Sex chromosome dosage compensation in non-model insects

Dissertation

Zur Erlangung des Grades

Doktor der Naturwissenschaften

Am Fachbereich Biologie

Der Johannes Gutenberg-Universität Mainz

Agata Kalita

geboren am 22 Juni 1994

in Rzeszów, Polen

Mainz, 2024

Tag der mündlichen Prüfung: 11.12.2024

Zusammenfassung

Die geschlechtliche Fortpflanzung ist bei Tieren weit verbreitet, und das Geschlecht wird oft durch Geschlechtschromosomen bestimmt. Diese entwickeln sich aus einem Paar autosomaler Vorfahren, wobei ein Chromosom eine geschlechtsbestimmende Region erhält und im Laufe der Zeit genetische Aktivität verliert. Die Degeneration des Y-Chromosoms führt bei Männchen effektiv zu einer Aneuploidie des X-Chromosoms. Dieses Ungleichgewicht in der Gendosierung kann schädlich sein und wird häufig durch einen Mechanismus namens Dosiskompensation (DC) ausgeglichen. Das Verständnis der DC und anderer molekularer Unterschiede zwischen den Geschlechtern ist von großer Bedeutung, da diese Mechanismen oft geschlechtsspezifische Eigenschaften und Funktionen steuern. Beispielsweise sind es nur die weiblichen Stechmücken, die den Malariaparasiten übertragen.

Zu Beginn meiner Doktorarbeit war der Mechanismus, der die DC in *Anopheles gambiae*, dem Hauptüberträger der Malaria, steuert, noch unbekannt. Um potenzielle DC-Faktoren zu identifizieren, erstellte ich einen geschlechtsspezifischen RNA-Expressionsatlas der den Verlauf der Embryogenese abdeckt. Ich stellte fest, dass der Dosierungsausgleich kurz nach der Aktivierung des zygotischen Genoms einsetzt. Dabei entdeckte ich ein zuvor uncharakterisiertes Gen mit durchweg männlich geprägter Expression: *AGAP005748*. Aufgrund seiner anzunehmenden Funktion nannte ich dieses Gen *SOA* (Sex Chromosome Activation). Es stellte sich heraus, dass *SOA* zwei geschlechtsspezifische, alternativ gespleißte Isoformen produziert. Bei Männchen wird ein kanonisches Transkript exprimiert, während bei Weibchen das zweite Intron erhalten bleibt. Der Einschluss dieses Introns führt zu einem vorzeitigen Stoppcodon und zur Produktion eines verkürzten Proteins bei Weibchen.

Ich konnte zeigen, dass SOA an die Promotoren von X-chromosomalen Genen in männlichen Moskitos bindet. Anschließend untersuchte ich die Auswirkungen von *SOA* auf die Genexpression. Dabei stellte ich fest, dass *SOA*, wenn es ektopisch in einer weiblichen Zelllinie exprimiert wird, an die Promotoren aktiver X-chromosomaler Gene bindet und deren Hochregulierung bewirkt. Um die physiologische Funktion von *SOA* zu verstehen, SOA zu verstehen, generierte unser Kollaborationspartner Dr. Eric Marois transgene Moskitos, bei denen, bei denen durch ein CRISPR-generiertes Knock-in vor der kodierenden Sequenz das Gen *SOA* ausgeschaltet wurde (*SOA-KI*). In diesen Mücken stellte ich einen Signalverlust an den SOA-Bindungsstellen fest. Darüber hinaus ergab eine RNA-seq-Analyse eine Herabregulierung von X-chromosomalen Genen in *SOA-KI*-Männchen, was bestätigt, dass SOA tatsächlich in vivo die DC vermittelt.

Überraschenderweise waren *SOA-KI*-Mücken beiderlei Geschlechts lebensfähig. Allerdings zeigten männliche *SOA-KI*-Puppen im Vergleich zum Wildtyp eine Entwicklungsverzögerung, während Weibchen nicht betroffen waren. Zusätzlich erzeugten wir eine transgene Linie namens *SOA-R*, die SOA-cDNA exprimiert. Weibliche Stechmücken, die dieses Allel tragen, banden SOA ebenfalls an das X-Chromosom und zeigten sie eine erhöhte X-chromosomale Genexpression.

Die Erkenntnisse aus meiner Forschung, insbesondere die Entdeckung der Nichtessentialität der DC bei *Anopheles*, haben mein Interesse an der Evolution dieses Prozesses weiter vertieft. Daher habe ich die verfügbaren Studien zum DC-Status und dessen Mechanismen in verschiedenen Arten untersucht. Auf dieser Basis verfasste ich einen Übersichtsartikel, der das Wissen über die Evolution der Geschlechtschromosomen und der Dosiskompensation bei Insekten zusammenfasst. Darin habe ich auch Erkenntnisse aus meiner eigenen Arbeit integriert, um anderen Forschenden, die den DC-Mechanismus bei weiteren Insektenarten erforschen möchten, eine Hilfestellung zu bieten.

In meiner Dissertation habe ich dargelegt, dass SOA der Hauptregulator der DC des X-Chromosoms in *A. gambiae* ist. Darüber hinaus habe ich das Wissen über den DC-Status bei Insekten zusammengefasst und einen Rahmen vorgeschlagen, um DC-Faktoren auch in Nicht-Modell-Insektenarten zu identifizieren.

Abstract

Sexual reproduction is common among animals. The sex of the animal is often determined by sex chromosomes. Differentiated sex chromosomes evolve from autosomal progenitors, when one chromosome from the pair gains a sex-determining region and loses its genetic activity over time. Degeneration of the Y chromosome effectively leads to aneuploidy of the X chromosome in males. The resulting gene dosage imbalance can be detrimental and is frequently corrected by a mechanism named dosage compensation (DC). Understanding dosage compensation and other molecular differences between sexes is highly relevant. For example, only the female mosquitos bite and transmit the malaria parasite.

The mechanism regulating DC in *Anopheles gambiae,* the major vector of malaria, was unknown when I began my PhD work. To identify putative DC factors, I generated a sex-specific RNA expression atlas along embryogenesis. I observed that DC initiates shortly after zygotic genome activation. I then discovered a previously uncharacterized gene with consistently male-biased expression: *AGAP005748*. Based on its suspected function, I named this gene *SOA* (sex chromosome activation). I discovered that *SOA* produces two sex-specific, alternatively spliced isoforms. Males express a canonical transcript, while in females the second intron is retained. Intron inclusion results in a premature stop codon and production of a truncated protein in females.

I discovered that SOA binds the promoters of expressed X-linked genes in mosquito males. Next, I aimed to assess the effect of SOA on gene expression. I observed that SOA expressed ectopically in a female cell line binds gene promoters of active X-linked genes resulting in their upregulation. To comprehend its physiological function, our collaborator dr. Eric Marois generated transgenic mosquitoes devoid of *SOA* through a CRISPR-mediated knock-in of a gene trap upstream of the coding sequence (*SOA-KI*). In these mosquitos, I detected a loss of signal at SOA binding sites. Additionally, RNA-seq revealed downregulation of the X chromosome in *SOA* mutant males, confirming SOA indeed mediates DC *in vivo*. Surprisingly, *SOA-KI* mosquitos of both sexes are viable. Male *SOA-KI* pupae exhibited a developmental delay compared to wild-type, while females were unaffected. We also generated a transgenic line called *SOA-R* that expresses the full-length SOA cDNA. In female mosquitoes carrying this allele SOA also binds and upregulates X-linked genes.

The insights from this work, especially the discovery of non-essentiality of DC in *Anopheles* have made me even more interested in the evolution of this process. Because of this, I surveyed the published works investigating the DC status and mechanism in different species. I wrote a review article summarizing the knowledge about sex chromosome evolution and DC in insects. In this article, I also used the insights from my previous work to provide guidance to researchers who aim to identify DC mechanisms in other insect species.

In my PhD work, I demonstrated that SOA is the master regulator of X chromosome dosage compensation in *A. gambiae*. I also summarized the knowledge about the DC status across insects and proposed a framework to uncover the DC factors in non-model insect species.

Preface

The results presented in this thesis have been previously published:

Publication 1: Kalita, A. I., Marois, E., Kozielska, M., Weissing, F. J., Jaouen, E., Möckel, M. M., Rühle, F., Butter, F., Basilicata, M. F., & Keller Valsecchi, C. I. (2023). The sex-specific factor SOA controls dosage compensation in Anopheles mosquitoes. *Nature*, *623*(7985), 175–182.

Publication 2: Kalita, A. I., & Keller Valsecchi, C. I. (2024). Dosage compensation in non-model insects – progress and perspectives. *Trends in Genetics*. https://doi.org/10.1016/j.tig.2024.08.010

Both publications are preceded by their summary and a statement of my contribution.

Table of contents

Chapter 1

Introduction

1.1. Sex chromosomes

Sexual reproduction is a common feature of eukaryotes (Goodenough & Heitman 2014). The most common sex determination systems are genotypic, i.e. the sex is determined by sex chromosomes (Tree of Sex Consortium 2014). These systems are classified based on which sex possesses two different sex chromosomes. In the male-heterogametic species, males have X and Y chromosomes, hence they produce two types of gametes: some containing the X, others the Y chromosome. In such species, the females carry two X chromosomes. In the female-heterogametic species, females harbor the Z and W chromosomes, while males carry two Z chromosomes.

Sex chromosomes evolve from autosomal progenitors (Charlesworth 1991). The classical model of sex chromosome evolution involves a gain of a sex-determining region followed by suppression of recombination, accumulation of mutations on the Y (or W) chromosome, and progressive loss of genetic activity of the Y/W (Charlesworth et al. 2005). The degeneration of the Y/W chromosome means that functionally, the heterogametic sex has only one copy of many genes on the X/Z chromosome. Surprisingly, this functional monosomy is well tolerated. This is in contrast to aneuploidies of other chromosomes. In humans, for example, monosomies of any of the autosomes are lethal (Hassold & Hunt 2001). Why is the functional monosomy of sex chromosomes so well tolerated and compatible with development? After all, since the X evolved from an autosomal progenitor, it is likely to contain genes relevant for both sexes, such as housekeeping genes. For at least some species, the answer is dosage compensation - a process that can equalize the expression of genes on the sex chromosomes (Gu & Walters 2017; Mank 2013).

1.2. Dosage compensation

Dosage compensation was first observed in insects. In 1932 H.J. Muller noticed that male flies with a partial loss-of-function mutant allele of a X-linked gene responsible for eye color had the same phenotype as females with two copies (Muller 1932). The phenotype was stronger than in females with one copy of the allele and one gene deletion. Since the males had only one copy of the gene, he concluded that there must be a mechanism that equalizes

the amount of X-linked gene products between the sexes. He called this process “dosage compensation”.

Much later, in 1961, Mary Lyon hypothesized that one of the female X chromosomes is randomly inactivated in mammals (Lyon 1961). In 1967, Susumu Ohno proposed his two-step model of dosage compensation evolution. He wrote that the degeneration of sex chromosomes leads to expression that is not optimal compared to the ancestral state - what he called “the peril of hemizygosity” (Ohno 1967). Hence, he posited that the expression of the X-linked genes becomes upregulated in the course of evolution to match the ancestral levels (step 1). This makes the expression optimal for males, but not females. Hence, in the second step, one of the female X chromosomes becomes silenced and brings the expression down to optimum in the females as well.

This theory was developed based on mammals, and the two-step process does not seem necessary in other species. For example, in *Drosophila*, the upregulation is specific to the single male X chromosome, with no silencing of the female X. Even in mammals, it is still controversial whether dosage compensation, defined as matching ancestral expression, is achieved. While X chromosome inactivation in females has been demonstrated repeatedly, the proof for upregulation posited in Ohno’s first step is still mixed (recently reviewed by (Cecalev et al. 2024)).

Dosage compensation (DC) has evolved independently in multiple lineages, but only a few mechanisms have been elucidated (Basilicata & Keller Valsecchi 2021). Dosage compensation mechanisms have been intensely studied in mammals, as well as the invertebrate model organisms *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster*. *D. melanogaster* has been the insect model of choice for research into dosage compensation. As a model organism, amenable to genetic manipulations and with many genetic tools available, it made it possible to study dosage compensation before genomics was widely applied.

1.3. Dosage compensation mechanism in *Drosophila melanogaster*

D. melanogaster is male-heterogametic, with males having XY and females XX chromosome complement. The expression of X-linked genes is balanced between the sexes and fully dosage compensated (Vicoso & Bachtrog 2015). The sex of *D. melanogaster* is determined based on the X-to-autosome ratio. In females, where the X:A ratio equals 1,

sex-lethal expression is high and initiates the female development program (Schütt & Nöthiger 2000). Sex-lethal also inhibits the productive splicing and translation of *male-specific lethal 2* (*msl2*). In effect, MSL2 protein is only present in the males. There, it interacts with two male-specific X-linked long-noncoding RNAs: *roX2* and *roX1*, which are important for targeting the MSL complex to the X chromosomes (Valsecchi et al. 2020). The complex also consists of other components (Figure 1). MLE (*maleless*) is a helicase which remodels the secondary structure of *roX* RNA (Maenner et al. 2013; Ilik et al. 2013). MSL1 (*male-specific lethal 1)* forms a scaffold needed for complex assembly (Kadlec et al. 2011). MOF (*males-absent-on-the-first*) is an acetyltransferase, which deposits an H4K16 acetylation mark on the single male X chromosome (Hilfiker et al. 1997). The activity of MOF is enhanced by MSL3 (*male-specific lethal 3)* and MSL1 (Morales et al. 2004). MOF-deposited H4K16ac increases gene expression from the single male X chromosome (Akhtar & Becker 2000).

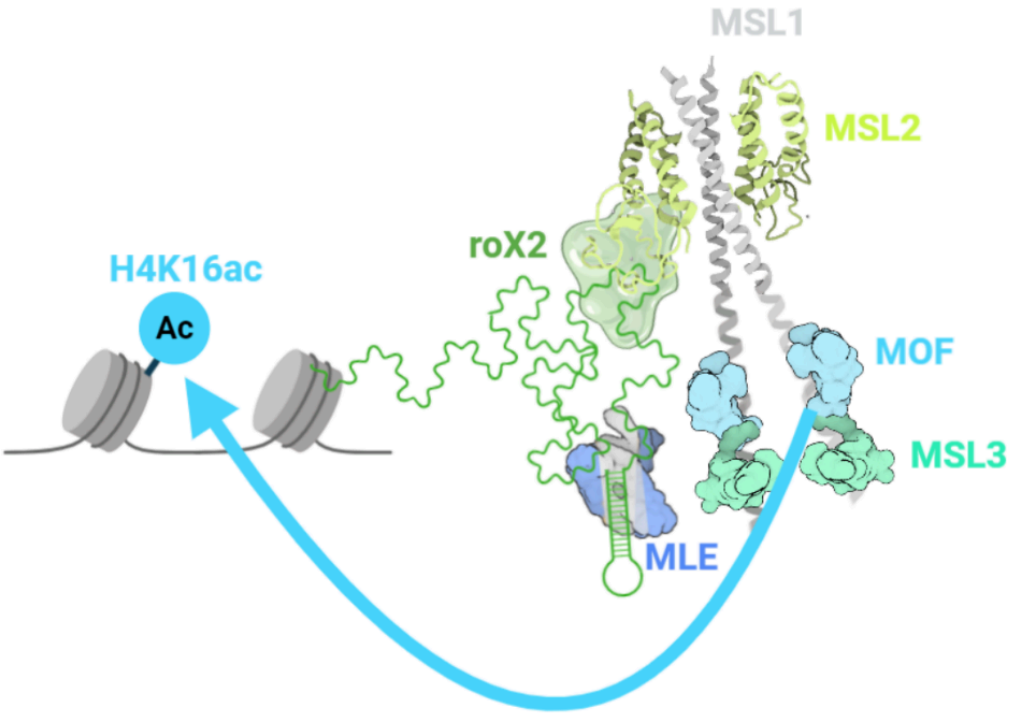


Figure 1. The dosage compensation complex (DCC) and its activity in *D. melanogaster*. Figure created in Biorender.

Dosage compensation can be important for proper development of the organism. In *Drosophila,* null mutation of *msl-2* or other MSL complex subunits is associated with male-specific lethality at the early pupae stage (Belote & Lucchesi 1980; Franke & Baker 1999). Lethality is also observed for males with constitutive *sex-lethal* expression (Cline 1978).

1.4. Relevance of research into dosage compensation in Anopheles gambiae

Malaria is a global public health problem, claiming over 608 000 lives globally in 2015, with 249 million infections that year (World Health Organization 2023). Malaria is caused by a parasite from the *Plasmodium* genus, transmitted most effectively by the *Anopheles gambiae* species. Mosquitos are highly sexually dimorphic species, where only females are blood-feeding and therefore can be a vector for malaria (Cox 2010). Vector-control approaches are often designed to take advantage of differences between sexes to decrease female fitness and fecundity, consequently curbing the spread of malaria. Hence, it is important to understand the molecular mechanisms underlying the differences between mosquito sexes. Despite their importance, little was known about such processes, especially dosage compensation, when I started my PhD.

1.5. Dosage compensation in *Anopheles* mosquitoes

Fruit flies and mosquitos are closely related insects and belong to the same order of *Diptera*. Like in flies, male *Anopheles* mosquitoes carry a single X chromosome, while females have two (Bonaccorsi et al., 1980). Sex chromosomes in these two species have evolved independently, yet from the same pair of ancestral autosomes (Vicoso & Bachtrog 2015). The differentiated sex chromosomes have a shared origin across the *Anopheles* genus and formed after the separation of *Anopheles* from *Culicinae* (between 160 and 100 million years ago) (Neafsey et al. 2015).

Unlike in *D. melanogaster,* the data on the molecular basis of dosage compensation in *A. gambiae* was limited when I first started the project. Published RNA-seq studies are consistent with complete dosage compensation of the single male X chromosome (Jiang et al. 2015; Rose et al. 2016) achieved by hyperactivation as in *Drosophila* (Deitz et al. 2018). Dosage compensation was shown to be active already in the larval stage of mosquito development, but the exact timing of the onset of DC during embryogenesis has not been described before.

In *Anopheles* mosquitoes, the sex is determined by a Y linked factor which is not conserved between species. It has been proposed that in the *Anopheles* genus, dosage compensation occurs downstream of such Y-linked master switch genes (Qi et al. 2019; Krzywinska & Krzywinski 2018). In *A. stephens*i, the master sex determination gene is Guy1.

Expression of Guy1 in female embryos leads to an upregulation of X-linked genes and female-specific lethality (Criscione et al. 2016; Qi et al. 2019). In *A. gambiae*, the signal triggering male development is the Y-linked factor *Yob*. Ectopic expression of *Yob* in the embryos has mixed results: when delivered to preblastoderm embryos as mRNA, no females developed to pupae (Krzywinska et al. 2016). In contrast, a stable line expressing Yob leads to lethality in a fraction of females - a percentage correlated with the expression levels of the transgene (Krzywinska & Krzywinski 2018). However, the expression of X-linked genes was not assessed in *Yob*-expressing females. A depletion of another component of the sex determination cascade, *femaleless*, leads to lethality of female embryos in *A. gambiae* and *A. stephens*i (Krzywinska et al. 2021). The extent of depletion is correlated with the penetrance of the lethality phenotype. Moreover, the depletion of *fle* leads to specific upregulation of X-linked genes in transgenic compared to wild-type females.

My PhD supervisor demonstrated that despite relative evolutionary proximity to *Drosophila*, *A. gambiae* species has evolved a distinct and entirely novel mechanism of DC (Keller Valsecchi et al. 2021). Although most subunits of the MSL complex in *Drosophila* are also conserved in *A. gambiae*, *msl-2* null mutant mosquitos do not show a sex-specific lethality at the larval stage, but rather a severe developmental arrest at embryonic stages that occurs in both sexes. *Anopheles msl-2* mutant embryos display no X-chromosome specific downregulation of gene expression by RNA-seq, which contrasts the chromosome-specific misregulation observed in the *msl-2* null mutant *Drosophila* embryos. The striking enrichment of H4K16 acetylation on the X chromosome of *Drosophila* males observed in ChIP-seq experiments was also not mirrored in *A. gambiae*.

The goal of my PhD project was to understand the molecular mechanism of dosage compensation in *A. gambiae.* We aimed to identify the timing of dosage compensation, characterize the factors involved, and determine the level of gene regulation responsible for this process.

Chapter 2

Publication 1: The sex-specific factor SOA controls dosage compensation in *Anopheles* mosquitoes

2.1. Summary

In many species, males carry a single X chromosome while females carry two. Dosage compensation is a mechanism equalizing the expression of X-linked genes between sexes. Very few dosage compensation mechanisms have been elucidated to date. In this publication, we discovered the master regulator of dosage compensation in the malaria-transmitting mosquito *Anopheles gambiae*.

We generated a sex-specific embryonic RNA expression atlas to identify putative dosage compensation factors. We discovered that a previously uncharacterised gene, AGAP005748, is consistently male-biased. We named the gene *SOA* (for Sex chrOmosome Activation). SOA produces two sex-specific splicing isoforms. Males express a canonical transcript while females retain the second intron, resulting in a truncated protein.

We discovered SOA specifically binds the promoters of expressed X-linked genes in males. Ectopic expression of the full-length SOA in the female cell line upregulated X-linked genes. SOA-lacking (*SOA-KI)* males lose SOA binding, resulting in X chromosome downregulation. Although viable, *SOA-KI* males exhibited a 4-hour developmental delay compared to wild-type males, while females were unaffected. Accordingly, expression of full-length SOA in female mosquitos resulted in increased X-linked gene expression and a female-specific developmental delay due to aberrant dosage compensation. Taken together, this work demonstrates that SOA is the master regulator of X chromosome dosage compensation in mosquitos.

2.2. Candidate's contribution

I participated in the conceptualization and planning of the study. I performed the cell culture experiments, as well as processed material from mosquitoes for RNA isolation and performed RT-PCR and RT-qPCR. I analyzed the RNA-seq data. I performed the ATAC-seq and CUT&Tag experiments, and plotted the data after the mapping and differential binding analysis was performed. I co-wrote the manuscript together with my supervisor, with input from other authors.

Supervisor's signature: _____

Article

The sex-specific factor SOA controls dosage compensation in *Anopheles* mosquitoes

https://doi.org/10.1038/s41586-023-06641-0

Received: 16 September 2022

Accepted: 13 September 2023

Published online: 28 September 2023

Open access

Check for updates

Agata Izabela Kalita[1], Eric Marois[2,6], Magdalena Kozielska[3], Franz J. Weissing[3], Etienne Jaouen[2], Martin M. Möckel[1], Frank Rühle[1], Falk Butter[1,4], M. Felicia Basilicata[1,5,6] & Claudia Isabelle Keller Valsecchi[1,6]✉

The *Anopheles* mosquito is one of thousands of species in which sex differences play a central part in their biology, as only females need a blood meal to produce eggs. Sex differentiation is regulated by sex chromosomes, but their presence creates a dosage imbalance between males (XY) and females (XX). Dosage compensation (DC) can re-equilibrate the expression of sex chromosomal genes. However, because DC mechanisms have only been fully characterized in a few model organisms, key questions about its evolutionary diversity and functional necessity remain unresolved[1]. Here we report the discovery of a previously uncharacterized gene (*sex chromosome activation* (*SOA*)) as a master regulator of DC in the malaria mosquito *Anopheles gambiae*. Sex-specific alternative splicing prevents functional SOA protein expression in females. The male isoform encodes a DNA-binding protein that binds the promoters of active X chromosomal genes. Expressing male SOA is sufficient to induce DC in female cells. Male mosquitoes lacking SOA or female mosquitoes ectopically expressing the male isoform exhibit X chromosome misregulation, which is compatible with viability but causes developmental delay. Thus, our molecular analyses of a DC master regulator in a non-model organism elucidates the evolutionary steps that lead to the establishment of a chromosome-specific fine-tuning mechanism.

Malaria is a life-threatening disease, with 241 million cases and 627,000 deaths reported by the World Health Organization in 2021 (ref. 2). It is caused by *Plasmodium* parasites and is transmitted most effectively by mosquitoes of the *A. gambiae* species complex. Mosquitoes are sexually dimorphic, with only females being able to take blood and thereby transmit malaria. However, despite the high relevance of understanding the molecular basis of sexual dimorphism in *Anopheles*, the onset and development of sexually distinct gene-expression pathways have been little studied to date.

Anopheles mosquitoes have heteromorphic sex chromosomes, in which males are XY and females are XX. Sex chromosomes generally evolve from a pair of ancestral autosomes, a process in which the Y chromosome typically becomes highly degenerated and is left with only few functional genes[1]. One of the Y-linked genes in *A. gambiae* is the master-switch gene of sexual differentiation *Yob*, which triggers maleness[3]. Along with sex chromosome differentiation, some species evolve DC, which corrects the expression imbalance of the X chromosomal genes (one in males compared with two in females; ZZ/ZW are not discussed here for simplicity)[1]. Transcriptome studies performed at the pupal and adult stages have revealed complete DC of the single male X chromosome in several *Anopheles* species[4–7].

Fruit flies and *Anopheles* mosquitoes belong to the same insect order Diptera. Their X chromosomes evolved independently but from the same ancestral autosome; hence, their X chromosomes and the encoded genes are similar[8,9]. *Drosophila melanogaster* is one of only three model organisms for which the molecular cascades that mediate DC have been elucidated[10]. The master regulator of *Drosophila* DC, the male-specific lethal 2 protein (MSL2) is only present in males. MSL2 recruits the MSL complex to the X chromosome, where the deposition of histone H4 lysine 16 acetylation (H4K16ac) contributes to an approximately twofold increase in gene expression. Loss of any MSL complex subunit causes male-specific lethality[11]. Conversely, ectopic expression of MSL2, but none of the other MSL subunits, is sufficient to induce X chromosome upregulation in females, which can trigger lethality[11,12].

Although *A. gambiae* and *D. melanogaster* have similar X chromosomes and both exhibit X chromosome upregulation, mosquitoes do not achieve DC through MSL2 and the H4K16ac pathway[13]. Until now, the genes and mechanisms that mediate DC in *Anopheles* remained unknown.

SOA produces sex-specific isoforms

To uncover *A. gambiae* DC factors, we determined the developmental window of DC onset using RNA sequencing (RNA-seq) (Fig. 1a). We observed a substantial imbalance between the sexes in the expression of X-linked but not autosomal genes shortly after zygotic genome activation (ZGA). This imbalance was compensated by 5–9 h of embryogenesis, with further fine-tuning at later stages. We then searched

[1]Institute of Molecular Biology (IMB), Mainz, Germany. [2]INSERM U1257, CNRS UPR9022, Université de Strasbourg, Strasbourg, France. [3]Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, Netherlands. [4]Institute of Molecular Virology and Cell Biology, Friedrich Loeffler Institute, Greifswald, Germany. [5]Institute of Human Genetics, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany. [6]These authors contributed equally: Eric Marois, M. Felicia Basilicata, Claudia Isabelle Keller Valsecchi. ✉e-mail: c.keller@imb-mainz.de

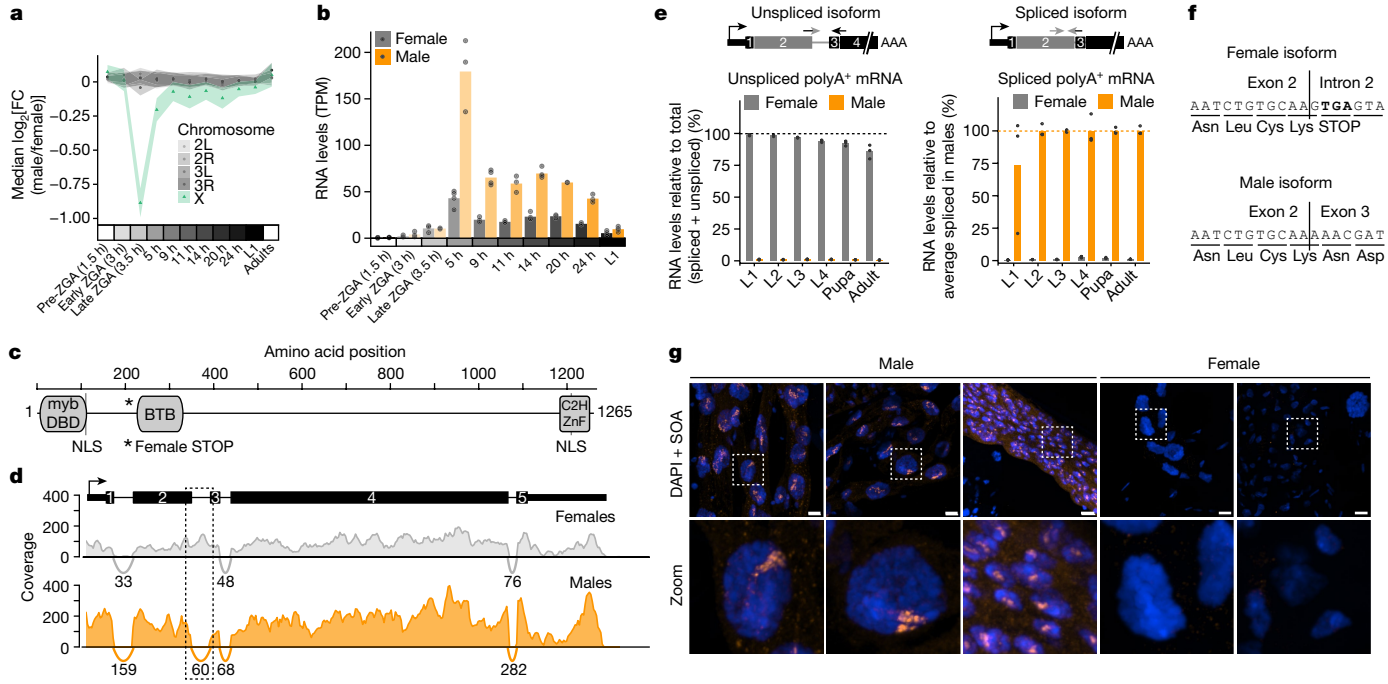


Fig. 1 | Identification of the sex-specifically spliced *SOA* gene. a, Dot plot showing the median $\log_2$ fold change ($\log_2$(FC)) of RNA levels between males and females from single-embryo RNA-seq (shading indicates 95% confidence intervals). Genes with read count > 0 were grouped on the basis of chromosomal location. Raw data points and replicate numbers provided in Supplementary Table 3. Adult dataset from ref. 4. L1, first instar larva. **b**, Bar plot showing *SOA* RNA levels from RNA-seq in transcripts per million (TPM). Overlaid data points are biological replicates. **c**, Scheme of the protein domain architecture of SOA. NLS, nuclear localization signal. **d**, RNA-seq coverage and splice junctions (arcs) at the *SOA* locus at 11 h of embryogenesis in females and males. Read numbers spanning respective exon–exon junctions are shown below the arcs (Supplementary Table 1). **e**, RT–qPCR quantification of polyadenylated (polyA$^+$) *SOA* mRNA isoform levels in females and males at larval (L1–L4), pupal and adult stages. The scheme (top) shows the primer strategy. Left, percentage unspliced relative to total (spliced and unspliced) mRNA levels. Right, percentage spliced mRNA relative to the average male spliced mRNA level at each stage. The bars represent the mean of $n = 2$ or $n = 3$ independent biological replicates indicated by overlaid data points. *Rp49* was used for normalization (Extended Data Fig. 4c and raw data in Supplementary Table 1). **f**, Nucleotide and amino acid sequence of the exon 2–intron 2 junction (female isoform) and exon 2–exon 3 junction (male isoform). **g**, Representative SOA immunostaining (orange) and DAPI (blue) conducted on adult mosquito tissues (Malpighian tubules or gut). Images on the bottom row are close-ups of the white square in the above images. Images represent 3D views of a *z*-stack. Scale bar, 10 µm. Complete panel with single channels and additional staining shown in Extended Data Fig. 5g.

for transcripts that were male-biased from 5 h onwards (Fig. 1b and Extended Data Fig. 1a). This analysis uncovered *Yob*, which encodes the Y-linked, male master sex determination gene[3], and *AGAP005748*, an uncharacterized protein-coding gene that we name after its putative function: *sex chromosome activation* (*SOA*). *SOA* encodes a 1,265 amino acid protein with three predicted domains: a myb DNA-binding domain; a broad-complex, tramtrack and bric à brac (BTB) (also known as POZ) domain; and a C2H2 zinc finger (ZnF) (Fig. 1c). It evolved through a tandem gene duplication event from *AGAP005747*. *SOA* orthologues are present in Anophelinae but not in Culicinae (for example, *Aedes aegypti)* (Extended Data Figs. 1b–h, 2 and 3a,b, Supplementary Table 1 and Supplementary Note 1). The lack of *SOA* in Culicinae is consistent with the absence of heteromorphic sex chromosomes in this subfamily, which therefore obviates the need for chromosome-wide DC.

SOA produces two sex-specific, alternatively spliced mRNA isoforms. Males express a canonical transcript, whereas females retain the second intron (Fig. 1d). This pattern is conserved among *Anopheles* (Extended Data Fig. 4a). We performed a gene-specific reverse transcription coupled to PCR (RT–PCR) experiment and found that after ZGA, *SOA* splicing seems identical between sexes, with both isoforms present. Shortly thereafter, a sex-specific pattern is established, which persisted in all post-embryonic stages (Extended Data Fig. 4b). Quantification of the polyadenylated *SOA* mRNA isoforms by quantitative RT–PCR (RT–qPCR) revealed that males express around 100-fold more spliced isoform than females (Fig. 1e, Extended Data Fig. 4c and Supplementary Table 1). Notably, intron retention led to the presence of an in-frame premature stop codon (Fig. 1f), which is evolutionarily conserved (Extended Data Fig. 4d) and only allows the production of a truncated 229 amino acid protein. We note that this in-frame stop codon could provide an explanation for the lower overall transcript levels in females (approximately 3–6-fold less; Extended Data Fig. 4c), as it could trigger the nonsense-mediated decay pathway[14].

To analyse the SOA protein, we generated an antibody against the amino-terminal myb domain compatible with detecting male and female isoforms (validation in Extended Data Fig. 5a–e; see also Supplementary Table 1 and Methods). Because endogenous SOA was below the detection limit of western blotting, we used mass spectrometry to capture SOA after immunoprecipitation (IP). As predicted, we only detected peptides corresponding to the short SOA(1–229) isoform in females, whereas peptides covering the full-length male SOA(1–1265) protein were exclusively found in males (Extended Data Fig. 5f and Supplementary Table 1). We then performed immunofluorescence (IF) stainings of adult mosquito tissues. SOA localized to a distinct subnuclear territory in males, whereas no specific staining could be detected in females (Fig. 1g ; full panel in Extended Data Fig. 5g). The male-specific SOA territory was also observed in imaginal discs of the fourth larval stage 4 (L4) and interphase cells of embryos (Extended Data Fig. 5h–j).

SOA binds X chromosomal gene promoters

Because localization in a nuclear territory is a hallmark of DC[15,16], we investigated whether SOA is associated with the X chromosome.

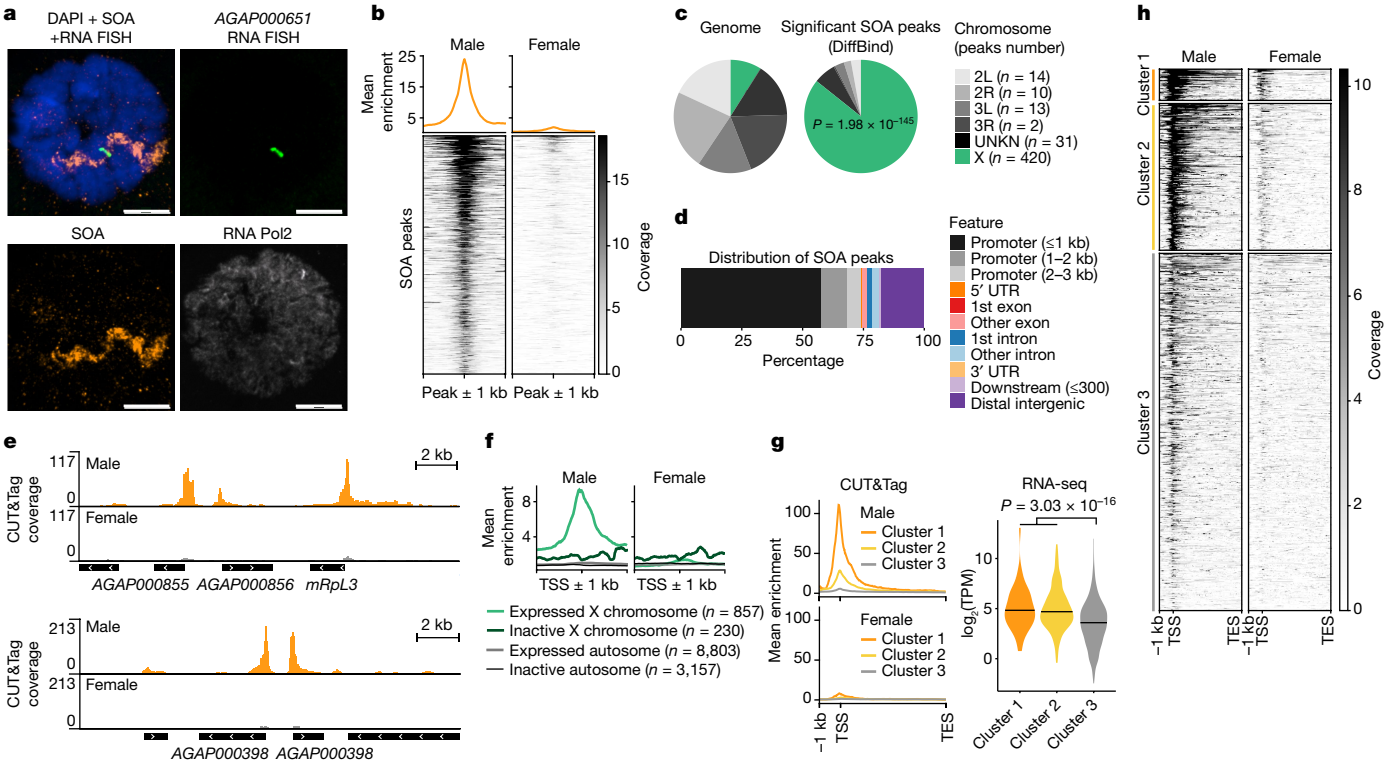


Fig. 2 | SOA binds to male X chromosomal gene promoters. a, Representative immunostaining of SOA (orange), RNA polymerase 2 (Pol2; grey) with RNA FISH (green) of a X-linked transcription site (*AGAP000651* intron). DAPI in blue. Scale bar, 10 µm. **b**, Heatmap showing normalized SOA CUT&Tag coverage for significant peaks (males versus females) and metaplot showing mean enrichment (top). **c**, Pie chart of the significant SOA peaks versus the *A. gambiae* genome. *P* value: one-sided Fisher's test for overrepresentation of peaks on the X chromosome.UNKN, scaffolds that could not be assigned to any chromosome. **d**, Bar plot of SOA peak annotations for genomic features. UTR, untranslated region. **e**, Genome browser snapshots of SOA CUT&Tag coverage. **f**, Metaplot of SOA CUT&Tag coverage at the TSS ± 1 kb (all genes). Lines reflect gene groups by chromosomal location and expression levels based on RNA-seq of wild-type male pupae. Genes with fewer than ten average read counts across replicates were considered as not expressed. **g**, Left, metaplot of SOA CUT&Tag coverage at 3 random *k*-means clusters generated from expressed, X-linked genes ($n = 857$ genes, see also **f**). The TSS is a reference point to plot 1 kb upstream; gene bodies (TSS to the transcription end site (TES)) were scaled to 5 kb. Right, violin plot of $\log_2$(TPM) values by RNA-seq of wild-type male pupae. The centre line indicates the median. *P* value: two-sided Wilcoxon rank-sum comparing combined clusters 1 and 2 versus cluster 3. **h**, As in **g**. Heatmap showing the SOA CUT&Tag coverage at expressed X-linked genes. Three random *k*-means clusters were generated that separated the groups on the basis of SOA binding strength. Biological replicates ($n = 4$ male, $n = 2$ female) were merged for visualization (**b**,**e**–**h**).

In stainings of polytene chromosome preparations from L4 larvae, SOA decorated one chromosome of males, but not females (Extended Data Fig. 6a). SOA staining overlapped with the transcription site of the X-linked *AGAP000651*, as visualized by RNA fluorescence in situ hybridization (FISH) and SOA IF (Fig. 2a). To investigate what genomic regions SOA binds to, we used the CUT&Tag method, in which a protein A (pA)–Tn5 transposase fusion protein is directed to an antibody-bound target (SOA) on chromatin[17]. In situ visualization of the DNA sequences tagmented by pA–Tn5 with fluorescent oligonucleotides (CUT&See) revealed an overlap with the male SOA territory by IF (Extended Data Fig. 6b). CUT&Tag sequencing was then performed using male and female pupae with the SOA antibody and an IgG control (Extended Data Fig. 6c and Methods). After differential binding analysis comparing males and females, we identified a total of 490 peaks with significant enrichment in males, but only 39 with significant enrichment in females (Fig. 2b and Supplementary Table 2). In total, 420 of the male-specific peaks were localized to the X chromosome (Fig. 2c and Extended Data Fig. 6d). The majority of them were found at gene promoters, typically residing within 1 kb of the transcription start site (TSS) (Fig. 2d,e and Extended Data Fig. 6e). Because DC is expected to affect expressed, but not inactive genes, we grouped all *A. gambiae* genes on the basis of their chromosomal location and expression status. Using this approach, which is independent of peak calling, we observed SOA binding exclusively at the promoters of X-linked expressed genes ($n = 857$), but at none of the other three groups (Fig. 2f). Further analysis of these 857 genes by unsupervised clustering distinguished them on the basis of the strength of SOA binding: $n = 50$ genes with strong binding, $n = 230$ genes with intermediate binding and $n = 577$ genes with weak binding (Fig. 2g,h). Cluster 3 (weak SOA binding) showed significantly lower RNA expression levels compared with cluster 1 and cluster 2 genes (Fig. 2g and Supplementary Table 3). To identify DNA sequence motifs bound by SOA, a MEME motif analysis of SOA peaks was performed. Three motifs were enriched, of which a simple CA dinucleotide repeat sequence was the most significant (Extended Data Fig. 6f). Last, investigation of the few autosomal peaks bound in males showed that they display specific but reduced enrichment levels (Extended Data Fig. 6g,h). Most of these peaks were located to genes close to telomeres (Supplementary Table 2). We speculate that the spatial proximity to the X chromosome territory could cause their binding.

