The Genome of *Armadillidium vulgare* (Crustacea, Isopoda) Provides Insights into Sex Chromosome Evolution in the Context of Cytoplasmic Sex Determination

Mohamed Amine Chebbi,[1] Thomas Becking,[1] Bouziane Moumen,[1] Isabelle Giraud,[1] Clément Gilbert,[†,1] Jean Peccoud,[1] and Richard Cordaux*[,1]

[1]Laboratoire Ecologie et Biologie des Interactions, Equipe Ecologie Evolution Symbiose, Université de Poitiers, UMR CNRS 7267, Poitiers, France

[†]Present address: Laboratoire Evolution, Génomes, Comportement, Ecologie, CNRS Université Paris-Sud UMR 9191, IRD UMR 247, Gif sur Yvette, France

***Corresponding author:** E-mail: richard.cordaux@univ-poitiers.fr.

Associate editor: John Parsch

Abstract

The terrestrial isopod *Armadillidium vulgare* is an original model to study the evolution of sex determination and symbiosis in animals. Its sex can be determined by ZW sex chromosomes, or by feminizing *Wolbachia* bacterial endosymbionts. Here, we report the sequence and analysis of the ZW female genome of *A. vulgare*. A distinguishing feature of the 1.72 gigabase assembly is the abundance of repeats (68% of the genome). We show that the Z and W sex chromosomes are essentially undifferentiated at the molecular level and the W-specific region is extremely small (at most several hundreds of kilobases). Our results suggest that recombination suppression has not spread very far from the sex-determining locus, if at all. This is consistent with *A. vulgare* possessing evolutionarily young sex chromosomes. We characterized multiple *Wolbachia* nuclear inserts in the *A. vulgare* genome, none of which is associated with the W-specific region. We also identified several candidate genes that may be involved in the sex determination or sexual differentiation pathways. The *A. vulgare* genome serves as a resource for studying the biology and evolution of crustaceans, one of the most speciose and emblematic metazoan groups.

Key words: **female heterogamety, recombination, homomorphy, hybrid de novo genome assembly, *Wolbachia*, repeats.**

Introduction

Terrestrial isopods (Crustacea, Oniscidea), also known as woodlice or pillbugs, are widely distributed on Earth across a variety of soil habitats, with about 3,600 named species (Schmalfuss 2003). Among crustaceans, terrestrial isopods are notable for having evolved total independence from their original aquatic environment through morphological, physiological, and behavioral adaptations to terrestrial ecosystems (Hornung 2011). They also constitute a prime model to study the evolution of sex determination in animals.

Sex determination is a fundamental biological pathway of animals (Bachtrog et al. 2014; Beukeboom and Perrin 2014). As such, it could be expected to be governed by highly conserved molecular mechanisms. On the contrary, many animal groups exhibit a high variability of sex determination systems, which raises the question of the underlying evolutionary forces driving transitions between these systems. Genetic sex determination is usually under the control of sex chromosomes, which present two major types: male heterogamety (XY males and XX females) and female heterogamety (ZZ males and ZW females). But unlike some groups that exclusively follow male (e.g., mammals) or female (e.g., birds) heterogamety, terrestrial isopods present both heterogametic types scattered across their phylogenetic tree (Juchault and Rigaud 1995; Becking et al. 2017). This pattern implies numerous transitions between heterogametic types in terrestrial isopods at various phylogenetic depths (Juchault and Rigaud 1995; Becking et al. 2017). Sex determination is so labile in terrestrial isopods that different systems sometimes coexist within species, making these crustaceans particularly well suited for understanding the factors driving the evolution of sex determination.

In this context, the best studied species is *Armadillidium vulgare* (Rigaud et al. 1997; Cordaux et al. 2011; Cordaux and Gilbert 2017). In this species, chromosomal sex determination follows ZW female heterogamety (Juchault and Legrand 1972). However, many *A. vulgare* females produce female-biased progenies caused by the presence of intracellular, feminizing *Wolbachia* bacterial endosymbionts (Rigaud et al. 1997; Cordaux et al. 2011; Cordaux and Gilbert 2017). These maternally inherited microorganisms induce feminization of nontransmitting ZZ genetic males into transmitting phenotypic females (Martin et al. 1973; Rigaud et al. 1991; Cordaux et al. 2004). Remarkably, the presence of feminizing *Wolbachia* ultimately drives the loss of the W sex chromosome in infected lines, such that all individuals are ZZ genetic

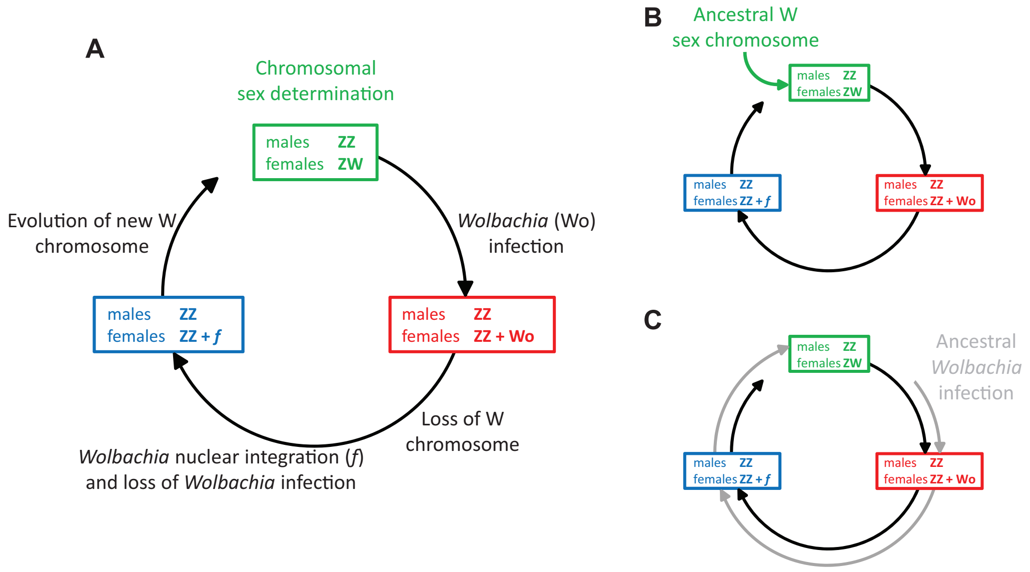

Fig. 1. Hypotheses on the evolution of sex determination in the terrestrial isopod *Armadillidium vulgare*, involving ZW sex chromosomes (green), feminizing *Wolbachia* (red), and the *f* element (blue). (*A*) Circular model proposed by Juchault and Mocquard (1993). Based on this model, there are two possibilities for the origin of the current ZW sex chromosomes (*B* and *C*). (*B*) The first cycle has not been completed yet, so that the current sex chromosomes are the ancestral sex chromosomes (i.e., existing prior to infection by feminizing *Wolbachia* endosymbionts). (*C*) The first cycle (at least) has been completed, meaning that *A. vulgare* has previously experienced at least one prior *Wolbachia* infection (in gray) that led to a turnover of ZW sex chromosomes, so that the current sex chromosomes have a female-specific region derived from *Wolbachia* sequence.

males. In these lines, embryos inheriting *Wolbachia* (∼90% of offspring) develop as females and those lacking *Wolbachia* develop as males. The *A. vulgare*/*Wolbachia* system thus constitutes a canonical example of cytoplasmic sex determination (Rigaud et al. 1997; Cordaux et al. 2011; Cordaux and Gilbert 2017).

Strikingly, in some *A. vulgare* lines lacking both the W sex chromosome and *Wolbachia*, female sex determination is under the control of the *f* element, a nuclear insert derived from the horizontal transfer of a feminizing *Wolbachia* genome into an *A. vulgare* chromosome (Leclercq et al. 2016; Cordaux and Gilbert 2017). Thus, the chromosome that carries the *f* element effectively is a W-like sex chromosome, leading to a turnover of sex chromosomes driven by *Wolbachia* endosymbionts in *A. vulgare* lines with the *f* element (Leclercq et al. 2016; Cordaux and Gilbert 2017). Overall, sex determination factors are highly variable and dynamic in *A. vulgare*. This led Juchault and Mocquard (1993) to suggest the "circular" model of evolution of sex determination mechanisms, involving repeated turnover events in which a system is replaced by another one (Juchault and Mocquard 1993; Rigaud et al. 1997; Cordaux et al. 2011) (fig. 1*A*). More generally, this scenario may also explain the frequent turnovers of sex chromosomes occurring in terrestrial isopods (Juchault and Rigaud 1995; Becking et al. 2017), as *Wolbachia* endosymbionts are widespread in this animal group (Bouchon et al. 1998; Cordaux et al. 2012).

The circular model of evolution of sex chromosomes raises the question of the origin of the current W sex chromosome of *A. vulgare*: Is it an ancestral W sex chromosome possessing a female-specific region unrelated to *Wolbachia* (fig. 1*B*)? Or is it derived from a prior *Wolbachia* infection with a female-specific region analogous to the *f* element (fig. 1*C*)? The available evidence suggests that the W and Z sex chromosomes of *A. vulgare* are genetically very similar. First, they show no apparent heteromorphy in karyotypic studies (Artault 1977). Second, WW female individuals are viable and fertile (Juchault and Legrand 1972), suggesting that the W and Z chromosomes share all genes required for survival and reproduction. Therefore, it is possible that the W and Z chromosomes are poorly differentiated and the female-specific region of the W sex chromosome is small.

Here, we used whole-genome sequencing to gain insight into the origin and evolution of the current W sex chromosome of *A. vulgare*. We report the assembly of the *A. vulgare* genome obtained using a hybrid strategy combining short Illumina and long Pacific Biosciences (PacBio) reads. This assembly was used to characterize female-specific genomic regions of the W chromosome, to test the hypothesis of poor differentiation of the W and Z chromosomes and to further evaluate the contribution of *Wolbachia* endosymbionts to the evolution of sex determination mechanisms in *A. vulgare*. As one of the very few available crustacean genomes, the *A. vulgare* genome opens new avenues for studying biology and evolution of this highly speciose and emblematic group of arthropods.

Results

De Novo Assembly of the *A. vulgare* Genome

We sequenced the female genome of an *A. vulgare* matriline derived from wild animals sampled in Nice, France. Females from this matriline have been consistently producing progenies with balanced sex ratios in our laboratory for ∼45 generations. Sex reversal and crossing experiments demonstrated that sex determination in this matriline follows female heterogamety (Juchault and Legrand 1972). Absence of *Wolbachia* endosymbionts and the *f* element in the individuals selected for sequencing was confirmed by polymerase chain reaction (PCR).

