



The First Myriapod Genome Sequence Reveals Conservative Arthropod Gene Content and Genome Organisation in the Centipede *Strigamia maritima*

Ariel D. Chipman^{1†}, David E. K. Ferrier^{2†}, Carlo Brena³, Jiaxin Qu⁴, Daniel S. T. Hughes^{5‡a}, Reinhard Schröder⁶, Montserrat Torres-Oliva^{3‡b}, Nadia Znassi^{3‡c}, Huaiyang Jiang⁴, Francisca C. Almeida^{7,8}, Claudio R. Alonso⁹, Zivkos Apostolou^{3,10}, Peshtewani Aqrawi⁴, Wallace Arthur¹¹, Jennifer C. J. Barna¹², Kerstin P. Blankenburg⁴, Daniela Brites^{13,14}, Salvador Capella-Gutiérrez¹⁵, Marcus Coyle⁴, Peter K. Dearden¹⁶, Louis Du Pasquier¹³, Elizabeth J. Duncan¹⁶, Dieter Ebert¹³, Cornelius Eibner^{11‡d}, Galina Erikson^{17,18}, Peter D. Evans¹⁹, Cassandra G. Extavour²⁰, Liezl Francisco⁴, Toni Gabaldón^{15,21,22}, William J. Gillis²³, Elizabeth A. Goodwin-Horn²⁴, Jack E. Green³, Sam Griffiths-Jones²⁵, Cornelis J. P. Grimmelikhuijzen²⁶, Sai Gubbala⁴, Roderic Guigó^{21,27}, Yi Han⁴, Frank Hauser²⁶, Paul Havlak²⁸, Luke Hayden¹¹, Sophie Helbing²⁹, Michael Holder⁴, Jerome H. L. Hui³⁰, Julia P. Hunn³¹, Vera S. Hunnekuhl³, LaRonda Jackson⁴, Mehwish Javaid⁴, Shalini N. Jhangiani⁴, Francis M. Jiggins³², Tamsin E. Jones²⁰, Tobias S. Kaiser³³, Divya Kalra⁴, Nathan J. Kenny³⁰, Viktoriya Korchina⁴, Christie L. Kovar⁴, F. Bernhard Kraus^{29,34}, François Lapraz³⁵, Sandra L. Lee⁴, Jie Lv²⁸, Christigale Mandapat⁴, Gerard Manning^{17‡e}, Marco Mariotti^{21,27}, Robert Mata⁴, Tittu Mathew⁴, Tobias Neumann^{33,36}, Irene Newsham^{4‡f}, Dinh N. Ngo⁴, Maria Ninova²⁵, Geoffrey Okwuonu⁴, Fiona Onger⁴, William J. Palmer³², Shobha Patil⁴, Pedro Patraquim⁹, Christopher Pham⁴, Ling-Ling Pu⁴, Nicholas H. Putman²⁸, Catherine Rabouille³⁷, Olivia Mendivil Ramos^{2‡g}, Adelaide C. Rhodes³⁸, Helen E. Robertson³⁵, Hugh M. Robertson³⁹, Matthew Ronshaugen²⁵, Julio Rozas⁷, Nehad Saada⁴, Alejandro Sánchez-Gracia⁷, Steven E. Scherer⁴, Andrew M. Schurko²⁴, Kenneth W. Siggins³, DeNard Simmons⁴, Anna Stief^{3,40}, Eckart Stolle²⁹, Maximilian J. Telford³⁵, Kristin Tessmar-Raible^{33,41}, Rebecca Thornton⁴, Maurijn van der Zee⁴², Arndt von Haeseler^{36,43}, James M. Williams²⁴, Judith H. Willis⁴⁴, Yuanqing Wu^{4‡h}, Xiaoyan Zou⁴, Daniel Lawson⁵, Donna M. Muzny⁴, Kim C. Worley⁴, Richard A. Gibbs⁴, Michael Akam³, Stephen Richards^{4*}

1 The Department of Ecology, Evolution and Behavior, The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Givat Ram, Jerusalem, Israel, **2** The Scottish Oceans Institute, Gatty Marine Laboratory, University of St Andrews, St Andrews, Fife, United Kingdom, **3** Department of Zoology, University of Cambridge, Cambridge, United Kingdom, **4** Human Genome Sequencing Center, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, United States of America, **5** EMBL - European Bioinformatics Institute, Hinxton, Cambridgeshire, United Kingdom, **6** Institut für Biowissenschaften, Universität Rostock, Abt. Genetik, Rostock, Germany, **7** Departament de Genètica and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain, **8** Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET), Universidad Nacional de Tucumán, Facultad de Ciencias Naturales e Instituto Miguel Lillo, San Miguel de Tucumán, Argentina, **9** School of Life Sciences, University of Sussex, Brighton, United Kingdom, **10** Institute of Molecular Biology & Biotechnology, Foundation for Research & Technology - Hellas, Heraklion, Crete, Greece, **11** Department of Zoology, National University of Ireland, Galway, Ireland, **12** Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom, **13** Evolutionsbiologie, Zoologisches Institut, Universität Basel, Basel, Switzerland, **14** Swiss Tropical and Public Health Institute, Basel, Switzerland, **15** Centre for Genomic Regulation, Barcelona, Barcelona, Spain, **16** Gravida and Genetics Otago, Biochemistry Department, University of Otago, Dunedin, New Zealand, **17** Razavi Newman Center for Bioinformatics, Salk Institute, La Jolla, California, United States of America, **18** Scripps Translational Science Institute, La Jolla, California, United States of America, **19** The Babraham Institute, Cambridge, United Kingdom, **20** Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, United States of America, **21** Universitat Pompeu Fabra (UPF), Barcelona, Spain, **22** Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain, **23** Department of Biochemistry and Cell Biology, Center for Developmental Genetics, Stony Brook University, Stony Brook, New York, United States of America, **24** Department of Biology, Hendrix College, Conway, Arkansas, United States of America, **25** Faculty of Life Sciences, University of Manchester, Manchester, United Kingdom, **26** Center for Functional and Comparative Insect Genomics, University of Copenhagen, Copenhagen, Denmark, **27** Center for Genomic Regulation, Barcelona, Spain, **28** Department of Ecology and Evolutionary Biology, Rice University, Houston, Texas, United States of America, **29** Institut für Biologie, Martin-Luther-Universität Halle-Wittenberg, Halle, Germany, **30** School of Life Sciences, The Chinese University of Hong Kong, Shatin, NT, Hong Kong SAR, China, **31** Department of Biochemistry and Cell Biology, Faculty of Veterinary Medicine, Utrecht University, Utrecht, The Netherlands, **32** Department of Genetics, University of Cambridge, Cambridge, United Kingdom, **33** Max F. Perutz Laboratories, University of Vienna, Vienna, Austria, **34** Department of Laboratory Medicine, University Hospital Halle (Saale), Halle (Saale), Germany, **35** Department of Genetics, Evolution and Environment, University College London, London, United Kingdom, **36** Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna, Medical University of Vienna, Vienna, Austria, **37** Hubrecht Institute for Developmental Biology and Stem Cell Research, Utrecht, The Netherlands, **38** Harte Research Institute, Texas A&M University Corpus Christi, Corpus Christi, Texas, United States of America, **39** Department of Entomology, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **40** Institute for Biochemistry and Biology, University Potsdam, Potsdam-Golm, Germany, **41** Research Platform "Marine Rhythms of Life", Vienna, Austria, **42** Institute of Biology, Leiden University, Leiden, The Netherlands, **43** Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, Austria, **44** Department of Cellular Biology, University of Georgia, Athens, Georgia, United States of America

Abstract

Myriapods (e.g., centipedes and millipedes) display a simple homonomous body plan relative to other arthropods. All members of the class are terrestrial, but they attained terrestriality independently of insects. Myriapoda is the only arthropod class not represented by a sequenced genome. We present an analysis of the genome of the centipede *Strigamia maritima*. It retains a compact genome that has undergone less gene loss and shuffling than previously sequenced arthropods, and many orthologues of genes conserved from the bilaterian ancestor that have been lost in insects. Our analysis locates many genes in conserved macro-synteny contexts, and many small-scale examples of gene clustering. We describe several examples where *S. maritima* shows different solutions from insects to similar problems. The insect olfactory receptor gene family is absent from *S. maritima*, and olfaction in air is likely effected by expansion of other receptor gene families. For some genes *S. maritima* has evolved paralogues to generate coding sequence diversity, where insects use alternate splicing. This is most striking for the *Dscam* gene, which in *Drosophila* generates more than 100,000 alternate splice forms, but in *S. maritima* is encoded by over 100 paralogues. We see an intriguing linkage between the absence of any known photosensory proteins in a blind organism and the additional absence of canonical circadian clock genes. The phylogenetic position of myriapods allows us to identify where in arthropod phylogeny several particular molecular mechanisms and traits emerged. For example, we conclude that juvenile hormone signalling evolved with the emergence of the exoskeleton in the arthropods and that RR-1 containing cuticle proteins evolved in the lineage leading to Mandibulata. We also identify when various gene expansions and losses occurred. The genome of *S. maritima* offers us a unique glimpse into the ancestral arthropod genome, while also displaying many adaptations to its specific life history.

Citation: Chipman AD, Ferrier DEK, Brena C, Qu J, Hughes DST, et al. (2014) The First Myriapod Genome Sequence Reveals Conservative Arthropod Gene Content and Genome Organisation in the Centipede *Strigamia maritima*. PLoS Biol 12(11): e1002005. doi:10.1371/journal.pbio.1002005

Academic Editor: Chris Tyler-Smith, The Wellcome Trust Sanger Institute, United Kingdom

Received: February 21, 2014; **Accepted:** October 15, 2014; **Published:** November 25, 2014

Copyright: © 2014 Chipman et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the following grants: NHGRI U54 HG003273 to R.A.G.; EU Marie Curie ITN #215781 “Evonet” to M.A.; a Wellcome Trust Value in People (VIP) award to C.B., a Wellcome Trust graduate studentship WT089615MA to J.E.G., and a Wellcome Trust Investigator Award (098410/Z/12/Z) to C.R.A.; “Marine Rhythms of Life” of the University of Vienna, an FWF (<http://www.fwf.ac.at/>) START award (#AY0041321) and HFSP (<http://www.hfsp.org/>) research grant (#RGY0082/2010) to K.T.-R.; MFPL Vienna International PostDoctoral Program for Molecular Life Sciences (funded by Austrian Ministry of Science and Research and City of Vienna, Cultural Department - Science and Research) to T.K.; Direct Grant (4053034) of the Chinese University of Hong Kong to J.H.L.H.; NHGRI HG004164 to G.M.; Danish Research Agency (FNU), Carlsberg Foundation, and Lundbeck Foundation to C.J.P.G.; U.S. National Institutes of Health R01AI55624 to J.H.W.; Royal Society University Research fellowship to F.M.J.; P.D.E. was supported by the BBSRC via the Babraham Institute. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: stephenr@bcm.edu

¶ ADC and DEKF are joint senior authors on this work.

¶a Current address: Human Genome Sequencing Center, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, United States of America

¶b Current address: Department of Developmental Biology, Georg-August-Universität Göttingen, Johann-Friedrich-Blumenbach-Institut für Zoologie und Anthropologie, Abteilung Entwicklungsbiologie, GZMB, Göttingen, Germany

¶c Current address: Centre de Recherche de Biochimie Macromoléculaire, Bioinformatique Structurale et Modélisation Moléculaire, Montpellier, France

¶d Current address: Department of Genetics, Friedrich Schiller University, Germany

¶e Current address: Genentech, Inc., South San Francisco, California, United States of America

¶f Current address: M.D. Anderson Cancer Center, Houston, Texas, United States of America

¶g Current address: Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, New York, United States of America

¶h Current address: Institute for Applied Cancer Science, MD Anderson Cancer Center, Houston, Texas, United States of America

Abbreviations: FGF, fibroblast growth factor; GR, gustatory receptor; GPCR, G protein-coupled receptor; JH, juvenile hormone; OR, odorant receptor; RR, Rebers and Riddiford; TGF, transforming growth factor.

Introduction

Arthropods are the most species-rich animal phylum on Earth. Of the four extant classes of arthropods (Insecta, Crustacea, Myriapoda, and Chelicerata) (Figure 1), only the Myriapoda (centipedes, millipedes, and their relatives) are currently not represented by any sequenced genome [1,2]. This absence is particularly unfortunate, as myriapods have recently been recognised as the living sister group to the clade that encompasses all insects and crustaceans [3–6]. Hence, the Myriapoda are particularly well placed to provide an outgroup for comparison, to determine ancestral character states and the polarity of evolutionary change within insects and crustaceans, which together represent the most diverse animal clade on Earth.

Although *Drosophila melanogaster* is the best studied arthropod, it lacks many genes present in the ancestral bilaterian gene set, and

chromosome rearrangements have disrupted all obvious evidence of synteny with other phyla [7]. Thus it is not fully representative of other arthropods. More comprehensive sampling of arthropod genomes will establish their basic structure, and determine when unique genomic characteristics of different taxa, such as the holometabolous insects, appear.

Phylogenetic Position of the Myriapods

Myriapods are today represented by two major lineages—the herbivorous millipedes (Diplopoda) and the carnivorous centipedes (Chilopoda), together with two minor clades, the Symphyla, which look superficially like small white centipedes, and the minute Pauropoda [8]. All are characterised by a multi-segmented trunk of rather similar (homonomous) segments, with no differentiation into thorax or abdomen. All recent studies, molecular and

Author Summary

Arthropods are the most abundant animals on earth. Among them, insects clearly dominate on land, whereas crustaceans hold the title for the most diverse invertebrates in the oceans. Much is known about the biology of these groups, not least because of genomic studies of the fruit fly *Drosophila*, the water flea *Daphnia*, and other species used in research. Here we report the first genome sequence from a species belonging to a lineage that has previously received very little attention—the myriapods. Myriapods were among the first arthropods to invade the land over 400 million years ago, and survive today as the herbivorous millipedes and venomous centipedes, one of which—*Strigamia maritima*—we have sequenced here. We find that the genome of this centipede retains more characteristics of the presumed arthropod ancestor than other sequenced insect genomes. The genome provides access to many aspects of myriapod biology that have not been studied before, suggesting, for example, that they have diversified receptors for smell that are quite different from those used by insects. In addition, it shows specific consequences of the largely subterranean life of this particular species, which seems to have lost the genes for all known light-sensing molecules, even though it still avoids light.

morphological, support the monophyly of myriapods [3–5,8–10] suggesting that they share a single common ancestor.

Myriapods, insects, and crustaceans have traditionally been identified as a clade of mandibulate arthropods, characterised by head appendages that include antennae and biting jaws [11]. Some molecular datasets have challenged this idea, suggesting instead that the myriapods are a sister group to the chelicerates [12,13]. The most comprehensive phylogenomic datasets thus far reject this, and strongly support the phylogeny that proposes that the chelicerates are the most basal of the four major extant arthropod clades, and the mandibulates represent a true monophyletic group [3,5,10,14–17].

Within the mandibulates, myriapods were believed until recently to share a common origin with insects as terrestrial arthropods. This view, based on a number of shared characters including uniramous limbs, air breathing through tracheae, the lack of a second pair of antennae, and excretion using Malpighian tubules, was widely supported by morphologically based phylogenies [9,18]. However, molecular phylogenies robustly reject the sister group relationship between insects and myriapods, placing the origin of myriapods basal to the diversification of crustaceans [5], and identifying insects as a derived clade within the Crustacea [19–21]. As crustaceans are overwhelmingly a marine group today, and were so ancestrally, this implies that myriapods and insects represent independent invasions of the land (with the chelicerates representing an additional, unrelated invasion). Their shared characteristics are striking convergences, not synapomorphies.

S. maritima as a Model Myriapod

We chose *S. maritima* as the species to sequence partly for pragmatic reasons: geophilomorph centipedes, such as *S. maritima*, have relatively small genome sizes, certainly compared to other centipedes [22]. More importantly, it is a species that has attracted interest for ecological and developmental studies [23–25], especially the process of segment patterning [26–32]. *S. maritima* is a common centipede of north western Europe, found along the coastline from France to the middle of Norway. It is a specialist of shingle beaches and rocky shores, occurring around the high tide mark, and feeding on the abundant crustaceans and insect larvae associated with the strand line. It is by far the most abundant centipede in these habitats around the British Isles, sometimes occurring at densities of thousands per square metre in suitable locations [25]. Eggs can be harvested from these abundant populations in large numbers with relatively little effort during the summer breeding season [27]. They can be reared in the lab from egg lay to at least the first free-living stage, adolescens I [24,33].

Some aspects of *S. maritima* biology are not common to all centipedes. Notable among these is epimorphic development, wherein the embryos hatch from the egg with the final adult number of leg-bearing segments. Epimorphic development is

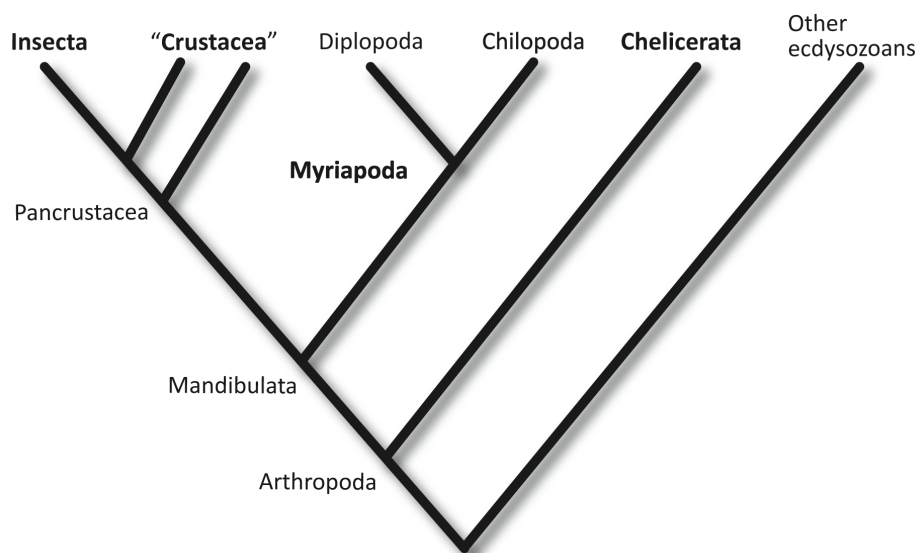


Figure 1. The phylogenetic position of the centipedes (Chilopoda), with respect to other arthropods, according to the currently best-supported phylogeny. (See text for details). The four traditionally accepted arthropod classes are marked in bold.

doi:10.1371/journal.pbio.1002005.g001

found in two centipede orders: geophilomorphs (including *S. maritima*) and scolopendromorphs. In contrast, more basal clades display anamorphic development and add segments post-embryonically [34]. These anamorphic clades have relatively few leg-bearing segments, generally 15, while geophilomorphs have many more, up to nearly 200 in some species [6]. These unique characteristics probably arose at least 300 million years ago, as the earliest fossils of the much larger scolopendromorph centipedes date to the Upper Carboniferous [35]. These share the same mode of development as the geophilomorphs, and are their likely sister group. Geophilomorphs are also adapted to a subsurface life style, the whole order having lost all trace of eyes [36,37], though apparently not photosensitivity [38].

We have sequenced the genome of *S. maritima* as a representative of the phylogenetically important myriapods. In contrast to the intensively sampled holometabolous insects, our analysis of this myriapod genome finds conservative gene sets and conserved synteny, shedding light on general genomic features of the arthropods.

Results and Discussion

Genome Assembly, Gene Densities, and Polymorphism

Genomic DNA from multiple individuals of a wild Scottish population of *S. maritima* was sequenced and assembled into a draft genome sequence spanning 176.2 Mb. This assembled sequence omits many repeat sequences including heterochromatin, which probably accounts for the difference between the assembly length and the total genome size estimate of 290 Mb. An analysis of repetitive elements within the assembly is presented in Text S1.

The assembly incorporates 14,992 automatically generated gene models, 1,095 of which have been additionally manually annotated. We re-sequenced four individuals comprising three females and one male. The frequency of identified polymorphism, with SNP density of 4.5 variants/kb, is comparable with the five variants per kb in the *Drosophila* genetic reference panel [39]. It is hard to say how typical this is for soil dwelling arthropods, as very little population data are available for such species.

Phylome Analysis and Phylogenomics

To understand general patterns of gene evolution in *S. maritima* we reconstructed the evolutionary histories of all of its genes, i.e., the phylome. The resulting gene phylogenies, available through phylomeDB [40], were analysed to establish orthology and paralogy relationships with other arthropod genomes [41], transfer functional knowledge from annotated orthologues, and to detect and date gene duplication events [42]. Some 32% of *S. maritima* genes can be traced back to duplications specific to this myriapod lineage since its divergence from other arthropod groups included in the analysis. Functions enriched among these genes include those related to, among other processes, catabolism of peptidoglycans, sodium transport, glutamate receptor, and sensory perception of taste. Related to this latter function, two of the largest gene expansions specific to the *S. maritima* lineage detected in our analysis are the gustatory receptor (GR) and ionotropic receptor (IR) families encoding putative membrane-associated gustatory and/or olfactory receptors (see Text S1, and Chemosensory section below).

Sex Chromosomes

No obviously differentiated sex chromosomes are apparent in the diploid *S. maritima* karyotype, which comprises one long pair of metacentric chromosomes, together with seven pairs of much

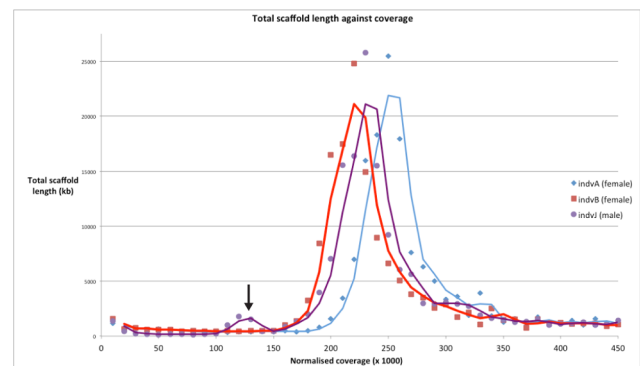


Figure 2. Plot showing that DNA from a male individual contains a distinct fraction of scaffolds that is underrepresented (black arrow), and presumably derives from heterogametic sex chromosomes. No such fraction is present in the sequenced DNA of two individual females. The data underlying this plot is presented in File S4.

doi:10.1371/journal.pbio.1002005.g002

shorter telocentric chromosomes (P. Woznicki, unpublished data; J. Green et al., unpublished). Read-depth data from the genome assembly show that a proportion of the genome is underrepresented compared to the bulk of the data. One obvious reason for underrepresentation would be sequences derived from sex chromosomes. To confirm this, the coverage of individual scaffolds from the assembly was examined in sequence obtained from single individuals. A distinct fraction of underrepresented scaffolds is present in DNA derived from a male, but absent in female sequence (Figure 2), implying an XY sex determination mechanism. Quantitative PCR from three scaffolds in the underrepresented fraction confirmed that they are present at approximately twice the copy number in females as in males, identifying them as X chromosome derived (J. Green et al., unpublished). Other scaffolds of this fraction contain male specific sequences, and therefore presumably derive from a Y chromosome (J. Green et al., unpublished) [31]. Combined with the karyotype data, this finding suggests that *S. maritima* possesses a weakly differentiated pair of X and Y chromosomes.

Mitochondrial Genome

From the whole genome assembly, *S. maritima* scaffold scf7180001247661 was found to contain a complete copy of the mitochondrial coding regions, flanked by a TY1/Copia-like retrotransposon, which all together spanned approximately 20 kb. This is unusually large for a metazoan mitochondrial genome and, as mis-assembly was suspected, PCR was used to clone the DNA between the genes at either end of the scaffold. This enabled us to close the circle of the mitochondrial genome, correct frameshifts, and confirm an unusual gene arrangement, resulting in a final circular assembly of 14,983 bp (Table S11). The gene arrangement in the *S. maritima* mitochondrial genome is striking (Figure S6). It diverges dramatically from the basic arthropod genome arrangement and differs from all other known centipede mitochondrial gene arrangements [43]. Although small sections of the *S. maritima* gene order are conserved with respect to the arthropod ground pattern found in *Limulus polyphemus* and the lithobiomorph centipede *Lithobius forficatus* (e.g., trnA-F-nad5-H-nad4-nad4L on the minus strand), other sections are completely rearranged to an extent unusual in arthropods, and metazoans (ACR and MJT, unpublished). This confounds attempts to use *S. maritima* mitochondrial gene order in phylogenetic reconstructions.

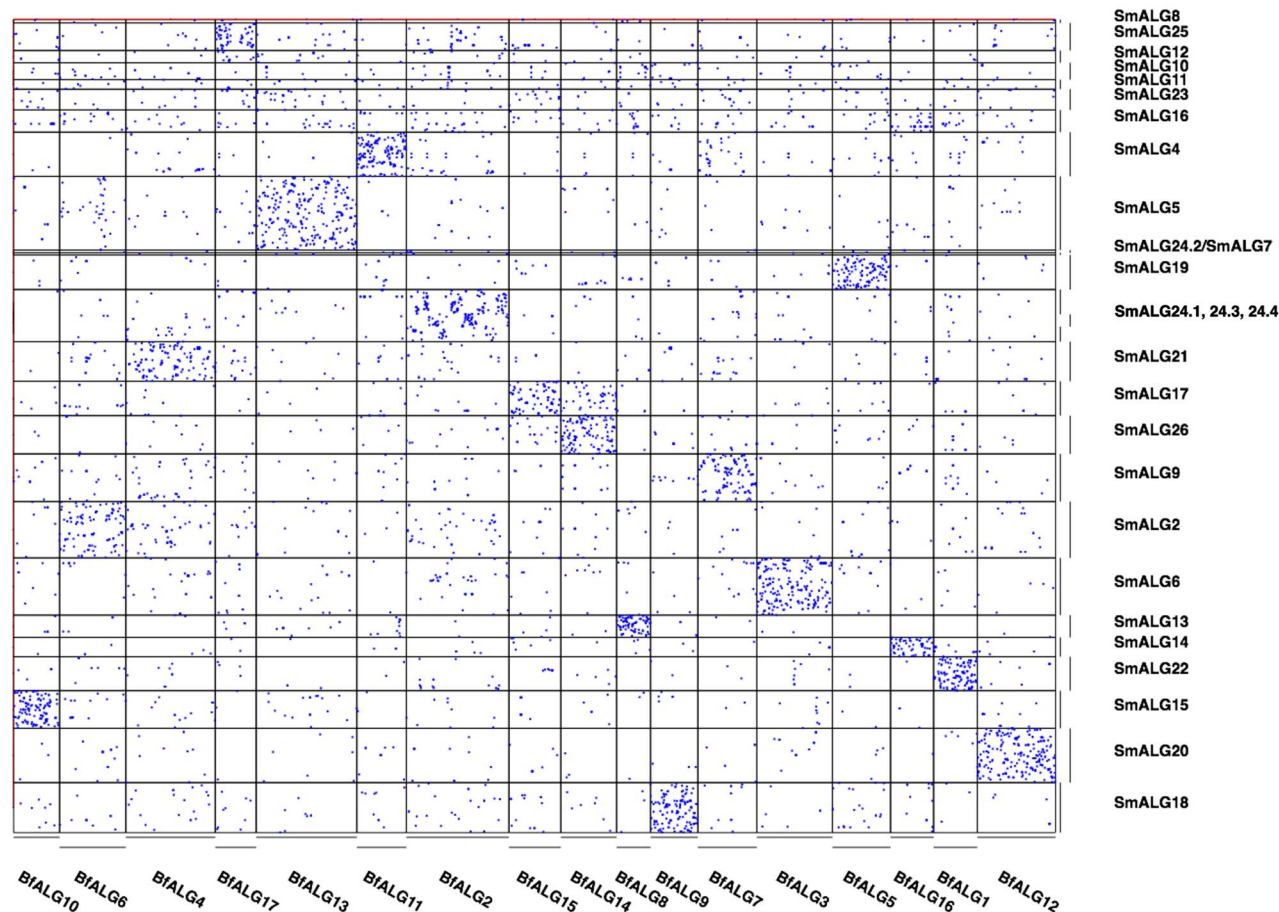


Figure 3. Conserved macro synteny signal between *S. maritima* and the chordate lancelet *B. floridae* clustered into ancestral linkage groups. Each dot represents a pair of genes, one in *B. floridae*, one in *S. maritima*, assigned to the same gene family by our orthology analysis. The ancestral linkage group identifiers refer to groups of scaffolds from the *S. maritima* (SmALG) or *B. floridae* (BfALG) assemblies, as detailed in File S2. The identification of ALGs is described in the SI. Note that two *S. maritima* scaffolds were divided across ALGs, and so appear multiple times in File S2. doi:10.1371/journal.pbio.1002005.g003

Conserved Synteny with Other Phyla

With the exception of some conserved local gene clusters, the location of genes on the chromosomes of *Drosophila* and other Diptera retains no obvious trace of the ancestral bilaterian gene linkage. Other holometabolous insects such as *Bombyx mori* and *Tribolium castaneum* do show significant conservation of large-scale gene linkage with other phyla, for example, in the chordate *Branchiostoma floridae* (amphioxus) and the cnidarian *Nematostella vectensis* [44,45]. The last common ancestor of these two lineages pre-dated the ancestor of all bilaterian animals, and yet the genomes of these species retain detectable conserved synteny: orthologous genes are found together on the same chromosomes, or chromosome fragments, far more often than would be expected by chance.

We find the *S. maritima* genome also retains significant traces of the large-scale genome organisation that was present in the bilaterian ancestor. Although the assignment of scaffolds to chromosomes is not determined in *S. maritima*, there are sufficient gene linkage data within scaffolds to reveal clear retained synteny between amphioxus and *S. maritima* (Figure 3), at a higher level than any of the Insecta or Pancrustacea we have examined.

Of the 62 scaffolds with at least 20 genes from ancestral bilaterian orthology groups, 37 show enrichment of shared orthologues with one or (in the case of a single scaffold) two

chordate ancestral linkage groups (ALGs) at a significance threshold of $p < 0.0001$ (after Bonferroni correction for 1,116 pairwise ALG-scaffold comparisons). Of these scaffolds' genes that have predicted human orthologues, 57% are found in a conserved macro-synteny context. At a more relaxed significance threshold ($p < 0.01$), 71% of these scaffolds have a significant association with at least one chordate ALG, and 17 of the 18 chordate ALGs hit at least one of these scaffolds.

Stronger synteny is also detected for the genome of the nematode *Caenorhabditis elegans* with *S. maritima* than with insects or other Metazoa. The *C. elegans* genome is highly rearranged, and shows low synteny with higher insects, or with chordates [7,46,47]. As members of the Ecdysozoa, nematodes last shared a common ancestor with the arthropods more recently than with chordates. This shared ancestry allows traces of conserved genome organisation to be detected with slowly rearranging arthropod genomes, even when it is only weakly apparent with chordates.

By implication, the last common ancestor of the arthropods retained significant synteny with the last common ancestor of bilaterians as well as the last common ancestors of other phyla, such as the Chordata. This conserved synteny is more complete with this *S. maritima* genome sequence, due to the relative scrambling of the genomes of those other arthropods that have been sequenced previously.

Homeobox Gene Clusters: Hox, ParaHox, SuperHox, and Mega-homeobox

The clustering of genes in a genome is often of functional significance (e.g., reflecting co-regulation), as well as providing important insights into the origins of particular gene families when clusters are composed of genes from the same class or family. Gene clusters can also be a useful proxy for the degree of genome rearrangement. The homeobox gene super-class is one type of gene for which clustering has been extensively explored. *S. maritima* has 113 homeobox-containing genes, which is slightly more than seen in other sequenced arthropods such as *D. melanogaster*, *T. castaneum*, and *Apis mellifera*. This is due to some lineage-specific duplications in *S. maritima* as well as the retention of some homeobox families that have been lost in other arthropods, including *Vax*, *Dmbx*, and *Hmbox* (see Text S1).

The homeobox-containing genes of the Hox gene cluster are renowned for their role in patterning the anterior-posterior axis of animal embryos. *S. maritima* has an intact, well-ordered Hox cluster containing one orthologue of each of the ten expected arthropod Hox genes, except for Hox3. There are two potential Hox3 genes elsewhere in the *S. maritima* genome [48], but the true orthology of these genes remains slightly ambiguous; it remains possible that they are the first example of ecdysozoan Xlox ParaHox genes (see Text S1). The Hox cluster spans 457 kb (*labial* to *eve*), a span similar to assembled Hox clusters in a range of other invertebrate groups (crustacean, mollusc, echinoderm, cephalochordate). This suggests that the contrasting very large (and frequently broken) Hox clusters of Drosophilids and some other insects are a derived characteristic. However, the spectrum of alternatively spliced and polyadenylated transcripts encoded by the Hox genes of *S. maritima* is comparable with what is known from *D. melanogaster* (details in Text S1). Exceptionally among protostomes, the *S. maritima* Hox cluster retains tight linkage to one orthologue of *evx/evenskipped*, as it does in some chordates and cnidarians.

Further instances of homeobox gene clustering and linkage, and reconstructions of ancestral states, are summarized in Figure 4 and Table 1 (and see Text S1). The Hox gene cluster is hypothesized to have evolved within the context of a Mega-homeobox cluster that existed before the origin of the bilaterians and consisted of an array of many ANTP-class genes [49–51]. By the time of the last common ancestor of bilaterians the Hox cluster existed within the context of a SuperHox cluster, containing the Hox genes themselves and at least eight further ANTP-class genes [52]. The conservative nature of the *S. maritima* genome has left several fragments from the Mega-homeobox and SuperHox clusters still intact (Figure 4; Table 1). Furthermore, homeobox linkages in *S. maritima* raise the possibility that further genes could have been members of the Mega-homeobox and SuperHox clusters, including the ANTP-class gene *Vax*, as well as the SINE-class gene *sine oculis* and the HNF-class gene *Hmbox* (see Text S1 for further details).

Chemosensory Gene Families (Gustatory Receptors, Ionotropic Receptors, Odorant Binding Proteins, Chemosensory Proteins)

The chemosensory system of arthropods is best known in insects. During the evolutionary transition from water to terrestrial environments, insects evolved a new set of genes to detect airborne molecules (odorants) [53–55]. The independent colonization of land by insects and myriapods raises two interesting questions: (1) what are the genes involved in chemosensation in non-insect arthropods, and (2) what genes are responsible for the detection of

airborne molecules in other terrestrial arthropods? We searched the *S. maritima* genome for homologues of the insect chemosensory genes, included in six gene families, three ligand binding protein families: odorant binding proteins (OBPs) [56,57], chemosensory proteins (CSPs) [58,59], and CheA/B [60,61]; and three membrane receptor families: GRs [62,63], odorant receptors (ORs) [64,65], and IRs [66,67].

Of the ligand binding proteins, we found only two genes belonging to the CSP family, but no representatives of the OBP or CheA/B families. Among the membrane receptor families, we identified a number of genes of both the GR and IR families, but no OR genes. The GR family in *S. maritima* is represented by 77 genes, 17 of which seem to be pseudogenes, with similar numbers of genes and pseudogenes being fairly typical features of this gene family in other arthropods. A phylogenetic tree revealed that none of the *S. maritima* GR genes have 1:1 orthology to other arthropod GRs. Instead, all *S. maritima* GRs cluster in a single clade, with six major subclades, representing separate expansions of the GR repertoire in the centipede lineage (Figure 5A and see Text S1). The IR family is known to be ancient [67], but *S. maritima* has a relative expansion of this family. The search for IRs led to the annotation of 69 genes, 15 of which belong to the IGluR subfamily, which is not involved in chemosensation, but is highly conserved among arthropods and animals in general. Among the remaining 54 IRs, three are orthologues of conserved IR genes that have been shown to have an olfactory function in *D. melanogaster*. However, 51 of the *S. maritima* IRs do not have orthologues either in *D. melanogaster* or in *Ixodes scapularis*, clustering together in a single clade (the expansion clade in Figure 5B). This finding suggests that most *S. maritima* IRs, as observed with GRs, have duplicated from a common ancestral gene exclusive to the centipede lineage.

The absence of the insect OR family agrees with the prediction of Robertson and colleagues [54] that this lineage of the insect chemoreceptor superfamily evolved with terrestriality in insects, and it is also missing from the water flea *Daphnia pulex* [53]. The same appears to be true for the OBPs. We therefore infer that, as centipedes adapted to terrestriality independently from the hexapods, they utilized a novel combination of expanded GR and IR protein families for olfaction, in addition to their more ancestral roles in gustation.

Light Receptors and Circadian Clock Genes

S. maritima, like all species of the order Geophilomorpha, is blind [37]. Nevertheless, it avoids open spaces and negative phototaxis has been demonstrated in other species of Geophilomorpha [38,68]. We searched the *S. maritima* genome for light receptor genes. Interestingly, we have found no opsin genes, no homologue of gustatory receptor 28b (GR28b), which is involved in larval light avoidance behaviour in *Drosophila* [69], and no cryptochromes. Thus, none of the known arthropod light receptors are present. Furthermore, there are no photolyases, which would repair UV light induced DNA damage. As a consequence, the critical avoidance of open spaces by *S. maritima* must either be mediated by other sensory instances than light perception, or *S. maritima* possesses yet unknown light receptor molecules.

The absence of light receptors, particularly cryptochromes, also raises the issue of the entrainment and composition of a potential *S. maritima* circadian clock. Strikingly, we could not identify any components of the major regulatory feedback loop of the canonical arthropod circadian clock (including *period*, *cycle*, *b-mal/clock*, *timeless*, *cryptochromes 1 and 2*, *jetlag* [70]). The only circadian clock genes found (*timeout*, *vrille*, *pdp1*, *clockwork orange*) are generally known to be involved in other physiological

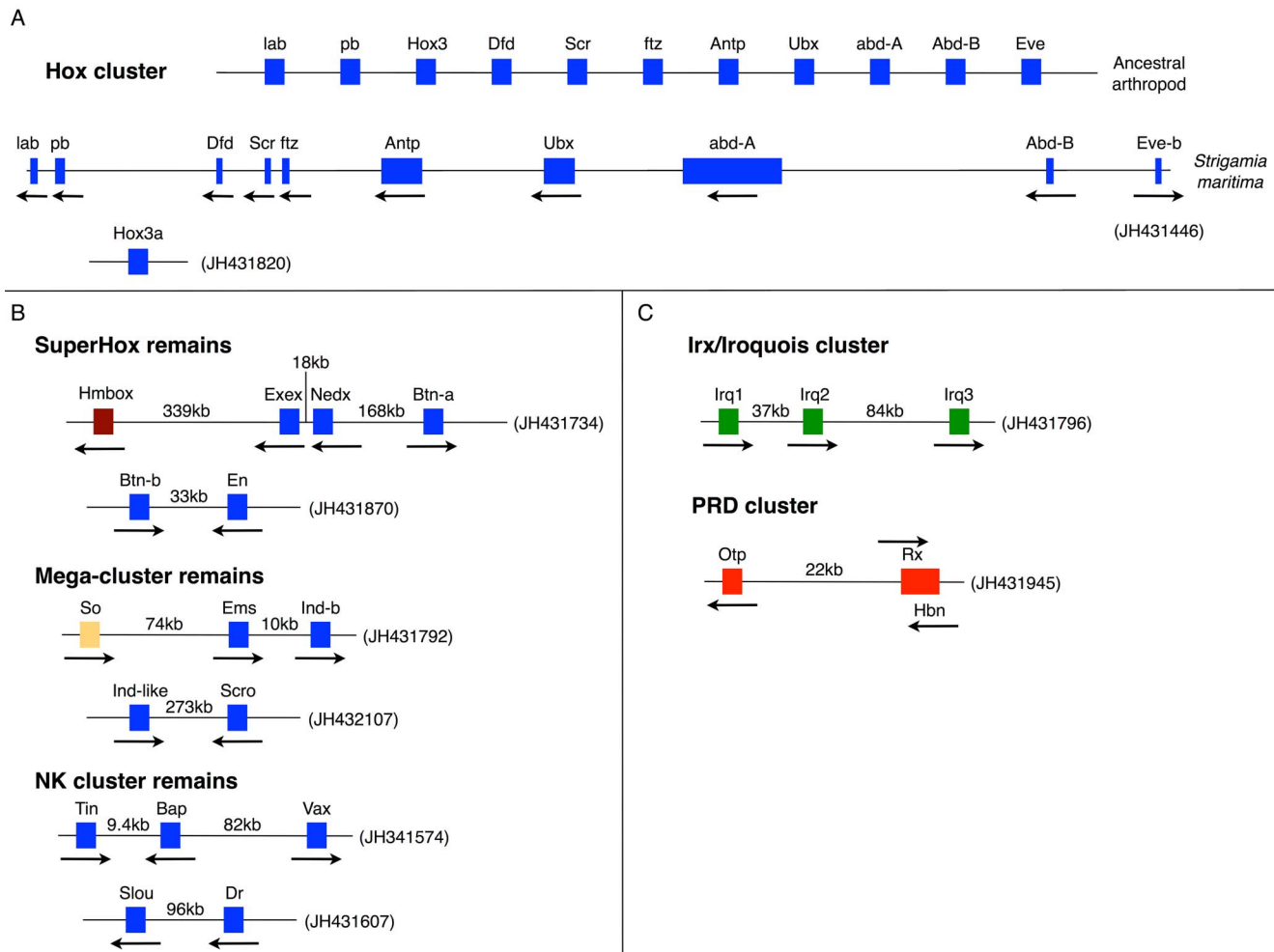


Figure 4. Homeobox gene clusters. (A) The Hox gene cluster of *S. maritima* compared to the cluster that can be deduced for the ancestral arthropod. *S. maritima* provides the first instance of an arthropod Hox cluster with tight linkage of an *Even-skipped* (*Eve*) gene (see text). Hox3 is the only gene missing from the *S. maritima* Hox cluster, but may be present elsewhere in the genome on a separate scaffold (see main text and Text S1 for details). The *S. maritima* cluster is drawn approximately to scale and spans 457 kb from the start codon of *labial* (*lab*) to the start codon of *Eve-b*. Arrows denote the transcriptional orientation. (B) Remains of clustering and linkage of ANTP class genes in *S. maritima*. The blue boxes are genes belonging to the ANTP class. The brown box is a gene belonging to the HNF class. The orange box is a gene belonging to the SINE class. The intergenic distances are indicated in kb. (C) Clusters of non-ANTP class homeobox genes in *S. maritima*. The green boxes are genes belonging to the TALE class. The red boxes are genes belonging to the PRD class. The intergenic distances are indicated in kb, except in the case of Rx-Hbn as these genes are overlapping but with opposite transcriptional orientations. All scaffold numbers are indicated in brackets.
doi:10.1371/journal.pbio.1002005.g004

processes as well [71–73]. The extensive secondary gene loss of both light receptors and circadian clock genes raises questions about the actual existence of a circadian clock in *S. maritima*. One could hypothesize that a circadian clock may not be required in *S. maritima*'s subsurface habitat, although other periodicities, such as tide cycles, might be important. If *S. maritima* does have a circadian clock then it must be operating via a mechanism distinct from the canonical arthropod system.

Other blind or subterranean animals do maintain a circadian rhythm, despite complete loss of vision and connection with the surface (e.g., *Spalax*) [74–76]. In other cases (e.g., blind cave crayfish [77]), despite the loss of vision, opsin proteins remain functional, and are hypothesized to have a role in circadian cycles. However, both these examples represent species that have become blind and subterranean relatively recently. To confirm that the loss of these genes is not general for all centipedes, we performed BLASTP analyses searching for the set of light sensing and circadian clock genes that are missing from *S. maritima* in

RNAseq data from the house centipede *Scutigera coleoptrata* (NCBI SRA accession SRR1158078), a species with well-developed eyes. We find homologs to period, cycle, b-mal/clock, jetlag, cryptochrome1, cryptochrome 2, (6-4)-photolyase, and nina-e (rhodopsin 1), suggesting that both light sensing and circadian clock systems were present in ancestor of myriapods. Although we have no direct information about photoreceptors or circadian genes in other geophilomorph species, the fact that all geophilomorphs are blind suggests that the loss of the related genes is very ancient, and may date back to the origin of the clade.

Putative Cuticular Proteins

A defining characteristic of arthropods is an exoskeleton with chitin and cuticular proteins as the primary components. Although several families of cuticular proteins have been recognized, the CPR family (Cuticular Proteins with the Rebers and Riddiford consensus) is by far the largest in every arthropod for which a complete genome is available, with 32 to >150 members [78].

Table 1. Instances of homeobox gene clustering and linkage.

Gene Cluster	Details	Conclusion or Hypothesis
Hox Cluster	Intact well ordered, but lacking <i>Hox3</i> (Figure 4A). Two potential <i>Hox3</i> genes elsewhere in the genome, but these could also be <i>Xlox</i> homologues	Has <i>Xlox</i> really been lost from all lineages of the ecdysozoan super phylum?
NK - <i>Vax</i> linkage	Centipede has gene pair remnants from the ancestral NK cluster <i>slouch</i> and <i>drop</i> , and <i>tinman</i> and <i>bagpipe</i> (now with <i>Vax</i> linkage, which also seen in mollusc) (Figure 4B)	<i>Vax</i> linkage likely ancestral, <i>Vax</i> a new member of the ancestral ANTP class mega-homeobox cluster.
IRX/Iroquois	Cluster of three <i>lrx</i> genes (Figure 4C)	Independent expansion from <i>Drosophila</i> by duplication of <i>mirror</i> .
<i>Orthopedia</i> , <i>Rax</i> , and <i>Homeobrain</i>	Cluster present in <i>S. maritima</i> (Figure 4C)	An ancestral cluster also found in insects, cnidarians, and molluscs.
SuperHox cluster remains	Linkage of <i>BtnN</i> and <i>En</i> on Scaffold JH431870. Linkage of <i>Exex-Nedx-BtnA</i> on scaffold JH431734 (Figure 4B) with <i>Hmbox</i> .	Remnants of the Super-Hox cluster?
ParaHox - NK linkage (Mega-cluster remains)	Tight linkage of <i>Ems</i> (NK gene) with <i>IndB</i> (ParaHox gene), and <i>Ind-like</i> (ParaHox like) with <i>scro</i> (NK gene) (Figure 4B)	Possible remnant of ParaHox and NK clusters from ancestral Mega-Cluster ^a
SINE-ANTP class linkage	linkage of <i>sine oculis</i> & <i>Ems</i>	Also seen in humans and zebrafish - thus linkage of SINE and ANTP genes in bilaterian ancestor

Further details are provided in Text S1.

^aNote these have become secondarily linked in vertebrates [50].

doi:10.1371/journal.pbio.1002005.t001

Proteins in the CPR family have a consensus region in arthropods of about 28 amino acids, first recognized by Rebers and Riddiford [79], which was subsequently extended to ~64 amino acid residues and shown to be necessary and sufficient for binding to chitin [80]. No clear instances of the Rebers and Riddiford (RR) consensus have been identified outside the arthropods. We identified 38 members of the CPR family in *S. maritima*. There are two main forms of the consensus, designated RR-1 and RR-2, with the former primarily associated with flexible cuticle, the latter with rigid cuticle. Interestingly, while chelicerates studied to date have no members of the RR-1 subfamily (as classified at CutProtFam-Pred, <http://aias.biol.uoa.gr/CutProtFam-Pred/home.php>), seven of the *S. maritima* CPR proteins clearly belong to this class. This would be consistent with the origin of the RR1-coding genes being in the mandibulate ancestor after this lineage had diverged from the chelicerate lineage. Further data are needed to verify that the identified proteins are indeed important constituents of the cuticle.

Neuro-endocrine Hormone Signalling

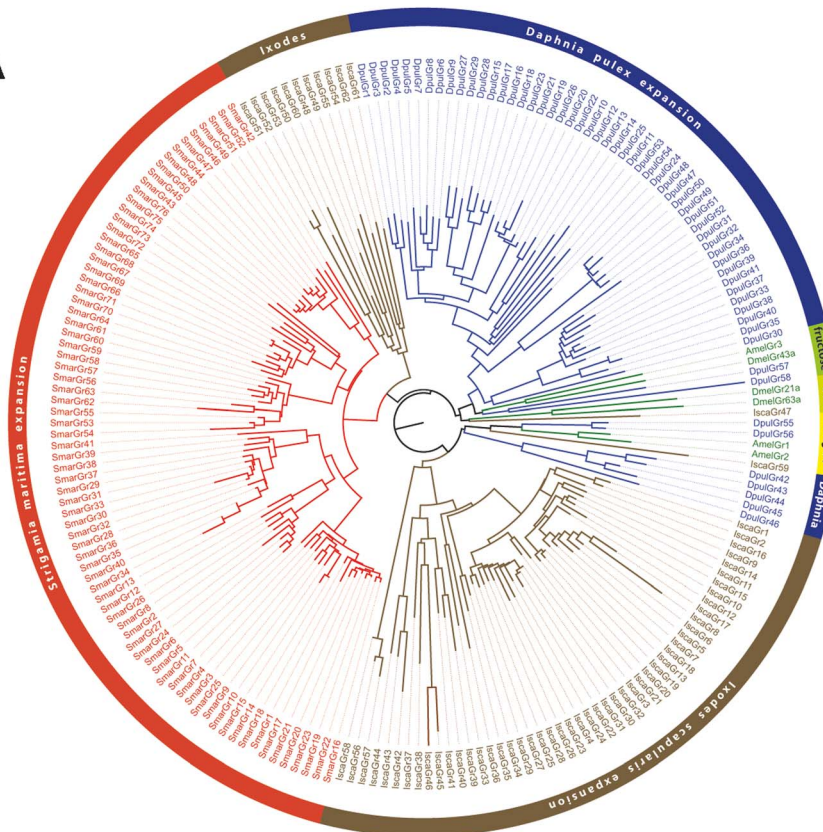
Cell-to-cell communication in arthropods occurs via a variety of neurotransmitters and neuro-endocrine hormones, including biogenic amines, neuropeptides, protein hormones, juvenile hormone (JH), and ecdysone. These signalling molecules and their receptors steer central processes such as growth, metamorphosis, feeding, reproduction, and behaviour. Most receptors for biogenic amines, neuropeptides, and protein hormones are G protein-coupled receptors (GPCRs) [81]. Intracellularly, the G proteins initiate second messenger cascades [82]. JH and ecdysone, however, are lipophilic and can diffuse through the cell membrane to bind with nuclear receptors [83,84]. In addition, ecdysone can also activate a specific GPCR, and initiate a second messenger cascade [85]. There is extensive cross-talk between these extracellular signal molecules.

S. maritima possesses 19 biogenic amine receptors, a number similar to the 18–22 biogenic amine receptors that have been identified in other arthropods (Table S19). In *S. maritima*, there are four octopamine GPCRs, one octopamine/tyramine, one tyramine, four dopamine and three serotonin GPCRs, three

GPCRs for acetylcholine, one GPCR for adenosine, and two orphan biogenic amine receptors. Although this distribution resembles very much that of *Drosophila* and other arthropods, there are some interesting differences with *Drosophila*, which expresses two additional β -adrenergic-like octopamine receptors compared to *S. maritima*, while *S. maritima* expresses two putative β -adrenergic-like octopamine receptors (Sm-OctBetaRHK and Sm-D1/OctBeta), which are expressed in a number of insect and tick species, but not in *Drosophila* (Table S20) [86]. The true functional identities of all the putative *S. maritima* biogenic amine GPCRs awaits their cloning, functional expression, and pharmacological characterization in cell lines.

