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

Marta Novo,*,‡,[1] Ana Riesgo,[1,2] Antoni Fernández-Guerra,[2] and Gonzalo Giribet[1]

[1]Museum of Comparative Zoology, Department of Organismic and Evolutionary Biology, Harvard University

[2]Centro de Estudios Avanzados de Blanes, CSIC, Girona, Spain

‡Present address: Cardiff School of Biosciences, Cardiff University, Cardiff, United Kingdom

***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 samnitica* and three different tissues of *H. elisae* 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 (Laverack 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 lamarckii* (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 (Extavour 2007). In annelids, the proteins encoded by the genes *vasa*, *PL10*, *piwi*, 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 (Suzuki et al. 2005), as many other metazoans do (Volff et al. 2003). After the formation of gametes 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, Riesgo 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; Pirooznia et al. 2007; Owen et al. 2008; Gong et al. 2010) and megascolecids (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 (Pirooznia 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*, Álvarez [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 Riesgo 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 58.37% 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 analyses 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.

Species	Tissue	Sequencer	Insert Size (bp)	Thin Limit	*N* Reads BT	*N* Reads AT	*N* Reads Trimmed	Avg. L AT	*N* Contigs	*N* Bases	Avg. L Contigs	SD	Max Contig L	N50	N50 Contig Count	Avg. L of N50	SD
H. samnitica	Posterior part	GAII	447	A	50,789,952	47,857,894	2,932,058	97.33	189,839	75,902,786	399.83	312.52	7,319	423	49,506	766.61	426.83
				B	50,789,952	45,094,358	5,695,594	83.77	169,050	68,825,148	407.13	320.15	5,661	436	43,377	793.34	432.37
H. elisae	REP	Hi-seq	331	A	107,887,540	105,784,930	2,102,610	100.58	301,292	135,462,453	449.61	474.45	17,619	497	64,973	1,042.46	757.65
				B	107,887,540	103,123,804	4,763,736	100.44	299,301	134,789,005	450.35	475.86	17,619	499	64,371	1,046.98	759.82
	DIG	Hi-seq	307	A	36,873,140	35,998,970	874,170	100.74	120,501	63,002,546	522.84	510.76	13,265	650	25,233	1,248.43	725.35
				B	36,873,140	35,196,193	1,676,947	100.67	119,745	62,694,709	523.57	511.55	13,265	651	25,052	1,251.29	725.96
	REST	Hi-seq	268	A	56,019,876	54,758,116	1,261,760	100.77	113,820	51,848,245	455.53	393.03	9,828	509	27,444	944.63	551.22
				B	56,019,876	53,713,426	2,306,450	100.7	113,519	51,708,606	455.51	393.59	9,828	509	27,371	944.6	552.83
	All together[a]	[a]	[a]	[a]	[a]	192,033,423	[a]	100.55	351,596	157,992,055	449.36	475.92	17,620	499	75,491	1,046.43	760.69

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.

[a] This assembly was performed with the selected thinned reads from the partial tissues all together.

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

	Species	*Hormogaster samnitica*				*Hormogaster elisae*											
	Tissue					REP				DIG				REST			
		A	Percentage	B	Percentage	A	Percentage	B	Percentage	A	Percentage	B	Percentage	A	Percentage	B	Percentage
Contig length	<300	98,728	52.01	86,493	51.16	151,384	50.24	150,082	50.14	48,703	40.42	48,350	40.38	49,172	43.20	49,066	43.22
	300-500	53,904	28.39	48,028	28.41	84,665	28.10	83,989	28.06	34,808	28.89	34,554	28.86	35,818	31.47	35,726	31.47
	500-1,000	27,796	14.64	25,518	15.09	43,034	14.28	42,949	14.35	24,079	19.98	24,061	20.09	20,491	18.00	20,380	17.95
	1,000-2,000	7,928	4.18	7,747	4.58	16,078	5.34	16,020	5.35	9,902	8.22	9,795	8.18	6,544	5.75	6,533	5.75
	2,000-3,000	952	0.50	879	0.52	3,708	1.23	3,698	1.24	2,092	1.74	2,086	1.74	1,071	0.94	1,060	0.93
	3,000-4,000	159	0.08	140	0.08	1,149	0.38	1,162	0.39	522	0.43	534	0.45	239	0.21	239	0.21
	4,000-5,000	28	0.01	26	0.02	374	0.12	371	0.12	163	0.14	161	0.13	62	0.05	65	0.06
	5,000-6,000	7	0.00	6	0.00	161	0.05	176	0.06	61	0.05	58	0.05	27	0.02	30	0.03
	>6,000	2	0.00	0	0.00	153	0.05	147	0.05	38	0.03	39	0.03	6	0.01	6	0.01
	Total	189,839		189,839		301,292		299,301		120,501		119,745		113,820		113,519	

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.

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

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

Note.—Individual tissues were only assembled with CLC (see tables 1 and 2). *N*, number; Avg., average; L, length; V/O, Velvet/Oases.

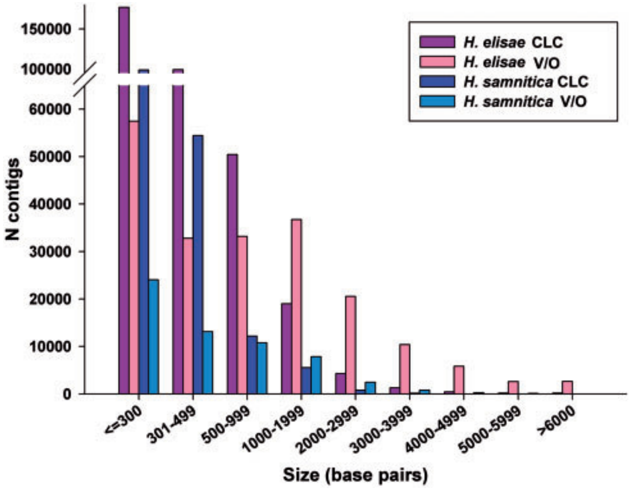

Fig. 1. Comparison of the size distribution (in base pairs) of contigs assembled by CLC and V/O with transcriptomic data from *Hormogaster elisae* and *Hormogaster samnitica*. *N* contigs, number of contigs.