We generated three types of Illumina libraries, which were used to obtain short paired-end reads, mate-pair reads, and long paired-end reads, with ~450-bp, ~5-kb, and ~20-kb insert sizes, respectively (supplementary table S1, Supplementary Material online). Genome size was estimated from the *k*-mer frequency distribution of filtered paired-end reads. This analysis estimated a haploid genome size of 1.52–1.75 Gb (supplementary table S2, Supplementary Material online), consistent with a previous estimate of 1.66–1.84 Gb based on flow cytometry and Feulgen image analysis densitometry (Jeffery and Gregory 2014). An initial assembly solely based on filtered Illumina reads was highly fragmented, comprising 6,484,519 contigs with an N_{50} contig size of 261 bp. Only 26.8% of the contigs formed scaffolds, which suggested that the *A. vulgare* genome harbors a high proportion of repeats that could not be resolved by Illumina reads alone. We therefore generated long PacBio reads (supplementary table S1, Supplementary Material online) and assembled them with Illumina reads in a hybrid approach. The resulting assembly was markedly improved, and now had an N_{50} contig size of 38,042 bp. The improved assembly was then processed through polishing, misassembly correction, several rounds of scaffolding by long paired-end and PacBio reads and a transcriptome assembly, and removal of sequences from bacteria, fungi, and other contaminants (see Materials and Methods section for details).

Our final assembly was composed of 43,541 contigs and scaffolds (hereafter collectively referred to as "contigs") with an N_{50} of 51,088 bp and containing only 0.43% undetermined nucleotides (table 1). The total length of the assembly was 1,725,108,002 bp, in excellent agreement with the predicted genome size. Genome completeness assessment using Benchmarking Universal Single-Copy Orthologs (BUSCO) (Simão et al. 2015; Waterhouse et al. 2018) revealed that 981 of 1,066 (92%) conserved specific arthropod genes were present in the final assembly (table 1). Furthermore, transcriptome assembly alignment on the constructed genome yielded 95.4% of transcripts longer than 1 kb aligned. Altogether, these results indicate that we have obtained a reliable assembly of the ZW female genome from *A. vulgare*.

A total of 19,051 gene models were predicted in the *A. vulgare* genome (supplementary table S3, Supplementary Material online). More than 90% of the predicted genes had over 50% of their exons supported by RNA-Seq data (supplementary fig. S1, Supplementary Material online). Mean gene size (including open reading frames, introns, and untranslated regions) was 8,636 bp and genes exhibited a mean of 6 exons by transcript and a mean intron size of 1,506 bp (supplementary table S3, Supplementary Material online). Of the 19,051 predicted genes, 10,462 (54.9%) had BlastP hits to the UniProt-SwissProt database (release September 2016) and 10,441 (54.8%) had InterProScan hits to Pfam domains (version 30.0) (Finn et al. 2016), with 6,637 (34.8%) having Gene Ontology (GO) terms (Ashburner et al. 2000; Gene Ontology Consortium 2017). The joint functional annotation procedure enabled to annotate 11,937 (62.7%) gene models. The annotated genome sequence of *A. vulgare* is available in DDBJ/ENA/GenBank under accession number SAUD00000000. The version described in this article is version SAUD01000000.

Table 1. Summary Statistics of *Armadillidium vulgare* Genome Assembly.

Assembly Features	Assembly Figures
Assembly statistics	
Number of contigs and scaffolds	43,541
Total size (bp)	1,725,108,002
Longest contig/scaffold (bp)	558,749
Number of contigs and scaffolds >1 kb	43,525
N_{50} contig size (bp)	38,434
N_{50} scaffold size (bp)	51,088
Undetermined nucleotides (%)	0.43
G + C content (%)	28.04
Analysis of BUSCO genes	
Complete genes	937/1,066 (87.9%)
Complete and single-copy genes	879/1,066 (82.5%)
Complete and duplicated genes	58/1,066 (5.4%)
Fragmented genes	44/1,066 (4.1%)
Missing genes	85/1,066 (8.0%)

Extreme Abundance of Repeats in the *A. vulgare* Genome

We found that 68.1% of the *A. vulgare* assembly was composed of repetitive DNA (fig. 2A and supplementary table S4, Supplementary Material online). Specifically, transposable elements (TEs) encompassed 49.9% (or ~861 Mb) of the *A. vulgare* genome. Of the ~2.2 million TE copies, the most abundant classes were long interspersed nuclear elements and DNA elements comprising respectively 24.2% and 16.2% of the genome. Divergence analyses between TE copies and the consensus sequences of their respective families (fig. 2B) suggested that the *A. vulgare* lineage experienced intense transposition activity 20–40 Ma. TE expansion was largely mediated by the CR1 family of long interspersed nuclear element retrotransposons (15.2% of the genome) and to a lower extent by the hAT superfamily of DNA transposons (4.9% of the genome). However, TE activity substantially decreased in the past few million years, and DNA transposons appear to be the major contributors of the current TE activity in the *A. vulgare* genome (fig. 2B). This is consistent with the identification of recent cases of horizontal transfers of *Mariner* DNA transposons involving *A. vulgare* (Dupeyron et al. 2014).

In addition to TEs, simple tandem repeats accounted for 17.4% of the *A. vulgare* genome. Interestingly, the microsatellite motif $(TTAGG)_n$ corresponding to the canonical telomeric microsatellite motif of arthropods (Vítková et al. 2005) comprised 601 kb (0.04%) of the *A. vulgare* genome. This observation suggests that *A. vulgare* chromosomes are likely to end with typical arthropod telomeric structures. Overall,

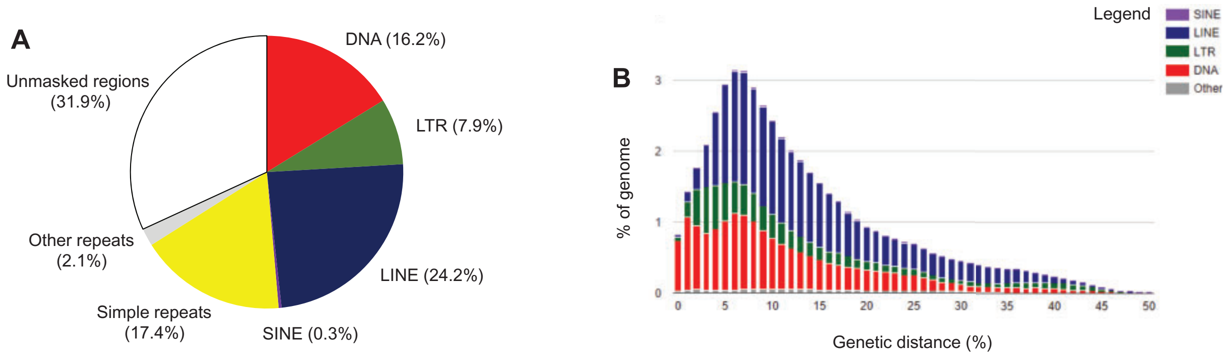

Fig. 2. Repeat content of the *Armadillidium vulgare* genome. (*A*) Repeats comprise 68.1% of the genome, the majority of which are TEs, including DNA transposons (DNA), long terminal repeat retrotransposons (LTR), long interspersed nuclear elements (LINE), and short interspersed nuclear elements (SINE). Unmasked regions correspond to nonrepeated regions of the genome. (*B*) Frequency distribution of TE types according to the divergence of individual copies to their respective family consensus sequences.

these findings reflect a very high abundance of expansive repeats in the *A. vulgare* genome.

Lack of Large-Scale Differentiation of Z and W Sex Chromosomes

To investigate the extent of genomic differentiation between the W and Z sex chromosomes, we Illumina-sequenced pools of ZZ males and ZW females and analyzed reads with the Chromosome Quotient (CQ) method (Hall et al. 2013). The rationale of this analysis is that W-specific contigs of the genome assembly (i.e., contigs with no detectable homolog on the Z chromosome) are expected to be mapped by female reads only (CQ ∼ 0). In addition, Z-specific contigs are expected to be mapped by male reads at twice the sequencing depth of female reads (CQ ∼ 2), whereas autosomal contigs should be mapped at similar sequencing depths by both female and male reads (CQ ∼ 1). Illumina reads generated from male and female pools were mapped onto the *A. vulgare* reference genome. The resulting frequency distribution of CQ scores was unimodal and centered at CQ ∼ 1 (fig. 3A), with no peak at CQ scores of ∼0 and ∼2. This analysis indicated that the *A. vulgare* assembly mostly contains autosomal contigs and apparently very few W- and Z-specific contigs.

To evaluate the robustness of this conclusion, we used the mapping-free Y chromosome Genome Scan (YGS) method (Carvalho and Clark 2013). YGS was initially designed to identify Y-specific sequences by computing the proportion of single-copy *k*-mers unmatched to female reads for each contig in a male genome assembly (Carvalho and Clark 2013). However, YGS can be used to identify W-specific sequences by searching for the proportion of single-copy *k*-mers in contigs of a female genome assembly unmatched to male reads. In this analysis, W-specific contigs are not expected to match to male reads (YGS ∼ 100%), whereas autosomal and Z-linked contigs are expected to match entirely (YGS ∼ 0%). The YGS analysis indicated that very few sequences of the assembly have high YGS values. Despite being based on different methodologies, the YGS and CQ analyses are consistent in suggesting that the *A. vulgare* genome contains very few W-specific contigs (fig. 3B), and that the *A. vulgare* W and Z sex chromosomes are very weakly differentiated.

The Female-Specific Region of the W Sex Chromosome Is Small

To identify candidate female-specific sequences of the W sex chromosome and minimize false positives, we intersected the results of the CQ and YGS analyses. Specifically, we selected sequences with scores of CQ $\leq$ 0.3 and YGS $\geq$ 40%. Although these two thresholds are not very stringent, only 27 contigs emerged as W-specific candidates (fig. 3C). Overall, the candidate contigs represent a total length of 673 kb, which corresponds to only 0.04% of the *A. vulgare* genome assembly (supplementary table S5, Supplementary Material online). To independently evaluate the female specificity of the candidate contigs, we designed PCR assays and tested them using the individual DNA samples of males and females sequenced as part of the CQ and YGS analyses. Reliable PCR assays were successfully designed for 19 of the 27 candidate contigs, most of which (12/19) exhibited female-specific amplification (supplementary table S5, Supplementary Material online). Globally, these results confirmed that the combined CQ and YGS analysis was efficient to identify female-specific contigs in the *A. vulgare* genome. Altogether, our results suggest that the W-specific region of the *A. vulgare* genome is very small, probably at most several hundreds of kb.

