Genetic mapping in the red mason bee implicates *ANTSR* as an ancient sex-determining locus in bees and ants

Tilman Rönneburg[1], Demetris Taliadoros[1], Anna Olsson[1], Sara Magnusson[1], Linn Huser[1], Muhammad Nafiz Ikhwan Bin Nor Fuad[1], Turid Everitt[1], Giselle C. Martín-Hernández[1], Panagiotis Theodorou[3], Björn Cederberg[2], Robert J. Paxton[3], Karsten Seidelmann[4], Matthew T. Webster[1*]

1) Department of Medical Microbiology and Biochemistry, SciLifeLab, Uppsala University, Uppsala, Sweden

2) Swedish Species Information Centre, Swedish University of Agricultural Sciences, Uppsala, Sweden

3) General Zoology, Institute of Biology, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

4) Animal Physiology, Institute of Biology, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

*correspondence to matthew.webster@imbim.uu.se

Short title: Sex determination in *Osmia bicornis*

Abstract

Haplodiploid inheritance, in which females are diploid and males are haploid, is found in all species of Hymenoptera. Sex in haplodiploids is commonly determined by the alleles present at a complementary sex determination (CSD) locus, with heterozygosity triggering the female developmental pathway. The identity of this locus differs among taxa and is only known in a few species. Here, we map a single CSD locus to a 2 kbp region in the genome of the red mason bee *Osmia bicornis.* It overlaps the long noncoding RNA *ANTSR*, which has been identified as the sex-determining gene in the invasive ant *Linepithema humile*. This locus is homozygous in diploid males and exhibits extremely high levels of haplotype diversity, consistent with the action of frequency-dependent selection. The elevated levels of heterozygosity in the CSD locus enable us to fine-map potentially functional genetic variation within it. We also identify elevated levels of genetic diversity in the ortholog of the CSD locus in five other bee and ant genera, suggesting that it may govern sex determination widely in Hymenoptera. Our data are consistent with the hypothesis that *ANTSR* evolved a role in sex determination over 150 million years ago and is the ancestral sex-determination locus of bees and ants.

Introduction

A huge variety of genetic sex determination systems exist in nature [1,2]. The molecular pathways involved in determining sex are highly conserved among animal and plant taxa, but the molecular switches that provide the initial signal to specify sex are highly diverse. Haplodiploidy is a common mode of sex determination, in which females are diploid and males are haploid. Up to 15% of arthropods have this mode of sex determination, including all Hymenoptera (wasps, bees, ants and sawflies) [3]. However, the molecular mechanisms that trigger the molecular cascade translating ploidy differences into sex differences are only known in a few haplodiploid species.

The most widespread mechanism of sex-determination in Hymenoptera is inferred to be complementary sex determination (CSD). This system is likely to be the ancestral state in Hymenoptera [4,5] although other mechanisms are responsible in some taxa. For example, in the parasitoid wasp, *Nasonia vitripennis*, sex is determined by different patterns of imprinting of a single gene in haploids and diploids [6]. Under CSD, on the other hand, sex is specified by the alleles present at one or more genetic loci at which heterozygotes become female and hemizygotes become male [7].

So far, CSD loci have been fully characterised in one bee and one ant species. In the honeybee *Apis mellifera*, the CSD locus is a single protein coding gene *complementary sex determiner (csd)* [8]. It has been shown that the presence of amino acid differences between two alleles of the Csd protein are necessary for its activation and to trigger female development [9]. In the Argentine ant *Linepithema humile* a single CSD locus has been mapped to a region overlapping the lncRNA gene *ANTSR* [10] (Figure 1A). It was demonstrated that heterozygosity at this locus leads to elevated expression of *ANTSR* which triggers the molecular cascade leading to female development. Both these molecular switches act by inducing alternative splicing of an ortholog of *transformer* (*tra*), which in turn interacts with *doublesex* (*dsx*) leading to female development. These two genes are highly conserved components of the sex determination pathway of insects [11].

In species with CSD, occasional homozygosity at one or more CSD loci results in diploid males [4,7] (Figure 1B). This is a relatively rare and deleterious outcome and diploid males are typically sterile. The fact that diploid males must be homozygous at CSD loci, whereas diploid females must have two different haplotypes, can be used to identify their location in the genome [8]. Furthermore, frequency-dependent selection at CSD loci results in the maintenance of a large number of haplotypes and minimises the chance of inheritance of

two copies of the same haplotype in diploids [12]. This leads to elevated levels of variation at CSD loci compared to the rest of the genome, also facilitating genome localisation of CSD.

Bees and ants share a common ancestor ~153 million years ago [13,14]. We know little about the diversity of sex-determination loci within these groups (Figure 1A). Analysis of species related to *A. mellifera* has revealed the presence of diverse haplotypes in the *csd* ortholog in other members of the *Apis* genus [15]. However, an ortholog of honeybee *csd* has not been identified in members of the closely related genera *Bombus* and *Melipona* [16]. This indicates that the *csd* locus is likely a recent innovation in honeybees. The mechanism of sex determination in almost all of the >20,000 bee species is therefore unknown.

Here we genetically map the CSD locus in the red mason bee *Osmia bicornis* using freely-mating populations. This species belongs to the insect family Megachilidae, which is the most closely related to the Apidae (including the genera *Apis*, *Bombus* and *Melipona*) [13,14,17,18]. We phenotypically characterised bees from nesting tubes to identify potential diploid males. We then used whole genome sequencing to identify the sex-determination locus based on homozygosity in diploid males and elevated genetic variation. Finally, we analyse genome-wide variation in heterozygosity in additional genera of bees (*Bombus*, *Anthophora*, *Lasioglossum*) and ants (*Formica* and *Solenopsis*) in order to search for CSD loci in these taxa.

Results

Identification of diploid males in natural populations of O. bicornis

We morphologically characterised *O. bicornis* individuals collected from 683 nesting tubes from two locations in Europe (Uppsala, Sweden and Halle, Germany). *O. bicornis* nests typically contain males closest to the entrance and females at the back, with males being lower in mass (Figure 1C). Potential diploid males were identified based on their occurrence among females in nests or due to elevated mass, or both (see Materials and Methods). We also collected females from nests containing potential diploid males, and from randomly selected nests. In total, we performed whole genome sequencing of 162 individuals (Supplemental Tables S1-S3). These comprised 88 potential diploid males and 74 females.

We mapped sequencing reads to the *O. bicornis* reference assembly iOsmBic2.1 to an average sequencing depth of 13X. We masked heterochromatic regions from the dataset (see Methods, Supplemental Figure S1A,B and Supplemental Table S1), resulting in a dataset consisting of 3,573,693 SNPs. We determined the ploidy of our samples by presence of a significant number of heterozygous sites based on the inbreeding coefficient, *F,* which is expected to equal one in haploids (when analysed in the same way as diploids). All females were confirmed as diploid. Out of the 88 potential diploid males, 18 were confirmed to be diploid, whereas 70 were haploid. Our final dataset therefore contained 18 diploid males, 70 haploid males and 74 females (Supplemental Figure S1C,D).

Females from nests containing diploid males had a slight but significantly higher *F* than those sampled from random nests (mean $F = 0.452$ compared to 0.230, Mann-Whitney U test, $p = 0.0042$). However, we found no significant difference in *F* between diploid males and their sisters (mean $F = 0.298$ compared to 0.452, $p = 0.356$). Overall, this indicates that matings that produce diploid males are between individuals that are more closely related than average, but are not close relatives.

After determining the ploidy of all potential diploid males, we were able to estimate the frequency of matched matings in the population in which the mother shares an allele at the CSD locus with the father, leading to diploid male production. We corrected for the probability that a diploid male is not observed in a matched mating based on the number of progeny in each nest. In total, we observed 6 nests containing diploid males from a total of 672 nests found in the nesting tubes. We estimate a matched mating frequency of 1.04%

(SD 0.44%; Supplemental Table S4). This frequency did not differ significantly between locations or sampling year (Fisher's exact test, p>0.05 for all comparisons). The mass of confirmed diploid males was not statistically different from female nestmates, or from all females (Supplemental Figure S2A; t-test, p>0.05).

Candidate CSD locus orthologous to ANTSR identified on chromosome 1

We next analysed the genomes of diploid males and females for runs of homozygosity (ROH). The percentage of the genome in ROH blocks >10 kbp ranged from 0.19% to 67.4% between samples (Supplemental Figure S2B). The distribution of this proportion was similar between diploid males and females. There was a slight but non-significant tendency for females from nests with diploid males to have higher homozygosity (average proportion of genome in ROH blocks >10 kbp for females from nests with diploid males = 34.4%, nests without diploid males = 31.6%; Supplemental Figure S2B).

We then used the genomic distribution of ROH to identify candidate regions for CSD. We divided the ROH into genomic segments of uneven size across all individuals. Using a linear model, we associated the presence or absence of ROH segments across the genome with the sex of the diploid individuals. This analysis identified a single 2 kbp candidate region on chromosome 1 with coordinates 14,404,983 - 14,406,957 in which the presence of ROH is highly associated with sex (Figure 2A; $p < 10^{-16}$ after adjustment for multiple testing). All diploid males, and none of the females, are completely homozygous across this region. This is the only region of the genome where strong differentiation is observed in our dataset and is therefore likely to contain the CSD locus.