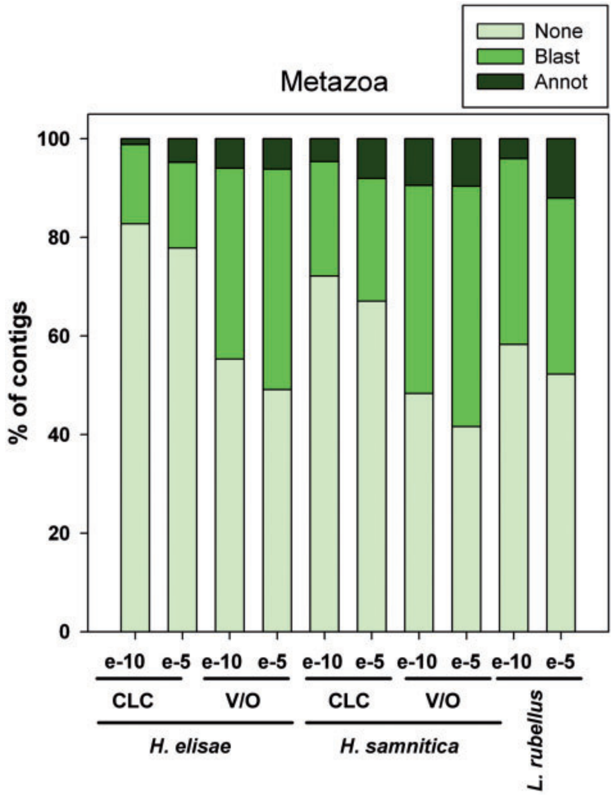

Fig. 2. Percentage of contigs without BLAST hit (None), with BLAST hit (Blast), and with GO assignment (Annot). Comparisons are shown among the three earthworm species, the two assembly methods for hormogastrids (CLC and V/O), and the used *e* values: least stringent (1e-5) and most stringent (1e-10).

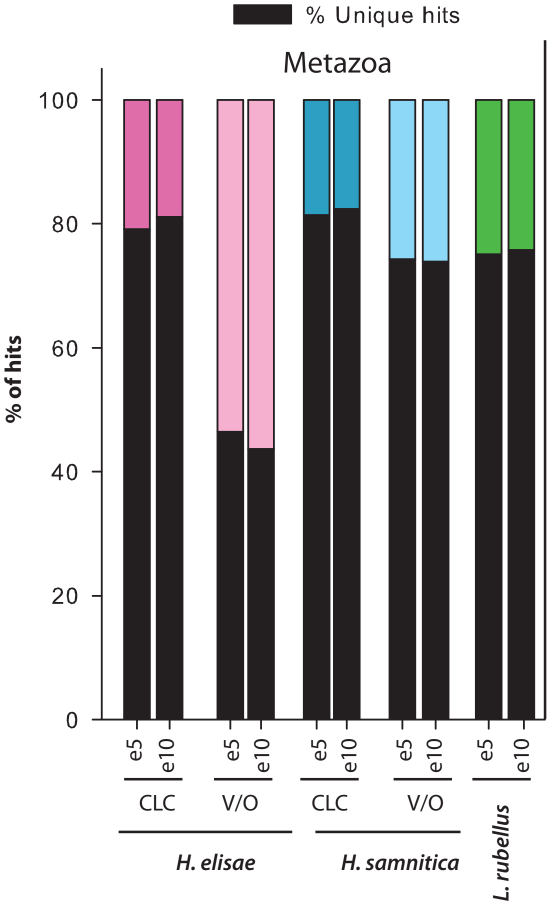

Fig. 3. Percentage of contigs that resulted in unique hits (only one contig matching to each protein) and redundant hits (two or more BLAST hits matching to each protein). Data are shown for the three earthworm species, the two assembly methods (CLC and V/O), and the used *e* values: less stringent (1e-5) and most stringent (1e-10).

In *H. samnitica*, between 4,246 and 7,420 contigs had GO terms assigned (*e* values 1e-10 and 1e-5, respectively), whereas in *L. rubellus*, 2,570 and 7,676 contigs had GO terms assigned (also *e* 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

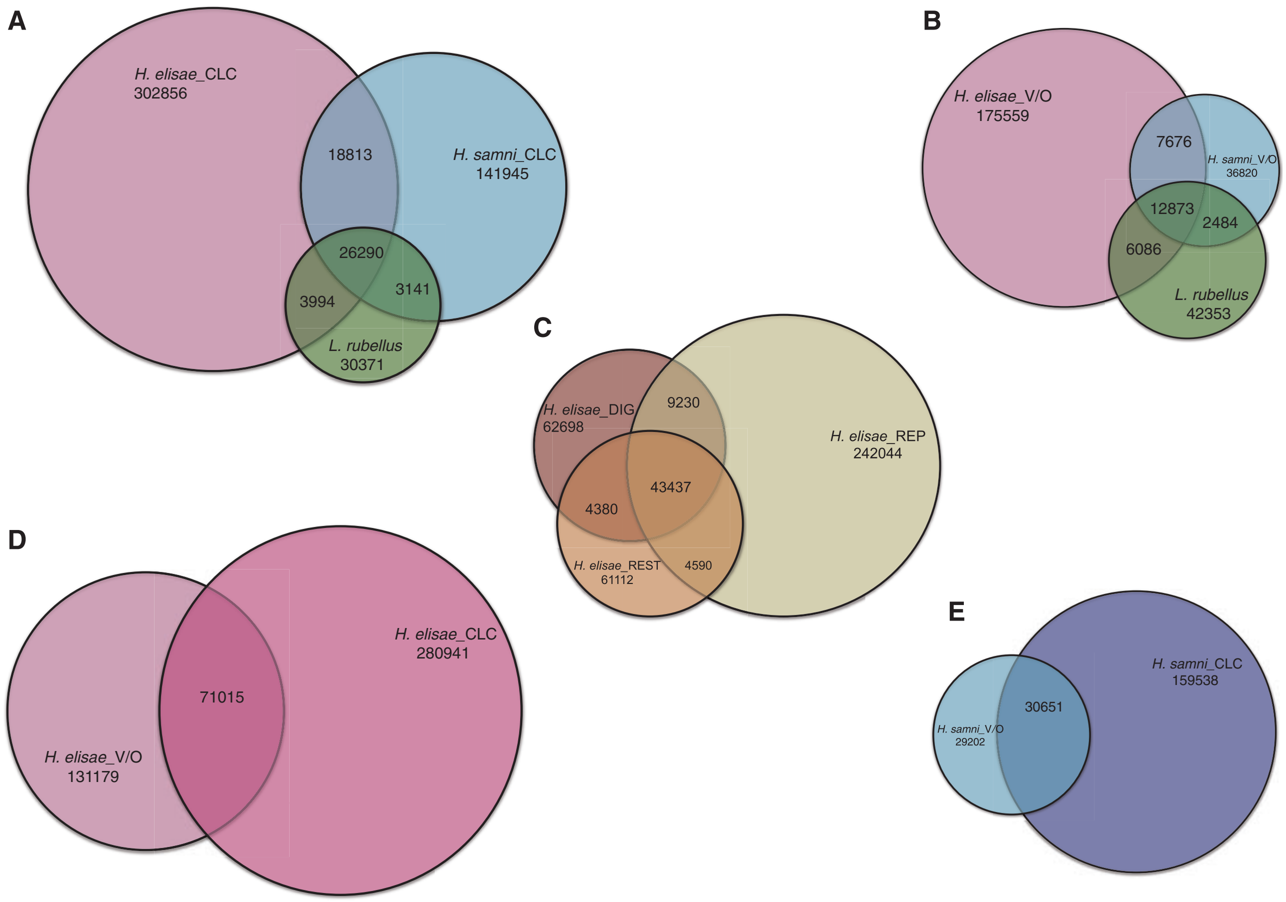

Fig. 4. Venn diagrams showing the overlap (number of contigs indicated) of the different assemblies comparing the three species (*A*: CLC, *B*: V/O), the overlap of contigs from different assemblies (CLC or V/O) for the same species (*D*: *Hormogaster elisae*; *E*: *H. samnitica*), and the comparison of the three tissues of *H. elisae* (*C*). The size of the circles is proportional to the number of contigs, but the overlap area is not exact. REP, reproductive tissue; DIG, digestive tissue; REST, rest of the tissues not included in REP and DIG.