In addition, 36 neuropeptide and protein hormone precursor genes are present in this centipede. Each neuropeptide precursor contains one or more (up to seven) immature neuropeptide sequences (Figure S20). Interestingly, the centipede contains two CCHamide-1, two eclosion hormone, and two FMRFamide genes, whereas these genes are only present as single copies in the genomes of most other arthropods [87]. In concert with the presence of 36 neuropeptide genes, we found 33 genes for neuropeptide receptors (31 GPCRs and two guanylyl cyclase receptors) (see Table S21). As observed for the neuropeptide genes, a number of the neuropeptide receptor genes, which are only found as single copies in most other arthropods, have also been duplicated. *S. maritima* has two inotocin GPCR genes, two SIFamide, two corazonin, two eclosion hormone guanylyl cyclase receptor genes, two eclosion triggering hormone GPCR genes, three sulfakinin GPCR genes, and three LGR-4 (Leu-rich-repeats-containing-GPCR-4) genes. The latter receptors are orphans (GPCRs without an identified ligand) and only present as single-copy genes in most other arthropods [88]. Several of these duplicated GPCR genes are located in close vicinity to each other in the genome (Figure S21, suggesting recent duplication events. Furthermore, duplications of both the eclosion hormone and its receptor genes and the duplication of the ecdysis triggering hormone receptor genes suggest that the process of ecdysis (moulting) has undergone some sort of modification, perhaps requiring more complex control in the lineage leading to centipedes.

A



B

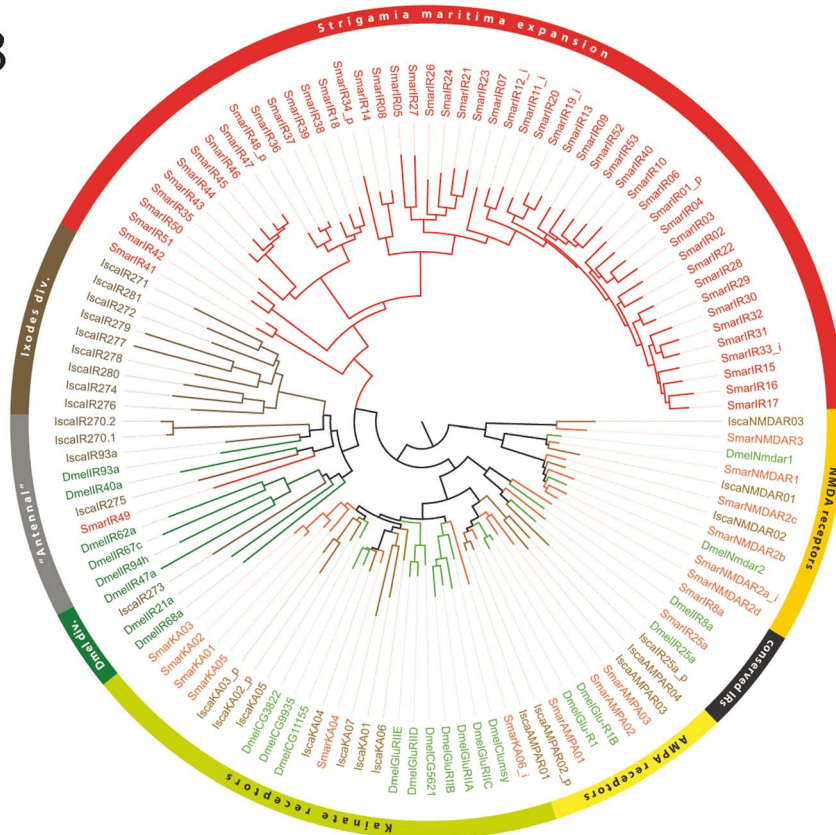


Figure 5. Expansion of chemosensory receptor families. (A) Phylogenetic relationships among *S. maritima* (Smar), *I. scapularis* (Isca), *D. pulex* (Dpul), and a few insect GRs that encode for sugar, fructose, and carbon dioxide receptors (Dmel, *D. melanogaster*, and Amel, *A. mellifera*). (B) Phylogenetic relationships among *S. maritima*, *I. scapularis*, and a few *D. melanogaster* IRs and Ig1uR genes (the suffix at the end of the protein names indicates: i, incomplete and p, pseudogene). doi:10.1371/journal.pbio.1002005.g005

We summarize in Table S22 the neuropeptide/protein hormone signalling systems that are present or absent in selected arthropod genome sequences. Each arthropod species, including *S. maritima*, has its own characteristic pattern, or “barcode,” of present/absent neuropeptide signalling systems. However, the relationship between the specific neuropeptide “barcode” and physiology remains to be elucidated.

Insect JH is important for growth, moulting, and reproduction in arthropods [84]. This hormone is a terpenoid (unsaturated hydrocarbon) that is synthesized from acetyl-CoA by several enzymatic steps (Figure S22). In several insects the production of JH is stimulated by the neuropeptide allatotropin, while it is inhibited by either allatostatin-A, -B, or -C [89,90]. We found that *S. maritima* has orthologues of many of the biosynthetic enzymes needed for JH biosynthesis in insects (Table S23). Also, the JH binding proteins are encoded in the centipede genome as well as JH degradation enzymes (Table S24). This implies that the complete JH system is present in this centipede. Similarly, neuropeptides that could stimulate or inhibit the synthesis and release of JH, such as allatotropin and the allatostatins -A, -B, and -C, are also present in *S. maritima* (Figure S22, suggesting that the overall functioning of the JH system in centipedes might be very similar to that of insects) (Table S23). To date, the existence of JH signalling systems has been demonstrated in insects, crustaceans, and recently in spider mites [89,91,92]. Its occurrence in *S. maritima* and spider mites (Chelicerata) indicates that JH signalling has deep evolutionary roots and we suggest that it might have evolved together with the emergence of the exoskeleton in arthropods.

Developmental Signalling Systems

Certain signalling systems, including transforming growth factor (TGF)-beta, Wnt, and fibroblast growth factor (FGF), are used throughout development across the animal kingdom. Various lineage-specific modifications of these systems have occurred, particularly within the arthropods. With regards to TGF-beta signalling we found single orthologues of all members of the Activin family, except Alp (Activin-like protein) (see Figure S23; Text S1). In the BMP-family, the *S. maritima* genome contains two divergent BMP sequences, as well as a clear orthologue of *glass-bottom boat* (*gbb*) and two *decapentaplegic* (*dpp*) orthologues. In addition, the *S. maritima* sequences confirm the ancestral presence of an anti-dorsalizing morphogenetic protein (ADMP) and a BMP9/10 orthologue in arthropods, which are both absent from *Drosophila* [93]. Most interestingly, the *S. maritima* genome includes the antagonistic BMP ligand BMP3 (previously suggested to be present only in deuterostomes [94]), a potential *gremlin/neuroblastoma suppressor of tumorigenicity*, and two nearly identical *bambi* genes (absent from *Drosophila*), and the BMP inhibitor *noggin* (present in vertebrates but lost in most holometabolous insects). The multiple BMP-agonists and -antagonists indicate that considerable changes have occurred in the TGF-beta signalling system during arthropod evolution, particularly in the Holometabola.

Reconstructions of Wnt gene evolutionary history suggest that the ancestral bilaterian possessed at least 13 distinct Wnt gene subfamilies [95,96]. This initial number has been secondarily

reduced in many taxa. This trend of secondary gene loss is readily apparent within the arthropods, with holometabolous insects such as *D. melanogaster* retaining only seven Wnt subfamilies [97,98]. In contrast, the Wnt signalling complement in *S. maritima* comprises 11 of the 13 Wnt-ligand subfamilies (Figure S24). Phylogenetic investigation has identified these genes as *wnt1*, *wnt2*, *wnt4*, *wnt5*, *wnt6*, *wnt7*, *wnt9*, *wnt10*, *wnt11*, *wnt16*, and *wntA*. *wnt3* and *wnt8* are missing from the *S. maritima* genome. While the absence of *wnt3* is common to protostomes, *wnt8* or *wnt8*-like sequences occur in other protostome genomes, including insects, spiders, and another myriapod, *Glomeris marginata* [97]. The Wnt genes are known to display a degree of linkage and clustering in many arthropods. Some conservation of this is also found in *S. maritima*, with *wnt1*, *wnt6*, and *wnt10* adjacent to each other on the same scaffold, possibly representing part of an ancient clustering (Table S25) [99].

The primary receptors for Wnt ligands in the canonical Wnt signalling pathway are the trans-membrane receptors of the Frizzled family. Five of these have been identified: *Frizzled1*, *Frizzled4*, *Frizzled5/8*, *Frizzled7*, and *Frizzled10*. As is the case for the *wnt* genes themselves, this is a larger number than is found in most arthropods. Other Fz-related genes are also present: *smoothened*, involved in Hedgehog signalling, and *secreted frizzled related protein*, which has inhibitory roles in Wnt signalling in other taxa. Putative non-canonical Wnt receptors are also encoded, including two subfamilies of *receptor tyrosine kinase-like orphan receptor* (*ror*). In addition to *ror2*, there is a lineage-specific duplication of *ror1*, making a total of three *ror* genes, as opposed to only one in *D. melanogaster*. Another Wnt agonist, the *R-spondin* orthologue was also found. As part of the Wnt-binding complex we found one *arrow*-LRP5/6-like Wnt-coreceptor gene in the genome: *lrp6*. Other LRP-molecules with potential Wnt-binding activity also exist: LRP1, LRP2, and LRP4. Because of the absence of an intracellular signalling domain these could potentially function as Wnt-inhibitors. Together, the large number of ligand and receptor genes point towards both the conservation of an ancestral Wnt signalling system and to a certain degree of unusual complexity in of this system in *S. maritima*.

Concerning the FGF pathway, we identified two closely related FGF receptors. These two *S. maritima* receptors are likely to stem from a duplication in the myriapod lineage that was independent from that which generated the two *Drosophila* FGFRs, *Heartless* and *Breathless* (Figure S25). The number of FGF ligands found in the genomes of insects such as *D. melanogaster* (three *fgf* genes) or *T. castaneum* (four *fgf* genes) is small when compared to 22 *fgf* genes found in the genomes of vertebrates. In the *S. maritima* genome, we identified three *fgf*-genes (Figure S26). One of them potentially represents an *fgf 18/8/24* orthologue to which the *fgf8*-like genes of *Tribolium* and of *Drosophila* (*pyramus* and *thisbe*) are associated. The second *S. maritima* *fgf* groups with the *fgf1* genes, while the third groups with the *fgf 16/9/20* clade (the first known arthropod member of this clade). Low support values for this grouping raise the possibility that it might actually be an orthologue of insect *branchless* genes. Other FGF-pathway genes present in *S. maritima* include *stumps* (Downstream-of-FGF-signalling [DOF]) and *sprouty related*.

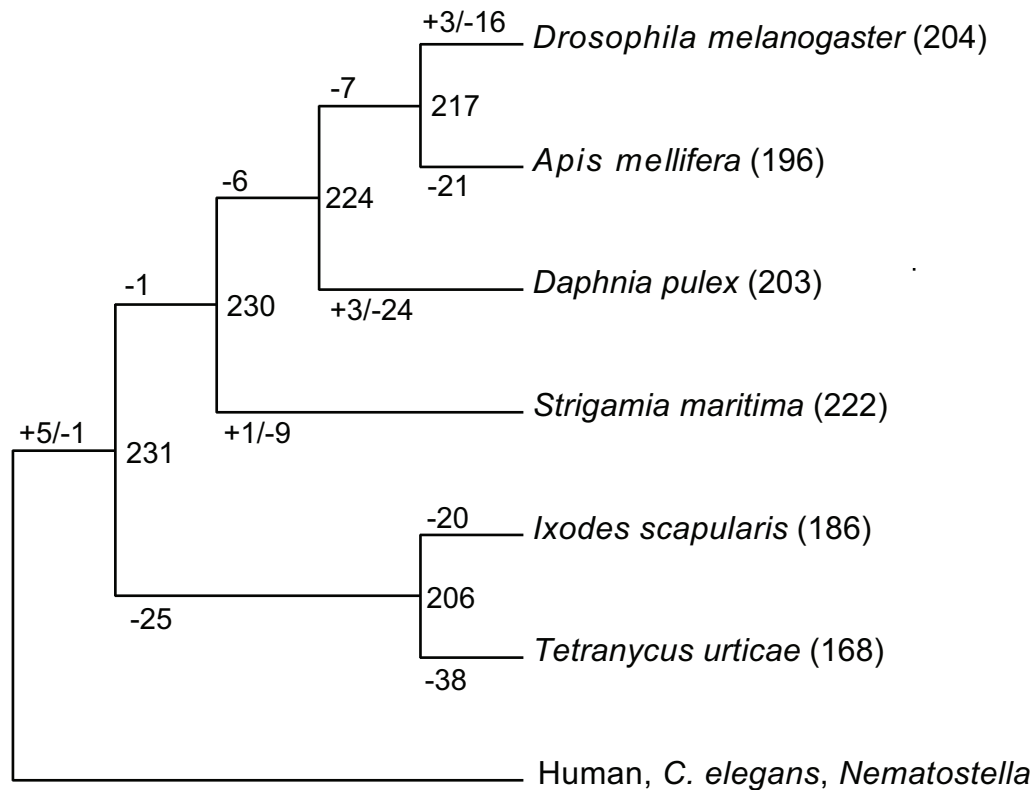


Figure 6. Ancestral protein kinases are extensively lost during arthropod evolution. *S. maritima* is an exception and retains the largest number of ancestral kinases. Numbers of kinase subfamilies in selected species are shown in parentheses after species names. The gains, losses, and inferred content of common ancestors are listed on internal branches. Kinases found in at least two species from human, *C. elegans* and *Nematostella vectensis* were used as an outgroup.
doi:10.1371/journal.pbio.1002005.g006

Protein Kinases

Kinases make up about 2% of all proteins in most eukaryotes, while they phosphorylate over 30% of all proteins and regulate virtually all biological functions. We identified 393 protein kinases in the *S. maritima* genome, representing 2.6% of the proteome. We classified these into conserved families and subfamilies, compared the kinome to those of 26 other arthropods and inferred the evolutionary history of all kinases across the arthropods (Figure 6). We predict that an early arthropod had at least 231 distinct kinases and see considerable loss of ancestral kinases in most extant species. *S. maritima* has the smallest number of losses among the arthropods, with only ten kinases lost relative to the arthropod ancestor. In contrast, the two chelicerates *T. urticae* and *I. scapularis* have lost 63 and 45 kinases, respectively, and *D. melanogaster* lost 30, giving *S. maritima* the richest repertoire of conserved kinases of any arthropod examined. All but one of the losses in *S. maritima* have been lost in other arthropods, suggesting that these genes may be partially redundant or particularly prone to loss. The one unique loss is NinaC, which in *Drosophila* is required for vision, likely associated with other vision related gene loss described above. As in many other species, we also see some novelties and expansions of existing families: the SRPK kinase family, involved in splicing and RNA regulation, has expanded to 36 members, and the nuclear VRK family is expanded to 16. A novel family of receptor guanylate cyclases (nine genes) and three clusters of unique protein-kinase-like (PKL) kinases, containing 28 genes in total, are also seen, though their functions are not known.

Developmental Transcription Factors

DNA binding proteins with the capacity to regulate the expression of other genes are central players in the control of development and many other processes. Since one of the original interests in *S. maritima* was for its developmental characteristics, we carried out a survey of developmentally relevant transcription factors, with an emphasis on transcription factors suspected to be involved in processes of axial specification, segmentation, mesoderm formation, and brain development. We identified orthologues of ~80 transcription factors of the Zinc finger and helix-loop-helix families in addition to the 113 homeobox-containing transcription factors already discussed (see Text S1). In no case did we fail to find at least one orthologue of the gene families expected from our knowledge of *Drosophila*, though individual duplications and losses among gene families were not uncommon. Among the set of pair-rule segmentation genes, for example, *S. maritima* has multiple homologues of *paired*, *even-skipped*, *odd-skipped*, *odd-paired*, and *hairy*-like genes, but only a single orthologue of *sloppy-paired* and *runt*-like genes, whereas *Drosophila* has multiple *runt* and *sloppy-paired* genes but only single orthologues of *even-skipped* and *odd-paired*. Where both lineages have multiple copies, (*paired*, *hairy*, *odd-skipped*), sequence alone rarely defines one-to-one orthologous relationships, and the evolutionary history remains unclear [29]. Other notable duplications include *caudal* (three genes) and *brachyury* (two genes). In a number of cases, transcription factors known to play a role in vertebrate development, but apparently missing from *Drosophila* and other insects, are retained in *S. maritima*. Examples include

the homeobox genes *Dmbx* and *Vax* noted above, and the FoxJ1, FoxJ2, and FoxL1 subfamilies of *forkhead/Fox* factors.

One of the developmental transcription factors provides an example where insects use isoforms to generate alternative proteins that are encoded by paralogous genes in *S. maritima*. Two centipede orthologues of the developmental transcription factor *cap'n'collar* encode isoforms that differ at their N-terminal end. The longer protein, encoded by the gene *cnc1*, contains sequence motifs that align to *Drosophila cnc* isoform C (Figure S27, which is broadly expressed throughout embryonic development) [100]. *S. maritima cnc1* is similarly expressed ubiquitously, whereas the other orthologue, *cnc2*, shows a segment specific pattern of expression similar to that of the shorter *Drosophila cnc* isoform B (VSH and MA, unpublished) [100].

Immune System

Arthropods can mount an innate immune response against pathogenic bacteria, fungi, viruses, and metazoan parasites. The nature of the responses to these invaders, such as phagocytosis, encapsulation, melanisation, or the synthesis of antimicrobial peptides, is often similar across arthropods [101]. Furthermore, key aspects of innate immunity are conserved between insects and mammals, which suggests an ancient origin of these defences. Previous studies have revealed extensive conservation of key pathways and gene families across the insects and crustaceans [102]. Beyond the Pancrustacea the extent of immunity gene conservation is unclear. Therefore, we searched the *S. maritima* genome for homologues of immunity genes characterised in other arthropods.

We found conservation of most immunity gene families between insects and *S. maritima* (Table S30), suggesting that the immune gene complement known from *Drosophila* was largely present in the most recent common ancestor of the myriapods and pancrustaceans. The humoral immune response of insects recognises infection using proteins that bind to conserved molecular patterns on pathogens [103]. Sequence homologues for the major recognition protein families found in *Drosophila*, peptidoglycan recognition proteins (PGRPs), and gram-negative bacteria-binding proteins (GNBPs), were found with the expected protein domains. These proteins then activate signalling pathways [103], and all four major insect immune signalling pathways (Toll, IMD, JAK/STAT, and JNK) are present in *S. maritima*, with 1:1 sequence homologues of most pathway members. The cellular immune response of insects relies on receptors and opsonins including thioester-containing proteins (TEPs), fibrinogen related proteins (FREPs), and scavenger receptors [103,104], and these are also present in *S. maritima*, often with protein domains in the same arrangement as *Drosophila*. We also find sequence homologues for effector gene classes including nitric oxide synthases (NOS) and prophenoloxidase (PPO). However, we failed to identify any antimicrobial peptide homologues, possibly as these genes are often short and highly divergent between species. In insects, it is common to find that certain immune gene families have undergone expansions in certain lineages [105]. Again, this is mirrored in *S. maritima*, where we found lineage-specific expansions of the PGRP and Toll-like receptor genes (TLRs) (Figure 7). Overall, the presence of the main families of immunity genes suggests that there is also functional conservation of the immune response.

The innate immune system is thought to rely on a small number of immune receptors that bind to conserved molecules associated with pathogens. This view was challenged by the discovery in *Drosophila* that the gene *Dscam* (Down syndrome cell adhesion molecule), which has the potential to generate over 150,000

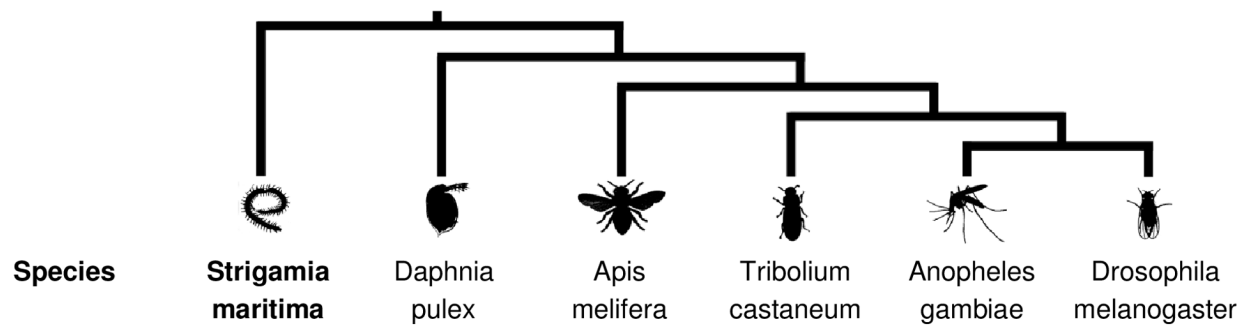
different protein isoforms by alternative splicing, functions as an immune receptor in addition to its roles in nervous system development [106]. *Dscam* family members are membrane receptors composed of several immunoglobulin (Ig) and fibronectin domains (FNIII). In pancrustaceans one member of the *Dscam* family has a large number of internal exon duplications and a sophisticated mechanism of mutually exclusive alternative splicing, which enables a single *Dscam* locus to somatically generate thousands of isoforms, which differ in half of two Ig domains (Ig2 and Ig3) and in another complete Ig domain (Ig7). This creates a high diversity of adhesion properties, useful for immune responses.

We found that *S. maritima* has evolved a different strategy to generate a diversity of *Dscam* isoforms [107]. The genome contains 60 to 80 canonical *Dscam* paralogues and over 20 other *Dscam* related incomplete or non-canonical genes (Figure 8). In 40 *Dscam* genes, the exon coding for Ig7 is duplicated two to five times (but not the exons coding for Ig2 and Ig3, which are duplicated in pancrustaceans). Our analysis of transcripts suggests that many of those duplicated exons might be alternatively spliced in a mutually exclusive fashion, supporting the hypothesis that the mechanism of mutually exclusive alternative splicing of *Dscam* probably evolved in the common ancestor of both pancrustaceans and myriapods. According to our phylogenetic analysis, which included 12 paralogues, the *S. maritima Dscams* share a common origin and arose by duplication in the centipede lineage [107]. In the chelicerate *I. scapularis*, *Dscam* has also been duplicated extensively, both by whole-gene and by domain duplications [107]. These *Dscam* homologues however do not have a canonical domain composition and whether or not alternative splicing is also present in chelicerates remains unknown. The independent evolution of *Dscam* diversification in different arthropod groups (one locus with dozens of exon duplications in pancrustaceans versus many gene duplications coupled with a few exon duplications in *S. maritima* (Figure 8) suggests that the functional diversity in adhesion properties was important in the early evolution of arthropods. Whether all of these genes function in the immune system or nervous system development remains to be determined.

The short-interfering RNA (siRNA) pathway is the primary defence of insects against RNA viruses, while the piRNA pathway silences transposable elements in the germ line and micro RNAs (miRNAs) function in gene regulation [108]. These RNAi pathways appear to be intact in *S. maritima*, as we found homologues of key genes, including *Ago1* and *Dicer-1* in the miRNA pathway, *Ago2* and *Dcr2* in the siRNA pathway, and *Ago3* and *piwi* in the piRNA pathway (Table S30). We found two paralogues of *Ago2* and three paralogues of *piwi*, suggesting that RNAi may be more complex than in *D. melanogaster*. In other arthropods, expansion of the *piwi* family has been linked to neo- or subfunctionalization of germ line and soma roles, and so it remains to be seen whether this is also the case for *S. maritima*.

Selenoproteins

Selenoproteins are peculiar proteins including a selenocysteine (Sec) residue, a very reactive amino acid typically found in the catalytic site of redox proteins, which is inserted through the recoding of a UGA codon [109]. While vertebrates possess 24–38 selenoproteins [110], insects have very few (*D. melanogaster* has three) or none at all. Several events of complete selenoproteome loss have been observed in insects [111]. These were ascribed to the fundamental differences in the insect antioxidant systems, which would favour selenoprotein loss or their conversion to standard proteins (cysteine homologues). The analysis of a



Subphylum	Myriapoda	Crustacea	Hexapoda	Hexapoda	Hexapoda	Hexapoda
Recognition and related						
PGRP	16	0	4	6	7	13
GNBP	3	11	2	3	7	3
TEP like	4	7	4	4	13	6
FREP like	13	-	-	7	46	13
SCR like	10	6	14	21	19	22
Dscam like	1	1	5	4	4	4
Signalling and Transduction						
Toll pathway						
Toll like	36	7	5	9	10	9
spz like	1	-	2	6	9	6
Myd88	1	1	1	1	1	1
tube	0	0	1	1	1	1
pelle	1	1	1	1	1	1
cactus	1	1	3	1	1	1
Dif	0	0	0	0	0	1
dorsal	1	1	1	1	1	1
IMD Pathway						
imd	~1	1	1	1	2	1
Fadd	1	-	1	1	1	1
Dredd	1	-	1	1	1	1
Tak1	1	-	1	1	1	1
Relish	2	1	1	1	1	1
Other						
domeless	1	-	1	1	1	1
JAK (hop)	1	-	1	1	1	1
Stat92E	1	1	1	1	1	1
JNK (bsk)	1	-	1	1	1	1
Hem	1	-	1	-	1	1
Effectors						
PPO	1	1	1	3	9	3
Nos	3	2	1	1	1	1

Figure 7. Presence and absence of immunity genes in different arthropods. Counts of immune genes are shown for *S. maritima*, *D. pulex* [131], *A. mellifera* [86], *T. castaneum*, *Anopheles gambiae*, and *D. melanogaster* [132]. ~, identity of the gene is uncertain; -, not investigated. doi:10.1371/journal.pbio.1002005.g007

myriapod selenoproteome is then crucial for a phylogenetic mapping of such differences.

The *S. maritima* genome was found to be surprisingly rich in selenoproteins: we have identified 20 predicted proteins (Table S26). Downstream of the coding sequence of each selenoprotein gene, we detected a selenocysteine insertion sequence (SECIS) element, the stem-loop structure necessary to target the Sec recoding machinery during selenoprotein translation. The full set of factors necessary for selenocysteine insertion and production was also found: tRNA-Sec, SecS, SBP2, eEFsec, pstk, secp43, SPS2. The centipede selenoproteome is rather similar to that predicted for the ancestral vertebrate (see [110]). This supports the idea that selenoprotein losses are specific to insects and can be ascribed to changes in that lineage, supporting the idea that a massive selenoproteome reduction occurred specifically in insects. A notable difference with vertebrates was found for the protein methionine sulfoxide reductase A (MsrA). This enzyme catalyzes the reduction of methionine-L-oxide to methionine, repairing proteins that were inactivated by oxidation. A selenoenzyme from this family has been previously characterized in the green alga *Chlamydomonas*, and selenocysteine containing forms were also observed in some non-insect arthropods [112]. In contrast, only cysteine homologues are present in vertebrate and insect genomes. We found a Sec-containing MsrA in the centipede genome, as well as in arthropods *D. pulex*, *I. scapularis*, and also in the chordate *B. floridae*. This, along with phylogenetic reconstruction analysis, supports the idea that the selenoprotein MsrA was present in their last common ancestor, and was later converted to a cysteine homologue independently in insects and vertebrates.

The two major antioxidant selenoprotein families in vertebrates, glutathione peroxidases (GPx), and thioredoxin reductases (TrxR), were also found with selenocysteine in the centipede genome. In contrast, all holometabolous insects possess only cysteine forms,

and consistently, important differences were noted in these and other enzymes in the glutathione and thioredoxin system (see [113] for an overview). Thus, on the basis of gene content, we expect the antioxidant systems of *S. maritima* to be more similar to vertebrates and other animals than to holometabolous insects like *D. melanogaster*.

DNA Methylation

Invertebrate DNA methylation occurs predominantly on gene bodies (exons and introns), via addition of a methyl group to a cytosine residue in a CpG context [114–116]. The exact function of gene body methylation is currently unknown, though it is correlated with active transcription in a wide range of species [116], and has been implicated in alternative splicing [117,118] and regulation of chromatin organization [118]. Methylated cytosines are susceptible to deamination, to form a uracil, which is recognized as a thymine. Thus, over evolutionary time, highly methylated genes (in germ-line cells) will have comparatively low CpG content. The “observed CpG/expected CpG” ($\text{CpG}_{(o/e)}$) ratio is an indicator of C-methylation: plots of $\text{CpG}_{(o/e)}$ for a gene set produce a bimodal distribution where a proportion of the genes have an evolutionary history of methylation [119]. In contrast, species without methylation systems, such as *D. melanogaster*, yield a unimodal distribution [119].

The *S. maritima* gene body $\text{CpG}_{(o/e)}$ plot has a trimodal distribution, with the majority of genes having a ratio close to 1 (Figure 9; Text S1). Underlying this major peak are two smaller peaks, one “low” and one “high” centred around ratios of 0.62 and 1.48, respectively. This “high” peak, that contains genes with higher than expected CpG content, is unusual and is not seen in this analysis of other arthropods [91,119–121]. Applying the same analysis to 1,000 bp windows across the entire genome (including both coding and non-coding regions) reveals a similar peak of high CpG content (Figure S29). This implies that the peak of “high” CpG content seen in gene bodies is due to unusually high CpG content in some regions of the genome rather than a specific feature of those coding regions. The “low” peak, however, indicates that 9.5% of genes have been methylated in the germ-line over evolutionary time. The number of genes contained within the “low” peak in *S. maritima* is smaller than observed in insect species with methylation, which can be as high as 40% in exceptional species such as the pea aphid and the honeybee [119,120], where the mechanism is likely involved in polyphenism and caste differences respectively. However, the number of genes methylated is less in non-social hymenopteran such as *Nasonia vitripennis*, in beetles, and in mites [91,121,122]. Consistent with the low-levels of germ-line methylation detected, the genome contains a single orthologue of the de novo DNA methylation enzyme Dnmt3 and four orthologues of the maintenance DNA methyltransferases Dnmt1(a–d). Two of the Dnmt1 orthologues have lost amino acids that are required for methyltransferase activity, but these genes are represented in the transcriptome data, and are thus unlikely to be pseudogenes. One Dnmt1 gene shows sex-specific splicing, with a shorter transcript producing a truncated protein seen in female-derived transcription libraries. We also find a single orthologue of Tet1, a putative DNA demethylation enzyme [123,124]. Taken together these data indicate that *S. maritima* has an active DNA methylation system, and that over evolutionary time a small number of genes have

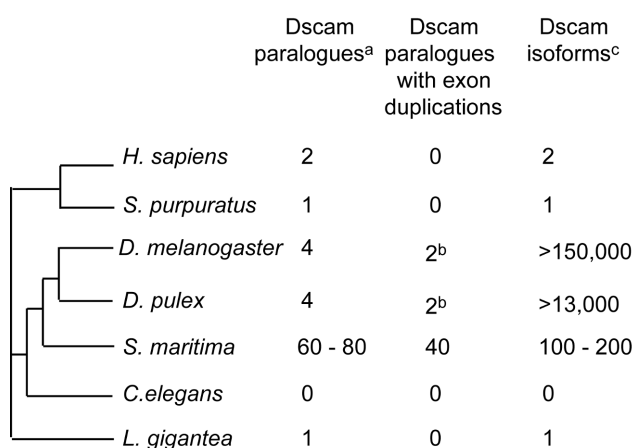


Figure 8. Dscam diversity caused either by gene and/or exon duplication in different Metazoa. ^aOnly canonical Dscam paralogues were considered. ^bIn *D. melanogaster* and *D. pulex* the paralogue Dscam-L2 has two Ig7 alternative coding exons. ^cPotential number of Dscam isoforms, circulating in one individual, produced by mutually exclusive alternative splicing of duplicated exons. doi:10.1371/journal.pbio.1002005.g008

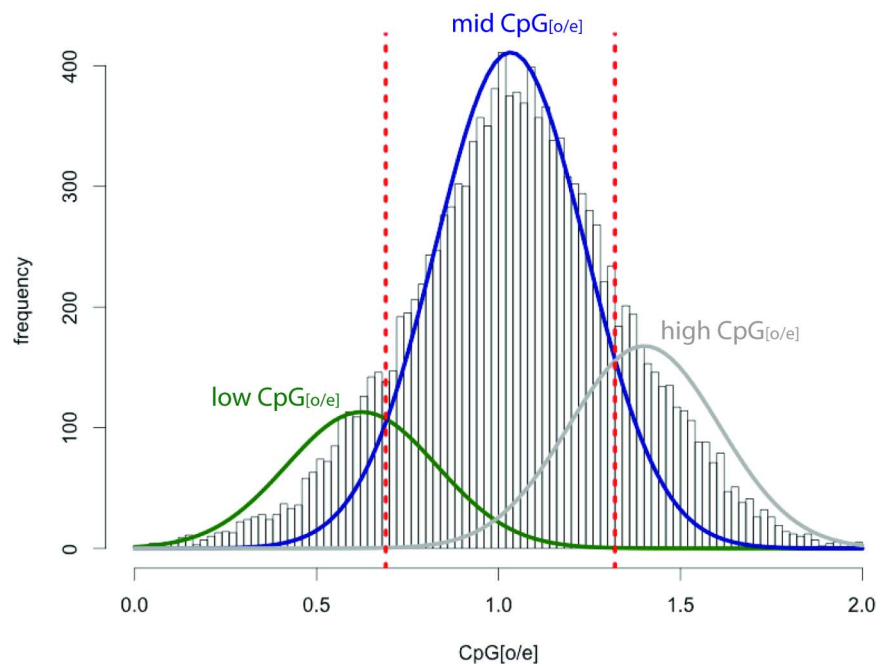


Figure 9. Frequency histogram of $\text{CpG}_{(o/e)}$ observed in *S. maritima* gene bodies. The y-axis depicts the number of genes with the specific $\text{CpG}_{(o/e)}$ values given on the x-axis. The distribution of $\text{CpG}_{(o/e)}$ in *S. maritima* is a trimodal distribution, with a low- $\text{CpG}_{(o/e)}$ peak consistent with the presence of historical DNA methylation in *S. maritima* and the presence of a high $\text{CpG}_{(o/e)}$ peak. The data underlying this plot are available in File S4. doi:10.1371/journal.pbio.1002005.g009

been methylated in the germ-line, resulting in a lower than expected CpG dinucleotide content.

Non-Protein-Coding RNAs in the *S. maritima* Genome

We annotated over 900 homologues of known non-coding RNAs in the *S. maritima* genome, including over 600 predicted tRNAs (plus an additional 300 tRNA pseudogenes), 71 copies of 5S rRNA and 12 5.8S rRNAs, 88 copies of RNA components of the major spliceosome, and three out of the four RNA components of the minor U12 spliceosome, and 54 microRNA genes. As is common for whole genome assemblies, we did not identify intact copies of the 18S or 28S rRNAs. Further details of our methodology are provided in Text S1.

The predicted tRNA gene set includes all anticodons necessary to code for the 21 amino acids, including four potential Sec tRNAs. We identify a massive expansion of the tRNA-Ala-GGC family, with 322 sequences classified as functional tRNAs by tRNAscan-SE and an additional 172 classified as pseudogenes. These appear scattered throughout the scaffolds of the genome assembly. It is highly likely that the majority of these genes are pseudogenes, and the expansion may represent co-option of the tRNA into a transposable element.

Three *S. maritima* microRNA genes have been reported previously, and are available in the miRBase database (version 18) [125]. Two of these, mir-282 and mir-965, have homologues in crustaceans and insects. The third, mir-3930, is specific to myriapods [15]. In addition, we found 52 homologues of known microRNAs (Figure S34). These include 28 homologues of the 34 ancient microRNA families found throughout the Bilateria [126]. Four of these families were previously reported to be lost at various stages during animal evolution and, consistent with this, we failed

to identify them in the *S. maritima* genome. Surprisingly, we also could not identify the *S. maritima* homologue of mir-125, a member of the ancient mir-100/let-7/mir-125 cluster, which is found in almost all bilaterians and has a well-established function in the regulation of development of many species [127–129]. Mir-100 and let-7 are well-conserved and localized within a 1 kb region on the same scaffold in *S. maritima*. Whilst we cannot rule out the possibility that the missing mir-125 is an artefact of the draft-quality genome assembly, the size of the scaffold strongly suggests that it is not present in the mir-100/let-7 cluster. We also identified 17 homologues of microRNAs common to ecdysozoans, and nine microRNAs known only from arthropods. Among the former, there are five homologues of mir-2 localized in close proximity to each other and downstream of mir-71. This clustering is conserved across protostomes, and it has previously been shown that the mir-2 family underwent various expansions during evolution [130]. Finally, we discovered a homologue of mir-2788, which was previously only known from insects, suggesting that this microRNA had an earlier origin.

Conclusions

The sequencing of the centipede genome extends significantly the diversity of available arthropod genomes, and provides novel information pertinent to a range of evolutionary questions. Myriapods show a simple body organization that has remained relatively unchanged in comparison to their ancestors from the Silurian or even earlier [6], leading to an expectation of general conservatism. The myriapods are descendants of an independent terrestrialisation event from the hexapods and chelicerates, opening the opportunity for studying convergent evolution in these taxa. Naturally, *S. maritima* itself has its own evolutionary

history, including both lineage specific features of the geophilomorphs and adaptations to their subterranean environment, allowing us to identify specific genomic signatures of ecological adaptations. Finally, the phylogenetic position of the myriapods within the arthropods has been the subject of intense debate for several years, and the availability of genomic data for a myriapod should contribute to the future resolution of this debate.

The morphological conservatism of centipedes is mirrored in many conservative aspects of the *S. maritima* genome. From the analyses of the various gene families outlined above it becomes clear that the *S. maritima* genome has undergone much less gene loss and rearrangement than the genomes of other sequenced arthropods, in particular those of the holometabolous insects such

as *D. melanogaster*. This prototypical nature of the *S. maritima* genome is illustrated by the conservation of synteny relative to the arthropod and bilaterian ancestors, and the conservation of some ancient gene linkages and clustering, as seen for numerous homeobox genes. As such, the *S. maritima* genome can serve as a guide to the ancestral state of the arthropod genomes, or as a reference in the reconstruction of evolutionary events in the history of arthropod genomes.

The independent terrestrialisation of the myriapods and insects is evidenced by the use of different evolutionary solutions to similar problems. Figure 10 summarizes some of the gene gains and losses observed. We see this most clearly in the independent expansions of gustatory receptor proteins in myriapods and insects and the

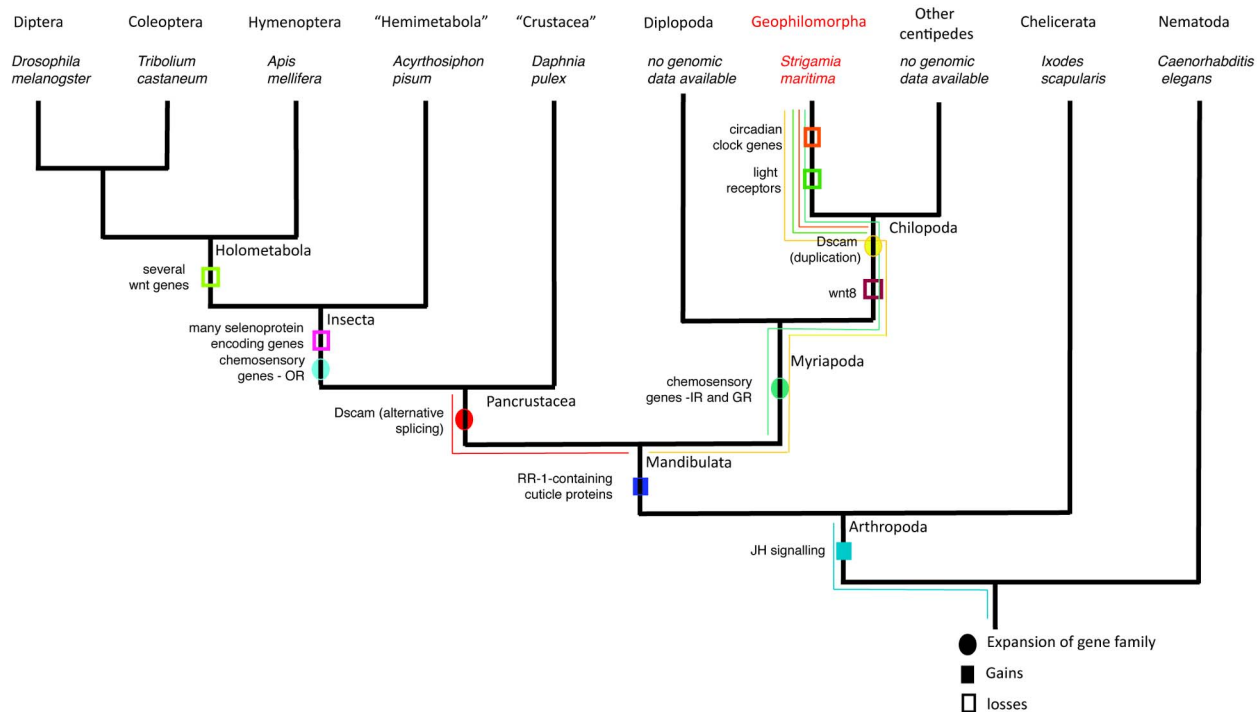


Figure 10. Arthropod phylogenetic tree (with nematode outgroup) showing selected events of gene loss, gene gain, and gene family expansions. Main taxa are listed on the tips, with representative species for which there is a fully sequenced genome listed below. Major nodes are also named. Data from the genome of *S. maritima* allow us to map when in arthropod evolution these events occurred, even when these events did not occur on the centipede lineage. A plausible node for the occurrence of each event is marked and colour-coded, with the possible range marked with a thin line of the same colour. The events, listed from left to right are: (1) Dscam alternative splicing as a strategy for increasing immune diversity is known from *D. melanogaster*, as well as the crustacean *D. pulex*, and thus probably evolved in the lineage leading to pancrustacea, after the split between centipedes. (2) Several wnt genes have been lost in holometabolous insects, leaving only seven of the 13 ancestral families. This loss occurred gradually over arthropod evolution, but reached its peak at the base of the Holometabola. (3) Selenoproteins are rare in insects. The presence of a large number of selenoproteins in *S. maritima* as well as in other non-insect arthropods suggests that the loss of many selenoproteins occurred at the base of the Insecta. (4) Expansion of chemosensory gene families occurred independently in different arthropod lineages as they underwent terrestrialisation. The OR family is expanded in insects only. (5) Chemosensory genes of the GR and IR genes have undergone a lineage specific expansion in the genome of *S. maritima*. As these are probably also linked with terrestrialisation we suggest that this expansion happened at the base of the Chilopoda, but it could have also occurred later in the lineage leading to *S. maritima*. (6) Cuticular proteins of the RR-1 family are numerous in the *S. maritima* genome. They are found in other arthropods, but not in chelicerates nor in any non-arthropod species. This suggests that the RR-1 subfamily evolved at the base of the Mandibulata. (7) The genome of *S. maritima* has a large complement of wnt genes, but is missing *wnt8*. Since this gene is found in the Diplopod *G. marginata* (a species without a fully sequenced genome), the loss most likely occurred at the base of the Chilopoda. (8) Unlike the situation in *D. melanogaster*, immune diversity in the *S. maritima* genome is achieved through multiple copies of the Dscam gene. This expansion of the family could have happened at any time after the split between Myriapoda and Pancrustacea. (9) Both circadian rhythm genes and many light receptors are missing in *S. maritima*. These losses are most likely due to the subterranean life style of geophilomorph centipedes and are probably specific to this group. However, we cannot rule out the possibility that they were lost somewhere in the lineage leading to myriapods. (10) The existence of JH signalling in *S. maritima* as well as in all other arthropods studied to date strengthens the idea that this signalling system evolved with the exoskeleton of arthropods, though its origins could be even more ancient and date back to the origin of moulting at the base of the Ecdysozoa.
doi:10.1371/journal.pbio.1002005.g010

differential expansions of ionotropic and odorant receptors to deal with terrestrial chemosensation in the two lineages. Similarly, though probably not for the same reasons, we see a divergent solution for the generation of Dscam diversity in the immune response through the use of paralogues instead of the insect strategy of alternative splicing. The chelicerates also attained terrestriality independently. However, our understanding of chelicerate genomes still lags behind our understanding of insect, and now myriapod, genomes. Thus, extending this comparison to chelicerates, intriguing as it may be, will have to await future analysis of their genomes.

Lineage specific features of the *S. maritima* genome include the apparent loss of all known photoreceptors and a loss of the canonical circadian clock system based around *period* and its associated gene network. The characterization of whether *S. maritima* does have a circadian clock, and if it does how this is controlled, awaits further work, as does the pinpointing of when in their evolutionary history these systems were lost. The absence of the microRNA miR-125 is another surprising evolutionary loss. The extensive rearrangement of the mitochondrial genome is striking in comparison with the general conservatism seen in other known arthropod mitochondrial genomes, and especially in contrast with the conservative nature of *S. maritima*'s nuclear genome.

Materials and Methods

The *S. maritima* raw sequence, and assembled genome sequence data are available at the NCBI under bioproject PRJNA20501 (<http://www.ncbi.nlm.nih.gov/bioproject/PRJNA20501>) Assembly ID GCA_000239455.1. The genome was sequenced using 454 sequencing technology, assembled using the celera assembler, annotated using a combination of the Maker 2.0 pipeline, and custom perl scripts followed by manual annotation of selected genes. Text S1 includes detailed methods for these steps, and additionally for the individuals sequenced, library construction and sequencing protocols used, repeat analysis, RNA sequencing, phylome db analysis, specific protocols for manual annotation of gene families, CpG analysis, and phylome and synteny re-construction.

Supporting Information

Figure S1 Frequency histogram showing the distribution of gene lengths in the *S. maritima* genome. Gene length data used in this plot are available in File S4. (PDF)

Figure S2 Multi-gene phylogeny for the 18 species included in the phylogenomics analysis. 1,491 widespread single-copy sets of orthologue sequences in at least 15 out of the 18 species were concatenated into a single alignment of 842,150 columns. Then, a maximum-likelihood tree was inferred using LG as evolutionary model by using PhyML. (PDF)

Figure S3 Multi-gene phylogeny for 12 species included in the phylogenomics analysis plus five additional Chelicerata species. 1,491 widespread single-copy sets of orthologue sequences were concatenated into a single alignment of 829,729 positions. Then, a maximum-likelihood tree was inferred using LG as the evolutionary model by using PhyML. (PDF)

Figure S4 Alternative topological placements of *S. maritima* relative to the main arthropod groups considered in the study: Chelicerata and Pancrustacea. Internal

organization of each group was initially collapsed and, therefore, optimized during maximum-likelihood reconstruction. (PDF)

Figure S5 Clusters of genes specifically expanded in the centipede lineage. On the plot, only clusters grouping five or more protein-coding genes were considered. The data underlying this plot are available in File S4. (PDF)

Figure S6 Mitochondrial gene organisation. Shaded regions represent differences from the ground pattern. Gene translocations in Myriapoda have been noted in *Scutigera caudata* (Myriapoda: Symphyla) [49]. The previous example of the small conserved region trnaF-nad5-H-nad4-nad4L on the minus strand between *Limulus*, *Lithobius*, and *Strigamia* is not a conserved feature in all Chilopoda, because *Scutigera coleoptrata* have an interruption between nad5 and H-nad4 with elements on the minus and plus strands accompanied by a translocation of nad4L to a position immediately preceding nad5. (PDF)

Figure S7 Classification of all *S. maritima* (Sma) homeodomains (excluding Pax2/5/8/sv) via phylogenetic analysis using *T. castaneum* (Tca) and *B. floridae* (Bfl) homeodomains. This phylogenetic analysis was constructed using neighbour-joining with a JTT distance matrix and 1,000 bootstrap replicates. Gene classes are indicated by colours. The genes coloured in grey are those genes that cannot be assigned to known classes. Further classification was performed using additional domains outside the homeodomain and by performing additional phylogenetic analysis for particular gene classes using maximum-likelihood and bayesian approaches. Pax2/5/8/sv is excluded due to the gene possessing only a partial homeobox. (PDF)

Figure S8 Phylogenetic analysis of ANTP class homeodomains of *S. maritima* (Sma) using *T. castaneum* (Tca) and *B. floridae* (Bfl) for comparison. These phylogenetic analyses were constructed using neighbour-joining with a JTT distance matrix, 1,000 bootstrap replicates (support values in black). Nodes with support equal to or above 500 in the maximum-likelihood (LG+G) analysis are in blue and nodes with posterior probabilities equal to or above 0.5 (LG+G) in the Bayesian analysis are in red. (PDF)

Figure S9 Phylogenetic analysis of PRD class homeodomains of *S. maritima* (Sma) using *T. castaneum* (Tca) and *B. floridae* (Bfl) for comparison. These phylogenetic analyses were constructed using neighbour-joining with a JTT distance matrix, 1,000 bootstrap replicates (support values in black). Nodes with support equal to or above 500 in the maximum-likelihood (LG+G) analysis are in blue and nodes with posterior probabilities equal to or above 0.5 (LG+G) in the Bayesian analysis are in red. (PDF)

Figure S10 Phylogenetic analysis of HNF class homeodomains of *S. maritima* (Sma) using *B. floridae* (Bfl), human (*Homo sapiens*, Hsa), and sea anemone (*N. vectensis*, Nve) for comparison. These phylogenetic analyses were constructed using neighbour-joining with a JTT distance matrix, 1,000 bootstrap replicates (support values in black). Nodes with support equal to or above 500 in the maximum-likelihood

(LG+G) analysis are in blue and nodes with posterior probabilities equal to or above 0.5 (LG+G) in the Bayesian analysis are in red. (PDF)

Figure S11 Phylogenetic analysis of Xlox/Hox3 genes of *S. maritima* (Sma) using a selection of Hox1, Hox2, Hox3, Hox4, and Xlox sequences. This analysis was based upon the whole coding sequence of the genes, and was constructed using neighbour-joining with a JTT distance matrix and 1,000 bootstrap replicates. The blue support value (of 333) is the node that reveals the affinity between the Xlox/Hox3 genes of *S. maritima* and Xlox sequences. Ame, *A. mellifera*; Bfl, *B. floridae*; Cte, *Capitella teleta*; Dme, *D. melanogaster*; Lgi, *Lottia gigantea*; and Tca, *T. castaneum*. (PDF)

Figure S12 Multiple alignment of relevant residues of the Hox1, Hox2, Hox3, Hox4, and Xlox sequences of different lineages compared to *S. maritima* Hox3a and Hox3b sequences. Three paired class genes are included as an outgroup. The grading of purple colouring of the amino acids shows the identity level of these sequences. The red rectangles in the multiple alignment delimit the core of the hexapeptide motif and the homeodomain. This is the alignment used to construct the phylogenetic tree in Figure S13. Ame, *A. mellifera*; Bfl, *B. floridae*; Cte, *Capitella teleta*; Dme, *D. melanogaster*; Lgi, *Lottia gigantea*; and Tca, *T. castaneum*. (PDF)

Figure S13 Phylogenetic analysis of *S. maritima* Xlox/Hox3 homeodomain and hexapeptide motifs using a selection of Hox1, Hox2, Hox3, Hox4, and Xlox sequences. This analysis used a section of the coding sequence including the hexapeptide and some flanking residues plus the homeodomain (alignment in Figure S12). Three paired class genes are included as an outgroup. This phylogeny was constructed using neighbour-joining with the JTT distance matrix and 1,000 bootstrap replicates. Maximum likelihood support values are shown in blue and Bayesian posterior probabilities in red. Ame, *A. mellifera*; Bfl, *B. floridae*; Cte, *Capitella teleta*; Dme, *D. melanogaster*; Lgi, *Lottia gigantea*; Tca, *T. castaneum*. (PDF)

Figure S14 Fisher's exact test to distinguish whether *S. maritima* scaffold 48457 has significant synteny conservation with ParaHox or Hox chromosomes of humans. No significant Hox or ParaHox association is found. (PDF)

Figure S15 Phylogenetic analysis of TALE class homeodomains of *S. maritima* (Sma) using *T. castaneum* (Tca) and *B. floridae* (Bfl), including the Iroquois/Irx genes. These phylogenetic analyses were constructed using neighbour-joining with a JTT distance matrix, 1,000 bootstrap replicates (support values in black). Nodes with support equal to or above 500 in the maximum-likelihood (LG+G) analysis are in blue and nodes with posterior probabilities equal to or above 0.5 (LG+G) in the Bayesian analysis are in red. (PDF)

Figure S16 RNA processing in the Hox cluster of *S. maritima*. The transcriptome of *S. maritima* (Sm) eggs (blue), females (green), and males (red) was mapped to the Hox gene cluster (top panel; see Figure 4 in the main text) and transcript models were inferred for each gene within the cluster (shaded area) taking into account the presence of ORF and polyadenylation signals (PAS) to support the existence of RNA processing events.