Male SOA is sufficient to induce DC

Having established that SOA specifically binds the X chromosome, we set out to assess its effect on gene expression and asked whether it is sufficient to induce DC. To this end, we ectopically expressed either the male or female isoform in a cell line without DC; that is, female Ag55 cells (Fig. 3a). We performed RNA-seq (Extended Data Fig. 6i and Methods) and found that after expression of the female SOA(1–229)

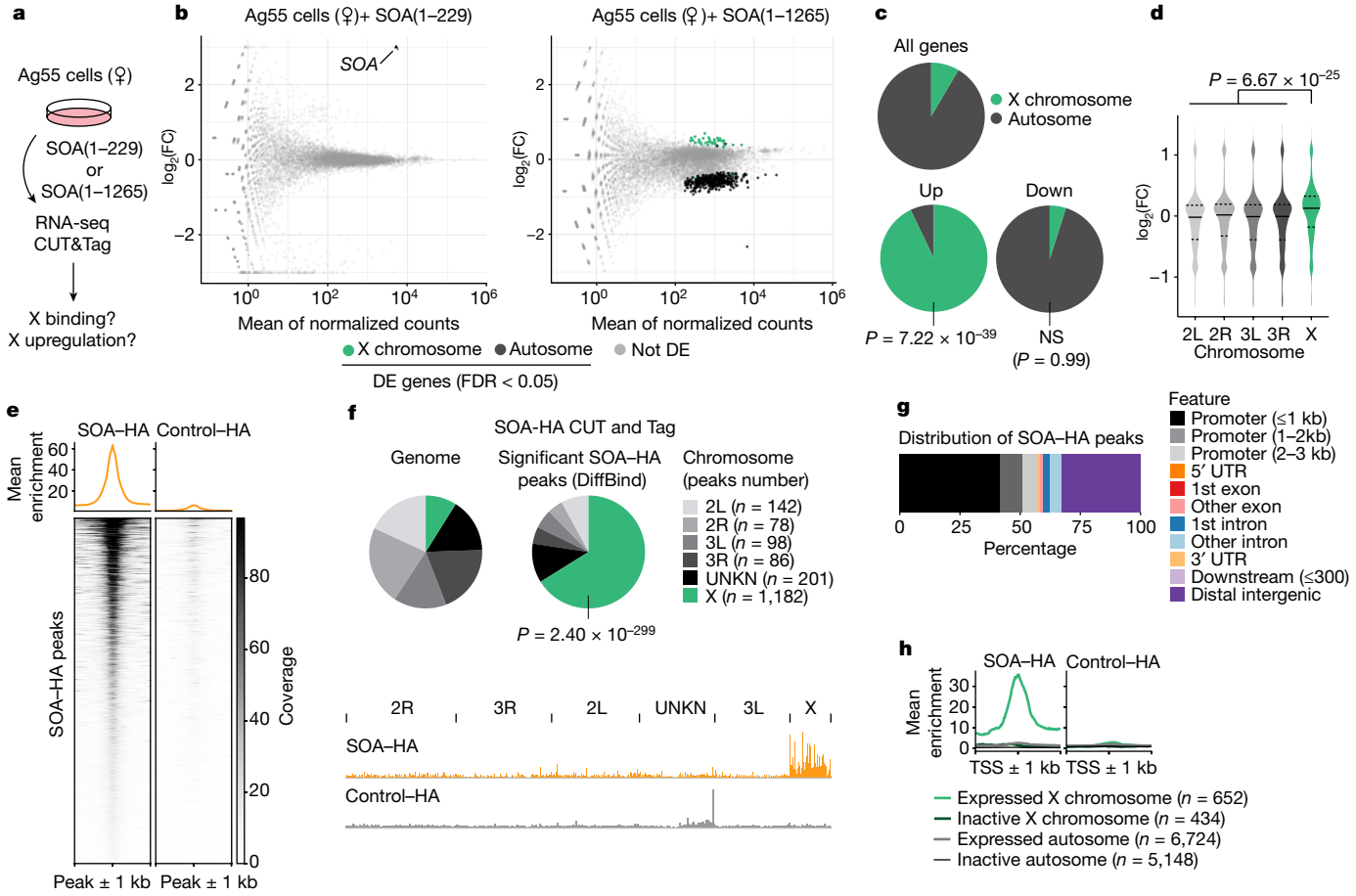


Fig. 3 | Expression of male SOA is sufficient to induce DC. a, Scheme illustrating transient expression of female isoform (SOA(1–229)–HA), male isoform (SOA(1–1265)–HA) or empty vector control with baculovirus in female Ag55 cells. **b**, MA plots from RNA-seq ($n = 3$ biological replicates) showing normalized read counts versus $\log_2$(FC) comparing SOA(1–229) with empty vector control (left) or SOA(1–1265) with SOA(1–229) (right). Differentially expressed (DE) genes are green (X chromosome) or black (autosomes), others are grey. Arrow indicates *SOA* (triangle) and cistronic *eGFP* (circle). FDR, false discovery rate. **c**, As in **b**. Pie charts of differentially expressed and all *A. gambiae* genes. *P* value: one-sided Fisher's test for overrepresentation of X-linked genes. NS, not significant. **d**, As in **b**. Violin plot of $\log_2$(FC) values of female Ag55 cells with SOA(1–1265). The centre line indicates the median. All genes with average read count > 0 were plotted. Median $\log_2$(FC) for X-chromosomal genes equals 0.122 (FC = 1.088). *P* value: two-sided Wilcoxon rank-sum test comparing X-linked versus autosomal genes. **e**, Heatmap showing normalized CUT&Tag coverage on significant peaks in Ag55 cells expressing SOA(1–1265) versus empty vector control ($n = 2$ biological replicates merged for visualization) and mean enrichment as a metaplot. **f**, As in **e**. Top, pie chart of significant CUT&Tag peaks. *P* value: one-sided Fisher's test for overrepresentation of peaks on the X chromosome. Bottom, genome browser snapshot of CUT&Tag coverage. **g**, As in **e**. Bar plot of SOA–HA peak annotations for genomic features. **h**, As in **e**. Metaplot of CUT&Tag coverage at the TSS ± 1 kb (all genes). Lines reflect gene groups by chromosomal location and expression levels based on RNA-seq of empty vector control Ag55 cells. Genes with fewer than ten average read counts across replicates were considered as not expressed.

isoform, there was only a single differentially expressed gene compared with the empty vector control-*SOA* itself (Fig. 3b and Extended Data Fig. 6j). By contrast, ectopic expression of male SOA(1–1265) induced a global upregulation of X chromosomal genes (Fig. 3b,c), irrespective of whether a gene was scored as differentially expressed or not (Fig. 3d). The differentially expressed genes upregulated by SOA were almost exclusively X-linked (Fig. 3c). This was accompanied by the downregulation of many genes on autosomes, probably as a secondary consequence of perturbed transcription regulators encoded on the X chromosome (for example, *AGAP000189*; Supplementary Table 2).

To analyse the SOA binding pattern in this ectopic system, we performed CUT&Tag using the HA tag present in our constructs (Extended Data Fig. 7a and Methods). A total of 1,787 peaks were scored significant for being more strongly bound by SOA(1–1265) compared with the empty vector control (Fig. 3e). Out of these, 1,182 (66%) localized to the X chromosome (Fig. 3f). As in the in vivo context (Fig. 2d,f), SOA–HA associated with active X chromosomal promoters (Fig. 3g,h and Extended Data Fig. 7b) and showed substantial enrichment at highly expressed genes (Extended Data Fig. 7c,d). Motif analysis also revealed binding to CA repeats (Extended Data Fig. 7e). Overall, the binding profiles of endogenous SOA in tissue and SOA–HA in cells were similar (Extended Data Fig. 7f,g). The improved signal-to-noise ratio explains the higher total number of significant peaks called in cells, whereas the non-endogenous *EF1a* promoter used in that context appeared to cause some spillover to autosomal genes, at which endogenous SOA is not found (Extended Data Fig. 6g,h).

We investigated whether SOA localization depended on an RNA co-factor such as roX1/roX2 (ref. 16) or Xist[18]. However, the SOA territory localization observed by IF remained intact after treatment with RNase A (Extended Data Fig. 7h). Similarly, X chromosome binding of SOA was insensitive to transcription inhibition by actinomycin D (Extended Data Fig. 7i,j). To investigate the potential involvement of a DNA-guided mechanism in X chromosome recruitment, we directed our attention towards the CA-repeat motif. First, we used the RepeatMasker annotation to analyse the distribution of repeats on the different chromosomal arms (Extended Data Fig. 8a–d). Second, we used

the FIMO tool to search the top-scoring $(CA)_7$ motif sequence in *A. gambiae* in comparison to *A. aegypti* (no DC, therefore used as a control) (Extended Data Fig. 8e,f). The RepeatMasker approach revealed that the X chromosome per se is repeat-rich (Extended Data Fig. 8a). Moreover, simple repeats such as $(CA)_n$ sequences were not only highly abundant, but were among the repeat families that are enriched on the X chromosome (Extended Data Fig. 8d). Both RepeatMasker and FIMO analyses showed that compared to autosomes, the frequency and length of X-linked CA repeats were significantly higher (Extended Data Fig. 8b,c,f). Such features are not observed in *A. aegypti*[19] (Extended Data Fig. 8e,f), which indicated that the SOA-bound motif is specific to the *Anopheles* X chromosome.

Next, we investigated how the different SOA protein domains (Extended Data Fig. 8g–i) contribute to CA-repeat binding. We used electrophoretic mobility shift assays (Extended Data Fig. 8j,k) and fluorescence polarization (Extended Data Fig. 8l) to quantify the binding affinity of recombinant SOA(1–112) (which contains the myb domain), SOA(1–331) (which contains the myb and BTB domains) and SOA(1195–1265) (which contains the ZnF domain) to CA-containing and non-CA-containing DNA sequences. The myb DNA-binding domain, but not the ZnF domain, associated with DNA in vitro (Extended Data Fig. 8j,l). In line with the fact that oligomerization provided by BTB domains can confer stable chromatin association[20], the DNA-binding property of the myb domain was enhanced in the presence of BTB (for CA_{10} dsDNA, K_d = 59 μM for SOA(1–112) compared with K_d = 40 nM for SOA(1–331)). Size-exclusion chromatography coupled to multi-angle light scattering confirmed the oligomerization function of the BTB domain, as SOA(1–122) and SOA(1–229) appeared as monomers, but SOA(1–331) was present in monomeric and multiple oligomeric species (Extended Data Fig. 8m). Nonetheless, in this in vitro setup with isolated domains, none of the fragments showed specificity towards CA-containing compared with non-CA containing sequences. To explore this effect in vivo, we expressed a SOA mutant without the myb domain in Ag55 cells and performed CUT&Tag (Extended Data Fig. 8n–p). In comparison to full-length SOA, SOA without the myb domain showed a substantial reduction in X chromosome association that was close to background levels.

Compromised DC in *SOA* mutant males

To understand its physiological roles, we generated transgenic mosquitoes that lack *SOA* by virtue of a CRISPR-mediated targeted knock-in in front of the *SOA* coding sequence (Extended Data Fig. 9a and Methods). The transgenic line, referred to as *SOA-KI*, was made homozygous and then verified by PCR and RT–qPCR (Extended Data Fig. 9a,b). The RT-qPCR assay showed substantially decreased *SOA* RNA levels in these mosquitoes. In CUT&Tag, the enrichment at male-specific SOA-binding sites was lost in *SOA-KI* compared with the wild-type mosquitoes (Fig. 4a and Extended Data Fig. 6d,e,g,h). IF showed that localization of SOA to the X chromosome territory was lost in *SOA-KI* males (Fig. 4b). RNA-seq analyses of gene expression changes (Extended Data Fig. 9c,d) revealed global downregulation of the X chromosome in SOA mutant males (Fig. 4c and Extended Data Fig. 9e). This result confirms that SOA mediates DC in vivo. Out of the 204 downregulated genes scored as differentially expressed (Supplementary Table 2), 164 were X-linked ($P = 6.73 \times 10^{-54}$, Fisher's exact test). We also analysed the expression changes in the three groups of genes that exhibited strong, intermediate and weak SOA association in CUT&Tag (clusters in Fig. 2g). The reduced gene expression in *SOA-KI* males correlated with the strength of SOA binding in wild-type males (Fig. 4d). Genes from cluster 1 with strong SOA binding were notable (median fold change of 0.608) providing support for a role for SOA in DC.

To investigate whether this effect is associated with changes in chromatin accessibility, we performed assay for transposase-accessible chromatin with sequencing (ATAC–seq) in wild-type and *SOA-KI* mosquitoes (Extended Data Fig. 9f,g). The accessibility of X-linked promoter regions remained unchanged, regardless of RNA expression changes in *SOA-KI* mosquitoes (Extended Data Fig. 9h) or direct SOA binding (Extended Data Fig. 9i). Furthermore, the male and female X chromosome displayed comparable accessibility (Extended Data Fig. 9j), which suggested that SOA binding at the TSS does not change the level of promoter opening per se, but presumably affects features after pre-initiation complex loading[21].

We next examined the phenotypic consequences of SOA loss. Homozygous *SOA-KI* mosquitoes of both sexes were viable and fertile. However, in a mixed mosquito culture of *SOA-KI* and wild-type genotypes, the mutant allele frequency diminished over time, which indicated a fitness defect (Fig. 4e; heterozygous *SOA-KI* males showed no phenotype). Of note, unlike the wild-type mosquitoes, adult male SOA mutants tended to emerge after females, which indicated a sex-specific developmental delay. Accordingly, a gene ontology (GO) term analysis of the differentially expressed genes based on RNA-seq revealed an enrichment of mitochondrial function and organization, oxidative phosphorylation and metabolic processes (Extended Data Fig. 9k and Supplementary Table 2). To quantify the developmental delay, we sorted neonate wild-type and *SOA-KI* larvae of both sexes (n = 100 for each of the 4 genotypes) and monitored their development in the same mixed culture. We precisely scored the timing of the appearance of pupae for all four genotypes indicating the time required to complete the larval stages (scheme in Fig. 4f). Male *SOA-KI* pupae emerged on average 4 h later than the wild-type males, whereas there was no effect on the development of the females (Fig. 4f, right, and Extended Data Fig. 9l).

Impact of ectopic *SOA* in female mosquitoes

We next wanted to explore the physiological consequences of expressing the male SOA isoform in female mosquitoes. In this transgenic line, referred to as *SOA-R* (for rescue), the spliced SOA(1–1265) cDNA (male isoform) was integrated immediately upstream of the *SOA-KI* cassette. The rationale behind this strategy was to express SOA in both sexes from its endogenous promoter while rescuing the loss-of-function condition in males (Fig. 5a). The transgenic *SOA-R* line was made homozygous and showed the same *SOA* mRNA expression levels in both sexes, which was slightly higher than the endogenous *SOA* mRNA levels in males (Fig. 5b and Extended Data Fig. 10a). In IF stainings of *SOA-R*, both sexes exhibited a subnuclear SOA territory, which overlapped with the transcription site of the X-linked *AGAP000651* (Fig. 5c and Extended Data Fig. 10b,c). SOA CUT&Tag corroborated that ectopic X chromosome binding was induced in female *SOA-R* pupae (Fig. 5d and Extended Data Fig. 10d,e). The majority of peaks were localized to the X chromosome (Fig. 5e), overlapped with the ones found in wild-type males (Extended Data Fig. 10f) and were more enriched at highly expressed genes (Extended Data Fig. 10g,h).

We performed RNA-seq (Extended Data Fig. 10i) and found that *SOA-R* females displayed a significant overrepresentation of X-linked genes among the upregulated population (upregulated, 300 on the X chromosome, 531 on autosomes, $P = 6.49 \times 10^{-43}$; downregulated, 51 on the X chromosome, 1,003 on autosomes, P = 0.9998, Fisher's exact test; Fig. 5f). The increase in RNA levels was most notable at genes with strong binding in CUT&Tag (cluster 1, median fold change of 1.53; Extended Data Fig. 10j), but significant upregulation was also observed when all expressed X-linked genes were taken into account (Extended Data Fig. 10k,l). We analysed the *SOA-R* transgenic line for developmental delay by scoring the timing of pupation. Compared with the parental *SOA-KI* line, the *SOA-R* males developed equally fast as the wild-type line. This rescue of the loss-of-function phenotype confirms the functionality of the *SOA-R* cDNA and that the *SOA-KI* phenotype was not caused by off-target mutations. By contrast, the *SOA-R* females showed a significant developmental delay of a few hours in comparison to all other genotypes (wild-type controls and *SOA-R* males) (Fig. 5g and Extended Data Fig. 10m).

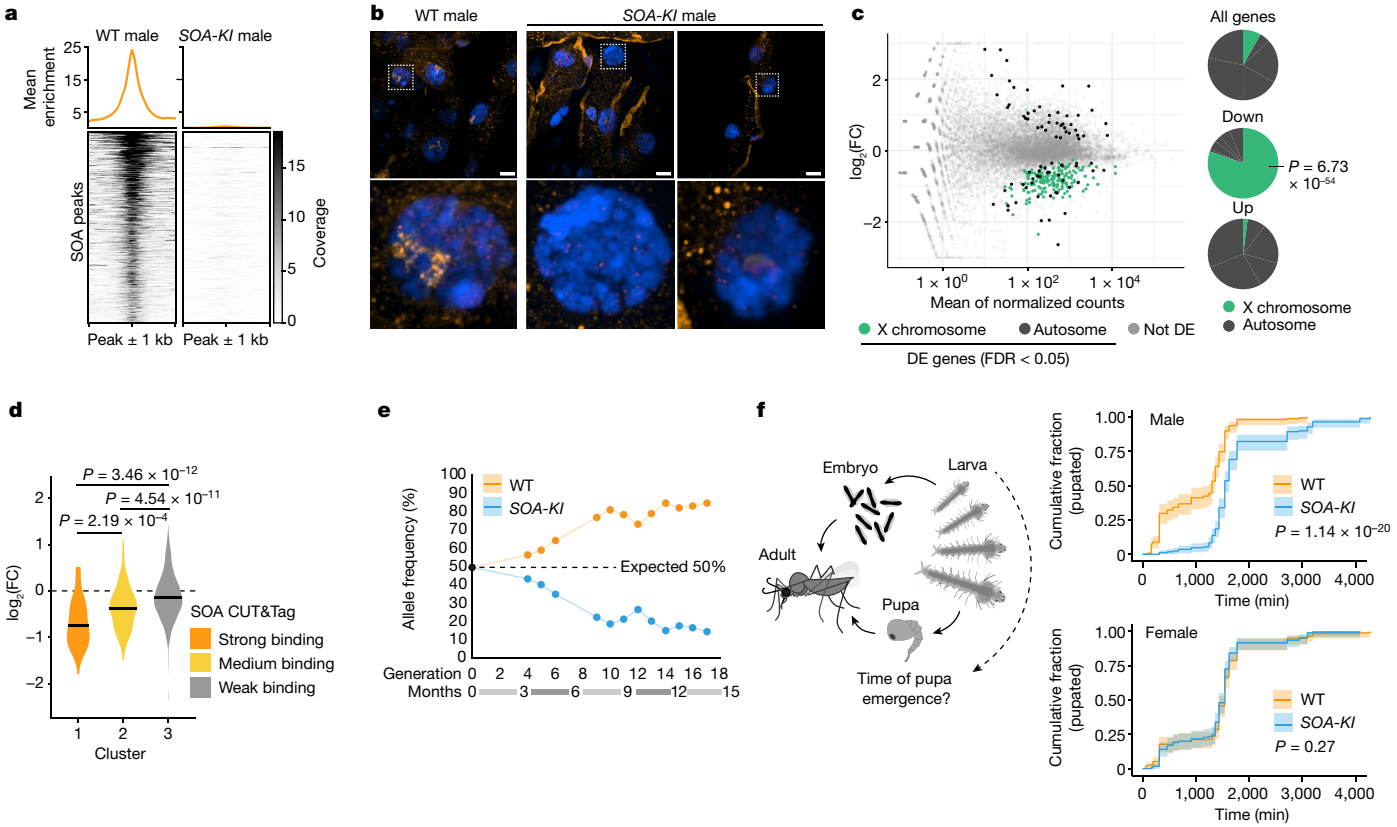


Fig. 4 | Loss of SOA-mediated DC leads to a male-specific developmental delay. a, Heatmap showing normalized CUT&Tag coverage in male wild-type (WT) and homozygous *SOA-KI* pupae ($n = 4$ and $n = 2$ biological replicates, respectively; merged for visualization) at significant peaks with binding in males > females. Metaplot (top) show mean enrichment. Datasets for Figs. 2 and 4 were generated together. **b**, Representative SOA immunostaining (orange) and DAPI (blue) conducted on WT and homozygous *SOA-KI* male adult mosquito Malpighian tubules. Images on the bottom row are close-ups of the white square in the top row. Images represent 3D views of a *z*-stack. Scale bar, 10 µm. **c**, Left, MA plots from RNA-seq showing normalized read counts versus $\log_2$(FC) comparing WT with homozygous *SOA-KI* male pupae ($n = 4$ biological replicates). DE genes are green (X chromosome) or black (autosomes), others are grey. Right, pie charts of DE and all *A. gambiae* genes. *P* value: one-sided Fisher's test for overrepresentation of X-linked genes. **d**, As in **c**. Violin plot of $\log_2$(FC) values obtained by DESeq2 analysis of RNA-seq in *SOA-KI* versus WT male pupae. Centre line indicates the median. X-linked genes with average read count > 0 were plotted and split into 3 groups according to the SOA-binding strength (Fig. 2g,h). Bonferroni-corrected *P* values: two-sided Wilcoxon rank-sum test; underlying data provided in Supplementary Table 3. **e**, Line plot illustrating allele frequencies observed in a mixed rearing of WT and *SOA-KI* transgenic mosquitoes ($n = 1$ population). Dashed line shows expected 50:50 allele frequencies. Raw values in Supplementary Table 1. **f**, Left, schematic of *Anopheles* development. Right, line plot (average of $n = 4$ replicate cultures with 95% confidence intervals) of developmental timing of WT and homozygous *SOA-KI* quantified as a cumulative distribution of pupa emergence over time. Each replicate culture reflects 100 neonate larvae of each genotype seeded for development through the larval stages (L1–L4). *P* value: log-rank test for stratified data (Mantel–Haenszel test), second independent experiment in Extended Data Fig. 10a.

In view of these results, we wanted to investigate how a developmental difference of only a few hours can explain the spread and fixation of the *SOA* allele in ancestral *Anopheles*. We considered the standard one-locus model for differential selection in the two sexes[22]. The fitness of males and females in a primordial *SOA*-less state was standardized to one. According to *Anopheles*-specific models, a 4-h acceleration in male development corresponds to a selection coefficient of $s_m = 0.0177$ in males (Methods), yielding a relative fitness of $1 + s_m = 1.0177$ of *SOA*-bearing males (assuming that SOA^+ is dominant over SOA^- in males). *SOA* would spread relatively rapidly and eventually reach fixation if it had no negative fitness effects in females (Fig. 5h, first panel). However, the results of the *SOA-R* transgenic line imply that before the 'invention' of alternative splicing, *SOA* was detrimental in females, as its presence may have led to dosage imbalance by overexpression of the entire X chromosome (Fig. 5f). This result is in line with the strict conservation of sex-specific splicing among Anophelinae, thereby preventing the expression of a full-length SOA protein in females (Extended Data Fig. 4a). We therefore assumed that the relative fitness of *SOA*-bearing females is $1 - s_f$ in homozygous females and $1 - h_f s_f$ in heterozygous females. The model predicts that the *SOA* allele will still spread until stable coexistence with the SOA^- allele is obtained, unless the selection coefficient s_f in females is much higher than the selection coefficient s_m in males (Fig. 5h and Extended Data Fig. 10n). When both alleles are present in the population, any factor alleviating the negative effect of *SOA* in females (such as alternative splicing, marked with an asterisk in Fig. 5h) will lead to the rapid fixation of *SOA* in the population, irrespective of how large the fitness benefit is in males.

Discussion

The expression of SOA in females is controlled through sex-specific alternative splicing, which parallels the regulatory mechanism of *msl-2* in *Drosophila*[23]. The female sex determination factor SXL binds to an alternatively spliced intron to prevent *msl-2* RNA export and translation. In contrast to MSL2, truncated *Anopheles* SOA protein was detectable in females by mass spectrometry, but it did not accumulate on the X chromosome and is nonfunctional for DC. A female protein present already during early embryogenesis could prevent intron 2 excision. One potential candidate is the sex determination factor *Femaleless* (*Fle*),

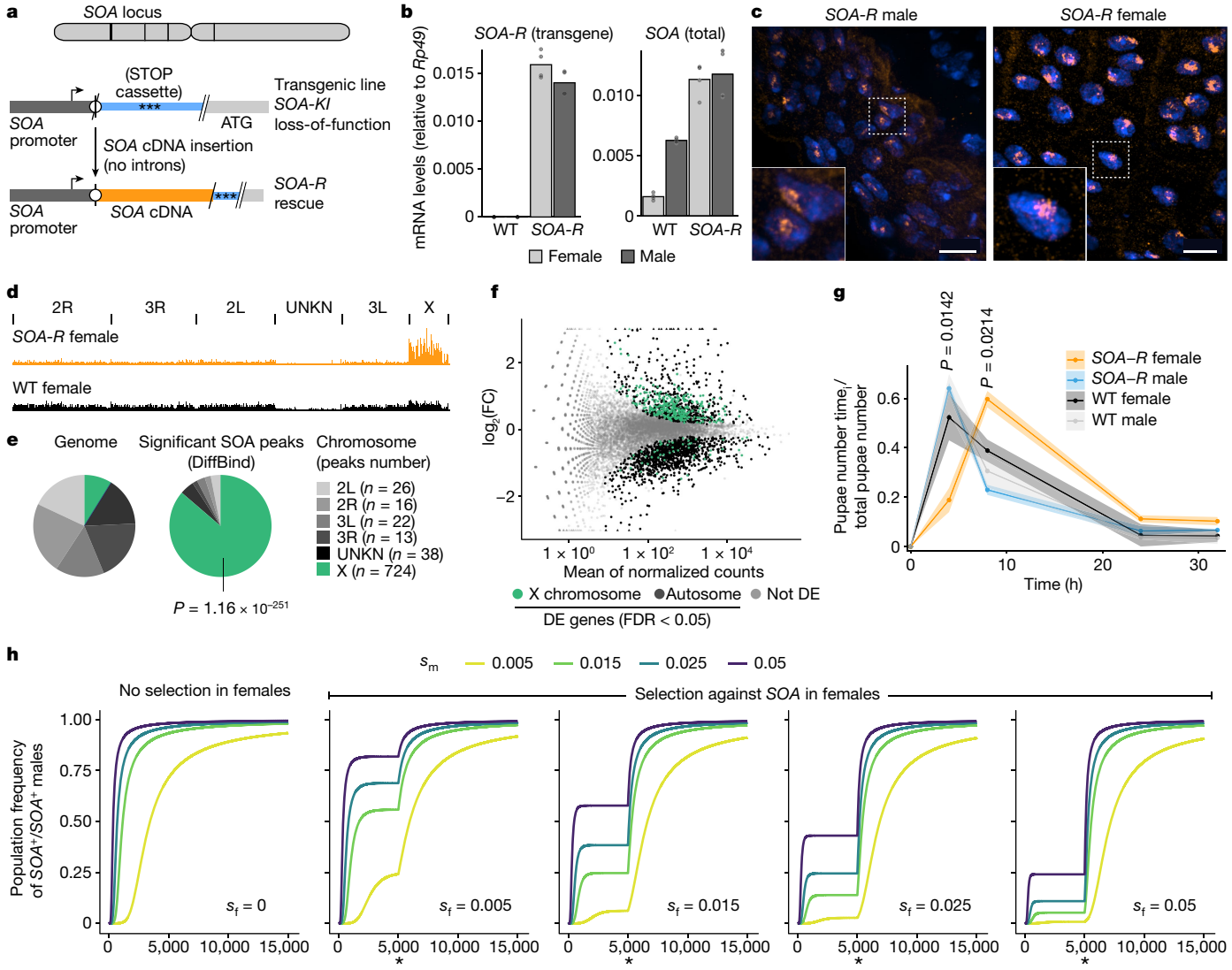


Fig. 5 | Spliced *SOA* isoform expression in female mosquitoes results in ectopic DC. a, Scheme outlining the strategy to create *SOA-R* transgenic mosquitoes. The *att*P landing site (circle) in the *SOA-KI* cassette was used to insert the *SOA* coding sequence. **b**, Bar plots (height: mean of $n = 4$ biological replicates) showing *SOA* mRNA levels normalized to *Rp49* in WT and homozygous *SOA-R* pupae measured by RT–qPCR. Left, expressed from the *SOA-R* cassette (SV40 terminator in the 3′ UTR). Right, total *SOA* mRNA. **c**, Representative SOA immunostainings (orange) and DAPI (blue) conducted on homozygous *SOA-R* male and female adult guts. Images on the bottom left are close-ups of the white square in the main images. Images represent 3D views of a *z*-stack. Scale bar, 10 µm (also see Extended Data Fig. 10c). **d**, Genome browser snapshot of SOA CUT&Tag coverage in homozygous *SOA-R* and WT female pupae ($n = 2$ biological replicates, merged for visualization). **e**, As in **d**. Pie charts of the significant CUT&Tag peaks versus the *A. gambiae* genome. *P* value: one-sided Fisher's test for overrepresentation of X-linked genes. **f**, MA plot from RNA-seq showing normalized read counts versus $\log_2$(FC) comparing homozygous *SOA-R* ($n = 4$ biological replicates) with WT female pupae ($n = 3$). DE genes are green (X chromosome) or black (autosomes), others are in grey. **g**, Line plot (average of $n = 3$ replicate cultures with shaded areas indicating the s.e.m.) of developmental progression of *SOA-R* quantified by pupa emergence over time. Benjamini–Hochberg-corrected *P* values: two-sided *t*-test with pairwise comparisons between the genotypes. Only significant *P* values (*SOA-R* versus WT females) shown. All data in Supplementary Table 1. **h**, Model predictions of the evolution of *SOA*. s_m, fitness increase of SOA^+ versus SOA^-/SOA^- males. s_f, fitness decrease of SOA^+/SOA^+ versus SOA^- females. Asterisk indicates evolution of alternative splicing at 5,000 generations.

which contains RNA-binding domains and the knockdown of which in females is associated with misregulation of X-linked transcripts[24]. FLE controls the sex-specific splicing of, for example, *fruitless* or *doublesex*[24], which are well conserved among insects[25]. Thus, SOA may have hijacked pre-existing sequences from such genes after duplication from its non-sex-specific paralogue.

By directly associating with the X chromosome, SOA joins a small list of master regulators that are sufficient to induce chromosome-wide expression alterations (MSL2 in *D. melanogaster*[12], SDC-2 in *Caenorhabditis elegans*[26] and Xist in mammals[18]). Unlike the *Drosophila* MSL complex, which initially targets high-affinity sites and then spreads to X-linked genes, SOA directly binds the promoters of active genes. Specificity may involve cooperative binding at CA dinucleotide repeats in a similar fashion as for *Drosophila* GAGA factor (GAF). GAF contains a BTB domain important for selecting proper GAF target sites, despite the relatively high abundance of individual GAGA motifs across the genome[20]. The SOA myb-BTB fragment alone is not sufficient for distinguishing CA sequences. We propose that co-factor recruitment through the carboxy-terminal part of SOA probably contributes to faithful target site recognition. After SOA recruitment to X-linked promoters, transcription itself (for example, pause release or elongation[21]) or co-transcriptional RNA processing events[27] may be altered to achieve DC.

In *Anopheles*, the loss of DC in males or its ectopic induction in females was associated with developmental delay. This effect differs from mutants in the sex determination pathway, which show sex reversal, sterility or lethality of variable penetrance[3,24,28]. The expression of *Guy1*, the Y-linked maleness gene in *Anopheles stephensi*, confers complete female-specific lethality accompanied by an upregulation of X-linked genes[29]. The molecular functions of Guy1 and Yob are not known yet, but our data showed that SOA directly binds to the X chromosome and that interfering with its function is not lethal. We favour a model in which Guy1 and Yob induce SOA, but also other yet to be identified factors, the latter of which or their combination with X-misregulation, is causal to lethality after their ectopic expression in females.

It is unclear why DC is essential in organisms such as *Drosophila*, but non-essential in *Anopheles*, whereas many species with heteromorphic sex chromosomes (for example, birds) do not exhibit chromosome-wide DC at all[1,10]. Despite an imbalance in X chromosomal expression already at early embryogenesis[30], *msl* mutants of *Drosophila* are viable for about 6 days and only die when they reach late larval/early pupal stages[31]. In *roX1*/*roX2* mutants, there are even rare survivors that reach adulthood[32]. Indeed, the molecular activities of the DC complexes have been studied in detail in model organisms, but the physiological consequences of their absence and the causation of lethality remain enigmatic. Hypotheses range from misregulation of a few, putative haplo-lethal genes encoded on the X chromosome to a global gene-dosage imbalance that causes perturbation of gene regulatory networks, overload of cellular machineries such as the ribosome and chaperones, leading to proteotoxicity[33]. This dosage-imbalance model attributes lethality to the degree of disequilibrium rather than the identity of X-linked genes. The difference in phenotypic outcome would accordingly be supported by the 2,500 protein-coding genes in *Drosophila* compared with 1,063 in *Anopheles* on the X chromosome, despite similar overall gene numbers[10]. In addition, autosomal retrocopies of X-linked genes could mitigate phenotypic consequences in *Anopheles* by allowing dosage-sensitive genes to evade the X chromosome and thus eliminating the need for DC[34]. Apparently, there is a continuum in phenotypic outcome, whereby non-essentiality may permit the evolution of a DC master regulator despite being beneficial for one sex but reducing the fitness of the other one. Our model predicts that under these circumstances, genes such as *SOA* can be polymorphic, which underscores the importance of a sufficient sampling rate, as DC alleles might be rare in a population. Alternative splicing would then be strongly selected, as it may alleviate or even resolve the conflict, whereupon DC can spread to fixation.

Last, we note that exploiting X chromosome misregulation has been proposed to artificially generate single-sex populations or sex ratio distortion gene drives for vector control programmes[29,35]. Our discovery that induction of the SOA–DC pathway—at least under the conditions studied by us—is not strongly detrimental for females warrants further studies to uncover factors and mechanisms that underlie sex-specific lethality to eventually harness them in malaria vector control programmes.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-06641-0.

1. Furman, B. L. S. et al. Sex chromosome evolution: so many exceptions to the rules. *Genome Biol. Evol.* **12**, 750–763 (2020).
2. The Lancet. Malaria in 2022: a year of opportunity. *Lancet* **399**, 1573 (2022).
3. Krzywinska, E., Dennison, N. J., Lycett, G. J. & Krzywinski, J. A maleness gene in the malaria mosquito *Anopheles gambiae*. *Science* **353**, 67–69 (2016).
4. Papa, F. et al. Rapid evolution of female-biased genes among four species of *Anopheles* malaria mosquitoes. *Genome Res.* **27**, 1536–1548 (2017).
5. Deitz, K. C., Takken, W. & Slotman, M. A. The effect of hybridization on dosage compensation in member species of the *Anopheles gambiae* species complex. *Genome Biol. Evol.* **10**, 1663–1672 (2018).
6. Rose, G. et al. Dosage compensation in the African malaria mosquito *Anopheles gambiae*. *Genome Biol. Evol.* **8**, 411–425 (2016).
7. Jiang, X., Biedler, J. K., Qi, Y., Hall, A. B. & Tu, Z. Complete dosage compensation in *Anopheles stephensi* and the evolution of sex-biased genes in mosquitoes. *Genome Biol. Evol.* **7**, 1914–1924 (2015).
8. Zdobnov, E. M. et al. Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* **298**, 149–159 (2002).
9. Vicoso, B. & Bachtrog, D. Numerous transitions of sex chromosomes in *Diptera*. *PLoS Biol.* **13**, e1002078 (2015).
10. Basilicata, M. F. & Keller Valsecchi, C. I. The good, the bad, and the ugly: evolutionary and pathological aspects of gene dosage alterations. *PLoS Genet.* **17**, e1009906 (2021).
11. Lucchesi, J. C. & Kuroda, M. I. Dosage compensation in *Drosophila*. *Cold Spring Harb. Perspect. Biol.* **7**, a019398 (2015).
12. Kelley, R. L. et al. Expression of *msl-2* causes assembly of dosage compensation regulators on the X chromosomes and female lethality in *Drosophila*. *Cell* **81**, 867–877 (1995).
13. Keller Valsecchi, C. I., Marois, E., Basilicata, M. F., Georgiev, P. & Akhtar, A. Distinct mechanisms mediate X chromosome dosage compensation in *Anopheles* and *Drosophila*. *Life Sci. Alliance* **4**, e202000996 (2021).
14. Karousis, E. D. & Mühlemann, O. The broader sense of nonsense. *Trends Biochem. Sci* **47**, 921–935 (2022).
15. Lyon, M. F. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* **190**, 372–373 (1961).
16. Valsecchi, C. I. K. et al. RNA nucleation by MSL2 induces selective X chromosome compartmentalization. *Nature* **589**, 137–142 (2021).
17. Kaya-Okur, H. & Henikoff, S. Bench top CUT&Tag. *protocols.io* https://www.protocols.io/view/bench-top-cut-amp-tag-kqdg34qdpl25/v3 (2020).
18. Brockdorff, N. et al. The product of the mouse *Xist* gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* **71**, 515–526 (1992).
19. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
20. Tang, X. et al. Kinetic principles underlying pioneer function of GAGA transcription factor in live cells. *Nat. Struct. Mol. Biol.* **29**, 665–676 (2022).
21. Ferrari, F. et al. 'Jump start and gain' model for dosage compensation in *Drosophila* based on direct sequencing of nascent transcripts. *Cell Rep.* **5**, 629–636 (2013).
22. Hartl, D. L. & Clark, A. G. *Principles of Population Genetics* 4th edn. (Sinauer and Associates, 2006).
23. Beckmann, K., Grskovic, M., Gebauer, F. & Hentze, M. W. A dual inhibitory mechanism restricts *msl-2* mRNA translation for dosage compensation in *Drosophila*. *Cell* **122**, 529–540 (2005).
24. Krzywinska, E. et al. *femaleless* controls sex determination and dosage compensation pathways in females of *Anopheles* mosquitoes. *Curr. Biol.* **31**, 1084–1091.e4 (2021).
25. Price, D. C., Egizi, A. & Fonseca, D. M. The ubiquity and ancestry of insect doublesex. *Sci. Rep.* **5**, 13068 (2015).
26. Dawes, H. E. et al. Dosage compensation proteins targeted to X chromosomes by a determinant of hermaphrodite fate. *Science* **284**, 1800–1804 (1999).
27. Rücklé, C. et al. RNA stability controlled by m^6A methylation contributes to X-to-autosome dosage compensation in mammals. *Nat. Struct. Mol. Biol.* **30**, 1207–1215 (2023).
28. Kyrou, K. et al. A CRISPR–Cas9 gene drive targeting *doublesex* causes complete population suppression in caged *Anopheles gambiae* mosquitoes. *Nat. Biotechnol.* **36**, 1062 (2018).
29. Qi, Y. et al. *Guy1*, a Y-linked embryonic signal, regulates dosage compensation in *Anopheles stephensi* by increasing X gene expression. *eLife* **8**, e43570 (2019).
30. Samata, M. et al. Intergenerationally maintained histone H4 lysine 16 acetylation is instructive for future gene activation. *Cell* **182**, 127–144.e23 (2020).
31. Belote, J. M. & Lucchesi, J. C. Male-specific lethal mutations of *Drosophila melanogaster*. *Genetics* **96**, 165–186 (1980).
32. Kim, M., Faucillion, M.-L. & Larsson, J. *RNA-on-X 1* and *2* in *Drosophila melanogaster* fulfill separate functions in dosage compensation. *PLoS Genet.* **14**, e1007842 (2018).
33. Lee, H. et al. Effects of gene dose, chromatin, and network topology on expression in *Drosophila melanogaster*. *PLoS Genet.* **12**, e1006295 (2016).
34. Miller, D. et al. Retrogene duplication and expression patterns shaped by the evolution of sex chromosomes in malaria mosquitoes. *Genes* **13**, 968 (2022).
35. Krzywinska, E. & Krzywinski, J. Effects of stable ectopic expression of the primary sex determination gene *Yob* in the mosquito *Anopheles gambiae*. *Parasit. Vectors* **11**, 648 (2018).