The presence of a very small W-specific region suggests that recombination suppression, which typically allows homologous Z- and W-linked sequences to molecularly diverge (Charlesworth et al. 2005), has not spread very far from the sex-determining locus, if at all, in *A. vulgare*. To investigate this hypothesis, we analyzed the 12 aforementioned loci in a panel of males and females from the sequenced matriline (originating from France), as well as two additional *A. vulgare* lines originally sampled in Greece and Denmark (table 2). Female-specific amplification or strong linkage to the female sex (defined as amplification in males at significantly lower frequency than in females) was confirmed for all 12 loci in the French samples. In the additional lines, four loci showed female-specific amplification or strong linkage to the female sex: two loci in both lines (contigs 60 and 19252) and two loci

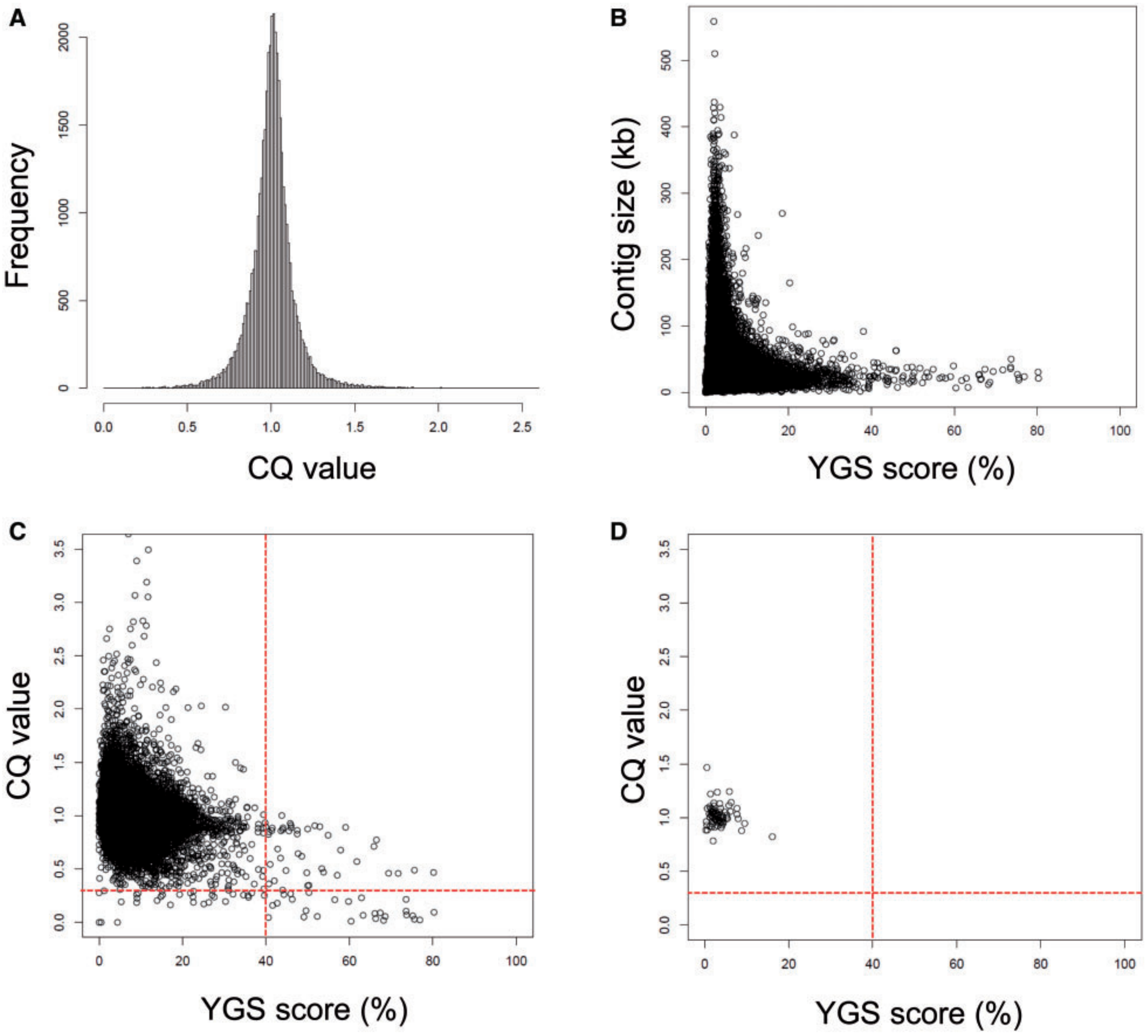

Fig. 3. Identification of sex-specific contigs in the *Armadillidium vulgare* genome assembly. (*A*) Frequency distribution of CQ values calculated for each contig and scaffold of the assembly. The unimodal distribution is centered at a CQ value of ~1. (*B*) Results of the YGS analysis. Most contigs and scaffolds of the assembly have low YGS scores. (*C*) Comparison of CQ and YGS scores. Only 27 contigs have both CQ ≤ 0.3 and YGS ≥ 40% (thresholds represented by dashed red lines), as expected of female-specific sequences. (*D*) Comparison of CQ and YGS scores for the 72 contigs containing *Wolbachia* nuclear inserts. None of them exhibit signature of female specificity.

in a single line (contig 24323 in Greece and contig 19613 in Denmark). Interestingly, amplification patterns indicated recombination occurred at some loci, as amplification occurred in both males and females at similar frequencies (e.g., contigs 35827 and 41355 in both lines). Overall, these results indicated that all three tested *A. vulgare* lines possess a homologous W sex chromosome. They also suggest that the nonrecombining region of the W sex chromosome is apparently extremely small in *A. vulgare*, as only two loci we tested showed strong linkage to the female sex across all tested *A. vulgare* lines (table 2).

Evolutionarily Young Sex Chromosomes?

The birth of sex chromosomes (reflecting the acquisition of a sex-determining locus by an autosome) may be dated in a species phylogenetic history. *Armadillidium vulgare* belongs to the monophyletic family Armadillidiidae, the last common ancestor of which is estimated to be ~35 My old (Becking et al. 2017). Interestingly, most Armadillidiidae species for which heterogametic systems are known have ZW sex chromosomes (Becking et al. 2017). This raises the possibility that *A. vulgare* shares the same ancestral sex chromosomes with at least some other Armadillidiidae species. If so, sex-linked loci in *A. vulgare* may be expected to be sex linked in other Armadillidiidae species as well, in particular the two loci showing strong linkage to the female sex across all tested *A. vulgare* lines (contigs 60 and 19252). We therefore tested sex linkage of the 12 sex-linked markers of *A. vulgare* in seven other Armadillidiidae species to gain insights into the homology of sex chromosomes, hence their age. We found no evidence that any of the 12 tested loci is linked to sex in any species other than *A. vulgare*, including *Armadillidium*

Table 2. Frequency of 12 Contigs in Males and females of Three *Armadillidium vulgare* Lines from France, Greece, and Denmark.

Population		France		Greece		Denmark	
Contig Id		n^a	**Amplif. Rate (%)**^{b}	n^a	**Amplif. Rate (%)**^{b}	n^a	**Amplif. Rate (%)**^{b}
60	Males	12	16.7	12	8.3	11	0
	Females	13	100 (***)	13	100 (***)	7	100 (***)
14967	Males	8	0	8	0	10	10.0
	Females	10	80.0 (***)	9	0	7	0
18557	Males	8	0	8	25.0	10	0
	Females	10	70.0 (**)	9	0	7	0
19252	Males	8	0	8	25.0	10	0
	Females	10	80.0 (***)	9	77.8 (*)	7	57.1 (**)
19613	Males	12	0	12	0	11	9.1
	Females	13	100 (***)	13	15.4	7	100 (***)
24323	Males	12	0	12	0	11	18.2
	Females	13	100 (***)	13	100 (***)	7	0
24894	Males	12	0	12	0	11	0
	Females	13	84.6 (***)	13	7.7	7	0
35827	Males	8	0	8	87.5	10	40.0
	Females	10	80.0 (***)	9	88.9	7	85.7
41355	Males	8	0	8	50.0	10	20.0
	Females	10	70.0 (**)	9	66.7	7	42.9
44847	Males	12	0	12	0	11	0
	Females	13	84.6 (***)	13	0	7	0
45225	Males	8	0	8	0	10	0
	Females	10	80.0 (***)	9	0	7	0
46089	Males	8	0	8	0	10	0
	Females	10	70.0 (**)	9	0	7	0

[a] Number of individuals tested.

[b] Amplification rate (%). Parentheses indicate a significant difference in amplification rate between males and females in a chi-square test, at <0.05 (*), <0.01 (**), or <0.001 (***) levels.

versicolor, which is the most closely related species to *A. vulgare* in our analysis (supplementary table S6, Supplementary Material online). This result is consistent with a recent origin of *A. vulgare* sex chromosomes, after the divergence between *A. vulgare* and *A. versicolor* which occurred ∼4 My ago (Becking et al. 2017).

To independently assess the age of *A. vulgare* sex chromosomes, we analyzed divergence between the Z and W sex chromosomes at the molecular level. To this end, we identified single nucleotide polymorphisms (SNPs) for which the reference female individual is heterozygous, using the Illumina paired-end data generated for genome sequencing (supplementary table S1, Supplementary Material online). Then, we calculated SNP density in the 27 W-specific contigs identified with the CQ and YGS analyses, after removal of hemizygous regions to focus on regions with orthologs on the Z chromosome (see Materials and Methods). We calculated an average SNP density of 5.3 SNP/kb between allelic Z/W regions of the 27 contigs, compared with 2.9 SNP/kb across all other contigs of the assembly. Thus, we observed a slight (1.8-fold) but significant (Mann–Whitney bilateral *U* test, $U = 272{,}360$, *P* value < 0.000003) excess of SNP density in allelic Z/W regions relative to other regions of the genome. Such a pattern of molecular divergence is typically expected of sex-specific sequences of the genome (Charlesworth et al. 2005), thereby providing additional support for W specificity of the candidate contigs identified with the CQ and YGS analyses. Moreover, based on a substitution rate of 0.178% per My (Becking et al. 2017), the 5.3 SNP/kb density (or 0.53% sequence divergence) in allelic Z/W regions corresponds to a divergence time of ∼3 My (in the absence of recombination), suggesting that the Z and W sex chromosomes probably started to diverge quite recently.

Thus, both analyses we performed support an evolutionary young age of *A. vulgare* Z and W sex chromosomes. Nevertheless, the true age of *A. vulgare* sex chromosomes could be underestimated, as linkage between sex-determining loci and neighboring loci can be broken by local recombination, possibly limiting our ability to test the homology between ZW systems across species and restraining the accumulation of sequence divergence between *A. vulgare* sex chromosomes.

