Identification of a micropeptide and multiple secondary cell genes that modulate *Drosophila* male reproductive success

Clément Immarigeon\(^{a,1,2,}\), Yohan Frei\(^{a,}\), Sofie Y. N. Delbare\(^{b,}\), Dragan Gligorov\(^{c,}\), Pedro Machado Almeida\(^{a,}\), Jasmine Grey\(^{b,}\), Léa Fabbro\(^{a,}\), Emi Nagoshi\(^{a,}\), Jean-Christophe Billeter\(^{a,}\), Mariana F. Wolfner\(^{b,}\), François Karch\(^{b,}\), and Robert K. Maeda\(^{a,}\)

\(^a\)Department of Genetics and Evolution, Sciences III, University of Geneva, 1211 Geneva 4, Switzerland; \(^b\)Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853-2703; and \(^c\)Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen 9700 CC, The Netherlands

Edited by Claude Desplar, New York University, New York, NY, and approved February 3, 2021 (received for review February 6, 2020)

Even in well-characterized genomes, many transcripts are considered noncoding RNAs (ncRNAs) simply due to the absence of large open reading frames (ORFs). However, it is now becoming clear that many small ORFs (smORFs) produce peptides with important biological functions. In the process of characterizing the ribosome-bound transcriptome of an important cell type of the seminal fluid-producing accessory gland of *Drosophila melanogaster*, we detected an RNA, previously thought to be noncoding, called *male-specific abdominal (msa)*. Notably, *msa* is nested in the HOX gene cluster of the Bithorax complex and is known to contain a micro-RNA within one of its introns. We find that this RNA encodes a “micropeptide” (9 or 20 amino acids, MSAmiP) that is expressed exclusively in the secondary cells of the male accessory gland, where it seems to accumulate in nuclei. Importantly, loss of function of this micropeptide causes defects in sperm competition. In addition to bringing insights into the biology of a rare cell type, this work underlines the importance of small peptides, a class of molecules that is now emerging as important actors in complex biological processes.

SmORF peptide | reproduction | accessory gland | postmating response | *Drosophila*

The genomes of higher eukaryotes appear to contain a paradox: most of their nonrepetitive DNA is transcribed, but only a small fraction codes for proteins. While some long noncoding RNAs (lncRNAs) are undoubtedly functional units, it remains controversial whether most lncRNAs have a function or simply reflect transcriptional noise (1). Further complicating this issue, some lncRNAs actually encode small peptides from small open reading frames (smORFs) that were overlooked in early scans (8, 11, 12). A growing number of cases have been reported where a smORF-encoded peptide is indeed functional (8, 11, 12). In *Drosophila*, several such peptides are critical for the control of development and physiology (13–18). However, the function of most smORF-encoded peptides is still unknown. Overall, the limits of the coding/noncoding and functional/nonfunctional parts of genomes are still blurry, even in the case of model organisms with well-annotated and studied genomes like *Drosophila melanogaster*.

In this study, we report the identification of a functional micropeptide encoded by a putative “noncoding” RNA, from the secondary cells of the *Drosophila* male accessory glands (AGs). AGs produce most of the seminal fluid proteins that are transferred to females during copulation and are critical for reproductive success. AGs are often referred to as functional analogs of the mammalian prostate gland and seminal vesicles (19, 20). In insects, secretions from AGs induce and maintain the physiological and behavioral changes occurring after mating in females, collectively called the postmating responses (PMRs). The PMR includes increased ovulation and egg laying, sperm storage and release, dietary changes, and gut growth. PMR also include decreased receptivity to remating, ensuring propagation of the first male’s genome at the expense of rivals (reviewed in refs. 21, 22). Due to their central role in insect reproduction and evolution, AGs and the individual seminal fluid proteins that they produce to mediate the PMR are topics of intense interest. Insights obtained from the *Drosophila* model have proven relevant for other problematic species (21, 23–25), and thus, identification of novel AG products offer important avenues for the benefit of human health and agriculture.

Here, we examined the overall transcriptome and ribosome-bound transcriptome of an important yet cryptic cell type in the AG: the secondary cells (SCs). SCs are large binucleate cells whose cytoplasm is filled with vacuole-like structures (26–29). These cells express specific genes (29) and proteins that are necessary to maintain a PMR past 24 h by allowing Sex Peptide, the major trigger for the *Drosophila* PMR (30–32), to bind to and be stored with sperm (28, 33–39). Unfortunately, due to their

**Significance**

In many species, mating induces physiological changes in the female that increase the reproductive success of the mating pair. The postmating response (PMR) is caused by male seminal fluid proteins interacting with the female reproductive system. Because of the importance of the PMR in many insect species relevant to human health and agriculture, we examined the gene expression profile of one important cell type of the gland that produces most of the seminal fluid proteins in *Drosophila*. Interestingly, among proteins necessary for the PMR in this species, we discovered that a micropeptide encoded by a supposedly noncoding transcript is produced and carries out important reproductive functions. Such micropeptides were previously unrecognized but are emerging as important actors in complex biological processes.

**Author contributions:** C.I., M.F.W., F.K., and R.K.M. designed research; C.I., Y.F., S.Y.N.D., D.G., J.G., L.F., J.-C.B., and R.K.M. performed research; C.I., Y.F., S.Y.N.D., P.M.A., L.F., E.N., M.F.W., and R.K.M. analyzed data; C.I. and R.K.M. wrote the paper; and M.F.W. and F.K. supervised the research. The authors declare no competing interest.

This article is a PNAS Direct Submission.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

**Present address:** Molecular, cellular and developmental biology (MCD) Unit, Centre de Biologie Intégrative, Université de Toulouse, CNRS, Bat 484, F-31062, Toulouse, France.