Methods

Mosquito rearing and *SOA* mutagenesis

A. gambiae mosquitoes were maintained in standard insectary conditions (26–28 °C, 75–80% humidity and 12–12-h light–dark cycle). To obtain the *SOA* mutant, we used the CRISPR–Cas9 system to insert a fluorescent marker cassette (3×P3-mTurquoise2) into the first *SOA* exon. In addition, an *att*P docking site for PhiC31-mediated plasmid integration was included at the start of the fluorescence marker cassette and at a position corresponding to the *SOA* initiator ATG codon to later allow the possibility of rescuing the mutation with a new copy of *SOA* (see below). The knocked-in fluorescent marker cassette was designed with a strong transcription terminator and multiple stop codons to halt the expression of *SOA* at both the transcriptional and translational level. For this, we built a gRNA-expressing and repair template donor plasmid in the pDSARN vector[36] as previously described[37]. This plasmid expressed two gRNAs under the control of the *AGAP013557* U6 promoter, recognizing target sites 5′-GTCAGCAGCCAGCTTGATGC-3′ and 5′-GCATCAAGCTGGCTGCTGAC-3′ in *SOA*. The 5′ and 3′ regions of homology from the *SOA* genomic sequence (each around 1.1-kb long) adjacent to the gRNA target sites were cloned in this plasmid, flanking the 3×P3-mTurquoise marker cassette. The sequence of the resulting genomic insertion is provided in Supplementary Table 1. The plasmid was microinjected into approximately 40–90 min-old embryos of an *A. gambiae* strain expressing *Cas9* in the germline from a YFP-marked transgene[37]. The progeny of surviving injected mosquitoes, backcrossed to WT, was screened for blue fluorescent larvae using a Nikon SMZ-18 binocular microscope equipped with a Lumencor Sola Light engine and CFP excitation and emission filters. Several dozens of mTurquoise-positive larvae were recovered, and the *SOA-KI* line was established from a single founder female. Junctions between the knocked-in synthetic sequence and the genome were amplified by PCR and sequence-verified. Homozygous and heterozygous *SOA-KI* lines were derived by COPAS sorting[38]. To track the natural dynamics of genotype frequencies across generations, the heterozygous (*WT/SOA-KI*) line was left to evolve naturally for >16 generations. At each generation, the entire population of newly hatched neonate L1 larvae was subjected to COPAS analysis to record the numbers of homozygous mutant, heterozygous and WT individuals as scored by the presence and intensity of mTurquoise marker present in the *SOA-KI* allele (WT is not fluorescent). Genetic crosses were used to combine the *SOA-KI* mutation with the T4 sexing transgene expressing GFP from the Y chromosome[39], allowing COPAS sorting of all-male or all-female populations of *SOA-KI* homozygous mutant and control mosquito larva populations for use in biochemistry experiments. To create the *SOA-R* transgenic mosquito line in which the *SOA* mutation is rescued with a *SOA* cDNA sequence encoding the male *SOA* isoform, we constructed a plasmid harbouring a PhiC31 *att*B site immediately preceding the full-length *SOA* coding sequence, itself followed by the SV40 3′ terminator sequence. A 3×P3-DsRed fluorescence marker was included in the plasmid as a transgenesis selection marker downstream of this *SOA* rescue cassette (the sequence of the rescue plasmid is provided in Supplementary Table 1). This plasmid was co-injected with a PhiC31 integrase-encoding helper plasmid[36] at a concentration of 320 and 80 ng μl^{-1}, respectively, in embryos of the *SOA-KI* line. Integration of the entire plasmid into the *SOA-KI att*P site placed the *SOA* male cDNA isoform under control of the endogenous *SOA* promoter. Transgenic mosquitoes were selected based on DsRed expression in addition to CFP, resulting in the *SOA-R* transgenic line. Work with genetically modified mosquitoes was evaluated by Haut Conseil des Biotechnologies and authorized by MESRI (déclaration d'utilisation d'OGM en milieu confiné no. 3243 and agreement no. 3912).

Developmental timing was scored by counting the appearance of pupae over time, starting from the moment when the first pupa appeared in the culture. At each sampling time, the newly formed pupae were removed from the culture.

Mice

Mice (CD-1 strain) were maintained in social groups of 4–5 individuals in Techniplast 2L type cages (365 × 207 × 140 mm) with Safe Select litter and nest-building wood, paper and cotton materials, 12–12-h dark–light cycle, 22 °C temperature and 50 ± 10% humidity and fed with Safe R04-25 pellets. For mosquito blood feeding, female CD-1 mice (>35 g) were anaesthetized with a mixture of Zoletil (42.5 mg kg^{-1}) and Rompun (8.5 mg kg^{-1}) in 0.9% NaCl solution, according to animal care procedures validated by regional CREMEAS ethics committee and by the French ministry of higher education, research and innovation under the agreement APAFIS no. 20562–2019050313288887 v.3. We complied with all relevant ethical regulations regarding the use of animals.

Genotyping

Pupae were homogenized in TRIzol (Fisher Scientific, 15-596-026). After adding chloroform and removing the aqueous phase, the phenol–chloroform phase was used for DNA isolation following the manufacturers' instruction manual. PCR was performed with LA Taq HS polymerase (Takara, RR042A). The PCR products were run on a 1% Tris-borate-EDTA (TBE) agarose gel and imaged using ChemiDoc MP v.3 (Bio-Rad).

RNA isolation, library generation and sequencing

RNA was extracted using TRIzol (Fisher Scientific, 15-596-026) and a Direct-zol RNA MicroPrep Kit (Zymo Research, R2062). For pupa samples, only the aqueous phase formed after phenol–chloroform separation was loaded on the column after mixing with 100% ethanol. NGS library preparation was performed using an Illumina Stranded mRNA Prep Ligation kit according to the Stranded mRNA Prep Ligation Reference Guide (June 2020; document no. 1000000124518 v00). For the Ag55 cell culture RNA-seq, libraries were prepared with a starting amount of 100 ng and 2 µl of ERCC spike-ins (Ambion, 4456740) in a 1:1,000 dilution and amplified in 12 PCR cycles. For the pupa RNA-seq, libraries were prepared with a starting amount of 1,000 ng and 2 µl of ERCC spike-ins (Ambion, 4456740) in a 1:100 dilution and amplified in 10 PCR cycles. Libraries were profiled in a High Sensitivity DNA on a 2100 Bioanalyzer (Agilent technologies), and quantified using a Qubit dsDNA HS Assay kit in a Qubit 2.0 Fluorometer (Life Technologies). Pooled samples were sequenced on a NextSeq 500 High Output, PE for 2× 73 cycles plus 2× 10 cycles for the dual index read.

RNA-seq data processing and visualization

For *SOA-KI* RNA-seq, the reads were mapped to the ribosomal RNA sequences extracted from the Ensembl AgamP4 genome using the Ensembl AgamP4 annotation (release 48) with STAR (v.2.7.3a) with the following parameters: outFilterMultimapNmax 1000000 outFilterMismatchNoverLmax 0.04 outFilterMismatchNmax 999. Reads mapping to rRNA were discarded, and unmapped reads were used in downstream processing. For the *SOA-R* and Ag55 RNA-seq, trimming and mapping against rRNA were not performed as there were few rRNA reads. In all experiments, the reads were mapped to the Ensembl AgamP4 genome using the Ensembl AgamP4 annotation (release 48) together with lncRNA annotation[40] and experiment-specific sequences (such as elements of the *SOA-KI* or *SOA-R* cassette, or sequences from the baculovirus in the Ag55 experiment to assess infection rates; more information is provided together with the uploaded data in the Genome Expression Omnibus database) with STAR (v.2.7.3a) using the following parameters: outFilterMismatchNoverLmax 0.04 outFilterMismatchNmax 999. Only uniquely mapped reads were used for downstream analysis. Coverage signal tracks (bigWigs) of primary alignments were generated using deepTools (v.3.1.0). Primary alignments were assigned to features using subread (v.1.6.5) with the AgamP4 annotation (release 48) combined with lncRNA annotation[40] as a reference. Differential

expression analysis was performed using DESeq2 (v.1.26.0), and only genes with FDR < 0.05 were considered as differentially expressed. The visualization of the RNA-seq data of *SOA* in *Anopheles gambiae*, *A. arabiensis*, *A. minimus* and *A. albimanus* was obtained using the genome browser tool from VectorBase (https://vectorbase.org).

CUT&Tag library generation and sequencing

CUT&Tag was performed as previously described[17]. In total, 0.4 million cells were used for each reaction. The pupa experiments were performed with flash-frozen tissue samples, which were homogenized in cold PBS and passed through a cell strainer (Corning, 352235). In the initial pupa experiment (WT and *SOA-KI* male and female pupae), the homogenate was fixed with 0.2% paraformaldehyde (PFA) for 2 min at room temperature. For the *SOA-R* CUT&Tag, no fixation was applied. The cell culture experiments were all performed on freshly collected cells with a native protocol. The antibodies used are listed in the Supplementary Table 4. We used pA–Tn5 prepared by the IMB Protein Production Core Facility and 15 PCR cycles in the library amplification step. Pooled samples were sequenced on NextSeq 500 High Output, PE for 2×75 cycles plus 2×8 cycles for the dual index read.

CUT&Tag data processing and analysis

Reads were trimmed using cutadapt (v.4.0) to remove Illumina adapter sequences and subsequently mapped to the reference genome with bowtie2 (v.2.4.5). For the WT male versus female pupa experiment, we performed an initial analysis to inspect the antibody specificity and therefore removed the multimapping and duplicate reads. We then called peaks using macs2 (v.2.1.2) with the corresponding IgG samples as controls, which identified 139 and 393 filtered peaks in female replicates 1 and 2, respectively, but 1,025, 653, 627 and 808 filtered peaks in males. Because we could not a priori exclude SOA binding to repetitive regions, we then performed a second analysis, in which multimapping and duplicate reads were retained for peak calling using macs2 (v.2.1.2). Note that CUT&Tag fragments can share exact starting and ending positions because the integration sites are affected by DNA accessibility. Therefore, duplicates observed in CUT&Tag are not necessarily a consequence of overamplification by PCR[41,42]. A greylist was generated on the basis of IgG samples using the R package GreyListChIP (v.1.22.0) and applied for peak filtering in the pupa experiments. This provided 7,742 consensus peaks for downstream analysis with DiffBind (v.3.4) to identify sites that were significantly (FDR < 0.05) differentially bound between samples (results in Supplementary Table 2). Note that the greylist was applied for the pupa datasets and the myb-less experiment in Ag55, whereas no greylist was applied to the long SOA versus empty Ag55 (cell culture) dataset, as this experiment contained almost no background. Background bins instead of library size were used for normalization. Downstream visualization of differentially bound peaks (for example, heatmaps) were generated using deepTools (v.3.5.1). To identify SOA-bound motifs, the sequences of peaks (±200 bp from the summit) with higher binding (FDR < 0.05) in males (pupa) or SOA(1–1265) were extracted using bedtools (v.2.29.2). Peak sequences were then used for motif discovery analysis using MEME-ChIP (MEME v.5.4.1), with the genome sequence as a background. The MEME output was then used in FIMO (v.5.4.1) with default settings and selecting the available metazoan upstream sequences for *A. gambiae* (AgamP4.34_2019-03-11) or *A. aegypti* (AaegL3.34_2019-03-11) databases. Overlapping CA motifs identified by FIMO were merged into a single CA motif using ggRanges. For the analysis of repeats, the RepeatMasker annotation was downloaded from https://www.repeatmasker.org/species/anoGam.html, RepeatMasker open-4.0.5-Repeat Library 20140131. Downstream analysis and statistical tests were performed using R studio.

ATAC–seq library generation and sequencing

ATAC–seq was performed as previously described[43] with the following changes. The starting material was flash-frozen pupae. After thawing, whole pupae were homogenized in cold PBS and passed through a cell strainer (Corning, 352235). The cell suspension was counted, and 50,000 cells were used for each reaction. We used 250 ng of Tn5 prepared by the IMB Protein Production Core Facility per reaction and 15 PCR cycles in the library amplification step. Pooled samples were sequenced on NextSeq 500 High Output, PE for 2×75 cycles plus 2×8 cycles for the dual index read.

ATAC–seq data processing and analysis

Reads were trimmed using cutadapt (v.4.0) to remove Illumina adapter sequences and subsequently mapped to the reference genome with bowtie2 (v.2.4.5). We excluded multimapping and duplicate reads from downstream analysis. We then called peaks using macs2 (v.2.1.2). Peaks with a length of at least 100 nt were used in downstream analysis with DiffBind (v.3.6.1) to identify sites that were significantly (FDR < 0.05) differentially bound between samples. Coverage signal tracks were generated using deepTools (v.3.5.1). The replicates were merged for visualization in heatmaps by calculating the mean normalized coverage using WiggleTools (v.1.2.8). multiBigwigSummary (Galaxy v.3.5.1.0.0.) was used to calculate the average scores for 20-kb bins on the merged bigwig files visualized in box plots. Heatmaps used to assess the changes in accessibility of SOA bound peaks or genes downregulated in *SOA-KI* males were generated using deepTools (v.3.5.1).

qPCR

RNA extracted as per the RNA-seq protocol was used for generating cDNA with oligo(dT) as primers. qPCR was performed with FastStart Universal SYBR Green Master (ROX) mix (Roche, 04913850001) in a 7 µl reaction at 300 nM final primer concentration. We used *SOA* as template and *Rp49* as an endogenous control. *SOA* expressed from the *SOA-R* cassette was specifically detected with a primer targeting a part of the exogenous SV40 terminator included in the mRNA 3′ UTR. Total *SOA* mRNA was detected with primers targeting the coding sequence, which enabled comparisons of *SOA* levels in homozygous *SOA-R* and WT conditions. Cycling conditions as recommended by the manufacturer were applied. We corrected for primer efficiency using serial dilutions.

RT–PCR

RT–PCR was conducted using a OneStep Reverse Transcription-PCR kit (Qiagen, 210212) according to the user manual. In this kit, the reaction mixture contains all of the reagents required for both RT and PCR. For each reaction, 2 ng of RNA was used with primers for *SOA* binding to exons 2 and 3 (rt15 + rt16, Supplementary Table 5). Hence, RT is primed in a gene-specific fashion from the primer in exon 3. *S7* was used as a loading control (rt01 + rt02). A total of 33 PCR cycles were used for *SOA*, 27 cycles for *S7*. The PCR products were separated on a 2% TBE agarose gel and imaged using ChemiDoc MP V3 (Bio-Rrad). Uncropped gel pictures are provided in Supplementary Fig. 1.

Cloning of plasmids for baculovirus expression

The expression cassettes for Ag55 cells were cloned into a pFastBac Dual backbone (Thermo Fisher, 10712024) used for baculovirus generation. Plasmids were generated by Gibson assembly and restriction cloning (details can be provided upon request). The *EF1a* promoter (approximately 1 kb upstream of the TSS of *AGAP007405*) was amplified from genomic DNA with primers s047 and s048 (Supplementary Table 5) using LA Taq polymerase (Takara, RR002A). The coding sequence of *SOA* was amplified from cDNA generated from an adult male RNA sample. Primstar GXL (Takara, R050A) was used to amplify the coding sequence from the start codon to the end, excluding the stop codon. The vector expressing *SOA*(*1–229*) was cloned from the vector with full-length *SOA* coding sequence, as was the vector expressing *SOA*(*112–1265*) (myb-less). All constructs contain a C-terminal 2×HA tag followed by a T2A cleavage site and eGFP, which enables assessment of the infection rate.

Generation of baculoviruses

pFastBac vectors with expression cassettes were transposed into the baculoviral genome using chemically competent DH10Bac cells (Thermo Fisher Scientific) according to the manufacturer's protocol. Preparation of the baculoviral genome, transfection/P0 virus generation and P1 virus amplification were performed as described in the Bac-to-Bac manual (Thermo Fisher Scientific), with the exception of using Cellfectin® II transfection reagent and Sf-900 III serum-free medium (Thermo Fisher Scientific).

Cell culture and baculovirus infections

Ag55 cells provided by M. Adang were cultured in Leibovitz L15 medium with 10% FBS (Gibco, 10270-10,6 lot: 2260092) and 1× penicillin–streptomycin (Gibco, 15140122) at 27 °C, 80% humidity. Ag55 cells were authenticated by RNA-seq. Cells were tested every 6 months for mycoplasma (MycoAlert PLUS Mycoplasma Detection kit, Lonza LT07-701). All tests were negative. For the CUT&Tag experiment, 2 million cells were seeded in a 6-well plate. After 16 h, 600 µl of baculovirus in Sf-900 III serum-free medium was added to the cells. For the RNA-seq experiment, 0.75 million cells were seeded per each well of a 24-well plate. After 16 h, 200 µl of baculovirus in Sf-900 III serum-free medium was added. In both experiments, after 6 h the medium was changed to fresh L15. For the western blotting, 20 million cells were seeded in a 10-cm dish and infected with 6 ml of baculovirus on the next day and the baculovirus was not removed. Cells were collected for further processing 48 h after the addition of the baculovirus.

Nuclear extracts and IP from Ag55 cells

Cells were collected and washed with PBS. The cell pellet was resuspended in hypotonic lysis buffer (25 mM HEPES, pH 7.6, 10 mM NaCl, 5 mM $MgCl_2$, 0.1 mM EDTA and 1× protease inhibitor cocktail) and incubated on ice for 15 min. Next, NP-40 was added to a final concentration of 0.1% and the cells were vortexed for 30 s. The nuclei were pelleted and washed with sucrose buffer (25 mM HEPES, pH 7.6, 2 mM $MgCl_2$, 3 mM $CaCl_2$, 0.3 M sucrose and 1× protease inhibitor cocktail). The nuclear pellet was then resuspended in HMG-K400 buffer (25 mM HEPES, pH 7.6, 2.5 mM $MgCl_2$, 10% glycerol, 0.2% Tween, 400 mM KCl and 1× protease inhibitor cocktail) and rotated for 30 min at 4 °C. After centrifugation, the supernatant was either used directly for western blotting or for IP with the HA antibody. IP was performed by incubating 0.160 mg of nuclear soluble protein extract with 2 µl of HA antibody overnight. The bound SOA–antibody complexes were captured using Protein G dynabeads (1 h at 4 °C) followed by 3 washes in HMGT-K400 buffer. IPs were eluted by incubation in 2× LDS buffer with 200 mM DTT (37 °C, 10 min). For the SOA antibody IP, chromatin extracts from Ag55 cells infected with male SOA(1–1265), female SOA(1–229) or empty baculovirus control, which are all tagged with a C-terminal 2×HA epitope, were prepared. Cells were fixed in 0.1% PFA and nuclei prepared by using a previously published Nexson protocol[44]. The chromatin was sheared by sonication and diluted into the final IP buffer (0.05% SDS, 125 mM NaCl, 10 mM Tris (pH 8), 1 mM EDTA). Next, 5% of the input was removed and the remaining material was incubated with SOA antibody overnight. The bound SOA–antibody complexes were captured using Protein G dynabeads (1 h at 4 °C) followed by 3 washes in RIPA (25 mM HEPES pH 7.6, 150 mM NaCl, 1 mM EDTA, 1% Triton-X 100, 0.1% SDS, 0.1% DOC and protease inhibitors), 1 wash in LiCl buffer (250 mM LiCl, 10 mM Tris-HCl, 1 mM EDTA, 0.5% NP-40 and 0.5% DOC) and 2 washes in TE buffer. IPs were boiled in 1× Laemmli buffer (95 °C, 10 min).

SDS–PAGE and western blotting

Proteins were separated by 4–12% NuPAGE gradient gels in 1× MOPS buffer. Gels were transferred to a 0.45 µm PVDF membrane in Tris-glycine transfer buffer with 10% methanol (16 h at 60 mA). Membranes were blocked for 1 h in 5% milk in PBS–0.2% Tween, then incubated with primary antibodies (Supplementary Table 4) overnight at 4 °C. For SOA antibody, 5% horse serum was used as a blocking agent. Secondary HRP-coupled antibodies were used at 1:5,000 dilution for 1 h. Blots were developed using Lumi-Light Western Blotting substrate (Roche, 12015200001) and/or SuperSignal West Femto (Thermo Fisher, 34094) and imaged on a ChemiDoc MP V3 (Bio-Rad). Uncropped western blots are provided in Supplementary Fig. 1.

Recombinant protein purification

The untagged SOA fragments were generated from His_6–GST-3C–SOA expression vectors and used for electrophoretic mobility shift assay (EMSA), size-exclusion chromatography coupled to multi-angle light scattering (SEC–MALS) and antibody generation. His_6–GST-3C–SOA fragments (1–122, 1–229 and 1–331) were expressed from pET vectors in *Escherichia coli* (BL21 DE3 codon$^+$) overnight at 18 °C using 1 mM IPTG in LB medium. Cells were lysed in lysis buffer (50 mM Tris-Cl pH 8.0, 800 mM NaCl, 1 mM EDTA, 1 mM DTT, 5% glycerol and EDTA-free complete protease inhibitor cocktail) using a Branson Sonifier 450 and cleared by centrifugation (40,000*g*, 30 min at 4 °C). Additional 250 mM NaCl was added to the cleared lysates and a PEI-based precipitation of nucleic acids (0.2% w/v polyethylenimine, 40 kDa, pH 7.4) for 5 min at 4 °C was performed, followed by a second round of centrifugation (4,000*g*, 4 °C, 15 min). Recombinant proteins were affinity-purified from cleared lysates using a NGC Quest Plus FPLC system (Bio-Rad) and a GSTrap HP 5 ml column (Cytiva) following the manufacturer's protocols. Proteins were digested with 3C protease (1:100 w/w) overnight at 4 °C during dialysis in 50 mM Tris-Cl pH 8.0, 800 mM NaCl, 1 mM DTT and 5% glycerol to cleave off the His_6–GST tag. Digested proteins were re-run over the GSTrap HP 5 ml column to absorb out the His_6–GST, concentrated using Amicon 15 ml spin concentrators (Merck Millipore) and subjected to gel filtration (Superdex 200 16/60 pg in 25 mM Na-HEPES, 800 mM NaCl, 1 mM DTT and 10% glycerol, pH 7.4). Peak fractions containing the recombinant proteins after gel filtration were pooled, and protein concentration was determined by using absorbance spectroscopy and the respective extinction coefficient at 280 nm before aliquots were flash-frozen in liquid nitrogen and stored at −80 °C. The His_6–MBP-tagged SOA fragments and His_6–MBP control were used in EMSA and fluorescence polarization (FP) experiments. His_6–MBP-tagged SOA fragments and His_6–MBP control were expressed from a pET vector in *E. coli* (BL21-CodonPlus(DE3)-RIL, Agilent) using LB medium and overnight incubation with 0.5 mM IPTG at 18 °C. Cells were lysed in lysis buffer (30 mM Tris-Cl, 500 mM NaCl, 10 mM imidazole, 0.5 mM TCEP, complete protease inhibitors, 2 mM $MgCl_2$ and 150 U ml^{-1} benzonase, pH 8.0) using a high-pressure homogenizer (constant systems CF1 at 1.9 kBar). The lysate was cleared by centrifugation (40,000*g*, 4 °C, 30 min) and loaded onto a HisTrap FF 5 ml column (Cytiva) using a NGC Quest Plus FPLC system (Bio-Rad). The column was washed with buffer A (30 mM Tris-Cl, 500 mM NaCl and 10 mM imidazole, pH 8.0), followed by a second wash with buffer A containing 1 M NaCl and a third wash with buffer A containing 25 mM imidazole. Recombinant proteins were eluted by applying a linear gradient of 25–500 mM imidazole (pH 8.0) in buffer A over 15 column volumes. Peak elution fractions were pooled and concentrated using an Amicon 15 ml spin concentrator with 10 kDa cut-off (Merck Millipore). Concentrated proteins were applied to a gel filtration column (Superdex 200 16/60 pg, Cytiva, in 10 mM Na-HEPES pH 7.4, 150 mM NaCl, 1 mM TCEP and 5% glycerol). Peak fractions containing recombinant proteins were pooled and concentrated to 200 µM using an Amicon 15 ml spin concentrator with 10 kDa cut-off. Aliquots of the recombinant proteins were snap-frozen in liquid nitrogen and stored at −80 °C. The recombinant proteins were analysed by SDS–PAGE and visualized by Coomassie staining.

Antibody generation

Tagless SOA(1–122) was re-buffered in PBS using a PD-10 column (Cytiva) for immunization. Immunization was carried out by Eurogentec using

their polyclonal 28-day speedy programme. For epitope purification of the SOA antibody from the serum, 2 ml sulfolink resin (Thermo Fisher Scientific) was covalently conjugated with 3 mg tagless SOA(1–122) according to the manufacturer's protocol. Next, 10 ml final bleed was incubated with the SOA(1–122)-conjugated sulfolink resin at 4 °C overnight while rotating. After incubation, the resin was washed with PBS containing 0.1% Triton X-100, followed by PBS in a gravity-flow poly-prep column (Bio-Rad). Elution was performed using low pH (100 mM glycine-Cl and 150 mM NaCl, pH 2.3) followed by immediate neutralization of elution fractions with Tris-Cl pH 8.0. The eluted antibody was re-buffered using a PD-10 column (PBS, 0.05% NaN_3 and 10% glycerol) and concentrated to 1 mg ml^{-1} using an Amicon spin-concentrator before flash-freezing in liquid nitrogen and storage at −80 °C.

Antibody validation

To validate the specificity of the SOA antibody described in this study, we performed western blotting comparing female Ag55 cells ectopically expressing full-length SOA(1–1265), SOA lacking the myb-domain epitope or an empty control. The SOA constructs additionally contained a C-terminal HA-tag. This revealed a specific band present in only full-length, but not the two control conditions (Extended Data Fig. 5a), and two nonspecific bands present in all conditions. Note that we were unable to detect endogenous SOA proteins by western blotting from Ag55 cells or from male/female tissues, which is probably due to the low abundance of the SOA protein. We conducted IP experiments with HA antibody or SOA antibody and detected the captured proteins by western blotting with the other antibody (SOA antibody for HA-IP and HA antibody for SOA-IP, respectively; Extended Data Fig. 5b,c). The specific SOA band detected in the input was also enriched by IP. Furthermore, SOA antibody could not recognize a SOA version lacking the myb domain (amino acids 1–112, the epitope used to raise the antibody), whereas the SOA(1–229) fragment (female isoform) could be successfully detected. We also conducted IP experiments with SOA antibody versus IgG control from male pupal extracts. The bound proteins in this endogenous setup were then identified in an unbiased fashion by mass spectrometry (MS) (Extended Data Fig. 5d,e and Supplementary Table 1). SOA was the only protein not detected in the control and displayed by far the highest enrichment relative to the few contaminants, both in terms of the number of identified unique peptides identified (n = 12, 11, 13 and 12 for the 4 replicates) as well as the intensity. We also validated the specificity of the antibody by CUT&Tag and IF using the *SOA-KI* loss-of-function mutants as a control. In both cases, the detected signals and peaks vanished (Fig. 4a,b), which directly supports specificity. Last, the CUT&Tag experiment from Ag55 cells expressing HA-tagged SOA(1–1265) was performed in parallel with SOA and HA-tag antibodies. The two profiles (HA antibody, SOA antibody) produced similar profiles (data not shown).

EMSA

The desired amount of protein was diluted into 10 µl of 1× EMSA buffer (20 mM HEPES-KOH (pH 7.5), 100 mM KCl and 0.05% NP-40). GST or MBP was used as a negative control. The protein amounts were 100 fmol (1×) to 12.5 pmol (125-fold excess over DNA). Next, 100 fmol of the DNA probe (601-sequence, 147 bp[45] or X-chromosome promoter sequences bound by SOA, 300 bp; Supplementary Table 1) was added, incubated at room temperature for 30 min and subjected to gel electrophoresis (1.6% TBE agarose). DNA was stained with SYBR Safe and detected using a Typhoon FLA9500 gel scanner. The experiment was repeated three times with similar results. Uncropped gel pictures are provided in Supplementary Fig. 1.

SEC–MALS measurement

SEC–MALS measurements were performed at 25 °C in 25 mM HEPES (pH 7.5), 500 mM NaCl and 1 mM DTT as the column buffer using a GE Healthcare Superdex 200 10/300 Increase column on an Agilent 1260 HPLC at a flow rate of 0.5 ml min^{-1}. Loading concentrations were 200 µM for the SOA(1–112) and SOA(1–229) fragments and 11 µM for the SOA(1–331) fragment. Elution was monitored using an Agilent multi-wavelength absorbance detector (data collected at 280 and 260 nm), a Wyatt Heleos II 8+ multi-angle light scattering detector and a Wyatt Optilab differential refractive index detector. The column was equilibrated overnight in the running buffer to obtain stable baseline signals from the detectors before data collection. Inter-detector delay volumes, band-broadening corrections and light-scattering detector normalization were calibrated using an injection of 2 mg ml^{-1} BSA solution (Thermo Pierce) and standard protocols in ASTRA 8. Weight-averaged molar mass (M_w), elution concentration and mass distributions of the samples were calculated using ASTRA 8 software (Wyatt Technology).

DNA oligomer interaction measurements in vitro using FP

To generate dsDNA oligonucleotide substrates, Cy5-labelled ssDNA 20-mers were annealed with reverse-complement 20-mer oligonucleotides at 50 µM in TE buffer by heating to 90 °C for 1 min and subsequent incubation on ice (all oligonucleotides synthesized and HPLC-purified by Integrated DNA Technologies, sequences in Supplementary Table 1). Using a 384-well plate (Corning, low-volume, polystyrene, black), Cy5-labelled ssDNA and dsDNA oligonucleotide substrates (5 nM) were incubated with varying concentrations of His_6–MBP-tagged SOA fragments or with a His_6–MBP control in a total volume of 20 µl FP buffer (10 mM Na-HEPES pH 7.4, 150 mM NaCl, 1 mM TCEP, 0.1 g l^{-1} BSA, 5% glycerol and 0.05% Triton X-100). After 10 min of incubation at 20 °C, FP of the Cy5-labelled oligonucleotides were analysed on a Tecan Spark 20M plate reader at 20 °C (excitation wavelength of 625 nm; emission wavelength of 665 nm; gain of 120; flashes of 15; integration time of 40 µs). Normalized FP values were calculated by subtracting the FP value of each oligonucleotide-only measurement from all conditions that contained variable amounts of the respective recombinant protein. The normalized FP values from three independent experiments, including standard deviations, were plotted using GraphPad Prism 8. EC_{50} values, which serve as a proxy for the binding constant (K_d), were determined by applying a four parameter [agonist] versus response fit with variable slope in GraphPad Prism 8 if applicable.

Sample preparation for MS

Approximately 0.2 ml (dry volume) of sex-separated pupae were homogenized for each replicate in 0.5 ml of cytoplasm isolation buffer (Cell Signaling Technologies, 9038S) using a handheld homogenizer. After 5 min of incubation on ice, the homogenate was cleaned by spinning through a cell strainer (Corning, 352235) on a FACS tube (500*g* for 5 min). Cell fractionation of nuclei was continued according to the manual using a Cell Fractionation kit (Cell Signaling Technologies, 9038S). The nuclei were resuspended in 0.125 ml of NIB (250 mM NaCl, 50 mM HEPES, pH 7.6, 0.1% IGEPAL, 10 mM $MgCl_2$, 10% glycerol and protease inhibitors complete, Roche). For the antibody validation experiment, NIB contained 600 mM NaCl. This was sonicated using a Bioruptor Plus, 5 cycles on/off (high), 30 s each followed by 5 min of centrifugation at 12,000*g*. The supernatant was quantified using Bradford reagent (Avantor PanReac AppliChem, A6932.0250) and 0.4 mg nuclear protein extract used per replicate with n = 5 males and n = 5 female extracts used in total. For the antibody validation experiment, n = 4 male replicates were used for each condition (SOA antibody, IgG control). Per IP and replicate, 20 µl of Protein G dynabeads (Thermo Fisher, 10004D) were washed 2× with NIB, then incubated with 4 µl of SOA antibody (rabbit polyclonal, clone 87) in 40 µl NIB for 45 min on a wheel. This was washed 2× with NIB and resuspended in 40 µl of NIB, which was then added to the nuclear extracts and incubated for 30 min at 4 °C on a wheel. Unbound proteins were removed by three washing steps with 200 µl NIB. Bound proteins eluted by heating beads in 30 µl

1×LDS buffer (Thermo Fisher Scientific) supplemented with 100 mM DTT for 10 min at 70 °C and 1,400 r.p.m. in a thermomixer (Eppendorf). Proteins were subsequently run on a 4–12% NOVEX NuPage gel (Thermo Fisher Scientific) for 8 min at 180 V in 1× MOPS buffer (Thermo Fisher Scientific). Proteins were fixed and stained with 0.25% Coomassie Blue G-250 (Roth) in 10% acetic acid (Sigma)–43% ethanol (Roth). The gel lane was minced and destained with a 50% ethanol–50 mM ammonium bicarbonate (ABC) pH 8.0 solution. Proteins were reduced in 10 mM DTT–50 mM ABC pH 8.0 for 1 h at 56 °C and then alkylated with 50 mM iodoacetamide–50 mM ABC pH 9.0 for 45 min at room temperature in the dark. Proteins were digested with mass-spectrometry-grade trypsin (Sigma) overnight at 37 °C. Peptides were extracted from the gel using twice a mixture of 30% acetonitrile (VWR) and 50 mM ABC pH 8.0 solution followed by two times with pure acetonitrile, which was ultimately evaporated in a concentrator (Eppendorf) and loaded on an activated self-made C18 mesh (AffiniSep) StageTips[46].

MS data acquisition and analysis

Peptides were separated on a 25 cm self-packed column (New Objective) with 75 µm inner diameter filled with ReproSil-Pur 120 C18-AQ (Dr. Maisch). The EASY-nLC 1000 (Thermo) column was mounted onto a Q Exactive Plus mass spectrometer (Thermo), and peptides were eluted from the column in an optimized 90 min gradient from 2 to 40% acetonitrile–0.1% formic acid solution at a flow rate of 200 nl min^{-1}. The mass spectrometer was operated in a data-dependent acquisition mode with one MS full scan and up to ten MS/MS scans using HCD fragmentation. MS raw data were searched against Anopheles_gambiae.AgamP4.pep.all (15,125 entries) with the Andromeda search engine[47] of the MaxQuant software suite (v.1.6.5.0)[48]. Cys-carbamidomethylation was set as fixed modification and Met-oxidation and protein N-acetylation were considered as variable modifications. Match between run option was activated. Before further processing, protein groups marked with reverse, only identified by site or with fewer than two peptides (one of them unique) were removed.

IF staining

In our initial IF stainings, tissues were dissected and then fixed in 4% formaldehyde in PEM (0.1 M PIPES (pH 6.9), 1 mM EGTA and 1 mM $MgCl_2$) for 20 min and washed three times with PBS. Samples were blocked for 1 h rocking with freshly prepared 0.5% BSA, 0.3% Triton X-100 in 1×PBS solution. The samples were washed with Basilicata-blocking (BB) buffer (0.5% BSA in PBS–0.2% Tween (Sigma Aldrich, P1379)), followed by overnight incubation with primary antibody (anti-SOA, rabbit polyclonal, 1:300 in BB). Samples were washed three times in BB and then stained with a secondary antibody (Alexa fluorophore-labelled goat anti-rabbit, ThermoFisher, A21430, 1:400 in BB). Samples were thoroughly washed with BB, then with 1×PBS–0.2% Tween. For the embryo staining, 19 h AEL-stage embryos were placed in small baskets (Falcon 40 µm cell strainers, 352340) and dechorionated in bleach (4.8% chlorine) for 1–2 min with visual monitoring of chorion dissolution under a binocular microscope. As soon as chorion disappeared, they were rinsed with PBS followed by fixation in PBS, 4% PFA and 0.1% Triton X-100 for 20 min at room temperature. They were then rinsed 3 times with PBS and then stored in methanol at −20 °C. Before IF staining, the black endochorion was then manually peeled off with a needle under a binocular microscope using a Petri dish with a double-sided tape with embryos submerged in 100% methanol. The peeled embryos were transferred using a 1.5 ml pipette into a 1.5 ml Eppendorf tube containing PBS. Blocking and antibody incubations were performed as for the dissected tissues. During the course of the project, we realized that lower PFA concentrations significantly improved the signal-to-noise of the SOA staining; therefore we changed the fixation step in our protocol to 1% PFA for 15 min. We also noted that prolonged incubation with primary antibody (60–72 h) improved signal-to-noise; for embryos prolonged incubation was crucial to obtain SOA staining. For the RNAseA experiment, midguts were dissected in PBS and then rinsed 2× with CSK buffer (10 mM PIPES-KOH, pH 7.0, 100 mM NaCl, 300 mM sucrose and 3 mM $MgCl_2$), then incubated for 10 min in CSK, 0.5% Triton X-100 and 1 mg ml^{-1} RNaseA (or control). The midguts were then rinsed 2× in CSK buffer. For each condition, 2 midguts (2 replicates) were then put in 0.15 ml TRIzol for RNA isolation to check the effectiveness of the RNase treatment versus control. Meanwhile, the remaining midguts were fixed with 1% PFA in PEM for 15 min at room temperature and stained as per the standard conditions described above. For actinomycin D treatment, the tissues were dissected and put into 0.5 ml of L15 tissue culture medium, 10% FBS and penicillin–streptomycin. Actinomycin D was added to a final concentration of 5 µg ml^{-1} to half of the samples, the other half was left untreated (control), and both conditions were incubated for 1 h at 26 °C in a tissue culture incubator. The tissues were then fixed in PEM and 1% PFA for 15 min at room temperature and the staining was conducted as described above. As a positive control, we co-stained for phosphorylated RNA Pol2, which has been previously described to increase after actinomycin D treatment[49].

Polytene chromosome preparations

Fourth instar larva were immobilized on ice for 15–20 min, then they were placed in a drop of 75 mM KCl and the head and abdomen was cut off with an ultrafine dissection scissor and discarded. The thorax was placed in a fresh drop of 75 mM KCl on a glass microscopy slide and the gut and tissues attached to it were gently pulled out with forceps and discarded. The remaining thorax piece containing the imaginal discs and salivary glands was gently opened and placed in a fresh drop of fixative (25% acetic acid, 1% methanol-free PFA in H_2O). Imaginal discs and salivary glands immediately turn white and are now easy to spot. They were dissected in approximately 5–7 min under a binocular microscope, attempting to completely remove the fat and cuticle. After 7–8 min, the fixative was removed and a fresh drop of PBS–0.1% Tween containing 1:1,000 of DAPI solution was added. A coverslip was put on the dissected discs and salivary glands and excess solution carefully removed with a Kimtech wipe. The coverslip was gently tapped with the rubber of a pencil while observing squashing under a fluorescent microscope. When spreading was sufficient, the slide was put in liquid nitrogen and the coverslip was flicked off with a razor blade. The slide was then placed in PBS and stored at 4 °C until staining. For the RNA FISH experiment, all solutions described above additionally contained RNasin Ribonuclease inhibitor (Promega N2511) at 1:1,000 dilution.

Staining of polytene chromosomes

The slides were incubated in a coplin jar containing PBS and 0.4% Triton X-100 for 30 min at room temperature on an orbital shaker set at 220 r.p.m. The slides were rinsed 2× with PBS and 0.1% Tween. The slides were then incubated on the orbital shaker with blocking buffer (PBS, 0.1% Tween, 0.2% BSA and 5% horse serum; filtered) for 30–60 min at room temperature. The slides were placed in a wet chamber, and incubation with primary antibody in blocking buffer (0.25 ml solution, slide covered with Parafilm) was conducted overnight at 4 °C. The slides were washed in a coplin jar on the orbital shaker 3× in PBS and 0.2% Tween. Secondary antibodies were incubated for 1–2 h in a wet chamber at room temperature (0.25 ml of solution, slide covered with Parafilm). The slides were washed in a coplin jar on the orbital shaker 2× in PBS and 0.2% Tween followed by a 15 min incubation with PBS, 0.1% Tween and DAPI (1:1,000) in a wet chamber as for the antibodies. The slides were rinsed with PBS and then mounted with Prolong Gold.

Co-immunostaining with RNA FISH

Polytene squashes were prepared as described above. RNA FISH was performed according to the manufacturer's protocol for IF followed by smFISH, referred to as the sequential protocol. PBS was prepared from a 5× sterile PBS solution with DEPC water and 1 µl RNAseIn per

50 ml of 1× buffer was added. Slides with squashes were briefly rinsed 2× in PBS, 0.1% Tween and RNAseIn for 10 min and 1× with PBS. Primary antibody in PBS incubation was performed 60–72 h at 4°C in a humidified chamber. Excess antibody was washed out 3× with PBS followed by secondary antibody incubation in PBS for at least 3 h. Unbound secondary was washed out 2× in PBS and the slide was then crosslinked in 4% PFA–PBS for 10 min at room temperature. Excess of fixative was removed using PBS washes and then the smFISH protocol was started using 1× wash buffer A (SMF-WA1-60-BS, LGC Biosearch Technologies) supplemented with 10% formamide. This was followed by hybridization in Stellaris RNA FISH hybridization buffer (SMF-WA1-60-BS, LGC Biosearch Technologies) supplemented with 10% with formamide containing 125 nM probe mix targeting the introns of the X-linked gene *act5c* (*AGAP000651*, sequences in Supplementary Table 1), which was incubated overnight in a humidified chamber at 37 °C. Excess probe was removed by two washes with wash buffer A, 30 min each at 37 °C, followed by a brief wash in wash buffer B (SMF-WB1-20-BS, LGC Biosearch Technologies). Slides were mounted in Vectashield vibrance with DAPI (H-1800, Vector Laboratories) and imaged after 1 h using Visiscope Microscope, ×63 water objective.

CUT&See

The protocol was based on the spatial CUT&Tag[50] with the following modifications. pA–Tn5 produced by the IMB Protein Production Core Facility was loaded with pre-annealed oligonucleotides Tn5MErev, Tn5ME-A-ATTO488 and Tn5ME-B-ATTO488. Adult male midguts were dissected, fixed with 0.2% PFA in PEM buffer with RNAseIN (1:1,000) at room temperature for 5 min. The fixation step was quenched with 2.5 M glycine (1:20). After quenching, the midguts were washed 2 times with the CUT&Tag wash buffer (20 mM HEPES pH 7.6, 150 mM NaCl, 0.5 mM spermidine and 1× protease inhibitor cocktail) and rinsed briefly with RNAse-free water. The midguts were then incubated for 5 min at room temperature in permeabilization buffer (0.1% NP40 and 0.05% digitonin in wash buffer) and washed once with the NP40–digitonin wash buffer (0.01% NP40 and 0.05% digitonin in wash buffer). Subsequently, the midguts were incubated overnight with the SOA antibody (1:100 dilution) at 4 °C on a Nutator in the antibody buffer (2 mM EDTA and 0.1% BSA in NP40–digitonin wash buffer). The next day, the midguts were rinsed once with NP40–digitonin wash buffer, then incubated on the Nutator for 1 h at room temperature with the secondary antibody (1:100 dilution of F(ab′)2-goat anti-rabbit IgG (H+L) cross-adsorbed secondary antibody, Alexa Fluor-555; 555A21430 ThermoFisher) in the same buffer. This was followed by a rinse with the NP40–digitonin wash buffer. Next, the pA–Tn5 complex pre-loaded with fluorescently labelled oligonucleotides was added into Dig-300 buffer (20 mM HEPES pH 7.6, 300 mM NaCl, 0.5 mM spermidine, 0.05% digitonin and 1× protease inhibitor cocktail) at a final concentration of 31 nM and incubated for 1 h at room temperature on the Nutator. After a 5-min wash with the Dig-300 buffer, the midguts were incubated in tagmentation buffer (10 mM $MgCl_2$ in Dig-300 buffer) for 1 h at 37 °C. The tagmentation step was stopped by adding EDTA to final concentration of 40 mM and incubating for 5 min on the Nutator. The midguts were finally washed with 1× NEBuffer 3.1 and then stained with DAPI.