H. elisae. 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 *seductin*), 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.

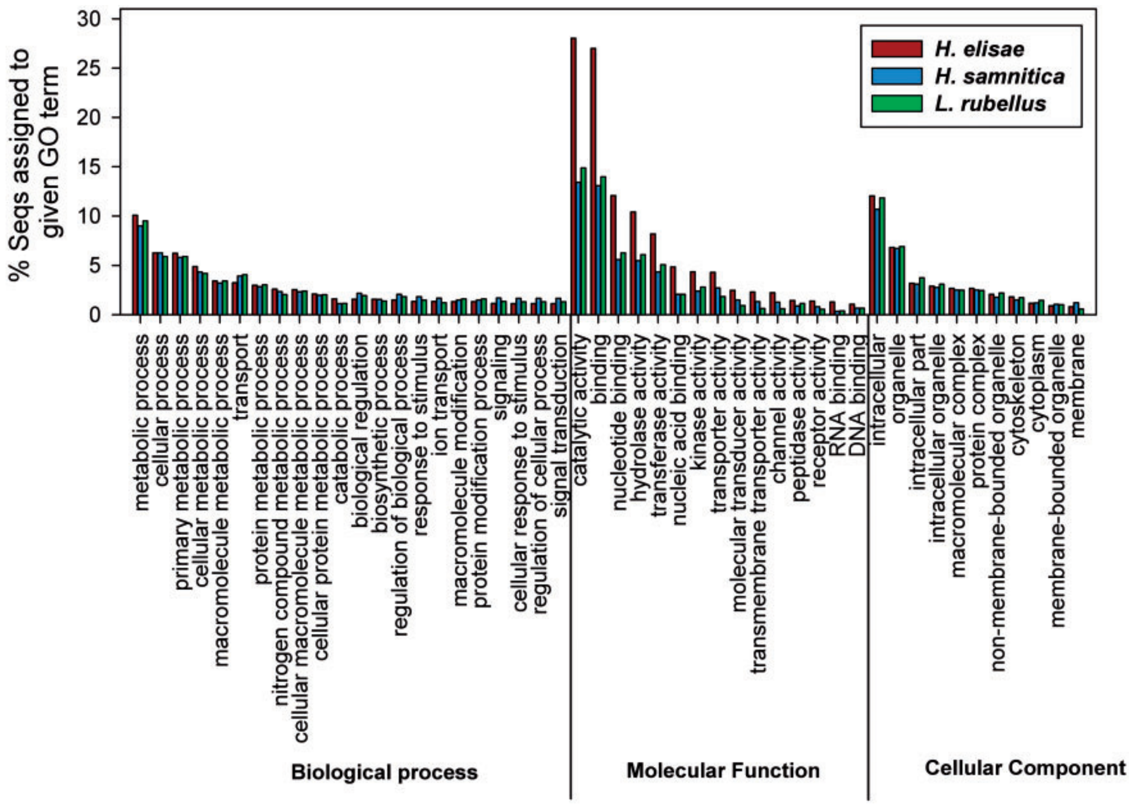

Fig. 5. Percentage of contigs mapped to given GO terms for the three earthworm species. CLC assemblies are used for hormogastrids, and data retrieved from www.earthworms.org (Elsworth B, personal communication) are used for *Lumbricus rubellus*.

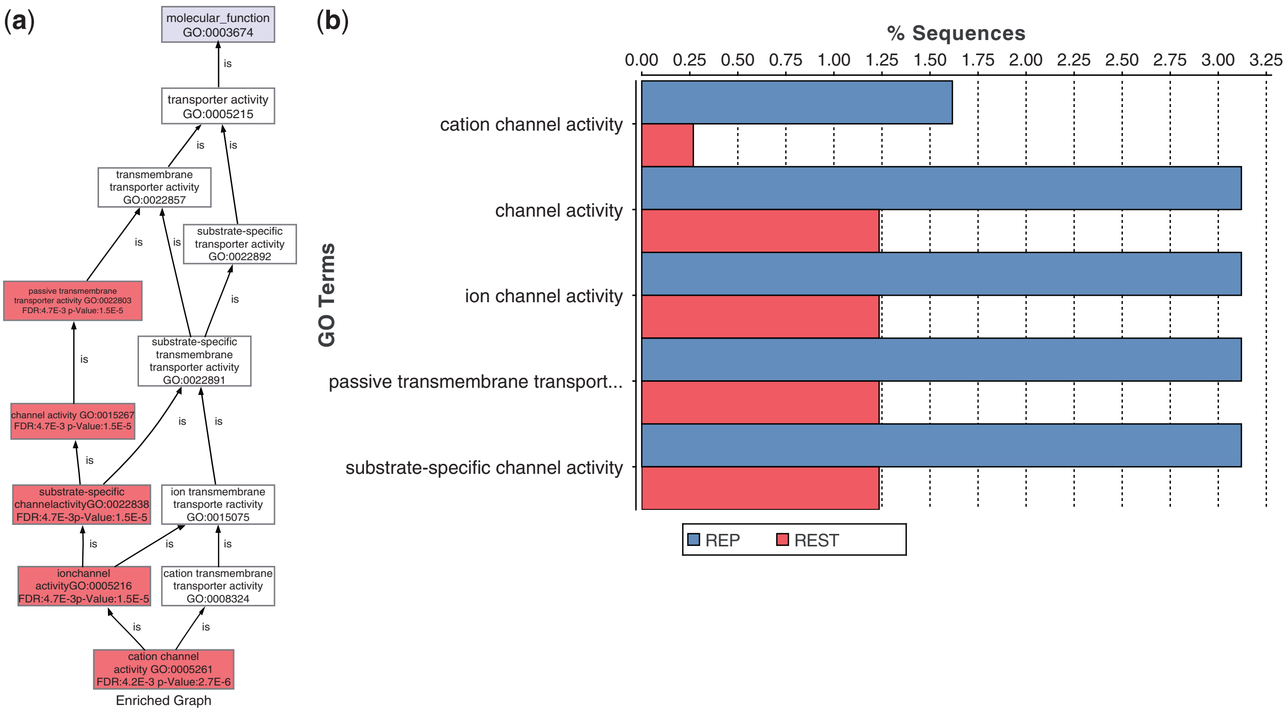

Fig. 6. Differentially enriched GO terms in the reproductive tissue (REP) of *Hormogaster elisae* compared with the normal tissue (REST). (*A*) Hierarchical diagram of differentially enriched GO terms in the REP (shaded) and their relationships with related terms. (*B*) For the differentially enriched GO terms, the percentage of sequences assigned in both tissue types is represented.