**To whom correspondence may be addressed.** Email: clem.immarigeon@gmail.com or robert.maeda@unige.ch.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2001897118/-/DCSupplemental.

Published April 5, 2021.
Dataset S1. As shown in Fig. 1 condition (somes and associated mRNAs from SCs but not from main cells verified the specificity of the technique by showing that ribosomes (∼msa protein by immunostaining accessory glands) (38). On the other TRAP data corroborate unsuccessful attempts to detect Ubx (21st most highly expressed RNA) but not translated in SCs (our as expected, as exemplified by TRAP is highly informative for specific genes that do not behave bosomes, suggesting that it may encode unknown peptides. GFP alone (GFP-tagged ribosomes (41, 42). For this, we expressed either characterized the translation of mRNAs by immunopurifying GFTagged ribosomes (41, 42). For this, we expressed either TRAPseq can reveal the coding potential of transcripts annotated as noncoding (lncRNAs) and can also reveal that some canonical mRNAs are not actually translated in a particular cell type. This method has previously been used to estimate the coding potential of transcripts associated with ribosomes and thus potentially translated into proteins. TRAPseq can reveal the coding potential of transcripts annotated as noncoding (lncRNAs) and can also reveal that some canonical mRNAs are not actually translated in a particular cell type. This method has previously been used to characterize the translation of mRNAs by immunopurifying GFTagged ribosomes (41, 42). For this, we expressed either GFF alone (“mock” condition) or GFF-RPL10Ab tagged ribosomes (“TRAP” condition) specifically in SCs (Fig. 1B). We verified the expression of the long-term PMR by showing that ribosomes and associated mRNAs from SCs but not from main cells were efficiently enriched in the TRAP compared to the mock condition (SI Appendix, Fig. S1). These datasets are accessible in Dataset S1. As shown in Fig. 1C, RNaseq and TRAPseq data are highly correlated (Pearson correlation coefficient ∼0.9), which suggests that RNaseq generally gives a relatively accurate estimation of mRNA translation for most genes. However, TRAP is highly informative for specific genes that do not behave as expected, as exemplified by Ubx and msa (Fig. 1C). On one hand, the HOX gene Ultrabithorax (Ubx), is highly transcribed (21st most highly expressed RNA) but not translated in SCs (our TRAP data corroborate unsuccessful attempts to detect Ubx protein by immunostaining accessory glands) (38). On the other hand, msa is annotated as a lncRNA, but is associated with ribosomes, suggesting that it may encode unknown peptides. Given the surprising finding of msa in our dataset, we sought to validate the accuracy and specificity of our screen by looking at the protein-coding genes. From the list of >7,000 genes robustly detected, we used a gating strategy to narrow the informative SC coding genes (see Fig. 1D for data analysis and cutoffs). After Gate#2, we obtained a list of 1,415 genes. Gene ontology using this gene list is consistent with secondary cell biology, as the most highly enriched terms are related to reproduction, postmaturing sperm, vesicle-mediated transport, secretion, and endocytosis, with many proteins predicted to be located at endomembranes, in vacuoles or in the extracellular space (Fig. 1E). This fits the current view of SCs as vacuole-filled secretory cells that produce, secrete, import, and modify proteins involved in the PMR. We then narrowed down the list of genes to the 151 transcripts that displayed the strongest transcription and ribosome association in secondary cells (Gate#3 genes, shown in red in Fig. 1C). These 151 genes include five of the six previously known SC markers msa (38), Rub19 (27), Dve (39), CG17575, and Lectin-46Ca (27, 35), validating both procedures and data analysis (Dataset S2), and many encode secreted proteins that are associated with the PMR (regulation of female receptivity) and sperm competition (Fig. 1E).

Strikingly, a small number of RNAs make up the majority of transcripts (in terms of number of reads): the top seven genes contribute to >70% of all the transcriptome reads (Figs. 1C and 24). Notably, these seven genes encode proteins known to be transferred to females during mating (33) (CG9029, Aep32CD, the CRISP CG17575, Lectin-46Cb, Lectin-46Ca, CG13965, and midline fasciclin/msas), including two previously known to be expressed in SCs. Three of these genes [CG17575 (38), Lectin-46Cb, and Lectin-46Ca] (36, 37, 43, 44) have already been characterized as regulators of the long-term PMR through transient binding to sperm (36, 37, 45), thus further validating the dataset as revealing important players in the SC’s reproductive role. To extend this validation, we selected five uncharacterized genes among the most highly expressed SC transcripts (Fig. 24) and assayed their function using standard PMR assays. We tested the ability of knockdown males to induce the decrease in female receptivity typically observed for many days after initial mating. We first tested CG9029, using two RNAi lines, as this was the most highly expressed gene and encodes a seminal fluid protein (43). While CG9029 was not required in SCs to initiate the PMR (no remating was observed at 24 h, not shown), we found that it is critical to extend the females’ unreceptive nature past 1 d (Fig. 2B). Confirming the role of CG9029 in creating a long-term PMR, we also found that mates of knockdown males display lower levels of egg deposition at times after day 1 (SI Appendix, Fig. S2 A and B). The long-term PMR is mediated by Sex Peptide (32, 33, 46), which must be stored with the sperm to allow the PMR to last for more than 1 d. We thus tested if CG9029 affected Sex Peptide storage in the sperm storage organ. Upon SC knockdown of CG9029, Sex Peptide was undetectable by 24 h postmating (Fig. 2C). This explains, at a molecular level, why females are receptive to remating and lay fewer eggs after 1 d. We found analogous results for three of the other genes we tested: Aep32CD, CG31145, and GILT-3 (SI Appendix, Fig. S2 C and D), showing that these genes must be expressed in SCs to allow proper Sex Peptide storage and reduce female receptivity (26, 27, 47, 48). Of our five candidates, only CG13965 did not show a detectable long-term PMR phenotype.