Microscopy

Slides were mounted using ProLong Gold Antifade mountant with DAPI (P36935, Thermo Fisher Scientific), unless otherwise stated, and imaged using a fluorescence spinning disc confocal microscope, VisiScope 5 Elements (Visitron Systems), which is based on a Ti-2E (Nikon) stand and equipped with a spinning disc unit (CSU-W1, 50 µm pinhole; Yokogawa). The set-up was controlled using VisiView 5.0 software, and images were acquired with a ×100/1.49 NA oil-immersion objective (CFI Apo SR TIRF ×100, Nikon) or ×60/1.2 NA water-immersion (CFI Plan Apo VC60x WI) and a sCMOS camera (BSI; Photometrics). 3D stacks of images were recorded for each sample. Confocal imaging was performed using a Stellaris 8 Falcon (Leica Microsystems) confocal system equipped with white light laser. Images (1,552 × 1,552 pixel format, 0.93 pixel size) were acquired using a HC PL APO CS2 ×63/1.40 NA oil-immersion lens, and fluorescence was detected using a detector HyD S for DAPI (emission band 427–460 nm), HyD X for Alexa488 (500–545 nm) and HyD R for Alexa555 (560–730 nm). Tissue images were acquired through 87 slices at 200-nm step intervals using a line accumulation of 3 times. 3D view of the *z*-stacks and image processing were obtained using Imaris software (v.9.9.1). The IF stainings were replicated in at least four independent experiments.

Modelling the evolution of *SOA*

Our results indicated that the SOA^+ allele speeds up male development by about 4 h. To investigate the evolutionary implications of such a progression of development, we used the standard one-locus-two-alleles model of viability selection, with different viabilities in males and females[22]. In this model, the relative viability of the three genotypes SOA^-/SOA^-, SOA^+/SOA^- and SOA^+/SOA^+ is $1, 1 + h_m \times s_m$ and $1 + s_m$, respectively, in males and $1, 1 - h_f \times s_f$ and $1 - s_f$, respectively, in females. Here s_m is the selection differential in favour of the SOA^+ allele in males, whereas s_f is the selection differential against SOA^+ in females. The factors h_m and h_f denote the degree of dominance of the SOA^+ allele. Throughout, we assumed that SOA^+ is dominant in males ($h_m = 1$) and recessive in females ($h_f = 0$) based on the general finding that selectively favoured alleles tend to be dominant in each sex[51]. However, we also considered other dominance values, and they led to the same conclusion (persistence of the SOA^+ allele at considerable frequencies for a wide range of selection coefficients) as long as $h_m > 0$.

Our estimate of s_m was based on the rationale that a shorter developmental time is favourable for survival to adulthood. According to population models specifically tailored to the life cycle of *Anopheles* mosquitoes[52], the daily survival probability of males is 0.9. Speeding up development by 4 h (which equates to one-sixth of a day) therefore corresponds to a survival benefit of $0.9^{5/6}/0.9 = 1.0177$. We therefore assume that the developmental advance of SOA^+-bearing males translates into the selection coefficient $s_m = 0.0177$. As this is a crude estimate, and sometimes different survival probabilities are used[53], we also considered other values of s_m, ranging from 0.005 to 0.05. We also considered a spectrum of selection coefficients s_f in females, ranging from 0 to 0.05. In Fig. 5h, s_f was set to zero in generation 5,000, corresponding to the assumption that alternative splicing (removing the negative fitness effects of SOA^+ in females) had evolved by then.

Evolutionary analyses, sequence analyses, alignments and visualizations

DNA and protein sequences were retrieved from VectorBase. Protein and DNA alignments were created using Clustal Omega. The pairwise percentage similarity of the SOA domains were obtained in Jalview (v.2.11.2.3). Alignments were visualized with ESPript. Lists of 1:1 orthologues were obtained using the Biomart tool from VectorBase. The *SOA* locus, its syntenic regions in other species and the analysis of its paralogue were obtained from VectorBase. The phylogeny and evolutionary distance calculations were performed using MEGA software (v.7.0). Figures were assembled using Adobe Illustrator and Adobe Photoshop (2021 version).

Bioinformatic and web resources

The following resources were used: cutadapt (https://github.com/marcelm/cutadapt); Bowtie2 (https://github.com/BenLangmead/bowtie2); macs2 (https://github.com/macs3-project/MACS); WiggleTools (https://github.com/Ensembl/WiggleTools); MEME (https://meme-suite.org/meme/); Gviz (https://bioconductor.org/packages/release/bioc/html/Gviz.html); STAR (https://github.com/alexdobin/

STAR); DiffBind (https://bioconductor.org/packages/DiffBind/); deepTools2 (https://deeptools.readthedocs.io/en/latest/); IGV (https://software.broadinstitute.org/software/igv/); R (https://www.r-project.org); DESeq2 (http://bioconductor.org/packages/DESeq2/); VectorBase (https://vectorbase.org/vectorbase/app); Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/); ESPript (https://espript.ibcp.fr/ESPript/ESPript/); Nuclear Localization Signal prediction (https://nls-mapper.iab.keio.ac.jp/cgi-bin/NLS_Mapper_form.cgi); IUPRED2 (https://iupred2a.elte.hu/); and DNA binding site predictor for Cys2His2 Zinc Finger Proteins (http://zf.princeton.edu/).

Statistics and reproducibility

All statistics were calculated using R Studio. In the violin plots, the centre line represents the median and the shape of the violin represents the distribution of underlying data. For all violin plots, *P* values were obtained using two-sided Wilcoxon rank-sum test (Extended Data Figs. 7j, 3d, 2g and 10h,k), with additional Bonferroni correction in Fig. 4d and Extended Data Figs. 7d 9e and 10j,l. In the box plots, the line that divides the box into two parts represents the median, box bottom, and top edges represent interquartile ranges (IQRs; 0.25th to 0.75th quartile (Q1–Q3)), whiskers represent Q1 − 1.5× IQR (bottom), Q3 + 1.5× IQR (top). Bar plots represent the mean with overlaid data points representing replicates. Results were considered significant at FDR below 0.05. NA, not analysed. For all pie charts, the *P* value was obtained with a one-sided Fisher's exact test for the overrepresentation on the X chromosome. For these, we compared SOA peaks to an equal number of peaks homogeneously distributed on all chromosomal arms (CUT&Tag, Figs. 2c, 3f and 5e) or analysed overrepresentation of X-linked genes in the upregulated and downregulated group in comparison with an equal number of genes homogeneously distributed on all chromosomal arms (RNA-seq, Figs. 3c and 4c). In Extended Data Fig. 8b, overrepresentation of CA-repeat-containing promoters on the X chromosome and autosomes were compared with all X-linked and autosomal genes. For scoring the developmental delay in Fig. 4f and Extended Data Fig. 9l, *P* values were obtained by a log-rank test for stratified data (Mantel–Haenszel test). In Fig. 5g, Benjamini–Hochberg-corrected *P* values were obtained with a two-sided *t*-test with pairwise comparisons between the genotypes. Further details are provided in the figure legends. Further data, DiffBind/DESeq2 and statistical test results are provided Supplementary Tables 1–3. The immunostainings were reproduced with similar results as follows: Fig. 1g and Extended Data Fig. 5g experiment (WT males, females) was conducted 7 times, each with tissues dissected from at least $n = 5$ adults of each sex (biological replicates); Extended Data Figs. 5h and 6a experiments (polytene squash, larval tissues) were conducted 3 times, each with at least 2 slides per sex, for which each slide contained tissues dissected from at least $n = 4$ larvae (biological replicates); Extended Data Fig. 5i,j (embryos) was conducted twice, each with at least $n = 30$ embryos (biological replicates); Fig. 2a experiment (SOA IF and co-FISH) was conducted 2 times with 2 slides each; each slide contained tissues dissected from at least $n = 4$ adults (8 biological replicates per experiment); Extended Data Fig. 6b experiment (CUT&See) was conducted once with tissue dissected from $n = 1$ adult (biological replicate); Extended Data Fig. 7h experiment (RNase A) was conducted 2 times, each with tissues dissected from at least $n = 5$ adults (biological replicates); Extended Data Fig. 7i experiment (actinomycin D) was conducted once with tissues dissected from at least $n = 5$ adults (biological replicates); Fig. 4b experiment (*SOA-KI*) was conducted 2 times, each with tissues dissected from at least $n = 5$ adults of each genotype (biological replicates); Extended Data Fig. 10b experiment (*SOA-R* IF and co-FISH) was conducted once with 2 slides, each slide contained tissues dissected from at least $n = 4$ larvae (biological replicates); and Fig. 5c and Extended Data Fig. 10c experiment (*SOA-R*) was conducted 2 times, each with tissues dissected from at least $n = 5$ adults of each sex (biological replicates).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

No restrictions apply and all data are available in the manuscript or the supplementary materials. RNA-seq, CUT&Tag and ATAC–seq data have been deposited into the Gene Expression Omnibus database (identifiers GSE210624 and GSE210630). MS data have been deposited into ProteomeXchange through the PRIDE database (project identifier PXD042353). DNA and protein sequences, and the Ensembl AgamP4 genome with the Ensembl AgamP4 annotation (release 48) were retrieved from VectorBase (www.vectorbase.org, publicly available). Metazoan upstream sequences for *A. gambiae* (AgamP4.34_2019-03-11) or *A. aegypti* (AaegL3.34_2019-03-11) databases used in FIMO are publicly available as part of the https://meme-suite.org/meme/tools/fimo search tool. RNA-seq data from ref. 4 is publicly available from the Sequence Read Archive under accession number SRP083856.

36. Volohonsky, G. et al. Tools for *Anopheles gambiae* transgenesis. *G3* **5**, 1151–1163 (2015).
37. Dong, Y., Simões, M. L., Marois, E. & Dimopoulos, G. CRISPR/Cas9 -mediated gene knockout of *Anopheles gambiae FREP1* suppresses malaria parasite infection. *PLoS Pathog.* **14**, e1006898 (2018).
38. Marois, E. et al. High-throughput sorting of mosquito larvae for laboratory studies and for future vector control interventions. *Malar. J.* **11**, 302 (2012).
39. Bernardini, F. et al. Site-specific genetic engineering of the *Anopheles gambiae* Y chromosome. *Proc. Natl Acad. Sci. USA* **111**, 7600–7605 (2014).
40. Jenkins, A. M., Waterhouse, R. M. & Muskavitch, M. A. T. Long non-coding RNA discovery across the genus *Anopheles* reveals conserved secondary structures within and beyond the *Gambiae* complex. *BMC Genomics* **16**, 337 (2015).
41. Henikoff, S., Henikoff, J. G., Kaya-Okur, H. S. & Ahmad, K. Efficient chromatin accessibility mapping in situ by nucleosome-tethered tagmentation. *eLife* **9**, e63274 (2020).
42. Zheng, Y., Ahmad, K. & Henikoff, S. CUT&Tag data processing and analysis tutorial. *protocols.io* https://www.protocols.io/view/cut-amp-tag-data-processing-and-analysis-tutorial-e6nvw93x7gmk/v1 (2020).
43. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1–21.29.9 (2015).
44. Arrigoni, L. et al. RELACS nuclei barcoding enables high-throughput ChIP–seq. *Commun Biol.* **1**, 214 (2018).
45. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
46. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
47. Cox, J. et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
48. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
49. Dubois, M. F., Nguyen, V. T., Bellier, S. & Bensaude, O. Inhibitors of transcription such as 5,6-dichloro-1-beta-D-ribofuranosylbenzimidazole and isoquinoline sulfonamide derivatives (H-8 and H-7) promote dephosphorylation of the carboxyl-terminal domain of RNA polymerase II largest subunit. *J. Biol. Chem.* **269**, 13331–13336 (1994).
50. Deng, Y. et al. Spatial-CUT&Tag: spatially resolved chromatin modification profiling at the cellular level. *Science* **375**, 681–686 (2022).
51. Kacser, H. & Burns, J. A. The molecular basis of dominance. *Genetics* **97**, 639–666 (1981).
52. Arifin, S. M. et al. An agent-based model of the population dynamics of *Anopheles gambiae*. *Malaria J.* **13**, 424 (2014).
53. White, M. T. et al. Modelling the impact of vector control interventions on *Anopheles gambiae* population dynamics. *Parasit. Vectors* **4**, 153 (2011).
54. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME suite. *Nucleic Acids Res.* **43**, W39–49 (2015).

Acknowledgements We thank V. Benes and the EMBL Genecore for the sequencing of the embryogenesis RNA-seq experiment; T. Sharpe and T. Mühlethaler at the Biozentrum Basel Biophysics Core for SEC–MALS; J. Cartano for technical support with the MS experiment; F. Kielisch (IMB Bioinformatics Core Facility) for help with statistical testing; A. Raj and M. Dunagin for help with RNA FISH probe design; J. H. G. Fritz García for help with optimization of the baculovirus experiments; G. Magnarini for technical assistance; A. Gautier and N. Schallon for insectary maintenance and blood feeding of mosquito lines; J. Barau for sharing expertise and reagents; staff at the IMB Microscopy and Genomics core facilities for support; staff at the IMB Protein Production Core Facility for the supply of recombinant enzymes; and M. Adang for gifting the Ag55 cell line. A.I.K. was supported by a Boehringer Ingelheim Fonds PhD Fellowship. The work of M.K. and F.J.W. is supported by the European Research Council (ERC Advanced grant no. 789240). M.F.B. received financial support for the work from the intramural High Potentials Grant programme of the University Medical Center Mainz. The

research of C.I.K.V. is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)–Individual Project Grant 513744403, Scientific Network Grant 531902894 and GRK GenEvo 407023052 and Forschungsinitiative Rheinland-Pfalz (ReALity). Mosquito breeding and transgenesis were supported by ANR grants ANR-11-EQPX-0022 and ANR-19-CE35-0007 GDaMO and by funding from INSERM, CNRS, the University of Strasbourg, and contrat triennal 'Strasbourg capitale européenne' 2018–2020. The IMB Genomics Core Facility, the Microscopy Core Facility and the use of the NextSeq500 (funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)–INST 247/870-1 FUGG) and spinning disc confocal system (VisiScope, 5-Elements, funded by the DFG–INST 247/912-1FUGG), and the confocal laser scanning microscope (Leica Stellaris 8 Falcon, funded by the DFG–497669232) are acknowledged.

Author contributions C.I.K.V. conceptualized the study with A.I.K., E.M. and M.F.B. E.M. and E.J. performed mosquito rearing, transgenesis and characterized the phenotype of the *SOA-KI* and *SOA-R* mosquitoes. M.F.B. performed IF, RNA FISH and microscopy. M.K. and F.W. performed the computational modelling of the spread of *SOA*. M.M.M. generated baculoviruses, recombinant proteins, conducted antibody purification and FP. F.B. performed MS. F.R. supported data processing and analyses. A.I.K. performed all other experiments and data analyses, including bioinformatics. C.I.K.V. and M.F.B. provided mentoring and guidance. C.I.K.V. and A.I.K coordinated the study, secured funding and wrote the manuscript with input from all authors.

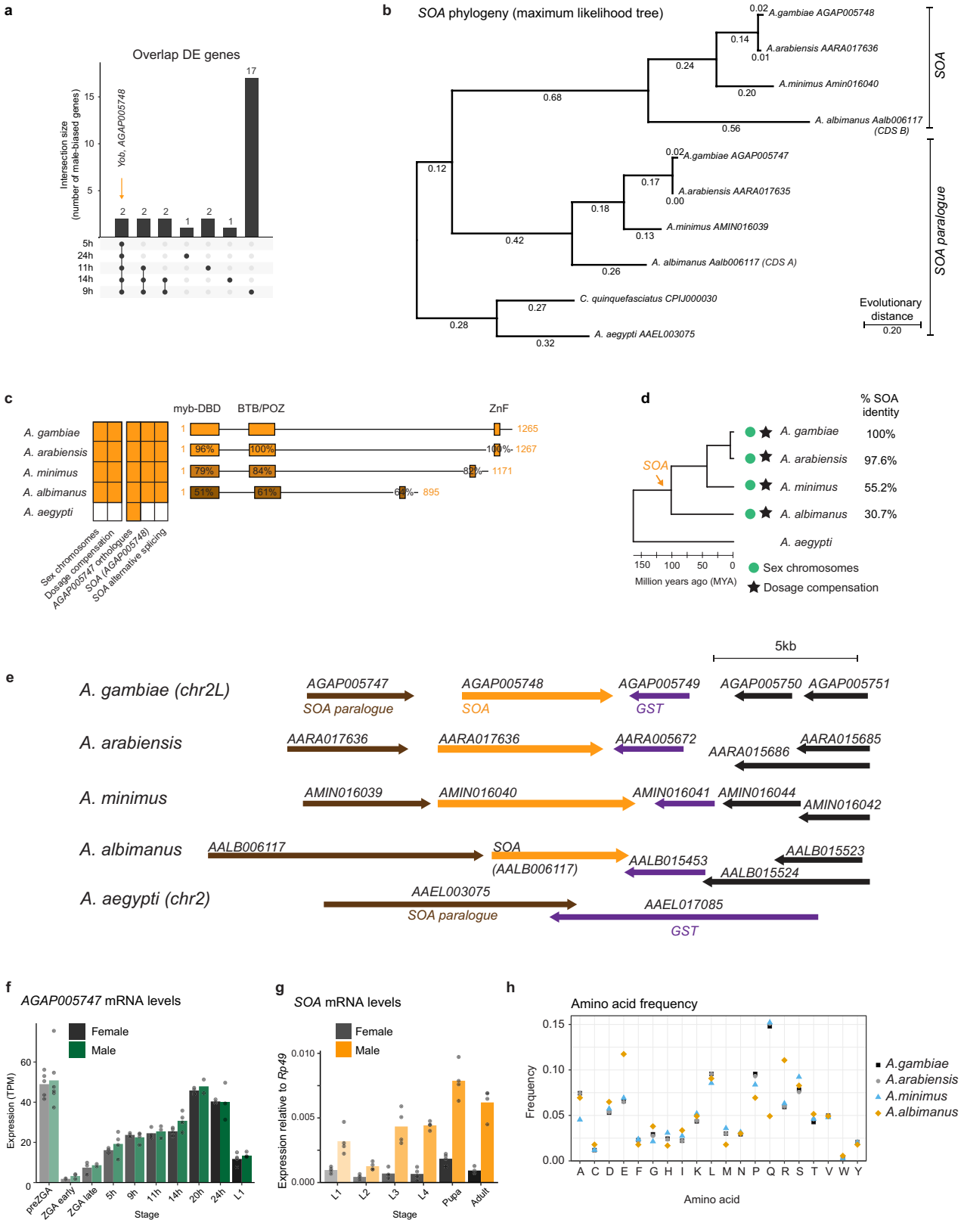
Competing interests The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-06641-0.

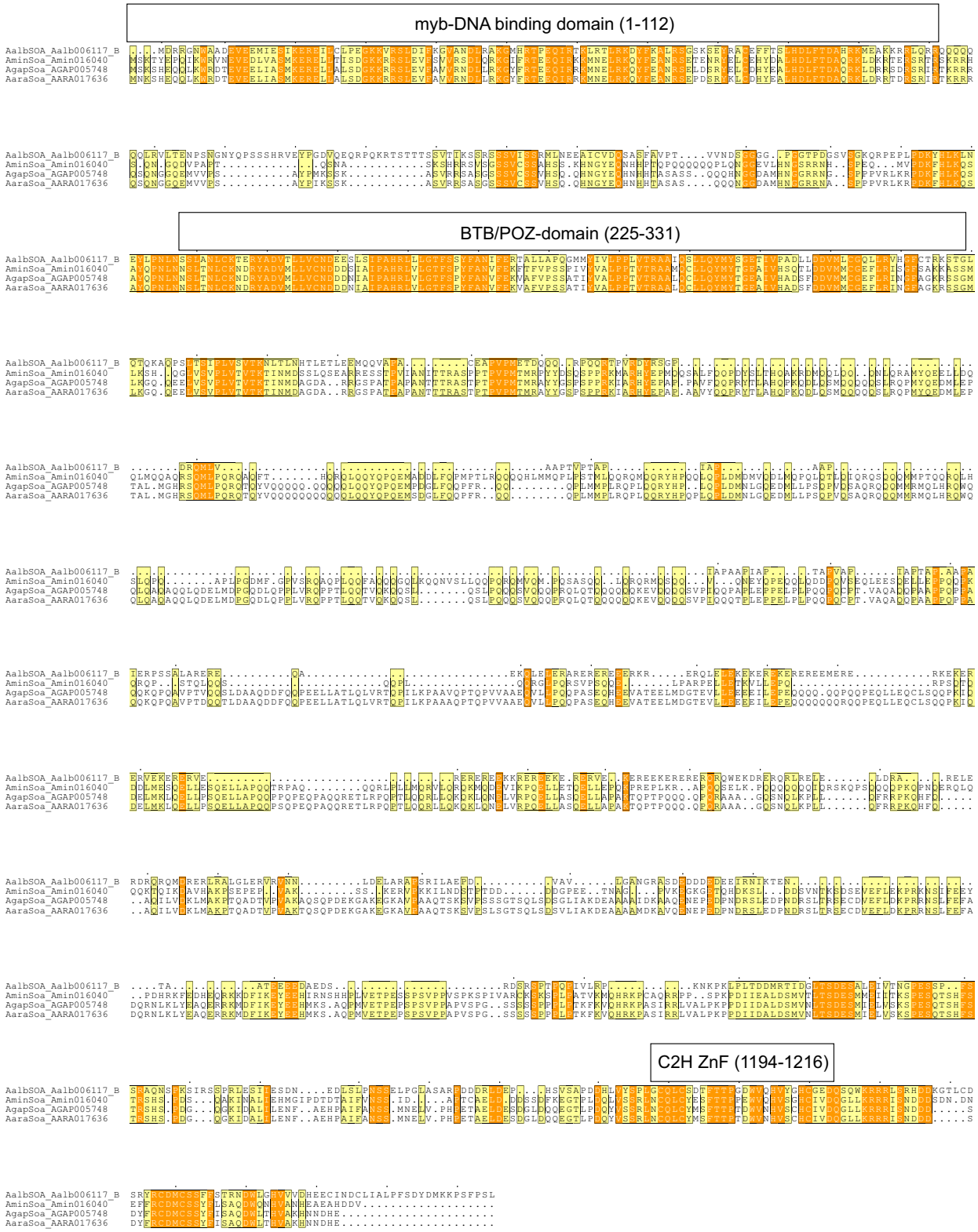
Correspondence and requests for materials should be addressed to Claudia Isabelle Keller Valsecchi.

Peer review information *Nature* thanks Jan Larsson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

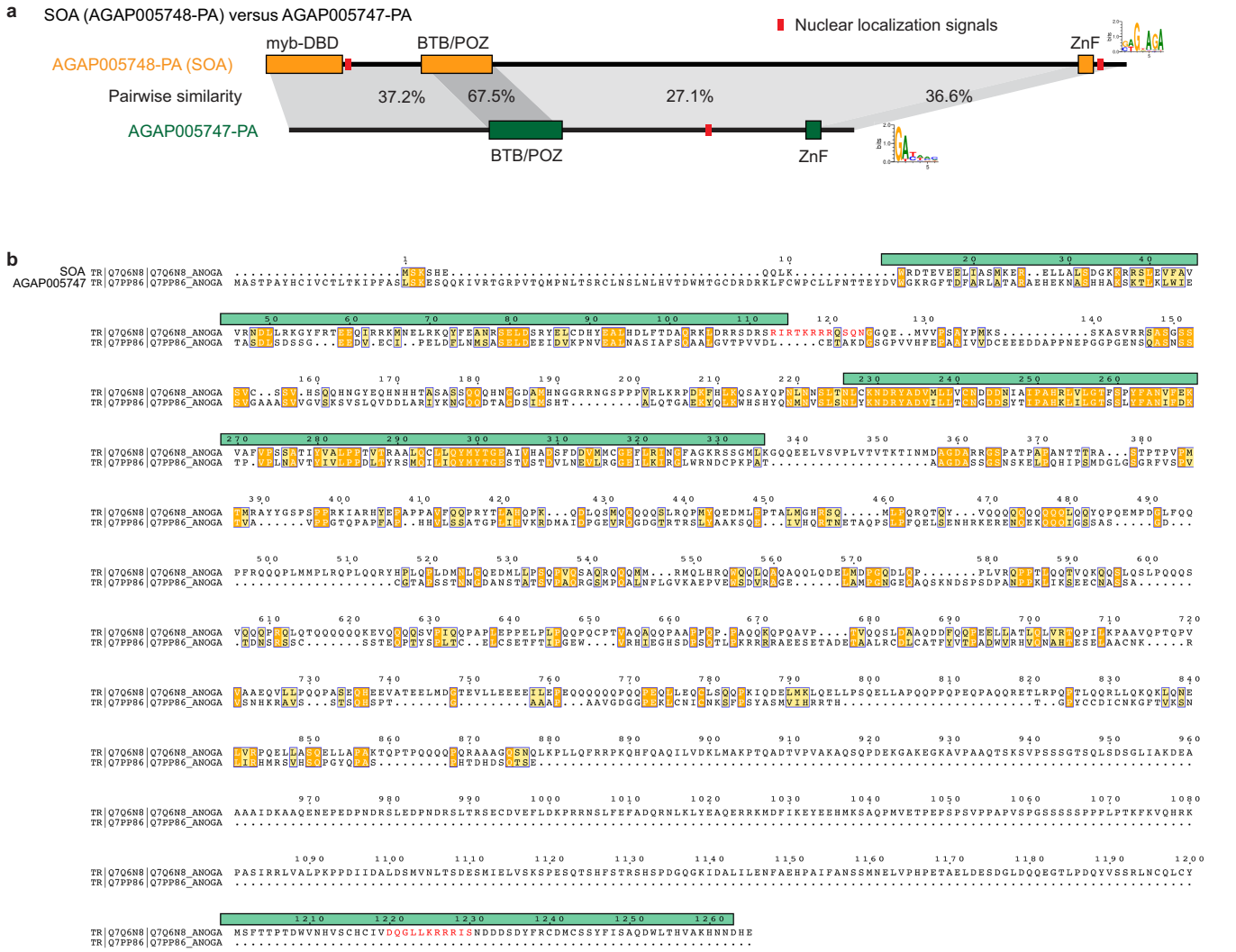


Extended Data Fig. 1 | See next page for caption.

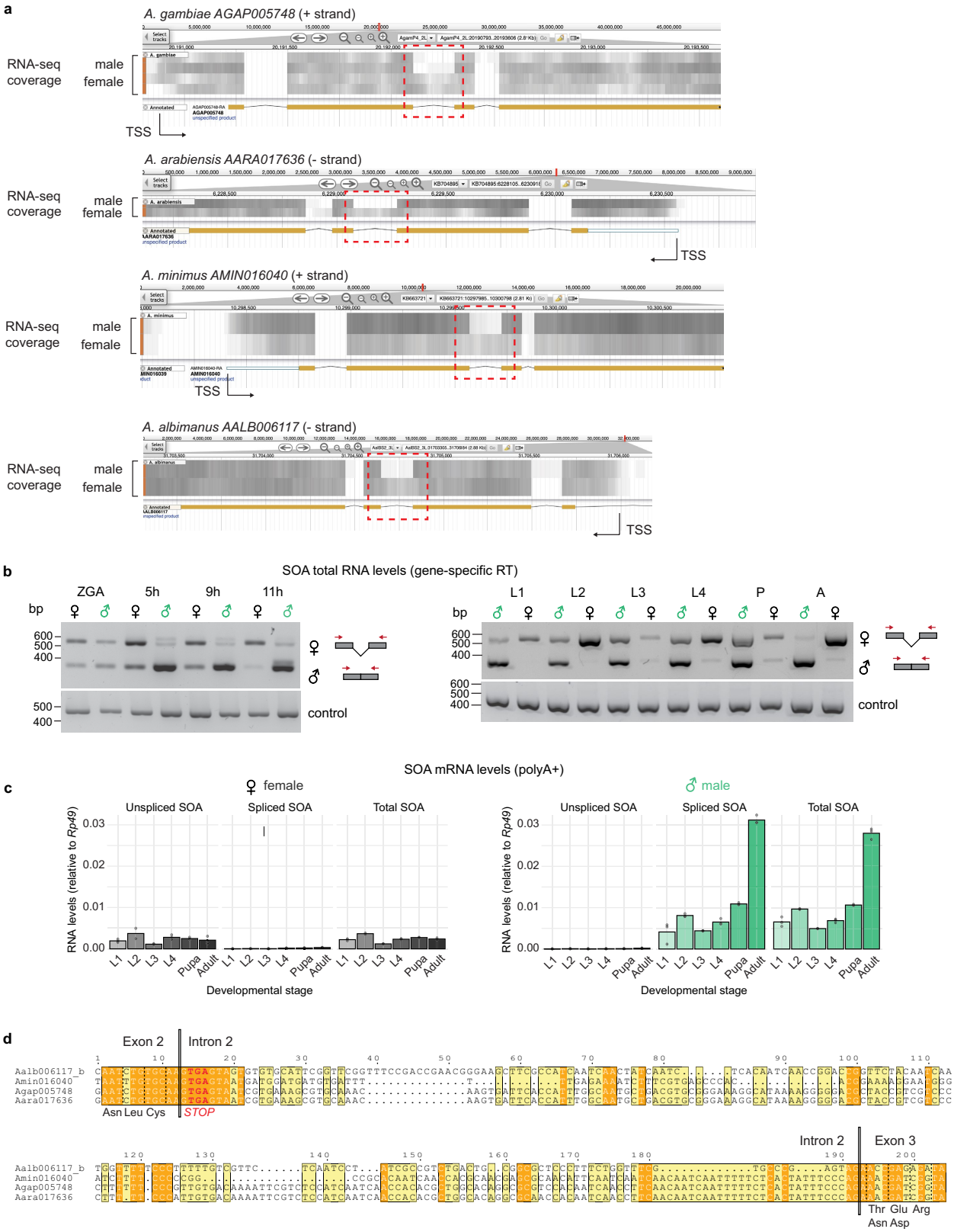
Extended Data Fig. 1 | Evolution of *SOA* by a tandem duplication in the *Anopheles* genus. (**a**) Upset plot showing the overlap between the male-biased differentially expressed (DE) genes obtained at the indicated timepoints in RNA-seq conducted from single male and female *A. gambiae* embryos at various hours (h) after egg laying (also see Fig. 1a). Differentially expressed genes between males and females were obtained with DESeq2. Only two genes were DE at several time-points from early to late embryogenesis, *AGAP005748* (*SOA*) and *AGAP029221* (Yob). **(b)** Maximum-likelihood tree of *SOA* (*AGAP005748*) orthologues and *SOA* paralogues (*AGAP005747*, with respective orthologues). The tree is based on the protein coding DNA sequences of the proteins, aligned with ClustalW in the MEGA 11 software and constructed with the Jones-Taylor-Thornton model. Based on these alignments, a maximum-likelihood tree was generated using default settings. The tree was rooted on the Culicinae outgroup branch. (**c**) Scheme regarding the evolution of *SOA* and its splicing in *Anopheles* genus after its separation from Culicinae. (Left:) Table indicating relevant characteristics for species spanning the *Anopheles* genus and *Aedes aegypti* as an outgroup with no heteromorphic sex chromosomes. (Right:) Schematic illustration of the protein domain architecture of SOA orthologues in the *Anopheles* genus, the conservation level is indicated by percent of identity and shades of respective structured domains. (**d**) Evolutionary tree of 5 representative mosquito species. Length of branches indicates separation of the Anopheline and Culicinae subfamilies based on molecular phylogeny. Additional information on the presence of the *SOA* gene (orange arrow), the presence of sex chromosomes (green dot) and DC (star symbol), as well as the percentage of SOA protein sequence identity (right) is included alongside the tree. (**e**) Synteny of the genomic regions surrounding *SOA* and its orthologues in *Anopheles* and *A. aegypti*. Data was obtained using the synteny tool from VectorBase. All *Anopheles* have both *SOA* and *SOA* paralogues, while *A. aegypti* only contains the paralogue and the *GST* gene in this region. Note that *AALB006117* is mis-annotated as a single gene. However, inspection of the RNA-seq data from[4] clearly reveals two distinct transcription units corresponding to *SOA* and *SOA* paralogue, respectively (also see Extended Data Fig. 4a). (**f**) Barplot showing *AGAP005747* (*SOA* paralogue) RNA levels from RNA-seq in transcript per million (TPM), overlaid data points represent values from biological replicates (single embryos). Raw datapoints and replicate numbers in Supplementary Table 3. No sex bias in expression of *AGAP005747* is observed. (**g**) Bar plot showing *SOA* mRNA levels normalized to *Rp49* in post-embryonic stages measured by RT-qPCR. Height of the bar plot is the mean of $n = 4$ independent experiments as overlaid individual data points. (**h**) Amino acid composition of four SOA orthologues in the *Anopheles* genus. Protein sequences were obtained from VectorBase.



Extended Data Fig. 2 | Sequence alignment of *SOA* orthologues among four *Anopheles* species. Alignment of full-length SOA protein sequences in *A. gambiae, A. arabiensis, A. minimus, A. albimanus* with Interpro domain architectures obtained in VectorBase. The alignment was generated in Clustal Omega and visualized with ESPript. Orange shaded residues are conserved in all 4, yellow shaded residues in 3 out of 4 species, respectively.

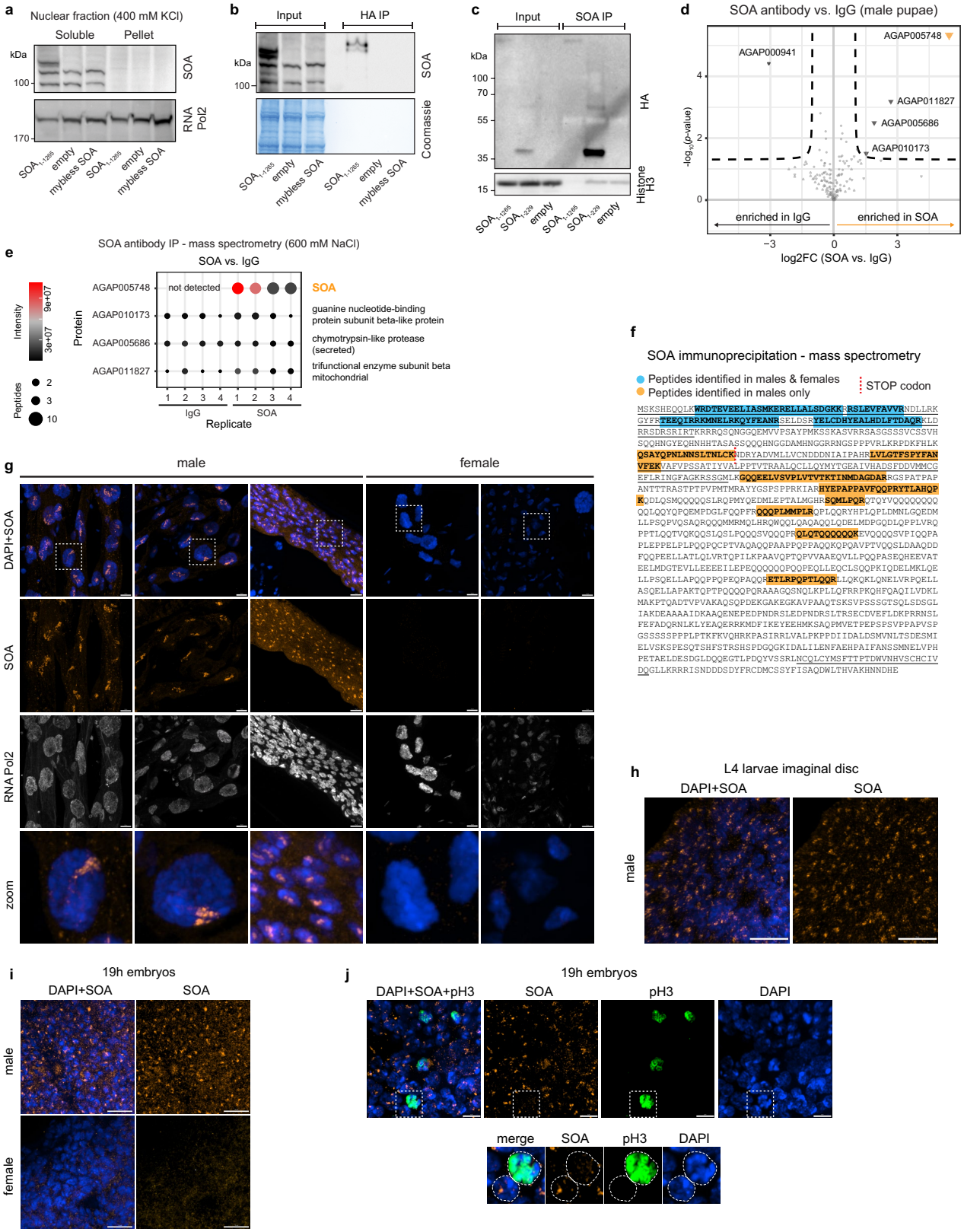


Extended Data Fig. 3 | Comparison of SOA protein with its paralogue. (**a**) Pairwise similarity between SOA (AGAP005748) and AGAP005747 proteins. Protein alignments were generated with Clustal Omega and pairwise similarity obtained in Jalview. Predicted nuclear localization signals are shown in red. The predicted sequence specificity of the C2H2-ZnF domains is shown as a motif logo. (**b**) Sequence alignment of SOA (AGAP005748) with SOA paralogue (AGAP005747) in *A. gambiae*. The alignment was generated in Clustal Omega and visualized with ESPript. Orange shaded residues are conserved, yellow shaded residues are similar, respectively. Green bars indicate the three structured domains of SOA (AGAP005748). Residues in red indicate the nuclear localization signals of SOA.



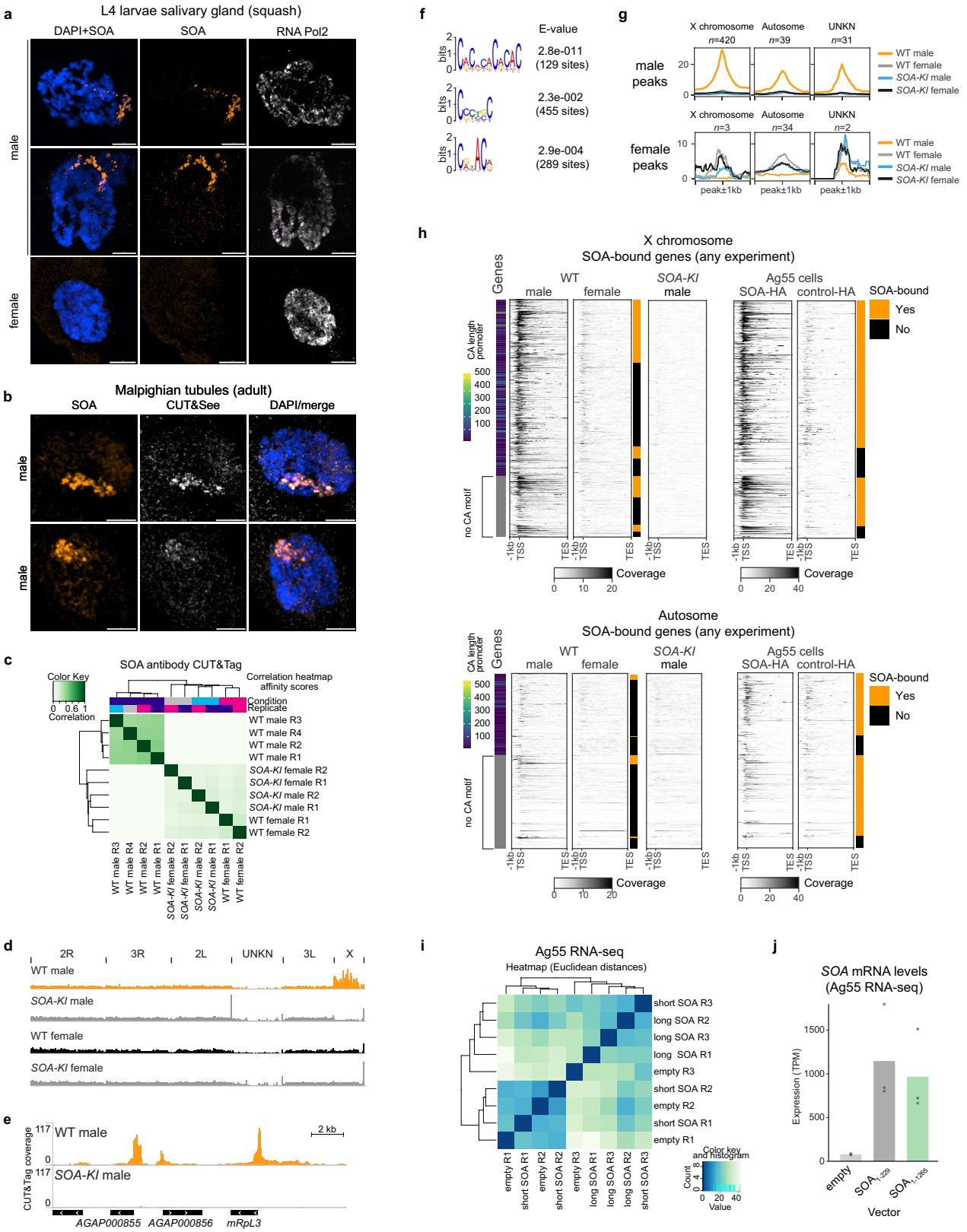
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Conservation of SOA alternative splicing in the *Anopheles* genus. (**a**) Genome browser snapshots of published RNA-seq data from adult male and female carcass[4] with RNA-seq coverage represented as density. The intron 2 is highlighted with a red box, indicating that sex-specific splicing of intron 2 of SOA orthologues in *Anopheles* genus is conserved. In *A. albimanus*, *SOA* and its paralogue are annotated as one long gene (*AALB006117*). However, inspection of the RNA-seq data clearly reveals two distinct transcription units with conserved alternative splicing in *SOA*. (**b**) Agarose gel showing RT-PCR products of the *SOA* intron 2 splicing in male and female (left:) embryos at zygotic genome activation (ZGA), 5h, 9h and 11h of embryogenesis or (right:) post-embryonic developmental stages: L1-L4 instar larvae, pupae (P), and adults (A). The reactions were conducted with a one-step RT-PCR kit, where reverse transcription is primed with the reverse primer in exon 3. The isoform with retained and excised intron 2 result in long and short RT-PCR products, respectively. S7 was used as a loading control. The experiment was conducted twice, results were confirmed with complementary methods: RNA-seq for embryogenesis and qPCR for post-embryonic stages (Fig. 1d, Extended Data Fig. 4c, Supplementary Table 1). (**c**) also see (Fig. 1e) qPCR quantification of polyadenylated SOA mRNA isoform levels in males and females. The barplot represents the mean levels of spliced, unspliced and total SOA relative to the *Rp49* reference gene. Overlaid data points represent the values of the biologically independent replicates, raw data is provided in Supplementary Table 1. (**d**) Alignment of pre-mRNAs of *SOA* (exon2-exon3) in four representative *Anopheles* species. Shaded nucleotides are conserved in all 4 species (orange) or 3 species (yellow), respectively.