We note the occurrence of more than one mRNA isoform of six *S. maritima* Hox genes (i.e., *Antp*, *Ubx*, *abd-A*, *lab*, *Dfd*, *pb*). In all these six cases alternative polyadenylation (APA) generates mRNAs bearing distinct 3' UTRs (alternative UTR sizes at the bottom). Alternative splicing (AS) with concomitant alternative promoter use (APU) events concern two *S. maritima* Hox genes *Dfd* and *ftz* (see alternative ORF sizes at the bottom). We also see that some genes such as *S. maritima* *Ubx* display high heterogeneity in 3'UTR sequences within the embryonic transcriptome ("eggs" data) suggesting the possibility that *S. maritima* *Ubx* APA might be developmentally controlled and/or display a tissue-specific pattern (see inset for further details on symbols). (PDF)

Figure S17 RNA processing in the *S. maritima* and *D. melanogaster* Hox clusters. (A) The incidence of alternatively processed mRNAs is comparable between *S. maritima* and *D. melanogaster*, in that over 75% of the *S. maritima* Hox genes undergo RNA processing of one type or another. Similarly, seven out of the eight *Drosophila* Hox genes produce different mRNA isoforms (FlyBase, <http://flybase.org/>). (B) Three *D. melanogaster* Hox genes undergo AS (blue) and five produce different transcripts via APA (red, FlyBase <http://flybase.org/>). In addition five fruit fly Hox genes form different RNA species by APU (green). (C) Classification of all alternatively processed mRNA events in the *S. maritima* Hox cluster based on the same categorisation as in (B). Note that patterns of AS and APA affecting *S. maritima* and *D. melanogaster* Hox genes are relatively comparable; in contrast, APU seems more prevalent in the *Drosophila* (five out of eight genes) than in the centipede (two out of nine genes) Hox genes. (PDF)

Figure S18 Phylogenetic tree of the *S. maritima*, *D. pulex*, *I. scapularis*, and representative insect GRs, part one. This is a corrected distance tree and was rooted at the midpoint in the absence of a clear outgroup, an approach that clearly indicates the distinctiveness of the centipede GRs. It is a more detailed version of Figure 5A. The *S. maritima*, *D. pulex*, *I. scapularis*, and representative insect gene/protein names are highlighted in red, blue, brown, and green, respectively, as are the branches leading to them to emphasize gene lineages. Bootstrap support levels in percentage of 10,000 replications of neighbour-joining with uncorrected distance is shown above major branches. Comments on major gene lineages are on the right. Suffixes after the gene/protein names are: PSE, pseudogene; FIX, sequence fixed with raw reads; JOI, gene model joined across scaffolds. Note that in Figure 5A for space reasons the IsGr47 and 59 proteins are included in the carbon dioxide and sugar receptor groupings, respectively; however, there is no bootstrap support for these branches, and no such functional assignment is claimed. Similarly, it is unlikely that the DpGr57/58 proteins are fructose receptors. (PDF)

Figure S19 Phylogenetic tree of the *S. maritima*, *D. pulex*, *I. scapularis*, and representative insect GRs, part two. This is a corrected distance tree and was rooted at the midpoint in the absence of a clear outgroup, an approach that clearly indicates the distinctiveness of the centipede GRs. It is a more detailed version of Figure 5A. The *S. maritima*, *D. pulex*, *I. scapularis*, and representative insect gene/protein names are highlighted in red, blue, brown, and green, respectively, as are the branches leading to them to emphasize gene lineages. Bootstrap support levels in percentage of 10,000 replications of neighbour-joining with uncorrected distance is shown above major branches. Comments on major gene lineages are on the right. Suffixes after the gene/protein names are: PSE, pseudogene; FIX, sequence

fixed with raw reads; JOI, gene model joined across scaffolds. Note than in Figure 5A for space reasons the IsGr47 and 59 proteins are included in the carbon dioxide and sugar receptor groupings, respectively; however, there is no bootstrap support for these branches, and no such functional assignment is claimed. Similarly, it is unlikely that the DpGr57/58 proteins are fructose receptors. (PDF)

Figure S20 Neuropeptide precursor sequences identified in the *S. maritima* genome. The putative signal peptides (predicted by SignalP) are underlined, the putative active neuropeptides or protein hormones (based on similarity to neuropeptides or protein hormones identified in other invertebrates) are marked in yellow. Green indicates putative basic cleavage sites flanking the putative neuropeptides. Glycines used for amidation are shown in blue, cysteines proposed to form cysteine bridges are shown in red. Dots indicate missing N- or C-termini. (DOCX)

Figure S21 Examples of tandem duplications of neuropeptide receptor genes. Structure of the two inotocin receptor genes found head-to-head on opposite strands of scaffold JH431865 (A). Structure of the two SIFamide receptor genes found tail-to-head on the same strand of scaffold JH432116 (B). (PDF)

Figure S22 Schematic diagram showing sesquiterpenoids/juvenoids synthesis (upper) and degradation (lower) pathways in arthropods. Molecules/hormones in synthesis are shown in bold, enzymes are shown in italics, and species/clades are shown in bold italics. (PDF)

Figure S23 Phylogenetic analysis of the TGF β ligands in arthropods. See Text S1 for details. Abbreviations: Ag, *Anopheles gambiae*; Am, *A. mellifera*; Ap, *Acyrtosiphon pisum*; Ca, *Clogmia albipunctata*; Dm, *Drosophila melanogaster*; Dp, *D. pulex*; Is, *I. scapularis*; Lg, *Lottia gigantea*; Ma, *Megaselia abdita*; Nv, *Nasonia vitripennis*; Ph, *Pediculus humanus*; Tc, *T. castaneum*. (EPS)

Figure S24 Range of Wnt genes present in *S. maritima*. Wnt genes present and number of Wnt subfamilies absent in *S. maritima* in comparison with other arthropods and three non-arthropod outgroups. (TIF)

Figure S25 Phylogeny of FGFR genes indicating that FGFR genes duplicated independently in *S. maritima* and *D. melanogaster*. See text for details. Alignment was performed using Clustal-Omega (<http://www.ebi.ac.uk/Tools/services/web>). The evolutionary history was inferred using the neighbour-joining method with bootstrapping to determine node support values (10,000 replicates). The evolutionary distances were computed using the Poisson correction method. Evolutionary analyses were conducted in MEGA5. (EPS)

Figure S26 Phylogeny including the three FGF genes of *S. maritima*. See text for details. Alignment was performed using Clustal-Omega (<http://www.ebi.ac.uk/Tools/services/web>). The evolutionary history was inferred using the neighbour-joining method with bootstrapping to determine node support values (10,000 replicates). The evolutionary distances were computed using the Poisson correction method. Evolutionary analyses were conducted in MEGA5. (EPS)

Figure S27 *Cap 'n' collar (cnc)* genes. (A) The two genes are located on different scaffolds. *Cnc1* is a long transcript consisting of 11 exons. *Cnc2* is shorter (eight exons), the three exons at the 3' end of the gene that encode the C-terminal region of the protein including the conserved domain (B) show a similar structure. (B) *S. maritima* Cnc protein structure. Both proteins contain the bZip domain in a similar position at the C-terminus. *Cnc1* encodes a long protein (925 amino acids). Bits of the N-terminal region (blue lines) align with *D. melanogaster* Cnc isoform C and *T. castaneum* Cnc variant A. (C) Cnc protein sequence alignment, only showing the aligning bits in the N-terminal region. Blue lines show short stretches of sequence that form a consensus motif. These motifs are not present in the proteins encoded by *Sm-cnc2*, *Dm-cnc* isoforms A and B, and *T. castaneum cnc* variant B. (JPG)

Figure S28 Frequency histograms of observed versus expected dinucleotide content in *S. maritima* gene bodies. (A–P) The y-axis depicts the number of genes with the specific dinucleotide_[o/e] values given on the x-axis. The distribution of all dinucleotide pairs, with the exception of CpG, is best described as a unimodal distribution. The distribution of CpG dinucleotides is best described as a trimodal distribution, with “high” and “low” CpG_[o/e] classes. The data underlying this figure are available in File S5. (TIF)

Figure S29 Frequency histogram of CpG_[o/e] observed in 1,000 bp windows of the *S. maritima* genome. The y-axis depicts the number of genes with the specific CpG_[o/e] values given on the x-axis. The distribution of CpG_[o/e] in *S. maritima* genome is a bimodal distribution, with a high CpG_[o/e] peak observed similar to that observed in the gene bodies (Figure 9). The data underlying this figure are available in File S6. (TIF)

Figure S30 Contrasting patterns of DNA methylation, as measured by over- and underrepresentation of CpG dinucleotides in coding regions (CpG_(o/e)), within arthropod species. In all graphs the y-axis depicts the number of genes with the specific CpG_(o/e) values given on the x-axis. (A) *D. melanogaster* coding regions show a unimodal peak reflective of the lack of DNA methylation in this species. (B) *A. mellifera* shows a bimodal peak consisting of genes with a lower than expected CpG_(o/e) (green distribution) and a higher than expected CpG_(o/e) (blue distribution). The presence of a bimodal distribution in this species is consistent with depletion of CpG dinucleotides in the coding regions of genes over evolutionary time as a result of DNA methylation. (C) A single unimodal peak is also observed for *Tetranychus urticae*, a species that has very low levels of DNA methylation. (D) The *S. maritima* distribution is best explained as a mixture of three distinct distributions that we have deemed “low” (green distribution), “medium” (blue distribution), and “high” (grey distribution). The genes within the low distribution likely contain genes that are historically methylated, whilst the “high” distribution can be explained by regions of the genome that are comparatively CpG-rich (as determined by the analysis of the *S. maritima* genome, Figure S29). The data underlying this figure are available in File S7. (PDF)

Figure S31 Chromosomal organisation of histone gene clusters in *S. maritima*. In insects such as *Drosophila* [115] and the pea aphid [109] histone encoding genes are present in quintet clusters, each cluster containing one gene from each of the five classes of histone. Only one such cluster could be identified in

S. maritima (A). The other four clusters identified in the *S. maritima* genome (B–D) all consist of two to three copies of a histone encoding gene of a single class. It is possible that these have arisen as a result of recent gene duplication.

(EPS)

Figure S32 *S. maritima* vasa DEAD-box helicase germ-line gene phylogeny. Maximum likelihood tree of *vasa/PL10* family genes. One gene is a likely *vasa* orthologue (SMAR015390), one groups with the *PL10* family (SMAR005518), and the majority group in an apparently distinct DEAD-box-containing clade. Bootstrap values for 2,000 replicates are shown at each node. Accession numbers for protein sequences are as follows: *Apis* Belle (XP_391829.3), *Apis* Vasa (NP_001035345.1), *Danio* PL10 (NP_571016.2), *Danio* Vasa (AAI29276.1), *Drosophila* Belle (NP_536783.1), *Drosophila* Vasa (NP_723899.1), *Gryllus* Vasa (BAG65665.1), *Mus* Mvh (NP_001139357.1), *Mus* PL10 (NP_149068.1), *Nasonia* Belle (XP_001605842.1), *Nasonia* Vasa (XP_001603956.2), *Nematostella* PL10 (XP_001627306.1), *Nematostella* Vasa 1 (XP_001628238.1), *Nematostella* Vasa 2 (XP_001639051.1), *Oncopeltus* Vasa (AGJ83330.1), *Parhyale* Vasa (ABX76969.1), *Tribolium* Belle (NP_001153721.1), *Tribolium* Vasa (NP_001034520.2), *Xenopus* PL10 (NP_001080283.1), *Xenopus* VLG1 (NP_001081728.1).

(EPS)

Figure S33 Phylogenomic inventory of meiotic genes in arthropods. Red genes are specific to meiosis in model species in which functional data are available. “+” and “–” indicate the presence and absence of orthologues, respectively. Numbers indicate copy number of duplicated genes.

(PDF)

Figure S34 Patterns of microRNA gain and loss across the animal kingdom with the inclusion of *S. maritima*.

The number of microRNAs that were gained or lost at each node are shown in green and red, respectively, and names are listed below each taxon. MicroRNAs that are found in the *S. maritima* genome are in bold, and families for which more than one homologue is found are marked with an asterisk. The tree depicts the Mandibulata hypothesis rather than the Myriochelata, as in [124].

(EPS)

Table S1 Detailed overview for the repetitive elements in *S. maritima*. For each group the number of elements (putative families), the number of their fragments or copies in the genome, the cumulative length, the proportion of the assembly, and some features are shown. This includes elements containing nested inserts of other elements (n), elements that appear to be complete (i.e., all typical structural and coding parts present, even if containing stop codons or frameshifts), elements with a RT or *Tase* domain detected (n), elements that potentially could be active as they contain an intact ORF with all the typical domains even though they could lack other structural features like terminal repeats, and elements that contain an intact ORF for the RT domain or parts of the *Tase* domain and could thus be partly active. The elements that could not be categorized or contained features of protein coding regions are shown at the bottom, whereby they probably do not belong to the transposable elements.

(XLSX)

Table S2 Set of species used in the comparative genomics analyses related to the *S. maritima* genome. Columns include, in this order, scientific names, the species code according to UNIPROT, the number of the longest unique

transcript used in the analyses, the data source, and the date in which data were retrieved.

(DOCX)

Table S3 Orthologues detected between a given species and *S. maritima*. First column indicates how many trees have been used to detect such orthologues. Columns “uniq” refers to the number of orthologues detected for each pair of species after removing redundancy. In one-to-many and many-to-many orthology relationships it is possible to count a given protein more than once. Regarding the ratios values, “all” column refers to the orthology ratio computed using all orthologue pairs meanwhile “uniq” refers to the ratio computed using “uniq” columns.

(DOCX)

Table S4 Orthology ratios for a given species related to *S. maritima*. This table is similar to Table S3, but in this case orthology relationships with ten or more proteins for any of the species are discarded in order to avoid biases introduced by species-specific gene family expansions.

(DOCX)

Table S5 Newly added Chelicerata species used to increase the taxon sampling for the species phylogeny.

First column indicates the scientific species name, the second one indicates which strategy has been used to identify single copy protein-coding genes. Third column shows how many single-copy genes have been identified in each species from the initial set of 1,491 used to reconstruct the species phylogeny. Last two columns show the data source and the date on which data were retrieved.

(DOCX)

Table S6 Results after applying the different statistical tests implemented in CONSEL for the alternative placement of *S. maritima* relative to Pancrustacea and Chelicerata groups of species (as shown in Figure S4) in the context of the 18 species used for the phylogenomics analyses.

The “item” column relates to Figure S4 as follows: (1) topology arrangement corresponding to Figure S4 left-hand panel, in which *S. maritima* was grouped with Chelicerata species. (2) Topology arrangement corresponding to Figure S4 central panel, in which *S. maritima* branches off before the split of Pancrustacea and Chelicerata. (3) Topology arrangement corresponding to Figure S4 right-hand panel, in which *S. maritima* was grouped with Pancrustacea species.

(DOCX)

Table S7 Results after applying the different statistical tests implemented in CONSEL for the alternative placement of *S. maritima* relative to the two arthropod groups, Pancrustacea and Chelicerata (as shown in Figure S4), with the inclusion of extra chelicerates.

Taxon sampling for the Chelicerata was increased after including sequences from five additional species. In order to reduce any potential bias introduced by distant and/or fast-evolving out-groups, six out-group species from the initial set were removed. The “item” column relates to Figure S4 as follows: (1) topology arrangement corresponding to Figure S4 left-hand panel, in which *S. maritima* was grouped with Chelicerata species. (2) Topology arrangement corresponding to Figure S4 central panel, in which *S. maritima* branches off before the split of Pancrustacea and Chelicerata. (3) Topology arrangement corresponding to Figure S4 right-hand panel, in which *S. maritima* was grouped with Pancrustacea species.

(DOCX)

Table S8 Enriched functional GO Terms for the ten largest clusters of duplicated *S. maritima* protein-coding genes specifically expanded in the centipede lineage, as compared with the whole genome.

(DOCX)

Table S9 Statistics regarding the duplications of centipede genes relative to seven specific ages detected using all available trees on the phylome.

(DOCX)

Table S10 Enriched functional GO terms for proteins duplicated at the different relative ages shown in Table S9. Columns show relative age, gene ontology namespace, the GO term id, and its name, respectively.

(DOCX)

Table S11 Overview of *S. maritima* mitochondrial genome.

(DOCX)

Table S12 Species used in the synteny analyses and the sources of their sequence data.

(DOCX)

Table S13 Block-synteny summary statistics for pairs of species. Hs, *Homo sapiens*; Bf, *B. floridae*; Sm, *S. maritima*; Lg, *Lottia gigantea*; Ct, *Capitella teleta*; Nv, *N. vectensis*; Ta, *Trichoplax adhaerens*; Ag, *Anopheles gambiae*; Bm, *B. mori*.

(DOCX)

Table S14 Summary of numbers of homeobox genes per class of *Strigamia*, *Branchiostoma*, and *Tribolium*.

(DOCX)

Table S15 Names and identification numbers of all *S. maritima* homeobox genes along with their orthologues from the beetle, *T. castaneum*, and amphioxus, *B. floridae*.

(XLS)

Table S16 One-to-one *S. maritima* to human orthologues starting from genes on *S. maritima* scaffold 48457, which contains *SmaHox3a*. The third column is the chromosomal location of the human orthologue. Human Hox chromosomes are 2, 7, 12, and 17 and the ParaHox chromosomes are 4, 5, 13, and X.

(DOCX)

Table S17 Evolutionary conservation of RNA processing modes in the *S. maritima* and *D. melanogaster* Hox clusters. Type of RNA processing event concerning each one of the *S. maritima* (left) and *D. melanogaster* (right) Hox genes. We note that orthologous genes in both species undergo similar types of RNA processing: the three posterior-most Hox genes: *Ubx*, *abd-a*, and *Abd-b* display a specific type of APA (tandem APA) in both *S. maritima* and *D. melanogaster* (conserved patterns highlighted by red asterisks) providing an example of what might be a feature present in the ancestral Hox cluster to insects and myriapods. Nonetheless, for most other Hox genes, RNA processing patterns differ markedly between *S. maritima* and *D. melanogaster*, indicating that the conserved incidence of alternative RNA processing across arthropods can only be proposed for the posterior-most Hox genes.

(PDF)

Table S18 Details of SmGr family genes and proteins. Columns are: Gene, the gene and protein name we are assigning (suffixes are PSE, pseudogene; FIX, assembly was repaired; JOI, gene model spans scaffolds); OGS, the official gene number in the

13,233 proteins (prefix is Smar_temp_); Scaffold, the genome assembly scaffold ID, prefix is scf718000 (amongst 14,739 scaffolds in assembly Smar05272011); Coordinates, the nucleotide range from the first position of the start codon to the last position of the stop codon in the scaffold; Strand – + is forward and – is reverse; introns, number of introns; ESTs, presence of an EST contig with appropriate splicing in one of the three transcriptome assemblies (F, female; M, male; E, eggs); AAs, number of encoded amino acids in the protein; comments, comments on the OGS gene model, repairs to the genome assembly, and pseudogene status (numbers in parentheses are the number of obvious pseudogenizing mutations).

(DOC)

Table S19 Total numbers of biogenic amine receptors in different species.

(DOCX)

Table S20 A comparison between the *D. melanogaster* and *S. maritima* biogenic amine receptors. The orthologues are given next to each other. When there is no orthologue, a dash (–) is written instead.

(XLSX)

Table S21 Genes encoding neuropeptide precursors and neuropeptide receptors annotated in *S. maritima*.

Abbreviations: ACP, adipokinetic hormone/corazonin-related neuropeptide; AKH, adipokinetic hormone; ADF, antidiuretic factor; AST, allatostatin; CCAP, crustacean cardio-active peptides; DH (Calc.-like), calcitonin-like diuretic hormone; DH (CRF-like), corticotropin releasing factor-like diuretic hormone; EH, eclosion hormone; ETH, ecdysis triggering hormone; GPA2, glycoprotein hormone A2; GPB5, glycoprotein hormone B5; ILP, insulin-like peptide; ITP, ion transport peptide; NPF, neuropeptide F; NPLP, neuropeptide-like precursor; PDF, pigment dispersing factor; PTTH, prothoracicotropic hormone; sNPF, short neuropeptide F.

(EPS)

Table S22 Presence or absence of neuropeptide signaling systems in arthropods. The centipede *S. maritima* contains two CCHamide-1, two eclosion hormone and two FMRFamide genes (2 p). In some cases neuropeptide precursors could not be identified, but the corresponding receptor genes are present (R). We assume that this is due to sequencing gaps. For abbreviations see Table S21.

(DOC)

Table S23 Genes commonly implicated in arthropod juvenoids biosynthesis (green) and degradation (blue), and their potential regulators (purple) [98–101]. Common abbreviations, and presence in the centipede *S. maritima*.

(DOCX)

Table S24 List of genes commonly implicated as potential regulators of arthropod juvenoids biosynthesis (purple) [98–101]. Common abbreviations, and presence in the centipede *S. maritima*.

(DOCX)

Table S25 Wnt genes in the genome of *S. maritima*. SMAR, the gene identification number, and scaffold, the scaffold identification number. Wnt 1, 6, and 10 are clustered together on the same scaffold (yellow highlighting), which is likely a remnant of the ancestral wnt gene cluster (see text for details).

(PDF)

Table S26 Selenoproteins in the *S. maritima* genome.

(DOCX)

Table S27 Histone encoding loci of *S. maritima*.
(DOCX)

Table S28 Number of loci within the genomes of arthropod species encoding the five classes of histones. Orthologues for *A. aegypti*, *D. pulex*, *T. urticae*, and *I. scapularis* were obtained by BLAST analysis. Orthologues for *A. mellifera* and *A. pisum* were obtained from published literature [108,109].
(DOCX)

Table S29 Germ line and RNAi genes annotated in the *S. maritima* genome. The name of the *Drosophila* orthologue is shown unless indicated with “(Mo),” for mouse.
(DOCX)

Table S30 Details of the manually annotated genes of *S. maritima*.
(XLSX)

File S1 One2One_GOTerms_GenomeIDs for Orthology-based functional annotation.
(XLSX)

File S2 Strigamia_pals for Figure 3.
(XLSX)

File S3 Gustatory receptor sequences.
(XLSX)

File S4 Raw data for Figure 2, Figure 9, Figure S1, and Figure S5.
(XLSX)

File S5 Raw data for Figure S28.
(XLSX)

File S6 Raw data for Figure S29.
(XLSX)

File S7 Raw data for Figure S30.
(XLSX)

Text S1 Supporting Methods Text.
(DOCX)

Acknowledgments

We thank Paul Kersey, Monica Munoz-Torres, and Jamie Walters for sharing their experience of community annotation projects; Rolf Sommer and Werner Mayer for assistance with the identification of *S. maritima* associated nematode sequences; Nipam Patel and all authors of the NHGRI Ecdysozoan Sequencing Proposal who initiated this project; P. Woznicki and F. Marec for sharing data on the karyotype of *S. maritima*; Geordie and Irene at BlarMhor for shelter and sustenance during the field collection of centipedes.

Author Contributions

The author(s) have made the following declarations about their contributions: Conceived and designed the experiments: MA SR. Performed the experiments: CB SES SNJ NS LLP SP XZ SG KPB SLL IN YW VK GO RM CP LF DS DNN PA MC LJ CM TM MJ RT CLK MH MJ FO YH JQ SR KCW HJ DSTH DL DK DMM MA RAG. Analyzed the data: HJ SR KCW DSTH DL CB TG SCG NHP PH JL MTO NZ JCJB DSTH DL DEKF OMR VSH KWS AS ZA RS JEG PP CRA WA LH CE PKD EJD LDP DE DB PDE CGE TEJ CJPG FH JHLH NJK FMJ WJP GM GE MM RG ACR MJT FL HER MR SGJ MN JR ASG FCA HMR ES FBK SH TSK TN AVH KTR MVDZ CR JPH JHW AMS EAGH JMW WJG ADC. Contributed reagents/materials/analysis tools: RAG MA DMM CB DSTH DL MTO NZ JCJB DSTH DL TG SCG NHP PH JL DK. Wrote the paper: ADC DEKF RS MA SR. Project management and senior authors: ADC DEKF RS MA SR. Sequencing PI: RAG. Centipede PI: MA. Specimen identification and preparation: CB. Sequencing operations manager: DMM. Sequencing project management: SES SNJ. Library: NS L-LP SP XZ SG KPB SLL IN YW. 454 sequencing: VK GO RM CP LF DS DNN PA MC LJ CM TM MJ RT CLK MH. Illumina sequencing: MJ FO YH. Assembly: JQ SR KCW. Automated annotation: HJ SR KCW DSTH DL. Submissions: DK. PhyloDB: TG SC-G. Chromosomal synteny conservation: NHP PH JL. Manual annotation: organization: MT-O NZ JCJB DSTH DL. Homeobox genes: DEKF OMR VSH KWS AS ZA. FGF signaling: RS. Sex chromosomes: JEG. Hox mRNA: PP CRA. wnt signaling: WA LH RS CE. Conserved gene clusters and methylation: PKD EJD. Dscam: LDP DE DB. Biogenic amine receptors: PDE. Germline genes: CGE TEJ. Neuropeptides and receptors: CJPG FH. Juvenile hormone systems: JHLH NJK. Immunity: FMJ WJP. Kinome: GM GE. Selenoproteins: MM RG. Mitochondria: ACR MJT FL HER. MiRNAs: MR SG-J MN. Chemosensory genes: JR AS-G FCA HMR. Repetitive elements: ES FBK SH. Light perception and circadian clock: TSK TN AVH KT-R. Innexins and TGF beta: MVDZ CR JPH. Cuticular proteins: JHW. Meiosis genes: AMS EAG-H JMW. Developmental transcription factors: WJG ADC VH JEG CB ZA.

References

1. Arthropod Genomes Consortium (2014) List of sequenced arthropod genomes. Available: http://arthropodgenomes.org/wiki/Sequenced_genomes.
2. Bracken-Grissom H, Collins AG, Collins T, Crandall K, Distel D, et al. (2014) The Global Invertebrate Genomics Alliance (GIGA): developing community resources to study diverse invertebrate genomes. *J Hered* 105: 1–18.
3. Edgecombe GD (2011) Phylogenetic relationships of Myriapoda. Minelli A, editor. *The Myriapoda*. Leiden: Brill. pp. 1–20.
4. Giribet G, Edgecombe GD, Wheeler WC (2001) Arthropod phylogeny based on eight molecular loci and morphology. *Nature*: 157–160.
5. Rota-Stabelli O, Telford MJ (2008) A multi criterion approach for the selection of optimal outgroups in phylogeny: recovering some support for Mandibulata over Myriochelata using mitogenomics. *Mol Phylogenet Evol* 48: 103–111.
6. Edgecombe GD, Giribet G (2007) Evolutionary biology of centipedes (Myriapoda: Chilopoda). *Ann Rev Entomol* 52: 151–170.
7. Simakov O, Marletaz F, Cho SJ, Edsinger-Gonzales E, Havlak P, et al. (2013) Insights into bilaterian evolution from three spiralian genomes. *Nature* 493: 526–531.
8. Edgecombe GD (2004) Morphological data, extant Myriapoda, and the myriapod stem-group. *Contrib Zool* 73: 207–252.
9. Bitsch C, Bitsch J (2004) Phylogenetic relationships of basal hexapods among the mandibulate arthropods: a cladistic analysis based on comparative morphological characters. *Zool Scr* 33: 511–550.
10. Rota-Stabelli O, Daley AC, Pisani D (2013) Molecular timetrees reveal a Cambrian colonization of land and a new scenario for ecdysozoan evolution. *Curr Biol* 23: 392–398.
11. Scholtz G, Edgecombe GD (2006) The evolution of arthropod heads: reconciling morphological, developmental and palaeontological evidence. *Dev Genes Evol* 216: 395–415.
12. Mallatt JM, Garey JR, Shultz JW (2004) Ecdysozoan phylogeny and Bayesian inference: first use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. *Mol Phylogenet Evol* 31: 178–191.
13. Pisani D, Poling LL, Lyons-Weiler M, Hedges SB (2004) The colonization of land by animals: molecular phylogeny and divergence times among arthropods. *BMC Biol* 2: 1.
14. Bourlat SJ, Nielsen C, Economou AD, Telford MJ (2008) Testing the new animal phylogeny: a phylum level molecular analysis of the animal kingdom. *Mol Phylogenet Evol* 49: 23–31.
15. Rota-Stabelli O, Campbell L, Brinkmann H, Edgecombe GD, Longhorn SJ, et al. (2011) A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. *Proc Roy Soc B* 278: 298–306.
16. Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, et al. (2010) Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature* 463: 1079–1083.
17. Rehm P, Meusemann K, Börner J, Misof B, Burmester T (2014) Phylogenetic position of Myriapoda revealed by 454 transcriptome sequencing. *Mol Phylogenet Evol*.
18. Kraus O, Kraus M (1994) Phylogenetic system of the Tracheata (Mandibulata): on “Myriapoda”: Insecta interrelationships, phylogenetic age and primary ecological niches. *Verh Naturwiss Ver Hambg* 34: 5–31.

19. Cook CE, Smith ML, Telford MJ, Bastianello A, Akam M (2001) Hox genes and the phylogeny of the arthropods. *Curr Biol* 11: 759–763.
20. Cook CE, Yue Q, Akam M (2005) Mitochondrial genomes suggest that hexapods and crustaceans are mutually paraphyletic. *Proc Biol Sci* 272: 1295–1304.
21. Regier JC, Shultz JW, Kambic RE (2005) Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic. *Proc Biol Sci* 272: 395–401.
22. Gregory TR (2014) Animal Genome Size Database. Available: <http://www.genomesize.com>.
23. Arthur W, Chipman AD (2005) The centipede *Strigamia maritima*: what it can tell us about the development and evolution of segmentation. *Bioessays* 27: 653–660.
24. Brena C, Akam M (2012) The embryonic development of the centipede *Strigamia maritima*. *Dev Biol* 363: 290–307.
25. Lewis JGE (1961) The life history and ecology of the littoral centipede *Strigamia* (= *Scolioptanes*) *maritima* (Leach). *Proc Zool Soc Lond* 137: 221–248.
26. Chipman AD, Akam M (2008) The segmentation cascade in the centipede *Strigamia maritima*: involvement of the Notch pathway and pair-rule gene homologues. *Dev Biol* 319: 160–169.
27. Chipman AD, Arthur W, Akam M (2004) Early development and segment formation in the centipede *Strigamia maritima* (Geophilomorpha). *Evol Dev* 6: 78–89.
28. Chipman AD, Arthur W, Akam M (2004) A double segment periodicity underlies segment generation in centipede development. *Curr Biol* 14: 1250–1255.
29. Green J, Akam M (2013) Evolution of the pair rule gene network: Insights from a centipede. *Dev Biol* 382: 235–245.
30. Kettle C, Johnstone J, Jowett T, Arthur H, Arthur W (2003) The pattern of segment formation, as revealed by *engrailed* expression, in a centipede with a variable number of segments. *Evol Dev* 5: 198–207.
31. Brena C, Green J, Akam M (2013) Early embryonic determination of the sexual dimorphism in segment number in geophilomorph centipedes. *Evodevo* 4: 22.
32. Brena C, Akam M (2013) An analysis of segmentation dynamics throughout embryogenesis in the centipede *Strigamia maritima*. *BMC Biology* 11: 112.
33. Vedel V, Apostolou Z, Arthur W, Akam M, Brena C (2010) An early temperature-sensitive period for the plasticity of segment number in the centipede *Strigamia maritima*. *Evol Dev* 12: 347–352.
34. Giribet G, Carranza S, Riutort M, Baguña J, Ribera C (1999) Internal phylogeny of the Chilopoda (Myriapoda, Arthropoda) using complete 18S rDNA and partial 28S rDNA sequences. *Phil Trans Roy Soc Lond B* 354: 215–222.
35. Mundel P (1979) The centipedes (Chilopoda) of the Mazon Creek. Nitecki MH, editor. Mazon Creek fossils. New York: Academic Press. pp. 361–378.
36. Minelli A (2011) Chilopoda – general morphology. Minelli A, editor. *The Myriapoda*. Leiden: Brill. pp. 43–66.
37. Müller CHG, Sombke A, Hilken G, Rosenberg J (2011) Chilopoda – sense organs. Minelli A, editor. *The Myriapoda*. Leiden: Brill. pp. 235–278.
38. Plateau F (1886) Recherches sur la perception de la lumière par les Myriapodes aveugles. *J Anat Physiol* 22: 431–457.
39. Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, et al. (2012) The *Drosophila melanogaster* genetic reference panel. *Nature* 482: 173–178.
40. Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, Denisov I, Kormes D, et al. (2011) PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nuc Acid Res* 39: D556–D560.
41. Gabaldón T (2008) Large-scale assignment of orthology: back to phylogenetics? *Genome Biol* 9: 235.
42. Huerta-Cepas J, Gabaldón T (2011) Assigning duplication events to relative temporal scales in genome-wide studies. *Bioinformatics* 27: 38–45.
43. Negrísolo E, Minelli A, Valle G (2004) The mitochondrial genome of the house centipede *Scutigera* and the monophyly versus paraphyly of myriapods. *Mol Biol Evol* 21: 770–780.
44. Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature* 453: 1064–1071.
45. Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, et al. (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* 317: 86–94.
46. Zdobnov EM, von Mering C, Letunic I, Bork P (2005) Consistency of genome-based methods in measuring metazoan evolution. *FEBS Lett* 579: 3355–3361.
47. Denoeud F, Henriot S, Mungpakdee S, Aury J-M, Da Silva C, et al. (2010) Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science* 330: 1381–1385.
48. Panfilio KA, Akam M (2007) A comparison of Hox3 and Zen protein coding sequences in taxa that span the *Hox3/zen* divergence. *Dev Genes Evol* 217: 323–329.
49. Garcia-Fernandez J (2005) The genesis and evolution of homeobox gene clusters. *Nat Rev Genet* 6: 881–892.
50. Hui JHL, McDougall C, Monteiro AS, Holland PWH, Arendt D, et al. (2012) Extensive chordate and annelid macrosynteny reveals ancestral homeobox gene organization. *Mol Biol Evol* 29: 157–165.
51. Pollard SL, Holland PWH (2000) Evidence for 14 homeobox gene clusters in human genome ancestry. *Curr Biol* 10: 1059–1062.
52. Butts T, Holland PWH, Ferrier DE (2008) The Urbilateria Super-Hox cluster. *Trends Genet* 24: 259–262.
53. Penalva-Arana DC, Lynch M, Robertson HM (2009) The chemoreceptor genes of the waterflea *Daphnia pulex*: many Grs but no Ors. *BMC Evol Biol* 9: 79.
54. Robertson HM, Warr CG, Carlson JR (2003) Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*. *P Natl Acad Sci U S A* 100: 14537–14542.
55. Vieira FG, Rozas J (2011) Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: Origin and evolutionary history of the chemosensory system. *Genome Biol Evol* 3: 476–490.
56. Pelosi P (1994) Odorant-binding proteins. *Crit Rev Biochem Mol* 29: 199–228.
57. Vogt RG, Riddiford LM (1981) Pheromone binding and inactivation by moth antennae. *Nature* 293: 161–163.
58. Angeli S, Ceron F, Scaloni A, Monti M, Monteforti G, et al. (1999) Purification, structural characterization, cloning and immunocytochemical localization of chemoreception proteins from *Schistocerca gregaria*. *Eur J Biochem* 262: 745–754.
59. Pelosi P, Zhou JJ, Ban LP, Calvello M (2006) Soluble proteins in insect chemical communication. *Cell Mol Life Sci* 63: 1658–1676.
60. Starostina E, Xu AG, Lin HP, Pikielny CW (2009) A *Drosophila* protein family implicated in pheromone perception is related to Tay-Sachs GM2-activator protein. *J Biol Chem* 284: 585–594.
61. Xu A, Park SK, D'Mello S, Kim E, Wang Q, et al. (2002) Novel genes expressed in subsets of chemosensory sensilla on the front legs of male *Drosophila melanogaster*. *Cell Tissue Res* 307: 381–392.
62. Clyne PJ, Warr CG, Carlson JR (2000) Candidate taste receptors in *Drosophila*. *Science* 287: 1830–1834.
63. Scott K, Brady R, Cravchik A, Morozov P, Rzhetsky A, et al. (2001) A chemosensory gene family encoding candidate gustatory and olfactory receptors in *Drosophila*. *Cell* 104: 661–673.
64. Clyne PJ, Warr CG, Freeman MR, Lessing D, Kim JH, et al. (1999) A novel family of divergent seven-transmembrane proteins: candidate odorant receptors in *Drosophila*. *Neuron* 22: 327–338.
65. Gao Q, Chess A (1999) Identification of candidate *Drosophila* olfactory receptors from genomic DNA sequence. *Genomics* 60: 31–39.
66. Benton R, Vannice KS, Gomez-Diaz C, Vossell LB (2009) Variant ionotropic glutamate receptors as chemosensory receptors in *Drosophila*. *Cell* 136: 149–162.
67. Croset V, Rytz R, Cummins SF, Budd A, Brawand D, et al. (2010) Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. *PLoS Genet* 6: e1001064.
68. Weil E (1958) Zur Biologie der einheimischen Geophiliden. *Z Angew Entomol* 42: 173–209.
69. Xiang Y, Yuan QA, Vogt N, Looger LL, Jan LY, et al. (2010) Light-avoidance-mediating photoreceptors tile the *Drosophila* larval body wall. *Nature* 468: 921–926.
70. Zhan S, Merlin C, Boore JL, Reppert SM (2011) The monarch butterfly genome yields insights into long-distance migration. *Cell* 147: 1171–1185.
71. Benna C, Bonaccorsi S, Wulbeck C, Helfrich-Forster C, Gatti M, et al. (2010) *Drosophila timeless2* Is required for chromosome stability and circadian photoreception. *Curr Biol* 20: 346–352.
72. George H, Terracol R (1997) The *vrrile* gene of *Drosophila* is a maternal enhancer of *decapentaplegic* and encodes a new member of the bZIP family of transcription factors. *Genetics* 146: 1345–1363.
73. Reddy KL, Rovani MK, Wohlwill A, Katzen A, Storti RV (2006) The *Drosophila* Par domain protein I gene, *Pdp1*, is a regulator of larval growth, mitosis and endoreplication. *Dev Biol* 289: 100–114.
74. Avivi A, Albrecht U, Oster H, Joel A, Beiles A, et al. (2001) Biological clock in total darkness: The Clock/MOP3 circadian system of the blind subterranean mole rat. *Proc Natl Acad Sci U S A* 98: 13751–13756.
75. Avivi A, Oster H, Joel A, Beiles A, Albrecht U, et al. (2004) Circadian genes in a blind subterranean mammal III: molecular cloning and circadian regulation of cryptochrome genes in the blind subterranean mole rat, *Spalax ehrenbergi* superspecies. *J Biol Rhyth* 19: 22–34.
76. Goldman BD, Goldman SL, Riccio AP, Terkel J (1997) Circadian patterns of locomotor activity and body temperature in blind mole-rats, *Spalax ehrenbergi*. *J Biol Rhyth* 12: 348–361.
77. Grandall KA, Hillis DM (1997) Rhodopsin evolution in the dark. *Nature* 387: 667–668.
78. Willis JH (2010) Structural cuticular proteins from arthropods: annotation, nomenclature, and sequence characteristics in the genomics era. *Insect Biochem Molec Biol* 40: 189–204.
79. Rebers JE, Riddiford LM (1988) Structure and expression of a *Manduca sexta* larval cuticle gene homologous to *Drosophila* cuticle genes. *J Mol Biol* 203: 411–423.
80. Rebers JE, Willis JH (2001) A conserved domain in arthropod cuticular proteins binds chitin. *Insect Biochem Molec Biol* 31: 1083–1093.
81. Fredriksson R, Schiöth HB (2005) The repertoire of G-protein-coupled receptors in fully sequenced genomes. *Mol Pharmacol* 67: 1414–1425.

82. Ritter SL, Hall RA (2009) Fine-tuning of GPCR activity by receptor-interacting proteins. *Nat Rev Mol Cell Bio* 10: 819–830.
83. Hill RJ, Billas IML, Bonneton F, Graham LD, Lawrence MC (2013) Ecdysone Receptors: from the Ashburner model to structural biology. *Annu Rev Entomol* 58: 251–271.
84. Jindra M, Palli SR, Riddiford LM (2013) The juvenile hormone signaling pathway in insect development. *Annu Rev Entomol* 58: 181–204.
85. Srivastava DP, Yu EJ, Kennedy K, Chatwin H, Reale V, et al. (2005) Rapid, nongenomic responses to ecdysteroids and catecholamines mediated by a novel *Drosophila* G-protein-coupled receptor. *J Neurosci* 25: 6145–6155.
86. Evans PD, Maqueira B (2005) Insect octopamine receptors: a new classification scheme based on studies of cloned *Drosophila* G-protein coupled receptors. *Invert Neurosci* 5: 111–118.
87. Hauser F, Neupert S, Williamson M, Predel R, Tanaka Y, et al. (2010) Genomics and peptidomics of neuropeptides and protein hormones present in the parasitic wasp *Nasonia vitripennis*. *J Proteome Res* 9: 5296–5310.
88. Hauser F, Cazzamali G, Williamson M, Park Y, Li B, et al. (2008) A genome-wide inventory of neurohormone GPCRs in the red flour beetle *Tribolium castaneum*. *Front Neuroendocrin* 29: 142–165.
89. Stay B, Tobe SS (2007) The role of allatostatins in juvenile hormone synthesis in insects and crustaceans. *Annu Rev Entomol* 52: 277–299.
90. Weaver RJ, Audsley N (2009) Neuropeptide regulators of juvenile hormone synthesis: structures, functions, distribution, and unanswered questions. *Trends Comp Endocrinol Neuro* 1163: 316–329.
91. Grbic M, Van Leeuwen T, Clark RM, Rombauts S, Rouze P, et al. (2011) The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. *Nature* 479: 487–492.
92. Hui JHL, Hayward A, Bendena WG, Takahashi T, Tobe SS (2010) Evolution and functional divergence of enzymes involved in sesquiterpenoid hormone biosynthesis in crustaceans and insects. *Peptides* 31: 451–455.
93. Van der Zee M, da Fonseca RN, Roth S (2008) TGF beta signaling in *Tribolium*: vertebrate-like components in a beetle. *Dev Genes Evol* 218: 203–213.
94. Lowery JW, LaVigne AW, Kokabu S, Rosen V (2013) Comparative genomics identifies the mouse *Bmp3* promoter and an upstream evolutionary conserved region (ECR) in mammals. *PLoS ONE* 8: e57840.
95. Cho SJ, Valles Y, Giani VC, Seaver EC, Weisblat DA (2010) Evolutionary dynamics of the wnt gene family: a lophotrochozoan perspective. *Mol Biol Evol* 27: 1645–1658.
96. Prud'homme B, Lartillot N, Balavoine G, Adoutte A, Vervoort M (2002) Phylogenetic analysis of the Wnt gene family: insights from lophotrochozoan members. *Curr Biol* 12: 1395–1400.
97. Janssen R, Le Gouar M, Pechmann M, Poulin F, Bolognesi R, et al. (2010) Conservation, loss, and redeployment of Wnt ligands in protostomes: implications for understanding the evolution of segment formation. *Bmc Evolutionary Biology* 10: 374.
98. Murat S, Hopfen C, McGregor AP (2010) The function and evolution of Wnt genes in arthropods. *Arthropod Struct Dev* 39: 446–452.
99. Nusse R (2001) An ancient cluster of Wnt paralogues. *Trends Genet* 17: 443–443.
100. McGinnis N, Ragnhildstveit E, Veraksa A, McGinnis W (1998) A cap 'n' collar protein isoform contains a selective Hox repressor function. *Development* 125: 4553–4564.
101. Iwanaga S, Lee BL (2005) Recent advances in the innate immunity of invertebrate animals. *J Biochem Mol Biol* 38: 128–150.
102. Hoffmann JA, Kafatos FC, Janeway CA, Ezekowitz RAB (1999) Phylogenetic perspectives in innate immunity. *Science* 284: 1313–1318.
103. Lemaitre B, Hoffmann J (2007) The host defense of *Drosophila melanogaster*. *Annu Rev Immunol* 25: 697–743.
104. Dong YM, Dimopoulos G (2009) *Anopheles* fibrinogen-related proteins provide expanded pattern recognition capacity against bacteria and malaria parasites. *J Biol Chem* 284: 9835–9844.
105. Waterhouse RM, Kriventseva EV, Meister S, Xi ZY, Alvarez KS, et al. (2007) Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science* 316: 1738–1743.
106. Watson FL, Puttmann-Holgado R, Thomas F, Lamar DL, Hughes M, et al. (2005) Extensive diversity of Ig-superfamily proteins in the immune system of insects. *Science* 309: 1874–1878.
107. Brites D, Brena C, Ebert D, Du Pasquier L (2013) More than one way to produce protein diversity: duplication and limited alternative splicing of an adhesion molecule gene in basal arthropods. *Evolution* 67: 2999–3011.
108. Obbard DJ, Gordon KHJ, Buck AH, Jiggins FM (2009) The evolution of RNAi as a defence against viruses and transposable elements. *Philos Trans Roy Soc B* 364: 99–115.
109. Squires JE, Berry MJ (2008) Eukaryotic selenoprotein synthesis: mechanistic insight incorporating new factors and new functions for old factors. *IUBMB Life* 60: 232–235.
110. Mariotti M, Ridge PG, Zhang Y, Lobanov AV, Pringle TH, et al. (2012) Composition and evolution of the vertebrate and mammalian selenoproteomes. *PLoS ONE* 7: e33066.
111. Chapple CE, Guigo R (2008) Relaxation of selective constraints causes independent selenoprotein extinction in insect genomes. *PLoS ONE* 3: e2968.
112. Kim HY, Fomenko DE, Yoon YE, Gladyshev VN (2006) Catalytic advantages provided by selenocysteine in methionine-S-sulfoxide reductases. *Biochemistry* 45: 13697–13704.
113. Corona M, Robinson GE (2006) Genes of the antioxidant system of the honey bee: annotation and phylogeny. *Insect Mol Biol* 15: 687–701.
114. Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, et al. (2010) Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A* 107: 8689–8694.
115. Suzuki MM, Kerr AR, De Sousa D, Bird A (2007) CpG methylation is targeted to transcription units in an invertebrate genome. *Genome Res* 17: 625–631.
116. Zemach A, McDaniel IE, Silva P, Zilberman D (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* 328: 916–919.
117. Foret S, Kucharski R, Pellegrini M, Feng S, Jacobsen SE, et al. (2012) DNA methylation dynamics, metabolic fluxes, gene splicing, and alternative phenotypes in honey bees. *Proc Natl Acad Sci U S A* 109: 4968–4973.
118. Laurent L, Wong E, Li G, Huynh T, Tsigiris A, et al. (2010) Dynamic changes in the human methylome during differentiation. *Genome Res* 20: 320–331.
119. Elango N, Hunt BG, Goodisman MA, Yi SV (2009) DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, *Apis mellifera*. *Proc Natl Acad Sci U S A* 106: 11206–11211.
120. Hunt BG, Brisson JA, Yi SV, Goodisman MAD (2010) Functional conservation of DNA methylation in the pea aphid and the honeybee. *Genome Biol Evol* 2: 719–728.
121. Park J, Peng ZG, Zeng J, Elango N, Park T, et al. (2011) Comparative analyses of DNA methylation and sequence evolution using *Nasonia* genomes. *Mol Biol Evol* 28: 3345–3354.
122. Richards S, Gibbs RA, Weinstock GM, Brown SJ, Denell R, et al. (2008) The genome of the model beetle and pest *Tribolium castaneum*. *Nature* 452: 949–955.
123. Kriauconis S, Heintz N (2009) The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* 324: 929–930.
124. Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, et al. (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* 324: 930–935.
125. Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nuc Acid Res* 39: D152–D157.
126. Wheeler BM, Heimberg AM, Moy VN, Sperling EA, Holstein TW, et al. (2009) The deep evolution of metazoan microRNAs. *Evol Dev* 11: 50–68.
127. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, et al. (2000) The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* 403: 901–906.
128. Christodoulou F, Raible F, Tomer R, Simakov O, Trachana K, et al. (2010) Ancient animal microRNAs and the evolution of tissue identity. *Nature* 463: 1084–1088.
129. Caygill EE, Johnston LA (2008) Temporal regulation of metamorphic processes in *Drosophila* by the *let-7* and *miR-125* heterochronic microRNAs. *Curr Biol* 18: 943–950.
130. Marco A, Hui JHL, Ronshaugen M, Griffiths-Jones S (2010) Functional shifts in insect microRNA evolution. *Genome Biol Evol* 2: 686–696.
131. McTaggart SJ, Conlon C, Colbourne JK, Blaxter ML, Little TJ (2009) The components of the *Daphnia pulex* immune system as revealed by complete genome sequencing. *BMC Genomics* 10.
132. Dasmahapatra KK, Walters JR, Briscoe AD, Davey JW, Whibley A, et al. (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* 487: 94–98.

Supporting Information index

- 1. Supplemental Material and Methods**
- 2. Repetitive element isolation and classification.**
- 3. Phylome analysis and phylogenomics.**
- 4. Synteny methodology.**
- 5. Homeobox genes.**
- 6. Non-homeobox gene clusters: innexin, Runt, E(spl)-C.**
- 7. Chemosensation: Gustatory receptors (GRs).**
- 8. Developmental signalling systems.**
- 9. Histones and Histone modifying enzymes in *S. maritima*.**
- 10. Germ line genes.**
- 11. Meiosis genes.**
- 12. CpG methylation.**
- 13. Non-protein-coding RNA genes in the *S. maritima* genome.**
- 14. mRNA purification and sequencing library construction.**
- 15. References.**
- 16. Supporting Information Figure Legends.**
- 17. Supporting Information Table Legends.**
- 18. Supporting Information data files.**

Supporting Information

1. Supplemental Material and Methods

Genome sequencing and annotation

S. maritima raw sequence, and assembled genome sequence data is available at the NCBI under bioproject [PRJNA20501](#), Assembly ID GCA_000239455.1. The genome was sequenced using 454 sequencing technology. Three whole genome shotgun libraries were used to produce the data, a 454 Titanium fragment library (made from DNA isolated from a single male) and 454 paired end libraries with targets of 3kb and 8 kb mate pair insert sizes (made from DNA from ~20 pooled individuals of both sexes). About 22.2 million reads were assembled, representing about 8,190 Mb of sequence and about 45.5x coverage of the *S. maritima* genome.

The genome was assembled using the CABOG Celera assembler (Celera 6.1) to yield a total of 173.6 MB of assembled sequence, to yield genome release Smar 1.0. Illumina data was used to correct homo-polymer errors originating from the 454 data. This assembly comprises 24,087 contiguous sequence fragments (contigs), with an N50 size of 24.7kb, linked by the paired end reads into 14,745 scaffolds with an N50 size of 139.4 kb. (The N50 size is the length such that 50% of the assembled genome lies in blocks of the N50 size or longer.) When the estimated gaps between contigs in scaffolds are included, the total span of the assembly is 176.2 Mb. This assembled sequence omits many repeat sequences, which probably account for the difference between the assembly length and the prior genome size estimate based on feulgen image analysis densitometry, of 290Mb [1]. In fact approximately 42% of the input reads remained un-assembled, which on a raw data basis would predict a total of 58% of the genome or ~168Mb would be accounted for in the assembly, close to our actual assembly size of 176.2Mb.