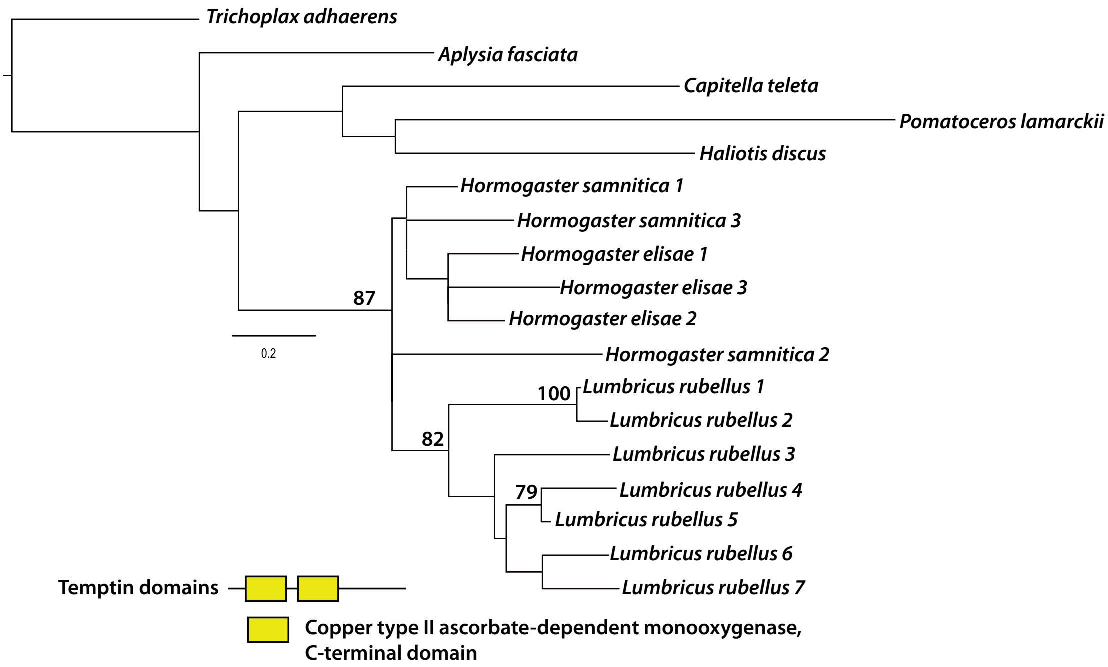

Fig. 7. Phylogenetic reconstruction showing the paralog sequences of the pheromone Temptin in earthworms. Bootstrap values above 50% are shown above the branches. GenBank accession numbers for all sequences used can be found in supplementary table S2, Supplementary Material online.

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 (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 trancriptomic 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

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

Genes	*Hormogaster elisae*						*H. samnitica*		
	Presence	Length (aa)	Expression (nRPKM)			Accession Number	Presence	Length (aa)	Accession Number
			Digestive	Rest	Reproductive				
Pheromones									
Attractin	Yes	1,177	9	3.5	3.3	GAHS01000001	Yes	912	GAHR01000001
Temptin 1	Yes	296	35.5	1.6	0.1	GAHS01000002	Yes	105	GAHR01000002
Temptin 2	Yes	96	5.34	0.6	0	GAHS01000003	Yes	97	GAHR01000003
Temptin 3	Yes	120	11.3	16.4	25.1	GAHS01000004	Yes	54	GAHR01000004
Temptin-like 1	Yes	278	34	3.8	0	GAHS01000005	Yes	378	GAHR01000005
Temptin-like 2	No	—	—	—	—	—	Yes	421	GAHR01000006
Other hormones									
Annetocin precursor	Yes	130	0.1	7.5	0	GAHS01000033	No	—	—
Annetocin receptor 1	Yes	140	0.4	1.5	0.1	GAHS01000006	Yes	120	GAHR01000007
Annetocin receptor 2	Yes	129	0.1	0.8	1.5	GAHS01000007	Yes	118	GAHR01000008
Annetocin receptor 3	Yes	132	0.1	0.2	0.4	GAHS01000008	No	—	—
Annetocin receptor 4	Yes	70	0	0	0.4	GAHS01000009	No	—	—
Sexual differentiation									
DMRT3	Yes	340	0.7	2.4	19.2	GAHS01000021	Yes	105	GAHR01000024
SAA	Yes	198	29.2	20.4	7.3	GAHS01000030	Yes	93	GAHR01000022
SPATA2	Yes	384	4.4	4.5	22.6	GAHS01000032	Yes	67	GAHR01000021
SOX 3	Yes	296	5.4	1.7	11.8	GAHS01000031	Yes	267	GAHR01000023
Germ line determination									
vasa	Yes	153	3.8	1.3	74.1	GAHS01000013	Yes	200	GAHR01000012
PL10 1	Yes	435	12.9	8	20.4	GAHS01000010	Yes	755	GAHR01000009
PL10 2	Yes	359	26.4	20	55.5	GAHS01000011	Yes	601	GAHR01000010
PL10 3	Yes	294	5.1	6.4	2.3	GAHS01000012	Yes	212	GAHR01000011
Germ cell-less	Yes	414	1.9	1.4	74.1	GAHS01000014	Yes	499	GAHR01000013
Tsunagi	Yes	178	22.3	23.1	32.6	GAHS01000016	Yes	133	GAHR01000015
Piwi 1	Yes	906	5.6	3.3	45.3	GAHS01000017	Yes	273	GAHR01000016
Piwi 2	Yes	745	3	3.3	13.7	GAHS01000018	Yes	200	GAHR01000017
Smaug	Yes	597	1	0.3	5	GAHS01000020	Yes	192	GAHR01000020
Nanos 1	Yes	145	0.6	0.3	1.6	GAHS01000019	Yes	101	GAHR01000018
Nanos 2	No	—	—	—	—	—	Yes	95	GAHR01000019
Mago-nashi	Yes	149	31.3	52.1	88.1	GAHS01000015	Yes	143	GAHR01000014
Fertilization									
Fertilin 1	Yes	400	4.4	3.3	2	GAHS01000022	No	—	—
Fertilin 2	Yes	611	5.6	3.4	0.5	GAHS01000023	No	—	—
Fertilin 3	Yes	781	2.8	2.9	2.5	GAHS01000024	No	—	—
Fertilin 4	Yes	288	0.5	1.2	1.5	GAHS01000025	No	—	—
Fertilin 5	Yes	287	1.6	0.7	1.2	GAHS01000026	No	—	—
Acrosin 1	Yes	216	5.7	2.3	2.3	GAHS01000027	No	—	—
Acrosin 2	Yes	205	4.9	6.3	43.6	GAHS01000028	No	—	—
Acrosin 3	Yes	337	22.7	23.8	7.8	GAHS01000029	No	—	—

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.