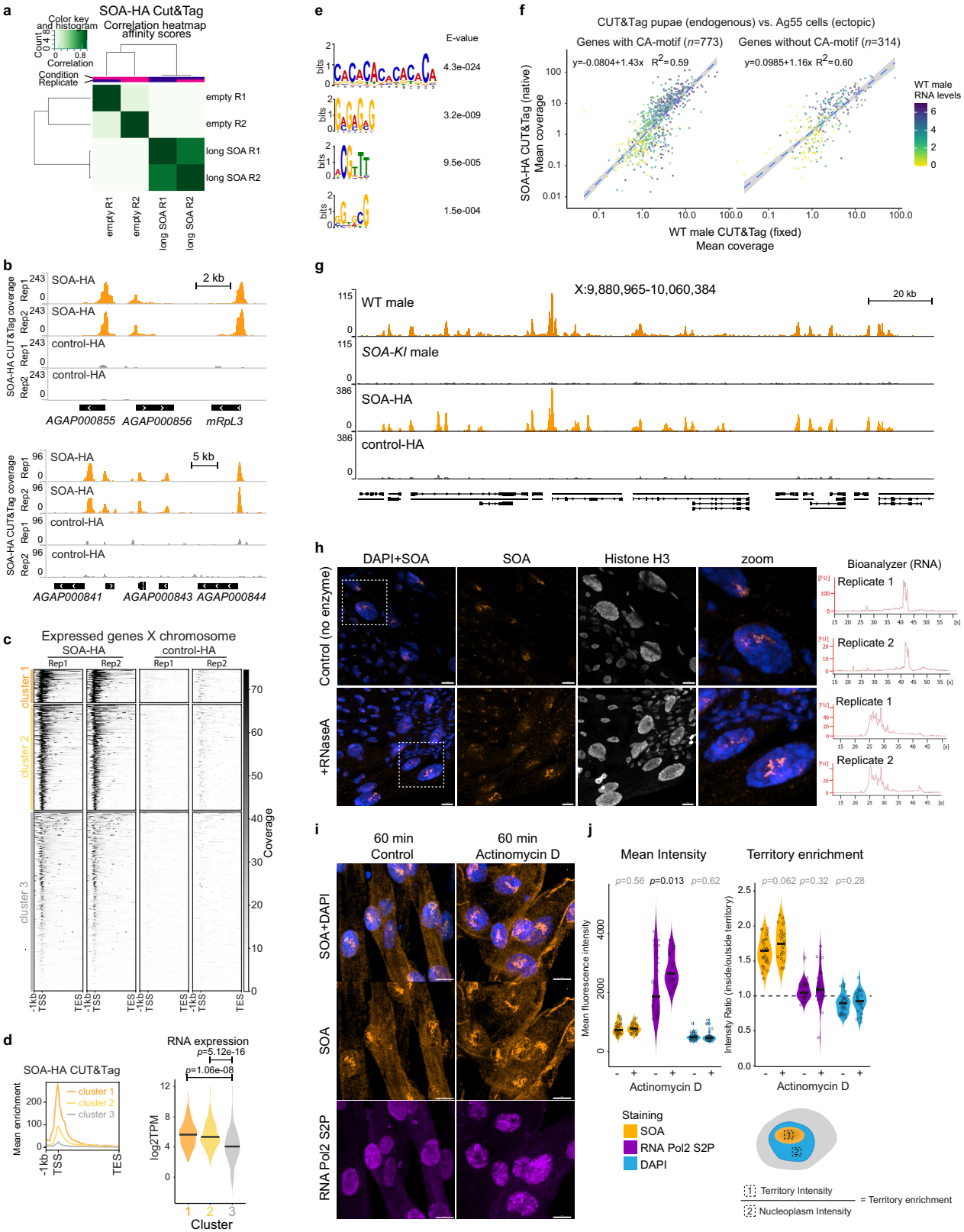
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Validation of the SOA antibody and SOA staining in embryonic and larval tissues. (**a**) Cropped immunoblot of ectopically expressed SOA and two negative controls. Nuclear soluble fraction extracted with 400 mM KCl was isolated from Ag55 cells expressing HA-tagged male SOA_{1-1265}, empty vector control, or mybless $SOA_{112-1265}$. Mybless SOA lacks the epitope (amino acids 1-112) used for immunization. RNA Polymerase 2 serves as a loading control. The experiment was repeated twice with similar results. (**b**) Cropped immunoblot of HA antibody immunoprecipitation (IP) with samples prepared as in (**a**). The SOA antibody was used for detecting the proteins immunoprecipitated by HA antibody, Coomassie serves as a loading and negative-IP control. The experiment was performed once. (**c**) Cropped immunoblot of SOA antibody IP with corresponding input samples. Chromatin extracted from Ag55 cells expressing HA-tagged male SOA_{1-1265}, female SOA_{1-229} or empty control were used. The HA antibody was used for detecting the proteins immunoprecipitated by SOA antibody, H3 antibody serves as a loading and negative-IP control. The experiment was performed once. (**d**) SOA IP-mass spectrometry experiment represented as a volcano plot, with log2 fold change (log2FC) between SOA and IgG on the *x*-axis and log*10* (*p*-value) on the *y*-axis. SOA (orange) and the 4 contaminant proteins (black) are highlighted in triangles, the remaining background noise proteins are shown in grey. IP was performed on nuclear extracts from male pupae using the SOA antibody ($n = 4$ biologically independent experiments) or IgG control ($n = 4$ biologically independent experiments). Raw data in Supplementary Table 1. (**e**) as in (**d**) Bubble plot representing the results of the SOA antibody IP-mass spectrometry experiment. The 4 significant proteins enriched in SOA versus IgG are shown in the plot. The color of the bubbles represents the measured intensity, and their size the number of unique detected peptides. SOA was the only protein not detected in IgG, while the other 3 were measured in both IPs. (**f**) Mass spectrometry was conducted on immunoprecipitated SOA from nuclear extracts of female and male pupae ($n = 5$ biologically independent experiments each). The panel shows the amino acid sequence of SOA, the peptides identified in male and female samples (blue shades) or in males only (orange shades). The position of the STOP codon is shown in red, the underlined amino acids correspond to the three structured domains. Raw data in Supplementary Table 1. Note that because SOA proteins were enriched via IP, this experiment cannot directly inform on the relative abundance of SOA protein isoforms in the sexes. Considering the mRNA quantification by qPCR (Extended Data Fig. 4c), SOA_{1-1265} and SOA_{1-229} proteins appear to be mutually exclusive in the two sexes and SOA_{1-1265} in males is at least 3-6 fold more abundant than SOA_{1-229} in females. (**g**) Representative pictures of SOA (orange) and RNA Polymerase 2 (grey) immunostaining conducted on male and female adult mosquito tissues with DAPI in blue. Pictures show Malpighian tubules or gut. The bottom shows a closeup (zoom) of the area highlighted with a white square. The pictures represent a 3D view of a z-stack, scale bar = 10 µm. This panel represents the complete panel of Fig. 1g, where a subset of the very same images (merged DAPI+SOA channels with close up) is presented. (**h**) Representative pictures of SOA immunostaining (orange) with DAPI (blue) conducted on imaginal discs from male L4 larvae. The pictures represent a 3D view of a z-stack. Scale bar = 10 µm. (**i**) Representative pictures of SOA immunostaining (orange) with DAPI (blue) conducted on male and female embryos at 19h after oviposition. The sexes were identified based on their clear differences in SOA staining. The pictures represent a 3D view of a z-stack. Scale bar = 10 µm. (**j**) as in (**i**) Representative pictures of SOA immunostaining (orange) with DAPI (blue) in a male embryo at 19h after oviposition. Mitotic cells were identified by a staining of phosphorylated Histone H3 (pH3, green). The bottom shows a closeup of the area in the white square highlighting two nuclei, where one undergoes mitosis, while the other one is in interphase. The SOA staining can only be detected in the latter. The pictures represent a 3D view of a z-stack. Scale bar = 10 µm.



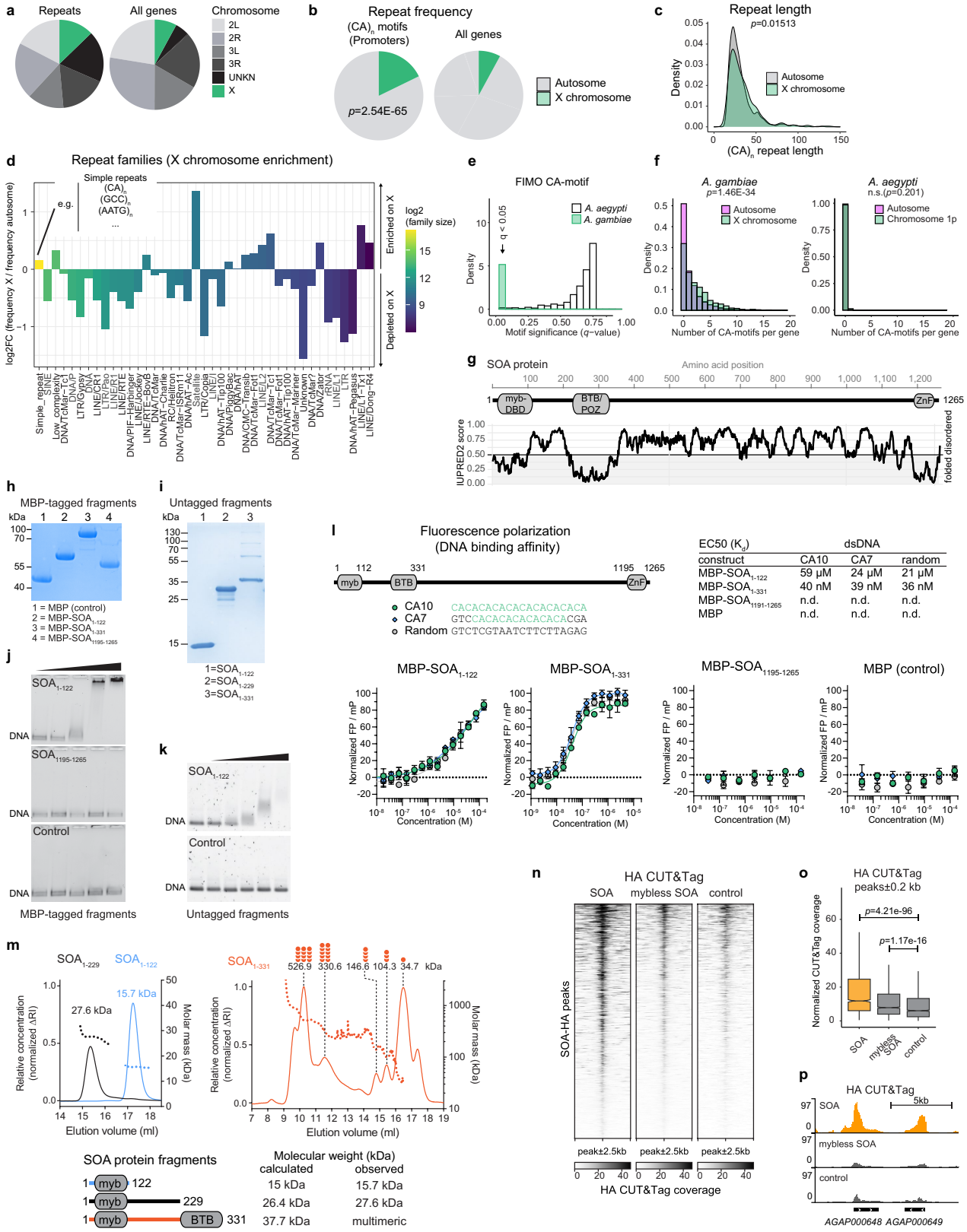
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | X chromosome binding and regulation by SOA. (**a**) Representative pictures of SOA (orange) and RNA Polymerase 2 (grey) immunostaining conducted on polytene squashes of salivary glands dissected from male and female L4 larvae. The pictures represent a 3D view of a z-stack. DAPI in blue, scale bar = 10 µm. (**b**) Pictures of CUT&See: SOA immunostaining (orange) combined with the visualization of SOA-targeted, pA-Tn5-mediated insertion of fluorescently labeled oligonucleotides (grey) conducted on wild-type male adult mosquito tissues. Pictures show Malpighian tubules, DAPI staining in blue. The pictures represent a 3D view of a z-stack. Scale bar = 10 µm. (**c**) Pearson correlation clustering of replicates based on affinity scores after peak calling of the SOA CUT&Tag data from pupae. The experiment was conducted with SOA antibody and IgG in wild-type (WT) male (n = 4 biological replicates) and female (n = 2), as well as homozygous *SOA* knock-in (*SOA-KI*) male (n = 2) and female (n = 2) pupae. The SOA antibody data was filtered using the IgG control and then subjected to clustering. (**d**) as in (**c**) Genome browser snapshot of the SOA CUT&Tag coverage on all chromosomal arms in the WT male and female as well as *SOA-KI* male and female genotypes. Duplicate reads were filtered out, replicates were merged for visualization. The enrichment is lost in the *SOA-KI* loss-of-function mutants. Note that the coverage in *SOA-KI* males is lower on the X due to copy number differences in comparison with XX females and autosomes. (**e**) as in (**c**) Genome browser snapshot of the SOA CUT&Tag coverage on a representative X-linked region in the WT and *SOA-KI* males. Replicates were merged for visualization. (**f**) MEME-ChIP motif analysis was conducted on all significant WT male CUT&Tag peaks. The position-weight matrix image of the three significant motifs (E-value ≤ 0.05) with obtained E-value from MEME is shown. (**g**) Metaplots showing the mean CUT&Tag enrichment at SOA peaks ± 1 kb identified with DiffBind (FDR<0.05) in a comparison of WT males vs. females. (top:) peaks enriched in males (fold>0); (bottom:) peaks enriched in females (fold<0). Each of the colored lines corresponds to a different genotype. The male peaks are specific, as the enrichment is lost in the *SOA-KI* loss-of-function mutant males. The female peaks do not vanish in the mutants and can be considered background. (**h**) Heatmap comparing the SOA CUT&Tag data from pupa with SOA-HA CUT&Tag data from cells. The analysis is focused on genes that have a significant peak called in any of the CUT&Tag experiments (Supplementary Table 2). X chromosomal genes are shown in the top heatmaps, autosomal genes at the bottom. For plotting the enrichments, the transcription start site (TSS) was used as a reference point with 1 kb upstream and gene bodies downstream scaled to 5 kb. To order the genes, they were sorted first according to the presence of CA-motif (Extended Data Fig. 7e) in their promoter as matched by FIMO. Then they were sorted based on their peak status (Yes/No, orange bars) in the Ag55 cells (SOA-HA) experiment and lastly based on peak status (Yes/No, orange bars) in pupae. For the genes that exhibit a CA-motif, a length heatmap indicating the total number of nucleotides that match the motif was created (left of the heatmap). The peak status associated with a gene (orange bar = Yes, black bar = No) was assigned based on the DiffBind (FDR < 0.05) output in a particular experiment. Due to differences in signal-to-noise the scale is different between pupae and Ag55 cells, but maintained in the top and bottom heatmaps, to be able to compare relative binding strengths between X and autosomes. The replicates were merged for visualization. (**i**) Euclidean distance heatmap obtained by DESeq2 representing the similarity of the samples in RNA-seq performed in female Ag55 cells that ectopically express male (long) $SOA_{1\text{-}1265}$, female (short) $SOA_{1\text{-}229}$ or empty vector control. (**j**) Bar plot showing the mean *SOA* mRNA levels from RNA-seq in transcript per million (TPM) with points showing the values of n = 3 biologically independent replicates.



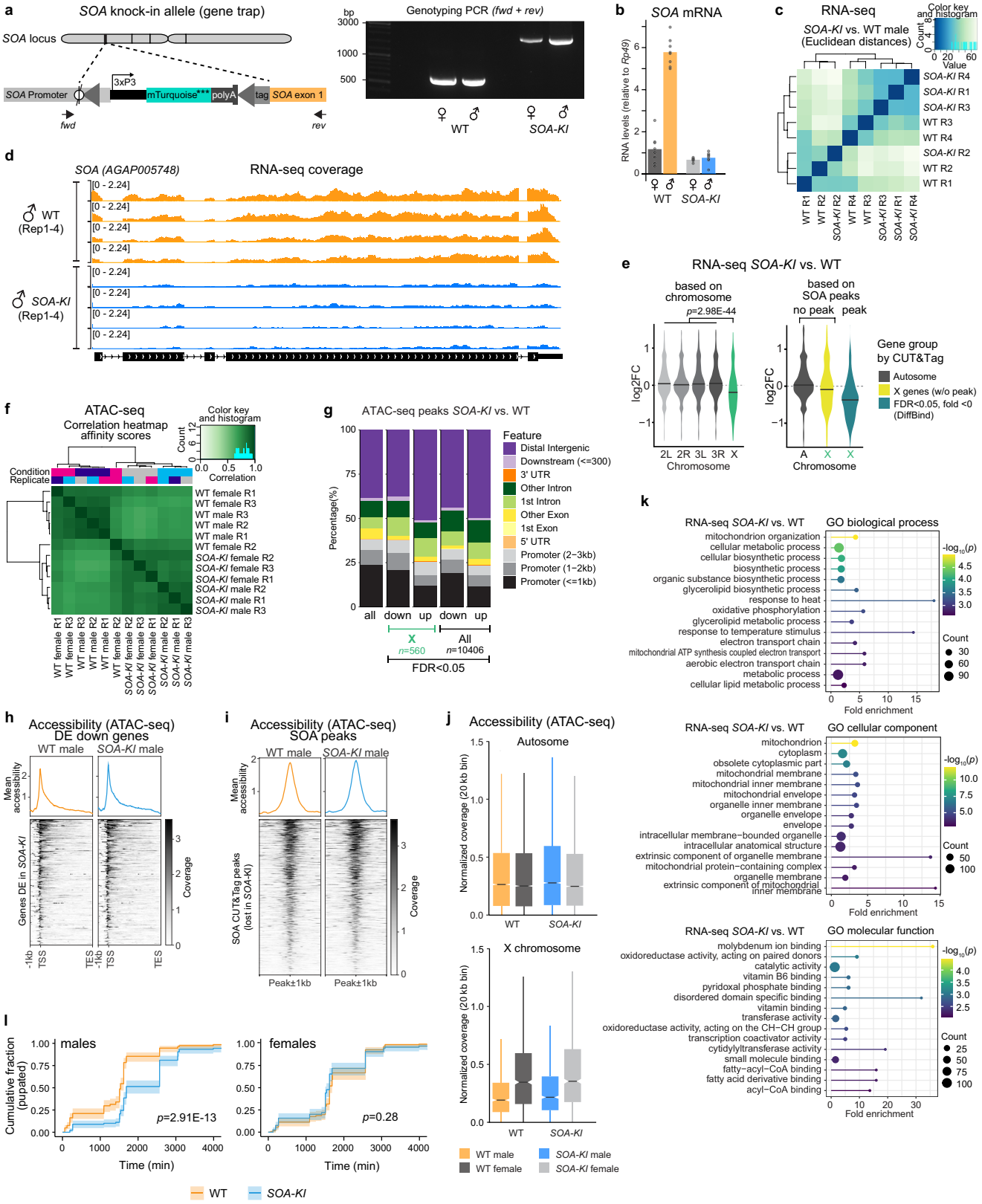
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Consequences of SOA expression in female Ag55 cells. (**a**) Pearson correlation of samples based on affinity scores after peak calling of the SOA CUT&Tag data from Ag55 cells infected with empty vector control or male (long) $SOA_{1\text{-}1265}$ baculovirus with the respective replicates. The experiment was conducted with HA antibody and IgG. The HA antibody data was filtered using the IgG control and then subjected to clustering. (**b**) Representative genome browser snapshots of the SOA-HA CUT&Tag coverage at two X-linked regions. (**c**) Heatmap showing the SOA-HA CUT&Tag enrichment with the TSS as reference point, 1 kb upstream and gene bodies scaled to 5 kb at expressed X-linked genes (≥10 average read counts in RNA-seq). 3 random *k*-means clusters were generated revealing three different groups with varying SOA binding strength. (**d**) as in (**c**), the mean enrichment in each of the 3 *k*-means clusters is shown as a metaplot. Replicates were merged for visualization. The bottom panel shows a violin plot with center line representing the median RNA expression in $\log_2$TPM from RNA-seq of Ag55 cells for each of the 3 clusters. The Bonferroni-corrected *p*-values were obtained with a two-sided Wilcoxon rank-sum with pairwise comparisons between the clusters. (**e**) MEME-ChIP motif analysis was conducted on all significant SOA-HA CUT&Tag peaks. The position-weight matrix image of the four significant motifs (*E*-value ≤ 0.05) with obtained *E*-value is shown. (**f**) Scatter plot showing the correlation between the mean CUT&Tag coverage in male pupae (*x*-axis) and mean CUT&Tag coverage of SOA-HA in in Ag55 cells. Each dot represents an expressed (≥10 average read counts) X chromosomal gene with the RNA levels $\log_2$(TPM+1) in WT males represented in color. The genes were further split based on the presence of a CA-motif in their promoters as assessed by a match in the FIMO search. The equation and R^2 value (coefficient of determination) of the fitted trend line was obtained by linear regression in R. (**g**) Representative genome browser snapshots of the SOA CUT&Tag data from pupae (WT and SOA-KI genotypes) and SOA-HA with empty vector control CUT&Tag data from Ag55 cells. The replicates were merged for visualization. (**h**) (left:) Representative pictures of SOA (orange) and H3 (grey) immunostaining conducted on male adult mosquito tissues after a 10 min treatment with buffer (control, top) or RNAseA (bottom). The pictures show Malpighian tubules. The right panel shows a closeup (zoom) of the area highlighted with a white square. The pictures represent a 3D view of a z-stack, DAPI in blue. Scale bar = 10 µm. (right:) Agilent bioanalyzer traces of RNA isolated from midguts ($n = 2$ biological replicates) that were undergoing the same treatment as the immunostaining. The control treatment samples show the characteristic doublet for insect rRNA, while in the RNaseA-treated sample degradation into smaller fragments can be observed. (**i**) Representative pictures of SOA immunostaining conducted on male adult mosquito tissues after treatment for 60 min with Control (top) or Actinomycin D (bottom). The treatment was conducted in L15 tissue culture medium followed by fixation. The pictures show Malpighian tubules with SOA in orange, Ser2-phosphorylated RNA Pol2 (RNA Pol2 S2P) in pink and DAPI in blue. The pictures represent a 3D view of a z-stack that was visualized with Imaris software. Scale bar = 10 µm. (**j**) Quantification of the staining in (**i**). The violin plots show the mean fluorescence intensity or territory enrichment. The territory was calculated by determining the ratio of the mean intensity in equally sized squares placed inside the territory and outside of the territory as visualized in the illustration on the right. The center line represents the median. SOA is represented in orange, RNA Pol2 S2P in pink and DAPI in blue. $n = 30$ nuclei were quantified for the control and $n = 26$ nuclei for the Actinomycin D treatment. *p*-values: two-sided Wilcoxon rank-sum for a comparison between control and Actinomycin D in each staining.



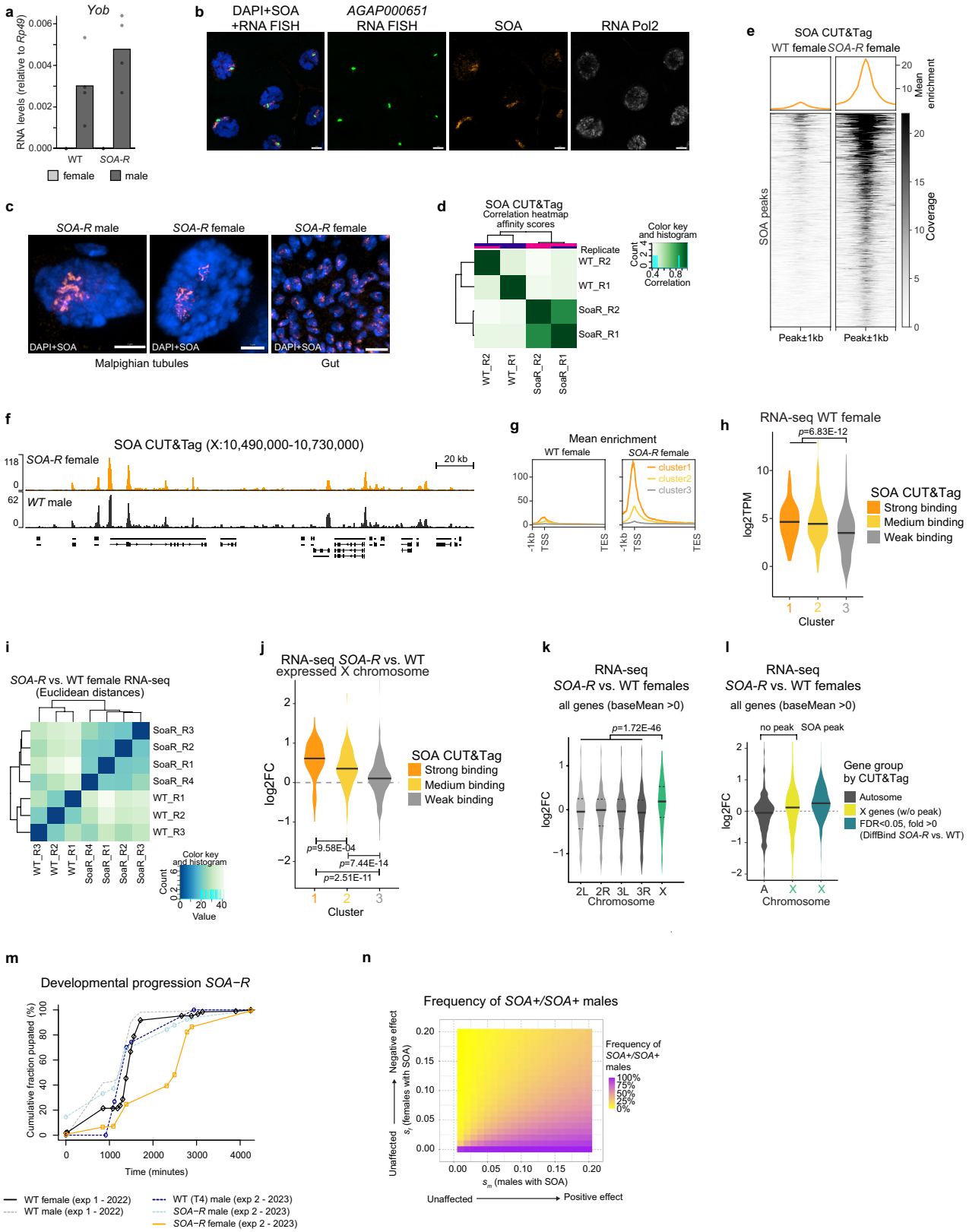
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Characterization of SOA X chromosome recruitment mechanism. (**a**) Pie chart indicating the number of repeats obtained by RepeatMasker on the X chromosome and other chromosomal locations (left) in comparison with the size of the respective regions in the genome (right). (**b**) Pie chart representing the number of (CA)n repeats localized at X-linked versus autosomal promoter region. The coordinates of the $(CA)_n$ simple repeats obtained from RepeatMasker were allocated to different feature classes (i.e. Promoter, intergenic, etc.) using the annnotatePeak function of ChIPseeker and the AgamP4.8.gtf annotation. *p*-value: one-sided Fisher's test for overrepresentation of X-linked genes containing $(CA)_n$ compared with the chromosomal localization of all *Anopheles* genes on X and autosomes, respectively. (**c**) as in (**b**) density distribution of the repeat length (difference between start and end coordinates) of the $(CA)_n$ motifs located at promoters on X and autosomes, respectively. *p*-value: two-sided Wilcoxon rank-sum test comparing X and autosomes. (**d**) The fraction of a given repeat class on the X chromosome was compared with the fraction of the same repeat class on autosomes. The log2 ratio fraction X/fraction A was obtained and shown as a barplot for the repeat classes, where the color indicates the family size, i.e. log2 overall number of the given repeat classes. Simple repeats (illustrated below the barplot), low complexity repeats, LINE/RTE–BovB and satellite are the top 4 (by family size) repeat classes enriched on the X. (**e**) Histogram showing the results of a 'Find Individual Motif Occurrences' (FIMO) search[54], in which the promoter regions of *A. aegypti* (control, no sex chromosomes) and *A. gambiae* were scanned for occurrences of the top scoring CA-motif (Extended Data Fig. 6f). The histogram shows the *q*-value of the obtained hits, which indicates the significance of the discovered loci to match the CA-motif used in the search. (**f**) FIMO motif searches as in (**e**). The histogram shows the number of motif hits found per gene promoter. The CA-motif tends to form clusters at X-linked promoters of *A. gambiae*, where often more than one motif per gene is present. Chromosome 1p in *A. aegypti* is homologous to the X of *A. gambiae*[19], but is not a differentiated sex chromosome. *p*-values: one-sided Fisher's exact test for overrepresentation of genes containing a FIMO-match on (left:) the X (*A. gambiae*) or (right:) chromosome 1p (*A. aegypti*). (**g**) Schematic illustration of the predicted domain architecture of SOA obtained on VectorBase from Interpro and intrinsically disordered scores from IUPRED2. (**h**) Coomassie stained gel of purified recombinant MBP and MBP-tagged SOA fragments. The purified fragments were used in the EMSA assay in (**j**) and fluorescence polarization assays (**l**). The SDS-PAGE was performed once to confirm the quality of the purified fragments. (**i**) Coomassie stained gel of purified recombinant N-terminal fragments of SOA without affinity tags. The purified fragments were used in the EMSA assay in (**k**) and Size exclusion multiangle light scattering (SEC-MALS, **m**). The SDS-PAGE was performed once to confirm the quality of the purified fragments. (**j**) EMSA assay of recombinant MBP-tagged myb-DNA binding domain ($SOA_{1\text{-}112}$), ZnF domain ($SOA_{1195\text{-}1265}$) and negative control protein (MBP). The protein amount in each lane was increased from 0 pmol (probe only) to 125-fold molar excess (12.5 pmol) over the probe (0.1 pmol). The probe was an equimolar mix of 300 bp-long X chromosomal promoter DNA sequences (sequences in Supplementary Table 1). After electrophoresis, the gel was stained with SYBRSafe. The experiment was performed twice with similar results. (**k**) EMSA assay of recombinant SOA myb-DNA binding domain. The protein amount in each lane was increased from 0 pmol (probe only) to 125-fold molar excess (12.5 pmol) over the probe (0.1 pmol). 147 bp 601-DNA sequence (Supplementary Table 1) was stained with SYBRSafe. GST protein was used as a negative control (bottom gel). The experiment was repeated three times with similar results. (**l**) Scheme and results of fluorescence polarization (FP) assay using Cy5-labeled DNA probes containing CA-motifs (CA10 - green circle, CA7 - blue diamond) or a random sequence (grey circle) that were incubated with various concentrations of MBP-$SOA_{1\text{-}122}$, MBP-$SOA_{1\text{-}331}$, or MBP-$SOA_{1195\text{-}1265}$. The mean relative FP values from three independent experiments including error bars indicating the standard deviation are shown over the indicated concentrations. Binding constants (K_d values) were determined by fitting a Michaelis–Menten non-linear regression to the relative FP values. The respective binding constants are given in the table (also see Supplementary Table 1). (**m**) (top:) Normalized differential refractive index (solid lines) and molar mass (dotted lines) from Size exclusion multiangle light scattering (SEC-MALS) for $SOA_{1\text{-}122}$, $SOA_{1\text{-}229}$ (left panel) and $SOA_{1\text{-}331}$ (right panel) with the elution volume by SEC displayed on the *x*-axis. The loading concentrations of the samples were 200 µM for the two short fragments (left) and 11 µM for the longest fragment (red) (bottom:) Schematic Illustration of the 3 purified SOA fragments analyzed by SEC-MALS. The calculated monomeric weight based on the protein sequence, as well as the observed weight-averaged molar mass are indicated. (**n**) Heatmap showing the normalized CUT&Tag coverage on all significant peaks (±2.5 kb) called in the $SOA_{1\text{-}1265}$ in comparison with empty vector control expressing Ag55 cells (Fig. 3e). The enrichment at these sites is shown for SOA-HA, SOA-HA lacking myb-domain (mybless) or empty vector control ($n = 2$ biologically independent replicates in all groups, merged for visualization). CUT&Tag was performed with HA antibody. (**o**) Box plot of the mean CUT&Tag enrichment of each peak ± 0.2 kb ($n = 1787$), as in (**n**) calculated with multiBigWigSummary with center line representing the median enrichment, box bottom, and top edges represent interquartile ranges (IQR, 0.25th to 0.75th quartile [Q1-Q3]), whiskers represent Q1 − 1.5*IQR (bottom), Q3 + 1.5*IQR (top). The Bonferroni-corrected *p*-values were obtained with a two-sided Wilcoxon rank-sum with pairwise comparisons between the groups. (**p**) Representative genome browser snapshots of the CUT&Tag data for SOA-HA, SOA-HA lacking myb-domain (mybless) or empty vector control ($n = 2$ biologically independent replicates in all groups, merged for visualization).



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Characterization of the *SOA-KI* mosquito transgenic line. (**a**) Schematic illustration of the *SOA* knock-in (*SOA-KI*) allele, in which the first exon of *SOA* and the coding sequence are interrupted by the eye and nervous system-specific 3xP3 promoter, mTurquoise coding sequence, a poly(A) site and a epitope tag. Two inverted loxP sites are illustrated by triangles, which were intended for marker cassette removal. The PhiC31 *att*P landing site is indicated with a circle. The position of the PCR screening primers is shown with arrows. The right panel shows a representative agarose gel of PCR products obtained in WT male, female or *SOA-KI* male, female homozygous transgenic line. (**b**) Bar plot showing *SOA* mRNA levels relative to *Rp49* in WT and homozygous *SOA-KI* pupae measured by RT-qPCR. Height of the bar plot is the mean of $n = 8$ biological replicates with overlaid individual data points. (**c**) Euclidean distance heatmap obtained by DESeq2 representing the similarity of the samples in RNA-seq conducted from WT and homozygous *SOA-KI* male pupae. (**d**) RNA-seq as in (**c**) Representative genome browser snapshot of the *SOA* locus with RNA-seq coverage for each of the $n = 4$ biological replicates. (**e**) RNA-seq as in (**c**) Violin plot with center line representing the median show the DESeq2-obtained log2FC in WT compared to homozygous *SOA-KI* mutant. Each gene with an average read count (baseMean) > 0 was taken into account, irrespective of whether it was scored as differentially expressed or not. (left:) The genes were grouped by chromosomal location. The *p*-value was obtained with a two-sided Wilcoxon rank-sum test comparing X (green) with all autosomes (grey). (right:) The genes were grouped by presence of a peak in CUT&Tag: All autosomal genes (grey), X-linked genes without peaks (yellow) and X-linked genes with a peak (blue) as scored by DiffBind (FDR <0.05, fold<0, *SOA-KI* versus WT) (Supplementary Table 2). Median log2FC values for each group are available in Supplementary Table 3. The Bonferroni-corrected *p*-values were obtained with a two-sided Wilcoxon rank-sum test comparing: autosomal versus X-linked genes without a SOA peak $p = 1.32E\text{-}10$; autosomal versus X-linked genes with a SOA peak $p = 1.02E\text{-}53$; X-linked genes without versus with a SOA peak $p = 1.03E\text{-}15$. (**f**) Heatmap showing the sample relatedness of the ATAC-seq replicates conducted from male WT and homozygous *SOA-KI* pupae based on Pearson correlation coefficient. (**g**) Barplot showing the % of ATAC-seq peaks in each of the genomic locations identified by ChIPseeker. (**h**) Heatmap of ATAC-seq coverage at each genomic region containing a SOA CUT&Tag peak scored as DiffBind in homozygous *SOA-KI* compared to WT males. The center of the peak is used as a reference point. The mean coverage is shown at the top of the heatmap. ATAC-seq replicates were merged for visualization by calculating the mean of normalized bigwigs using WiggleTools. (**i**) The accessibility of each gene with significantly decreased expression in homozygous *SOA-KI* (Fig. 4c) is visualized as a heatmap with the normalized ATAC-coverage using the TSS as reference point, 1kb upstream of the TSS and the scaled gene body (downstream of the TSS). The mean coverage is shown at the top of the heatmap. ATAC-seq replicates were merged for visualization by calculating the mean of normalized bigwigs using WiggleTools. (**j**) Box plot showing the normalized ATAC-seq coverage per 20 kb bin in the indicated chromosomal locations and genotypes. The line that divides the box into two parts represents the median, box bottom, and top edges represent interquartile ranges (IQR, 0.25th to 0.75th quartile [Q1-Q3]), whiskers represent Q1 − 1.5*IQR (bottom), Q3 + 1.5*IQR (top). The experiment was conducted in WT and homozygous *SOA-KI* pupae of both sexes ($n = 4$ biological replicates each). Note that accessibility of autosomes is equal between sexes and genotypes. Due to copy number differences, the expected 2-fold difference between males (XY) and females (XX) is observed on the X chromosome. Since this ratio is not substantially different from 2, we conclude that (regardless of chromosomal location and genotype) accessibility between males and females is highly similar. (**k**) Gene Ontology (Biological Process, top; Cellular Component, middletop; Molecular Function, bottom) analysis of the differentially expressed genes from RNA-seq from WT and homozygous *SOA-KI* male pupae. The lollipop plot shows the fold enrichment of genes in the various classes, with the point size indicative of the gene count and color indicative of the *p*-value. The analyses were conducted with the GO-Term tool on VectorBase. (**l**) 100 neonate larvae of each of the 4 scored genotypes (WT males, WT females, homozygous *SOA-KI* males, homozygous *SOA-KI* females) were seeded in the same culture for development through larval stages. The developmental timing of each of the 4 genotypes was scored by counting the appearance of pupae, which is represented as a cumulative distribution. The $t = 0$ of the x-axis represent the time when the first pupa appeared in the culture. The line represents the average of 4 replicates with shaded 95% confidence interval and *p*-value obtained by a log-rank test for stratified data (Mantel-Haenszel test). A second independent experiment with an additional 4 replicate cultures is presented in Fig. 4f.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Characterization of *SOA-R* mosquitoes and computational modelling for the spread of SOA. (**a**) Bar plot showing *Yob* mRNA levels relative to *Rp49* in WT and homozygous *SOA-R* pupae measured by RT-qPCR. *Yob* mRNA levels confirm the sex of the pupae used in Fig. 5b. The height of the bar plot is the mean of $n = 4$ biological replicates with overlaid individual data points. (**b**) Representative pictures of SOA immunostaining (orange), RNA Polymerase 2 immunostaining (grey) and co-RNA FISH (green) of a X-linked transcription site (*AGAP000651*) on salivary gland nuclei of a homozygous male *SOA-R* L4 larva. The RNA-FISH probes were designed against the introns of the *AGAP000651* gene. DAPI is shown in blue, scale bar = 10 µm. (**c**) Representative pictures of SOA immunostaining conducted on homozygous *SOA-R* male and female adult mosquito tissues. Pictures show nuclei of Malpighian tubules (left, scale bar = 5 µm) or gut (right, scale bar = 10 µm) with SOA in orange and DAPI in blue. The pictures represent a 3D view of a z-stack. Further images in Fig. 5c. (**d**) Pearson correlation clustering of SOA CUT&Tag samples based on affinity scores after peak calling. The experiment was conducted with SOA antibody and IgG in WT and homozygous *SOA-R* female pupae. The SOA antibody data was filtered using the IgG control and then subjected to clustering. (**e**) Heatmap showing the normalized CUT&Tag coverage on all significant peaks (FDR<0.05, fold-change >0) in *SOA-R* in comparison with WT female pupae. The mean enrichment is shown as a metaplot on top ($n = 2$ biological replicates, merged for visualization). (**f**) Genome browser snapshot of the SOA CUT&Tag enrichment obtained in *SOA-R* females in comparison with WT males on a representative region of the X-chromosome. Duplicate reads were filtered out and the replicates were merged for visualization. (**g**) CUT&Tag as in (**e**) Metaplot showing the mean CUT&Tag enrichment on expressed X-linked genes (≥10 average read counts), which were further grouped by unsupervised k-means clustering in 3 groups with strong, medium and weak SOA binding strength. The coverage was calculated using the TSS as a reference point with 1 kb upstream and the gene bodies downstream scaled to 5 kb. The replicates were merged for visualization. (**h**) Violin plot with center line representing the median RNA expression in log2 TPM (transcripts per million) from RNA-seq of WT females for each of the 3 clusters (based on binding in *SOA-R*, see (**g**)). *p*-value: two-sided Wilcoxon rank-sum comparing combined clusters 1 and 2 versus cluster 3. (**i**) Euclidean distance heatmap obtained by DESeq2 representing the similarity of the samples in RNA-seq conducted from WT ($n = 3$ biological replicates) and homozygous *SOA-R* ($n = 4$ biological replicates) female pupae. (**j**) RNA-seq as in (**i**) Violin plots showing the log2FC on expressed X-linked genes (≥10 average read counts), which were further grouped by unsupervised *k*-means clustering in 3 groups with strong, medium and weak SOA binding strength, see (**g**). The center line represents the median log2FC, which equals 0.613, 0.355, and 0.117 (strong, intermediate and weak binding) and corresponds to fold changes of 1.529, 1.279, and 1.084, respectively. (**k**) RNA-seq as in (**j**) but plotting the log2FC for all genes according to the chromosomal location in WT compared to homozygous *SOA-R* pupae as a violin plot. Each gene with an average read count (baseMean) > 0 was taken into account, irrespective of whether it was scored as DE or not. The Bonferroni-corrected *p*-value was obtained with a two-sided Wilcoxon rank-sum test comparing X with all autosomes. The center line represents the median (also see Supplementary Table 3). (**l**) RNA-seq as in (**i**) but plotting the log2FC distribution of autosomal (grey) and X-linked genes. The X-linked genes were split into two groups based on SOA binding in CUT&Tag (Supplementary Table 2). The yellow violin plot shows X chromosomal genes without SOA peaks, the blue violin plot shows peaks that were scored as differentially bound by DiffBind (*SOA-R* versus WT females, FDR<0.05, fold>0). Median log2FC values for each group are available in Supplementary Table 3. The Bonferroni-corrected *p*-values obtained with a two-sided Wilcoxon rank-sum test comparing all groups between each other are: autosomal versus X-linked genes without SOA peak $p = 6.57E\text{-}12$; autosomal versus X-linked genes with SOA peak $p = 1.45E\text{-}44$; X-linked genes without versus with SOA peak $p = 5.48E\text{-}06$. (**m**) A single culture of *SOA-R* (males + females) was conducted in parallel to WT males (T4 strain), cultured separately. For both, the developmental timing of each of the 3 genotypes was scored by counting the appearance of pupae. Pupa appearance is represented as a cumulative distribution with dots representing a given time-point when pupa numbers were scored. The $t = 0$ on the *x*-axis represents the time when the first pupa appeared in the culture. The data represents one experiment. For comparison, the mean WT male and female pupation timings scored in Fig. 4f (exp 1-2022) are plotted in the panel. A separate experiment with additional $n = 3$ independent replicate cultures for *SOA-R* grown together with WT is presented in Fig. 5g. (**n**) Checkerboard plot indicating the relative frequency of *SOA+/SOA+* males (colour-coded) after 10,000 generations of selection depending on the selection coefficients s_m in males (*x*-axis) and s_f in females (*y*-axis). Fitness is normalized to 1 in *SOA−/SOA−* males and females. Moreover, we assume that *SOA+* is dominant over *SOA−* in males and recessive in females. Hence, the fitness of *SOA+* bearing males is $1 + s_m$, while the fitness of *SOA+/SOA+* females is $1 - s_f$. Even if selection against *SOA+* in females is stronger than selection in favour of *SOA+* in males, *SOA+* is, for most parameter combinations, maintained in the population at considerable frequencies.