Population sequencing

We sequenced at 25X coverage four individuals (females A, B and D – SRA accessions SRX326837, SRX326839 and SRX326840 respectively – and male J, accession SRX326841) collected in Brora, northern Scotland June 2009, and subsequently starved for 30 days. DNA was extracted from whole individuals using Qiagen Genomic DNA extraction kit. Sequence was generated on the Illumina GAI and HiSeq platforms. Paired 95bp reads were aligned to Smar1.0 using BWA, indels were locally realigned using GATK. SNPs and indels were called using the GATK unified genotyper with standard parameters. SNP density was calculated using VCF tools using the –SNPdensity option.

RNA sequencing

Total RNA was isolated from adult males and females collected and treated as in the previous paragraph, and liquid nitrogen frozen mixed eggs (107) from 7 clutches, collected during the 2006, 2007 and 2009 collections, using Qiagen mini RNA extraction kit. mRNA was purified from total RNA using Dynabeads® mRNA Purification Kit, first

strand cDNA was reverse transcribed from poly-A mRNA using random hexamer and SuperScript® First-Strand Synthesis kit. The second strand cDNA was synthesized using DNA polymerase I and purified with 1.8X Agencourt AMPure XR beads. Double stranded cDNA was constructed into Illumina paired-end libraries, and assembled using Bowtie/Tophat.

Gene Annotation

A first pass automated annotation was generated via 3 iterations of a modified version of the Maker2 annotation pipeline [2], using ab initio gene prediction, protein homology (using as evidence the entire UniProt Metazoa database) and mapped EST evidence from the assembled transcriptomes with ab initio re-training between each iteration. This yielded 13,233 putative gene models. 1,377 of these automated gene models were manually checked and annotated. This identified very few assembly errors, but a small number of sequence errors (largely in homopolymer runs). However, in a significant number of cases, the automated annotation had fused adjacent genes, largely on the basis of confounding RNASeq evidence.

To avoid such gene model merging, greedy extension was used to cluster the BLAST alignments to the entire UniProt Metazoa protein database into discrete genomic loci. Putatively merged RNASeq-derived transcripts were then identified as those that spanned multiple protein clusters. For these protein-cluster-spanning transcripts, we examined the original Bowtie/Tophat derived splice-junction mappings for the presence of poorly supported splice-junctions (using read-coverage as a measure). Where identified, such transcripts were split, in some cases losing 5' and 3' UTR information. The resulting transcripts were used to re-predict the gene set using the same modified Maker2 pipeline as described above, yielding 14,911 putative gene models. Manual annotation, quality control and tracking was performed by uploading new submissions to a centralised instance of the VectorBase Community Annotation Pipeline (CAP) system [3]. These were subsequently integrated into the new annotation set using the VectorBase Patch-build system [3] to yield a final gene set of 14,992 gene models of which 1,095 had been subjected to manual reappraisal. To allow for gene identifier consistency with the original gene set release the Ensembl stable identifier pipeline was used to allocate identifiers for the final gene set.

A notable contaminant identified in the original assembly was ribosomal RNA sequence closely similar to that of nematodes in the genus *Pristionchus*, which are known arthropod parasites. No single-copy genes from this nematode were identified, suggesting the abundance of the contamination is low, and only the multi-copy rRNA genes had enough sequence to assemble. Scaffolds containing identified contaminants were removed from the annotated assembly, but it remains possible that some sequences of nematode origin remain.

The final gene-set contains 14,992 coding genes, 1,202 non-coding genes, 16,215 transcripts, and is available from the Ensembl Metazoa website: http://metazoa.ensembl.org/Strigamia_maritima/Info/Index. To assess the completeness of gene recovery we looked for “core” genes identified by [4]. Comparing

these models to our gene models we identify 95.1% of the CEGMA core genes, and at the bp level we have a median and mean sensitivity of 0.99 and 0.94 respectively (coding bases of CEGMA core genes overlapping Smar gene models) and median and mean specificity of 0.99 and 0.87 respectively. Additionally, 14,090 gene models (89.9%) have 10 or more overlapping RNAseq reads from the three tissues (adult male and female and mixed sex embryos) we used to support the annotation (the percentage increases to 93.4% if you require only 1 or more RNAseq reads). This annotated genome was then used to deduce the *S. maritima* phylome as well as phylogenomic analyses as described below.

2. Repetitive element isolation and classification.

Methodology

We aimed to detect and annotate repetitive elements in the assembled portion of the genome, and have left the likely repetitive heterochromatic 40% of the genome that could not be assembled for future work. Thus the analysis and results described here are for the assembled genome only. Repetitive elements were detected and annotated with the REPET software package ([5], version 2.0) consisting of two pipelines integrating a set of bioinformatics programs. First, repeated sequences were detected by similarity (all-by-all BLAST using BLASTER) and LTR retrotransposons were detected by structural search (LTRharvest). The similarity matches were clustered with GROUPER, RECON and PILER, and the structural matches with single-linkage NCBI Blastclust. From each cluster a consensus sequence is generated by multiple alignment with Map. The consensus sequences were analyzed for terminal repeats (TRsearch), tandem repeats (TRF), open reading frames (dbORF.py, REPET) and poly-A tails (polyAtail, REPET). In addition, the consensus sequences were screened for matches to nucleotide and amino acid sequences from known transposable elements (RepBase 17.01, [6]) using BLASTER (tblastx, blastx) as well as searched for HMM profiles (Pfam database 26.0, [7]) using hmmer3. Based on the detected structural features and homologies, the consensus sequences are classified by PASTEC according to [8]. Redundancies are removed (BLASTER, MATCHER) as well as elements classified as SSRs (>0.75 SSR coverage) or unclassified elements built from less than 10 fragments.

This set of *de novo* detected repetitive elements was used to mine the genome in the second pipeline with BLASTER (NCBI BLAST, sensitivity 4, followed by MATCHER), RepeatMasker (NCBI BLAST/ CrossMatch, sensitivity q, cutoff at 200) and CENSOR (NCBI BLAST). False positive matches were removed by an empirical statistical filter. Satellites were detected with TRF, MREPS and RepeatMasker and were then merged. In addition, the genomic sequences were screened for matching nucleotide and amino acid sequences from known transposable elements (RepBase 17.01, [6]) via BLASTER (tblastx, blastx) followed by MATCHER. Finally a removal of redundant TEs, removal of SSR annotations included within TE annotations and "long join procedure" to connect distant fragments was performed. Sequences from the *de novo* repetitive element library which were found to have at least one perfect match in the genome were then used to rerun the whole analysis.

To ensure compatibility and to avoid introducing a bias, we refrained from a manual curation or clustering of the *denovo* detected elements before mining the genome. However, post hoc we manually analyzed all elements which were previously classified into class I retrotransposon or class II DNA transposon elements or unclassified elements with detected coding element features (similarity to known transposable elements) due to potential chimeric insertion. We excluded at this stage derivative elements (LARD, TRIM, MITE) from detailed further inspection unless carrying such a feature. Elements classified as "*potential Hostgene*" or unclassified elements (*noCat*)

were also excluded at this stage. Manual inspection was done with ORF Finder (NCBI, <http://www.ncbi.nlm.nih.gov/gorf/gorf.html>), CDD search (NCBI, [9]), with a search in the most up-to-date online RepBase database (accessed December 2012-February 2013) via CENSOR ([10]) and phylogenetic analysis for LINE RT domains with RTclass1 ([11]) in order to achieve a detailed classification for each element, determine its potential relation to a family of known elements, to evaluate the completeness and to detect potential active elements. We defined an element to be complete if it possessed the relevant coding parts with the element-typical domains and the structural features (LTR, TIR). The potential activity was defined according to the region an intact ORF, if present, covered. If an intact ORF seemed to cover a complete region, including the typical domains (e.g. GAG as well as POL, Tase), then the element is considered to be potentially active. If a Tase domain is covered by a truncated ORF or the Tase itself appears to be truncated but is covered by an intact ORF, or if the RT domain is covered by an active ORF but not the remaining element-typical domains, then the element is considered to be maybe potentially active. During the manual classification to at least superfamily level, novel transposable element types not covered by the system of [8] were also considered: *Kolobok*, *Sola*, *Chapaev*, *Ginger*, *Academ*, *Novosib* and *ISL2EU* class II DNA transposons ([12], [13]).

Simple sequence repeats and other low complexity regions were extracted from the REPET pipeline database and processed with a custom Perl script to calculate the total coverage of these types of repetitive DNA by omitting overlaps with transposable element or other repetitive element annotations.

Results

Processing the centipede *S. maritima* assembly with the REPET pipeline yielded 7463 *de novo* predicted repetitive elements, of which 3715 were validated by annotation of at least one complete copy. In total 48.82% (86.03 Mb) of the genome assembly appears to be repetitive. Non-interspersed repeats (SSR, low complexity) accounted for 6.38% (11.24 Mb), whereas interspersed repeats represented 42.44% (74.79 Mb) of the centipede assembly.

All orders and most of the superfamilies of retro-transposable elements were detected in the genome of *S. maritima*. In comparison to other animals (e.g. human, insects, nematodes), LTR retrotransposons are very abundant: they account for 22.06 % of the assembly (38.86 Mb). By far the most frequent are elements from the *Copia* and *Gypsy* superfamilies, whereas *BelPao* elements are rare. Also elements from the orders DIRS, PLE, LINE and SINE were rare (each below 1% of the assembly). The small amount of LINEs is different to other organisms in which elements of this type are typically much more frequent. TRIMs and LARDs are derivatives of retro-elements and were detected in larger numbers, occupying 23.83 Mb (13.52%) of the assembly (Table S1).

Class II DNA transposons were less frequent and account for 2.3% of the assembly (4.06 Mb). The majority of these elements were TIR Transposons, especially of the *Mariner* and *Mutator* superfamilies. Interestingly, no fragments or elements of the *PiggyBac* superfamily could be found. Elements of this type are common in some insect genomes.

Other types of DNA transposons, *Maverick* (Polinton) and *Helitron* could be found in small numbers only. The DNA transposon derivatives (MITEs) that were detected account for less than 0.74 % (1.3 Mb) (Table S1).

Besides the well-classified sequences, numerous elements could not be assigned to a superfamily or even class. The latter contains a larger number of elements (2.39%, 4.2 Mb), which could represent novel types but need further investigation. 'Not categorized' elements or detected elements which contained no typical transposable element feature, but had profiles from protein coding genes, were separately annotated and accounted. Both together comprise 13.04 % (23 Mb) of the assembly (Table S1).

Most of the elements appear to be fragmented and incomplete. Although some still contain sequences of typical transposable element protein domains, they seem to be inactive due to stop codons and frameshift mutations. However, we detected more than 700 retro- and 18 DNA transposons with RT or *Tase* domains, respectively. In particular, a high number of elements of the *Copia* and *Gypsy* superfamily appeared to be complete (n=75) and/or possess active ORFs containing at least the RT domain (n=189). Such elements were also found among the LINEs (complete n=7 / potentially active n=15; especially from the *RTE* superfamily), and were found from the PLE order (n=2 / n=2), as well as from TIR DNA Transposons (n=7 / n=8, especially *Mariner*, *hAT* and *Mutator*). These elements also appear to have higher abundance and a higher number of chimeric inserts (cf. e.g. *Copia*, *Gypsy*, Table S1), which would be consistent with recent activity.

If compared to other organisms, the genomic coverage of transposable elements is rather high, and is most striking for the retro-transposons. Other animal species have lower contents of such elements. Especially if compared to insects, the centipede shows a high amount of transposable elements in the genome (48 vs. 2-37%: [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25]). However, the genome of the mosquito *Aedes aegypti* appears to contain amounts similar to the centipede (50%, [26]). It is noteworthy, however, that the *S. maritima* assembly does not contain much of the repeat-rich heterochromatin, introducing a degree of ambiguity into simple comparisons of repeat density between such draft genomes.

Some superfamilies of DNA transposons could not be found or only in small quantities. For example *PiggyBac*, *hAT* and *P* elements are frequent in genomes of *Bombus impatiens* and *Drosophila*, the pea aphid, a lizard or *Atta cephalotes* ([27], [21], [28], [14], [29]), but were barely detected here.

We did not perform a particular scan for known Viruses, but while inspecting the transposable element sequences, some conserved protein domains or sequences similar to Baculoviridae were found.

Abbreviations

ORF	open reading frame
LTR	long terminal repeat
TIR	terminal inverted repeat
SSR	simple sequence repeat

RT	reverse transcriptase
TASE	Transposase
GAG	GAG-Protein of retrotransposons
POL	POL-polyprotein of retrotransposons
LARD	large retrotransposon derivative
TRIM	terminal repeat retrotransposon in miniature
MITE	miniature inverted-repeat transposable element
SINE	short interspersed element
LINE	long interspersed element
DIRS	Dictyostelium intermediate repeat sequence
PLE	Penelope

3. Phylome analysis and phylogenomics.

Phylome reconstruction.

Proteins encoded in 18 fully-sequenced genomes, including the *S. maritima* genome, were downloaded from various sources (Table S2). The final database used for the phylome reconstruction contained 14,959 unique protein sequences for the centipede *S. maritima*. The resulting phylome comprises 11,112 single trees, which represents 74.28% of the used proteins.

To perform the phylome reconstruction, a Smith-Waterman [30] search was used to retrieve homologous sequences using an e-value cut-off of 1e-5, and considering only sequences that aligned with a continuous region representing at least 50% of the query sequence. Then, selected homologous sequences were aligned using three different programs: MUSCLE v3.8 [31], MAFFT v6.712b [32], and KAlign v2.08 [33]. Alignments were performed in forward and reverse direction (i.e. using the Head or Tail approach [34]), and the six resulting alignments were combined using M-Coffee [35]. The resulting combined alignment was subsequently trimmed with trimAl v1.4 [36], using a consistency score cut-off of 0.1667 and a gap score cut-off of 0.1, to remove poorly aligned regions.

Phylogenetic trees based on the Maximum Likelihood (ML) approach were inferred from these alignments. ML trees were reconstructed using the best-fitting evolutionary model. The selection of the evolutionary model best fitting each protein family was performed as follows: A phylogenetic tree was reconstructed using a Neighbour Joining (NJ) approach as implemented in BioNJ [37]; The likelihood of this topology was computed, allowing branch-length optimization, using nine different models (JTT, WAG, MtREV, VT, LG, Blosum62, DCMut, MtArt and Dayhoff), as implemented in PhyML v3 [38]; The two evolutionary models best fitting the data were determined by comparing the likelihood of the used models according to the AIC criterion [39]. Then, ML trees were derived using the two best-fitting models with the default tree topology search method NNI (Nearest Neighbor Interchange), and the one with best likelihood was used for further analyses. A similar approach based on NJ topologies to select the best-fitting model for a subsequent ML analysis has been shown previously to be highly accurate [40]. Branch support was computed using an aLRT (approximate likelihood ratio test) parametric test based on a chi-square distribution, as implemented in PhyML. In all cases, a discrete gamma-distribution with four rate categories plus invariant positions was used, estimating the gamma parameter and the fraction of invariant positions from the data.

Orthology/paralogy predictions.

Orthology and paralogy relationships among *S. maritima* genes and those encoded by the other considered genomes were inferred using a phylogenetic approach [41] (summarized in Tables S3 and S4). In brief, a species-overlap algorithm, as implemented in ETE v2 [42], was used to label each node in the phylogenetic tree as duplication or

speciation depending on whether the descendant partitions have, at least one, common species or not (i.e. using a Species Overlap Score of 0). The resulting orthology and paralogy predictions can be accessed through phylomeDB.org [43]. These predictions have been used in subsequent analyses such as orthology-based functional annotation, identification of gene expansions, or duplication dating.

Phylogenomics.

We took the opportunity provided by the first complete myriapod genome to investigate arthropod relationships from a genome-wide perspective. A possible advantage of using complete genomes to reconstruct evolutionary relationships among arthropods is that large data sets minimise stochastic or sampling error. A multi-gene phylogeny for the species included in the phylome was inferred using 1,491 gene families with a clear, phylogeny-based, one-to-one orthology present in at least 15 out of the 18 species included in the analyses (Figure S2). Protein sequence alignments were performed as described above and then concatenated into a single alignment of 842,150 columns. Species relationships were inferred from this alignment using a Maximum Likelihood (ML) approach as implemented in PhyML [39], using LG as the evolutionary model, since in 1,330 out of 1,491 gene families this model was the best-fitting, with the tree topology search method set to SPR (Subtree Pruning and Regrafting). Branch supports were computed using an aLRT (approximate likelihood ratio test) parametric test based on a chi-square distribution.

Increasing taxon sampling for phylogenetic inference.

In order to increase the taxon sampling for the Chelicerata, 5 additional species were used to infer a species phylogeny. Depending on the current status of each genome, two different strategies were used to identify the original 1,491 sets of widespread single-copy proteins in these newly considered species. If only the assembly was available then an exonerate [44] *protein2genome* search was executed using all sequences from each dataset as queries. Only 5 best-hits were retrieved and predictions were filtered out to keep only those with a single copy on the target genome with introns with sizes smaller than 10,000 bp. If a complete proteome was available, then a Bi-directional Best Hit (BBH) search using BLAST [30] with similar parameters to the ones used during phylome reconstruction was performed. Table S5 shows the newly added species as well as how many protein-coding genes were identified using the two strategies.

*Investigating *S. maritima* phylogenetic position in the context of Arthropoda evolution.*

The link between myriapods and chelicerates (Myriochelata) suggested by some molecular studies is in conflict with morphological characters linking the Myriapoda with the Pancrustacea. As mentioned in the main text, current consensus suggests that myriapods, insects and crustaceans form a monophyletic group, the Mandibulata. Support for Myriochelata is widely held to stem from difficulty in resolving the short Mandibulata node coupled with subtle effects of systematic biases in the data. As the difference between the two hypotheses hinges on the placement of the outgroup taxa,

the use of a closely related outgroup that does not exhibit obvious systematic bias is desirable.

To further investigate the phylogenetic position of *S. maritima* in the context of Arthropoda evolution, a new species phylogeny was reconstructed using 5 additional Chelicerata species and removing 6 distant and fast-evolving species from the initial set (Figure S3). Alignments were reconstructed for this new set of species and best-fitting evolutionary models determined as described above. Then, a multi-gene phylogeny was reconstructed based on the concatenation of the 1,491 sets of widespread single-copy protein-coding genes. A maximum-likelihood tree was derived from the concatenated alignment of 829,729 columns by using PhyML [39] with LG as the evolutionary model, since in 1,229 out of 1,491 gene families this model was the best-fitting, with the tree topology search method set to SPR (Subtree Pruning and Regrafting). Branch supports were computed using an aLRT (approximate likelihood ratio test) parametric test based on a chi-square distribution.

To investigate the statistical support for the current placement of *S. maritima* regarding the different groups of arthropod species in the two reconstructed species phylogenies, different topologies were evaluated (see Figure S4). Using ETE v2 [42] three different topologies were generated with all possible placements of *S. maritima* relative to the two arthropod groups considered in this analysis: Chelicerata and Pancrustacea. In order to avoid any potential bias on the likelihood values introduced by a specific organization within each group, only specific nodes were constrained and, therefore, the groups' internal organization was inferred in a later step.

Maximum likelihood trees were reconstructed with PhyML v3.0 [39] using as input the alignment corresponding to the 1,491 marker genes and the three different alternative topologies evaluated. LG was used as the evolutionary model since it best fits most of the individual marker genes in both cases, and the SPR (Subtree Pruning and Regrafting) algorithm was used as the tree topology search method. Branch support was computed using an aLRT (approximate likelihood ratio test) parametric test based on a chi-square distribution. In all cases, a discrete gamma-distribution with four rate categories plus invariant positions was used, estimating the gamma parameter and the fraction of invariant positions from the data. PhyML was set to follow constraints on the input topologies while the internal organization of the different collapsed groups was optimized. Likelihood values for each alternative topology were used to evaluate the statistical support of alternative positions of *S. maritima* with CONSEL [45]. CONSEL evaluates, using 8 statistical tests, the likelihood values for each of the input topologies and decides whether the observed differences, in terms of likelihood, are significant or not and, therefore, if alternative topologies to the most supported one should be considered. Tables S6 and S7 shows the results after applying CONSEL to the likelihood values generated for the initial set of species (Table S6) and for the new set of species after including 5 additional Chelicerata species and removing 6 distant and fast-evolving out-group species (Table S7).

The most recent analysis, combining many genes and a denser taxon sampling than we can achieve using whole genomes, recovers the Mandibulata with significant support. It is in the context of this phylogeny that we interpret our phylogenomics data as inconclusive. While our large number of genes is likely to have removed stochastic error, systematic error may remain. Additional genomic data from slowly evolving ecdysozoan outgroups, such as priapulids, and from additional myriapods would likely help in resolving this issue.

Orthology-based functional annotation.

To complement genome functional annotation, we searched for centipede proteins that had one-to-one orthology relationships with 9 arthropod species: *A. pisum*, *A. gambiae*, *B. mori*, *D. pulex*, *D. melanogaster*, *I. scapularis*, *N. vitripennis*, *P. humanus* and *T. castaneum*, for which GO terms are available. Of the 5,984 one-to-one orthologues (~40% of centipede genome), 4,930 of them mapped to at least one arthropod gene with some GO annotation. Annotated GO terms using this strategy are provided in File S1.

Lineage specific expansions in S. maritima.

We focused on lineage-specific expansions in the centipede genome for which 4,796 protein-coding genes (~32%) were mapped to such events. Since many protein-coding genes were detected as part of expansions across several single-gene trees, a clustering step was performed in order to group such genes into unique events. Genes were assigned to the same cluster if the overlap among expansions, in terms of shared genes, was at least of 50%. Using this cut-off 76.5% of the genes mapped to lineage-specific expansions were assigned to a unique cluster. Figure S5 shows the frequency of number of protein-coding genes per cluster in those cases with 5 or more members.

Functional categorization of the largest lineage specific expansions.

Clusters of duplicated centipede protein-coding genes specifically expanded in this lineage were analysed, looking for any statistically significant functional enrichment. Functional enrichment is provided for the 10 biggest clusters with statistically significant enriched terms. Enrichment analyses of over-represented GO terms for these expanded families compared with the annotated *S. maritima* genes were performed by using FatiGo as implemented in Babelomics webserver [46] using the Fisher exact test for genome comparison and e-value cut-off of 0.001. GO terms redundancy was reduced using REVIGO webserver [47] with default parameters. Table S8 shows these enriched functional terms for the 10 largest clusters.

Dating of duplications.

We scanned the phylome to detect and date duplication events, using a previously described algorithm [48]. We focused on events assigned to seven different relative evolutionary periods: Age (01) *S. maritima* specific, Age (02) Arthropoda I which groups

S. maritima and *I. scapularis* according to the most likely species tree, Age (03) Arthropoda II which groups *S. maritima*, *I. scapularis* with all Pancrustacea species included in the study, Age (04) duplications at Ecdysozoa level, Age (05) duplications dated at Protostomia level, Age (06) duplications mapped to the Bilateria group of species, and Age (07) which includes duplications dated at the base of all species used in this study, equating to the base of Eumetazoa. Individual trees were scanned and all duplication events that involved the seed protein and others centipede proteins were dated. Summary about such analysis can be found in Table S9.

Functional enrichment for dated duplicates.

S. maritima proteins duplicated at different relative ages were analyzed looking for any functional enrichment. Enrichment analyses for over-represented GO terms for the dated duplicated protein-coding genes compared to the whole set of annotated centipede proteins were performed using FatiGO as implemented in Babelomics webserver [46]. A Fisher exact test looking for overrepresented terms in specific sets of proteins against the whole annotated genome was used with an e-value cut-off of 0.001. GO terms redundancy was reduced using the webserver REVIGO [47] with default parameters. Table S10 shows over-represented terms grouped by age and ontology. Notably, given the few duplications detected at age 2: Arthropod I, there are no enriched functional terms for this category.

4. Synteny methodology.

Synteny analysis tested for linkage of orthologous genes on the same chromosomes (or scaffolds in the case of incomplete assembly) in pairs of species. This is sometimes called *macro-synteny* to distinguish it from analysis of more localized *micro-synteny*. A separate gene orthology analysis was performed than that in section 2 of this supporting information, as described in [50] and its supplemental data, except with a larger species tree (see list of species used here in Table S12). The clustering method performs two merging steps at each node of the species tree, working from leaves to root. In the first step (omitted at the leaves), two gene clusters from different sides of a branch are merged based on mutual best BLASTP hits with each other's members (without considering outgroups). In the second step, clusters within the current node's subtree are merged if they have mutual BLASTP hits not blocked by better hits to genes in the outgroup.

To test for significant conservation of macro-synteny we made comparisons to a null model of the number of orthologous genes that two regions in different genomes would share by chance. If the effects of gene duplication and loss are ignored, the number of shared orthologue groups would follow a hypergeometric distribution (applying Fisher's exact test). Differences in gene family size resulting from gene duplication and loss make this distribution only approximate, and we limit the effect of such changes by excluding orthologue groups with more than ten members from the analysis.

We determine the block-synteny summary statistic, P , as follows. Given the computed clusters of orthologous genes, within each genome we pre-grouped the scaffolds (or chromosome segments, in the case of *Homo sapiens* and *Trichoplax adhaerens*) into Putative Ancestral Linkage groups (PALs) as described in [50]. P then represents the percentage of genes in the two species having cross-species orthologues and having them in the PAL homologous to their own PALs.

To assign PAL homology relationships in pairwise genome comparisons, we used the log-likelihood score, $\log(mp)$, to measure the orthologue concentration for each pair of PALs, where m is the number of pairwise PAL comparisons between each pair of genomes (*i.e.* a multiple test correction) and p is the probability of the observed number of shared orthologues relative to the null model that the two PALs draw their genes independently from their common ancestor. Each PAL is considered homologous to the PAL with which it has its lowest log-likelihood score in the other genome.

Comparing the block-synteny summary statistic, P , *S. maritima* shows the greatest synteny of any non-chordate (Table S13). In particular, it shows upwards of 60 percent synteny with chordates such as humans and *B. floridae*, while the moth *Bombyx mori* shows only slightly more than 40 percent synteny with the analysed species other than *S. maritima*.

5. Homeobox genes

Homeobox gene inventory and retention of Dmbx, Vax and Hmbox

We used the complete homeobox catalogues of an insect and chordate (*Tribolium castaneum* and *Branchiostoma floridae* respectively) as queries for a saturated search of the whole genome assembly as well as the unassembled reads of the *S. maritima* genome sequencing project. We found 113 homeobox-containing genes. This compares to 133 homeobox genes in the chordate amphioxus and 104, 103, and 93 in insects such as *Drosophila melanogaster*, *Tribolium castaneum* and *Apis mellifera*. Of these 113 *S. maritima* homeobox genes, seven are very divergent and it is difficult to determine their orthology precisely. However, with a combination of molecular phylogenetics with Neighbour-Joining, Maximum-likelihood and Bayesian approaches, and using additional information from domains or sequence conservation outside of the homeodomain, we can include three of the seven genes in the ANTP class (two) and PRD class (one). Apart from the remaining four unclassified sequences, we find 54 ANTP-class genes, 26 PRD-class genes and 29 distributed amongst the nine remaining classes that are usually recognized. We found two genes with more than one homeobox, one in the Zinc Finger (ZF) class (containing four homeoboxes) and one in the Cut class (containing two homeoboxes) (Figure S7, Figure S8, Figure S9; Table S14, Table S15).

The number of *S. maritima* homeobox genes is slightly larger than the numbers found in most other arthropods analysed so far. This, at least in part, may be due to several instances of lineage-specific duplications alongside a distinct lack of homeobox gene loss in *S. maritima*. We find multiple copies (usually two to three) of Eve, Not, Vnd, BarH, Btn, Cad, Ind, Unc4, Otd and Irq. There is also a duplication of a potential Hox3 gene, discussed below. A further distinctive feature of the *S. maritima* homeobox complement is the presence of Vax and Dmbx, which have not previously been found in an arthropod genome. These genes can no longer be thought of as representing losses from the Arthropoda as a whole. Also, we find a *S. maritima* Hmbox gene, which is a member of the HNF-class. This is interesting on two counts. Firstly, the HNF class as a whole is missing from other arthropod genomes like those of the insects, and so this represents the first example of an arthropod HNF class gene described to date. Secondly, Hmbox genes have previously been proposed as chordate-specific, in contrast to more ancient members of the HNF class like HNF1/Tcf (a gene present in diploblasts as well as several bilaterians) [51]. Thus, this *S. maritima* Hmbox gene (which possesses a POU-like domain, the typical insertion for HNF-class genes of 15-20 amino acids between the second and third helix in the homeodomain, and bootstrap support of 92.6% for a grouping with chordate Hmbox genes in a HNF-class tree Figure S10) implies that Hmbox genes are not chordate-specific but have been widely lost in multiple lineages of the animal kingdom. Also, the ancient HNF1/Tcf family has instead been lost from *S. maritima*.

The first ecdysozoan Xlox ParaHox gene?

The clustering and linkage of homeobox genes is often of functional significance (e.g. the Hox genes) or provides an important insight into the origins of this gene family as well as a useful proxy for the degree of genome rearrangement relative to other species. In contrast to the intact Hox cluster, its evolutionary sister the ParaHox gene cluster is not intact, which reflects the situation found in other ecdysozoans as well. In addition to the break-up of the ParaHox cluster, the ParaHox genes of *S. maritima* have undergone duplications, producing two copies of Ind (and a third Ind-like gene) and three of Cad, which is likely to have implications for their roles in early development of the ectoderm, nervous system and gut. No ecdysozoan Xlox, which is the third ParaHox gene, has been described to date. The counterpart to the Xlox ParaHox gene from the Hox cluster (following the ProtoHox to Hox/ParaHox model of [52]) is Hox3. In *S. maritima* Hox3 is absent from the Hox cluster, but elsewhere within the genome there are two genes with sequence affinities to Hox3/Xlox. It is thus interesting to try to determine whether these two *S. maritima* Hox3/Xlox genes are either Hox genes that have somehow translocated out of the Hox cluster (and Xlox is absent from *S. maritima* as with other ecdysozoans), or instead these genes are the first examples of ecdysozoan Xlox genes (and Hox3 has been deleted from the *S. maritima* Hox cluster and genome). The further possibility that one of these genes is a Hox3 orthologue and the other is *S. maritima* Xlox is unlikely, due to the highly supported grouping of both genes together in phylogenetic trees, which implies that they arose from a duplication specific to the *S. maritima* lineage. A Neighbour-Joining phylogenetic tree of the entire coding sequences of these *S. maritima* Hox3/Xlox genes along with a selection of Hox1, Hox2, Hox3, Hox4 and Xlox genes reveals some affinity of the *S. maritima* genes with the Xlox genes of amphioxus, *Lottia* and *Capitella*. However, it is noteworthy that the bootstrap support value for this association is very low (only 33%) and so the grouping of the *S. maritima* genes with Xlox genes of other species cannot be considered as significant (Figure S11). Further phylogenetic analysis, focusing on the most similar regions of the Xlox and Hox sequences, including the hexapeptide and homeodomain regions (Figure S12) and rooting the trees with some members of the PRD class, now reveals a possible affinity with Hox3 genes rather than Xlox (Figure S13). But again there are no significant support values for this Hox3 grouping (the 42.9% support value is not shown in the tree as the threshold is 50%).

An alternative approach to phylogenetic trees that can sometimes help with resolving gene orthology is comparison of synteny [53]. One of the *S. maritima* Hox3/Xlox genes (*Hox3b_Sma*) is on a small scaffold with no gene neighbours and so comparative synteny cannot be analysed, but the second gene (*Hox3a_Sma*) is on a scaffold with 94 other genes (scaffold JH431820). We find that by reciprocal best BLAST searches against the human genome (v68 from ENSEMBL) we retrieve 24 one-to-one *S. maritima* to human orthologues (Table S16). Examining the locations in the human genome of these 24 genes reveals that five genes are located within chromosomes bearing human Hox clusters, five within chromosomes bearing human ParaHox loci and 14 in chromosomes with neither a Hox or ParaHox association (non-Hox/ParaHox chromosomes). Using Fisher's Exact Tests we find no significant associations with Hox, ParaHox or non-Hox/ParaHox chromosomes (Figure S14) (all the tests $p \geq 0.6$). As with the phylogenetic analyses, the synteny analyses also unfortunately do not resolve whether these *S. maritima* genes are orthologues of Hox3 or Xlox. Further sampling of other

ecdysozoan lineages is thus required in order to determine whether Xlox really has been lost from all lineages of this super-phylum.

Homeobox gene clusters: NK, Irx, Otp-Rx-Hbn

In addition to the clustering of Hox genes, some arthropods also contain an NK gene cluster, which is involved in mesoderm development and provides an additional example of gene clustering that is likely due to the regulatory mechanisms operating on the genes (which as yet are poorly characterized) [54, 55]. *S. maritima* does not possess an intact NK cluster, but does have some gene pairs that are remains from the ancestral cluster, potentially reflecting the retention of some shared regulatory mechanism(s). These pairs are tinman and bagpipe (often known as NK4 and NK3 in chordates), and slouch (NK1) and Drop (Msx) (Figure 4). In addition, the NK cluster remnant of *bagpipe* (*bap*) and *tinman* (*tin*) is linked with Vax (Figure 4), this linkage being relatively tight as there are only seven intervening genes. This linkage is also conserved in the mollusc *L. gigantea*, however, the number of intervening genes is larger as is the distance between *bap* and Vax (which is 747 Kb). Thus, the linkage of Vax with the NK cluster is likely an ancient aspect of the organisation of these genes, dating to at least the divergence of the Ecdysozoa and Lophotrochozoa. Vax can thus be included as a new member of the ancestral ANTP-class Mega-homeobox cluster that arose deep in animal ancestry [56, 57] (see below).

There is also a cluster of three *Irx/Iroquois* homeobox genes in *S. maritima* (Figure 4). The three-gene *Irx/Iroquois* clusters of insects and chordates are independently derived [51, 58, 59]. The three-gene cluster of *D. melanogaster* arose from an ancestral state (still present in most other insects) of two genes, one being orthologous to *mirror* and the second being pro-orthologous to *araucan* and *caupolican*. Two of the *S. maritima* *Irx* genes have affinity with the insect *mirror* gene in phylogenetic trees (Figure S15). This may indicate that the three-gene cluster of this myriapod arose by duplication of the *mirror* gene rather than the *araucan/caupolican* gene, in contrast to the route to the three-gene cluster of *Drosophila*. The *S. maritima* *Irx/Iroquois* cluster thus represents a further example of the repeated independent expansion of this gene cluster in multiple lineages of the animal kingdom [51, 58, 59].

An additional example of an ancient homeobox gene cluster is the PRD class cluster involving *Orthopedia* (*Otp*), *Rax* (*Rx*) and *Homeobrain* (*Hbn*). This cluster, which is present in *S. maritima* (Figure 4), is also found in cnidarians, insects and molluscs [60].

Homeobox gene clusters: SuperHox and Mega-homeobox

The ANTP-class of genes (including the Hox, ParaHox and NK genes) evolved very early in animal evolution, probably via states in which many of the genes were clustered into a Mega-homeobox cluster before the origin of the bilaterians and a SuperHox cluster in the Urbilaterian [56, 57, 61]. We have found some remains of this SuperHox cluster in *S. maritima* (Figure 4), represented by the linkage of Exex (Mnx)-Nedx-BtnA (Mox) in scaffold JH431734 and the linkage of BtnB (Mox) with En in scaffold JH431870. The Hmbox gene is linked to the Exex-Nedx-BtnA SuperHox remnant in *S. maritima* (Figure

4). It remains to be seen, following further phylogenetically widespread genome sequencing, whether such a linkage represents a remnant of an ancestral state, and hence a new member of the SuperHox cluster.

The tight linkage of *Ems* with the *IndB* ParaHox gene is potentially revealing with regards to the evolution of the Mega-homeobox cluster. *Ems/Emx* is a member of the ancestral NK linkage group [57, 62], whilst *IndB* is a ParaHox gene. This tight linkage of these two genes in *S. maritima* may thus be a remnant of their existence in the Mega-cluster from early in animal evolution, with *S. maritima* thus providing new evidence in support of the Mega-homeobox cluster hypothesis. We note, however, that NK and ParaHox genes have become secondarily linked again in vertebrates (having been on distinct chromosomes in the chordate and lophotrochozoan ancestors) [62]. Whilst this tight *Ems* – *IndB* linkage is intriguing, further, phylogenetically widespread examination of ANTP-class homeobox linkage patterns is certainly required to establish the veracity (or otherwise) of the Mega-cluster hypothesis. Similarly, the linkage of the ParaHox-like gene, *Ind-like*, with the NK gene *scro* may also be indicative of an ancestral linkage in the Mega cluster. However, this *Ind-like* – *scro* linkage in *S. maritima* is looser than the linkage of *Ems* – *IndB* (273kb versus 10kb (Figure 4)) and so a secondary association cannot presently be excluded.

Finally for the homeobox super-class, the linkage of the SINE class gene, *sine oculis* (*so*), with *Ems* is not unique to *S. maritima*. Humans have two semi-orthologues of *so*, namely *SIX1* and *SIX2*, and two semi-orthologues of *Ems*, namely *EMX1* and *EMX2*. *SIX2* is linked with *EMX1* on human chromosome 2, a linkage that is also echoed on zebrafish linkage group 13. A linkage of these SINE and ANTP-class genes at least as old as the bilaterian ancestor thus seems likely.

The assumptions underpinning the deductions about the Mega-homeobox and SuperHox clusters are that these homeobox genes are most likely to have arisen via tandem duplications and that close linkage of these genes in multiple lineages is most likely indicative of an ancestral condition, rather than reflecting a secondary ‘coming together’ of the genes. The various alternative hypotheses are discussed in Hui et al (2012) [62], which concludes that more needs to be known about the evolutionary dynamics of genome organisation in order to more reliably assess the true likelihood of the Mega-homeobox and SuperHox cluster hypotheses. For the present time, however, the Mega-homeobox and SuperHox hypotheses remain the most parsimonious frameworks for understanding the evolution of homeobox gene linkage patterns.

Hox mRNA processing in S. maritima

The generation of alternative RNA isoforms through RNA processing mechanisms such as alternative splicing (AS), alternative polyadenylation (APA) and alternative promoter usage (APU) is a prominent feature of the *Drosophila* Hox genes [63 – 75]. Several studies have looked at the functional implications of Hox RNA processing in fruit flies and concluded that AS (leading to the generation of distinct protein isoforms) as well as APA (which produces mRNA transcripts bearing different 3’ un-translated regions, 3’UTRs) can influence gene expression and function during fruit fly development [75 –

80]. Although recent work has identified alternative RNA isoforms for some of the Hox genes in other insects (Bomtorim et al., *pers. comm.*) and even mammals (Patraquim et al., *pers. comm.*) the evolutionary origin and developmental roles of Hox RNA processing within the arthropods remain very poorly understood. The availability of genomic and RNA sequencing data from *S. maritima* offers an unusual opportunity to explore these questions.

Based on the information currently available from the *S. maritima* genome and transcriptome project we note the existence of at least nine Hox genes in this organism (see main text Figure 4A). Of these, RNA sequencing data indicate that at least six *S. maritima* (*Sm*) Hox genes (i.e. *Antp*, *Ubx*, *abd-A*, *lab*, *Dfd*, *pb*) produce more than one mRNA isoform (Figure S16, Figure S17). In all these six cases APA generates mRNAs bearing distinct 3'UTR sequences which might interact differentially with RNA regulators such as RNA binding proteins (RBPs) and microRNAs (miRNAs). Differential splicing with concomitant APU events concern two *S. maritima* Hox genes *Dfd* and *ftz* (Figure S16, Figure S17).

All in all, more than three quarters of the *S. maritima* Hox genes undergo RNA processing of one type or other (Figure S17, Panel A). Similarly, seven out of the eight *D. melanogaster* Hox genes produce different mRNA isoforms (Figure S17, A) (FlyBase, <http://flybase.org/>). Three *D. melanogaster* Hox genes undergo AS and five produce different transcripts via APA (Figure S17, Panel B) (FlyBase <http://flybase.org/>). In addition 5 fruit fly Hox genes form different RNA species by APU (Figure S17, Panel B). From this comparison we conclude that the patterns of AS and APA affecting the centipede and *Drosophila* Hox genes are relatively similar to one another; in contrast, APU seems more prevalent in the *Drosophila* (5 out of 8 genes) than in the centipede (2 out of 9 genes) Hox genes.

Regarding the developmental progression of Hox RNA processing patterns in *S. maritima* we note that some genes such as *Ubx* display high heterogeneity in 3'UTR sequences within the embryonic transcriptome ("eggs" data) suggesting the possibility that *S. maritima* *Ubx* APA might be developmentally controlled and/or display a tissue-specific pattern (Figure S16). However this is not a general case as the data available for most other *S. maritima* Hox genes do not support developmentally variable APA patterns. In contrast, during *D. melanogaster* embryogenesis several Hox genes undergo APA (Thomsen et al. 2010).

We also see that the *S. maritima* transcriptome data supports a previously described bicistronic Hox mRNA unit bearing the coding sequences for *Ubx* and *Antp* [81]; interestingly, the transcriptome data would also be consistent with a similar bicistronic structure concerning *ftz* and *Scr*, however this signal could also be explained as the product of antisense transcription over these genes.

A possible reason underlying the similarities between *S. maritima* and *D. melanogaster* Hox AS and APA patterns might be that these RNA processing patterns represent an ancestral feature of the arthropod Hox clusters retained in both organisms. Alternatively, both organisms might have developed similar molecular strategies concerning the RNA regulation of their Hox genes as a result of convergent evolutionary

processes. To discriminate among these alternative scenarios a possibility is to look in higher detail at particular RNA processes affecting specific genes, scanning for molecular signatures that could imply common ancestry (or highly improbable convergent processes). We see such a signature in the three posterior-most *Hox* genes: *Ubx*, *abd-a*, *Abd-b* which undergo a specific type of APA (tandem APA) in both *S. maritima* and *D. melanogaster*, providing an example of what might be a feature present in the ancestral Hox cluster to insects and myriapods (Table S17). Nonetheless for most other Hox genes RNA processing patterns differ markedly between *S. maritima* and *D. melanogaster*.

Emerging genomic and transcriptomic information from spiders, crustaceans and non-Drosophilid insects should provide important elements to deduce the most likely evolutionary sequences concerning the molecular control of Hox gene expression by RNA processing.

6. Non-homeobox gene clusters: innexin, Runt, E(spl)-C

Clustering is not confined to homeobox genes. Innexins are a family of gap junction proteins, related to the vertebrate Pannexins [82]. We identified thirteen *innexin* genes in *S. maritima*. Five of them are located in a cluster composed of an innexin2, two innexin7s, an innexin1 and an innexin8 orthologue. This cluster is also present in *N. vitripennis* and *T. castaneum*, but is broken up in *D. melanogaster*.

In insect genomes, besides the clusters described above, Runt and Enhancer of Split (E(spl)-C) complexes exist. In contrast to the widespread occurrence of the various homeobox gene clusters, the Runt and E(spl)-C complexes appear to be arthropod specific. In most insects, the Runt complex comprises four Runt domain transcription factors [83]. In *Daphnia pulex*, an orthologue of one of these genes is present, clustered with two out of three other *D. pulex* Runt domain genes that are difficult to classify by phylogenetics. The chelicerate *I. scapularis* has two Runt domain genes, neither clear orthologues of the genes in the insect cluster. In contrast to these species, the *S. maritima* genome has only a single Runt domain transcription factor, providing evidence that the *Drosophila* Runt complex was an insect innovation not found in other arthropods.

E(spl)-C is a conserved Notch responsive element comprising four genes of both basic helix-loop-helix (bHLH) and bearded class genes [84]. The complex is greatly expanded in *D. melanogaster*, present in *D. pulex*, but absent from *I. scapularis*. *S. maritima* possesses 12 bHLH genes, most not found in complexes. Two complexes of these genes do exist, one made up of hairy and deadpan-like genes, the other comprising two E(spl)-like genes, but with no clear orthology relationship to E(spl)-C genes, which have characteristic bHLH-orange domains, and no bearded class genes. These data imply that E(spl)-C is a crustacean/insect complex with no orthologous complex in other arthropods.

7. Chemosensation: Gustatory receptors (GRs).

The gustatory receptor (GR) family of seven-transmembrane proteins mediates most of insect gustation (e.g. [85, 86]), as well as some aspects of olfaction, for example, the carbon dioxide receptors in flies (e.g. [87]). It ranges in size from 10-200 genes, but most insects examined so far have 50-100 genes. The GR family is more ancient than the OR family, which was clearly derived from within it, and is found in the crustacean *Daphnia pulex* [88], the tick *Ixodes scapularis* (HM Robertson, unpublished), and many other animals (HM Robertson, unpublished).

The GR family was manually annotated using methods employed for insect, *Daphnia*, and tick genomes. Briefly, TBLASTN searches were performed using major lineages of insect, *Daphnia*, and tick GRs as queries, and gene models were manually assembled in TextWrangler. Iterative searches were conducted with each new centipede protein as query until no new genes were identified in each major subfamily or lineage. When available, contigs of ESTs from RNA-seq experiments on whole animals of each sex and eggs were employed to confirm or refine gene models (Table S18). Two checks for possible divergent genes/proteins were performed. The first was a PSI-BLASTP search of the automated annotations with two iterations, and the second was TBLASTN searches of the three transcriptome assemblies with all of the existing GRs. Neither revealed additional GR lineages, although the presence of only a few of the identified GRs in the automated gene models and in the EST contigs means these checks are not conclusive. All of the SmGr genes and encoded proteins are detailed in Table S18. All SmGr proteins are provided in FASTA format (SI_file3).

Several difficulties with the genome assembly were encountered in this gene family. These were primarily length differences in homopolymer regions that in the assembly appeared to cause frameshifts within exons, but on examination of the raw reads these could be corrected. These presumably result from the known homopolymer length difficulties encountered with 454 pyro-sequencing. Seven gene models were corrected (suffix FIX in the figure, table, and FASTA). One gene model (Gr68) was designed that spans scaffolds, with no support other than the agreement of the available exons on both scaffolds, and their appropriate relatedness to similar genes in the tandem array in scaffold scf7180001247276. These problems are noted in Table S18.

Pseudogenes were translated as best possible to provide an encoded protein that could be aligned with the intact proteins for phylogenetic analysis, and attention was paid to the number of pseudogenizing mutations in each pseudogene. A 200 amino acid minimum was enforced for including pseudogenes in the analysis (roughly half the length of a typical GR), and there are several shorter fragments of genes that were not included in Table S18 or the analysis. All *Daphnia* and *Ixodes* GRs, and representative carbon dioxide and sugar receptors from insects (the most highly conserved GR lineages in insects), were aligned in CLUSTALX v2.0 [89] using default settings. Problematic gene models and pseudogenes were refined in light of these alignments.

For phylogenetic analysis, the poorly aligned and variable length N-terminal and C-terminal regions were excluded (specifically 15 amino acids before a conserved G

residue in the N-terminus and immediately after the conserved TYhhhhhQF motif in the C-terminal TM7 domain, which somewhat unusually in these SmGRs has S or T instead of the final F, and this is often the final or penultimate amino acid), as was a major internal region of length differences, specifically a long length difference region in the internal loop 2. Other regions of potentially uncertain alignment between these highly divergent proteins were retained, because while potentially misleading for relationships of major subfamilies (which are poorly supported anyway), they provide important information for relationships within subfamilies.

Phylogenetic analysis of this set of 202 proteins was carried out in the same fashion as for previous GR analyses (e.g. [90, 91], involving a combination of model-based correction of distances between each pair of proteins, and distance-based phylogenetic tree building. Pairwise distances were corrected for multiple changes in the past using the BLOSUM62 amino acid exchange matrix in the maximum likelihood phylogenetic program TREEPUZZLE v5.2 [92]. These corrected distances were fed into PAUP*v4.0b10 [93] where a full heuristic distance search was conducted with tree-bisection-and-reconnection branch swapping to search for the shortest tree. The resultant tree is shown in Figure S18 and Figure S19. Bootstrap analysis with 10,000 replications of neighbour-joining using uncorrected distances was performed to assess the confidence of major branches in the tree, and is shown above major branches in the tree. The tree was manually coloured and labels attached to lineages and subfamilies in Adobe Illustrator. The circular tree in Figure 5A has the same structure, but less detail.

The SmGr gene set consists of 76 models, comparable to that for *Daphnia* and *Ixodes*, and many insects such as *Drosophila* flies. Thirteen (17%) of these are apparent pseudogenes, seven gene models required repair of the assembly, and one was joined across scaffolds. The result is 62 apparently intact GR proteins. Less obvious pseudogenes (for example with small in-frame deletions or insertions, crucial amino acid changes, or promoter defects) would not be recognized, so this total might be high. Approximately eight gene fragments remain so short and incomplete they were not included, but some might represent intact genes.

The automated gene modeling had access to all available arthropod GRs in GenBank, for comparative information, but succeeded in building gene models for just nine of these 76 genes, only one of which was precisely correct. All others required at least one change, while 49 new gene models were generated (not including pseudogenes or those requiring repair of the assembly) (Table S18). Unfortunately, because these genes are typically expressed at low levels in only a few cells, only nine genes are represented by appropriately spliced EST contigs in the three transcriptomes (Table S18), nevertheless these manually built gene models are highly confident, because there are representative EST contigs for most subfamilies, and the basic gene structure for the entire SmGr set is a long first exon, followed by three short C-terminal exons separated by three phase 0 introns. The locations of these introns and their phases are the same as predicted by [94] to be ancestral to the entire insect chemoreceptor superfamily, and are also shared with Gr genes in other animals (HM Robertson unpublished). There were only two exceptions: Gr52 lost the first intron, and Gr76 gained a N-terminal phase 1 intron.

None of the major gene subfamilies known in the insects GRs are present in this centipede GR family, consistent with the large phylogenetic distance of centipedes from insects. Thus there are no obvious members of the sugar receptor subfamily (e.g. [90]), the fructose receptor (e.g. [95]), or relatives of the otherwise highly conserved carbon dioxide receptor subfamily (e.g. [91]). Instead, the centipede, *Daphnia*, and tick GRs form exclusive lineages in the tree (Figure S18 and Figure S19). The centipede GRs form six recently amplified subfamilies, with only a few older divergent proteins (GRs 12/13, 41/42, 52, and 75/76). This pattern of multiple recent gene subfamily expansions suggests that this centipede lineage has recently adapted to new chemical ecologies that have led to the retention and differentiation of new genes in multiple subfamilies. This pattern is reinforced by the presence of multiple pseudogenes within most of these subfamilies, presumably as some genes became redundant for the changing chemical ecology. Furthermore, most of these subfamily expansions involve tandem duplications, which is presumably how these new genes arose through unequal crossing over. The largest of these is Gr1-13 (Table S18), although the phylogenetic relationships of the genes in this expansion are complicated (Figure S18 and Figure S19), suggesting that this tandem array predates the divergence of subfamilies A and B. Rather strangely, this array was apparently duplicated at some point and separately expanded as Grs 14-27 in the same two subfamilies, but most of these genes are instead now singletons spread around the genome. Similarly, most of subfamily C (Gr28-38) is in a single tandem array. Unfortunately, given the extreme divergence of these centipede GRs from all insect GRs with known ligands or functions, no inferences of function can be ascribed to them, indeed it is possible that some are expressed in the antennae and involved in olfaction, as adaptation to terrestriality occurred independently in myriapods and insects.

8. Developmental signalling systems

Phylogenetic analysis of the Transforming Growth Factor β (TGF β) ligands in Arthropods.