No Evidence That the Female-Specific Region of the Native W Sex Chromosome Is Derived from *Wolbachia*

The circular model of evolution of sex chromosomes in *A. vulgare* raises the possibility that the current W chromosome is derived from a *Wolbachia* insertion into the nuclear genome, generating a female-specific region analogous to the *f* element (fig. 1C). To test this hypothesis, versus the alternative hypothesis that the female-specific region of the W chromosome may not originate from *Wolbachia* (fig. 1B), we searched for *Wolbachia* nuclear inserts potentially resulting from horizontal transfer into the *A. vulgare* genome.

Although the sequenced individuals originated from a line lacking *Wolbachia* endosymbionts and the *f* element

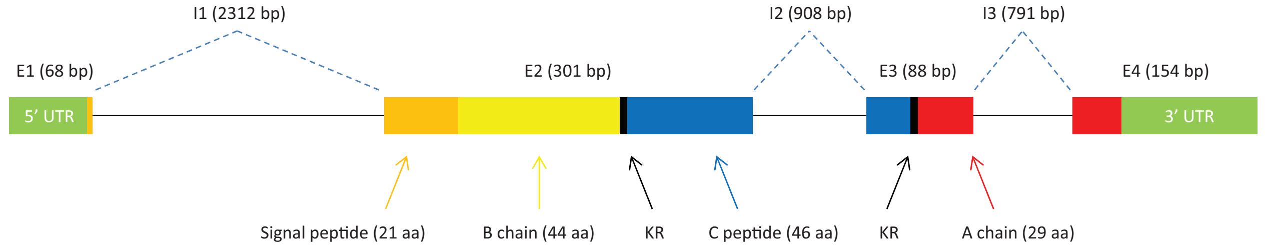

Fig. 4. Schematic representation of the gene coding for the androgenic gland hormone, which is the final effector in the male sexual differentiation cascade. There is a single copy of this gene in the *Armadillidium vulgare* genome (located in scaffold 682). The four exons (E1–E4) and three introns (I1–I3) are shown, along with untranslated regions (UTR) in green. Protein domains are displayed with different colors (orange, yellow, black, blue, and red). Sizes of the different components are given for the gene in base pairs (bp) and for the protein in amino acids (aa).

(independently confirmed by PCR), nucleotide and protein similarity searches of the *A. vulgare* genome assembly revealed the presence of 75 genomic regions showing high similarity to *Wolbachia* (supplementary table S7, Supplementary Material online). Given the co-occurrence within genomic contigs of 73 of these sequences with non-*Wolbachia* sequences, and their interspersed genomic distribution, the *Wolbachia*-like sequences are interpreted as nuclear inserts. This conclusion is further supported by the fact that the 75 regions, ranging in size from 69 to 64,179 bp, account for a total of 192,904 bp, which is far smaller than expected for a complete cytoplasmic *Wolbachia* genome (typically around 1 Mb). A total of 35 apparently intact genes were identified from 10 nuclear inserts, suggesting that the sequenced *A. vulgare* matriline may have carried a *Wolbachia* infection in its recent evolutionary history. Phylogenetic analyses of the *Wolbachia* nuclear inserts aligning to ≥10 *Wolbachia* genomes over >1,500 bp (supplementary fig. S2, Supplementary Material online) all placed *w*VulC, a feminizing *Wolbachia* strain currently known to infect other *A. vulgare* lines (Cordaux et al. 2004), as the most closely related to the nuclear inserts. This analysis suggests that the sequenced *A. vulgare* matriline may have been infected by a *Wolbachia* strain closely related to *w*VulC in the past.

The 72 contigs carrying the 75 nuclear inserts showed CQ scores ranging from 0.78 to 1.46 (mean 1.02) and YGS scores ranging from 0.3% to 16.1% (mean 3.4%), which are typical of autosomal contigs (supplementary table S7, Supplementary Material online and fig. 3*D*). Thus, the female-specific region of the W chromosome apparently does not originate from *Wolbachia*.

Sex Determination and Sexual Differentiation Candidate Genes in the *A. vulgare* Genome

To further explore the genetic basis of sex determination and sexual differentiation in *A. vulgare*, we searched for homologs of genes that have previously been shown to be involved in sex determination and/or sexual differentiation in related organisms. Specifically, we focused on sex-regulator genes characterized in the insect model *Drosophila melanogaster* that have previously been shown to have homologs in malacostracans (the crustacean class of *A. vulgare*), including *Sex-lethal*, *Transformer*, *Transformer-2*, *Doublesex* (and other representatives of the DMRT family) and *Fruitless* (reviewed in Chandler et al. [2018]). We also searched for peptide hormone genes known to be involved in sexual differentiation in malacostracans, including those belonging to the crustacean hyperglycemic hormone (CHH) family and the androgenic gland hormone (AGH) (reviewed in Chandler et al. [2018]).

We identified a total of 31 homologs of the seven queried genes in the *A. vulgare* genome (supplementary table S8, Supplementary Material online). At least one homolog was identified for each gene except *Transformer*, which had none. We identified two *Sex-Lethal*, two *Transformer-2* and seven *Fruitless* homologs respectively sharing 81%, 65%, and 64–100% similarity at the amino acid level, suggesting complex evolutionary histories of gene duplications. The *A. vulgare* genome also contains three DMRT genes. A phylogenetic analysis indicated that they are homologs of *Doublesex*, *DMRT93B*, and *DMRT99B* (supplementary fig. S3, Supplementary Material online). Interestingly, homologs of these three genes have been previously identified in crustaceans (Chandler et al. 2018), including the branchiopod *Daphnia magna* (Kato et al. 2008, 2011), suggesting conserved function albeit not necessarily related to sex determination. The *A. vulgare* genome also contains 16 peptide hormone genes belonging to the CHH family. A phylogenetic analysis indicated that *A. vulgare* CHH genes form a monophyletic group within the type I CHH family sensu Montagné et al. (2010) (supplementary fig. S4, Supplementary Material online). Notably, six of the 16 CHH are highly similar to Arv-CHH, a CHH from the sinus gland of *A. vulgare*, which has been shown to exhibit high hyperglycemic activity (Martin et al. 1993). Finally, a single copy of the AGH gene was identified in scaffold 682 of the *A. vulgare* assembly (fig. 4). It is a 4,622-bp-long gene (including 5′ and 3′ untranslated regions) composed of four exons (ranging in size from 68 to 301 bp) and three introns (ranging in size from 791 to 2,312 bp). The AGH gene encodes a 144 amino acid (aa) protein that differs at three aa positions (N98H, E99D, and V101E) from the sequence previously inferred from cDNA and peptide sequencing (Martin et al. 1999; Okuno et al. 1999). The three aa changes are located in the C peptide, which is cleaved in the mature AGH (Martin et al. 1999; Okuno et al. 1999).

All 31 contigs carrying candidate genes exhibited autosomal CQ and YGS scores (supplementary table S8, Supplementary Material online), with one exception, contig 46089, which was previously identified as one of the 27 W-specific sequences. The CHH gene in contig 46089 is one of seven annotated genes among these 27 sequences (supplementary table S5, Supplementary Material online) and it represents a prime candidate for master sex determination in *A. vulgare*.

Discussion

Reliable Assembly of a Highly Repeated, Large Crustacean Genome

We established the draft genome assembly of the common woodlouse *A. vulgare*, a large crustacean genome, built by de novo assembly using both Illumina and PacBio sequencing technologies. The reliability and completeness of this genome assembly is supported by 1) the very low portion (0.43%) of unidentified nucleotides, 2) consistency of assembly size (1.72 Gb) with inferences based on flow cytometry and densitometry (1.66–1.84 Gb) (Jeffery and Gregory 2014) and *k*-mer spectrum analysis (1.52–1.75 Gb), and 3) the high portion (92%) of identified conserved arthropod BUSCO genes (release September 2016) (Simão et al. 2015; Waterhouse et al. 2018).

The genome sequence of *A. vulgare* constitutes one of a very few available for crustaceans. Indeed, fewer than 20 crustacean genome assemblies are listed in the "Genome" section of NCBI (as of September 2018) despite this subphylum being one of the most speciose metazoan groups, with >50,000 described species (Martin and Davis 2001). One possible reason for this paucity in genome resources is the large genome size of many crustacean species, often exceeding 1 Gb and sometimes reaching tens of Gb (Jeffery and Gregory 2014), making it difficult to obtain good-quality genome assemblies. In this context, the *A. vulgare* assembly we obtained is among the most contiguous of all large-genome crustaceans sequenced to date (table 3). Based on its contiguity and completeness, the *A. vulgare* genome serves as a useful resource for the study of crustacean biology and evolution.

Sex Chromosome Evolution in the Context of Cytoplasmic Sex Determination

We did not find *Wolbachia*-derived sequences in the W-specific contigs of the *A. vulgare* assembly, as all *Wolbachia* nuclear inserts we identified in the *A. vulgare* genome show signatures of autosomal location. Therefore, there is no evidence that the female-specific region of the W sex chromosome of *A. vulgare* is derived from *Wolbachia*. It follows that, according to the circular model of evolution of sex determination mechanisms (proposed by Juchault and Mocquard [1993]) (fig. 1), the ancestral W sex chromosome has not been eliminated from the *A. vulgare* lineage by feminizing *Wolbachia*. This may imply that none of the populations constituting the *A. vulgare* ancestry was exclusively under a regime of cytoplasmic sex determination. Instead, ZW chromosomal and cytoplasmic sex determination were probably coexisting in these populations, similarly to most extant *A. vulgare* populations, in which feminizing *Wolbachia* are far from prevalent (Verne et al. 2012). Clarifying why feminizing *Wolbachia* have not reached high prevalence in *A. vulgare* populations (e.g., due to very recent spread in populations or sex ratio selection) will require additional investigations. This may shed new light on the evolutionary dynamics of sex chromosomes in the context of cytoplasmic sex determination.

We concluded that the W-specific region of the *A. vulgare* genome is extremely small, probably at most hundreds of kb. The W and Z chromosomes could therefore be genetically very similar in sequence and gene content, consistent with apparent lack of heteromorphy in karyotypic studies (Artault 1977) and viability and fertility of WW individuals (Juchault and Legrand 1972). Regions around sex-determining loci often stop recombining (Beukeboom and Perrin 2014), possibly because sexually antagonistic mutations (which are beneficial to one sex but harmful to the other) establish polymorphisms, favoring recombination reduction or suppression (Bachtrog et al. 2014). This may lead to so-called degeneration, with formation of pseudogenes and accumulation of repetitive DNA (Bergero and Charlesworth 2009; Bachtrog 2013). In this context, the lack of evidence for differentiation of sex chromosomes in *A. vulgare* suggests that they have not experienced substantial degeneration. Indeed, allelic Z/W regions exhibit low nucleotide divergence. Furthermore, most W-specific contigs in the *A. vulgare* line used for sequencing are not female specific in other *A. vulgare* lines we analyzed, although they share a homologous W sex chromosome, as attested by four loci showing linkage to the female sex in all or a subset of the lines.