Chapter 3

Publication 2: Dosage compensation in non-model insects – progress and perspectives

3.1. Summary

Dosage compensation frequently evolves to address gene expression imbalances arising from sex chromosome differentiation. Insects are an excellent model for studying this process, due to the diversity of their sex chromosomes and their ecological significance.

While extensively studied in a few model organisms, the knowledge about dosage compensation in non-model insects was limited until recently. Because of the increased availability of genomics, many insect species have been tested for their dosage compensation status. In this review article, we surveyed the literature to find all insect species where dosage compensation has been studied. We summarized the findings of these studies to underline the similarities and differences between groups with different sex chromosome origins and ages. We also discuss how dosage compensation mechanisms can spread to newly sex-linked chromosomes. Additionally, we discuss the best practices and propose a framework researchers can apply to find the dosage compensation mechanism in their species of interest.

3.2. Candidate's contribution

I surveyed the published literature, summarized the collected information, wrote the initial draft, and prepared the figures.

Supervisor's signature: _____

Dosage compensation in non-model insects - progress and perspectives

Agata Izabela Kalita[1], Claudia Isabelle Keller Valsecchi[1]

[1]Institute of Molecular Biology, Mainz, Germany

Correspondence: c.keller@imb-mainz.de

Abstract

In many multicellular eukaryotes, heteromorphic sex chromosomes are responsible for determining the sexual characteristics and reproductive functions of individuals. Sex chromosomes can cause a dosage imbalance between sexes, which in some species is re-equilibrated by dosage compensation (DC). Recent genomic advancements have extended our understanding of DC mechanisms in insects beyond model organisms like *Drosophila melanogaster*. Here we review the current knowledge of insect DC, focusing on its conservation and divergence across orders, evolutionary dynamics on neo-sex chromosomes, and diversity of molecular mechanisms. We propose a framework to uncover DC regulators in non-model insects that relies on integrating evolutionary, genomic and functional approaches. This comprehensive approach will facilitate a deeper understanding of the evolution and essentiality of gene regulatory mechanisms.

Sex chromosomes and dosage compensation: An overview

Sex chromosome evolution leads to expression imbalance

Sexual reproduction is ubiquitous in animals and was present in the last eukaryotic common ancestor [1]. The most common mechanism of sex determination both in vertebrate and invertebrate species is genotypic [2]. In genotypic sex determination, the combination of sex chromosomes determines the organism's sex upon fertilization.

Genetic sex determination systems are categorized based on their morphology and which sex carries the **sex-limited chromosome** (see Glossary and Box 1). Degeneration of the sex-limited chromosome (Y or W) effectively leads to a situation where the **heterogametic sex** (XY or XO males; ZW or ZO females, see Glossary) has only one functional copy of the many genes located on the other sex chromosome (X or Z). As gene expression levels tend to scale with the gene copy number [3], this could potentially result in decreased gene expression compared to the ancestral state. Sex chromosomes originate from autosomes and thus contain genes that are important for both sexes, such as **housekeeping genes** or **haploinsufficient genes** (see Glossary). As such, the regulatory elements of these genes cannot freely evolve to return gene expression in the heterogametic sex back to ancestral levels, because this would result in unoptimal expression levels in the **homogametic sex** (XX females or ZZ males, see Glossary) [4]. If a sex chromosome-linked gene encodes a component of a multisubunit complex, its expression would not match expression of other subunits encoded on the autosomes. This in turn would disrupt the stoichiometry and the functioning of the whole complex [5,6]. Of note, this dosage imbalance does not occur in species with homomorphic sex chromosomes or different sex determination mechanisms (such as **haplodiploidy** or environmental sex determination, see Glossary).

Dosage compensation

The resulting gene dosage imbalance is often believed to be the driving force behind the evolution of **dosage compensation (DC,** see Glossary**)**, a cellular mechanism that equalizes the expression of sex chromosome-linked genes. Complete DC refers to a state where the expression of monoallelic, sex chromosomal genes is equal to their ancestral diploid expression i.e. before sex chromosomes have diverged. Because expression levels in ancestral proto-sex chromosomes can usually not be experimentally determined, complete DC is typically assessed by either comparing sex chromosomal gene expression to the same gene set in an outgroup species with undifferentiated sex chromosomes or, when this is not possible, comparing sex chromosomal gene expression to overall autosomal gene expression [7]. **Dosage balance** refers to the expression of sex chromosome-linked genes that is overall equal between the sexes, but - different from DC - not necessarily equal to autosomal expression (see Glossary and Figure 1, [8]).

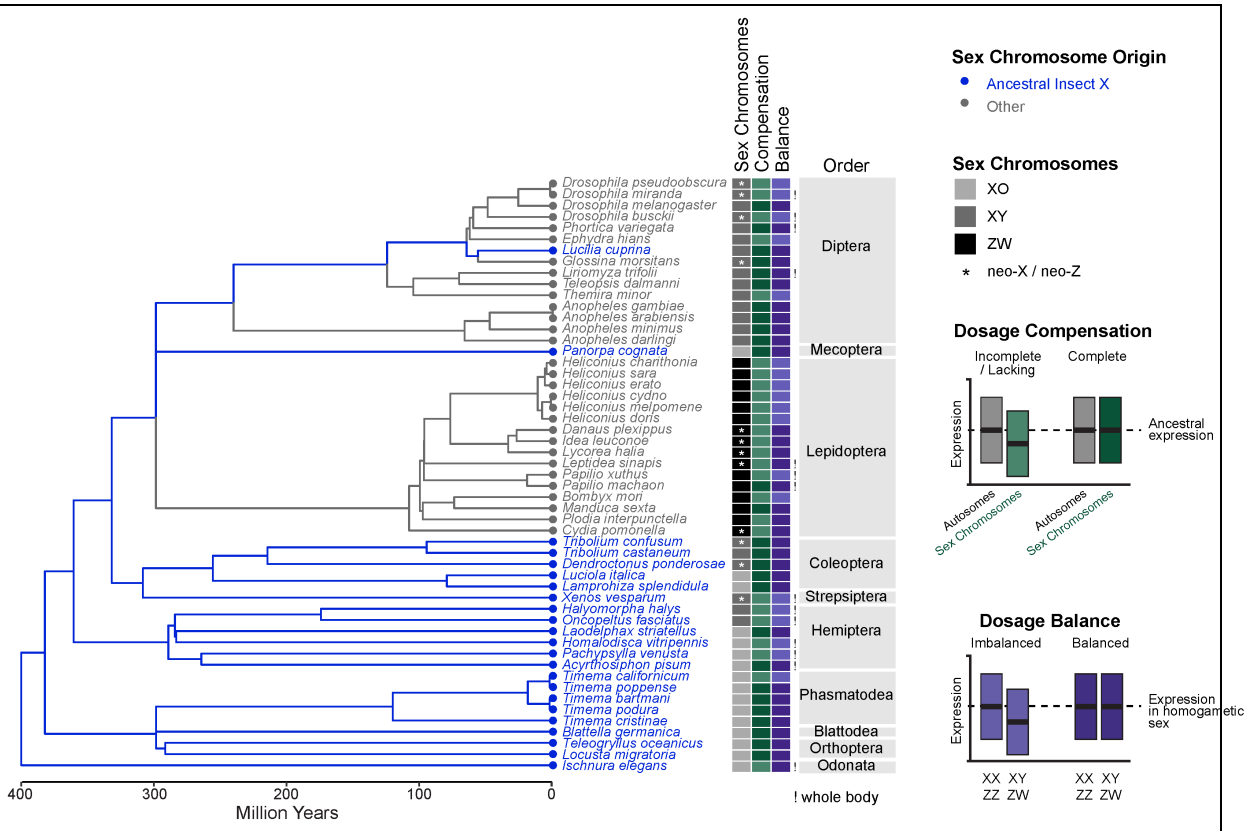


Figure 1 Phylogenetic tree of insect species where DC has been assessed.

The sex chromosome system, compensation, and balance status is shown beside the tree for each species. For species with neo-sex chromosomes formed by recent fusion events (indicated by an asterisk), the compensation status presented is based on the whole X chromosome. The exclamation marks indicate the species where the information is based solely on analyses from whole-body samples and could thus be cofounded by the gonad transcriptome. The diagrams on the right show explanations of the different classifications of DC and balance status. DC status is assessed by comparing expression in the heterogametic sex to ancestral expression (or autosomes, as an imperfect proxy). Dosage balance is assessed based on whether expression of sex chromosome-linked genes is equal on average between the sexes. Phylogenetic trees were obtained from TimeTree [116] and visualized using ggtree [117].

Additional information and references in Table 1. References supporting the ancestral origin of the X in Box 2.

The study of insects has been crucial for understanding the evolution of sex chromosomes and sex-specific gene regulatory mechanisms. Indeed, studies of DC in the dipteran model organism *Drosophila melanogaster* have contributed significantly to our knowledge of chromatin modifications and regulatory RNAs. The abundance of insect species and the diversity of their sex-determination systems make this group particularly interesting for investigating the evolution of sex-specific gene regulatory mechanisms beyond *Drosophila*. Insect sex chromosomes have both evolutionarily young and old sex origins. They can be exceptionally stable (the

ancestral insect X is conserved over 400 million years [9]), but also very dynamic (numerous transitions in Diptera [10]).

The increased availability and reduced price of genomic methods in the last 10 years have facilitated broader investigations across many insect orders, deepening our understanding of the DC phenomenon beyond *Drosophila.* In this review, we will summarize the current state of knowledge about DC in Insects. Many insect species have recently been assessed for their DC status, but the insights into new molecular mechanisms are still limited. Hence, we also propose approaches to discover new pathways and regulatory molecules.

Evolution of sex chromosomes in insects

Surprising "longevity" of the ancestral insect X chromosome

The ancestors of insects around 480 million years ago were likely male heterogametic and exhibited an XY/XX or XO/XX sex determination system [11,12]. Since then, insects radiated into approximately 5.5 million species with diverse karyotypes [13]. Recent analyses have revealed that multiple extant insect species spanning distant orders have an X chromosome of shared origin (see Figure 1, Box 2) [9,14]. These deeply conserved ancestral insect X chromosomes exhibit DC, which in some species is incomplete (see Table 1). The ancestral X persists in many hemimetabolous species, and might also be the origin of the X in beetles (Coleoptera, see Box 2) [9,15]. While the origin of the X seems conserved across beetle evolution, some species show **neo-sex chromosome** formation via fusion with autosomes [16] (see Glossary). Although Coleoptera is estimated to be the most species-rich insect order [17], the study of DC in this group is substantially underrepresented. Data on DC is only available for five Coleopteran species (see Figure 1).

The Z chromosome is conserved but prone to fusions in Lepidoptera

While the ancestral X persists in many insect lineages, in others it was replaced by a different sex chromosome system. For example, a new chromosomal element evolved into the Z chromosome in Lepidoptera and its sister lineage, Trichoptera. Recent efforts to track the karyotype evolution within Lepidoptera have revealed the conservation of a specific **Merian element** (see Glossary) as the sex chromosome [16,18,19]. The Z chromosome has undergone a higher number of fusion events than any of the autosomes in Lepidoptera karyotype evolution, leading to the formation of many neo-Z chromosomes. In most Lepidopteran species studied, the expression of Z-linked genes is characterized by dosage balance between the sexes and lacking or incomplete DC as compared to autosomes (type II according to Gu and Walters [8], Z=ZZ < AA, see Figure 1) [20].

The X chromosomes in Diptera evolved independently multiple times

While in Lepidoptera the Z chromosome formed before their radiation and is shared across the order, new sex chromosomes evolved independently multiple times in Diptera [10]. Based on the comparison to the sister lineage, Mecoptera, the ancestral dipteran X chromosome shrank and lost many genes [21]. The loss of gene content on the X preceded the radiation of dipterans and made the X chromosome prone to revert to an autosome. This allowed for subsequent evolution of new sex chromosome pairs from ancestral dipteran autosomes termed **Muller elements** (see Glossary) [22]. Even though the new sex chromosomes evolved independently and from different Muller elements, multiple dipterans exhibit expression patterns consistent with DC achieved via X upregulation in the males (see Figure 1) [10].

In summary, the insect sex chromosomes can display remarkable stability, but also undergo dynamic changes. In insect species with heteromorphic sex chromosomes, expression of sex-linked genes is at least partially equalized between the sexes.

DC mechanisms

Insights from Dipterans into ancestral X chromosome DC

In *D. melanogaster*, DC has been extensively studied and is achieved by MSL-complex mediated deposition of histone H4 lysine 16 acetylation on the X leading to increased gene expression (see Figure 2A, reviewed in [23]).

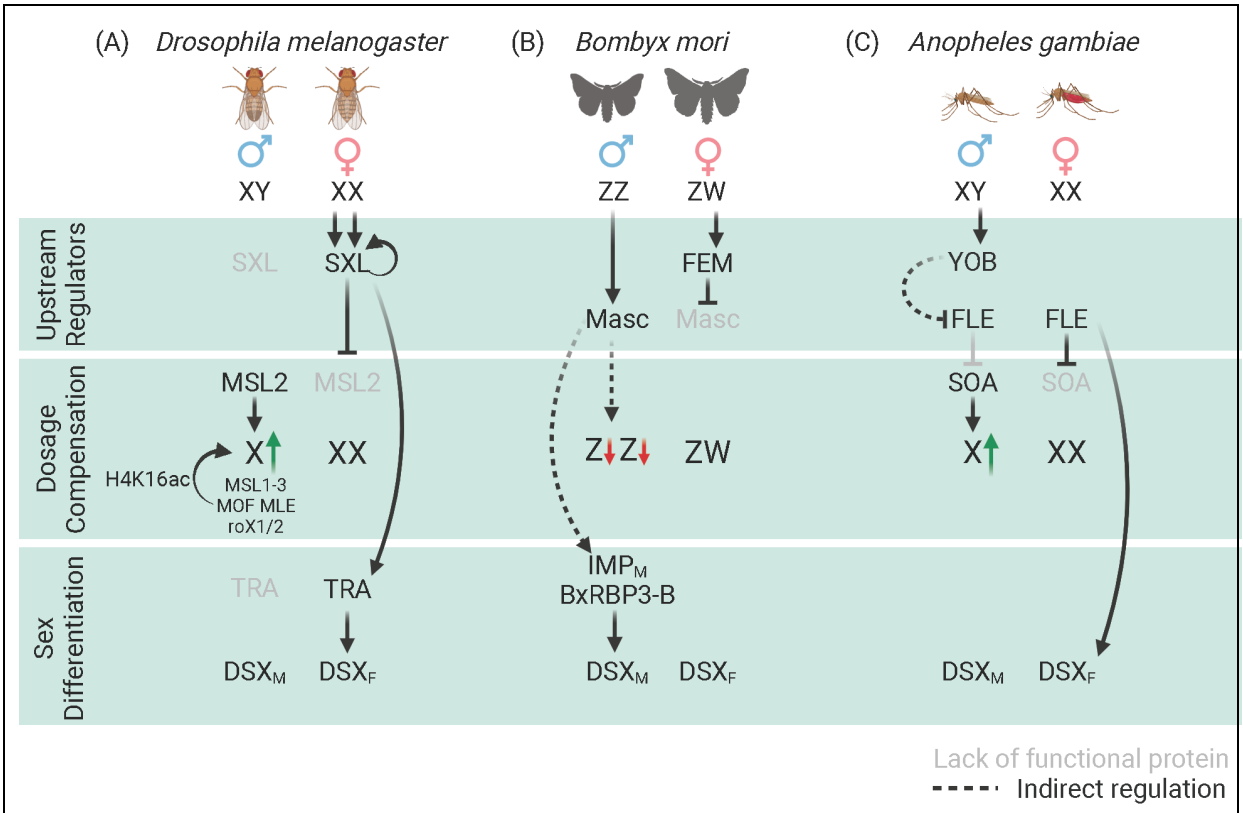


Figure 2. Insect sex determination cascades and their interplay with DC.

In most insects, DC and sex determination are regulated by the same upstream factors. The figure depicts three examples of species in which the sex determination cascade and its interaction with DC has been resolved in more detail. They represent systems where sex is determined by (A) a sex chromosome counting mechanism (B) dominant female-specifying factor (C) dominant male-specifying factor. The pathways for *B. mori* [118] and *A. gambiae* are described in the main text.

(A) *D. melanogaster* has been the insect model of choice that enabled the study of DC before the advent of genomics. In *D. melanogaster* the ratio of X chromosomes to autosomes is "transmitted" into SXL expression in the females (XX, X to autosome ratio 1:1) [66]. The master switch SXL initiates a cascade of female-specific expression and splicing patterns as well as the inhibition of MSL2 translation and export [119]. Thereby, MSL2 only accumulates in males. MSL2 in males targets the MSL complex to the single X chromosome by the interaction with two sex-specifically transcribed long non-coding RNAs, roX1/2 [120,121]. Another subunit of the MSL complex, the histone acetyltransferase MOF deposits an activating histone mark, H4K16 acetylation. This leads to approximately a twofold increase in gene expression of the single male X chromosome as compared to each of two female X chromosomes in *Drosophila* [23]. Created with BioRender.com

SXL - sex lethal, FEM - feminizer, FLE - femaleless; DSX - doublesex, TRA - transformer, MSL - male specific lethal; MOF - male on the first, MLE - maleless, SOA - sex chromosome activation, IMP - insulin-like growth factor II mRNA-binding protein.

The presence of H4K16ac on sex chromosomes has been assessed in species other than *D. melanogaster* (see paragraph "DC spreading to a new sex-linked

chromosomal arm"). In the following section, however, we focus on factors inducing differential transcriptional regulation on the sex chromosomes.

It may appear counterintuitive that the study of *D. melanogaster* can be informative for investigating the DC mechanism that controls the ancestral insect X, despite the evolution of a new set of sex chromosomes. In *D. melanogaster*, the ancient insect X has reverted to an autosome (chromosome 4, derived from Muller Element F; also known as the dot chromosome). Chromosome 4 has its own chromosome-wide regulation mechanism, likely retained as one of the "remnants of its former life as a sex chromosome" (quote from [22]). Chromosome 4 is specifically bound by the Painting of Fourth (POF) protein [24]. POF binding leads to decreased RNA polymerase II pausing and increased transcription rates [25]. *Pof* null mutants have decreased expression of genes on chromosome 4, and, unlike wild-type flies, they do not survive if haploid for chromosome 4 [26]. This suggests that POF can "compensate" the chromosome 4 linked genes. Another dipteran, the blowfly *Lucilia cuprina,* maintained the original F-element as the X chromosome. DC in *L. cuprina* is controlled by an ortholog of *Pof*: *no blokes* (*nbl*), as evidenced by decreased expression of X-linked genes in the *nbl* mutant males [27]. It is currently unclear if the mechanism of action is conserved between NBL and POF and whether this protein is involved in DC of the ancestral insect X chromosome in more evolutionarily distant species.

The Anopheles DC mechanism is different from Drosophila

The X chromosome of the *Anopheles* genus evolved independently from the *Drosophila* X, but from the same Muller element A [10]. The single X of male *Anopheles* is hypertranscribed and fully dosage compensated compared to females, autosomes and the proto-X [28–30]. Despite these similarities to *Drosophila*, neither H4K16ac nor the master regulator MSL2 are implicated in *Anopheles* DC [31]. Recently, the master regulator of *Anopheles gambiae* DC has been identified (*AGAP005748*; *SOA* [32], also named "*007*" [33]). *SOA* displays male-biased expression and sex-specific alternative splicing. Retention of intron 2 in females leads to a premature stop codon, preventing the production of a full-length protein. This sex-specific splicing, which is conserved among the *Anopheles* genus, may potentially be controlled by an RNA binding protein, Femaleless (*fle*) [32–34]. Males with loss-of-function *SOA* alleles exhibit reduced expression of X-linked genes. It is currently unclear what other regulatory molecules SOA interacts with and if it acts through histone modifications to achieve DC. Of note, MSL2 in *Drosophila* binds high affinity sites on the X, from which it uses a "spreading" mechanism to upregulate genes, but SOA, instead, binds promoters of X-linked genes directly.

Downregulation of both male Z chromosomes in the silkworm

In *Bombyx mori*, dosage balance is achieved by the downregulation of both Z chromosomes in the males, which is initiated by the expression of *Masculinizer* [35].

In females, *Masculinizer* is targeted for degradation by a W-linked piRNA *Feminizer* [36]. The role of *Masculinizer* in male development is conserved in many Lepidopteran species [37–43]. It is not known whether *Masc* directly binds the Z chromosome or is the upstream activator of such binding factor, nor what (if any) histone modifications are responsible for hypotranscription of Z in males. In another Lepidopteran, downregulation of the ancient Z does not involve modulation by H4K20me1, as one would perhaps expect based on the sex-chromosome dampening mechanism in *Caenorhabditis elegans [44]*. Identifying the molecular mechanism would be useful to disentangle the roles of *Masc* in sex differentiation and achieving dosage balance.

In summary, equalization of gene expression between the sexes is a common feature of regulation of heteromorphic sex chromosomes in insects. However, the mechanisms and regulatory molecules by which this goal is achieved differ between species with chromosomes of independent origin. Because of the limited number of known DC and dosage balancing mechanisms, further research is needed to reveal their full molecular diversity and to determine whether certain mechanisms are more prevalent or evolve repeatedly (see Outstanding Questions).

DC spreading to a new sex-linked chromosomal arm

X chromosome fusions in the Drosophila genus

The ancestral *Drosophila* X (**Muller element** A) has undergone multiple fusion events in this genus resulting in the formation of neo-X chromosomes of different ages and origins (see Figure 3). This allows for investigating how DC spreads to the newly sex-linked chromosome. In *D. miranda,* Muller element D fused to the X approximately 15-18 million years ago and has characteristics of a fully differentiated sex chromosome [45]. In contrast, additional Muller elements that started segregating with the X around 1 million years ago in *D. busckii* (element F) and *D. miranda* (element C) are still in the process of evolving into fully differentiated sex chromosomes [46]. These young neo-X and neo-Y chromosomes retain considerable sequence similarity, which would suggest that not all genes on neo-X necessarily require compensation.

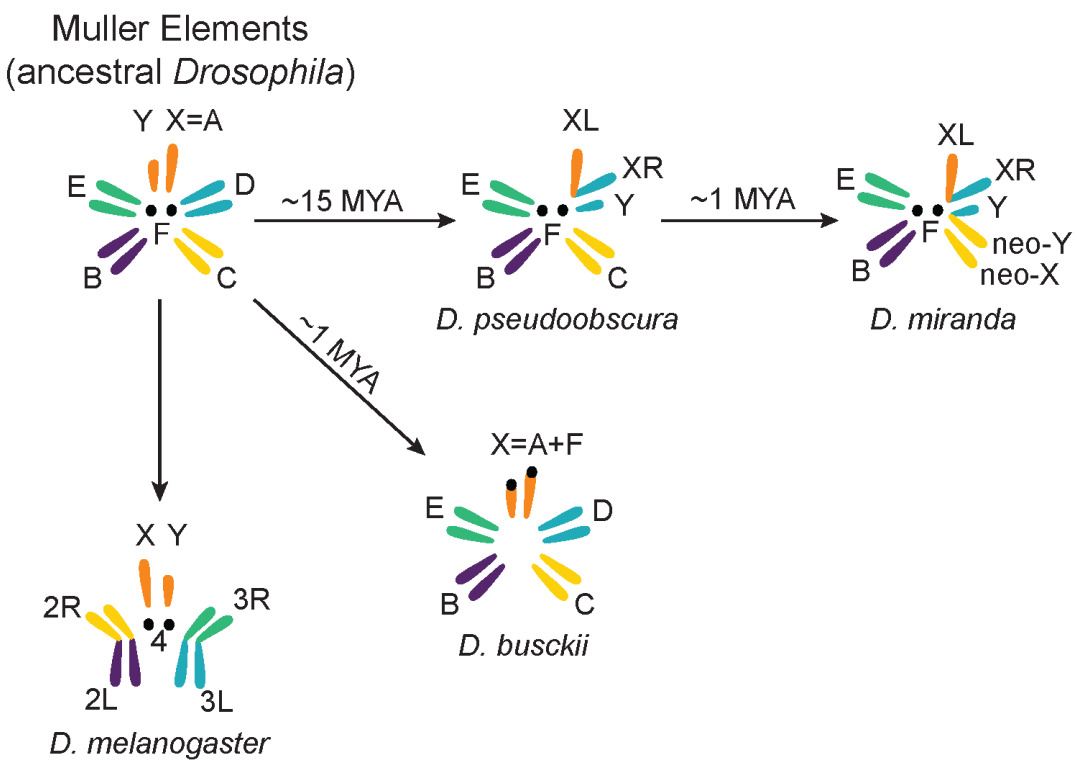


Figure 3. Karyotype evolution in Drosophila genus.

The figure depicts how chromosomal arms evolved via fusions in different *Drosophila* species. While in *Drosophila melanogaster* the fusions occurred between the autosomes, in *D. miranda, D. pseudoobscura,* and *D. busckii* the fusions involved the sex chromosomes, leading to formation of neo-X and neo-Y chromosomes. In *D. pseudoobscura*, the X chromosome fused to Muller element D. After the split of *D. pseudoobscura* and *D. miranda*, the Y chromosome of *D. miranda* additionally fused to Muller element C. This led to a formation of a neo-X chromosome that is not fused to the ancestral X, but segregates with it. In *D. busckii,* the neo-X chromosome formed by fusion to Muller element F, which was the ancestral dipteran X chromosome.

Already in 1940 Muller discovered that the gene content of chromosomal arms was conserved between *Drosophila* species and proposed that the ancestral *Drosophila* karyotype consisted of 5 chromosomal arms and the dot chromosome (termed Muller elements A-F). This discovery was later confirmed when many *Drosophila* genomes were sequenced. While inversions within the chromosomal arms may disrupt the order of the genes, the movement of genes between chromosomal arms is less common than intrachromosomal rearrangements. This nomenclature provided a useful framework for comparative studies revealing how chromosomes evolve over time (reviewed in [122]). Later the use of Muller elements was expanded to denote chromosomal arms in other dipteran species. Recently, similar syntenic elements were proposed for additional insect orders: Stevens elements for Coleoptera [16] and Merian elements for Lepidoptera [19].

In both *D. busckii* and *D. miranda*, the DC mechanism (MSL2 binding and H4K16ac) is conserved for the ancestral part of the X [47]. In *D. miranda,* it additionally binds and compensates the part originating from Muller element D (XR in Figure 3) [48].

For the more recent neo-X in *D. busckii* and *D. miranda,* the expression patterns are consistent with incomplete DC, however, the mechanisms differ between the two species. In *D. miranda*, the neo-X shows intermediate MSL2 binding and H4K16ac enrichment as it started evolving binding sites for the MSL complex via a combination of transposon activity and mutagenesis [49]. In contrast, H4K16ac enrichment was not detected on the neo-X of *D. busckii*. Observed partial DC implies a separate compensation mechanism for the neo-X. This could be a dedicated ancestral mechanism of Muller element F, mediated by POF, which coats the entire X in *D. busckii* [24]. It has been proposed that neo-Y genes that were already heterochromatic on ancestral Muller Element C are prone to downregulation, while the neo-X genes located in regions with previously active chromatin marks were most likely to be dosage compensated [50]. These analyses imply that the ancestral chromatin state of the proto-X/Y can influence the speed of the Y degeneration and DC evolution.

DC mechanisms can be co-opted or evolve anew on neo-Z chromosomes

Apart from the *Drosophila* genus, the spreading of balancing mechanisms on neo-sex chromosomes has also been studied in Coleoptera and Lepidoptera. Like the ancestral part of the X, the neo-X of *Tribolium confusum* (~20 MYA) is fully compensated [16]. In the Lepidopteran species *Cydia pomonella*, the neo-Z is partially downregulated to achieve dosage balance, similar to the ancestral part of the Z chromosome [51]. However, in three Nymphalidae species *(Idea leuconoe, Lycorea halia*, and *Danaus plexippus*), the neo-Z has evolved a more complete compensation status than the ancestral Z [44,52,53]. While the neo-Z of *I. leuconoe* and *D. plexippus* have a common origin (at least 26 MYA), the neo-Z of *L. halia* evolved via a fusion to a different autosome. In all three species, the expression of the ancestral part of the Z chromosome is balanced between the sexes but reduced compared to autosomes. The neo-Z chromosomes on the other hand have evolved almost complete DC, resulting in expression levels similar to the autosomes. In *D. plexippus*, the neo-Z chromosome in females is enriched for H4K16ac, an activating histone mark, whereas the ancestral part of the X is depleted for this mark in males [44]. Taken together, these results demonstrate that while the DC mechanism of the ancestral sex chromosome can be co-opted on the newly sex-linked chromosomal arm, it is also possible for a new, distinct DC mechanism to evolve.

While the neo-sex chromosomes can inform us about the spreading of DC mechanism to whole chromosomal arms, the mode of spreading at smaller distances can be inferred from insertions of transgenes on the sex chromosomes. Such transgenes are dosage compensated in *D. melanogaster* [54], but not in *L. cuprina* [55]. More specifically, in *Drosophila* the degree of DC is correlated with the proximity to the MSL high affinity sites [56,57]. These results suggest different spreading modes at smaller distances: sequence-independent in *D. melanogaster* and sequence-dependent in *L. cuprina*. Transgenic toolkits in species beyond *D. melanogaster* remain underdeveloped, which limits our understanding of these spreading phenomena in other insects.

Essentiality of DC

DC perturbation is detrimental but not always lethal

DC is essential for the heterogametic sex in *D. melanogaster*, *Mus musculus,* while the balancing mechanism in *C. elegans* and *Bombyx mori* is essential in the homogametic sex. In contrast, loss of DC is not lethal in *A. gambiae* and some evolutionary lineages lack chromosome-wide DC. Thus, the phenotypes of DC misregulation in different insects appear perplexing at first sight and deserve a more careful discussion of their potential causes.

The embryonic onset of the balancing mechanism has been demonstrated in *B. mori* [58], *D. melanogaster* [59,60], and *A. gambiae* [32]. Despite that, the developmental stages at which the phenotype of their misregulation manifests differ between species. The disturbance of dosage balance is embryonic lethal in *B. mori* [36]. In dipterans the phenotype is observed at later stages: a male-specific developmental delay in *Anopheles* larvae [32,33], and lethality of *Drosophila* males at the late larvae stage [61,62]. In *L. cuprina*, males die at the pupal stage. In this case, this phenotype only occurs if the mother is also homozygous for *nbl* mutation, implying that the maternally deposited protein or transcript is sufficient to allow normal male development. Besides these loss-of-function phenotypes, there are also insights on what happens when DC is ectopically induced in the other sex. *A. gambiae* females expressing the male *SOA* isoform show a developmental delay at larval stages [32,33]. In contrast, MSL2 expressing *Drosophila* females present with a reduction in viability scored at adult stage, with the penetrance depending on the allele used and the gene product levels [63–65].

Taken together these observations suggest that dosage balance is important already in embryogenesis, but the phenotypic consequences of its misregulation may manifest later in development.

Phenotypes of sex determination mutants and their connection to DC

The primary sex determination factors in insects commonly also regulate DC (see Figure 2 for their interactions). In *D. melanogaster* the phenotypes of *Sxl* mutants are broadly consistent with the *msl-2* mutants. Constitutive *Sxl* expression is lethal for males [66], and the lethality can be rescued by expressing *msl-2* ectopically [67]. Consistently, a loss-of-function *Sxl* mutation is lethal for females [66]. Because of this compelling evidence, the lethality observed in other insects upon disturbing the sex determination factors upstream of DC has often been assumed to be caused by DC misregulation. However, more recent studies in non-model insects challenged this view indicating that sex chromosome imbalance-related phenotypes can be variable and often difficult to disentangle from roles in sexual differentiation.

In *Anopheles* mosquitoes, the sex is determined by a dominant Y-linked factor (see Figure 2C, [68,69]). Expression of this factor in embryos leads to an upregulation of X-linked genes and complete [68–70] or partial [71] female-specific lethality. The same phenotype is observed in female embryos upon depletion of another component of the sex determination cascade, *femaleless* [34]. The observed lethality phenotypes were assumed to be the result of inducing aberrant X upregulation. However, increased expression of X-linked genes resulting from expression of full SOA isoform did not reproduce the same phenotype [32].

Masc mutation leads to increased expression from Z chromosomes and embryonic lethality in *B. mori* males (see Figure 2B, [36]). Conversely, the expression of *Masc* transcript that is resistant to *Feminizer*-derived piRNAs results in lethality of *B. mori* females at the larva/pupa stage [72]. The role of *Masc* orthologs in regulating male development has been confirmed in multiple Lepidopteran species [37–42,73]. Its ability to regulate dosage balance has not been tested directly outside of *B. mori*, but has been implied based on a variety of phenotypes. The difference between complete lethality in *B. mori* and phenotypes in other Lepidoptera could be biological, but could also be explained by different *Masc* depletion strategies, efficiencies, and life stages. For example, *Masc* knockdowns of comparable efficiencies cause partial male lethality when induced in embryos of *C. pomonella* [42] or *Ephestia kuehniella* [40], but result in a developmental delay and decreased body size of pupa when induced at the third instar larva stage of *Helicoverpa armigera* [39]. More recently, CRISPR-mediated targeting approaches of *Masc* have been reported. In the G0 generation, which likely consists of chimeric or heterozygous individuals, phenotypes ranging from sterility in *Agrotis ipsilon* [37] to sterility combined with partial male lethality in *Ostrinia furnacalis* were observed [41]

Lastly, we note that DC mechanisms can be targeted by endosymbionts such as *Spiroplasma* or *Wolbachia* to induce male-killing. For example, *Wolbachia* produces a protein Oscar that inhibits *Masc* accumulation [74,75], while Spiroplasma-produced protein *Sp*AID targets and damages the X chromosome by associating with the Drosophila DC complex [76,77]. Hence, studying their mode of interfering with their insect host might reveal new insights into DC [77,78].

In conclusion, a careful investigation of the links between sex determination pathways and DC is needed in insects. Although genetics in non-models is not as straightforward as in *D. melanogaster*, one possibility is to assess the phenotypic outcome of sex determination factor expression (e.g. *A. gambiae Yob*) when the DC machinery is not functional (e.g. in a *SOA* mutant).

Why is lack of DC detrimental?

Two main theories aim to explain the detrimental effects of disrupting DC [79]. The first theory claims that the phenotype is an effect of the combined misregulation of many genes. In this theory, the strength of the phenotype scales with the number of genes on the X and the degree of X chromosome misregulation. The difference in the

number of X-linked genes could partially contribute to the strength of phenotypes: in *Drosophila*, the X carries approximately twice as many genes as in *Anopheles*. This cannot however explain the difference between e.g. *A. gambiae* and *L. cuprina*. The misregulation of the X in *A. gambiae* (approximately 1100 genes) leads to a developmental delay, while disrupted expression of fewer X-linked genes (around 300) in *L. cuprina* results in lethality at the pupal stage. The second theory claims that lethality is a result of the disrupted expression of a few crucial genes (e.g. housekeeping or hapoinsufficient genes). The *Anopheles* X carries a lower fraction of housekeeping genes than autosomes, while in *Drosophila* housekeeping genes are present on the X at the same frequency as on autosomes [80].

We suggest that the two aforementioned theories are complementary rather than exclusive. The combined effect of decreased expression of many sex chromosome-linked genes and insufficient expression of a few haploinsufficient genes could be the cause of fitness defects and even lethality for the DC mutants. Further research into the question of essentiality of DC will require characterizing additional mechanisms. Since most of our knowledge about DC in insects comes from *D. melanogaster*, it is tempting but inadvisable to generalize these insights to insects as a whole. As shown by the example of *Anopheles*, we need to be careful when assuming the causative role of sex chromosome misregulation in the lethality observed in mutants of the sex determination cascade.

Considerations and methodological framework to study DC

With more studies published on DC across different insect species, adoption of common practices and analytical approaches can improve the comparability of the results between studies. One of the challenges in studying DC is its variability across tissues, especially between somatic cells and the germline. When gonads are analyzed separately, their expression patterns are compatible with the lack of DC (or meiotic sex chromosome inactivation) in the heterogametic sex (see Table 1). Because the gonads can constitute a significant part of the adult insect body, the imbalance in expression of sex chromosome-linked genes between males and females is often more pronounced in whole body samples than in somatic tissues [7]. For example, most studies that use whole-body samples of hemimetabolous insects result in DC assessed as incomplete [81–83]. On the other hand, when somatic samples are analyzed separately from the gonads, complete DC is present in somatic tissues, but absent in the gonads [15,84–86]. The lack of DC is also intertwined with the observed sex-biased expression in the gonads. Sex-biased genes are nonrandomly distributed between the autosomes and sex chromosomes. Although not universal, the X in insects is often enriched for genes with female-biased expression and/or depleted in genes with male-biased expression. Conversely, the Z is often enriched for genes with male-biased expression (see Table 1).

Sexual dimorphism and the extent of DC can also vary between different somatic tissues [52,84,85], life stages [29], and depend on gene expression levels. Applying

an absolute expression threshold to exclude lowly expressed genes tends to skew the results towards complete DC [80,87]. To avoid this bias, it is advisable to test multiple thresholds and explore whether the results change with more stringent criteria. An alternative, more robust approach is to apply a percentile cutoff (e.g. analyze the 80% most highly expressed X-linked genes rather than genes above 1 FPKM) [80].

Taken together, the sex-bias, tissue specificity of DC and filtering strategy can confound the analysis. As such, the future studies should analyze somatic tissues separately from the gonads whenever possible. Ideally, tissues with limited sexual dimorphism should be used to limit the influence of the aforementioned confounding variables on the results. If only whole body samples are available, exclusion of strongly sex-biased genes can improve the DC analysis [88]. Lastly, whenever possible, outgroup species with undifferentiated sex chromosomes should be included to estimate the ancestral gene expression levels.

DC mechanisms and how to find them

While genomic tools made it possible to study the extent of DC in a range of insect species, this has not yet led to the discovery of many new DC mechanisms. The identification of DC mechanisms in model organisms like *D. melanogaster* and *C. elegans* relied on extensive genetic resources: balancer chromosomes, isogenic lines, chromosome markers and the ability to perform thousands of crosses [61,89]. These mutagenesis screens also scored a strong phenotype: sex-specific lethality. Applying similar high throughput strategies to non-model organisms is challenging, even with the availability of genome engineering tools such as CRISPR. Additionally, they rely on an assumption that DC is essential in the screened species, whereas milder phenotypes such as the one observed in *Anopheles* [32,33] are not amenable to screening. We thus suggest alternative strategies to discover new DC mechanisms (see Box 3).