Having validated our list of SC signature genes on the basis of scientific literature and experimental data for protein-coding genes, we were then confident to examine the supposedly noncoding transcript msa (38). Although this transcript is annotated as lncRNA, its detection at high levels in our TRAPseq led us to investigate its potential to encode previously unknown peptides. The msa “lncRNA” Encodes SC-Specific Micropeptides. msa is a SC-specific isoform of the ia8 lncRNA that spans the intergenic region between the HOX genes Abhd-B and abd-A in the Bithorax complex (Fig. 3A). Previously, we showed that msa is critical for SC function (38) and that much of its function can be attributed to a miRNA, located within its fourth intron. As our data suggested that this transcript is also peptide-encoding, we verified our TRAP results using RT-qPCR on the ribosome immunopurified RNAs. Indeed, RT-qPCR confirms a strong enrichment of this transcript associated with SC ribosomes (SI Appendix, Fig. S1).
This RNA contains a block of sequence located within its last exon that we previously identified based on its conservation in several Drosophila species (49). This block starts with an ATG and finishes with two, in-frame, putative stop codons and could encode a 9-amino acid peptide. As this analysis was based on a limited number of species, we extended the prior evolutionary analysis to more Drosophila species (Fig. 3B and Dataset S3). This revealed that the block of sequence conservation extends to another upstream ATG. We also note sharp drops in conservation both upstream of the first ATG (M) and downstream of the stop codons (*), as evidenced by the "amino-acid conservation" plot and DNA sequence consensus line in Fig. 3B. Thus, the conserved region consists of a putative smORF with up to three in-frame ATGs that could code for overlapping small peptides of 20, 9, or 3 amino acids, respectively. Although we could not find any convincing homologs outside of the Drosophila genus, we find this sequence in Drosophilidae from multiple subgenera (Fig. 3B and Dataset S3). This sequence has thus been conserved for over 60 million years of evolution (the last common ancestor of D. melanogaster and Drosophila willistoni has been estimated to have lived 62.2 Mya) (50). We call this multi-ATG putative smORF "MSAmiP", and will refer to the encoded peptides as MSAmiP20, MSAmiP9, and MSAmiP3 depending on their size in amino acids. Conservation of the putative MSAmiP smORF shows that the amino acid sequence of MSAmiP9 has mostly been unchanged in all the species tested, with most mutations being silent or conservative (e.g., L to F, Fig. 3B). The MSAmiP20 sequence, by comparison has accumulated more mutations, as exemplified for Drosophila busckii whose MSAmiP20 is the most divergent with 6/20 amino acid changes. However, protein structure prediction...
software (PEP-FOLD 3.5) suggests a helical conformation for the MSAmiP9 portion, which is retained in the analysis of the MSAmiP20 sequences. This predicted structural conservation suggests that selective pressure may have occurred at the protein level and could indicate that the MSAmiP-encoded peptides are functional.

To investigate the translation potential and function of this smORF in vivo, we generated genetic tools to modify its endogenous sequence within the Bithorax complex (Fig. 3A). Using CRISPR-mediated cassette exchange, we replaced the exon of msa containing the putative smORF with a site-specific integration platform (attP) (51). Using this "Δexon8 attP" platform, we could then insert any modified version of the exon into its normal genomic locus via PhiC31-mediated recombination: we generated a rescue construct (exon8wt rescue), a clean MSAmiP smORF knockout (ΔMSAmiP), and a C-terminal GFP fusion construct (MSAmiP-GFP). Note that the GFP portion of the MSAmiP-GFP fusion lacks its start codon to prevent GFP translation independent of the smORF. Examination of the MSAmiP-GFP flies revealed GFP expression in the SCs (Fig. 3C), confirming that msa encodes at least one small peptide translated from this MSAmiP smORF. GFP expression from MSAmiP-GFP seems to accumulate in SC nuclei (as revealed by colocalization with DAPI staining) (Fig. 3C) but some was also observed in the cytoplasm where it was excluded from the large vacuoles. As GFP signal was not seen elsewhere in any of the other tissues examined at any developmental stages, we conclude that MSAmiP is only translated in the SCs.

MSAmiP Mutant Males Have Increased Sperm Competition Defensive Abilities. To investigate MSAmiP function without affecting the other functions of msa (e.g., miR-iab-8), we compared the ΔMSAmiP line, in which the 60 base pairs of the MSAmiP smORF are deleted, to the exon8wt rescue line (control), in which the intact exon8 was placed back at its normal location (Fig. 3A). These two lines contain virtually identical genomes, except for the presence or absence of the MSAmiP smORF. Previously, we showed that a deletion of the promoter of msa results in phenotypes affecting fertility, maintenance of the PMR, and sperm competition (35). To test if any of these phenotypes can be attributed to the loss of MSAmiP, we tested ΔMSAmiP males for their ability to mediate these processes. We first tested fertility and the PMR. Upon single mating, MSAmiP-deficient males produce as many offspring as control males (SI Appendix, Fig. S4A). After 4 d, females mated to these males were tested for receptivity to courting males. In this assay, ΔMSAmiP mated females are unresponsive to remating at 4 d, similar to females mated to control males (SI Appendix, Fig. S4B). Based on these assays, we conclude that MSAmiP is not involved in fecundity or in triggering the PMR.