Genomic regions responsible for CSD are expected to evolve under frequency-dependent selection, which results in the presence of a large number of divergent alleles to avoid the occurrence of homozygosity in diploids [15]. This process leads to high levels of nucleotide diversity at CSD loci. We found that the candidate CSD region coincides with a local maximum in nucleotide diversity estimated in 10 kbp windows across the genome (Figure 2B). Nucleotide diversity in the region is 0.017, which is an order of magnitude higher than the chromosome average ($mean_{CHR}$=0.0014, $stdev_{CHROM}$=0.00087). It is the highest peak on chromosome 1, the fourth highest peak genome-wide, and in the top 99.86% of 10 kbp bins genome-wide. This supports the action of frequency-dependent selection in the region.

The candidate CSD region overlaps a single uncharacterised lncRNA in the genome annotation (Figure 2C,D,E). The identity and order of genes flanking this lncRNA are identical to those flanking *ANTSR*, the lncRNA implicated in sex determination in the ant *L. humile* [10]. These include COPA, CRELD2, miR-315, and THUMPD3 (Figure 2E). This indicates that the CSD locus identified here overlaps the *ANTSR* ortholog. ANTSR has 2 exons annotated in the *O. bicornis* genome assembly and the 2 kbp CSD locus overlaps 272 bp of the intron, 354 bp of exon 2 and 1348 bp of downstream sequence.

High genetic diversity in CSD concentrated in specific subregions

In total we identified 197 SNPs and 26 indels in the CSD locus. We identified 188 distinct haplotypes in the full sample of 254. A tree showing the relationships between haplotypes reveals high diversity across several clades (Figure 3A). We examined the distribution of heterozygous sites in females across the CSD region to attempt to identify critical regions where heterozygosity is required to activate female development (Figure 3B). Heterozygosity in the *ANTSR* exons within the CSD locus does not seem to be required to trigger female development as 7/74 females do not have any heterozygous sites in the overlap between the candidate region and exonic regions. The distribution of heterozygous sites in females is variable across the CSD locus. We define three regions where heterozygosity is particularly elevated (Supplemental Table S5): in an intron (P1), overlapping the final exon and downstream noncoding region (P2), and a downstream noncoding region (P3). These regions are particularly notable for a high number of indel polymorphisms. These candidate

173 functional regions could potentially affect female development by influencing transcription
174 levels of *ANTSR* (Supplementary Table S6).

175 We found that the two haplotypes present in females at this locus differ on average at 24.6
176 SNPs and 9.5 indels, but there is a large variation in this number (Figure 3C). We examined
177 the importance of heterozygosity in the three regions we defined (Supplemental Figure S3).
178 We identified 11 females that were homozygous across one of these regions and one that
179 was invariable across two regions (Supplemental Table S6). The minimum number of
180 differences observed between the two haplotypes in a female in our dataset is 3 SNPs,
181 which is found in two individuals. These observations suggest that none of these regions is
182 critical for determining sex on its own, and that heterozygosity at one or more of a large
183 number of critical sites across these regions is required for triggering female development,
184 indicative of genetic heterogeneity.

185 The majority of haplotypes (n = 152) are seen only once in the sample, with 36 haplotypes
186 observed multiple times (Figure 3D). The frequency of diploid male production in a randomly
187 mating population is expected to be approximately 1/k, where k is the number of sex alleles
188 [4]. In our populations this proportion is ~0.53% (Supplemental Table S4), which is the same
189 as the theoretical expectation of 0.53% assuming that there are 188 haplotypes at the CSD
190 locus, although it is likely that additional haplotypes exist in the population that are not
191 observed in our sample.

192 *Elevated nucleotide diversity implicates ANTSR in sex-determination in other bee and ant*
193 *species*

194 We identified the location of the *ANTSR* ortholog in the genomes of nine other bee species
195 (*Apis mellifera*, *Anthophora quadrimaculata, Anthophora retusa, Anthophora plagiata,*
196 *Bombus monticola, Bombus lapponicus, Bombus balteatus*, *Lasioglossum albipes*, and
197 *Lasioglossum pauxillum*) and three ant species (*Formica polyctena*, *Formica aquilonia* and
198 *Solenopsis invicta*; Figure 4A,B; Supplemental Table S7) In each case, we could identify a
199 candidate *ANTSR* ortholog including the same set of flanking genes as in *O. bicornis*,
200 indicating a conserved genomic neighbourhood.

201 We assayed variation in heterozygosity in 10 kbp windows across the genomes of these
202 species. There is no elevated variation in the orthologous region of *Apis mellifera*, in which
203 sex determination is known to be controlled at a locus unrelated to *ANTSR* [8,16]. In all other
204 species of bees and ants that we investigated, we find markedly elevated genetic variation in
205 the vicinity of the *ANTSR* ortholog (Supplemental Table S8). In *A. quadrimaculata, A. retusa,*
206 *B. monticola, B. lapponicus*, *L. pauxillum* and *L. albipes,* we observed a peak of elevated
207 genetic variation in the 10 kbp window containing *ANTSR* (genetic variation in the top 0.3%
208 of windows genome wide). In *A. plagiata*, *B. balteatus*, *S. invicta, F. polyctena* and *F.*
209 *aquilonia,* such a peak was found between 0.5 and 45 kbp (Supplementary Table S8), away
210 from *ANTSR*.

211 These findings are consistent with the action of frequency-dependent selection in the vicinity
212 of *ANTSR* in all of the bee and ant species we investigated, with the exception of *Apis*
213 *mellifera*. This lncRNA gene is therefore likely to be the sex-determining locus in all of these
214 species in addition to *Osmia bicornis*. Together, our findings suggest that *ANTSR* is the
215 predominant sex-determining gene in bees and ants, and likely had this role in their common
216 ancestor ~153 million years ago [13,14].

217

218 **Discussion**

219 Single-locus complementary sex determination (CSD) is found widely in Hymenoptera and is
220 likely to be the ancestral mode of sex determination [4,5,7,15]. CSD loci have been
221 characterised in one ant (*L. humile*) and one bee (*A. mellifera*) species [8,10], mapping to
222 two non-orthologous loci. Here we identify the CSD locus in the solitary bee *O. bicornis* and

223 analyse haplotype variation at this locus. Our main results are: 1) a single CSD locus in *O.*
224 *bicornis* maps to an ortholog of the *ANTSR* lncRNA gene, which is known to determine sex
225 in the ant *L. humile*, 2) a large number of highly divergent haplotypes are observed at the *O.*
226 *bicornis* CSD locus, with elevated diversity in specific sub-regions, 3) Elevated levels of
227 nucleotide diversity are found in or near orthologs of the *O. bicornis* CSD locus in three other
228 bee genera: *Anthophora*, *Bombus* and *Lasioglossum*, and two ant genera: *Formica* and
229 *Solenopsis*. Together the results indicate that *ANTSR* is likely to be the ancestral sex-
230 determination gene of bees and ants and could be the predominant mechanism of sex
231 determination in Hymenoptera.

232 Downstream sex determination in insects acts through expression of alternative splice
233 variants of the gene transformer (*tra*) or a related homolog, leading to male or female
234 development [19]. In most Hymenoptera, splicing of *tra* is determined by zygosity at a CSD
235 locus [4,7]. Two possible modes of action of CSD loci were originally suggested: 1) formation
236 of a heteropolymer consisting of the expressed product of two alleles at CSD, and 2)
237 production of mRNA defective in one or more cistrons, with different alleles containing
238 different defective cistrons [4]. The first mechanism has been demonstrated in honeybee,
239 where formation of a heterodimer of the Csd protein is required to generate alternative splice
240 variants of *fem* (ortholog of *tra*) [9].

241 The CSD locus identified here overlaps the lncRNA *ANTSR*, which was identified as the sex-
242 determining locus in the ant *L. humile* [10]. This locus has also been implicated in sex
243 determination by a genetic mapping study in the clonal raider ant, *Ooceraea biroi* [20], which
244 also identified orthology to a CSD QTL locus previously found in the ant *Vollenhovia emeryi*
245 [21]. Its mechanism of action is not fully understood, but it is not consistent with either of the
246 two originally proposed mechanisms above. In *L. humile, ANTSR* has four exons and
247 encodes a 793-bp transcript. The CSD locus was identified as a 5 kbp region overlapping
248 two of the *ANTSR* exons. The exonic sequences are relatively low in variation, and are
249 identical in some haplotypes, indicating that heterozygosity in exons does not generate the
250 sex determination signal. Variability in CSD is mainly restricted to downstream flanking and
251 intronic sequence of *ANTSR* and heterozygosity in these regions appears to regulate
252 expression of *ANTSR*, which in turn affects splicing of *tra* [10]. Our findings in *O. bicornis* are
253 consistent with those in *L. humile*. The *ANTSR* ortholog in *O. bicornis* has only two exons,
254 which also completely lack variability in some females, with variability in CSD also focussed
255 on intronic and downstream flanking sequence. It therefore seems likely that heterozygosity
256 at the CSD locus increases expression of *ANTSR* in both *O. bicornis* and *L. humile* [10].

257 We identified 188 distinct haplotypes in the 2 kbp CSD identified in *O. bicornis*. The high
258 haplotype diversity found in *O. bicornis* makes it possible to more precisely define
259 hypervariable regions of CSD that might be important in its function. We identified three sub-
260 regions with particularly elevated density of SNPs and indels, which overlap an intron and
261 downstream noncoding sequence. However, we identified females that lack variability in one
262 or two of these regions, indicating that heterozygosity at a small number of many possible
263 critical sites distributed across these three sub-regions is sufficient to trigger female
264 development. A plausible mechanism that has been proposed is transvection [10], where the
265 regulatory region of a gene can influence expression in *trans*. This mechanism is likely to be
266 common in insects in which homologous chromosomes remain paired in somatic cells [22]
267 and is responsible for sex-biased expression of an X-linked gene in *Drosophila* [23].

