supported by a grant from the Fundación Caja Madrid and by a Postdoctoral Fellowship from the Spanish Government. A.R. was recipient of a European Commission Marie Curie IOF grant. This research was funded by internal funds from the Museum of Comparative Zoology and the Faculty of Arts and Sciences to G.G.

References
Darwin C. 1881. The formation of vegetable mould through the actions of worms. London: John Murray.