We then tested ΔMSAmiP males to see if the mutant recapitulated the sperm-competition phenotype that we had previously...
observed for males lacking the msa transcript (35). Generally, if a Drosophila female mates with two males, the second male will sire most of the progeny ("last male sperm precedence") (52–55); the first male’s sperm are displaced by those of the second male (56–59). Previously, we had shown that if the first male lacked msa, last male precedence was reduced (35). To test the role of MSAmiP in this phenomenon, we used the assay schematized in Fig. 4A to score the relative proportion of progeny sired by two different males successively mated to a female. Fig. 4B shows the result from ΔMSAmiP and control males. As expected, in the control condition there is a strong preference for the sperm from the last male (mean P1 = 0.26 for exon8wt rescue). In contrast, when the first male lacks MSAmiP, last male precedence is greatly reduced (mean P1 = 0.45). Importantly, the total number of

Fig. 3. MSAmiP smORF is conserved and encodes a micropeptide translated in SCs. (A) The top line represents the abdominal region of the Bithorax complex marked off in kilobases, with the structure of the Abd-B and abd-A transcription units. The structures of the iab-8 ncRNA and msa are drawn below. Exons are indicated as black boxes (not to scale). Arrows represent the last exon and the sense of transcription. The blue hairpin shows the approximate location of miR iab8. The MSAmiP smORF is indicated in red in the last exon (exon8) of the msa RNA and of the iab-8 ncRNA. Note that the msa RNA originates from an alternative promoter (active only in SCs) and shares all its downstream exons with the 3′ exons of the the iab8 ncRNA. The genetic tools generated to study MSAmiP smORF are presented below. Dotted vertical lines show the limits of the CRISPR-mediated deletion (Δexon8) where exon8 is replaced by an attP integration site. This J exon8 attP platform was used to generate several modifications of the endogenous locus, including exon8 wt rescue, ΔMSAmiP deletion, and MSAmiP-GFP knockin (green rectangle represents GFP ORF fused to MSAmiP smORF). (B) MSAmiP smORF conservation among Drosophilidae. Amino acid sequence alignment of the genomic region surrounding MSAmiP smORF in Drosophila species from multiple subgenera shows that the conserved block starts at MSAmiP20 methionine (M) and stops after the stop codons (*). Level of each amino acid conservation is represented underneath as a bar chart (Jalview 2.10.5 version (http://www.jalview.org/) using clustalx coloring. Structure predictions using PEP-FOLD 3.5 show that resulting polypeptides could adopt very similar tridimensional conformations with helical C terminus. (C) MSAmiP is translated in secondary cells. The genital tract from a MSAmiP-GFP knockin male is imaged live (no fixation or staining) and GFP is visible in SCs. AG, accessory glands; SCs, secondary cells; ED, ejaculatory duct. (Scale bars, 50 μm.) C′ and C″ show GFP accumulation in SC nuclei (stained with DAPI, arrowheads), in fixed tissue.
Accumulates in SC Nuclei. MSAmiP Seems to Primarily Be a 9 Amino Acid Micropeptide That

progeny from each female is not affected (Fig. 4C, gray bars) but the proportion of progeny from each male changes (Fig. 4C, white and black bars). Thus, we conclude that the sperm-competition phenotype observed in males lacking msa (and in control mutants) (35) is likely due to the loss of MSAmiP and is related to sperm storage/usage by the female in a competitive context, and not due to a fecundity issue. Interestingly, the relative success of MSAmiP-GFP knockin males is similar to that of control exon8wt rescue males, indicating that the MSAmiP-GFP fusion is functional (Fig. 4B).

The latter options were ruled out by transcriptome analysis of flies carrying the ΔMSAmiP deletion, which showed that this mutation does not destabilize the msa transcript (SI Appendix, Fig. S3A), or remove critical cis-regulatory elements required for expression of any genes nearby (SI Appendix, Fig. S3B). To confirm directly that the sperm-competition phenotype observed from the loss of MSAmiP peptides, we assayed a frameshift mutation of MSAmiP-GFP in which only a single nucleotide was removed (MSAmiP*3-GFP, schematized in Fig. 5A). We observed that MSAmiP*3-GFP mutant males also display a significantly higher P1 ratio than MSAmiP-GFP controls (SI Appendix, Fig. S4C). These results verify that MSAmiP smORF-encoded peptides are likely responsible for the sperm-competition phenotype.

MSAmiP*3-GFP, MSAmiP*9-GFP, and MSAmiP*20-GFP produce weak fluorescence in SCs, and MSAmiP-GFP produce more fluorescence than the other lines. This result can also be seen on Western blots of knockin male accessory glands using an antibody against GFP (Fig. 5C). Notably, both MSAmiP*3-GFP and MSAmiP*20-GFP produce weak fluorescence in SCs, and

Fig. 4. MSAmiP is involved in sperm competition. (A) Experimental setup to test the sperm-competition defensive ability of the three mutant males indicated in blue, red, and green. Individual cn bw females with white eyes are crossed first to a red-eyed male (test male 1) and then remated to white-eyed males (cn bw male 2). Eggs are collected over 8 d in four successive vials. Paternity is scored based on offspring eye color. (B) Proportion of the progeny sired by male 1 (P1) depends on MSAmiP. Results are shown as box and whiskers plots showing 10 to 90 percentile and the median. A P value of <0.0001 (****) was observed for ΔMSAmiP mutant vs. exon8wt rescue. For exon8wt rescue vs. MSAmiP-GFP knockin, the P value was 0.9 (not significant, ns). Finally, the P value for ΔMSAmiP mutant vs. MSAmiP-GFP knockin was 0.0065 (**). Number of individual, doubly mated females is given (n =). See Materials and Methods for statistical analysis. (C) The number of progeny sired by males 1 and 2. Statistical tests are similar to previous ones but assuming a Poisson distribution. The total number of offspring is not affected by the genotype of the first male (ns, P = 0.9187), but the respective progeny of each male is significantly affected (P < 0.0001 for both comparisons).
MSAmiP*9-GFP alone does not elicit as much fluorescence as the full-length knockin. Thus, our results using both tissue culture cells and fly accessory glands suggest that the 9 amino acid form of MSAmiP is the predominant form of the peptide (Fig. 5B and C and SI Appendix, Fig. S5). However, other isoforms are probably produced at lower levels.