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

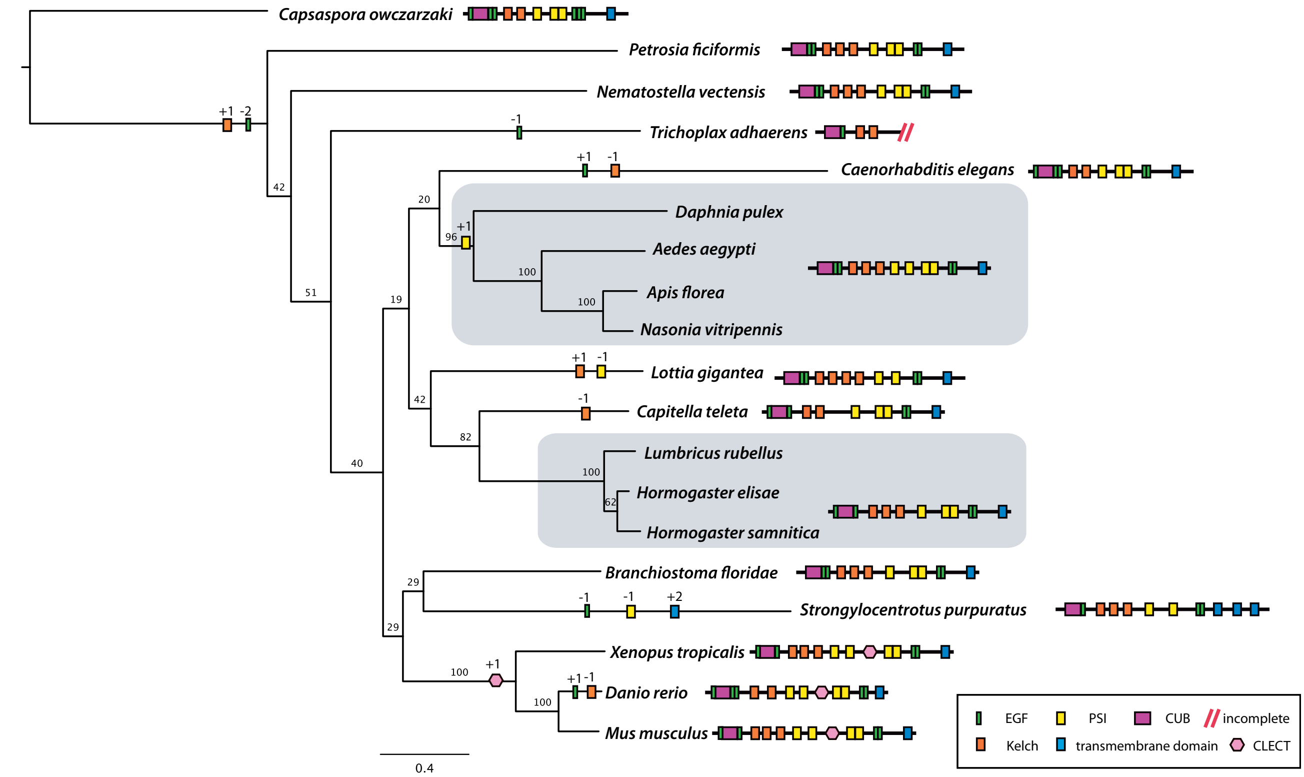

Fig. 8. Phylogenetic reconstruction of metazoans using the protein sequence of the pheromone Attractin. Bootstrap support values are shown above the branches. Evolution of protein domains is shown along the tree. GenBank accession numbers for all sequences used can be found in supplementary table S2, Supplementary Material online.

(Schulz et al. 2012). 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 assemblies 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, Temptin, 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 Temptin is what induces attraction. Their

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.

Contig Number	Protein Name in NCBI	Species	*E* Value	Accession Number	nRPKM		
					Digestive	Rest	Reproductive
254294	Cytochrome P450 2J6	*Crassostrea gigas*	2.00E-65	EKC41577.1	0	0.1	169.4
241136	hypothetical protein Smp_086420	*Schistosoma mansoni*	8.60E + 00	CCD80869.1	0.4	1.23	151.4
325591	EF-hand domain-containing protein 1	*Crassostrea gigas*	0.00E + 00	EKC30173.1	0.4	1.28	146.4
255355	Ankyrin repeat and protein kinase domain-containing protein 1	*Crassostrea gigas*	8.00E-10	EKC18036.1	0.3	2.1	142.4
254411	phosphoglucomutase 1	*Mustela putorius furo*	8.00E-145	AES03896.1	0.5	1.8	136.5
254080	signal peptide, CUB and EGF-like domain-containing protein 2	*Sus scrofa*	3.00E-29	XP_003129439.3	0.5	1	136.2
325605	similar to voltage-dependent anion-selective channel isoform 2	*Tribolium castaneum*	8.00E-67	XP_976150.1	0.9	2.1	129.9
254082	methenyltetrahydrofolate synthase domain-containing protein-like	*Oreochromis niloticus*	3.00E-49	XP_003442372.1	0.7	1	127.3
253469	DnaJ-18	*Bombyx mori*	7.00E-56	AFC01232.1	0.4	1	126.9
253629	hypothetical protein PANDA_005434	*Ailuropoda melanoleuca*	0.00E + 00	EFB15915.1	0.5	1.2	126.5
253603	Poly [ADP-ribose] polymerase 14	*Crassostrea gigas*	9.00E-09	EKC39322.1	0.5	1.3	125
94896	hypothetical protein CGI_10027320	*Crassostrea gigas*	3.00E-35	EKC43221.1	0.9	0.9	120.1
105021	GA14893 U605_DROPS	*Drosophila pseudoobscura*	4.00E-07	XP_001361258.2	0.8	0.6	119.1
325760	glycogen phosphorylase	*Belgica antarctica*	4.00E-126	AFS17314.1	0.6	1.7	117.5
253514	High mobility group protein B3	*Crassostrea gigas*	4.00E-11	EKC41956.1	0.8	2.1	114.8
253900	glycogen phosphorylase	*Marupenaeus japonicus*	0.00E + 00	BAJ23879.1	0.7	0.7	113.7
238696	glycogen [starch] synthase, muscle-like	*Strongylocentrotus purpuratus*	1.00E-113	XP_783574.3	0.8	1.35	113.5
238838	Tektin-3	*Crassostrea gigas*	2.00E-67	EKC40032.1	0.5	0.6	108.3
312866	proteasome activator complex subunit 3-like, partial	*Amphimedon queenslandica*	4.20E-01	XP_003392061.1	0.4	0.5	108.3
240618	Coiled-coil domain-containing protein 89, partial	*Crassostrea gigas*	3.00E-03	EKC31850.1	0.3	1.7	103.7
325671	Glycoprotein 3-alpha-L-fucosyltransferase A	*Crassostrea gigas*	5.00E-35	EKC41098.1	0.3	1.8	103.8
328238	Deleted in malignant brain tumors 1 protein	*Crassostrea gigas*	2.30E-33	EKC31891.1	0.5	0.8	103.4

Note.—The expression level is presented in nRPKM (Reads Per Kilobase of exon model per Million mapped reads, normalized values).