Our results suggest that recombination suppression has not spread very far from the sex-determining locus, if at all. A possible explanation is that *A. vulgare* sex chromosomes are too young to have accumulated sexually antagonistic mutations yet. This scenario is consistent with the low divergence recorded between allelic Z/W regions and the absence of evidence that *A. vulgare* W-specific sequences are also female specific in any other Armadillidiidae species we tested, leading to the conclusion that current *A. vulgare* sex chromosomes may be just a few million years old. A recent origin of *A. vulgare* sex chromosomes would support the proposition by Becking et al. (2017) that the magnitude of sex chromosome turnover in terrestrial isopods (inferred by Becking et al. [2017]) has been underestimated. Alternatively, it remains formally possible that current *A. vulgare* sex chromosomes may be substantially older, possibly as old as ∼35 My if they are ancestral to the Armadillidiidae family (Becking et al. 2017). This would imply that they constitute old sex chromosomes still undergoing extensive recombination (as reported in frogs [Rodrigues et al. [2018]), thereby making them an original model for studying the molecular evolution of sex chromosomes. Investigating the patterns of recombination in *A. vulgare* thus represents a promising area of research, for further understanding genome and sex chromosome evolution in isopods and, more generally, in crustaceans.

Insights into Genetic Basis of Sex Determination and Sexual Differentiation

The major molecular effector identified so far in the *A. vulgare* sex determination and sexual differentiation cascade is the AGH, which is the final effector in the cascade that triggers male sexual differentiation (Katakura 1984; Legrand et al. 1987; Martin et al. 1999; Okuno et al. 1999). In addition to AGH (fig. 4), our targeted search for genes involved in sex determination or sexual differentiation in the *A. vulgare* genome identified 30 candidates. Interestingly, we identified three members of the DMRT gene family in the *A. vulgare* genome, including *Doublesex*, *DMRT93B*, and *DMRT99B* homologs. DMRT genes have been found to be frequently involved in sex determination or sexual differentiation in animals (Kopp 2012; Matson and Zarkower 2012; Beukeboom and Perrin 2014), including crustaceans (Kato et al. 2011; Chandler et al. 2018). Likewise, they may be part of the sex determination or sexual differentiation cascade in *A. vulgare*, but they cannot be considered as master sex determination genes because they exhibit autosomal locations in the genome.

The prime candidate gene that displayed a signature of W specificity was a CHH gene homolog located in contig 46089. CHH genes are multifunctional peptide hormone genes known to be involved in sexual differentiation in malacostracans (Nagaraju 2011; Webster et al. 2012; Chandler et al. 2018). These combined features make the CHH gene in contig 46089 a promising candidate master gene for sex determination in *A. vulgare*. Yet, it should be noted that, although W specificity of contig 46089 in the sequenced *A. vulgare* line was independently confirmed by PCR, it did not display W specificity in other tested *A. vulgare* lines. The CHH gene in contig 46089 may therefore simply locate close to the locus determining sex in *A. vulgare*, without being the master gene itself. In any event, we identified several candidate genes whose investigation may provide new insight into the molecular genetic basis of sex determination and sexual differentiation in *A. vulgare*. Combined with the identification of candidate genes of *Wolbachia*-induced feminization (Pichon et al. 2012; Badawi et al. 2018), studies of nuclear genes may ultimately lead to the elucidation of the molecular mechanism that endosymbionts use to mediate the development of genetic males into phenotypic females in their animal hosts.

Materials and Methods

DNA Library Preparation and Sequencing

All *A. vulgare* individuals used for sequencing were from our inbred laboratory matriline BF, which is originally derived from wild animals caught in Nice, France, in 1967. Specifically, we used ZW genetic females descended from a single female from family BF2787 for genome sequencing and ZW sisters and ZZ brothers from family BF2875 for analyses of sex-linked sequences. Total genomic DNA was extracted using the Qiagen DNeasy Blood and Tissue Kit, according to the protocol for animal tissues (3 h of incubation in proteinase K at 56 °C and 30 min of RNase treatment at 37 °C). Absence of *Wolbachia* endosymbionts and the *f* element in all samples was confirmed by PCR, as described previously (Leclercq et al. 2016). For genome sequencing, three types of Illumina libraries were generated, including short paired-end reads, mate-pair reads, and long paired-end reads, with ∼450-bp, ∼5-kb, and ∼20-kb insert sizes, respectively (supplementary table S1, Supplementary Material online). In addition, PacBio RS II sequencing (P6C4 chemistry) was performed to obtain long sequencing reads (supplementary table S1, Supplementary Material online). For the analysis of sex-linked sequences, we sequenced two pools each made of equimolar DNA amounts from 10 brothers or sisters, using Illumina paired-end libraries with ∼250-bp insert sizes (supplementary table S1, Supplementary Material online). GenBank accession numbers for Illumina and PacBio sequence data sets are provided in supplementary table S1, Supplementary Material online.

Illumina reads were subjected to quality control with FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/; last accessed February 7, 2019) and low quality bases were filtered out using Trimmomatic (version 0.33) (Bolger et al. 2014), by setting the quality thresholds to a minimum Phred score mean of 15 across each four-base window and to 20 for heading and tailing bases. Paired-end reads were also trimmed for adaptors with Trimmomatic setting seed mismatches at 2, palindrome clip threshold at 30, and simple clip threshold at 10. For mate-pair reads, we used Skewer (version 0.1.125) (Jiang et al. 2014) to remove adaptors as this tool is better suited to middle junction linkers. Long paired-end reads were cleaned for adaptors by the sequencing provider. Reads were then filtered for PCR duplicates with PRINSEQ-lite (version 0.20.3) (Schmieder and Edwards 2011). Base errors of raw PacBio subreads were corrected with LoRDEC (version 0.4) (Salmela and Rivals 2014), by constructing a *k*-mer de Bruijn graph from Illumina reads, using a *k*-mer size of 23.

Reference Genome Assembly

After filtering, Illumina reads were composed of 397,021,412 paired reads (77.36 Gb), 20,704,514 mate pairs (3.65 Gb), and 1,463,027 long paired reads (0.35 Gb). Genome size was estimated from counting canonical 17-, 21-, 25-, and 31-mer frequencies of quality-filtered paired-end reads with Jellyfish (version 2.2.6) (Marçais and Kingsford 2011). Based on binomial distributions obtained with GenomeScope (Vurture et al. 2017), we inferred genome size by dividing the total number of *k*-mers by the coverage observed at the highest peak (Sohn and Nam 2018). High frequency *k*-mers (>15,000) were excluded, as they usually represent contaminants that can artificially inflate genome size (Vurture et al. 2017).

Genome assembly was performed following two different approaches: Illumina-only versus hybrid assembly (involving both Illumina and PacBio reads). The Illumina-only assembly was obtained from paired-end, mate-pair, and long paired-end reads, using SOAPdenovo (version 2.04) (Luo et al. 2012) with a *k*-mer size of 61. A workflow of the hybrid assembly is shown in supplementary figure S5, Supplementary Material online. First, PacBio reads longer than 5 kb (including circular consensus sequences and corrected subreads) were selected for the assembly process (1,590,883 long reads, or 17.39 Gb, in total). Genome assembly was performed with DBG2OLC (Ye et al. 2016), using both PacBio long reads and Illumina unitigs as input. Illumina unitigs had been generated beforehand by SparseAssembler (Ye et al. 2012) using a *k*-mer size of 71, with maximal coverage (NodeCovTh) and linking (EdgeCovTh)

Table 3. Comparison of Assembly Statistics for *Armadillidium vulgare* and the 17 Crustacean Assemblies >20 Mb in Size Available in the Genome Section of NCBI[a] (as of september 2018).

Species	Class or Subclass	Assembly Size (Mb)	Undetermined Nucleotides (%)	Number of Scaffolds	Scaffold N_{50} (kb)	Number of Contigs	Contig N_{50} (kb)	GenBank Accession Number
Daphnia magna	Branchiopoda	130	18.12	28,801	398	38,559	10	LRGB00000000.1
Daphnia pulex	Branchiopoda	197	19.57	5,186	642	18,989	49	ACJG00000000.1
Eulimnadia texana	Branchiopoda	121	<0.01	108	18,070	110	10,428	NKDA00000000.1
Triops cancriformis	Branchiopoda	109	0.00	0	0	60,629	13	BAYF00000000.1
Acartia tonsa	Copepoda	989	0.31	>100,000	4	>100,000	3	OETC00000000.1
Caligus rogercresseyi	Copepoda	398	0.00	0	0	>100,000	2	LBBV00000000.1
Eurytemora affinis	Copepoda	389	0.65	6,171	252	14,526	68	AZAI00000000.2
Lepeophtheirus salmonis	Copepoda	665	0.00	0	0	83,165	167	LBBX00000000.1
Oithona nana	Copepoda	85	3.46	4,626	401	7,437	39	FTRT00000000.1
Hyalella azteca	Eumalacostraca[b]	551	0.48	18,000	215	23,426	114	JQDR00000000.2
Parhyale hawaiensis	Eumalacostraca[b]	4,024	27.40	>100,000	69	>1,000,000	4	LQNS00000000.1
Caridina multidentata	Eumalacostraca[c]	1,949	<0.01	>1,000,000	<1	>1,000,000	<1	BDMR000000000.1
Eriocheir sinensis	Eumalacostraca[c]	1,549	2.69	6,500	490	48,470	45	GCA_003336515.1
Penaeus japonicus	Eumalacostraca[c]	1,660	2.13	>1,000,000	<1	>1,000,000	<1	NIUR000000000.1
Penaeus monodon	Eumalacostraca[c]	1,447	5.54	>1,000,000	<1	>1,000,000	<1	NIUS000000000.1
Procambarus virginalis	Eumalacostraca[c]	3,290	50.51	>1,000,000	39	>1,000,000	1	MRZY000000000.1
Armadillidium vulgare	Eumalacostraca[d]	1,725	0.43	43,541	51	52,671	38	SAUD00000000
Ligia exotica	Eumalacostraca[d]	954	<0.01	>1,000,000	<1	>1,000,000	<1	BDMT000000000.1

[a] https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/crustacea, accessed on September 18, 2018.
[b] Amphipoda order.
[c] Decapoda order.
[d] Isopoda order.

thresholds respectively of 2 and 1. The DBG2OLC assembler aligns Illumina unitigs on the long reads and uses them as anchors to deflate the long read overlap graph complexity, thereby decreasing the assembly computational time. For the alignment, we used a matching *k*-mer size of 17 with an adaptive coverage threshold of 0.01 of the unitig length (AdaptiveTh) for a minimum of 2 (kmerCovTh). As we had a moderate sequencing depth of PacBio reads and 1–2% of long subreads were expected to be chimeric (Fichot and Norman 2013), we disabled the RemoveChimera parameter to avoid over removal of reads due to abundant repeats. However, we selected the most stringent MinOverlap length (150 bp) allowed to construct accurate overlaps. This led to the backbone raw assembly. Next, reads and unitigs included in the assembly were realigned on each related backbone, and a consensus module Sparc (Ye and Ma 2016) was then called to build the consensus of the most likely contig sequences. A second polishing step was performed by mapping Illumina paired-end reads on the DBG2OLC-generated assembly with Bowtie 2 (version 2.2.9) (Langmead and Salzberg 2012), and scanning the resulting alignment with Pilon (version 1.18) (Walker et al. 2014) to reduce indel and substitution errors. Larger structural errors in the draft genome assembly were also evaluated, so that misassemblies were first detected by aligning long paired-end reads and then breaking at coverage drops, using REAPR (version 1.0.18) (Hunt et al. 2013).