Upon examining the localization of MSAmiP-GFP, we found that GFP signal was largely enriched in SC nuclei with some cytoplasmic staining, but no signal in the large vacuoles (Fig. 3C). This staining is restricted to SCs. Although MSAmiP-GFP is functional according to our sperm-competition assay (Fig. 4B), we worried that the subcellular localization could be partially artifactual due to the GFP tag, which has been shown to accumulate in the nucleus (60). Because our attempts to make specific antibodies to MSAmiP were unsuccessful, we confirmed the subcellular localization of MSAmiP using a different tag. For this, we generated an N- and C-terminal Flag-HA-tagged version of MSAmiP (FLAG-HA-MSAmiP20 and MSAmiP-Flag-HA).

Note that for the N-terminal-tagged version we used the miP20 methionine as the start codon. Although miP20 does not seem to be the most abundant form of MSAmiP, it does seem to be expressed and we reasoned that its localization might be less affected by the addition of a tag. For both tagged constructs, we observed accumulation of signal in SC nuclei (Fig. 6, see the

Fig. 5. MSAmiP9 seems to be the most abundant peptide encoded by msa in vivo. (A) Design of the isoform-specific MSAmiP-GFP* knockins. The amino acid sequence encoded by MSAmiP smORF is shown with methionines highlighted in green (M). Horizontal double arrows represent the different MSAmiP isoforms that may produce a GFP fusion in each construct. Blue triangles labeled fs show the position of the frameshift mutations, and red asterisks show point mutations of methionine-into arginine-codon (R). (B) Confocal image (Z slice) of a live AG distal tip representative of each genotype. (C) GFP fusion protein detected by Western blot from male accessory glands.

Fig. 6. MSAmiP localizes preferentially to secondary cell nuclei. Confocal image (1z) of male accessory glands expressing GFP and MSAmiP-FLAG-HA (A) or FLAG-HA-MSAmiP20 (B) in secondary cells. Close-up of a single SC is shown with its contours indicated by the dashed lines. GFP is in green (A’ and B’), DAPI in blue (A’’ and B’’) or gray (A’’’ and B’’’), and Flag (A, A’, and A’’’ or HA, B, B’, and B’’) in red. While most of the MSAmiP-FLAG-HA/Flag-HA-MSAmiP20 signal is seen in nuclei, a weak signal is also visible in the cytoplasm. (Scale bars in A and B, 30 μm; 5 μm in close-ups.)
close-ups on individual, binucleate SCs). Furthermore, this result could be confirmed in cultured cells (SI Appendix, Fig. 6b), suggesting that the nuclear enrichment is an intrinsic property of MSAmiP.

**Discussion**

**MSAmiP, a Micropeptide Required for Proper Sperm Competition.**

The data presented in this study show that the msa transcript, annotated as a lncRNA, encodes a microRNA involved in sperm competition. We previously reported a similar sperm-competition phenotype for the iab-6(cocuD1) mutant using the same assay shown in Fig. 4 (35). Subsequently, we showed that the iab-6(cocuD1) mutant was a deletion of a region (the DI enhancer) that includes the msa promoter, and that several iab-6(cocuD1) mutant phenotypes were attributable to msa, and in particular to mirl-iab-8, a microRNA nested in one of the introns of msa (38). So far, the loss of MSAmiP is the only clearly documented reason for the iab-6(cocuD1) mutant phenotype in sperm competition.

Here, we find that ΔMSAmiP males sire a larger proportion of offspring than the wild-type control, when it is the first male to mate in a sperm-competitive situation. This initially seems counterintuitive, as one might think that the relative competitive success of males without MSAmiP might have led to the loss of the peptide-coding sequence due to selection. However, the surprising phenomenon of a deficient male doing better as first male has been seen in mutants for a number of seminal protein genes or secondary cell regulators: Sex Peptide, ACP62F, and Dad (BMP signaling); as with ΔMSAmiP, loss of function of these genes enhanced sperm defensive ability in sperm-competition assays (56, 61–63). A potential explanation for these results is based on findings with Sex Peptide and Dad. Their increased sperm defensive ability appears to be a consequence of inefficient release of the mutant male’s sperm from its storage in the female: both Sex Peptide and Dad mutant sperm are overretained in the sperm storage organs (48, 56). Thus, there are more of them left to compete with the second male’s sperm when second mating occurs. In light of these and recent data (64), we propose that while overretention of sperm (such as those from ΔMSAmiP males) in storage organ may enhance success in sperm-competition assays, the failure to properly release those sperm in the first place may be detrimental to the male’s production of offspring over time, and hence the genes are not lost during evolution.

