Characterization of a Sex-Determining Region and Its Genomic Context via Statistical Estimates of Haplotype Frequencies in Daughters and Sons Sequenced in Pools

Richard Cordaux, Mohamed Amine Chebbi, Isabelle Giraud, David R.J. Pleydell, Jean Peccoud

Characterization of a Sex-Determining Region and Its Genomic Context via Statistical Estimates of Haplotype Frequencies in Daughters and Sons Sequenced in Pools

Richard Cordaux[1], Mohamed Amine Chebbi[1], Isabelle Giraud[1], David Richard John Pleydell[2], and Jean Peccoud[1,*]

[1]Laboratoire Écologie et Biologie des Interactions, Équipe Écologie Évolution Symbiose, UMR CNRS 7267, Université de Poitiers, France
[2]UMR Animal, Santé, Territoires, Risques et Écosystèmes, INRAE, CIRAD, Montpellier SupAgro, Université de Montpellier, France

*Corresponding author: E-mail: jeanpeccoud@gmail.com.

Accepted: 25 May 2021

Abstract

Sex chromosomes are generally derived from a pair of autosomes that have acquired a locus controlling sex. Sex chromosomes may evolve reduced recombination around this locus and undergo a long process of molecular divergence. At that point, the original loci controlling sex may be difficult to pinpoint. This difficulty has affected many model species from mammals to birds to flies, which present highly diverged sex chromosomes. Identifying sex-controlling loci is easier in species with molecularly similar sex chromosomes. Here we aimed at pinpointing the sex-determining region (SDR) of *Armadillidium vulgare*, a terrestrial isopod with female heterogamety (ZW females and ZZ males) and whose sex chromosomes appear to show low genetic divergence. To locate the SDR, we assessed single-nucleotide polymorphism (SNP) allele frequencies in F1 daughters and sons sequenced in pools (pool-seq) in several families. We developed a Bayesian method that uses the SNP genotypes of individually sequenced parents and pool-seq data from F1 siblings to estimate the genetic distance between a given genomic region (contig) and the SDR. This allowed us to assign more than 43 Mb of contigs to sex chromosomes, and to demonstrate extensive recombination and very low divergence between these chromosomes. By taking advantage of multiple F1 families, we delineated a very short genomic region (~65 kb) that presented no evidence of recombination with the SDR. In this short genomic region, the comparison of sequencing depths between sexes highlighted female-specific genes that have undergone recent duplication, and which may be involved in sex determination in *A. vulgare*.

Key words: sex chromosomes, terrestrial isopods, pool-seq, recombination, SNP, gene duplication.

Significance

Identifying loci controlling the sex of individuals (male or female) has remained difficult due to high levels of divergence between sex chromosomes in many species. We attempt this using sequenced pools of same-sexed individuals of the common pillbug (*Armadillidium vulgare*), a species thought to present relatively undifferentiated sex chromosomes. A statistical method designed specifically for such data enabled us to confirm a very low level of divergence between sex chromosomes and to identify a short genomic region that may contain the sex-determining locus in *A. vulgare*.

Introduction

The existence of males and females (gonochorism) constitutes a phenotypic variation found in many taxa which exerts a profound impact on their evolution. Despite gonochorism being both common and ancient, the mechanisms initiating the developmental cascade towards distinct male and female

phenotypes appear to be highly variable (Bachtrog et al. 2014; Beukeboom and Perrin 2014). In some species, sex is solely or partially determined by environmental factors, such as temperature (Merchant-Larios and Diaz-Hernandez 2013) and social interactions (Brante et al. 2016). In many other species, sex is determined entirely by genotype (reviewed in Bachtrog et al. [2014]; Beukeboom and Perrin [2014]).

The most well-known categories of sex-determining genotypes are XX/XY, in which males are heterozygotes (also called heterogametic, as in therian mammals and *Drosophila*), and ZZ/ZW, in which females are heterogametic (as in birds and lepidopterans). Other sex-determining genotypes are known, see, for example, Bachtrog et al. (2014). Chromosomes carrying a locus (i.e., a unit of heredity) whose variants determine sex can be defined as "sex chromosomes." Under this definition, sex chromosomes are not required to be a pair of chromosomes that appear morphologically different (from each other) under the microscope—a property referred to as heteromorphism. In fact, most sex chromosomes must initially be homomorphic, since they typically evolve from a pair of autosomes that have acquired a sex determining locus (Muller 1918; Wright et al. 2016; Furman et al. 2020). However, in contrast to homologous autosomes, sex chromosomes may diverge under the possible influence of alleles with sex-antagonistic effects, which favor reduced crossing over rates around the sex-determining locus (Rice 1987; Bergero and Charlesworth 2009; Bachtrog et al. 2014; Wright et al. 2016, 2017). As a result, the chromosomal region whose alleles associate with the sex phenotype (hereafter referred to as the sex-determining region or SDR) tends to increase in size. The reduced rates of recombination lead to divergence of the two sex chromosomes, due to an inability to efficiently purge deleterious mutations (Bergero and Charlesworth 2009; Bachtrog 2013). Through this divergence process, sex chromosomes may become visually recognizable in a karyotype. This heteromorphism helps the identification of genetic sex-determining systems (reviewed in Bachtrog et al. [2014]).

The ease of identifying pairs of heteromorphic sex chromosomes contrasts however with the difficulty of locating the original sex-determining locus when it is part of a large SDR. The fact that most model organisms—including mammals, birds, and fruit flies—possess large SDRs may partly explain why sex-determining genes have been identified in relatively few taxa (reviewed in Beukeboom and Perrin [2014]) compared with the large diversity of taxa possessing sex chromosomes.

In taxa where sex chromosomes undergo rapid turnover, such as teleost fishes (Mank and Avise 2009), sex chromosomes are evolutionarily young, hence SDRs are likely to be short. In several taxa, short SDRs have facilitated the identification of sex-determining genes (Kamiya et al. 2012; Akagi et al. 2014), some of which differ among species of the same genus (Matsuda et al. 2002; Nanda et al. 2002). These taxa therefore emerge as useful models to study the appearance and early evolution of sex chromosomes (Charlesworth et al. 2005) and to learn about the diversity of genes and mechanisms leading to the development of sex phenotypes.

Terrestrial isopods, also known as woodlice or pillbugs, provide an interesting group of organisms for studying sex-determining loci. These crustaceans appear to have undergone multiple evolutionary transitions between XY systems and ZW systems (Becking et al. 2017), which implies that the sex chromosomes of several isopod species may be evolutionarily young. The isopod for which sex determination has been studied the most is the common pillbug *Armadillidium vulgare* (Cordaux et al. 2011). Although the sex chromosomes are visually undistinguishable among the 27 chromosome pairs of the *A. vulgare* genome (Artault 1977), crossing experiments using genetic females masculinized via hormone treatments (Juchault and Legrand 1972) have demonstrated that this species has female heterogamety. Homomorphism of the *A. vulgare* sex chromosomes is also consistent with the reported viability and fertility of WW individuals generated through these crossing experiments. The ~1.72 Gb genome of an *A. vulgare* female has recently been assembled (Chebbi et al. 2019). Although none of the 43,541 assembled contigs and scaffolds are anchored to a chromosome, comparison of sequencing depths from data obtained from ten males and ten females, combined with a *k*-mer-based approach, located 27 contigs representing ~673 kb of W-specific sequences (sequences that are dissimilar to those found in ZZ males) (Chebbi et al. 2019). It is unknown whether W-specific sequences represent large insertions (such as gene duplications or transposable element insertions) in the W-linked haplotype of the SDR, large deletions in the Z-linked haplotype and/or the accumulation of smaller mutations between the Z and W haplotypes. At any rate, the total W-specific sequence is much shorter than the average chromosome size in *A. vulgare* (~1.72 Gb/27, or ~60 Mb), which is consistent with homomorphism in the sex chromosomes. However, the *A. vulgare* SDR might not be restricted to a region of complete absence of similarity between the Z and W alleles—in some species, sex has been shown to be controlled by a single-nucleotide polymorphism (SNP) (Kamiya et al. 2012), a hypothesis that cannot be excluded in *A. vulgare*.

In situations such as these, where sex chromosomes present very low molecular divergence and undergo crossing overs, methods based on SNPs can be useful for locating SDRs and genetically linked loci. Clearly, SNP data can be used to locate loci for which individuals of a given sex all have expected genotypes, according to the type of heterogamety at hand (e.g., loci with SNPs that are heterozygous in all ZW daughters and homozygous in all ZZ sons). However, when no information on the location of the SDR is available (e.g., in the absence of a genetic map) and if no candidate loci for sex determination are suspected, the whole genome must be analyzed, and hence re-sequenced or scanned. Doing so

can be very costly if whole-genome sequencing is undertaken on many individuals, especially if a large sample size is required for reliable statistical inference. Such considerations often motivate the use of techniques permitting transcriptome sequencing or partial genome sequencing (reviewed in Palmer et al. [2019]), such as Restriction-site Associated DNA marker (RAD) sequencing (Baird et al. 2008). However, these approaches reduce sequencing costs at the expense of an increase in DNA library preparation costs. An alternative strategy is to pool the DNA from multiple individuals prior to sequencing (Futschik and Schlötterer 2010), a method referred to as "pool-seq." Pool-seq can drastically cut DNA library preparation costs, and sequencing costs too given that the sequencing effort per individual is generally lower than in the case of individual whole-genome sequencing. Pool-seq substitutes the obtention of individual genotypes with estimates of allele frequencies within pools. The main drawback is that allele frequencies among the pooled individuals are inferred from allele frequencies among sequenced fragments ("reads") covering a SNP. This inference adds a degree of uncertainty that diminishes with increased sequencing effort (Gautier et al. 2013). Since each SNP is analyzed separately, somewhat arbitrary thresholds on estimated frequencies among reads and on sequencing depths have been used when determining whether individual SNPs are associated with the sex phenotype (Pan et al. 2019). These approximations reflect a lack of a probabilistic framework for analyzing pool-seq data in the context of genetic sex determination. Such methods are indeed only available for the analysis of individual genomes or transcriptomes (Gautier 2014; Muyle et al. 2016). In addition, these methods assign a given locus to discrete segregation types (sex-linked, autosomal), which limits their ability to characterize genomic regions that are genetically close (in terms of centimorgans), but not fully linked, to the SDR. Unfortunately, if a pool-seq approach is used, the absence of individual genotypes prevents the construction of a genetic map, thereby limiting knowledge about the genomic context of the SDR.

Here, we developed a statistical approach that overcomes these limitations and is based on the individual genotyping of parents and pooled sequencing of progeny from several crosses. This method allowed us to identify a short genomic region that likely contains the SDR of *A. vulgare*. Beyond the SDR, our approach allowed us to assign more than 43 Mb of contigs to sex chromosomes, even though these chromosomes appeared no different from autosomes with respect to molecular divergence and showed uniform recombination rates.

Materials and Methods

General Approach

We infer the genetic distance between a locus and the SDR from inheritance patterns among SNP alleles in controlled crosses (fig. 1). In our experiment, siblings are not individually genotyped, but their whole genome is sequenced after pooling DNA samples from siblings of the same sex. This allows us to estimate allele frequencies at inherited SNPs, in daughters and sons, from the frequencies of reads carrying either allele at a SNP. These estimates in turn allow inference regarding the genetic distance between a focal locus and the SDR, provided this locus has similar (alignable) sequences between the sex chromosomes.