For Figure S23, Maximum likelihood analysis using the WAG+i+g amino acid substitution model was carried out as described in [102]. Bootstrap values (1000 replicates) are indicated in percentages. The TGF β protein family is divided in a Bone Morphogenetic Protein (BMP) and an Activin subfamily. In our analysis, the Mavericks belong to the Activins. Concerning the BMP subfamily, the *S. maritima* genome contains two closely related *decapentaplegic* (*dpp*) duplicates. *S. maritima* reveals that arthropods ancestrally possess a BMP10 orthologue, a protein that has been lost in *Drosophila*, and an ADMP (anti-dorsalizing morphogenetic protein) orthologue, a protein that was lost in an ancestor of the beetles and flies [102]. Interestingly, we found an orthologue of the inhibitory BMP3 ligand that was suggested only to be present in deuterostomes [103]. We propose that this gene was lost in the holometabolous insects. *S. maritima* possesses one clear Glass bottom boat (*Gbb*) orthologue. As previously shown in detail [104, 105], *Drosophila* and *Megaselia scw* are diverged duplicates of *gbb*. Two *S. maritima* BMPs (SMAR009587 and SMAR007428) do not group with significant support to any particular TGF β family in our phylogeny and were not given a name. SMAR009587 clusters with the *Gbbs*, albeit with very low bootstrap values. SMAR007428 sometimes even clusters with the Activins. Phylogenetic analyses did not detect any close relation to vertebrate Nodal or Lefty of these BMPs; future analyses should reveal more about their evolutionary origin and function. Concerning the Activin subfamily, *S. maritima* possesses a clear Activin β orthologue, but no orthologue of the Activin-like protein (*Alp*) Dawdle. A clear Myostatin and a clear Maverick orthologue were identified in the *S. maritima* genome. The branching order of the Myostatins is reversed and of the Mavericks slightly disturbed, possibly because of incomplete *Acyrtosiphon* sequences. Addition of the molluscan *Lottia gigantea* Myostatin sequence did not alter the directionality. The alignment is available upon request.

Abbreviations: Is=*Ixodes scapularis*; Dp=*Daphnia pulex*; Ap=*Acyrtosiphon pisum*; Ph=*Pediculus humanus*; Nv=*Nasonia vitripennis*; Am=*Apis mellifera*; Tc=*Tribolium castaneum*; Ag=*Anopheles gambiae*; Dm=*Drosophila melanogaster*; Ca=*Clogmia albipunctata*; Ma=*Megaselia abdita*; Lg=*Lottia gigantea*.

Genbank accession numbers for non-*Strigamia* genes used in the analyses:

IsDpp=ISCW023553; DpDpp=EFX89580 (DAPPUDRAFT_347232);
ApDpp1=XP_001945626; ApDpp2=XP_001946010; ApDpp3=XP_001944147; ApDpp4=XP_003245371; PhDpp=PHUM346320; NvDpp=XM_001607627;
AmDpp=XP_001122815; TcDpp=EFA02913; AgDpp=AGAP007987; DmDpp=NP_477311; DpBMP10=EFX72705(DAPPUDRAFT_346932); TcBMP10=XP_973577;
AmBMP10=XP_001120039; IsADMP=ISCW019844; DpADMP=EFX77345 (DAPPUDRAFT_225730); AmADMP=XP_392320; NvADMP=XM_001604700;
IsBMP3=ISCW021631; ApBMP3=XP_001944767; DpBMP3=EFX74191 (DAPPUDRAFT_346933); IsGbb=ISCW019587; DpGbb=EFX74626 (DAPPUDRAFT_347233); ApGbb=XP_001947957; PhGbb=PHUM150910;
AmGbb=XP_394252; NvGbb=XP_001603876; TcGbb1=EFA04645; TcGbb2=EFA04646;

AgGbb1= XM_316789; AgGbb2= XM_320599; CaGbb, MaGbb and MaScw were obtained from <http://diptex.crg.es> [105]; DmGbb= NP_477340; DmScw= NP_524863; DpAlp=EFX87955 (DAPPUDRAFT_305466); IsAlp=ISCW010227; PhAlp=PHUM033950; AmAlp=XP_001122210; NvAlp=XP_003425497; TcAlp=XM_965262; DmAlp=NP_523461; IsAct=ISCW016200; ApAct=XM_003246878; PhAct=PHUM193490; NvAct=XM_001602234; AmAct=XP_001123044; TcAct=EFA05602; DmAct=NP_651942; AgAct=AGAP000342; LgMyo=ESO82089; IsMyo=ISCW005998; DpMyo=EFX67990 (DAPPUDRAFT_130202); PhMyo=PHUM135650; ApMyo1=ACYPI20476; ApMyo2=ACYPI49127; ApMyo3=ACYPI38027; TcMyo=XP_966819; NvMyo=XM_001602205; AgMyo=AGAP005289; DmMyo=NP_726604; DpMav=EFX89436 (DAPPUDRAFT_17212); AmMav=XM_001122118; NvMav was predicted from the genome sequence by extending XM_001606098; AgMav was predicted from the genomic sequence using XM_001656165 and AGAP012076. These two predictions are available upon request. ApMav=XM_003240719; AmMav=NP_524626; TcMav=XM_001811382.

9. Histones and Histone modifying enzymes in *S. maritima*

The core unit of chromatin is the nucleosome, a highly conserved repeating unit composed of two copies of each of the four 'core' histone proteins (H2A, H2B, H3, H4) assembled into an octamer and wound around 146-147 bp of DNA. The linker histone H1 binds the nucleosome and locks the DNA into place by binding the entry and exit sites of the DNA.

Modification of the histone proteins by methylation, acetylation and phosphorylation dynamically influences the structure of the chromatin. Chromatin structure regulates gene expression by influencing the recruitment of transcription factors, the recruitment of RNA polymerase, and also additional histone modifying enzymes.

Histone genes in S. maritima.

The core of the nucleosome is made up of the histone proteins H2A, H2B, H3 and H4. The 'core' histones are highly conserved and orthologues can be reliably identified by BLAST analysis (Table S27). The histone H1 family are more divergent at the sequence level but we have identified three orthologues in *S. maritima*.

In general, *S. maritima* has fewer histone encoding loci than dipterans such as *Aedes aegypti* and *D. melanogaster* [106, 107] but number of loci encoding each class of histone are consistent with other arthropods (Table S28, [108, 109]). There are, however, more genes encoding the H2B core histone than observed in non-dipteran arthropods (Table S28).

In *Drosophila* the histone genes are present in the genome in large numbers of quintet clusters, each cluster having one gene from each of the five classes of histones. This arrangement of genes is observed in other insects such as the pea aphid [109], and we see one quintet cluster of histone genes in *S. maritima* (Figure S31, panel A). The remainder of the histone genes are only present as single copies on a scaffold, are interrupted by non-histone encoding genes (Figure S31, panel B) or are the result of recent gene duplications (Figure S31, panels C, D).

Two loci encoding the variant histones H2A.X and H3.3 were identified. H2A.X/H2A.Z (His2AV in *Drosophila*) is found throughout eukaryotes and is associated with heterochromatin and collapsed replication forks. Phosphorylation of this histone variant is associated with double stranded breaks in the DNA [110]. The H3.3 variant histone is also evolutionarily conserved and is associated with diverse regions of the genome in eukaryotes, including pericentromeric and telomeric regions. H3.3 is also enriched in actively transcribed genes where it is thought to replace the canonical histone H3 proteins during gene transcription [110].

We could not identify an orthologue of the male specific gene *mst77F*, which encodes a sperm specific linker histone in *Drosophila* [111].

Histone modifying genes in S. maritima.

Histone acetyltransferases (HATs)

These enzymes catalyse the addition of acetyl groups to lysine residues on core histones. This favours a chromatin conformation that is accessible to transcriptional machinery and thus tends to favour active gene expression.

HATs are divided into three classes

- 1) MYST-type acetyltransferases: *S. maritima* has orthologues of all four *D. melanogaster* MYST acetyltransferase enzymes (Males absent on the first, Tip60, Enoki mushroom and Chameau).
- 2) GNAT (GCN5-type N-acetyltransferase)-type HATs: Orthologues of four *D. melanogaster* GNAT enzymes were found in *S. maritima* (CG2051, ATAC complex component 2, elongator complex protein 3 and Pcaf/GCN5)
- 3) p300/CBP(CREB binding protein) HATs: *S. maritima* has two orthologues of CREB binding protein, one (Smar_008296) has very high sequence similarity to *D. melanogaster* CBP (Neijre), the second (Smar_011410) is more diverged, but is most similar to *D. melanogaster* CBP.

HAT activity has also been ascribed to TBP-associated factor 1 [112] which has RNA polymerase II transcription factor activity and involved in pre-initiation complex assembly. *S. maritima* has two orthologues of Taf1 (Smar_007466 and Smar_005203).

The HAT gene complement of *S. maritima* is similar to that of other arthropods [109].

Histone deacetylases (HDACs)

HDAC enzymes remove the acetyl groups added to lysine residues on histones by the HAT enzymes. There are two classes of HDAC enzymes in animals; RPD3-type and Sir2 type (silent information regulator 2 or sirtuin 2).

The *S. maritima* genome encodes seven Rpd-type HDACs and four Sir2-type HDACs. Phylogenetic analysis indicates that the Rpd-type HDAC family of *S. maritima* includes two orthologues of Rpd3 (HDAC1), one orthologue each of HDAC3 and 4, two orthologues of HDAC8, and one orthologue of the class IV HDAC, HDACX.

Notably *S. maritima* does not have an orthologue of the HDAC6 gene. HDAC6 is an unusual histone deacetylase as it is located in the cytoplasm. HDAC6 binds to ubiquitin and deacetylates tubulin, and is functionally distinct from other HDACs. HDAC6 appears to function as a sensor of stressful environmental stimuli and an effector, which mediates and coordinates appropriate cell responses [113]. HDAC6 is highly conserved and is present in the genomes of *T. urticae*, *D. melanogaster* and *A. pisum*.

The sirtuin genes are NAD⁺ dependent deacetylase enzymes that have been hypothesised to be potentially responsive to environmental perturbation (including diet [114]). *S. maritima* has two orthologues of Sirt2, one of Sir2 and one of Sirt6. Sirt6 is an unusual HDAC as it is present in the cytoplasm.

10. Germ line genes

A small number of genes play conserved roles in germ line specification and development in all metazoans. As germ line specification relies heavily on post-transcriptional regulation, many of these genes encode RNA binding proteins, piRNA interacting proteins and translational regulators. In basally branching metazoans, multiple copies of genes such as *vasa*, *piwi* and *nanos* are present, whereas most of these genes are present only in a single copy in bilaterian genomes, barring genome-wide duplications. The evolution of such gene families is unclear for the arthropods, as genome data are available principally for insects.

We searched the *S. maritima* genome for the presence of 32 genes with known germ line function in at least *D. melanogaster* or mouse (see Table S29). For six of these genes (*c-Myc*, *fear of intimacy*, *aubergine*, *valois*, *Stella* and *oskar*) we failed to find any likely orthologues. However, we have found at least one putative *S. maritima* orthologue for each of the remaining 26 genes, most of which are found in a single copy. The *vasa* family of DEAD-box helicases comprises one *vasa-like* gene, one *PL10/belle* orthologue, and eight additional DEAD-box-containing genes, designated *Smar DEAD Box 1* through to *Smar DEAD Box 8*, which potentially represent a *S. maritima* specific expansion (Figure S32). The *piwi/Argonaute* family, which plays conserved roles in metazoan piRNA and stem cell regulation, is discussed in the section on Immunity and RNAi. In contrast to most other arthropod genomes examined to date, we found two *nanos* paralogues, both of which contain the conserved CCHC Zn-finger domains characteristic of *nanos* genes, and one of which is significantly shorter than most metazoan *nanos* orthologues. Duplication of the arthropod *nanos* has previously only been documented in the pea aphid *A. pisum*. The functional significance of these lineage specific duplications of *nanos* remains to be tested.

11. Meiosis genes

Among arthropod lineages, the diversity of reproductive modes often requires modifications to the key processes of meiosis. For example, parthenogenesis requires meiotic innovations to produce diploid eggs from asexual females (*i.e.* cyclical parthenogenesis) or sperm from haploid males (*i.e.* arrhenotokous parthenogenesis). In addition, *Drosophila* males undergo achiasmate meiosis (*i.e.* the absence of chiasmata between homologue pairs), which is reflected by the absence of meiotic recombination. *S. maritima* is an obligate sexual species, but little is known about meiosis within the Myriapoda. We have surveyed the genome of *S. maritima* for the presence of >50 meiosis-related genes. These genes are involved in many processes of meiosis, including cell-cycle regulation, homologue pairing, meiotic recombination and DNA repair. We also searched for these genes in the genomes of 23 additional arthropods, including members of the Hexapoda, Crustacea and Chelicerata (Figure S33). We performed phylogenetic analyses to confirm the identity of orthologues and to distinguish paralogues. For *S. maritima*, the majority of meiosis-related genes (including several meiosis specific genes) were identified. Genes absent in *S. maritima* have also been sporadically lost in other arthropods, suggesting that certain genes are dispensable for meiosis. Gene duplications found in *S. maritima* were not unique to that lineage, as paralogues were also identified in other arthropod genomes.

12. CpG methylation.

Gene body sequences were extracted from the predicted *S. maritima* gene set using CLC genomics workbench (version 5). For analysis of whole genome CpG[o/e] the genome sequence was split into 1000 nt non-overlapping fragments using a custom perl script. Nucleotide and dinucleotide content of gene body sequences and whole genome sequences were calculated using a custom perl script. The number of components in these distributions was estimated in R (www.r-project.org) using mclust [116] model-based clustering. The best fitting model was identified among several non-nested models using Bayesian information criteria (BIC). See text for further details and Figure S28, Figure S29 and Figure S30.

13. Non-protein-coding RNA genes in the *S. maritima* genome

Centipede microRNAs were computationally identified by two independent approaches, which produced >90% overlapping results. First, we retrieved precursor sequences of microRNA families conserved in all bilaterian animals, in invertebrates, in arthropods or in insects from miRBase (v 18; [117]). For each family, we searched for homologous sequences of its members in the *S. maritima* genome using BLASTN with the following parameters: -word_size=4 -reward=5 -penalty=-4 -gapopen=8 -gapextend=6. We then used INFERNAL 1.0.2 [118] to build covariance models based on the multiple sequence alignments of each microRNA family, and searched for similar profiles in the regions of *S. maritima* genome determined by BLASTN. Significant hits were added to the existing alignments, and results were manually inspected. In addition, we used MapMi [119] to map all known animal mature microRNAs to the *S. maritima* genome allowing three mismatches. Results scoring 35 or above were aligned and inspected manually for good sequence conservation and folding into microRNA-like hairpin using RALEE [120]. tRNA genes were predicted using tRNAscan-SE 1.23 with default parameters [121], and other non non-coding RNAs with the Rfam annotation pipeline (version 10; [122]) using INFERNAL 1.0.2 [118] and BLAST [123].

14. mRNA purification and sequencing library construction.

mRNA is purified from total RNA using Dynabeads® mRNA Purification Kit (Life tech, catalog number: 610-06). Briefly, 2.5 ug total RNA in 50ul DEPC-treated H₂O was denatured at 65°C for 5 minutes to disrupt the secondary structure, and immediately cooled on ice for 1 minute. 100 ul of oligo (dT)₂₅ Dynabeads was washed and resuspended in 50 ul of binding buffer before use. The denatured total RNA was added into prewashed Dynabeads and incubated at room temperature for 5 minutes on a rotary shaker. mRNA-Bead complex was captured on magnet rack and washed twice with 200 ul washing buffer. The mRNA was eluted by adding 11 ul of H₂O to mRNA-Bead complex followed by heated to 75°C for 2 minutes. Tube was placed on magnet rack immediately for 30 seconds. The supernatant containing the purified mRNA was transfer to a fresh RNase-free PCR tube while the tube was on magnet rack.

The first strand cDNA was reverse transcribed from poly-A mRNA using random hexamer and SuperScript® First-Strand Synthesis kit (Life Tech, catalogue number: 11904-018). Random hexamer was annealed to mRNA by heating mRNA/random hexamer mixture to at 65°C for 5 minutes then cooled on ice. 8.5 ul of The first strand synthesis reaction mix containing 500 uM dNTP, 20 units RNaseOut, 10 mM DTT, 200 unit Superscript II as added to mRNA/random hexamer. The first strand cDNA was synthesized by incubation reaction at 25°C 10 minutes, 42°C 60 minutes, 70°C 15 minutes and hold at 4°C. The second strand cDNA was synthesized using DNA polymerase I (life Tech, catalog number: 18010-025). The second strand cDNA synthesis was incubated at 16°C for 2 hrs in a thermocycler. The double strand cDNA was purified with 1.8X Agencourt AMPure XR beads (Beckman coulter, catalogue number: A63882).

Double stranded cDNA was constructed into Illumina paired-end libraries according to the manufacturer's protocol (Illumina Inc.). Double strand cDNA was sheared to fragments of approximately 400 bp with the Covaris S2 or E210 system (Covaris, Inc. Woburn, MA). The setting was 10% Duty cycle, Intensity of 4,200 Cycles per Burst, for 55 seconds. Fragments were processed through DNA End-Repair in 100ul containing sheared DNA, 10ul 10X buffer, 5ul End-Repair Enzyme Mix and H₂O (NEBNext End-Repair Module; Cat. No. E6050L) at 20°C for 30 minutes; A-tailing was performed in 50ul containing End-Repaired DNA, 5ul 10X buffer, 3ul Klenow Fragment (NEBNext dA-Tailing Module; Cat. No. E6053L) at 37°C for 30 minutes, each step followed by purification using a QIAquick PCR purification kit (Cat. No. 28106). Resulting fragments were ligated with Illumina PE adapters and the NEBNext Quick Ligation Module (Cat. No. E6056L). After ligation, size selection was carried out by using 2% low-melt agarose gel running in 1X TBE. Gel slices were excised from 290bp to 340bp and the size-selected DNA was purified using a Qiagen MinElute gel extraction kit and eluted in 30ul EB buffer. PCR with Illumina PE 1.0 and 2.0 primers was performed in 25-µl reactions containing 12.5 ul of 2x Phusion High-Fidelity PCR master mix, 2.5ul size-selected fragment DNA, 0.3ul each primer and H₂O. The standard thermocycling for PCR was 30 s at 98°C for the initial denaturation followed by 12 cycles of 10 s at 98°C, 30 s at 65°C and 30 s at 72°C and a final extension of 5 min. at 72°C. 1.8X Agencourt® XP® Beads (Beckman Coulter Genomics, Inc.; Cat. No. A63882) were used to purify the PCR

products. After Bead purification, PCR products were quantified using PicoGreen (Cat. No. P7589) and their size distribution analyzed using the Agilent Bioanalyzer 2100 DNA Chip 7500 (Cat. No. 5067-1506). Then, 15ul of 10nM final library was sequenced on Illumina's Genome Analyzer IIx system according to the manufacturer's specifications. Briefly, cluster generations were performed on an Illumina cluster station. 36-76 cycles of sequencing were carried out with each library in a separate, single flow cell lane on the Illumina GA II. Sequencing analysis was first done with Illumina analysis pipeline. Sequencing image files were processed to generate base calls and phred-like base quality scores and to remove low-quality reads.

15. References.

1. Gregory TR (2014) Animal Genome Size Database. <http://www.genomesize.com>.
2. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, et al. (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* 18: 188-196.
3. Megy K, Emrich SJ, Lawson D, Campbell D, Dialynas E, et al. (2012) VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics. *Nuc Acid Res* 40: D729-D734.
4. Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23: 1061-1067.
5. Flutre T, Duprat E, Feuillet C, Quesneville H (2011) Considering transposable element diversification in de novo annotation approaches. *PLoS ONE*, 6:e16526.
6. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* 110: 462-467.
7. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The Pfam protein families database. *Nucleic Acids Res* 40: D290-301.
8. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, et al. (2007) A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* 8: 973-982.
9. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, et al. (2007) CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res* 35: D237-240.
10. Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* 7:474.
11. Kapitonov V, Tempel S, Jurka J (2009) Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. *Gene* 448: 207-213.
12. Yuan Y-W, Wessler SR (2011) The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc Natl Acad Sci USA* 108: 7884-7889.
13. Kapitonov VV, Jurka J (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature Reviews Genetics* 9: 411-412; author reply 414.
14. Suen G, Teiling C, Li L, Holt C, Abouheif E, et al. (2011) The genome sequence of the leaf-cutter ant *Atta cephalotes* reveals insights into its obligate symbiotic lifestyle. *PLoS Genetics* 7:e1002007.
15. Werren JH, Richards S, Desjardins C, Niehuis O, Gadau J, et al. (2010) Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science* 327: 343-348.
16. Smith CR, Smith CD, Robertson HM, Helmkampf M, Zimin A, et al. (2011) Draft genome of the red harvester ant *Pogonomyrmex barbatus*. *Proc Natl Acad Sci USA* 108: 5667-5672.
17. Nygaard S, Zhang G, Schiøtt M, Li C, Wurm Y, et al. (2011) The genome of the leaf-cutting ant *Acromyrmex echinatior* suggests key adaptations to advanced social life and fungus farming. *Genome Res* 21: 1339-1348.

18. Bonasio R, Zhang G, Ye C, Mutti NS, Fang X, et al. (2010) Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science* 329: 1068–1071.
19. Smith CD, Zimin A, Holt C, Abouheif E, Benton R, et al. (2011) Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*). *Proc Natl Acad Sci USA* 108: 5673–5678.
20. Wurm Y, Wang J, Riba-Grognuz O, Corona M, Nygaard S, et al. (2011) The genome of the fire ant *Solenopsis invicta*. *Proc Natl Acad Sci USA* 108: 5679–5684.
21. The International Aphid Genomics Consortium (2010) Genome sequence of the pea aphid *Acyrtosiphon pisum*. *PLoS Biology* 8:e1000313.
22. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, et al. (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* 298: 129–149.
23. Arensburger P, Megy K, Waterhouse RM, Abrudan J, Amedeo P, et al. (2010) Sequencing of *Culex quinquefasciatus* establishes a platform for mosquito comparative genomics. *Science* 330: 86–88.
24. Xia Q, Zhou Z, Lu C, Cheng D, Dai F, et al. (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science* 306: 1937–1940.
25. Marinotti O, Cerqueira GC, de Almeida LG, Ferro MI, Loreto EL, et al. (2013) The genome of *Anopheles darlingi*, the main neotropical malaria vector. *Nucleic Acids Res* 41(15): 7387–7400.
26. Nene V, Wortman JR, Lawson D, Haas B, Kodira C, et al. (2007) Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science* 316: 1718–1723.
27. Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, et al. (2002) The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biology* 3:RESEARCH0084.
28. Novick PA, Smith JD, Floumanhaft M, Ray DA, Boissinot S (2011) The evolution and diversity of DNA transposons in the genome of the Lizard *Anolis carolinensis*. *Genome Biol Evol* 3: 1–14.
29. Pace JK, Feschotte C (2007) The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome Res* 17: 422–432.
30. Smith TF, Waterman MS. (1981) Identification of common molecular subsequences. *J Mol Biol* 147: 195–197.
31. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.
32. Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Briefings in bioinformatics* 9: 286–298.
33. Lassmann T, Frings O, Sonnhammer ELL (2009) Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic Acids Res* 37: 858–865.
34. Landan G, Graur D (2007) Heads or tails: a simple reliability check for multiple sequence alignments. *Mol Biol Evol* 24: 1380–1383.
35. Wallace IM, O'Sullivan O, Higgins DG, Notredame C (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucl Acid Res* 34: 1692–1699.
36. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25: 1972–1973.
37. Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* 14: 685–695.

38. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59: 307–321.
39. Akaike H (1974) A new look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* 19: 716–723.
40. Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, Denisov I, Kormes D, et al. (2011) PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res* 39: D556–560.
41. Gabaldón T (2008) Large-scale assignment of orthology: back to phylogenetics? *Genome Biology* 9:235.
42. Huerta-Cepas J, Dopazo J, Gabaldón T (2010) ETE: a python Environment for Tree Exploration. *BMC Bioinformatics* 11:24.
43. Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, Marcet-Houben M, Gabaldón T, et al. (2014) PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.* 42: D897-902.
44. Slater GSC, Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6(1):31.
45. Shimodaira H, Hasegawa M (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17(12): 1246.
46. Medina I, Carbonell J, Pulido L, Madeira SC, Goetz S, et al. (2010) Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Res* 38: W210–213.
47. Supek F, Bošnjak M, Škunca N, Šmuc T (2011) REVIGO summarizes and visualizes long lists of Gene Ontology terms. *PLoS ONE* 6:e21800.
48. Huerta-Cepas J, Gabaldón T (2011) Assigning duplication events to relative temporal scales in genome-wide studies. *Bioinformatics* 27(1): 38-45.
49. Podsiadlowski L, Kohlhagen H, Koch M (2007) The complete mitochondrial genome of *Scutigera marseulii* (Myriapoda: Symphyla) and the phylogenetic position of Symphyla. *Mol Phylogenet Evol* 45(1): 251-260.
50. Simakov O, Marletaz F, Cho SJ, Edsinger-Gonzales E, Havlak P, et al. (2013) Insights into bilaterian evolution from three spiralian genomes. *Nature* 493: 526-531.
51. Takatori N, Butts T, Candiani S, Pestarino M, Ferrier DEK, et al. (2008) Comprehensive survey and classification of homeobox genes in the genome of amphioxus, *Branchiostoma floridae*. *Dev Genes Evol* 218: 579-590.
52. Brooke NM, Garcia-Fernández J, Holland PWH (1998) The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature* 392: 920-922.
53. Hui JHL, Holland PWH, Ferrier DEK (2008) Do cnidarians have a ParaHox cluster? Analysis of synteny around a *Nematostella* homeobox gene cluster. *Evol Dev* 10: 725-730.
54. Jagla K, Bellard M, Frasch M (2001) A cluster of *Drosophila* homeobox genes involved in mesoderm differentiation programs. *Bioessays* 23(2): 125-133.
55. Doran Cande J, Chopra VS, Levine M (2009) Evolving enhancer-promoter interactions within the tinman complex of the flour beetle, *Tribolium castaneum*. *Development* 136: 3153-3160.
56. Pollard SL, Holland PWH (2000) Evidence for 14 homeobox gene clusters in human genome ancestry. *Curr Biol* 10(17): 1059-1062.

57. Garcia-Fernàndez J (2005) The genesis and evolution of homeobox gene clusters. *Nature Reviews Genetics* 6: 881-892.
58. Irimia M, Maeso I, Garcia-Fernàndez J (2008) Convergent evolution of clustering of Iroquois homeobox genes across metazoans. *Mol Biol Evol* 8: 1521-1525.
59. Kerner P, Ikmi A, Coen D, Vervoort M (2009) Evolutionary history of the Iroquois/Irx genes in metazoans. *BMC Evol Biol* 9:74.
60. Mazza ME, Pang K, Reitzel AM, Martindale MQ, Finnerty JR (2010) A conserved cluster of three PRD-class homeobox genes (*homeobrain*, *rx* and *orthopedia*) in the Cnidaria and Protostomia. *EvoDevo* 1: 3.
61. Butts T, Holland PWH, Ferrier DEK (2008) The Urbilaterian Super-Hox cluster. *Trends in Genetics* 24: 259-262.
62. Hui JHL, McDougall C, Monteiro AS, Holland PWH, Arendt D, et al. (2012) Extensive chordate and annelid macrosynteny reveals ancestral homeobox gene organization. *Mol Biol Evol* 29: 157-165.
63. Akam ME, Martinez-Arias A (1985) The distribution of Ultrabithorax transcripts in *Drosophila* embryos. *EMBO J* 4: 1689-1700.
64. Celniker SE, Keelan DJ, Lewis EB (1989) The molecular genetics of the bithorax complex of *Drosophila*: characterization of the products of the Abdominal-B domain. *Genes Dev* 3: 1424-1436.
65. Celniker SE, Sharma S, Keelan DJ, Lewis EB (1990) The molecular genetics of the bithorax complex of *Drosophila*: cis-regulation in the Abdominal-B domain. *EMBO J* 9: 4277-4286.
66. Kornfeld K, Saint RB, Beachy PA, Harte PJ, Peattie DA et al. (1989) Structure and expression of a family of Ultrabithorax mRNAs generated by alternative splicing and polyadenylation in *Drosophila*. *Genes Dev* 3: 243-258.
67. Kuziora MA, McGinnis W (1988) Different transcripts of the *Drosophila Abd-B* gene correlate with distinct genetic sub-functions. *EMBO J* 7: 3233- 3244.
68. Laughon A, Boulet AM, Bermingham JR, Laymon RA, Scott MP (1986) Structure of transcripts from the homeotic *Antennapedia* gene of *Drosophila melanogaster*: two promoters control the major protein-coding region. *Mol Cell Biol* 6: 4676-4689.
69. O'Connor MB, Binari R, Perkins LA, Bender W (1988) Alternative RNA products from the Ultrabithorax domain of the bithorax complex. *EMBO J* 7: 435-445.
70. Rowe A, Akam ME (1988) The structure and expression of a hybrid homeotic gene. *EMBO J* 7: 1107-1114.
71. Sanchez-Herrero E, Crosby MA (1988) The *Abdominal-B* gene of *Drosophila melanogaster*: overlapping transcripts exhibit two different spatial distributions. *EMBO J* 7: 2163-2173.
72. Schneuwly S, Kuroiwa A, Baumgartner P, Gehring WJ (1986) Structural organization and sequence of the homeotic gene *Antennapedia* of *Drosophila melanogaster*. *EMBO . 5*: 733-739.
73. Scott MP, Weiner AJ, Hazelrigg TI, Polisky BA, Pirrotta V, et al. (1983) The molecular organization of the *Antennapedia* locus of *Drosophila*. *Cell* 35: 763-776.
74. Stroeher VL, Jorgensen EM, Garber RL (1986) Multiple transcripts from the *Antennapedia* gene of *Drosophila melanogaster*. *Mol Cell Biol* 6: 4667-4675.
75. Tyler DM, Okamura K, Chung WJ, Hagen JW, Berezikov E, et al. (2008) Functionally distinct regulatory RNAs generated by bidirectional transcription and processing of microRNA loci. *Genes Dev* 22: 26-36.

76. Subramaniam V, Bomze HM, López AJ. (1994) Functional differences between Ultrabithorax protein isoforms in *Drosophila melanogaster*: evidence from elimination, substitution and ectopic expression of specific isoforms. *Genetics* 136: 979-991.
77. Reed HC, Hoare T, Thomsen S, Weaver TA, White RAH, et al. (2010) Alternative splicing modulates Ubx protein function in *Drosophila melanogaster*. *Genetics* 184: 745-758.
78. Thomsen S, Azzam G, Kaschula R, Williams LS, Alonso CR (2010) Developmental RNA processing of 3'UTRs in Hox mRNAs as a context-dependent mechanism modulating visibility to microRNAs. *Development* 137: 2951-2960.
79. De Navas LF, Reed HC, Akam ME, Barrio R, Alonso CR, et al. (2011) Integration of RNA processing and expression level control modulates the functions of the *Drosophila* Hox gene *Ultrabithorax* during adult development. *Development* 138: 107-116.
80. Venables JP, Tazi J, Juge F (2012) Regulated functional alternative splicing in *Drosophila*. *Nucleic Acids Res.* 40(1): 1-10.
81. Brena C, Chipman AD, Minelli A, Akam ME (2006) Expression of trunk Hox genes in the centipede *Strigamia maritima*: sense and anti-sense transcripts. *Evol Dev* 8: 252-265.
82. Abascal F, Zardoya R (2013) Evolutionary analyses of gap junction protein families. *Biochim Biophys Acta* 1828: 4-14.
83. Duncan EJ, Wilson MJ, Smith JM, Dearden PK (2008) Evolutionary origin and genomic organisation of runt-domain containing genes in arthropods. *BMC Genomics* 9: 558.
84. Duncan EJ, Dearden PK (2010) Evolution of a genomic regulatory domain: The role of gene co-option and gene duplication in the Enhancer of Split Complex. *Genome Res* 20: 917-928.
85. Su CY, Menuz K, Carlson JR (2009) Olfactory perception: receptors, cells, and circuits. *Cell* 139: 45-59.
86. Touhara K, Vosshall LB (2009) Sensing odorants and pheromones with chemosensory receptors. *Annu Rev Physiol* 71: 307-332.
87. Jones WD, Cayirlioglu P, Kadow IG, Vosshall LB (2007) Two chemosensory receptors together mediate carbon dioxide detection in *Drosophila*. *Nature* 445: 86-90.
88. Penalva-Arana DC, Lynch M, Robertson HM (2009) The chemoreceptor genes of the water flea *Daphnia pulex*: many GRs but no ORs. *BMC Evol Biol* 9: e79.
89. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23: 2947-2948.
90. Kent LB, Robertson HM (2009) Evolution of the sugar receptors in insects. *BMC Evol Biol* 9: e41.
91. Robertson HM, Kent LB (2009) Evolution of the gene lineage encoding the carbon dioxide heterodimeric receptor in insects. *J Insect Sci* 9: e19.
92. Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18: 502-504.
93. Swofford DL (2003) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
94. Robertson HM, Warr CG, Carlson JR (2003) Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A* 100 Suppl 2: 14537-14542.
95. Miyamoto T, Slone J, Song X, Amrein H (2012) A fructose receptor functions as a

- nutrient sensor in the *Drosophila* brain. *Cell* 151: 1113-1125.
96. Fredriksson R, Schiöth HB (2005) The repertoire of G-protein-coupled receptors in fully sequenced genomes. *Mol Pharmacol* 67: 1414-1425.
 97. Hauser F, Cazzamali G, Williamson M, Park Y, Li B, et al. (2008) A genome-wide inventory of neurohormone GPCRs in the red flour beetle *Tribolium castaneum*. *Front Neuroendocrin* 29: 142-165.
 98. Konopova and Jindra (2008) Broad-Complex acts downstream of Met in juvenile hormone signalling to coordinate primitive holometabolan metamorphosis. *Development* 135: 559-568.
 99. Li Y, Zhang Z, Robinson GE, Palli SR (2007) Identification and characterization of a Juvenile Hormone response element and its binding proteins. *J Biol Chem* 282: 37605-37617.
 100. Zhou X, Tarver MR, Scharf ME (2007) Hexamerin-based regulation of juvenile hormone-dependent gene expression underlies phenotypic plasticity in a social insect. *Development* 134: 601-610.
 101. Wainwright G, Webster SG, Wilkinson MC, Chung JS, Rees HH (1996) Structure and significance of mandibular organ-inhibiting hormone in the crab, *Cancer pagurus*. Involvement in multihormonal regulation of growth and reproduction. *J Biol Chem* 271: 12749-12754.
 102. Van der Zee M, da Fonseca RN, Roth S (2008) TGFbeta signaling in *Tribolium*: vertebrate-like components in a beetle. *Dev Genes Evol* 218: 203-213.
 103. Lowery JW, Lavigne AW, Kokabu S, Rosen V (2013) Comparative genomics identifies the mouse Bmp3 promoter and an upstream evolutionary conserved region (ECR) in mammals. *PLoS ONE* 8: e57840.
 104. Fritsch C, Lanfear R, Ray RP (2010) Rapid evolution of a novel signalling mechanism by concerted duplication and divergence of a BMP ligand and its extracellular modulators. *Dev Genes Evol* 220: 235-250.
 105. Wotton KR, Alcaine Colet A, Jaeger J, Jimenez-Guri E (2013) Evolution and expression of BMP genes in flies. *Dev Genes Evol* 223: 335-340.
 106. Lifton RP, Goldberg ML, Karp RW, Hogness DS (1977) Organization of histone genes in *Drosophila melanogaster* - functional and evolutionary implications. *Cold Spring Harb Sym* 42: 1047-1051.
 107. Matsuo Y, Yamazaki T (1989) Transfer-RNA derived insertion element in histone gene repeating unit of *Drosophila melanogaster*. *Nucleic Acids Res* 17: 225-238.
 108. Lyko F, Foret S, Kucharski R, Wolf S, Falckenhayn C, et al. (2010) The honey bee epigenomes: differential methylation of brain DNA in queens and workers. *PLoS Biology* 8: e1000506.
 109. Rider SD, Srinivasan DG, Hilgarth RS (2010) Chromatin-remodelling proteins of the pea aphid, *Acyrtosiphon pisum* (Harris). *Insect Mol Biol* 19: 201-214.
 110. Millar CB (2013) Organizing the genome with H2A histone variants. *The Biochemical Journal* 449: 567-579.
 111. Jayaramaiah Raja S, Renkawitz-Pohl R (2005) Replacement by *Drosophila melanogaster* protamines and Mst77F of histones during chromatin condensation in late spermatids and role of sesame in the removal of these proteins from the male pronucleus. *Mol Cell Biol* 25: 6165-6177.
 112. Mizzen CA, Yang XJ, Kokubo T, Brownell JE, Bannister AJ, et al. (1996) The TAF(II)250 subunit of TFIID has histone acetyltransferase activity. *Cell* 87: 1261-1270.

113. Matthias P, Yoshida M, Khochbin S (2008) HDAC6 a new cellular stress surveillance factor. *Cell Cycle* 7: 7-10.
114. Feil R, Fraga MF (2011) Epigenetics and the environment: emerging patterns and implications. *Nature Reviews Genetics* 13: 97-109.
115. Nagel S, Grossbach U (2000) Histone H1 genes and histone gene clusters in the genus *Drosophila*. *J Mol Evol* 51: 286-298.
116. Fraley C, Raftery AE (2003) Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST. *J Classif* 20(2):263-286.
117. Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* 39: D152-D157.
118. Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics* 25: 1335-1337.
119. Guerra-Assunção JA, Enright AJ (2010) MapMi: automated mapping of microRNA loci. *BMC bioinformatics* 11: 133.
120. Griffiths-Jones S (2005) RALEE--RNA ALignment editor in Emacs. *Bioinformatics* 21: 257-259.
121. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25: 955-964.
122. Gardner PP (2009) The use of covariance models to annotate RNAs in whole genomes. *Briefings in functional genomics & proteomics* 8: 444-450.
123. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403-410.
124. Rota-Stabelli O, Campbell L, Brinkmann H, Edgecombe GD, Longhorn SJ, et al. (2011) A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. *Proc Roy Soc Biol Sci* 278: 298-306.

16. Supporting Information Figure Legends

Figure S1. Frequency histogram showing the distribution of gene lengths in the *S. maritima* genome.

Gene length data used in this plot are available in File S4.

Figure S2. Multi-gene phylogeny for the 18 species included in the phylogenomics analysis.

1,491 widespread single-copy sets of orthologue sequences in at least 15 out of the 18 species were concatenated into a single alignment of 842,150 columns. Then, a maximum-likelihood tree was inferred using LG as evolutionary model by using PhyML.

Figure S3. Multi-gene phylogeny for 12 species included in the phylogenomics analysis plus 5 additional Chelicerata species.

1,491 widespread single-copy sets of orthologue sequences were concatenated into a single alignment of 829,729 positions. Then, a maximum-likelihood tree was inferred using LG as the evolutionary model by using PhyML.

Figure S4. Alternative topological placements of *S. maritima* relative to the main arthropod groups considered in the study: Chelicerata and Pancrustacea.

Internal organization of each group was initially collapsed and, therefore, optimized during Maximum-Likelihood reconstruction.

Figure S5. Clusters of genes specifically expanded in the centipede lineage.

On the plot, only clusters grouping 5 or more protein-coding genes were considered. The data underlying this plot is available in File S4.

Figure S6. Mitochondrial gene organisation.

Shaded regions represent differences from the ground pattern. Gene translocations in Myriapoda have been noted in *Scutigera causeyae* (Myriapoda: Symphyla) [49]. The previous example of the small conserved region trnaF-nad5-H-nad4-nad4L on the minus strand between *Limulus*, *Lithobius* and *Strigamia* is not a conserved feature in all Chilopoda, because *Scutigera coleoptrata* have an interruption between nad5 and H-nad4 with elements on the minus and plus strands accompanied by a translocation of nad4L to a position immediately preceding nad5.

Figure S7. Classification of all *S. maritima* (Sma) homeodomains (excluding Pax2/5/8/sv) via phylogenetic analysis using *T. castaneum* (Tca) and *B. floridae* (Bfl) homeodomains.

This phylogenetic analysis was constructed using Neighbour-Joining with a JTT distance matrix and 1000 bootstrap replicates. Gene classes are indicated by colours. The genes coloured in grey are those genes that cannot be assigned to known classes. Further classification was performed using additional domains outside the homeodomain and by performing additional phylogenetic analysis for particular gene classes using maximum-likelihood and bayesian approaches. Pax2/5/8/sv is excluded due to the gene possessing only a partial homeobox.

Figure S8. Phylogenetic analysis of ANTP class homeodomains of *S. maritima* (Sma) using *T. castaneum* (Tca) and *B. floridae* (Bfl) for comparison.

These phylogenetic analyses were constructed using Neighbour-Joining with a JTT distance matrix, 1000 bootstrap replicates (support values in black). Nodes with support equal to or above 500 in the maximum-likelihood (LG+G) analysis are in blue and nodes with posterior probabilities equal to or above 0.5 (LG+G) in the Bayesian analysis are in red.

Figure S9. Phylogenetic analysis of PRD class homeodomains of *S. maritima* (Sma) using *T. castaneum* (Tca) and *B. floridae* (Bfl) for comparison.

These phylogenetic analyses were constructed using Neighbour-Joining with a JTT distance matrix, 1000 bootstrap replicates (support values in black). Nodes with support equal to or above 500 in the maximum-likelihood (LG+G) analysis are in blue and nodes with posterior probabilities equal to or above 0.5 (LG+G) in the Bayesian analysis are in red.

Figure S10. Phylogenetic analysis of HNF class homeodomains of *S. maritima* (Sma) using *B. floridae* (Bfl), human (*Homo sapiens*, Hsa) and sea anemone (*Nematostella vectensis*, Nve) for comparison.

These phylogenetic analyses were constructed using Neighbour-Joining with a JTT distance matrix, 1000 bootstrap replicates (support values in black). Nodes with support equal to or above 500 in the maximum-likelihood (LG+G) analysis are in blue and nodes with posterior probabilities equal to or above 0.5 (LG+G) in the Bayesian analysis are in red.

Figure S11. Phylogenetic analysis of Xlox/Hox3 genes of *S. maritima* (Sma) using a selection of Hox1, Hox2, Hox3, Hox4 and Xlox sequences.

This analysis was based upon the whole coding sequence of the genes, and was constructed using Neighbour-Joining with a JTT distance matrix and 1000 bootstrap replicates. The blue support value (of 333) is the node that reveals the affinity between the Xlox/Hox3 genes of *S. maritima* and Xlox sequences. Ame = *Apis mellifera*, Bfl = *Branchiostoma floridae*, Cte = *Capitella teleta*, Dme = *Drosophila melanogaster*, Lgi = *Lottia gigantea* and Tca = *Tribolium castaneum*.

Figure S12. Multiple alignment of relevant residues of the Hox1, Hox2, Hox3, Hox4 and Xlox sequences of different lineages compared to *S. maritima* Hox3a and Hox3b sequences.

Three Paired class genes are included as an outgroup. The grading of purple colouring of the amino acids shows the identity level of these sequences. The red rectangles in the multiple alignment delimit the core of the hexapeptide motif and the homeodomain. This is the alignment used to construct the phylogenetic tree in Figure S13. Ame = *Apis mellifera*, Bfl = *Branchiostoma floridae*, Cte = *Capitella teleta*, Dme = *Drosophila melanogaster*, Lgi = *Lottia gigantea* and Tca = *Tribolium castaneum*.

Figure S13. Phylogenetic analysis of *S. maritima* Xlox/Hox3 homeodomain and hexapeptide motifs using a selection of Hox1, Hox2, Hox3, Hox4 and Xlox sequences.

This analysis used a section of the coding sequence including the hexapeptide and some flanking residues plus the homeodomain (alignment in Figure S12). Three Paired class genes are included as an outgroup. This phylogeny was constructed using Neighbor-Joining with the JTT distance matrix and 1000 bootstrap replicates. Maximum Likelihood support values are shown in blue and Bayesian posterior probabilities in red. Ame = *Apis mellifera*, Bfl = *Branchiostoma floridae*, Cte = *Capitella teleta*, Dme = *Drosophila melanogaster*, Lgi = *Lottia gigantea* and Tca = *Tribolium castaneum*.

Figure S14. Fisher's Exact Test to distinguish whether *S. maritima* scaffold 48457 has significant synteny conservation with ParaHox or Hox chromosomes of humans.

No significant Hox or ParaHox association is found.

Figure S15. Phylogenetic analysis of TALE class homeodomains of *S. maritima* (Sma) using *T. castaneum* (Tca) and *B. floridae* (Bfl), including the Iroquois/Irx genes.

These phylogenetic analyses were constructed using Neighbour-Joining with a JTT distance matrix, 1000 bootstrap replicates (support values in black). Nodes with support equal to or above 500 in the maximum-likelihood (LG+G) analysis are in blue and nodes with posterior probabilities equal to or above 0.5 (LG+G) in the Bayesian analysis are in red.

Figure S16. RNA processing in the Hox cluster of *S. maritima*.

The transcriptome of *S. maritima* (*Sm*) eggs (blue), females (green) and males (red) was mapped to the Hox gene cluster (top panel – see Figure 4 in the main text) and transcript models were inferred for each gene within the cluster (shaded area) taking into account the presence of open-reading frames (ORF) and polyadenylation signals (PAS) to support the existence of RNA processing events. We note the occurrence of more than one mRNA isoform of six *S. maritima* Hox genes (i.e. *Antp*, *Ubx*, *abd-A*, *lab*, *Dfd*, *pb*). In all these six cases alternative polyadenylation (APA) generates mRNAs bearing distinct 3' untranslated regions (UTR; alternative UTR sizes at the bottom). Alternative splicing (AS) with concomitant alternative promoter use (APU) events concern two *S. maritima* Hox genes *Dfd* and *ftz* (see alternative ORF sizes at the bottom). We also see that some genes such as *S. maritima* *Ubx* display high heterogeneity in 3'UTR sequences within the embryonic transcriptome ("eggs" data) suggesting the possibility that *S. maritima* *Ubx* APA might be developmentally controlled and/or display a tissue-specific pattern (see inset for further details on symbols).

Figure S17. RNA processing in the *S. maritima* and *D. melanogaster* Hox clusters.

A) The incidence of alternatively processed mRNAs is comparable between *S. maritima* and *D. melanogaster*, in that over 75% of the *S. maritima* Hox genes undergo RNA processing of one type or another. Similarly, seven out of the eight *Drosophila* Hox genes produce different mRNA isoforms (FlyBase, <http://flybase.org/>). B) Three *D. melanogaster* Hox genes undergo AS (blue) and five produce different transcripts via APA (red, FlyBase <http://flybase.org/>). In addition 5 fruit fly Hox genes form different RNA species by APU (green). C) Classification of all alternatively processed mRNA events in the *S. maritima* Hox cluster based on the same categorisation as in B). Note

that patterns of AS and APA affecting *S. maritima* and *D. melanogaster* Hox genes are relatively comparable; in contrast, APU seems more prevalent in the *Drosophila* (5 out of 8 genes) than in the centipede (2 out of 9 genes) Hox genes.

Figure S18. Phylogenetic tree of the *S. maritima*, *D. pulex*, *I. scapularis*, and representative insect GRs, part one.

This is a corrected distance tree and was rooted at the midpoint in the absence of a clear outgroup, an approach that clearly indicates the distinctiveness of the centipede GRs. It is a more detailed version of Figure 5A. The *S. maritima*, *D. pulex*, *I. scapularis*, and representative insect gene/protein names are highlighted in red, blue, brown, and green, respectively, as are the branches leading to them to emphasize gene lineages. Bootstrap support levels in percentage of 10,000 replications of neighbour-joining with uncorrected distance is shown above major branches. Comments on major gene lineages are on the right. Suffixes after the gene/protein names are: PSE – pseudogene; FIX – sequence fixed with raw reads; JOI – gene model joined across scaffolds. Note than in Figure 5A for space reasons the IsGr47 and 59 proteins are included in the carbon dioxide and sugar receptor groupings, respectively, however there is no bootstrap support for these branches, and no such functional assignment is claimed. Similarly, it is unlikely that the DpGr57/58 proteins are fructose receptors.

Figure S19. Phylogenetic tree of the *S. maritima*, *D. pulex*, *I. scapularis*, and representative insect GRs, part two.

This is a corrected distance tree and was rooted at the midpoint in the absence of a clear outgroup, an approach that clearly indicates the distinctiveness of the centipede GRs. It is a more detailed version of Figure 5A. The *S. maritima*, *D. pulex*, *I. scapularis*, and representative insect gene/protein names are highlighted in red, blue, brown, and green, respectively, as are the branches leading to them to emphasize gene lineages. Bootstrap support levels in percentage of 10,000 replications of neighbour-joining with uncorrected distance is shown above major branches. Comments on major gene lineages are on the right. Suffixes after the gene/protein names are: PSE – pseudogene; FIX – sequence fixed with raw reads; JOI – gene model joined across scaffolds. Note than in Figure 5A for space reasons the IsGr47 and 59 proteins are included in the carbon dioxide and sugar receptor groupings, respectively, however there is no bootstrap support for these branches, and no such functional assignment is claimed. Similarly, it is unlikely that the DpGr57/58 proteins are fructose receptors.

Figure S20. Neuropeptide precursor sequences identified in the *S. maritima* genome.

The putative signal peptides (predicted by SignalP) are underlined, the putative active neuropeptides or protein hormones (based on similarity to neuropeptides or protein hormones identified in other invertebrates) are marked in yellow. Green indicates putative basic cleavage sites flanking the putative neuropeptides. Glycines used for amidation are shown in blue, cysteines proposed to form cysteine bridges are shown in red. Dots indicate missing N- or C-termini.

Figure S21. Examples of tandem duplications of neuropeptide receptor genes.

Structure of the two inotocin receptor genes found head-to-head on opposite strands of scaffold JH431865 (A). Structure of the two SIFamide receptor genes found tail-to-head on the same strand of scaffold JH432116 (B).

Figure S22. Schematic diagram showing sesquiterpenoids/juvenoids synthesis (upper) and degradation (lower) pathways in arthropods.

Molecules/hormones in synthesis are shown in **bold**, enzymes are shown in *italics*, and species/clades are shown in ***bold italics***.

Figure S23. Phylogenetic analysis of the Transforming Growth Factor β (TGF β) ligands in Arthropods.

See Text S1 for details. Abbreviations: Is=*Ixodes scapularis*; Dp=*Daphnia pulex*; Ap=*Acyrtosiphon pisum*; Ph=*Pediculus humanus*; Nv=*Nasonia vitripennis*; Am=*Apis mellifera*; Tc=*Tribolium castaneum*; Ag=*Anopheles gambiae*; Dm=*Drosophila melanogaster*; Ca=*Clogmia albipunctata*; Ma=*Megaselia abdita*; Lg=*Lottia gigantea*.

Figure S24. Range of Wnt genes present in *S. maritima*.

Wnt genes present and number of *Wnt* subfamilies absent in *S. maritima* in comparison with other arthropods and three non-arthropod outgroups.

Figure S25. Phylogeny of FGF receptor (FGFR) genes indicating that FGFR genes duplicated independently in *S. maritima* and *D. melanogaster*.

See text for details. Alignment was performed using Clustal-Omega (<http://www.ebi.ac.uk/Tools/services/web>). The evolutionary history was inferred using the Neighbor-Joining method with bootstrapping to determine node support values (10000 replicates). The evolutionary distances were computed using the Poisson correction method. Evolutionary analyses were conducted in MEGA5.

Figure S26. Phylogeny including the three FGF genes of *S. maritima*.

See text for details. Alignment was performed using Clustal-Omega (<http://www.ebi.ac.uk/Tools/services/web>). The evolutionary history was inferred using the Neighbor-Joining method with bootstrapping to determine node support values (10000 replicates). The evolutionary distances were computed using the Poisson correction method. Evolutionary analyses were conducted in MEGA5.

Figure S27. *Cap 'n' collar (cnc)* genes.

A) The two genes are located on different scaffolds. *Cnc1* is a long transcript consisting of 11 exons. *Cnc2* is shorter (eight exons), the three exons at the 3' end of the gene that encode the C-terminal region of the protein including the conserved domain (B) show a similar structure. (B) *S. maritima* Cnc protein structure. Both proteins contain the bZip domain in a similar position at the C-terminus. *Cnc1* encodes a long protein (925 amino acids). Bits of the N-terminal region (blue lines) align with *D. melanogaster* Cnc isoform C and *T. castaneum* Cnc variant A. (C) Cnc protein sequence alignment, only showing the aligning bits in the N-terminal region. Blue lines show short stretches of sequence that form a consensus motif. These motifs are not present in the proteins encoded by *Sm-cnc2*, *Dm-cnc* isoforms A and B, and *T. castaneum* cnc variant B.