268 Only seven haplotypes were identified at the 5 kbp CSD locus identified in *L. humile* [10],
269 likely because the population of *L. humile* used to map CSD was highly inbred, which leads
270 to a high frequency of diploid male production. The production of diploid males represents a
271 significant cost to complementary sex determination, particularly in inbred populations. This
272 cost could be partly ameliorated by more efficient purging of deleterious mutations in
273 haplodiploids [24]. We estimate that diploid male production occurs at a rate of ~0.5% of
274 matings in our *O. bicornis* population. This low level of diploid male production is consistent

with high haplotype diversity, but nonetheless represents a substantial genetic load associated with complementary sex determination.

We have identified elevated genetic variation close to *ANTSR* in 13 species belonging to six bee and ant genera: *Anthophora, Bombus, Lasioglossum, Osmia, Formica,* and *Solenopsis*. This indicates that the CSD locus is associated with *ANTSR* in these genera. The observation that the peak does not always correspond exactly to *ANTSR* (e.g. in *A. quadrimaculata, B. balteatus, F. polyctena, F. aquilonia,* and *S. invicta*), supports the view that CSD is a non-coding regulatory element that regulates *ANTSR*, which has shifted location in evolution.

Our results are consistent with previous research demonstrating that the CSD locus in *Apis mellifera* (*Csd*) is a recent innovation, which arose from a duplication of the *feminizer* gene [16]. As *Bombus* is a close relative of *Apis* (~54 MYA split) [14,25] but likely uses *ANTSR* for sex-determination, it is likely that it represents the ancestral state. Taken together, our results suggest that *ANTSR* is the ancestral sex determination locus of Apoidea (bees and sphecoid wasps) and Formicidea (ants). These superfamilies encompass around 50,000 species, or around one third of Hymenoptera. It is also plausible that *ANTSR* could determine sex much more widely in Hymenoptera.

Materials and Methods

Sample collection

O. bicornis is a univoltine, protandrous, sexually size-dimorphic, polylectic, and monandrous spring bee. Each individual hibernates as an adult inside a cocoon in the maternal nest. Females construct typically mixed nests of line type containing both sexes. A series of larger cells for daughters is constructed at the back of the tube and a series of smaller cells for sons at the entrance [26] (Figure 1C). Body size of an individual is limited by the provisions in the brood cell [27]. Males developing in a female-dedicated brood cell with more provisions reach a female-like body mass. This situation occurs when the eggs are diploid male (due to homozygosity at the CSD locus) or due to fertilisation errors, which result in a haploid egg. The rate of fertilisation errors has previously been estimated to be about 6% in female-dedicated eggs [28].

Diploid males occur rarely and are detectable as adults with phenotypic male characteristics in female-dedicated brood cells. We analysed the contents of 683 *Osmia bicornis* nests from the botanical gardens Halle (Saale), Germany and from various locations in the vicinity of Uppsala, Sweden during the period 2021-2023 to identify potential diploid males. The nests were contained in Japanese knotweed in Halle (internal diameter 8-10 mm, length 133.4 ± 13.2 mm) or cardboard tubes in Uppsala (internal diameter 7 mm, length 150 mm). They were analysed during the winter, when adult bees are dormant inside cocoons.

Males and females were distinguished by morphological examination of the contents of each cocoon. Bees with black hairs on the clypeus were considered female, whereas the presence of long white hairs indicated a male. Potential diploid males were identified as morphological males that satisfied these criteria: 1) presence in a female-dedicated brood cell or 2) mass greater than 0.1 g, or both. Ploidy of all potential diploid males was determined by subsequent genome sequencing. We also collected females from nests containing potential diploid males and from random nests for sequencing.

Library preparation and sequencing

Samples were stored in 95% ethanol at -20 °C prior to DNA extraction. They were dissected to extract wing muscle from the thorax. DNA was extracted using the Qiagen DNeasy Blood & Tissue Kit. We performed sequencing library preparation using the Illumina Nextera Flex kit. Sequencing was performed using an Illumina NovaSeq 6000 (output per sample shown on Supplemental Table S1).

Read mapping, variant calling and filtering

Reads from all samples were mapped to the *Osmia bicornis* genome assembly iOsmBic2.1 using bwa v0.7.17 [29]. The resulting files were sorted and indexed using samtools 1.19 [30] before using Picard v2.25.0 [31] to mark duplicates and add read groups. Subsequently, GATK HaplotypeCaller v4.2.0.0 [32] was used to call variants. We considered all samples as diploids for variant calling. The software vcftools v0.1.16 [33] was used to filter the dataset, retaining only biallelic SNPs, MQ>20, AF≥0.01, as well as missingness<0.1. We also filtered markers for substandard or excess depth (1000>DP>3000) and quality by depth (QD>20).

We noted that in certain regions of the genome, genotype calls had a large excess of heterozygosity, read mapping depth was significantly higher than average and quality-by-depth was significantly lower (Supplemental Figure S1A,B). This observation is consistent with the presence of long stretches of highly-repetitive heterochromatin in the genome assembly, in which genetic variants cannot be reliably assayed. This has been reported in a wide range of other genomes [34]. We identified regions with suspected heterochromatin based on elevated sequencing depth, reduced quality-by-depth and excess heterozygosity. To define these regions, we ran vcf-tools with the *--hardy* option, plotting -log(P_{het_excess}). We manually excluded regions of excess heterozygosity that were observed at the distal ends of each of the 16 chromosomes (Supplemental Figure S1; Supplemental Table S9). These blocks were excluded for the determination of ploidy and all further genome-wide summary statistics. Additionally, we filtered all markers identified as heterozygous in at least two haploid individuals from the dataset for mapping in the diploid individuals. The regions comprise 23.8% of the total assembled chromosomes. We also excluded 3 samples due to low average sequencing depth (<2 million reads).

Identification of ploidy

We used the --het option in vcftools [33] to calculate the inbreeding coefficient F, the ratio of observed and expected homozygotes per individual. This statistic revealed the ploidy of all samples, and determined whether potential diploid males were haploid or diploid. The male samples clustered into clear groups, with those with ($F > 0.95$) considered to be haploid males due to lack of heterozygous sites (Supplemental Figure S1C,D).

Estimation of the frequency of matched matings

We used the number and distribution of confirmed diploid males among nests to estimate the frequency of matched matings in the populations. A matched mating is defined as one where a female mates with a male that shares one of its CSD haplotypes, which results in a probability of 0.5 that each fertilised egg will become a diploid male. We estimated the proportion of matched matings by dividing the number of nests containing diploid males with the total number of nests containing females, with a correction based on the likelihood of detecting a matched mating based on independent segregation of alleles at a single CSD locus. This likelihood is defined as $d = 1 - 0.5^n$, where n is the number of diploid offspring present in a nest. We used this estimate to define the effective number of nests, by summing the values of d across all nests analysed according to the number of diploid offspring per nest (Supplemental Table S4). This was performed using data collected in Halle over two consecutive years, and from one year in Uppsala.

Differences in mass

Using the ploidy confirmed by sequencing and the recorded sex, we tested for mass differences between cocoons containing females, haploid males and diploid males. To compare diploid males to their female nestmates, we used a subset of 15 diploid males with multiple sequenced sisters (n = 37). We used statsmodels in python to make a linear mixed effect model to test for an association between mass in grams and sex encoded as 0 and 1, with NestID as a random effect. We tested for differences between the mean mass of all confirmed diploid males (n = 18), females (n = 1745), and haploid males, for which data was available (n = 1659), using two-sided unequal variance t-tests.

SNP density and nucleotide diversity

Nucleotide diversity (**π**) and SNP density were calculated in 10 kbp windows basis using vcftools [33] with the --window-pi option. For fine mapping nucleotide diversity in the identified putative CSD region, **π** was calculated per site using vcftools and the --site-pi option, and subsequently evaluated in a 1 kbp-window rolling mean using numpy v1.26.0 [35].

Identification of runs of homozygosity associated with sex

Runs of homozygosity (ROH) were calculated for each individual using plink v1.9 [36] with the *---homozyg group --homozyg-window-snp 50 --homozyg-window-het 3 --homozyg-window-missing 5 --homozyg-window-threshold 0.05 --homozyg-snp 10 --homozyg-kb 5 --allow-extra*-chr and otherwise default options. Then, using pandas (2.2.2; [37]) and numpy (1.26.0; [35]) in python 3.10, we segmented the resulting ROH into a binary square matrix across the individuals and the genome, with the borders across the latter being defined by the beginning or end of a ROH within any of the investigated individuals. Subsequently, we tested the association between presence or absence of a ROH segment with sex in diploid males and females with a linear model, using statsmodels (0.14.2; [38]) and controlled for the rate of false positive testing using the Benjamini-Hochberg [39] procedure implemented in statsmodels with $\alpha = 0.05$.

Haplotype analysis

In order to investigate the haplotype in the candidate region, we extracted the highest scoring ROH-block (1974 bp), phased it using beagle v5.5 [40], and generated a sequence in fasta format for each haplotype. These were subsequently aligned using MAFFT (v7.520; [41]) and used to generate a tree using RAxML v8.2.12 [42], which was plotted using FigTree v1.4.4 [43].