Our analysis relies on biallelic SNPs which, like the SDR, are heterozygous in the mother and homozygous in the father, hereafter called "informative SNPs." Genotypes at SNPs that are heterozygous in the father arise from random selection of paternal gametes, which do not determine sex. The W allele of the SDR is transmitted to all daughters (hence has a frequency of 0.5 among them, given that they also inherit the Z allele) and to no sons. A SNP showing these patterns would not have recombined with the SDR during crosses, whereas another SNP that only slightly deviates from these patterns may be genetically close to the SDR on the sex chromosome pair. These considerations form the basis of our methodology.

Crosses, Sequencing, and Mapping

We applied our approach to three *A. vulgare* lines: WXa, ZM, and BF (table 1), which have been shown to harbor the same ZW locus (Chebbi et al. 2019). For each line, a single virgin female was crossed with a single male until it showed evidence for gravidity and then isolated to lay progeny. DNA was extracted from gonads, heads and legs of ten descendants of each sex with the Qiagen 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 heritable elements controlling sex other than the ZW locus—*Wolbachia* endosymbionts and the *f* element (Leclercq et al. 2016)—was confirmed in all samples by PCR, as described previously (Leclercq et al. 2016). DNA concentration was estimated for each sample by Qubit fluorometric quantification, to enable pooling DNA samples in equimolar proportions. DNA samples from ten same-sex individuals constituted a pool containing 7 µg of DNA. Each pool was sequenced on an Illumina HiSeq2500 platform (125-bp paired ends) by Beckman Coulter Genomics. We aimed at a sequencing depth of 30× per pool to ensure that most parts of the 20 chromosome doses (i.e., from ten diploid individuals) in the pools were sequenced.

To identify informative SNPs, the whole genome of each parent was sequenced individually. The DNA of parents from lines WXa and ZM was extracted as described above and sequenced on an Illumina HiSeqX platform (150-bp paired ends) by Génome Québec. To enable reliable SNP genotyping, we targeted an average sequencing depth of 30× per parent. However, technical reasons unrelated to our approach prevented the sequencing of parents from line BF.

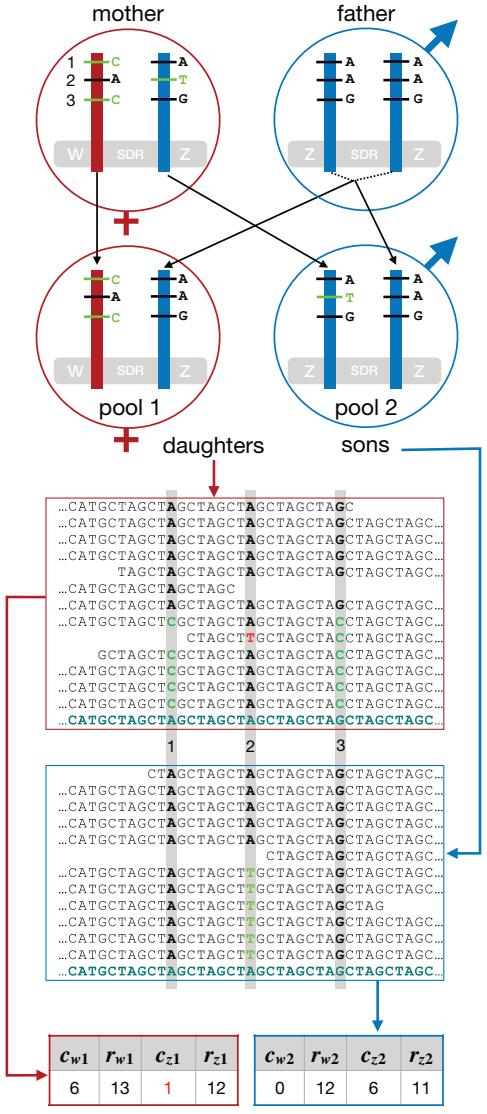

Fig. 1.—Locating the ZW SDR via a cross where the genomes of several F1 siblings per sex are sequenced in pools and parental genomes sequenced individually. Red/blue rods represent chromosomes carrying the W/Z alleles. They possess informative SNPs (heterozygous for mothers and homozygous for fathers) shown as horizontal segments. For each SNP, the maternal allele (as defined in the "Materials and Methods" section) appears in green. In this example, no crossing over occurred between the SNPs and the SDR. The mapping of reads obtained from the pools shows sequences (reads) aligned on the reference genome (bottom sequence), with the three informative SNPs outlined. A sequencing error is shown in red. Tables at the bottom show the values of the variables used to estimate the frequency of the W-linked haplotype (C-A-C) in each pool, based on the mapped reads. See "Statistical Estimation of Haplotype Frequencies" section for the definition of these variables.

Sequencing reads were trimmed, to remove low-quality parts, using trimmomatic version 0.33 (Bolger et al. 2014). For each F1 pool and parent, trimmed reads were aligned on the female reference *A. vulgare* genome (Chebbi et al. 2019) using bwa_mem (Li 2013) with default settings. In the resulting alignment (bam) file, reads sequenced from the same original DNA fragments (PCR or optical duplicates) were flagged by picardtools MarkDuplicates version 2.12.0 (http://broadinstitute.github.io/picard/, last accessed June 9, 2020). Reads containing indels were realigned on the reference genome using GATK's IndelRealigner.

To establish the whole-genome genotype of each parent, we followed the GATK best practices (Van der Auwera et al. 2013) as described in Chebbi et al. (2019). This involved recalibrating base quality scores of mapped reads to reduce the risk of considering sequencing errors as variants, followed by SNP genotype calling with HaplotypeCaller. Genotyping was performed independently on each parent, recording all positions in a gvcf file. The four gvcf files were merged, using GenotypeGvcf, into a single vcf file, discarding putative SNPs not passing GATK's built-in quality check. We used this file to select informative SNPs, excluding positions with indels, lack of sequence data in any individual or with more than two alleles.

Statistical Estimation of Haplotype Frequencies

Because frequencies at SNPs of the same locus (here, a contig) are totally interdependent in the absence of crossing over within the locus, we did not analyze each SNP independently. Instead, we developed a statistical method to estimate the frequency of a whole haplotype linking SNP alleles. Doing so accounts for interdependance within each locus and should therefore provide greater accuracy.

In the mother, a locus on the sex chromosomes comprises a haplotype that is linked to (i.e., on the same chromosome as) the Z allele (which we call the Z-linked haplotype) and another haplotype linked to the W allele (W-linked haplotype). The phasing (reconstruction) of these haplotypes is addressed in the next section.

In a given F1 pool i, of n_i individuals born from n_i oocytes, we denote the unknown frequency of the W-linked haplotype as f, and the set of potential values of f as $\boldsymbol{F}_i = \{(0\ldots n_i)/(2n_i)\}$. The proportion of oocytes that underwent recombination between the locus and the SDR, which is the distance to the SDR in Morgans, is $1–2f$ for daughters and $2f$ for sons.

To estimate f, we count sequenced DNA fragments (e.g., read pairs) that can be attributed to the W-linked haplotype. We also count fragments that can be attributed to the Z-linked maternal haplotype as its frequency in the pool is exactly 0.5—f, the other half of the $2n_i$ chromosomes of the pool being inherited from the father. For these counts, we first designate as "maternal allele" the allele of an informative SNP that is only carried by the mother (fig. 1). We then let c_{wi} denote the number of DNA fragments sequenced from pool i and carrying the maternal alleles from the W-linked haplotype. In figure 1, these alleles belong to the first and third SNPs, as their maternal alleles are linked to the W haplotype. We let

Table 1
Characteristics of the Three *Armadillidium vulgare* Lines Used in This Study

Line/Family	WXa/1591	ZM/544	BF/2875
Source location	Helsingør, Denmark	Heraklion, Greece	Nice, France
Collection date	1982	1989	1967
Sequence Read Archive accession numbers:			
• Mother (ZW)	SRR13605041	SRR13605043	Not sequenced
• Father (ZZ)	SRR13605040	SRR13605042	Not sequenced
• Pool of ten daughters (ZW)	SRR13582783	SRR13582781	SRR8238986
• Pool of ten sons (ZZ)	SRR13582782	SRR13582780	SRR8238987

r_{wi} represent the number of fragments from pool i carrying either maternal or paternal alleles for the corresponding SNPs. We finally let c_{zi} denote the number of fragments carrying the maternal alleles from the Z-linked haplotype, and r_{zi} represent the number of fragments carrying alleles from either parent at the corresponding SNPs (the second SNP in fig. 1).

Given these specifications, we can express the posterior probability of the W-linked haplotype frequency given the observed data via Bayes' theorem:

$$P(f|\boldsymbol{cr}_i) = \frac{P(f).P(c_{wi}, c_{zi}|f, r_{wi}, r_{zi})}{P(c_{wi}, c_{zi}|r_{1wi}, r_{zi})}, \quad (1)$$

where $\boldsymbol{cr}_i$ refers to the set of variables $\{c_{wi}, r_{wi}, c_{zi}, r_{zi}\}$.

To specify $P(c_{wi}, c_{zi} \mid f, r_{wi}, r_{zi})$, we assume c_{wi} and c_{zi} to be independent since the values of these variables arise from counting different DNA fragments (fragments from the W-linked haplotype for c_{wi} and from the Z-linked haplotype for c_{zi}). Hence,

$$P(c_{wi}, c_{zi}|f, r_{wi}, r_{zi}) = P(c_{wi}|f, r_{wi})P(c_{zi}|f, r_{zi}). \quad (2)$$

We then consider c_{wi} as the realization of a *Binomial*(r_{wi}, f) distribution, that is, that the c_{wi} read pairs carrying the maternal allele from the W maternal chromosome arise with frequency f from r_{wi} independent draws among DNA molecules containing the W-linked haplotype in the pool. Following the same principle, c_{zi} is the realization of a *Binomial*(r_{zi}, 0.5—f) distribution.

For a locus that segregates perfectly with the SDR, c_{zi} should be zero in daughters (fig. 1). However, a nucleotide indicating the maternal allele in a read may result from an "error." Potential errors include mutations between parents and offspring, in vitro mutations and sequencing or mapping errors. An error causing c_{zi} to be positive would nullify the posterior probability that $f = 0.5$, that is, the probability that no crossing over occurred between the locus and the SDR. To avoid this, we introduce a constant ε to represent the probability that a maternal allele appears in a read due to error. The total probability that a read carries the maternal allele from the W-linked haplotype therefore becomes $(1 - \varepsilon)(1/2 - f) + \varepsilon = (1 + \varepsilon)/2 + \varepsilon f - f$.

We do not apply this correction to c_{wi} in daughters because a maternal allele linked to the W allele should have a frequency of 0.5 for a locus linked to the SDR, such that errors leading to the detection of maternal alleles are compensated by error leading to the detection of nonmaternal alleles.