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 Fibrillin 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) 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 (Volff et al. 2003). The protein SPATA2 is involved in the meiotic progression of male and female gametes in vertebrates, but it is absent in *Drosophila* and *Caenorhabditis* (La Salle 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 hormogastrid 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 phosphorilase, 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 anecic 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*), *SPATA22,* and *acrosin 2* 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 foregut) 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 transcriptomes 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 transcriptomes, 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, oesophagus, stomach, intestine, typhlosole); and 3) remaining tissue (REST, i.e., nervous system, circulatory system, integument, 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 N_2 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 bromochloropropane (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*, *seductin*, 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' trancriptomes. 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/gorf/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 *Capsaspora 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

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

Aguadé M, Miyashita N, Langley CH. 1992. Polymorphism and divergence in the Mst26A male accessory gland gene region in *Drosophila*. *Genetics* 132:755–770.

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

Álvarez J. 1977. El género *Hormogaster* en España. *Publ Cent Piren Biol Exp.* 9:27–35.

Brennan PA, Keverne EB. 2004. Something in the Air? New insights into mammalian pheromones. *Curr Biol.* 14:R81–R89.

Buckley TR, James S, Allwood J, Bartlam S, Howitt R, Prada D. 2011. Phylogenetic analysis of New Zealand earthworms (Oligochaeta: Megascolecidae) reveals ancient clades and cryptic taxonomic diversity. *Mol Phylogenet Evol.* 58:85–96.

Bundey JG, Sidhu JK, Faisal R, Spurgeon DJ, Svendsen C, Wren JF, Stürzenbaum SR, Morgan AJ, Kille P. 2008. "Systems toxicology" approach identifies coordinated metabolic responses to copper in a terrestrial non-model invertebrate, the earthworm *Lumbricus rubellus*. *BMC Genomics* 6:25.

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

Cardé RT, Millar JG. 2009. Pheromones. In: Resg VH, Cardé R, editors. Encyclopedia of insects, 2nd ed. San Diego (CA): Academic Press/Elsevier Science. p. 766–772.

Caro G, Abourachid A, Decaëns T, Buono L, Mathieu J. 2012. Is earthworms' dispersal facilitated by the ecosystem engineering activities of conspecifics? *Biol Fertil Soils* 48:961e965.

Chang CH, Rougerie R, Chen JH. 2009. Identifying earthworms through DNA barcodes: pitfalls and promise. *Pedobiologia* 52:171–180.

Cho SJ, Lee MS, Tak ES, Lee E, Koh KS, Ahn CH, Park SC. 2009. Gene expression profile in the anterior regeneration of the earthworm using expressed sequence tags. *Biosci Biotechnol Biochem.* 73:29–34.

Christensen TA, Hildebrand JG. 1994. Neuroethology of sexual attraction and inhibition in heliothine moths. In: Schildberger K, Elsner N, editors. Neural basis of behavioural adaptations—progress in zoology. Stuttgart (Germany): G Fisher. p. 37–46.

Cobolli-Sbordoni M, de Matthaeis E, Alonzi A, Matoccia M, Omodeo P, Rota E. 1992. Speciation, genetic divergence and paleogeography in the Hormogastridae. *Soil Biol Biochem.* 24:1213–1221.

Cognetti ML. 1914. Escursioni Zoologiche del Dr. Enrico Festa nei monti della Vallata del Sangro (Abruzzo). Nota sugli Oligocheti degli Abruzzi. *Boll Mus Zool Comp Anat Univ Torino.* 29:689.

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

Cummins SF, Nichols AE, Amare A, Hummon AB, Sweedler JV, Nagle GT. 2004. Characterization of *Aplysia* Enticin and Temptin, two novel water-borne protein pheromones that act in concert with attractin to stimulate mate attraction. *J Biol Chem.* 279: 25614–25622.

Cummins SF, Nichols AE, Shein CH, Nagle GT. 2006. Newly identified water-borne protein pheromones interact with attractin to stimulate mate attraction in *Aplysia*. *Peptides* 27:597–606.

Cummins SF, Xie F, de Vries MR, Annangudi SP, Misra M, Degnan BM, Sweedler JV, Nagle GT, Shein CH. 2007. *Aplysia* temptin—the "glue" in the water-borne attractin pheromone complex. *FEBS J.* 274: 5425–5437.

Darszon A, Labarca P, Beltrán C, García-Soto J, Liévano A. 1994. Sea urchin sperm: an ion channel reconstitution study case. *Methods* 6:37–50.

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

Dill KK, Seaver EC. 2008. *Vasa* and *nanos* are coexpressed in somatic and germ line tissue from early embryonic cleavage stages through adulthood in the polychaete *Capitella* sp. I. *Dev Genes Evol.* 218: 453–463.

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

Edwards CA, editor 2004. Earthworm ecology, 2nd ed. Boca Raton (FL): CRC Press LLC.

Ewen-Campen B, Shaner N, Panfilio KA, Suzuki Y, Roth S, Extavour CG. 2011. The maternal and early embryonic transcriptome of the milkweed bug *Oncopeltus fasciatus*. *BMC Genomics* 12:61.

Extavour CGM. 2007. Evolution of the bilaterian germ line: lineage origin and modulation of specification mechanisms. *Int Comp Biol.* 47: 770–785.

Feldmeyer B, Wheat CW, Krezdorn N, Rotter B, Pfenninger M. 2011. Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (*Radix balthica*, Basommatophora, Pulmonata), and a comparison of assembler performance. *BMC Genomics* 12:317.

Fernández R, Almodóvar A, Novo M, Gutiérrez M, Díaz Cosín DJ. 2011. A vagrant clone in a peregrine species: phylogeography, high clonal diversity and geographic distribution in *Aporrectodea trapezoides* (Dugès, 1828). *Soil Biol Biochem.* 43:2085–2093.

García-Reyero N, Habib T, Pirooznia M, Gust KA, Gong P, Warner C, Willbanks M, Perkins E. 2011. Conserved toxic responses across divergent phylogenetic lineages: a meta-analysis of the neurotoxic effects of RDX among multiple species using toxicogenomics. *Ecotoxicology* 20:580–594.

Giani VC, Yamaguchi E, Boyle MJ, Seaver EC. 2011. Somatic and germline expression of *piwi* during development and regeneration in the marine polychaete annelid *Capitella teleta*. *Evodevo* 2:10.

Gong P, Pirooznia M, Guan X, Perkins EJ. 2010. Design, validation and annotation of transcriptome-wide oligonucleotide probes for the oligochaete annelid *Eisenia fetida*. *PLoS One* 5:e14266.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol.* 27:221–224.