Comparison with other species

We used blast hits identified with the blastn tool of BLAST+ v. 2.16.0 [44] to locate the ortholog of *ANTSR* in several other bee species for which genetic variation data were available. We created a query list using the *ANTSR* sequence of *L. humile* and its orthologous sequences in *A. mellifera* (LOC100576403) and *Bombus affinis* (LOC125386106) reported by Pan et al. [10]. We also included the *O. bicornis* lncRNA sequence studied here. We blasted the four query sequences against two *Bombus*, three *Anthophora,* two *Lasioglossum and three ant* genome assemblies. These were *Bombus sylvicola* (BioProject PRJNA646847) [45]*, Bombus balteatus* (PRJNA704506) [46]*, Anthophora retusa, Anthophora quadrimaculata*, and *Anthophora plagiata* (PRJNA1255204) [47]. *Lasioglossum pauxillum* (PRJNA13330011), *Lasioglossum albipes (*PRJNA413432) [48], *Formica polyctena*, *Formica aquilonia* (PRJEB51899) [49] and *Solenopsis invicta* (PRJNA685290) [50]. The region with the highest query cover and identity was identified for each species.

We then examined the annotations of protein-coding and non-coding RNA genes in these regions for each target genome to look for the presence of an annotated lncRNA gene and conservation of the flanking genes. Functional annotations were not available for all of the species. We therefore performed functional annotation for these genome assemblies based on orthology using eggNOG-mapper v.2 [51]. This was performed with eggNOG-mapper default values against the eggNOG 5 database.

We identified the most likely position of ANTSR in each genome based on the annotation of a lncRNA in the vicinity of the best BLAST hit. A candidate gene was present in the annotation of some species, but not available for *A. quadrimaculata*, *L. pauxillum, L. albipes, F. polyctena, F. aquilonia* and *S. invicta*. For these species, the best BLAST hit was used as the putative location. For *A. mellifera*, we used the coordinates previously reported by Pan et al. [10]. We analysed genetic diversity in sequenced populations of each of these species.

We used 65 and 49 individuals for *Bombus lapponicus* and *Bombus monticola*, respectively, mapped to the reference genome of the closely related species *B. sylvicola* [45]. For *B. balteatus* we used 299 bees mapped to the *B. balteatus* genome [46]. We used 57, 43, and 27 individuals for *A. retusa, A. quadrimaculata*, and *A. plagiata* [47]. We used 55 *Apis mellifera* individuals (BioProject PRJNA357367) [52]. We used 71 individuals for *L. pauxillum*, 160 for *L. albipes,* 10 individuals for each of the two *Formica* species, and 368 for *S. invicta*. A summary of datasets is presented in Table S7. We used vcftools v.0.1.16 [33] to calculate genetic diversity based on the pairwise differences (π) in windows of 10 kbp across the genome using vcftools --window-pi. We generated a phylogeny of these species using BeeTree [25], plotted using FigTree 1.4.4 [43].

Data availability

All scripts and code associated with the project are available on GitHub (https://github.com/troe27/osmia-sex-locus). Illumina population sequencing data for *O. bicornis* are available at NCBI BioProject PRJEB88603. Data from *A. mellifera*, *Bombus* spp. and *Anthophora* spp. are available at NCBI BioProjects PRJNA1255204, PRJNA646847, and PRJNA357367 respectively. Data from *L. pauxillum* and *L. albipes* are available at NCBI BioProjects PRJNA13330011 and PRJNA413432 respectively. Data for the *Formica* species are at NCBI BioProject PRJEB51899, and *S. Invicta* at PRJNA685290. Spreadsheets of the contents of each nesting tube and cocoon mass are available as Supplementary Tables. Supplementary Data S01-S25 underlying the figures as well as a snapshot of the code can be found on Zenodo (DOI: 10.5281/zenodo.17135872).

Funding

This study was funded by Erik Philip-Sörensens Stiftelse. No funders had any role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors acknowledge support from the National Genomics Infrastructure in Stockholm funded by Science for Life Laboratory, the Knut and Alice Wallenberg Foundation and the Swedish Research Council, and NAISS/Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure. The computational analysis was enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

Author contributions

M.T.W. designed the project with input from B.C., R.J.P. and K.S. S.M., B.C., K.S. and M.T.W. analysed nest contents. A.O. and S.M. performed DNA extractions. T.R., D.T., L.H., M.N.I.B.N.F., T.E., P.T., K.S. and M.T.W. performed data analysis. T.R. and G.C.M.-H. produced the figures. T.R. and M.T.W. wrote the manuscript with input from K.S., B.C. and R.J.P.

Declaration of interests

The authors declare no competing interests.

Acknowledgements

We thank Adriana M Cintrón Santiago and Tuuli Larva for help with DNA extractions. We thank Hanna Sigeman and Miguel Carneiro for helpful comments on the manuscript.

475 **References**

476

477 1. Bachtrog D, Mank JE, Peichel CL, Kirkpatrick M, Otto SP, Ashman T-L, et al. Sex
478 Determination: Why So Many Ways of Doing It? PLOS Biology. 2014;12: e1001899.
479 doi:10.1371/journal.pbio.1001899

480 2. Beukeboom LW, Perrin N. The Evolution of Sex Determination. Oxford University
481 Press; 2014. doi:10.1093/acprof:oso/9780199657148.001.0001

482 3. de la Filia AG, Bain SA, Ross L. Haplodiploidy and the reproductive ecology of
483 Arthropods. Current Opinion in Insect Science. 2015;9: 36–43.
484 doi:10.1016/j.cois.2015.04.018

485 4. Cook JM, Crozier RH. Sex determination and population biology in the hymenoptera.
486 Trends Ecol Evol. 1995;10: 281–286. doi:10.1016/0169-5347(95)90011-x

487 5. Asplen MK, Whitfield JB, De Boer JG, Heimpel GE. Ancestral state reconstruction
488 analysis of hymenopteran sex determination mechanisms. Journal of Evolutionary
489 Biology. 2009;22: 1762–1769. doi:10.1111/j.1420-9101.2009.01774.x

490 6. Zou Y, Geuverink E, Beukeboom LW, Verhulst EC, van de Zande L. A chimeric gene
491 paternally instructs female sex determination in the haplodiploid wasp Nasonia.
492 Science. 2020;370: 1115–1118. doi:10.1126/science.abb8949

493 7. Cook JM. Sex determination in the Hymenoptera: a review of models and evidence.
494 Heredity. 1993;71: 421–435. doi:10.1038/hdy.1993.157

495 8. Beye M, Hasselmann M, Fondrk MK, Page RE, Omholt SW. The Gene csd Is the
496 Primary Signal for Sexual Development in the Honeybee and Encodes an SR-Type
497 Protein. Cell. 2003;114: 419–429. doi:10.1016/S0092-8674(03)00606-8

498 9. Otte M, Netschitailo O, Weidtkamp-Peters S, Seidel CAM, Beye M. Recognition of
499 polymorphic Csd proteins determines sex in the honeybee. Science Advances. 2023;9:
500 eadg4239. doi:10.1126/sciadv.adg4239

501 10. Pan Q, Darras H, Keller L. LncRNA gene ANTSR coordinates complementary sex
502 determination in the Argentine ant. Science Advances. 2024;10: eadp1532.
503 doi:10.1126/sciadv.adp1532

504 11. Bopp D, Saccone G, Beye M. Sex determination in insects: variations on a common
505 theme. Sex Dev. 2014;8: 20–28. doi:10.1159/000356458

506 12. Hasselmann M, Beye M. Signatures of selection among sex-determining alleles of the
507 honey bee. Proc Natl Acad Sci U S A. 2004;101: 4888–4893.
508 doi:10.1073/pnas.0307147101

509 13. Blaimer BB, Santos BF, Cruaud A, Gates MW, Kula RR, Mikó I, et al. Key innovations
510 and the diversification of Hymenoptera. Nat Commun. 2023;14: 1212.
511 doi:10.1038/s41467-023-36868-4

512 14. Peters RS, Krogmann L, Mayer C, Donath A, Gunkel S, Meusemann K, et al.
513 Evolutionary History of the Hymenoptera. Current Biology. 2017;27: 1013–1018.
514 doi:10.1016/j.cub.2017.01.027

515 15. Cho S, Huang ZY, Green DR, Smith DR, Zhang J. Evolution of the complementary sex-
516 determination gene of honey bees: Balancing selection and trans-species
517 polymorphisms. Genome Res. 2006;16: 1366–1375. doi:10.1101/gr.4695306

518 16. Hasselmann M, Gempe T, Schiott M, Nunes-Silva CG, Otte M, Beye M. Evidence for
519 the evolutionary nascence of a novel sex determination pathway in honeybees. Nature.
520 2008;454: 519–22. doi:10.1038/nature07052

521 17. Danforth BN, Cardinal S, Praz C, Almeida EAB, Michez D. The Impact of Molecular
522 Data on Our Understanding of Bee Phylogeny and Evolution. Annual Review of
523 Entomology. 2013;58: 57–78. doi:10.1146/annurev-ento-120811-153633

524 18. Danforth BN, Sipes S, Fang J, Brady SG. The history of early bee diversification based
525 on five genes plus morphology. Proceedings of the National Academy of Sciences.
526 2006;103: 15118–15123. doi:10.1073/pnas.0604033103