From the above considerations, for daughters we have:

$$P(c_{wi}|f, r_{wi}) = \binom{r_{wi}}{c_{wi}} f^{c_{wi}}(1 - f)^{r_{wi} - c_{wi}}$$

and

$$P(c_{zi}|f, r_{zi}) = \binom{r_{zi}}{c_{zi}} \left(\frac{1 + \varepsilon}{2} + \varepsilon f - f\right)^{c_{zi}} \left(\frac{1 - \varepsilon}{2} - \varepsilon f + f\right)^{r_{zi} - c_{zi}}.$$

For a pool of sons, c_{wi} should be zero in the absence of recombination with the SDR (fig. 1), hence the probability that a read carries a maternal allele from the W-linked haplotype becomes $(1 - \varepsilon)f + \varepsilon = \phi - \varepsilon f + \varepsilon$. Hence, for sons:

$$P(c_{wi}|f, r_{wi}) = \binom{r_{wi}}{c_{wi}} (f - \varepsilon f + \varepsilon)^{c_{wi}}(1 - f + \varepsilon f - \varepsilon)^{r_{wi} - c_{wi}}$$

and

$$P(c_{zi}|f, r_{zi}) = \binom{r_{zi}}{c_{zi}} (0.5 - f)^{c_{zi}}(0.5 + f)^{r_{zi} - c_{zi}}.$$

The marginal probability of the maternal allele read counts, $P(c_{wi}, c_{zi} \mid r_{wi}, r_{zi})$, integrates over all the values that f can take, hence

$$\begin{aligned} P(c_{wi}, c_{zi}|r_{wi}, r_{zi}) &= \sum_{f \in F_i} P(f).P(c_{wi}, c_{zi}|f, r_{wi}, r_{zi}) \\ &= \sum_{f \in F_i} P(f).P(c_{wi}|f, r_{wi})P(c_{zi}|f, r_{zi}). \end{aligned}$$

We use a discrete uniform prior for f, hence $P(f) = 1/(n_i + 1)\ \forall\ f \in \boldsymbol{F}_i$. Although a prior based on *Binomial*(n_i, 0.5) would more accurately represent the inheritance of maternal haplotypes for most loci, such a prior would not be suitable for loci at less than 50 cM to the SDR, the frequency of which is unknown. In particular, a binomial prior assumes that f is quite unlikely to equal 0.5 (with a probability of 0.5^{n_i}), increasing the risk of false negatives when it comes to the selection of loci that are completely linked to the SDR. We would rather include false positives in our selection of candidates, as this selection is only a first step in the search for the sex-determining locus.

Substituting these terms into equation (1) yields the following in daughters, after cancellation of terms present in both the numerator and the denominator:

$$P(f|\boldsymbol{cr}_i) = \frac{f^{c_{wi}}(1-f)^{r_{wi}-c_{wi}}\left(\frac{1+\varepsilon}{2}+\varepsilon f-f\right)^{c_{zi}}\left(\frac{1-\varepsilon}{2}-\varepsilon f+f\right)^{r_{zi}-c_{zi}}}{\sum_{\varphi\in F_i}\varphi^{c_{wi}}(1-\varphi)^{r_{wi}-c_{wi}}\left(\frac{1+\varepsilon}{2}+\varepsilon\varphi-\varphi\right)^{c_{zi}}\left(\frac{1-\varepsilon}{2}-\varepsilon\varphi+\varphi\right)^{r_{zi}-c_{zi}}}.$$

In sons, the equivalent equation is

$$P(f|\boldsymbol{cr}_i) = \frac{(f-\varepsilon f+\varepsilon)^{c_{wi}}(1-f+\varepsilon f-\varepsilon)^{r_{wi}-c_{wi}}(0.5-f)^{c_{zi}}(0.5+f)^{r_{zi}-c_{zi}}}{\sum_{\varphi\in F_i}(\varphi-\varepsilon\varphi+\varepsilon)^{c_{wi}}(1-\varphi+\varepsilon\varphi-\varepsilon)^{r_{wi}-c_{wi}}(0.5-\varphi)^{c_{zi}}(0.5+\varphi)^{r_{zi}-c_{zi}}}.$$

These posterior probabilities are used to estimate the expected number of recombination events, denoted here as n_{rec}, that occurred between a given contig and the SDR in the oocytes of several pools. This is done by calculating the sum of the expected number of recombination events over pools as follows:

$$n_{rec} = \sum_{i\in D}\sum_{f\in F_i} n_i(1-2f)P(f|\boldsymbol{cr}_i) + \sum_{i\in S}\sum_{f\in F_i} n_i 2fP(f|\boldsymbol{cr}_i), \quad (3)$$

where D and S are the sets of indices for daughter and son pools, respectively. This formula accounts for variable levels of uncertainty in the true value of f among pools. Note that although the true number of recombination events is a discrete random variable, its expected value n_{rec} is a weighted average and is therefore a continuous variable.

Estimation of Recombination with the Sex-Determining Locus during Crosses

To implement this approach, we used a custom R script (R Development Core Team 2020) that scans each F1 bam file via samtools version 1.10 (Li et al. 2009) and retrieves the base carried by each read at informative SNPs, associated with a unique read-pair identifier. Read pairs marked as duplicates were ignored as well as secondary alignments and those with mapping quality score <20. For each pool and SNP, we counted reads carrying the parental alleles. Reads carrying other alleles were ignored.

To phase Z- and W-linked haplotypes, we used the fact that a maternal allele that is linked to the W allele in the mother should be more frequent in daughters than in sons (fig. 1). The opposite is true for a maternal allele that is linked to the Z allele. We thus attributed the maternal allele of a SNP to the W-linked haplotype whenever it was more frequent in daughters than in sons. Otherwise, the maternal allele was assigned to the Z-linked haplotype. If the maternal allele was equifrequent in both sexes, we attributed haplotypes at random. The frequency of the maternal allele in the pool was estimated by the proportion of reads carrying this allele. Simulations showed that the accuracy of our haplotype phasing method is very high for contigs that are genetically close to the SDR, and decreases with the genetic distance from the SDR, leading to an underestimate of the true number of recombination events (supplementary text, fig. S1, Supplementary Material online). We investigated the use of haplotype phasing tools based on read pair overlaps (Martin et al. 2016; Edge et al. 2017) to help the phasing of maternal haplotypes, but phasing errors made their use more detrimental than helpful (supplementary text, Supplementary Material online).

For each contig in each pool i, we established variables of the $\boldsymbol{cr}_i$ set by counting read pairs according to the definitions for these variables. We then computed $P(f \mid \boldsymbol{cr}_i)$ for every possible value of f. We set the error probability ε at 0.01, which is higher than the typical Illumina technology sequencing error rate. Using results from the four pools of the WXa and ZM lines (for which we sequenced the parents), we estimated the number of recombination events between the contig and the SDR, n_{rec}.

For these computations, and all subsequent analyses, we excluded any informative SNP that failed to pass the following criteria, which we applied independently for both families. First, genotyping quality in the mother (determined by GATK's haplotype caller) had to be higher than 10 and that of the father higher than 40 (the presence of the rarer maternal allele in the F1 allowed us to be less restrictive on the mother's genotype quality, while we wanted to ensure that the father was not heterozygous). Second, at least one read had to carry either parental allele in each F1 pool, and the total number of reads carrying either allele in both F1 pools combined had to not exceed the 95% quantile of this variable. We reasoned that excessive sequencing depth may reflect the alignment of reads from several loci on the same genomic region, due to paralog collapsing during genome assembly. Third, the maternal allele had to be present in at least one F1 read and both parental alleles had to be present in at least 75% of the F1 reads covering the SNP (both pools combined).

Preliminary results revealed a problem in which some contigs showed very low probabilities of perfect segregation with the SDR in a given pool i. If i is a pool of daughters, this probability is $P(f=0.5|\ \boldsymbol{cr}_i)$. It can be greatly reduced by rare SNPs whose maternal alleles were assigned to the Z-linked haplotype, leading to aberrant c_{zi}/r_{zi} ratios (in sons, the equivalent problem is due to alleles assigned to the W-linked haplotype, but we do not detail it here for the sake of brevity). Accurate estimation of $P(f=0.5 \mid \boldsymbol{cr}_i)$ is critical as we use it to select contigs that may contain the SDR (see "Localization of Genomic Regions That May Contain the SDR" section). We attribute the negative influence of "suspicious" SNPs on $P(f=0.5 \mid \boldsymbol{cr}_i)$ to mapping or assembly errors (c_{zi} being much too high to result from sequencing errors). These errors would lead to reads from different loci aligning on the same genomic region. Hence, the apparent variation between reads

would not represent allelic variation (SNPs), but another type of variation. To locate these suspicious SNPs, we computed $P(f = 0.5 | \boldsymbol{cr}_i)$ on each individual SNP as if it constituted its own haplotype, and ignored SNPs yielding much lower posterior probabilities than the rest of the SNPs of each contig (see supplementary text, fig. S2, Supplementary Material online).

Contig Assignment to Sex Chromosomes and Analysis of Recombination

We used our estimates of the number of oocytes that underwent recombination between target contigs and the SDR during both crosses (n_{rec}, eq. 3) to isolate contigs that are significantly closer to the SDR than expected assuming an autosomal location. To account for uncertainty in n_{rec} and the approximations of our method (in particular, haplotype phasing), we compared the observed values of n_{rec} with those obtained by simulating sequencing data in the F1 pools. These simulations used the actual genomic positions of the informative SNPs and the identifiers of reads covering these SNPs, and only changed the bases that reads carried at SNPs to reflect a given genetic distance to the SDR.

The simulations were performed using the following procedure, which we applied to every contig. First, we assigned the contig a genetic distance (in Morgans) to the SDR, which we call d. For each informative SNP in each family, the maternal allele was randomly linked to the Z or to the W allele with equal probability.

We then applied the following to each of pool i of the family. We defined as n_{Zi} the number of chromosomes carrying haplotype Z in the pool of $2n_i$ chromosomes. To simulate linkage to the SDR, n_{Zi} was sampled from $Binomial(n_i, d)$ if the pool contained daughters or from $Binomial(n_i, 1 - d)$ if the pool contained sons. To simulate the sequencing of the maternal haplotypes, each read was randomly attributed to maternal or paternal DNA with the same probability. Then, each read of maternal origin was attributed to the Z-linked haplotype with probability n_{Zi}/n_i or to the W-linked haplotype otherwise. At each SNP, each read was set to carry the maternal allele if the read and the maternal allele of this SNP were both attributed to the same haplotype (Z- or W-linked). Otherwise, the read was set to carry the alternative allele. Based on these artificial reads, we phased maternal haplotypes and computed $P(f | \boldsymbol{cr}_i)$ as for the real data. We repeated the procedure 1,000 times for every contig, with d set to 0.5 to simulate autosomal contigs. If, for any contig, the value of n_{rec} obtained from the real data was lower than the 1/1,000th quantile of values obtained from simulations, we considered the contig as located on the sex chromosomes with a 1/1,000 risk of false positive.