We scaffolded the assembly with mate-pair and long paired-end reads by three iterations of SSPACE (Boetzer et al. 2011). Subsequently, we applied PBJelly 2 (English et al. 2012), which aligned corrected PacBio subreads and circular consensus sequences to merge scaffolds and to fill gaps. The remaining undetermined nucleotides were then minimized with GapFiller v1.11 (Boetzer and Pirovano 2012) using paired-end reads. Scaffolding and gap filling steps were applied twice according to the above description. Finally, we used BLAT (version 36x2) (Kent 2002) to align an *A. vulgare* transcriptome assembly to the full genome assembly, and confident alignments (identity $\geq$98%) were exploited for scaffolding using L_RNA_scaffolder (Xue et al. 2013). The *A. vulgare* transcriptome was assembled de novo with the Trinity package (version 2.1.1) (Haas et al. 2013) setting "–SS_lib_type F," from Illumina reads generated previously (Romiguier et al. 2014). To this end, low quality reads and sequencing adaptors had been filtered out using Trimmomatic (version 0.33) (Bolger et al. 2014) and poly-A tails trimmed at five positions from the ends with PRINSEQ-lite (version 0.20.4) (Schmieder and Edwards 2011).

To remove potential contaminants from the genome assembly, contigs were searched for similarities against the non-redundant NCBI nucleotide (nt) database (release July 2016) using BlastN (version 2.2.30+) (Camacho et al. 2009)) "-task megablast" and compared with Uniref90 (release September 2016) using diamond (version 0.7.12) (Buchfink et al. 2015) following the BlastX search. For both tasks, *e*-value cutoff was set to 10^{-25} and taxa were assigned according to the highest-scoring matches sum across all hits for each taxonomic rank in the two databases. Sequencing coverage was estimated by mapping paired reads to the genome with Bowtie2 (version 2.2.9) (Langmead and Salzberg 2012) in "very-sensitive" mode. Contigs were then visualized with Blobtools (version 0.9.17) (Kumar et al. 2013) using taxon-annotated-GC-coverage plots. We retained all contigs annotated as metazoans as likely *A. vulgare* sequences. Therefore, archaea, prokaryote, fungi, and streptophyte hits were conservatively excluded, except those assigned to *Wolbachia* and viruses, which may be naturally inserted in the genome (Thézé et al. 2014;

Leclercq et al. 2016). Unassigned contigs were also retained, as the absence of hits for these contigs could be explained by the lack of sequenced genomes from closely related species in public databases.

The completeness of the genome assembly was assessed by searching for similarities against highly conserved genes among arthropods. For this purpose, we ran BUSCO (version 3.0.1) (Simão et al. 2015; Waterhouse et al. 2018) in “genome” mode specifying the arthropod profile library containing 1,066 arthropod core proteins (release September 2016) (Simão et al. 2015; Waterhouse et al. 2018).

Genome Annotation

Repeats were identified de novo and classified with RepeatModeler (version open-1.0.8) (Tarailo-Graovac and Chen 2009), allowing the runs of RECON (Bao and Eddy 2002), RepeatScout (Price et al. 2005), and Tandem Repeats Finder (Benson 1999) to build an *A. vulgare*-specific repeat library. Remaining unidentified repeats were marked as unknown and were classified by TEclass (version 2.1.3) (Abrusan et al. 2009). The classifiers were built using the version 21.08 of RepBase (Jurka et al. 2005). The repeats were then searched against a custom UniProt-SwissProt database (version September 2016), from which transposases were discarded to exclude potential gene fragments in the TEs. The transposase database was obtained by merging the RepeatPeps library from RepeatMasker (version open-4.0.6) (Tarailo-Graovac and Chen 2009) and the TELibrary proteins from the TESeeker package (Kennedy et al. 2011) (https://repository.library.nd.edu/view/24/TELibrary.zip; last accessed February 7, 2019). We used BlastX (version 2.2.29+) (Camacho et al. 2009) to exclude elements with significant hits (*e*-value $\leq 10^{-10}$) to genes in the custom protein database. Those sequences were also excluded from the repeat library with 50-bp upstream and downstream of the blast hit using a Perl script ProtExcluder1.2 (http://www.hrt.msu.edu/uploads/535/78637/ProtExcluder1.2.tar.gz; last accessed February 7, 2019). Only repeats larger than 50 bp were retained in the specific *A. vulgare* repeat library. Then, de novo repeats were combined with the RepBase library (Update 20150807) included in the RepeatMasker package. A final repeat masking step was performed with RepeatMasker (version open-4.0.6) (Tarailo-Graovac and Chen 2009) in the more sensitive slow search mode (-s) using ncbi-RMBlastN (version 2.2.27, http://www.repeatmasker.org/RMBlast.html; last accessed February 7, 2019) (-e ncbi) and specifying the combined repeat library (-lib). Pairwise nucleotide distances between repeat family copies and their respective consensus sequences were corrected using the Kimura 2-parameter model. We used the “createRepeatLandscape.pl” Perl script included in the RepeatMasker package to plot repeat landscape with relative abundance and divergence for each repeat class.

Gene annotation was performed with Maker (version 2.31.8) (Holt and Yandell 2011) in two iterations. The genome assembly was masked for repetitive elements using RepeatMasker (by providing the de novo *A. vulgare* repeat library) and RepeatRunner included in Maker. Organism model was set to “all” in the configuration file to also use the whole RepBase library by RepeatMasker. The first iteration of Maker used transcriptome assembly (est2genome = 1) and UniProt-SwissProt database (version September 2016) (protein2genome = 1) alignments as sources of evidence for homology-based gene prediction. Maker optimized the alignment step by identifying splice sites using Exonerate (Slater and Birney 2005) to produce a first draft of gene models. Ab initio gene prediction was performed with SNAP (version 2006-07-28) (Korf 2004) and AUGUSTUS (version 3.2.2) (Stanke et al. 2006). SNAP was trained to generate hidden Markovian models for the 1,000 best genes. AUGUSTUS was trained by sampling SNAP output and optimized to construct a new species gene model. The second run of Maker included the SNAP hidden Markovian model file and the optimized species gene model for AUGUSTUS. Considering these elements, Maker refined the final gene models in a GFF3 output file. The reliability of gene annotations was evaluated by measuring their annotation edit distances (Eilbeck et al. 2009).

Predicted genes were functionally annotated by combining two approaches. First, InterProScan (version 5.21-60.0) (Quevillon et al. 2005) was used to identify functional protein domains using the PfamA database (version 30.0) (Finn et al. 2016). GO terms were defined by running “-goterms” option. Next, we used BlastP (version 2.2.30+) (Camacho et al. 2009) to search best hits (*e*-value $\leq 10^{-6}$) of predicted genes to the UniProt-SwissProt database (release September 2016). Finally, we ran “ipr_update_gff” and “maker_functional_gff” tools distributed with Maker to update the GFF3 file with the functional annotations.

Analyses of W-Specific Sequences

The CQ method (Hall et al. 2013) was used to calculate male to female mean per-site coverage ratio of aligned reads on female reference genome contigs. Reads were aligned using Bowtie2 (version 2.2.9) (Langmead and Salzberg 2012) in the –very-fast mode. We then retained only reads aligned with no mismatch for CQ analysis using bamtools “filter” command by setting “-tag NM: 0.” These stringent criteria were defined to reduce false negative or positive candidates which could occur by over mapping of male or female reads containing sequencing errors. Even if mismatches could reflect allelic variations between siblings, these should be rarer than sequencing errors. A homemade R script was then run to calculate CQ ratios and the maximum CQ score was set to 0.3 to retain contigs as W-specific candidates, as recommended by CQ authors (Hall et al. 2013). The YGS method (Carvalho and Clark 2013) based on *k*-mer counts ($k = 18$) was also used as a complementary approach. The YGS method was performed to scan female contigs by counting their respective proportion of unmatched single-copy *k*-mers to male reads. The minimum YGS score was set to 40% to retain contigs as W-specific candidates. This threshold was selected to account for the high repetitive nature of the *A. vulgare* genome. We beforehand replaced *k*-mer bases with Phred scores <10 by N’s and filtered low-frequency *k*-mers with Jellyfish (version

2.2.6) (Marçais and Kingsford 2011) setting (–lower count = 2).