Figure S28. Frequency histograms of observed versus expected dinucleotide content in *S. maritima* gene bodies.

(A – P) The y-axis depicts the number of genes with the specific dinucleotide_[o/e] values given on the x-axis. The distribution of all dinucleotide pairs, with the exception of CpG, is best described as a unimodal distribution. The distribution of CpG dinucleotides is best described as a trimodal distribution, with ‘high’ and ‘low’ CpG_[o/e] classes. The data underlying this figure is available in File S5.

Figure S29. Frequency histogram of CpG_[o/e] observed in 1000 bp windows of the *S. maritima* genome.

The y-axis depicts the number of genes with the specific CpG_[o/e] values given on the x-axis. The distribution of CpG_[o/e] in *S. maritima* genome is a bimodal distribution, with a high CpG_[o/e] peak observed similar to that observed in the gene bodies (Figure 9). The data underlying this figure is available in File S6.

Figure S30. Contrasting patterns of DNA methylation, as measured by over- and under-representation of CpG dinucleotides in coding regions (CpG_(o/e)), within arthropod species.

In all graphs the y-axis depicts the number of genes with the specific CpG_(o/e) values given on the x-axis. A) *D. melanogaster* coding regions show a unimodal peak reflective of the lack of DNA methylation in this species. B) *Apis mellifera* shows a bimodal peak consisting of genes with a lower than expected CpG_(o/e) (green distribution) and a higher than expected CpG_(o/e) (blue distribution). The presence of a bimodal distribution in this species is consistent with depletion of CpG dinucleotides in the coding regions of genes over evolutionary time as a result of DNA methylation. C) A single unimodal peak is also observed for *Tetranychus urticae*, a species that has very low levels of DNA methylation. D) The *S. maritima* distribution is best explained as a mixture of three distinct distributions that we have deemed ‘low’ (green distribution), ‘medium’ (blue distribution) and ‘high’ (grey distribution). The genes within the low distribution likely contain genes that are historically methylated, whilst the ‘high’ distribution can be explained by regions of the genome that are comparatively CpG-rich (as determined by the analysis of the *S. maritima* genome, Figure S29). The data underlying this figure is available in File S7.

Figure S31. Chromosomal organisation of histone gene clusters in *S. maritima*.

In insects such as *Drosophila* [115] and the pea aphid [109] histone encoding genes are present in quintet clusters, each cluster containing one gene from each of the five classes of histone. Only one such cluster could be identified in *S. maritima* (A). The other four clusters identified in the *S. maritima* genome (B-D) all consist of a 2 – 3 copies of a histone encoding gene of a single class. It is possible that these have arisen as a result of recent gene duplication.

Figure S32. *S. maritima* vasa DEAD-box helicase germline gene phylogeny.

Maximum likelihood tree of *vasa*/*PL10* family genes. One gene is a likely *vasa* orthologue (SMAR015390), one groups with the *PL10* family (SMAR005518), and the majority group in an apparently distinct DEAD-box-containing clade. Bootstrap values for 2000

replicates are shown at each node. Accession numbers for protein sequences are as follows: *Apis* Belle (XP_391829.3), *Apis* Vasa (NP_001035345.1), *Danio* PL10 (NP_571016.2), *Danio* Vasa (AAI29276.1), *Drosophila* Belle (NP_536783.1), *Drosophila* Vasa (NP_723899.1), *Gryllus* Vasa (BAG65665.1), *Mus* Mvh (NP_001139357.1), *Mus* PL10 (NP_149068.1), *Nasonia* Belle (XP_001605842.1), *Nasonia* Vasa (XP_001603956.2), *Nematostella* PL10 (XP_001627306.1), *Nematostella* Vasa 1 (XP_001628238.1), *Nematostella* Vasa 2 (XP_001639051.1), *Oncopeltus* Vasa (AGJ83330.1), *Parhyale* Vasa (ABX76969.1), *Tribolium* Belle (NP_001153721.1), *Tribolium* Vasa (NP_001034520.2), *Xenopus* PL10 (NP_001080283.1), *Xenopus* VLG1 (NP_001081728.1).

Figure S33. Phylogenomic inventory of meiotic genes in arthropods.

Red genes are specific to meiosis in model species in which functional data is available. “+” and “-” indicate the presence and absence of orthologues respectively. Numbers indicate copy number of duplicated genes.

Figure S34. Patterns of microRNA gain and loss across the animal kingdom with the inclusion of *S. maritima*.

The number of microRNAs that were gained or lost at each node are shown in green and red, respectively, and names are listed below each taxon. MicroRNAs that are found in the *S.maritima* genome are in bold, and families for which more than one homologue is found are marked with an asterisk. The tree depicts the Mandibulata hypothesis rather than the Myriochelata, as in [124].

17. Supporting Information Table Legends.

Table S1. Detailed overview for the repetitive elements in *S. maritima*.

For each group the number of elements (putative families), the number of their fragments or copies in the genome, the cumulative length, the proportion of the assembly and some features are shown. This includes elements containing nested inserts of other elements (n), elements which appear to be complete (i.e. all typical structural and coding parts present, even if containing stop codons or frameshifts), elements with a RT or *Tase* domain detected (n), elements which potentially could be active as they contain an intact ORF with all the typical domains even though they could lack other structural features like terminal repeats, and elements which contain an intact ORF for the RT domain or parts of the *Tase* domain and could thus be partly active. The elements which could not be categorized or contained features of protein coding regions are shown at the bottom, whereby they probably do not belong to the transposable elements.

Table S2. Set of species used in the comparative genomics analyses related to the *S. maritima* genome.

Columns include, in this order, scientific names, the species code according to UNIPROT, the number of the longest unique transcript used in the analyses, the data source and the date in which data was retrieved.

Table S3. Orthologues detected between a given species and *S. maritima*.

First column indicates how many trees have been used to detect such orthologues. Columns “uniq” refers to the number of orthologues detected for each pair of species after removing redundancy. In one-to-many and many-to-many orthology relationships it is possible to count a given protein more than once. Regarding the ratios values, “all” column refers to the orthology ratio computed using all orthologue pairs meanwhile “uniq” refers to the ratio computed using “uniq” columns.

Table S4. Orthology ratios for a given species related to *S. maritima*.

This table is similar to Table S3, but in this case orthology relationships with 10 or more proteins for any of the species are discarded in order to avoid biases introduced by species-specific gene family expansions.

Table S5. Newly added Chelicerata species used to increase the taxon sampling for the species phylogeny.

First column indicates the scientific species name, the second one indicates which strategy has been used to identify single copy protein-coding genes. Third column shows how many single-copy genes have been identified in each species from the initial set of 1,491 used to reconstruct the species phylogeny. Last two columns show the data source and the date on which data was retrieved.

Table S6. Results after applying the different statistical tests implemented in CONSEL for the alternative placement of *S. maritima* relative to Pancrustacea and Chelicerata groups of species (as shown in Figure S4) in the context of the 18 species used for the phylogenomics analyses. The ‘item’ column relates to Figure S4

as follows, (1) topology arrangement corresponding to Figure S4 left-hand panel, in which *S. maritima* was grouped with Chelicerata species. (2) Topology arrangement corresponding to Figure S4 central panel, in which *S. maritima* branches off before the split of Pancrustacea and Chelicerata. (3) Topology arrangement corresponding to Figure S4 right-hand panel, in which *S. maritima* was grouped with Pancrustacea species.

Table S7. Results after applying the different statistical tests implemented in CONSEL for the alternative placement of *S. maritima* relative to the two arthropod groups, Pancrustacea and Chelicerata (as shown in Figure S4), with the inclusion of extra chelicerates. Taxon sampling for the Chelicerata was increased after including sequences from 5 additional species. In order to reduce any potential bias introduced by distant and/or fast-evolving out-groups, 6 out-group species from the initial set were removed. The 'item' column relates to Figure S4 as follows, (1) topology arrangement corresponding to Figure S4 left-hand panel, in which *S. maritima* was grouped with Chelicerata species. (2) Topology arrangement corresponding to Figure S4 central panel, in which *S. maritima* branches off before the split of Pancrustacea and Chelicerata. (3) Topology arrangement corresponding to Figure S4 right-hand panel, in which *S. maritima* was grouped with Pancrustacea species.

Table S8. Enriched functional GO Terms for the 10 largest clusters of duplicated *S. maritima* protein-coding genes specifically expanded in the centipede lineage, as compared with the whole genome.

Table S9. Statistics regarding the duplications of centipede genes relative to seven specific ages detected using all available trees on the phylome.

Table S10. Enriched functional GO terms for proteins duplicated at the different relative ages shown in Table S9.

Columns show relative age, gene ontology namespace, the GO term id and its name, respectively.

Table S11. Overview of *Strigamia maritima* mitochondrial genome.

Table S12. Species used in the synteny analyses and the sources of their sequence data.

Table S13. Block-synteny summary statistics for pairs of species.

Hs = *Homo sapiens*, Bf = *Branchiostoma floridae*, Sm = *Strigamia maritima*, Lg = *Lottia gigantea*, Ct = *Capitella teleta*, Nv = *Nematostella vectensis*, Ta = *Trichoplax adhaerens*, Ag = *Anopheles gambiae*, Bm = *Bombyx mori*.

Table S14. Summary of numbers of homeobox genes per class of *Strigamia*, *Branchiostoma* and *Tribolium*.

Table S15. Names and identification numbers of all *S. maritima* homeobox genes along with their orthologues from the beetle, *T. castaneum*, and amphioxus, *B. floridae*.

Table S16. One-to-one *S. maritima* to human orthologues starting from genes on *S. maritima* scaffold 48457, which contains *SmaHox3a*.

The third column is the chromosomal location of the human orthologue. Human Hox chromosomes are 2, 7, 12 and 17 and the ParaHox chromosomes are 4, 5, 13 and X.

Table S17. Evolutionary conservation of RNA processing modes in the *S. maritima* and *D. melanogaster* Hox clusters.

Type of RNA processing event concerning each one of the *S. maritima* (left) and *D. melanogaster* (right) Hox genes. We note that orthologous genes in both species undergo similar types of RNA processing: the three posterior-most Hox genes: *Ubx*, *abd-a* and *Abd-b* display a specific type of APA (tandem APA) in both *S. maritima* and *D. melanogaster* (conserved patterns highlighted by red asterisks) providing an example of what might be a feature present in the ancestral Hox cluster to insects and myriapods. Nonetheless, for most other Hox genes, RNA processing patterns differ markedly between *S. maritima* and *D. melanogaster*, indicating that the conserved incidence of alternative RNA processing across arthropods can only be proposed for the posterior-most Hox genes.

Table S18. Details of SmGr family genes and proteins.

Columns are: Gene – the gene and protein name we are assigning (suffixes are PSE – pseudogene, FIX – assembly was repaired; JOI – gene model spans scaffolds); OGS – the official gene number in the 13,233 proteins (prefix is Smar_temp_); Scaffold – the genome assembly scaffold ID, prefix is scf718000 (amongst 14,739 scaffolds in assembly Smar05272011); Coordinates – the nucleotide range from the first position of the start codon to the last position of the stop codon in the scaffold; Strand – + is forward and - is reverse; Introns – number of introns; ESTs – presence of an EST contig with appropriate splicing in one of the three transcriptome assemblies (F - female, M - male, E - eggs); AAs – number of encoded amino acids in the protein; Comments – comments on the OGS gene model, repairs to the genome assembly, and pseudogene status (numbers in parentheses are the number of obvious pseudogenizing mutations).

Table S19. Total numbers of biogenic amine receptors in different species.

Table S20. A comparison between the *D. melanogaster* and *S. maritima* biogenic amine receptors.

The orthologues are given next to each other. When there is no orthologue, a dash (–) is written instead.

Table S21. Genes encoding neuropeptide precursors and neuropeptide receptors annotated in *S. maritima*.

Abbreviations: ACP, adipokinetic hormone/corazonin-related neuropeptide; AKH, adipokinetic hormone; ADF, antidiuretic factor; AST, allatostatin; CCAP, crustacean cardio-active peptides; DH (Calc.-like), calcitonin-like diuretic hormone; DH (CRF-like),

corticotropin releasing factor-like diuretic hormone; EH, eclosion hormone; ETH, ecdysis triggering hormone; GPA2, glycoprotein hormone A2; GPB5, glycoprotein hormone B5; ILP, insulin-like peptide; ITP, ion transport peptide; NPF, neuropeptide F; NPLP, neuropeptide-like precursor; PDF, pigment dispersing factor; PTTH, prothoracicotropic hormone; sNPF, short neuropeptide F.

Table S22. Presence or absence of neuropeptide signaling systems in arthropods.

The centipede *S. maritima* contains two CCHamide-1, two eclosion hormone and two FMRFamide genes (2 p). In some cases neuropeptide precursors could not be identified, but the corresponding receptor genes are present (R). We assume that this is due to sequencing gaps. For abbreviations see Table S21.

Table S23. Genes commonly implicated in arthropod juvenoids biosynthesis (green) and degradation (blue), and their potential regulators (purple)[98 - 101].

Common abbreviations, and presence in the centipede *S. maritima*.

Table S24. List of genes commonly implicated as potential regulators of arthropod juvenoids biosynthesis (purple)[98 - 101].

Common abbreviations, and presence in the centipede *S. maritima*.

Table S25. Wnt genes in the genome of *S. maritima*.

SMAR = the gene identification number, and scaffold = the scaffold identification number. Wnt 1, 6 and 10 are clustered together on the same scaffold (yellow highlighting), which is likely a remnant of the ancestral wnt gene cluster (see text for details).

Table S26. Selenoproteins in the *S. maritima* genome.

Table S27. Histone encoding loci of *S. maritima*.

Table S28. Number of loci within the genomes of arthropod species encoding the five classes of histones.

Orthologues for *A. aegypti*, *D. pulex*, *T. urticae* and *I. scapularis* were obtained by BLAST analysis. Orthologues for *A. mellifera* and *A. pisum* were obtained from published literature [108, 109].

Table S29. Germ line and RNAi genes annotated in the *S. maritima* genome.

The name of the *Drosophila* orthologue is shown unless indicated with "(Mo)", for mouse.

Table S30. Details of the manually annotated genes of *S. maritima*.

18. Supporting Information data files.

File S1. One2One_GOTerms_GenomeIDs for Orthology-based functional annotation.

File S2. Strigamia_pals for Figure 3.

File S3. Gustatory receptor sequences.

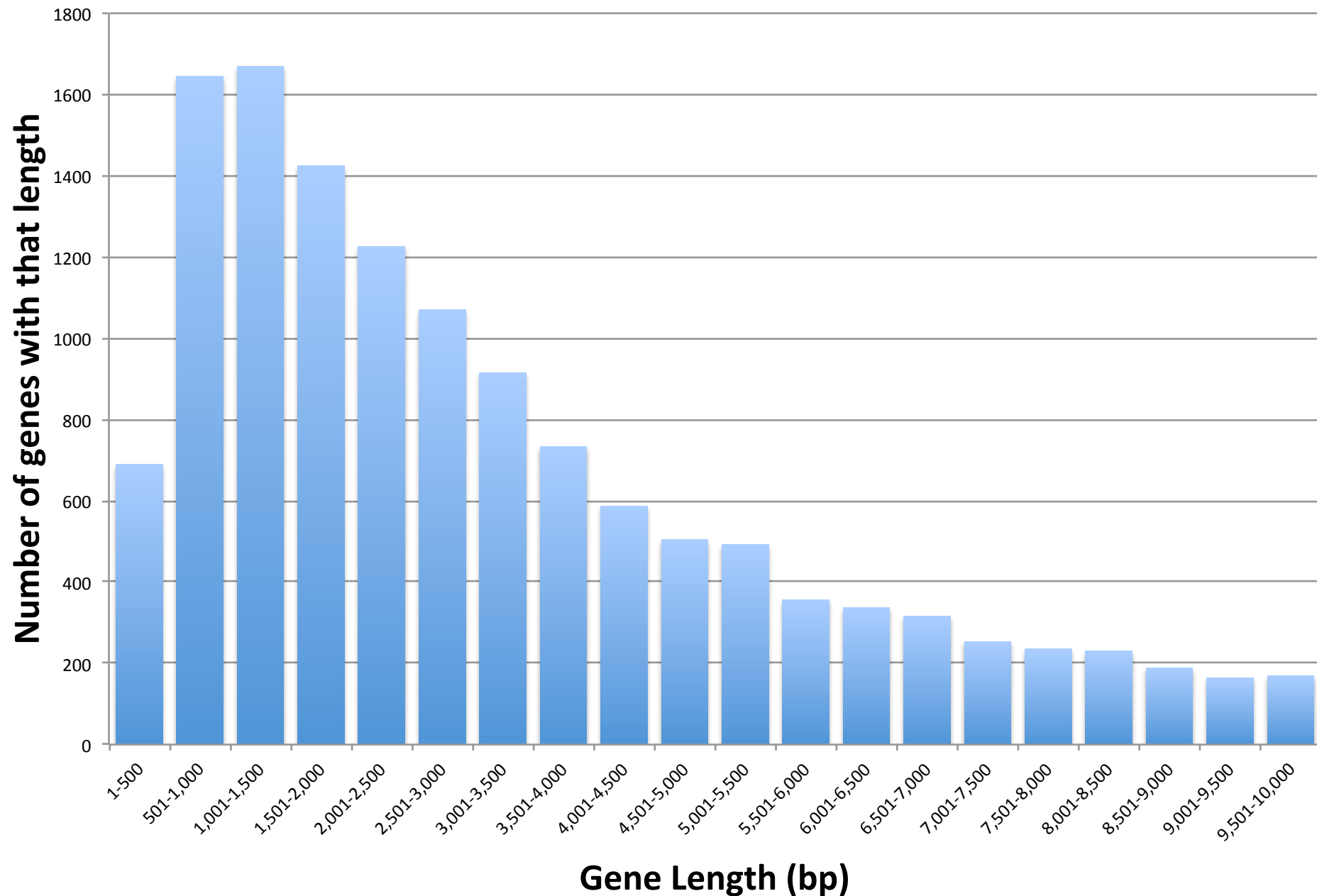
File S4. Raw data for Figure 2, Figure 9, Figure S1 and Figure S5.

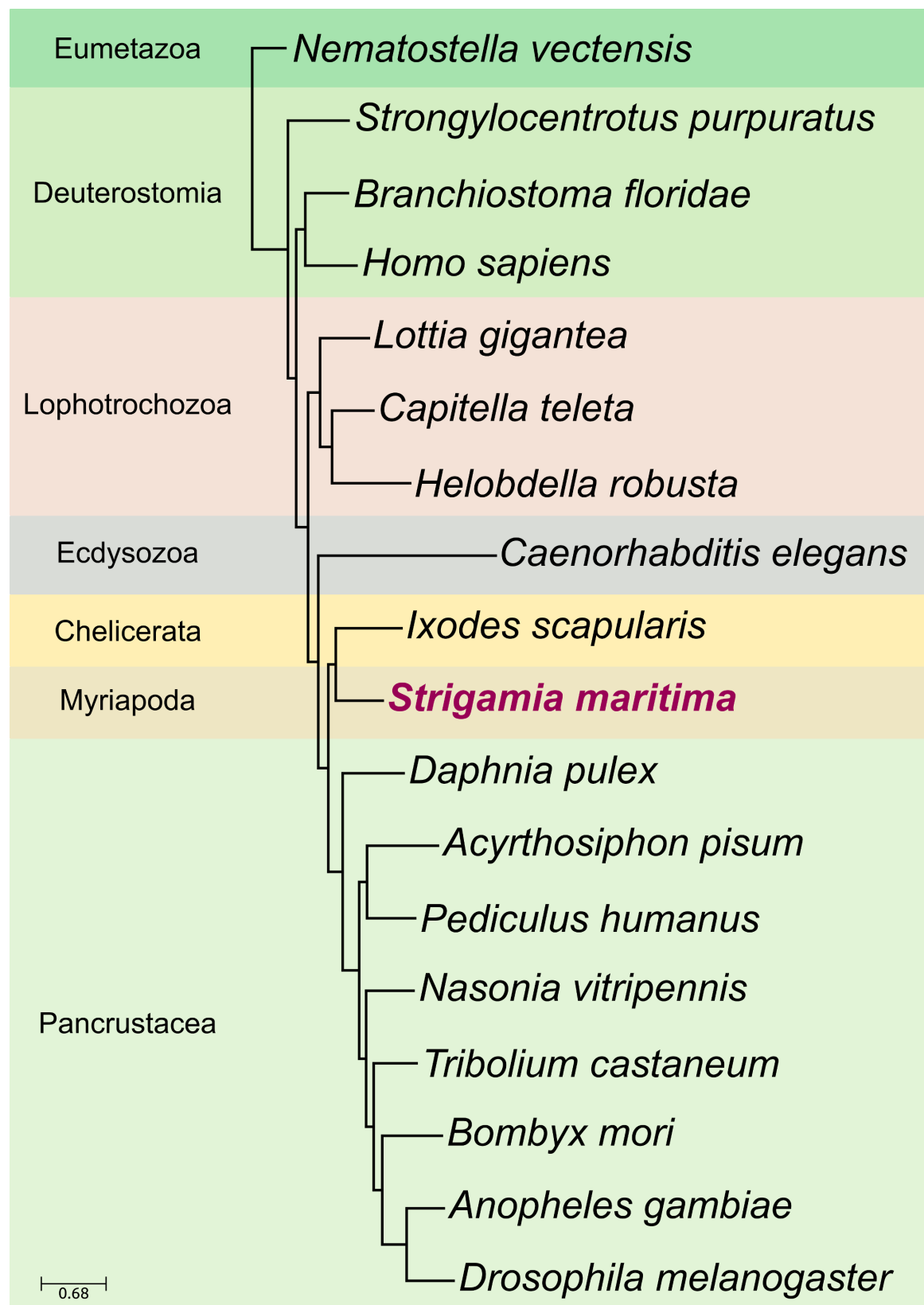
File S5. Raw data for Figure S28.

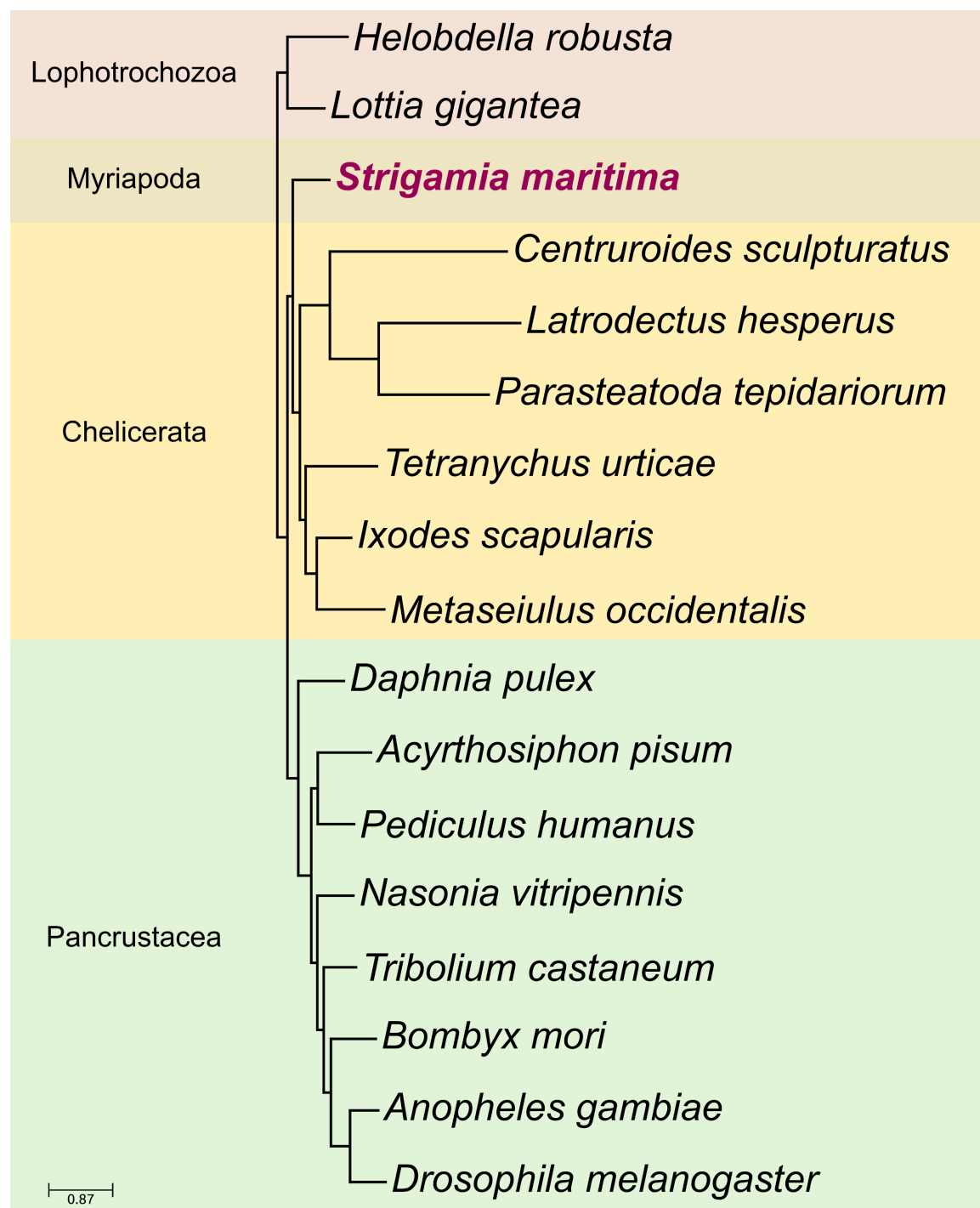
File S6. Raw data for Figure S29.

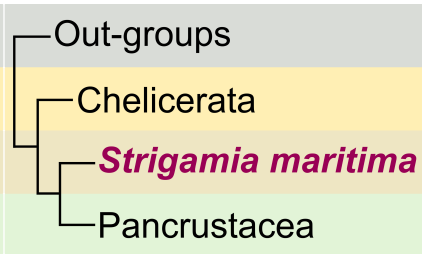
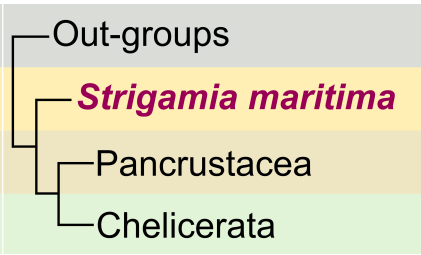
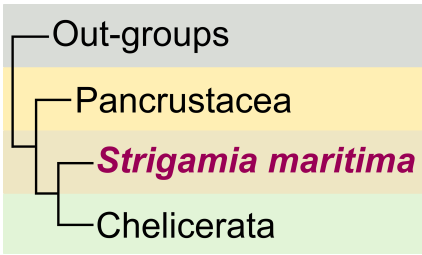
File S7. Raw data for Figure S30.