We also performed simulations to investigate a potential reduction of recombination rate near the SDR, which may evolve during sex chromosome divergence. In an approach analogous to building a Marey map (Chakravarti 1991), we plotted the cumulated length of contigs (a proxy for physical distance on chromosomes) as a function of the inferred genetic distance to the SDR (derived from n_{rec}). In order to compare this curve with expectations under uniform recombination rates along the sex chromosomes, we created an envelope as follows. Uniformity in recombination rates was ensured by attributing every contig of the genome a value of d sampled from $Uniform(0, 0.5)$. We assigned these simulated contigs to sex chromosomes as we did for real contigs. After discarding simulated contigs not assigned to sex chromosomes, we randomly sampled a number of simulated contigs, ensuring that their total length was as close as possible to the total length of observed contigs assigned to sex chromosomes (which we call L). To do so, we created a table of randomly sorted contigs with two columns specifying their length and inferred distance to the SDR. In a third column, we computed the cumulated length of contigs from the start of the table, and we located the row for which the absolute value of the difference between the cumulated length and L was the smallest. We then discarded all the contigs after that row. We repeated this procedure 1,000 times to build an envelope for the cumulated length of contigs as a function of the inferred genetic distance to the SDR.

Investigation of Heterozygous SNP Density

The SDR and nearby genomic regions are predicted to show increased allelic divergence compared with autosomal regions due to balancing selection potentially combined with reduced recombination rates. We investigated this hypothesis by measuring the density of heterozygous SNPs in the two individually genotyped mothers. For each mother, we ignored SNPs 1) of genotype quality <40, 2) of sequencing depth <5, 3) or of sequencing depth higher than the 95% quantile for the genotyped individual (only considering positions reported in the vcf file). For each contig, we counted the number of unique heterozygous positions passing these filters when considering both mothers combined. To estimate SNP density given the sequencing effort, we recorded the number of unique contig positions belonging to the aforementioned range of sequencing depths for each contig, again combining both mothers. Sequencing depth was computed with samtools, excluding duplicate reads, secondary alignments, alignments with mapping quality zero and bases with PHRED score <10, to mimic the parameters used by the SNP caller.

We then analyzed how heterozygous SNP density varied according to the inferred genetic distance to the SDR (n_{rec}). For these analyses, we opted to ignore contigs for which n_{rec} could not be inferred with sufficient certainty. We did so by discarding contigs for which the highest posterior probability of f, $\max(P(f | \boldsymbol{cr}_i))$, was lower than 0.5 in any pool i. Doing so considers that contigs with fewer informative SNPs, hence with lower heterozygous SNP density in females on average,

are less likely to be assigned to sex chromosomes due to reduced statistical power. This bias might lead to a spurious correlation between female heterozygosity and assignment to sex chromosomes. Selecting contigs for which that data allowed estimating n_{rec} with a certain level of confidence should mitigate this bias.

Localization of Genomic Regions That May Contain the SDR

We tested two criteria for identifying genomic regions that may contain the SDR. The first criterion identified contigs showing little evidence for recombination with the SDR during our crosses. We based this criterion on the posterior probability of absence of recombination, which we multiplied across pools for each contig as follows:

$$\prod p = \prod_{i \in D} P(f = 0.5 | \boldsymbol{cr}_i) \times \prod_{i \in S} P(f = 0 | \boldsymbol{cr}_i). \quad (4)$$

We considered that contigs for which $\prod p$ exceeded 0.5 were unlikely to have recombined with the SDR during the crosses. For the second criterion, we considered that the contigs which were more likely to perfectly segregate with the SDR were those for which n_{rec} was closer to zero than to one ($n_{rec} < 0.5$). This second criterion was fulfilled by all but three of the contigs identified by the first criterion (109 vs. 112) and did not identify any additional contig. We opted for the more inclusive criterion.

We then assessed whether each of these selected genomic regions recombined with the SDR at any time after the divergence of the WXa and ZM lines (fig. 2). This task relied on SNPs whose alleles can be assigned to W- or Z-linked parental haplotypes. These were the informative SNPs or those that were homozygous in mothers (fig. 2). For each SNP that was not informative in at least one family, we imposed a minimal genotype quality of 40 and a maximum sequencing depth equaling the 95% quantile of this variable, for each parent of the family (or families). We then discarded non-informative SNPs for which the rarer allele was carried by only one parental chromosome among the four parents, as such SNPs cannot inform on recombination. We refer to the remaining set of SNPs as "selected SNPs." Recombination between a selected SNP and the SDR was inferred if these two loci constituted four different haplotypes in the parents, considering the SDR as a biallelic locus with Z and W alleles. We refer to selected SNPs that recombined with the SDR as "recombinant SNPs" (fig. 2).

Among selected SNPs, we looked for nonrecombinant SNPs that were informative in both families (e.g., SNPs #1 and #4 in fig. 2). We reasoned that close linkage to the SDR should maintain female heterozygosity by balancing selection, hence increase the frequency of this category of SNP. Within this category, we looked for SNPs that may functionally contribute to sex determination and which may be located in the sex-determining locus. Such a SNP must be heterozygous in all females and males of all families, including parents and F1s, must be homozygous at the same base. This SNP must therefore have its maternal allele linked to the W allele and it must present the same maternal and alternative alleles across families (e.g., SNP #1 in fig. 2). Hereafter, we call any such SNP a "potentially causal SNP."

To more reliably determine that a given SNP is potentially causal, we estimated the probability that this SNP has not recombined with the SDR in our third family (BF). Because parental genotypes were missing, we assumed that the candidate SNP was informative in this family, its maternal allele linked to the W allele and that the SNP presented the same maternal and alternative alleles as the WXa and ZM families. Based on these assumptions, we computed $\prod p$ (eq. 4) on the two BF pools for each selected SNP as if it constituted its own haplotype. Not fulfilling these assumptions or having recombined with the SDR would result in a low value of $\prod p$. We discarded a SNP as potentially causal if the $\prod p$ of the BF pools was <0.01. This threshold was chosen after considering that 98.9% of the informative SNPs carried by the studied contigs in the WXa and ZM families had a value exceeding 1% for this variable. We therefore considered the 1% threshold as rather permissive in the selection of potentially causal SNPs.

Beyond SNPs, we aimed at defining larger regions that may or may not have recombined with the SDR. To do so, we delineated contig regions (hereafter called "blocks") within which no SNP showed evidence for recombination with any other, by applying the aforementioned four-haplotype criterion on every possible pair of selected SNPs (see supplementary text, Supplementary Material online for details). Note that a potentially causal SNP and a recombinant SNP cannot be in the same block as the former segregates identically with the ZW locus.

We ignored every block containing a single selected SNP, unless this SNP was potentially causal, as we considered such a short block as possibly resulting from a genotyping error rather than from recombination. As blocks were initially delineated by SNP coordinates, we extended block boundaries up to contig edges, or up to midpoints between consecutive blocks, as appropriate. We then considered any block harboring at least two recombinant SNPs or at least 50% of recombinants among selected SNPs as having recombined with the SDR.

Search for W-Specific Sequences

The Z and W alleles of the SDR may not show detectable homology due to excessive molecular divergence. This possibility implies that the most divergent parts of the Z and W alleles constitute different contigs in the reference genome assembly, with some contig(s) containing W-specific regions and other contigs (or another contig) containing Z-specific sequences. Such regions would not present informative SNPs since W-derived and Z-derived reads would not map on the same locations. However, the sequencing depth of a W-specific region should be close to zero for male-derived

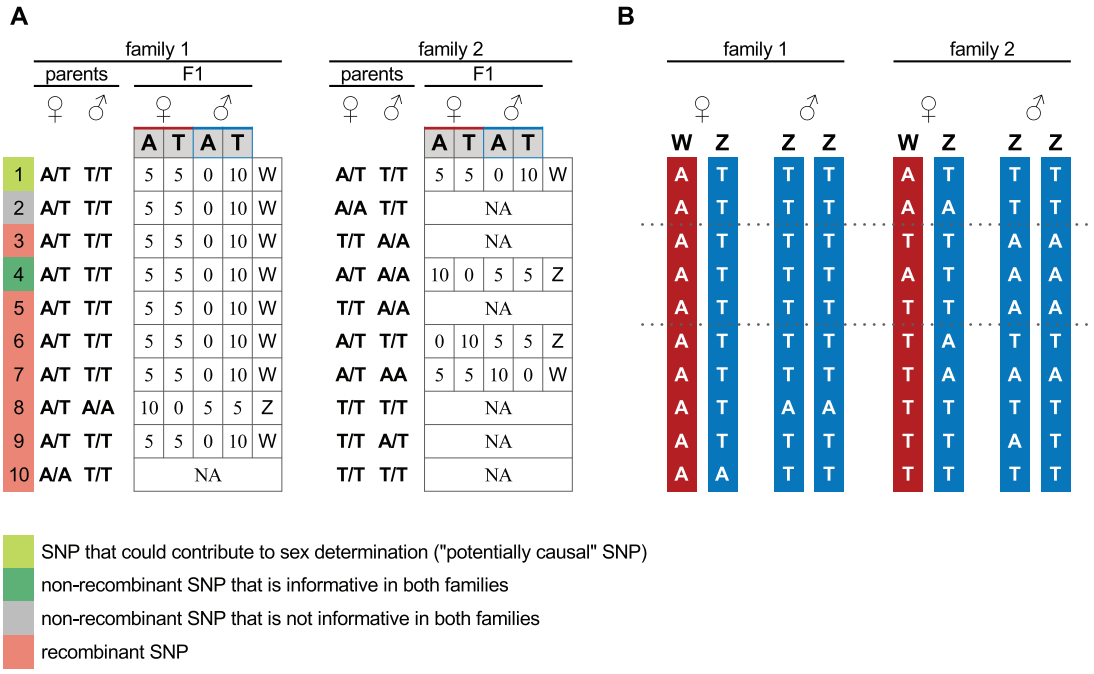

Fig. 2.—Ten hypothetical SNPs at a locus that has not recombined with the SDR during crosses involving two families. Letters A/T indicate DNA bases (alleles) at the SNPs. (*A*) Parental genotypes and data from F1 pools of five siblings. Numbers (0, 5, 10) in the F1 tables indicate the number of chromosomes that carry each allele and are only recorded for informative SNPs (otherwise, "NA" is noted). The SDR allele that is linked to the maternal allele (rightmost column of each table) is inferred from allele frequencies in the F1 (see "Estimation of Recombination with the Sex-Determining Locus" section). This inference permits the phasing of parental haplotypes, shown as vertical rods in panel (*B*). Recombination must have occurred between certain SNPs and the SDR during the divergence of families, barring homoplasy in the SNPs. SNP #4 may not have recombined with the SDR, but it is flanked by two SNPs that must have. Because there is no evidence of recombination between SNP #4 and these two others (these three SNPs constitute three different haplotypes, not four), they constitute a single genomic block delineated by the dotted lines.

sequencing reads and it should be higher (half the autosomal sequencing depth on average) for female-derived reads. We used this criterion to locate such regions.