Female sex linkage of the candidate contigs was assessed by designing diagnostic PCR tests (supplementary table S5, Supplementary Material online). Primers were designed in unique regions of the contigs (by masking annotated repeats and confirming primer sequence specificity by blast searches against the entire assembly) showing no coverage by male reads, using Primer3Plus (version 2.3.7) (Untergasser et al. 2012). PCR reactions were carried out in 25 µl with 0.75 µl of DMSO, 12.5 µl of Master Mix 2x containing Phusion High-Fidelity DNA polymerase, 1.25 µl each primer (10 µM), and 1 µl of DNA. PCRs were conducted using the following temperature cycling: initial denaturation at 98 °C for 30 s, followed by 35 cycles of denaturation at 98 °C for 15 s, annealing at 55 °C for 30 s and elongation at 72 °C for 30 s, ending with a 2-min elongation step at 72 °C. An initial PCR screen was performed using individual DNA samples from 18 individuals (eight males and 10 females) from the BF2875 family sequenced as part of the CQ and YGS analyses. Contigs passing the initial screen were then tested in 50 additional individuals, including 1) seven individuals (four males and three females) from the BF matriline, 2) 25 individuals (12 males and 13 females) from our laboratory matriline ZM (derived from wild animals caught in Heraklion, Greece, in 1989), and 3) 18 individuals (11 males and 7 females) from our laboratory matriline WXa (derived from wild animals caught in Helsingor, Denmark, in 1982). In addition, the markers were tested in five males and five females from the seven following Armadillidiidae species: *A. versicolor*, *Armadillidium maculatum*, *Armadillidium depressum*, *Armadillidium granulatum*, *Armadillidium assimile*, *Armadillidium nasatum*, and *Eluma purpurascens*. Two autosomal controls were successfully amplified in all samples: beta actin (primer sequences: 5′-GATTCTGGTGATGG TGTATCTC and 5′-CGGTGGTAGTGAAAGTGTAAC, annealing temperature: 60 °C, product size: 150 bp) and 18S rRNA (primer sequences: 5′-AATAAAAAGACCGA TTTCCG and 5′-TTTTGTAACTACGAAGCCG, annealing temperature: 55 °C, product size: 615 bp). Absence of *Wolbachia* endosymbionts and the *f* element in all samples was confirmed by PCR, as described previously (Leclercq et al. 2016).

To identify heterozygous SNPs in the BF2787 reference female, we applied the Genome Analysis ToolKit (GATK) pipeline (version 3.8-0-ge9d806836) (Van der Auwera et al. 2013). Short paired-end reads were aligned to the genome assembly using BWA-MEM (version 0.7.16a-r1181) (Li 2013) with default settings. Picard tools (version 2.12.0) (http://broadinstitute.github.io/picard; last accessed February 7, 2019) were then used to mark PCR duplicates in the alignment file. We realigned reads around indels and recalibrated base quality scores using GATK tools and guidelines. We then called SNPs using GATK's HaplotypeCaller followed by hard filtering using the following parameters: QualByDepth (QD) < 2.0, FisherStrand (FS) > 60.0, MSMappingQuality (MQ) $<$ 40.0, MappingQualityRankSumTest (MQRankSum) < -12.5, ReadPosRankSumTest (ReadPosRankSum) < -8.0.

To compare heterozygosity (SNP density) between candidate W-linked contigs and other contigs, we counted SNPs for which the BF2787 reference female was heterozygous, discarding those with quality score <200 and ignoring contigs shorter than 400 bp. As the candidate W-linked contigs may be partly hemizygous in a WZ female, we specifically computed heterozygosity of nonhemizygous regions, by retaining only parts of these contigs that were covered by >5 reads sequenced from males. To estimate male-read coverage, we aligned the short reads from males on the reference genome using the procedure described above for the reference female. Regions covered by male reads were obtained from depth-of-coverage data computed by SAMtools (Li et al. 2009), ignoring reads with mapping quality 0. Regions distant by <10 bp were merged, and those shorter than 100 bp after merging were discarded.

Analyses of *Wolbachia* Nuclear Inserts

The analysis of potential DNA integrations of *Wolbachia* sequences into the *A. vulgare* genome was carried out using BlastN (version 2.2.30+) (Camacho et al. 2009) to search for nucleotide similarities between *A. vulgare* contigs and 36 *Wolbachia* genomes listed in supplementary figure S2, Supplementary Material online. The BlastN alignments (High-scoring Segment Pairs [HSPs]) were parsed as follows. HSPs nested within other HSPs (in *A. vulgare* contig coordinates) were discarded. Partially overlapping HSPs were combined, as were HSPs separated by ≤50 bp. Sequences hitting to *Wolbachia* larger than 100 bp were then retrieved with their upstream and downstream 50 bp. To avoid retrieving sequence regions originating from other bacteria or mitochondria, retained sequences were searched for similarities against a custom nucleotide database containing NCBI reference prokaryote and mitochondrial genomes as well as the 36 aforementioned *Wolbachia* genomes. BlastN was set to retain at most ten hits per query and we retained the HSP with the highest score when multiple HSPs totally overlapped. We then retained each HSP hitting to *Wolbachia* with an *e*-value $\leq 10^{-25}$. *Wolbachia* inserts were also screened by searching for protein similarities of predicted *A. vulgare* proteins to nonredundant NCBI nucleotide (nr) database (version Mars 2017) using BlastP (version 2.2.30+) (Camacho et al. 2009) set to an *e*-value of $\leq 10^{-6}$ and to retain at most one alignment per query. Alignment output was manually parsed and newly identified *Wolbachia* genes were added to the annotation GFF3 file. Retained sequences hitting to *Wolbachia* were then visualized with GenomeView (Abeel et al. 2012) to identify *Wolbachia* insert boundaries.

To identify *Wolbachia* strains most closely related to nuclear inserts in the *A. vulgare* genome, we performed phylogenetic analyses separately for nuclear inserts aligning to at least 10 out of the 36 aforementioned *Wolbachia* genomes over >1,500 bp. Each insert sequence was searched for similarities against the 36 *Wolbachia* genomes with BlastN (version 2.2.30+) (Camacho et al. 2009). HSPs of ≥30 bp were retained. For each *Wolbachia* strain, we retained the largest HSP when HSPs were nested within others and the HSP with the highest score when multiple HSPs partially overlapped.

Pairwise alignments were performed using the Biostrings package (http://bioconductor.org/packages/release/bioc/html/Biostrings.html; last accessed February 7, 2019). Then, to produce a multiple-sequence alignment, pairwise alignments were trimmed to retain the longest shared region aligning to the various *Wolbachia* strains and stripped from all positions corresponding to deleted nucleotides in the nuclear insert. Stacking the resulting alignments produced an alignment of all *Wolbachia* strains to the nuclear insert. A neighbor-joining tree was then built from the resulting alignment using MEGAX (Kumar et al. 2018), with the Kimura 2-parameter substitution model. Robustness of tree topology was assessed by bootstrap analysis using 500 replicates.

Analyses of Sex Determination and Sexual Differentiation Candidate Genes

Genes homologous to *Sex-lethal*, *Transformer*, *Transformer-2*, *Doublesex* (and other representatives of the DMRT family), *Fruitless*, CHH, and AGH genes were identified from the *A. vulgare* genome annotation. Phylogenetic analyses of DMRT and CHH homologs were performed using amino acid sequences. For DMRT genes, we aligned the conserved DM domain of *A. vulgare* homologs with those listed in Kato et al. (2011). Prottest (version 3.4) (Darriba et al. 2011) was used to identify the best substitution model (JTT + G) and a maximum likelihood phylogeny was reconstructed using RAxML (version 7.4.6) (Stamatakis 2006), with 100 independent replicates followed by 1,000 replicates of bootstrap resampling. For CHH genes, we aligned *A. vulgare* homologs with those listed in Montagné et al. (2010). A neighbor-joining tree was built using MEGAX (Kumar et al. 2018), with the JTT + G substitution model and 1,000 bootstrap replicates.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Sébastien Leclercq for discussions at an early stage of the project, and the Associate Editor and two anonymous reviewers for their constructive comments on a previous version of the manuscript. This work was funded by European Research Council Starting grant 260729 (EndoSexDet) and Agence Nationale de la Recherche grant ANR-15-CE32-0006-01 (CytoSexDet) to R.C., the 2015–2020 State-Region Planning Contracts (CPER) and European Regional Development Fund (FEDER), and intramural funds from the Centre National de la Recherche Scientifique and the University of Poitiers. C.G. was supported by Agence Nationale de la Recherche grant ANR-15-CE32-0011-01 (TransVir).

References

Abeel T, Van Parys T, Saeys Y, Galagan J, Van de Peer Y. 2012. GenomeView: a next-generation genome browser. *Nucleic Acids Res.* 40(2): e12.

Abrusan G, Grundmann N, DeMester L, Makalowski W. 2009. TEclass—a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* 25(10): 1329–1330.

Artault J-C. 1977. Contribution à l'étude des garnitures chromosomiques chez quelques Crustacés Isopodes. Thèse de 3ème cycle. Université de Poitiers. Poitiers, France.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 25(1): 25–29.

Bachtrog D. 2013. Y chromosome evolution: emerging insights into processes of Y chromosome degeneration. *Nat Rev Genet.* 14(2): 113–124.

Bachtrog D, Mank JE, Peichel CL, Kirkpatrick M, Otto SP, Ashman T-L, Hahn MW, Kitano J, Mayrose I, Ming R, et al. 2014. Sex determination: why so many ways of doing it? *PLoS Biol.* 12(7): e1001899.

Badawi M, Moumen B, Giraud I, Grève P, Cordaux R. 2018. Investigating the molecular genetic basis of cytoplasmic sex determination caused by *Wolbachia* endosymbionts in terrestrial isopods. *Genes* 9(6): 290.

Bao Z, Eddy SR. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* 12(8): 1269–1276.

Becking T, Giraud I, Raimond M, Moumen B, Chandler C, Cordaux R, Gilbert C. 2017. Diversity and evolution of sex determination systems in terrestrial isopods. *Sci Rep.* 7(1): 1084.

Benson G. 1999. Tandem Repeats Finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27(2): 573–580.

Bergero R, Charlesworth D. 2009. The evolution of restricted recombination in sex chromosomes. *Trends Ecol Evol.* 24(2): 94–102.

Beukeboom L, Perrin N. 2014. The evolution of sex determination. Oxford (NY): Oxford University Press.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27(4): 578–579.

Boetzer M, Pirovano W. 2012. Toward almost closed genomes with GapFiller. *Genome Biol.* 13:56.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15): 2114–2120.

Bouchon D, Rigaud T, Juchault P. 1998. Evidence for widespread *Wolbachia* infection in isopod crustaceans: molecular identification and host feminization. *Proc Biol Sci.* 265(1401): 1081–1090.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods.* 12(1): 59–60.

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

Carvalho AB, Clark AG. 2013. Efficient identification of Y chromosome sequences in the human and *Drosophila* genomes. *Genome Res.* 23(11): 1894–1907.

Chandler JC, Elizur A, Ventura T. 2018. The decapod researcher's guide to the galaxy of sex determination. *Hydrobiologia.* 825: 61–80.

Charlesworth D, Charlesworth B, Marais G. 2005. Steps in the evolution of heteromorphic sex chromosomes. *Heredity* 95(2): 118–128.

Cordaux R, Bouchon D, Grève P. 2011. The impact of endosymbionts on the evolution of host sex-determination mechanisms. *Trends Genet.* 27(8): 332–341.