Recent models propose that DC may contribute to the degeneration of the sex-limited chromosome, rather than simply evolving as a consequence of degeneration [90,91]. This would imply that species with a common origin of sex chromosomes also share similar mechanisms of DC. For example, core components of the MSL complex have maintained their function within the *Drosophila* genus, but not in *Anopheles*, where the same Muller element evolved into the X independently [10]. Hence, understanding the evolutionary history of sex chromosomes is crucial. Recent efforts to track the karyotype evolution within Lepidoptera and Coleoptera have revealed a conservation of a specific Merian or **Stevens element** (see Glossary) as the sex chromosome in these orders [16,18,19].

As the DC mechanisms likely evolve alongside the sex chromosomes, the DC factors can be identified in a representative species with the most developed genomic and transgenic tools and later validated in less tractable species that share sex chromosome origins. It is important to first understand the extent and timing of DC. The mechanisms identified so far in insects start in early embryogenesis. In theory,

this should make them easier to identify, as the sexual dimorphism at this stage is very limited, so the few differentially expressed genes are likely to be involved in DC or sex determination. In practice, determining the sex of embryos can be challenging as it cannot be done based on morphology. Due to recent improvements in protocols of library generation for low-input samples, RNA from a single embryo can now be sequenced. Additionally, it is feasible to extract DNA and RNA from the same embryo, and use the gDNA for molecular genotyping [32,59] or potentially whole genome sequencing to determine the embryo's sex. Differential expression analysis between the sexes might reveal a list of candidate genes. The list could be narrowed down based on the predicted structure or nuclear localization of the gene products.

Next, the function of the putative DC factors should be validated experimentally. Cell lines from diverse insect species are increasingly available and a valuable tool to narrow down candidates before more complex investigations *in vivo* would start [92,93]. However, generating new species cell lines remains challenging and time-consuming, making initiatives like the Tick Cell Biobank essential for storing and disseminating existing arthropod cell lines and sharing the expertise needed to create new ones [94]. Cell lines can also be a crucial tool when non-model insect species cannot be propagated in the laboratory setting, making it difficult to perform tractable crosses and generating homozygous knockouts [95]. RNAi can be a useful first approach as it is more easily applicable to a wider range of insect species. Nonetheless, only by assessing complete loss-of-function mutant insects can the question of DC essentiality be unequivocally resolved. It is important not to infer the role in DC based solely on the presence or absence of lethality phenotype in mutants. Instead, the global gene expression should be profiled to test for misregulation of Z or X-linked genes, which would indicate a failure to dosage compensate. The advent of CRISPR and recent improvements in Cas9 ribonucleoprotein delivery methods promise to make genome editing feasible in a wide variety of non-model insect species [96,97]. The challenges and the progress in applying molecular techniques to non model insects have been recently reviewed [98] and strategies to establish new model organisms have been suggested [99].

In summary, while mutagenesis screens have been powerful for discovering DC mechanisms in model organisms, a combination of genomic, transcriptomic, developmental and comparative evolutionary approaches can be used to study DC in a wider range of species.

Concluding remarks and future perspectives

Studying the DC mechanism in *D. melanogaster* has been extremely powerful in discovering new molecular principles for gene regulation (e.g. non-coding RNAs, chromatin modification, histone acetylation, transcription factor specificity, higher order chromosome conformation). In Diptera, sex chromosomes evolve rapidly and repeatedly, but more recently we learnt that other orders comprise deeply conserved sex-chromosomes. Thus, insects offer a rich canvas to discover entirely new

molecular mechanisms and exciting biology (see Outstanding Questions). After the characterizations of genomes and transcriptomes, we anticipate that in the following years the field will focus on exploring the molecular basis of DC mechanisms taking advantage of recently developed gene editing tools. Of note, despite the long history of studying DC in *Drosophila*, there are still many open questions, for which insights from non-model insects can be fruitful (e.g. on the essentiality of DC). We are looking forward to the upcoming discoveries of new DC mechanisms that will reveal the common themes and unique innovations in DC evolution.

Acknowledgements

We thank Ann-Kathrin Huylmans as well as the two peer reviewers for critical reading and constructive feedback of the manuscript. We apologize to colleagues whose work could not be covered in this review due to space constraints. AIK was supported by a Boehringer Ingelheim Fonds PhD Fellowship. CIKV is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Individual Project Grant 513744403, Scientific Network Grant 531902894, GRK GenEvo 407023052, GRK 4R 491145305, Forschungsinitiative Rheinland-Pfalz (ReALity) and institutional funding from the IMB.

Declaration of interests

The authors declare no conflicts of interests.

Additional elements

Glossary

Dosage compensation (process): cellular mechanism that specifically regulates the expression of genes located on the sex chromosomes (X or Z). This refers to all mechanisms that can achieve varying degrees of compensation and balancing.

Complete dosage compensation: a state where the expression of monoallelic, sex chromosomal genes is equal to their ancestral diploid expression i.e. before sex chromosomes have diverged. Because measuring expression levels from ancestral proto-sex chromosomes is usually not possible, complete dosage compensation is typically assessed by 1) comparing sex chromosomal gene expression to the same gene set in an outgroup species with undifferentiated sex chromosomes or when this is not possible 2) comparing sex chromosomal gene expression to overall autosomal gene expression.

Dosage balance: a state in which expression of the sex chromosome linked genes is, on average, equal between males and females.

Haploinsufficient genes: genes which need both alleles for proper function, i.e. a single copy is not sufficient to produce a normal phenotype; also called dosage sensitive genes.

Haplodiploidy: a sex determination system where the sex is determined by the ploidy of the egg, i.e. unfertilized eggs develop into haploid males, while fertilized eggs develop into diploid females.

Heterogametic sex: the sex that produces gametes of two types that differ by whether they contain the X/Z chromosome or the Y/W chromosome.

Heteromorphic sex chromosomes: sex chromosomes that have homology but differ in size, shape and gene content.

Homogametic sex: the sex that produces only one type of gametes, all containing the X (in female heterogametic species: Z) chromosome.

Homomorphic sex chromosomes: sex chromosomes that are not distinguishable morphologically based on their size and shape.

Housekeeping genes: genes that are essential for basic cellular functions and ubiquitously expressed across tissues.

Merian elements: ancestral linkage groups corresponding to the 32 ancestral chromosomes that have remained largely intact throughout the diversification of Lepidoptera.

Muller elements: conserved chromosomal elements that contain similar gene content across *Drosophila* evolution.

Neo-sex chromosomes (e.g. neo-X): sex chromosomes that recently evolved via a fusion of autosomes with ancestral sex chromosomes.

Sex-limited chromosome: the chromosome present only in the heterogametic sex, e.g. W in female-heterogametic species or Y in male-heterogametic species.

Stevens elements: ancestral chromosomal linkage groups identified in Coleoptera that contain a conserved set of genes across beetle species.

Box 1 - Sex Chromosome Evolution and Classification

Sex chromosome evolution begins with the appearance of a sex-determining gene on one of the two homologous autosomes [100]. Homomorphic sex chromosomes are characterized by similar shape, size, and gene content. This makes it difficult to differentiate them cytogenetically, i.e. with karyotype data alone. Homomorphic sex chromosomes can remain stable over long evolutionary periods as observed in Isoptera [101,102] and *Culicinae* [103]. Heteromorphic sex chromosomes, also referred to as differentiated sex chromosomes, show distinct morphology by karyotyping. They evolve from homomorphic sex chromosomes when recombination suppression between them spreads from the sex determination locus. While recombination suppression was originally proposed to be selected for as a way to resolve sexual antagonism, recent theoretical models and data from non-model organisms point to alternative forces driving its emergence and long-term maintenance (reviewed in [104]). For example, neutral accumulation of sequence divergence around the inversion sites and recombination suppression could drive each other via a positive feedback loop [105]. An alternative explanation posits that random inversions could become beneficial if they have fewer deleterious mutations than average [90,106]. The fixation of such inversion can only happen if the remaining recessive mutations they carry are sheltered by always being heterozygous (as is the case when they appear on the Y or W chromosome, but not autosomes or X and Z; sheltering hypothesis [106]). The sheltering model relies on the assumption that the series of consecutive inversions is very unlikely to revert in the same order and reinstate recombination between sex chromosomes. In the regulatory model, complete reinstatement of recombination is possible, but it has a high fitness cost because of a mismatch in gene regulatory elements that evolve early in sex chromosome evolution. These nascent sex-specific regulatory mechanisms (such as early DC) decrease the potential fitness defects resulting from Y/W degeneration and increase the fitness cost of inversions reverting. Whatever the cause, such processes make the sex-limited chromosome (i.e. Y or W) prone to the accumulation of mutations and transposable elements, making it repeat-rich and more heterochromatic. Eventually, the degenerated chromosome shrinks due to gene loss and deletions. In some lineages it can be lost entirely, leading to the formation of systems such as ZO/ZZ or XO/XX.

Another important classification of genotypic sex determination systems relies on understanding which sex carries the sex-limited chromosome. This categorization divides the species into male-heterogametic and female-heterogametic. The heterogametic sex is the one in which the gametes can differ in their haploid chromosome complement. In the male heterogametic species, the Y is the male-limited chromosome, while the X is present in both sexes. In the female heterogametic species, the W is the female-limited chromosome, while the Z chromosome is present in both sexes.

Box 2 - Insect genomes and conservation of their sex chromosomes

DC mechanisms in the model species *D. melanogaster*, *C. elegans* and *M. musculus* were identified and characterized before their genomes were available. However, in non-model species the availability of genome assemblies and accessibility of sequencing has been (and will be) crucial to assess their DC status and potentially uncover novel mechanisms. Recent methodological advances in genome sequencing [107–109] have made it possible to now have more than a thousand insect genomes assembled at chromosome-level. The abundance of insect genomes already allowed for the discovery of surprising patterns in sex chromosome evolution.

Early work in dipterans revealed a remarkable turnover of sex chromosomes [10] but more recently, it was shown that the ancient X is remarkably conserved among non-dipteran insects [9] (also see Figure 1). Since the gene order is not stable over such long evolutionary distances, the common origin is assessed by an excess of shared genes between X chromosomes. The following studies show the common origin of the ancestral X across many insect orders (also see Figure 1).

Bachtrog and Vicoso [22] have initially shown that Muller element F is an ancestral X chromosome in Brachycera. Next, the same authors discovered a conservation of element F as the X in multiple families across Diptera, with others evolving new sex chromosome complement. They suggested the ancestral dipterans either had Muller element F as their X chromosome or had homomorphic sex chromosomes [10].

The shared gene content between Muller element F (chromosome 4 in *Drosophila melanogaster*) and the X chromosome of hemimetabolous insects species has been demonstrated in *Blattella germanica* [14] and *Ischnura elegans* [15]. The second publication has also revealed excess shared gene content between *Ischnura* and *Tribolium castaneum* X. Later it was shown that the X chromosome seems largely conserved across evolution in Coleoptera [16]. Specifically, *Tribolium confusum* and *Dendroctonus ponderosae* share the ancestral part of their X with *T. castaneum*. The origin of the X in *Xenos vesparum* has also been inferred based on the homology of its ancestral part to the X in *T. castaneum* [110].

A direct comparison of the origin of the X across multiple insect orders has been made recently [9]. It revealed ancestral origin of the X in *Ischnura elegans, Timema cristinae, Lucilia cuprina, Acyrthosiphon pisum* and additional species. Of note, the shared origin of the X extends to Collembola - the closest outgroup of Insects. Since *A. pisum* was shown to have X of ancestral origin, the same is likely for additional Hemipteran species: *Halyomorpha halys, Homalodisca vitripennis,* and *Oncopeltus fasciatus,* as their sex chromosomes are homologous [82]. The same inference can be made for Timema species: *T. bartmani, T. californicum and T. poppensi,* as their X chromosomes are homologous to the X of *T. cristinae* [111]. The ancestral origin of X has also been demonstrated in *Locusta migratoria* and *Panorpa cognata* [21], as well as *Locusta*

migratoria, Teleogryllus oceanicus, Pachypsylla venusta, and *Laodelphax striatellus* [112].

Box 3 - Steps towards discovering the DC mechanism

1. Assess if there is DC and/or balance in this species.
    a. RNA-seq in male and female somatic tissues.
2. Analyze the origin of the sex chromosomes (SC).
    a. Check for excess shared X/Z-linked gene orthologs between the species of interest and species with established SC origin.
    b. If the DC mechanism has been elucidated in a species with a shared SC origin, find orthologues of the DC master regulator and go to step 5.
3. Check for a sex-specific enrichment of histone modifications already known to play a role in DC (by chromatin profiling or staining).
    i. H4K16ac
    ii. H2AK118/9ub
    iii. H3K27me3
    iv. H3K9me2/3
    v. H4K20me1

- No → step 4

- Yes → find orthologues of histone modification writers, look for sex-specific expression or splicing. If sex-specific → go to step 5
    i. MOF
    ii. RING1B
    iii. Ezh1
    iv. Su(var)3-9
    v. DPY-21/Jumonji

Caveat: there can be no a priori assumption that different lineages will use the same (or any) histone marks to achieve DC. However, recent examples of H4K16ac as the DC mark in distant species [44,113] suggests that some modifications have been harnessed for this role repeatedly.

4. Identify and validate early sex-specific genes
    a. Isolate both RNA and DNA from single embryos; use gDNA-qPCR or genome sequencing to determine embryo sex [114] →RNA-seq
    b. Identify DC onset and genes with sex-specific expression or splicing. If many candidates are found:
        i. Narrow down the list bioinformatically based on predicted nuclear localization signals, gene expression or chromatin-regulating domains.
        ii. Use available cell lines for pre-screening (knockdown if the sex of the cell line is compatible, overexpression if not).

Caveats: apart from DC, genes with early sex-specific expression can be involved in the sex determination cascade [36,115].

5. Validate the candidate *in vivo*.
    a. a. Full knock-out of the candidate gene → RNA-seq: SC misexpression?
    b. Ectopic expression by knock-in of coding sequence in opposite sex
        i. RNA-seq → SC misexpression?
        ii. Chromatin binding assay (possible without a custom antibody, if a epitope tag is also encoded in a transgene)

Caveat: if the candidate also regulates sexual development, mutants could have reproductive defects or altered phenotypic sex, making crossing to obtain homozygous knock-outs challenging.

6. If downregulation in homogametic sex is observed, differentiate between a single SC being silenced or both SC being dampened:
    i. Allele-specific expression analysis in the homogametic sex. The heterozygosity of SNPs in RNA-seq: most genes expressed monoallelically - silencing; biallelically - dampening.
    ii. Intronic RNA fluorescence in situ hybridization - one dot at the transcription site expected for X/Z-linked genes in the heterogametic sex, two in homogametic.

Caveat: the allele-specific analysis requires high heterozygosity, therefore it can only work on non-inbred specimens or hybrids resulting from crossing two distant inbred strains [35]. This approach also requires the W or Y chromosome to be degenerated, otherwise the gametologs can confound the heterozygosity analysis.

These steps are a starting point for new species and based on what we know about DC in insects so far. Completely new mechanisms might exist, which will not be captured with these approaches (such as a DC factor regulated exclusively at the protein level). Further methodological progress (e.g. in proteomics) will open up new ways to identify DC mechanisms.

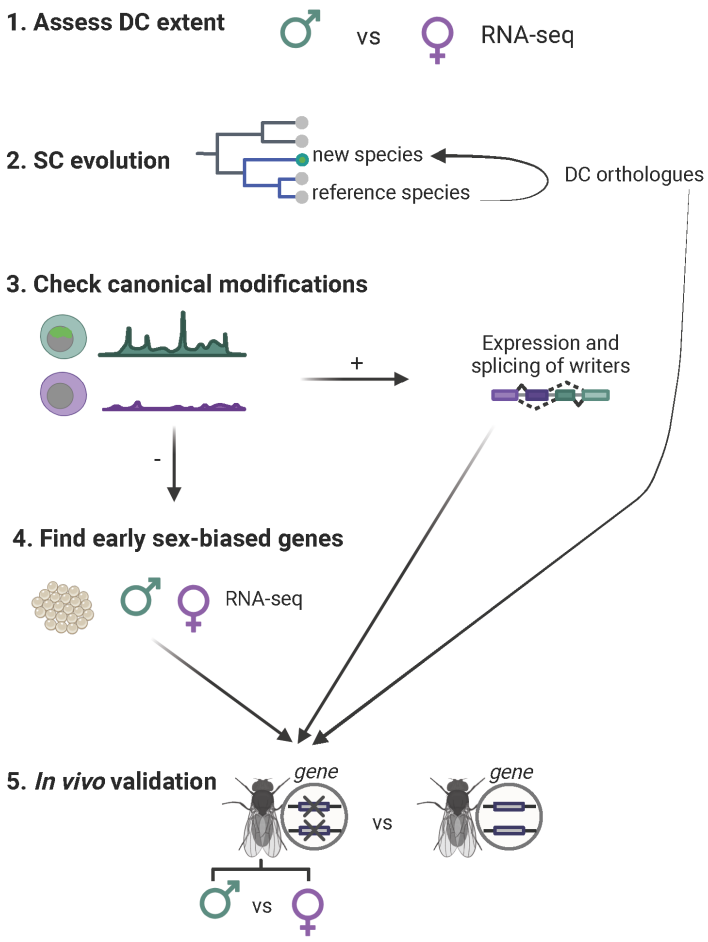


Figure I. A graphical summary of approaches to identify DC mechanism in a new species.

The figure shows a list of steps, in order of priority. If steps 2 or 3 reveal a promising candidate, the researcher can go directly to step 5 to validate the candidates *in vivo.* Created with BioRender.com

Table 1

Dosage compensation status of species depicted in Figure 1

Abbreviations

ME: Muller element
anc-X: ancestral part of the X chromosome
neo-X: newly fused part of the X chromosome
SGB: sexually biased genes
M: male
F: female
M:F: ratio of male to female expression
no DC (in gonads): expression patterns are consistent with lack of DC in gonads, though this expression pattern could alternatively result from strong sex bias or meiotic sex chromosome inactivation (MSCI)
NA: no information available
WB: whole body samples
P: pupae
Ad: adults
L: larvae
E: embryos

Species / Order	Sex-chromosomes	DC/balance status	Symbolic	Values	Tissues tested	Expression pattern in gonads	Stage	Enrichment in sex-biased genes	Reference
Acyrthosiphon pisum Hemiptera	XO	[81] incomplete [82,123] complete	X=XX=AA	[81] M:F: X=1.31, A=1.02; [82] (M:F X)/(M:F A) =1.02 [123] X/A F=0.58 M=0.94	WB	NA	Ad	masculinization of X[1]	[81] [82] [123]
Anopheles albimanus Diptera	XY	complete DC	X=XX>AA	A:X F=0.76 M=0.86	carcass	no DC	Ad	demasculinized X in gonads, feminized in the carcass	[30]
Anopheles arabiensis Diptera	XY	complete DC	X=XX>AA	A:X F=0.84 M=0.85	carcass	no DC	Ad	demasculinized and feminized X in gonads	[30]

[1] The unusual pattern compared to male heterogametic species could result from a reproduction system that includes asexual females

Species / Order	Sex-chromosomes	DC/balance status	Symbolic	Values	Tissues tested	Expression pattern in gonads	Stage	Enrichment in sex-biased genes	Reference
Anopheles gambiae Diptera	XY	complete DC	X=XX>AA	[29] L M:F (X=0.89–1.08, A=0.89–0.99); P M:F (X=0.92–0.96, A=0.97–1.01); [30] Ad A:X F=0.85 M=0.87	[29] WB [30] carcass	[29] no DC in P testes (premeiotic) [30] no DC in Ad gonads	[29] L, P [30] Ad	[30] demasculinized X in P	[29] [30]
Anopheles minimus Diptera	XY	complete DC	X=XX=AA	A:X F=0.97 M=1.00	carcass	no DC	Ad	demasculinized X in gonads	[30]
Anopheles stephensi Diptera	XY	complete DC	X=XX=AA	X:A F=0.92 M=0.96	WB	NA	Ad	demasculinized X	[28]
Blattella germanica Blattodea	XO	complete DC	X=XX=AA	ME-F vs A-E: p=0.15 for M vs mixed-sex heads; p=0.3 for M head vs WB F	head	NA	Ad	NA	[14]
Bombyx mori Lepidoptera	ZW	balanced, incomplete DC	Z ≈ ZZ < AA	[124] mid stage E: Z:A M=0.53,F=0.49; late E: Z:A M=0.62 F=0.66; L: Z:A M=0.77 F=0.71; [88] Ad thorax Z:A M=0.858 F=0.789, Heads M=0.833, F=0.81	[124] whole E, L head [88] thorax, head	no DC	[124] E, L, [88] Ad	[88] masculinization and defeminization of the Z in gonads, masculinization of Z in thorax	[124] [88]
Cydia pomonella Lepidoptera	ZW	complete DC of neo-Z ancZ balanced	ancZ ≈anc ZZ < AA neoZ=neoZZ=AA	M:F: head (A=1.04 neoZ=0.97 ancZ=1.05) midgut (A=0.97 neoZ=0.95 ancZ=1.01)	head, midgut, gonads	no DC	Ad	masculinization of anc-Z in gonads	[51]
Danaus plexippus Lepidoptera	ZW	neo-Z: complete DCa nc-Z: balanced in head, abdomen; imbalanced in thorax, WB	NeoZ=neoZZ=AA ancZ: heads+ abdomens Z=ZZ<AA thorax+WB: Z<ZZ=AA	[44] ancZ:A (F=0.57 M =0.56) neoZ:A (F=1.09 M =1.10) [52] F:M WB: [ancZ:0.87 neoZ:0.97 A:1.05] thorax: [ancZ:0.85 neoZ:1 A:1.05] head: [ancZ:0.96 neoZ:1.01 A:1] abdomen: [ancZ:1.04 neoZ:1.02 A:1]	[44] head [52] WB, thorax, head, abdomen	NA	Ad	[52] WB: masculinization and defeminization of anc-Z	[44] [52]
Dendroctonus ponderosae Coleoptera	XY	complete DC	ancX=ancXX=AA	NA	head	ancX: no DC	Ad	demasculinized and feminized anc-X in gonads	[16]

Species / Order	Sex-chromosomes	DC/balance status	Symbolic	Values	Tissues tested	Expression pattern in gonads	Stage	Enrichment in sex-biased genes	Reference
Drosophila busckii Diptera	XY	complete DC for ancX, partial for neo-X	ancX=ancXX=AA neoX<neoXX=AA	NA	WB	NA	3rd instar L, Ad	NA	[47]
Drosophila melanogaster Diptera	XY	complete DC	X=XX=AA	NA	[125] WB [10] WB, somatic	[10,126] no DC	Ad	[125] demasculinized X [10] feminized X in heads and gonads, demasculinized X in gonads	[126] [125] [10]
Drosophila miranda Diptera	XY	complete DC for anc-X and older neo-X, incomplete for youngest neo-X	ME-A: ancX=ancXX=AA; ME-D: neoX=neoXX=AA; ME-C: neoX<neoXX=AA	NA	[127] whole E	[128] no DC	[127] E	NA	[127] [128]
Drosophila pseudoobscura Diptera	XY	[129] ancX complete, neo-X: complete in heads, incomplete in WB and WB w/o gonads	ancX=ancXX=AA	[129] neo-X/proto-X: WB=0.85, WB w/o gonads = 0.9, heads=1.0	[125] WB [129] WB, WB w/o gonads, head		Ad	[125] demasculinized anc-X and neo-X	[125] [129]
Ephydra hians Diptera	XY	slightly incomplete DC	X≈XX=AA	M X/proto-X $p<0.05$ in somatic tissues, $p<0.001$ in WB	WB, somatic	no DC	Ad	feminized X in WB, somatic tissues and gonads	[10]
Glossina morsitans Diptera	XY	complete DC	X=XX=AA	NA	WB, head, antenna	NA	Ad	slightly feminized X in WB, somatic tissues and gonads	[10]
Halyomorpha halys Hemiptera	XY	partial DC	X<XX=AA	(M/F X)/(M/F A) =0.83	WB	NA	Ad	demasculinization of the X	[10]
Heliconius melpomene Lepidoptera	ZW	[130,131] slightly imbalanced, variable by tissue	Z≈ZZ<AA	Z:A Head: M=0.7615, F=0.7298; Antenna: M=0.7085 F=0.6872; Leg: M=0.7337 F=0.6493; Mouth: M=0.7444 F=0.653	[131] head, abdomen [130] antenna, leg, mouth [132] brain+eye	NA	Ad	[131] masculinization and defeminization of Z in abdomen [132] no sex-bias on Z	[131] [130] [132]

Species / Order	Sex-chromosomes	DC/balance status	Symbolic	Values	Tissues tested	Expression pattern in gonads	Stage	Enrichment in sex-biased genes	Reference
Heliconius charithonia Lepidoptera	ZW	imbalanced lowly and mid-expressed genes, balanced highly expressed	Z≈ZZ<AA	Z:A (M=0.93 F=0.89); A:Z (F:1.1-1.16 M=1.06-1.14)	brain+eyes	NA	Ad	Z-chromosome enrichment in both female-and male-biased genes	[132]
Heliconius cydno Lepidoptera	ZW	slightly imbalanced	Z≈ZZ<AA	Head Z:A M=0.7044, F=0.6689	head, abdomen	NA	Ad	masculinization and defeminization of Z in abdomen	[131]
Heliconius doris Lepidoptera	ZW	imbalanced lowly expressed genes, balanced mid and highly expressed	Z≈ZZ<AA	F:M Z=0.98	brain+eyes	NA	Ad	Z-chromosome enriched in female- and male-biased genes	[132]
Heliconius erato Lepidoptera	ZW	imbalanced lowly and mid expressed genes, balanced highly expressed	Z≈ZZ<AA	F:M (A=1.005, Z=0.95)	brain+eyes	NA	Ad	masculinization of Z	[132]
Heliconius sara Lepidoptera	ZW	imbalanced lowly and mid expressed genes, balanced highly expressed	Z≈ZZ<AA	Z:A (M=0.85 F=0.83)	brain+eyes	NA	Ad	masculinization of Z	[132]
Homalodisca vitripennis Hemiptera	XO	partial	X≈XX=AA	(M/F X)/(M/F A) =0.94	WB	NA	Ad	demasculinization of the X	[82]
Idea leuconoe Lepidoptera	ZW	neoZ: complete DC ancZ: balanced	ancZ=ancZZ<AA neoZ=neoZZ>AA	soma: [anc-Z/A M=0.825 F=0.82 neo-Z/A M=1.05 F=1.14] head [anc-Z/A M=0.893 F=0.929 neo-Z/A M=1.1 F=1.13]	thorax, head	no DC	Ad	masculinization of Z in gonads	[53]
Ischnura elegans Odonata	XO	balanced, incomplete DC	X=XX<AA	(M:F X)/(M:F A) = 1.00; X=XX=0.93 AA	WB	NA	Ad	no enrichment of SBG on the X	[15]

Species / Order	Sex-chromosomes	DC/balance status	Symbolic	Values	Tissues tested	Expression pattern in gonads	Stage	Enrichment in sex-biased genes	Reference
Lamprohiza splendidula Coleoptera	XO	complete DC	X=XX>=AA	NA	head, abdomen	NA	Ad	autosomal enrichment of SBG	[133]
Laodelphax striatellus Hemiptera	XO	complete DC	X=XX=AA	X:A M=0.952-0.976, F=0.965-1.002 in somatic tissues.	head, antenna, leg	no DC	Ad	feminization of the X in somatic tissues and gonads, demasculinization in the gonads	[85]
Leptidea sinapis Lepidoptera	ZW	balanced	Z=ZZ<AA	L: M:F A=1.08, Z=1.08; Z:A F=0.5 M=0.39; P: M:F A=0.92, Z=1.09; Z:A F=0.46 M=0.53 Ad: M:F A=1.11, Z=1.67; Z:A F=0.44, M=0.65 (includes SBG)	WB	NA	L, P, Ad	masculinization of Z in all stages	[73]
Liriomyza trifolli Diptera	XY	complete, X slightly upregulated in females	X≈XX=AA	NA	WB	NA	Ad	NA	[10]
Locusta migratoria Orthoptera	XO	complete DC	X=XX>AA	NA	hindleg, brain	no DC	Ad	NA	[112]
Lucilia cuprina Diptera	XY	complete DC	X=XX=AA	NA	head, thoraxes	NA	Ad	NA	[55]
Luciola italica Coleoptera	XO	complete DC	X=XX>=AA	NA	head, abdomen	NA	Ad	autosomal enrichment of SBG	[133]
Lycorea halia Lepidoptera	ZW	neoZ: complete DC in head, almost in soma; ancZ: balanced	ancZ=ancZZ<AA neoZ=neoZZ≈AA	soma: [anc-Z/A M=0.617 F=0.57 neo-Z/A M=0.846 F=0.875] head [anc-Z/A M=0.719 F=0.721 neo-Z/A M=0.912 F=0.941]	thorax, head	no DC	Ad	masculinization and defeminization of Z in the gonads	[53]
Manduca sexta Lepidoptera	ZW	complete DC	Z = ZZ = AA	Z M:F 0.81 in Table 1; Z: (M=8.62 F=8.78) A: (M=10.6 F=10.82) [FPKM] Calculated: M(Z/A)/ F(Z/A) =1.002, Z M:F = 0.98	heads	NA	Ad	no enrichment of SBG on Z	[134]
Oncopeltus fasciatus	XY	partial	X<XX=AA	(M/F X)/(M/F A) =0.87	WB	NA	Ad	demasculinization of the X	[82]

Species / Order	Sex-chromosomes	DC/balance status	Symbolic	Values	Tissues tested	Expression pattern in gonads	Stage	Enrichment in sex-biased genes	Reference
Hemiptera									
Pachypsylla venusta Hemiptera	XO	partial	X<XX=AA	M:F A=0.96 X=0.81	WB	NA	Ad	Male-biased genes enriched on autosomes	[83]
Panorpa cognata Mecoptera	XO	complete DC in heads, slightly incomplete in carcass	X=XX=AA	NA	head, carcass	no DC	Ad	feminization and demasculinization of X in gonads	[21]
Papilio machaon Lepidoptera	ZW	complete DC	Z=ZZ=AA	Ad [SGB included Z:A (M=1.015, F=1.044); M:F (A=1, Z=0.972) SGB excluded Z:A (F=0.909, M=0.996); M:F (A=0.925, Z=1.014)] P [Z:A (M=1.022, F=0.869); M:F (A=0.925, Z=1.014)]	WB	NA	Ad	feminization of chrZ in Ad	[88]
Papilio xuthus Lepidoptera	ZW	balanced	Z≈ZZ<AA	Ad: [SGB included Z:A (M=1.043, F=0.581); M:F (A=0.973, Z=1.748) SGB excluded Z:A (F=0.578, M=0.657); M:F (A=1.045, Z=1.187)] P: [Z/A M=0.880 F=0.999]	WB	NA	Ad, P	masculinization and defeminization of chrZ in Ad, masculinization in P	[88]
Phortica variegata Diptera	XY	complete DC	X=XX=AA	NA	WB	NA	Ad	NA	[10]
Plodia interpunctella Lepidoptera	ZW	[135] lack of DC [88] balanced	Z=ZZ<AA	[135] WB: Z:A M=0.954 F=0.534 [88] Ad [Thorax: Z:A (F=0.91, M=0.895) Head Z:A (F=0.836 M=0.772)] L heads: Z:A (F=0.746 M=0.757)	[135] WB [88] thorax, head	[88] no DC	[135] Ad [88] Ad, L	[88] masculinization and defeminization of the Z in gonads	[135] [88]
Teleogryllus oceanicus Orthoptera	XO	variable	X=XX>AA	X F:M=1.05	somatic (neural, thoratic, wingbud)	no DC	Ad	feminization of the X in the gonads	[84]

Species / Order	Sex-chromosomes	DC/balance status	Symbolic	Values	Tissues tested	Expression pattern in gonads	Stage	Enrichment in sex-biased genes	Reference
Teleopsis dalmani Diptera	XY	[10] incomplete [136] complete	X=XX=AA	[136] X: (L: M=9.3, F=9.36, Ad: M=8.23, F=8.22); M: (L: A=9.48, X=9.3; Ad A =8.21, X=8.23) [mean]	[10] WB; [136] A: head, L: eye disc	NA	[10] Ad [136] Ad, L	[136] feminized X and defeminized autosomes	[10] [136]
Themira minor Diptera	XY	incomplete	X≈XX=AA	NA	WB, head	no DC	Ad	feminized and demasculinized X in WB, gonads; feminized in somatic tissues	[10]
Timema bartmani Phasmatodea	XO	complete in legs, almost complete in heads	X≈XX=AA	NA	head, leg	no DC	Ad	feminization and demasculinization of X in heads and in gonads	[111]
Timema californicum Phasmatodea	XO	almost complete	X≈XX=AA	NA	head, leg	no DC	Ad	feminization and demasculinization in gonads	[111]
Timema cristinae Phasmatodea	XO	complete in legs, almost complete in heads	X≈XX=AA	NA	head, leg	no DC	Ad	feminization and demasculinization in gonads	[111]
Timema podura Phasmatodea	XO	complete in legs, almost complete in heads	X≈XX=AA	NA	head, leg	no DC	Ad	feminization and demasculinization in gonads	[111]
Timema poppense Phasmatodea	XO	complete	X=XX=AA	NA	[111] head, leg [137] antenna, brain, gut, leg	[111] no DC [137] DC before meiosis, MSCI after meiosis	[111] Ad [137] Nymphs, Ad	feminization and demasculinization in gonads	[111] [137]
Tribolium castaneum Coleoptera	XY	complete	X=XX=AA	X M:F =0.93 (p=0.74); M X:A=0.91 (p=0.11)	gonadectomized body	no DC	Ad	demasculinized and feminized X in gonads	[138]
Tribolium confusum Coleoptera	XY	complete	X=XX=AA	NA	head	no DC (anc-X and neo-X)	Ad	anc-X demasculinized, no bias on neo-X	[16]
Xenos vesparum Strepsiptera	XY	anc-X: almost complete DC neo-X partial	ancX=ancXX<AA neoX<neoXX=AA	NA	WB	NA	Ad F, L F, P M	NA	[110]

References

1. Goodenough, U. and Heitman, J. (2014) Origins of eukaryotic sexual reproduction. *Cold Spring Harb. Perspect. Biol.* 6
2. Tree of Sex Consortium (2014) Tree of Sex: a database of sexual systems. *Sci Data* 1, 140015
3. Stingele, S. *et al.* (2012) Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. *Mol. Syst. Biol.* 8, 608
4. Furman, B.L.S. *et al.* (2020) Sex Chromosome Evolution: So Many Exceptions to the Rules. *Genome Biol. Evol.* 12, 750–763
5. Birchler, J.A. and Veitia, R.A. (2010) The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytol.* 186, 54–62
6. Papp, B. *et al.* (2003) Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424, 194–197
7. Mank, J.E. (2013) Sex chromosome dosage compensation: definitely not for everyone. *Trends Genet.* 29, 677–683
8. Gu, L. and Walters, J.R. (2017) Evolution of Sex Chromosome Dosage Compensation in Animals: A Beautiful Theory, Undermined by Facts and Bedeviled by Details. *Genome Biol. Evol.* 9, 2461–2476
9. Toups, M.A. and Vicoso, B. (2023) The X chromosome of insects likely predates the origin of Class Insecta. *Evolution* DOI: 10.1093/evolut/qpad169
10. Vicoso, B. and Bachtrog, D. (2015) Numerous transitions of sex chromosomes in Diptera. *PLoS Biol.* 13, e1002078
11. Misof, B. *et al.* (2014) Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346, 763–767
12. Blackmon, H. *et al.* (2017) Sex Determination, Sex Chromosomes, and Karyotype Evolution in Insects. *J. Hered.* 108, 78–93
13. Stork, N.E. (2018) How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth? *Annu. Rev. Entomol.* 63, 31–45
14. Meisel, R.P. *et al.* (2019) The X chromosome of the German cockroach, Blattella germanica, is homologous to a fly X chromosome despite 400 million years divergence. *BMC Biol.* 17, 100
15. Chauhan, P. *et al.* (2021) Genome assembly, sex-biased gene expression and dosage compensation in the damselfly Ischnura elegans. *Genomics* 113, 1828–1837
16. Bracewell, R. *et al.* (2023) Sex chromosome evolution in beetles*bioRxiv*, 2023.01.18.524646
17. Grove, S.J. and Stork, N.E. (2000) An inordinate fondness for beetles. *Invertebr. Syst.* 14, 733–739
18. Chen, X. *et al.* (2023) Unraveling the complex evolutionary history of lepidopteran chromosomes through ancestral chromosome reconstruction and novel chromosome nomenclature. *BMC Biol.* 21, 265
19. Wright, C.J. *et al.* (2024) Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera. *Nat Ecol Evol* DOI: 10.1038/s41559-024-02329-4
20. Walters, J.R. and Hardcastle, T.J. (2011) Getting a full dose? Reconsidering sex chromosome dosage compensation in the silkworm, Bombyx mori. *Genome Biol. Evol.* 3, 491–504
21. Lasne, C. *et al.* (2023) The Scorpionfly (Panorpa cognata) Genome Highlights Conserved and Derived Features of the Peculiar Dipteran X Chromosome. *Mol. Biol. Evol.* 40
22. Vicoso, B. and Bachtrog, D. (2013) Reversal of an ancient sex chromosome to an autosome in Drosophila. *Nature* 499, 332–335
23. Samata, M. and Akhtar, A. (2018) Dosage Compensation of the X Chromosome: A Complex Epigenetic Assignment Involving Chromatin Regulators and Long Noncoding RNAs. *Annu. Rev. Biochem.* 87, 323–350

24. Larsson, J. *et al.* (2001) Painting of fourth, a chromosome-specific protein in *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A.* 98, 6273–6278
25. Riddle, N.C. *et al.* (2012) Enrichment of HP1a on Drosophila chromosome 4 genes creates an alternate chromatin structure critical for regulation in this heterochromatic domain. *PLoS Genet.* 8, e1002954
26. Johansson, A.-M. *et al.* (2007) Painting of fourth and chromosome-wide regulation of the 4th chromosome in Drosophila melanogaster. *EMBO J.* 26, 2307–2316
27. Davis, R.J. *et al.* (2018) no blokes Is Essential for Male Viability and X Chromosome Gene Expression in the Australian Sheep Blowfly. *Curr. Biol.* 28, 1987–1992.e3
28. Jiang, X. *et al.* (2015) Complete dosage compensation in Anopheles stephensi and the evolution of sex-biased genes in mosquitoes. *Genome Biol. Evol.* 7, 1914–1924
29. Rose, G. *et al.* (2016) Dosage compensation in the African malaria mosquito Anopheles gambiae. *Genome Biol. Evol.* 8, 411–425
30. Papa, F. *et al.* (2017) Rapid evolution of female-biased genes among four species of Anopheles malaria mosquitoes. *Genome Res.* 27, 1536–1548
31. Keller Valsecchi, C.I. *et al.* (2021) Distinct mechanisms mediate X chromosome dosage compensation in Anopheles and Drosophila. *Life Sci Alliance* 4
32. Kalita, A.I. *et al.* (2023) The sex-specific factor SOA controls dosage compensation in Anopheles mosquitoes. *Nature* 623, 175–182
33. Krzywinska, E. *et al.* (2023) A novel factor modulating X chromosome dosage compensation in Anopheles. *Curr. Biol.* 33, 4697–4703.e4
34. Krzywinska, E. *et al.* (2021) femaleless Controls Sex Determination and Dosage Compensation Pathways in Females of Anopheles Mosquitoes. *Curr. Biol.* 31, 1084–1091.e4
35. Tomihara, K. *et al.* (2022) Masculinizer-induced dosage compensation is achieved by transcriptional downregulation of both copies of Z-linked genes in the silkworm, Bombyx mori. *Biol. Lett.* 18, 20220116
36. Kiuchi, T. *et al.* (2014) A single female-specific piRNA is the primary determiner of sex in the silkworm. *Nature* 509, 633–636
37. Wang, Y.-H. *et al.* (2019) The Masc gene product controls masculinization in the black cutworm, Agrotis ipsilon. *Insect Sci.* 26, 1037–1044
38. Harvey-Samuel, T. *et al.* (2020) Identification and characterization of a Masculinizer homologue in the diamondback moth, Plutella xylostella. *Insect Mol. Biol.* 29, 231–240
39. Deng, Z. *et al.* (2021) Identification and Characterization of the Masculinizing Function of the Helicoverpa armigera Masc Gene. *Int. J. Mol. Sci.* 22
40. Visser, S. *et al.* (2021) A conserved role of the duplicated Masculinizer gene in sex determination of the Mediterranean flour moth, Ephestia kuehniella. *PLoS Genet.* 17, e1009420
41. Bi, H. *et al.* (2022) Masculinizer and Doublesex as Key Factors Regulate Sexual Dimorphism in Ostrinia furnacalis. *Cells* 11
42. Pospíšilová, K. *et al.* (2023) Masculinizer gene controls male sex determination in the codling moth, Cydia pomonella. *Insect Biochem. Mol. Biol.* 160, 103991
43. Fukui, T. *et al.* (2023) Masculinizer is not post-transcriptionally regulated by female-specific piRNAs during sex determination in the Asian corn borer, Ostrinia furnacalis. *Insect Biochem. Mol. Biol.* 156, 103946
44. Gu, L. *et al.* (2019) Dichotomy of Dosage Compensation along the Neo Z Chromosome of the Monarch Butterfly. *Curr. Biol.* 29, 4071–4077.e3
45. Carvalho, A.B. and Clark, A.G. (2005) Y chromosome of D. pseudoobscura is not homologous to the ancestral Drosophila Y. *Science* 307, 108–110
46. Bachtrog, D. and Charlesworth, B. (2002) Reduced adaptation of a non-recombining neo-Y chromosome. *Nature* 416, 323–326
47. Zhou, Q. and Bachtrog, D. (2015) Ancestral Chromatin Configuration Constrains Chromatin Evolution on Differentiating Sex Chromosomes in Drosophila. *PLoS Genet.* 11, e1005331
48. Alekseyenko, A.A. *et al.* (2013) Conservation and de novo acquisition of dosage

compensation on newly evolved sex chromosomes in Drosophila. *Genes Dev.* 27, 853–858
49. Ellison, C.E. and Bachtrog, D. (2013) Dosage compensation via transposable element mediated rewiring of a regulatory network. *Science* 342, 846–850
50. Zhou, Q. *et al.* (2013) The epigenome of evolving Drosophila neo-sex chromosomes: dosage compensation and heterochromatin formation. *PLoS Biol.* 11, e1001711
51. Gu, L. *et al.* (2017) Conserved Patterns of Sex Chromosome Dosage Compensation in the Lepidoptera (WZ/ZZ): Insights from a Moth Neo-Z Chromosome. *Genome Biol. Evol.* 9, 802–816
52. Ranz, J.M. *et al.* (2021) A de novo transcriptional atlas in Danaus plexippus reveals variability in dosage compensation across tissues. *Commun Biol* 4, 791
53. Mora, P. *et al.* (2024) Sex-biased gene content is associated with sex chromosome turnover in Danaini butterflies. *Mol. Ecol.*
54. Gorchakov, A.A. *et al.* (2009) Long-range spreading of dosage compensation in Drosophila captures transcribed autosomal genes inserted on X. *Genes Dev.* 23, 2266–2271
55. Linger, R.J. *et al.* (2015) Dosage Compensation of X-Linked Muller Element F Genes but Not X-Linked Transgenes in the Australian Sheep Blowfly. *PLoS One* 10, e0141544
56. Huylmans, A.K. and Parsch, J. (2015) Variation in the X:Autosome Distribution of Male-Biased Genes among Drosophila melanogaster Tissues and Its Relationship with Dosage Compensation. *Genome Biol. Evol.* 7, 1960–1971
57. Belyi, A. *et al.* (2020) The Influence of Chromosomal Environment on X-Linked Gene Expression in Drosophila melanogaster. *Genome Biol. Evol.* 12, 2391–2402
58. Gopinath, G. *et al.* (2017) RNA sequencing reveals a complete but an unconventional type of dosage compensation in the domestic silkworm Bombyx mori. *R Soc Open Sci* 4, 170261
59. Lott, S.E. *et al.* (2011) Noncanonical compensation of zygotic X transcription in early Drosophila melanogaster development revealed through single-embryo RNA-Seq. *PLoS Biol.* 9
60. Prayitno, K. *et al.* (2019) Progressive dosage compensation during Drosophila embryogenesis is reflected by gene arrangement. *EMBO Rep.* 20, e48138
61. Belote, J.M. and Lucchesi, J.C. (1980) Male-specific lethal mutations of Drosophila melanogaster. *Genetics* 96, 165–186
62. Franke, A. and Baker, B.S. (1999) The rox1 and rox2 RNAs are essential components of the compensasome, which mediates dosage compensation in Drosophila. *Mol. Cell* 4, 117–122
63. Kelley, R.L. *et al.* (1995) Expression of msl-2 causes assembly of dosage compensation regulators on the X chromosomes and female lethality in Drosophila. *Cell* 81, 867–877
64. Lim, C.K. and Kelley, R.L. (2012) Autoregulation of the Drosophila Noncoding roX1 RNA Gene. *PLoS Genet.* 8, e1002564
65. Valsecchi, C.I.K. *et al.* (2018) Facultative dosage compensation of developmental genes on autosomes in Drosophila and mouse embryonic stem cells. *Nat. Commun.* 9, 3626
66. Cline, T.W. (1978) Two closely linked mutations in Drosophila melanogaster that are lethal to opposite sexes and interact with daughterless. *Genetics* 90, 683–698
67. Bhadra, U. *et al.* (2000) Histone acetylation and gene expression analysis of sex lethal mutants in Drosophila. *Genetics* 155, 753–763
68. Criscione, F. *et al.* (2016) GUY1 confers complete female lethality and is a strong candidate for a male-determining factor in Anopheles stephensi. *Elife* 5
69. Krzywinska, E. *et al.* (2016) A maleness gene in the malaria mosquito Anopheles gambiae. *Science* 353, 67–69
70. Qi, Y. *et al.* (2019) Guy1, a Y-linked embryonic signal, regulates dosage compensation in Anopheles stephensi by increasing X gene expression. *Elife* 8
71. Krzywinska, E. and Krzywinski, J. (2018) Effects of stable ectopic expression of the primary sex determination gene Yob in the mosquito Anopheles gambiae. *Parasit. Vectors* 11, 648