Sequencing depth of all six F1 pools was measured with samtools, using the same mapping quality threshold as that used for the SNP analysis. For each pool, sequencing depth was averaged over a 2-kb sliding window that moved by 500 bp increments, leading to a 1,500-bp overlap between successive windows. To standardize results, given differences in sequencing effort, we multiplied depths by the highest mean sequencing depth over the six pools (averaged over all windows) and divided them by the mean sequencing depth of the pool under consideration. We then computed the "Chromosome Quotient" (CQ) (Hall et al. 2013) by dividing the sequencing depth obtained from sons by that obtained from daughters, for each window in each family. We selected candidate W-specific sequences as genomic windows with CQ <0.3 and female sequencing depth >5 in all three families. This rather permissive filter allows a certain proportion of male reads to be mapped, although possibly only onto part of a genetic window, and excludes regions that are not well sequenced in both sexes for reasons unrelated to sex determination.

To independently validate the female specificity of these sequences, we designed PCR primers that should yield amplicons only in females, in a fashion similar to Chebbi et al. (2019). Given our results (see next section), we developed primers for a single contig: contig 20397. We targeted regions for which sequencing depth was positive in all daughter pools and zero for all son pools of the three families, and for which no sequence variation was detected among daughters. We used the top pair of primers returned by the Primer3Plus web interface (Rozen and Skaletsky 2000): 5′-GGCAGCTGAAAAACACCAGG-3′ and 5′-ACTTTAGGGGT TTCAGTGGTGA-3′, yielding an amplicon of expected size 588 bp centered on position 10324 of contig 20397. We performed PCRs on all sequenced F1 siblings of the three studied lines for which DNA was available. These represent 6, 10 and 9 daughters, and 6, 8 and 7 sons from the WXa, ZM and BF families, respectively.

The amplification of each DNA sample took place in a 15-µl mix containing 0.6 µl of DNA solution, Promega PCR buffer (1× final concentration), Promega GoTaq polymerase (0.75 units), 43 µM of each dNTP and 0.28 µM of each primer. PCRs used the following temperature cycling: initial denaturation at 94 °C for 3 min, followed by 35 cycles of denaturation at

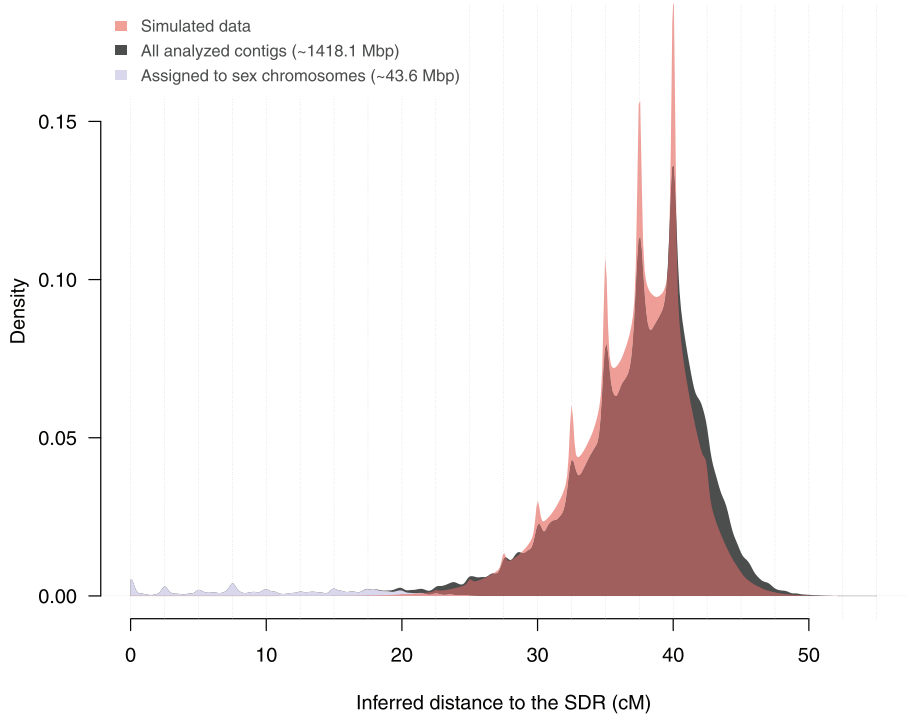

Fig. 3.—Distributions of the inferred genetic distance of *A. vulgare* contigs to the SDR for real data and for simulated data assuming that all contigs are located on autosomes. Genetic distances are inferred from simulated or observed genetic data from 40 F1 siblings belonging to two families. Vertical dotted lines represent genetic distances corresponding to integer numbers of recombination events during the crosses. Distributions only consider contigs for which both families present informative SNPs (33,875 contigs). Contigs whose inferred distance to the SDR was sinificantly lower than the distance yielded by simulations were assigned to sex chromosomes and constitute the blue area (see text). The modes of the distributions are lower than 50 cM, despite this value being the expectancy for autosomal contigs, because the phasing of maternal haplotypes is less reliable for contigs that are distant from the SDR (supplementary text, fig. S1, Supplementary Material online).

94 °C for 30 s, annealing at 55 °C for 30 s and elongation at 72 °C for 60 s, ending with a 10-min elongation step at 72 °C.

Results

Sex Chromosomes Constitute at Least 43 Mb of the *A. vulgare* Genome

For the WXa and ZM families combined, more than 5.1 million individual positions of the genome were homozygous in fathers and heterozygous in mothers, constituting potentially informative SNPs. The ~3.7 million SNPs that passed our filters were carried by 40,640 contigs constituting ~96.7% of the genome assembly length (1.72 Gb). Among these, 30,875 contigs (~82.2% of the genome assembly length) carried informative SNPs in both families.

Sequence data from the F1 pools at informative SNPs were used to compute n_{rec} (eq. 3), the estimated number of oocytes that have recombined with the SDR in our crosses. Detailed results for each contig are provided in supplementary file S1, Supplementary Material online. Figure 3 shows the distribution of the percentage of recombinant oocytes (n_{rec}/40 × 100), which is the inferred distance to the SDR in cM. The distribution obtained from real data is similar to that obtained from simulated autosomal contigs, but is slightly shifted to the right, possibly due to errors that we did not simulate (supplementary text, Supplementary Material online). Despite this slight shift, a tail is visible to the left, denoting contigs located closer to the SDR than expected for autosomes. In particular, 1,004 contigs, representing the pale blue distribution in figure 3 and totaling ~43.6 Mb, present significantly lower distances to the SDR than expected from autosomal contigs. These 1,004 contigs were thus assigned to sex chromosomes. Considering that these results implicate ~82.2% of the genome assembly (contigs showing informative SNPs in both families), we extrapolate that ~53 Mb of contigs (43.6/0.822) are located on sex chromosomes. These contigs should include about one thousandth of the autosomal contigs (false positives) at our significance level of 1/1,000. However, false negatives are likely to be more frequent than 1/1,000 according to additional simulations we performed to evaluate our method (supplementary text, fig. S3, Supplementary Material online). Therefore, the estimated length of *A. vulgare* sex chromosomes can be considered as conservative.

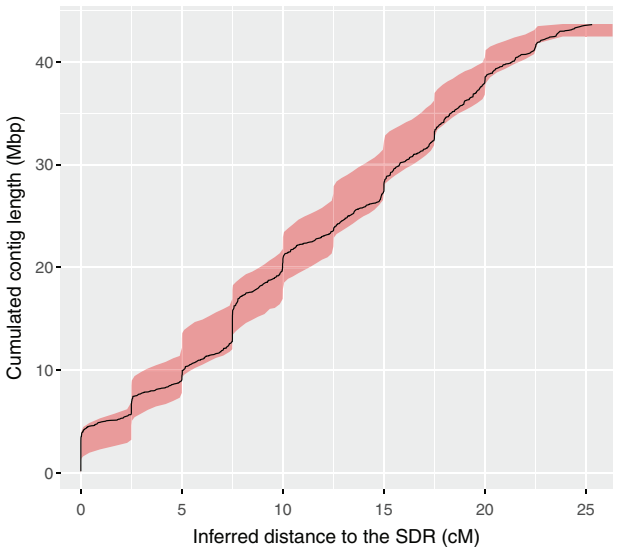

Fig. 4.—Cumulated length of 1004 *A. vulgare* contigs that locate below or at a given genetic distance from the SDR. Genetic distances are inferred from genetic data from 40 F1 siblings belonging to two families. The black curve represents observed data, and the colored area is the envelope constructed from the 0.005 and 0.995 quantiles of the cumulated length of contigs simulated under the assumption of uniform crossing over rates along sex chromosomes.

The SDR of *A. vulgare* Is Located within Less than 1 Mb

The sex chromosomes of *A. vulgare* appear to show relatively uniform crossing over rates. The cumulated length of contigs assigned to sex chromosomes indeed increases regularly with the inferred genetic distance to the SDR. Moreover, the curve remains within the envelope obtained from simulations assuming uniform crossing over rates (fig. 4). There is consequently no evidence for a reduction in crossing over rates near the SDR. If it were the case, the cumulated length of contigs would have been higher than expected at short genetic distances.

Among the 1,004 contigs assigned to sex chromosomes, 112 collectively accounting for ~5.1 Mb (fig. 5) presented little evidence of recombination with the SDR during our crosses ($\prod p > 0.5$, eq. 4). However, our SNP-based analysis leveraging the use of the two *A. vulgare* (WXa and ZM) lines indicated that most of these contigs have recombined with the SDR at some point after the divergence of the WXa and ZM chromosomes. Indeed, the 112 selected contigs were largely composed of genomic blocks harboring recombinant SNPs. Overall, the genomic regions that did not show evidence of recombination with the SDR are rare, totaling ~895 kb.

We detected only ten SNPs that could functionally determine sex (i.e., potentially causal SNPs for which all males

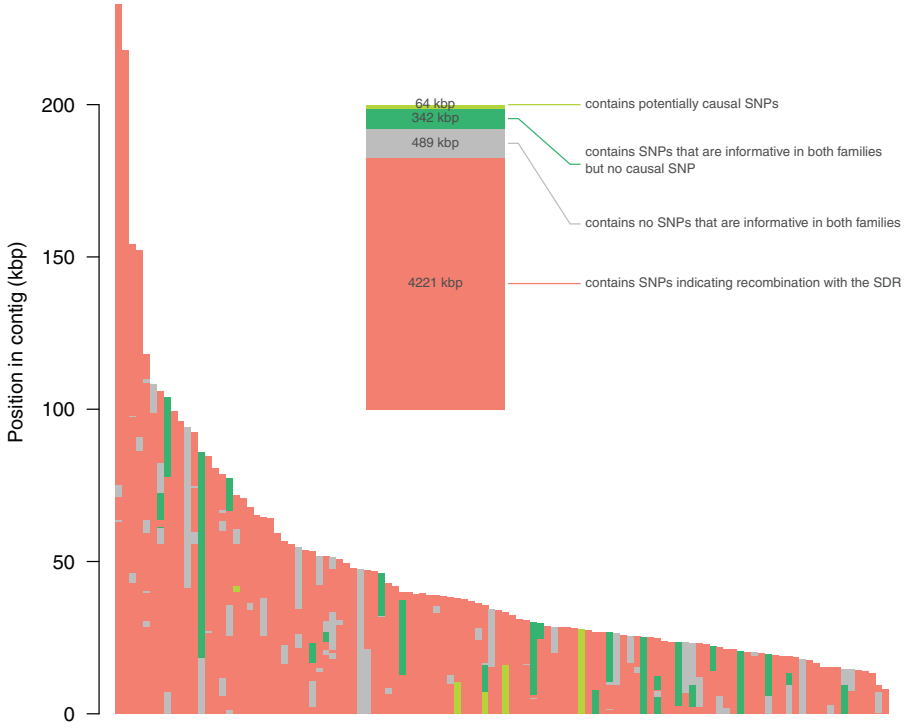

Fig. 5.—The 112 *A. vulgare* contigs that are inferred not to have recombined with the SDR in 40 F1s from two crosses. Each sectored vertical bar of the larger plot represents a contig. Contigs are ranked according to their length. Sectors within bars represent the genomic blocks constituting contigs (see "Localization of Genomic Regions That May Contain the SDR" section). Bar colors represent the SNPs that genomic blocks carry and use the same color codes as in figure 2 and in the inset. The inset shows the total lengths of different categories of genomic blocks according to the SNPs they carry. Blocks belonging to first three categories (from the top) contain no more than one recombinant SNP and less than 50% of recombinant SNPs.