How MSAmiP mediates its function remains unclear. Our data show that MSAmiP accumulates in SC nuclei (Figs. 3D and 5), which suggests that it could play a role in gene expression. However, examination of the ΔMSAmiP SC transcriptome reveals little difference compared to wild-type cells (SI Appendix, Fig. S3), with only six genes significantly misexpressed (CG9119, ppk20, CG13360, CG33080, nSyb, and CG7582, see Dataset S4 and Materials and Methods for statistical analysis). None of these genes point to an obvious explanation for the phenotype we observed, nor to additional phenotypes to explore. Importantly, none is significantly misexpressed in the same direction in the iab-6(cocuD1) (msa mutant) transcriptome (Dataset S1), so we consider it unlikely that these genes are the functionally relevant MSAmiP targets. Thus, based on our statistical analysis of differentially expressed genes, we conclude that the effect of MSAmiP expression is subtle and does not dramatically affect transcript levels in SCs.

Based on our GFP fusion data, MSAmiP seems to predominantly be a 9-amino acid peptide with some additional expression coming from other in-frame ATG(s) (Fig. 5). Examining the peptide sequence in different Drosophila species, we can see that depending on the species, the 9-amino acid methionine or the 20-amino acid methionine has been lost (see Drosophila erecta and Drosophila ananassae in Fig. 3B). This could indicate that both forms are functional. Although we favor a model in which the 9-amino acid form is more abundant in D. melanogaster, we do see the 20-amino acid form expressed at a noticeable level (Fig. 5) and we cannot rule out the possibility that our mutations change the usage ratio of each methionine of the peptide stability. Regardless of the exact size, MSAmiP is a small peptide. Thus, like other micropeptides (16, 18) MSAmiP would probably need to interact with other factors to be biologically active. As our GFP fusion protein seems to be functional (Fig. 4B), it could provide a useful tool to search for binding partners.

**smORF-Encoded Peptides Should Receive More Attention.** Our findings with MSAmiP provide a strong argument for the consideration of smORFs as a source of biologically significant molecules. A growing number of such peptides are now known to be involved in the development of cancers and are being investigated as potential tumor markers, therapeutic targets, or drugs (65). Interestingly, one of these peptides is produced from a transcript annotated as a lncRNA from the human HOX B cluster and has been shown to suppress colon cancer growth (66). Like MSAmiP, this peptide was found through a combination of transcriptome and translatome methods. Although these peptides show no identifiable similarity at the sequence level, the finding of two peptides expressed from Hox gene complexes in two distant organisms is of note and may indicate that the Hox complexes are somehow an evolutionary environment conducive to the appearance/repetition of smORFs. Alternatively, and perhaps more probably, this finding may simply reflect the vast underestimation of the number of functional peptides stemming from smORFs in eukaryotic genomes.

In line with this, five other SC signature genes in our dataset are annotated as ncRNAs: Hsromega, CR40963, CR42862, CR41602, and CR40677 (Dataset S2). Interestingly, Hsromega encodes a conserved 27-aa peptide that is translated upon stress induction and has been shown to be associated with translating ribosomes (67). Although it remains to be determined why CR40963 and CR42862 are associated with ribosomes, it seems very possible, based on our data, that they may code for novel small peptides. CR41602 and CR40677 are annotated as ribosomal RNA pseudogenes. Thus, their high TRAP-qPCR counts may simply reflect the integration of these RNAs into the purified ribosomes and not their translation potential (like our 18S rRNA TRAP-qPCR control, SI Appendix, Fig. S1A).

**Genomic Insights into the Translatome of an Important Secretory Cell Type.** Secondary cells of Drosophila male accessory glands play a critical role in reproduction (27, 28, 35, 38, 39), but the nature of their products, and the function of those molecules is not yet well understood. Here, we shed light on this rare cell population, using transcriptome and TRAPseq analysis as an entry point. We find that SCs are highly specialized cells, whose most abundantly translated RNAs encode proteins necessary for the long-term PMR: the top seven SC signature-coding genes represent more than 70% of total transcriptome reads, and encode secreted proteins critical for the long-term PMR (Fig. 2 and SI Appendix, Fig. S2). In addition, our translatomics analysis of SC reveals an enrichment for GO terms related to import and endocytosis (Fig. 1E). SCs might thus import proteins from neighboring main cells and use their unique network of vacuole-like structures (27) to modify them posttranslationally, prior to secretion into the lumen. Curiously, we also observed some surprising about the relative transcription and translation of RNAs in these cells. Although highly transcribed genes are expected to produce large amounts of protein and lncRNAs are not, comparing the transcriptome and TRAPseq of SCs shows that these assumptions can be wrong. The Ubx (Ultrabithorax) mRNA is highly represented in the SC transcriptome (21st most highly expressed gene), yet it appears to be untranslated in SCs (Dataset S1 and Fig. 1C), confirming results obtained by the immunoostaining of accessory glands (38). Conversely, the msaf transcript has long
been annotated as an lncRNA but our TRAPseq showed that it was ribosome-associated, and our further analysis reported here shows that it encodes a micropeptide in SCs. These findings highlight the importance of examining both translational and transcriptional data before the selection of potential functional candidates.

At the other end of the lncRNA spectrum, some lncRNAs may play a role in SC biology through their RNAs. For example, microRNA precursor CR43314 and iab-4 are expressed (but not translated) in SCs (Dataset S1) and probably play a role in the posttranscriptional regulation of gene expression. Notably, mirR-1ab-4 is the complementary miRNA to mirR-1ab-8 made by msa, and both of these miRNAs target Ubx (68–72), possibly responsible for the Ubx mRNA being untranslated in SCs.