72. Sakai, H. *et al.* (2016) Transgenic Expression of the piRNA-Resistant Masculinizer Gene Induces Female-Specific Lethality and Partial Female-to-Male Sex Reversal in the Silkworm, Bombyx mori. *PLoS Genet.* 12, e1006203
73. Höök, L. *et al.* (2019) Multilayered Tuning of Dosage Compensation and Z-Chromosome Masculinization in the Wood White (Leptidea sinapis) Butterfly. *Genome Biol. Evol.* 11, 2633–2652
74. Katsuma, S. *et al.* (2022) A Wolbachia factor for male killing in lepidopteran insects. *Nat. Commun.* 13, 6764
75. Fukui, T. *et al.* (2024) Expression of the Wolbachia male-killing factor Oscar impairs dosage compensation in lepidopteran embryos. *FEBS Lett.* 598, 331–337
76. Harumoto, T. *et al.* (2016) Male-killing symbiont damages host's dosage-compensated sex chromosome to induce embryonic apoptosis. *Nat. Commun.* 7, 12781
77. Harumoto, T. and Lemaitre, B. (2018) Male-killing toxin in a bacterial symbiont of Drosophila. *Nature* 557, 252–255
78. Sugimoto, T.N. *et al.* (2015) Misdirection of dosage compensation underlies bidirectional sex-specific death in Wolbachia-infected Ostrinia scapulalis. *Insect Biochem. Mol. Biol.* 66, 72–76
79. Basilicata, M.F. and Keller Valsecchi, C.I. (2021) The good, the bad, and the ugly: Evolutionary and pathological aspects of gene dosage alterations. *PLoS Genet.* 17, e1009906
80. Chen, J. *et al.* (2020) The evolution of sex chromosome dosage compensation in animals. *J. Genet. Genomics* 47, 681–693
81. Jaquiéry, J. *et al.* (2013) Masculinization of the x chromosome in the pea aphid. *PLoS Genet.* 9, e1003690
82. Pal, A. and Vicoso, B. (2015) The X Chromosome of Hemipteran Insects: Conservation, Dosage Compensation and Sex-Biased Expression. *Genome Biol. Evol.* 7, 3259–3268
83. Li, Y. *et al.* (2020) The Aphid X Chromosome Is a Dangerous Place for Functionally Important Genes: Diverse Evolution of Hemipteran Genomes Based on Chromosome-Level Assemblies. *Mol. Biol. Evol.* 37, 2357–2368
84. Rayner, J.G. *et al.* (2021) Variable dosage compensation is associated with female consequences of an X-linked, male-beneficial mutation. *Proc. Biol. Sci.* 288, 20210355
85. Hu, Q.-L. *et al.* (2022) Chromosome-level Assembly, Dosage Compensation and Sex-biased Gene Expression in the Small Brown Planthopper, Laodelphax striatellus. *Genome Biol. Evol.* 14
86. Li, X. *et al.* (2024) The grasshopper genome reveals long-term gene content conservation of the X Chromosome and temporal variation in X Chromosome evolution. *Genome Res.* DOI: 10.1101/gr.278794.123
87. Castagné, R. *et al.* (2011) The choice of the filtering method in microarrays affects the inference regarding dosage compensation of the active X-chromosome. *PLoS One* 6, e23956
88. Huylmans, A.K. *et al.* (2017) Global Dosage Compensation Is Ubiquitous in Lepidoptera, but Counteracted by the Masculinization of the Z Chromosome. *Mol. Biol. Evol.* 34, 2637–2649
89. Nusbaum, C. and Meyer, B.J. (1989) The Caenorhabditis elegans gene sdc-2 controls sex determination and dosage compensation in XX animals. *Genetics* 122, 579–593
90. Lenormand, T. and Roze, D. (2022) Y recombination arrest and degeneration in the absence of sexual dimorphism. *Science* 375, 663–666
91. Lenormand, T. *et al.* (2020) Sex Chromosome Degeneration by Regulatory Evolution. *Curr. Biol.* 30, 3001–3006.e5
92. Katsuma, S. *et al.* (2019) Masc-induced dosage compensation in silkworm cultured cells. *FEBS Open Bio* 9, 1573–1579
93. He, X. *et al.* (2023) Insect Cell-Based Models: Cell Line Establishment and Application in Insecticide Screening and Toxicology Research. *Insects* 14
94. Bell-Sakyi, L. *et al.* (2018) The Tick Cell Biobank: A global resource for in vitro research on ticks, other arthropods and the pathogens they transmit. *Ticks Tick Borne Dis.* 9,

1364–1371
95. Watanabe, K. *et al.* (2020) Establishment and characterization of novel cell lines derived from six lepidopteran insects collected in the field. *In Vitro Cell. Dev. Biol. Anim.* 56, 425–429
96. Shirai, Y. *et al.* (2022) DIPA-CRISPR is a simple and accessible method for insect gene editing. *Cell Rep Methods* 2, 100215
97. De Rouck, S. *et al.* (2024) SYNCAS: Efficient CRISPR/Cas9 gene-editing in difficult to transform arthropods. *Insect Biochem. Mol. Biol.* 165, 104068
98. Sieriebriennikov, B. *et al.* (2021) A molecular toolkit for superorganisms. *Trends Genet.* 37, 846–859
99. Matthews, B.J. and Vosshall, L.B. (2020) How to turn an organism into a model organism in 10 “easy” steps. *J. Exp. Biol.* 223
100. Vicoso, B. (2019) Molecular and evolutionary dynamics of animal sex-chromosome turnover. *Nat Ecol Evol* 3, 1632–1641
101. Blackmon, H. and Demuth, J.P. (2015) Genomic origins of insect sex chromosomes. *Curr Opin Insect Sci* 7, 45–50
102. Sylvester, T. *et al.* (2020) Lineage-specific patterns of chromosome evolution are the rule not the exception in Polyneoptera insects. *Proc. Biol. Sci.* 287, 20201388
103. Ryazansky, S.S. *et al.* (2024) The chromosome-scale genome assembly for the West Nile vector Culex quinquefasciatus uncovers patterns of genome evolution in mosquitoes. *BMC Biol.* 22, 16
104. Jay, P. *et al.* (2024) Why do sex chromosomes progressively lose recombination? *Trends Genet.* 40, 564–579
105. Jeffries, D.L. *et al.* (2021) A neutral model for the loss of recombination on sex chromosomes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 376, 20200096
106. Jay, P. *et al.* (2022) Sheltering of deleterious mutations explains the stepwise extension of recombination suppression on sex chromosomes and other supergenes. *PLoS Biol.* 20, e3001698
107. Hotaling, S. *et al.* (2021) Long Reads Are Revolutionizing 20 Years of Insect Genome Sequencing. *Genome Biol. Evol.* 13, evab138
108. Li, H. and Durbin, R. (2024) Genome assembly in the telomere-to-telomere era. *Nat. Rev. Genet.*
109. Yamaguchi, K. *et al.* (2021) Technical considerations in Hi-C scaffolding and evaluation of chromosome-scale genome assemblies. *Mol. Ecol.* 30, 5923–5934
110. Mahajan, S. and Bachtrog, D. (2015) Partial dosage compensation in Strepsiptera, a sister group of beetles. *Genome Biol. Evol.* 7, 591–600
111. Parker, D.J. *et al.* (2022) X chromosomes show relaxed selection and complete somatic dosage compensation across Timema stick insect species. *J. Evol. Biol.* 35, 1734–1750
112. Li, X. *et al.* (2022) Grasshopper genome reveals long-term conservation of the X chromosome and temporal variation in X chromosome evolution*bioRxiv*, 2022.09.08.507201
113. Marin, R. *et al.* (2017) Convergent origination of a Drosophila-like dosage compensation mechanism in a reptile lineage. *Genome Res.* 27, 1974–1987
114. Pérez-Mojica, J.E. *et al.* (2023) Continuous transcriptome analysis reveals novel patterns of early gene expression in Drosophila embryos. *Cell Genom* 3, 100265
115. Meccariello, A. *et al.* (2019) Maleness-on-the-Y (MoY) orchestrates male sex determination in major agricultural fruit fly pests. *Science* 365, 1457–1460
116. Kumar, S. *et al.* (2017) TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* 34, 1812–1819
117. Yu, G. *et al.* (2017) Ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* 8, 28–36
118. Katsuma, S. *et al.* (2018) Unique sex determination system in the silkworm, Bombyx mori: current status and beyond. *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci.* 94, 205–216
119. Kelley, R.L. *et al.* (1997) Sex lethal controls dosage compensation in Drosophila by a

non-splicing mechanism. *Nature* 387, 195–199
120. Meller, V.H. and Rattner, B.P. (2002) The roX genes encode redundant male-specific lethal transcripts required for targeting of the MSL complex. *EMBO J.* 21, 1084–1091
121. Valsecchi, C.I.K. *et al.* (2020) RNA nucleation by MSL2 induces selective X chromosome compartmentalization. *Nature*
122. Schaeffer, S.W. (2018) Muller “Elements” in Drosophila: How the Search for the Genetic Basis for Speciation Led to the Birth of Comparative Genomics. *Genetics* 210, 3–13
123. Richard, G. *et al.* (2017) Dosage compensation and sex-specific epigenetic landscape of the X chromosome in the pea aphid. *Epigenetics Chromatin* 10, 30
124. Rosin, L.F. *et al.* (2022) Dosage compensation in *Bombyx mori* is achieved by partial repression of both Z chromosomes in males. *Proc. Natl. Acad. Sci. U. S. A.* 119, e2113374119
125. Sturgill, D. *et al.* (2007) Demasculinization of X chromosomes in the Drosophila genus. *Nature* 450, 238–241
126. Meiklejohn, C.D. *et al.* (2011) Sex chromosome-specific regulation in the Drosophila male germline but little evidence for chromosomal dosage compensation or meiotic inactivation. *PLoS Biol.* 9, e1001126
127. Lott, S.E. *et al.* (2014) Sex-specific embryonic gene expression in species with newly evolved sex chromosomes. *PLoS Genet.* 10, e1004159
128. Wei, K.H.-C. *et al.* (2022) Single cell RNA-seq in Drosophila testis reveals evolutionary trajectory of sex chromosome regulation*bioRxiv*, 2022.12.07.519494
129. Nozawa, M. *et al.* (2014) Tissue- and stage-dependent dosage compensation on the neo-X chromosome in Drosophila pseudoobscura. *Mol. Biol. Evol.* 31, 614–624
130. Briscoe, A.D. *et al.* (2013) Female behaviour drives expression and evolution of gustatory receptors in butterflies. *PLoS Genet.* 9, e1003620
131. Walters, J.R. *et al.* (2015) Sex Chromosome Dosage Compensation in Heliconius Butterflies: Global yet Still Incomplete? *Genome Biol. Evol.* 7, 2545–2559
132. Catalán, A. *et al.* (2018) Evolution of Sex-Biased Gene Expression and Dosage Compensation in the Eye and Brain of Heliconius Butterflies. *Mol. Biol. Evol.* 35, 2120–2134
133. Catalán, A. *et al.* (2024) Two novel genomes of fireflies with different degrees of sexual dimorphism reveal insights into sex-biased gene expression and dosage compensation. *Commun. Biol.* 7, 906
134. Smith, G. *et al.* (2014) Complete dosage compensation and sex-biased gene expression in the moth Manduca sexta. *Genome Biol. Evol.* 6, 526–537
135. Harrison, P.W. *et al.* (2012) Incomplete sex chromosome dosage compensation in the Indian meal moth, Plodia interpunctella, based on de novo transcriptome assembly. *Genome Biol. Evol.* 4, 1118–1126
136. Wilkinson, G.S. *et al.* (2013) Sex-biased gene expression during head development in a sexually dimorphic stalk-eyed fly. *PLoS One* 8, e59826
137. Djordjevic, J. *et al.* (2024) Dynamics of X chromosome hyper-expression and inactivation in male tissues during stick insect development*bioRxiv*, 2024.07.01.601468
138. Whittle, C.A. *et al.* (2020) Absence of a Faster-X Effect in Beetles (Tribolium, Coleoptera). *G3* 10, 1125–1136

Discussion

4.1. The definition of dosage compensation

In my literature search, I realized what authors mean when they state the sex-chromosome linked genes are "dosage compensated" differs a lot. When Muller coined the term "dosage compensation", his focus was on the ability of this mechanism to equalize the expression between the sexes, as this was what he could observe in *Drosophila* (Muller 1932). However, the term gained more recognition after Susumu Ohno's publication in 1967 (Ohno 1967). In this work, Ohno focused on the mammalian dosage compensation system. The second step proposed by Ohno (XCI; X chromosome inactivation) has been reproducibly demonstrated, the first step (upregulation of the single active X chromosome) has been more controversial. This leads to a linguistically confusing situation: there undeniably is a dosage compensation mechanism in mammals (XCI), but it is unclear whether the X is fully dosage compensated, instead of its expression just being balanced between the sexes.

The trouble with applying this term inconsistently is especially noticeable when comparing groups with different compensation modes, for example Diptera and Lepidoptera. In *Drosophila melanogaster* the X is fully dosage compensated, meaning the expression of X-linked genes matches the ancestral expression before the sex chromosomes differentiated, since the X is upregulated in the males (Vicoso & Bachtrog 2015). In most Lepidoptera, e.g. *Bombyx mori*, the genes on the two Z chromosomes are downregulated in the males, hence they are not dosage compensated *sensu stricto:* expression of Z chromosomal genes is usually half that of autosomes in both sexes (Rosin et al. 2022). Yet, some publications describe these expression patterns as full dosage compensation (Huylmans et al. 2017). The argument could be made that the low expression levels are a natural characteristic of the proto-Z. Since the Z chromosome in this lineage is old, it is challenging to assess the ancestral proto-Z expression levels. In order for the observed expression patterns to be called "complete dosage compensation", the expression of the progenitor of the Z chromosome would need to have been half that of other autosomes. Considering the expression distribution of each autosome separately, this is highly unlikely. Examples of such analyses are presented in the following figures: Supplementary Figure 22 in (Ranz et al. 2021), Supplementary Figure S1 in (Gu et al. 2019), and Figure S1 in (Walters et al. 2015). In these plots, chromosome Z usually has the lowest median expression compared to all the autosomes, making it an outlier. While it is possible for these lower expression levels to be a random occurrence or a natural

characteristic of the proto-Z chromosome, it is highly unlikely, as it deviates from expected distributions.

The inconsistent use of the term is a hindrance in comparing the dosage compensation patterns across evolutionary lineages. In the review article, I provided in Table 1 the underlying numbers which formed the basis of dosage compensation and dosage balance assessment in Figure 1 (Kalita & Keller Valsecchi 2024). This allows other researchers to make their own judgements based on the numbers directly, irrespective of the definition or the cutoff used.

4.2. DC mechanism in *A. gambiae* compared to other species

In our work, we discovered the master regulator of dosage compensation in *Anopheles gambiae*. Since SOA does not have any predicted catalytic domains, I hypothesize that it brings about dosage compensation by recruiting its interaction partners to the X chromosome. Identifying these interaction partners and their activity is likely the key to uncovering how exactly the X-linked genes get upregulated. While much work is still needed to fully characterize the mechanism, the data we have allows us to make some comparisons to other dosage compensation mechanisms.

One of the interesting features of the *Anopheles* dosage compensation mechanism is that it seems to be a synthesis of two modes: global and gene-by-gene. The global mode is represented best by the X chromosome inactivation in mammals, where *Xist* is expressed from the X inactivation center and later coats the entire X chromosome in a sequence-independent manner. In contrast, the mechanism in birds fits the gene-by-gene mode better, as the dosage-sensitive genes that became upregulated to achieve dosage compensation in the females (step 1 in Ohno's hypothesis) have a binding site for a male-specific microRNA (step 2) (Fallahshahroudi et al. 2024). In mosquitoes, the effects of the dosage compensation are global: the expression of the whole X chromosome is upregulated in the males. However, SOA does not coat the entire X chromosome; rather, it binds the promoters of many X-linked genes. How it recognizes the X-linked promoters specifically is still unclear. One explanation could be sequence specificity: SOA binding sites were enriched in a simple CA repeat motif. The CA motif is enriched on promoters of X linked genes as compared to autosomes. Of note, these simple repeats are reminiscent of the GAGA motif in the MSL2 binding sites (Kuzu et al. 2016). In both cases however, the levels of motif enrichment on the X are not sufficient to

fully explain how specifically these factors are targeted to the X chromosomes. Another example of how motif abundance can affect the whole chromosome in a gene-by-gene fashion was reported recently for the upregulation of mammalian X. The X-linked transcripts have lower abundance of the DRACH motif in comparison to autosomes. The DRACH motif is necessary for the epitranscriptomic modification m^6A (N^6-methyladenosine) to be deposited (Rücklé et al. 2023). As a consequence, X-derived transcripts are depleted in m^6A and have higher stability.

Dosage compensation is achieved by the upregulation of the single male X chromosome in multiple dipteran insects, even though their sex chromosomes have evolved independently and from different autosomal progenitors. Therefore, it seems that factors other than the gene content of the autosomal progenitor predict the mode of dosage compensation that evolves in a given lineage. One such factor could be the number of chromosomes in this group. Dipterans typically have 6 chromosomal arms (Morelli et al. 2022). Hence, when one of these becomes the sex chromosome, a significant portion of the insect's genes becomes sex-linked. This contrasts with the high number of chromosomes in the ancestral karyotype of Lepidoptera: 31 (Wright et al. 2024). In this insect group, the sex chromosome carries a much lower fraction of all genes. Thereby, the physiological consequences of sex chromosome degeneration might not be as severe as in dipterans and they can be resolved by equalizing the expression between the sexes, even if the expression does not match the ancestral levels.

An alternative explanation for the recurrent evolution of DC by hyperactivation in dipterans relies on the fact that the vast majority of species in this group are male-heterogametic. Evolutionary forces exert their influence differently in male- versus female-heterogametic species. To test this hypothesis, it would be worth understanding the dosage compensation mode in one of the rare female-heterogametic dipterans from the Tephritidae family, e.g. *Tephritis californica* (Vicoso & Bachtrog 2015).

In summary, despite different molecules being responsible in *Anopheles*, there are certain similarities to other dosage compensation mechanisms.

4.3. The non-essentiality of DC in Anopheles

One of the surprising discoveries we made in Publication 1 was that loss of SOA and therefore dosage compensation is not lethal for mosquito males. This was unexpected for a couple reasons. First, loss of dosage compensation has been shown to be lethal in all the

(three) species where the mechanism was known and the knockout of the master regulator of DC was tested for viability: mice, *C. elegans, D. melanogaster.*

In mice, *Xist* is the long non-coding RNA that initiates X chromosome inactivation (Penny et al. 1996). *Xist* lacking female mice embryos die in embryogenesis (Marahrens et al. 1997). However, this is caused by the failure in the development of the placenta, not the embryo itself . When *Xist* is depleted specifically in the embryo proper, the female embryos develop to birth but do not survive to adulthood (Yang et al. 2016). In *C. elegans*, dosage balance is achieved by dampening the expression of two X chromosomes in the hermaphrodites. Disruption of this process by mutating SDC-2, the master regulator that brings the dosage compensation complex to the X (Chu et al. 2002), results in hermaphrodite lethality (Nusbaum & Meyer 1989). Finally, *D. melanogaster* males lacking MSL2 die at the early pupae stage (Belote & Lucchesi 1980) . Hence, there was a widespread assumption that dosage compensation (or at least balance between the sexes) is essential for viability. Based on research in these species alone, it was not however possible to understand the exact causes of the lethality in DC factor mutants. The existence of species that lack chromosome-wide dosage compensation (Basilicata & Keller Valsecchi 2021) stands in contradiction to its indispensable role in the aforementioned model species. It is possible that the essentiality observed in all earliest tested species is partially due to chance, partially due to selection bias. In *C. elegans* and *D. melanogaster* the way the DC factors were discovered involved high-throughput genetic screens that used sex-specific lethality as the scored phenotype.

The second reason why the viability of mosquitoes with aberrant dosage compensation was unexpected was that it stood in a seeming contradiction to the previously observed lethality of mosquitoes with disturbed expression of their upstream sex determination factors. It was proposed that the lethality of *Yob*-expressing females is due to the aberrant upregulation of the X chromosome (Krzywinska et al. 2016; Krzywinska & Krzywinski 2018). However, the expression of X linked genes in these publications was not actually measured. In females depleted for *femaleless*, both X chromosome upregulation and female-specific lethality were observed (Krzywinska et al. 2021). However, the expression of full SOA isoform in the females did not reproduce the same phenotype. Yob and *femaleless* likely have other targets apart from SOA. In order to determine if misregulation of those targets is responsible for lethality, one would have to assess the viability in Yob-expressing females with a knock out of SOA. The most likely explanation for the lethality phenotype is that results from the combined effects of aberrant DC and the incompatibility of the sex determination cascade with the chromosomal sex.

Another common feature of *Drosophila* and *Anopheles* is that sex-specific splicing plays a role in ensuring male-specific activity of their respective DC master regulators. There is however an interesting difference: in *Drosophila* females, the MSL2 protein is not produced, while in mosquitoes the short isoform of SOA was detected by mass spectrometry in female samples. It is therefore possible that SOA has a role to play in females as well. One idea is that the short isoform has a dominant negative effect - i.e. it can bind SOA target sites and outcompete the full length isoform. This would be an interesting safe-fail mechanism in case any mRNA particles get incorrectly spliced in the females, thereby allowing a for a few full-length SOA protein molecules to be produced. The short isoform in this scenario could prevent any aberrant activation of genes on the two female X chromosomes.

4.4. Methodological progress allows for studying DC outside model organisms

The dosage compensation mechanisms in model organisms were discovered before any animal genomes were sequenced. However, with the wide availability and affordability of genomics it became possible to study DC in a much wider range of species.

The first insect genome was sequenced in 2000. Significant acceleration in genome sequencing occurred in the past five years. Apart from the increased availability and decreased price of whole genome sequencing, crucial advances were made recently in genome assembly (Li & Durbin 2024). This is an important step in determining the chromosomal location of genes and assigning them to the sex chromosomes or autosomes. The first advance was the availability of long-read sequencing, which significantly increased the length of scaffolds (Hotaling et al. 2021). The next improvement that led to high-quality genome assemblies was applying Hi-C for scaffolding (Yamaguchi et al. 2021). Initiatives like i5k and The Darwin Tree of Life have played pivotal roles in increasing the numbers of available insect genomes (i5K Consortium 2013; The Darwin Tree of Life Project Consortium et al. 2022). The dosage compensation status has now been assessed with RNA-seq in more than 50 insect species.

Like the genomic revolution, the advent of CRISPR/Cas9 has also made it easier to study non-model organisms. The validation of potential dosage compensation master regulators should be done *in vivo* and, ideally, with a stable expression or knockout to understand the essentiality of dosage compensation in a given species. CRISPR has been successfully applied to many insects. The initial method required embryo injections that can

be technically challenging or incompatible with embryo development for some species. Because of this, a method involving direct parental delivery, DIPA-CRISPR, has been developed and has already been successfully applied to insects such as *Tribolium castaneum, Blatella germanica* (Shirai et al. 2022), *Aedes aegypti* (Shirai et al. 2023), *Sogatella furcifera* (Zhang et al. 2023), and even outside of arthropods, in the tardigrade (Kondo et al. 2024). This approach involves injecting the Cas9 protein alongside the sgRNA into the mother and relies on the uptake of the ribonucleoprotein particles by the oocytes. SYNCAS CRISPR improves upon the gene-editing efficiency of DIPA-CRISPR by injecting Cas9 together with Branched Amphiphilic Peptide Capsules (BAPC) and saponins to increase cell uptake and endosomal escape, respectively (De Rouck et al. 2024). These advancements in Cas9 delivery will hopefully expand the applicability of CRISPR to an even wider array of insect species.

The methodological progress was also crucial in our discovery of SOA as the master regulator of dosage compensation. We took advantage of the improvements in library preparation methods that allowed us to sequence low-input samples (single mosquito embryos). This was key to our discovery of SOA, which we could see differentially expressed and spliced already in embryos 5 hours after fertilization. Only *SOA* and Yob were differentially expressed at this time point, while in pupae and adult stages there are thousands of genes differentially expressed between males and females. We also benefited from the recently developed chromatin profiling method CUT&Tag (Cleavage Under Targets and Tagmentation) (Kaya-Okur et al. 2019). Biochemical properties of SOA protein likely make it prone to aggregate in standard chromatin precipitation experiments, while CUT&Tag avoids these issues by performing chromatin profiling inside the cell, not in a lysate. Of note, CUT&Tag requires less input material than standard chromatin precipitation (ChIP), which is especially important for species that cannot be reared in the laboratory setting. Last but not least, our *in vivo* validation would not be feasible without efficient CRISPR/Cas9 genome editing.

Taken together, the methodological advances in sequencing, chromatin profiling and genome editing have been crucial for our discovery. I therefore think that the wide availability and applicability of these tools to a broader range of insect species will bring about an increase in new discoveries into dosage compensation and its mechanisms in diverse insect groups.

4.5. Conclusions and outlook

The discovery of SOA as the master regulator of dosage compensation in *Anopheles gambiae* opens up new avenues of research. It forms a foundation that will hopefully lead to full characterization of the mosquito DC mechanism in the future. As SOA itself is not predicted to have any domains with enzymatic activity that could affect gene expression (e.g. by modifying histones), identifying and validating its interaction partners is the next step to elucidate the molecular details of the mechanism. The specific aspects that still need to be addressed are the targeting mechanism, how increased transcriptional output is achieved and if there are histone modifications involved.

As *SOA* and its sex-specific splicing are conserved across the *Anopheles* genus, our results could inform the DC mechanism in related malaria-transmitting mosquitoes. For example in *Anopheles stephensi*, which separated from *A. gambiae* a relatively short time ago and is becoming an increasing burden in Africa's urban populations (Sinka et al. 2020).

In summary, my PhD work focused on dosage compensation in non-model insect species. In my first publication, together with my collaborators, we discovered a master regulator of dosage compensation in *Anopheles gambiae.* The review article has summarized the state of knowledge on sex chromosome dosage compensation across all insects. Additionally, the review introduced a framework to discover new dosage compensation mechanisms in other insect species. Taken together, my PhD work has been an important contribution to the field of sex chromosome dosage compensation in insects and I hope it will help other researchers discover new mechanisms.

References

Akhtar, A. & Becker, P.B., 2000. Activation of transcription through histone H4 acetylation by MOF, an acetyltransferase essential for dosage compensation in Drosophila. *Molecular cell*, 5(2), pp.367–375.

Basilicata, M.F. & Keller Valsecchi, C.I., 2021. The good, the bad, and the ugly: Evolutionary and pathological aspects of gene dosage alterations. *PLoS genetics*, 17(12), p.e1009906.

Belote, J.M. & Lucchesi, J.C., 1980. Male-specific lethal mutations of Drosophila melanogaster. *Genetics*, 96(1), pp.165–186.

Cecalev, D., Viçoso, B. & Galupa, R., 2024. Compensation of gene dosage on the mammalian X. *Development* , 151(15). Available at: http://dx.doi.org/10.1242/dev.202891.

Charlesworth, B., 1991. The evolution of sex chromosomes. *Science*, 251(4997), pp.1030–1033.

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

Chu, D.S. et al., 2002. A molecular link between gene-specific and chromosome-wide transcriptional repression. *Genes & development*, 16(7), pp.796–805.

Cline, T.W., 1978. Two closely linked mutations in Drosophila melanogaster that are lethal to opposite sexes and interact with daughterless. *Genetics*, 90(4), pp.683–698.

Cox, F.E., 2010. History of the discovery of the malaria parasites and their vectors. *Parasites & vectors*, 3(1), p.5.

Criscione, F., Qi, Y. & Tu, Z., 2016. GUY1 confers complete female lethality and is a strong candidate for a male-determining factor in Anopheles stephensi. *eLife*, 5. Available at: http://dx.doi.org/10.7554/eLife.19281.

Deitz, K.C., Takken, W. & Slotman, M.A., 2018. The Effect of Hybridization on Dosage Compensation in Member Species of the Anopheles gambiae Species Complex D. Sloan, ed. *Genome biology and evolution*, 10(7), pp.1663–1672.

De Rouck, S. et al., 2024. SYNCAS: Efficient CRISPR/Cas9 gene-editing in difficult to transform arthropods. *Insect biochemistry and molecular biology*, 165, p.104068.

Fallahshahroudi, A. et al., 2024. A male-essential microRNA is key for avian sex chromosome dosage compensation. Available at: https://doi.org/10.1101/2024.03.06.581755.

Franke, A. & Baker, B.S., 1999. The rox1 and rox2 RNAs are essential components of the compensasome, which mediates dosage compensation in Drosophila. *Molecular cell*, 4(1), pp.117–122.

Goodenough, U. & Heitman, J., 2014. Origins of eukaryotic sexual reproduction. *Cold Spring Harbor perspectives in biology*, 6(3). Available at: http://dx.doi.org/10.1101/cshperspect.a016154.

Gu, L. et al., 2019. Dichotomy of Dosage Compensation along the Neo Z Chromosome of the Monarch Butterfly. *Current biology: CB*, 29(23), pp.4071–4077.e3.

Gu, L. & Walters, J.R., 2017. Evolution of Sex Chromosome Dosage Compensation in Animals: A Beautiful Theory, Undermined by Facts and Bedeviled by Details. *Genome biology and evolution*, 9(9), pp.2461–2476.

Hassold, T. & Hunt, P., 2001. To err (meiotically) is human: the genesis of human aneuploidy. *Nature reviews. Genetics*, 2(4), pp.280–291.

Hilfiker, A. et al., 1997. mof, a putative acetyl transferase gene related to the Tip60 and MOZ human genes and to the SAS genes of yeast, is required for dosage compensation in Drosophila. *The EMBO journal*, 16(8), pp.2054–2060.

Hotaling, S. et al., 2021. Long Reads Are Revolutionizing 20 Years of Insect Genome Sequencing. *Genome biology and evolution*, 13(8), p.evab138.

Huylmans, A.K., Macon, A. & Vicoso, B., 2017. Global dosage compensation is ubiquitous in Lepidoptera, but counteracted by the masculinization of the Z chromosome. *Molecular biology and evolution*, 34(10), pp.2637–2649.

i5K Consortium, 2013. The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *The Journal of heredity*, 104(5), pp.595–600.

Ilik, I.A. et al., 2013. Tandem stem-loops in roX RNAs act together to mediate X chromosome dosage compensation in Drosophila. *Molecular cell*, 51(2), pp.156–173.

Jiang, X. et al., 2015. Complete dosage compensation in Anopheles stephensi and the evolution of sex-biased genes in mosquitoes. *Genome biology and evolution*, 7(7), pp.1914–1924.

Kadlec, J. et al., 2011. Structural basis for MOF and MSL3 recruitment into the dosage compensation complex by MSL1. *Nature structural & molecular biology*, 18(2), pp.142–149.

Kalita, A.I. & Keller Valsecchi, C.I., 2024. Dosage compensation in non-model insects – progress and perspectives. *Trends in genetics: TIG*. Available at: http://dx.doi.org/10.1016/j.tig.2024.08.010 [Accessed September 28, 2024].

Kaya-Okur, H.S. et al., 2019. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature communications*, 10(1), p.1930.

Keller Valsecchi, C.I. et al., 2021. Distinct mechanisms mediate X chromosome dosage compensation in Anopheles and Drosophila. *Life science alliance*, 4(9). Available at: http://dx.doi.org/10.26508/lsa.202000996.

Kondo, K., Tanaka, A. & Kunieda, T., 2024. Single-step generation of homozygous knock-out/knock-in individuals in an extremotolerant parthenogenetic tardigrade using DIPA-CRISPR. *bioRxiv*, p.2024.01.10.575120. Available at: https://www.biorxiv.org/content/10.1101/2024.01.10.575120v1 [Accessed April 2, 2024].

Krzywinska, E. et al., 2016. A maleness gene in the malaria mosquito Anopheles gambiae. *Science*, 353(6294), pp.67–69.

Krzywinska, E. et al., 2021. femaleless Controls Sex Determination and Dosage Compensation Pathways in Females of Anopheles Mosquitoes. *Current biology: CB*, 31(5), pp.1084–1091.e4.

Krzywinska, E. & Krzywinski, J., 2018. Effects of stable ectopic expression of the primary sex determination gene Yob in the mosquito Anopheles gambiae. *Parasites & vectors*, 11(Suppl 2), p.648.

Kuzu, G. et al., 2016. Expansion of GA dinucleotide repeats increases the density of CLAMP binding sites on the X-chromosome to promote Drosophila dosage compensation. *PLoS genetics*, 12(7), p.e1006120.

Li, H. & Durbin, R., 2024. Genome assembly in the telomere-to-telomere era. *Nature reviews. Genetics*, pp.1–13.

Lyon, M.F., 1961. Gene action in the X-chromosome of the mouse (Mus musculus L.). *Nature*, 190(4773), pp.372–373.

Maenner, S. et al., 2013. ATP-dependent roX RNA remodeling by the helicase maleless enables specific association of MSL proteins. *Molecular cell*, 51(2), pp.174–184.

Mank, J.E., 2013. Sex chromosome dosage compensation: definitely not for everyone. *Trends in genetics: TIG*, 29(12), pp.677–683.

Marahrens, Y. et al., 1997. Xist-deficient mice are defective in dosage compensation but not spermatogenesis. *Genes & development*, 11(2), pp.156–166.

Morales, V. et al., 2004. Functional integration of the histone acetyltransferase MOF into the dosage compensation complex. *The EMBO journal*, 23(11), pp.2258–2268.

Morelli, M.W., Blackmon, H. & Hjelmen, C.E., 2022. Diptera and Drosophila karyotype databases: A useful dataset to guide evolutionary and genomic studies. *Frontiers in ecology and evolution*, 10. Available at: https://doi.org/10.3389/fevo.2022.832378.

Muller, H.J., 1932. Further studies on the Nature and causes of gene mutations. *Proc 6th Int Congr Genet*, 1: 213–255. Available at: http://www.esp.org/books/6th-congress/facsimile/contents/6th-cong-p213-muller.pdf.

Neafsey, D.E. et al., 2015. Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. *Science*, 347(6217), p.1258522.

Nusbaum, C. & Meyer, B.J., 1989. The Caenorhabditis elegans gene sdc-2 controls sex determination and dosage compensation in XX animals. *Genetics*, 122(3), pp.579–593.

Ohno, S., 1967. *Sex chromosomes and sex-linked genes* 1966th ed., Berlin, Germany: Springer.

Penny, G.D. et al., 1996. Requirement for Xist in X chromosome inactivation. *Nature*, 379(6561), pp.131–137.

Qi, Y. et al., 2019. Guy1, a Y-linked embryonic signal, regulates dosage compensation in Anopheles stephensi by increasing X gene expression. *eLife*, 8. Available at: http://dx.doi.org/10.7554/eLife.43570.

Ranz, J.M. et al., 2021. A de novo transcriptional atlas in Danaus plexippus reveals variability in dosage compensation across tissues. *Communications biology*, 4(1), p.791.

Rose, G. et al., 2016. Dosage compensation in the African malaria mosquito Anopheles gambiae. *Genome biology and evolution*, 8(2), pp.411–425.

Rosin, L.F. et al., 2022. Dosage compensation in Bombyx mori is achieved by partial repression of both Z chromosomes in males. *Proceedings of the National Academy of Sciences of the United States of America*, 119(10), p.e2113374119.

Rücklé, C. et al., 2023. RNA stability controlled by m6A methylation contributes to X-to-autosome dosage compensation in mammals. *Nature structural & molecular biology*, 30(8), pp.1207–1215.

Schütt, C. & Nöthiger, R., 2000. Structure, function and evolution of sex-determining systems in Dipteran insects. *Development* , 127(4), pp.667–677.

Shirai, Y. et al., 2023. DIPA-CRISPR gene editing in the yellow fever mosquito Aedes aegypti (Diptera: Culicidae). *Applied entomology and zoology*, 58(3), pp.273–278.

Shirai, Y. et al., 2022. DIPA-CRISPR is a simple and accessible method for insect gene editing. *Cell*

reports methods, 2(5), p.100215.

Sinka, M.E. et al., 2020. A new malaria vector in Africa: Predicting the expansion range of Anopheles stephensi and identifying the urban populations at risk. *Proceedings of the National Academy of Sciences of the United States of America*, 117(40), pp.24900–24908.

The Darwin Tree of Life Project Consortium et al., 2022. Sequence locally, think globally: The Darwin Tree of Life Project. *Proceedings of the National Academy of Sciences*, 119(4), p.e2115642118.

Tree of Sex Consortium, 2014. Tree of Sex: a database of sexual systems. *Scientific data*, 1, p.140015.

Valsecchi, C.I.K. et al., 2020. RNA nucleation by MSL2 induces selective X chromosome compartmentalization. *Nature*. Available at: https://www.ncbi.nlm.nih.gov/pubmed/33208948.

Vicoso, B. & Bachtrog, D., 2015. Numerous transitions of sex chromosomes in Diptera. *PLoS biology*, 13(4), p.e1002078.

Walters, J.R., Hardcastle, T.J. & Jiggins, C.D., 2015. Sex Chromosome Dosage Compensation in Heliconius Butterflies: Global yet Still Incomplete? *Genome biology and evolution*, 7(9), pp.2545–2559.

World Health Organization, 2023. *World malaria report 2023*, World Health Organization. Available at: https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2023 [Accessed September 30, 2024].

Wright, C.J. et al., 2024. Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera. *Nature ecology & evolution*. Available at: http://dx.doi.org/10.1038/s41559-024-02329-4.

Yamaguchi, K. et al., 2021. Technical considerations in Hi-C scaffolding and evaluation of chromosome-scale genome assemblies. *Molecular ecology*, 30(23), pp.5923–5934.

Yang, L. et al., 2016. Female mice lacking Xist RNA show partial dosage compensation and survive to term. *Genes & development*, 30(15), pp.1747–1760.

Zhang, M.-Q. et al., 2023. Efficient DIPA-CRISPR-mediated knockout of an eye pigment gene in the white-backed planthopper, Sogatella furcifera. *Insect science*. Available at: http://dx.doi.org/10.1111/1744-7917.13286.