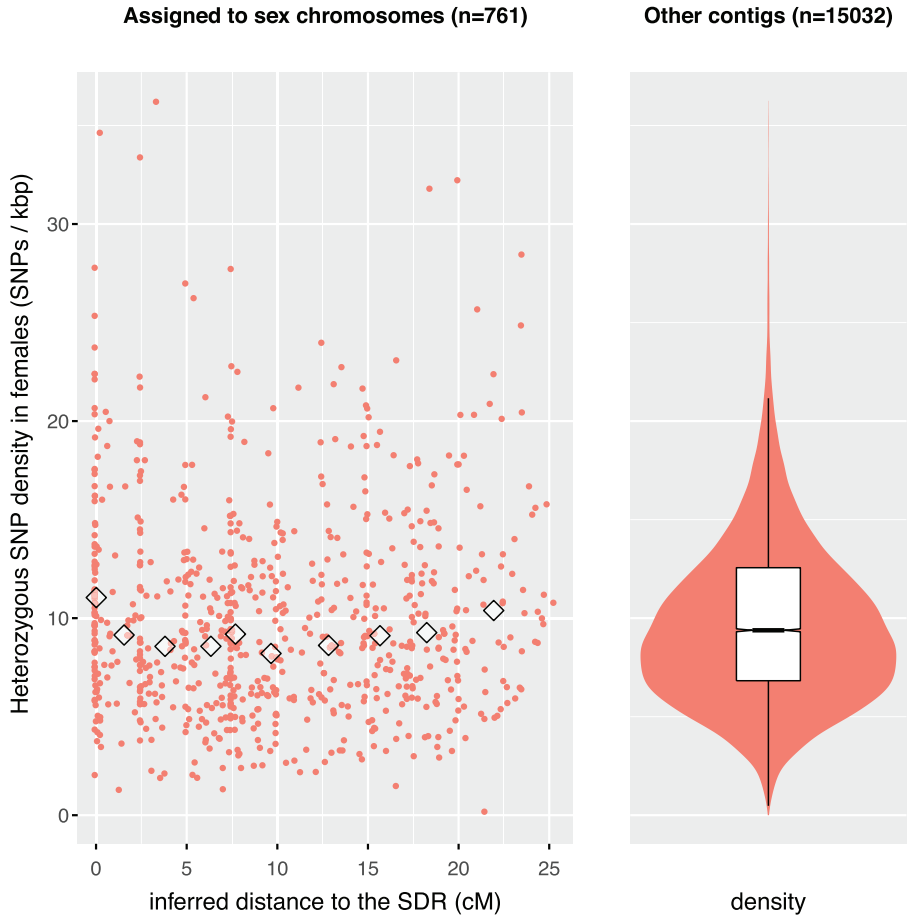

Fig. 6.—Density of heterozygous SNPs in females as a function of the inferred genetic distance to the SDR for contigs assigned to sex chromosomes (left-hand plot) and its distribution for other contigs (right-hand plot). The diamonds on the left-hand plot represent medians computed for ten classes of genetic distance. Classes are delimited by the deciles and therefore comprise ~70 contigs each. For the density computation (right-hand plot), contigs were assigned weights equal to the lengths of regions with sufficient sequencing depth to measure heterozygosity (see "Materials and Methods" section).

appear to be homozygous and females heterozygous for the same bases) in all three families (120 SNPs if we ignore the BF family). The ten SNPs correspond to five genomic blocks that are located on as many different contigs and that represent ~64 kb. None of these SNPs are located in an exon.

The SDR had a modest impact on the molecular divergence of the *A. vulgare* sex chromosomes. Indeed, female heterozygosity did not significantly decrease with the inferred genetic distance to the SDR (one-sided Spearman's rank sum test $S = 74{,}540{,}329$, $p = 0.3416$) (fig. 6). Yet, female heterozygosity increased at the closest distance to the SDR—its median was indeed significantly higher for the 112 contigs that showed little evidence of recombination with the SDR than the other contigs assigned to sex chromosomes (~10.32 SNPs/kb vs. ~9.01 SNPs/kb, one-sided Mann–Whitney's $U = 41{,}839$, $p \approx 0.005$). Despite this difference, the median female heterozygosity of the contigs assigned to sex chromosomes did not differ from that of the other contigs (two-sided $U = 5{,}598{,}974$, $p = 0.3253$). Overall, there is no evidence that the divergence between sex chromosomes in *A. vulgare* is higher than between autosomes of the same pair. It should be kept in mind that these comparisons do not include the whole genome as we discarded contigs for which *f* could not be inferred with a posterior probability of at least 0.5 (see "Investigation of Heterozygous SNP Density" section).

W-Specific Sequences Are Rare

The ratios of sequencing depths obtained from sons to those obtained from daughters (CQ scores) presented distributions that are typical of autosomes for all three families (supplementary fig. S4, Supplementary Material online), showing little evidence for heteromorphic sex chromosomes. Only seven contigs contained W-specific sequences, as defined by genomic windows with CQ <0.3 and female sequencing depth >5 in all three families. These windows added up to ~92 kb.

Informative SNPs present in these seven contigs indicated that six have recombined with the SDR during the crosses, with a minimum n_{rec} of ~3.5. The one exception was contig 20397 (fig. 7). Remarkably, it is also the contig that possesses

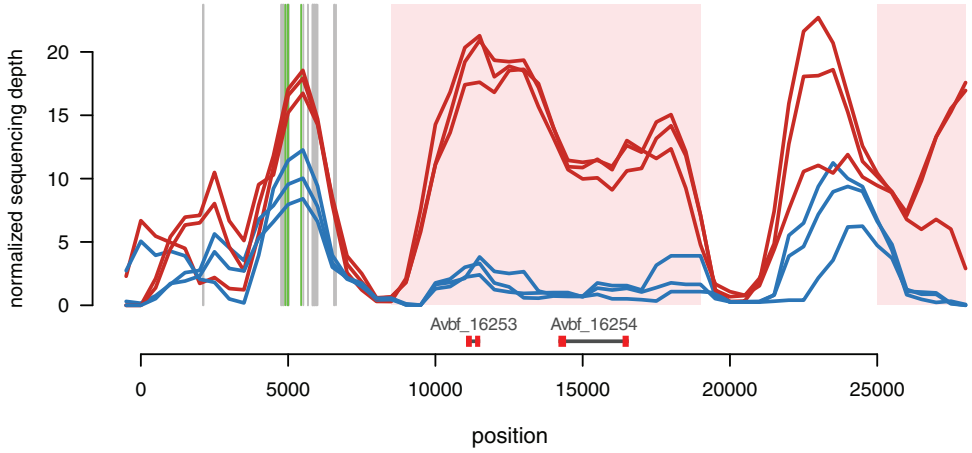

Fig. 7.—Normalized female (red curves) and male (blue curves) sequencing depths on contig 20397, presenting low chromosome quotient (CQ), based on sequenced DNA from the progeny of three *A. vulgare* families, constituting six pools. Regions of CQ < 0.3 and of female sequencing depth ≥ 5 in all families are represented as light pink areas. Vertical lines represent informative SNPs, including (gray) SNPs that are variable in a single family and (green) SNPs that are compatible with the control of sex ("potentially causal SNPs," see text). Annotated genes are represented by horizontal dark gray lines under gene identifiers. Exons are shown as thick red bars.

the most potentially causal SNPs ($n = 4$) in a single block spanning the whole contig (fig. 5). A PCR assay targeting a 588-bp region near position 10300 on contig 20397 yielded an amplicon of expected size in each of the 25 tested daughters, and no amplicon in the 21 tested sons. These results corroborate the hypothesis that the low-CQ regions of contig 20397 contain sequences that are specific to females in the three studied families.

Information about the two annotated genes in contig 20397 (fig. 7) is provided in table 2. For gene *Avbf_16253*, which encodes a putative protein of unknown function, a sequence homology search against the nonredundant protein database of the National Center for Biotechnology Information using the BLASTp online tool did not return any hit outside *Armadillidium* proteins.

A DNA sequence similarity search using BLASTn (Camacho et al. 2009) with default settings showed that the exons of these genes were similar to exons of other annotated genes of the *A. vulgare* genome. These genes therefore present paralogs. The sequence identity of the most similar copy of each gene (table 2) was higher than the identity measured with the most similar annotated gene of *A. nasatum* (Becking et al. 2019), which was inferred to have diverged from the *A. vulgare* lineage ~25 Ma (Becking et al. 2017). This result suggests relatively recent divergence of the two genes from their closest paralogs. The low sequence divergence between the exons of each of these two genes and its closest relative, representing at most six substitutions, prevented a meaningful analysis of natural selection acting on its specific evolutionary branch.

Discussion

Benefits of Our Method

Segregation/association analysis methods aimed at locating sex-associated regions generally rely on individual genotyping by partial genome sequencing (e.g., Gamble et al. 2015; Jeffries et al. 2018, see Palmer et al. [2019] for a review). Pool-seq appears to be less utilized for this task (but see Michalovova et al. [2015]). However, when obtaining a genetic map is not essential, whole-genome pool-seq has several advantages: it is easy to implement, it is applicable to any species, and perhaps more importantly, it can yield millions of SNPs. In comparison, partial genome sequencing using RAD tags yields tens of thousands of reliable SNPs at most. If this number does not largely exceed the number of contigs in the genome assembly, as might be the case for a nonmodel organism, a high false negative risk may affect the selection of contigs that contain (or compose) the SDR. RNA sequencing (as used in Muyle et al. [2016] and Michalovova et al. [2015]) would also suffer this problem if exon density is low (~1.4% of the genome assembly in the case of *A. vulgare*), added to the fact that only a subset of genes is expressed during an experiment. This could be a concern in species with short SDRs that encompass only few expressed genes and/or few contigs of a genome assembly.