**Newly Characterized SC Genes Can Help Uncouple the Sequence of Events Leading to the Long-Term PMR.** In addition to discovering that a micropeptide is translated and important in SC function, our transcriptome studies identified SC signature-coding genes (Dataset S2 and Fig. 2), including some with known function and expression (verifying our screening) and several for which we found reproductive function. Interestingly, SC genes appear to be involved at different steps of the cascade of events leading to the long-term PMR, confirming that an important role of SCs is to lengthen the PMR through the storage of Sex Peptide (Fig. 2C and SI Appendix, Fig. S2).

Notably, three of the five most-expressed SC genes (CG17575, Lectin-46Ch, and Lectin-46Ca) belong to the previously known network of accessory gland proteins that allow Sex Peptide to bind to sperm for storage (36, 37, 43, 44). However, these proteins bind sperm only transiently, and hence are not the actual scaffold on which Sex Peptide anchors (36). Here we show that additional proteins secreted by SCs are also critical for Sex Peptide storage past 1 d (CG9029, ACP32CD, and CG31145) and thus represent candidate scaffold proteins, or sperm modifiers that allow Sex Peptide binding. For example, CG9029 sequence does not show strict homology to proteins with known functions; it is predicted to encode a protein of 122 aa, high in mucin-like, O-linked glycosylation sites (S/T rich PxP repeats). As mucin proteins form polymers and biofilms, this might be consistent with a scaffolding role.

**Concluding Remarks.** In this study we used transcriptome and TRAPseq as entry points to study secondary cells (SCs), a rare secretory cell type. In addition to gaining insights into SCs' products and functions, we made a surprising discovery: a transcript classified as noncoding from the Bithorax HOX gene cluster was found to code for a set of small peptides that we call MSAmiP. These peptides, produced specifically in a few cells, affect male reproductive success, and have been conserved through Drosophila evolution. MSAmiP is a peptide produced from an miRNA precursor transcript in animals. This study shows that small peptides, including ones encoded by RNAs previously designated as ncRNA, are a potential reservoir of novel functional proteins, whose importance may only manifest in specialized cell types.

**Materials and Methods**

**Plasmid and Transgenic Fly Line Generation.** Cloning strategies to generate all plasmids used in this study are detailed in SI Appendix. The successive steps to obtain the exonB attP platform using CRISPR, followed by PhiC31 site-specific integrations to obtain the multiple knock-in lines and the ΔMSAmiP are described in SI Appendix. See Dataset S5 for PCR primers, Gblocks, and CRISPR guides.

**Secondary Cell Transcriptomics and TRAPseq.** SC sorting and transcriptome are described in detail in ref. 40. TRAPseq was performed following the protocol from ref. 41, with adaptations detailed in SI Appendix. Briefly, for each biological replicate (n = 3) 250 male genital tracts expressing GFP-RPl10Ab (TRAP) or GFP (mock) in SCs were recovered, lysed, and submitted to anti-GFP immunoprecipitation. RNAs were extracted and polyA+ RNAs were sequenced. Sequencing procedure, gene counts, and data analysis are detailed in SI Appendix.

**Fly Stocks.** To sort SCs by FACS, UAS-GFP expression was driven in SCs using AbdB-Gal4 and D1:Gal4 (35, 38) and the iab-6mutai (35). For TRAP UAS GFP-RPL10Ab (41) recombined with D1:Gal4 was used. UAS-RNAi lines to test new candidate genes were obtained from the TRIP collection and VDRC (Vienna Drosophila Resource Center) collections: CG9029 RNAi1 = TRIP HM22752 and RNAi2 = TRIP HM502425, GL73 RNAi1 = 102104KK and RNAi2 = 38069GD; CG12365 RNAI = 41223CD and RNAI2 = 106357KK; CG31145 RNAI1 = 108878KK and RNAI2 = 25036GD; and Acp32CD RNAI1 = 102687KK. All transgenic lines generated in this study are detailed in SI Appendix.

**Antibodies.** Antibodies used in the course of this study include: rabbit anti-GFP (Torrey Pines), mouse anti-Flag M2 (Sigma), rabbit anti-Sex Peptide (32), goat anti-tubulin (ab6160, Abcam), and sheep anti-rat-HRP (ab6852, Abcam).

**Microscopy and GFP Fluorescence Quantification.** Micrographs used in this study were acquired using a Leica SP confocal microscope (Figs. 1, 3, 5, and 6) or a regular fluorescence/phase contrast microscope (SI Appendix, Fig. S5 E and F). MSAmiP-GFP knockin images are taken from untrained, freshly dissected live glands. Images from Figs. 1 and 3C and SI Appendix, Fig. 36 are maximal projection of multiple confocal Z (stack). Images in Figs. 3C and C', 5, and 6 are single Z images. The procedure to obtain quantifications shown in SI Appendix, Fig. S5 is detailed in SI Appendix.

**Western Blots.** All detail can be found in SI Appendix. Briefly, for Fig. 2C and SI Appendix, Fig. S2D, single pair matings were performed and the mated females were snap frozen in liquid nitrogen after 2 to 4 h, 24 h, or 4 d after mating. The equivalent of approximately five female glands were run per well. For Fig. 5C, five pairs of accessory glands from 1-d-old males are loaded for each knockin and wild-type lane and one pair of glands is loaded for the da->>MSAmiP-GFP positive control (daughterless:Gal4 > UAS:MSAmiP-GFP).

**Receptivity Assays.** We set up single pair matings in vials at room temperature using 5-d-old virgin Canton S females and 5-d-old tester males. Males were removed after copulation ended. Females were left in the vial and after 4 d, were presented with one 5-d-old Canton S male. Matings were observed and scored for 1 h (Fig. 2B) or 2 h (SI Appendix, Figs. S2C and S4B).

Statistical significance of our results was assessed using Fisher’s exact test.