Grassa CJ, Kulathinal RJ. 2011. Elevated Evolutionary rates among functionally diverged reproductive genes across deep vertebrate lineages. *Int J Evol Biol.* 2011:274975.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 59:307–321.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C. 2004. The gene ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32:D258.

Hartmann S, Helm C, Nickel B, Meyer M, Struck TH, Tiedemann R, Selbig J, Bleidorn C. 2012. Exploiting gene families for phylogenomic analysis of myzostomid transcriptome data. *PLoS One* 7:e29843.

Haerty W, Jagadeeshan S, Kulathinal RJ, et al. (11 co-authors). 2007. Evolution in the fast lane: rapidly evolving sex-related genes in *Drosophila*. *Genetics* 177:1321–1335.

Heinbockel T, Christensen TA, Hildebrand JG. 2004. Representation of binary pheromone blends by glomerulus-specific olfactory projection neurons. *J Comp Physiol A.* 190:1023–1037.

Hernández P, Fernández R, Novo M, Trigo D, Díaz Cosín DJ. 2007. Geostatistical and multivariate analysis of the horizontal distribution of an earthworm community in El Molar (Madrid, Spain). *Pedobiologia* 51:13–21.

Houck LD. 2009. Pheromone communication in amphibians and reptiles. *Annu Rev Physiol.* 71:161–176.

Howes L, Jones R. 2002. Interactions between zona pellucida glycoproteins and sperm proacrosin/acrosin during fertilization. *J Reprod Immunol.* 53:181–192.

James S, Porco D, Decäens T, Richard B, Rougerie R, Erséus C. 2010. Barcoding reveals cryptic diversity in *Lumbricus terrestris* L., 1758 (Clitellata): resurrection of *L. herculeus* (Savigny, 1826). *PLoS One* 5: e15629.

Jiang XC, Inouchi J, Wang D, Halpern M. 1990. Purification and characterization of a chemoattractant from electric shock-induced earthworm secretion, its receptor binding, and signal transduction through the vomeronasal system of garter snakes. *J Biol Chem.* 265: 8736–8744.

Juliano CE, Wessel GM. 2010. Versatile germline genes. *Science* 329: 640–641.

Kaissling KE. 1996. Peripheral mechanisms of pheromone reception in moths. *Chem Senses.* 21:257–268.

Kang D, Pilon M, Weisblat DA. 2002. Maternal and zygotic expression of a nanos-class gene in the leech *Helobdella robusta*: primordial germ cells arise from segmental mesoderm. *Dev Biol.* 245:28–41.

Kimura T, Yomogida K, Iwai N, Kato Y, Nakano T. 1999. Molecular cloning and genomic organization of mouse homologue of *Drosophila* germ cell-less and its expression in germ lineage cells. *Biochem Biophys Res Commun.* 262:223–230.

King RA, Tibble AL, Symondson WOC. 2008. Opening a can of worms: unprecedented sympatric speciation within British lumbricid earthworms. *Mol Ecol.* 17:4684–4698.

Kodama T, Hisatomi T, Kanemura T, Mokubo K, Tsuboi M. 2003. Molecular cloning and DNA analysis of a gene encoding alpha mating pheromone from the yeast *Saccharomyces naganishii*. *Yeast* 20:109–115.

La Salle S, Palmer C, O'Brien M, Schimenti JC, Epigg J, Handel A. 2011. *Spata22*, a novel vertebrate-specific gene, is required for meiotic progress in mouse germ cells. *Biol Reprod.* 86:45.

Lavelle P, Spain AV. 2001. Soil ecology. London: Kluwer Academic Publishers.

Laverack MS. 1960. Tactile and chemical perception in earthworms. I. Responses to touch, sodium chloride, quinine and sugars. *Comp Biochem Physiol.* 1:155–163.

Leatherman JL, Levin L, Boero J, Jongens TA. 2002. germ cell-less acts to repress transcription during the establishment of the *Drosophila* germ cell lineage. *Curr Biol.* 12:1681–1685.

Lee CE, Frost BW. 2002. Morphological stasis in the *Eurytemora affinis* species complex (Copepoda: Temoridae). *Hydrobiologia* 480: 111–128.

Lee MS, Cho SJ, Tak ES, Lee JA, Cho HJ, Park BJ, Shin C, Kim DK, Park SC. 2005. Transcriptome analysis in the midgut of the earthworm (*Eisenia andrei*) using expressed sequence tags. *Biochem Biophys Res Commun.* 328:1196–1204.

Letunic I, Doerks T, Bork P. 2012. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.* 40: D302–D305.

Li Y, Wang N, Perkins EJ, Zhang C, Gong P. 2010. Identification and optimization of classifier genes multi-class earthworm microarray dataset. *PLoS One* 5:13715.

Lu F, Teal PEA. 2001. Sex Pheromone components in oral secretions and crop of male Caribbean fruit flies, *Anastrepha suspensa* (Loew). *Arch Insect Biochem Physiol.* 48:144–154.

Luporini P, Alimenti C, Ortenzi C, Vallesi A. 2005. Ciliate mating types and their specific protein pheromones. *Acta Protozool.* 44:89–101.

Mohr SE, Dillon ST, Boswell RE. 2001. The RNA-binding protein Tsunagi interacts with Mago Nashi to establish polarity and localize *oskar* mRNA during *Drosophila* oogenesis. *Genes Dev.* 15: 2886–2899.

Mueller JL, Ram KR, McGraw LA, Qazi MCB, Siggia ED, Clark AG, Aquadro CF, Wolfner MF. 2005. Cross-species comparison of *Drosophila* male accessory gland protein genes. *Genetics* 171: 131–143.

Nayernia K, Reim K, Oberwinkler H, Engel W. 1994. Diploid expression and translational regulation of rat *acrosin* gene. *Biochem Biophys Res Commun.* 202:88–93.

Novo M, Almodóvar A, Díaz Cosín DJ. 2009. High genetic divergence of hormogastrid earthworms (Annelida, Oligochaeta) in the central

Iberian Peninsula: evolutionary and demographic implications. *Zool Script.* 38:537–552.

Novo M, Almodóvar A, Fernández R, Giribet G, Díaz Cosín DJ. 2011. Understanding the biogeography of a group of earthworms in the Mediterranean basin—the phylogenetic puzzle of Hormogastridae (Clitellata: Oligochaeta). *Mol Phylogenet Evol.* 61:125–135.

Novo M, Almodóvar A, Fernández R, Trigo D, Díaz Cosín DJ. 2010. Cryptic speciation of hormogastrid earthworms revealed by mitochondrial and nuclear data. *Mol Phylogenet Evol.* 56:507–512.

Novo M, Almodóvar A, Fernández R, Trigo D, Díaz Cosín DJ, Giribet G. 2012. Appearances can be deceptive: different diversification patterns within a group of Mediterranean earthworms (Oligochaeta: Hormogastridae). *Mol Ecol.* 21:3776–3793.