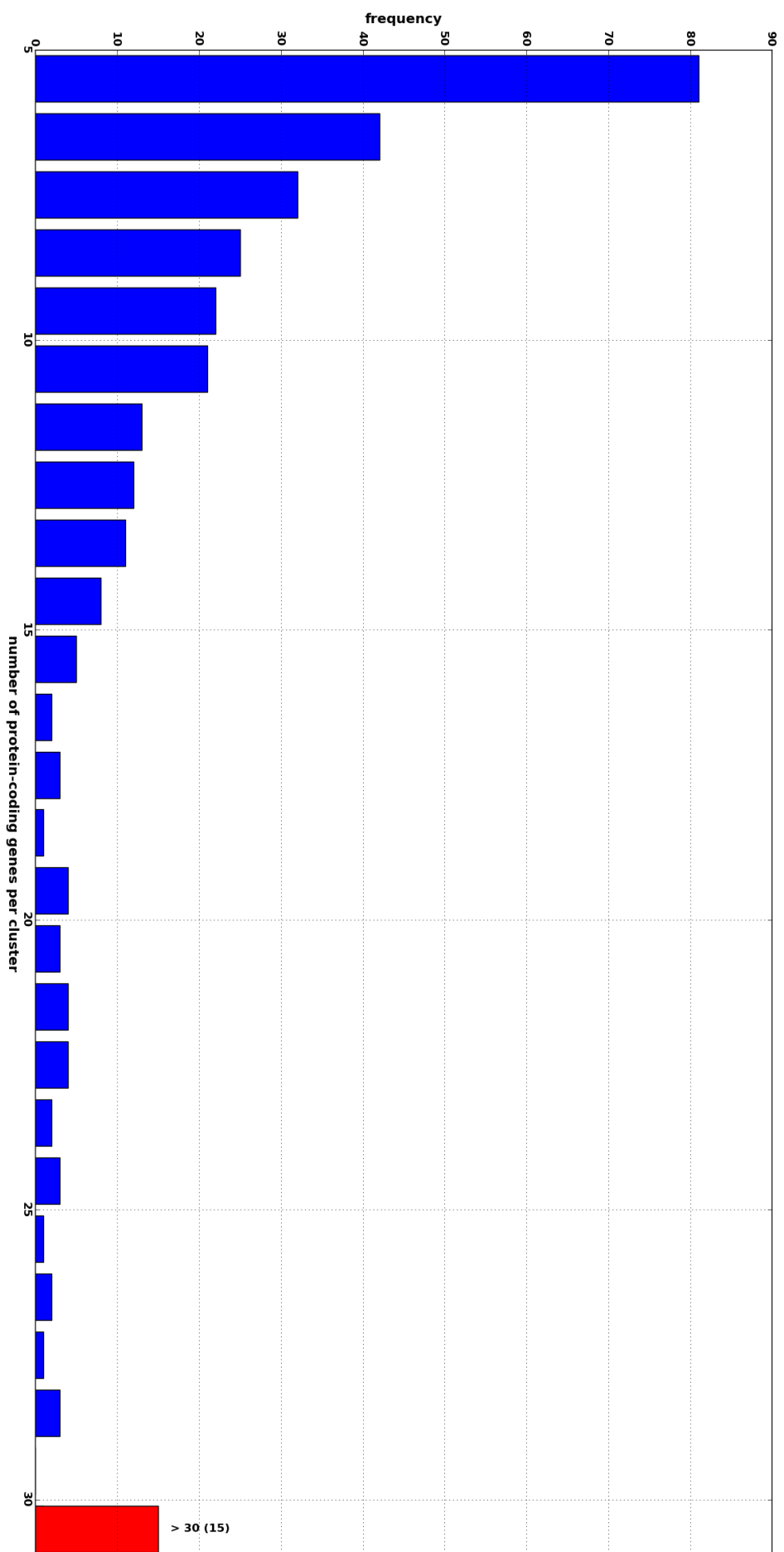
S. maritima gene lengths

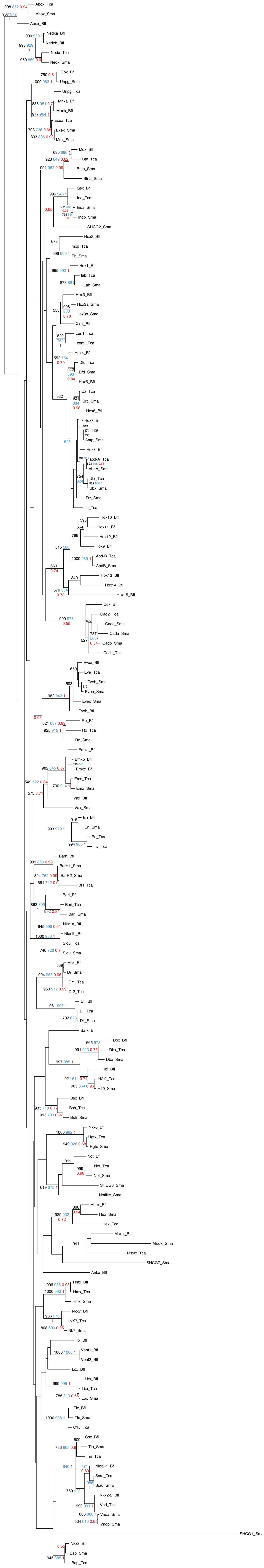


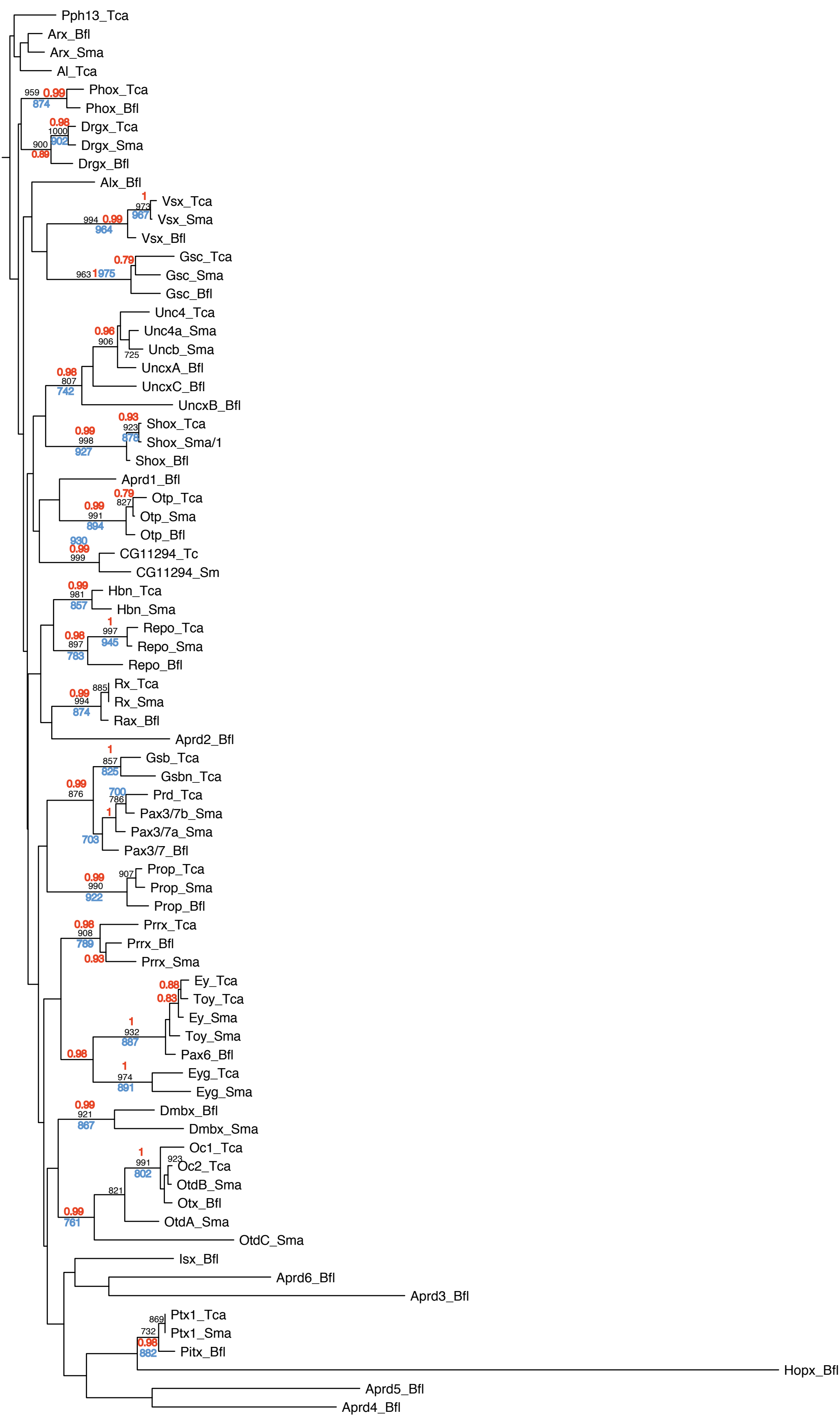




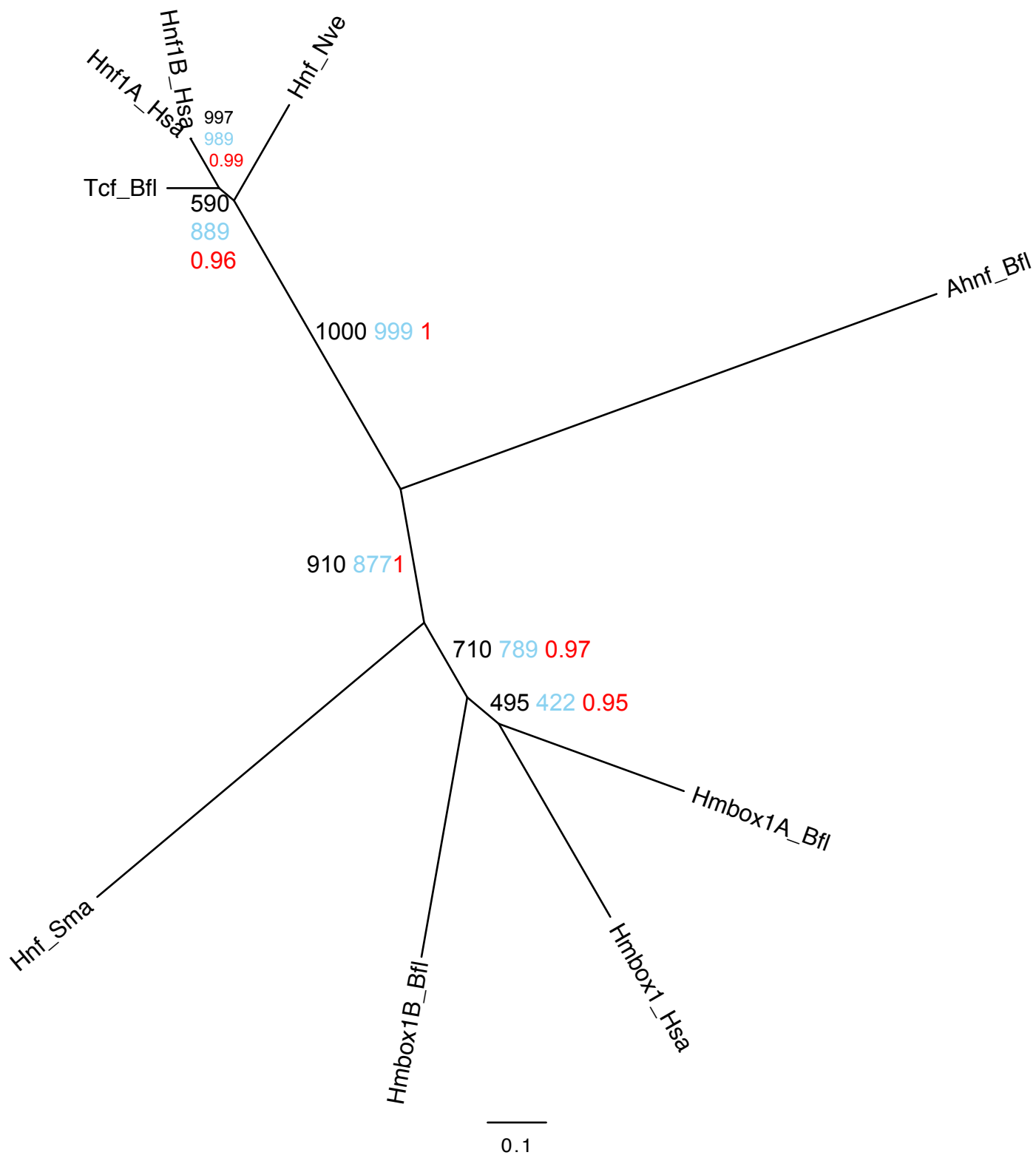


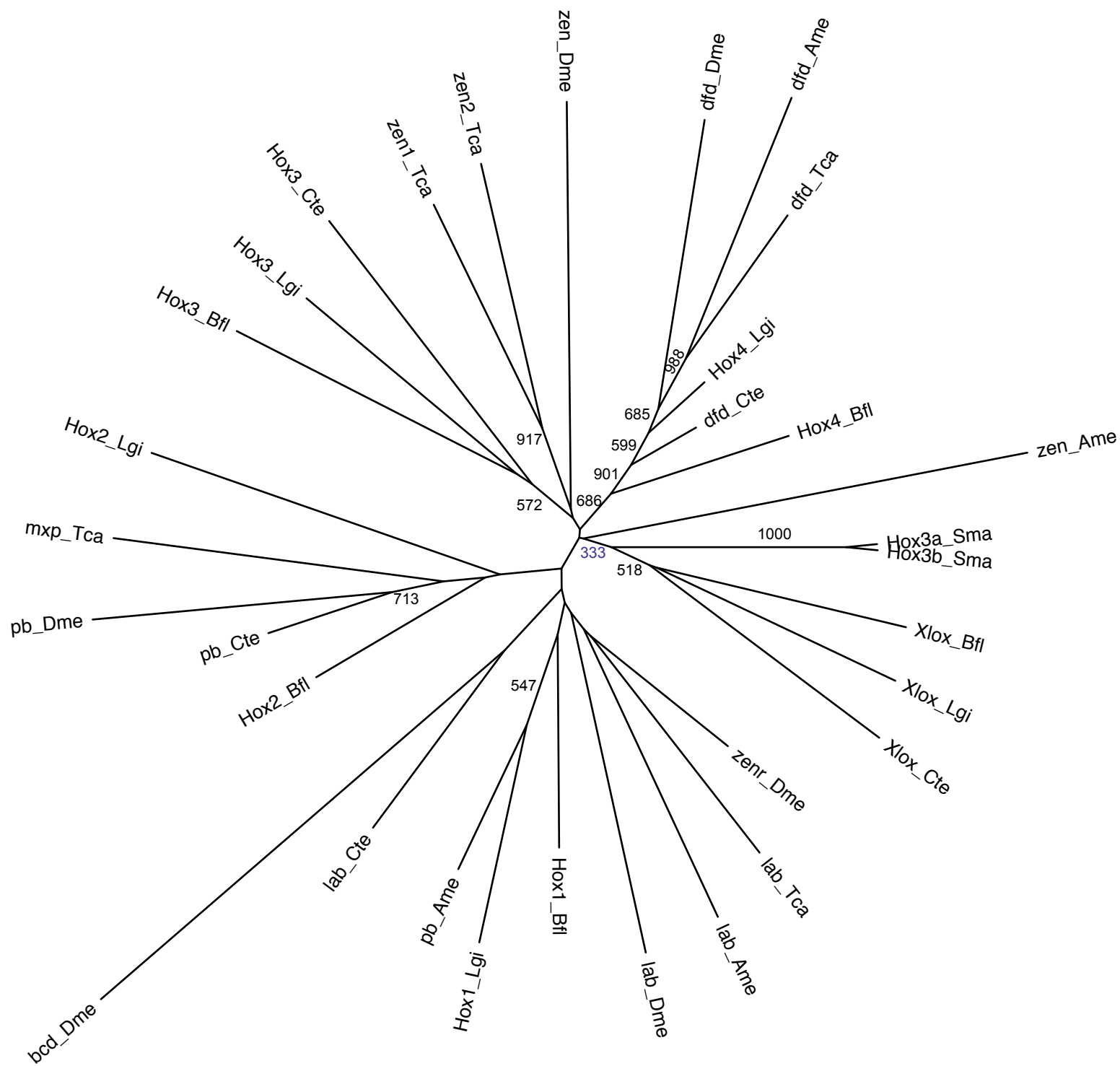




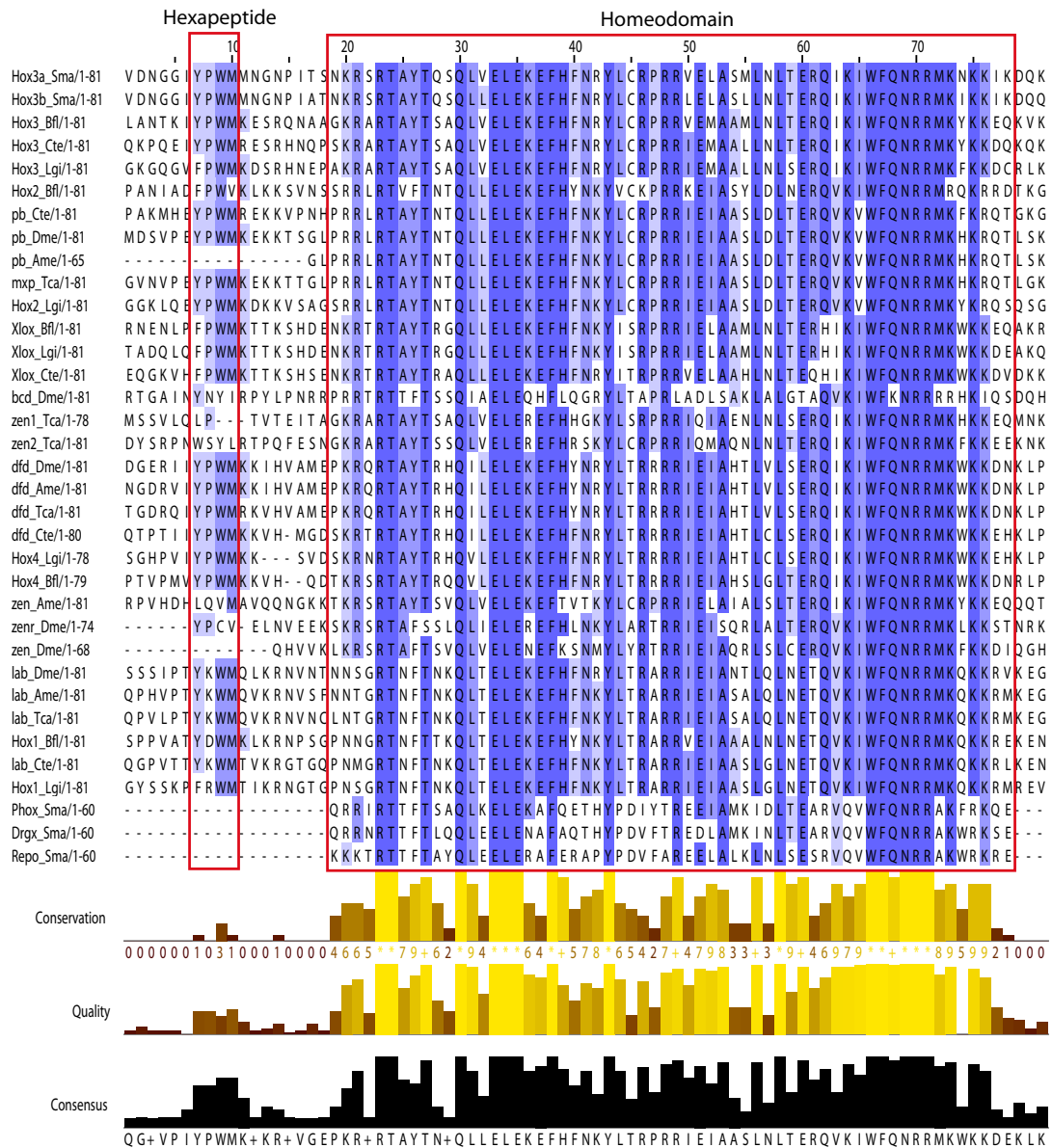


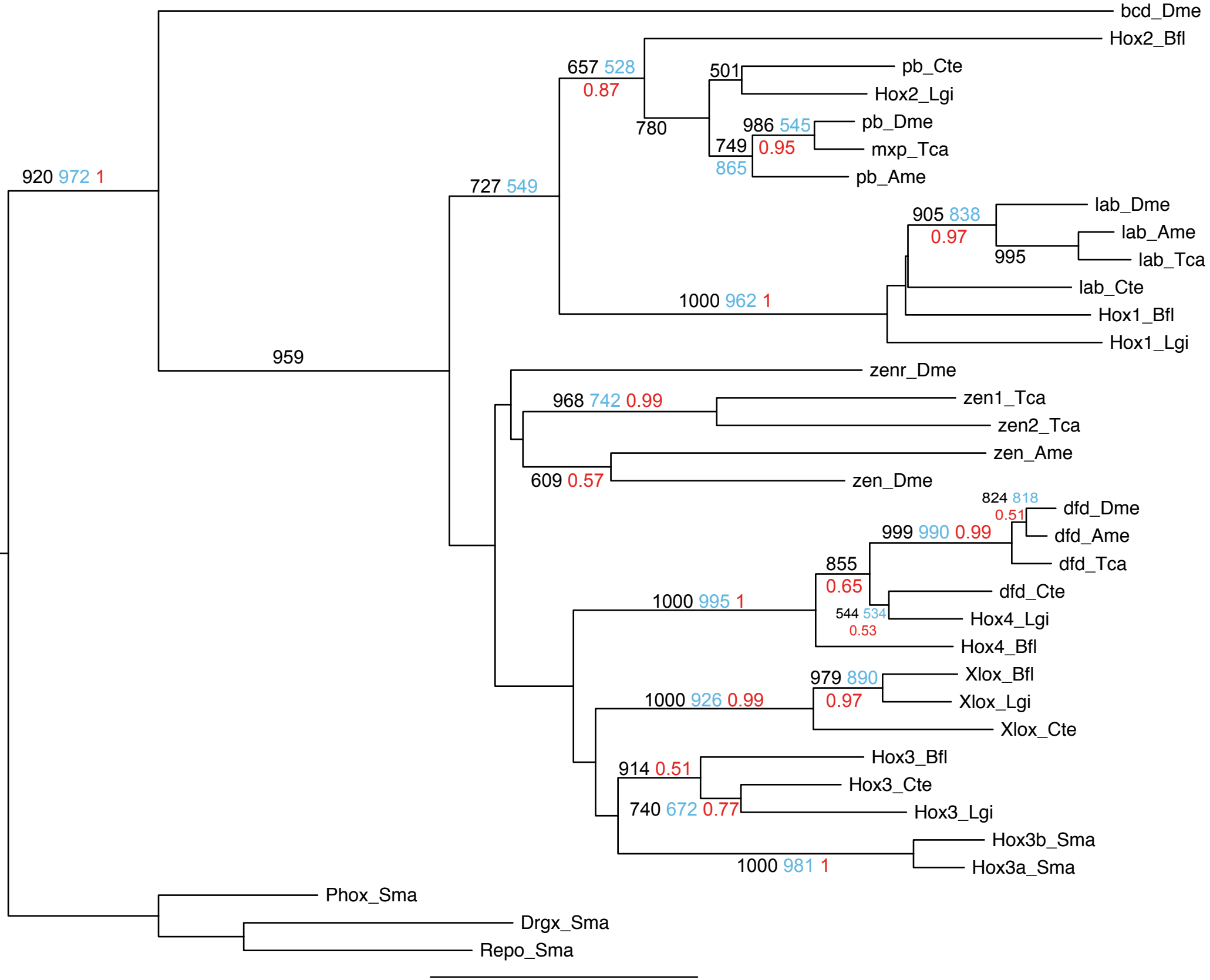
0.2

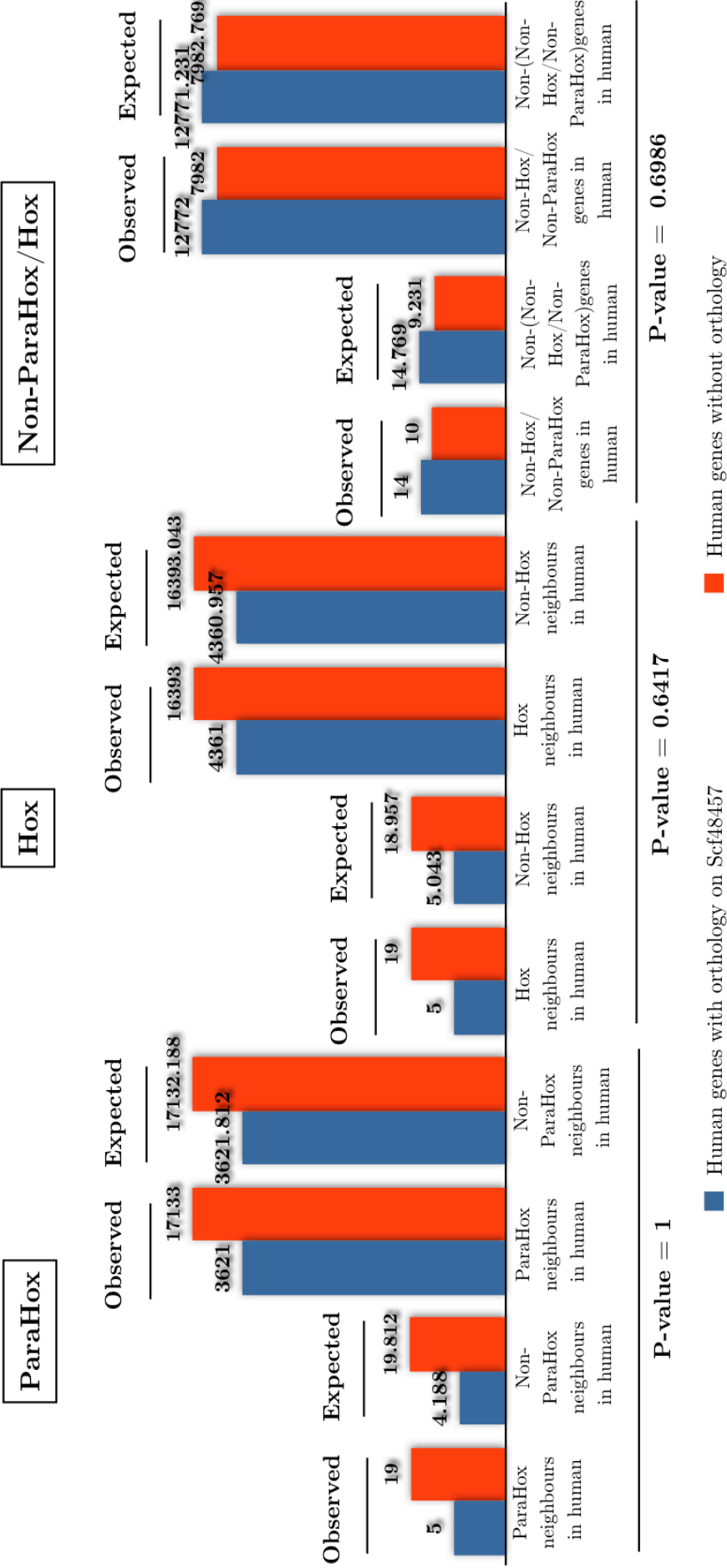


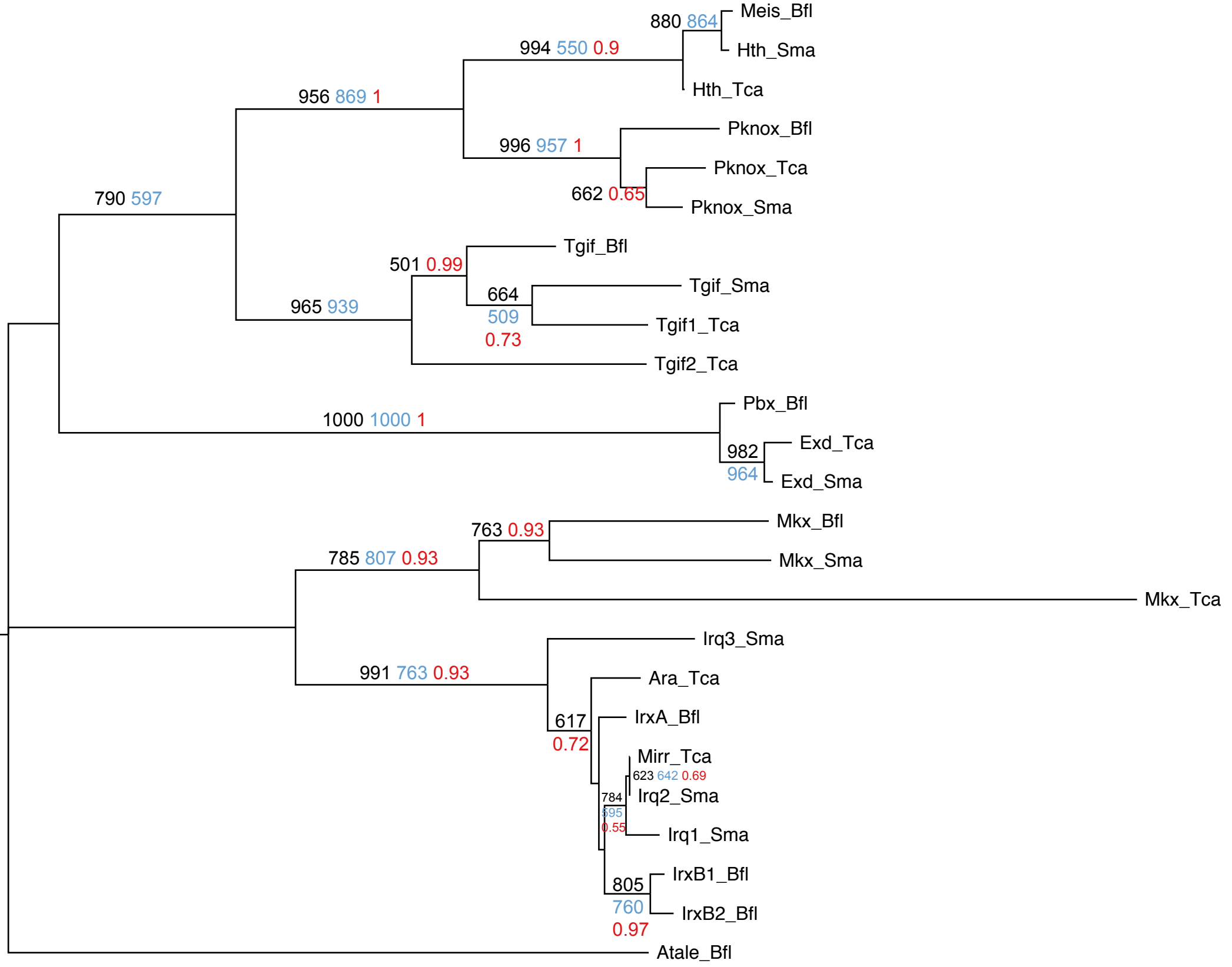


0.3

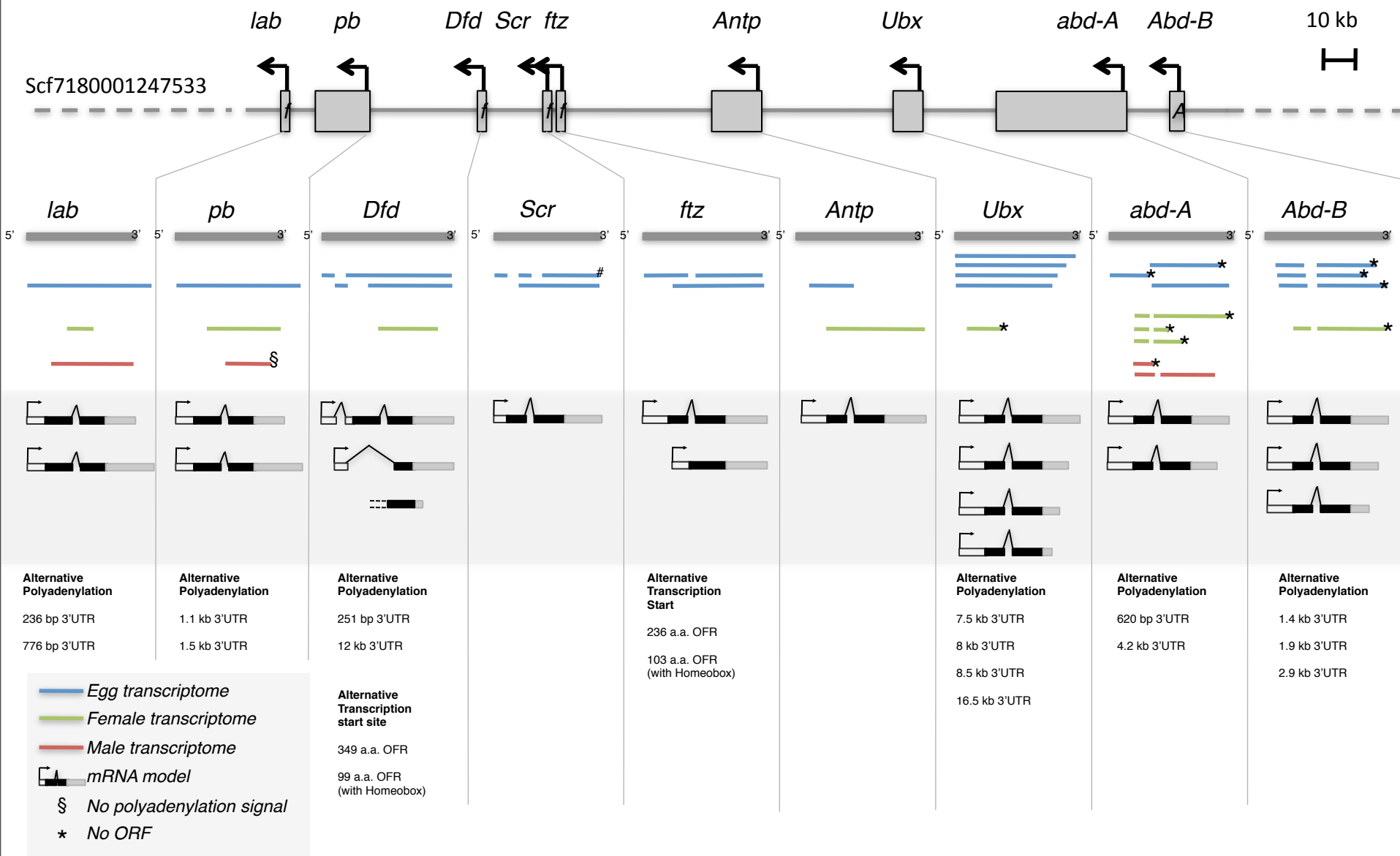


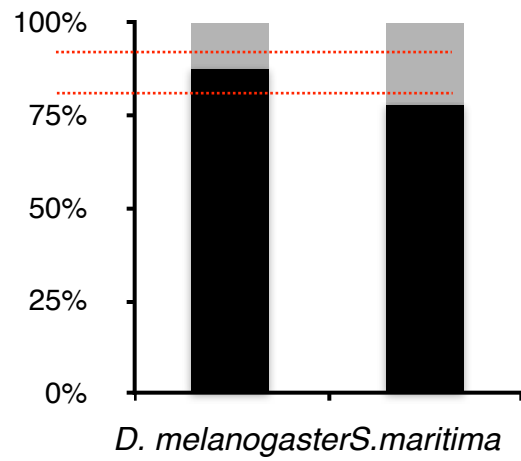
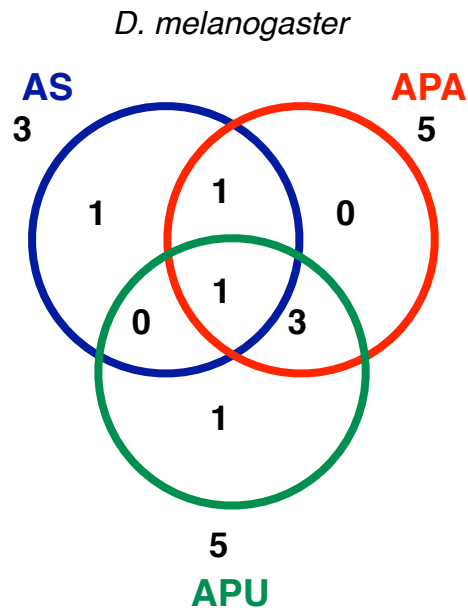
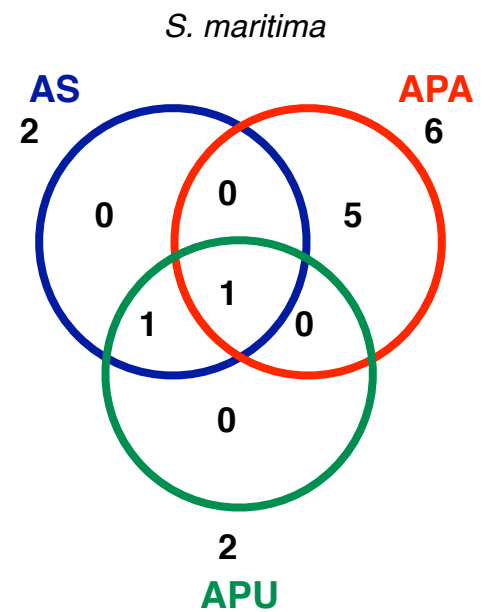






0.1



A**B****C**

50% corrected distance

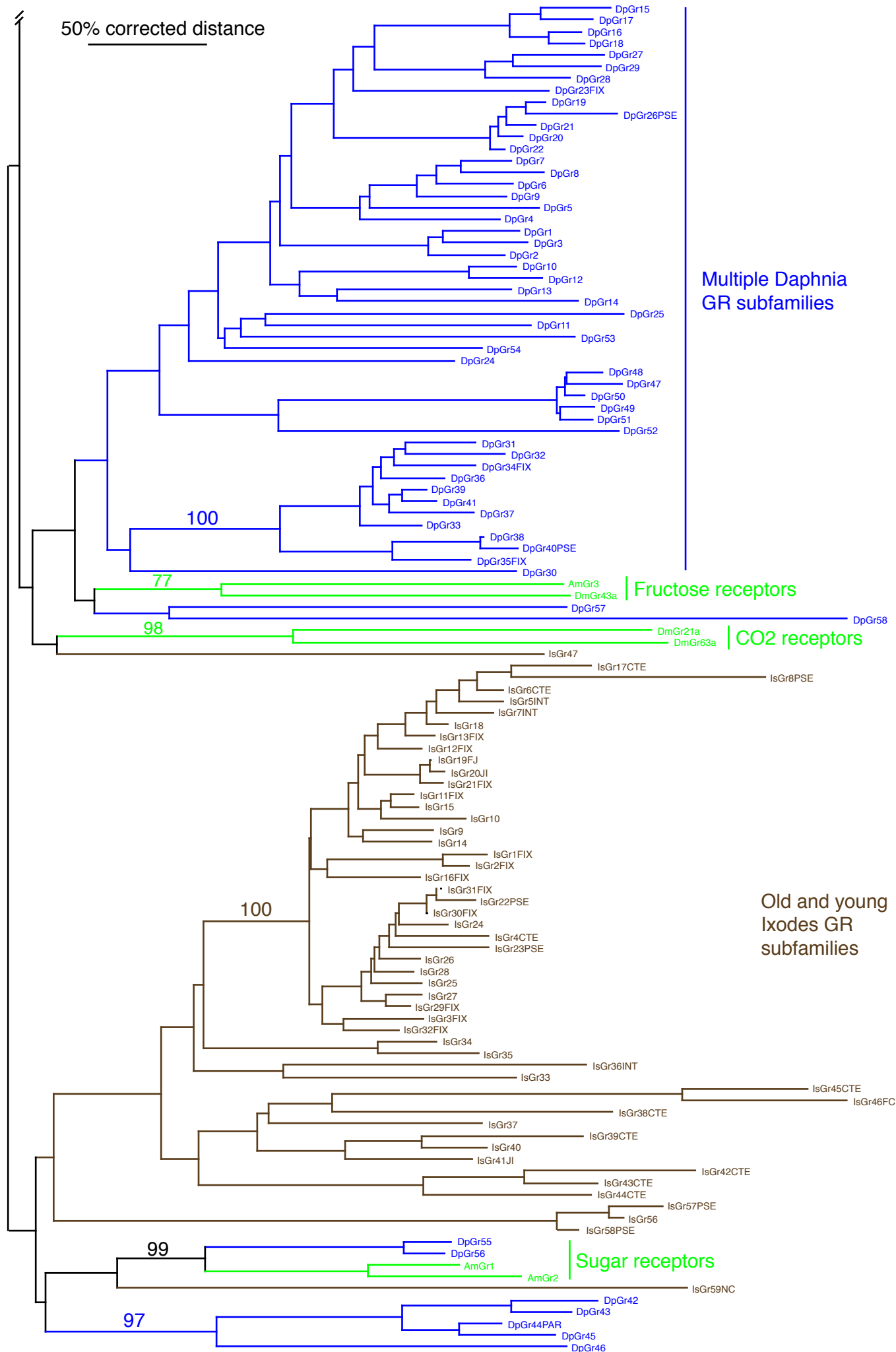
Multiple Daphnia
GR subfamilies

Fructose receptors

CO2 receptors

Old and young
Ixodes GR
subfamilies

Sugar receptors



Six young Strigamia GR subfamilies

An old Ixodes GR subfamily

>Smar-ACP

MKWIAVYLLLTIIIVLTIVAPVEGQVTFSRDWT**PAGKR**GMDCGFVKTKLLRDI~~AVLLQVKYHFAELCMLTLA~~
WLMLDGGQFTEILWTSCDGAFF

>Smar-AKH

MTKFTWLSMTLLVLMVFITVDVNGQ**IN**FSPGWG**Q****GKR**SLSDDKPVNGYSDCSETMIEVYRLLK

>Smar-Allatostatin A

MTFAVTWWCLLLTAP**TLLMSEYIYDISSES**DSNDQ**EKR**GLNTPWKLPEGYVYLRKVPSTGEYQIGKKDNQ
Y**RRRFSFGLGKR**FHSDV**LLENNEDEVGKRG**SQRNRHF**GLGKR**QPDY**LQGRIGRYNMGLGKR**SVDSREAE
VAIEEM**KRGASKFN**F**GLGKR**TKPY**SFGLGKR**WDDRGVEENLIEEY**KRAKTYGFGLGKR**DEEEMEE**EKKNR**P
Y**QFGLGKR**DRSY**SFGLGKR**MEEEEKKT

>Smar-Allatostatin B

MLLSWTSSVTIVLVIASVLAASASE**DKRA**WSDLN**GAWGKR**NWDQLRGV**WGKR**GAGQLPNSV**WGKR**EDAPSD
WNAFRGS**WGKR**NNWNKLQGV**WGKR**DSSDWNKLQGL**WGKR**ASWTHST

>Smar-Allatostatin C

MASSGKFCILIFALVLTLSHVTSK**SIGEH**EPNFNTDLSLVDDDGSM**D**TALIN**YLFARQMIKRL**QSSMDVT
DLQ**RKRSYWKQCAFNAVS****CFGKK**

>Smar-Allatostatin CC

MYSLILVFCVCMLTLPYVSCQ**IE**MNSLKNMAKTFHLSSDSNYFQHPV**KR**STMLLDRLVTALQAFKQETQ
EVTGMELQ**RRRPNGRVYWR****CYFNAVS****CFRRKK**

>Smar-Allatotropin

MKPVCLVILLFGLLVSATSSSTDEPANRV**QTRGF**KNSALAT**ARGFGKR**TMLNDLVDSTDRAIMSNEQLAD
LMSRNPQFAQQILTKFVDTDGDGVLSFRE

>Smar-Bursicon alpha

MVTLVFMVAAMMCIFSSRLATA**DECHIT**PV**I**HVLKYPGC**NQKPIPSFACQGR**CTSYVSGSKMWQ**MERSCMC**
CQEMGVREANVTLHC**PHARPGE**PKFRKVTT**TRAPVDCMRP**CT**SVEKHLVQPQESAPWLSDPNFNDA**ILSV

>Smar-Bursicon beta

MTNTWSAFAFFTIITFVLFITTKSLRA**FPEST****CETL**PSFIH**IIKEEYDSRTKLVRTCEGD**VAVN**KCEGTCT**
SQMQPSVTT**STGFLKECYCC**RESYLQ**EREVILQRCF**NFDGETLAGDMSLMKIRLKEPA**ECQCYKCGE**

>Smar-Corazonin

MGFQKTKLLLVASILVFIICTSG**QTFQYSKGWEPGRKR**AVDRSYQVRDWDAGRKRENIRGLDSSTAWILG
LKRAAFN

>Smar-CCAP

MQYVTIHGFVTL**LLIVF**SCTIGCAAQ**KRRPF****CNAFTGCGRKR**SELPASNVNDLETDLLLDKLSHQILGLVH
VLEALHMRIETSRQ**Q**QPLPMIETDSRIPNYILDRKRRSTKI

>Smar-CCHamide 1-1

MHGCRCNTALFILLVLSLLVSSATG**RRGCLNYGHSCLGAHGKR**SSNRQAMRRDIANFLPLHKTA**AEYNTIS**
EENLLRDESDTRKWN**DKWRDM**VISALEKDD

>Smar-CCHamide 1-2

MSTLKFI**FALLTLAVYVYQVQGLRGCTNYGHS****CFGAHGKR**TPKNDEKDQTSFLDSTENTKHVNQIT**TNDVP**
GVKNGLSAFLRKWMTALRQSGNDEILQ

>Smar-DH31

...D **KRNLDLGF**SRGFSGSQAAKHLMGLAAANFAGGPGRRKRSEE

>Smar-DH44 (CRF-like)

MLLRLFSFIVCVCAIACVGARSLRLOECTDCTLVPADDNRYSIQDDFNNKGFILKARRIPKWSLFANSPD
EVSSERMIHGFSMTRLDGT **KKRNDGTNLSIVNPVEVIRQRM**IDAARERQ**QIDANAEM**LEI**GKR**QPKAW
RSDWH

>Smar-EFLamide

MKLLCIDPSLFIFTYLVLVSLIPSTNLVSSQEIERVPESYGMDSSVL **KRSARISQEDVYRMFVLLNKIKDRG**
GI **KRIGSEFI**GKRYSQAEDIE **KKLGSEFLGKRGIGSEFLGKR**SQRSQAEN

>Smar-EH

MRQSESQRAAVSTVVILLLLHLDRASS**RSINL** **CIONCAQCKKMFGPYFEGQLCAETCIELRAKMTPD**CADA
NSISPFLNKFER

>Smar-EH-like

MACLLSAACVLASVACISEESSLGVCIRN**CGQCKRMYGDFFLGQHCAEECLQTEGRGDLPC**NNPKSLYRF
LGKT

>Smar-Elevenin

MKIVIRSVIPLLICLILLLVFVSHSEQVDCRIYVFAPK**CRGISAKR**GIDPPGRQQPNQKLDINADPYAPY
EYNPTDYSWDERASDSNYGSRISDFVAPIFDSSSPSRFKNSDDRDQGRFLRALVKMYFDQRQDSED

>Smar-ETH

MTTFTSSFFVPTQMYILVLFAVFLLQTEA**OFFAKTSKNLPRI**GRRVDHQEVSEIIPPSFKHLLAFVRKFD
SDANGCLSPHEELIEIPLFQMAFENEDFTPLEIAADVVEEYKSEVEGNLKDIVAKVLAASYAEKK

>Smar-FMRamide

MQQVLLSFLVVTIPALAIPVSARCFLOPSLGGVPNENSSPLLCTLMSRESADEVSEDDDKLMDSDALEPI
NNELTSEENSMDDDDQKDALL **RS**LRQEPGHKFLRFGRASSNHNFLRFGRDPEHKFLRFGRDQHKFLRFGRSV
ANGEENRLRLRENQSKMTVLPMFMR**LRGRG**PEHVFMRFGRQGSNSEGHKFMRFGR**TQNDTPVQDENTQQEKA**

>Smar-FMRamide-like

..TEQKFARFD **RS**NFDENGDPQENWLDSENKLLSLRSAR**LI**REL**GRKRG**SLEKNFLRFGRSLENGNDINT
KRDTSSLLDDSHMSYLKSDQLNQNKLV**PKRN**KLENNFLRFGRWK**TENLNEQYE**

>Smar-ILP

MSSHLRTAAAVLILLSVVLSIAAQESAYFQELFTPHEIKIRAKQKYCGRN**LVEVLQLVCARNVLNNLEAEE**
ISDLLGNDFARSLMGIKHKIQTRGITDE**CCRKGCSF**NELKSY**CAEEP**

>Smar-Inotocin

MKSTHFVNIFIYSVFIFIMADGCYITN**CPPGGKR**SGNEKSGRGVRQ**CTP**CGPGGIGR**CYGPDI**CCGANVGC
FVG**TRESAI**CRLENLYSLP**CQ**NEGRAC**GT**DGT**CS**ADGF**CC**STDQ**CKADES**CRGKVHHTNNLQ**RVL**DGEIDL
NDVMGPQR

>Smar-ITP

MNHTLFFRVFVVLSAIIASTCLVSARSLNLEDVSGVHRIN**KRSFHTLGCLGDYDTAGFSRLDRLCEDCYDM**
YRDSQVRAMCRSSCFTTDTFKKCAEALLVNMEEEKLGDVVNRLYGRD

>Myosuppressin

MKLLTLYKILLVISVVIIPSIFSHPPPQ**CD**TDDPLPPRL**LRICNALRTISEYTELMEDYLDEEVMHTLAVS**
DM**KRDERDTGHVFM**RFG

>Smar-NPF

MSLSISTRSTLALICVIVVLYVFCAPAQATSGPDQAVSMTEALKYLQALDKYYSQVARPRFGRSLPMRYPS
KDMSVEAEANRLLEQRRR

>Smar-NPLP1

MQGRTCLRMLLLAIVFSLTQGALQDTGPEATTISGRLTELSGQQERIKPPVKRYVGSVARAGGLPPFFHG
KRHEIESTDDEDENNEIIKRYLGSIVRQGIFSHNKRQDEEMLDDAVEKRHLGSVLRAGDSRLVGRDLFSS
LQDKRFMGSLARAGELGPGGRMSGKKRYLGSVARVDGIPFRAKRTPODTESPDWESEENLDDNEEDVGNEE
EWEEDDLLIPFKRNIAASLARNGWLPHTRSLRRHDAPTYSVSEEAPKRNIYYRPASASGRSRLLGWLRREEA
RLKEEAEARNGQGKRTIAALARSDELPRYRFSRSADRLPAPPFMRMATYRGGGGGGYSIRSPHAFASLHE
GGWNRFKRSFEQLDRMQMHLDALDDLCDSDWDLERCGRPHKGEKKSMEHSSSRKDEGGEMSMVT

>Smar-Proctolin

MTVKCAVVFALLMTLMYCWSAQCRYLPSRADNTRAEEIREILRE...

>Smar-Pyrokinin

MWVSLCVWFGIFCSGFLSLFEVDKRQGLIPFPRIGRALPVDNSILLSSQEIRDALYQGLFRSNRLKRDVE
VEGGDDWAEITNDGFQGGPDEKKTVEKTSKLAPRLGRSRYRIGASPFQRLGRAYVSFGPRLGRSKLPPRL
GKRLNRN

>Smar-RYamide

MFSSKQSTLFYLLTILGMCALFWRVESQQFYPNGRYGRSDKMPALSDVRGTREMTVSFFGDGTVQCTYTG
PDFYRCK...

>Smar-SIFamide

MASKTTIILLVVAIVAICLVVDVTSANYRKPPFNCSIFGKRAPEDSTAEKLFAMCAIATDACSQWFPASEA
K

>Smar-sNPF

MHSIITCSFLLFATLFIITFPLENASPAPYTDYDNIRELYELLRNEALNDRTGHQVVRKGRDPSLRLRF
GRRSDPAWQEGREPSLRLRFGRSADDVRGYSKQLRFGRSDDTAWQHDVASENDVIEN

>Smar-Sulfakinin

MNCTVIFLVHYLVLVCTFVSSNGSPSVARSSKHDSRLANIMAPYLYLKLHDQAARRPVDSESESVKDEVE
SEFDDFFESFDKTGRNFDDYGHMRFGKREFDDYGHPRYGRSA

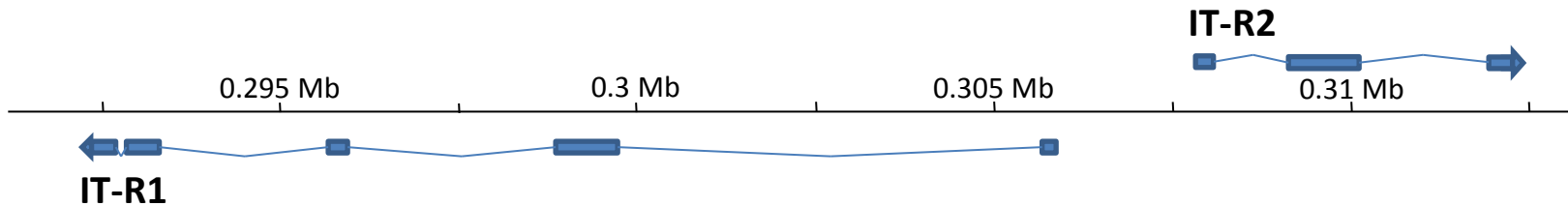
>Smar-Tachykinin

MMNFRGIWSLRLSIFIIFFGPIFGQOVQSLNDKIKNMQLDEDNLLRKS RNVFMGMRGKKMSMDQDRDLQR
AATSIAEI KRINGFMAMRGNKINSIHNIFFDEVENPYIFPGDKRAKGFLGMRGKKQPSGSEMNRWSNSKFFA
MRGKR

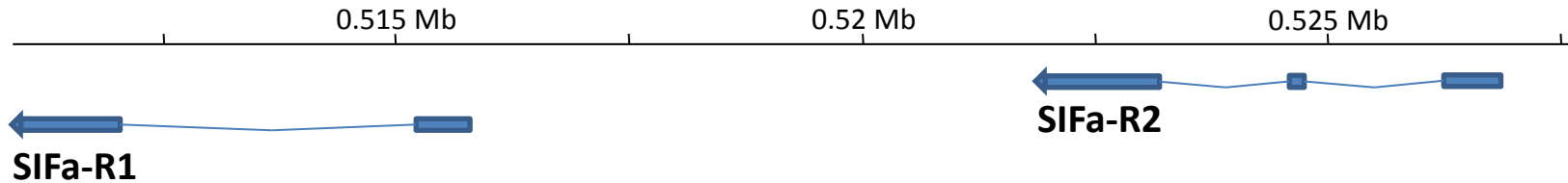
>Smar-Trunk (PTTH?)

MMKRKKVSFLFGFVLGFYWLGGQGNATGENVMTNFLETMGKATSKTSKTTETNDRLWTCEWSEHWLDLGN
DYFPRYIRTAKCTTEKCFWFNFFKCEGRAFTVKVLRKRQDECI SHVNNKTVILLEDWVFEERAVNFCCVCVP
VWKKKKATLIFFFFSS

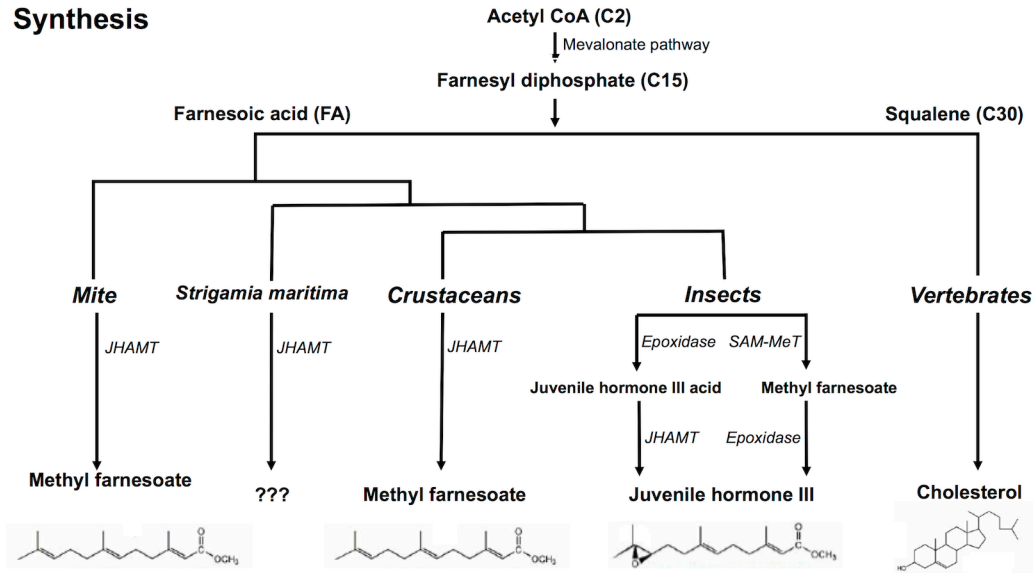
A



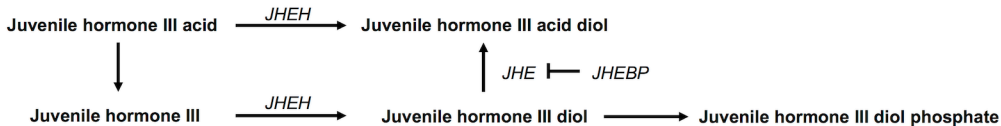
B



Synthesis



Degradation



Bone Morphogenetic Protein subfamily

Decapentaplegic

BMP10

ADMP

BMP3

Glass bottom boat/Scw

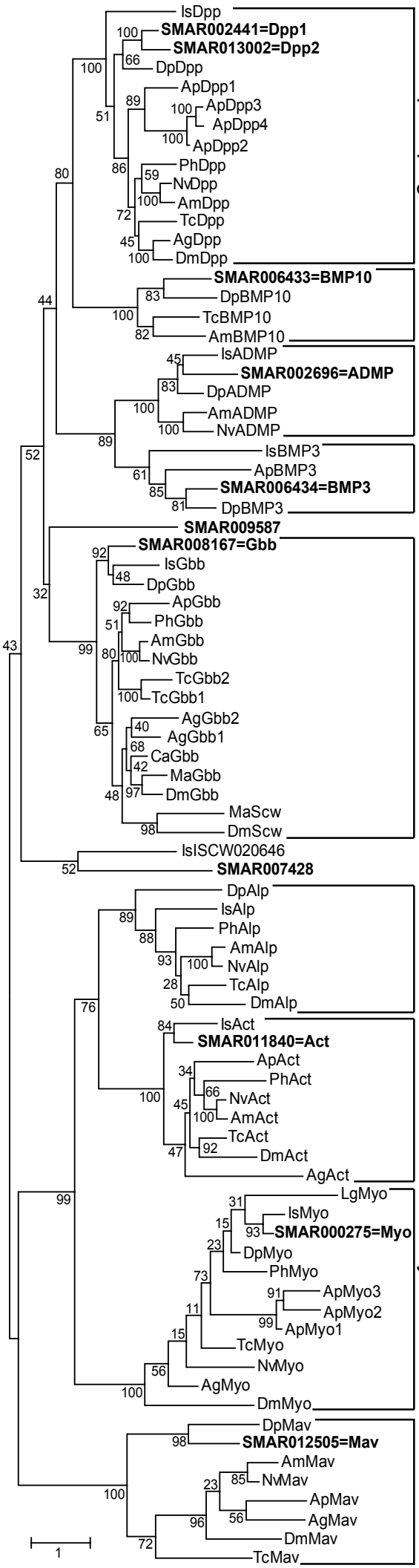
Activin-like

Activin-beta

Myostatin

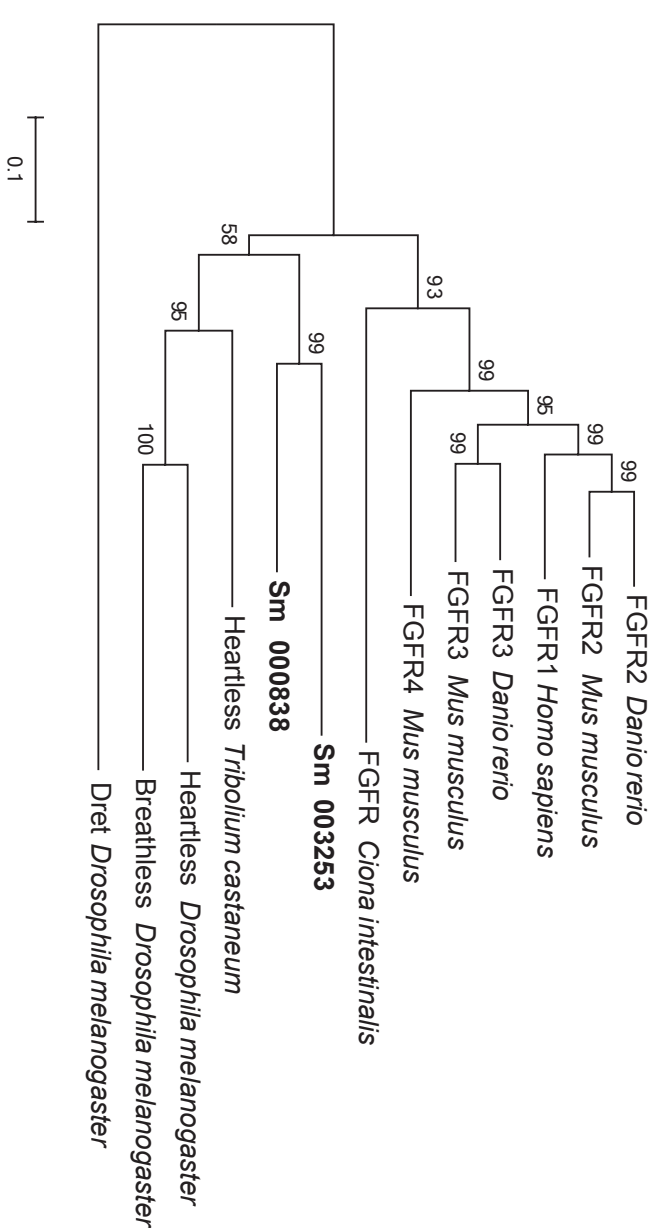
Maverick

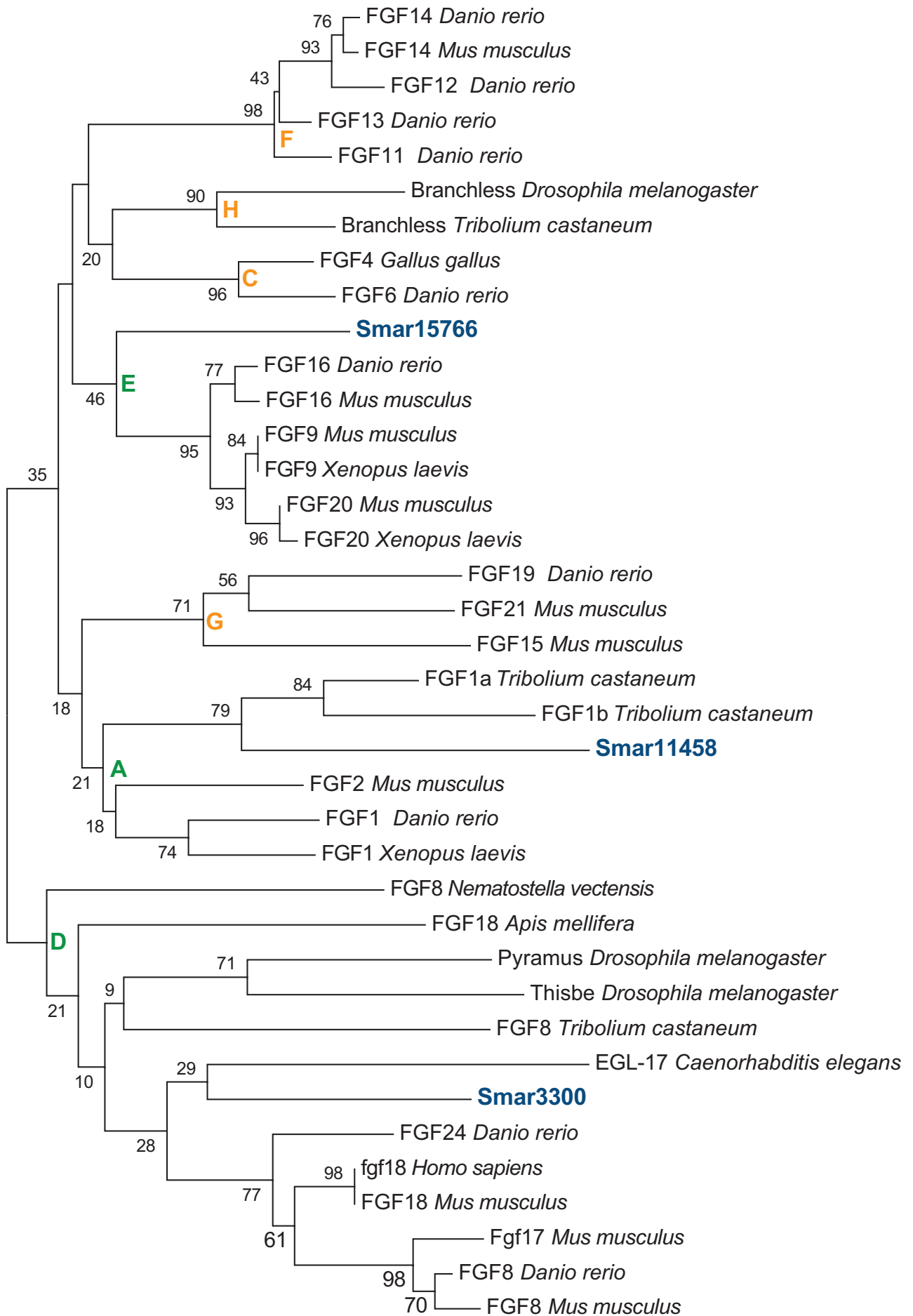
Activin subfamily





FGF-Receptor





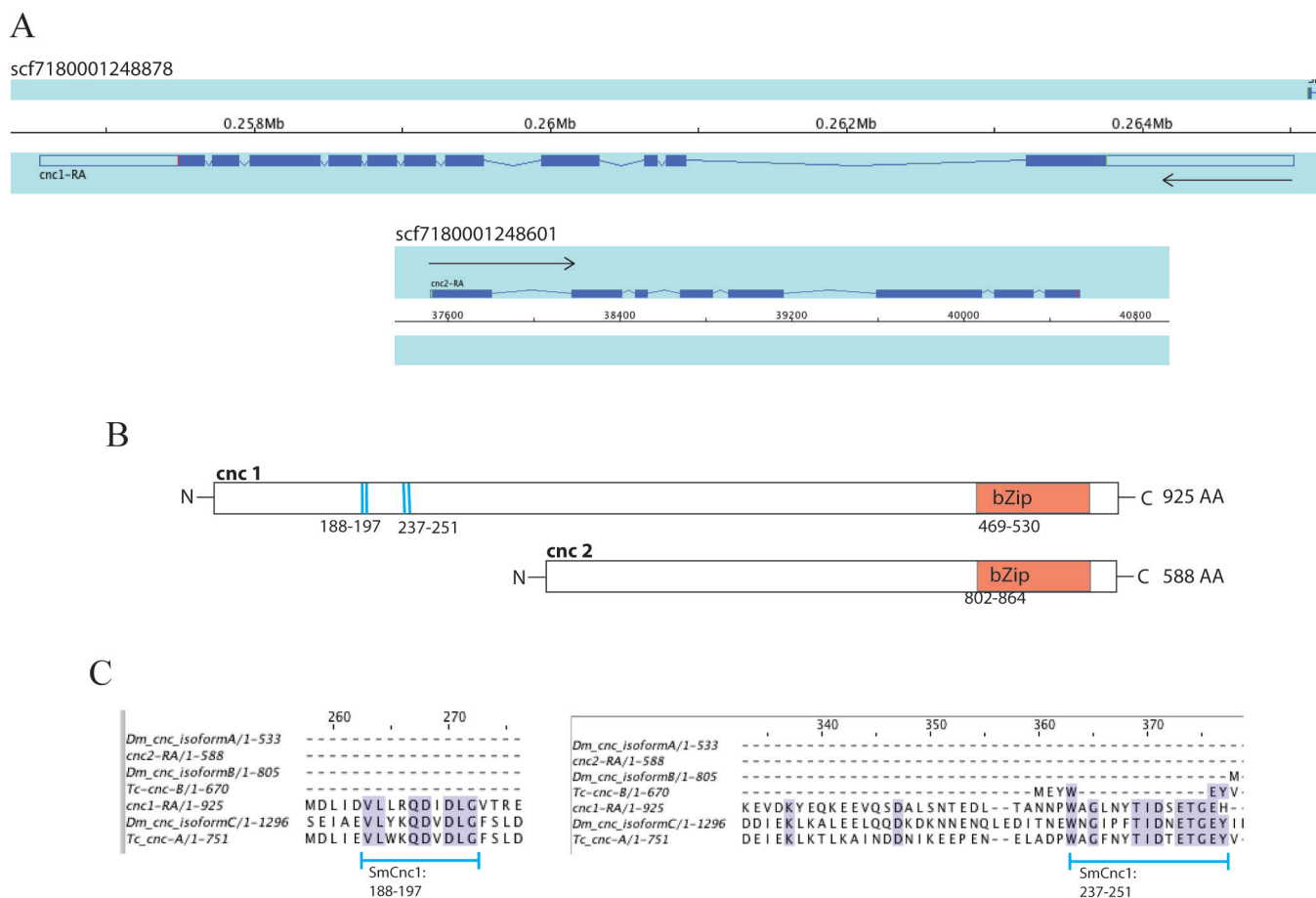
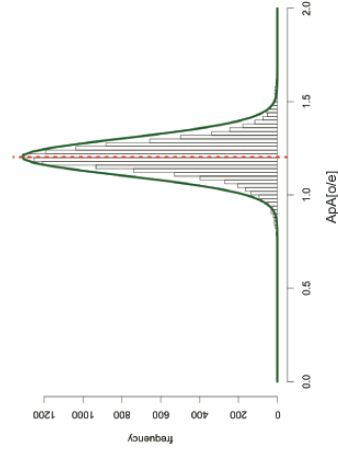
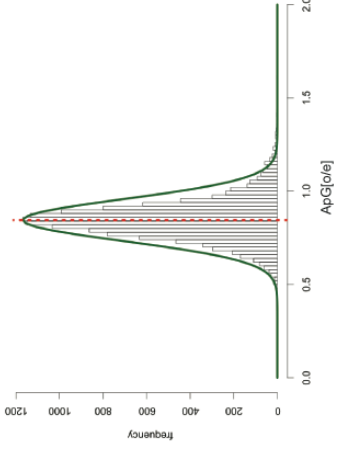


Figure 1.10: The *Strigamia cap'n'collar* (*cnc*) genes. **A** The two genes are located on different scaffolds. *Cnc1* is a long transcript consisting of 11 exons. *Cnc2* is shorter (8 exons), the 3 exons at the 3' end of the gene that encode the C-terminal region of the protein including the conserved domain (**B**) show a similar structure. **B** *Strigamia* Cnc protein structure. Both genes contain the bZip domain in a similar position at the C-terminus. *Cnc1* encodes a long protein (925 amino acids). Bits of the N-terminal region (blue lines) align with *Drosophila* Cnc isoform C and *Tribolium* Cnc variant A. **C** Cnc protein sequence alignment, only showing the aligning bits in the N-terminal region. Blue lines shows short stretches of sequence that form a consensus motif. These motifs are not present in the proteins encoded by *Sm-cnc2*, *Dm-cnc isoforms A and B*, and *Tribolium cnc variant B*.