**Concluding Remarks.** In summary, single pair matings of cn bw females to tester males, secondary rematings to cn bw male (3 d later), fly handling, and progeny counting was performed as in ref. 35. Experiments presented in Fig. 4B were performed in two blocks. For the first block, we compared P1 between ΔMSAmiP, exonB wt rescue and MSAmiP-GFP knockin. We performed two replicates for block 1, with sample sizes: MSAmiP-mutant n = 40 for replicate 1 and n = 22 for replicate 2; and WT rescue n = 38 for replicate 1 and n = 29 for replicate 2 MSAmiP. For block 2, we performed two replicates, each with a smaller sample size (total sample size MSAmiP mutant n = 28; MSAmiP-GFP n = 35; WT rescue n = 25).

We performed a statistical test in R using the packages lim4 and emmeans (73–75). The test was performed within each cell type, i.e., proportion of offspring sired by the first male. The test was performed depending on the genotype of the first male (ΔMSAmiP, MSAmiP-GFP, or exonB wt rescue) using a generalized linear mixed model (GLMM), assuming a binomial distribution of the data, adding “female ID,” “repli cate,” and “block” as random effects. A subsequent experiment presented in SI Appendix, Fig. S4C was carried out, following similar procedures although in different laboratories, on the MSAmiP3-GFP mutant (SI Appendix, Methods).

**Data Availability.** The data discussed in this publication have been deposited in NCBI’s Gene Expression Omnibus (76) and are accessible through GEO Series accession number GSE165066 (77).
ACKNOWLEDGMENTS. We are indebted to Daniel Pauli for stimulating discussions and to Fabienne Clécard, Annick Mutter, Jorge Faustin, and Eva Favre for technical help. We thank the Cornell Statistical Consulting Unit for consultation, the Flow Cytometry core facility of the University of Geneva for secondary cell sorting, the iGE3 genomics platform for RNA sequencing, and Federico Miozzo for his help with RNA-Seq data analysis. C.L. thanks the members of François Payre laboratory and Muriel Boure for sequencing, and Federico Miozzo for his help with RNASeq data analysis. C.I. and Eva Favre for technical help. We thank the Cornell Statistical Consulting Unit for consultation, the Flow Cytometry core facility of the University of Geneva for secondary cell sorting, the iGE3 genomics platform for RNA sequencing, and Federico Miozzo for his help with RNA-Seq data analysis. C.L. thanks the members of François Payre laboratory and Muriel Boure for sequencing, and Federico Miozzo for his help with RNASeq data analysis. C.I. and Eva Favre for technical help. We thank the Cornell Statistical Consulting Unit for consultation, the Flow Cytometry core facility of the University of Geneva for secondary cell sorting, the iGE3 genomics platform for RNA sequencing, and Federico Miozzo for his help with RNA-Seq data analysis. C.L. thanks the members of François Payre laboratory and Muriel Boure for sequencing, and Federico Miozzo for his help with RNASeq data analysis. C.I. and Eva Favre for technical help. We thank the Cornell Statistical Consulting Unit for consultation, the Flow Cytometry core facility of the University of Geneva for secondary cell sorting, the iGE3 genomics platform for RNA sequencing, and Federico Miozzo for his help with RNA-Seq data analysis. C.L. thanks the members of François Payre laboratory and Muriel Boure for sequencing, and Federico Miozzo for his help with RNASeq data analysis. C.I. and Eva Favre for technical help. We thank the Cornell Statistical Consulting Unit for consultation, the Flow Cytometry core facility of the University of Geneva for secondary cell sorting, the iGE3 genomics platform for RNA sequencing, and Federico Miozzo for his help with RNA-Seq data analysis. C.L. thanks the members of François Payre laboratory and Muriel Boure for sequencing, and Federico Miozzo for his help with RNASeq data analysis. C.I. and Eva Favre for technical help. We thank the Cornell Statistical Consulting Unit for consultation, the Flow Cytometry core facility of the University of Geneva for secondary cell sorting, the iGE3 genomics platform for RNA sequencing, and Federico Miozzo for his help with RNA-Seq data analysis. C.L. thanks the members of François Payre laboratory and Muriel Boure for sequencing, and Federico Miozzo for his help with RNASeq data analysis. C.I. and Eva Favre for technical help. We thank the Cornell Statistical Consulting Unit for consultation, the Flow Cytometry core facility of the University of Geneva for secondary cell sorting, the iGE3 genomics platform for RNA sequencing, and Federico Miozzo for his help with RNA-Seq data analysis. C.L. thanks the members of François Payre laboratory and Muriel Boure for sequencing, and Federico Miozzo for his help with RNASeq data analysis. C.I. and Eva Favre for technical help. We thank the Cornell Statistical Consulting Unit for consultation, the Flow Cytometry core facility of the University of Geneva for secondary cell sorting, the iGE3 genomics platform for RNA sequencing, and Federico Miozzo for his help with RNA-Seq data analysis. C.L. thanks the members of François Payre laboratory and Muriel Boure for sequencing, and Federico Miozzo for his help with RNASeq data analysis. C.I. and Eva Favre for technical help. We thank the Cornell Statistical Consulting Unit for consultation, the Flow Cytometry core facility of the University of Geneva for secondary cell sorting, the iGE3 genomics platform for RNA sequencing, and Federico Miozzo for his help with RNA-Seq data analysis. C.L. thanks the members of François Payre laboratory and Muriel Boure for sequencing, and Federico Miozzo for his help with RNASeq data analysis.
64. S. Misra, M. F. Wolfner, Drosophila seminal sex peptide associates with rival as well as own sperm, providing SP function in polyandrous females. eLife 9, 1–21 (2020).