Cordaux R, Gilbert C. 2017. Evolutionary significance of *Wolbachia*-to-animal horizontal gene transfer: female sex determination and the f element in the isopod *Armadillidium vulgare*. *Genes* 8(7): 186.

Cordaux R, Michel-Salzat A, Frelon-Raimond M, Rigaud T, Bouchon D. 2004. Evidence for a new feminizing *Wolbachia* strain in the isopod *Armadillidium vulgare*: evolutionary implications. *Heredity* 93(1): 78–84.

Cordaux R, Pichon S, Hatira HBA, Doublet V, Grève P, Marcadé I, Braquart-Varnier C, Souty-Grosset C, Charfi-Cheikhrouha F, Bouchon D. 2012. Widespread *Wolbachia* infection in terrestrial isopods and other crustaceans. *Zookeys* 176:123–131.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27(8): 1164–1165.

Dupeyron M, Leclercq S, Cerveau N, Bouchon D, Gilbert C. 2014. Horizontal transfer of transposons between and within crustaceans and insects. *Mob DNA* 5(1): 4.

Eilbeck K, Moore B, Holt C, Yandell M. 2009. Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics* 10:67.

English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, et al. 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* 7(11): e47768.

Fichot EB, Norman RS. 2013. Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. *Microbiome* 1(1): 10.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44(D1): D279–D285.

Gene Ontology Consortium. 2017. Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res*. 45:D331–D338.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-Seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. 8(8): 1494–1512.

Hall AB, Qi Y, Timoshevskiy V, Sharakhova MV, Sharakhov IV, Tu Z. 2013. Six novel Y chromosome genes in Anopheles mosquitoes discovered by independently sequencing males and females. *BMC Genomics.* 14:273.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12:491.

Hornung E. 2011. Evolutionary adaptation of oniscidean isopods to terrestrial life: structure, physiology and behavior. *Terr Arthropod Rev*. 4(2): 95–130.

Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. 2013. REAPR: a universal tool for genome assembly evaluation. *Genome Biol*. 14(5): R47.

Jeffery NW, Gregory TR. 2014. Genome size estimates for crustaceans using Feulgen image analysis densitometry of ethanol-preserved tissues. *Cytometry A* 85(10): 862–868.

Jiang H, Lei R, Ding S-W, Zhu S. 2014. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* 15:182.

Juchault P, Legrand J-J. 1972. Croisements de néo-mâles expérimentaux chez *Armadillidium vulgare* Latr. (Crustacé Isopode Oniscoïde). Mise en évidence d'une hétérogamétie femelle. *C R Acad Sci Paris* 1387–1389:274–276.

Juchault P, Mocquard JP. 1993. Transfer of a parasitic sex factor to the nuclear genome of the host: a hypothesis on the evolution of sex-determining mechanisms in the terrestrial isopod *Armadillidium vulgare* Latr. *J Evol Biol*. 6(4): 511–528.

Juchault P, Rigaud T. 1995. Evidence for female heterogamety in two terrestrial crustaceans and the problem of sex chromosome evolution in isopods. *Heredity* 75(5): 466–471.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*. 110(1–4): 462–467.

Katakura Y. 1984. Sex differentiation and androgenic gland hormone in the terrestrial isopod *Armadillidium vulgare*. In: Symposia of the zoological society of London. Cambridge: Cambridge University Press. p. 127–142.

Kato Y, Kobayashi K, Oda S, Colbourn JK, Tatarazako N, Watanabe H, Iguchi T. 2008. Molecular cloning and sexually dimorphic expression of DM-domain genes in *Daphnia magna*. *Genomics* 91(1): 94–101.

Kato Y, Kobayashi K, Watanabe H, Iguchi T. 2011. Environmental sex determination in the branchiopod crustacean *Daphnia magna*: deep conservation of a doublesex gene in the sex-determining pathway. *PLoS Genet*. 7(3): e1001345.

Kennedy RC, Unger MF, Christley S, Collins FH, Madey GR. 2011. An automated homology-based approach for identifying transposable elements. *BMC Bioinformatics* 12:130.

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Res*. 12(4): 656–664.

Kopp A. 2012. Dmrt genes in the development and evolution of sexual dimorphism. *Trends Genet*. 28(4): 175–184.

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* 5:59.

Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M. 2013. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front Genet.* 4:237.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol*. 35(6): 1547–1549.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 9(4): 357–359.

Leclercq S, Thézé J, Chebbi MA, Giraud I, Moumen B, Ernenwein L, Grève P, Gilbert C, Cordaux R. 2016. Birth of a W sex chromosome by horizontal transfer of *Wolbachia* bacterial symbiont genome. *Proc Natl Acad Sci.U S A*. 113(52): 15036–15041.

Legrand J-J, Legrand-Hamelin E, Juchault P. 1987. Sex determination in Crustacea. *Biol Rev*. 62(4): 439–447.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv* 1303.3997.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16): 2078–2079.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1: 18.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* 27(6): 764–770.

Martin G, Juchault P, Legrand JJ. 1973. Mise en évidence d'un micro-organisme intracytoplasmique symbiote de l'oniscoïde *Armadillidium vulgare* Latr., dont la présence accompagne l'intersexualité ou la féminisation total des mâles génétiques de la lignée thélygène. *C R Acad Sci*. 276:2213–2216.

Martin G, Sorokine O, Dorsselaer A. 1993. Isolation and molecular characterization of a hyperglycemic neuropeptide from the sinus gland of the terrestrial isopod *Armadillidium vulgare* (Crustacea). *Eur J Biochem*. 211(3): 601–607.

Martin G, Sorokine O, Moniatte M, Bulet P, Hetru C, Van Dorsselaer A. 1999. The structure of a glycosylated protein hormone responsible for sex determination in the isopod, *Armadillidium vulgare*. *Eur J Biochem*. 262(3): 727–736.

Martin J, Davis G. 2001. An updated classification of the recent Crustacea. *Nat Hist Mus Los Angel Cty Sci Ser*. 39:1–124.

Matson CK, Zarkower D. 2012. Sex and the singular DM domain: insights into sexual regulation, evolution and plasticity. *Nat Rev Genet*. 13(3): 163–174.

Montagné N, Desdevises Y, Soyez D, Toullec J-Y. 2010. Molecular evolution of the crustacean hyperglycemic hormone family in ecdysozoans. *BMC Evol Biol*. 10:62.

Nagaraju GPC. 2011. Reproductive regulators in decapod crustaceans: an overview. *J Exp Biol*. 214(Pt 1): 3–16.

Okuno A, Hasegawa Y, Ohira T, Katakura Y, Nagasawa H. 1999. Characterization and cDNA cloning of androgenic gland hormone of the terrestrial isopod *Armadillidium vulgare*. *Biochem Biophys Res Commun*. 264(2): 419–423.

Pichon S, Bouchon D, Liu C, Chen L, Garrett RA, Grève P. 2012. The expression of one ankyrin pk2 allele of the WO prophage is correlated with the *Wolbachia* feminizing effect in isopods. *BMC Microbiol*. 12:55.

Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. *Bioinformatics* 21(1 Suppl): i351–i358.

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. 2005. InterProScan: protein domains identifier. *Nucleic Acids Res*. 33(Web Server): W116–W120.

Rigaud T, Juchault P, Mocquard J-P. 1997. The evolution of sex determination in isopod crustaceans. *Bioessays* 19(5): 409–416.

Rigaud T, Souty Grosset C, Raimond R, Mocquard JP, Juchault P. 1991. Feminizing endocytobiosis in the terrestrial crustacean *Armadillidium vulgare* Latr. (Isopoda): Recent acquisitions. *Endocytobiosis Cell Res.* 7:259–273.

Rodrigues N, Studer T, Dufresnes C, Perrin N. 2018. Sex-chromosome recombination in common frogs brings water to the fountain-of-youth. *Mol Biol Evol.* 35(4): 942–948.

Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, Chiari Y, Dernat R, Duret L, Faivre N, et al. 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* 515(7526): 261–263.

Salmela L, Rivals E. 2014. LoRDEC: accurate and efficient long read error correction. *Bioinformatics* 30(24): 3506–3514.

Schmalfuss H. 2003. World catalog of terrestrial isopods (Isopoda: Oniscidea). *Stuttg Beitr Nat A Biol.* 638: 341.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27(6): 863–864.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19): 3210–3212.

Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31.

Sohn J, Nam J-W. 2018. The present and future of de novo whole-genome assembly. *Brief Bioinformatics* 19(1): 23–40.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21): 2688–2690.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* 34(Web Server): W435–W439.

Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinform*. 25: 4.10.1–4.10.14.

Thézé J, Leclercq S, Moumen B, Cordaux R, Gilbert C. 2014. Remarkable diversity of endogenous viruses in a crustacean genome. *Genome Biol Evol*. 6(8): 2129–2140.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3—new capabilities and interfaces. *Nucleic Acids Res*. 40(15): e115.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From FastQ data to high-confidence variant calls: the Genome Analysis ToolKit best practices pipeline. In: Bateman A, Pearson WR, Stein LD, Stormo GD, Yates JR, editors. Current protocols in bioinformatics. Hoboken (NJ): John Wiley & Sons, Inc. p. 11.10.1–11.10.33.

Verne S, Johnson M, Bouchon D, Grandjean F. 2012. Effects of parasitic sex-ratio distorters on host genetic structure in the *Armadillidium vulgare*–*Wolbachia* association: effect of *Wolbachia* on mtDNA in *A. vulgare*. *J Evol Biol*. 25(2): 264–276.

Vítková M, Král J, Traut W, Zrzavý J, Marec F. 2005. The evolutionary origin of insect telomeric repeats, (TTAGG) n. *Chromosome Res.* 13(2): 145–156.

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. 2017. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33(14): 2202–2204.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9(11): e112963.

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol*. 35(3): 543–548.

Webster SG, Keller R, Dircksen H. 2012. The CHH-superfamily of multifunctional peptide hormones controlling crustacean metabolism, osmoregulation, moulting, and reproduction. *Gen Comp Endocrinol*. 175(2): 217–233.

Xue W, Li J-T, Zhu Y-P, Hou G-Y, Kong X-F, Kuang Y-Y, Sun X-W. 2013. L_RNA_scaffolder: scaffolding genomes with transcripts. *BMC Genomics*. 14:604.

Ye C, Hill CM, Wu S, Ruan J, Ma Z. 2016. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci Rep.* 6: 31900.

Ye C, Ma ZS. 2016. Sparc: a sparsity-based consensus algorithm for long erroneous sequencing reads. *PeerJ* 4:e2016.

Ye C, Ma ZS, Cannon CH, Pop M, Douglas WY. 2012. Exploiting sparseness in de novo genome assembly. *BMC Bioinformatics* 13(Suppl 6): S1.