Novo M, Fernández R, Fernández-Marchán D, Gutiérrez M, Díaz Cosín DJ. 2012. Compilation of morphological and molecular data, a necessity for taxonomy: the case of *Hormogaster abbatissae* sp. n. (Annelida, Clitellata, Hormogastridae). *Zookeys* 242:1–16.

Oumi T, Ukena K, Matsushima O, Ikeda T, Fujita T, Minakata H, Nomoto K. 1996. Annetocin, an annelid oxytocin-related peptide, induces egg-laying behaviour in the earthworm *Eisenia foetida*. *Exp Zool.* 276:151–156.

Owen J, Hedley BA, Svendsen C, et al. (12 co-authors). 2008. Transcriptome profiling of developmental and xenobiotic responses in a keystone soil animal, the oligochaete annelid *Lumbricus rubellus*. *BMC Genomics* 9:266.

Oyama A, Shimizu T. 2007. Transient occurrence of vasa-expressing cells in nongenital segments during embryonic development in the oligochaete annelid *Tubifex tubifex*. *Dev Genes Evol.* 217:675–690.

Painter SD, Clough B, Garden RW, Sweedler JV, Nagle GT. 1998. Characterization of *Aplysia* attractin, the first water-borne peptide pheromone in invertebrates. *Biol Bull.* 194:120–131.

Parma DH, Bennett PE Jr, Boswell RE. 2007. Mago Nashi and Tsunagi/Y14, respectively, regulate *Drosophila* germline stem cell differentiation and oocyte specification. *Dev Biol.* 308:507–519.

Pirooznia M, Gong P, Guan X, Inouye LS, Yang K, Perkins EJ, Deng Y. 2007. Cloning, analysis and functional annotation of expressed sequence tags from the earthworm *Eisenia fetida*. *BMC Bioinformatics* 8:S7.

Protasio AV, Tsai IJ, Babbage A, et al. (24 co-authors). 2012. A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. *PLoS Negl Trop Dis.* 6: e1455.

Quail M, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13:341.

Ram JL, Müller CT, Beckmann M, Hardege JD. 1999. The spawning pheromone cysteine-glutathione disulfide ("nereithione") arouses a multicomponent nuptial behavior and electrophysiological activity in *Nereis succinea* males. *FASEB J.* 13:945–952.

Rebscher N, Zelada-González F, Banisch TU, Raible F, Arendt D. 2007. *Vasa* unveils a common origin of germ cells and of somatic stem cells from the posterior growth zone in the polychaete *Platynereis dumerilii*. *Dev Biol.* 306:599–611.

Reich D, Green RE, Kircher M, et al. (28 co-authors). 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468:1053–1060.

Ressler RH, Cialdini RB, Ghoca ML, Kleist SM. 1968. Alarm pheromone in the earthworm *Lumbricus terrestris*. *Science* 161:597–599.

Riesgo A, Andrade SCS, Sharma PP, Novo M, Pérez-Porro AR, Vahtera V, González VL, Kawauchi GY, Giribet G. 2012. Comparative transcriptomics of newly sequenced invertebrates and efficiency estimation of genomic sampling in non-model taxa. *Front Zool.* 9:33.

Roelofs WL, Liu W, Hao G, Jiao H, Rooney AP, Linn CE Jr. 2002. Evolution of moth sex pheromones via ancestral genes. *Proc Natl Acad Sci U S A.* 99:13621–13626.

Rosenkoetter JS, Boice R. 1975. Earthworm pheromones and T-maize performance. *J Comp Physiol Psychol.* 88:904–910.

Rougerie R, Decaëns T, Deharveng L, Porco D, James SW, Chang CH, Richard B, Potapov M, Suhardjono Y, Hebert PDN. 2009. DNA barcodes for soil animal taxonomy. *Pesq Agropec Bras.* 44:789–801.

Saudan P, Hauck K, Soller M, et al. (12 co-authors). 2002. Ductus ejaculatorius peptide 99B (DUP99B), a novel *Drosophila melanogaster* sex-peptide pheromone. *Eur J Biochem.* 269:989–997.

Schultz J, Milpetz F, Bork P, Ponting CP. 1998. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A.* 95:5857–5864.

Schulz MH, Zerbino DR, Vingron M, Birney E. 2012. *Oases*: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 28:1086–1092.

Siebert S, Robinson MD, Tintori SC, Goetz F, Helm RR, Smith SA, Shaner N, Haddock SHD, Dunn CW. 2011. Differential gene expression in the siphonophore *Nanomia bijuga* (Cnidaria) assessed with multiple next-generation sequencing workflows. *PLoS One* 6:e22953.

Smith S, Wilson NG, Goetz F, Feehery C, Andrade SCS, Rouse GW, Giribet G, Dunn CW. 2011. Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature* 480:364–367.

Sorensen PW. 2004. Brief review of fish pheromones and discussion of their possible uses in the control of non-indigenous teleost fishes. *NZ J Mar Freshw Res.* 38:399–417.

Spitzer M, Lorkowski S, Cullen P, Sczyrba A, Fuellen G. 2006. IsoSVM - Distinguishing isoforms and paralogs on the protein level. *BMC Bioinformatics* 7:110.

Sugio M, Takeuchi K, Kutsuna J, Tadokoro R, Takahashi Y, Yoshida-Noro C, Tochinai S. 2008. Exploration of embryonic origins of germline stem cells and neoblasts in *Enchytraeus japonensis* (Oligochaeta, Annelida). *Gene Expr Patterns.* 8:227–236.

Susswein AJ, Nagle GT. 2004. Peptide and protein pheromones in molluscs. *Peptides* 25:1523–1530.

Suzuki T, Honda M, Matsumoto S, Stürzenbaum SR, Gamou S. 2005. Valosine-containing proteins (VCP) in an annelid: identification of a novel spermatogenesis related factor. *Gene* 362:11–18.

Takahashi T, McDougall C, Troscianko J, Chen W, Jayaraman-Nagarajan A, Shimeld SM, Ferrier DEK. 2009. An EST screen from the annelid *Pomatoceros lamarckii* reveals patterns of gene loss and gain in animals. *BMC Evol Biol.* 9:240.

Vacquier VD. 1998. Evolution of gamete recognition proteins. *Science* 281:1995–1998.

Volff J, Zarkower D, Bardwell VJ, Schartl M. 2003. Evolutionary dynamics of the DM domain gene family in metazoans. *J Mol Evol.* 57: S241–S249.

Weiss J, Meeks JJ, Hurley L, Raverot G, Frassetto A, Jameson JL. 2003. *Sox3* Is required for gonadal function, but not sex determination, in males and females. *Mol Cell Biol.* 23:8084–8091.

Wyatt TD. 2003. Pheromones and animal behaviour: communication by smell and taste. Cambridge: Cambridge Press.

Zeeck E, Muller CT, Beckmann M, Hardege JD, Papke U, Sinnvell V, Schroeder FC, Francke W. 1998. Cysteine-gluthaione disulfide, the sperm-release pheromone of the marine polychaete *Nereis succinea* (Annelida: Polychaeta). *Chemoecology* 8:33–38.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18:821–829.