527 19. Gempe T, Beye M. Function and evolution of sex determination mechanisms, genes
528 and pathways in insects. BioEssays. 2011;33: 52–60. doi:10.1002/bies.201000043

529 20. Lacy KD, Lee J, Rozen-Gagnon K, Wang W, Carroll TS, Kronauer DJ. Heterozygosity
530 at a conserved candidate sex determination locus is associated with female
531 development in the clonal raider ant (Ooceraea biroi). eLife. 2025;14.
532 doi:10.7554/eLife.106913.1

533 21. Miyakawa MO, Mikheyev AS. QTL Mapping of Sex Determination Loci Supports an
534 Ancient Pathway in Ants and Honey Bees. PLOS Genetics. 2015;11: e1005656.
535 doi:10.1371/journal.pgen.1005656

536 22. Mellert DJ, Truman JW. Transvection Is Common Throughout the Drosophila Genome.
537 Genetics. 2012;191: 1129–1141. doi:10.1534/genetics.112.140475

538 23. Galouzis CC, Prud'homme B. Transvection regulates the sex-biased expression of a fly
539 X-linked gene. Science. 2021;371: 396–400. doi:10.1126/science.abc2745

540 24. Charlesworth D, Willis JH. The genetics of inbreeding depression. Nat Rev Genet.
541 2009;10: 783–796. doi:10.1038/nrg2664

542 25. Henríquez-Piskulich P, Hugall AF, Stuart-Fox D. A supermatrix phylogeny of the world's
543 bees (Hymenoptera: Anthophila). Molecular Phylogenetics and Evolution. 2024;190:
544 107963. doi:10.1016/j.ympev.2023.107963

545 26. Raw A. The biology of the solitary bee Osmia rufa (L.) (Megachilidae). Transactions of
546 the Royal Entomological Society of London. 1972;124: 213–229. doi:10.1111/j.1365-
547 2311.1972.tb00364.x

548 27. Seidelmann K. Optimal resource allocation, maternal investment, and body size in a
549 solitary bee, Osmia bicornis. Entomologia Experimentalis et Applicata. 2018;166: 790–
550 799. doi:10.1111/eea.12711

551 28. Raw A, O'Toole C. Errors in the Sex of Eggs Laid By the Solitary Bee Osmia rufa
552 (Megachilidae). Behaviour. 1979;70: 168–171. doi:10.1163/156853979X00043

553 29. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler
554 transform. Bioinformatics. 2009;25: 1754–1760. doi:10.1093/bioinformatics/btp324

30. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079. doi:10.1093/bioinformatics/btp352

31. Picard Tools. Available: http://broadinstitute.github.io/picard/

32. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20: 1297–1303. doi:10.1101/gr.107524.110

33. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27: 2156–2158. doi:10.1093/bioinformatics/btr330

34. Dallaire X, Bouchard R, Hénault P, Ulmo-Diaz G, Normandeau E, Mérot C, et al. Widespread Deviant Patterns of Heterozygosity in Whole-Genome Sequencing Due to Autopolyploidy, Repeated Elements, and Duplication. Genome Biology and Evolution. 2023;15: evad229. doi:10.1093/gbe/evad229

35. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585: 357–362. doi:10.1038/s41586-020-2649-2

36. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81: 559–575. doi:10.1086/519795

37. McKinney W. Data Structures for Statistical Computing in Python. scipy. 2010 [cited 26 May 2025]. doi:10.25080/Majora-92bf1922-00a

38. Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. Austin, Texas; 2010. pp. 92–96. doi:10.25080/Majora-92bf1922-011

39. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57: 289–300.

40. Browning BL, Tian X, Zhou Y, Browning SR. Fast two-stage phasing of large-scale sequence data. The American Journal of Human Genetics. 2021;108: 1880–1890. doi:10.1016/j.ajhg.2021.08.005

41. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution. 2013;30: 772–780. doi:10.1093/molbev/mst010

42. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30: 1312–1313. doi:10.1093/bioinformatics/btu033

43. FigTree. [cited 9 May 2025]. Available: https://tree.bio.ed.ac.uk/software/figtree/

44. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421. doi:10.1186/1471-2105-10-421

45. Christmas MJ, Jones JC, Olsson A, Wallerman O, Bunikis I, Kierczak M, et al. Genetic Barriers to Historical Gene Flow between Cryptic Species of Alpine Bumblebees Revealed by Comparative Population Genomics. Mol Biol Evol. 2021;38: 3126–3143. doi:10.1093/molbev/msab086

46. Christmas MJ, Jones JC, Olsson A, Wallerman O, Bunikis I, Kierczak M, et al. A genomic and morphometric analysis of alpine bumblebees: Ongoing reductions in tongue length but no clear genetic component. Mol Ecol. 2022;31: 1111–1127. doi:10.1111/mec.16291

47. Taliadoros D, Soares AER, Dias G, Bunikis I, Pippel M, Olsson A, et al. Genome variation in three Anthophora bee species reflects divergent demographic histories. [cited 5 May 2025]. Available: https://www.authorea.com/users/917380/articles/1290047--1-p-1-genome-variation-in-three-anthophora-bee-species-reflects-divergent-demographic-histories?commit=f707f2073ef4098c25ec5395b7ab7ee843d86513

48. Jones BM, Webb AE, Geib SM, Sim S, Schweizer RM, Branstetter MG, et al. Repeated Shifts in Sociality Are Associated With Fine-tuning of Highly Conserved and Lineage-Specific Enhancers in a Socially Flexible Bee. Mol Biol Evol. 2024;41: msae229. doi:10.1093/molbev/msae229

49. Portinha B, Avril A, Bernasconi C, Helanterä H, Monaghan J, Seifert B, et al. Whole-genome analysis of multiple wood ant population pairs supports similar speciation histories, but different degrees of gene flow, across their European ranges. Molecular Ecology. 2022;31: 3416–3431. doi:10.1111/mec.16481

50. Stolle E, Pracana R, López-Osorio F, Priebe MK, Hernández GL, Castillo-Carrillo C, et al. Recurring adaptive introgression of a supergene variant that determines social organization. Nat Commun. 2022;13: 1180. doi:10.1038/s41467-022-28806-7

51. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol. 2021;38: 5825–5829. doi:10.1093/molbev/msab293

52. Wallberg A, Schöning C, Webster MT, Hasselmann M. Two extended haplotype blocks are associated with adaptation to high altitude habitats in East African honey bees. PLoS Genetics. 2017;13: e1006792. doi:10.1371/journal.pgen.1006792

53. Romiguier J, Borowiec ML, Weyna A, Helleu Q, Loire E, Mendola CL, et al. Ant phylogenomics reveals a natural selection hotspot preceding the origin of complex eusociality. Current Biology. 2022;32: 2942-2947.e4. doi:10.1016/j.cub.2022.05.001