The millions of SNPs yielded by whole-genome sequencing is leveraged by our approach through the combination of all the informative SNPs of a contig into haplotypes. Haplotype frequencies are more precisely inferred than allele frequencies based on single SNPs. A high density of SNPs is also

Table 2
Annotated Genes in Contig 20397 Containing Genomic Windows of Low Chromosome Quotient in the Three *Armadillidium vulgare* Lines

Name	Number of Exons	Exon Length	Description	Number of Paralogs[a]	Highest Identity with Paralogs[b]
Avbf_16253	2	204	Hypothetical protein	11	97.0%
Abvf_16254	2	228	Putative tRNA N6-adenosine threonylcarbamoyltransferase, mitochondrial	2	98.7%

[a]Genes that harbor a region of at least 100 bp having a BLASTn *e*-value $<10^{-4}$ with the coding sequence of the gene listed in the first column.
[b]Identity considers the alignments reported by BLASTn, not the whole gene lengths.

particularly useful for pinpointing the SDR within contigs that did not recombine with this locus during the crosses, thanks to a multifamily setup (figs. 2 and 5). Here, the resolution corresponds to the typical distance between SNPs that inform on recombination with the SDR and would be quite limited if partial genome sequencing were employed. In fact, our approach can be used to locate a genomic region controlling any qualitative trait that depends on a single locus. Doing so would simply require treating the phenotype that associates with the heterozygous genotype as the ZW females of our study.

We emphasize that the number of read pairs covering all the informative SNPs of a contig (variables r_{wi} and r_{zi}) determines the certainty of the estimated allele frequencies. These numbers of reads, which should ideally constitute hundreds, positively correlate with contig length, informative SNP density and sequencing depth. Contigs with more informative SNPs, hence longer contigs for a given SNP density, thus require lower average sequencing depth for a similar degree of certainty in estimated haplotype frequencies. Reciprocally, a higher sequencing effort may be required if the density of heterozygous SNPs in the heterogametic parent(s) is suspected to be low, and/or if contigs are short.

With respect to the crossing scheme, the total number of F1 siblings (N) is inversely proportional to the resolution of the estimated genetic distance to the SDR (100/N cM). For a desired resolution (in base pairs), the choice of N should be made with consideration to recombination rates in the heterogametic sex, that is, the genome size represented by 100/N cM. Use of large F1 pools may also improve the accuracy of haplotype phasing, since phasing requires comparing allele frequencies between daughters and sons. However, the size of a pool, n, must be chosen to ensure that 1/n is at least twice as high as the sequencing error rate, to clearly differentiate rare alleles from errors. Cost considerations aside, a larger number of families (hence of pools) can circumvent this limitation and also increase the ability to detect past recombination with the SDR during the divergence of studied lineages, and hence to exclude genomic regions that do not contain the SDR.

Field approaches can constitute powerful alternatives to pinpoint loci controlling certain phenotypes, especially for species that cannot be bred in the laboratory, and pool-seq has been frequently used in this context (Kofler et al. 2011). It is not yet clear which approach, family-based or population-based, yields better precision. More recombination events should have occurred between a given SNP and the target locus in natural populations than in the past history of a couple of laboratory-bred families. However, because haplotypes cannot be phased on field-collected individuals analyzed in pools, recombination events cannot be identified. Hence, loci that do not control the studied phenotype may be harder to exclude with certainty. Confronting our results with those from an association study on wild *A. vulgare* populations would therefore be of interest.

Although a SNP-based method (be it conducted in the laboratory or in the field) is restricted to genomic regions that are similar between sex chromosomes, whole-genome pool-seq also enables coverage-based analyses designed to find regions that are not. This brings clear advantages over partial genome sequencing here as well. Coverage-based analyses (reviewed in Palmer et al. [2019]) include subtraction-based methods (e.g., the CQ method of Hall et al. [2013] that we used) or methods that look for sex-specific *k*-mers (Carvalho and Clark 2013). The use of sliding windows, which would not be permitted under partial genome sequencing, allows low-CQ regions to be shorter than contigs. In this case, the statistical analysis of nearby SNPs provides a useful complement to the CQ scores. Indeed, the absence of reads covering DNA sequences from just one sex (CQ = 0) may not imply the complete absence of such sequences from the genome(s) of the analyzed individuals, due to the odds of DNA sequencing. Also, it is unclear whether a low CQ value indicates low genetic similarity between sex chromosomes at the focal genomic window (allowing a small portion of reads from a chromosome to map on the sequence of the other) or the rare occurrence of an allele in the sex where this allele is supposed to be absent. The analysis of nearby SNPs reduces these uncertainties by providing a probabilistic assessment of the association between genetic variants and sexes, which may extend to low-CQ regions of the same contig. This

combination of approaches allowed us to pinpoint contig 20397 among the seven contigs showing low-CQ windows in *A. vulgare*.

Low Heteromorphism and Sequence Divergence between *A. vulgare* Sex Chromosomes

Our previous study (Chebbi et al. 2019) outlined 27 contigs containing W-specific sequences, only two of which (including contig 20397) presented low-CQ windows in the present study. This difference can be explained by the fact that Chebbi et al. (2019) investigated a single family, which reduced the probability of recombination with the SDR, and computed CQ at the scale of whole contigs rather than genomic windows. Indeed, 14 of the 27 contigs outlined by Chebbi et al. (2019) are among those we assigned to sex chromosomes (considering that four of the 27 contigs did not have informative SNPs and could not be assigned). Due to partial linkage to the SDR, these 14 contigs may have harbored genetic differences that associated with the sex of individuals studied in Chebbi et al. (2019), which are the siblings from line BF that we reused here. This association did not hold in the other two lines we analyzed, except for contig 20397. The low CQ scores reported by Chebbi et al. (2019) at these contigs would therefore correspond to simple polymorphism on the sex chromosomes. The other nine contigs that we did not assign to sex chromosomes may be more distant from the SDR. Their median inferred genetic distance to the SDR of ~31.5 cM is still significantly lower than the median of other contigs not assigned to sex chromosomes (39.3 cM) (two-sided Mann and Whitney's $U = 64{,}752$, $p \approx 8.5 \times 10^{-4}$). Thus, some of the nine contigs may belong to sex chromosomes despite not having passed the assignment test.

The scarcity of W-specific sequences in the *A. vulgare* genome is mirrored by the rarity of Z-specific sequences. Indeed, the CQ distributions obtained from the three studied lines (supplementary fig. S4, Supplementary Material online) show no increase near the value of 2, which is the expected CQ value for Z-specific regions. We did not specifically study regions with high CQ, because elevated CQ on short genomic windows is subject to high sampling variance, hence to false positives/negatives, as opposed to low CQ which involves low sequencing depth, hence low variance. At any rate, these results demonstrate the very low divergence of the *A. vulgare* sex chromosomes. In our previous study (Chebbi et al. 2019), the inference of low divergence was not definitive as the size of sex chromosomes was undetermined. Here we show that the sex chromosomes have a minimal size of 53 Mb, that is, 83% the average size of *A. vulgare* chromosomes (~64 Mb) based on genome size and number of chromosomes. Even though our estimate of sex chromosome length is conservative, it is orders of magnitude above that of W-specific sequences.

The fact that sex chromosomes did not show evidence for nonuniform crossover rates (fig. 4) and are not distinguishable from autosomes in terms of heterozygous SNP density in females (fig. 6) suggests comparable levels of recombination between chromosome types. The higher density of heterozygous SNPs in contigs locating the closest to the SDR (fig. 6) could just be the byproduct of balancing selection, increasing coalescent times of SDR-linked alleles over a relatively short genomic region. As for any balanced polymorphism, the regions with elevated genetic divergence are expected to be very narrow (Charlesworth et al. 1997; Innan and Nordborg 2003). Hence, we do not consider this observation as conclusive evidence for a reduction of crossing over rate around the SDR.

The apparent absence of recombination reduction in *A. vulgare* sex chromosomes could reflect their recent origin, consistent with the apparent rapid turnover of sex chromosomes in terrestrial isopods (Becking et al. 2017). An evolutionary scenario for this renewal involves feminizing bacterial endosymbionts of the genus *Wolbachia* (Rigaud et al. 1997; Cordaux et al. 2011). *Wolbachia* endosymbionts that infect terrestrial isopods can improve their maternal transmission by feminizing their carriers, as commonly observed in *A. vulgare* populations (Juchault et al. 1993; Verne et al. 2012; Valette et al. 2013). Theoretical models and field surveys on *A. vulgare* indicate that invasion of a ZW population by feminizing *Wolbachia* leads to the loss of the W allele, as feminized ZZ individuals take the role of mothers (reviewed in Cordaux and Gilbert [2017]). Sex then becomes entirely determined by the presence of *Wolbachia* and is generally biased toward females, reflecting the prevalence of the feminizing bacteria. New sex chromosomes may emerge via the selection of masculinizing nuclear genes that may reestablish an even sex ratio (Caubet et al. 2000; Becking et al. 2019) or through horizontal gene transfer of feminizing *Wolbachia* genes into host genomes, producing new W-type chromosomes (Leclercq et al. 2016; Cordaux and Gilbert 2017).

Despite the lack of evidence for a recombination-suppressed region, we cannot strictly exclude that the ZW chromosomes of *A. vulgare* are old. Indeed, the conditions that theoretically favor a reduction of recombination rates around the sex-determining locus have seldom been verified (Wright et al. 2017) and might not apply to the majority of taxa. Given that female heterogamety prevails among Armadillidiidae and related families (Becking et al. 2017), the *A. vulgare* sex chromosomes may have been maintained at a low level of molecular divergence for a long period through sustained recombination, similarly to what is observed in palaognaths (Xu et al. 2019), European tree frogs (Stöck et al. 2011), or guppies (Bergero et al. 2019; Darolti et al. 2020). Characterizing the SDRs of ZW species that are closely related to *A. vulgare* should allow evaluating the homology, hence the age, of sex chromosomes among these lineages.

The SDR of *A. vulgare*

Barring false negatives, the sex-controlling locus of *A. vulgare* should lie within the ~0.9 Mb of genomic blocks that did not show evidence for recombination with the SDR (fig. 5). As these blocks span several contigs that also harbor recombinant SNPs, the total length of the SDR, as defined by the nonrecombining chromosomal region surrounding the sex-determining gene(s), is likely to be much shorter than 0.9 Mb. Where the SDR is located in these 0.9 Mb cannot yet be determined, but we expect this locus to harbor Z-specific and W-specific alleles that have been maintained for a long time by balancing selection. We therefore do not consider as likely SDR candidates the 489 kb of genomic regions harboring no SNP whose alleles could be associated to the Z and W alleles in both *A. vulgare* lines. On the other hand, the 64 kb of genomic regions containing potentially causal SNPs are of greater interest as they contain sex-associated variation in all three investigated families.