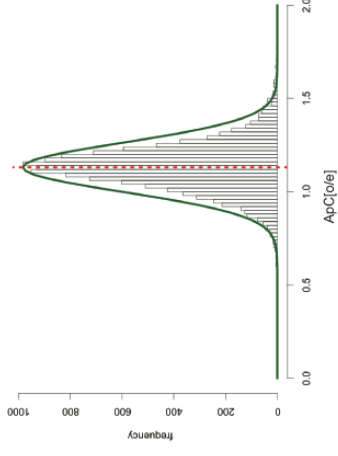
A)



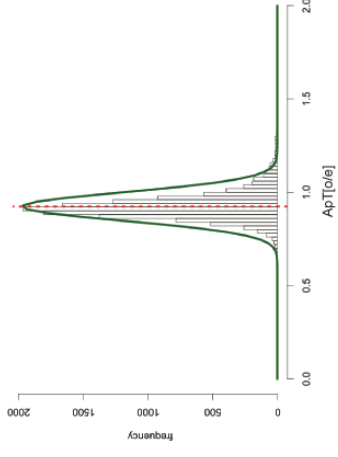
B)



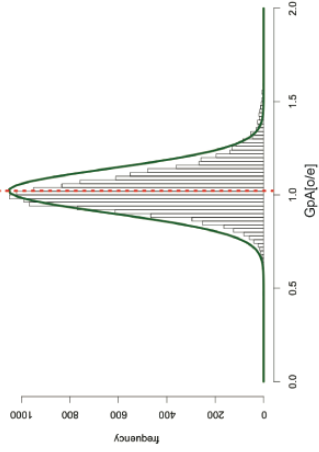
C)



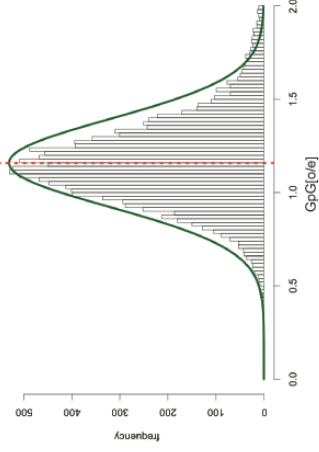
D)



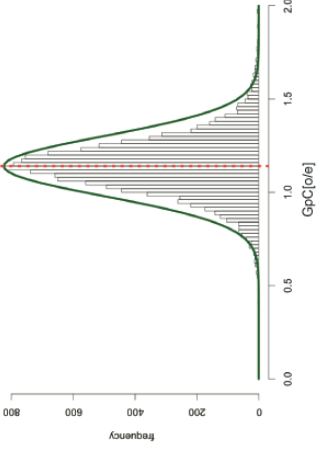
E)



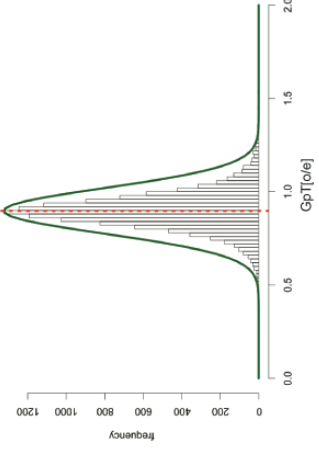
F)



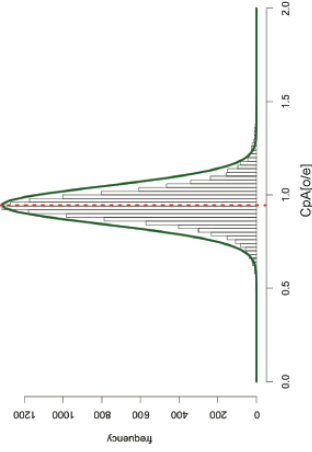
G)



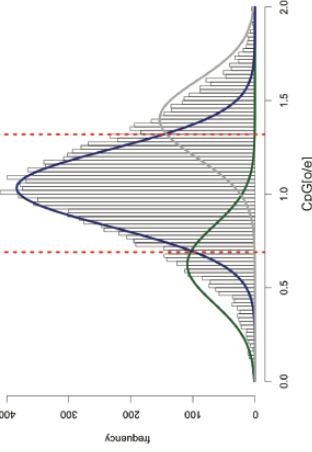
H)



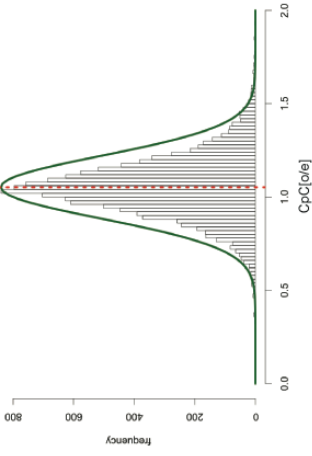
I)



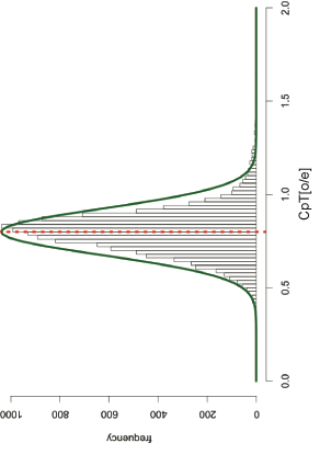
J)



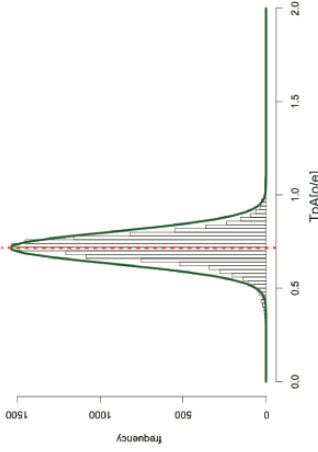
K)



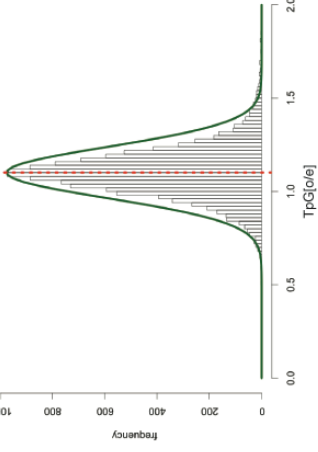
L)



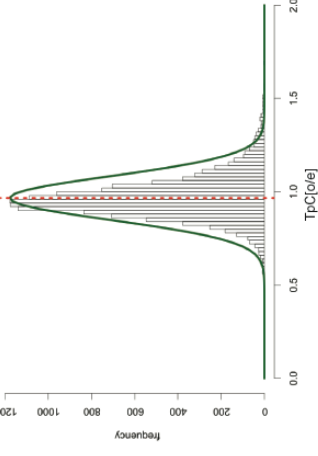
M)



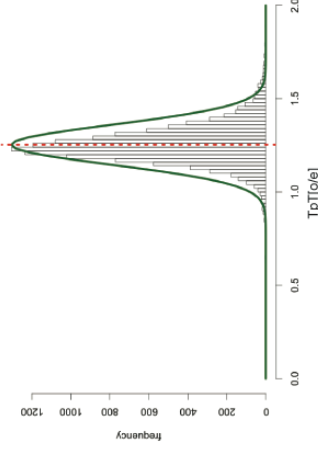
N)

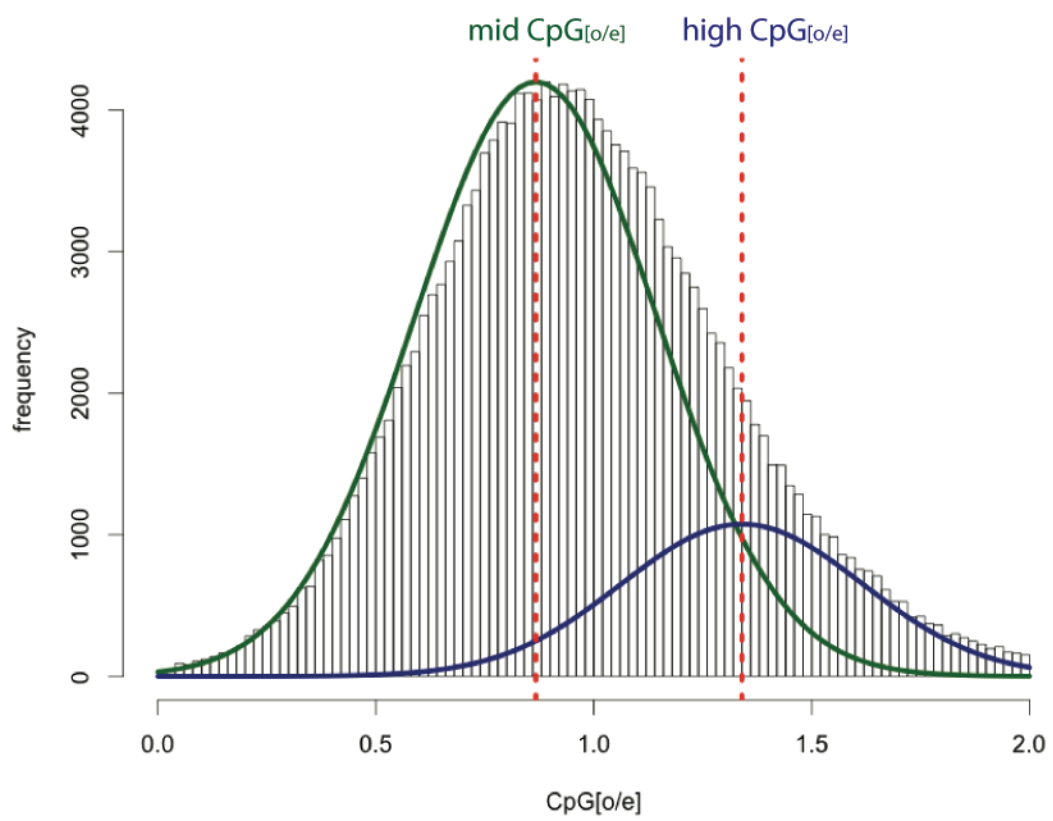


O)

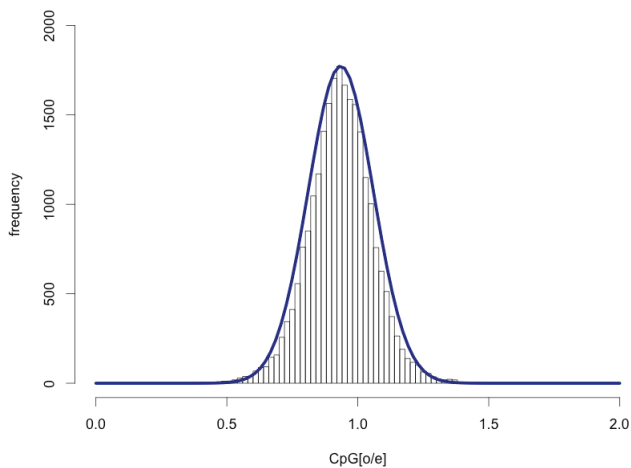


P)

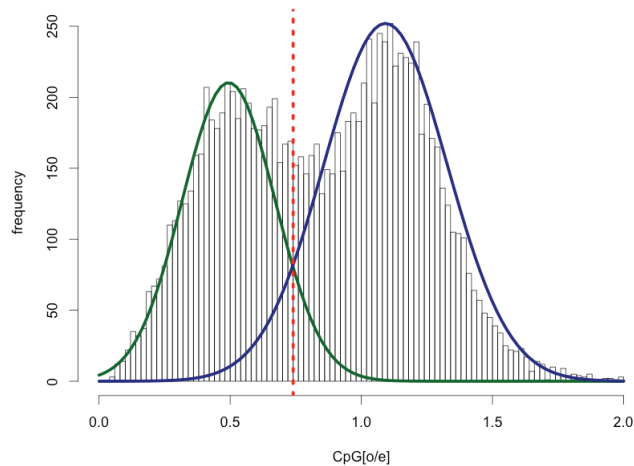




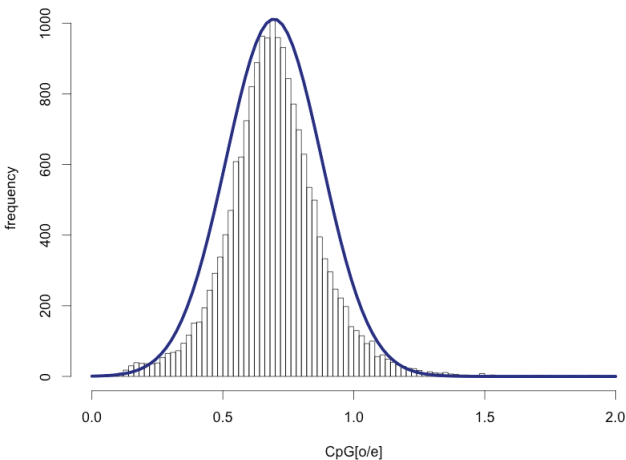
A. *Drosophila melanogaster*



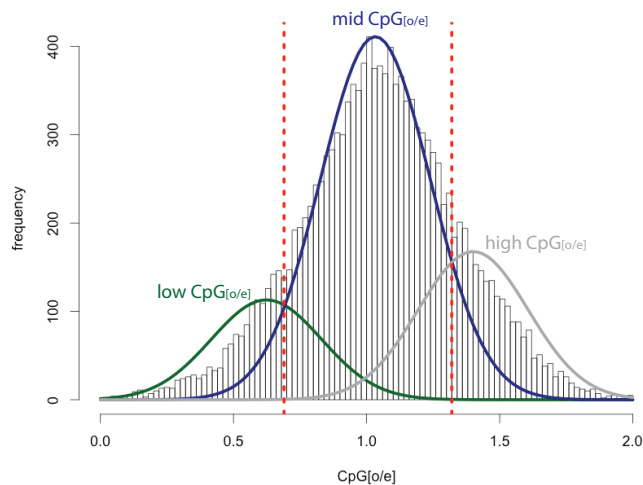
B. *Apis mellifera*



C. *Tetranychus urticae*



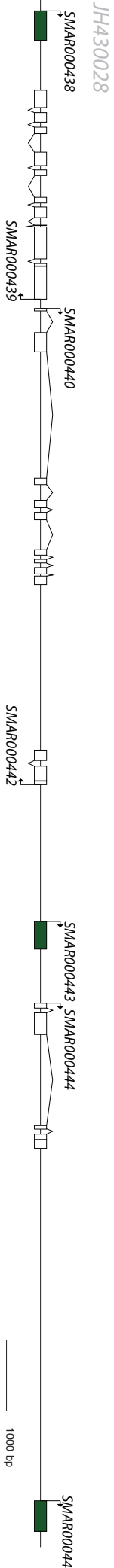
D. *Strigamia maritima*



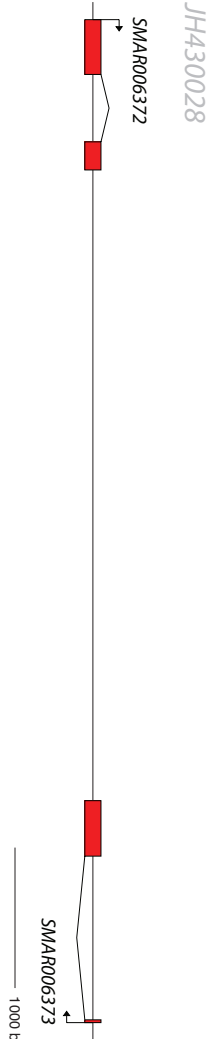
A.



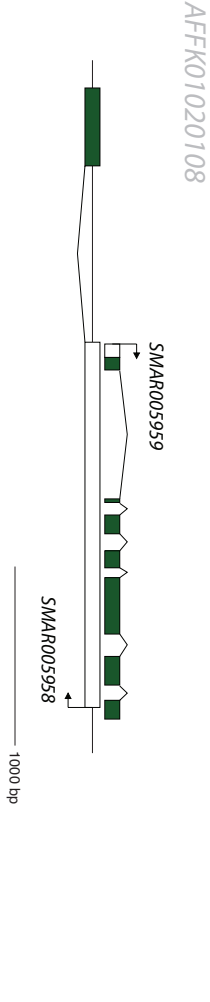
B.



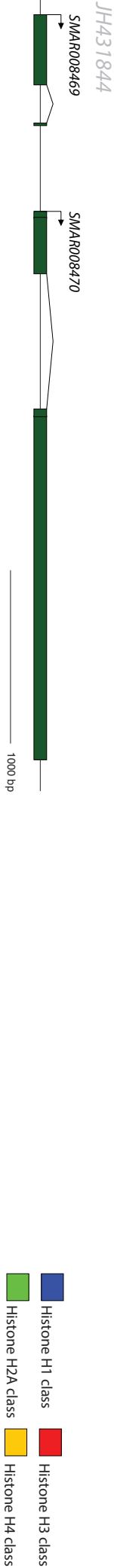
C.

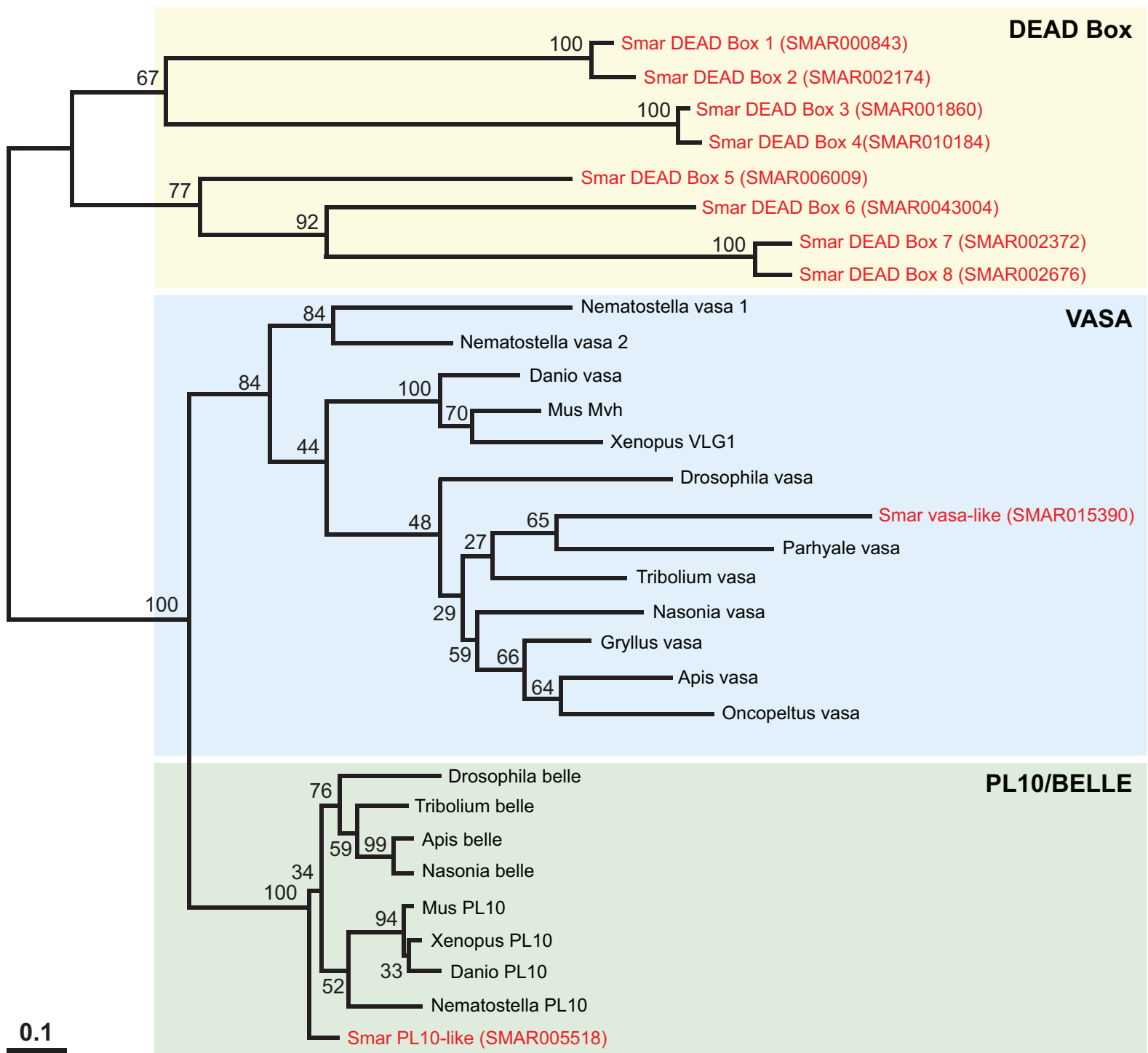


D.



E.





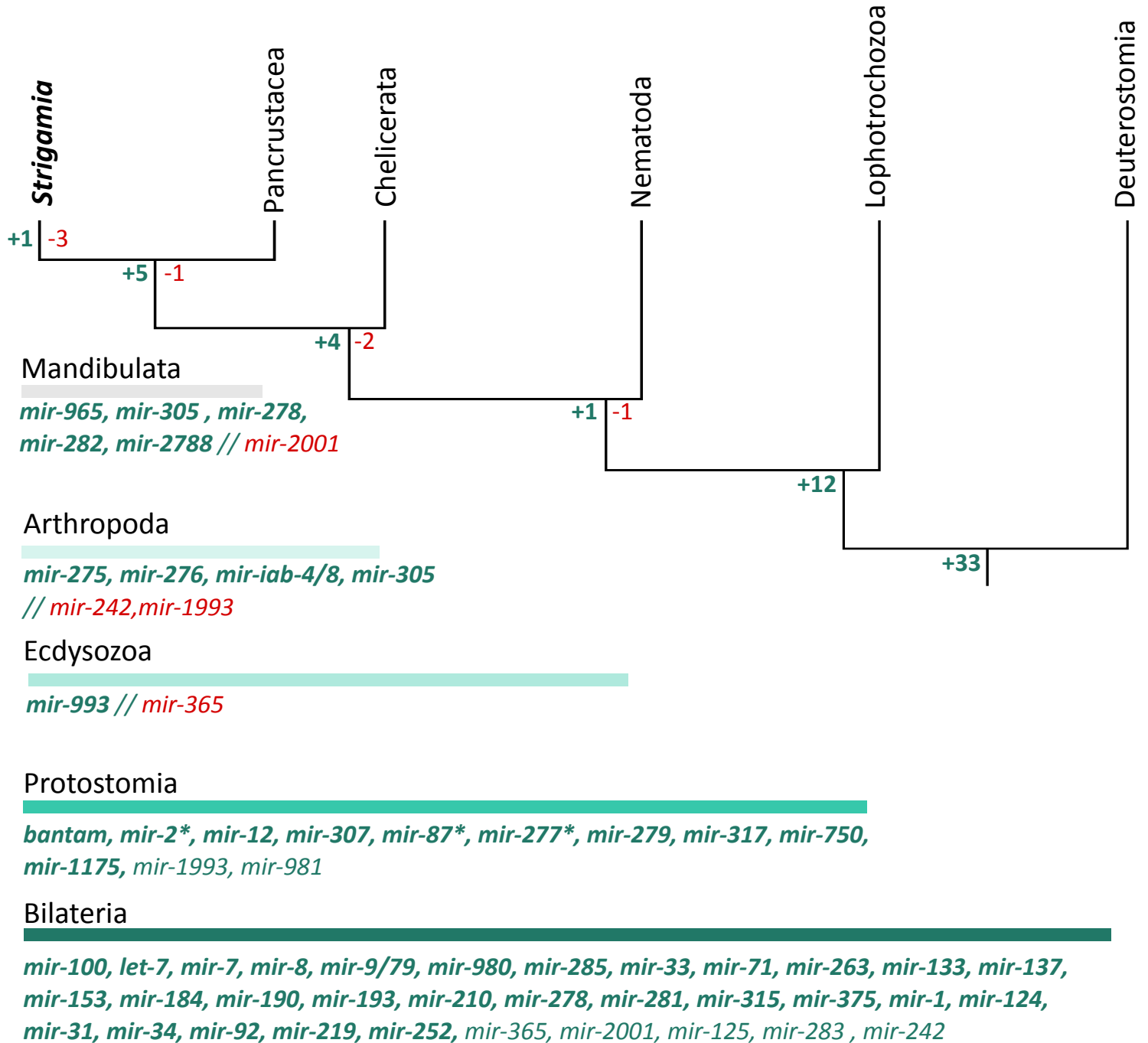


Table S2. Set of species used in the comparative genomics analyses related to the *S. maritima* genome.

Species name	Species Code	Unique longest transcripts	Source	As on
<i>Strigamia maritima</i>	STRMM	14,959	Ensembl - Metazoa 20	09/2013
<i>Pediculus humanus</i>	PEDHC	10,761	Vectorbase	01/2011
<i>Lottia gigantea</i>	LOTGI	23,701	JGI	09/2010
<i>Capitella teleta</i>	283909	31,857	Ensembl - Metazoa 20	09/2013
<i>Nematostella vectensis</i>	NEMVE	24,424	Quest for Orthologs - 2011.04	10/2011
<i>Caenorhabditis elegans</i>	CAEEL	20,333	WormBase	05/2012
<i>Helobdella robusta</i>	HELRO	23,327	JGI	04/2012
<i>Daphnia pulex</i>	DAPPU	30,335	JGI	12/2011
<i>Ixodes scapularis</i>	IXOSC	20,473	Quest for Orthologs - 2011.04	12/2012
<i>Acyrtosiphon pisum</i>	ACYPI	27,584	Aphid	11/2011
<i>Tribolium castaneum</i>	TRICA	16,573	BeetleBASE - HGSC	12/2011
<i>Bombyx mori</i>	BOMMO	14,593	SilkDB	12/2011
<i>Anopheles gambiae</i>	ANOGA	12,580	VectorBASE	05/2012
<i>Drosophila melanogaster</i>	DROME	13,755	Quest for Orthologs - 2012.05	07/2012
<i>Nasonia vitripennis</i>	NASVI	15,073	BCM	11/2011
<i>Strongylocentrotus purpuratus</i>	STRPU	28,760	HGSC	12/2011
<i>Branchiostoma floridae</i>	BRAFL	28,394	Quest for Orthologs - 2011.04	12/2012

<i>Homo sapiens</i>	HUMAN	19,997	Quest for Orthologs - 2012.05	07/2012
---------------------	-------	--------	----------------------------------	---------

Table S3. Orthologues detected between a given species and *S. maritima*.

number of trees used	<i>S. maritima</i>		Other species			ratios	
	orthologues	uniq	Sp. Code	orthologues	uniq	all	uniq
5726	16944	7050	NEMVE	7770	5660	2.18	1.25
6404	18891	7707	TRICA	8889	6291	2.13	1.23
6116	16729	7195	LOTGI	7895	5924	2.12	1.21
4946	12607	6004	BOMMO	6058	4649	2.08	1.29
5231	13335	6346	IXOSC	6775	5133	1.97	1.24
6359	17196	7590	283909	8858	6680	1.94	1.14
5862	12841	6721	PEDHC	6645	5309	1.93	1.27
4649	11604	5629	HELRO	6088	4809	1.91	1.17
5802	16228	7230	STRPU	9065	6956	1.79	1.04
5555	19058	6679	NASVI	11170	7577	1.71	0.88
5446	12060	6256	ANOGA	7078	5384	1.70	1.16
5478	12482	6272	DROME	7543	5642	1.65	1.11
5848	14966	6918	ACYPI	9199	6798	1.63	1.02
4204	9401	4992	CAEEL	5810	4465	1.62	1.12
6207	16226	7269	HUMAN	10170	7871	1.60	0.92
5918	14604	7007	DAPPU	9511	6509	1.54	1.08
5804	15252	6945	BRAFL	11364	6957	1.34	1.00

Table S4. Orthology ratios for a given species related to *S. maritima*.

number of trees used	<i>S. maritima</i>		Other species			ratios	
	orthologues	uniq	Sp. code	orthologues	uniq	all	uniq
5382	9869	6397	NEMVE	6883	5429	1.43	1.18
4700	7065	5288	BOMMO	5314	4502	1.33	1.17
5637	8039	6142	PEDHC	6149	5252	1.31	1.17
5799	9148	6531	LOTGI	7021	5739	1.30	1.14
4996	7687	5643	IXOSC	5918	4925	1.30	1.15
6019	8979	6700	TRICA	7085	5899	1.27	1.14
6036	9632	6770	283909	7623	6204	1.26	1.09
5217	7545	5682	ANOGA	6104	5203	1.24	1.09
4048	6365	4556	CAEEL	5178	4231	1.23	1.08
4438	6917	5033	HELRO	5663	4720	1.22	1.07
5447	8834	6279	STRPU	7233	5998	1.22	1.05
5452	8505	6288	BRAFL	7007	5868	1.21	1.07
5622	8561	6210	DAPPU	7184	5901	1.19	1.05
5237	7538	5720	DROME	6407	5444	1.18	1.05
5529	8208	6122	ACYPI	7219	6092	1.14	1.00
5905	8897	6505	HUMAN	9054	7613	0.98	0.85
5095	7678	5771	NASVI	8076	6818	0.95	0.85

Table S5. Newly added Chelicerata species used to increase the taxon sampling for the species phylogeny.

Species name	Strategy used	Identified proteins	Source	As on
<i>Centruroides sculpturatus</i>	Exonerate	756	BCM - HGSC	07/2013
<i>Latrodectus hesperus</i>	Exonerate	512	BCM - HGSC	02/2013
<i>Parasteatoda tepidariorum</i>	Exonerate	1,058	BCM - HGSC	01/2013
<i>Metaseiulus occidentalis</i>	BBH - Blast	699	BCM - HGSC	07/2013
<i>Tetranychus urticae</i>	BBH - Blast	659	BEG - UGent	11/2012

Table S6. Results after applying the different statistical tests implemented in CONSEL for the alternative placement of *S. maritima* relative to Pancrustacea and Chelicerata groups of species (as shown in Fig. S4) in the context of the 18 species used for the phylogenomics analyses.

rank	item	obs	au	np	bp	pp	kh	sh	wkh	wsh
1st	(1)	-479.7	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2nd	(3)	-479.7	1e-05	1e-06	0	5e-209	0	0	0	0
3rd	(2)	1089.7	9e-09	2e-07	0	0	0	0	0	0

Table S7. Results after applying the different statistical tests implemented in CONSEL for the alternative placement of *S. maritima* relative to the two arthropod groups, Pancrustacea and Chelicerata (as shown in Fig. S4), with the inclusion of extra chelicerates.

rank	item	obs	au	np	bp	pp	kh	sh	wkh	wsh
1st	(1)	-68.5	0.786	0.718	0.723	1.000	0.739	0.872	0.739	0.871
2nd	(3)	68.5	0.316	0.245	0.237	2e-30	0.261	0.390	0.261	0.394
3rd	(2)	153.1	0.072	0.037	0.040	3e-67	0.069	0.118	0.069	0.119

Table S8. Enriched functional GO Terms for the 10 largest clusters of duplicated *S. maritima* protein-coding genes as compared with the whole genome.

Number of protein in cluster	Ontology	Go Term	Go Term Name
50	Cellular component	GO:0016020	membrane
50	Cellular component	GO:0030288	outer membrane-bounded periplasmic space
50	Cellular component	GO:0030312	external encapsulating structure
50	Molecular function	GO:0004871	signal transducer activity
50	Molecular function	GO:0005215	transporter activity
50	Molecular function	GO:0005234	extracellular-glutamate-gated ion channel activity
50	Molecular function	GO:0022857	transmembrane transporter activity
52	Molecular function	GO:0005515	protein binding
52	Molecular function	GO:0008270	zinc ion binding
52	Molecular function	GO:0043167	ion binding
52	Biological process	GO:0006259	DNA metabolic process
52	Biological process	GO:0006278	RNA-dependent DNA replication
52	Biological process	GO:0009058	biosynthetic process
52	Biological process	GO:0034641	cellular nitrogen compound metabolic process
52	Molecular function	GO:0003723	RNA binding
52	Molecular function	GO:0003964	RNA-directed DNA polymerase activity
52	Molecular function	GO:0016779	nucleotidyltransferase activity
54	Molecular function	GO:0005515	protein binding
63	Cellular component	GO:0016021	integral to membrane
63	Biological process	GO:0050877	neurological system process
63	Biological process	GO:0050912	detection of chemical stimulus involved in sensory perception of taste
63	Molecular function	GO:0008527	taste receptor activity
76	Cellular component	GO:0005634	nucleus
76	Cellular component	GO:0043226	organelle
76	Molecular function	GO:0003677	DNA binding

79	Cellular component	GO:0005634	nucleus
79	Cellular component	GO:0043226	organelle
79	Biological process	GO:0006259	DNA metabolic process
79	Biological process	GO:0015074	DNA integration
79	Molecular function	GO:0003676	nucleic acid binding
79	Molecular function	GO:0003677	DNA binding
79	Molecular function	GO:0008270	zinc ion binding
79	Molecular function	GO:0043167	ion binding
98	Cellular component	GO:0016020	membrane
98	Biological process	GO:0006810	transport
98	Biological process	GO:0006814	sodium ion transport
98	Molecular function	GO:0005272	sodium channel activity
98	Molecular function	GO:0022857	transmembrane transporter activity
201	Biological process	GO:0006259	DNA metabolic process
201	Biological process	GO:0015074	DNA integration
201	Biological process	GO:0034641	cellular nitrogen compound metabolic process
201	Molecular function	GO:0003676	nucleic acid binding
292	Cellular component	GO:0005622	intracellular
292	Cellular component	GO:0005623	cell
292	Cellular component	GO:0005634	nucleus
292	Cellular component	GO:0043226	organelle
292	Biological process	GO:0006259	DNA metabolic process
292	Biological process	GO:0015074	DNA integration
292	Molecular function	GO:0003676	nucleic acid binding
292	Molecular function	GO:0003677	DNA binding
292	Molecular function	GO:0008270	zinc ion binding
292	Molecular function	GO:0043167	ion binding

Table S9. Statistics regarding the duplications of centipede genes relative to seven specific ages detected using all available trees on the phylome.

Age	Events	Trees with events (trees: 11,112)	Ratio
1: <i>S. maritima</i>	10767	4097	0.9690
2: Arthropoda I	152	127	0.0137
3: Arthropoda II	1214	871	0.1093
4: Ecdysozoa	125	111	0.0112
5: Protostomia	486	391	0.0437
6: Bilateria	2333	1517	0.2100
7: Eumetazoa	6065	2645	0.5458

Table S10. Enriched functional GO terms for proteins duplicated at the different relative ages shown in Table S9.

Age	Ontology	Go Term	Go Name
1	Biological process	GO:0006259	DNA metabolic process
1	Biological process	GO:0006278	RNA-dependent DNA replication
1	Biological process	GO:0006313	transposition
1	Biological process	GO:0006810	transport
1	Biological process	GO:0006814	sodium ion transport
1	Biological process	GO:0009253	peptidoglycan catabolic process
1	Biological process	GO:0015074	DNA integration
1	Biological process	GO:0032196	transposition
1	Biological process	GO:0034641	cellular nitrogen compound metabolic process
1	Biological process	GO:0050877	neurological system process
1	Molecular function	GO:0003676	nucleic acid binding
1	Molecular function	GO:0003964	RNA-directed DNA polymerase activity
1	Molecular function	GO:0004523	ribonuclease H activity
1	Molecular function	GO:0004803	transposase activity
1	Molecular function	GO:0005234	extracellular-glutamate-gated ion channel activity
1	Molecular function	GO:0005272	sodium channel activity
1	Molecular function	GO:0008745	N-acetylmuramoyl-L-alanine amidase activity
1	Molecular function	GO:0016705	oxidoreductase activity
1	Molecular function	GO:0022857	transmembrane transporter activity
1	Cellular Component	GO:0016020	membrane
3	Biological process	GO:0006030	chitin metabolic process
3	Biological process	GO:0006066	alcohol metabolic process
3	Biological process	GO:0006259	DNA metabolic process
3	Biological process	GO:0006508	proteolysis
3	Biological process	GO:0006810	transport
3	Biological process	GO:0006814	sodium ion transport
3	Biological process	GO:0015074	DNA integration

3	Biological process	GO:0055085	transmembrane transport
3	Molecular function	GO:0003676	nucleic acid binding
3	Molecular function	GO:0004252	serine-type endopeptidase activity
3	Molecular function	GO:0004601	peroxidase activity
3	Molecular function	GO:0005215	transporter activity
3	Molecular function	GO:0005234	extracellular-glutamate-gated ion channel activity
3	Molecular function	GO:0008061	chitin binding
3	Molecular function	GO:0008233	peptidase activity
3	Molecular function	GO:0008812	choline dehydrogenase activity
3	Molecular function	GO:0020037	heme binding
3	Molecular function	GO:0022857	transmembrane transporter activity
3	Cellular Component	GO:0005921	gap junction
3	Cellular Component	GO:0016020	membrane
4	Biological process	GO:0006208	pyrimidine nucleobase catabolic process
4	Biological process	GO:0006810	transport
4	Biological process	GO:0006811	ion transport
4	Molecular function	GO:0005230	extracellular ligand-gated ion channel activity
4	Molecular function	GO:0022857	transmembrane transporter activity
4	Cellular Component	GO:0016020	membrane
4	Cellular Component	GO:0045211	postsynaptic membrane
5	Biological process	GO:0005975	carbohydrate metabolic process
5	Biological process	GO:0006508	proteolysis
5	Biological process	GO:0006810	transport
5	Biological process	GO:0006835	dicarboxylic acid transport
5	Biological process	GO:0006836	neurotransmitter transport
5	Biological process	GO:0007166	cell surface receptor signaling pathway
5	Biological process	GO:0007186	G-protein coupled receptor signaling pathway
5	Biological process	GO:0009253	peptidoglycan catabolic process
5	Molecular function	GO:0004930	G-protein coupled receptor activity
5	Molecular function	GO:0005328	neurotransmitter:sodium symporter activity

5	Molecular function	GO:0008233	peptidase activity
5	Molecular function	GO:0008745	N-acetylmuramoyl-L-alanine amidase activity
5	Molecular function	GO:0016810	hydrolase activity
5	Molecular function	GO:0022857	transmembrane transporter activity
5	Cellular Component	GO:0016020	membrane
5	Cellular Component	GO:0016021	integral to membrane
6	Biological process	GO:0006508	proteolysis
6	Biological process	GO:0006810	transport
6	Biological process	GO:0006811	ion transport
6	Biological process	GO:0006836	neurotransmitter transport
6	Biological process	GO:0007186	G-protein coupled receptor signaling pathway
6	Biological process	GO:0015074	DNA integration
6	Biological process	GO:0055085	transmembrane transport
6	Molecular function	GO:0004252	serine-type endopeptidase activity
6	Molecular function	GO:0004871	signal transducer activity
6	Molecular function	GO:0004872	receptor activity
6	Molecular function	GO:0005234	extracellular-glutamate-gated ion channel activity
6	Molecular function	GO:0005328	neurotransmitter:sodium symporter activity
6	Molecular function	GO:0008233	peptidase activity
6	Molecular function	GO:0022857	transmembrane transporter activity
6	Molecular function	GO:0050254	rhodopsin kinase activity
6	Cellular Component	GO:0005886	plasma membrane
6	Cellular Component	GO:0016020	membrane
6	Cellular Component	GO:0016021	integral to membrane
6	Cellular Component	GO:0030054	cell junction
6	Cellular Component	GO:0030288	outer membrane-bounded periplasmic space
6	Cellular Component	GO:0030312	external encapsulating structure
6	Cellular Component	GO:0045211	postsynaptic membrane
7	Biological process	GO:0006184	GTP catabolic process
7	Biological process	GO:0006468	protein phosphorylation

7	Biological process	GO:0006508	proteolysis
7	Biological process	GO:0006754	ATP biosynthetic process
7	Biological process	GO:0006810	transport
7	Biological process	GO:0006812	cation transport
7	Biological process	GO:0006836	neurotransmitter transport
7	Biological process	GO:0006913	nucleocytoplasmic transport
7	Biological process	GO:0007017	microtubule-based process
7	Biological process	GO:0007018	microtubule-based movement
7	Biological process	GO:0007156	homophilic cell adhesion
7	Biological process	GO:0007165	signal transduction
7	Biological process	GO:0007223	Wnt receptor signaling pathway
7	Biological process	GO:0007264	small GTPase mediated signal transduction
7	Biological process	GO:0008152	metabolic process
7	Biological process	GO:0009056	catabolic process
7	Biological process	GO:0015031	protein transport
7	Biological process	GO:0016055	Wnt receptor signaling pathway
7	Biological process	GO:0043401	steroid hormone mediated signaling pathway
7	Biological process	GO:0043687	post-translational protein modification
7	Biological process	GO:0044281	small molecule metabolic process
7	Biological process	GO:0051246	regulation of protein metabolic process
7	Biological process	GO:0051276	chromosome organization
7	Biological process	GO:0051603	proteolysis involved in cellular protein catabolic process
7	Biological process	GO:0055085	transmembrane transport
7	Biological process	GO:0055114	oxidation-reduction process
7	Biological process	GO:0065003	macromolecular complex assembly
7	Molecular function	GO:0003707	steroid hormone receptor activity
7	Molecular function	GO:0003777	microtubule motor activity
7	Molecular function	GO:0003924	GTPase activity
7	Molecular function	GO:0003995	acyl-CoA dehydrogenase activity
7	Molecular function	GO:0004222	metalloendopeptidase activity

7	Molecular function	GO:0004298	threonine-type endopeptidase activity
7	Molecular function	GO:0004386	helicase activity
7	Molecular function	GO:0004674	protein serine/threonine kinase activity
7	Molecular function	GO:0004702	receptor signaling protein serine/threonine kinase activity
7	Molecular function	GO:0004767	sphingomyelin phosphodiesterase activity
7	Molecular function	GO:0004871	signal transducer activity
7	Molecular function	GO:0005328	neurotransmitter:sodium symporter activity
7	Molecular function	GO:0005509	calcium ion binding
7	Molecular function	GO:0005524	ATP binding
7	Molecular function	GO:0005525	GTP binding
7	Molecular function	GO:0008026	ATP-dependent helicase activity
7	Molecular function	GO:0008233	peptidase activity
7	Molecular function	GO:0008568	microtubule-severing ATPase activity
7	Molecular function	GO:0016301	kinase activity
7	Molecular function	GO:0016491	oxidoreductase activity
7	Molecular function	GO:0016772	transferase activity
7	Molecular function	GO:0016887	ATPase activity
7	Molecular function	GO:0016905	myosin heavy chain kinase activity
7	Molecular function	GO:0019787	small conjugating protein ligase activity
7	Molecular function	GO:0019829	cation-transporting ATPase activity
7	Molecular function	GO:0019899	enzyme binding
7	Molecular function	GO:0042624	ATPase activity
7	Molecular function	GO:0046872	metal ion binding
7	Molecular function	GO:0046982	protein heterodimerization activity
7	Cellular Component	GO:0000786	nucleosome
7	Cellular Component	GO:0005839	proteasome core complex
7	Cellular Component	GO:0005856	cytoskeleton
7	Cellular Component	GO:0005875	microtubule associated complex
7	Cellular Component	GO:0016021	integral to membrane

Table S11. Overview of *Strigamia maritima* mitochondrial genome.

Gene	Strand	Start position	End position	Length (bp)	Start Codon	Stop Codon	Intergenic nucleotides
cox1	+	1	1557	1557	ATT	TAA	
cox2	+	1532	2215	684	ATG	TAG	-25
cox3	+	2221	3063	843	ATG	TAA	6
nad6	+	3060	3524	465	ATA	TAG	-3
nad2	+	3525	4487	963	ATT	TAA	3
trnF	-	4527	4606	79			40
nad5	-	4606	6306	1701	ATG	TAG	0
trnH	-	6287	6347	60			-9
nad4	-	6348	7664	1317	ATG	TAA	1
nad4l	-	7658	7921	264	ATT	TAA	-6
trnP	-	7909	7972	63			-12
NC1	+	7923	8414	491			0
trnD	+	8415	8489	74			0
atp8	+	8461	8622	162	ATA	TAA	-28
atp6	+	8616	9281	666	ATG	TAA	-6
trnR	+	9331	9375	44			50
trnE	+	9409	9454	45			34
trnT	+	9455	9506	51			1
cob	+	9508	10641	1134	ATC	TAG	2
trnM	+	10652	10710	58			2
trnI	+	10706	10759	53			11
trnY	-	10765	10852	87			6
trnL1	+	10788	10848	60			-64
trnV	-	10871	10939	68			23
NC2	+	10940	11331	391			0
trnS2	+	11332	11387	55			0
nad3	+	11382	11732	351	ATT	TAA	-5
trnN	+	11752	11799	47			20
trnK	-	11852	11895	43			53
nad1	-	11942	12862	921	ATT	TAG	47
rrnL	-	12888	14258	1370			26
trnL2	+	14078	14148	70			-180
rrnS	-	14111	14850	739			-37

Species	Proteome build	File
Arabidopsis thaliana	TAIR10.62	ftp://ftp.ensemblgenomes.org/pub/plants/release-9/fasta/arabidopsis_thaliana/pep/Arabidopsis_thaliana.TAIR10.10.pep.all.fa.gz
Naegleria gruberi	Naegr1_best	ftp://ftp.jgi-psf.org/pub/JGI_data/Naegleria_gruberi/Naegr1_best_proteins.fasta.gz
Thalassiosira pseudonana	Thaps3_chromosomes_Filtered2	ftp://ftp.jgi-psf.org/pub/JGI_data/Thalassiosira_pseudonana/v3.0/Thaps3_chromosomes_geneModels_FilteredModels2_aa.fasta.gz
Chlamydomonas reinhardtii	Chlre4_best	ftp://ftp.jgi-psf.org/pub/JGI_data/Chlamydomonas_reinhardtii/v4.0/annotation/Chlre4_best_proteins.fasta.gz
Phycomyces blakesleeanus	Phyb11_best	ftp://ftp.jgi-psf.org/pub/JGI_data/Phycomyces_blakesleeanus/annotation/v1.0/Phyb11_best_proteins.fasta.gz
Mucor circinelloides	Muccil_best	ftp://ftp.jgi-psf.org/pub/JGI_data/Mucor_circinelloides/v1.0/annotation/Muccil_best_proteins.fasta.gz
Schizosaccharomyces octosporus	Schizosaccharomyces_octosporus_protein	schizosaccharomyces_octosporus_6_proteins.fasta
Schizosaccharomyces cryophilus	Schizosaccharomyces_cryophilus_protein	schizosaccharomyces_cryophilus_4_proteins.fasta
Schizosaccharomyces pombe	Schizosaccharomyces_pombe_protein	schizosaccharomyces_pombe_972h-2_proteins.fasta
Schizosaccharomyces japonicus	Schizosaccharomyces_japonicus_protein	schizosaccharomyces_japonicus_yfs275_5_proteins.fasta
Saccharomyces	EF2.62	ftp://ftp.ensembl

cerevisiae		1.org/pub/release-62/fasta/saccharomyces_cerevisiae/pep/Saccharomyces_cerevisiae.EF2.62.pep.all.fasta.gz
Tremella mesenterica	Tremel1_best	ftp://ftp.jgi-psf.org/pub/JGI_data/Tremella_mesenterica/v1.0/annotation/Tremel1_best_proteins.fasta.gz
Batrachochytrium dendrobatidis	Batde5_best	ftp://ftp.jgi-psf.org/pub/JGI_data/Batrachochytrium_dendrobatidis/annotation/v1.0/Batde5_best_proteins.fasta.gz
Acropora digitifera	nomask_110621.TEremove	nomask_110621.prot.t1.TEremove.fasta
Nematostella vectensis	Nemvel1FilteredModels1	ftp://ftp.jgi-psf.org/pub/JGI_data/Nematostella_vectensis/v1.0/annotation/proteins.Nemvel1FilteredModels1.fasta.gz
Hydra magnipapillata	hydra_Hma2	ftp://ftp.jgi-psf.org/pub/JGI_data/Hydra_magnipapillata/annotation/hydra_Hma2.pep.fasta.gz
Capitella teleta	FilteredModelsv1.0	ftp://ftp.jgi-psf.org/pub/JGI_data/Capitella/v1.0/FilteredModelsv1.0.aa.fasta.gz
Helobdella robusta	Helro1_FilteredModels3	ftp://ftp.jgi-psf.org/pub/JGI_data/Helobdella_robusta/v1.0/proteins.Helro1_FilteredModels3.fasta.gz
Lottia gigantea	Lotgil_GeneModels_FilteredModels1	ftp://ftp.jgi-psf.org/pub/JGI_data/Lottia_gigantea/v1.0/Lotgil_GeneModels_FilteredModels1_aa.f

		asta.gz
Schistosoma mansoni	sma_v3.1	ftp://ftp.ensemblgenomes.org/pub/metazoa/release-9/fasta/schistosoma_mansoni/pep/Schistosoma_mansoni.sma_v3.1.1a.pep.all.fa.gz
Pristionchus pacificus	pp1.62	ftp://ftp.ensemblgenomes.org/pub/metazoa/release-9/fasta/pristionchus_pacificus/pep/Pristionchus_pacificus.pp1.1a.pep.all.fa.gz
Caenorhabditis elegans	Caenorhabditis_elegans.WS220.220	ftp://ftp.ensemblgenomes.org/pub/metazoa/release-9/fasta/caenorhabditis_elegans/pep/Caenorhabditis_elegans.WS220.220.pep.all.fa.gz
Drosophila grimshawi	dgri_r1.3_FB2008_07	ftp://ftp.ensemblgenomes.org/pub/metazoa/release-11/fasta/drosophila_grimshawi/pep/Drosophila_grimshawi.dgri_r1.3_FB2008_07.pep.all.fa.gz
Drosophila melanogaster	BDGP5.25.62	ftp://ftp.ensembl.org/pub/release-62/fasta/drosophila_melanogaster/pep/Drosophila_melanogaster.BDGP5.25.62.pep.all.fa.gz
Aedes aegypti	2009-06-VectorBase	ftp://ftp.ensemblgenomes.org/pub/metazoa/release-11/fasta/aedes_aegypti/pep/Aedes_aegypti.AegL1.pep.all.fa.gz
Culex quinquefasciatus	2008-05-VectorBase	ftp://ftp.ensemblgenomes.org/pub/metazoa/release

		- 11/fasta/culex_q uinquefasciatus/ pep/Culex_quinqu efasciatus.CpipJ 1.pep.all.fa.gz
Anopheles gambiae	AgamP3.6	ftp://ftp.ensem blgenomes.org/pub /metazoa/release - 11/fasta/anophel es_gambiae/pep/A nopheles_gambiae .AgamP3.pep.all. fa.gz
Bombyx mori	silkworm glean pep v2.0	silkpep.fa
Apis mellifera	2005-BeeBase	ftp://ftp.ensem blgenomes.org/pub /metazoa/release - 11/fasta/apis_me llifera/pep/Apis _mellifera.Amel_ 2.0.pep.all.fa.g z
Tribolium castaneum	3.0_Tribolium_Official_Gene_ sequences	ftp://bioinforma tics.ksu.edu/pub /BeetleBase/3.0/ Sequences/Tribol ium_Official_Gen e_Sequences/pept ide.fa
Pediculus humanus	PhumU1.2	ftp://ftp.ensem blgenomes.org/pub /metazoa/release - 11/fasta/pedicul us_humanus/pep/P ediculus_humanus .PhumU1.pep.all. fa.gz
Daphnia pulex	FilteredModelsv1.0	ftp://ftp.jgi- psf.org/pub/JGI_ data/Daphnia_pul ex/v1.0/Filtered Modelsv1.0.aa.fa sta.gz
Strigamia maritima	Strigamia_6.1	http://www.hgsc. bcm.tmc.edu/coll aborations/insec ts/strigamia/Mak er_results/centi pede_maker_sept_ 2011/all_maker_p rotains.fa
Ixodes scapularis	IscaW1.62	ftp://ftp.ensem blgenomes.org/pub /metazoa/release -

		9/fasta/ixodes_scapularis/pep/Ixodes_scapularis.IscaW1.1a.pep.all.fa.gz
Tetranychus urticae	PEP_20120618	https://bioinformatics.psb.ugent.be/gdb/tetranychus/Tetur_PEP_20120618.tfa.gz
Strongylocentrotus purpuratus	SpBase_SPU	SpBase_SPU_peptide.fasta.gz
Saccoglossus kowalevskii	SkowalevskiiJGIv3.0	ftp://ftp.jgi-psf.org/pub/JGI_data/Saccoglossus_kowalevskii/v3/annotation/SkowalevskiiJGIv3.0.longestTrs.pep.fasta.gz
Gasterosteus aculeatus	Ensembl_64	ftp://ftp.ensembl.org/pub/release-64/fasta/gasterosteus_aculeatus/pep/Gasterosteus_aculeatus.BROADS1.64.pep.all.fasta.gz
Oryzias latipes	Ensembl_64	ftp://ftp.ensembl.org/pub/release-64/fasta/oryzias_latipes/pep/Oryzias_latipes.MEDAKA1.64.pep.all.fasta.gz
Danio rerio	Zv9.62	ftp://ftp.ensembl.org/pub/release-62/fasta/danio_rerio/pep/Danio_rerio.Zv9.62.pep.all.fasta.gz
Anolis carolinensis	Ensembl_64	ftp://ftp.ensembl.org/pub/release-64/fasta/anolis_carolinensis/pep/Anolis_carolinensis.AnoCar2.0.64.pep.all.fasta.gz
Gallus gallus	WASHUC2.62	ftp://ftp.ensembl.org/pub/release-62/fasta/gallus_gallus/pep/Gallus_gallus.WASHUC2.62.pep.all.fasta.g

		z
Homo sapiens	Homo_sapiens.GRCh37.61	ftp://ftp.ensembl.org/pub/current/fasta/homo_sapiens/pep/Homo_sapiens.GRCh37.61.pep.all.fa.gz
Xenopus (Silurana) tropicalis	Ensembl_61	ftp://ftp.ensembl.org/pub/current/fasta/xenopus_tropicalis/pep/Xenopus_tropicalis.JGI4.1.61.pep.all.fa.gz
Ciona intestinalis	Ciona_intestinalis.JGI2.52	ftp://ftp.ensembl.org/pub/release-52/fasta/ciona_intestinalis/pep/Ciona_intestinalis.JGI2.52.pep.all.fa.gz
Branchiostoma floridae	Brafl1	ftp://ftp.jgi-psf.org/pub/JGI_data/Branchiostoma_floridae/v1.0/proteins.Brafl1.fasta.gz
Trichoplax adhaerens	Triad1_best_proteins	ftp://ftp.jgi-psf.org/pub/JGI_data/Trichoplax_adhaerens_Grell-BS-1999/annotation/v1.0/Triad1_best_proteins.fasta.gz
Mnemiopsis leidyi	ML2.2	ML2.2.aa
Amphimedon queenslandica	Aqu1	ftp://ftp.jgi-psf.org/pub/JGI_data/Amphimedon_queenslandica/annotation/Aqu1.pfp.fa.gz
Monosiga brevicollis	Monbr1_best_proteins	ftp://ftp.jgi-psf.org/pub/JGI_data/Monosiga_brevicollis/annotation/v1.0/Monbr1_best_proteins.fasta.gz
Capsaspora owczarzaki	capsaspora_owczarzaki_atcc_30864_2_proteins	capsaspora_owczarzaki_atcc_30864_2_proteins.fasta
Dictyostelium discoideum	created_03-15-2011	dicty_primary_protein_20110315.gz

File S1.

One2One_GOTerms_GenomeIDs for Orthology-based functional annotation.

<https://doi.org/10.1371/journal.pbio.1002005.s065>

(XLSX)

File S2.

Strigamia_pals for [Figure 3](#).

<https://doi.org/10.1371/journal.pbio.1002005.s066>

(XLSX)

File S3.

Gustatory receptor sequences.

<https://doi.org/10.1371/journal.pbio.1002005.s067>

(XLSX)

File S4.

Raw data for [Figure 2](#), [Figure 9](#), [Figure S1](#), and [Figure S5](#).

<https://doi.org/10.1371/journal.pbio.1002005.s068>

(XLSX)

File S5.

Raw data for [Figure S28](#).

<https://doi.org/10.1371/journal.pbio.1002005.s069>

(XLSX)

File S6.

Raw data for [Figure S29](#).

<https://doi.org/10.1371/journal.pbio.1002005.s070>

(XLSX)

File S7.

Raw data for [Figure S30](#).

<https://doi.org/10.1371/journal.pbio.1002005.s071>

(XLSX)

Table S15.

Names and identification numbers of all *S. maritima* homeobox genes along with their orthologues from the beetle, *T. castaneum*, and amphioxus, *B. floridae*.

<https://doi.org/10.1371/journal.pbio.1002005.s049>

(XLS)

Table S20.

A comparison between the *D. melanogaster* and *S. maritima* biogenic amine receptors. The orthologues are given next to each other. When there is no orthologue, a dash (–) is written instead.

<https://doi.org/10.1371/journal.pbio.1002005.s054>

(XLSX)

Table S30.

Details of the manually annotated genes of *S. maritima*.

<https://doi.org/10.1371/journal.pbio.1002005.s064>

(XLSX)