Figures

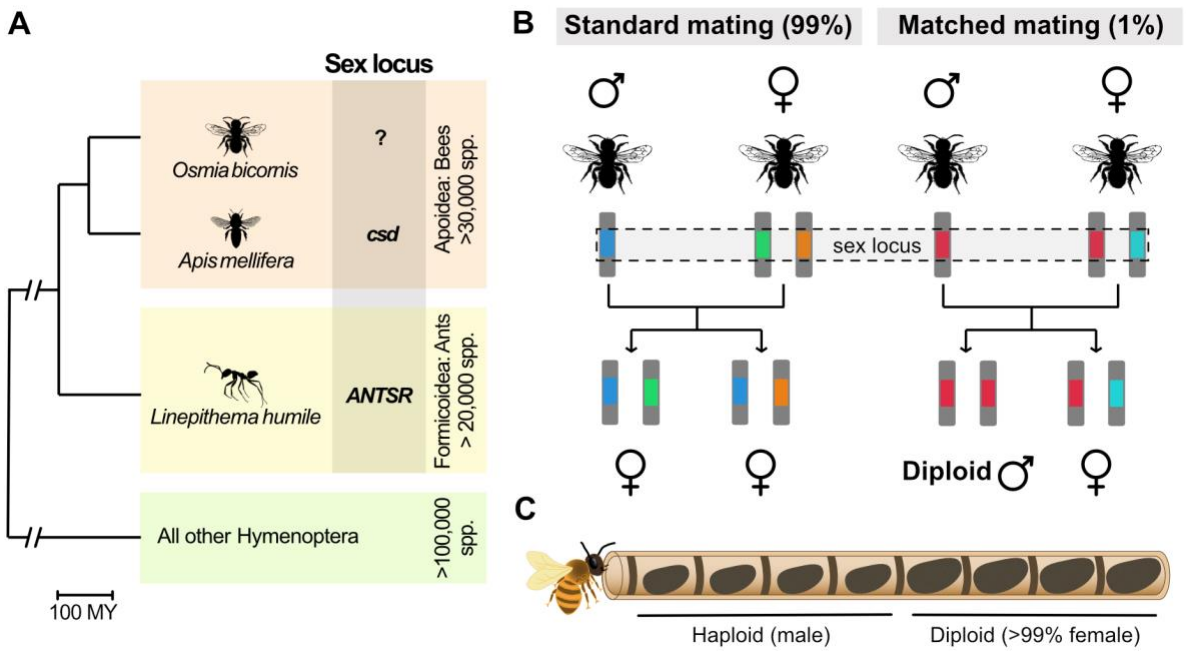

Figure 1: Complementary sex determination (CSD) in bees and ants: A. Single-locus CSD is likely to be the predominant mode of sex determination in Hymenoptera. CSD loci have been identified in two species: the honeybee *Apis mellifera* and the invasive ant *Linepithema humile*. The tree shows evolutionary relationships between these species and *Osmia bicornis*, the main subject of this study. The origin of Hymenoptera is estimated as 386 million years ago and the common ancestor of Apoidea and Formicoidea is estimated as 153 million years ago [14]. **B.** Under single-locus CSD, females are heterozygous at the CSD locus, whereas males are haploid and therefore hemizygous. A large number of CSD alleles are expected to segregate in outbred populations. A small proportion of homozygotes at the CSD locus are produced by chance, when parents share an allele, which results in the production of sterile diploid males. **C.** An *O. bicornis* nest consists of a row of cocoons in a tube. The mother lays fertilised diploid eggs deepest in the nest and unfertilised haploid eggs closer to the entrance. Eggs develop into adults inside cocoons. Diploid eggs typically develop into females whereas haploid eggs become male. A small proportion of diploids that are homozygous at the CSD locus develop into diploid males, which occur amongst females and are larger than haploid males. The silhouettes for *Osmia bicornis*, *Apis mellifera* and *Linepitherma humile* were obtained from PhyloPic (https://www.phylopic.org ,T. Michael Keesey, 2025) and were contributed by Thomas Hegna (2012; Public Domain Mark 1.0 Universal), Melissa Broussard (2023; Attribution 4.0 International), and Jitte Groothuis (2022; Attribution 4.0 International).

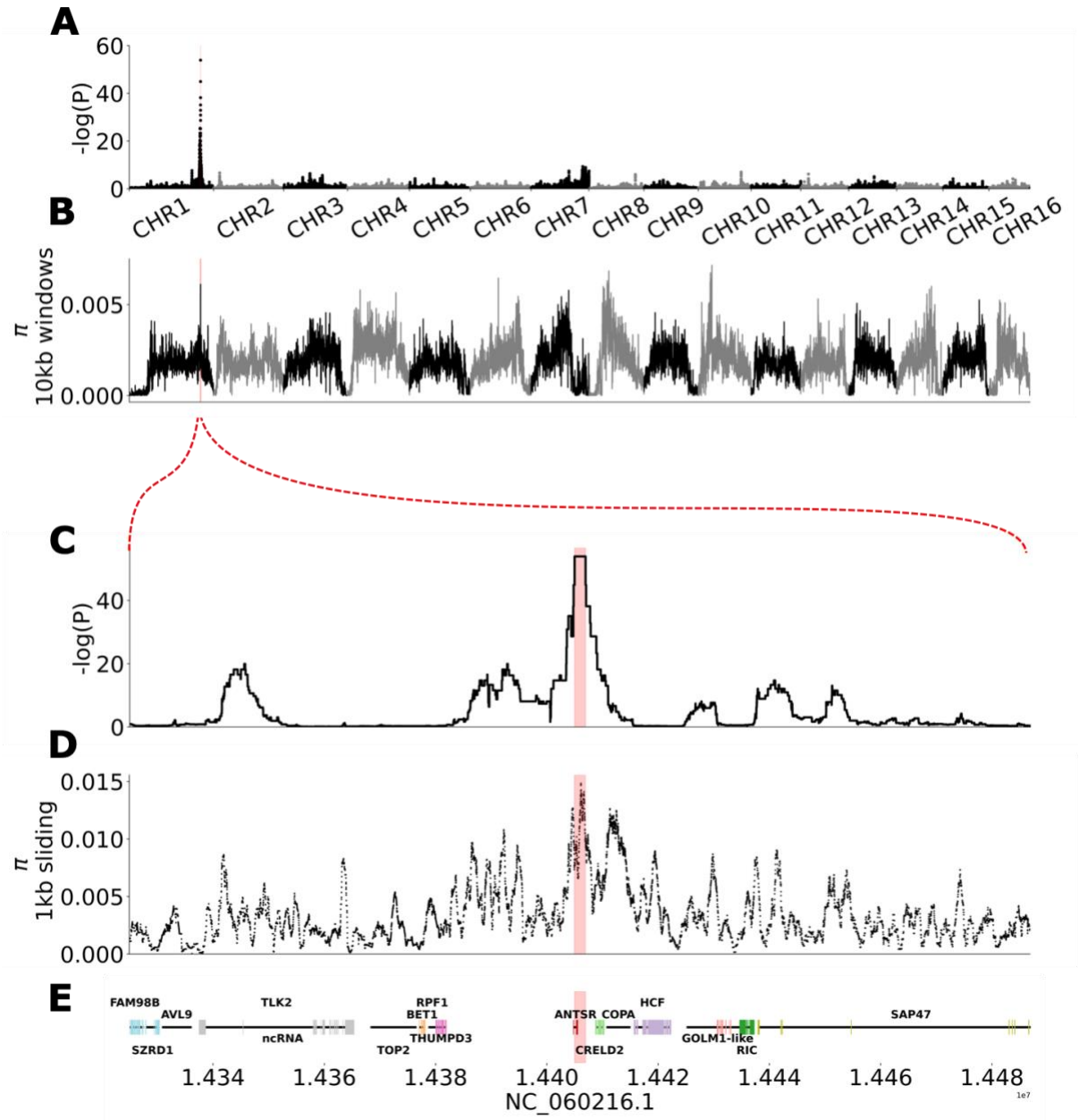

Figure 2: Mapping the CSD locus. A. Association of presence/absence of runs of homozygosity (ROH) with sex in diploid samples using a linear model (sex ~ ROH; n = 18 diploid males, n = 74 females; correction for genome wide significance using Benjamin-Hochberg procedure). **B.** Nucleotide diversity in 10 kbp windows across the genome. **C**. Close-up of association results on chromosome 1. The red curly bracket between panel B and C indicates the location on Panel A and B. **D.** Nucleotide diversity across the candidate region estimated in 1 kbp sliding windows. **E.** Genomic features in and around the candidate region. The red vertical bar indicates the best-scoring ROH segment across all subplots. Data from panels A-E can be found in Supplemental Data S01-S03 and Supplementary Table S5.

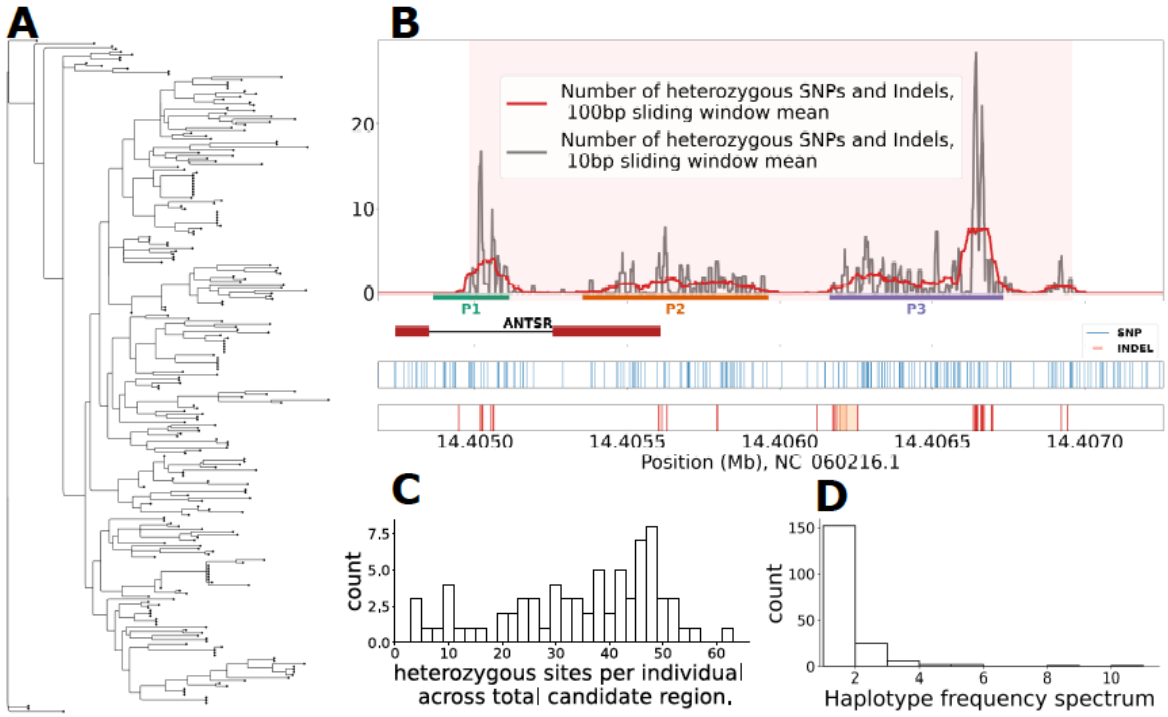

Figure 3: Haplotypes across the 2 kbp CSD locus. A. Maximum-likelihood tree of all unique haplotypes across the mapping population (74 females, 70 haploid males, 18 diploid males, 254 total haplotypes). Circles at branch tips indicate the number of observations of each haplotype. **B.** Count of differences between the two haplotypes carried by females across the candidate region. The top panel shows differences across all females, with the candidate region highlighted in light red. Three regions with larger differences within the region are labeled P1, P2, P3, highlighted with blue, green and orange bars at the bottom of the panel, respectively. The middle panel shows how ANTSR is located within the candidate region. The bottom two panels show a rug plot of heterozygous SNPs across the region in blue, with indel starting positions shown as red bars, and their extent in orange. **C.** Histogram of the total number of heterozygous sites across the candidate region per individual. **D.** Histogram of the haplotype frequency spectrum showing the number of times each haplotype is observed in the population samples. Data for panels A,B,C and D can be found in Supplementary Data S04-S07.