Among the candidate genomic regions, contig 20397 emerges as the most interesting one. The low CQ scores on this contig (fig. 7) may indicate the presence of large indels between sex chromosomes, like in medaka fish that present a Y-specific insertion (Myosho et al. 2015) and/or the accumulation of smaller mutations that make male reads unable to align on the reference sequence (i.e., the W allele of the SDR). As the two annotated genes of contig 20397 appear to be female specific, hence absent from the Z allele, their expression may be required for development into a female phenotype. Interestingly, these genes present highly similar paralogs. Evolution of master sex-determination genes by duplication of existing genes has been reported in several taxa, such as medaka fish (Matsuda et al. 2002; Nanda et al. 2002) and clawed frog (Yoshimoto et al. 2008). Unfortunately, the available functional annotation for the two genes on contig 20397 (table 2) does not inform us about possible mechanisms of action. We could not find a documented role for a mitochondrial tRNA processing enzyme (the putative function of Abvf_16254) in sex determination. As we cannot determine whether these two genes have evolved under natural selection after their divergence from their closest paralog, we cannot exclude that these genes are redundant copies that happen to be linked to a nearby feminizing allele. Their absence from males may have no phenotypic consequence. We also keep in mind that female heterogamety may not necessarily involve feminizing gene transcripts or proteins encoded by the W allele and that sex could be determined by the dosage of a Z-encoded protein, as in the chicken (Smith et al. 2009). Investigating these hypotheses requires searching for Z-specific sequences in the *A. vulgare* genome. CQ scores obtained by comparing laboratory produced WW individuals (Juchault and Legrand 1972) with ZZ or ZW individuals should be close to zero for Z-specific sequences, similarly to the CQ scores we used to locate W-specific sequences.

To conclude, the low molecular divergence of the *A. vulgare* sex chromosomes, and their apparently uniform recombination rates, allowed us to pinpoint a limited set of regions that could contain the SDR and to identify two potential feminizing genes. Their strict association with the female sex in additional *A. vulgare* lines and their expression levels during sex differentiation will be the focus of future research. This research will also address the possibility that the sex-determining locus was missed. Although the risk of this locus simply not being in the current genome assembly appears low (supplementary text, Supplementary Material online), many contigs, especially short ones, had insufficient sequencing coverage and/or number of analyzable SNPs. Mapping our sequence data on a more contiguous (chromosome-scale) genome assembly and applying our approach to long genomic windows will greatly lessen these risks and ensure that the SDR of *A. vulgare* is characterized in its entirety.

Supplementary Material

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

Acknowledgments

We thank Clément Gilbert for discussions at an early stage of the project, Bouziane Moumen for providing support for our computing infrastructure and the genotoul bioinformatics platform Toulouse Midi-Pyrénées (bioinfo.genotoul.fr/) for providing computing and storage resource. We thank Dr O'Neill and three anonymous reviewers for their comments on the manuscript. This work was funded by the European Research Council Starting (Grant No. 260729) (EndoSexDet) and Agence Nationale de la Recherche (Grant No. ANR-15-CE32-0006-01) [CytoSexDet] to R.C. and Grant No. ANR-20-CE02-0004 [SymChroSex] to J.P.), 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.

Data Availability

The custom code used to perform the data analysis is available at https://github.com/jeanlain/ZWAvulgare (last accessed June 2021). The genetic data generated in this study are deposited to NCBI under accession number PRJNA697978.

Literature Cited

Akagi T, Henry IM, Tao R, Comai L. 2014. A Y-chromosome–encoded small RNA acts as a sex determinant in persimmons. Science 346:646.

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

Bachtrog D. 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nat Rev Genet. 14:113–124.
Bachtrog D, et al. 2014. Sex determination: why so many ways of doing it? PLoS Biol. 12(7):e1001899.
Baird NA, et al. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One. 3:e3376.
Becking T, et al. 2017. Diversity and evolution of sex determination systems in terrestrial isopods. Sci Rep. 7(1):1084.
Becking T, et al. 2019. Sex chromosomes control vertical transmission of feminizing Wolbachia symbionts in an isopod. PLoS Biol. 17(10):e3000438.
Bergero R, Charlesworth D. 2009. The evolution of restricted recombination in sex chromosomes. Trends Ecol Evol. 24:94–102.
Bergero R, Gardner J, Bader B, Yong L, Charlesworth D. 2019. Exaggerated heterochiasmy in a fish with sex-linked male coloration polymorphisms. Proc Natl Acad Sci U S A. 116(14):6924–6931.
Beukeboom L, Perrin N. 2014. The evolution of sex determination. Oxford (UK): Oxford University Press.
Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.
Brante A, Quinones A, Silva F. 2016. The relationship between sex change and reproductive success in a protandric marine gastropod. Sci Rep. 6(1):26439.
Camacho C, et al. 2009. BLAST plus: architecture and applications. BMC Bioinform. 10:Article number 421.
Carvalho AB, Clark AG. 2013. Efficient identification of Y chromosome sequences in the human and Drosophila genomes. Genome Res. 23:1894–1907.
Caubet Y, Hatcher MJ, Mocquard J-P, Rigaud T, 2000. Genetic conflict and changes in heterogametic mechanisms of sex determination. J Evol Biol. 13:766–777.
Chakravarti A. 1991. A graphical representation of genetic and physical maps: the Marey map. Genomics 11:219–222.
Charlesworth B, Nordborg M, Charlesworth D. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet Res. 70:155–174.
Charlesworth D, Charlesworth B, Marais G. 2005. Steps in the evolution of heteromorphic sex chromosomes. Heredity 95:118–128.
Chebbi MA, et al. 2019. The genome of *Armadillidium vulgare* (Crustacea, Isopoda) provides insights into sex chromosome evolution in the context of cytoplasmic sex determination. Mol Biol Evol. 36:727–741.
Cordaux R, Bouchon D, Greve P. 2011. The impact of endosymbionts on the evolution of host sex-determination mechanisms. Trends Genet. 27:332–341.
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.
Darolti I, Wright AE, Mank JE. 2020. Guppy Y chromosome integrity maintained by incomplete recombination suppression. Genome Biol Evol. 12:965–977.
Edge P, Bafna V, Bansal V. 2017. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res. 27:801–812.
Furman BLS, et al. 2020. Sex chromosome evolution: so many exceptions to the rules. Genome Biol Evol. 12:750–763.
Futschik A, Schlötterer C. 2010. The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics 186:207.
Gamble T, et al. 2015. Restriction site-associated DNA sequencing (RAD-seq) reveals an extraordinary number of transitions among gecko sex-determining systems. Mol Biol Evol. 32:1296–1309.
Gautier M. 2014. Using genotyping data to assign markers to their chromosome type and to infer the sex of individuals: a Bayesian model-based classifier. Mol Ecol Resour. 14:1141–1159.
Gautier M, et al. 2013. Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. Mol Ecol. 22:3766–3779.
Hall AB, et al. 2013. Six novel Y chromosome genes in Anopheles mosquitoes discovered by independently sequencing males and females. BMC Genom. 14:273.
Innan H, Nordborg M. 2003. The extent of linkage disequilibrium and haplotype sharing around a polymorphic site. Genetics 165:437.
Jeffries DL, et al. 2018. A rapid rate of sex-chromosome turnover and non-random transitions in true frogs. Nat Commun. 9:4088.
Juchault P, Legrand JJ. 1972. Croisements de neo-mâles expérimentaux chez *Armadillium vulgare* Latr. (Crustacé, Isopode, Oniscoide). Mise en évidence d'une hétérogamétie femelle. Comptes Rendus de L'Academie Des Sciences, Paris. 274:1387–1389.
Juchault P, Rigaud T, Mocquard JP. 1993. Evolution of sex determination and sex-ratio variability in wild populations of *Armadillidium vulgare* (latr) (crustacea, isopoda) – a case-study in conflict-resolution. Acta Oecol. 14:547–562.
Kamiya T, et al. 2012. A trans-species missense SNP in *Amhr2* is associated with sex determination in the tiger pufferfish, *Takifugu rubripes* (Fugu). PLoS Genet. 8:e1002798.
Kofler R, et al. 2011. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS One. 6:e15925.
Leclercq S, et al. 2016. Birth of a W sex chromosome by horizontal transfer of *Wolbachia* bacterial symbiont genome. Proc Natl Acad Sci U S A. 113:15036–15041.
Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 1303.3997.
Li H, et al. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079.
Mank JE, Avise JC. 2009. Evolutionary diversity and turn-over of sex determination in teleost fishes. Sex Dev. 3:60–67.
Martin M, et al. 2016. WhatsHap: fast and accurate read-based phasing. *bioRxiv*. 085050.
Matsuda M, et al. 2002. *DMY* is a Y-specific DM-domain gene required for male development in the medaka fish. Nature 417:559–563.
Merchant-Larios H, Diaz-Hernandez V. 2013. Environmental sex determination mechanisms in reptiles. Sex Dev. 7:95–103.
Michalovova M, Kubat Z, Hobza R, Vyskot B, Kejnovsky E. 2015. Fully automated pipeline for detection of sex linked genes using RNA-Seq data. BMC Bioinform. 16:78.
Muller HJ. 1918. Genetic variability, twin hybrids and constant hybrids. In a case of balanced lethal factors. Genetic. 3(5):422–499.
Muyle A, et al. 2016. SEX-DETector: a probabilistic approach to study sex chromosomes in non-model organisms. Genome Biol Evol. 8:2530–2543.
Myosho T, Takehana Y, Hamaguchi S, Sakaizumi M. 2015. Turnover of sex chromosomes in celebensis group medaka fishes. G3 5:2685–2691.
Nanda I, et al. 2002. A duplicated copy of *DMRT1* in the sex-determining region of the Y chromosome of the medaka, Oryzias latipes. Proc Natl Acad Sci U S A. 99(18):11778–11783.
Palmer DH, Rogers TF, Dean R, Wright AE. 2019. How to identify sex chromosomes and their turnover. Mol Ecol. 28:4709–4724.
Pan Q, et al. 2019. Identification of the master sex determining gene in Northern pike (Esox lucius) reveals restricted sex chromosome differentiation. PLoS Genet. 15(8):e1008013.
R Development Core Team, 2020. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.
Rice WR. 1987. The accumulation of sexually antagonistic genes as a selective agent promoting the evolution of reduced recombination between primitive sex chromosomes. Evolution 41:911–914.
Rigaud T, Juchault P, Mocquard J-P. 1997. The evolution of sex determination in isopod crustaceans. Bioessays 19(5):409–416.

Rozen S, Skaletsky HJ. 2000. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S, editors. Bioinformatics methods and protocols: methods in molecular biology. Totowa (NJ): Humana Press. p. 365–386.

Smith CA, et al. 2009. The avian Z-linked gene *DMRT1* is required for male sex determination in the chicken. Nature 461:267.

Stöck M, et al. 2011. Ever-young sex chromosomes in European Tree Frogs. PLoS Biol. 9:e1001062.

Valette V, et al. 2013. Multi-infections of feminizing Wolbachia strains in natural populations of the terrestrial isopod *Armadillidium Vulgare*. PLoS One. 8:e82633.

Van der Auwera GA, et al. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bionformatics. 43:11.10.11–11.10.33.

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. J Evol Biol. 25:264–276.

Wright AE, et al. 2017. Convergent recombination suppression suggests role of sexual selection in guppy sex chromosome formation. Nat Commun. 8:14251.

Wright AE, Dean R, Zimmer F, Mank JE. 2016. How to make a sex chromosome. Nat Commun. 7:12087.

Xu L, Wa Sin SY, Grayson P, Edwards SV, Sackton TB. 2019. Evolutionary dynamics of sex chromosomes of paleognathous birds. Genome Biol Evol. 11:2376–2390.

Yoshimoto S, et al. 2008. A W-linked DM-domain gene, *DM-W*, participates in primary ovary development in *Xenopus laevis*. Proc Natl Acad Sci U S A. 105:2469–2474.

Associate editor: Rachel O'Neill