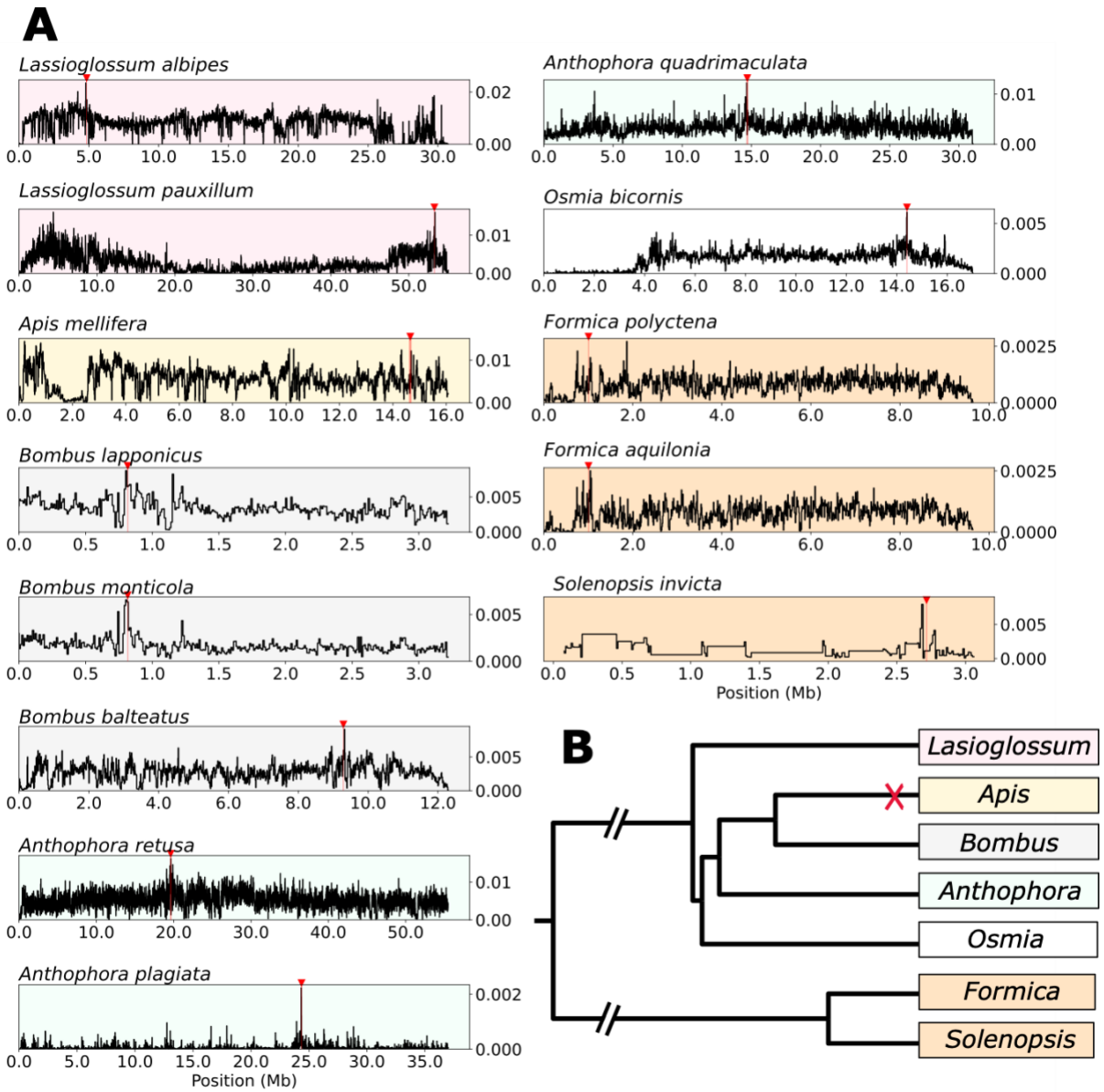

Figure 4: Nucleotide diversity around *ANTSR* in other bee and ant species. A: Variation in nucleotide diversity (π) is shown in 10 kbp windows across the contig containing *ANTSR* in each of the following species**:** *Lasioglossum albipes* and *Lasioglossum pauxillum* (both shaded red) *Apis mellifera* (shaded yellow)*, Bombus lapponicus*, *Bombus monticola*, *Bombus balteatus* (all *Bombus* shaded grey), *Anthophora retusa*, *Anthophora plagiata*, *Anthophora quadrimaculata* (all *Anthophora* shaded green), *Osmia bicornis (shaded white)* and three ant species *(Formica polyctena, Formica aquilonia and Solenopsis invicta,* all shaded orange*).* The vertical red line and red arrow above the plots indicate the location of the *ANTSR* ortholog for each species. **B:** Phylogenetic tree indicating the relationship between the investigated genera. The tree is based on the distance matrix from BeeTree [25], and Romiguier et al. [53]. The data underlying the nucleotide diversity plots can be found in Supplementary Data S02 and S08-S19

Supplementary Figures

Figure S1: Presence of heterochromatin at the distal ends of chromosomes and its effect on the inbreeding coefficient

Figure S2: Plots of weight and ROH coverage for sampled male and female *Osmia bicornis*

Figure S3: Number of heterozygous sites across the candidate region

Supplementary Tables

Table S1: Summary of all samples sequenced

Table S2: Summary of *Osmia bicornis* cocoons containing potential diploid males collected in Halle, Germany over 2 consecutive years

Table S3: Summary of *Osmia bicornis* cocoons containing potential diploid males collected in Uppsala

Table S4: Estimation of frequency of matched mating

Table S5: Coordinates of elements within the candidate region of *Osmia bicornis*

Table S6: Number of heterozygous sites (SNP and INDELS) per female individual across peaks in candidate-region and total candidate region

Table S7: Information on Hymenoptera species used in comparative analysis.

Table S8: Nucleotide diversity in the Orthologous region to ANTSR, by species

Table S9: Estimated heterochromatin regions on each *Osmia bicornis* chromosome

Supplementary Data: available at zenodo (DOI:10.5281/zenodo.17135872)

Data S01: p-values from modelling sex in diploid individuals as a function of presence/absence of runs of homozygosity, adjusted by Benjamini-Hochberg procedure

Data S02: 10kb windowed nucleotide diversity across all *Osmia bicornis* females included in this study

Data S03: per-site nucleotide diversity across all *Osmia bicornis* females included in this study

Data S04: Raxml tree of all haplotypes found in the candidate region of *Osmia bicornis*.

Data S05: Frequency of all haplotypes found in the candidate region of *Osmia bicornis*.

Data S06: Number of heterozygous sites in the candidate region and subregions per female individual of *Osmia bicornis* included in this study.

Data S07: Imputed and phased candidate-region VCF-file.

Data S08: 10kb windowed nucleotide diversity across *Anthophora plagiata*

Data S09: 10kb windowed nucleotide diversity across *Anthophora quadrimaculata*

Data S10: 10kb windowed nucleotide diversity across *Anthophora retusa*

Data S11: 10kb windowed nucleotide diversity across *Apis mellifera*

Data S12: 10kb windowed nucleotide diversity across *Bombus balteatus*

Data S13: 10kb windowed nucleotide diversity across *Bombus monticola*

Data S14: 10kb windowed nucleotide diversity across *Bombus lapponicus*

Data S15: 10kb windowed nucleotide diversity across *Formica aquilonia*

Data S16: 10kb windowed nucleotide diversity across *Formica polyctena*

Data S17: 10kb windowed nucleotide diversity across *Lasioglossum albipes*

Data S18: 10kb windowed nucleotide diversity across *Lasioglossum pauxillum*

Data S19: 10kb windowed nucleotide diversity across *Solenopsis invicta*

Data S20: P-value for heterozygosity excess and deviation from the Hardy-Weinberg equilibrium across the *Osmia bicornis* genome.

Data S21: Inbreeding coefficient, *F,* for all individuals, either filtered for heterochromatic regions or unfiltered.

Data S22: Depth, quality per depth, mapping quality and p-value for excessive heterozygosity across the largest chromosome for *Osmia bicornis.*

Data S23: Snapshot of the Github repository at the time of submission.

Data S24: Fraction of the *Osmia bicornis* genome covered by runs of homozygosity larger than 10kbp, per individual.

SUPPLEMENTARY FIGURES:

Genetic mapping in the red mason bee *Osmia bicornis* implicates *ANTSR* as an ancient sex-determining locus in bees and ants

Tilman Rönneburg[1], Demetris Taliadoros[1], Anna Olsson[1], Sara Magnusson[1], Linn Huser[1], Muhammad Nafiz Ikhwan Bin Nor Fuad[1], Turid Everitt[1], Giselle C. Martín-Hernández[1], Panagiotis Theodorou[3], Björn Cederberg[2], Robert J. Paxton[3], Karsten Seidelmann[4], Matthew T. Webster[1*]

1) Department of Medical Microbiology and Biochemistry, SciLifeLab, Uppsala University, Uppsala, Sweden

2) Swedish Species Information Centre, Swedish University of Agricultural Sciences, Uppsala, Sweden

3) General Zoology, Institute of Biology, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

4) Animal Physiology, Institute of Biology, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

*correspondence to matthew.webster@imbim.uu.se

Short title: Sex determination in *Osmia bicornis*

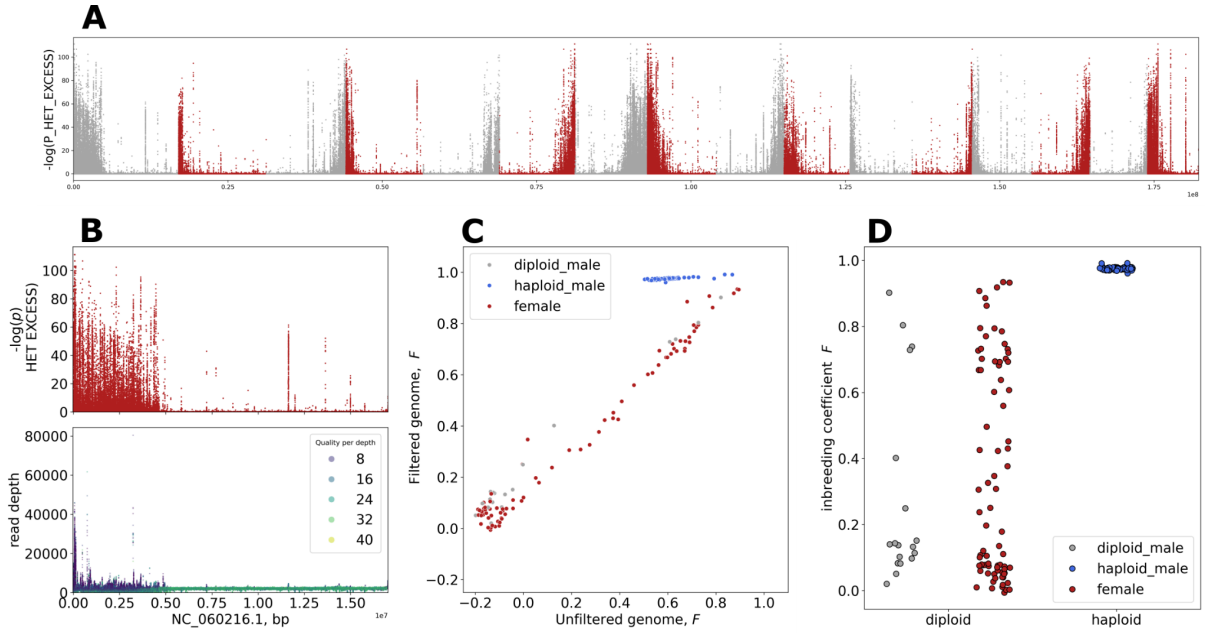

Figure S1. Presence of heterochromatin at the distal ends of chromosomes and its effect on the inbreeding coefficient, *F*. A: Manhattan plot of p-values for excess heterozygosity per marker across all individuals and the 16 largest chromosomes. **B:** Top panel: Manhattan plot of p-values for excess heterozygosity per marker across all individuals and Chromosome 1. Bottom panel: read depth (y-axis) and Quality-by-depth (color) across chromosome 1. **C:** Effect of filtering heterochromatin-regions on *F*. *F* derived from unfiltered genome on the x-axis, *F* derived from the filtered genome on the y-axis. Females colored in red, haploid males initially suspected to be diploid in blue, correctly identified diploid males in grey. **D:** Stripplot of *F* after filtering. Females colored in red, haploid males initially suspected to be diploid in blue, correctly identified diploid males in grey. Data underlying panels A-D can be found in Supplementary Data S20-S22.

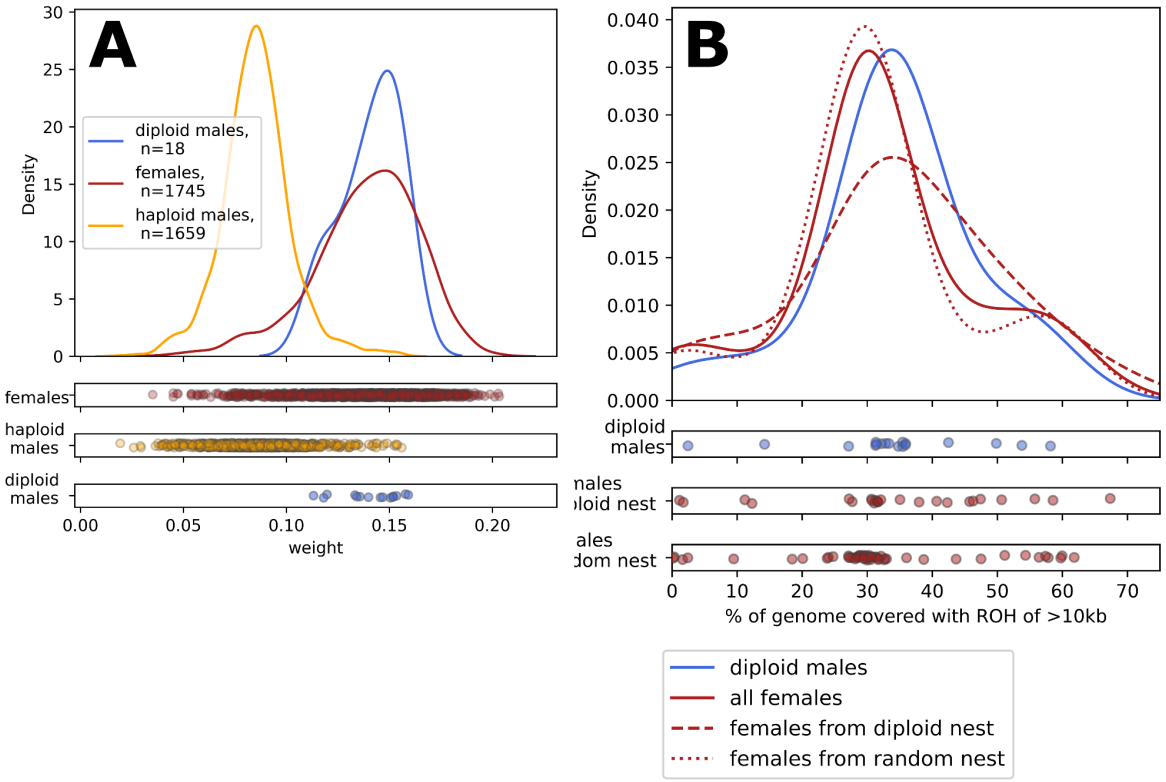

Figure S2. A. Kernel density estimate plot and strip plot of weight distribution of all females and haploid males from nests collected in Uppsala. Red line and dots: Females. Yellow line and dots: haploid males. Blue line and dots: diploid males. **B:** Kernel density and strip plot of the percentage of the genome covered with ROH bigger than 10 kbp using the full dataset of 162 sequenced samples. solid blue line: diploid males.solid red line: all females. Dashed red line: females from nests containing a diploid male. Dotted red line: females collected from nests without diploid males. Data underlying panel A can be found in Supplementary Table S3. Data underlying Panel B can be found in Supplementary Data S24.

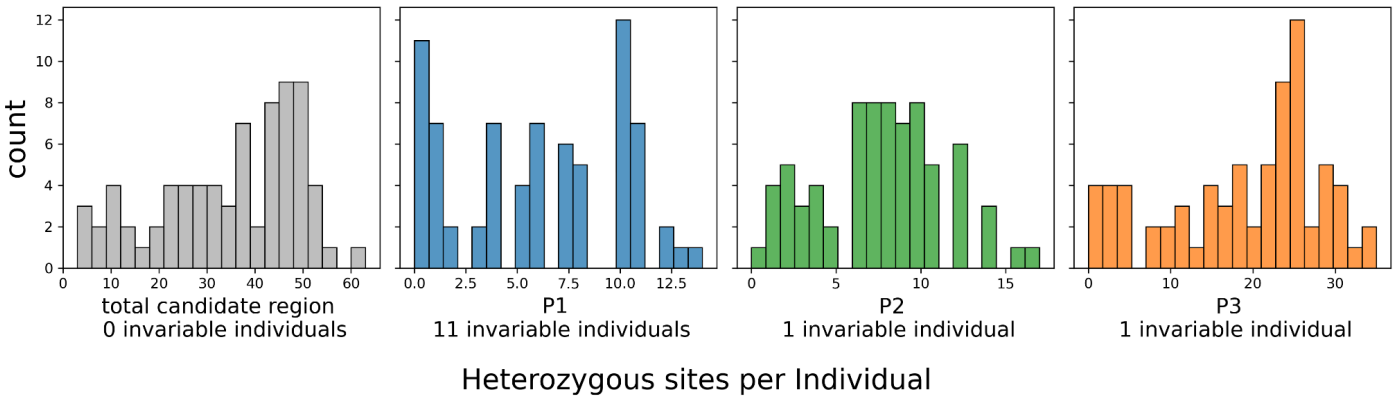

Figure S3. Number of heterozygous sites across the candidate region: from left to right: total candidate region, region P1, P2 and P3. See Figure 3 and Supplementary Table S5 for coordinates. Data underlying this figure can be found in Supplementary Data S5.